CS-GY 6763/CS-UY 3943: Lecture 7
Submodularity

NYU Tandon School of Engineering, R. Teal Witter

Consider a set function $f : 2^{[n]} \to R$.

Example: There are $n = 3$ classes and f represents the knowledge gained from a set of classes.

For $e \in [n]$ and $S \subseteq [n]$, the marginal gain of element e with respect to set S is

$$f(e|S) = f(\{e\} \cup S) - f(S).$$

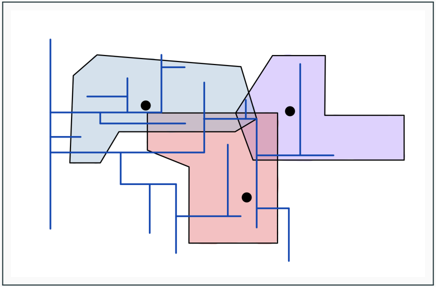
Definition (Submodular set function)

A set function $f : 2^{[n]} \to \mathbb{R}$ is submodular if, for all $e \in [n]$ and $S \subseteq S' \subseteq [n]$,

$$f(e|S) \geq f(e|S').$$

APPLICATION: COVERAGE PROBLEM[1]

In coverage problem, $f(S)$ is the amount of water "covered" by S.

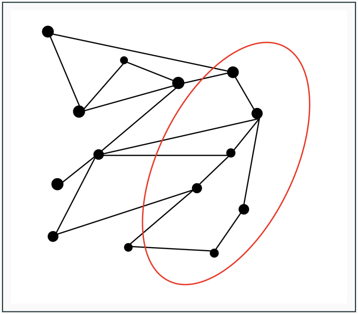


Why is coverage *submodular*?

[1] Finding a maximum of at most k hyper-edges is NP-Hard.

APPLICATION: GRAPH CUT[2]

In graph cut, $f(S)$ is the number of edges between S and $[n] \setminus S$.



Why is graph cut *submodular*?

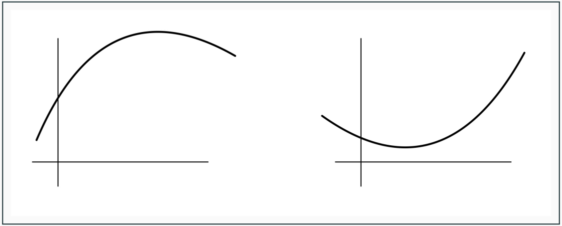
How is graph cut different from set cover?

[2]Finding a maximum graph cut is NP-Hard.

application: others!

- **Combinatorial optimization** (1970-)
  - rank of a matroid
  - submodular flows
- **Algorithmic game theory** (2000-)
  - marketing on networks
  - combinatorial auctions
- **Machine learning** (2005-)
  - document summarization
  - active learning

IS SUBMODULARITY MORE LIKE CONCAVITY OR CONVEXITY?



Arguments for concavity:

- Non-increasing derivative.

Arguments for convexity:

- Max cover and max cut are NP-hard.
- Exact minimization can be done in polynomial time.[3]

[3] M. Grötschel, L. Lovász & A. Schrijver. *Combinatorica* (1981).

Let $f : 2^{[n]} \to \mathbb{R}$ be a normalized, monotone, submodular set function.

We want to find

$$\arg\max_{S \subseteq [n]} f(S) \text{ subject to } |S| \leq k.$$

... any ideas?

Let $\{\} = S_0 \subset S_1 \subset S_2 \subset \ldots \subset S_n = [n]$ with $S_i = \{s_1, s_2, \ldots, s_i\}$.

Theorem (Nemhauser-Wolsey 1981)

Let $f : 2^{[n]} \to \mathbb{R}$ be a normalized, monotone, submodular set function. Fix positive integers ℓ and k. Choose $s_i = \arg\max_{e \in [n] \setminus S_i} f(e|S_i)$. Then

$$f(S_\ell) \geq (1 - e^{-\ell/k}) f(S^*)$$

where

$$S^* = \arg\max_{S:|S|=k} f(S).$$

Claim: $f(S^*) - f(S_i) \leq k\,[f(S_{i+1}) - f(S_i)]$

greedy proof: step 2

Claim: $f(S^*) - f(S_i) \leq k\,[f(S_{i+1}) - f(S_i)]$

If we use the same number of sensors as optimal ($\ell = k$), then we get a $(1 - e^{-1}) \approx .63$ approximate solution.

If we use five times as many sensors as optimal ($\ell = 5k$), then we get a $(1 - e^{-5}) \approx .99$ approximate solution.

What if we have a set function is only *close* to submodular?

Definition (Weak Submodularity)

Fix a positive integer k. A set function $f : 2^{[n]} \to \mathbb{R}$ is γ_k-weakly submodular for k if, for all $S' \in [n]$ and $S \subset [n] \setminus S'$ where $|S| \leq k$,

$$\gamma_k(f) \leq \frac{\sum_{e \in S} f(e|S')}{f(S|S')}.$$

Intuition: How much f can increase by adding a set of size k vs. combined increase of each element.

Definition (Weak Submodularity)

Fix a positive integer k. A set function $f : 2^{[n]} \to \mathbb{R}$ is γ_k-weakly submodular for k if, for all $S' \in [n]$ and $S \subset [n] \setminus S'$ where $|S| \leq k$,

$$\gamma_k(f) \leq \frac{\sum_{e \in S} f(e|S')}{f(S|S')}.$$

Sanity check: What is $\gamma_k(f)$ if f is submodular?

Theorem (Das-Kempe 2011)

Let $f : 2^{[n]} \to \mathbb{R}$ be a normalized, monotone, γ_k-weakly submodular set function. Choose $s_i = \arg\max_{e \in [n] \setminus S_i} f(e|S_i)$. Then

$$f(S_\ell) \geq (1 - e^{-\gamma_k}) f(S^*).$$

We want to minimize $||\mathbf{Ax} - \mathbf{b}||^2$ over $\mathbf{x} \in \mathbb{R}^d$.

Recall from last class that the Hessian $\mathbf{H}$ of least squares regression is $2\mathbf{A}^T\mathbf{A}$ and so

$$\alpha \mathbf{I}_{\mathbf{d}\times\mathbf{d}} \preceq 2\mathbf{A}^T\mathbf{A} \preceq \beta \mathbf{I}_{\mathbf{d}\times\mathbf{d}}$$

where we say $\mathbf{H}$ is α-strongly convex and β-smooth. In particular, we argued $\alpha = \lambda_{\min}(2\mathbf{A}^T\mathbf{A})$ and $\beta = \lambda_{\max}(2\mathbf{A}^T\mathbf{A})$.

Question: What if we can only choose k features?

We want to minimize $||\mathbf{A}\mathbf{x} - \mathbf{b}||^2$ over $\mathbf{x} \in \mathbb{R}^d$ where $k << d$ entries in $\mathbf{x}$ are non-zero. Let $\mathbf{x}'$ be the 'condensed' $k \times 1$ vector and $\mathbf{A}'$ be the 'condensed' $n \times k$ matrix.

Then H' is α'-strongly convex and β'-smooth.

Exercise: Why is $\alpha \leq \alpha'$ and $\beta' \leq \beta$?

Theorem (Elenberg-Khanna-Dimakis-Negahban 2018)

Let $\max_{S:|S|\leq k} f(S) = \max_{\mathsf{x}'} -||\mathsf{A}'\mathsf{x}' - \mathsf{b}||^2$. *Then*

$$\gamma_k \geq \frac{\alpha'}{\beta'}.$$

Corollary: Greedily choosing k features gives a $1 - e^{-\lambda_{\min}(2A'^TA')/\lambda_{\max}(2A'^TA')}$ -approximation to the optimal features.

TAKEAWAYS

- Greedy solutions often work well
- Our tools (bound progress, $(1-1/x)^x \leq 1/e$) are versatile
- Submodularity research is shallow (rather than deep)

Thank you!