45. Uncertainty principles and ideal atomic decomposition
Donoho, D.L., Huo, X. and others,, 2001. IEEE transactions on information theory, Vol 47(7), pp. 2845--2862. Citeseer.

46. Compressed sensing and best k-term approximation
Cohen, A., Dahmen, W. and DeVore, R., 2009. Journal of the American mathematical society, Vol 22(1), pp. 211--231.

47. A remark on compressed sensing
Kashin, B.S. and Temlyakov, V.N., 2007. Mathematical notes, Vol 82(5), pp. 748--755. Springer.

48. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting
Wainwright, M., 2007. 2007 IEEE International Symposium on Information Theory, pp. 961--965.

49. Lower bounds for sparse recovery
Do Ba, K., Indyk, P., Price, E. and Woodruff, D.P., 2010. Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms, pp. 1190--1197.

50. Neighborly polytopes and sparse solution of underdetermined linear equations
Donoho, D.L., 2005.

51. Compressed sensing: How sharp is the RIP
Blanchard, J.D., Cartis, C. and Tanner, J., 2009. SIAM Rev., accepted, Vol 10, pp. 090748160.

52. A deep learning approach to structured signal recovery
Mousavi, A., Patel, A.B. and Baraniuk, R.G., 2015. 2015 53rd annual allerton conference on communication, control, and computing (Allerton), pp. 1336--1343.

53. Learned D-AMP: Principled neural network based compressive image recovery
Metzler, C., Mousavi, A. and Baraniuk, R., 2017. Advances in Neural Information Processing Systems, Vol 30.

54. Compressed Sensing using Generative Models [HTML]
Bora, A., Jalal, A., Price, E. and Dimakis, A.G., 2017. Proceedings of the 34th International Conference on Machine Learning, Vol 70, pp. 537--546. PMLR.

55. Average firing rate rather than temporal pattern determines metabolic cost of activity in thalamocortical relay neurons
Yi, G. and Grill, W., 2019. Scientific reports, Vol 9(1), pp. 6940. DOI: 10.1038/s41598-019-43460-8

56. Distributed representations
Plate, T., 2003. Cognitive Science, pp. 1-15.

57. Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis [link]
Ganguli, S. and Sompolinsky, H., 2012. Annual Review of Neuroscience, Vol 35(1), pp. 485-508. DOI: 10.1146/annurev-neuro-062111-150410

Nonlinear Compression

This paper focuses on the assumption that representations are linear. But what if models don't use linear feature directions to represent information? What might such a thing concretely look like?

Neural networks have nonlinearities that make it theoretically possible to compress information even more compactly than a linear superposition. There are reasons we think models are unlikely to pervasively use nonlinear compression schemes:

- The model needs to decompress things before it can compute with them naturally: Most of the computation in the model is linear, so this kind of compression is likely only worth it to save space in the residual stream across many layers before being decompressed to be computed with linearly again.
- They're probably difficult to learn: Nonlinear compression schemes may require finely tuned approximations of discontinuities, and for the compression and decompression to line up, and may be difficult for gradient descent to learn.
- They probably take enough neurons that the benefit over superposition isn't worth it:
    - Representing the piecewise linear functions in the simple example with ReLU neurons using the universal function approximation result that each line segment takes two neurons, would require 12 neurons per Z segment, so only starts to beat linear compression at a combined 36 neurons for the compression and decompression.
    - This comparison has only one hidden dimension and dense features, which is somewhat of a degenerate case for superposition. Superposition is much more powerful for compression of sparse features in many dimensions. We suspect in large models the scaling is in favor of superposition, although this is just intuition and it's possible that scaling nonlinear compression is competitive.