Greedy Optimization Over Infinite Atoms Low-rank matrix estimation, submodularity, and super-resolution

Sahand Negahban

Yale University
Department of Statistics and Data Science

Joint work with: Rajiv Khanna, Ethan R. Elenberg, and Alex Dimakis

Sparse Approximations

- Estimate from set $\mathcal{C}_k = \{\sum_{i=1}^{r} c_i a_i \mid a_i \in \mathcal{A}\}$

$$\widehat{\Theta} = \arg\max_{\Theta \in \mathcal{C}_k} \ell(\Theta)$$

- Optimization over atomic sets or dictionaries ...,DeVore and Temlyakov '96; Barron, Cohen, Dahmen, DeVore '08; Chandrasekaran, Recht, Parrilo, Willsky '10; Candes and Fernandez-Granda '12; Bhaskar, Tang, Recht '12; Rao, Shah, Wright '15,...
- Examples $\{\exp(-j\omega t)\}$, $\{uv^T \mid \|u\| = \|v\| = 1\}$, $\{e_i\}$

Set Function Optimization

- Fix $\mathsf{S} = \{a_1, a_2, \ldots, a_k\}$ let $\mathcal{C}_\mathsf{S} = \{\sum_{i=1}^{k} c_i a_i \mid a_i \in \mathsf{S}\}$
- $\widehat{\Theta} = \arg\min_{\Theta \in \mathcal{C}_\mathsf{S}} \ell(\Theta)$
- Define set function $f : 2^{\mathcal{A}} \mapsto \mathbb{R}$

$$f(\mathsf{S}) = \max_{\Theta \in \mathcal{C}_\mathsf{S}} \ell(\Theta) - \ell(0)$$

Set Function Optimization

- Fix $\mathsf{S} = \{a_1, a_2, \dots, a_k\}$ let $\mathcal{C}_\mathsf{S} = \{\sum_{i=1}^k c_i a_i \mid a_i \in \mathsf{S}\}$
- $\widehat{\Theta} = \arg\min_{\Theta \in \mathcal{C}_\mathsf{S}} \ell(\Theta)$
- Define set function $f : 2^{\mathcal{A}} \mapsto \mathbb{R}$

$$f(\mathsf{S}) = \max_{\Theta \in \mathcal{C}_\mathsf{S}} \ell(\Theta) - \ell(0)$$

- Optimize over all sets

$$\mathsf{S}_k^* = \underset{\mathsf{S} \subset 2^{\mathcal{A}} \,|\, |\mathsf{S}| \le k}{\arg\max} f(\mathsf{S})$$

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
- Can we do similar things for our atoms?

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*

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

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

Weak Submodularity

Version for low-rank matrices

Weak Submodularity

Version for low-rank matrices

Definition (Submodularity Ratio)

Let $\mathsf{S}, \mathsf{L} \subset \mathcal{A}$ *be two disjoint sets where the elements of* S *are orthogonal with respect to* L, $|\mathsf{L}| = k$, $|\mathsf{S}| = r$, *and* $f(\cdot)$ *a set function. The submodularity ratio of* L *with respect to* S *is given by*

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

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

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

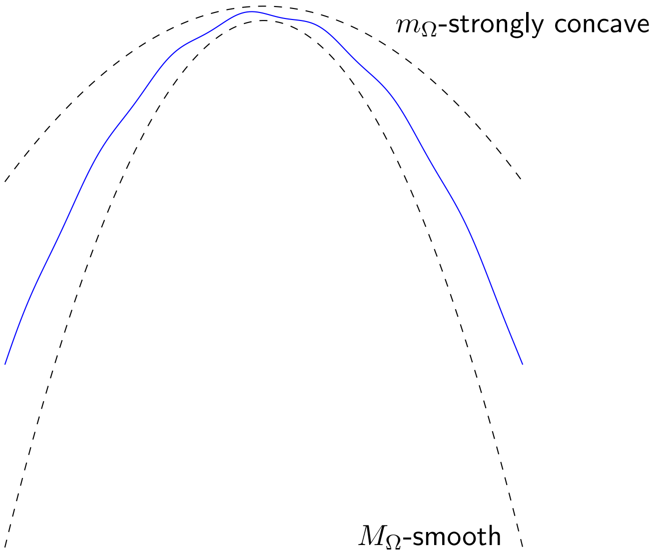
Restricted Strong Convexity/Smoothness

Definition (Restricted Strong Concavity, Restricted Smoothness)

A function $\ell : \mathbb{R}^p \to \mathbb{R}$ *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:

$$f(\mathsf{S}) = \max_{\Theta \in \mathcal{C}_{\mathsf{S}}} \ell(\Theta) - \ell(0)$$

Theorem (RSC/RSM Implies Weak Submodularity)

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

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

Greedy Algorithm Guarantees

- Two greedy algorithms:
  - Orthogonal Matching Pursuit (Approximate Greedy/GECO/Admira*)
  - Forward Stepwise Selection (Greedy)

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

*Lee and Bresler '09; Shalev-Shwartz, Gonen, Shamir '11; like fully corrective Frank-Wolfe; Dudik, Harchaoui, Malick '11; Khanna, Jaggi '16

Forward Stepwise Selection

Choose the next low-rank with the largest marginal gain after re-optimizing

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

Forward Stepwise Selection

Theorem (Forward Stepwise Algorithm Guarantee)

The objective ℓ is M-smooth between rank one matrices and m-strongly concave on all rank $2k$ matrices. 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 rank-one update that correlates the most with gradient of the loss

- **Input:** sparsity parameter k, objective function $\ell(\cdot)$
- $\mathsf{S}_0^P \leftarrow \emptyset$
- $\mathbf{r} \leftarrow \nabla \ell(0)$
- for $i = 1 \ldots k$
  - $s \leftarrow \arg\max_{a \in \mathcal{A}} |\langle a, \mathbf{r} \rangle|$
  - $\mathsf{S}_i^P \leftarrow \mathsf{S}_{i-1}^P \cup \{s\}$
  - $\Theta^{(\mathsf{S}_i^P)} \leftarrow \arg\max_{\Theta \in \mathcal{C}_{\mathsf{S}_i^P}} \ell(\Theta)$
  - $\mathbf{r} \leftarrow \nabla \ell(\boldsymbol{\beta}^{(\mathsf{S}_i^P)})$
- **Output:** S_k^P, $\ell(\boldsymbol{\beta}^{(\mathsf{S}_k^P)})$

Orthogonal Matching Pursuit

Theorem (OMP Algorithm Guarantee)

Objective l is M-smooth across rank one matrices and m-strongly concave on all rank $2k$ matrices. Let S_k^P be the set of features selected by the OMP algorithm and S_k be the optimal feature set on r variables corresponding to values f^{OMP} and f^{OPT}. Then

$$f^{OMP} \geq \left(1 - e^{-m/M}\right) f^{OPT}.$$

Improving Bounds

Run algorithms for $k > r$ steps:

Improving Bounds

Run algorithms for $k > r$ steps:

Corollary

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

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

Improving Bounds

Run algorithms for $k > r$ steps:

Corollary

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

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

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

Comparison to other bounds

- Define atomic norm also norm in total variation with respect to the dictionary

$$\|v\|_{\mathcal{A}} := \inf \left\{ \sum_i |c_i| \ \text{ s.t. } \ v = \sum c_i a_i \right\}$$

- Bounds of the form $\ell(\widehat{\Theta}_k) \geq \ell(\Theta^*) - \epsilon$

Comparison to other bounds

- Bounds of the form $\ell(\widehat{\Theta}_k) \geq \ell(\Theta^*) - \epsilon$
- Three types of bounds

$$\epsilon = \begin{cases} \frac{\|\Theta^*\|_{\mathcal{A}}^2}{k} & \text{general case} \\ \alpha^k \ell(\Theta^*) & \text{strongly concave } \alpha \approx \exp(-\frac{1}{d_1}) \\ \frac{\ell(0) r}{k} & \text{restricted strong concavity} \end{cases}$$

Comparison to other bounds

- Bounds of the form $\ell(\widehat{\Theta}_k) \geq \ell(\Theta^*) - \epsilon$
- Three types of bounds

$$\epsilon = \begin{cases} \frac{\|\Theta^*\|_{\mathcal{A}}^2}{k} & \text{general case} \\ \alpha^k \ell(\Theta^*) & \text{strongly concave } \alpha \approx \exp(-\frac{1}{d_1}) \\ \frac{\ell(0) r}{k} & \text{restricted strong concavity} \end{cases}$$

- Often $\epsilon = O(r(d_1 + d_2)/n)$. In each case k must grow linearly in n or d_1.

Comparison to other bounds

- Bounds of the form $\ell(\widehat{\Theta}_k) \geq \ell(\Theta^*) - \epsilon$
- Three types of bounds

$$\epsilon = \begin{cases} \frac{\|\Theta^*\|_{\mathcal{A}}^2}{k} & \text{general case} \\ \alpha^k \ell(\Theta^*) & \text{strongly concave } \alpha \approx \exp(-\frac{1}{d_1}) \\ \frac{\ell(0) r}{k} & \text{restricted strong concavity} \end{cases}$$

- Often $\epsilon = O(r(d_1 + d_2)/n)$. In each case k must grow linearly in n or d_1.
- Our bound $\epsilon = \exp\left(-\frac{\gamma k}{r}\right)(\ell(\Theta^*) - \ell(0))$

Bounding parameter recovery

Corollary

Take any rank r matrix and denote it Θ^*. Then

$$\|\widehat{\Theta}_k - \Theta^*\|_F^2 \leq (e^{-\gamma(r/k)})\ell(0) + \frac{4(r+k)\|\nabla \ell(\Theta^*)\|_2^2}{\gamma^2}$$

Bounding parameter recovery

Corollary

Take any rank r matrix and denote it Θ^*. Then

$$\|\widehat{\Theta}_k - \Theta^*\|_F^2 \leq (e^{-\gamma(r/k)})\ell(0) + \frac{4(r+k)\|\nabla \ell(\Theta^*)\|_2^2}{\gamma^2}$$

- Other infinite dimensional atoms?

Conclusions

- Use idea of submodularity to understand greedy low-rank matrix optimization
- RSC/RSM imply weak submodularity
- New bounds for greedy low-rank matrix estimation

Conclusions

- Use idea of submodularity to understand greedy low-rank matrix optimization
- RSC/RSM imply weak submodularity
- New bounds for greedy low-rank matrix estimation
- `https://arxiv.org/abs/1703.02721`

Thank you!