Sparse sign-consistent Johnson–Lindenstrauss matrices: Compression with neuroscience-based constraints

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1419100111">http://dx.doi.org/10.1073/pnas.1419100111</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Dec 27 16:21:07 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/97241">http://hdl.handle.net/1721.1/97241</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Sparse sign-consistent Johnson–Lindenstrauss matrices: Compression with neuroscience-based constraints

Zeyuan Allen-Zhu, Rati Gelashvili, Silvio Micali, and Nir Shavit

Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

Contributed by Silvio Micali, October 7, 2014 (sent for review February 13, 2014)

Johnson–Lindenstrauss (JL) matrices implemented by sparse random synaptic connections are thought to be a prime candidate for how convergent pathways in the brain compress information. However, to date, there is no complete mathematical support for such implementations given the constraints of real neural tissue. The fact that neurons are either excitatory or inhibitory implies that every so implementable JL matrix must be sign consistent (i.e., all entries in a single column must be either all nonnegative or all nonpositive), and the fact that any given neuron connects to a relatively small subset of other neurons implies that the JL matrix should be sparse. We construct sparse JL matrices that are sign consistent and prove that our construction is essentially optimal. Our work answers a mathematical question that was triggered by earlier work and is necessary to justify the existence of JL compression in the brain and emphasizes that inhibition is crucial if neurons are to perform efficient, correlation-preserving compression.

1. Introduction

The existence of some form of compression in the brain is well accepted among neurobiologists. Its biological “evidence” proceeds from the brain’s numerous convergent pathways, where information coming from a large number of neurons must be compressed into a small number of axons or neurons. Classical examples are the optic nerve fibers that carry information about the activity of 100 times as many photoreceptors* (1) or the converge to the spinal cord (2).

As far back as 1961, Barlow (3) hypothesized that the role of early sensory neurons is to remove statistical redundancy in sensory input. This “efficient encoding” theory has been studied by many, as surveyed in depth by Simoncelli and Olshausen (4).

A recent survey by Ganguli and Sompolinsky (5) highlights the importance of compression and compressed sensing in the neural system for reducing the dimensionality of the activity pattern. A fundamental question they pose is “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?” They identify, as a minimal requirement, the importance of preserving the similarity structure of the neuronal representations at the source area, to capture the idea (6, 7) that in higher perceptual or association areas in the brain, semantically similar objects elicit similar neural activity patterns.

Ganguli and Sompolinsky suggest that such compression can be achieved in the brain via random synaptic-connectivity matrices implementing Johnson–Lindenstrauss (JL) matrices. However, because each neuron is either excitatory or inhibitory, an additional constraint, sign consistency, is necessary for this implementation to work.

In this paper, we show for the first time to our knowledge that JL matrices can be simultaneously compression efficient, sparse, and sign consistent and are thus implementable by biologically plausible neural networks.

JL Compression and Synaptic Connectivity. Informally, JL compression (8) uses a random matrix A to map a long vector of reals, x, the input, to a much shorter vector of reals, y = Ax, the JL output. The JL result shows that if the number of input vectors one may ever need to compress is reasonably upper bounded, then the following property is satisfied: Inner-product preservation. ⟨x, x′⟩ ≈ ⟨Ax, Ax′⟩ for all envisaged x and x′.

Note that inner-product preservation implies the aforementioned “similarity property” of biological interest; that is,

Correlation preservation.

That is, similar JL inputs correspond to similar JL outputs. The mentioned insight for implementing JL compression in the brain is random connectivity. An m ×d JL matrix A is biologically implemented via the synaptic connections among (the axons of) d “input” neurons and (the dendrites of) m < d “output” neurons. In essence,

Significance

Significant biological evidence indicates that the brain may perform some form of compression. To be meaningful, such compression should preserve pairwise correlation of the input data. It is mathematically well known that multiplying the input vectors by a sparse and fixed random matrix A achieves the desired compression. But, to implement such an approach in the brain via a synaptic-connectivity matrix, A should also be sign consistent: that is, all entries in a single column must be either all nonnegative or all nonpositive. This is so because most neurons are either excitatory or inhibitory. We prove that sparse sign-consistent matrices can deliver the desired compression, lending credibility to the hypothesis that correlation-preserving compression occurs in the brain via synaptic-connectivity matrices.

Author contributions: S.M. and N.S. designed research; Z.A.-Z. and R.G. performed research; Z.A.-Z. and R.G. contributed new reagents/analytic tools; and S.M. and N.S. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

1To whom correspondence should be addressed. Email: silvio@csail.mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1419100111/-/DCSupplemental.

‡Tan C (2013) From vision to memory to concept learning. Lecture, MIT CBCL-CSAIL Brains, Minds and Machines Seminar Series.

*As usual, |x⟩ represents the inner product of x and x′, that is, ⟨x, x′⟩; inner-product preservation immediately implies (and is in fact equivalent to) norm preservation: Namely, |x| ≈ |Ax| for all envisaged x.

‡As usual, |x⟩ represents the ℓ2 norm of x, that is, \(\sqrt{\sum_i x_i^2}\).
1.1. Three Interrelated Challenges.

Sign consistency. Let us explain the constraint in reference 16 of Rajan, Abbott, and Sompolinsky (10). According to Dale’s principle, almost all neurons have one of the following two types: excitatory or inhibitory, but not both. The type of a neuron essentially determines the “sign of the signal” it can transmit to a postsynaptic neuron. p. As a standard approximation, an excitatory neuron can only increase the activity of p, and an inhibitory one can only decrease it. Thus, a synaptic-connection matrix must be sign consistent (10, 11). That is, the nonzero entries of a column j must be either (i) all positive, if the jth input neuron is excitatory, or (ii) all negative, if n is inhibitory.

Unfortunately, typical JL matrices are not sign consistent. Sparsity. Let us emphasize another fundamental biological constraint: sparsity. A neuron may be connected to up to a few thousand postsynaptic neurons (12). (Furthermore, two neurons typically share multiple connections.) Thus, no synaptic-connectivity matrix could implement a dense JL matrix when m is large.3 As originally constructed, JL matrices were dense. Sparse JL matrices have been recently constructed (14–18), but they are far from being sign consistent. Therefore, although the sign consistency of synaptic action may have a few exceptions, the extent to which the above mathematical constructions may be biologically relevant is not clear.

Efficiency. As we mentioned at the start of Section 1, Introduction, implementing an m × d JL matrix in the brain is interesting only if m is significantly smaller than d. (Of course, achieving such efficiency is more challenging with sign consistency, but Rajan and Abbott (10) have expressed optimism about the general ability to satisfy the latter constraint.)

Three prior approaches. Let us explain why these three challenges have not been simultaneously met. A first and simplest way for JL matrices to be sign consistent is for them not to have any negative entries, corresponding to synaptic-connectivity matrices without inhibitory neurons. However, it is not hard to prove that nonnegative JL matrices must be extremely inefficient (e.g., m ≥ d/2 for typical choices of parameters; a Appendix D). This strong lower bound actually provides additional evidence for the cruciality of inhibition for neural functions.

The result of Rajan and Abbott (10) on the eigenvalue spectra of square matrices implies a way to transform JL matrices into sign-consistent ones (subject to mild assumptions on the inputs). However, the sign-consistent JL matrices they obtained were very dense: Half of their entries had to be nonzero (details in Section 2, Related Mathematical Work).

A third approach to sign-consistent JL matrices is implicitly provided by a transformation of Krahmer and Ward (19). Indeed, when applied to nonnegative matrices satisfying the restricted isometry property, their transformation yields sign-consistent JL matrices, but this construction can be proved to be much less efficient than ours.

3Although it may be easier to biologically construct a large random matrix, billions of years of evolution may not suffice for the emergence of a very special and very large matrix of neural connections. Moreover, this random construction need not be first found by evolution and then preserved genetically. That is, a good matrix A need not be the same across different individuals of the same species. It suffices that our DNA ensures that each individual, during development, randomly constructs his own matrix A.

4Moreover, even if m were small—e.g., m = 1,000—it seems hard to find in the brain a complete bipartite graph with d “inputs” and m “outputs” (ref. 13, p. 35).

1.2. Our Contributions. The mentioned biological constraints motivate the following purely mathematical question: How efficient can sparse (randomly constructed) and sign-consistent JL matrices be?

We answer this question exactly by providing tight upper and lower bounds.

We begin by formally stating the classical JL lemma (using norms rather than inner products):

Letting m = Θ(e^2 log(1/δ)), there exists a distribution A over m × d matrices such that, for any x ∈ ℝ^d, with probability at least 1 − δ, ∥Ax∥_2 = (1 ± δ)∥x∥_2.

The parameter ε measures the distortion introduced by the JL compression; in particular, one may consider ε = 10% (5). The parameter δ measures the confidence with which the 1 ± ε distortion is guaranteed. Because one may only need to compress polynomially many (rather than exponentially many) vectors in his lifetime, one typically chooses δ = 1/ poly(d) and thus log(1/δ) = O(log d). [With this choice of δ, after applying union bound, a matrix A generated according to A is capable of compressing poly(d) envisaged vectors from ℝ^d, with high confidence.]

We prove two main results:

Theorem 1. “A Construction of Sparse, Efficient, and Sign-Consistent JL Matrices.” That is, letting m = Θ(e^{-2 log(1/δ)}), there exists a distribution A over m × d sign-consistent matrices such that, for any x ∈ ℝ^d, with probability at least 1 − δ, ∥Ax∥_2 = (1 ± ε)∥x∥_2.

More precisely, a matrix A generated according to A enjoys the following properties:

• Sparsity: Each column has Θ(e^{-1 log(1/δ)}) nonzero entries.

• Same magnitude: All nonzero entries have the same absolute value.

• Simplicity: The positions of the nonzero entries in a column and the sign of a column itself are both randomly selected, independent from other columns.

Note that Theorem 1 shows the norm preservation up to a multiplicative error 1 ± ε. This implies correlation preservation up to an additive error ±O(ε); that is,

\[ \frac{\langle x, x' \rangle}{\|x\| \cdot \|x'\|} - \frac{\langle Ax, Ax' \rangle}{\|Ax\| \cdot \|Ax'\|} = O(\epsilon). \]

Note also that we cannot hope for a multiplicative error on correlation preservation because the correlation value is between −1 and 1, and thus a multiplicative error would imply the ability to recover orthogonal vectors (i.e., vectors with correlation zero) precisely.

Theorem 2. “Output-Length Optimality Among All Sign-Consistent JL Matrices.” That is, let A be a distribution over m × d sign-consistent matrices such that, for any x ∈ ℝ^d, with probability at least 1 − δ, ∥Ax∥_2 = (1 ± ε)∥x∥_2. Then, m = Ω(e^{-2 log(1/δ) · min \{log d, log(1/δ)\} \#}).

Note that in the interesting parameter regime of δ = 1/ poly(d), our lower bound becomes Ω(e^{-2 log^2(1/δ)}); that is, it essentially matches our upper bound.

1.3. In Sum. Our work closes an open mathematical question that is necessary to justify the existence of JL compression in the brain. Our work provides the missing support by constructing JL matrices that are simultaneously sparse and sign consistent and offer the most efficient JL compression possible; moreover, our work interestingly implies that inhibition is crucial if neurons are to perform efficient, correlation-preserving compression.

5Recall that the notation of Ω(N) signifies that logarithmic factors of N are ignored. Thus, in our case, factors of log(1/ε) and log(1/δ) are ignored in this lower bound.
Looking forward, the brain has inspired several models of computation, from perceptrons (20) to neural networks (21), which have already proved fruitful in many fields and in machine learning in particular.

Computer scientists have started studying computational models that are increasingly biologically relevant, for fundamental tasks such as concept representation and formation (22) and memory allocation (23–25). We consider our paper a further step in this direction.

2. Related Mathematical Work

To the best of our knowledge, the only mathematical analysis of a random sign-consistency matrix is the one suggested by Rajan and Abbott (10) and followed by ref. 11. Although their results are about the eigenvalue spectra of a random sign-consistent square matrix, it implies1 the following way of constructing an \( m \times d \) sign-consistent JL matrix \( A' \).

- First, construct an \( m \times d \) JL matrix \( A \), by randomly assigning each entry of \( A \) from \( \{-1/\sqrt{m}, 1/\sqrt{m}\} \).
- Second, construct an \( m \times d \) special matrix \( M \), by assigning each entry of (a random) half of the columns of \( M \) to be \(-1/\sqrt{m}\) and each entry of the remaining half to be \(1/\sqrt{m}\).
- At last, set \( A' = (1/\sqrt{2})(A + M) \).

Then, \( A' \) is sign consistent, and \( A'x = Ax \) (so \( A' \) is JL), assuming that \( x \) satisfies \( \sum_i x_i = 0 \). This assumption aside, however, the resulting matrix \( A' \) must be very dense.

The classical JL construction requires a distribution over dense matrices (e.g., i.i.d. Gaussian or Rademacher entries), but achieves a target dimension of \( m = \Theta(e^{-2}\log(1/\delta)) \), which is essentially optimal (26). A beautiful line of work (19, 27–31) has made use of the Hadamard or Fourier matrices in the JL construction, to speed up the matrix-vector multiplication to nearly linear time. However, their matrices are dense too. Recent constructions (14–18) yield sparse JL matrices that have \( \Theta(e^{-2}\log(1/\delta)) \) rows, which have been shown to be essentially optimal (32).

Although not applicable to JL matrices, Clarkson and Woodruff have shown how to construct sign-consistent and optimally sparse (namely, a single nonzero entry per column) random matrices (33). Their matrices preserve correlation for inputs satisfying an algebraic constraint, namely, coming from a hidden subspace. By contrast, we want to compress arbitrary inputs.

For numerous applications of JL compression in computer science, see refs. 34 and 35.

A JL matrix \( A \) can be easily constructed (with very high probability) by choosing each entry at random. Of course, such a randomly constructed matrix, it would be nice to reconstruct, with meaningful approximation, the original JL-compressed input \( x \) from \( Ax \); but this cannot be done without assuming that the inputs are of a restricted type [e.g., close to vectors with few nonzero entries (36, 37)]. However, even without reconstructing the inputs, inner-product preservation allows us to perform a variety of fundamental computations on the JL outputs directly and thus with great efficiency. This includes nearest neighbors (38), classification (39), regression (40), and many others.

3. A Simple Experimental Illustration

Let us consider a simple experiment to numerically verify the dependency \( m = \Theta(e^{-2}\log(1/\delta)) \) in the classical JL construction and the dependency \( m = \Theta(e^{-2}\log(1/\delta)) \) in our novel construction. Rather than fixing the distortion \( \varepsilon \) and computing the target dimension \( m \), we find it more convenient to fix \( m \) and \( \delta \) first and then compute \( \varepsilon \).

Specifically, we fix \( d = 3.000 \) and \( \delta = 0.1 \). Consider the following values of \( m \): \( m = 10, 20, 40, \ldots, 1,280 \); and then numerically compute the distortion \( \varepsilon \) for each value of \( m \), both for the classical and for our new JL construction.

In the classical JL construction, the \( m \times d \) dimension matrix \( A \) is chosen so that each entry is either \( 1/\sqrt{m} \) or \(-1/\sqrt{m} \), each with half probability.

In our construction, we first define the column sparsity \( s = \sqrt{m} \). [Note that this is consistent with the parameters suggested by Theorem 1: \( s = \Theta(e^{-1}\log(1/\delta)) = \Theta(\sqrt{m}) \).] Then, for each column of \( A \), we randomly choose \( s \) entries of this column and then flip a fair coin: If it is heads, we set each of these \( s \) entries to \( 1/\sqrt{s} \); if it is tails, we set each of them to \(-1/\sqrt{s} \).

For the above two constructions, we apply the JL transformation \( Ax \) on 1,000 randomly chosen inputs \( x_1, \ldots, x_{1,000} \in \mathbb{R}^d \). For each construction, we compute the 1,000 distortions \( \|Ax\|_2/\|x\|_2 \); call \( \varepsilon \) the 100th highest distortion, and plot \( \varepsilon \) in Fig. 1. (This process is equivalent to choosing \( \delta = 0.1 \) and throwing out the highest \( \delta \) fraction of the distortions. Indeed, \( 100 = 0.1 \times 1,000 \).)

The experiment illustrates that for both curves, whenever \( m \) is enlarged by a factor of 4, the error \( \varepsilon \) decreases approximately by a factor of 2. This corresponds to the dependency \( m \propto e^{-2} \) in both constructions. Also note that the blue curve falls slightly below the red curve, corresponding to the difference between the dimension choice of \( m = \Theta(e^{-2}\log(1/\delta)) \) in the classical JL construction and \( m = \Theta(e^{-2}\log(1/\delta)^2) \) in ours.

4. Proof Sketch of Theorem 1

Let \( A_{m,d} \) be the distribution of \( m \times d \) matrices defined as follows. For each of the \( d \) columns, we choose uniformly at random \( s \) distinct entries [of \( \binom{m}{s} \) possibilities] and assign a random value between \([-1/\sqrt{s}, 1/\sqrt{s}] \) (with half probability each) to these \( s \) entries, while leaving it zero in other entries of the same column.†

Theorem 1. Letting \( m = \Theta(e^{-2}\log^2(1/\delta)) \) and \( s = \Theta(e^{-1}\log(1/\delta)) \), for any \( x \in \mathbb{R}^d \), with probability at least \( 1 - \delta \), we have \( \|Ax\|_2 = (1 \pm \varepsilon)\|x\|_2 \) over the choice of \( A \sim A_{m,d} \).

The proof of Theorem 1 is quite complex and is given in SI Appendix A. (In particular, the classical technique of the Hanson-Wright inequality fails to give a tight upper bound in our case, just like ref. 18.) Below, we outline just the important ingredients of the proof.

According to the proof of Theorem 1, our theorem remains true if one divides each column into \( m/\sqrt{s} \) blocks and chooses one random entry from each block and/or if one uses \( \Theta(\log(1/\delta)) \)-wise independent hash functions to generate \( A_{m,d} \).

Fig. 1. The classical and our sign-consistent JL constructions.

---

1In fact, given any random matrix whose eigenvalues are randomly distributed on the complex unit disk, a random subset of its rows forms a JL matrix.

2**To be sure, inner-product preservation always implies a weak form of reconstructability. Namely, each entry \( x \) of an input vector \( x \) can be reconstructed up to an additive error of \( \|x\|_2 \).

3Our theorem remains true if one divides each column into \( m/\sqrt{s} \) blocks and chooses one random entry from each block and/or if one uses \( \Theta(\log(1/\delta)) \)-wise independent hash functions to generate \( A_{m,d} \).
Proof Sketch. Observe that the entries of a matrix $A \in \mathbb{R}^{m \times d}$ that we construct can be written as $A_{ij} = \eta_{ij}/\sqrt{s}$, where $\sigma_{ij} \in \{-1, 1\}$ is chosen uniformly at random, and $\eta_{ij} \in \{0, 1\}$ is an indicator variable for the event $A_{ij} \neq 0$. All of the $\{\sigma_{ij}\}_{i \in [d]}$ are independent; $\{\eta_{ij}\}_{i \in [d], j \in [m]}$ are independent across columns, but not independent (and in fact negatively correlated) in the same column, because there are exactly $s$ nonzero entries per column.

Given any fixed $x \in \mathbb{R}^d$ with $|x|^2 = 1$, let us study the following random variable:

$$Z \triangleq \sqrt{\|Ax\|^2 - 1} = \frac{1}{s} \sum_{r=1}^{m} \sum_{i \in [d]} \eta_{ir} \sigma_{ir} x_i x_r.$$

To show that $|Z| \leq \varepsilon$ with probability at least $1 - \delta$, we need a good upper bound on the th moment of $Z$ [note that we will eventually choose $t = \Theta(\log(1/\delta))$]:
\[
\sum_{j_1, \ldots, j_k} \sum_{v_i \in G} \frac{1}{\sigma} \prod_{i=1}^{d} \frac{d_i}{\sigma} \leq \sum_{j_1, \ldots, j_k} \frac{1}{\sigma} \prod_{i=1}^{d} \left( \frac{d_i}{\sigma} \right) \frac{v_i}{v_j}.
\]

getting a weaker upper bound as it replaces \( \left( \frac{v - 1}{v - 1} \right) \) with \( \left( \frac{v}{v_j} \right) \) in Lemma 2. However, even such a simple replacement leads to a \( \log(1/\delta) \) factor loss! Finally, after enduring layers of algebraic simplifications we prove the following:

**Lemma 3.** \( \text{dist}(\mathbb{E}[Z]) \leq 2^{O(1)} \sigma \left( \frac{v}{\sigma^2} \right) \).

By Lemma 3, there exists a constant \( C \) such that \( \text{dist}(\mathbb{E}[Z]) \leq (Ct/m)^{1/2} \). Using Markov’s inequality, we have \( \text{Pr}[\text{dist}(Z) > \epsilon] \leq \text{E}[\text{dist}(Z)]/\epsilon^t \leq (Ct/m)^{1/2} \). We now set parameters \( t = \log(1/\delta), \epsilon = \epsilon^{-1/2} \), and \( m \geq e^{2t/2C} \). Plugging them in we get \( \text{Pr}[\text{dist}(Z) > \epsilon] \leq \delta \) as desired, finishing the proof of Theorem 1.

5. **Statement of Theorem 2**

Here we formally state our second theorem and defer its full proof to SI Appendix B.

**Theorem 2.** There is some fixed \( \epsilon_0 \in (0, 1/2) \) such that for all \( \epsilon \in (1/\sqrt{d}, \epsilon_0) \), all \( m \leq O(d/\log(1/\epsilon)) \), and all \( \delta \leq \epsilon^{12} \), the following holds. Let \( A \) be a distribution over \( \mathbb{R}^m \) such that for all \( x \in \mathbb{R}^n \), with probability at least \( 1 - 1/\delta \), the \( \ell_2 \) embedding \( \|Ax\|_2 = (1 \pm \epsilon)\|x\|_2 \) has \( \epsilon \) distortion. Then,

\[
m \geq \Omega \left( e^{-2 \log(1/\delta)} \min \left( \log d, \log \frac{1}{\delta} \right) \right).
\]