Restricted Strong Convexity implies weak Submodularity

Sahand Negahban

Yale
Department of Statistics

Joint work with: Alex Dimakis*, Ethan R Elenberg*, and Rajiv Khanna*
*UT Austin, Department of Electrical and Computer Engineering

Set function optimization

- Many problems can be cast as an optimization over a finite set
- Examples:
  - data summarization: K-medians or K-mediods
  - subset cover
  - most explanatory variables
- Take $\mathsf{V} = \{1, 2, \ldots, p\}$ and a set function $f : 2^{\mathsf{V}} \mapsto \mathbb{R}$
- Goal:

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

- K-mediods: given $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$

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

Subset selection

- high-dimensional statistics: $n \gg p$
- variable selection
- Lasso, Graphical Lasso, sparse PCA
- reduce to lower-dimensional structure
- sparse optimization: goal to maximize $l(\beta)$

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

- e.g. $l(\beta) = -\text{log-likelihood}$

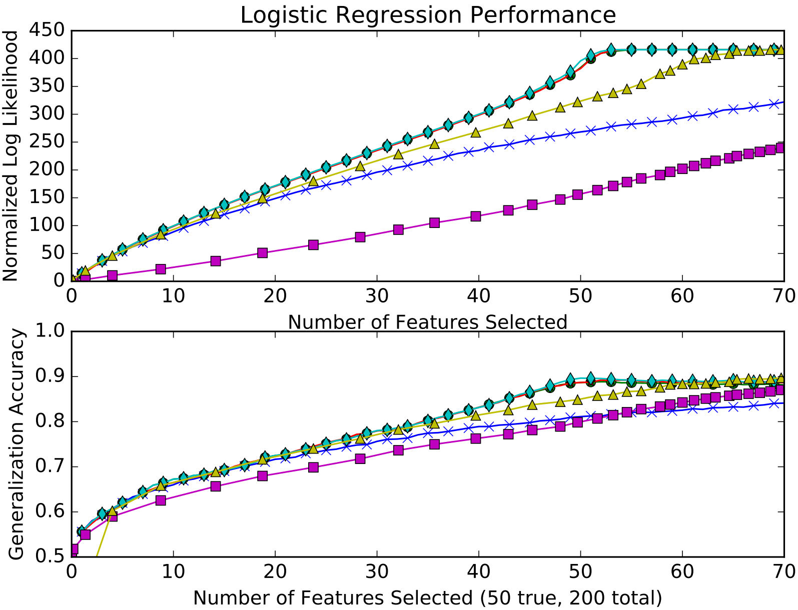
Computational challenges

- set function optimization is in general NP-complete
- subset selection for sparse linear regression, k-medians, subset-cover, etc...
- sometimes subset selection for regression is tractable
- what settings for general problems?
- what structural assumptions can we exploit in order to alleviate computational challenges?
- for sparse linear regression use ideas such as Restricted Isometry Property, Restricted Strong Convexity, or convex relaxations

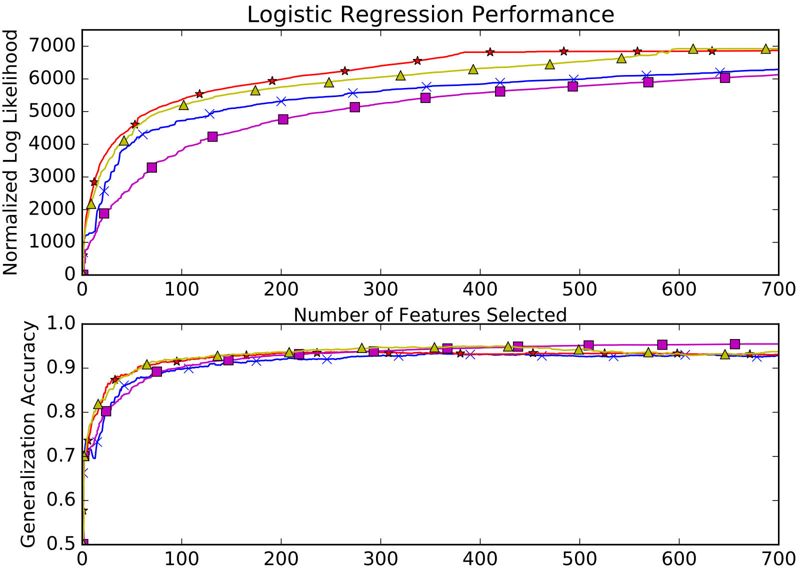
Computational Answers for Sparse Regression Problems

- Long line of work
- Early methods based on Greedy heuristics (OMP, COSAMP, Forward Stagewise Selection, etc...)
- Theoretical guarantees under structural assumptions: Zhang; Needell and Tropp; Jalali et. al.
- Das and Kempe use weak submodularity to provide guarantees for Greedy methods under linear regression
- More recent focus on convex relaxations
  - algorithm converges without any assumptions
  - can provide theoretical guarantees
- In practice heuristic greedy methods perform as well or better

Synthetic Experimental Results



RCV1 Binary Text Classification



$n = 10,000, p = 47,236, k = 700$

Submodular Functions

- Submodular functions analogous to convex ones
- "diminishing rewards" if $A \subset B$ then

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

- f monotone: $f(A \cup \{x\}) \geq f(A)$
- k-medians, k-mediods, column subset selection under log-det maximization all monotone submodular
- Generalized Linear Model (GLM) set function generally *not* submodular
    - Logistic Regression, Linear Regression, Poisson Regression

Submodular Optimization

- *maximize* a submodular function
- greedy optimization is a heuristic
  - add elements to set that improves result the most
- for submodular set function $f(S)$, can prove

$$f(S_k) \geq (1 - 1/e) f(S_k^*)$$

- improving upon the $(1 - 1/e)$ is NP-complete in general
- under strong "suppressor" condition, linear regression satisfies submodularity

Weak Submodularity

- relax submodularity

Weak Submodularity

- relax submodularity

Definition (Submodularity Ratio (Das and Kempe))

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

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

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 $\mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^p$,

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

Main Theorem

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

Theorem (RSC/RSM Implies Weak Submodularity)

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

$$\gamma_{\mathsf{U},k} \geq \left(\frac{m}{M}\right)^2$$

Main Theorem

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

Theorem (RSC/RSM Implies Weak Submodularity)

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

$$\gamma_{\mathsf{U},k} \geq \left(\frac{m}{M}\right)^2$$

- does NOT imply submodularity

Pure Greedy Optimization for Sparse Optimization

- **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**,** $f(\mathsf{S}_k^G)$**.**

Approximation Guarantees

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

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

Orthogonal Matching Pursuit (Approximate greedy)

- **Input:** sparsity parameter k, observations $\mathbf{X}$, $\mathbf{y}$, 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, $l(\boldsymbol{\beta}^{(\mathsf{S}_k^P)})$

Approximation Guarantees for OMP

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

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

Improving bounds

- Can improve approximation results

Improving bounds

- Can improve approximation results

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

- Can improve approximation results

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

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

e.g. $r = ck$ corresponds to a $(1 - e^{-c\gamma})$-approximation, and setting $r = k \log n$ corresponds to a $(1 - n^{-\gamma})$-approximation.

Conclusions

- Extend sub-modularity ratio framework to general loss functions
- RSC/RSM imply weak submodularity
- New bounds for OMP and Forward Stage-wise Regression independent of specific model

Conclusions

- Extend sub-modularity ratio framework to general loss functions
- RSC/RSM imply weak submodularity
- New bounds for OMP and Forward Stage-wise Regression independent of specific model

Thank you!