Restricted Strong Convexity Implies Weak Submodularity

Sahand Negahban[†], Alex Dimakis[⋆]
Ethan R. Elenberg[⋆], Rajiv Khanna[⋆]

[⋆]UT Austin,
Department of Electrical and Computer Engineering
[†]Yale University,
Department of Statistics

Set Function Optimization

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

Set Function Optimization

- Examples:
  - Data summarization (k-medians, k-medoids)
  - Subset cover
  - Sparse regression
- k-medoids: given $\mathsf{V} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$

$$\operatorname*{argmax}_{\mathsf{S}:|\mathsf{S}|\leq k} \max_{\pi:\mathsf{V}\mapsto\mathsf{S}} \sum_{j=1}^{n} -\|x_{\pi(j)} - x_j\|_1$$

Set Function Optimization

- Examples:
  - Data summarization (k-medians, k-medoids)
  - Subset cover
  - Sparse regression
- k-medoids: given $\mathsf{V} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$

$$\operatorname*{argmax}_{\mathsf{S}:|\mathsf{S}|\leq k} \max_{\pi:\mathsf{V}\mapsto\mathsf{S}} \sum_{j=1}^{n} -\|x_{\pi(j)} - x_j\|_1$$

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

$$\operatorname*{argmax}_{\mathsf{S}:|\mathsf{S}|\leq k} f(\mathsf{S})$$

Subset (Support) Selection

- High-dimensional statistics: $p \gg n$
- Variable selection
- Lasso, Graphical Lasso, sparse PCA
- Reduce to lower-dimensional structure
- Sparse optimization: goal to maximize $l(\beta)$

$$\max_{\mathsf{S}||\mathsf{S}|\leq k} \max_{\beta_{\mathsf{S}^c}=0} l(\beta) - l(0)$$

- e.g. $l(\beta) =$ log-likelihood
- $f(\mathsf{S}) = \max_{\beta_{\mathsf{S}^c}=0} l(\beta) - l(0)$

Computational Challenges

- 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
- Greedy heuristics
  - OMP, CoSaMP, Forward Stagewise/Stepwise Selection, ...
  - Theoretical guarantees under structural assumptions
  - Zhang; Cai and Wang; Needell and Tropp; Jalali et. al.
- Convex relaxations
  - Algorithm converges without any assumptions
  - Can provide theoretical guarantees
  - In practice, greedy methods perform as well or better

Computational Answers for Sparse Regression Problems

- Das and Kempe ('11): Use **weak submodularity** to provide guarantees for greedy methods under *linear* regression and RSC
- Bach ('13): Use submodularity with suppressors
- Krause and Cevher ('10): Use submodularity with incoherence
- **This talk:** Guarantees for general, greedy support selection
  - Connect weak submodularity to Restricted Strong Convexity/Smoothess

Submodular Functions

- Analogous to convex, concave functions
- *Diminishing Returns*: if $\mathsf{A} \subseteq \mathsf{B}$ then

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

- f monotone: $f(\mathsf{A} \cup \{x\}) \geq f(\mathsf{A})$

Submodular Functions

- Analogous to convex, concave functions
- *Diminishing Returns*: if $\mathsf{A} \subseteq \mathsf{B}$ then

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

- f monotone: $f(\mathsf{A} \cup \{x\}) \geq f(\mathsf{A})$

- **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

Submodular Maximization

- *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 $f(\mathsf{S})$,

$$f(\mathsf{S}_k) \geq (1 - {}^1\!/\!_e) f(\mathsf{S}_k^*)$$

- Cannot improve upon $(1 - {}^1\!/\!_e)$ in polynomial time
- Under "incoherence" assumptions, does linear regression satisfy submodularity?

Weak Submodularity

Relax the previous definitions

Weak Submodularity

Relax the previous definitions

Definition (Submodularity Ratio (Das-Kempe '11))

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

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

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

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

Weak Submodularity

Relax the previous definitions

Definition (Submodularity Ratio (Das-Kempe '11))

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

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

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

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

$$f(\cdot) \text{ submodular} \quad \Leftrightarrow \quad \gamma_{\mathsf{U},k} \geq 1, \;\; \forall\, \mathsf{U}, k$$

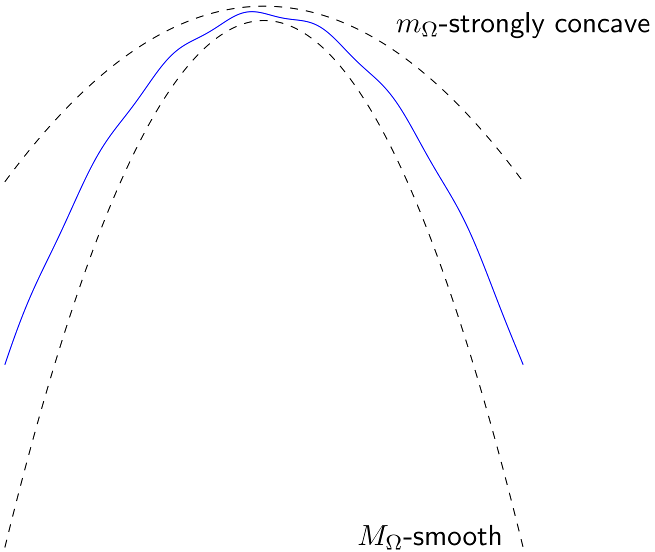
Restricted Strong Convexity/Smoothness

Definition (Restricted Strong Concavity, Restricted Smoothness)

A function $l : \mathbb{R}^p \to \mathbb{R}$ is said to be restricted strong concave with parameter m_Ω and restricted smooth with parameter M_Ω if for all $\mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^p$,

$$-\frac{m_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \geq l(\mathbf{y}) - l(\mathbf{x}) - \langle \nabla l(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq -\frac{M_\Omega}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

Restricted Strong Convexity/Smoothness



Main Theorem

Normalized support function:

$$f(\mathsf{S}) = \max_{\beta_{\mathsf{S}^c}=0} l(\beta) - l(0)$$

Theorem (RSC/RSM Implies Weak Submodularity)

$l(.)$ is M-smooth on all $(|\mathsf{U}|+1)$-sparse vectors, and m-strongly concave on all $(|\mathsf{U}|+k)$-sparse vectors. Then the submodularity ratio $\gamma_{\mathsf{U},k}$ is lower bounded by

$$\gamma_{\mathsf{U},k} \geq \frac{m}{M}\ .$$

Main Theorem

Normalized support function:

$$f(\mathsf{S}) = \max_{\beta_{\mathsf{S}^c}=0} l(\beta) - l(0)$$

Theorem (RSC/RSM Implies Weak Submodularity)

$l(.)$ is M-smooth on all $(|\mathsf{U}|+1)$-sparse vectors, and m-strongly concave on all $(|\mathsf{U}|+k)$-sparse vectors. Then the submodularity ratio $\gamma_{\mathsf{U},k}$ is lower bounded by

$$\gamma_{\mathsf{U},k} \geq \frac{m}{M} .$$

- Does NOT imply submodularity

Main Theorem

Normalized support function:

$$f(\mathsf{S}) = \max_{\beta_{\mathsf{S}^c}=0} l(\beta) - l(0)$$

Theorem (RSC/RSM Implies Weak Submodularity)

$l(.)$ is M-smooth on all $(|\mathsf{U}|+1)$-sparse vectors, and m-strongly concave on all $(|\mathsf{U}|+k)$-sparse vectors. Then the submodularity ratio $\gamma_{\mathsf{U},k}$ is lower bounded by

$$\gamma_{\mathsf{U},k} \geq \frac{m}{M} .$$

- Does NOT imply submodularity
- Matches Das-Kempe '11 in the case of linear regression

Greedy Algorithm Guarantees

- 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 $i = 1 \ldots p$
  - $\mathbf{v}[i] \leftarrow f(\{i\})$
- $\mathsf{S}_k \leftarrow$ indices corresponding to the top k values of $\mathbf{v}$
- **Output:** S_k**,** $f(\mathsf{S}_k)$**.**

Oblivious Selection

Theorem (Oblivious Algorithm Guarantee)

$l(.)$ 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.

$$f^{OBL} \geq \max\left\{\frac{m}{kM}, \frac{3m^2}{4M^2}, \frac{m^3}{M^3}\right\} f^{OPT}.$$

Forward Stepwise Selection

Choose the next feature with the largest marginal gain

- **Input:** sparsity parameter k, set function $f(\cdot)$
- $\mathsf{S}_0^G \leftarrow \emptyset$
- for $i = 1 \ldots k$
  - $s \leftarrow \arg\max_{j \in [p] \setminus \mathsf{S}_{i-1}} f(\mathsf{S}_{i-1}^G \cup \{j\}) - f(\mathsf{S}_{i-1}^G)$
  - $\mathsf{S}_i^G \leftarrow \mathsf{S}_{i-1}^G \cup \{s\}$
- **Output:** S_k^G**,** $f(\mathsf{S}_k^G)$**.**

Forward Stepwise Selection

Theorem (Forward Stepwise Algorithm Guarantee)

l is M-smooth and m-strongly concave on all $2k$-sparse vectors. Let S_k^G 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

$$f^G \geq \left(1 - e^{-\gamma_{\mathsf{S}_k^G, k}}\right) f^{OPT} \geq \left(1 - e^{-m/M}\right) f^{OPT}.$$

Orthogonal Matching Pursuit

Choose the next feature that correlates the most with residual

- **Input:** sparsity parameter k, objective function $l(\cdot)$
- $\mathsf{S}_0^P \leftarrow \emptyset$
- $\mathbf{r} \leftarrow \nabla l(0)$
- for $i = 1 \ldots k$
  - $s \leftarrow \arg\max_j |\langle e_j, \mathbf{r} \rangle|$
  - $\mathsf{S}_i^P \leftarrow \mathsf{S}_{i-1}^P \cup \{s\}$
  - $\boldsymbol{\beta}^{(\mathsf{S}_i^P)} \leftarrow \mathrm{argmax}_{\boldsymbol{\beta}:\mathrm{supp}(\boldsymbol{\beta}) \subseteq \mathsf{S}_i^P} \, l(\boldsymbol{\beta})$
  - $\mathbf{r} \leftarrow \nabla l(\boldsymbol{\beta}^{(\mathsf{S}_i^P)})$
- **Output:** S_k^P, $l(\boldsymbol{\beta}^{(\mathsf{S}_k^P)})$

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

$$f^{OMP} \geq \left(1 - e^{-(3m/4M)\gamma_{\mathsf{S}_k^P,k}}\right) f^{OPT} \geq \left(1 - e^{-3m^2/4M^2}\right) f^{OPT}.$$

Improving Bounds

Run algorithms for $r > k$ steps:

Improving Bounds

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 $\gamma = ({}^{3m}/_{4M})\gamma_{\mathsf{S}_r^P, k}$ be the submodularity ratio associated with the output of f^{P+} and k. Then

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

Improving Bounds

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 $\gamma = (3m/4M)\gamma_{\mathsf{S}_r^P,k}$ 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

Bounding parameter recovery

- $f(S_r^G) = l(\widehat{\beta}^G) - l(0)$

Corollary

Take any k sparse vector and denote it β^*. Then

$$\|\widehat{\beta}^G - \beta^*\|_2^2 \leq (e^{-\gamma(r/k)})l(0) + \frac{4(r+k)\|\nabla l(\theta^*)\|_\infty^2}{m^2}$$

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

- `eelenberg.github.io/weak-submodular-preprint.pdf`

Thank you!

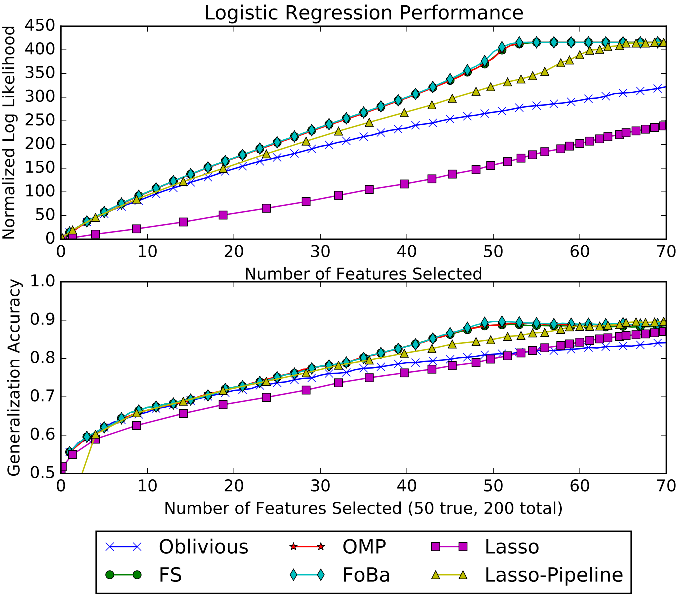
Experiments

- 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
  - $n = 10{,}000, \quad p = 47{,}236, \quad k = 700$

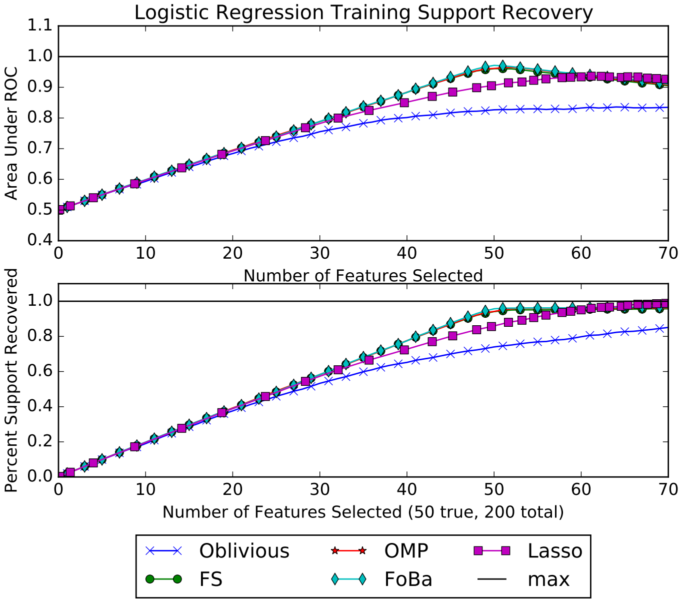
Experiments

- 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 (ℓ_1-regularization)
  - Lasso support selection + final unregularized regression

Results: Synthetic (20 runs)



Results: Synthetic (20 runs)



Results: RCV1