# Provable Algorithms for Nonlinear Models in Machine Learning and Signal Precessing

## ABSTRACT

In numerous signal processing and machine learning applications, the problem of signal recovery from a limited number of nonlinear observations is of special interest.

These problems also called inverse problem have recently received attention in signal processing, machine learning, and high-dimensional statistics. The inverse problems are inherently ill-posed since the number of measurements are typically less than the number of unknowns. As a result, one needs to assume some structures on the underlying signal such as sparsity, structured sparsity, low-rank and so on. In addition, having a nonlinear map from the signal space to the measurement space may add more challenges to the problem. For instance, the assumption on the nonlinear function such as known/unknown, invertibility, smoothness, even/odd, and so on can change the tractability of the problem dramatically. The nonlinear inverse problems are also a special interest in the context of neural network and deep learning as each layer can be casted as an instance of the inverse problem. Thus, understanding of an inverse problem can serve as a building block for more general and complex networks. In this thesis, we focus on the signal model, the nonlinear function, and the connection of inverse problems to the analysis of some specific neural networks.

First, we start with the superposition signal model in which the underlying signal is assumed to be the superposition of two components with sparse representation (i.e., their support is arbitrary sparse) in some specific domains. Initially, we assume that the nonlinear function also called link function is not known. Then, the goal is defined as recovering the components of the superposition signal from the nonlinear observation model. This problem which is called signal demixing is of special importance in several applications ranging from astronomy to computer vision. Our first contribution is a simple, fast algorithm that recovers the component signals from the nonlinear measurements. We support our algorithm with rigorous theoretical analysis and provide upper bounds on the estimation error as well as the sample complexity of demixing the components (up to a scalar ambiguity). Next, we remove the assumption on the link function and studied the same problem when the link function is known and monotonic, but the observation is corrupted by some additive noise. We proposed an algorithm under this setup for recovery of the components of the superposition signal, and derive nearly-tight upper bounds on the sample complexity of the algorithm to achieve stable recovery of the components. Moreover, we showed that the algorithm enjoys a linear convergence rate. Chapter 1 includes this part.

In chapter 2, we target two assumptions made in the first chapter: the first assumption which is concerned about the underlying signal model considers the case that the constituent components have arbitrary sparse representations in some incoherent domains. While having arbitrary sparse support can be a good way of modeling of many natural signals, it is just a simple and not realistic assumption. Many real signals such as natural images show some specific structure on their support. That is, when they are represented in a specific domain, their support comprises non-zero coefficients which are grouped or classified in a specific pattern. For instance, it is well-known that many natural images show so-called *tree sparsity* structure when they are represented in the wavelet domains. This motivates us to study other signal models in the context of our demixing problem introduced in chapter 1. In particular, we study certain families of structured sparsity models in the constituent components and propose a method which provably recovers the components given (nearly) $\mathcal{O}(s)$ samples where $s$ denotes the sparsity level of the underlying components. This strictly improves upon previous nonlinear demixing techniques and asymptotically matches the best

possible sample complexity. The second assumption, we made in the first chapter is about having a smooth monotonic link function when the link function is known. In chapter 2, we go beyond this assumption, and we study the bigger class of nonlinear link functions and consider the demixing problem from a limited number of nonlinear observations where this nonlinearity is due to either periodic function or aperiodic one. For both of these considerations, we propose new robust algorithms and equip them with statistical analysis.

In chapter 3, we continue our investigation about choosing a proper underlying signal model in the demixing framework. In chapters 1 and 2, our approach is based on *hard-coded* approach. That is, we assume some prior knowledge in the signal domain and exploit the structure of this prior in designing efficient algorithms. However, many real signals including natural images have a more complicated structure than just simple sparsity (arbitrary or structured). Towards choosing a proper structure, some research works try to automate the process of choosing prior knowledge on the underlying signal models by using deep learning. This line of research considers selecting the structure of the signal as a representation learning problem. Given the success of deep learning in some recent works, in chapter 3, we use deep learning techniques to model the low-dimension structure of the constituent components and consequently, estimating these components from their superposition. Our approach in this chapter is empirical, and we defer more theoretical investigation of this approach as our future direction.

In chapter 4, we study other low-dimension signal models. In particular, we focus on the common low-rank matrix as our signal model. In this case, our interest signal to estimate (recover) is a low-rank matrix. Specifically, we formulate the general problem of optimizing a convex function over the set of matrices, subject to rank constraints. Recently, different algorithms have been proposed for the low-rank matrix estimation problem. However, existing first-order methods for solving such problems either are too slow to converge, or require multiple invocations of singular value decompositions.

On the other hand, factorization-based non-convex algorithms, while being much faster, and has a provable guarantee, require stringent assumptions on the condition number of the optimum. Here, we provide a novel algorithmic framework that achieves the best of both worlds: as fast as factorization methods, while requiring no dependency on the condition number. We instantiate our general framework for three important and practical applications; nonlinear affine rank minimization (NLARM), Logistic PCA, and precision matrix estimation (PME) in the probabilistic graphical model. We then derive explicit bounds on the sample complexity as well as the running time of our approach and show that it achieves the best possible bounds for both cases. We also provide an extensive range of experimental results for all of these applications.

Finally, we extend our understanding of nonlinear models to the problem of learning neural network in chapter 5. In particular, we shift gear to study the problem of (provably) learning the weights of a two-layer neural network with quadratic activations (sometimes called shallow networks). Our shallow network comprises of the input layer, one hidden layer, and the output layer with a single neuron. We focus on the under-parametrized regime where the number of neurons in the hidden layer is (much) smaller than the dimension of the input. Our approach uses a lifting trick, which enables us to borrow algorithmic ideas from low-rank matrix estimation (fourth chapter). In this context, we propose three novel, non-convex training algorithms which do not need any extra tuning parameters other than the number of hidden neurons. We support our algorithms with rigorous theoretical analysis and show that the proposed algorithms enjoy linear convergence, fast running time per iteration, and near-optimal sample complexity. We complement our theoretical results with several numerical experiments.