Summary of Thesis: Non-convex Optimization for Machine Learning: Design, Analysis, and Understanding

Tengyu Ma

October 15, 2018

Non-convex optimization is ubiquitous in modern machine learning: recent breakthroughs in deep learning require optimizing non-convex training objective functions; problems that admit accurate convex relaxation can often be solved more efficiently with non-convex formulations. However, the theoretical understanding of non-convex optimization remained rather limited.

The thesis developed novel tools for analyzing and designing non-convex optimization algorithms and apply them to various machine learning problems including sparse coding, topic models, matrix completion, linear dynamical systems, and word embeddings.

The first part of the thesis proposes a framework for analyzing non-convex algorithms that start from a coarse initialization. The framework is used to analyze sparse coding (dictionary learning), which is one of the first results on the convergence of alternating minimization algorithms. The result gives the first efficient algorithm for sparse coding that works almost up to the information theoretic limit for sparse recovery on incoherent dictionaries. All previous algorithms that approached or surpassed this limit run in time exponential in some natural parameter. The algorithm also gives the state-of-the-art sample complexity guarantees for sparse coding. The proposed framework has been adopted to analyze algorithms for various ML problems after its first publication (http://arxiv.org/abs/1503.00778) in COLT 2016.

The second part of the thesis takes a geometric viewpoint: optimization is feasible when all the local minima of the non-convex function are also global minima. Based on the conference paper (http://arxiv.org/abs/1605.07272) (which wins the NIPS 2016 best student paper award), we show that such a surprising property holds for the non-convex formulation of the matrix completion problem, explaining why empirically it reliably outperforms convex relaxation.

The main technical idea of the analysis is to prove the empirical risk from the noisy observations is sufficiently concentrated around the population risk in terms of not only the value but also the geometric properties. It's crucial to prove such concentration for just a minimal set of geometric properties so that we achieve the correct dependency on the dimension.

The thesis also shows that such geometric property holds for the non-convex objective for learning linear dynamical systems (under structural assumptions on the data.) A milder assumption is needed if the model is over-parameterized (with more parameters than statistically necessary), which provides another evidence that over-parameterization can improve the loss surface and make optimization easier. The result is based on the JMLR paper Gradient Descent Learns Linear Dynamical Systems (https://arxiv.org/abs/1609.05191).

The last part of the thesis lays the foundations of word embeddings in natural language processing, summarizing two TACL papers RAND-WALK: A Latent Variable Model Approach to Word Embeddings (https://arxiv.org/abs/1502.03520) and Linear Algebraic Structure of Word Senses, with Applications to Polysemy https://arxiv.org/abs/1601.03764. Empirically word embeddings are successfully trained with several different non-convex objectives but their workings were not well understood. Towards understanding word embeddings, we propose a new generative model, a dynamic version of the log-linear topic model of Mnih and Hinton (2007). We show that word2vec and GloVe are two different learning algorithms for learning the generative model and that the learned word vectors can solve analogy tasks (under a definition of analogy based on word-word co-occurrence matrix.) This provides a theoretical justification for nonlinear models like PMI, word2vec, and GloVe, as well as some hyperparameter choices. It also helps explain why low-dimensional semantic embeddings contain linear algebraic structure that allows solution of word analogies, as shown by Mikolov et al. (2013a) and many subsequent papers. This generative model and its variants are later used to design new embeddings for senses, sentences, and documents.