

# RESEARCH STATEMENT

Xi Chen

Machine Learning Dept. School of Computer Science  
Carnegie Mellon University  
xichen@cs.cmu.edu

## 1 Motivation

My principal research interests encompass *machine learning and data mining*, focusing on learning algorithms for *high-dimensional* and *complex structured* data. The development of modern technology has enabled collecting data of unprecedented size and complexity. Examples include web text data, user preference profiles, images & videos, microarrays & proteomics, etc. These data exhibit the following properties which present new challenges for the current machine learning techniques:

1. The data are often ultra-high dimensional (e.g. text data, genetic data) but also extremely sparse in that only a small subset of features are relevant for specific learning tasks and these need to be identified.
2. The data are large-scale and may be collected in a streaming fashion.
3. The relationship among inputs and responses could be very complicated, e.g. highly nonlinear as in brain image data.
4. There are often hidden structures and complex relations among data. Examples include potentially implicit group structures, such as synonyms in text data; and graph structures, such as dependency and associative relationships in social network data, gene regulatory network in microarray data, etc.
5. The hidden structures (groups, networks) of the data are seldom static, as relations and dependencies change over time, location or task.

Recently developed convex programming based sparse learning techniques provide us a suite of powerful tools for understanding and exploring high-dimensional data. By exploring sparsity, we can always learn a parsimonious and compact model which is more interpretable and computationally tractable. When it is known that the underlying model is indeed sparse, sparse learning methods can provide us a more consistent model and much improved prediction performance. Despite these merits, many of existing sparse learning methods either rely on rigid assumptions (e.g. linear assumption), or fail to incorporate rich structural information, or are computationally very expensive. My research theme is to contribute to the development of sparse learning algorithms to incorporate and learn the complicated dynamic structures for large-scale high-dimensional data with applications in text mining, genetic data analysis, fMRI data analysis and social network analysis.

## 2 Recent Research Projects

### 2.1 *Structured Sparse Learning & Efficient Optimization*

In many real-world applications, it is crucial to explore structure information among features. For example, in genetic data analysis, biologists need to select relevant pathways (a collection of genes and promoter regions with similar function or at the same location as a group) to find controlling factors of a disease; or use the gene regulatory network as a graph to guide the selection of relevant genes and promoters. For document classification with word features, using WordNet, one can group nouns, verbs, adjectives and adverbs into different sets of cognitive synonyms (synsets) and organize them into hierarchical structure.

To utilize available structure information to guide the learning procedure, one can formulate the problem of *structured sparse learning* into an optimization problem, which minimizes a loss function with structured sparsity-inducing regularization. However, due to the non-smoothness and non-separability of structured regularizers, the corresponding optimization has been widely recognized as a challenging problem; and traditional methods can only handle very simple structures (e.g. non-overlapping groups or chains) with poor scalability. To address this challenge, we proposed a *unified smoothing proximal-gradient (SPG) optimization framework* which can deal with a wide spectrum of structured regularizers, including overlapping groups, hierarchical trees and graphs as special cases [AOAS 11, UAI 11]. Utilizing the dual norm, we made a key observation that a variety of structured regularizers can be reformulated into a special form. Capitalizing on this form, we introduced its smooth approximation and solved this approximation via an accelerated gradient descent scheme. The method is very efficient with fast theoretical convergence rate guarantees. Since it only uses the gradient information, it can easily scale up to large dataset with millions of instances and features. Moreover, we proposed several variants of SPG, tailored to address *multi-task learning* where the correlated responses also lie in a high dimensional space with certain kind of structures, and *structured sparse canonical correlation analysis* [AISTATS 11 Oral].

Utilizing the stochastic gradient, we further extended SPG to an *online learning scenario*, which can provide real time service for web data arriving at a high-rate. We also proposed a *distributed version* of our online learning algorithm to deal with a more challenging case where a single processor cannot keep up with the high rate at which data arrive. We have implemented the distributed algorithms on IBM NIMBLE, which is a general platform in IBM for conducting parallel learning, and applied them to categorize large volumes of web-documents in real time.

### 2.2 Learning Sparse *Dynamic* Networks

In addition to incorporating the available structure information, we strive to automatically learn hidden structures in the data. One important structure learning problem is to uncover the dependency relationship among various features (e.g. profession, education, hobby, etc in social network data). Graphical models, particularly Markov Random Fields, which combine statistical machine learning theory and graph theory, are widely used in modeling the *dependency relationships* between features as *sparse*

*graph structures*. Most of the existing work in learning graphical models assumes that the graph structures are static. However, this assumption is clearly violated in the real world, and instead, the graph structures are often changing gradually but smoothly according to other conditions as input, (e.g. time, space, etc). To learn dynamic sparse graph structures, we proposed the following two different approaches:

1. Using the kernel smoothing technique, we estimate the graph structure at each input point (e.g. time, space, etc) as a kernel weighted combination of the graph structures estimated at all other different points [AAAI 10].
2. We split the input space using the dyadic tree model and at each leaf of the tree, we estimate a sparse graph structure. The oracle inequalities on risk minimization and graph estimation consistency are established [NIPS 10 Spotlight].

Both methods can efficiently learn *dynamic sparse graph structures* evolving over a set of dimensions and can scale up to very large networks.

### 2.3 More Flexible *Nonparametric* Sparse Learning

Many existing sparse learning methods (e.g. Lasso), heavily rely on the linear assumption, i.e. responses are linear combinations of inputs. However, this assumption is too constraining for many modern applications. To address this challenge, we proposed several large-scale *nonparametric* sparse learning methods, which assume that responses are smooth functions of inputs. And the inference is directly conducted in infinite-dimensional space. We showed that nonparametric methods are more flexible with many potential benefits for modern scientific applications:

1. While the convex optimization based methods for variable selection are well studied, greedy based algorithms have not received much attention. In fact, greedy methods are very simple, easy to use and can scale up to ultra-high dimensional large datasets. We established a general greedy framework [NIPS 09] for sparse *nonparametric* learning, which can be applied to both additive model and generalized multivariate regression. Our method outperforms state-of-the-art competitors based on convex optimization, e.g. sparse additive model (SpAM), with rigorous theoretical justifications. This work has attracted some attentions in reviving investigations of greedy approaches for large-scale sparse learning.
2. One of the most popular nonparametric methods in machine learning and data mining is the classification and regression trees (CART). Tree models have many virtues, e.g. simplicity of design, good interpretability, easy to implement and good practical performance. However, traditional CART does not explicitly enforce the sparsity constraint for high dimensional data. We proposed a multivariate dyadic regression tree model with a novel sparsity-inducing regularization term. It can simultaneously conduct function estimation and variable selection in high dimensions and achieve nearly optimal rates of convergence. [NIPS 10 & Winner of the American Statistical Association (ASA) Student Paper Competition 10].

## 2.4 Applications of Sparse Learning

In addition to developing new methodologies for sparse learning, I am also interested in designing new sparse learning methods with specific applications:

1. *Text Data Analysis*: One of the most remarkable properties of text data is that it is highly *sparse* in the sense that many features (words, links, phrases, named entities, etc) are irrelevant for the modeling task. Therefore, as we demonstrated, sparse learning could be a very powerful tool for various text learning tasks.
  - (a) *Learning to Rank*: Most existing *learning to rank* methods use a small number of hand-crafted features. A drawback of using hand-crafted features is that they are often expensive and specific to datasets and tasks, requiring domain knowledge in preprocessing. In contrast, the raw features (e.g., words, n-grams, or tags such as POS and named-entities), are easily available and carry strong semantic information. One of the main reason that raw features have not been widely used in ranking is due to the prohibitive computational cost and the memory requirement to store a large amount of parameters scaling according to the vocabulary size or even worse (e.g. for n-grams). We proposed a novel ranking model with sparsity constraint which captures synonymy and polysemy among query and document features. The empirical performance of our method matches the state-of-the-art performance of previous approaches that use very carefully hand-crafted feature engineering. Due to the sparsity-inducing regularization, our method also achieves fast convergence rate, better generalization ability and takes much less memory for real web ranking with millions of features [ICDM 10].
  - (b) *Dimension Reduction with LSA*: Latent semantic analysis (LSA) has been one of the most important unsupervised dimensional reduction techniques for the text data analysis. It is well known that higher dimensional latent space can capture more information of the original data and hence achieves better reconstruction performance. However, when the dimension of latent space is large, it is required much more memory to store the projection matrix; and given a large corpus with millions of documents, the computational cost for projecting these documents is also expensive. To address this challenge, we proposed *sparse LSA* which formulates LSA as an optimization problem and enforces the sparsity constraint on the projection matrix. With a novel optimization algorithm, the highly sparse projection matrix can be efficiently learned. It requires small amount of memory with very low computational cost for projecting billions of documents. Moreover, as an important by-product, sparse LSA can provide us a compact representation of the topic-world relationship as latent Dirichlet allocation. [SDM 10].
2. *Genome-wide Association Study*: we applied structured sparse learning method to study associations between genomic and phenotypic variations, known as eQTLs mapping, which is a fundamental problem in genome-wide association study. Our method fully utilizes the pathway and regulatory network as prior knowledge in

the sparse learning and hence leads to more interpretable and biological meaningful results [AOAS 11, AISTATS 12].

3. *Neural Semantic Basis Discovery*: we proposed *adaptive multi-task sparse learning methods* and applied it to predict fMRI images when presenting a stimulus word. It leads to interesting discovery of the relationship between words' semantic meanings and activation regions in our brain [SDM 12].
4. *Time-evolving Recommending System*: we developed an algorithm for modeling the temporal pattern of user preference over items (e.g., blogs, games, movies) and further proposed time-evolving collaborative filtering for large-scale recommending system [SDM 10].

### 3 Research Plans

The following are my plans for the future research:

1. An immediate future research is to combine the idea of structured sparse learning (Section 2.1) and nonparametric sparse learning (Section 2.3). We would like to incorporate group and graph structures in *nonparametric* sparse learning, which can provide more flexible models using the structured information as a guide. However, it presents new challenges on the optimization algorithms since learning smooth functions is much harder than learning a set of parameters. I plan to address this problem by using more advanced functional optimization and convex analysis tools and to propose structured nonparametric sparse learning methods which can scale up to millions of instances and features.
2. Another important research direction is to automatically learn different hidden structures from the data itself. Although learning the conditional dependency relationship as an undirected graph structure as been studied (Section 2.2), there are many others structures, e.g., group structure, forest structure, directed graph structure, need to be further explored.
3. To make the sparse learning be applied to real web learning task, I will further improve my current work both in scalability and speed. For example, I will explore map-reduce type abstractions for large-scale parallel computation of various sparse learning tasks, where the challenge is how to de-couple the computation among different machines. In Section 2.1, we introduced the distributed gradient method. I would also like to explore other types of distributed optimization techniques, e.g., distributed coordinate descent, for general structured sparse learning.

I believe it is the right time for the industry applications of sparse learning methodologies. In the future, I would like to collaborate with industry in understanding and developing solutions for more practical problems. In the long term, I would like to develop practical, scalable and flexible models to explore, understand, and learn from complex high-dimensional data with dynamic hidden structures and apply the models with efficient learning algorithms to real world applications.