Fast Training of Graph-Based Algorithms for Nonlinear Dimensionality Reduction

Max Vladymyrov and Miguel A. Carreira-Perpiñán
EECS, University of California, Merced

Introduction Dimensionality reduction algorithms have long been used either for exploratory analysis of a high-dimensional dataset, to reveal structure such as clustering, or as a preprocessing step, by extracting low-dimensional features that are useful for classification or other tasks. Here we focus on dimensionality reduction algorithms where a dataset consisting of N objects is represented by a weighted graph, rather than by a high-dimensional feature vector for each object. One type of these methods are spectral methods such as Laplacian eigenmaps [1], Isomap and others, which have become extremely popular in the last 10 years because of their ease of implementation. Their objective function is quadratic (and subject to quadratic constraints), so that the minimum can be obtained by solving an eigenproblem using numerical linear algebra routines. Using a cluster of several hundred machines and various approximations, it is possible to deal with millions of points with Laplacian eigenmaps [2]. Spectral methods can sometimes find a good low-dimensional representation of a nonlinear dataset, but they severely distort the data in the presence of noise, variations in density, or when it contains multiple manifolds. Recently, several methods have been developed that define a non-quadratic objective function, and which are able to achieve much better low-dimensional representations than linear or spectral methods. These methods include Stochastic Neighbor Embedding (SNE; [3]), t-SNE [4] and the Elastic Embedding (EE; [5]). However, their practical applicability has been limited to small datasets so far because of the slowness of its optimization. Current implementations have training times of the order of hours even for datasets of a few thousand points. Our goal is to develop fast training algorithms that can allow interactive use with small or medium datasets (for example, to update incrementally an existing embedding as new high-dimensional data is added to it or removed from it), and relatively short times with large datasets (as well as the ability to create a quick "preview" of a large dataset and then refine it as needed).

General formulation of embedding algorithms and fast optimization One first step towards this goal is using partial-Hessian algorithms [6]. We consider an objective function $E(\mathbf{X};\lambda) = E^+(\mathbf{X}) + \lambda E^-(\mathbf{X})$ where $\mathbf{X}$ is the desired set of low-dimensional coordinates (an L-dimensional vector for each of the N objects, e.g. $L=2$ for 2D visualization). The terms $E^+(\mathbf{X})$ and $E^-(\mathbf{X})$ are defined from the weighted graph of the dataset. They ensure that the projections of similar objects are close and the projections of dissimilar objects are distant, respectively, and the parameter λ controls the relative importance of each of these two pieces of information. This generalized formulation includes as particular cases existing algorithms, such as SNE, t-SNE and EE, Laplacian eigenmaps, and also suggests new algorithms which, while still producing high-quality low-dimensional representations, can simplify the optimization considerably. It also can be extended to algorithms that embed different data types, such as authors and documents they write, in the same low-dimensional space.

Previously, the optimization has been based on gradient descent (with improvements such as a momentum term) [3] or a fixed-point iteration [5]. While the large number of parameters prevents computing the Hessian, in [6] we show that it is possible to identify important parts of the Hessian that are sparse and positive definite, so that each iteration makes much more progress at a cost similar to that of the gradient, and global convergence is guaranteed. In particular, the use of a *spectral direction* combined with caching a (sparsified) Cholesky factorization strikes the best compromise between ease of implementation, robustness to parameters and fastest overall runtime. It achieves speedups of 1–2 orders of magnitude over previous methods. Fig. 1 shows a 2D embedding of $N = 20\,000$ MNIST digits obtained in just 15 minutes in a workstation, while the original optimization techniques show very little progress.

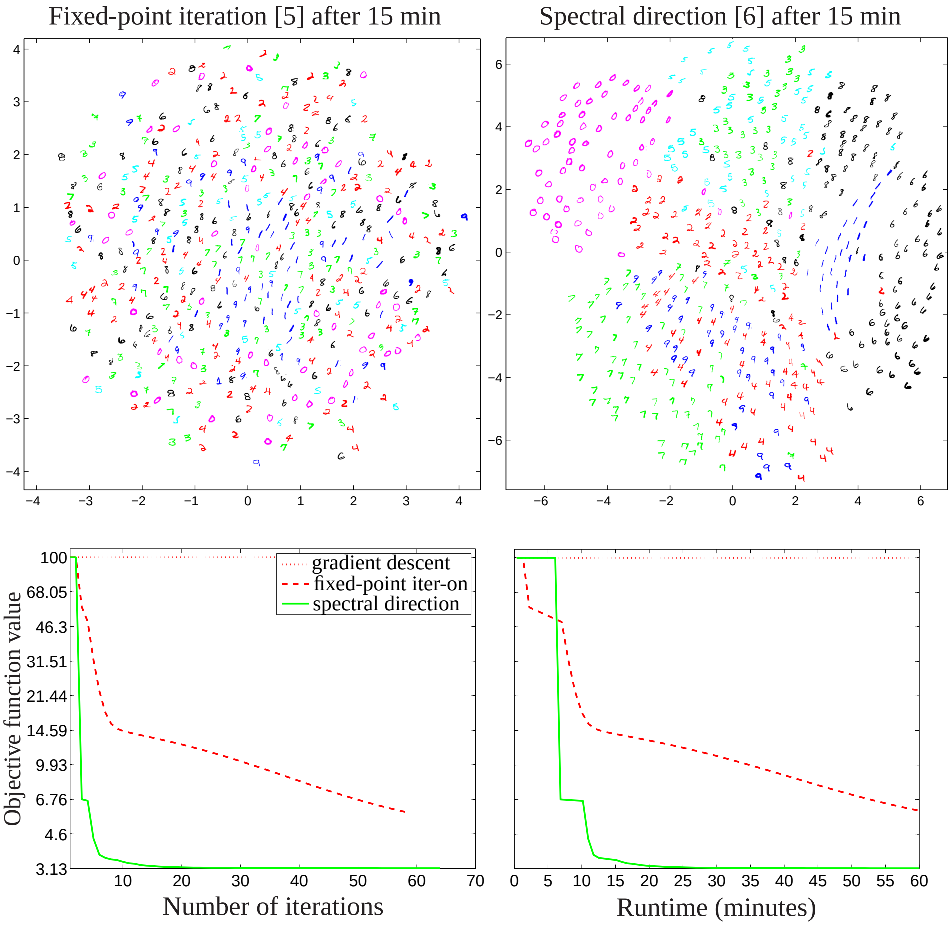

Figure 1: Comparison of optimization techniques for an MNIST data subset of $N = 20\ 000$ handwritten digit images of size 28×28 (784 dimensions), projecting to a 2D low-dimensional space using the elastic embedding (EE). *Top*: resulting embedding after 15 minutes for the fixed-point iteration method of [5] (left) and our spectral direction (right). *Bottom*: decrease of the objective function in number of iterations (left) and run time (right). Gradient descent [3] barely decreases the objective function.

Current research directions While the spectral direction reduces the number of iterations necessary for convergence, each iteration is still quadratic on the number of points N (since the objective function has $\mathcal{O}(N^2)$ terms) and therefore does not scale. We are exploring additional optimization techniques, subsampling, the use of parallel computation, and the use of fast multipole methods [7] to approximate the kernel sums in each iteration in linear time.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[2] A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *CVPR*, 2008.

[3] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *NIPS*, 2003.

[4] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 2008.

[5] M. Á. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In *ICML*, 2010.

[6] M. Vladymyrov and M. Á. Carreira-Perpiñán. Partial-Hessian strategies for fast learning of nonlinear embeddings. In *ICML*, 2012.

[7] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comp. Phys.*, 1987.