Restricted Strong Convexity Implies Weak Submodularity

Alex Dimakis*, Sahand Negahban†, Ethan R. Elenberg*, Rajiv Khanna*

*UT Austin,
Department of Electrical and Computer Engineering
†Yale University,
Department of Statistics
Set Function Optimization

- Many problems can be cast as an optimization over a finite set
- Examples:
  - Data summarization ($k$-medians, $k$-medoids)
  - Subset cover
  - Sparse regression
Set Function Optimization

- Many problems can be cast as an optimization over a finite set
- Examples:
  - Data summarization ($k$-medians, $k$-medoids)
  - Subset cover
  - Sparse regression

$k$-medoids: given $V = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$

$$\arg\max_{S:|S| \leq k} \max_{\pi:V \rightarrow S} \sum_{j=1}^n -\|x_{\pi(j)} - x_j\|_1$$
Many problems can be cast as an optimization over a finite set.

Examples:
- Data summarization ($k$-medians, $k$-medoids)
- Subset cover
- Sparse regression

$k$-medoids: given $V = \{x_i\}_{i=1}^{n} \subset \mathbb{R}^d$

$$\arg\max_{S:|S| \leq k} \max_{\pi:V \rightarrow S} \sum_{j=1}^{n} -\|x_{\pi(j)} - x_j\|_1$$

In general, take $V = \{1, 2, \ldots, p\}$ and set function $f : 2^V \mapsto \mathbb{R}$

$$\arg\max_{S:|S| \leq k} f(S)$$
High-dimensional statistics: $p \gg n$

Variable selection

Lasso, Graphical Lasso, sparse PCA

Reduce to lower-dimensional structure

Sparse optimization: goal to maximize $l(\beta)$

$$f(S) = \max_{\beta_{Sc}=0} l(\beta) - l(0)$$

e.g. $l(\beta) = \log$-likelihood
Set function optimization is in general NP-hard
- $k$-medians, subset cover, facility location, etc.
- Sometimes subset selection for regression is tractable
  - What settings for general problems?
  - What structural assumptions can we exploit?
  - For sparse linear regression, use ideas such as Restricted Isometry Property, Restricted Strong Convexity, or convex relaxations
Computational Answers for Sparse Regression Problems

- Long line of work
- Early methods based on greedy heuristics
  - OMP, CoSaMP, Forward Stagewise/Stepwise Selection, ... 
  - Theoretical guarantees under structural assumptions
  - Zhang; Needell-Tropp; Jalali et. al.
- More recent focus on convex relaxations
  - Algorithm converges without any assumptions
  - Can provide theoretical guarantees
  - In practice, greedy methods perform as well or better
Das and Kempe (’11): Use **weak submodularity** to provide guarantees for greedy methods under *linear* regression

**This talk:** Guarantees for general, greedy support selection
  - Connect weak submodularity to Restricted Strong Convexity/Smoothness
Submodular Functions

- Analogous to convex, concave functions
- **Diminishing Returns**: if $A \subseteq B$ then
  \[
  f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)
  \]
- $f$ monotone: $f(A \cup \{x\}) \geq f(A)$
Submodular Functions

- Analogous to convex, concave functions
- **Diminishing Returns**: if $A \subseteq B$ then
  \[
  f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)
  \]
- $f$ monotone: $f(A \cup \{x\}) \geq f(A)$

- **Submodular**: maximize $\log \det$ of a principle submatrix
- **Monotone submodular**: $k$-medians, $k$-medoids
- **NOT submodular**: Generalized Linear Model (GLM)
  - Logistic Regression, Linear Regression, Poisson Regression
Maximize a submodular function under cardinality constraints

- Greedy optimization is a family of heuristics
  - Add elements to set that improve incremental result the most
- Fact (Nemhauser '78): Monotone, submodular function $f(S)$,
  \[ f(S_k) \geq (1 - \frac{1}{e}) f(S_k^*) \]
- Cannot improve upon $(1 - \frac{1}{e})$ in polynomial time
- Under "incoherence" assumptions, does linear regression satisfy submodularity?
Weak Submodularity

Relax the previous definitions
Relax the previous definitions

**Definition (Submodularity Ratio (Das-Kempe ’11))**

Let $S, L \subseteq [p]$ be two disjoint sets, and $f(\cdot) : [p] \to \mathbb{R}$. The submodularity ratio of $L$ with respect to $S$ is given by

$$
\gamma_{L,S} := \frac{\sum_{j \in S} [f(L \cup \{j\}) - f(L)]}{f(L \cup S) - f(L)}.
$$

The submodularity ratio of a set $U$ with respect to an integer $k$ is given by

$$
\gamma_{U,k} := \min_{L,S:L \cap S = \emptyset, L \subseteq U, |S| \leq k} \gamma_{L,S}.
$$
Relax the previous definitions

**Definition (Submodularity Ratio (Das-Kempe ’11))**

Let $S, L \subset [p]$ be two disjoint sets, and $f(\cdot) : [p] \to \mathbb{R}$. The submodularity ratio of $L$ with respect to $S$ is given by

$$
\gamma_{L,S} := \frac{\sum_{j \in S} [f(L \cup \{j\}) - f(L)]}{f(L \cup S) - f(L)}.
$$

The submodularity ratio of a set $U$ with respect to an integer $k$ is given by

$$
\gamma_{U,k} := \min_{L,S:L \cap S = \emptyset, L \subseteq U, |S| \leq k} \gamma_{L,S}.
$$

$f(\cdot)$ submodular $\iff \gamma_{U,k} \geq 1, \ \forall U, k$
A function $l : \mathbb{R}^p \to \mathbb{R}$ is said to be restricted strong concave with parameter $m_\Omega$ and restricted smooth with parameter $M_\Omega$ if for all $x, y \in \Omega \subset \mathbb{R}^p$, 

$$-rac{m_\Omega}{2} \|y - x\|^2 \geq l(y) - l(x) - \langle \nabla l(x), y - x \rangle \geq -\frac{M_\Omega}{2} \|y - x\|^2$$
Normalized support function:

\[ f(S) = \max_{\beta SC=0} l(\beta) - l(0) \]

Theorem (RSC/RSM Implies Weak Submodularity)

\( l(.) \) is \( M \)-smooth and \( m \)-strongly concave on all \( (|U| + k) \)-sparse vectors. Then the submodularity ratio \( \gamma_{U,k} \) is lower bounded by

\[ \gamma_{U,k} \geq \left( \frac{m}{M} \right)^2. \]
Main Theorem

Normalized support function:

\[ f(S) = \max_{\beta SC = 0} l(\beta) - l(0) \]

Theorem (RSC/RSM Implies Weak Submodularity)

\( l(.) \) is \( M \)-smooth and \( m \)-strongly concave on all \( (|U| + k) \)-sparse vectors. Then the submodularity ratio \( \gamma_{U,k} \) is lower bounded by

\[ \gamma_{U,k} \geq \left( \frac{m}{M} \right)^2. \]

- Does NOT imply submodularity
Three greedy algorithms:
- Oblivious (Univariate)
- Orthogonal Matching Pursuit (Approximate Greedy)
- Forward Stepwise Selection (Greedy)

If $l(\cdot)$ is a log-likelihood function for a statistical model, guarantees for greedy feature selection
Oblivious Selection

Rank features individually by their improvement over a null model

- **Input**: sparsity parameter $k$, set function $f(\cdot)$
- for $i = 1 \ldots p$
  - $v[i] \leftarrow f(\{i\})$
- $S_k \leftarrow$ indices corresponding to the top $k$ values of $v$
- **Output**: $S_k$, $f(S_k)$. 
Theorem (Oblivious Algorithm Guarantee)

\( l(\cdot) \) is \( M \)-smooth and \( m \)-strongly concave on all \( k \)-sparse vectors. Let \( f^{OBL} \) be the value at the set selected by the Oblivious algorithm, and let \( f^{OPT} \) be the optimal value over all sets of size \( k \).

\[
f^{OBL} \geq \max \left\{ \frac{m^2}{kM^2}, \frac{m^4}{4M^4} \right\} f^{OPT}.
\]
Choose the next feature with the largest marginal gain

- **Input:** sparsity parameter \( k \), set function \( f(\cdot) \)
- \( S_0^G \leftarrow \emptyset \)
- for \( i = 1 \ldots k \)
  - \( s \leftarrow \arg \max_{j \in [p] \setminus S_{i-1}} f(S_{i-1} \cup \{j\}) - f(S_{i-1}) \)
  - \( S_i^G \leftarrow S_{i-1}^G \cup \{s\} \)
- **Output:** \( S_k^G, f(S_k^G) \)
Forward Stepwise Selection

**Theorem (Forward Stepwise Algorithm Guarantee)**

$l$ is $M$-smooth and $m$-strongly concave on all $2k$-sparse vectors. Let $S^G_k$ be the set selected by the FS algorithm and $S^*$ be the optimal set of size $k$ corresponding to values $f^G$ and $f^{OPT}$. Then

$$f^G \geq \left(1 - e^{-\gamma S^G_k,k}\right) f^{OPT} \geq \left(1 - e^{-\left(m/M\right)^2}\right) f^{OPT}.$$
Orthogonal Matching Pursuit

Choose the next feature that correlates the most with residual

- **Input:** sparsity parameter $k$, objective function $l(\cdot)$
- $S_0^P \leftarrow \emptyset$
- $r \leftarrow \nabla l(0)$
- for $i = 1 \ldots k$
  - $s \leftarrow \text{arg max}_j |\langle e_j, r \rangle|$
  - $S_i^P \leftarrow S_{i-1}^P \cup \{s\}$
  - $\beta(S_i^P) \leftarrow \text{argmax}_{\beta: \text{supp}(\beta) \subseteq S_i^P} l(\beta)$
  - $r \leftarrow \nabla l(\beta(S_i^P))$
- **Output:** $S_k^P$, $l(\beta(S_k^P))$
Orthogonal Matching Pursuit

**Theorem (OMP Algorithm Guarantee)**

Function $l$ is $M$-smooth and $m$-strongly concave on all $2k$-sparse vectors. Let $S^k_P$ be the set of features selected by the OMP algorithm and $S^k$ be the optimal feature set on $k$ variables corresponding to values $f^{OMP}$ and $f^{OPT}$. Then

$$f^{OMP} \geq \left(1 - e^{-\frac{m}{4M}} \gamma_{S^P, k}\right) f^{OPT} \geq \left(1 - e^{-\frac{m^3}{4M^3}}\right) f^{OPT}.$$


Run algorithms for $r > k$ steps:
Run algorithms for $r > k$ steps:

**Corollary**

Let $f^{P+}$ denote the solution obtained after $r$ iterations of the OMP algorithm, and let $f^{OPT}$ be the objective at the optimal $k$-subset of features. Let $\gamma = \left(\frac{m}{4M}\right)\gamma_{SP,r,k}$ be the submodularity ratio associated with the output of $f^{P+}$ and $k$. Then

$$f^{P+} \geq (1 - e^{-\gamma(r/k)}) f^{OPT}.$$
Run algorithms for $r > k$ steps:

**Corollary**

Let $f^{P+}$ denote the solution obtained after $r$ iterations of the OMP algorithm, and let $f^{OPT}$ be the objective at the optimal $k$-subset of features. Let $\gamma = (m/4M)\gamma_{S^r_k}$ be the submodularity ratio associated with the output of $f^{P+}$ and $k$. Then

$$f^{P+} \geq (1 - e^{-\gamma(r/k)})f^{OPT}.$$ 

- $r = ck \quad \rightarrow \quad (1 - e^{-c\gamma})$-approximation
- $r = k \log n \quad \rightarrow \quad (1 - n^{-\gamma})$-approximation
Experiments

- **Synthetic data**: Correlated design matrix (AR process), true support is normalized $\pm 1$ Bernoulli, 50 of 200 features
  - Response computed with logistic model
  - 600 training and test samples
- **Real data**: RCV1 binary text classification dataset
  - $n = 10,000$, $p = 47,236$, $k = 700$
Experiments

- Synthetic data: Correlated design matrix (AR process), true support is normalized ±1 Bernoulli, 50 of 200 features
  - Response computed with logistic model
  - 600 training and test samples
- Real data: RCV1 binary text classification dataset
  - \( n = 10,000, \quad p = 47,236, \quad k = 700 \)
- Fit logistic regression, compare to 3 additional algorithms:
  - Forward-Backward greedy
  - Lasso (\( \ell_1 \)-regularization)
  - Lasso support selection + final unregularized regression
Results: Synthetic (20 runs)

Logistic Regression Performance

Number of Features Selected (50 true, 200 total)

Generalization Accuracy

Oblivious, OMP, Lasso, Lasso-Pipeline, FS, FoBa
Results: Synthetic (20 runs)

Logistic Regression Training Support Recovery

Number of Features Selected (50 true, 200 total)

Area Under ROC

Percent Support Recovered

Oblivious, OMP, Lasso, max

FS, FoBa
Results: RCV1

Logistic Regression Performance

Number of Features Selected
0
100
200
300
400
500
600
700

Normalized Log Likelihood
0
100
200
300
400
500
600
700

Generalization Accuracy

Oblivious
OMP
Lasso
Lasso-Pipeline

Number of Features Selected
0
100
200
300
400
500
600
700

Generalization Accuracy

Oblivious
OMP
Lasso
Lasso-Pipeline
Conclusions

- Extend submodularity ratio framework to general likelihood functions
- RSC/RSM imply weak submodularity
- New bounds for Oblivious, OMP, and Forward Stepwise Regression, independent of specific model
Conclusions

- Extend submodularity ratio framework to general likelihood functions
- RSC/RSM imply weak submodularity
- New bounds for Oblivious, OMP, and Forward Stepwise Regression, independent of specific model

Thank you!