EFFICIENT MACHINE LEARNING USING RANDOM PROJECTIONS

Chinmay Hegde,[r] *Mark A. Davenport,*[r] *Michael B. Wakin*[m] *and Richard G. Baraniuk* [r]

[r] Department of Electrical and Computer Engineering, Rice University

[m] Department of Electrical Engineering and Computer Science, The University of Michigan at Ann Arbor

ABSTRACT

As an alternative to cumbersome nonlinear schemes for dimensionality reduction, the technique of random linear projection has recently emerged as a viable alternative for storage and rudimentary processing of high-dimensional data. We invoke new theory to motivate the following claim: the random projection method may be used in conjunction with standard algorithms for a multitude of machine learning tasks, with virtually no degradation in performance. Thus, random projections can been shown to result in both significant computational savings and provably good performance..

1. BACKGROUND AND MOTIVATION

The unimpeded growth in the size of datasets generated by signal acquisition systems (e.g., sensor networks, 3D imaging systems) poses a significant challenge to machine learning algorithms. This effect – frequently referred to as the "curse of dimensionality" – usually forces an algorithm designer to sacrifice accuracy in order to make the problem computationally feasible. Luckily, in many cases we can avoid this difficult decision. Suppose our dataset X consists of points $x \in \mathbb{R}^N$. Often, points in X, although N-dimensional, can be described using some model with only K pieces of information, where $K \ll N$. In these cases, we would like to be able to obtain, store, and work with a K-dimensional representation of the dataset, as opposed to handling the original N-dimensional dataset.

Approach 1. Construct data-adaptive nonlinear mappings (see [1, 2] among many others) that transform the data into a low-dimensional representation while preserving certain desirable properties. Such algorithms attempt to construct an embedding $f : \mathbb{R}^N \rightarrow \mathbb{R}^K$ that maps elements of X to $\mathbb{R}^K$, where K is the intrinsic dimension of the dataset. The mapping f is invariably data dependent; often, the dependence is global [1] (implying that the low-dimensional representations can be obtained only after processing *every* element of X.)

Approach 2. Compute a non-adaptive linear projection of the N-dimensional dataset into a *random* M-dimensional subspace of $\mathbb{R}^N$. In this case, the mapping f can be represented as an $M \times N$ matrix Φ where the entries of Φ are independently drawn from a specified probability distribution. The simplicity of this dimensionality reduction procedure is striking; it is clear that the mapping is data independent, and the process of obtaining the image of any given data vector x under the mapping Φ is a stand-alone computation. In addition, the powerful Johnson-Lindenstrauss (JL) Lemma [3] guarantees that, provided $M = O(\log |X|)$, then there exists an $\epsilon \in (0, 1)$ such that

$$(1 - \epsilon)\|x - y\| \leq \|\Phi x - \Phi y\| \leq (1 + \epsilon)\|x - y\| \qquad (1)$$

for all $x, y \in X$. In other words, the distance between any pair of points is approximately preserved by the mapping Φ.

Either approach has its own share of advantages and disadvantages. While nonlinear dimensionality reduction techniques adaptively construct the most parsimonious representation of X, they are expensive to implement when $|X|$ is large or when X is high-dimensional. At the same time, despite its easy implementation, the random projection approach described above fails to consider the geometric inter-relationships between data points, and merely depends on the total *number* of input data vectors, which could potentially be large or even infinite. Thus, random projections would seem to be an inefficient dimensionality reduction technique when our dataset is very large but has a relatively small *intrinsic dimension*. The key to breaking this impasse is to realize that we are not operating on an *arbitrary* set of points, and for structured data sets the bound on M from the JL Lemma can be extremely pessimistic.

2. AN OPTIMISTIC LOOK AT RANDOM PROJECTIONS

The concept of random projections has generated renewed interest in recent years. Data stream algorithms employ the random projections approach in the form of *sketches* [4] for tasks like histogram maintenance and ℓ_p-norm computation. More recently, random projections have been exploited in the field of *compressive sensing* [5, 6], which deals with efficient acquisition and recovery of sparse signals. In fact, efforts have been made to develop inexpensive hardware realizations [7, 8] that *directly acquire* random projections of analog signals, and thus the computational cost of applying Φ to the data can be essentially zero. In cases where this is not possible, fast algorithms for computing random projections are also being explored [9].

We now observe that both data stream algorithms and compressive sensing employ random projections for acquiring a compressed representation of signals which are governed by *low-complexity models*. Thus, Approach 2 is being utilized in a scenario that is seemingly appropriate for the algorithms described in Approach 1. In particular, random projections are being used *not* to embed a set of points into a lower dimension, but rather to preserve the geometric structure of low-complexity signal classes. However, there is a close relationship between the two approaches. For the case of compressive sensing, the precise connection between the JL Lemma and the preservation of the geometry of the class of sparse signals was only recently established [10]. In particular, the same techniques used to establish the JL Lemma can be exploited to show that provided $M = O(K \log N)$, there exists an $\epsilon \in (0, 1)$ such that (1) holds for any x, y that are *K-sparse* (i.e., have at most K non-zero entries). The important difference is that (1) now describes a stable embedding of infinitely many points, namely, the set of *all* K-sparse signals.

It is possible to use the same techniques to consider other low-complexity signal models. For example, we may consider the random projection of a K-dimensional compact manifold $\mathcal{M}$ residing

in $\mathbb{R}^N$ into $\mathbb{R}^M$ [11]. In particular, if $M = O(K \log N)$, (1) now holds for *every* $x, y \in \mathcal{M}$; further, the set of *all* pairwise geodesic distances between points on the manifold are preserved with small distortion.

Notice that in both cases the dimension of the embedding subspace now depends only on the *complexity* (or information content K) of the dataset, not on its *cardinality*. Also, the dependence of M on N is only logarithmic, and hence we see that $K < M \ll N$. Thus, by a simple random projection operation, we obtain a compressed representation of a sparse or manifold-modeled point cloud, while preserving the geometric structure of the dataset.

3. LEARNING IN THE COMPRESSED DOMAIN

Obtaining a low-dimensional representation of a dataset is an important component in the machine learning process. However, it is usually not the ultimate goal. For instance, suppose the objective is to perform a nonlinear inference task (involving some form of detection/classification/estimation), with the given data as the input.

Let $\mathcal{L}$ denote some machine-learning algorithm tailored to the problem we wish to solve. Our claim is as follows: for a wide variety of machine-learning algorithms $\mathcal{L}$, the performance of $\mathcal{L}$ when given access to only a randomly projected (i.e., M-dimensional) version of X is *essentially the same as* its performance on the original dataset X. The implications of this are significant; this implies that the machine is oblivious to whether it works with the original data, or with only a low-dimensional, easily obtainable representation. In other words, random projections can be used as a *universal, inexpensive* preprocessing step to almost any machine learning task. Further, in compressive sensing the data is *directly acquired* in the form of low-dimensional random projections; in fact, recovering the original high-dimensional data points expends considerable computational resources. Thus, in this setting the random projections approach to machine learning could lead to tremendous savings in processing and memory costs incurred during the learning process. In [12], the above claim is rigorously proved for two special cases: 1) when $\mathcal{L}$ is the Grassberger-Procaccia algorithm for estimating intrinsic dimension of a point cloud; 2) when $\mathcal{L}$ is the Isomap algorithm for nonlinear dimensionality reduction of Euclidean manifolds.

We reinforce the claim by presenting results on the performance of binary classification algorithms using compressive projections. Suppose that a network of compressive imaging cameras [7] with resolution N are observing a scene, with $\mathcal{L}$ being the task of performing automatic target recognition among P classes (each class being modeled by a K-dimensional manifold). The classifier is assumed to be provided with a sufficient number of labeled (training) points from each of the P classes. This is dubbed as the "smashed filter" [13], and it can be proved that the number of measurements required per sample point is given by $M = O(K \log(NP))$. Again, only a very small number of measurements (relative to the original resolution N) are required for reliable classification performance.

As a final example, let $\mathcal{L}$ be the task of determining the class of an object present in a smoothly varying video sequence. The smashed filter implements a nearest-neighbor (NN) solution for the multi-class labelling problem. In cases where the training data is derived from an insufficient sampling of the underlying manifolds, this might yield incorrect results. To handle such situations, we develop a manifold learning-based algorithm that exploits the smooth geometric structure of the *unlabeled* data points to make more robust decisions as compared to naive NN based classification followed by majority voting. Next, we use Theorem 3.2 of [12] to claim that the algorithm works provably well under random projections.

4. FUTURE DIRECTIONS

Performing learning tasks in the compressed domain certainly seems attractive. Yet, there are several unanswered questions. It is not precisely clear how the presence of data noise affects the learning performance in the compressed domain; machine learning with irregularly sampled point clouds is in general a hard problem; we would like to possess an overarching framework which unifies the analysis of different kinds of manifold learning algorithms on compressive measurements; and finally, the bounds on the minimum number of measurements are not tight. Answering these questions will lead to the development of a coherent theory and even broader applications.

5. REFERENCES

[1] J. B. Tenenbaum, V.de Silva, and J. C. Landford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[2] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[3] W. B Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Proc. Conf. in Modern Analysis and Probability*, 1984, pp. 189–206.

[4] P. Indyk, "Stable distributions, pseudorandom generators, embeddings and data stream computation," *Journal of the ACM*, 2006.

[5] D.L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, September 2006.

[6] E.J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[7] M.B. Wakin, J.N. Laska, M.F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K.F. Kelly, and R.G. Baraniuk, "An architecture for compressive imaging," in *IEEE International Conference on Image Processing (ICIP)*, Atlanta, GA, Oct. 2006, pp. 1273–1276.

[8] J.N. Laska, S. Kirolos, M.F. Duarte, T. Ragheb, R.G. Baraniuk, and Y. Massoud, "Theory and implementation of an analog-to-information conversion using random demodulation," in *Proc. IEEE Int. Symposium on Circuits and Systems (ISCAS)*, New Orleans, LA, May 2007, To appear.

[9] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast johnson-lindenstrauss transform," in *Sym. Theory of Computing (STOC)*. Seattle, WA, May 2006.

[10] R.G. Baraniuk, M.A. Davenport, R.A. DeVore, and M.B. Wakin, "A simple proof of the restricted isometry property for random matrices," 2006, To appear in *Constructive Approximation*.

[11] R.G. Baraniuk and M.B. Wakin, "Random projections of smooth manifolds," 2007, To appear in *Found. Computational Mathematics*.

[12] C. Hegde, M.B. Wakin, and R.G. Baraniuk, "Random projections for manifold learning," in *Neural Information Processing Systems (NIPS)*, 2007.

[13] M.A. Davenport, M.F. Duarte, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, and R.G. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proc. IS&T/SPIE Sym. on Electronic Imaging: Computational Imaging*, 2007, vol. 6498.