Submodular Functions
Properties – Algorithms – Machine Learning

Rémi Gilleron

Inria Lille - Nord Europe & LIFL & Univ Lille

Jan. 12 – revised Aug. 14

History and Context

- Origin: many set functions in economics have the diminishing returns property leading to so-called submodular functions. Optimization problems first studied in the early 70's [Edmonds'71].
- Maximization of submodular functions: a greedy algorithm with performance guarantees [Nemhauser et al'78] with applications to machine learning: information gain, network diffusion, active learning.
- Exact discrete minimization problem: a fully combinatorial, strongly polynomial algorithm for exact minimization of submodular functions [Iwata & Orlin, SODA'09]. Many particular cases.
- Approximate minimization of submodular functions: related to convex optimization problems using the Lovász extension [Lovász'82, Bach'11]. Many applications to machine learning: clustering, subset selection, sparsity.

Motivation for Magnet

- Many functions defined on graphs are submodular: cover functions, cut functions, ...
- Many machine learning settings lead to optimization of submodular functions, thus
- if you can prove that the function to be optimized is submodular, then you get for free: algorithms, performance guarantees, complexity analysis, but, it can be the case that more efficient algorithms exist for particular submodular functions.
- There are strong connections between minimization of submodular functions and convex minimization problems.
- Also, submodular functions allow to define structured sparsity inducing norms.

Plan

Plan

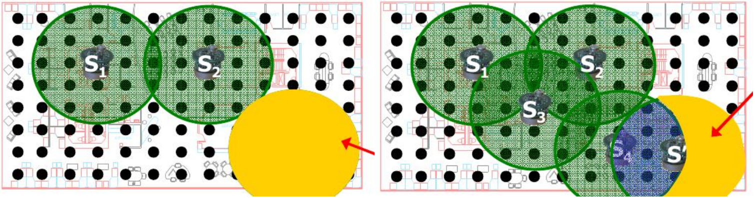
Economics

- A factory has the capability of producing any subset S of a given set E of products.
- Producing subset S induces a setup cost $c(S)$ to make the factory ready to produce S.
- Suppose that we have decided to produce S and we consider whether to add a product e to S, then we will have to pay an additional cost (or marginal cost) $c(S \cup \{e\}) - c(S)$.
- Economics suggests that additional cost is a non-increasing function of S, i.e. adding e to a larger set should produce an additional cost no more than adding e to a smaller set.
- The cost function c is submodular, i.e.

$$\forall S \subset T \subset T \cup \{e\},\ c(S \cup \{e\}) - c(S) \geq c(T \cup \{e\}) - c(T).$$

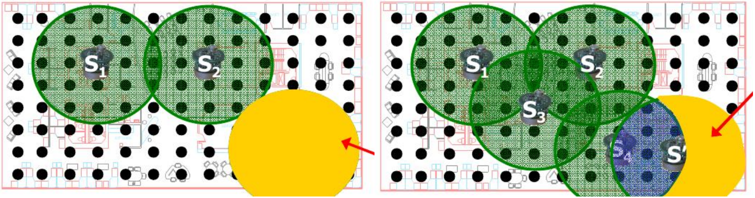
Set Cover

- We have to place sensors in a room with a set E of possible locations. We suppose that each sensor covers a disc of fixed radius.
- Choosing a subset S of possible locations induces a covered area $f(S)$ by sensors placed at S
- Suppose that we have chosen S and T, and a new location s,



Set Cover

- We have to place sensors in a room with a set E of possible locations. We suppose that each sensor covers a disc of fixed radius.
- Choosing a subset S of possible locations induces a covered area $f(S)$ by sensors placed at S
- Suppose that we have chosen S and T, and a new location s,



Then, it is easy to show that the covering function f is submodular:

$$\forall S \subset T \subset T \cup \{s\},\ f(S \cup \{s\}) - f(S) \geq f(T \cup \{s\}) - f(T)$$

Cut function for undirected graphs

Let $G = (V, E)$ be an undirected graph, let S be a subset of V, let $\delta(S) = \{(i,j) \in E \mid i \in S,\ j \notin S\}$, the cut capacity function c is defined by $cut(S) = \#\delta(S)$ (or by $cut(S) = w(\delta(S))$ for weighted graphs). We have

$$cut(S \cup \{v\}) - cut(S) = \#\{(v,j) \in E \mid j \notin S\} - \#\{(v,j) \in E \mid j \in S\},$$
$$cut(T \cup \{v\}) - cut(T) = \#\{(v,j) \in E \mid j \notin T\} - \#\{(v,j) \in E \mid j \in T\},$$

as $\#\{(v,j) \in E \mid j \notin T\} \leq \#\{(v,j) \in E \mid j \notin S\}$, and $\#\{(v,j) \in E \mid j \in T\} \geq \#\{(v,j) \in E \mid j \in S\}$, therefore we get

$$\forall S \subset T \subset T \cup \{v\},\ cut(S \cup \{v\}) - f(S) \geq cut(T \cup \{v\}) - f(T).\ \text{Thus,}$$

the capacity cut function is submodular. Moreover, the capacity cut function is symmetric, i.e. $\forall S\ cut(S) = cut(\bar{S})$, and hence non monotone.

Submodular functions for graphs

- Let $G = (V, E)$ be an undirected graph,
  - cut functions,
  - Let A be a subset of E, the rank $r(A)$ of A is the maximal size of a subset F of A such that (V, F) has no cycle. The rank function r is submodular.
  - Let A be a subset of E, the function nc which gives the number of connected components of the subgraph induced by A is supermodular (its opposite is modular) because $nc(A) = |V| - r(A)$
- Let $G = (V, E)$ be a weighted directed graph, let S be a subset of V, define the cut function $cut^+(S) = \sum_{\{(i,j)\in E | i \in S, j \notin S\}} w_{(i,j)}$, the cut function cut^+ is submodular. The same holds for cut^- (ingoing edges), cut (outgoing and ingoing edges), and also for s,t-cut functions (source and sink in different subsets), i.e. cut functions are submodular

Note: also many functions in matroids are submodular 8-)

Plan

Definitions

We consider E a set, and we set $E = \{1, \ldots, p\}$. We consider real-valued set functions $f : 2^E \to \mathbb{R}$, and $f(\emptyset) = 0$. Set function f is submodular if

Definition with Diminishing Returns Property (first order differences)

(1) $\forall S \subset T \subset T \cup \{e\}$, $f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T)$

Alternative Definition

(2) $\forall X, Y \subseteq E$, $f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y)$

(2) $\Rightarrow$ (1): let $S \subset T \subset T \cup \{e\}$, take $X = S \cup \{e\}$ and $Y = T$, then
$f(S \cup \{e\}) + f(T) \geq f(S \cup \{e\} \cup T) + f((S \cup \{e\}) \cap T)$, thus
$f(S \cup \{e\}) + f(T) \geq f(T \cup \{e\}) + f(S)$, and then
$f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T)$ $\square$

Definitions

We consider E a set, and we set $E = \{1, \ldots, p\}$. We consider real-valued set functions $f : 2^E \to \mathbb{R}$, and $f(\emptyset) = 0$. Set function f is submodular if

Definition with Diminishing Returns Property (first order differences)

(1) $\forall S \subset T \subset T \cup \{e\},\ f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T)$

Alternative Definition

(2) $\forall X, Y \subseteq E,\ f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y)$

(1) $\Rightarrow$ (2): (1) can be rewritten $f(S) - f(T) \leq f(S \cup \{e\}) - f(T \cup \{e\})$, enumerate $Y \setminus X$ as $e_1, e_2, \ldots, e_k$, then

$f(X \cap Y) - f(X) \leq f((X \cap Y) \cup \{e_1\}) - f(X \cup \{e_1\})$

$f(X \cap Y) - f(X) \leq f((X \cap Y) \cup \{e_1, e_2\}) - f(X \cup \{e_1, e_2\})$

...

$f(X \cap Y) - f(X) \leq f((X \cap Y) \cup \{e_1, \ldots, e_k\}) - f(X \cup \{e_1, \ldots, e_k\})$

$f(X \cap Y) - f(X) \leq f(Y) - f(X \cup Y)$ □

The case of cardinality functions: $f(S)$ as a function of $|S|$

Example: let $E = \{1,2\}$, $A \subseteq E$, $x = |A|$, $f(A) = c(x)$, and

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	Submodular ?
$c(x) = x$	0	1	1	2	
$c(x) = x^2$	0	1	1	4	
$c(x) = -x^2$	0	-1	-1	-4	
$c(x) = \sqrt{x}$	0	1	1	$\sqrt{2}$	
$c(x) = Gini(\frac{x}{\lvert E \rvert}) = 4\frac{x}{2}(1 - \frac{x}{2})$	0	1	1	0	

submodular: $\forall S \subset T \subset T \cup \{e\},\ f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T)$

The case of cardinality functions: $f(S)$ as a function of $|S|$

Example: let $E = \{1,2\}$, $A \subseteq E$, $x = |A|$, $f(A) = c(x)$, and

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	Submodular ?
$c(x) = x$	0	1	1	2	modular
$c(x) = x^2$	0	1	1	4	No [1]
$c(x) = -x^2$	0	-1	-1	-4	Yes
$c(x) = \sqrt{x}$	0	1	1	$\sqrt{2}$	Yes
$c(x) = Gini(\frac{x}{\lvert E\rvert}) = 4\frac{x}{2}(1 - \frac{x}{2})$	0	1	1	0	Yes

submodular: $\forall S \subset T \subset T \cup \{e\},\ f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T)$

Theorem

Let a set function f be defined as a cardinality function: $f(A) = c(|A|)$ for a scalar function c. Then, f is submodular if and only if c is concave.

Note: submodular defined as the (discrete) first derivative is non increasing, i.e. the function is concave.

[1] For the square function: $f(\{2\}) - f(\emptyset) = 1$, $f(\{1,2\}) - f(\{1\}) = 3$

Closure properties

Example: let $E = \{1, 2\}$,

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$	Submodular ?
f	0	1	0	1	Yes
g	0	0	1	1	Yes
$h = 3f$	0	3	0	3	Yes
$i = 3f + 2g$	0	3	2	5	Yes
$j = min(f, g)$	0	0	0	1	No
$j = max(f, g)$	0	1	1	1	Yes But[2]

- Submodular functions are closed under positive linear combinations.
- They are not closed under Min and Max

[2]Not in general. For a proof, use concavity of cardinality functions

So far

- Many examples of submodular functions.
- Submodular cardinality functions are concave.
- Submodular functions are closed under positive linear combinations, not under Min and Max.
- If f is submodular, $-f$ is supermodular and its dual d is supermodular, where d is defined by $d(A) = f(E) - f(E \setminus A)$

Let us now consider optimization of submodular functions.
Note that maximizing a submodular function $\equiv$ minimizing its dual

Plan

Exact Submodular Function Maximization

Given a submodular function f, the problem is find $S^* = argmax_{S \subseteq E} f(S)$

- Note that we suppose given a value oracle for f. Also, it can be restricted to a family $\mathcal{F}$ of feasible sets, and then we suppose given a membership oracle for $\mathcal{F}$.
- The max-cut problem for graphs is a particular case. Recall that the decision problem "given a graph G and an integer k, determine whether there is a cut of size at least k in G" is NP-complete.
- In general, (exact) submodular function maximization is NP-hard

Approximate Submodular Function Maximization

Find $S^* = argmax_{S \subseteq E} f(S)$, where f is submodular
Let us denote $opt = max_{S \subseteq E} f(S)$

When the function is non negative, several approximation algorithms with theoretical guarantees exist:

- Random: return R a uniformly random subset of E
  The expected value of $f(R)$ is shown to be greater than $\frac{1}{4} opt$
- Local Search:
  1. $S \leftarrow \{e\}$ where $f(\{e\})$ is maximal over singleton sets
  2. If $\exists e \in E \setminus S \ f(S \cup \{e\}) \geq (1 + \frac{\epsilon}{n^2}) f(S)$, $S \leftarrow S \cup \{e\}$ Goto 2
  3. If $\exists e \in E \setminus S \ f(S \setminus \{e\}) \geq (1 + \frac{\epsilon}{n^2}) f(S)$, $S \leftarrow S \setminus \{e\}$ Goto 2
  4. Return the maximum between $f(S)$ and $f(X \setminus S)$

  The expected value of $f(R)$ is shown to be greater than $(\frac{1}{3} - \frac{\epsilon}{n}) opt$.
  The output is a local optimum of LS, it satisfies: if $I \subseteq S$ or $S \subseteq I$, then $f(I) \leq f(S)$.

Greedy Submodular Function Maximization

Find $S^* = argmax_{S \subseteq E} f(S)$, where f is submodular and non decreasing[3]

Let us denote $opt = max_{S \subseteq E} f(S)$, $opt_k = max_{S \subseteq E, |S| \leq k} f(S)$

Greedy Algorithm for Non Decreasing Submodular Functions

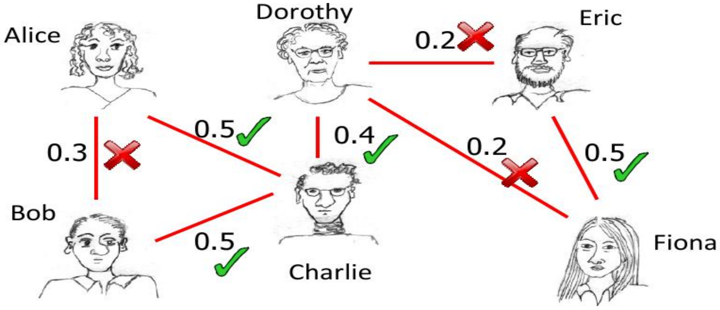
1. $S \leftarrow \emptyset$
2. repeat until $|A| = k$
3. select $e \in E \setminus S$ s.t. $f(S \cup \{e\}) - f(S)$ is maximal, $S \leftarrow S \cup \{e\}$

Theorem (Performance Guarantee) [Nemhauser et al'78]

Let f be a non decreasing submodular function. Greedy output a set S such that $f(S) \geq [1 - (1 - \frac{1}{k})^k] opt_k$, and $f(S) \geq (1 - \frac{1}{e}) opt_k$. The $(1 - \frac{1}{e})$-approximation is optimal: for any ϵ, it is NP-hard to achieve a $(1 - \frac{1}{e} + \epsilon)$-approximation. I.e. Greedy is the best we can do.

[3] $f(S \cup \{e\}) - f(S) \geq 0$

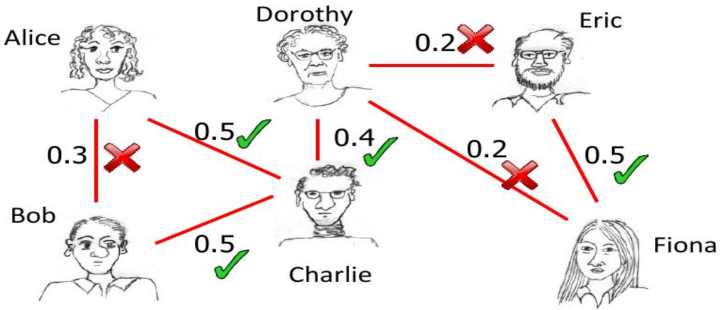
Spread of Influence in Social Networks



Independent Cascade Model

Start with a seed S of active nodes. Unfold in discrete step according to: when u becomes active at step t, it is given a chance to active its neighbours according to p_{uv}.

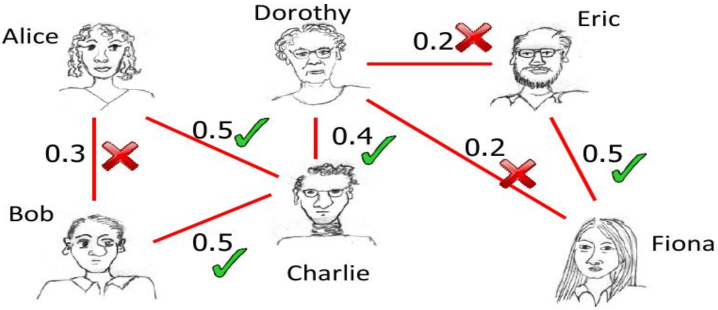
Spread of Influence in Social Networks



Influence Maximization Problem

Given k, find initial seed S of k active nodes of maximum influence $f(S)$, where $f(S)$ is the expected number of nodes active at the end of the process.

Spread of Influence in Social Networks



Theorem [KempeKleinbergTardos'03]

The influence function is submodular. hint: one can flip coins in advance and to get live edges. Then $f(S)$ can be computed from paths of live edges).

As f is non negative and non decreasing, we get that the Greedy algorithm solves the Influence Maximization Problem (find a seed S) within $1 - \frac{1}{e}$ of optimal influence.

So far

- Examples and properties of submodular functions
- Maximization of submodular functions
  - NP-hard problem
  - Approximate maximization algorithms with performance guarantees
  - Applications to influence networks, document summarization, variable selection using information gain, active learning, ...

  and not presented here:
  - Randomized maximization algorithms with performance guarantees
  - Approximate maximization algorithms with constraints (more general than $|S| \leq k$). For instance, when the set of feasible sets has a structure (for instance a matroid).

Let us now consider minimization of submodular functions.

Plan

Minimization of Submodular Functions

- While (exact) maximization is NP-hard, there is an algorithm that computes the minimum of any submodular function $f : \{0,1\}^n \to \mathbb{R}$ in poly(n) time [GrõtschelLovászSchriver'88].
- The combinatorial algorithms are sophisticated: a fully combinatorial, strongly polynomial algorithm for (exactly) minimizing submodular functions by [Iwata & Orlin, SODA'09] in time $O(n^8 logn)$; algorithm in time $O(n^3)$ for symmetric[4] submodular functions; many other particular cases.
- Thus, (approximate) minimization algorithms for submodular functions have been defined using continuous optimization. Most of them are based on the convexity of the Lovàsz extension of any submodular function.
- Many applications to machine learning: clustering (with entropy or cut function), MAP inference in Markov Random Fields, speaker segmentation, image denoising, among others.

[4] $f(A) = f(E \setminus A)$

Lovász Extension

Let f be a set function, equivalently let f be a function from $\{0,1\}^n$ into $\mathbb{R}$, the Lovász Extension g of f is a function from $[0,1]^n$ into $\mathbb{R}$ defined as: let $x \in [0,1]^n$, $g(x) = \sum_{i=0}^{i=n} \lambda_i f(S_i)$, where $\emptyset = S_0 \subset S_1 \subset \ldots \subset S_n$ is a chain such that $\sum_{i=0}^{i=n} \lambda_i 1_{S_i} = x$ and $\sum_{i=0}^{i=n} \lambda_i = 1$ and $\lambda_i \geq 0$. Intuition:

- An input of g is a point in the n-dimensional cube, an input of f is one of the 2^n corners of this cube.
- The cube can be divided into $n!$ simplices where a simplex is the set of points whose coordinates are ordered according to a permutation of $\{1,\ldots,n\}$.
  x=(0.1,0.8,0.3) lies in the simplex $\{x \in \mathbb{R}^3 \mid x_2 \geq x_3 \geq x_1\}$ with corners $(0,0,0)$, $(0,1,0)$, $(0,1,1)$ and $(1,1,1)$.
- Then, each point x lies in a simplex, x is a weighted average of the corners of this simplex, thus define $f(x)$ as the weighted average of the f-values at the corners. f(x) = f(0.1,0.8,0.3) = 0.2 f(0,0,0) + 0.1 f(1,1,1) + 0.2 f(0,1,1) + 0.5 f(0,1,0)

g can be easily computed from f-values on the corners

Submodularity and Convexity

Let $E = \{1, \ldots, n\}$.
Let $A \subseteq E$, let 1_A be the indicator vector of set A.
The base results [Edmonds'71, Lovasz'83] are:

- Lovász extension: every submodular function f induces its Lovász extension g on $[0,1]^n$ such that $g(1_A) = f(A)$ for every $A \subseteq E$.
- Submodularity and Convexity: f is submodular if and only if its Lovász extension g is convex.
- Minimization: let f be a submodular function and let g be its Lovász extension, then $min_{x \in [0,1]^n} g(x) = min_{x \in \{0,1\}^n} g(x) = min_{A \subseteq E} f(A)$ where $1_A = x$.

Structured Sparsity and Submodular Functions

Let us consider estimators obtained by the regularized empirical risk minimization: $argmin_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{i=n} \ell(y_i, w^T x_i) + \lambda\Omega(w)$, where l is a loss function and Ω is a norm.

- sparsity: the ℓ_1-norm is fequently used to promote sparsity
- structured sparsity goes further with the notion of sparsity patterns. For instance, grouped ℓ_1-norm with overlapping groups
- using submodular functions: an alternative is to consider penalty functions of the form $f : w \mapsto f(Supp(w))$ where $Supp(w) = \{i \in E \mid w_i \neq 0\}$ is the support of w. Then, if f is submodular and non decreasing, if the values of all singletons is strictly positive, the function $\Omega : w \mapsto \Omega(w) = g(|w|)$ is a norm and it is the convex envelope of f. Thus new structured sparsity-inducing norms together with minimization algorithms.

Plan

Conclusion

- useful knowledge for machine learners because submodular functions appear naturally in many ML settings,
- also NLP settings 8-),
- they come with well studied maximization and minimization algorithms often with complexity analysis and performance guarantees,
- they allow to define new sparsity-inducing norms.
- Convex or concave ?
- More: constrained optimization algorithms such as cardinality constraints ($|S| \geq k$), optimization algorithms for subclasses of submodular functions, matroids 8-), ...