HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

An Investigation into Two-Stage Supermodular Minimization

Taylor Killian and Leonhard Spiegelberg
AM 221 Final Project

IACS Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Introduction

Here we summarize our excursion into the area of sub-/supermodular optimization during the Spring '16 semester. We attempt to develop a method by which to solve dictionary selection under a supermodular assumption. There has been significant work in the realm of submodular maximization, from which we extend to supermodular formulations, which includes the development of algorithms and optimization strategies that guarantee (1-1/e) approximations of the true optimal value.

Background and Definitions

Typical optimization problems that use sub-/supermodular functions are of the form: Given a set of objects $V = \{v_1, \ldots, v_n\}$ and a function $f : 2^V \to \mathbb{R}$ that returns a real value for any subset. Suppose we are interested in finding the subset that either maximizes or minimizes the function, e.g., $\arg\max_{S \subseteq V} f(S)$, possibly subject to some constraints on the size of S. In dictionary selection these constraints are on the number of items that the subset S can contain, where the function encodes the the accuracy of representation by S.

We begin by establishing key definitions and properties of sub-/supermodularity.

A function $f : 2^V \to \mathbb{R}$ is ***submodular*** if for any $A, B \subseteq V$ we have:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

An alternate and equivalent definition that aids in developing intuition is, in the event that $A \subseteq B \subseteq V$ and $v \in V \setminus B$ we have:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$$

This means that the marginal increase of any additional element diminishes with the overall size of the set. This *"diminishing returns"* property is a key property of submodular functions.

Now, a function $f : 2^V \to \mathbb{R}$ is ***supermodular*** if for any sets $A, B \subseteq V$ we have:

$$f(A) + f(B) \leq f(A \cup B) + f(A \cap B)$$

This leads to a similar *"increasing returns"* property given sets $A \subseteq B \subseteq V$ and $v \in V \setminus B$ we have:

$$f(A \cup \{v\}) - f(A) \leq f(B \cup \{v\}) - f(B)$$

Sub-/Supermodular functions are are the functional inverse of each other. That relation is what motivates excursion into the development of two-stage supermodular minimization.

Dictionary Selection Problem Definition

Given data $X = \{x_1, \ldots, x_k\}, x_i \in \mathbb{R}^d$ we aim to find a dictionary $\mathcal{D} \in \mathbb{R}^{d \times n}$ and a representation $\mathcal{R} \in \mathbb{R}^{n \times k}$ such that we can solve the following optimization problem:

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

$$s.t. \;\; \|d_j\|_2 \leq 1, \; \forall j = 1, \ldots, n$$

Here the penalty term enforces that the r_i are sparse enough to facilitate efficient information transfer as well as reliable data reconstruction. This kind of dictionary selection problem has many areas of application where a sparse representation of data can aid in further processing; e.g. classical machine learning, signal processing, network analysis, etc.

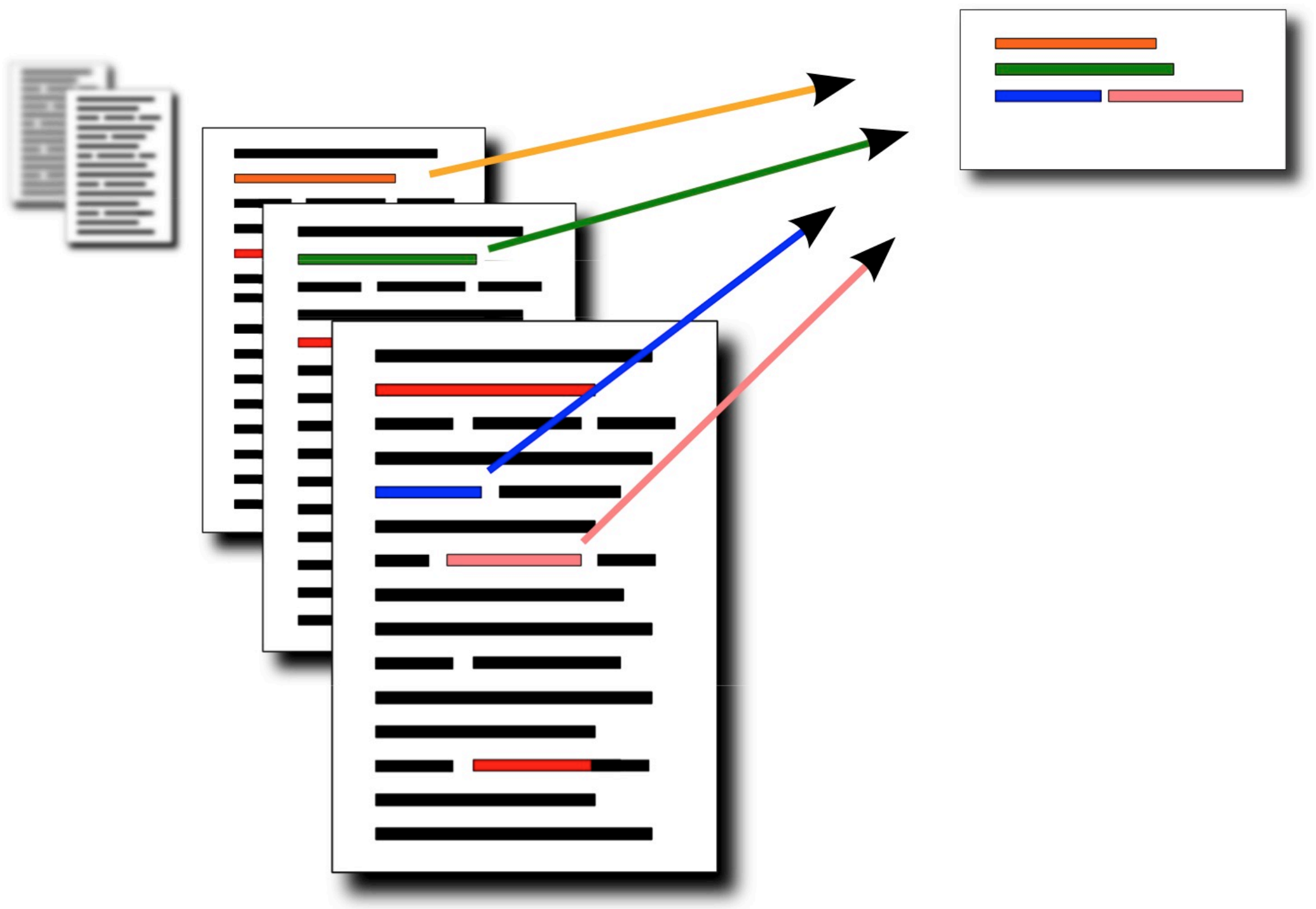


Image courtesy of :
http://www.cse.chalmers.se/~kageback/extractive-summarization-using-continuous-vector-space-models/

One application that we highlight is document summarization where we try to represent a collection of text documents with sentence fragments. Here we would attempt to minimize topic reconstruction error subject to the number of fragments we choose to represent a document by. This problem is functionally equivalent to sparse multi-linear regression for which we leverage results from Boutsidis et al. [1] to begin the development of a two-stage supermodular minimization algortithm.

Inherently, the dictionary selection problem is a two stage optimization. We first solve for the dictionary, then optimize the sparse representation of the data. We iterate between the two stages until convergence.

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:*** 1102.3875, 2011.
[4] Yaron Singer, Eric Balkanski, et. al. *Learning sparse combinatorial representations via two-stage submodular maximization.* ***(preprint) Submitted to 2016 ICML,*** 2016

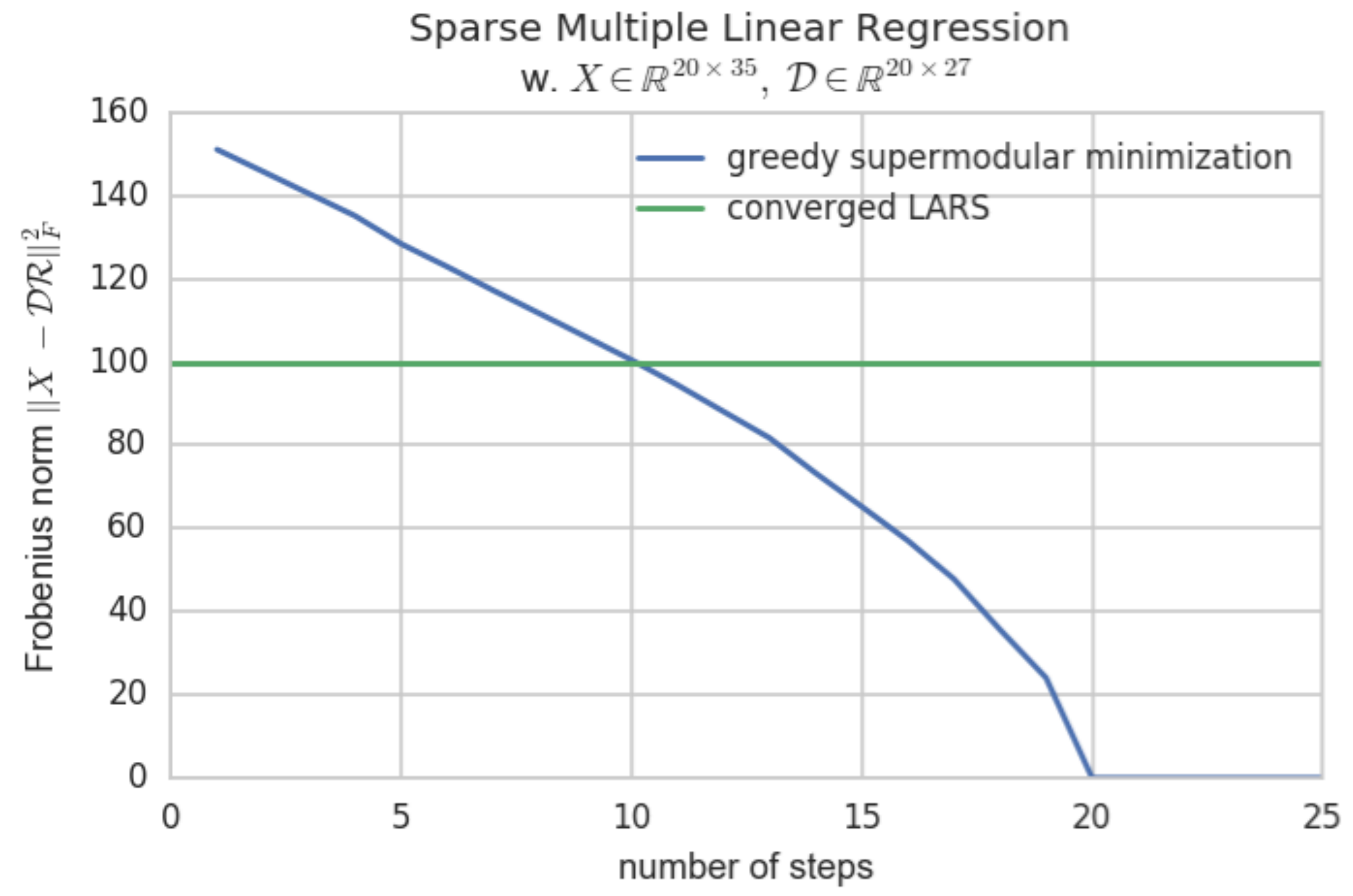
Weakly-α Supermodularity

Previous work to solve the Dictionary Learning problem has been focusing on trying to relax the mathematically intractable pseudo-"ℓ_0"-norm by higher norms. Such a relaxation does not guarantee any bounds on the original problem. Given a known dictionary, [1] have shown that Sparse Multiple Linear Regression (SMLR) can have a α-weakly supermodular objective function. A set function $f : 2^V \to \mathbb{R}$ is α-weakly supermodular if for $S, T \subseteq V \;\; \alpha > 0$

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

The SMLR can then be solved by an adapted greedy algorithm guaranteeing a bound on optimality.

$$\min_{S:|S| \leq k} \|X - D_S D_S^+ X\|_F^2$$



Two-stage Minimization

The development of a two-stage supermodular minimization algorithm follows [4] where a two-stage submodular maximization procedure is laid out. We utilize the weakly-α approximation highlighted above in order to select the sparse representation.

1. Initialize $\mathcal{D}$ by selecting the first k rows of X.
2. Iterate until convergence
   a. Solve for $\mathcal{R}$ according to the weakly-α algorithm above
   b. Improve $\mathcal{D}$ via local search
3. After approximate $\mathcal{D}, \mathcal{R}$ are found we ensure all constraints are met. If not we re-intialize and run the procedure again

Future Work: -Complete development of algorithm. -Investigate continuous relaxation of the above sub optimization problems. -Establish formal convergence, approximation criteria of our algorithm.

Acknowledgements

We would be remiss without thanking and acknowledging the direction, advice and encouragement from Prof. Yaron Singer, Thibaut Horel and Rajko Radanovic.