On Approximation Guarantees for Greedy Low Rank Approximation

Sahand Negahban

Yale University
Department of Statistics and Data Science

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

Motivation

- Goal: Provide a novel analysis of greedy low rank approximation by establishing connections with combinatorial submodular optimization
- Optimize

$$\max_{\mathsf{rank}(\Theta)\leq r} \ell(\Theta)$$

- Greedily add low-rank components

Informal Result

Goal:

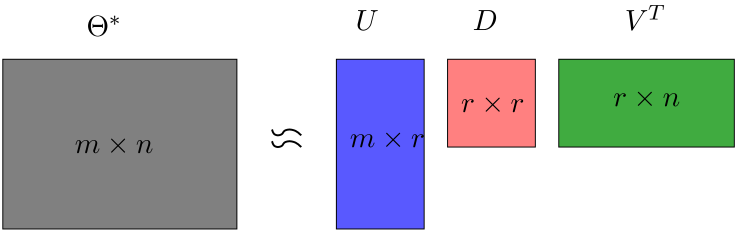
$$\max_{\mathrm{rank}(\Theta) \leq r} \ell(\Theta)$$

We show that

$$\ell(\Theta_k) - \ell(\mathbf{0}) \geq (1 - \exp(-ck/r))(\ell(\Theta^*) - \ell(\mathbf{0}))$$

- Θ_k is obtained by k calls to a greedy algorithm
- $\Theta^\star$ is the best rank r approximation.
- The constant c depends on the properties of the function $\ell(\cdot)$

Example: Low-rank matrix approximation



Set-up: Noisy observations of $\phi(\Theta_{i,j})$ for link function ϕ

Exponential Family PCA, Collins et. al. '02

Atomic 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,...
- Our set $\{uv^T \mid \|u\| = \|v\| = 1\}$
- Low-rank optimization as a set optimization problem over rank-one matrices

Writing a set function

- Given atom selection algorithm. L set of indices of selected atoms
- Take $\mathbf{U}_\mathsf{L}, \mathbf{V}_\mathsf{L}$ by stacking the vectors selected.
- Define set function:

$$f(\mathsf{L}) := \max_{\mathbf{H} \in \mathbb{R}^{|\mathsf{L}| \times |\mathsf{L}|}} \ell(\mathbf{U}_\mathsf{L} \mathbf{H} \mathbf{V}_\mathsf{L}) - \ell(\mathbf{0})$$

- The internal maximization over $\mathbf{H}$ is analogous to fitting weights for chosen support in classic sparsity
- The equivalent set function optimization problem:

$$\max_{\mathsf{S} \leq k} f(\mathsf{S})$$

Submodular functions

Definition

A set function $f(\cdot) : [p] \to \mathbb{R}$ *is submodular if for all* $\mathsf{A}, \mathsf{B} \subseteq [p]$*,*

$$f(\mathsf{A}) + f(\mathsf{B}) \geq f(\mathsf{A} \cup \mathsf{B}) + f(\mathsf{A} \cap \mathsf{B}).$$

Submodular functions

Definition

A set function $f(\cdot) : [p] \to \mathbb{R}$ *is submodular if for all* $\mathsf{A}, \mathsf{B} \subseteq [p]$,

$$f(\mathsf{A}) + f(\mathsf{B}) \geq f(\mathsf{A} \cup \mathsf{B}) + f(\mathsf{A} \cap \mathsf{B}).$$

Normalized: $f(\emptyset) = 0$
Monotone: $\mathsf{A} \subset \mathsf{B} \implies f(\mathsf{A}) \leq f(\mathsf{B})$

Submodular functions

Definition

A set function $f(\cdot) : [p] \to \mathbb{R}$ *is submodular if for all* $\mathsf{A}, \mathsf{B} \subseteq [p]$,

$$f(\mathsf{A}) + f(\mathsf{B}) \geq f(\mathsf{A} \cup \mathsf{B}) + f(\mathsf{A} \cap \mathsf{B}).$$

Normalized: $f(\emptyset) = 0$
Monotone: $\mathsf{A} \subset \mathsf{B} \implies f(\mathsf{A}) \leq f(\mathsf{B})$

Definition

Greedy Selection: $\max_{s \in [p] \setminus \mathsf{S}_{i-1}} f(\mathsf{S}_{i-1} \cup \{s\}) - f(\mathsf{S}_{i-1})$

Submodular functions

Definition

A set function $f(\cdot) : [p] \to \mathbb{R}$ *is submodular if for all* $\mathsf{A}, \mathsf{B} \subseteq [p]$,

$$f(\mathsf{A}) + f(\mathsf{B}) \geq f(\mathsf{A} \cup \mathsf{B}) + f(\mathsf{A} \cap \mathsf{B}).$$

Normalized: $f(\emptyset) = 0$
Monotone: $\mathsf{A} \subset \mathsf{B} \implies f(\mathsf{A}) \leq f(\mathsf{B})$

Definition

Greedy Selection: $\max_{s \in [p] \setminus \mathsf{S}_{i-1}} f(\mathsf{S}_{i-1} \cup \{s\}) - f(\mathsf{S}_{i-1})$

Theorem (Nemhauser 1978)

For normalized monotone submodular functions, greedy selections guarantee $(1 - \frac{1}{e})$ *approximation.*

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*

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

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*

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

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

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

Guarantees $(1 - \frac{1}{e^{\gamma_{\mathsf{G},k}}})$ approximation, where G is the algo output

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*

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

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

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

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

Defined for finite sets

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*

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

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)

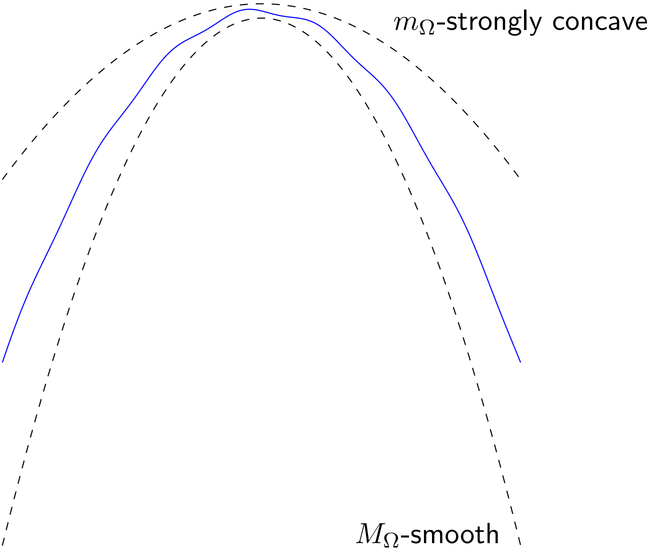
A matrix-variate function ℓ is said to be restricted strong concave with parameter m_Ω and restricted smooth with parameter M_Ω if for all $\mathbf{x}, \mathbf{y} \in \Omega \subset \mathbb{R}^p$,

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

RSC/RSM Assumptions on $\ell(\cdot)$

1. $\ell(\cdot)$ is m_i-strongly concave over matrices of rank i
2. $\ell(\cdot)$ is $\tilde{M}_1$-smooth over $\Omega := \{(\mathbf{X}, \mathbf{Y}) :\ \text{rank}(\mathbf{X} - \mathbf{Y}) = 1\}$.

Restricted Strong Convexity/Smoothness



RSC/RSM $\implies$ weak submodularity

- Recall:
$$f(\mathsf{L}) := \max_{\mathbf{H} \in \mathbb{R}^{|\mathsf{L}| \times |\mathsf{L}|}} \ell(\mathbf{U}_{\mathsf{L}} \mathbf{H} \mathbf{V}_{\mathsf{L}}) - \ell(\mathbf{0})$$
- Lower bounding the submodularity ratio provides provable approximation guarantees

Theorem

If L is set of k rank 1 atoms and up to r additional atoms all orthogonal to all atoms in L are greedily added, then under assumptions 1 and 2,

$$\gamma_{\mathsf{L},r} \geq \frac{m_{r+k}}{\tilde{M}_1}$$

RSC/RSM $\implies$ weak submodularity

- Recall:

$$f(\mathsf{L}) := \max_{\mathbf{H} \in \mathbb{R}^{|\mathsf{L}| \times |\mathsf{L}|}} \ell(\mathbf{U}_{\mathsf{L}} \mathbf{H} \mathbf{V}_{\mathsf{L}}) - \ell(\mathbf{0})$$

- Lower bounding the submodularity ratio provides provable approximation guarantees

Theorem

If L is set of k rank 1 atoms and up to r additional atoms all orthogonal to all atoms in L are greedily added, then under assumptions 1 and 2,

$$\gamma_{\mathsf{L},r} \geq \frac{m_{r+k}}{\tilde{M}_1}$$

Extends the previous result of Elenberg et. al. (2016) to case of matrices. Does NOT imply submodularity

Greedy approximation bounds

- 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

Greedy approximation bounds

- Can plugin γ obtained above to get greedy bounds
- However, greedy is infeasible because of the infinite number of atoms

OMP – Approximation results

- OMP Selection (Shalev-Shwartz et. al. 2011) : Choose the next atom that satisfies:

$$\langle \nabla \ell(\mathbf{B}^{(\mathsf{L})}), \mathbf{u}_s \mathbf{v}_s^\top \rangle \geq \tau \max_{(\mathbf{u},\mathbf{v}) \in (\mathcal{U} \times \mathcal{V}) \perp \mathsf{S}_{i-1}^O} \langle \nabla \ell(\mathbf{B}^{(\mathsf{L})}), \mathbf{u}\mathbf{v}^\top \rangle.$$

Theorem

Let S be the solution set obtained using OMP selections for k iterations, and let $\mathsf{S}^\star$ be the optimum size r support set. Then, under the assumptions 1 and 2,

$$f(\mathsf{S}) \geq \left(1 - \exp(\tau^2 \frac{m_{r+k}}{\tilde{M}_1} \frac{k}{r})\right) f(\mathsf{S}^\star).$$

OMP – Approximation results

- OMP Selection (Shalev-Shwartz et. al. 2011) : Choose the next atom that satisfies:

$$\langle \nabla \ell(\mathbf{B}^{(\mathsf{L})}), \mathbf{u}_s \mathbf{v}_s^\top \rangle \geq \tau \max_{(\mathbf{u},\mathbf{v}) \in (\mathcal{U} \times \mathcal{V}) \perp \mathsf{S}_{i-1}^{O}} \langle \nabla \ell(\mathbf{B}^{(\mathsf{L})}), \mathbf{u}\mathbf{v}^\top \rangle.$$

Theorem

Let S be the solution set obtained using OMP selections for k iterations, and let $\mathsf{S}^\star$ be the optimum size r support set. Then, under the assumptions 1 and 2,

$$f(\mathsf{S}) \geq \left(1 - \exp(\tau^2 \frac{m_{r+k}}{\tilde{M}_1} \frac{k}{r})\right) f(\mathsf{S}^\star).$$

Improves upon earlier bounds by Shalev-Shwartz et. al. by an exponential factor.

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

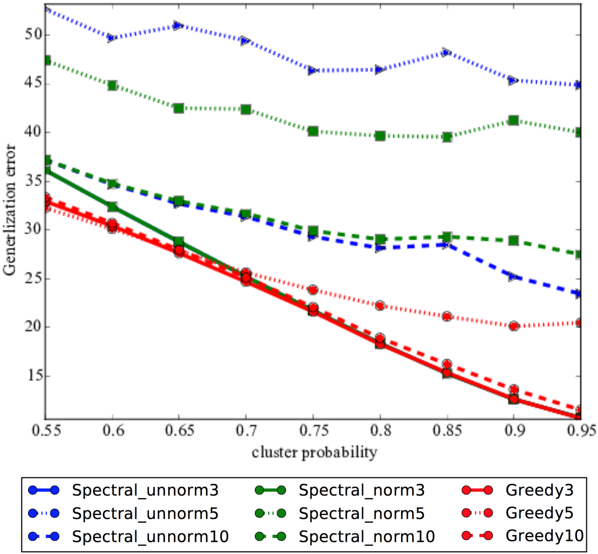
$$\|\widehat{\Theta}_k - \Theta^*\|_F^2 \leq (e^{-\gamma(r/k)})\ell(0) + \frac{4(r+k)\|\nabla\ell(\Theta^*)\|_2^2}{\gamma^2}$$

Experiments - Clustering under Stochastic Block Model

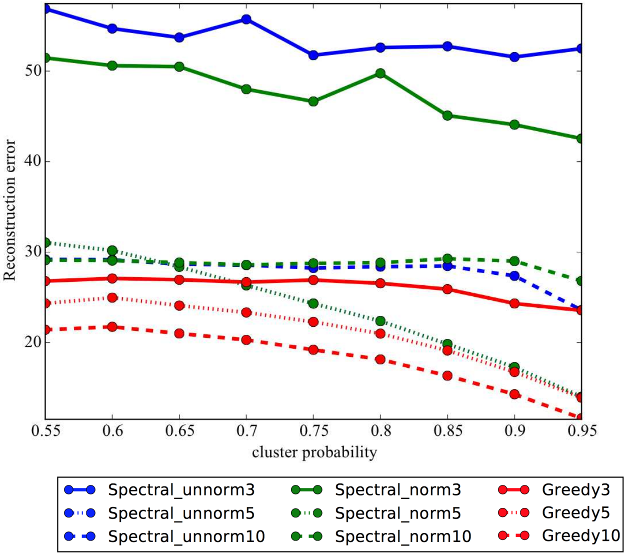
- Form a generating matrix as $\mathbf{C} = p * \mathbf{M} + (1-p) * (\mathbf{1} - \mathbf{M})$, where $\mathbf{M}$ is block diagonal with 1s for nodes in the same cluster, 0s elsewhere
- For each cell $\mathbf{C}_{ij}$, draw a Bernoulli(p)
- The resulting matrix is noisily low rank
- Use greedy selections on

$$\ell(\Theta) = \langle \Theta, \mathbf{X} \rangle - \sum_{i,j} \log G(\Theta_{ij}),$$

Experiments - Clustering (Generalization error)



Experiments - Clustering (Reconstruction error)



Conclusions

- Low rank optimization can be re-interpreted as set optimization over infinite number of atoms
- A greedy algorithm can be used for an efficient search
- We provide new approximation bounds by establishing connections to the weak submodularity.
- Additional Experiments on Word Embeddings in the paper.

Conclusions

- Low rank optimization can be re-interpreted as set optimization over infinite number of atoms
- A greedy algorithm can be used for an efficient search
- We provide new approximation bounds by establishing connections to the weak submodularity.
- Additional Experiments on Word Embeddings in the paper.
- `https://arxiv.org/abs/1703.02721`

Thank you!