

# AM221 Final Project Status Report

## Dictionary Selection Under a Supermodular Assumption

Taylor Killian & Leonhard Spiegelberg

April 29, 2016

### **Abstract**

To be filled with an interesting summary of our work and results

## **1 Introduction**

## **2 Related Work**

## **3 Problem Definition**

## **4 Results**

## **5 Evaluation**

## **6 Discussion**

## **7 Model/Problem Refocusing**

We don't have any reason to refocus or redefine our project. There was initially some concern that our problem had already been solved in a satisfiable manner. We're happy to report that there is significant room for contribution in the realm of supermodular optimization. We began by developing intuition in supermodular and submodular optimization, highlighted here in this section, and by reviewing recent research in the development of algorithms for both supermodular minimization [?] and for a two-stage formulation of submodular maximization [?].

Sparse Regression / Dictionary Selection is a class of problems that are variants of representation learning. The goal of these kinds of problems is to determine a sparse representation of input data in the form of a linear combination of basis elements as well as the basis elements themselves. That is, one essentially factors the input matrix into a sparse catalog and a dictionary of basis elements, both of which need to be inferred from the data.

More formally this problem has been defined as follows:

Given data  $X = [x_1, \dots, x_k]$ ,  $x_i \in \mathbb{R}^d$ , we wish to find a dictionary  $\mathcal{D} \in \mathbb{R}^{d \times n}$  and a representation  $\mathcal{R} = [r_1, \dots, r_k]$ ,  $r_i \in \mathbb{R}^n$  such that:

$$\|X - \mathcal{D}\mathcal{R}\|_F^2 \text{ is minimized and } \mathcal{R} \text{ is sparse enough.}$$

Depending on the aim of (a) finding a minimal representation for given precision or (b) finding a best approximation using up to  $t$  variables the problem can be stated more generally as:

$$\min_{\mathcal{D}, \mathcal{R}} f(\mathcal{D}, \mathcal{R}; X) + g(\mathcal{D}) + h(\mathcal{R})$$

where  $f$  represents the data-fidelity term (usually a measure for goodness of the approximation) with regularizers  $g$ ,  $h$  for the dictionary and representatives respectively.

One way of stating the problem for (a) is

$$\min_{\mathcal{D}, \mathcal{R}} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \lambda \sum_{i=1}^k \|r_i\|_0 \quad (1)$$

$$\text{s.t.} \quad \|d_j\|_2 \leq 1, \forall j = 1, \dots, n \quad (2)$$

with a regularizer on the amount of variables used and a further restriction on the dictionary vectors to avoid scaling problems. The  $\|\cdot\|_0$  is hereby defined as the number of non-zero entries making the whole problem computationally difficult to track[?]. A formulation for the case (b) is

$$\min_{\mathcal{R}} \sum_{i=1}^k \|r_i\|_0 \quad (3)$$

$$\text{s.t.} \quad \|X - \mathcal{D}\mathcal{R}\|_F^2 \leq \epsilon \quad (4)$$

$$\|r_i\|_0 \leq t, \forall i = 1, \dots, k \quad (5)$$

We abstract some of the functional definitions from the above statement for case (b) to represent the optimization problem in a more general form:

$$\begin{aligned} & \text{minimize} && F(a) \\ & \text{subject to} && \text{supp}(a) \subseteq S \\ & && |S| \leq t \end{aligned}$$

## 7.1 Related Work

In order to tackle the problem whose complexity mainly arises from the  $\|\cdot\|_0$  norm, one approach is to relax the norm to a more feasible one (i.e. LASSO, LARS for convex-relaxation or log penalty for non-convex relaxation). However, this might lead to over-penalization[?]. Another ansatz is to utilize a greedy approach which constantly adds or removes variables based on some measure[?]. This can be stated as a maximization problem of a submodular function. A submodular function thereby can be viewed as a discrete analog to

a convex function in the continuous case[?] which encodes the principle of “diminishing returns”. These submodular approximations to Sparse Dictionary Selection have become nearly ubiquitous in the literature with slight algorithmic variations that approach the theoretical bounds,  $(1 - 1/e)$ , set in [?].

Applications of Sparse Dictionary Learning—in particular the utilization of submodular approximations—are broad and varied; with uses found in classical Machine Learning (i.e. Sparse Linear Regression), Computer Vision(Image reconstruction, inpainting, denoising), Signal Processing[?][?], Social Network Analysis, Statistics[?], and Economics[?] among many others.

There has been little work in Sparse Regression and Dictionary Learning that utilize the functional inverse of submodular approximations, known as supermodular functions. Supermodularity can facilitate the potential formulation of a dual to the native Dictionary Learning problem. Recently there has been some attempts at developing foundations for extending the suite of algorithms used in submodular maximization to supermodular formulations. In order to accomplish this, several caveats have been made on the functional representations themselves [?]. We are continuing our research and study and are closer to the development of a two-stage supermodular extension of the continuous greedy algorithm used in the submodular case [?]. This project still intends to be the summary of our explorations within supermodular functions and their extension to this class of problems.

## 8 Data

So far, we haven’t yet determined how we will use data to solidify our project as much of our work at this point is in understanding and developing the theory necessary to develop a new or modified algorithm. As we are looking to develop a novel algorithm to provide a new benchmark for approximate solutions to this class of problems, we will likely assimilate the data and experiments used in [?], [?], [?], [?], among others.

## 9 Completed Steps

In our proposal we outlined the following course of action:

Our first step in this project will be to understand the supermodular minimization framework in which sparse regression can be defined. Then, after having connected and identified sparse regression as supermodular minimization problem we want to show that dictionary selection is a 2-stage sparse regression.

In the following step we need to understand the greedy algorithm for submodular maximization and its improvement, the continuous greedy algorithm. After this, understanding 2-stage submodular optimization would ideally allow us to connect the pieces, gain insight and finally propose a new algorithm of dealing with the problem.

We have successfully completed the first of these stages (as will be shown in ??) and are on our way toward completing the second.

## 9.1 Dictionary Selection as Two-Stage Supermodular Minimization

In [?], the authors step through a few examples to demonstrate the application of supermodular functions to relevant problems. One that we highlight is that of Sparse Multiple Linear Regression (SMLR). We focus on this problem due to its close relation to Dictionary Selection (in fact we aim to show that these problems are functionally equivalent) as both problems are under the umbrella of Representation Learning. The goal of demonstrating the equivalence of SMLR and Dictionary Selection is that we can extend the theoretical guarantees and algorithmic foundations in [?] in our future work.

SMLR is defined as follows:

Given two matrices  $X \in \mathbb{R}^{m \times n}$ ,  $Y \in \mathbb{R}^{m \times l}$  and an integer  $k$ , find a matrix  $W \in \mathbb{R}^{n \times l}$  that minimizes  $\|XW - Y\|_F^2$  subject to  $W$  having at more  $k$  non-zero rows. This problem is usually paired with a common assumption/simplification, where the columns of  $X$  are adjusted to have unit norm.

As outlined in Section ??, here is a general definition of the Dictionary Selection problem.

$$\underset{\mathcal{D}, \mathcal{R}}{\operatorname{argmin}} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \lambda \sum_{i=1}^k \|r_i\|_0 \quad (6)$$

$$\text{s.t.} \quad \|d_j\|_2 \leq 1, \forall j = 1, \dots, n \quad (7)$$

Thus, given an input dataset  $X = [x_1, \dots, x_k]$ ,  $x_i \in \mathbb{R}^d$ ,  $X \in \mathbb{R}^{d \times k}$  we wish to find the dictionary  $\mathcal{D} \in \mathbb{R}^{d \times n}$ ,  $\mathcal{D} = [d_1, \dots, d_n]$  and a representation  $\mathcal{R} = [r_1, \dots, r_k]$ ,  $r_i \in \mathbb{R}^n$ ,  $\mathcal{R} \in \mathbb{R}^{n \times k}$ , such that both  $\|X - \mathcal{D}\mathcal{R}\|_F^2$  is minimized and the representations  $r_i$  are "sparse enough" (can be specified column by column as in [?], or be extended as defined in SMLR).

In SMLR, the data matrix  $X$  with the determined, sparse  $W$ , are used to approximate  $Y$  while in Dictionary Selection the matrices  $\mathcal{D}$ ,  $\mathcal{R}$  are determined to approximate the data  $X$ . In both problems, sparse representations are used to select a combination of data that closely replicates known data. For all intents and purposes these two problems are functionally equivalent with consideration being made to ensure that the sparsity requirements of Dictionary Selection (where there is a limit on the sparsity of each row/column) port into those of SMLR (where the sparsity requirements are placed on the matrix in full). This isn't of too much concern as this can be handled in the definition of each specific application of the Dictionary Selection problem.

Now, you'll note that the definition of the dictionary selection problem (Equation 6) requires the minimization over the matrices  $\mathcal{D}$  and  $\mathcal{R}$ . When taken together, this problem is combinatorially infeasible [?] and non-convex. However, if we separate the problem into a two-stage optimization (where we fix one matrix and solve for the other and then iterate) we can apply well understood convex solution strategies. [?], [?], [?], [?]. These two stages are:

1. Fix  $\mathcal{D}$ , find a sparse coding between it and  $X$ .
2. Solve the Dictionary Optimization problem:  $\mathcal{D} = X\mathcal{R}^+$  following which we renormalize  $\mathcal{D}$ .

Upon iteration we can expect convergence, using guarantees in the continuous greedy algorithm [?], [?], [?] as well as those made in [?]. The commonly known sparse regression can be seen as an instance of sparse

multiple linear regression with  $Y$  existing only of one column. I.e.  $XW - Y$  is a vector.

## 9.2 Weakly- $\alpha$ Supermodularity

Weakly- $\alpha$ -supermodularity is a relaxation of supermodularity. With this relaxation, it can be shown that problems that are  $\alpha$ -weakly supermodular can be solved using a slightly adapted standard greedy algorithm. However, convergence typically requires more steps than if the problem was supermodular only.

A non-negative, non-increasing set function  $f(S) : 2^{[n]} \rightarrow \mathbb{R}^+$  is weakly- $\alpha$ -supermodular if there exists  $\alpha \geq 1$  such that for any two sets  $S, T \subseteq [n]$

$$f(S) - f(S \cup T) \leq \alpha |T \setminus S| \max_{i \in T \setminus S} f(S) - f(S \cup \{i\})$$

To understand the definition better, assume we are given two disjoint sets  $A, B \subseteq [n]$  with  $A \cap B = \emptyset$ . Then  $\alpha$ -weakly-supermodularity means that we can find an element  $i \in B$  s.t.

$$\frac{f(A) - f(A \cup B)}{\alpha |B|} \leq f(A) - f(A \cup \{i\})$$

In words, we can find an element  $i$  that is better (in terms of lowering the objective function) than  $\frac{1}{\alpha}$  times the average gain of an element of  $B$ . Note that for a supermodular function  $\alpha = 1$ .

In order to solve the SMLR problem

$$\min\{f(S) : |S| \leq k\}$$

for some  $k$ , the greedy extension algorithm needs at most

$$\left\lceil \alpha k \ln \left( f \left( \frac{S_0}{\epsilon} \right) \right) \right\rceil$$

steps given a start solution  $S_0$  (usually the empty set), a threshold error  $\epsilon$  (i.e.  $f(S_k) \leq (1 + \epsilon)f(S^*)$  for the optimal solution  $S^*$ ) and the weakness-parameter  $\alpha$  of the function  $f$  [?].

As in [?] the authors show that sparse regression can be seen as minimizing a  $\alpha$ -weakly supermodular function with  $\alpha = \|X^+\|_F$ , the extended greedy algorithm presented can be used to solve sparse regression.

## 10 Next Steps

Our primary next step is in taking measures to develop the structure for an extension of the continuous greedy algorithm from submodular functions. After developing this algorithm, which we intend to pattern after that done in [?], we will need to develop guarantees for convergence, accuracy and efficiency. There are theoretical benchmarks that have been made in [?], [?] and [?] that we also need to be sure that we meet in the “dual”-like formulation of supermodular optimization for dictionary selection. Once we have this algorithm in place we will need to apply it some data sets like those in [?].

In another step, we would like to implement the extended greedy algorithm presented in [?] and com-

pare its performance to other known approximative algorithms in order to better understand its limits or potential flaws. Eventually, this can also serve as a benchmark for the extension of the continuous greedy algorithm we seek to develop.