Minimizing submodular functions

Nikolaj Tatti

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Submodular function

- Assume that f is a set function.
- Given two sets $A \subseteq B$ and $z \notin B$, we have

$$f(A \cup z) - f(A) \geq f(B \cup z) - f(B) \quad .$$

- diminishing returns: the gains of adding z do not increase with the base set.

Submodular function

- Assume that f is a set function.
- Given two sets $A \subseteq B$ and $z \notin B$, we have

$$f(A \cup z) - f(A) \geq f(B \cup z) - f(B) \quad .$$

- diminishing returns: the gains of adding z do not increase with the base set.
- Alternative definition:

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B)$$

- We will assume throughout the whole presentation that $f(\emptyset) = 0$.

Submodular minimization

- Unlike maximization, minimization can be done in polynomial time
- ...but most algorithms are not practical
- We will show an algorithm based on minimal norm theorem
- Even though it's not polynomial...it's more practical
- An iterative algorithm where each step is polynomial
- ...but the number of steps can be large

Vector notations

- We can assume that $U = 1, \ldots, n$.
- This allows us to write x_u, where $x \in \mathbb{R}^n$ and $u \in U$.
- Given, $x \in \mathbb{R}^n$ and a set $A \subseteq U$, we define

$$x(A) = \sum_{a \in A} x_a \quad .$$

Polyhedra, Base polyhedra, Tight sets

Polyhedra $P(f)$ is a set of points $x \in \mathbb{R}^n$ such that

$$x(A) \le f(A) \quad \text{for every} \quad A \subseteq U \quad .$$

Base polyhedra $B(f)$ is a subset of $P(f)$ such that $x(U) = f(U)$.

Given $x \in P(f)$ we say that a set A is tight for x if $x(A) = f(A)$.

Tight set lattice

Lemma

If S and T are tight for $x \in P(f)$, then $S \cup T$ and $S \cap T$ are tight for x.

Tight set lattice

Lemma

If S and T are tight for $x \in P(f)$, then $S \cup T$ and $S \cap T$ are tight for x.

Proof.

$$
\begin{aligned}
x(S \cup T) + x(S \cap T) &\leq f(S \cup T) + f(S \cap T) \\
&\leq f(S) + f(T) = x(S) + x(T) = x(S \cup T) + x(S \cap T)
\end{aligned}
$$

That is,

$$
f(S \cup T) + f(S \cap T) = x(S \cup T) + x(S \cap T) \quad .
$$

Tight set lattice

Lemma

If S and T are tight for $x \in P(f)$, then $S \cup T$ and $S \cap T$ are tight for x.

Proof.

$$
\begin{aligned}
x(S \cup T) + x(S \cap T) &\leq f(S \cup T) + f(S \cap T) \\
&\leq f(S) + f(T) = x(S) + x(T) = x(S \cup T) + x(S \cap T)
\end{aligned}
$$

That is,

$$f(S \cup T) + f(S \cap T) = x(S \cup T) + x(S \cap T) \quad .$$

Since $x(S \cup T) \leq f(S \cup T)$ and $x(S \cap T) \leq f(S \cap T)$, we have $x(S \cup T) = f(S \cup T)$ and $x(S \cap T) = f(S \cap T)$. □

Smallest tight set

Lemma

Fix $x \in B(f)$. Let $u \in U$. There is a tight set D_u such that $u \in D_u$ and any tight set containing u is a superset of D_u.

Smallest tight set

Lemma

Fix $x \in B(f)$. Let $u \in U$. There is a tight set D_u such that $u \in D_u$ and any tight set containing u is a superset of D_u.

Proof.

Since $x \in B(f)$, U is a tight set for x.

If there are two tight sets A and B containing u, then $A \cap B$ is also tight set containing u.

There is a minimal tight set containing u. □

Minimum norm theorem

Theorem

Let

$$x^* = \arg\min_x \{\|x\|_2 \mid x \in B(f)\} \quad .$$

Define

$$A = \{i \mid x_i^* < 0\} \quad .$$

Then A minimizes f.

We prove two claims

- $x^*(A) \leq f(S)$ for any S.
- $x^*(A) = f(A)$.

Proof of Claim 1

$$x^*(A) \leq x^*(A \cap S) \leq x^*(S) \leq f(S) \quad .$$

- first inequality: $x^*_a < 0$ for $a \in A$,
- second inequality: $x^*_a \geq 0$ for $a \notin A$,
- third inequality: $x^* \in P(f)$.

Proof of Claim 2

Select $a \in A$. Let D_a be the smallest tight set using x^*.
We claim that $D_a \subseteq A$.

Proof of Claim 2

Select $a \in A$. Let D_a be the smallest tight set using x^*.
We claim that $D_a \subseteq A$.
Assume otherwise: there is $b \in D_a$ such that $x_b^* \geq 0$.
We can increase x_a^* and decrease x_b^* by $\epsilon > 0$. Let x' be the new vector.

- Let S a non-tight set for x^*. Then $x'(S) \leq x^*(S) + \epsilon < f(S)$, if we select ϵ small enough.
- Let S a tight set for x^*. If $a \in S$, then since $D_a \subseteq S$, we have $x'(S) = x^*(S) = f(S)$. If $a \notin S$, then $x'(S) \leq x^*(S)$.

Proof of Claim 2

Select $a \in A$. Let D_a be the smallest tight set using x^*.
We claim that $D_a \subseteq A$.
Assume otherwise: there is $b \in D_a$ such that $x^*_b \geq 0$.
We can increase x^*_a and decrease x^*_b by $\epsilon > 0$. Let x' be the new vector.

- Let S a non-tight set for x^*. Then $x'(S) \leq x^*(S) + \epsilon < f(S)$, if we select ϵ small enough.
- Let S a tight set for x^*. If $a \in S$, then since $D_a \subseteq S$, we have $x'(S) = x^*(S) = f(S)$. If $a \notin S$, then $x'(S) \leq x^*(S)$.

That is, $x' \in B(f)$, also $\|x'\| < \|x^*\|$.

So, we must have $D_a \subseteq A$.

Proof of Claim 2

Note that

$$A \subseteq \bigcup_{a \in A} D_a \subseteq \bigcup_{a \in A} A = A \quad .$$

Each D_a is a tight set, the so the union is also tight.
Thus, $f(A) = x^*(A)$.

Solving miminal norm problem

Testing for the optimal point

- How to test whether $x \in B(f)$ has the smallest norm.
- $B(f)$ is a convex shape and the norm is a convex function.
- The local minimum is a global minimum.
- x is a local minimum if there is no vector $y \in B(f)$ such that

$$\frac{\partial}{\partial \lambda} \left\| x + \lambda(y - x) \right\|^2 < 0 \quad \text{at} \quad \lambda = 0 \quad .$$

- ...otherwise move slightly towards y from x to get a better point.

Finding direction

Note that at $\lambda = 0$, we have

$$\frac{\partial}{\partial \lambda} \|x + \lambda(y - x)\|^2 = 2(x + \lambda(y - x))^T(y - x) = 2x^T(y - x) = 2x^T y - 2x^T x \quad .$$

- Find $\min_y x^T y$ such that $y \in B(f)$.
- If $x^T y \geq x^T x$, then x is optimal.
- Otherwise, there is a better point between x and y.
- We will solve finding y later but for now assume it's possible.

Wolfe's algorithm

- $B(f)$ is a polytope, so there is a finite number of corner points such that $B(f)$ lies between these points.
- Write $m(S)$ to be the vector with the smallest norm in the affine space spanned by S
- Key insight: there is a set of corner points S such that $m(S)$ is the optimal solution.
- Enumerate sets of corner points such that
  1. $m(S)$ is in convex hull of S (and also in $B(f)$)
  2. $\|m(S)\|$ is decreasing
- There are finite number of sets of corner points so we will converge

Wolfe's algorithm

Algorithm maintains:

- A candidate set S of corner points
- A point x in convex hull of S that in the end will be optimal

Wolfe's algorithm

Algorithm maintains:

- A candidate set S of corner points
- A point x in convex hull of S that in the end will be optimal

Step (a): test for optimality of x

- Compute $y = \arg\min x^T y$ such that $y \in B(f)$
- either x will be optimal
- ...or we have a new corner point y

Wolfe's algorithm

Algorithm maintains:

- A candidate set S of corner points
- A point x in convex hull of S that in the end will be optimal

Step (a): test for optimality of x

- Compute $y = \arg\min x^T y$ such that $y \in B(f)$
- either x will be optimal
- ...or we have a new corner point y

Step (b): find new x

- Compute $m(S)$
- If inside the simplex, set $x = m(S)$ repeat Step (a)
- ...otherwise, find face S' intersecting with the segment $x - m(S)$
- Set S to S'. Set x to the intersection point. Repeat Step (b).

Solving linear program

Solving linear program

We need compute $y = \arg\min x^T y$ such that $y \in B(f)$

- This is an example of linear program...
- ...and there are many solvers for linear programs
- ...but we cannot use any of them
- ...because we have exponential number of constraints.
- Luckily, there is a closed solution due to submodularity.

Solution

Order x such that

$$x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n}$$

Define

$$y_j = f(i_1, \ldots, i_j) - f(i_1, \ldots, i_{j-1}) \quad .$$

Corner case: $y_1 = f(i_1) - f(\emptyset) = f(i_1)$ (we assume wlog that $f(\emptyset) = 0$)

We claim that y is optimal and in $B(f)$.
The optimality follows from Langrange duality, doesn't rely on submodularity

The point y is valid

Lemma

$y(U) = f(U)$

Proof.

$$y(U) = \sum_{j=1}^{n} y_j = \sum_{j=1}^{n} f(i_1, \ldots, i_j) - f(i_1, \ldots, i_{j-1}) = f(i_1, \ldots, i_n) - f(\emptyset) = f(U) \quad .$$

□

The point y is valid

Lemma

$y(B) \leq f(B)$ for any $B \subseteq U$

Proof.

Write $B_j = B \cap \{i_1, \ldots, i_j\}$.

The point y is valid

Lemma

$y(B) \leq f(B)$ *for any* $B \subseteq U$

Proof.

Write $B_j = B \cap \{i_1, \ldots, i_j\}$. Then

$$
\begin{aligned}
y(B) &= \sum_{j \in B} f(i_1, \ldots, i_j) - f(i_1, \ldots, i_{j-1}) \\
&\leq \sum_{j \in B} f(B_{j-1} \cup i_j) - f(B_{j-1}) \\
&= \sum_{j \in B} f(B_j) - f(B_{j-1}) \\
&= f(B) \quad .
\end{aligned}
$$

□

Summary

Minimum norm approach

- Find $x \in B(f)$ with the smallest norm
- Negative components of x minimize f
- Wolfe's algorithm: iterative algorithm, looks like gradient descent
- ...but stops in finite number of steps
- Requires solving linear program
- ...which can be done easily since f is submodular

Encore: optimality of y

Langrange duality

Minimize $p(z)$ such that $q_i(z) \leq 0$ and $r_j(z) = 0$.

Langrange duality

Minimize $p(z)$ such that $q_i(z) \leq 0$ and $r_j(z) = 0$. Define Langrangian

$$\Lambda(z, \lambda, \mu) = p(z) + \sum_i \lambda_i q_i(z) + \sum_j \mu_j r_j(z) \quad .$$

Langrange duality

Minimize $p(z)$ such that $q_i(z) \leq 0$ and $r_j(z) = 0$. Define Langrangian

$$\Lambda(z, \lambda, \mu) = p(z) + \sum_i \lambda_i q_i(z) + \sum_j \mu_j r_j(z) \quad .$$

Define dual

$$d(\lambda, \mu) = \min_z \Lambda(z, \lambda, \mu) \quad .$$

Langrange duality

Minimize $p(z)$ such that $q_i(z) \leq 0$ and $r_j(z) = 0$. Define Langrangian

$$\Lambda(z, \lambda, \mu) = p(z) + \sum_i \lambda_i q_i(z) + \sum_j \mu_j r_j(z) \quad .$$

Define dual

$$d(\lambda, \mu) = \min_z \Lambda(z, \lambda, \mu) \quad .$$

Let y such that $q_i(z) \leq 0$ and $r_j(y) = 0$ and let $\lambda \geq 0$ and μ. Then

$$p(y) \geq d(\lambda, \mu)$$

If we can find y, λ and μ such that $p(y) = d(\lambda, \mu)$, then y is optimal.

Langrangian for our case

$$\Lambda(z, \lambda, \mu) = x^T z + \mu\left[z(U) - f(U)\right] + \sum_{B \subset U} \lambda_B \left[z(B) - f(B)\right]$$

$$= -\mu f(U) - \sum_{B \subset U} f(B) + \sum_{i=1}^{n} z_i (x_i + \mu + \sum_{i \in B \subset U} \lambda_B)$$

Langrangian for our case

$$\Lambda(z, \lambda, \mu) = x^T z + \mu \left[z(U) - f(U) \right] + \sum_{B \subset U} \lambda_B \left[z(B) - f(B) \right]$$

$$= -\mu f(U) - \sum_{B \subset U} f(B) + \sum_{i=1}^{n} z_i (x_i + \mu + \sum_{i \in B \subset U} \lambda_B)$$

The dual is

$$d(\lambda, \mu) = \min_z -\mu f(U) - \sum_{B \subset U} \lambda_B f(B) + \sum_{i=1}^{n} z_i (x_i + \mu + \sum_{i \in B \subset U} \lambda_B) \quad .$$

Langrangian for our case

$$\Lambda(z, \lambda, \mu) = x^T z + \mu [z(U) - f(U)] + \sum_{B \subset U} \lambda_B [z(B) - f(B)]$$

$$= -\mu f(U) - \sum_{B \subset U} f(B) + \sum_{i=1}^{n} z_i (x_i + \mu + \sum_{i \in B \subset U} \lambda_B)$$

The dual is

$$d(\lambda, \mu) = \min_z -\mu f(U) - \sum_{B \subset U} \lambda_B f(B) + \sum_{i=1}^{n} z_i (x_i + \mu + \sum_{i \in B \subset U} \lambda_B) \quad .$$

If $x_i + \mu + \sum_{i \in B \subset U} \lambda_B = 0$ for every i, then

$$d(\lambda, \mu) = -\mu f(U) - \sum_{B \subset U} \lambda_B f(B)$$

Otherwise, $d(\lambda, \mu) = -\infty$ (that is λ, μ are not optimal).

Langrangian for our case

We need to find $\lambda \geq 0$ and μ such that for every i

$$x_i + \mu + \sum_{i \in T \subset U} \lambda_T = 0 \quad .$$

and

$$x^T y = -\mu f(U) - \sum_{T \subset U} \lambda_T f(T) \quad .$$

Langrangian for our case

We need to find $\lambda \geq 0$ and μ such that for every i

$$x_i + \mu + \sum_{i \in T \subset U} \lambda_T = 0 \quad .$$

and

$$x^T y = -\mu f(U) - \sum_{T \subset U} \lambda_T f(T) \quad .$$

Set

$$\mu = -x_n, \quad \lambda_{i_1, \ldots, i_j} = x_{i_{j+1}} - x_{i_j}$$

and $\lambda_T = 0$ for the remaining sets.

Langrangian for our case

First, since x_{i_j} are ordered, $\lambda_{i_1, \ldots, i_j} = x_{i_{j+1}} - x_{i_j} \geq 0$.

Langrangian for our case

First, since x_{i_j} are ordered, $\lambda_{i_1, \ldots, i_j} = x_{i_{j+1}} - x_{i_j} \geq 0$. Next,

$$\sum_{i_k \in T \subset U} \lambda_T = \sum_{j=k}^{n-1} x_{i_{j+1}} - x_{i_j} = x_n - x_{i_k} = -\mu - x_{i_k}.$$

Langrangian for our case

First, since x_{i_j} are ordered, $\lambda_{i_1,\ldots,i_j} = x_{i_{j+1}} - x_{i_j} \geq 0$. Next,

$$\sum_{i_k \in T \subset U} \lambda_T = \sum_{j=k}^{n-1} x_{i_{j+1}} - x_{i_j} = x_n - x_{i_k} = -\mu - x_{i_k}.$$

Finally,

$$\begin{aligned}
\mu f(U) + \sum_{T \subset U} \lambda_T f(T) &= -x_n f(U) + \sum_{j=1}^{n-1} (x_{i_{j+1}} - x_{i_j}) f(i_1, \ldots, i_j) \\
&= \sum_{j=1}^{n-1} x_{i_{j+1}} f(i_1, \ldots, i_j) - \sum_{j=1}^{n} x_{i_j} f(i_1, \ldots, i_j) \\
&= \sum_{j=1}^{n} x_j (f(i_1, \ldots, i_{j-1}) - f(i_1, \ldots, i_j)) = -\sum_{j=1}^{n} x_j y_j \quad .
\end{aligned}$$