Review in Advance first posted online on April 5, 2012. (Changes may still occur before final publication online and in print.)

Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis

Surya Ganguli[1] and Haim Sompolinsky[2,3]

[1]Department of Applied Physics, Stanford University, Stanford, California 94305; email: sganguli@stanford.edu

[2]Edmond and Lily Safra Center for Brain Sciences, Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel; email: haim@fiz.huji.ac.il

[3]Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138

Annu. Rev. Neurosci. 2012. 35:485–508

The *Annual Review of Neuroscience* is online at neuro.annualreviews.org

This article's doi: 10.1146/annurev-neuro-062111–150410

0147-006X/12/0721-0485$20.00

Keywords

random projections, connectomics, imaging, memory, communication, learning, generalization

Abstract

The curse of dimensionality poses severe challenges to both technical and conceptual progress in neuroscience. In particular, it plagues our ability to acquire, process, and model high-dimensional data sets. Moreover, neural systems must cope with the challenge of processing data in high dimensions to learn and operate successfully within a complex world. We review recent mathematical advances that provide ways to combat dimensionality in specific situations. These advances shed light on two dual questions in neuroscience. First, how can we as neuroscientists rapidly acquire high-dimensional data from the brain and subsequently extract meaningful models from limited amounts of these data? And second, how do brains themselves process information in their intrinsically high-dimensional patterns of neural activity as well as learn meaningful, generalizable models of the external world from limited experience?

Contents

INTRODUCTION 486
ADVANCES IN THE THEORY OF HIGH-DIMENSIONAL STATISTICS 488
The Compressed Sensing Framework: Incoherence and Randomness 488
L_1 Minimization: A Nonlinear Recovery Algorithm 490
Dimensionality Reduction by Random Projections 491
Compressed Computation 492
Approximate Sparsity and Noise .. 493
Sparse Models of High-Dimensional Data 494
Dictionary Learning 495
COMPRESSED SENSING OF THE BRAIN 495
Rapid Functional Imaging 495
Fluorescence Microscopy 496
Gene-Expression Analysis 496
Compressed Connectomics 497
COMPRESSED SENSING BY THE BRAIN 497
Semantic Similarity and Random Projections 498
Short-Term Memory in Neuronal Networks 498
SPARSE EXPANDED NEURONAL REPRESENTATIONS 499
Neuronal Implementations of L1 Minimization 500
Compression and Expansion in Long-Range Brain Communication 501
LEARNING IN HIGH-DIMENSIONAL SYNAPTIC WEIGHT SPACES 501
Neural Learning of Classification 502
Optimality and Sparsity of Synaptic Weights 502
DISCUSSION 503
Dimensionality Reduction: CS versus Efficient Coding 503
Expansion and Sparsification: Compressed Sensing versus Independent Components Analysis 503
Beyond Linear Projections: Neuronal Nonlinearities 504

INTRODUCTION

For most of its history, neuroscience has made wonderful progress by considering problems whose descriptions require only a small number of variables. For example, Hodgkin & Huxley (1952) discovered the mechanism of the nerve impulse by studying the relationship between two variables: the voltage and the current across the cell membrane. But as we have started to explore more complex problems, such as the brain's ability to process images and sounds, neuroscientists have had to analyze many variables at once. For example, any given gray-scale image requires N analog variables, or pixel intensities, for its description, where N could be on the order of 1 million. Similarly, such images could be represented in the firing-rate patterns of many neurons, with each neuron's firing rate being a single analog variable. The number of variables required to describe a space of objects is known as the dimensionality of that space; i.e., the dimensionality of the space of all possible images of a given size equals the number of pixels, whereas the dimensionality of the space of all possible neuronal firing-rate patterns in a given brain area equals the number of neurons in that area. Thus our quest to understand how networks of neurons store and process information depends crucially on our ability to measure and understand the relationships between high-dimensional spaces of stimuli and neuronal activity patterns.

However, the problem of measuring and finding statistical relationships between patterns becomes more difficult as their dimensionality increases. This phenomenon is known as the curse of dimensionality. One approach to addressing this problem is to somehow reduce the number of variables required to describe the patterns in question, a process known as dimensionality reduction. We can do this, for example, with natural images, which are a highly restricted subset of all possible images, so that they can be described by many fewer variables than the number of pixels. In particular, natural images are often sparse in the sense that if you view them in the wavelet domain (roughly as a superposition of edges), only a very small number of K wavelet coefficients will have significant power, where K can be on the order of 20,000 for a 1-million-pixel image. This observation underlies JPEG compression, which computes all possible wavelet coefficients and keeps only the K largest (Taubman et al. 2002). Similarly, neuronal activity patterns that actually occur are often a highly restricted subset of all possible patterns (Ganguli et al. 2008a, Yu et al. 2009, Machens et al. 2010) in the sense that they often lie along a low K-dimensional manifold embedded in N-dimensional firing-rate space; by this we mean that only K numbers are required to uniquely specify any observed activity pattern across N neurons, where K can be much smaller than N. As a concrete example, consider the set of visual activity patterns in N neurons in response to a bar presented at a variety of orientations. As the orientation varies, the elicited firing-rate responses trace out a circle, or a one-dimensional manifold in N-dimensional space.

More generally, given a class of apparently high-dimensional stimuli, or neuronal activity patterns, how can either we or neural systems extract a small number of variables to describe these patterns without losing too much important information? Machine learning provides a variety of algorithms to perform this dimensionality reduction, but they are often computationally expensive in terms of running time. Moreover, how neuronal circuits could implement many of these algorithms is not clear. However, recent advances in an emerging field of high-dimensional statistics (Donoho 2000, Baraniuk 2011) have revealed a surprisingly simple yet powerful method of performing dimensionality reduction: One can randomly project patterns into a lower-dimensional space. To understand the central concept of a random projection (RP), it is useful to think of the shadow of a wire-frame object in three-dimensional space projected onto a two-dimensional screen by shining a light beam on the object. For poorly chosen angles of light, the shadow may lose important information about the wire-frame object. For example, if the axis of light is aligned with any segment of wire, that entire length of wire will have a single point as its shadow. However, if the axis of light is chosen randomly, it is highly unlikely that the same degenerate situation will occur; instead, every length of wire will have a corresponding nonzero length of shadow. Thus the shadow, obtained by this RP, generically retains much information about the wire-frame object.

In the context of image acquisition, an RP of an image down to an M-dimensional space can be obtained by taking M measurements of the image, where each measurement consists of a weighted sum of all the pixel intensities, and allowing the weights themselves to be chosen randomly (for example, drawn independently from a Gaussian distribution). Thus the original image (i.e., the wire-frame structure) is described by M measurements (i.e., its shadow) by projecting against a random set of weights (i.e., a random light angle). Now, the field of compressed sensing (CS) (Candes et al. 2006, Candes & Tao 2006, Donoho 2006; see Baraniuk 2007, Candes & Wakin 2008, Bruckstein et al. 2009 for reviews) shows that the shadow can contain enough information to reconstruct the original image (i.e., all N pixel values) as long as the original image is sparse enough. In particular, if the space of the images in question can be described by K variables, then as long as M is slightly larger than K, CS

provides an algorithm (called L_1 minimization, described below) to reconstruct the image. Thus for typical images, we can simultaneously sense and compress 1-million-pixel images with ~20,000 random measurements. As we review below, these CS results have significant implications for data acquisition in neuroscience.

Furthermore, in the context of neuronal information processing, an RP of neuronal activity in an upstream brain region consisting of N neurons can be achieved by synaptic mapping to a downstream region consisting of $M < N$ neurons, where the downstream neurons' firing rates are obtained by linearly summing the firing rates of the upstream neurons through a set of random synaptic weights. Thus the downstream activity constitutes a shadow of the upstream activity through an RP determined by the synaptic weights (i.e., angle of light). As we review below, the theory of CS and RPs can provide a theoretical framework for understanding one of the most salient aspects of neuronal information processing: radical changes in the dimensionality, and sometimes sparsity, of neuronal representations, often within a single stage of synaptic transformation.

Finally, another application of CS is the problem of modeling high-dimensional data. This is challenging because such models have high-dimensional parameter spaces, necessitating many example data points to learn the correct parameter values. Neural systems face a similar challenge in searching high-dimensional synaptic weight spaces to learn generalizable rules from limited experience. We review how regularization techniques (Tibshirani 1996, Efron et al. 2004) closely related to CS allow statisticians and neural systems alike to rapidly learn sparse models of high-dimensional data from limited examples.

ADVANCES IN THE THEORY OF HIGH-DIMENSIONAL STATISTICS

Before we describe the applicability of CS and RPs to the acquisition and analysis of data and to neuronal information processing and learning, we first give in this section a more precise overview of recent results in high-dimensional statistics. We begin by giving an overview of the CS framework and define the mathematical notation we use throughout this review. Subsequently, a reader who is interested mainly in applications can skip the rest of this section. Here, we discuss how to recover the sparse signals from small numbers of measurements, even in the presence of approximate sparsity and noise, and we discuss RPs and sparse regression in more detail. Finally, we discuss dictionary learning, an approach to find bases in which ensembles of signals are sparse.

The Compressed Sensing Framework: Incoherence and Randomness

We now formalize the intuitions given in the introduction and describe the mathematical notation that we use throughout this review (see also **Figure 1**). We let $\mathbf{u}^0$ be an N-dimensional signal that we wish to measure. Thus $\mathbf{u}^0$ is a vector with components $\mathbf{u}_i^0$ for $i = 1, \ldots, N$, where each $\mathbf{u}_i^0$ can take an analog value. In the example of an image, $\mathbf{u}_i^0$ would be the gray-scale intensity of the ith pixel. The M linear measurements of $\mathbf{u}^0$ are of the form $x_\mu = \mathbf{b}^\mu \cdot \mathbf{u}^0$ for $\mu = 1, \ldots, M$. Here we think of x_μ as an analog outcome of measurement μ obtained by computing the overlap or dot product between the unknown signal $\mathbf{u}^0$ and a measurement vector $\mathbf{b}^\mu$. We can summarize the relationship between the signal and the measurements via the matrix relationship $\mathbf{x} = \mathbf{B}\mathbf{u}^0$. Here $\mathbf{B}$ is an $M \times N$ measurement matrix, whose μth row is the vector $\mathbf{b}^\mu$, and $\mathbf{x}$ is a measurement vector whose μ'th component is x_μ. Now the true signal $\mathbf{u}^0$ is sparse in a basis given by the columns of an $N \times N$ matrix $\mathbf{C}$. By this we mean that $\mathbf{u}^0 = \mathbf{C}\mathbf{s}^0$, where $\mathbf{s}^0$ is a sparse N-dimensional vector, in the sense that it has a relatively small number K of nonzero elements, though we do not know ahead of time which K of the N components are nonzero. For example, when $\mathbf{u}^0$ is an image in the pixel basis, $\mathbf{s}^0$ could be the wavelet coefficients of that same image, and the columns of $\mathbf{C}$ would comprise a complete basis

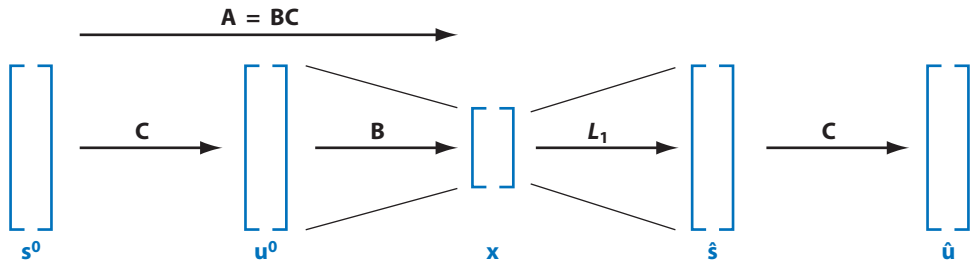

Figure 1

Framework of compressed sensing (CS). A high-dimensional signal $\mathbf{u}^0$ is sparse in a basis given by the columns of a matrix **C** so that $\mathbf{u}^0 = \mathbf{C}\mathbf{s}^0$, where $\mathbf{s}^0$ is a sparse coefficient vector. Through a set of measurements given by the rows of B, $\mathbf{u}^0$ is compressed to a low-dimensional space of measurements **x**. If the measurements are incoherent with respect to the sparsity basis, then L_1 minimization can recover a good estimate $\hat{\mathbf{s}}$ of the sparse coefficients $\mathbf{s}^0$ from **x**, and then an estimate of $\mathbf{u}^0$ can be recovered by expanding in the basis C.

of orthonormal wavelets. Finally, the overall relationship between the measurements and the sparse coefficients is given by $\mathbf{x} = \mathbf{A}\mathbf{s}^0$, where $\mathbf{A} = \mathbf{B}\mathbf{C}$. We often refer to **A** also as the measurement matrix.

An important question is, given a sparsity basis **C**, what should we choose as our measurement basis **B**? Consider what might happen if we measured signals in the same basis in which they were sparse. For example, in the case of an image, one could directly measure M randomly chosen wavelet coefficients of the image in which M is just a little larger than K. The problem, of course, is that for any given image, it is highly unlikely that all the K coefficients with large power coincide with the M coefficients we chose to measure. So unless the number of measurements M equals the dimensionality of the image, N, we will inevitably miss important coefficients. In the wire-frame shadow example above, this is the analog of choosing a poor angle of light (i.e., measurement basis) that aligns with a segment of wire (i.e., sparsity basis), which causes information loss.

To circumvent this problem, one of the key ideas of CS is that we should make our measurements as different as possible from the domain in which the signal is sparse (i.e., shine light at an angle that does not align with any segment of wire frame). In particular, the measurements should have many nonzero elements in the domain in which the image is sparse. This notion of difference is captured by the mathematical definition of incoherence, or a small value of the maximal inner product between rows of **B** and columns of **C**, so that no measurement vector should look like any sparsity vector. CS provides mathematical guarantees that one can achieve perfect recovery with a number of measurements M that is only slightly larger than K, as long as the M measurement vectors are sufficiently incoherent with respect to the sparsity domain (Candes & Romberg 2007).

An important observation is that any set of measurement vectors, which are themselves random, will be incoherent with respect to any fixed sparsity domain. For example, the elements of each such measurement vector can be drawn independently from a Gaussian distribution. Intuitively, it is highly unlikely for a random vector to look like a sparsity vector (i.e., just as it is unlikely for a random light angle to align with a wire segment). One of the key results of CS is that with such random measurement vectors, only

$$M > O(K \log(N/K)) \qquad 1.$$

measurements are needed to guarantee perfect signal reconstruction with high probability (Candes & Tao 2005, Baraniuk et al. 2008, Candes & Plan 2010). Thus random measurements constitute a universal measurement

strategy in the sense that they will work for signals that are sparse in any basis. Indeed, the sparsity basis need not even be known yet when the measurements are chosen. Its knowledge is required only after measurements are taken, during the nonlinear reconstruction process. And remarkably, investigators have further shown that no measurement matrices and no reconstruction algorithm can yield sparse signal recovery with substantially fewer measurements (Candes & Tao 2006, Donoho 2006), than that shown in Equation 1.

L_1 Minimization: A Nonlinear Recovery Algorithm

Given only our measurements $\mathbf{x}$, how can we recover the unknown signal $\mathbf{u}^0$? One could potentially do this by inverting the relationship between measurements and signal by solving for an unknown candidate signal $\mathbf{u}$ in the equation $\mathbf{x} = \mathbf{Bu}$. This is a set of M equations, one for each measurement, with N unknowns, one for each component of the candidate signal $\mathbf{u}$. If the number of independent measurements M is greater than or equal to the dimensionality N of the signal, then the set of equations $\mathbf{x} = \mathbf{Bu}$ has a unique solution $\mathbf{u} = \mathbf{u}^0$; thus, solving these equations will recover the true signal $\mathbf{u}^0$. However, if $M < N$, the set of equations $\mathbf{x} = \mathbf{Bu}$ no longer has a unique solution. Indeed there is generically an $N - M$ dimensional space of candidate signals $\mathbf{u}$ that satisfy the measurement constraints. How might we find the true signal $\mathbf{u}^0$ in this large space of candidate signals?

If we know nothing further about the true signal $\mathbf{u}^0$, then the situation is indeed hopeless. However, if $\mathbf{u}^0 = \mathbf{Cs}^0$ where $\mathbf{s}^0$ is sparse, we can try to exploit this prior knowledge as follows (see **Figure 1**). First, the measurements are linearly related to the sparse coefficients $\mathbf{s}^0$ through the M equations $\mathbf{x} = \mathbf{As}^0$, where $\mathbf{A} = \mathbf{BC}$ is an $M \times N$ matrix. Again, when $M < N$, there is a large $N - M$ dimensional space of solutions $\mathbf{s}$ to the measurement constraint $\mathbf{x} = \mathbf{As}$. However, not all of them will be sparse, as we expect the true solution $\mathbf{s}^0$ to be. Thus one might try to construct an estimate $\hat{\mathbf{s}}$ of $\mathbf{s}^0$ by solving the optimization problem

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s}} \sum_{i=1}^{N} V(s_i) \quad \text{subject to } \mathbf{x} = \mathbf{As}, \quad 2.$$

where $V(s)$ is any cost function that penalizes nonzero values of s. A natural choice, is $V(s) = 0$ if $s = 0$ and $V(s) = 1$ otherwise. With this choice, Equation 2 says that our estimate $\hat{\mathbf{s}}$ is obtained by searching, in the space of all candidate signals $\mathbf{s}$ that satisfy the measurement constraints $\mathbf{x} = \mathbf{As}$, for the one that has the smallest number of nonzero elements. This approach, while reasonable given the prior knowledge that the true signal $\mathbf{s}^0$ has a small number of nonzero coefficients, unfortunately yields a computationally intractable combinatorial optimization problem; to solve it, one must essentially search over all subsets of possible nonzero elements in $\mathbf{s}$.

An alternative approach, adopted by CS, is to solve a related and potentially easier problem, by choosing $V(s) = |s|$. The quantity $\sum_{i=1}^{N} |s_i|$ is known as L_1 norm of $\mathbf{s}$; hence, this method is called L_1 minimization. The advantage of this choice is that the L_1 norm is a convex function on the space of candidate signals, which implies that the optimization problem in Equation 2, with $V(s) = |s|$, has no (nonglobal) local minima, and there are efficient algorithms for finding the global minimum using methods of linear programming (Boyd & Vandenberghe 2004), message passing (Donoho et al. 2009), and neural circuit dynamics (see below). CS theory shows that with an appropriate choice of $\mathbf{A}$, L_1 minimization exactly recovers the true signal so that $\hat{\mathbf{s}} = \mathbf{s}^0$, with a number of measurements that is roughly proportional to the number of nonzero elements in the source, K, which can be much smaller than the dimensionality N of the signal.

A popular and even simpler reconstruction algorithm is L_2 minimization in which $V(s) = s^2$ in Equation 2. This result can arise as a consequence of oft-used Gaussian priors on the unknown signal and leads to an estimate that is simply linearly related to the measurements through the pseudoinverse relation

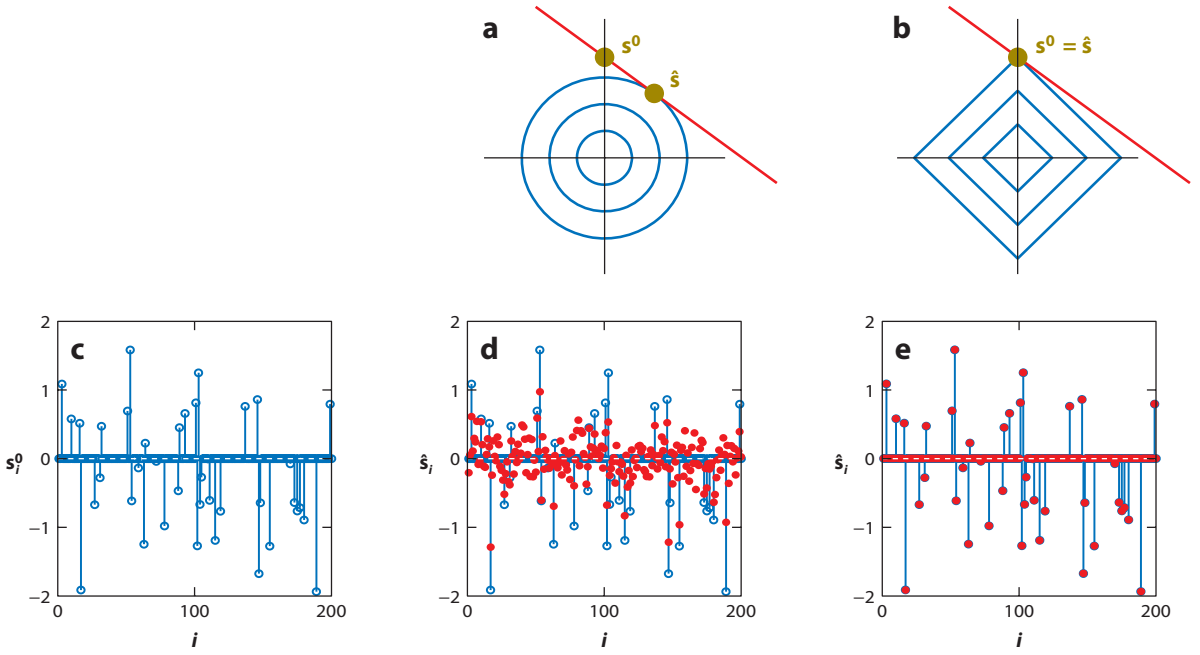

Figure 2

Geometry of compressed sensing (CS). (*a*) A geometric interpretation of L_2 minimization. An unknown $N = 2$ dimensional sparse signal $\mathbf{s}^0$ with $K = 1$ nonzero components is measured using $M = 1$ linear measurements, yielding a one-dimensional space of candidate signals consistent with the measurement constraints (*red line*). The estimate $\hat{\mathbf{s}}$ is the candidate signal with the smallest L_2 norm and can be found geometrically by expanding the locus of points with a fixed and increasing L_2 norm (*the blue circles*) until the locus first intersects the allowed space of candidate signals. This intersection point is the L_2 estimate $\hat{\mathbf{s}}$, which is different from the true signal $\mathbf{s}^0$. (*b*) In the identical scenario as in panel *a*, L_1 minimization recovers an estimate by expanding the locus of points with the same L_1 norm (*blue diamonds*), and in this case, the expanding locus first intersects the space of candidate signals at the true signal $\mathbf{s}^0$ so that perfect recovery $\hat{\mathbf{s}} = \mathbf{s}^0$ is achieved. Of course, a sparse signal could also have been located on the other coordinate axis, in which case L_1 minimization would have failed to recover $\mathbf{s}^0$ accurately. (*c*) An unknown sparse signal $\mathbf{s}^0$ of dimension $N = 200$, with $f = K/N = 0.2$, i.e., 20% of its elements are nonzero. (*d*) An estimate $\hat{\mathbf{s}}$ (*red dots*)recovered from $M = 120$ random linear measurements of $\mathbf{s}^0$ ($\alpha = N/T = 0.6$, or 60% subsampling) by L_2 minimization superimposed on the true signal $\mathbf{s}^0$. (*e*) From the same measurements in panel *d*, L_1 minimization yields an estimate $\hat{\mathbf{s}}$ (*red dots*) that coincides with the true signal. Note that the parameters of $f = 0.2$ and $\alpha = 0.6$ lie just above the phase boundary for perfect recovery in **Figure 3**.

$\hat{\mathbf{s}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$. **Figure 2** provides heuristic intuition for the utility of L_1 minimization and its superior performance over L_2 minimization in the case of sparse signals.

An interesting observation is that the bound in Equation 1 represents a sufficient condition on the number of measurements M for perfect signal recovery. Alternately, recent work on the typical behavior of CS in the limit where M and N are large has revealed that the performance of CS is surprisingly insensitive to the details of the measurement matrix $\mathbf{A}$ and the unknown signal $\mathbf{s}^0$ and depends only on the degree of subsampling $\alpha = M/N$ and the signal sparsity $f = K/N$. In the $\alpha - f$ plane, there is a universal, critical phase boundary $\alpha_c(f)$ such that if $\alpha > \alpha_c(f)$, then L_1 minimization will typically yield perfect signal reconstruction, whereas if $\alpha < \alpha_c(f)$, it will yield a nonzero error (see **Figure 3**) (Donoho & Tanner 2005a,b, Donoho et al. 2009, Kabashima et al. 2009, Ganguli & Sompolinsky 2010b).

Dimensionality Reduction by Random Projections

The above CS results can be understood using the theory of RPs. Geometrically, the mapping

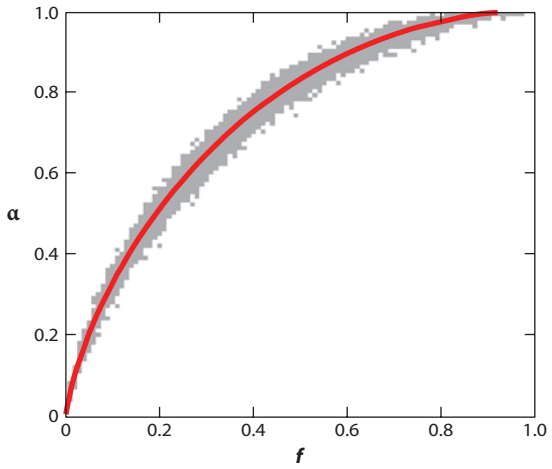

Figure 3

Phase transition in compressed sensing (CS) (reproduced from Ganguli & Sompolinsky 2010b). We use linear programming to solve Equation 2 50 times for each value of α and f in increments of 0.01, with $N = 500$. The grey transition region shows when the fraction of times perfect recovery occurs is neither 0 nor 1. The red curve is the theoretical phase boundary $\alpha_c(f)$. As $f \to 0$, this boundary is of the form $\alpha_c(f) = f \log 1/f$.

$\mathbf{x} = \mathbf{As}$ through a measurement matrix $\mathbf{A}$ can be thought of as a linear projection from a high N-dimensional space of signals down to a low M-dimensional space of measurements. In this geometric picture, the space of K-sparse signals consists of a low-dimensional (non-smooth) manifold, which is the union of all K-dimensional linear spaces characterized by K nonzero values at specific locations, as in **Figure 4*a***. Candes & Tao (2005) show that any projection that preserves the geometry of all K-sparse vectors allows one to reconstruct these vectors from the low-dimensional projection efficiently and robustly using L_1 minimization. The power of compression by RPs lies in the fact that they preserve the geometrical structure of this manifold. In particular, Baraniuk et al. (2008) show that RPs down to an $M = O(K \log(N/K))$ dimensional space preserve the distance between any pair of K-sparse signals up to a small distortion.

However, we can move beyond sparsity and consider how well RPs preserve the geometric structure of other signal or data patterns that lie on more general low-dimensional manifolds embedded in a high-dimensional space. An extremely simple manifold is a point cloud consisting of a finite set of points, as in **Figure 4*b***. Suppose this cloud consists of P points $\mathbf{s}^\alpha$, for $\alpha = 1, \ldots, P$, embedded in an N-dimensional space, and we project them down to the points $\mathbf{x}^\alpha = \mathbf{As}^\alpha$ in a low M-dimensional space through an appropriately normalized RP. How small can we make M before the point cloud becomes distorted in the low-dimensional space so that pairwise distances in the low-dimensional space are no longer similar to the corresponding distances in the high-dimensional space?

The celebrated Johnson-Lindenstrauss (JL) lemma (Johnson & Lindenstrauss 1984, Indyk & Motwani 1998, Dasgupta & Gupta 2003) provides a striking answer. It states that RPs with $M > O(\log P)$ will yield, with high probability, only a small distortion in distance between all pairs of points in the cloud. Thus the number of projected dimensions M needs only be logarithmic in the number of points P independent of the embedding dimension of the source data, N.

Finally, we consider data distributed along a nonlinear K-dimensional manifold embedded in N-dimensional space, as in **Figure 4*c***. An example might be a set of images of a single object observed under different lighting conditions, perspectives, rotations, and scales. Another example would be the set of neural firing-rate vectors in a brain region in response to a continuous family of stimuli. Baraniuk & Wakin (2009) and Baraniuk et al. (2010) show that $M > O(K \log NC)$ RPs preserve the geometry of the manifold with small distortion. Here C is a number related to the curvature of the manifold so that highly curved manifolds require more projections. Overall, these results show that surprisingly small numbers of RPs, which can be chosen without any knowledge of the data distribution, can preserve geometric structure in data.

Compressed Computation

Although CS emphasizes the reconstruction of sparse high-dimensional signals from

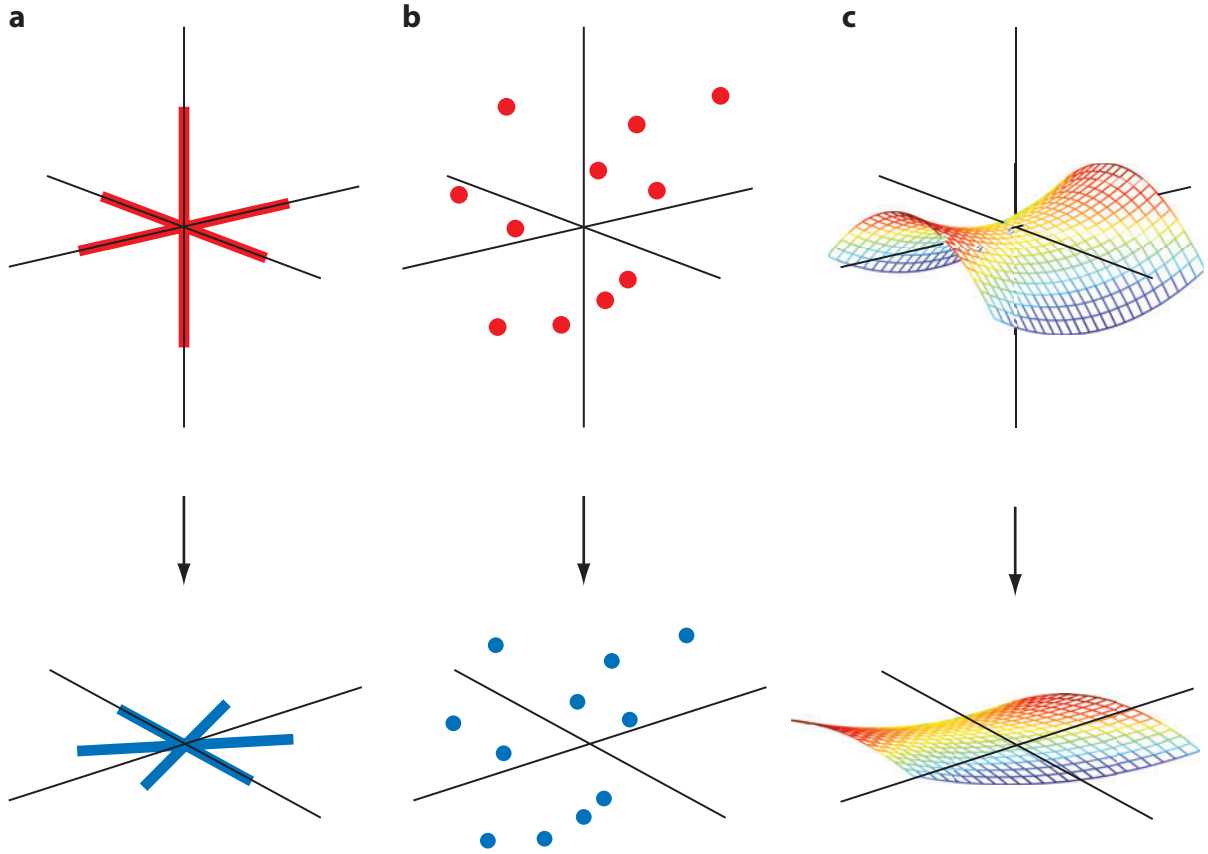

Figure 4

Random Projections. (*a*) A manifold of *K*-sparse signals (red) in *N*-dimensional space is randomly projected down to an *M*-dimensional space (here $K = 1$, $N = 3$, $M = 2$). (*b*,*c*) Projection of a point cloud, and a nonlinear manifold respectively.

low-dimensional projections, many important problems in signal processing and learning can be accomplished by performing computations directly in the low-dimensional space without the need to first reconstruct the high-dimensional signal. For example, regression (Zhou et al. 2009), signal detection (Duarte et al. 2006), classification (Blum 2006, Haupt et al. 2006, Davenport et al. 2007, Duarte et al. 2007), manifold learning (Hegde et al. 2007), and nearest neighbor finding (Indyk & Motwani 1998) can all be accomplished in a low-dimensional space given a relatively small number of RPs. Moreover, task performance is often comparable to what can be obtained by performing the task directly in the original high-dimensional space. The reason for this remarkable performance is that these computations rely on the distances between data points, which are preserved by RPs. Thus RPs provide one way to cope with the curse of dimensionality, and as we discuss below, this can have significant implications for neuronal information processing and data analysis.

Approximate Sparsity and Noise

Above, we have assumed a definition of sparsity in which an *N*-dimensional signal $\mathbf{s}^0$ has $K < N$ nonzero elements, with the other elements being exactly 0. In reality, many of the coefficients of a signal may be small, but they are unlikely to be exactly zero. We thus expect signals not to be exactly sparse but to be well

approximated by a K-sparse vector $\mathbf{s}^0_K$, which is obtained by keeping the K largest coefficients of $\mathbf{s}^0$ and setting the rest of them to 0. In addition, we have to allow for measurement noise so that $\mathbf{x} = \mathbf{A}\mathbf{s}^0 + \mathbf{z}$, where $\mathbf{z}$ is a noise vector whose μ'th component is zero mean Gaussian noise with a fixed variance.

In the presence of noise, it no longer makes sense to enforce perfectly the measurement constraints $\mathbf{x} = \mathbf{A}\mathbf{s}$. Instead, a common approach, known as the LASSO method, is to solve the alternate optimization problem

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \left\{ \|\mathbf{x} - \mathbf{A}\mathbf{s}\|^2 + \lambda \sum_{i=1}^{T} V(s_i) \right\}, \quad 3.$$

where $V(s) = |s|$ (the absolute value function) and λ is a parameter to be optimized. The cost function minimized here allows deviations between $\mathbf{A}\mathbf{s}$, which are the noise-free measurement outcomes generated by a candidate signal $\mathbf{s}$, and the actual noisy measurements $\mathbf{x}$. However, such deviations are penalized by the quadratic term in Equation 3.

Several works (see e.g., Candes et al. 2006, Wainwright 2009, Bayati et al. 2010, Candes & Plan 2010) have addressed the performance of the LASSO in the combined situation of noise and departures from perfect sparsity. The main outcome is roughly that for an appropriate choice of λ, which depends on the signal-to-noise ratio (SNR), the same conditions that guaranteed exact recovery of K-sparse signals by L_1 minimization in the absence of noise also ensure good performance of the LASSO for approximately sparse signals in the presence of noise. In particular, whenever $\mathbf{s}^0_K$ is a good approximation to $\mathbf{s}^0$, the LASSO estimate $\hat{\mathbf{s}}$ in Equation 3 is a good approximation to $\mathbf{s}^0$, up to a level of precision that is allowed by the noise.

Sparse Models of High-Dimensional Data

L_1-based minimization can also be applied to the modeling of high-dimensional data. A simple example is sparse linear regression. Suppose that our data set consists of M N-dimensional vectors, $\mathbf{a}^\mu$, along with M scalar response variables x_μ. The regression model assumes that on each observation, μ, $x_\mu = \mathbf{a}^\mu \cdot s^0 + z_\mu$, where $\mathbf{s}^0$ is an N-dimensional vector of unknown regression coefficients and z_μ is Gaussian measurement noise. This can be summarized in the matrix equation $\mathbf{x} = \mathbf{A}\mathbf{s}^0 + \mathbf{z}$, where the M rows of the $M \times N$ matrix $\mathbf{A}$ are the N-dimensional data points, $\mathbf{a}^\mu$. Now if the number of data points M is fewer than the dimensionality of the data N, it would seem hopeless to infer the regression coefficients. However, in many high-dimensional regression problems, we expect that the regression coefficients will be sparse. For example, $\mathbf{a}^\mu$ could be a vector of expression levels of $N = O(1000)$ genes measured in a microarray under experimental condition μ, and x_μ could be the response of a biological signal of interest. However, only a small fraction of genes are expected to regulate any given signal of interest, and hence we expect the regression coefficients $\mathbf{s}^0$ to be sparse.

This scenario is exactly equivalent to the case of CS with noise. Here the regression coefficients $\mathbf{s}^0$ play the role of an unknown sparse signal to be recovered, the input data points $\mathbf{a}^\mu$ play the role of the measurement vectors, and the scalar output or response x_μ plays the role of the measurement outcome in CS. The same LASSO algorithm described in Equation 3 can be used to infer the regression coefficients (Tibshirani 1996). Here, the parameter λ is not set by the SNR but rather is chosen to minimize some measure of the prediction error on a new input. This estimate can be obtained through cross validation, for example. Efron et al. (2004) have proposed efficient algorithms to compute $\hat{\mathbf{s}}$, optimizing over λ for a given data set ($\mathbf{A}$, $\mathbf{x}$).

The technique of L_1 regularization generalizes beyond linear regression to the problem of learning large statistical models with expected sparse parameter sets. Indeed it has been used successfully in learning logistic regression (Lee et al. 2006b) and in various graphical models (Lee et al. 2006a, Wainwright et al. 2007), as well as in point process models of neuronal spike trains (Kelly et al. 2010).

Dictionary Learning

As Equations 2 and 3 imply, to reconstruct a signal from a small number of random measurements using L_1 minimization, we need to know $\mathbf{A} = \mathbf{BC}$, which means that we need to know the basis $\mathbf{C}$ in which the signal is sparse. What if we have to work with a new ensemble of signals and we do not yet know of a basis in which these signals are sparse?

One approach is to perform dictionary learning (Olshausen et al. 1996; Olshausen & Field 1996, 1997) on the ensemble of signals. Suppose $\{\mathbf{x}^\alpha\}$ for $\alpha = 1, \ldots, P$ is a collection of P M-dimensional signals. We imagine that each signal is well approximated by a sparse linear combination of the columns of an unknown $M \times N$ matrix $\mathbf{A}$, i.e., $\mathbf{x}^\alpha \approx \mathbf{As}^\alpha$ for all $\alpha = 1, \ldots, P$, where $\mathbf{s}^\alpha$ is an unknown sparse N-dimensional vector. We refer to the columns of $\mathbf{A}$ as the dictionary elements. Thus, the nonzero coefficients of $\mathbf{s}^\alpha$ indicate which dictionary elements linearly combine to form the signal $\mathbf{x}^\alpha$. Here N can be larger than M, in which case we are looking for an overcomplete basis, or dictionary, to represent the ensemble of signals. Given our training signals $\mathbf{x}^\alpha$, we wish to find the sparse codes $\mathbf{s}^\alpha$ and dictionary $\mathbf{A}$. These can potentially be found by minimizing the following energy function:

$$E(\mathbf{s}^1, \ldots, \mathbf{s}^P, \mathbf{A}) = \sum_{\alpha=1}^{P} (\|\mathbf{x}^\alpha - \mathbf{As}^\alpha\|^2 + \lambda \|\mathbf{s}^\alpha\|_1), \quad 4.$$

where $||\mathbf{s}^\alpha||_1$ denotes the L_1 norm of $\mathbf{s}^\alpha$. For each α, this second term enforces the sparsity of the code, whereas the first quadratic cost term enforces the fidelity of the code and the dictionary. Subsequent work (Kreutz-Delgado et al. 2003; Aharon et al. 2006a,b) has extended this basic formalism as well as derived efficient algorithms for solving Equation 4. Moreover, Aharon et al. (2006b), Isely et al. (2010), and Hillar & Sommer (2011) have recently shown that if the signals $\mathbf{x}^\alpha$ are indeed generated by sparse noiseless codes through a dictionary $\mathbf{A}$, under certain conditions related to CS, dictionary learning will recover $\mathbf{A}$, up to permutations and scalings of its columns.

COMPRESSED SENSING OF THE BRAIN

Rapid Functional Imaging

In many ways, magnetic resonance imaging (MRI) is a well-suited application for CS (Lustig et al. 2008). In MRI, a strong static magnetic field with a linear spatial gradient, $\Delta\mathbf{H}$, causes magnetic dipoles in a tissue sample to align with the magnetic field. A radio frequency excitation pulse then generates a transverse complex magnetic moment at location $\mathbf{r}$, with amplitude $m(\mathbf{r})$ and a phase $\phi(\mathbf{r})$ proportional to $\mathbf{r} \cdot \Delta\mathbf{H}$. Depending on the sample preparation, the amplitudes $m(\mathbf{r})$ correlate with various local properties of interest. For example, in functional MRI, it correlates with the concentration of oxygenated hemoglobin, which in turn increases in response to neural activity. Thus, the measurement goal is to extract the spatial profile of $m(\mathbf{r})$. A detector coil measures the spatial integral of the complex magnetization. Hence, it essentially measures a spatial Fourier transform of the profile with a Fourier wave vector $\mathbf{k} = (\mathbf{k_x}, \mathbf{k_y}, \mathbf{k_z}) \propto \Delta\mathbf{H}$.

The traditional approach to MR imaging has been to sample the image densely through a regular lattice in Fourier wave vector space, or $\mathbf{k}$-space, by generating a sequence of static linear gradient fields and radio frequency pulses. If the Fourier space is sampled at the Nyquist-Shannon rate, then one can perform a linear reconstruction of the image $m(\mathbf{r})$ simply by performing an inverse Fourier transform of the measurements. However, acquiring each Fourier sample can take time, so any method to reduce the number of such samples can dramatically reduce patient time in scanners, as well as increase the temporal resolution of dynamic imaging.

CS provides an interesting approach to reducing the number of measurements. In the CS framework, the measurement basis $\mathbf{B}$ in **Figure 1** consists of Fourier modes. CS will work well if the MRI image is sparse in a basis $\mathbf{C}$ that is incoherent with respect to $\mathbf{B}$. For example, many MRI images, such as angiograms, are sparse in the position, or pixel basis. For such

images, one can subsample random trajectories in **k**-space and use nonlinear L_1 reconstruction to recover the image. For appropriately chosen random trajectories, one can obtain high-quality images using a tenth of the number of measurements required in the traditional approach (Lustig et al. 2008). Similarly, brain images are often sparse in a wavelet basis, and for such images, random trajectories in **k**-space can be found that speed up the rate at which images can be acquired by a factor of 2.4 compared with the traditional approach (Lustig et al. 2007). Moreover, dynamic movies of oscillatory phenomena that are sparse in the temporal frequency domain can be obtained at high temporal resolution by sampling randomly both in **k**-space and in time (Parrish & Hu 1995).

Fluorescence Microscopy

Simultaneously imaging the dynamics of multiple molecular species at both high spatial and temporal resolution is a central goal of cellular microscopy. CS-inspired technologies such as single-pixel cameras (Takhar et al. 2006, Duarte et al. 2008) combined with fluorescence microscopy techniques (Wilt et al. 2009, Taraska & Zagotta 2010) provide one promising route toward such a goal (Coskun et al. 2010; E. Candes, personal communication). In fluorescence imaging, multiple molecular species can be tagged with markers capable of emitting light at different frequencies. Imaging the molecules then requires two key steps: First, the sample must be illuminated with light, causing the tagged species to fluoresce, and second, the emitted photons from the fluorescent species must be detected. Traditionally, two main methods have been used to accomplish both steps. In widefield (WF) microscopy, the entire image is illuminated at once, and a large array of detectors records the emitted photons. In raster scan (RS) microscopy, each point of the image is illuminated in sequence, so only one detector is required to collect the emitted photons at any given time.

WF can achieve high temporal resolution but requires many photodetectors for high spatial resolution. This is problematic for imaging applications in which photons at many different frequencies, corresponding to different molecules, need to be simultaneously measured. This requires a prohibitively expensive high-density array of photodetectors that can perform hyperspectral imaging, i.e., measure many spectral channels at once. One could employ a single such detector in RS mode, but then achieving high spatial resolution comes at the cost of low temporal resolution because of the required number of raster scans.

The single-pixel-camera approach exploits the potential spatial sparsity of a fluorescence image to achieve both high spatial and temporal resolution. In this approach, the image is illuminated using a sequence of random light patterns. This can be achieved by a digital micromirror device (DMD), which consists of a spatial array of micrometer scale mirrors whose angles can be rapidly and individually adjusted. Light is reflected off this array into the sample, and on each trial, a different configuration of mirrors leads to a different pattern of illumination. A single hyperspectral photodetector (the single pixel) then measures the total emitted fluorescence. Owing to the randomness of the light patterns, the image can be reconstructed at the micrometer spatial resolution of the DMD using a number of measurements that is much smaller than the number of pixels (or resolvable spatial locations) in the image. Thus compressive imaging retains the relative speed and resolution of WF and the simplicity and achievable spectral range of RS. As such, this rapidly evolving method has the potential to open up new experimental windows into the dynamics of intracellular molecular cascades within neurons.

Gene-Expression Analysis

The use of microarrays to collect large-scale data sets of gene-expression levels across many brain regions is now a well-established enterprise in neuroscience. Suppose we want to measure a vector $\mathbf{s}^0$ of concentrations of N genetic sequences in a sample. A microarray consists of N spots, indexed by $i = 1, \ldots, N$, where

each spot i contains a unique complementary sequence that will specifically bind with the sequence i in the sample. All N genetic sequences of interest in the sample are fluorescently tagged and exposed to all the spots. Each spot binds a specific sequence, and after the excess unbound DNA is washed off, the vector of concentrations $\mathbf{s}^0$ can be read off by imaging the fluorescence levels of the spots.

Often this procedure is highly inefficient because any particular sample will contain only a few genetic sequences of interest, i.e., the concentration vector $\mathbf{s}^0$ is sparse. Dai et al. (2009) proposed a CS-based approach in which one can use $M < N$ spots, where each spot contains a random subset of the N sequences of interest. Thus each spot, now indexed by $\mu = 1, \ldots, M$, is characterized by an N-dimensional measurement vector $\mathbf{a}^\mu$, where the component $\mathbf{a}_i^\mu$ reflects the binding affinity of sequence i in the sample to the contents of spot μ. After the CS microarray is exposed to the sample, the M-dimensional vector of fluorescence levels $\mathbf{x}$ is approximately related to the sample concentration $\mathbf{s}^0$ through the linear relation $\mathbf{x} = \mathbf{A}\mathbf{s}^0$, where the rows of $\mathbf{A}$ are the measurement vectors $\mathbf{a}^\mu$. Thus if each spot contains enough randomly chosen complementary sequences, such that the measurements are incoherent with regard to the basis of sequences, one can use the LASSO method in Equation 3 to recover the concentrations $\mathbf{s}^0$ from the fluorescence measurements $\mathbf{x}$. Dai et al. (2009) do a thorough analysis of this basic framework. Overall, reducing the number of spots required to collect gene expression data reduces both the cost and the size of the array, as well as the amount of biological sample material required to make accurate concentration measurements.

Compressed Connectomics

The problem of reconstructing functional circuit connectivity from recordings of neuronal postsynaptic responses presents a considerable challenge to neuroscience. Consider, for example, a simple scenario in which we have a population of N neurons that are potentially presynaptic to a given neuron whose membrane voltage x we can record intracellularly. The synaptic strengths from the N neurons to the recorded neuron is an unknown N-dimensional vector $\mathbf{s}^0$. The traditional approach to estimating this set of synaptic strengths is to excite each potential presynaptic neuron one by one and record the resultant postsynaptic membrane voltage x. Each such measurement reveals the strength of one synapse. This brute-force approach is highly inefficient because the synaptic connectivity $\mathbf{s}^0$ is often sparse, with only $K < N$ nonzero elements, where K/N is $\sim$10%. Thus most measurements would simply yield 0.

Hu & Chklovskii (2009) propose a CS-based approach to recovering $\mathbf{s}^0$ by randomly stimulating F neurons out of N on any given trial μ. This method corresponds to a random measurement matrix $\mathbf{A}$ characterized by F nonzero entries per row. Given that the true weight vector $\mathbf{s}^0$ is sparse, Hu & Chklovskii (2009) propose to use L_1 minimization in Equation 2 to recover $\mathbf{s}^0$ from knowledge of the inputs $\mathbf{A}$ and outputs $\mathbf{x}$. The authors find for a wide range of parameters that $F/N = 0.1$ minimizes the required number of measurements, M, and for this value of F, $M = O(K \log N)$ measurements are required to recover $\mathbf{s}^0$. Thus random stimulation of 10% of the population constitutes an effective measurement basis for CS of synaptic connectivity (Hu & Chklovskii 2009). Alternative ideas have been proposed for CS of connectivity using fluorescent synaptic markers (Mishchenko 2011).

COMPRESSED SENSING BY THE BRAIN

The problem of storing, communicating, and processing high-dimensional neural activity patterns, or external stimuli, presents a fundamental challenge to any neural system. This challenge is complicated by the widespread existence of convergent pathways, or bottlenecks, in which information stored in a large number of neurons is often compressed into a small number of axons, or neurons in a

downstream system. For example, 1 million optic nerve fibers carry information about the activity of 100 times as many photoreceptors. Only 1 million pyramidal tract fibers carry information from motor cortex to the spinal cord. And corticobasal ganglia pathways undergo a 10–1,000-fold convergence. In this section we review how the theory of CS and RPs yields theoretical insight into how efficient storage, communication, and computation are possible despite drastic reductions in the dimensionality of neural representations through information bottlenecks.

Semantic Similarity and Random Projections

How much can a neural system reduce the dimensionality of its activity patterns without incurring a large loss in its ability to perform relevant computations? A plausible minimal requirement is that any reduction through a convergent pathway should preserve the similarity structure of the neuronal representations at the source area. This requirement is motivated by the observation that in higher perceptual or association areas in the brain semantically similar objects elicit similar neural activity patterns (Kiani et al. 2007). This similarity structure of the neural code is likely the basis of our ability to categorize objects and generalize appropriate responses to new objects (Rogers & McClelland 2004). Moreover, this similarity structure is remarkably preserved across monkeys and humans, for example, in image representations in the inferotemporal (IT) cortex (Kriegeskorte et al. 2008).

When a semantic task involves a finite number of activity patterns, or objects, the JL lemma discussed above implies that the required communication resources vary only logarithmically with the number of patterns, independent of how many neurons are involved in the source area. For example, suppose 20,000 images can be represented by the corresponding population activity patterns in the IT cortex. Then the similarity structure between all pairs of images can be preserved to 10% precision in a downstream area using only ∼1000 neurons. Furthermore, this result can be achieved with a very simple dimensionality-reduction scheme, namely by a random synaptic connectivity matrix. Moreover, any computation that relies on similarity structure, and can be solved by the IT cortex, can also be solved by the downstream region.

A more stringent challenge occurs when convergent pathways must preserve the similarity structure of not just a finite set of neuronal activity patterns, but an arbitrarily large, possibly infinite, number of patterns, as is likely the case in any pathway that represents information about continuous families of stimuli. The theories of CS and RPs of manifolds discussed above reveal that again drastic compression is possible if the corresponding neural patterns are sparse or lie on a low-dimensional manifold (for example, as in **Figure 4*a–c***). In this case, the number of required neurons in a randomly connected downstream area is proportional to the intrinsic dimension of the ensemble of neural activity patterns and depends only weakly (logarithmically) on the number of neurons in the source area.

Hidden low-dimensional structure in neural activity patterns has been found in several systems (Ganguli et al. 2008a, Yu et al. 2009, Machens et al. 2010), and moreover, intrinsic spatiotemporal fluctuations exhibited in many models of recurrent neuronal circuits, including chaotic networks, are low dimensional (Rajan et al. 2010, Sussillo & Abbott 2009). The ubiquity of this low-dimensional structure in neuronal systems may be intimately related to the requirement of communication and computation through widespread anatomical bottlenecks.

Short-Term Memory in Neuronal Networks

Another bottleneck is posed by the task of working memory, where streams of sensory inputs must presumably be stored within the dynamic reverberations of neuronal circuits. This is a bottleneck from time into space: Long

temporal streams of input must be stored in the instantaneous spatial activity patterns of a limited number of neurons. The influential idea of attractor dynamics (Hopfield 1982) suggests how single stimuli can be stored as stable patterns of activity, or fixed points, but such simple fixed points are incapable of storing temporal sequences of information, like an ongoing sentence, song, or motion trajectory. More recent proposals (Jaeger 2001, Maass et al. 2002, Jaeger & Haas 2004) suggest that recurrent networks could store temporal sequences of inputs in their ongoing, transient activity. This new paradigm raises several theoretical questions about how long memory traces can last in such networks, as functions of the network size, connectivity, and input statistics. Several studies have addressed these questions in the case of simple linear neuronal networks and Gaussian input statistics. These studies show that the duration of memory traces in any network cannot exceed the number of neurons (in units of the intrinsic time constant) (Jaeger 2001, White et al. 2004) and that no network can outperform an equivalent delay line or a nonnormal network, characterized by a hidden feedforward structure (Ganguli et al. 2008b).

However, a more ethologically relevant temporal input statistic is that of a sparse, non-Gaussian sequence. Indeed a wide variety of temporal signals of interest are sparse in some basis, for example, human speech in a wavelet basis. Recent work (Ganguli & Sompolinsky 2010a) has derived a connection between CS and short-term memory by showing that recurrent neuronal networks can essentially perform online, dynamical compressed sensing of an incoming sparse sequence, yielding sequence memory traces that are longer than the number of neurons, again in units of the intrinsic time constant. In particular, neuronal circuits with M neurons can remember sparse sequences, which have a probability f of being nonzero at any given time for an amount of time that is $O(\frac{M}{f \log(1/f)})$. This enhanced capacity cannot be attained by purely feedforward networks, or random Gaussian network connectivities, but requires antisymmetric connectivity matrices that generate complex transient activity patterns and diverse temporal filtering properties.

SPARSE EXPANDED NEURONAL REPRESENTATIONS

In the previous section, we have discussed how CS and RPs can explain how convergent pathways can compress neuronal representations. However, in many computations, neural systems may need to expand these low-dimensional compressed representations back into high-dimensional sparse ones. For example, such representations reduce the overlap between activity patterns, thereby simplifying the tasks of learning, discrimination, categorization, noise filtering, and multiscale stimulus representation. Indeed, like convergence, the expansion of neural representations through divergent pathways is a widespread anatomical motif. For example, information in 1 million optic nerve fibers is expanded into more than 100 million primary visual cortical neurons. Also in the cerebellum, a small number of mossy fibers target a large number of granule cells, creating a 100-fold expansion.

How do neural circuits transform compressed dense codes into expanded sparse ones? A simple mechanism would be to project the dense activity patterns into a larger pool of neurons via random divergent projections and use high spiking thresholds to ensure sparsity of the target activity patterns. Indeed, Marr (1969) suggested this mechanism in his influential hypothesis that the granule cell layer in the cerebellar cortex performs sparse coding of dense stimulus representations in incoming mossy fibers to facilitate learning of sensorimotor associations at the Purkinje cell layer. Although random expansion may work for some computations, sparse codes are generally most useful when they represent essential sparse features of the compressed signal. In the next sections, we review how CS methods for generating sparse expanded representations, which faithfully capture hidden structures in compressed data, can operate within neural systems.

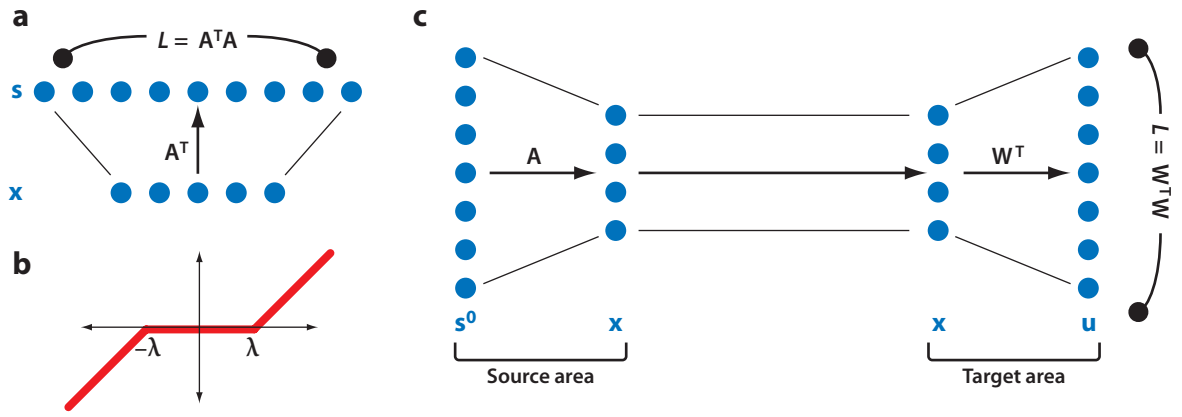

Figure 5

Neural L_1 minimization and long-range brain communication. (*a*) A two-layer circuit for performing L_1 minimization and dictionary learning. (*b*) Nonlinear transfer function from inputs to firing rates of neurons in the second layer in panel *a*. (*c*) A scheme for efficient long-range brain communication in which sparse activity $\mathbf{s}^0$ is compressed to a low-dimensional dense representation $\mathbf{x}$ in a source area and efficiently communicated downstream to a target area with a small number of axons, where it could re-expanded into a new sparse representation $\mathbf{u}$ through a dictionary learning circuit as in panel *a*.

Neuronal Implementations of L1 Minimization

Given that solving the optimization problem in Equation 3 with $V(s) = |s|$ has proven to be an efficient method for sparse signal reconstruction, whether neuronal circuits can perform this computation is a natural question. Here we describe one plausible two-layer circuit solution (see **Figure 5*a***) proposed in Rozell et al. (2008), inspired by gradient descent in $\mathbf{s}$ on the cost function in Equation 3. Suppose that the low M-dimensional input $\mathbf{x}$ is represented in the first layer by the firing rates of a population of M neurons such that the μth input neuron has a firing rate $\mathbf{x}_\mu$. Now suppose that the reconstructed sparse signal is represented by a larger population of N neurons where s_i is the firing rate of neuron i. In this population, we denote the synaptic potential for each neuron by v_i, which determines the neuron's firing rate via a static nonlinearity F, $s_i = F(v_i)$.

The synaptic connectivity from the M input neurons to the N second-layer neurons computing the sparse representation $\mathbf{s}$ is given by the $N \times M$ matrix $\mathbf{A}^T$ such that the ith column of $\mathbf{A}$, $\mathbf{a}^i$, denotes the set of M synaptic weights from the input neurons to neuron i in the second layer. Finally, assume there is lateral inhibition between any pair of neurons i and j in the second layer, governed by synaptic weights L_{ij}, which are related to the feedforward weight vectors to the pair of neurons, through $L_{ij} = \mathbf{a}^i \cdot \mathbf{a}^j$. Then the internal dynamics of the second-layer neurons obey the differential equations

$$\tau \frac{dv_i}{dt} = -v_i + \mathbf{a}^i \cdot \mathbf{x} - \sum_{j=1}^{T} L_{ij} s_j, \qquad 5.$$

where $\mathbf{x}$ is the activity of the input layer. Rozell et al. (2008) found that for an appropriate choice of the static nonlinearity, this dynamic is similar to a gradient descent on the cost function given by Equation 3. In particular, for L_1 minimization, the static nonlinearity F is simply a threshold linear function with threshold λ and gain 1 (see **Figure 5*b***).

To obtain a qualitative understanding of this circuit, consider what happens when the second-layer activity pattern is initially inactive so that $\mathbf{s} = 0$ and an input $\mathbf{x}$ occurs in the first layer. Then the internal variable $v_i(t)$ of each second-layer neuron i will charge up with a rate controlled by the overlap of the input $\mathbf{x}$ with the synaptic weight vector, $\mathbf{a}^i$, which is closely related to the receptive field (RF) of neuron

i. As neuron *i*'s internal activation crosses the threshold λ, it starts to fire and inhibits neurons with RFs similar to $\mathbf{a}^i$. This sets up a competitive dynamic in which a small number of neurons with RFs similar to the input $\mathbf{x}$ come to represent it, yielding a sparse representation $\hat{\mathbf{s}}$ of the input $\mathbf{x}$, which is the solution to Equation 3. In the case of zero noise, the above circuit dynamic needs to be supplemented with an appropriate dynamic update of the threshold λ, which eventually approaches zero at the fixed point (Donoho et al. 2009). Finally, we note that several works (Olshausen et al. 1996, Perrinet 2010) have proposed synaptic Hebbian plasticity and homeostasis rules that supplement Equation 5 and allow the circuit to solve the full dictionary learning problem, Equation 4, without prior knowledge of $\mathbf{A}$.

An intriguing feature of the above dynamic is that the inhibitory recurrent connections are tightly related to the feedforward excitatory drive. Koulakov & Rinberg (2011) suggest that exactly this computation may be implemented in the rodent olfactory bulb. They propose that reciprocal dendrodendritic synaptic coupling between mitral cells and granule cells yields an effective lateral inhibition between granule cells that is related to the feedforward drive from mitral cells to granule cells, in accordance with the requirements of Equation 5. Thus the composite olfactory circuit builds up a sparse code for odors in the granule cell population. Likewise, Hu et al. (2011) proposed that sparse coding is implemented within the amacrine/horizontal cell layers in the retina.

Compression and Expansion in Long-Range Brain Communication

A series of papers (Coulter et al. 2010, Isely et al. 2010, Hillar & Sommer 2011) have integrated the dual aspects of CS theory: dimensionality reduction of sparse neural representations, and the recoding of stimuli in sparse overcomplete representations into a theory of efficient long-range brain communication (see also Tarifi et al. 2011). According to this theory (see **Figure 5*c***), each area in a long-range communication pathway has both dense and sparse representations. Local sparse representations are first compressed to communicate them using a small number of axons and potentially re-expanded in a downstream area.

Where in the brain might these transformations occur? Coulter et al. (2010) predict that this could occur within every cortical column, with compressive projections, possibly random, occurring between more superficial cortical layers and the output layer 5. A key testable physiological prediction would then be that activity in more superficial layers is sparser than activity in deeper output layers. Another possibility is the transformation from sparse high-dimensional representations of space in the CA3/CA1 fields of the hippocampus to denser, lower-dimensional representations of space in the subiculum, which constitutes the major output structure of the hippocampus. A functional explanation for this representational dichotomy could be that the hippocampus is performing an RP from CA3/CA1 to the subiculum, thereby minimizing the number of axons required to communicate the results of hippocampal computations to the rest of the brain.

Overall, these works suggest more generally that random compression and sparse coding can be combined to yield computational strategies for efficient use of the limited bandwidth available for long-range brain communication.

LEARNING IN HIGH-DIMENSIONAL SYNAPTIC WEIGHT SPACES

Learning new skills and knowledge is thought to be achieved by continuous synaptic modifications that explore the space of possible neuronal circuits, selecting through experience those that are well adapted to the given task. We review how regularization techniques used by statisticians to learn high-dimensional statistical models from limited amounts of data can also be employed by synaptic learning rules to search efficiently the high-dimensional space of synaptic patterns to learn appropriate rules from limited experience.

Neural Learning of Classification

A simple model of neural decision making and classification is a single-layer feedforward network in which the postsynaptic potential of the readout neuron is a sum of the activity of its afferents, weighted by a set of synaptic weights, and the decision is signaled by firing or not firing depending on whether the potential reaches threshold. Such a model is equivalent to the classical perceptron (Rosenblatt 1958). Computationally, this model classifies N-dimensional input patterns into two categories separated by a hyperplane determined by the synaptic weights. These weights are learned through experience-dependent modifications based on a set of M training input examples and their correct classifications. Of course, the goal of any organism is not to classify past experience correctly, but rather to generalize to novel experience. Thus an important measure of learning performance is the generalization error, or the probability of incorrectly classifying a novel input, and a central question of learning theory is how many examples M are required to achieve a good generalization error given a number of N unknown synaptic weights that need to be learned.

This question has been studied exhaustively (Gardner 1988, Seung et al. 1992) (see Engel & den Broeck 2001 for an overview), and the general consensus finds that for a wide variety of learning rules, a small generalization error can occur only when the number of examples M is larger than the number of synapses N. This result has striking implications because it suggests that learning may suffer from a curse of dimensionality: Given the large number of synapses involved in any task, this theory suggests we need an equally large number of training examples to learn any task.

Recent work (Lage-Castellanos et al. 2009) has considered the case when a categorization task can be realized by a sparse synaptic weight vector, meaning that only a subset of inputs are task relevant, though which subset is apriori unknown. The authors showed that a simple learning rule that involves minimization of the classification error on the training set, plus an L_1 regularization on the synaptic weights of the perceptron, yields a good generalization error even when the number of examples can be less than the number of synapses. Thus a sparsity prior is one route to combat the curse of dimensionality in learning tasks that are realizable by a sparse rule.

Optimality and Sparsity of Synaptic Weights

Consider again the perceptron learning to classify a finite set of M input patterns. In general, many synaptic weight vectors will classify these inputs correctly. We can, however, look for the optimal weight vector that maximizes the margin, or the minimal distance between input patterns and the category boundary. For such weights, the induced synaptic potentials are as far as possible from threshold, and the resultant classifications yield good generalization and noise tolerance (Vapnik 1998).

A remarkable theoretical result is that if synapses are constrained to be either excitatory or inhibitory, then near capacity, the optimal solution is sparse, with most of the synapses silent (Brunel et al. 2004), even if the input patterns themselves show no obvious sparse structure. This result has been proposed as a functional explanation for the abundance of silent synapses in the cerebellum and other brain areas.

When the sign of the weights are unconstrained, the optimal solutions are still sparse, but not in the basis of neurons. Instead, the optimal weight vector can be expressed as a linear combination of a small number of input patterns, known as support vectors, the number of support vectors being much smaller than their dimensionality. Indeed, several powerful learning algorithms, including support vector machines (SVMs) (see Burges 1998, Vapnik 1998, Smola 2000 for reviews), exploit this form of sparsity to achieve good generalization from relatively few high-dimensional examples.

Finally, because a sufficiently large number of RPs preserve Euclidean distances, they incur only a modest reduction in the margin of the optimal category boundary separating classes (Blum 2006). Hence, classification problems can also be learned directly in a low-dimensional space. In summary, there is an interesting interplay among sparsity, dimensionality, and the learnability of high-dimensional classification problems: Any such rapidly learnable problem (i.e., one with a large margin), is both (*a*) sparse, in the sense that its solution can be expressed in terms of a sparse linear combination of input patterns, and (*b*) low-dimensional in the sense that it can be learned in a compressed space after a RP.

DISCUSSION

Dimensionality Reduction: CS versus Efficient Coding

Efficient coding theories (Barlow 1961, Atick 1992, Atick & Redlich 1992, Barlow 2001) suggest that information bottlenecks in the brain perform optimal dimensionality reduction by maximizing mutual information between the low-dimensional output and the high-dimensional input (Linsker 1990). The predictions of such information maximization theories depend on assumptions about input statistics, neural noise, and metabolic constraints. In particular, infomax theories of early vision, based on Gaussian signal and noise assumptions, predict that high-dimensional spatiotemporal patterns of photoreceptor activation should be projected onto the linear subspace of their largest principal components. Furthermore, the individual projection vectors, i.e., retinal ganglion cell (RGC) RFs, depend on the stimulus SNR; in particular, at a high SNR, RFs should decorrelate or whiten the stimulus. This is consistent with the center-surround arrangement of RFs, which removes much of the low-frequency correlations in natural images (Atick 1992, Atick & Redlich 1992, Borghuis et al. 2008).

What is the relation between infomax theories and CS? According to CS theory, for sparse inputs, close to optimal dimensionality reduction is achieved when the projection vectors are maximally incoherent with respect to the basis in which the stimulus is sparse. Assuming visual stimuli are approximately sparse in a wavelet or Gabor-like basis, incoherent projections are likely to be spatially distributed. If sparseness is a prominent feature of natural visual spatiotemporal signals, how can we reconcile the observed RGC center-surround RFs with the demand for incoherence? Incoherent or random projections are optimal for signal ensembles composed of a combination of a few feature vectors in which the identity of these vectors varies across signals. This may be an adequate description of natural images after whitening. However, prewhitened natural images have strong second-order correlations, implying that they lie close to a low-dimensional linear space given by their principal components. Thus, the ensemble of natural images is characterized by both linear low-dimensional structure and sparse structure imposed by higher-order statistics. In such ensembles, whether sensory stimuli or neuronal activity patterns, when second-order correlations are strong enough, the optimal dimensionality reduction may indeed be close to that predicted by Gaussian-based infomax, as has been argued in recent work (Weiss et al. 2007).

Expansion and Sparsification: Compressed Sensing versus Independent Components Analysis

What does efficient coding theory predict regarding the recoding of signals through expansive transformations, for example, from the optic nerve to visual cortex? Several modern efficient coding theories, such as basis pursuit, independent components analysis (ICA), maximizing non-Gaussianity, and others, suggest that even after decorrelation, natural images include higher-order statistical dependencies that arise through linear mixing of statistically independent sources. The role of the cortical representation is to further reduce the

redundancy of the signal by separating the mixed signal into its independent causes (i.e., an unmixing operation), essentially generating a factorial statistical representation of the signal.

The application of ICA to natural images and movies yields at the output layer, single-neuron response histograms, which are considerably sparser than those in the input layer. These responses have Gabor-like RFs similar to those of simple cells in V1 (Olshausen et al. 1996, Bell & Sejnowski 1997, van Hateren & Ruderman 1998, van Hateren & van der Schaaf 1998, Simoncelli & Olshausen 2001, Hyvarinen 2010). ICA algorithms have also been applied to natural sounds (Lewicki 2002), yielding a set of temporal filters, resembling auditory cortical RFs.

Although the algorithms and results of ICA and source extraction by CS are often similar, there are important differences. First, CS results in signals that are truly sparse, i.e., most of the coefficients are zero, whereas ICA algorithms generally yield signals with many small values, i.e., distributions with high kurtosis but no coefficients vanish (Olshausen et al. 1996, Bell & Sejnowski 1997, Hyvarinen 2010). Second, ICA emphasizes the statistical independence of the unmixed sources (Barlow 2001). Sparseness is a special case; ICA can be applied to reconstruct dense sources as well. In contrast, signal extraction by CS relies only on the assumed approximate sparseness of the signal, and not on any statistical priors, and is similar in spirit to the seminal work of Olshausen et al. (1996). Indeed, a recent study suggests that sparseness may be a more useful notion than independence and that the success of ICA in some applications is due to its ability to generate sparse representations rather than to discover statistically independent features (Daubechies et al. 2009).

Beyond Linear Projections: Neuronal Nonlinearities

The abundance of nonlinearities in neuronal signaling raises the question of the relevance of the CS linear projections to neuronal information processing. One fundamental nonlinearity is the input-output relation between synaptic potentials and action potential firing of individual neurons. This nonlinearity is often approximated by the linear-nonlinear (LN) model (Dayan & Abbott 2001, Ostojic & Brunel 2011) in which the firing rate of a neuron, x, is related to its input activity $\mathbf{a}$ through $x = \sigma(\mathbf{a} \cdot \mathbf{s}^0)$, where $\mathbf{s}^0$ is the neuron's spatiotemporal linear filter and $\sigma(\cdot)$ is a scalar sigmoidal function. As long as $\sigma(\cdot)$ is an invertible function of its input, the nonlinearity in the measurement can be undone to recover the fundamental linear relation between the synaptic input to the neuron and the source, given by $\mathbf{As}^0$; hence, the results of CS should hold. More generally, it will be an important challenge to evaluate the role of dimensionality reduction, expansion, and sparse coding in neuronal circuit models that incorporate additional nonlinearities, including nonlinear temporal coding of inputs, synaptic depression and facilitation, and nonlinear feedback dynamics through recurrent connections.

In summary, we have reviewed a relatively new set of surprising mathematical phenomena related to RPs of high-dimensional patterns. But far from being a set of intellectual curiosities, these phenomena have important practical implications for data acquisition and analysis and important conceptual implications for neuronal information processing. It is likely that more surprises await us, lurking in the properties of high-dimensional spaces and mappings, properties that could further change the way we measure, analyze, and understand the brain.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

S.G. and H.S. thank the Swartz Foundation, Burroughs Wellcome Foundation, Israeli Science Foundation, Israeli Defense Ministry (MAFAT), the McDonnell Foundation, and the Gatsby Charitable Foundation for support, and we thank Daniel Lee for useful discussions.

LITERATURE CITED

Aharon M, ElRzad M, Bruckstein A. 2006a. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.* 54(11):4311

Aharon M, Elad M, Bruckstein A. 2006b. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebr. Appl.* 416(1):48–67

Atick J. 1992. Could information theory provide an ecological theory of sensory processing? *Netw. Comput. Neural Syst.* 3(2):213–51

Atick J, Redlich A. 1992. What does the retina know about natural scenes? *Neural Comput.* 4(2):196–210

Baraniuk R. 2007. Compressive sensing. *Signal Proc. Mag. IEEE* 24(4):118–21

Baraniuk R. 2011. More is less: signal processing and the data deluge. *Science* 331(6018):717–19

Baraniuk R, Cevher V, Wakin M. 2010. Low-dimensional models for dimensionality reduction and signal recovery: a geometric perspective. *Proc. IEEE* 98(6):959–71

Baraniuk R, Davenport M, DeVore R, Wakin M. 2008. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* 28(3):253–63

Baraniuk R, Wakin M. 2009. Random projections of smooth manifolds. *Found. Comput. Math.* 9(1):51–77

Barlow H. 1961. Possible principles underlying the transformation of sensory messages. *Sensory Commun.* pp. 217–34

Barlow H. 2001. Redundancy reduction revisited. *Netw. Comput. Neural Syst.* 12(3):241–53

Bayati M, Bento J, Montanari A. 2010. The LASSO risk: asymptotic results and real world examples. *Neural Inf. Process. Syst.* (*NIPS*)

Bell A, Sejnowski T. 1997. The independent components of natural scenes are edge filters. *Vis. Res.* 37(23):3327–38

Blum A. 2006. Random projection, margins, kernels, and feature-selection. In *Subspace, Latent Structure and Feature Selection*, ed. C Saunders, M Grobelnik, S Gunn, J Shawe-Taylor, pp. 52–68. Heidelberg, Germ.: Springer

Borghuis B, Ratliff C, Smith R, Sterling P, Balasubramanian V. 2008. Design of a neuronal array. *J. Neurosci.* 28(12):3178–89

Boyd S, Vandenberghe L. 2004. *Convex Optimization*. New York: Cambridge Univ Press

Bruckstein A, Donoho D, Elad M. 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. *Siam Rev.* 51(1):34–81

Brunel N, Hakim V, Isope P, Nadal J, Barbour B. 2004. Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* 43(5):745–57

Burges C. 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2(2):121–67

Candes E, Plan Y. 2010. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory* 57(11)7235–54

Candes E, Romberg J. 2007. Sparsity and incoherence in compressive sampling. *Invers. Probl.* 23(3):969–85

Candes E, Romberg J, Tao T. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* 59(8):1207–23

Candes E, Tao T. 2005. Decoding by linear programming. *IEEE Trans. Inf. Theory* 51:4203–15

Candes E, Tao T. 2006. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* 52(12):5406–25

Candes E, Wakin M. 2008. An introduction to compressive sampling. *IEEE Sig. Proc. Mag.* 25(2):21–30

Coskun A, Sencan I, Su T, Ozcan A. 2010. Lensless wide-field fluorescent imaging on a chip using compressive decoding of sparse objects. *Opt. Express* 18(10):10510–23

Coulter W, Hillar C, Isley G, Sommer F. 2010. Adaptive compressed sensing—a new class of self-organizing coding models for neuroscience. Presented at *IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 5494–97

Dai W, Sheikh M, Milenkovic O, Baraniuk R. 2009. Compressive sensing DNA microarrays. *EURASIP J. Bioinf. Syst. Biol.* 2009:162824

Dasgupta S, Gupta A. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms* 22(1):60–65

Daubechies I, Roussos E, Takerkart S, Benharrosh M, Golden C, et al. 2009. Independent component analysis for brain fMRI does not select for independence. *Proc. Natl. Acad. Sci.* 106(26):10415–20

Davenport M, Duarte M, Wakin M, Laska J, Takhar D, et al. 2007. The smashed filter for compressive classification and target recognition. *Proc. Comput. Imaging V SPIE Electron. Imaging*, San Jose

Dayan P, Abbott L. 2001. *Theoretical Neuroscience. Computational and Mathematical Modelling of Neural Systems*. Cambridge, MA: MIT Press

Donoho D. 2000. High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pp. 1–32

Donoho D. 2006. Compressed sensing. *IEEE Trans. Inf. Theory* 52(4):1289–306

Donoho D, Maleki A, Montanari A. 2009. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* 106(45):18914–19

Donoho D, Tanner J. 2005a. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* 102:9452–57

Donoho D, Tanner J. 2005b. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* 102:9446–51

Duarte M, Davenport M, Takhar D, Laska J, Sun T, et al. 2008. Single-pixel imaging via compressive sampling. *Signal Proc. Mag. IEEE* 25(2):83–91

Duarte M, Davenport M, Wakin M, Baraniuk R. 2006. Sparse signal detection from incoherent projections. *Proc. Acoust. Speech Signal Process.* (*ICASSP*) 3:III–III

Duarte M, Davenport M, Wakin M, Laska J, Takhar D, et al. 2007. Multiscale random projections for compressive classification. Presented at *IEEE Int. Conf. Image Process* (*ICIP*) *Int. Conf.* 6:VI161–64, San Antonio, TX

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Ann. Stat.* 32(2):407–99

Engel A, den Broeck CV. 2001. *Statistical Mechanics of Learning*. Cambridge Univ. Press, London

Ganguli S, Bisley J, Roitman J, Shadlen M, Goldberg M, Miller K. 2008a. One-dimensional dynamics of attention and decision making in lip. *Neuron* 58(1):15–25

Ganguli S, Huh D, Sompolinsky H. 2008b. Memory traces in dynamical systems. *Proc. Natl. Acad. Sci. USA* 105(48):18970–74

Ganguli S, Sompolinsky H. 2010a. Short-term memory in neuronal networks through dynamical compressed sensing. *Neural Inf. Process. Syst.* (*NIPS*) 23:667–75

Ganguli S, Sompolinsky H. 2010b. Statistical mechanics of compressed sensing. *Phys. Rev. Lett.* 104(18):188701

Gardner E. 1988. The space of interactions in neural network models. *J. Phys. A* 21:257–70

Haupt J, Castro R, Nowak R, Fudge G, Yeh A. 2006. Compressive sampling for signal classification. Presented at *Conf. Signals, Syst. Comput.* (*ACSSC*), 40th, Asilomar, pp. 1430–34

Hegde C, Wakin M, Baraniuk R. 2007. Random projections for manifold learning. *Neural Inf. Process. Syst.* **http://books.nips.cc/papers/files/nips20/NIPS2007_1100.pdf**

Hillar CJ, Sommer FT. 2011. Ramsey theory reveals the conditions when sparse coding on subsampled data is unique. *ArXiv* abs/1106.3616

Hodgkin A, Huxley A. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117:500–44

Hopfield J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* 79(8):2554–59

Hu T, Chklovskii D. 2009. Reconstruction of sparse circuits using multi-neuronal excitation (rescume). *Adv. Neural Inf. Proc. Syst.* 22:790–98

Hu T, Druckmann S, Chklovskii D. 2011. Early sensory processing as predictive coding: subtracting sparse approximations by circuit dynamics. *Front. Neurosci. Conf. Abs: COSYNE*

Hyvarinen A. 2010. Statistical models of natural images and cortical visual representation. *Top. Cogn. Sci.* 2:251–64

Indyk P, Motwani R. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. *Proc. Annu. ACM Symp. Theory Comput.*, 30th, pp. 604–13

Isely G, Hillar CJ, Sommer FT. 2010. Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. *Adv. Neural Inf. Proc. Syst.* (*NIPS*) 23:910–18

Jaeger H. 2001. Short term memory in echo state networks. *GMD Rep. 152*. Germ. Natl. Res. Cent. Inf. Technol., Bremen

Jaeger H, Haas H. 2004. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304(5667):78–81

Johnson W, Lindenstrauss J. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* 26(189–206):1–1

Kabashima Y, Wadayama T, Tanaka T. 2009. A typical reconstruction limit for compressed sensing based on l p-norm minimization. *J. Stat. Mech.* L09003

Kelly R, Smith M, Kass R, Lee T. 2010. Accounting for network effects in neuronal responses using l1 regularized point process models. *Neural Inf. Proc. Syst.* (*NIPS*) 23:1099–107

Kiani R, Esteky H, Mirpour K, Tanaka K. 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97(6):4296–309

Koulakov A, Rinberg D. 2011. Sparse incomplete representations: a novel role for olfactory granule cells. *Neuron* 72(1):124–36

Kreutz-Delgado K, Murray J, Rao B, Engan K, Lee T, Sejnowski T. 2003. Dictionary learning algorithms for sparse representation. *Neural Comput.* 15(2):349–96

Kriegeskorte N, Mur M, Ruff D, Kiani R, Bodurka J, et al. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60(6):1126–41

Lage-Castellanos A, Pagnani A, Weigt M. 2009. Statistical mechanics of sparse generalization and graphical model selection. *J. Stat. Mech.: Theory Exp.* 2009:P10009

Lee S, Ganapathi V, Koller D. 2006a. Efficient structure learning of markov networks using L1 regularization. *Neural Inf. Process. Syst.* (*NIPS*) 19:817–24

Lee S, Lee H, Abbeel P, Ng A. 2006b. Efficient l1 regularized logistic regression. *Proc. Natl. Conf. on Artif. Intell.* 21:401

Lewicki M. 2002. Efficient coding of natural sounds. *Nat. Neurosci.* 5(4):356–63

Linsker R. 1990. Perceptual neural organization: some approaches based on network models and information theory. *Annu. Rev. Neurosci.* 13(1):257–81

Lustig M, Donoho D, Pauly J. 2007. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.* 58(6):1182–95

Lustig M, Donoho D, Santos J, Pauly J. 2008. Compressed sensing MRI. *Signal Proc. Mag. IEEE* 25(2):72–82

Maass W, Natschlager T, Markram H. 2002. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* 14(11):2531–60

Machens C, Romo R, Brody C. 2010. Functional, but not anatomical, separation of what and when in prefrontal cortex. *J. Neurosci.* 30(1):350–60

Marr D. 1969. A theory of cerebellar cortex. *J. Physiol.* 202(2):437–70

Mishchenko Y. 2011. Reconstruction of complete connectivity matrix for connectomics by sampling neural connectivity with fluorescent synaptic markers. *J. Neurosci. Methods* 196(2):289–302

Olshausen B, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–9

Olshausen B, Field D. 1996. Natural image statistics and efficient coding. *Netw. Comput. Neural Syst.* 7(2):333–39

Olshausen B, Field D. 1997. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vis. Res.* 37(23):3311–25

Ostojic S, Brunel N. 2011. From spiking neuron models to linear-nonlinear models. *PLoS Comput. Biol.* 7(1):e1001056

Parrish T, Hu X. 1995. Continuous update with random encoding (cure): a new strategy for dynamic imaging. *Magn. Reson. Med.* 33(3):326–36

Perrinet L. 2010. Role of homeostasis in learning sparse representations. *Neural Comput.* 22(7):1812–36
Rajan K, Abbott L, Sompolinsky H. 2010. Stimulus-dependent suppression of chaos in recurrent neural networks. *Phys. Rev. E* 82(1):011903
Rogers T, McClelland J. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press
Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65(6):386–408
Rozell C, Johnson D, Baraniuk R, Olshausen B. 2008. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* 20(10):2526–63
Seung H, Sompolinsky H, Tishby N. 1992. Statistical mechanics of learning from examples. *Phys. Rev. A* 45(8):6056–91
Simoncelli E, Olshausen B. 2001. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24(1):1193–216
Smola A. 2000. *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press
Sussillo D, Abbott L. 2009. Generating coherent patterns of activity from chaotic neural networks. *Neuron* 63(4):544–57
Takhar D, Laska J, Wakin M, Duarte M, et al. 2006. A new compressive imaging camera architecture using optical-domain compression. *Proc. Comput. Imaging IV* 6065
Taraska J, Zagotta W. 2010. Fluorescence applications in molecular neurobiology. *Neuron* 66(2):170–89
Tarifi M, Sitharam M, Ho J. 2011. Learning hierarchical sparse representations using iterative dictionary learning and dimension reduction. *ArXiv* 1106:0357
Taubman D, Marcellin M, Rabbani M. 2002. Jpeg2000: image compression fundamentals, standards and practice. *J. Electron. Imaging* 11:286
Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* (*Methodol.*) 58:267–88
van Hateren J, Ruderman D. 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. Ser. B: Biol. Sci.* 265(1412):2315–20
van Hateren J, van der Schaaf A. 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. Ser. B: Biol. Sci.* 265(1394):359–66
Vapnik V. 1998. *Statistical Learning Theory*. New York: Wiley-Interscience
Wainwright M. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programming (LASSO). *Inf. Theory IEEE Trans.* 55(5):2183–202
Wainwright M, Ravikumar P, Lafferty J. 2007. High-dimensional graphical model selection using L1-regularized logistic regression. *Adv. Neural Inf. Proc. Syst.* 19:1465–72
Weiss Y, Chang H, Freeman W. 2007. Learning compressed sensing. Presented at *Allerton Conf.*, Urbana-Champaign
White O, Lee D, Sompolinsky H. 2004. Short-term memory in orthogonal neural networks. *Phys. Rev. Lett.* 92(14):148102–5
Wilt B, Burns L, Ho E, Ghosh K, Mukamel E, Schnitzer M. 2009. Advances in light microscopy for neuroscience. *Annu. Rev. Neurosci.* 32:435–506
Yu B, Cunningham J, Santhanam G, Ryu S, Shenoy K, Sahani M. 2009. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* 102(1):614–35
Zhou S, Lafferty J, Wasserman L. 2009. Compressed and privacy-sensitive sparse regression. *Inf. Theory IEEE Trans.* 55(2):846–66