AM221 Final Project

Dictionary Selection Under a Supermodular Assumption

Taylor Killian
Institute for Applied Computational Science
52 Oxford Street
Cambridge, MA
taylorkillian@g.harvard.edu

Leonhard Spiegelberg
Institute for Applied Computational Science
52 Oxford Street
Cambridge, MA
spiegelberg@g.harvard.edu

May 1, 2016

Abstract

We attempt to develop a method by which to solve dictionary selection under a supermodular assumption. This method is motivated primarily by the work of Singer et al. (2016) [11] where a two-stage submodular maximization generalization was developed. The development of a two-stage supermodular minimization algorithm to address Sparse Dictionary Selection utilizes approximation results derived by Boutsidis et al. (2015) [1] for supermodular optimization. We show that Sparse Dictionary Selection is functionally equivalent to Sparse Multi Linear Regression which allows us to leverage [1]. We analyze and experiment with an implementation of our two-stage optimization routine and compare it to current methods. With this project we have laid the ground for further exploration into the formalization of the two-stage method that we propose here.

Contents

1 Introduction

Dictionary selection and sparse regression (problems that we show to be functionally equivalent) 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, or design matrix, into a sparse catalog and a dictionary of basis elements, both of which need to be inferred from the data.

1.1 Problem Definition

Dictionary learning, in its general form, can be seen as

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

where f describes the objective function used to measure goodness of approximation of X through $\mathcal{D},\mathcal{R}$, g describing suitable constraints on the dictionary, h on the representation respectively.

With an input dataset $X = [x_1,\ldots,x_k], x_i \in \mathbb{R}^d, X \in \mathbb{R}^{d\times k}$ we wish to find a dictionary $\mathcal{D} \in \mathbb{R}^{d\times n},\ \mathcal{D} = [d_1,\ldots,d_n]$ and a representation $\mathcal{R} \in \mathbb{R}^{n\times k},\ \mathcal{R} = [r_1,\ldots,r_k],\ r_i \in \mathbb{R}^n$ such that both $\|X - \mathcal{D}\mathcal{R}\|_F^2$ is minimized and the representations r_i are "sparse enough". To limit the dictionary becoming infinitely large (or small) we introduce a constraint on the dictionary's columns. For this any sufficient norm can be used. An often used norm is the l_2-norm. Thus for the dictionary learning problem we introduce the problem with constraints

$$\begin{aligned} &\min_{\mathcal{D},\mathcal{R}} \ \|X - \mathcal{D}\mathcal{R}\|_F^2 \\ \text{s.t.} \quad &\|d_j\|_2 \le 1,\ \forall j = 1,...,n \\ &\|r_i\|_0 \le t,\ \forall i = 1,...,k \end{aligned}$$

Here the constraints express the expectation that both $\mathcal{D}$ is controllably determined and that $\mathcal{R}$ is adequately sparse. The difficulty in solving this problem comes from the mathematical challenge the $\|.\|_0$ norm yields. This norm is defined as:

Let $x \in \mathbb{R}^d$

$$\ell_0(x) := \big|\{x_i : x_i \neq 0\}\big|$$

That is, any column of $\mathcal{R}$ must have no more than t non-zero elements. By using Lagrange multipliers this can be

brought to the general form above

$$\min_{\mathcal{D},\mathcal{R},\theta,\lambda} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \sum_{j=1}^{m} \theta_j(\|d_j\|_2 - 1) + \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

I.e. the functions are

$$f(\mathcal{D}, \mathcal{R}, X) := \|X - \mathcal{D}\mathcal{R}\|_F^2$$
$$g(\mathcal{D}, \theta) := \sum_{j=1}^{n} \theta_j(\|d_j\|_2 - 1)$$
$$h(\mathcal{R}, \lambda) := \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

1.2 Related Work

In order to tackle the problem whose complexity mainly arises from the $\|.\|_0$ 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 [10]. Another ansatz is to utilize a greedy approach which constantly adds or removes variables based on some measure [3]. Another method attempts to adjust the dictionary and sparse representation in an online fashion [9]. 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 [7] 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 [8].

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 [7] [10], Social Network Analysis, Statistics [4], and Economics [6] 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 [1]. We leverage the gains made in submodular optimization and use recent results in approximating supermodular functions in constructing a two-stage supermodular minimization extension of the continuous greedy algorithm used in

the submodular case [11].

1.3 Supermodular Assumption

Typical optimization problems that use submodular or supermodular functions are generally of the form:

Given a set of objects $V = \{v_1, \ldots, v_n\}$ and a function $f : 2^V \rightarrow \mathbb{R}$ that returns a real value for any subset. Since we are interested in finding a dictionary $\mathcal{D}$ and coding $\mathcal{R}$, our objective function becomes the error that arises when reconstructing X with $\mathcal{D}$ and $\mathcal{R}$ and the set V is represented by the columns of X we choose to determine $\mathcal{D}$, and later use to solve for $\mathcal{R}$, with. In solving the minimization we iteratively select a subset $S \subseteq X$ according to $arg\min_{S \subseteq X} f(S)$, subject to some constraints on the size of S. As was expressed in 1.1, we can see that our general formulation of Dictionary Selection:

$$\min_{\mathcal{D},\mathcal{R},\theta,\lambda} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \sum_{j=1}^{m} \theta_j(\|d_j\|_2 - 1) + \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

can be expressed in supermodular forms.

The remainder of this paper is focused on the development of a two-stage supermodular minimization algorithm. We build up to this algorithm via discussion and analysis of recent results. We present a framework of our two-stage minimization method for dictionary selection and compare it to other commonly implemented algorithms, e.g. MOD and "Online" learning.

2 Analysis and Extension of Recent Results

This section is dedicated to a brief analysis and extension of two recent results that we leverage in the development of a two-stage supermodular minimization algorithm. The first is the development of a generalized two-stage submodular maximization algorithm [11], that we use as a guide in the development of our own supermodular algorithm. The second is the development of supermodular minimization techniques based on an approximation of the supermodular functions used to express various objectives [1]. We include a brief discussion of these approximation methods. The primary example that the authors use to demonstrate their approximation method is that of Sparse Multiple Linear Regression (SMLR). We show that SMLR is functionally equivalent to Dictionary Selection and that Dictionary Selection is inherently a two-stage problem. This fact allows us to directly use the results from [1] as a fundamental part of solving for the sparse representation $\mathcal{R}$ as shown below in Section 3.

2.1 Weakly-α Supermodularity

Weakly-α-supermodularity is a relaxation of supermodularity. With this relaxation, it can be shown that problems that are α-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 $f(S) : 2^{[n]} \to \mathbb{R}^+$ is weakly-α-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 α-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 ϵ (i.e $f(S_k) \leq (1+\epsilon) f(S^*)$ for the optimal solution S^*) and the weakness-parameter α of the function f [1].

As in [1] the authors show that sparse regression can be seen as minimizing a α-weakly supermodular function with $\alpha = \|X^+\|_F$, the extended greedy algorithm presented can be used to solve sparse regression.

2.2 Dictionary Selection as Two-Stage Supermodular Minimization

Here we highlight the functional equivalence of Dictionary Selection to SMLR, aided by the fact that 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 [1] 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.

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

$$\text{s.t.} \qquad \|d_j\|_2 \leq 1, \forall j = 1, ..., n \tag{2}$$

Thus, given an input dataset $X = [x_1, \ldots, x_k],\ x_i \in \mathbb{R}^d,\ X \in \mathbb{R}^{d\times k}$ we wish to find the dictionary $\mathcal{D} \in \mathbb{R}^{d\times n},\ \mathcal{D} = [d_1, \ldots, d_n]$ and a representation $\mathcal{R} = [r_1, \ldots, r_k],\ r_i \in \mathbb{R}^n,\ \mathcal{R} \in \mathbb{R}^{n\times k}$, 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 [4], 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 [12] 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. [3], [2], [4], [11]. 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 [2], [3], [7] as well as those made in [1]. 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.

3 Two-stage Supermodular Minimization Algorithm

Recall the general form of the Dictionary Selection Problem:

$$\min_{\mathcal{D},\mathcal{R},\theta,\lambda} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \sum_{j=1}^{m} \theta_j(\|d_j\|_2 - 1) + \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t)$$

We now relax the problem by formulating it into the two stages described above:

1. solve for fixed $\mathcal{D}, \theta$ the minimization to obtain optimal $\mathcal{R}, \lambda$.
2. solve optimization for $\mathcal{D}, \theta$ with obtained values
3. repeat previous steps until done

In generic terms, this is similar to the method of optimal directions. In the traditional method of optimal directions [5] the subproblem $\min_{\mathcal{R},\lambda} f(\mathcal{D}, \mathcal{R}, X) + g(\mathcal{D}, \theta) + h(\mathbb{R}, \lambda)$ is solved either by various relaxation methods (i.e. LASSO, Matching Pursuit, etc.), however for our approach we want to embed the problem into a combinatorial optimization framework, which guarantees us bounds on optimality. Using [1] we reformulate the Dictionary Selection with α-weakly supermodular functions. We will now show how to obtain this formulation: Recall that

$$\begin{aligned} &\min_{\mathcal{R},\lambda} && f(\mathcal{D}, \mathcal{R}, X) && + && g(\mathcal{D}, \theta) && + && h(\mathcal{R}, \lambda) \\ =&\min_{\mathcal{R},\lambda} && \|X - \mathcal{D}\mathcal{R}\|_F^2 && + && \sum_{j=1}^{n} \theta_j(\|d_j\|_2 - 1) && + && \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t) \end{aligned}$$

then let $S \subseteq [n]$ be a set of column indices for the dictionary. From now onwards we will assume that the given, fixed dictionary is normalized, i.e. $\forall d_i : \|d_i\|_2 = 1$. This allows us to remove the term $g(\mathcal{D}, \theta)$. Define now $\mathcal{D}_S$ as the matrix obtained from $\mathcal{D}$ with all columns not indexed by S to be set to zero.

$$(\mathcal{D}_S)_{ij} := \begin{cases} 0 & j \notin S \\ \mathcal{D}_{ij} & j \in S \end{cases}$$

$\mathcal{D}_S^+$ is the pseudoinverse which can be obtained i.e. via singular value decomposition when $\mathcal{D}_S = U\Sigma V^T$ through $\mathcal{D}_S^+ = V\Sigma^+ U^T$. Then the problem can be equivalently stated as

$$\begin{aligned} &\min_{\mathcal{R},\lambda} \|X - \mathcal{D}\mathcal{R}\|_F^2 + \sum_{i=1}^{k} \lambda_i(\|r_i\|_0 - t) \\ \iff &\min_{S \subseteq [n], |S| \leq t} \|X - \mathcal{D}_S \mathcal{D}_S^+ X\|_F^2 \end{aligned}$$

[1] have shown that the objective function $f(S) := \|X - \mathcal{D}_S \mathcal{D}_S^+ X\|_F^2$ of this optimization problem is α-weakly

supermodular with

$$\alpha = \max_{\tilde{S} \subseteq [n]} \|\mathcal{D}_{\tilde{S}}^{+}\|_F^2$$

3.1 How to combine it

For the two stage formulation we can generally think of the problem as of adapting the idea of the method of optimal directions but instead of projecting the given representation onto the dictionary we strive for an additional minimization problem: Let $f_1, ..., f_m$ be instances of the SMLR problem with different dictionaries D_k

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

Then the dictionary problem can be seen as

$$\min_{\mathcal{D}, \theta} q(f_1, ..., f_m) + \sum_{j=1}^{n} \theta_j (\|d_j\|_2 - 1)$$

for some combination function q. We now ask, which properties such a combination function ideally should satisfy.

One choice for q could be

$$q(f_1(\mathcal{D}, \theta), ..., f_m(\mathcal{D}, \theta)) := \sum_{i=1}^{m} f_i(\mathcal{D}, \theta)$$

another one to minimize the product of the SMLR instances. Reformulating it with a logarithm might give even better results with increased computationally tractability

$$q(f_1(\mathcal{D}, \theta), ..., f_m(\mathcal{D}, \theta)) = \sum_{i=1}^{m} \log(f_i(\mathcal{D}, \theta))$$

Algorithm proposals

In the following we proprose a general randomized greedy algorithm to solve the dictionary learning problem.

Algorithm 1: General randomized greedy algorithm for Sparse Dictionary Learning

input : A data matrix $X \in \mathbb{R}^{d \times k}$

output: A dictionary $\mathcal{D} \in \mathbb{R}^{d \times n}$ and representation $\mathcal{R} \in \mathbb{R}^{n \times k}$

1 $D_0 \leftarrow$ `RandomDictionary(`-1, 1`)` `// assert` $\mathcal{D}_0 \neq 0$

2 normalize s.t. $\|(\mathcal{D}_0)_{*j}\|_2 = 1$

3 $S_0, \mathcal{R}_0 \leftarrow \alpha$-`GreedySolver(`$X$, $\mathcal{D}$, t`)`

4 $m \leftarrow 0$

5 **while** $m <$ MAXSTEPS **do**

6 $\quad \mathcal{D}_{m+1} \leftarrow X_{S_m \setminus \{q\} \cup \{p\}}$

7 $\quad$ normalize s.t. $\|(\mathcal{D}_{m+1})_{*j}\|_2 = 1$

8 $\quad S, \mathcal{R} \leftarrow \alpha$-`GreedySolver(`$X$, $\mathcal{D}_{m+1}$, t`)`

9 $\quad$ **if** $\|X - \mathcal{D}_{m+1}\mathcal{R}\|_F^2 < \|X - \mathcal{D}_m\mathcal{R}_m\|_F^2$ **then**

10 $\quad\quad S_{m+1}, \mathcal{R}_{m+1} \leftarrow S, \mathcal{R}$

11 $\quad$ **end**

12 $\quad m \leftarrow m + 1$

13 **end**

We start by initializing a non-zero random dictionary. To allow for all directions, negative and positive values should be picked. To restrict the dictionary size, normalization along columns is done in lines 2 and 8. We use the greedy algorithm as proposed in [1] to solve for a given dictionary the SMLR problem by using t steps (yielding t columns for representation). The indices of the columns that best solve the SMLR are returned as set S along a representation $\mathcal{R}$. We then iteratively refine the solution. Therefore one observation is crucial: Given a set of columns S that represents the indices of non-zero columns of R, only $|S|$ corresponding columns in D obtained through matrix multiplication will contribute something to the norm. Thus, in order to improve the solution we need to exchange one (or more) columns of the current set S with some other column that might give a better solution. To obtain a new dictionary candidate we construct it by using a similar argument via the data matrix X. I.e. choose randomly a column index $q \in S$ that will be removed and add another random column index $p \in \{1, ..., k\} \setminus S$. This step is performed in line 7, 1. We then obtain the dictionary candidate by l_2-column normalization and solve the resulting SMLR problem via the α-weakly supermodular greedy algorithm. If the new solution is better than the old one, we keep it, else we perform another round of local search.

4 Results

HERE WE WILL TALK ABOUT THE RESULTS WE GET FROM LEOs EXCURSION INTO THE ANALYSIS

5 Discussion

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 [11], we will need to develop guarantees for convergence, accuracy and efficiency. There are theoretical benchmarks that have been made in [2], [8] and [10] 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 [11].

In another step, we would like to implement the extended greedy algorithm presented in [1] and compare 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.

References

[1] Christos Boutsidis, Edo Liberty, and Maxim Sviridenko. Greedy minimization of weakly supermodular set functions. *CoRR*, abs/1502.06528, 2015.

[2] Volkan Cevher and Andreas Krause. Greedy dictionary selection for sparse representation. *Selected Topics in Signal Processing, IEEE Journal of*, 5(5):979–988, 2011.

[3] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[4] F. Doshi-Velez and S. A. Williamson. Restricted Indian Buffet Processes. *ArXiv e-prints*, August 2015.

[5] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. On 1999 IEEE International Conference - Volume 05*, ICASSP '99, pages 2443–2446, Washington, DC, USA, 1999. IEEE Computer Society.

[6] U. Feige. On maximizing welfare when utility functions are subadditive. In *STOC '06 Proceedings of the thirty-eigth annual ACM symposium on Theory of computing*, pages 41–50, 2006.

[7] Zhuolin Jiang, Guangxiao Zhang, and L.S. Davis. Submodular dictionary learning for sparse coding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3418–3425, June 2012.

[8] Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *TWENTY-FIRST CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE (UAI*, 2005.

[9] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *26th International Conference on Machine Learning*, 2009.

[10] Jianping Shi, Xiang Ren, Guang Dai, Jingdong Wang, and Zhihua Zhang. A non-convex relaxation approach to sparse dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1809–1816, June 2011.

[11] Yaron Singer, Eric Balkanski, and et. al. Learning sparse combinatorial representations via two-stage submodular maximization. *(preprint) Submitted to 2016 ICML*, 2016.

[12] A.M. Tillmann. On the computational intractability of exact and approximate dictionary learning. *Signal Processing Letters, IEEE*, 22(1):45–49, Jan 2015.