6.883 Learning with Combinatorial Structure
Note for Lecture 9
Author(s): Yeesian Ng

1 Minimizing a submodular function

1.1 Lovász extension as a continuous relaxation

Previously, we spoke about constructing the Lovász extension, which is a continuous extension of the original submodular function. To minimize the submodular function

$$\min_{S \subseteq \mathcal{V}} F(S) = \min_{x \in \{0,1\}^n} F(x) \tag{1}$$

we might minimize the Lovász extension (section 2.1 of Lecture 7)

$$\min_{x \in [0,1]^n} f(x) \tag{2}$$

which is a convex optimization problem, that can be solved using a subgradient method. It can be shown[1] that the relaxation is exact, so we can recover an optimal set S^* to (1) from an optimal solution x^* to (2).

1.2 Difficulties with solving the Dual

Alternatively, we can consider the dual of (1)

$$f(x) = \max_{y \in \mathcal{B}_F} \sum_{i=1}^{n} \min\{y_i, 0\} \tag{3}$$

as an optimization over the base polytope $\mathcal{B}_F$.

To test membership in the polytope is to test whether $y(S) \leq F(S)$ for all S. One possibility is to check whether $F(S) - y(S) \geq 0$ for all S. But this is equivalent to

$$\min_{S \subseteq \mathcal{V}} [F(S) - y(S)] \geq 0 \tag{4}$$

which requires the minimization of another submodular function $g(S) = F(S) - y(S)$.

By the ed of the lecture, we will discover that projections onto the base polytope are equivalent to solving a parametric submodular minimization problem.

[1] See Problem 2 of Homework 2

2 Parametric Submodular Minimization

Consider the following formulation

$$\min_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{n} \psi_i(x_i) \tag{5}$$

where $\psi_i(\cdot)$ is strictly convex and continuously differentiable, and $\lim_{x\to\infty} \psi_i'(x) = \infty$ and $\lim_{x\to-\infty} \psi_i'(x) = -\infty$.

To see its connection to submodular minimization[2], we study the problem of solving

$$S_\alpha^* \in \arg\min_{S \subseteq \mathcal{V}} F(S) + \alpha|S| \tag{6}$$

where we introduce a penalty $\alpha|S|$ on the cardinality of the function.

Observe that $\lim_{\alpha\to\infty} S_\alpha^* = \emptyset$ and $\lim_{\alpha\to-\infty} S_\alpha^* = \mathcal{V}$. More generally, we consider some weight function $w_\alpha : \mathcal{V} \mapsto \mathbb{R}$, which must strictly increase with respect to α.[3,4]

Proposition 1 (Monotonicity)**.** *The set of solutions is going to be monotone, i.e.*

$$\alpha < \beta \implies S^\beta \subseteq S^\alpha$$

where $S^\alpha \in \arg\min_{S \subseteq \mathcal{V}} F(S) + w_\alpha(S)$.

Proof. By the optimality of S^α, we have

$$\begin{aligned} F(S^\alpha) + w_\alpha(S^\alpha) &\le F(S^\alpha \cup S^\beta) + w_\alpha(S^\alpha \cup S^\beta) \quad \forall S^\beta \\ F(S^\beta) + w_\beta(S^\beta) &\le F(S^\alpha \cap S^\beta) + w_\beta(S^\alpha \cup S^\beta) \quad \forall S^\alpha \end{aligned}$$

By summing them up, and rearranging terms, we get

$$\begin{aligned} 0 &\ge F(S^\alpha \cup S^\beta) + F(S^\alpha \cap S^\beta) - (F(S^\alpha) + F(S^\beta)) \\ &\ge w_\alpha(S^\alpha) + w_\beta(S^\beta) - w_\alpha(S^\alpha \cup S^\beta) - w_\beta(S^\alpha \cap S^\beta) \\ &= -w_\alpha(S^\beta \setminus S^\alpha) + w_\beta(S^\beta \setminus S^\alpha) \\ &= |S^\beta \setminus S^\alpha|(w_\beta - w_\alpha) \end{aligned}$$

Since $w_\beta - w_\alpha > 0$, it must be that $S^\beta \subseteq S^\alpha$. □

[2]Optimizing (6) does not always lead to the same optimal solution as that for (5)
[3]Although the w's can be made coordinate-dependent, we'll skip that development for this lecture
[4]If we let $w_\alpha(S) = \alpha|S|$, we recover (6)

Proposition 2. *Define $u \in \mathbb{R}^n$ such that $u_j = \sup\{\alpha \in \mathbb{R} \mid j \in S^\alpha\}$. Then u is the unique optimal solution of (5), i.e.*

$$u = \arg\min_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{n} \psi_i(x_i)$$

Proof. Let z be an arbitrary solution, and $\beta = \min\{z_i, u_i\}_{1 \le i \le n}$. The idea is to write the Lovász extension as an integral.

$$\begin{aligned}
\psi(u_j) &= \psi(\beta) + \int_\beta^{u_j} \psi'(\alpha)\,\mathrm{d}\alpha = \psi(\beta) + \int_\beta^{\infty} \psi'(\alpha)\mathbb{I}[u_j \ge \alpha]\,\mathrm{d}\alpha \\
f(u) &= \int_\beta^0 [F(\{u \ge \alpha\}) - F(\mathcal{V})]\,\mathrm{d}\alpha + \int_0^\infty F(\{u \ge \alpha\})\,\mathrm{d}\alpha
\end{aligned}$$

Letting $w_\alpha(S) = \sum_{j \in S} \psi'_j(\alpha)$, we get

$$\begin{aligned}
f(u) + \sum_{j=1}^n \psi(u_j) &= \int_\beta^\infty \left[F(\{u \ge \alpha\}) + \sum_{j=1}^n \psi'(\alpha)\mathbb{I}[u_j \ge \alpha] \right] \mathrm{d}\alpha + n\psi(\beta) - \int_\beta^0 F(\mathcal{V})\mathrm{d}\alpha \\
&\le \int_\beta^\infty \left[F(\{z \ge \alpha\}) + \sum_{j=1}^n \psi'(\alpha)\mathbb{I}[z_j \ge \alpha] \right] \mathrm{d}\alpha + n\psi(\beta) - \int_\beta^0 F(\mathcal{V})\mathrm{d}\alpha \\
&= \int_\beta^\infty \left[F(\{z \ge \alpha\}) + \sum_{j=1}^n \psi'(\alpha)\mathbb{I}[z_j \ge \alpha] \right] \mathrm{d}\alpha + n\psi(\beta) - \int_\beta^0 F(\mathcal{V})\mathrm{d}\alpha \\
&= f(z) + \sum_{j=1}^n \psi(z_j)
\end{aligned}$$

for any arbitrary $z_j \in \{\alpha \in \mathbb{R} \mid j \in S^\alpha\}$, where the first inequality follows from the monotonicity of S^α, and the last inequality comes from rearranging terms. Hence u is optimal to (5). $\square$

Remark: Observe that for any α, u encodes the parametric solution S^α through

$$\{u > \alpha\} \subseteq S^\alpha \subseteq \{u \ge \alpha\}$$

where $\{u > \alpha\}$ is a shorthand for the minimal minimizer $\{j \mid u_j > \alpha\}$, and $\{u \ge \alpha\}$ for the maximal minimizer $\{j \mid u_j \ge \alpha\}$.

3 Minimum norm Problem

Now consider the case $\psi_j(x) = x^2$ for all j. Then $\psi_j(\alpha) = \alpha$, and we have

$$f(x) + \frac{1}{2}||x||^2 \equiv F(S) + \alpha|S| \tag{7}$$

by propositions (1) and (2). Thus, by solving the LHS, we can read out all the solutions of the RHS, and the entries of the vector x will have the form

$$x_j^* = \frac{F(S_i) - F(S_{i-1})}{|S_i \setminus S_{i-1}|}$$

for all $j \in S_i \setminus S_{i-1}$ ($S_k \subset S_{k-1} \subset \dots$).

Taking the dual, we have

$$\begin{aligned}
\min_x f(x) + \frac{1}{2}||x||^2 &= \min_x \max_{y \in \mathcal{B}_F} y^\mathsf{T} x + \frac{1}{2}||x||^2 && (8)\\
&= \max_{y \in \mathcal{B}_F} \min_x y^\mathsf{T} x + \frac{1}{2}||x||^2 && (9)\\
&= \max_{y \in \mathcal{B}_F} -\frac{1}{2}||y||^2 && (10)\\
&= -\min_{y \in \mathcal{B}_F} \frac{1}{2}||y||^2 && (11)
\end{aligned}$$

where the second equality follows from strong duality, by observing that the expression is concave (linear) in y and convex in x, which satisfies the saddle-point property.

Therefore, it can be seen that solving the projection problem (i.e. by projecting 0 onto the polytope) is as difficult as solving the parametric submodular problem (6), which by setting $\alpha = 0$, which is at least as hard as solving the original problem (1).

Remark: The optimal solution to the dual problem $\max_{y \in \mathcal{B}_F} -\frac{1}{2}||y||^2$ (which by negation is same as that of the primal problem) has the following properties

- nested sublevel sets $S_1 \subset S_2 \subset \dots \subset S_k$
- "tight sets": for all sublevel sets, we have $y^*(S_j) = F(S_j)$.
- for $i \in S_j \setminus S_{j-1}$, we have $y_i^* = \frac{F(S_j) - F(S_{j-1})}{|S_j \setminus S_{j-1}|}$
- y^* is the lexicographically maximal base (Fujishige 1980)

4 Convergence Bounds

Suppose we are interested in solving (1), and have an iterative method for the minimization of (7), which generated a primal-dual candidate (x, y). From duality, we have

$$\tilde{f}(x) - \tilde{f}(x^*) \leq \tilde{f}(x) - \tilde{g}(y)$$

where the RHS is called the *duality gap*, with $\tilde{f}(x) = f(x) + \frac{1}{2}||x||^2$ and $\tilde{g}(y) = -\frac{1}{2}||y||^2$.

Then we have the following result (stated here without proof):

Theorem 1 (Bach 2013). *If $\tilde{f}(x) - \tilde{g}(y) \leq \epsilon'$, then there exists an α such that the discrete duality gap is bounded as*

$$F(S^\alpha) - F(S^*) \leq F(S^\alpha) - y_-(\mathcal{V}) \leq \sqrt{(2n\epsilon')}$$

where $S^\alpha = \{e \mid -y_e \geq \alpha\}$.

5 Min-norm point Algorithm

Here we describe an active set method for solving the min-norm problem. Briefly: it starts with a polytope with many vertices[5,6]. It maintains an *active set* of vertices, and iteratively alternates between optimizing over the convex hull of those vertices, and updating the active set of vertices to optimize over.

1. Pick any corner point v_i, and set $y = v_i$, and $S = \{v_i\}$.
2. Find $v' \in \arg\min_{v \in \mathcal{B}_F} y^\mathsf{T} v$, and test for optimality (i.e. $\langle v', y \rangle \stackrel{?}{=} ||y||)$[7]. If y is not optimal, set $S = S \cup \{v'\}$ and continue.
3. Find a min-norm point z in the *affine hull* of S, i.e.

$$\begin{aligned} &\min_\eta \frac{1}{2}||\sum_{v_i \in S} \eta_i v_i||^2 \\ &\text{s.t.} \sum_i \eta_i = 1 \end{aligned}$$

4. Let η be the solution obtained. If $z := \sum_{v_i \in S} \eta_i v_i \in \text{conv}(S)$, set $y = z$ and continue with (2). Otherwise "correct" by

[5]potentially exponential in the number of dimensions of the polytope

[6]for n dimensions, you need at most n vertices to represent any point in the polytope

[7]The optimality conditions follow from looking at the saddle point conditions for the optimization problem (9): we need that $\nabla_x \phi(x, y) = 0$ and $\nabla_y \phi(x, y) = 0$, so $\langle x^*, y^* \rangle = \langle y^*, y^* \rangle$ here.

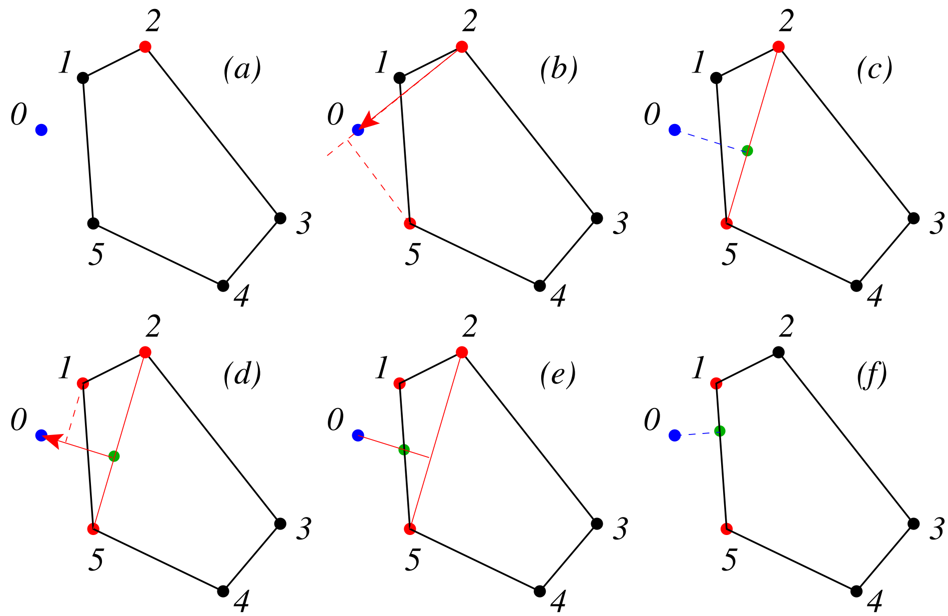

Figure 9.1: Illustration of Frank-Wolfe minimum-norm-point algorithm: (a) initialization with $J = \{2\}$ (step (1)), (b) check optimality (step (5)) and take $J = \{2, 5\}$, (c) compute affine projection (step (2)), (d) check optimality and take $J = \{1, 2, 5\}$, (e) perform line search (step (3)) and take $J = \{1, 5\}$, (f) compute affine projection (step (2)) and obtain optimal solution.

Figure copied from Bach, *Learning with Submodular Functions – A Convex Optimization Perspective*

a) Set y to $y_{\text{new}} = \alpha y + (1-\alpha)z = \alpha(\sum_i \eta_i v_i) + (1-\alpha)(\sum_{i \in S} \gamma_i v_i)$ for the smallest α satisfying $\alpha\eta_i + (1-\alpha)\gamma_i \geq 0$ for all i.

b) Drop points with $\eta_{\text{new},i} = 0$ from S.

and continue with (3).