Learning low rank matrices online using retracions

Uri Shalit [1,2] Daphna Weinshall [2] Gal Chechik [1,3]

[1] Bar Ilan University, Israel [2] Hebrew University, Israel [3] Google, CA

Many learning problems involve models represented in matrix form. These include metric learning, collaborative filtering, and multi-task learning where all tasks operate over the same set of features. In many of these models, a natural way to regularize the model is to limit the rank of the corresponding matrix. In metric learning, a low rank constraint allows to learn a low dimensional representation of the data in a discriminative way. In multi-task problems, low rank constraints provide a way to tie together different tasks. In all cases, low-rank matrices can be represented in a factorized form that dramatically reduces the memory and run-time complexity of learning and inference with that model. Low-rank matrix models could therefore scale to handle substantially many more features and classes than with full rank dense matrices.

As with many other problems, the rank constraint is non-convex, and in the general case, minimizing a convex function subject to a rank constraint is NP-hard. As a result, two main approaches have been commonly used. Sometimes, a low-rank matrix $W \in \mathbb{R}^{n\times m}$ is represented as a product of two low dimension matrices $W = AB^T, A \in \mathbb{R}^{n\times k}, B \in \mathbb{R}^{m\times k}$ and simple gradient descent techniques are applied to each of the product terms. Second, projected gradient algorithms can be applied by repeatedly taking a gradient step and projecting back to the manifold of low-rank matrices. Unfortunately, computing the projection to that manifold becomes prohibitively costly for large matrices and cannot be computed after every gradient step.

In this work we describe a new algorithms for online learning on low-rank manifolds. It is based on an operation called *retraction*, which maps from a vector space that is tangent to the manifold, into the manifold. The common Euclidean projection operator is a special case of a retraction, which can be shown to be a second order approximation of the ideal geodesic flow along the manifold towards the minimum.

Importantly, our algorith, LORETA, is based on a retraction that can be computed very efficiently. When the gradient steps are rank-1, as commonly happen in online learning, the retraction can be computed in a time that is linear in the size of A and B, with small constants.

We test Loreta in two different domains and learning tasks. First, we learn a bilinear similarity measure among pairs of text documents, where the number of features (text terms) representing each document could become very large. Loreta performed better than other techniques that operate on a factorized model, and also improves retrieval precision by 33% as compared with training a full rank model over pre-selected most informative features, using comparable memory footprint. Second, we applied Loreta to image multi-label ranking, a problem in which the number of classes could grow to millions. Loreta significantly improved over full rank models, using a fraction of the memory required. These two experiments suggest that low-rank optimization could become very useful for learning high-dimensional problems.

Presenter: Gal Chechik Preference: poster