

Convex Analysis and Optimization with Submodular Functions: a Tutorial

Francis Bach
INRIA - Willow project-team
Laboratoire d'Informatique de l'École Normale Supérieure
Paris, France
francis.bach@ens.fr

November 14, 2010

Introduction

Set-functions appear in many areas of computer science and applied mathematics, such as machine learning [1, 2, 3, 4], computer vision [5, 6], operations research [7] or electrical networks [8]. Among these set-functions, submodular functions play an important role, similar to convex functions on vector spaces. In this tutorial, the theory of submodular functions is presented, in a self-contained way, with all results shown from first principles. A good knowledge of convex analysis is assumed (see, e.g., [9, 10]).

Several books and tutorial articles already exist on the same topic and the material presented in this tutorial rely mostly on those [11, 8, 12, 13]. However, in order to present the material in the simplest way, ideas from related research papers have also been used.

Notation. We consider the set $V = \{1, \dots, p\}$, and its power set 2^V , composed of the 2^p subsets of V . Given a vector $s \in \mathbb{R}^p$, s also denotes the modular set-function defined as $s(A) = \sum_{k \in A} s_k$. Moreover, $A \subset B$ means that A is a subset of B , potentially equal to B . For $q \in [1, +\infty]$, we denote by $\|w\|_q$ the ℓ_q -norm of w , by $|A|$ the cardinality of the set A , and, for $A \subset V = \{1, \dots, p\}$, 1_A denotes the indicator vector of the set A . If $w \in \mathbb{R}^p$, and $\alpha \in \mathbb{R}$, then $\{w \geq \alpha\}$ (resp. $\{w > \alpha\}$) denotes the subset of $V = \{1, \dots, p\}$ defined as $\{k \in V, w_k \geq \alpha\}$ (resp. $\{k \in V, w_k > \alpha\}$). Similarly if $v \in \mathbb{R}^p$, we have $\{w \geq v\} = \{k \in V, w_k \geq v_k\}$.

Tutorial outline. In Section 1, we give the different definitions of submodular functions and of the associated polyhedra. In Section 2, we define the Lovász extension and give its main properties. Associated polyhedra are further studied in Section 3, where support functions and the associated maximizers are computed (we also detail the facial structure of such polyhedra). In Section 4, we provide some duality theory for submodular functions, while in Section 5, we present several operations that preserve submodularity. In Section 6, we consider separable optimization problems associated with the Lovász extension; these are reinterpreted in Section 7 as separable optimization over the submodular or base polyhedra.

In Section 8, we present various approaches to submodular function minimization (without all details of algorithms). In Section 9, we specialize some of our results to non-decreasing submodular functions. Finally, in Section 10, we present classical examples of submodular functions.

Contents

1	Definitions	3
2	Lovász extension	4
3	Support function of submodular and base polyhedra	8
4	Minimizers of submodular functions	11
5	Operations that preserve submodularity	12
6	Proximal optimization problems	15
7	Optimization over the base polyhedron	18
7.1	Optimality conditions	18
7.2	Lexicographically optimal bases	20
7.3	Optimization for proximal problems	21
8	Submodular function minimization	21
8.1	Minimum-norm point algorithm	22
8.2	Combinatorial algorithms	22
8.3	Minimizing posimodular functions	23
8.4	Line search in submodular polyhedron	23
8.5	Homotopy method for proximal problems	23
8.6	Decomposition algorithm for proximal problems	24
9	Polymatroids (non-increasing submodular functions)	25
10	Examples of submodular functions	27
10.1	Cardinality-based functions	28
10.2	Cut functions	28
10.3	Set covers	30

10.4 Flows	31
10.5 Entropies	34
10.6 Spectral functions of submatrices	34
10.7 Best subset selection	34
10.8 Matroids	35

1 Definitions

Throughout this tutorial, we consider $V = \{1, \dots, p\}$, $p > 0$ and its power set (i.e., set of all subsets) 2^V , which is of cardinality 2^p . We also consider a real-valued set-function $F : 2^V \rightarrow \mathbb{R}$ such that $F(\emptyset) = 0$. As opposed to the common convention with convex functions, we do not allow infinite values for the function F .

Definition 1 (Submodular function) *A set-function $F : 2^V \rightarrow \mathbb{R}$ is submodular if and only if, for all subsets $A, B \subset V$, we have: $F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$.*

The simplest example of submodular function is the cardinality (i.e., $F(A) = |A|$ where $|A|$ is the number of elements of A), which is both submodular and supermodular (i.e., its opposite is submodular), which we refer to as *modular*.

From Def. 1, it is clear that the set of submodular functions is closed under addition and multiplication by a positive scalar. The following proposition shows that a submodular has the “diminishing return” property, and that this is sufficient to be submodular. Thus, submodular functions may be seen as a discrete analog to *concave* functions. However, in terms of optimization they behave more like *convex* functions (e.g., efficient minimization, duality theory, linked with convex Lovász extension).

Proposition 1 (Equivalent definition with first order differences) *F is submodular if and only if for all $A, B \subset V$ and $k \in V$, such that $A \subset B$ and $k \notin B$, we have $F(A \cup \{k\}) - F(A) \geq F(B \cup \{k\}) - F(B)$.*

Proof Let $A \subset B$, and $k \notin B$, $F(A \cup \{k\}) - F(A) - F(B \cup \{k\}) + F(B) = F(C) + F(D) - F(C \cup D) - F(C \cap D)$ with $C = A \cup \{k\}$ and $D = B$, which shows that the condition is necessary. To prove the opposite, we assume that the condition is satisfied; one can first show that if $A \subset B$ and $C \cap B = \emptyset$, then $F(A \cup C) - F(A) \geq F(B \cup C) - F(B)$ (this can be obtained by summing the m inequalities $F(A \cup \{c_1, \dots, c_k\}) - F(A \cup \{c_1, \dots, c_{k-1}\}) \geq F(B \cup \{c_1, \dots, c_k\}) - F(B \cup \{c_1, \dots, c_{k-1}\})$ where $C = \{c_1, \dots, c_m\}$).

Then for any $X, Y \subset V$, take $A = X \cap Y$, $C = X \setminus Y$ and $B = Y$ to obtain $F(X) + F(Y) \geq F(X \cup Y) + F(X \cap Y)$, which shows that the condition is sufficient. ■

The following proposition gives the tightest condition for submodularity (easiest to show in practice).

Proposition 2 (Equivalent definition with second order differences) *F is submodular if and only if for all $A \subset V$ and $j, k \in V \setminus A$, we have $F(A \cup \{k\}) - F(A) \geq F(A \cup \{j, k\}) - F(A \cup \{j\})$.*

Proof This condition is weaker than the one from previous proposition. To prove that it is still sufficient, simply apply it to subsets $A \cup \{b_1, \dots, b_{s-1}\}$, $j = b_s$ for $B = A \cup \{b_1, \dots, b_m\} \supset A$ with $k \notin B$, and sum the m inequalities $F(A \cup \{b_1, \dots, b_{s-1}\} \cup \{k\}) - F(A \cup \{b_1, \dots, b_{s-1}\}) \geq F(A \cup \{b_1, \dots, b_s\} \cup \{k\}) - F(A \cup \{b_1, \dots, b_s\})$, to obtain the condition in Prop. 1. ■

A vector $s \in \mathbb{R}^p$ naturally leads to a modular set-function defined as $s(A) = \sum_{k \in A} s_k = s^\top 1_A$, where $1_A \in \mathbb{R}^p$ is the indicator vector of the set A . We now define specific polyhedra in \mathbb{R}^p . These play a crucial role in submodular analysis, as most results may be interpreted or proved using such polyhedra.

Definition 2 (Submodular and base polyhedra) *Let F be a submodular function such that $F(\emptyset) = 0$. The submodular polyhedron $P(F)$ and the base polyhedron $B(F)$ are defined as:*

$$\begin{aligned} P(F) &= \{s \in \mathbb{R}^p, \forall A \subset V, s(A) \leq F(A)\} \\ B(F) &= \{s \in \mathbb{R}^p, s(V) = F(V), \forall A \subset V, s(A) \leq F(A)\} = P(F) \cap \{s(V) = F(V)\}. \end{aligned}$$

As shown in the following proposition, the submodular polyhedron $P(F)$ has non empty-interior and is unbounded. Note that the other polyhedron (the base polyhedron) will be shown to be non-empty and bounded as a consequence of Prop. 5. It has empty interior since it is included in the subspace $s(V) = F(V)$. See Figure 1 for examples with $p = 2$ and $p = 3$.

Proposition 3 (Properties of submodular polyhedron) *Let F be a submodular function such that $F(\emptyset) = 0$. If $s \in P(F)$, then for all $t \in \mathbb{R}^p$, such that $t \leq s$, we have $t \in P(F)$. Moreover, $P(F)$ has non-empty interior.*

Proof The first part is trivial, since $t(A) \leq s(A)$ if $t \leq s$. For the second part, we only need to show that $P(F)$ is non-empty, which is true since the constant vector equal to $\min_{A \subset V, A \neq \emptyset} \frac{F(A)}{|A|}$ belongs to $P(F)$. ■

2 Lovász extension

We consider a set-function F such that $F(\emptyset) = 0$, which is not necessary submodular. We can define its Lovász extension [14], which is often referred to as its Choquet integral [15]. The Lovász extension allows to draw links between submodular set-functions and regular convex functions, and transfer known results from convex analysis, such as duality.

Definition 3 (Lovász extension) *Given a set-function F such that $F(\emptyset) = 0$, the Lovász extension $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is defined as follows; for $w \in \mathbb{R}^p$, order the components $w_{j_1} \geq \dots \geq$*

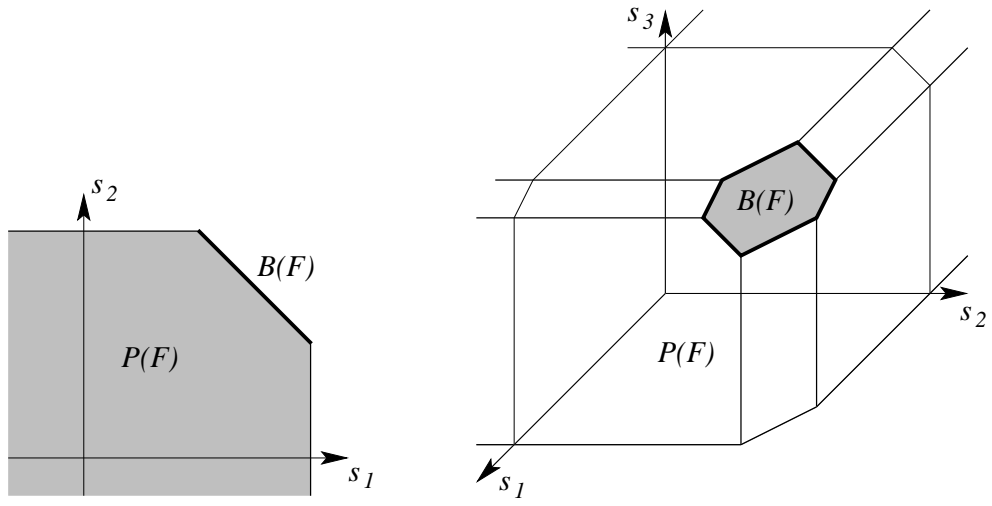


Figure 1: Submodular polyhedron $P(F)$ and base polyhedron $B(F)$ for $p = 2$ (left) and $p = 3$ (right), for a non-decreasing submodular function.

w_{j_p} , and define $f(w)$ through any of the following equations:

$$f(w) = w_{j_1}F_{j_1} + \sum_{k=2}^p w_{j_k} [F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})], \quad (1)$$

$$= \sum_{k=1}^{p-1} F(\{j_1, \dots, j_k\})(w_{j_k} - w_{j_{k+1}}) + F(V)w_{j_p}, \quad (2)$$

$$= \int_{\min\{w_1, \dots, w_p\}}^{+\infty} F(\{w \geq z\})dz + F(V) \min\{w_1, \dots, w_p\}, \quad (3)$$

$$= \int_0^{+\infty} F(\{w \geq z\})dz + \int_{-\infty}^0 [F(\{w \geq z\}) - F(V)]dz. \quad (4)$$

Proof To prove that we actually define a function, one needs to prove that the definition is independent of the non unique ordering $w_{j_1} \geq \dots \geq w_{j_p}$, which is trivial from the last formulation in Eq. (4). The first and second formulations in Eq. (1) and Eq. (2) are equivalent (by integration by parts, or Abel summation formula). To show equivalence with Eq. (3), one may notice that $z \mapsto F(\{w \geq z\})$ is piecewise constant, with value zero for $z > w_{j_1} = \max\{w_1, \dots, w_p\}$, and equal to $F(\{j_1, \dots, j_k\})$ for $z \in (w_{j_{k+1}}, w_{j_k})$, $k = \{1, \dots, p-1\}$, and equal to $F(V)$ for $z < w_{j_p} = \min\{w_1, \dots, w_p\}$. What happens at break points is irrelevant for integration.

To prove Eq. (4), notice that for $\alpha \leq \min\{0, w_1, \dots, w_p\}$, Eq. (3)

$$\begin{aligned} f(w) &= \int_{\alpha}^{+\infty} F(\{w \geq z\})dz - \int_{\alpha}^{\min\{w_1, \dots, w_p\}} F(\{w \geq z\})dz + F(V) \min\{w_1, \dots, w_p\} \\ &= \int_{\alpha}^{+\infty} F(\{w \geq z\})dz - \int_{\alpha}^{\min\{w_1, \dots, w_p\}} F(V)dz + \int_0^{\min\{w_1, \dots, w_p\}} F(V)dz \\ &= \int_{\alpha}^{+\infty} F(\{w \geq z\})dz - \int_{\alpha}^0 F(V)dz, \end{aligned}$$

and we get the result by letting α tend to $-\infty$. ■

Note that for modular functions $A \mapsto s(A)$, with $s \in \mathbb{R}^p$, then the Lovász extension is the linear function $w \mapsto w^\top s$. The following proposition details classical properties of the Choquet integral. Property (e) below implies that the Lovász extension is equal to the original set-function on $\{0, 1\}^p$ (which can canonically be identified to 2^V), and hence is indeed an *extension* of F .

Proposition 4 (Properties of Lovász extension) *Let F be any set-function such that $F(\emptyset) = 0$. We have:*

(a) *if F and G are set-functions with Lovász extensions f and g , then $f + g$ is the Lovász extension of $F + G$, and for all $\lambda \in \mathbb{R}_+$, λf is the Lovász extension of λF ,*

(b) *for $w \in \mathbb{R}_+^p$, $f(w) = \int_0^{+\infty} F(\{w \geq z\}) dz$,*

(c) *if $F(V) = 0$, for all $w \in \mathbb{R}^p$, $f(w) = \int_{-\infty}^{+\infty} F(\{w \geq z\}) dz$,*

(d) *for all $w \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$, $f(w + \alpha 1_V) = f(w) + \alpha F(V)$,*

(e) *the Lovász extension f is positively homogeneous,*

(f) *for all $A \subset V$, $F(A) = f(1_A)$,*

(g) *if F is symmetric (i.e., $\forall A \subset V$, $F(A) = F(V \setminus A)$), then f is even,*

(h) *if $V = A_1 \cup \dots \cup A_m$ is a partition of V , and $w = \sum_{i=1}^m v_i 1_{A_i}$ (i.e., is constant on each set A_i), with $v_1 \geq \dots \geq v_m$, then $f(w) = \sum_{i=1}^{m-1} (v_i - v_{i+1}) F(A_1 \cup \dots \cup A_i) + v_{i+1} F(V)$.*

Proof Properties (a), (b) and (c) are immediate from Eq. (4) and Eq. (2). (d), (e) and (f) are straightforward from Eq. (2). If F is symmetric, then $F(V) = 0$, and thus $f(-w) = \int_{-\infty}^{+\infty} F(\{-w \geq z\}) dz = \int_{-\infty}^{+\infty} F(\{w \leq -z\}) dz = \int_{-\infty}^{+\infty} F(\{w \leq z\}) dz = \int_{-\infty}^{+\infty} F(\{w > z\}) dz = f(w)$ (because we may replace strict inequalities by regular inequalities), i.e., f is even. ■

Note that when the function is a cut function, then the Lovász extension is related to the total variation and property (c) is often referred to as the co-area formula (see [16] and references therein, as well as Section 10.2).

The next result relates the Lovász extension with the support function of the submodular polyhedron $P(F)$ which is defined in Def. 2. This is the basis for many of the theoretical results and algorithms related to submodular functions. It shows that maximizing a linear function with non-negative coefficients on the submodular polyhedron may be obtained in closed form, by the so-called “greedy algorithm” (see [14] for an intuitive explanation), and the optimal value is equal to the value $f(w)$ of the Lovász extension. Note that otherwise, solving a linear programming problem with 2^p constraints would then be required.

Proposition 5 (Greedy algorithm) *Let F be a submodular function such that $F(\emptyset) = 0$. Let $w \in \mathbb{R}_+^p$. A maximizer of $\max_{s \in P(F)} w^\top s$ may be obtained by the following algorithm: order the components of w , as $w_{j_1} \geq \dots \geq w_{j_p} \geq 0$ and define $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$. Moreover, for all $w \in \mathbb{R}_+^p$, $\max_{s \in P(F)} w^\top s = f(w)$.*

Proof By convex duality (which applies because $P(F)$ has non empty interior from Prop. 3), we have, by introducing Lagrange multipliers $\lambda_A \in \mathbb{R}_+$ for the constraints $s(A) \leq F(A)$, $A \subset V$:

$$\begin{aligned} \max_{s \in P(F)} w^\top s &= \min_{\lambda_A \geq 0, A \subset V} \max_{s \in \mathbb{R}^p} \left\{ w^\top s - \sum_{A \subset V} \lambda_A [s(A) - F(A)] \right\} \\ &= \min_{\lambda_A \geq 0, A \subset V} \max_{s \in \mathbb{R}^p} \left\{ \sum_{A \subset V} \lambda_A F(A) + \sum_{k=1}^p s_k (w_k - \sum_{A \ni k} \lambda_A) \right\} \\ &= \min_{\lambda_A \geq 0, A \subset V} \sum_{A \subset V} \lambda_A F(A) \text{ such that } \forall k \in V, w_k = \sum_{A \ni k} \lambda_A. \end{aligned}$$

If we take the (primal) solution s of the greedy algorithm, we have $f(w) = w^\top s$ from Eq. (1), and s is feasible (i.e., in $P(F)$), because of the submodularity of F . Indeed, without loss of generality, we assume that $j_k = k$ for all $k \in \{1, \dots, p\}$. We can decompose $A = A_1 \cup \dots \cup A_m$, where $A_k = (u_k, v_k]$ are *integer intervals*. We then have:

$$\begin{aligned} s(A) &= \sum_{k=1}^m \{F((0, v_k]) - F((0, u_k])\} \\ &\leq \sum_{k=1}^m \{F((u_1, v_k]) - F((u_1, u_k])\} \text{ by submodularity} \\ &= F((u_1, v_1]) + \sum_{k=2}^m \{F((u_1, v_k]) - F((u_1, u_k])\} \\ &\leq F((u_1, v_1]) + \sum_{k=2}^m \{F((u_1, v_1] \cup (u_2, v_k]) - F((u_1, v_1] \cup (u_2, u_k])\} \text{ by submodularity} \\ &= F((u_1, v_1] \cup (u_2, v_2]) + \sum_{k=3}^m \{F((u_1, v_1] \cup (u_2, v_k]) - F((u_1, v_1] \cup (u_2, u_k])\}. \end{aligned}$$

By pursuing applying submodularity, we finally obtain that $S(A) \leq F((u_1, v_1] \cup \dots \cup (u_m, v_m]) = F(A)$, i.e., $s \in P(F)$.

Moreover, we can define dual variables $\lambda_{\{j_1, \dots, j_k\}} = w_{j_k} - w_{j_{k+1}}$ for $k \in \{1, \dots, p-1\}$ and $\lambda_V = w_{j_p}$ with all other λ_A equal to zero. Then they are all non negative (notably because $w \geq 0$), and satisfy the constraint $\forall k \in V, w_k = \sum_{A \ni k} \lambda_A$. Finally, the dual cost function has also value $f(w)$ (from Eq. (2)). Thus by duality (which holds, because $P(F)$ is not empty), s is an optimal solution. Note that it is not unique (see Prop. 27 for a description of the set of solutions). ■

The next proposition draws precise links between convexity and submodularity, by showing that a set-function F is submodular if and only if its Lovász extension f is convex. This is further developed in Prop. 7 where it is shown that minimizing F on 2^V (which is equivalent to minimizing f on $\{0, 1\}^p$ since f is an extension of F) and minimizing f on $[0, 1]^p$ is equivalent (when F is submodular).

Proposition 6 (Convexity and submodularity) *A set-function F is submodular if and only if its Lovász extension f is convex.*

Proof Let $A, B \subset V$. The vector $1_{A \cup B} + 1_{A \cap B} = 1_A + 1_B$ has components equal to 0 (on $V \setminus (A \cup B)$), 2 (on $A \cap B$) and 1 (on $A \Delta B = (A \setminus B) \cup (B \setminus A)$). Therefore, $f(1_{A \cup B} + 1_{A \cap B}) = \int_0^2 F(1_{\{w \geq z\}}) dz = \int_0^1 F(A \cup B) dz + \int_1^2 F(A \cap B) dz = F(A \cup B) + F(A \cap B)$.

If f is convex, then by homogeneity, $f(1_A + 1_B) \leq f(1_A) + f(1_B)$, which is equal to $F(A) + F(B)$, and thus F is submodular.

If F is submodular, then by Proposition 5, for all $w \in \mathbb{R}_+^p$, $f(w)$ is a maximum of linear functions, thus, it is convex on \mathbb{R}_+^p . Moreover, because $f(w + \alpha 1_V) = f(w) + \alpha F(V)$, it is convex on \mathbb{R}^p . ■

The next proposition completes Prop. 6 by showing that minimizing the Lovász extension on $[0, 1]^p$ is equivalent to minimizing it on $\{0, 1\}^p$, and hence to minimizing the set-function F on 2^V (when F is submodular).

Proposition 7 (Minimization of submodular functions) *Let F be a submodular function and f its Lovász extension; then $\min_{A \subset V} F(A) = \min_{w \in [0, 1]^p} f(w)$.*

Proof Because f is an extension from $\{0, 1\}^p$ to $[0, 1]^p$ (property (d) from Proposition 4), then we must have $\min_{A \subset V} F(A) = \min_{w \in \{0, 1\}^p} f(w) \geq \min_{w \in [0, 1]^p} f(w)$. For the other inequality, any $w \in [0, 1]^p$ may be decomposed as $w = \sum_{i=1}^p \lambda_i 1_{A_i}$ where $A_1 \subset \dots \subset A_p = V$, where λ is nonnegative and has a sum smaller than or equal to one (this can be obtained by considering A_i the set of indices of the i largest values of w). We then have $f(w) = \sum_{i=1}^p \int_{\sum_{k=1}^{i-1} \lambda_k}^{\sum_{k=1}^i \lambda_k} F(A_i) dz = \sum_{i=1}^p \lambda_i F(A_i) \geq \sum_{i=1}^p \lambda_i \min_{A \subset V} F(A) \geq \min_{A \subset V} F(A)$ (because $\min_{A \subset V} F(A) \leq 0$). This leads to the desired result. ■

3 Support function of submodular and base polyhedra

The next proposition completes Prop. 5 by computing the full support function of $B(F)$ and $P(F)$ (see [9, 10] for definitions of support functions), i.e., computing $\max_{s \in B(F)} w^\top s$ and $\max_{s \in P(F)} w^\top s$ for all possible w (with positive and/or negative coefficients). Note the different behaviors for $B(F)$ and $P(F)$.

Proposition 8 (Support function of submodular and base polyhedra) *Let F be a submodular function such that $F(\emptyset) = 0$. We have:*

- (a) for all $w \in \mathbb{R}^p$, $\max_{s \in B(F)} w^\top s = f(w)$,
- (b) if $w \in \mathbb{R}_+^p$, $\max_{s \in P(F)} w^\top s = f(w)$,
- (c) if there exists j such that $w_j < 0$, then $\max_{s \in P(F)} w^\top s = +\infty$.

Proof (a) From the proof of Prop. 5, for $w \in \mathbb{R}_+^p$, then the result of the greedy algorithm satisfies $s(V) = F(V)$, and hence (a) is true on \mathbb{R}_+^p . For all w , for α large enough, $w + \alpha 1_V \geq 0$, and thus $f(w) + \alpha F(V) = f(w + \alpha 1_V) = \max_{s \in B(F)} (w + \alpha 1_V)^\top s = \alpha F(V) + \max_{s \in B(F)} w^\top s$, i.e., (a) is true.

Property (b) is shown in Proposition 5. For (c), notice that $s(\lambda) = s_0 - \lambda\delta_j \in P(F)$ for $\lambda \rightarrow +\infty$ and $s_0 \in P(F)$ and that $w^\top s(\lambda) \rightarrow +\infty$. ■

The next proposition shows necessary and sufficient conditions for optimality in the definition of support functions. Note that Prop. 5 gave one example obtained from the greedy algorithm, and that we can now characterize all maximizers. Moreover, note that the maximizer is unique only when w has distinct values, and otherwise, the ordering of the components of w is not unique, and hence, the greedy algorithm may have multiple outputs (and all convex combinations of these are also solutions). The following proposition essentially shows what is exactly needed to be a maximizer.

Proposition 9 (Maximizers of the support function of submodular polyhedron)

Let F be a submodular function such that $F(\emptyset) = 0$. Let $w \in (\mathbb{R}_+^*)^p$, with unique values $v_1 > \dots > v_m > 0$, taken at sets A_1, \dots, A_m (i.e., $V = A_1 \cup \dots \cup A_m$ and $\forall i \in \{1, \dots, m\}, \forall k \in A_i, w_k = v_i$). Then s is optimal for $\max_{s \in P(F)} w^\top s$ if and only if for all $i = 1, \dots, m$, $s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$.

Proof Let $B_i = A_1 \cup \dots \cup A_i$, for $i = 1, \dots, m$. From the optimization problems defined in the proof of Prop. 5, let $\lambda_V = v_m > 0$, and $\lambda_{B_i} = v_i - v_{i+1} > 0$ for $i < m$, with all other $\lambda_A, A \subset V$, equal to zero. Such λ is optimal (because the dual function is equal to $f(w)$).

Let $s \in B(F)$. We have:

$$\begin{aligned} \sum_{A \subset V} \lambda_A F(A) &= v_m F(V) + \sum_{i=1}^{m-1} F(B_i)(v_i - v_{i+1}) \\ &= v_m(F(V) - s(V)) + \sum_{i=1}^{m-1} [F(B_i) - s(B_i)](v_i - v_{i+1}) \\ &\quad + v_m s(V) + \sum_{i=1}^{m-1} s(B_i)(v_i - v_{i+1}) \\ &\geq v_m s(V) + \sum_{i=1}^{m-1} s(B_i)(v_i - v_{i+1}) = s^\top w. \end{aligned}$$

Thus s is optimal, if and only if the primal objective value $s^\top w$ is equal to the optimal dual objective value $\sum_{A \subset V} \lambda_A F(A)$, and thus, if and only if there is equality in all above inequalities, hence the desired result. ■

Note that if $v_m = 0$ in Prop 9 (i.e., we take $w \in \mathbb{R}_+^p$ and there is a w_k equal to zero), then the optimality condition is that for all $i = 1, \dots, m-1$, $s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$ (i.e., we don't need that $s(V) = F(V)$, i.e., the optimal solution is not necessarily in the base polyhedron).

Proposition 10 (Maximizers of the support function of base polyhedron) Let F be a submodular function such that $F(\emptyset) = 0$. Let $w \in \mathbb{R}^p$, with unique values $v_1 > \dots > v_m$, taken at sets A_1, \dots, A_m . Then s is optimal for $\max_{s \in B(F)} w^\top s$ if and only if for all $i = 1, \dots, m$, $s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$.

Proof The proof follows the same arguments than for Prop. 9. ■

Given the last proposition, we may now give necessary and sufficient conditions for characterizing faces of the base polyhedron. We first characterize when the base polyhedron $B(F)$ has full relative interior.

Definition 4 (Inseparable set) *Let F be a submodular function such that $F(\emptyset) = 0$. A set $A \subset V$ is said separable if and only there is a set $B \subset A$, such that $B \neq \emptyset$, $B \neq A$ and $F(A) = F(B) + F(A \setminus B)$. If A is non separable, A is said inseparable.*

Proposition 11 (Full-dimensional base polyhedron) *Let F be a submodular function such that $F(\emptyset) = 0$. The base polyhedron has full relative interior if and only if V is not separable.*

Proof If V is separable into A and $V \setminus A$, then for all $s \in B(F)$, we must have $s(A) = F(A)$ and hence the base polyhedron is included in the intersection of two affine hyperplanes, i.e., $B(F)$ does not have full relative interior in $\{s(V) = F(V)\}$.

We now assume that $B(F)$ is included in $\{s(A) = F(A)\}$, for A as a non-empty strict subset of V . Then $B(F)$ can be factorized in to $B(F_A) \times B(F^A)$ where F_A is the restriction of F to A and F^A the contraction of F on A . Indeed, if $s \in B(F)$, then $s_A \in B(F_A)$ because $s(A) = F(A)$, and $s_{V \setminus A} \in B(F^A)$, because for $B \subset V \setminus A$, $s_{V \setminus A}(B) = s(B) = s(A \cup B) - s(A) \leq F(A \cup B) - F(A)$. Similarly, if $s \in B(F_A) \times B(F^A)$, then for all set $B \subset V$, $s(B) = s(A \cap B) + s((V \setminus A) \cap B) \leq F(A \cap B) + F(A \cup B) - F(A) \leq F(B)$ by submodularity, and $s(A) = F(A)$.

This shows that $f(w) = f_A(w_A) + f^A(w_{V \setminus A})$, which implies that $F(V) = F(A) + F(V \setminus A)$, when applied to $w = 1_V$, i.e., V is separable. ■

We can now detail the facial structure of the base polyhedron, which will be dual to the one of the polyhedron defined by $\{w \in \mathbb{R}^p, f(w) \leq 1\}$ (i.e., level set of the Lovász extension). As the base polyhedron $B(F)$ is a polytope in dimension $p - 1$ (because it is bounded and contained in the affine hyperplane $\{s(V) = F(V)\}$), one can define a set of *faces*. Faces are the intersections of the polyhedron $B(F)$ with any of its supporting hyperplanes. Supporting hyperplanes are themselves defined as the hyperplanes $w^\top s = \max_{s \in B(F)} w^\top s = f(w)$ for $w \in \mathbb{R}^p$. From Prop. 10, faces (which potentially empty relative interior) are obtained as the intersection of $B(F)$ with $s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$ for an ordered partition of V . Together with Prop. 11, we can now provide characterization of the faces of $B(F)$.

Proposition 12 (Faces of the base polyhedron) *Let $A_1 \cup \dots \cup A_m$ be an ordered partition of V , such that for all $j \in \{1, \dots, m\}$, A_j is inseparable for the function $G_j : B \mapsto F(A_1 \cup \dots \cup A_{j-1} \cup B) - F(A_1 \cup \dots \cup A_{j-1})$ defined on subsets of A_j , then the set of bases $s \in B(F)$ such that for all $j \in \{1, \dots, m\}$, $s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$ is a proper face of $B(F)$ with non-empty relative interior.*

Proof We have a face from Prop. 10, and it has non empty interior by applying Prop. 11 on each submodular function G_j . ■

The next proposition computes the Fenchel conjugate of the Lovász extensions restricted to $[0, 1]^p$, noting that by Prop. 8, the regular Fenchel conjugate of the unrestricted Lovász extension is the indicator function of the base polyhedron (for a definition of Fenchel conjugates, see [9, 10]). This allows a form of conjugacy between set-functions and convex functions.

Proposition 13 (Conjugate of a submodular function) *Let F be a submodular function such that $F(\emptyset) = 0$. The conjugate $\tilde{f} : \mathbb{R}^p \rightarrow \mathbb{R}$ of F is defined as $\tilde{f}(s) = \max_{A \subset V} s(A) - F(A)$. Then, the conjugate function \tilde{f} is convex, and is equal to the Fenchel-conjugate of the Lovász extension restricted to $[0, 1]^p$. Moreover, for all $A \subset V$, $F(A) = \max_{s \in \mathbb{R}^p} s(A) - \tilde{f}(s)$.*

Proof The function \tilde{f} is a maximum of linear functions and thus it is convex. We have for $s \in \mathbb{R}^p$:

$$\max_{w \in [0,1]^p} w^\top s - f(w) = \max_{A \subset V} s(A) - F(A) = \tilde{f}(s)$$

because $F - s$ is submodular and because of Proposition 7, which leads to first the desired result. The last assertion is a direct consequence of the fact that $F(A) = f(1_A)$. ■

4 Minimizers of submodular functions

In this section, we review some relevant results for submodular function minimization (for which algorithms are presented in Section 8).

Proposition 14 (Property of minimizers of submodular functions) *Let F be a submodular function such that $F(\emptyset) = 0$. The set $A \subset V$ is a minimizer of F on 2^V if and only if A is a minimizer of the function from 2^A to \mathbb{R} defined as $B \subset A \mapsto F(B)$, and if \emptyset is a minimizer of the function from $2^{V \setminus A}$ to \mathbb{R} defined as $B \subset V \setminus A \mapsto F(B \cup A) - F(A)$.*

Proof The set of two conditions is clearly necessary. To show that it is sufficient, we let $B \subset V$, we have: $F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \geq F(A) + F(A)$, by using the submodularity of F and then the set of two conditions. This implies that $F(A) \leq F(B)$, for all $B \subset V$, hence the desired result. ■

The following proposition provides a useful step towards submodular function minimization. In fact, it is the starting point of most polynomial-time algorithms presented in Section 8.

Proposition 15 (Dual of minimization of submodular functions) *Let F be a submodular function such that $F(\emptyset) = 0$. We have:*

$$\min_{A \subset V} F(A) = \max_{s \in B(F)} s_-(V), \quad (5)$$

where $s_- = \min\{s, 0\}$. Moreover, given $A \subset V$ and $s \in B(F)$, we always have $F(A) \geq s_-(V)$ with equality if and only if $\{s < 0\} \subset A \subset \{s \leq 0\}$ and A is tight for s , i.e., $s(A) = F(A)$.

We also have

$$\min_{A \subset V} F(A) = \max_{s \in P(F), s \leq 0} s(V). \quad (6)$$

Moreover, given $A \subset V$ and $s \in P(F)$ such that $s \leq 0$, we always have $F(A) \geq s(V)$ with equality if and only if $\{s < 0\} \subset A$ and A is tight for s , i.e., $s(A) = F(A)$.

Proof We have, by convex duality, and Props. 7 and 8:

$$\min_{A \subset V} F(A) = \min_{w \in [0,1]^p} f(w) = \min_{w \in [0,1]^p} \max_{s \in B(F)} w^\top s = \max_{s \in B(F)} \min_{w \in [0,1]^p} w^\top s = \max_{s \in B(F)} s_-(V).$$

Strong duality indeed holds because of Slater's condition ($[0,1]^p$ has non empty interior). Moreover, we have, for all $A \subset V$ and $s \in B(F)$:

$$F(A) \geq s(A) = s(A \cap \{s < 0\}) + s(A \cap \{s > 0\}) \geq s(A \cap \{s < 0\}) \geq s_-(V)$$

with equality if there is equality in the three inequalities. The first one leads to $s(A) = F(A)$. The second one leads to $A \cap \{s > 0\} = \emptyset$, and the last one leads to $\{s < 0\} \subset A$. Moreover,

$$\begin{aligned} \max_{s \in P(F), s \leq 0} s(V) &= \max_{s \in P(F)} \min_{w \geq 0} s^\top 1_V - w^\top s = \min_{w \geq 0} \max_{s \in P(F)} s^\top 1_V - w^\top s \\ &= \min_{1 \geq w \geq 0} f(1_V - w) \text{ because of property (c) in Prop. 8} \\ &= \min_{A \subset V} F(A) \text{ because of Prop. 7.} \end{aligned}$$

Moreover, given $s \in P(F)$ such that $s \leq 0$ and $A \subset V$, we have:

$$F(A) \geq s(A) = s(A \cap \{s < 0\}) \geq s(V)$$

with equality if and only if A is tight and $\{s < 0\} \subset A$. ■

5 Operations that preserve submodularity

In this section, we present several ways of building submodular functions from existing ones. For all of these, we describe how the Lovász extensions and the submodular polyhedra are affected. Note that in many cases, operations are simpler in terms of polyhedra.

Proposition 16 (Restriction of a submodular function) *let F be a submodular function such that $F(\emptyset) = 0$ and $A \subset V$. The restriction of F on A , denoted F_A is a set-function on A defined as $F_A(B) = F(B)$ for $B \subset A$. The function f_A is submodular. Moreover, if we can write the Lovász extension of F as $f(w) = f(w_A, w_{V \setminus A})$, then the Lovász extension of F_A is $f_A(w_A) = f(w_A, 0)$. Moreover, the submodular polyhedron $P(F_A)$ is simply the projection of $P(F)$ on the components indexed by A , i.e., $s \in P(F_A)$ if and only if $\exists t$ such that $(s, t) \in P(F)$.*

Proof Submodularity and the form of the Lovász extension are straightforward from definitions. To obtain the submodular polyhedron, notice that we have $f_A(w_A) = f(w_A, 0) = \max_{(s,t) \in P(F)} w_A^\top s + 0^\top t$, which implies the desired result, this shows that the Fenchel-conjugate of the Lovász extensions is the indicator function of a polyhedron. ■

Proposition 17 (Contraction of a submodular function) *let F be a submodular function such that $F(\emptyset) = 0$ and $A \subset V$. The contraction of F on A , denoted F^A is a set-function on $V \setminus A$ defined as $F^A(B) = F(A \cup B) - F(A)$ for $B \subset V \setminus A$. The function F^A is submodular. Moreover, if we can write the Lovász extension of F as $f(w) = f(w_A, w_{V \setminus A})$, then the Lovász extension of F^A is $f^A(w_{V \setminus A}) = f(1_A, w_{V \setminus A}) - F(A)$. Moreover, the submodular polyhedron $P(F^A)$ is simply the projection of $P(F) \cap \{s(A) = F(A)\}$ on the components indexed by $V \setminus A$, i.e., $t \in P(F^A)$ if and only if $\exists s \in P(F) \cap \{s(A) = F(A)\}$, such that $s_{V \setminus A} = t$.*

Proof Submodularity and the form of the Lovász extension are straightforward from definitions. Let $t \in \mathbb{R}^{|V \setminus A|}$. If $\exists s \in P(F) \cap \{s(A) = F(A)\}$, such that $s_{V \setminus A} = t$, then we have for all $B \subset V \setminus A$, $t(B) = t(B) + s(A) - F(A) \leq F(A \cup B) - F(A)$, and hence $t \in P(F^A)$. If $t \in P(F^A)$, then take any $v \in B(F_A)$ and concatenate v and t into s . Then, for all subsets $C \subset V$, $s(C) = s(C \cap A) + s(C \cap (V \setminus A)) = v(C \cap A) + t(C \cap (V \setminus A)) \leq F(C \cap A) + F(A \cup (C \cap (V \setminus A))) - F(A) = F(C \cap A) + F(A \cup C) - F(A) \leq F(C)$ by submodularity. Hence $s \in P(F)$. ■

The next proposition shows how to build a new submodular function from an existing one, by partial minimization. Note the similarity (and the difference) between the submodular polyhedra for a partial minimum (Prop. 18) and for the restriction defined in Prop. 16.

Proposition 18 (Partial minimum of a submodular function) *We consider a submodular function G on $V \cup W$, where $V \cap W = \emptyset$ (and $|W| = q$), with Lovász extension $g : \mathbb{R}^{p+q} \rightarrow \mathbb{R}$. We consider, for $A \subset V$, $F(A) = \min_{B \subset W} G(A \cup B) - \min_{B \subset W} G(B)$. The set-function F is submodular and such that $F(\emptyset) = 0$. Its Lovász extension is such that for all $w \in [0, 1]^p$, $f(w) = \min_{v \in [0, 1]^q} g(w, v) - \min_{v \in [0, 1]^q} g(0, v)$. Moreover, if $\min_{B \subset W} G(B) = 0$, we have for all $w \in \mathbb{R}_+^p$, $f(w) = \min_{v \in \mathbb{R}_+^q} g(w, v)$, and the submodular polyhedron $P(F)$ is the set of $s \in \mathbb{R}^p$ such that there exists $t \in \mathbb{R}_+^q$, such that $(s, t) \in P(G)$.*

Proof Define $c = \min_{B \subset W} G(B)$, which is independent of A . We have, for $A, A' \subset V$, and any $B, B' \subset W$, by definition of F :

$$\begin{aligned} F(A \cup A') + F(A \cap A') &\leq -2c + G([A \cup A'] \cup [B' \cup B']) + G([A \cap A'] \cup [B' \cap B']) \\ &= -2c + G([A \cup B] \cup [A' \cup B']) + G([A \cup B] \cap [A' \cup B']) \\ &\leq -2c + G(A \cup B) + G(A' \cup B') \text{ by submodularity.} \end{aligned}$$

Minimizing with respect to B and B' leads to the submodularity of F .

Following Prop. 13, we can get the conjugate function \tilde{f} from the one \tilde{g} of G . For $s \in \mathbb{R}^p$, we have, by definition, $\tilde{f}(s) = \max_{A \subset V} s(A) - F(A) = \max_{A \cup B \subset V \cup W} s(A) + c - G(A \cup B) =$

$c + \tilde{g}(s, 0)$. We thus get from Prop. 13 that for $w \in [0, 1]^p$,

$$\begin{aligned}
f(w) &= \max_{s \in \mathbb{R}^p} w^\top s - \tilde{f}(s) \\
&= \max_{s \in \mathbb{R}^p} w^\top s - \tilde{g}(s, 0) - c \\
&= \max_{s \in \mathbb{R}^p} \min_{(\tilde{w}, v) \in [0, 1]^{p+q}} w^\top s - \tilde{w}^\top s + g(\tilde{w}, v) - c \text{ by applying Prop. 13,} \\
&= \min_{(\tilde{w}, v) \in [0, 1]^{p+q}} \max_{s \in \mathbb{R}^p} w^\top s - \tilde{w}^\top s + g(\tilde{w}, v) - c \\
&= \min_{v \in [0, 1]^q} g(w, v) - c \text{ by maximizing with respect to } s.
\end{aligned}$$

Note that $c = \min_{B \subset W} G(B) = \min_{v \in [0, 1]^q} g(0, v)$.

For any $w \in \mathbb{R}_+^p$, for any $\lambda \geq \|w\|_\infty$, we have $w/\lambda \in [0, 1]^p$, and thus

$$\begin{aligned}
f(w) &= \lambda f(w/\lambda) = \min_{v \in [0, 1]^q} \lambda g(w/\lambda, v) - c\lambda = \min_{v \in [0, 1]^q} g(w, \lambda v) - c\lambda \\
&= \min_{v \in [0, \lambda]^q} g(w, v) - c\lambda.
\end{aligned}$$

Thus, if $c = 0$, we have $f(w) = \min_{v \in \mathbb{R}_+^q} g(w, v)$, by letting $\lambda \rightarrow +\infty$. We then also have:

$$\begin{aligned}
f(w) &= \min_{v \in \mathbb{R}_+^q} g(w, v) = \min_{v \in \mathbb{R}_+^q} \max_{(s, t) \in P(G)} w^\top s + v^\top t \\
&= \max_{(s, t) \in P(G), t \in \mathbb{R}_+^q} w^\top s.
\end{aligned}$$

■

The following propositions give an interpretation of the intersection between the submodular polyhedron and sets of the form $\{s \leq z\}$ and $\{s \geq z\}$.

Proposition 19 (Convolution of a submodular function and a modular function)

Let F be a submodular function such that $F(\emptyset) = 0$ and $z \in \mathbb{R}^p$. Define $G(A) = \min_{B \subset A} F(B) + z(A \setminus B)$. Then G is submodular and the submodular polyhedron $P(G)$ is equal to $P(F) \cap \{s \leq z\}$. Moreover, for all $A \subset V$, $G(A) \leq F(A)$ and $G(A) \leq z(A)$.

Proof Let $A, A' \subset V$, and B, B' the corresponding minimizers defining $G(A)$ and $G(A')$. We have:

$$\begin{aligned}
G(A) + G(A') &= F(B) + z(A \setminus B) + F(B') + z(A' \setminus B') \\
&\geq F(B \cup B') + F(B \cap B') + z(A \setminus B) + z(A' \setminus B') \text{ by submodularity} \\
&= F(B \cup B') + F(B \cap B') + z([A \cup A'] \setminus [B \cup B']) + z([A \cap A'] \setminus [B \cap B']) \\
&\geq G(A \cup A') + G(A \cap A') \text{ by definition of } G,
\end{aligned}$$

hence the submodularity of G . If $s \in P(G)$, then $\forall B \subset A \subset V$, $s(A) \leq G(A) \leq F(B) + z(A \setminus B)$. From $B = A$, we get that $s \in P(F)$; from $B = \emptyset$, we get $s \leq z$, and hence $s \in P(F) \cap \{s \leq z\}$. If $s \in P(F) \cap \{s \leq z\}$, for all $\forall B \subset A \subset V$, $s(A) = s(A \setminus B) + s(B) \leq z(A \setminus B) + F(B)$; by minimizing with respect to B , we get that $s \in P(G)$.

We get $G(A) \leq F(A)$ by taking $B = A$ in the definition of $G(A)$, and we get $G(A) \leq z(A)$ by taking $B = \emptyset$. ■

Proposition 20 (Monotonization of a submodular function) *Let F be a submodular function such that $F(\emptyset) = 0$. Define $G(A) = \min_{B \supseteq A} F(B) - \min_{B \subset V} F(B)$. Then G is submodular such that $G(\emptyset) = 0$, and the base polyhedron $B(G)$ is equal to $B(F) \cap \{s \geq 0\}$. Moreover, G is non-decreasing, and for all $A \subset V$, $G(A) \leq F(A)$.*

Proof Let $c = \min_{B \subset V} F(B)$. Let $A, A' \subset V$, and B, B' the corresponding minimizers defining $G(A)$ and $G(A')$. We have:

$$\begin{aligned} G(A) + G(A') &= F(B) + F(B') - 2c \\ &\geq F(B \cup B') + F(B \cap B') - 2c \text{ by submodularity} \\ &\geq G(A \cup A') + G(A \cap A') \text{ by definition of } G, \end{aligned}$$

hence the submodularity of G . It is obviously non-decreasing. We get $G(A) \leq F(A)$ by taking $B = A$ in the definition of $G(A)$. Since G is increasing, $B(G) \subset \mathbb{R}_+^p$ (because all of its extreme points, obtained by the greedy algorithm, are in \mathbb{R}_+^p). By definition of G , $B(G) \subset B(F)$. Thus $B(G) \subset B(F) \cap \mathbb{R}_+^p$. The opposite inclusion is trivial from the definition. ■

6 Proximal optimization problems

In this section, we consider separable convex functions and the minimization of such functions penalized by the Lovász extension of a submodular function. When the separable functions are all quadratic functions, those problems are often referred to as *proximal problems* (see, e.g., [17] and references therein). We make the simplifying assumption that the problem is strictly convex and differentiable (but not necessarily quadratic), but sharp statements could also be made in the general case. The next proposition shows that it is equivalent to the maximization of a separable concave function over the base polyhedron.

Proposition 21 (Dual of proximal optimization problem) *Let ψ_1, \dots, ψ_p be p continuously differentiable strictly convex functions on \mathbb{R} , with Fenchel-conjugates $\psi_1^*, \dots, \psi_p^*$. The two following optimization problems are dual of each other:*

$$\min_{w \in \mathbb{R}^p} f(w) + \sum_{j=1}^p \psi_j(w_j), \quad (7)$$

$$\max_{s \in B(F)} - \sum_{j=1}^p \psi_j^*(-s_j). \quad (8)$$

The pair (w, s) is optimal if and only if $s_k = -\psi'_k(w_k)$ for all $k \in \{1, \dots, p\}$, and $s \in B(F)$ is optimal for the maximization of $w^\top s$ over $s \in B(F)$ (see Prop. 10 for optimality conditions).

Proof We have:

$$\begin{aligned}
\min_{w \in \mathbb{R}^p} f(w) + \sum_{j=1}^p \psi_j(w_j) &= \min_{w \in \mathbb{R}^p} \max_{s \in B(F)} w^\top s + \sum_{j=1}^p \psi_j(w_j) \\
&= \max_{s \in B(F)} \min_{w \in \mathbb{R}^p} w^\top s + \sum_{j=1}^p \psi_j(w_j) \\
&= \max_{s \in B(F)} - \sum_{j=1}^p \psi_j^*(-s_j),
\end{aligned}$$

where ψ_j^* is the Fenchel-conjugate of ψ_j (which may in general have a domain strictly included in \mathbb{R}). Thus the separably penalized problem defined in Eq. (7) is equivalent to a separable maximization over the base polyhedron (i.e., Eq. (8)). Moreover, the unique optimal s for Eq. (8) and the unique optimal w for Eq. (7) are related through $s_j = -\psi_j'(w_j)$ for all $j \in V$. ■

For simplicity, we now assume that for all $j \in V$, functions ψ_j are such that $\sup_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = +\infty$ and $\inf_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = -\infty$. This implies that the Fenchel-conjugates ψ_j^* are defined and finite on \mathbb{R} . Following [16], we also consider a sequence of set optimization problems, parameterized by $\alpha \in \mathbb{R}$:

$$\min_{ACV} F(A) + \sum_{j \in A} \psi_j'(\alpha) \tag{9}$$

We denote by A^α any minimizer of Eq. (9). Note that A^α is a minimizer of a submodular function $F + \psi'(\alpha)$, where $\psi'(\alpha) \in \mathbb{R}^p$ is the vector of components $\psi_k'(\alpha)$.

The main property, as shown in [16], is that solving Eq. (7), which is a convex optimization problem, is equivalent to solving Eq. (9) for all possible α , which are submodular optimization problems. We first show a monotonicity property of solutions of Eq. (9).

Proposition 22 (Monotonicity of solutions) *If $\alpha > \beta$, then any solutions A^α and A^β of Eq. (9) for α and β satisfy $A^\alpha \subset A^\beta$.*

Proof We have, by optimality of A^α and A^β :

$$\begin{aligned}
F(A^\alpha) + \sum_{j \in A^\alpha} \psi_j'(\alpha) &\leq F(A^\alpha \cup A^\beta) + \sum_{j \in A^\alpha \cup A^\beta} \psi_j'(\alpha) \\
F(A^\beta) + \sum_{j \in A^\beta} \psi_j'(\beta) &\leq F(A^\alpha \cap A^\beta) + \sum_{j \in A^\alpha \cap A^\beta} \psi_j'(\beta),
\end{aligned}$$

and by summing the two inequalities and using the submodularity of F ,

$$\sum_{j \in A^\alpha} \psi_j'(\alpha) + \sum_{j \in A^\beta} \psi_j'(\beta) \leq \sum_{j \in A^\alpha \cup A^\beta} \psi_j'(\alpha) + \sum_{j \in A^\alpha \cap A^\beta} \psi_j'(\beta),$$

which is equivalent to $\sum_{j \in A^\alpha \setminus A^\beta} (\psi_j'(\beta) - \psi_j'(\alpha)) \geq 0$, which implies, since for all $j \in V$, $\psi_j'(\beta) < \psi_j'(\alpha)$ (because of strict convexity), that $A^\alpha \setminus A^\beta = \emptyset$. ■

The next proposition shows that we can obtain the unique solution of Eq. (7) from all solutions of Eq. (9).

Proposition 23 (Proximal problem from submodular function minimizations) *Given any solutions A^α of problems in Eq. (9), for all $\alpha \in \mathbb{R}$, we define the vector $u \in \mathbb{R}^p$ as*

$$u_j = \sup(\{\alpha \in \mathbb{R}, j \in A^\alpha\}).$$

Then u is the unique solution of the proximal problem in Eq. (7).

Proof Because $\inf_{\alpha \in \mathbb{R}} \psi'_j(\alpha) = -\infty$, for α small enough, we must have $A^\alpha = V$, and thus u_j is well-defined and finite for all $j \in V$.

If $\alpha > u_j$, then, by definition of u_j , $j \notin A^\alpha$. This implies that $A^\alpha \subset \{j \in V, u_j \geq \alpha\} = \{u \geq \alpha\}$. Moreover, if $u_j > \alpha$, there exists $\beta \in (\alpha, u_j)$ such that $j \in A^\beta$. By the monotonicity property of Prop. 22, A^β is included in A^α . This implies $\{u > \alpha\} \subset A^\alpha$.

We have for all $w \in \mathbb{R}^p$, and β less than the smallest of $(w_j)_-$ and the smallest of $(u_j)_-$:

$$\begin{aligned} & f(u) + \sum_{j=1}^p \psi_j(u_j) \\ &= \int_0^\infty F(\{u \geq \alpha\}) d\alpha + \int_\beta^0 (F(\{u \geq \alpha\}) - F(V)) d\alpha + \sum_{j=1}^p \left\{ \int_\beta^{u_j} \psi'_j(\alpha) d\alpha + \psi_j(\beta) \right\} \\ &= C + \int_\beta^\infty \left[F(\{u \geq \alpha\}) + \sum_{j=1}^p 1_{u \geq \alpha} \psi'_j(\alpha) \right] d\alpha \text{ with } C = \int_0^\beta F(V) d\alpha + \sum_{j=1}^p \psi_j(\beta) \\ &\leq C + \int_\beta^\infty \left[F(\{w \geq \alpha\}) + \sum_{j=1}^p 1_{w \geq \alpha} \psi'_j(\alpha) \right] d\alpha \text{ by optimality of } A^\alpha \\ &= f(w) + \sum_{j=1}^p \psi_j(w_j). \end{aligned}$$

This shows that u is the unique optimum of problem in Eq. (7). ■

From the previous proposition, we also get the following corollary, i.e., all solutions of Eq. (9) may be obtained from the single solutions of Eq. (7).

Proposition 24 (Submodular function minimizations from proximal problem) *If u is the unique minimizer of Eq. (7), then for all $\alpha \in \mathbb{R}$, the minimal minimizer of Eq. (9) is $u > \alpha$ and the maximal minimizer is $\{u \geq \alpha\}$, that is, the minimizers A^α are the sets such that $\{u > \alpha\} \subset A^\alpha \subset \{u \geq \alpha\}$.*

Given the previous propositions, we can solve a sequence of problems in Eq. (9), with decreasing α 's, in order to obtain the unique minimizer w of Eq. (7). Note that because of the monotonicity, the sets A^α can only increase. When a certain $j \in V$ enters A^α , then w_j is exactly equal to the corresponding α . Once we know the largest values of w , we may redefine the problem by restricting on the unknown indices of w , which is valid for smaller values of α .

7 Optimization over the base polyhedron

Optimization of separable functions over the base polyhedron has many applications, e.g., minimization of a submodular function (from Prop. 15), proximal methods described in Section 6 (e.g., Prop 21). In this section, we study these problems in more details.

7.1 Optimality conditions

We first show that when optimizing on the base polyhedron $B(F)$, then one only needs to look at directions of the form $\delta_k - \delta_q$ for certain pairs (k, q) , which will be said *exchangeable* ($\delta_k \in \mathbb{R}^p$ is the vector which is entirely equal to zero, except a component equal to one at position k , which can also denote $1_{\{k\}}$).

Definition 5 (Tight sets) *Given a base $s \in B(F)$, a set $A \subset V$ is said tight if $s(A) = F(A)$.*

Proposition 25 (Lattice of tight sets) *If A and B are tight for $s \in B(F)$, then $A \cap B$ and $A \cup B$ are also tight for s .*

Proof We have:

$$F(A \cup B) + F(A \cap B) \geq s(A \cup B) + s(A \cap B) = s(A) + s(B) = F(A) + F(B) \geq F(A \cup B) + F(A \cap B).$$

Thus there is equality everywhere, which leads to the desired result. Note that this shows that the set of tight sets for $s \in \mathbb{R}^p$ is a lattice. ■

We now define the notion of exchangeable pairs, which we allow us to describe the tangent cone of the base polyhedron in Prop. 28.

Definition 6 (Dependence function and exchangeable pairs) *Given a base $s \in B(F)$ and $k \in A$, the dependence function $\text{Dep}(s, k)$ is the (non-empty) smallest tight set that contains k . If $g \in \text{Dep}(s, k)$, then the pair (k, g) is said exchangeable.*

Prop. 25 shows that $\text{Dep}(s, k)$ is indeed well-defined because V is tight and contains k , and the set of tight sets containing k is a lattice. The following proposition details the most important properties of exchangeable pairs, which are straightforward given the definition (in fact, the conjunction of these two properties is equivalent to the definition of exchangeable pairs).

Proposition 26 (Properties of exchangeable pairs) *Let $s \in B(F)$ and (k, q) is an exchangeable pair for s . Then:*

- (a) *there exists $A \subset V$ such that $k, q \in A$ and A is tight for s ,*
- (b) *if $A \subset V$ is tight for s , then $k \in A \Rightarrow q \in A$.*

The next proposition shows that only these exchangeable pairs need to be considered for checking optimality conditions for optimization over the base polyhedron.

Proposition 27 (Maximizers of support function of the base polyhedron) *Let $w \in \mathbb{R}^p$. The base $s \in B(F)$ is a maximizer of $\max_{s \in B(F)} s^\top w$ if and only for all $k \in V$ and $q \in \text{Dep}(s, k)$, $w_k \leq w_q$ (i.e., for all exchangeable pairs).*

Proof If s is optimal, then if $k \in V$ and $q \in \text{Dep}(s, k)$, then for $\alpha > 0$ small enough, $s' = s + \alpha(\delta_k - \delta_q)$ is in $B(F)$ (indeed, if A is not tight, then a small modification of s does not change the constraint, and if A is tight, if $A \ni k$, then $q \in A$ by Prop. 26 and thus $s'(A) = F(A)$; finally, if A tight and $k \notin A$, then $s'(A)$ can only decrease). Optimality of s implies that $w_k \leq w_q$.

If the condition is true, we can order values of w , as $w_{B_1} > \dots > w_{B_m}$ (where $w_k = w_{B_j}$ for $k \in B_j$). Let $A_j = B_1 \cup \dots \cup B_j$, so that $k \in A_j$ if and only if $w_k \geq w_{B_j}$. This implies, because of the condition, that $A_j = \bigcup_{k \in A_j} \text{Dep}(s, k)$, and thus that A_j is tight (as a union of tight sets), i.e., $s(A_j) = F(A_j)$. Then, for any $t \in B(F)$,

$$\begin{aligned} s^\top w - t^\top w &= \sum_{k \in V} w_k (s_k - t_k) = \sum_{i=1}^m w_{B_i} [s(B_i) - t(B_i)] \\ &= \sum_{i=1}^m w_{B_i} [(s - t)(A_i) - (s - t)(A_{i-1})] \\ &= \sum_{i=1}^m w_{B_i} [(F - t)(A_i) - (F - t)(A_{i-1})] \\ &= \sum_{i=1}^m [F(A_i) - t(A_i)] (w_{B_i} - w_{B_{i+1}}) \geq 0. \end{aligned}$$

Thus s is optimal. Note that this also a consequence of Prop. 10. ■

From Prop. 27, we may now deduce the tangent cone of the base polyhedron, from which we then obtain optimality conditions.

Proposition 28 (Tangent cone of base polyhedron) *Let $s \in B(F)$, the tangent cone of $B(F)$ at s is generated by vectors $\delta_k - \delta_q$ for all $k \in V$ and $q \in \text{Dep}(s, k)$, i.e., for all exchangeable pairs (k, q) .*

Proof Given the proof of Prop. 27, each of the vectors $\delta_k - \delta_q$ belongs to the tangent cone. If the tangent cone strictly contains the conic hull of these vectors, by Farkas lemma (see, e.g., [10]), there exists y in the tangent cone and $w \in \mathbb{R}^p$, such that for all exchangeable pairs (k, q) , $w^\top (\delta_k - \delta_q) \leq 0$ and $w^\top y > 0$. By the last proposition, s is an optimal base for the weight vector w , however, $s + \alpha y \in P(F)$ for $\alpha > 0$ sufficiently small and $(s + \alpha y)^\top w > s^\top w$, which is a contradiction. ■

Proposition 29 (Optimality conditions for separable optimization) *Let g_j be convex functions on \mathbb{R} , $j = 1, \dots, p$. Then $s \in B(F)$ is a minimizer of $\sum_{j \in V} g_j(s_j)$ over $s \in B(F)$ if and only if for all exchangeable pairs (k, g) , $\partial_+ g_k(s_k) \geq \partial_- g_q(s_q)$, where $\partial_+ g_k(s_k)$ is the right-derivative of g_k at s_k and $\partial_- g_q(s_q)$ is the left-derivative of g_q at s_q .*

Proof This is immediate from Prop. 28 related to the tangent cone of $B(F)$. ■

We can give an alternative description of optimality conditions based on Prop. 10, which we give only for differentiable functions for simplicity.

Proposition 30 (Alternative optimality conditions for separable optimization) *Let g_j be differentiable convex functions on \mathbb{R} , $j = 1, \dots, p$. Let $s \in B(F)$ and $w \in \mathbb{R}^p$ defined as $\forall k \in V, w_k = g'_k(s_k)$; define $B(\alpha) = \{w \leq \alpha\}$ for $\alpha \in \mathbb{R}$. Then, s is a minimizer of $\sum_{j \in V} g_j(s_j)$ over $s \in B(F)$ if and only if for all $\alpha \in \mathbb{R}$, the sets $B(\alpha)$ are tight.*

Proof Note that the condition has to be checked only for α belonging to the of values taken by w . We consider the unique values $v_1 < \dots < v_m$, taken at sets A_1, \dots, A_m (i.e., $V = A_1 \cup \dots \cup A_m$ and $\forall k \in A_i, w_k = v_i$). The condition then becomes that all $B_i = A_1 \cup \dots \cup A_i$ are tight for s . This is immediate from Prop. 10. Indeed, s is optimal if and only if s is optimal for the problem $\min_{s \in B(F)} w^\top s$. ■

7.2 Lexicographically optimal bases

We can give another interpretation to optimality conditions in Prop. 29. Given a vector $s \in \mathbb{R}^p$, we denote by $T(s) \in \mathbb{R}^p$, the sequence of components of s in order of increasing magnitude. That is, if $s_{j_1} \leq s_{j_2} \leq \dots \leq s_{j_p}$, then $T(s) = (s_{j_1}, \dots, s_{j_p})$. Given two vectors s and s' in \mathbb{R}^p , s is said lexicographically greater than or equal to s' , if either (a) $s = s'$, or, (b) $s \neq s'$, and for the minimum index i such that $s_i \neq s'_i$, then $s_i \geq s'_i$.

We now show that finding a base $s \in B(F)$ that lexicographically maximizes the ordered vector of derivatives $g'_k(s_k)$ is equivalent to minimizing $\sum_{k \in V} g_k(s_k)$ over the base polyhedron. Many algorithms for proximal problems are in fact often cast as maximization for such lexicographical orders (see, e.g. [18]).

Proposition 31 (Lexicographically optimal base) *Let g_j be differentiable strictly convex functions on \mathbb{R} , $j = 1, \dots, p$. Then $s \in B(F)$ lexicographically maximizes the vector $T(g'(s)) = T[(g'_1(s_1), \dots, g'_p(s_p))]$ over $s \in B(F)$ if and only if s is a minimizer of $\sum_{k \in V} g_k(s_k)$ over the base polyhedron $B(F)$.*

Proof First assume that $s \in B(F)$ lexicographically maximizes the vector $T(g'(s)) = T[(g'_1(s_1), \dots, g'_p(s_p))]$ over $s \in B(F)$. Then, for any exchangeable pair (k, q) associated with s , we have that $t = s + \alpha(\delta_k - \delta_q) \in B(F)$ for α sufficiently small (from Prop. 28). Moreover, all components $g'_j(s_j)$ are unchanged, except the k -th and q -th position, for which we have $g'_k(t_k) > g'_k(s_k)$ and $g'_q(t_q) < g'_q(s_q)$. Thus, if $g'_k(s_k) < g'_q(s_q)$, $T(g'(t))$ is lexicographically strictly greater than $T(g'(s))$, which is a contradiction. This implies that for all exchangeable pairs, $g'_k(s_k) \geq g'_q(s_q)$, which implies, by Prop. 29 that s is indeed a minimizer.

Let now s be a minimizer of $\sum_{k \in V} g_k(s_k)$ over the base polyhedron $B(F)$. Let t be a base in $B(F)$ such that $T(g'(t))$ is lexicographically greater than or equal to $T(g'(s))$. We consider $v = g'(t) \in \mathbb{R}^p$ and $w = g'(s) \in \mathbb{R}^p$. We denote by $w_{B_1} < \dots < w_{B_m}$ the m distinct values of $w \in \mathbb{R}^p$, taken on the subsets A_j , $j = 1, \dots, m$. From Prop. 10, the sets $B_j = A_1 \cup \dots \cup A_j$ are tight for s . We show by induction on j that for $k \in B_j$, $s_k = t_k$, which will show that we must have $s = t$, and thus that $T(g'(s))$ is lexicographically optimal.

This is true for $j = 0$, and if we assume it is true for j , then, since $T(v)$ is lexicographically greater than or equal to $T(w)$, we have for all $k \in A_{j+1}$, $v_k \geq w_k$ (since all the smaller ones are equal by the induction assumption), which implies, by strict convexity of g_k that $t_k \geq s_k$. Moreover, since B_{j+1} is tight, we have $F(B_{j+1}) \geq t(B_{j+1}) \geq s(B_{j+1}) = F(B_{j+1})$, which implies that $t_k = s_k$ for $k \in A_{j+1}$. ■

7.3 Optimization for proximal problems

We can now obtain from the base polyhedron perspective the previous results linking problems in Eq. (7) and Eq. (9), i.e., give an alternative proof of Prop. 24 from Section 6.

Indeed, from Prop. 29, s is optimal for $\max_{s \in B(F)} -\sum_{j=1}^p \psi_j^*(-s_j)$ if and only if for all exchangeable pairs (k, q) for s , $(\psi_k^*)'(-s_k) \leq (\psi_q^*)'(-s_q)$. If we denote $w_k = (\psi_k^*)'(-s_k)$ (which is equivalent to $s_k = -\psi_k'(w_k)$), then s is optimal if $w_k \leq w_q$ for all exchangeable pairs (k, q) .

Let $\alpha \in \mathbb{R}$, we consider the optimization problem

$$\max_{s \in B(F)} \sum_{k \in V} (s_k + \psi_k'(\alpha))_- = (s + \psi'(\alpha))_-(V). \quad (10)$$

From Prop. 29 and the fact that the right-derivative of $s_k \mapsto (s_k + \psi_k'(\alpha))_-$ is -1 for $s_k < \psi_k'(\alpha)$ and zero otherwise, and its left-derivative of $s_k \mapsto (s_k + \psi_k'(\alpha))_-$ is -1 for $s_k \leq -\psi_k'(\alpha)$ and zero otherwise, s is optimal if and only if for all exchangeable pairs (k, q) for s , we have $1_{\{s_k < -\psi_k'(\alpha)\}} \leq 1_{\{s_q \leq -\psi_q'(\alpha)\}}$, which is equivalent to the fact that $s_k < -\psi_k'(\alpha)$ implies that $s_q \leq -\psi_q'(\alpha)$.

If s is optimal for Eq. (8), then $(\psi_k^*)'(-s_k) \leq (\psi_q^*)'(-s_q)$ for all exchangeable pairs. Thus, if s is optimal for Eq. (8), then s is optimal for the maximization of Eq. (10) for all $\alpha \in \mathbb{R}$.

Finally, from Prop. 15, solving Eq. (10) is equivalent to minimizing the submodular function $F + \psi'(\alpha)$, which is exactly Eq. (9). Also, from Prop. 15, we have that any optimal A^α satisfies $\{s + \psi'(\alpha) < 0\} \subset A^\alpha \subset \{s + \psi'(\alpha) \leq 0\}$. Moreover, since at the optimum, $w_k + \psi_k'(s_k) = 0$, we thus have $s_k + \psi_k'(\alpha) < 0$ if and only if $w_k > \alpha$, and $s_k + \psi_k'(\alpha) \leq 0$ if and only if $w_k \geq \alpha$. We thus get back Prop. 24.

8 Submodular function minimization

Several generic algorithms may be used for the minimization of a submodular function. They are all based on a sequence of evaluations of $F(A)$ for certain subsets $A \subset V$. For specific functions, such as the ones defined from cuts, faster algorithms exist (see, e.g., [19, 6] and Section 10.2).

Note that maximizing submodular functions is a hard combinatorial problem in general. However, when maximizing a non-decreasing submodular function under a cardinality constraint, the simple greedy method allows to obtain a $(1 - 1/e)$ -approximation [20].

In this section, we first review classical approaches for submodular function minimization. The first approach presented in Section 8.1 is the most efficient in practice, but has no complexity bound. We briefly mention in Section 8.2 existing combinatorial algorithms with theoretical complexity bounds, but these are not used in practice. In Section 8.3, we consider certain submodular functions, so-called posimodular functions, for which simple combinatorial algorithms exist with better complexity.

We then present algorithms which are based on a sequence of submodular function minimization, and that can be used for problems such as line search in the submodular polyhedron or proximal problems.

8.1 Minimum-norm point algorithm

From Eq. (9) or Prop. 24, we obtain that if we know how to minimize $f(w) + \frac{1}{2}\|w\|_2^2$, or equivalently, minimize $\frac{1}{2}\|s\|_2^2$ such that $s \in B(F)$, then we get all minimizers of F from the negative components of s .

The minimum-norm point algorithm computes the minimum of $\|s\|_2^2$ for $s \in B(F)$. It uses an old algorithm from [21] that will find a minimum-norm base $s \in B(F)$ in a finite number of steps. This is made possible by the fact that we know how to efficiently maximize linear functions over $B(F)$, where solutions are obtained by the greedy algorithm from Prop. 5.

The complexity of each step of the algorithm is essentially $O(p)$ function evaluations and operations of order $O(p^3)$. However, there are no known upper bounds on the number of iterations.

Note that once we know which values of the optimum values s should be equal, greater or smaller, then, we obtain in closed form all values. Indeed, let $c_1 < c_2 < \dots < c_m$ the m different values taken by s (or w), and A_i the corresponding sets such that $w_k = c_j$ for $k \in A_j$. We then have:

$$c_j = \frac{f(A_1 \cup \dots \cup A_j) - f(A_1 \cup \dots \cup A_{j-1})}{|A_j|}$$

which allows to compute the values c_j knowing only the sets A_j .

8.2 Combinatorial algorithms

Algorithms are based on Prop. 15, i.e., on the identity $\min_{A \subset V} F(A) = \max_{s \in B(F)} s_-(V)$. Combinatorial algorithms will usually output the subset A and a base $s \in B(F)$ such that A is tight for s and $\{s < 0\} \subset A \subset \{s \leq 0\}$, as a certificate of optimality.

Most algorithms, will also output the largest minimizer A of F , or sometimes describe the entire lattice of minimizers. Best algorithms have polynomial complexity [22, 23, 24], but still have high complexity (typically $O(p^6)$ or more).

8.3 Minimizing posimodular functions

A submodular function F is said symmetric if for all $B \subset V$, $F(V \setminus B) = F(B)$. By applying submodularity, get that $2F(B) = F(V \setminus B) + F(B) \geq F(V) + F(\emptyset) = 2F(\emptyset) = 0$, which implies that F is non-negative. Hence its global minimum is attained at V and \emptyset .

Such functions can be minimized in time $O(p^3)$ over all *non-trivial* (i.e., different from \emptyset and V) subsets of V [25]. Moreover, the algorithm is valid for the regular minimization of *posimodular* functions [26], i.e., of functions that satisfies

$$\forall A, B \subset V, F(A) + F(B) \geq F(A \setminus B) + F(B \setminus A).$$

These include symmetric submodular functions as well as modular functions, and hence the sum of any of those (in particular, cuts with sinks and sources, as presented in Section 10.2).

8.4 Line search in submodular polyhedron

The general line search problem in the submodular polyhedron amounts to start from $s \in P(F)$ and search on the direction $t \in \mathbb{R}^p$, i.e., find the maximal $\lambda \geq 0$ such that $s + \lambda t \in P(F)$, which is equivalent to $\lambda t \in P(F - s)$. Note that since $s \in P(F)$, $F - s$ is submodular and non-negative.

We thus now assume that F is non-negative and that $s = 0$. Given $t \in \mathbb{R}^p$, we consider the problem of finding the largest $\lambda \geq 0$ such that $\lambda t \in P(F)$. We denote by μ the optimal value (which is finite, as soon as there is at least one $t_k > 0$, which we assume). We have $\lambda \leq \mu$ if and only if $g(\lambda) = \min_{A \subset V} F(A) - \lambda t(A) \geq 0$. More precisely, $g(\lambda) \geq 0$ if and only if $\lambda t \in P(F)$. Moreover, $g(0) = 0$ and g is non-increasing, which implies that g is zero on $[0, \mu]$ and then strictly negative.

We thus need to find the zero of the function $g(\lambda)$, which is piecewise affine. This can be done with the secant method, once we have a $\lambda > 0$ such that $g(\lambda) < 0$. Such a λ can be obtained by noting that $P(F)$ is included in $\{s, \forall k \in V, s_k \leq F(\{k\})\}$, which implies that if $\lambda > \min_{k \in V} \frac{F(\{k\})}{t_k}$, then $g(\lambda) < 0$.

The secant method is simply starting with a λ such that $g(\lambda) > 0$, and then find the minimizer A in the definition of $g(\lambda)$, and set $\lambda = F(A)/t(A)$, and start again in $g(\lambda) < 0$ (see [27] for more details). Note that if the minimum-norm point algorithm is used for submodular function minimization, then we obtain instead a minimizer of $w \mapsto f(w) - \lambda w^\top t + \frac{1}{2} \|w\|_2^2$, and we can also update λ as $\lambda = (f(w) + \frac{1}{2} \|w\|_2^2) / (w^\top t)$.

8.5 Homotopy method for proximal problems

We review in Section 8.5 and Section 8.6 two strategies for maximizing separable concave functions on the base polyhedron. One strategy is based on the equivalence with the sequence of minimizations of submodular functions (Prop. 23). The other one is based on a decomposition strategy.

The first method is based on the fact that if α is large enough, then $A^\alpha = \emptyset$ is optimum for Eq. (9). From Prop. prop:dualmin, this is valid as long as $0 \in P(F + \psi^l(\alpha))$, i.e., $-\psi^l(\alpha) \in P(F)$. The minimum $\alpha \in \mathbb{R}$ such that this is valid can be obtained by line search.

Once the minimal α is found, and A is the maximal tight set associated with $-\psi'(\alpha)$, then if $A = V$, $w = \alpha 1_V$. Otherwise, we let $w_A = \alpha 1_A$, and in order to determine $w_{V \setminus A}$ we recursively apply the same procedure to the function $F_{V \setminus A} : 2^{V \setminus A} \rightarrow \mathbb{R}$, defined as $F_{V \setminus A}(B) = F(B)$ (i.e., restriction of F to $V \setminus A$).

This algorithm, adapted from [28] (see also [11, Sec. 9.2]), requires to be able to find the minimum α such that $-\psi'(\alpha) \in P(F)$. This may be done as follows (same procedure as in Section 8.4, but extended to non quadratic functions).

Consider $g(\alpha) = \min_{A \subset V} F(A) + \psi'(\alpha)(A)$. The function is piecewise smooth and strictly increasing. It is equal to zero if and only if $-\psi'(\alpha) \in P(F)$, and it is strictly negative otherwise. We start with a point α_0 such that $g(\alpha_0) < 0$, we let A_0 be a minimizer in the definition of $g(\alpha_0)$. We find the unique α_1 such that $F(A_0) + \psi'(\alpha_1)(A_0) = 0$ and we start again, until we have $g(\alpha_1) = 0$.

In order to find α_0 such that $g(\alpha_0) < 0$, we use the fact that $P(F) \subset \prod_{k \in V} (-\infty; F(\{k\})]$, and thus if there exists $k \in V$, $\psi'_k(\alpha) > -F(\{k\})$, then $-\psi'(\alpha) \notin P(F)$. We can thus consider $\alpha_0 = \min_{k \in V} (\psi'_k)^{-1}(-F(\{k\}))$.

8.6 Decomposition algorithm for proximal problems

We adapt the algorithm of [29] and [11, Sec. 8.2]. Note that it can be slightly modified for problems with non-decreasing submodular functions [29] (see also Section 9).

For simplicity, we consider *strictly convex differentiable* functions g_j , $j = 1, \dots, p$, and the following algorithm:

1. Find the unique minimizer $t \in \mathbb{R}^p$ of $\sum_{j \in V} g_j(t_j)$ such that $t(V) = F(V)$.
2. Minimize the submodular function $F - t$, i.e., find the *largest* $A \subset V$ that minimizes $F(A) - t(A)$.
3. If $A = V$, then t is optimal. Exit.
4. Find a minimizer s_A of $\sum_{j \in A} g_j(s_j)$ over s in the base polyhedron associated to F_A , the restriction of F to A .
5. Find a minimizer $s_{V \setminus A}$ of $\sum_{j \in V \setminus A} g_j(s_j)$ over s in the base polyhedron associated to the contraction F^A of F on A , defined as $F^A(B) = F(A \cup B) - F(A)$.
6. Concatenate s_A and $s_{V \setminus A}$. Exit.

The algorithm must stop after *at most* p iterations. Indeed, if $A \neq V$ in Step 3, then we must have $A \neq \emptyset$ (indeed, $A = \emptyset$ implies that $t \in P(F)$, which in turns implies that $A = V$ because by construction $t(V) = F(V)$, which leads to a contradiction). Thus we actually split V into two non-trivial parts A and $V \setminus A$.

We now need to prove optimality. Let s be the output of the algorithm. We first show that $s \in B(F)$. We have for any $B \subset V$:

$$\begin{aligned} s(B) &= s(B \cap A) + s(B \cap (V \setminus A)) \\ &\leq F(B \cap A) + F(A \cup B) - F(A) \text{ by definition of } s_A \text{ and } s_{V \setminus A} \\ &\leq F(B) \text{ by submodularity.} \end{aligned}$$

Thus s is indeed in the submodular polyhedron $P(F)$. Moreover, we have $s(V) = s_A(A) + s_{V \setminus A}(V \setminus A) = F(A) + F(V) - F(A) = F(V)$, i.e., s is in the base polyhedron $B(F)$.

We now construct a second base $\bar{s} \in B(F)$ as follows: \bar{s}_A is the minimizer of $\sum_{j \in A} g_j(s_j)$ over s in the base polyhedron associated to the submodular polyhedron $P(F_A) \cap \{s_A \leq t_A\}$. From Prop. 19, the associated submodular function is $H_A(B) = \min_{C \subset B} F(C) + t(B \setminus C)$. We have $H_A(A) = \min_{C \subset A} F(C) - t(C) + t(A) = F(A)$ because A is the largest minimizer of $F - t$. Thus, the base polyhedron associated with H_A is simply $B(F_A) \cap \{s_A \leq t_A\}$. Moreover, from Prop. 19, we have that $H_A \leq F_A$, and thus if s_A is tight for F_A then s_A is tight for H_A .

Moreover, we define $\bar{s}_{V \setminus A}$ as the minimizer of $\sum_{j \in V \setminus A} g_j(s_j)$ over the base polyhedron $B(J^A)$ where we define the submodular function J^A on $V \setminus A$ as follows: $J^A(B) = \min_{C \supset B} F(C \cup A) - F(A) - t(C) + t(B)$. Then $J^A - t$ is non-decreasing and submodular (by Proposition 20). Moreover, $J^A(V \setminus A) = F(V) - F(A)$ and $J^A \leq F^A$. Finally $B(F^A) \cap \{s_{V \setminus A} \geq t_{V \setminus A}\} = B(J^A)$ and thus if s_A is tight for F^A then s_A is tight for J^A .

We now show that \bar{s} is optimal for the problem. Since \bar{s} has a higher objective value than s , the base s will then be optimal as well. If we take an exchangeable pair (k, q) for \bar{s} . Then, we have several cases (note that A is tight for \bar{s}):

- $k \in A$, implies $q \in A$ (by Prop. 26, since A is tight), and thus the optimality condition stems from the sub-problem on A (since being tight for F_A implies being tight for H)
- $k \notin A$, $q \in A$, it comes from $\bar{s}_A \leq t_A$ and $\bar{s}_{V \setminus A} \geq t_{V \setminus A}$, which implies $g'_k(\bar{s}_k) \geq g'_q(\bar{s}_q)$ (since all $g'_k(t_k)$ are equal by definition of t).
- $k \notin A$, $q \notin A$, it comes from the optimality of the subproblem on $V \setminus A$, (since being tight for F^A implies being tight for J^A).

In all cases, for exchangeable pairs (k, q) , we have $g'_k(\bar{s}_k) \geq g'_q(\bar{s}_q)$ and thus, by Prop. 29, \bar{s} is optimal and hence s is optimal. Note that we could also have used Prop 30 to show optimality.

Note finally that similar algorithms may be applied when we restrict s to be integers (see, e.g., [29, 6]).

9 Polymatroids (non-increasing submodular functions)

When the submodular function F is also *non-decreasing*, i.e., when for $A, B \subset V$, $A \subset B \Rightarrow F(A) \leq F(B)$, then a truncated greedy algorithm may be applied for all linear functions (i.e., with potentially negative coefficients). Such non-decreasing and submodular functions are often referred to as *polymatroid set-functions* [11] or *β -functions* [30]. Note that in this situation, the Lovász extension is non-decreasing with respect to all components, i.e., if $w \leq w'$, then $f(w) \leq f(w')$.

Proposition 32 (Truncated greedy algorithm) *Assume F is submodular and non-decreasing. Let $w \in \mathbb{R}^p$; a maximizer of $\max_{s \in P(F), s \geq 0} w^\top s$ may be obtained by the following algorithm: order all the strictly positive components of w , as $w_{j_1} \geq \dots \geq w_{j_m} > 0$ and de-*

fine $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$ for $k \leq m$, and zero otherwise. Moreover, $\max_{s \in P(F), s \geq 0} w^\top s = f(w_+)$.

Proof The proof is similar to that of Prop. 5. The constraint $w_k = \sum_{A \ni k} \lambda_A$ is simply replaced by $w_k \leq \sum_{A \ni k} \lambda_A$ (because of the new constraint $s \geq 0$). The vector s is then feasible because of the monotonicity of F . ■

We can also specialize several other results to polymatroids. In this setting, it is easy to see that the base polyhedron $B(F)$ is included in positive orthant \mathbb{R}_+^p (this is for example a consequence of the greedy algorithm from Prop. 5). However, $P(F)$ is not included in the positive orthant, and it is common to consider the positive polyhedron

$$P_+(F) = P(F) \cap \mathbb{R}_+^p = \{s \geq 0, \forall A \subset V, s(A) \leq F(A)\},$$

which is compact (while $P(F)$ is never, as it is unbounded).

We now extend Prop. 10 and Prop. 9 related to support functions, to the independence polyhedron $P_+(F)$, as well as proposition Prop. 12, related to faces of the polyhedron.

Proposition 33 (Maximizers of the support function of independence polyhedron)

Let F be a non-decreasing submodular function such that $F(\emptyset) = 0$. Let $w \in \mathbb{R}^p$, with unique values $v_1 > \dots > v_m$, taken at sets A_1, \dots, A_m . Then s is optimal for $\max_{s \in P_+(F)} w^\top s$ if and only if for all $i = 1, \dots, m$, $v_i < 0 \Rightarrow s_{A_i} = 0$, and $v_i \geq 0 \Rightarrow s(A_1 \cup \dots \cup A_i) = F(A_1 \cup \dots \cup A_i)$.

Proof The proof follows the same arguments than for Prop. 9, with a special treatment for the negative values of w . ■

Proposition 34 (Faces of the independence polyhedron) Let F be a non-decreasing submodular function such that $F(\emptyset) = 0$. Let B be a stable set (i.e., such that all strict larger subsets have strictly greater function values), and $A_1 \cup \dots \cup A_m$ an ordered partition of B , such that for all $j \in \{1, \dots, m\}$, A_j is inseparable for the function $G_j : B \mapsto F(A_1 \cup \dots \cup A_{j-1} \cup B) - F(A_1 \cup \dots \cup A_{j-1})$ defined on subsets of A_j , then the set of $s \in P_+(F)$ such that for all $j \in \{1, \dots, m\}$, $s(A_1 \cup \dots \cup A_j) = F(A_1 \cup \dots \cup A_j)$, and $s_{V \setminus B} = 0$, is a proper face of $P_+(F)$ with non-empty relative interior.

Proof We have a face from Prop. 33, and it has non empty interior by applying Prop. 11 on each submodular function G_j , and using the stability of B . ■

We now show how to minimize a separable convex function on the submodular polyhedron or the positive submodular polyhedron (rather than on the base polyhedron). We first show the following proposition for the submodular polyhedron of any submodular function (non necessarily non-decreasing).

Proposition 35 (Separable optimization on the submodular polyhedron) Assume that F is submodular. Let ψ_j , $j = 1, \dots, p$ be p convex functions such that ψ_j^* is defined and finite on \mathbb{R} . Let (v, t) be a primal-dual optimal pair for the problem

$$\min_{v \in \mathbb{R}^p} \max_{t \in B(F)} t^\top v + \sum_{k \in V} \psi_k(v_k) = \min_{v \in \mathbb{R}^p} f(v) + \sum_{k \in V} \psi_k(v_k) = \max_{t \in B(F)} - \sum_{k \in V} \psi_k^*(-t_k).$$

For $k \in V$, let s_k be a maximizer of $-\psi_k^*(-s_k)$ on $(-\infty, t_k]$. Define $w = v_+$. Then (w, s) is a primal-dual optimal pair for the problem

$$\min_{w \in \mathbb{R}^p} \max_{s \in P(F)} s^\top w + \sum_{k \in V} \psi_k(w_k) = \min_{w \in \mathbb{R}_+^p} f(w) + \sum_{k \in V} \psi_k(w_k) = \max_{s \in P(F)} - \sum_{k \in V} \psi_k^*(-s_k).$$

Proof The pair (w, s) is optimal if and only if $w_k s_k + \psi_k(w_k) + \psi_k^*(-s_k) = 0$, i.e., (w_k, s_k) is a Fenchel-dual pair for ψ_k , and $f(w) = s^\top w$. The first statement is true by construction (indeed, if $s_k = t_k$, then this is a consequence of optimality for the first problem, if $s_k < t_k$, then $w_k = (\psi_k^*)'(-s_k) = 0$).

For the second statement, notice that s is obtained from t by keeping the components of t corresponding to strictly positive values of v (let K denote that subset), and lowering the ones for $V \setminus K$. For $\alpha > 0$, the level sets $\{w \geq \alpha\}$ are equal to $\{v \geq \alpha\} \subset K$. Thus, by Prop. 10, all of these are tight for t and hence for s because these sets are included in K , and $s_K = t_K$. This shows, by Prop. 9, that $s \in P(F)$ is optimal for $\max_{s \in P(F)} w^\top s$. ■

Note that Prop. 35 involves primal-dual pairs (w, s) and (v, t) , but that we can define w from v only, and define s from t only; thus, primal-only views and dual-only views are possible. This also applies to Prop. 36.

Proposition 36 (Separable optimization on the positive submodular polyhedron)

Assume that F is submodular and non-increasing. Let ψ_j , $j = 1, \dots, p$ be p convex functions such that ψ_j^* is defined and finite on \mathbb{R} . Let (v, t) be a primal-dual optimal pair for the problem

$$\min_{v \in \mathbb{R}^p} \max_{t \in B(F)} t^\top v + \sum_{k \in V} \psi_k(v_k) = \min_{v \in \mathbb{R}^p} f(v) + \sum_{k \in V} \psi_k(v_k) = \max_{t \in B(F)} - \sum_{k \in V} \psi_k^*(-t_k).$$

For $k \in V$, let s_k be a maximizer of $-\psi_k^*(-s_k)$ on $[0, t_k]$. For all k , define w_k through $s_k + \psi_k'(w_k) = 0$. Then (w, s) is a primal-dual optimal pair for the problem

$$\min_{w \in \mathbb{R}^p} \max_{s \in P_+(F)} s^\top w + \sum_{k \in V} \psi_k(w_k) = \min_{w \in \mathbb{R}^p} f(w_+) + \sum_{k \in V} \psi_k(w_k) = \max_{s \in P_+(F)} - \sum_{k \in V} \psi_k^*(-s_k).$$

Proof We first apply Prop 35 to the convex functions $\tilde{\psi}_k(w_k) = \min_{v_k \leq w_k} \psi_k(v_k)$, which Fenchel-conjugates equal to $\psi_k^*(s_k)$ if $s_k \leq 0$ and $+\infty$ otherwise. We obtain the minimum over \mathbb{R}_+^p of $f(w) + \sum_{j \in V} \tilde{\psi}_k(w_k)$. Since f non-decreasing with respect to each variable taken separately (because F is non-decreasing), it is equivalent to minimizing on \mathbb{R}^p , $\min_{w \in \mathbb{R}^p} f(w_+) + \sum_{k \in V} \psi_k(w_k)$. ■

10 Examples of submodular functions

We now present classical examples of submodular functions. For each of these, we also describe the corresponding Lovász extensions, and, when appropriate, the associated submodular polyhedra.

10.1 Cardinality-based functions

We consider functions that depend only on $s(A)$ for a certain $s \in \mathbb{R}_+^p$. If $s = 1_V$, these are functions of the cardinality. The next proposition shows that only concave functions lead to submodular functions, and is coherent with the diminishing return property from Section 1 (Prop. 1).

Proposition 37 (Submodularity of cardinality-based set-functions) *If $s \in \mathbb{R}_+^p$ and $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a concave function, then $F : A \mapsto g(s(A))$ is submodular. If $F : A \mapsto g(s(A))$ is submodular for all $s \in \mathbb{R}_+^p$, then g is concave.*

Proof The function $F : A \mapsto g(s(A))$ is submodular if and only if for all $A \subset V$ and $j, k \in V \setminus A$: $g(s(A) + s_k) - g(s(A)) \geq g(s(A) + s_k + s_j) - g(s(A) + s_j)$. If g is concave and $a \geq 0$, $t \mapsto g(a + t) - g(t)$ is non-increasing, hence the first result. Moreover, if $t \mapsto g(a+t) - g(t)$ is non-increasing for all $a \geq 0$, then g is concave, hence the second result. ■

Proposition 38 (Lovász extension of cardinality-based set-functions) *Let $s \in \mathbb{R}_+^p$ and $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a concave function such that $g(0) = 0$, the Lovász extension of the submodular function $F : A \mapsto g(s(A))$ is equal to*

$$f(w) = \sum_{k=1}^p w_{j_k} [g(s_{j_1} + \dots + s_{j_k}) - g(s_{j_1} + \dots + s_{j_{k-1}})].$$

If $s = 1_V$, i.e., $F(A) = g(|A|)$, then $f(w) = \sum_{k=1}^p w_{j_k} [g(k) - g(k-1)]$.

The Lovász extension is thus a function of order statistics.

10.2 Cut functions

Given a set of (non necessarily symmetric) weights $d : V \times V \rightarrow \mathbb{R}_+$, define

$$F(A) = \sum_{k \in A, j \in V \setminus A} d(k, j),$$

which we denote $d(A, V \setminus A)$. Note that for a cut function and disjoint subsets A, B, C , we always have:

$$\begin{aligned} F(A \cup B \cup C) &= F(A \cup B) + F(A \cup C) + F(B \cup C) - F(A) - F(B) - F(C) + F(\emptyset) \\ F(A \cup B) &= d(A \cup B, (A \cup B)^c) = d(A, A^c \cap B^c) + d(B, A^c \cap B^c) \\ &\leq d(A, A^c) + d(B, B^c) = F(A) + F(B), \end{aligned}$$

where we denote $A^c = V \setminus A$. We then have, for any sets $A, B \subset V$:

$$\begin{aligned} F(A \cup B) &= F([A \cap B] \cup [A \setminus B] \cup [B \setminus A]) \\ &= F([A \cap B] \cup [A \setminus B]) + F([A \cap B] \cup [B \setminus A]) + F([A \setminus B] \cup [B \setminus A]) \\ &\quad - F(A \cap B) - F(A \setminus B) - F(B \setminus A) + F(\emptyset) \\ &= F(A) + F(B) + F(A \Delta B) - F(A \cap B) - F(A \setminus B) - F(B \setminus A) \\ &= F(A) + F(B) - F(A \cap B) + [F(A \Delta B) - F(A \setminus B) - F(B \setminus A)] \\ &\leq F(A) + F(B) - F(A \cap B), \end{aligned}$$

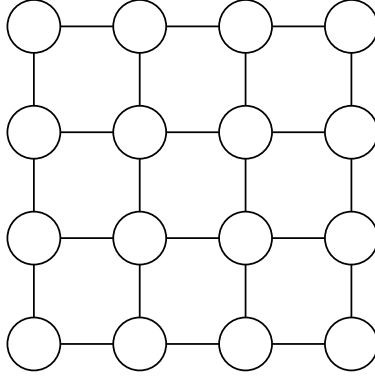


Figure 2: Two-dimensional grid with 4-conenctivity.

which shows submodularity. Moreover, the Lovász extension is equal to

$$f(w) = \sum_{k,j \in V} d(k,j)(w_k - w_j)_+.$$

Then, if the weight function d is symmetric, then the submodular function is also symmetric and the Lovász extension is even (from Prop. 4). Examples of such cuts are shown in Figure 3 (left and middle). A instance of these Lovász extensions plays a crucial role in signal and image processing; indeed, for a graph composed a two-dimensional grid with 4-connectivity (see Figure 2), we obtain the total variation. In fact, some of the results presented in this tutorial were first tackled on this particular case (see, e.g., [16] and references therein).

We can also consider partial minimization to obtain “regular functions” [5]. Examples lead to $f(w) = \max_{k \in G} w_k - \min_{k \in G} w_k$, which corresponds to $F(A) = 1_{A \cap G \neq \emptyset} - 1_{A \cap G = \emptyset}$.

It may also lead to “noisy cuts”, i.e., for a given a weight function $d : V \times V \rightarrow \mathbb{R}_+$, we add p nodes, each of them associated to the original nodes, and consider the convex and submodular functions

$$\begin{aligned} f(w) &= \min_{v \in \mathbb{R}^p} \sum_{k,j \in V} d(k,j)(v_k - v_j)_+ + \lambda \sum_{k \in V} \alpha_k |v_k - w_k|, \\ F(A) &= \min_{B \subset V} \sum_{k \in B, j \in B^c} d(k,j) + \lambda \sum_{k \in V} \alpha_k |1_{k \in A} - 1_{k \in B}|, \end{aligned}$$

which are associated to each other due to Prop. 18. An example of such cut is shown in Figure 3 (right).

This example is particularly interesting, because it leads to a family of submodular functions for which dedicated fast algorithms exist. Indeed, minimizing the cut functions or the partially minimized cut, plus a modular function defined by $z \in \mathbb{R}^p$, may be done with a min-cut/max-flow algorithm (see, e.g., [31]). Indeed, following [5, 16], we add two nodes to the graph, a source s and a sink t . All original edges have non-negative capacities $d(k,j)$, while, the edge that links the source s to the node $k \in V$ has capacity $(z_k)_+$ and the edge that links the node $k \in V$ to the sink t has weight $-(z_k)_-$ (see bottom line of Figure 3). Finding a minimum cut or maximum flow in this graph leads to a minimizer of $F - z$.

For proximal methods, such as defined in Eq. (9) (Section 6), we have $z = \psi(\alpha)$ and we need to solve an instance of a *parametric max-flow* problem, which may be done using efficient

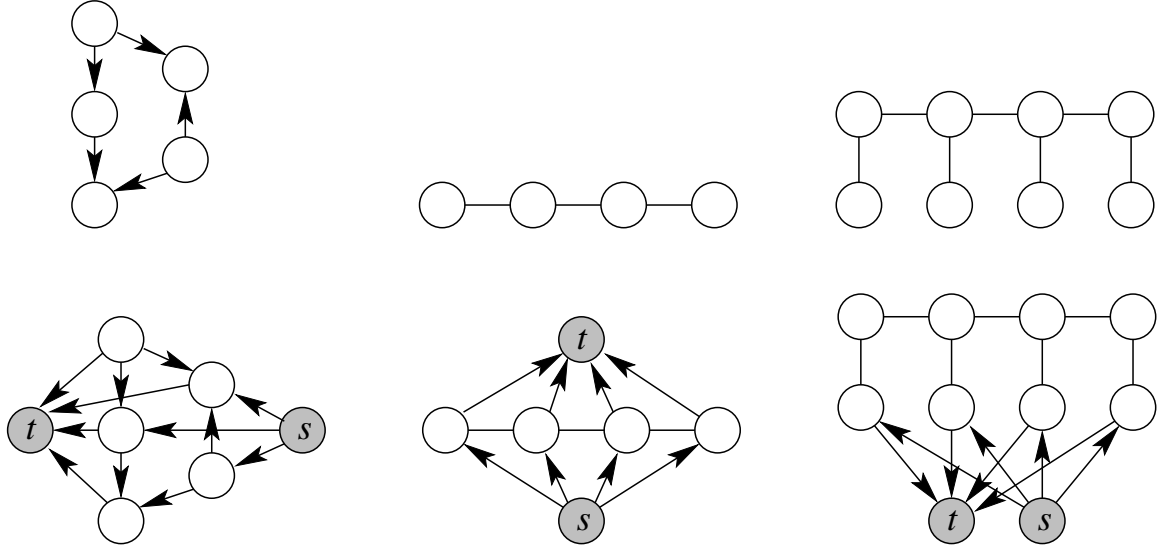


Figure 3: Top: graphs for symmetric (left) and non-symmetric cost functions. Bottom: corresponding networks (note that for the right plot, this corresponds to a partial minimization, we refer to in the text as noisy cuts).

dedicated algorithms [19, 6, 16]. See also Section 8.5 for generic algorithms based on a sequence of singular function minimizations.

10.3 Set covers

Given a *non-negative* function $D : 2^V \rightarrow \mathbb{R}_+$, then we can define

$$F(A) = \sum_{G \subset V, G \cap A \neq \emptyset} \text{Dep}(G),$$

with $f(w) = \sum_{G \subset V} \text{Dep}(G) \max_{k \in G} w_k$. The submodularity and the Lovász extension can be obtained using linearity and the fact that the Lovász extension of $A \mapsto 1_{G \cap A = \emptyset}$ is $w \mapsto \max_{k \in G} w_k$.

Möbius inversion. Note that any set-function F may be written as

$$F(A) = \sum_{G \subset V, G \cap A \neq \emptyset} \text{Dep}(G) = \sum_{G \subset V} \text{Dep}(G) - \sum_{G \subset V \setminus A} \text{Dep}(G),$$

for a certain set-function D , which is not usually non-negative. Indeed, by Möbius inversion formula (see, e.g., [32]), we have:

$$\text{Dep}(G) = \sum_{A \subset G} (-1)^{|G|-|A|} [F(V) - F(A)].$$

Thus, functions for which D is non-negative are a specific subset of submodular functions. Moreover, these functions are always non-decreasing. Such functions are used in the context of sparsity-inducing norms [4, 33, 34].

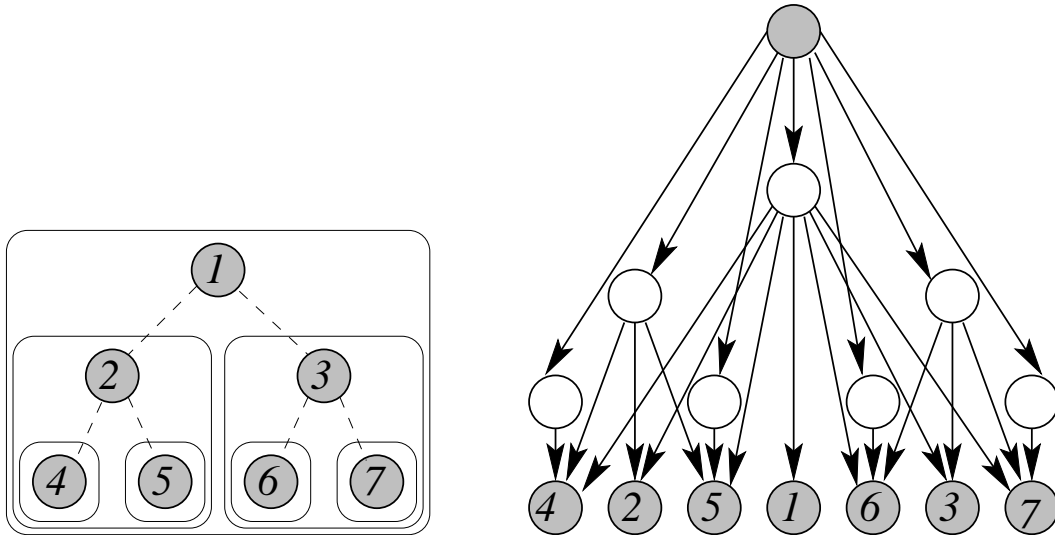


Figure 4: Left: Groups corresponding to a hierarchy. Right: network flow interpretation of same submodular function.

Reinterpretation in terms of set-covers. Let W be any “base” set. Given for each $k \in V$, a set $S_k \subset W$, we define $F(A) = |\bigcup_{k \in A} S_k|$. More generally, we can define $F(A) = \sum_{j \in W} \Delta(j) 1_{\exists k \in A, S_k \ni j}$, if we have weights $\Delta(j) \in \mathbb{R}_+$ for $j \in W$ (this corresponds to replace the cardinality function on W , by a weighted cardinality function, with weights Δ). Then, F is submodular (as a consequence of the equivalence with the previously defined functions, which we now prove).

These two types of functions are in fact equivalent. Indeed, for a weight function $D : 2^V \rightarrow \mathbb{R}_+$, we let $W = 2^V$ and $S_k = \{G \subset V, G \ni k\}$, and $\Delta(G) = \text{Dep}(G)$, to obtain a set cover.

For a certain set cover define by W , $S_k \subset W$, $k \in V$, and δ , define

$$\text{Dep}(G) = \sum_{j \in W} \Delta_j 1_{G_j = \bigcup_{k \in V, S_k \ni j} S_k},$$

to obtain a set-function expressed in terms of groups and non-negative weight functions.

Examples. In Figure 4, we show a set of groups (i.e., only the groups $G \subset V$ for which $\text{Dep}(G) > 0$), which can be embedded into a hierarchy, as well as the corresponding flow interpretation from Section 10.4. We also show in Figure 5 and Figure 6 examples in one dimension.

10.4 Flows

Following [18], we can obtain a family of non-decreasing submodular set-functions (which include set covers) from multi-sink multi-source networks. We define a weight function on a set W , which includes a set S of sources and a set V of sinks (which will be the set on which the submodular function will be defined). We assume that we are given capacities, i.e., a function c from $W \times W$ to \mathbb{R}_+ . For all functions $\varphi : W \times W \rightarrow \mathbb{R}$, we use the notation $\varphi(A, B) = \sum_{k \in A, j \in B} \varphi(k, j)$.

A flow is a function $\varphi : W \times W \rightarrow \mathbb{R}_+$ such that (a) $\varphi \leq c$ for all arcs, (b) for all $w \in W \setminus (S \cup V)$, the net-flow at w , i.e., $\varphi(W, \{w\}) - \varphi(\{w\}, W)$, is null, (c) for all sources $s \in S$, the net-flow at s is non-positive, i.e., $\varphi(W, \{s\}) - \varphi(\{s\}, W) \leq 0$, (d) for all sinks $t \in V$, the net-flow at t is non-negative, i.e., $\varphi(W, \{t\}) - \varphi(\{t\}, W) \geq 0$. We denote by \mathcal{F} the set of flows.

For $A \subset V$ (the set of sinks), we define

$$F(A) = \max_{\varphi \in \mathcal{F}} \varphi(W, A) - \varphi(A, W),$$

which is the maximal net-flow getting out of A . From the max-flow/min-cut theorem (see, e.g., [31]), we have immediately that

$$F(A) = \min_{X \in W, S \subset X, A \subset W \setminus X} c(X, W \setminus X).$$

One then obtain that F is submodular (as the partial minimization of a cut function) and non-decreasing by construction. One particularity is that for this type of submodular non-decreasing functions, we have an explicit description of the positive submodular polyhedron. Indeed, $x \in \mathbb{R}_+^p$ belongs to $P(F)$ if and only if, there exists a flow $\varphi \in \mathcal{F}$ such that for all $k \in V$, $x_k = \varphi(W, \{k\}) - \varphi(\{k\}, W)$ is the net-flow getting out of k .

Similarly to other cut-derived functions, there are dedicated algorithms for proximal methods and submodular minimization [35]. See also [34] for applications to sparsity-inducing norms.

Flow interpretation of set-covers. Following [34], we now show that the submodular functions defined in this section includes the ones defined in Section 10.3. Indeed, consider a non-negative function $D : 2^V \rightarrow \mathbb{R}_+$, and define $F(A) = \sum_{G \subset V, G \cap A \neq \emptyset} \text{Dep}(G)$. The Lovász extension may be written as, for all $w \in \mathbb{R}_+^p$,

$$\begin{aligned} f(w) &= \sum_{G \subset V} \text{Dep}(G) \max_{k \in G} w_k \\ &= \sum_{G \subset V} \max_{t^G \in \mathbb{R}_+^p, t_{V \setminus G}^G = 0, t^G(G) = \text{Dep}(G)} w^\top t^G \\ &= \max_{t^G \in \mathbb{R}_+^p, t_{V \setminus G}^G = 0, t^G(G) = \text{Dep}(G)} \sum_{G \subset V} w^\top t^G \\ &= \max_{t^G \in \mathbb{R}_+^p, t_{V \setminus G}^G = 0, t^G(G) = \text{Dep}(G)} \sum_{k \in V} \left(\sum_{G \subset V} t_k^G \right) w_k. \end{aligned}$$

Thus $s \in P(F)$, if and only there exists $t^G \in \mathbb{R}_+^p$, $t_{V \setminus G}^G = 0$, $t^G(G) = \text{Dep}(G)$ for all $G \subset V$, such that $s = \sum_{G \subset V} t^G$. This can be given a network flow interpretation on the graph composed of a single source s , one node per subset $G \subset V$ such that $\text{Dep}(G) > 0$, and the sink set V . The source is connected to all subsets G , with capacity $\text{Dep}(G)$, and each subset is connected to the variables it contains, with infinite capacity. We give examples of such networks in Figure 5 and Figure 6.

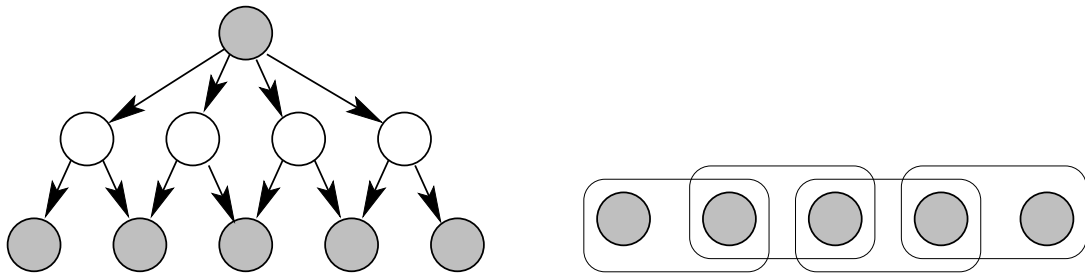


Figure 5: Flow (left) and set of groups (right).

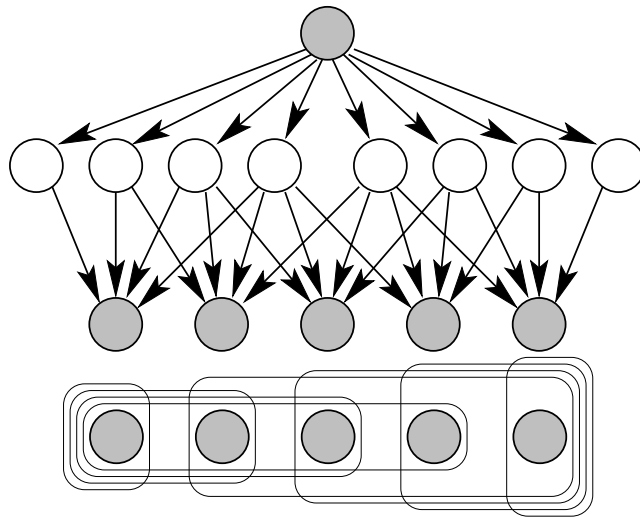


Figure 6: Flow (top) and set of groups (bottom).

10.5 Entropies

Given p random variables X_1, \dots, X_p which all take a finite number of values, we define $F(A)$ as the joint entropy of the variables $(X_k)_{k \in A}$. This function is submodular because, if $A \subset B$ and $k \notin B$, $F(A \cup \{k\}) - F(A) = H(X_A, X_k) - H(X_A) = H(X_k | X_A) \geq H(X_k | X_B) = F(B \cup \{k\}) - F(B)$ (by the data processing inequality [36]).

This can be extended to any distribution by considering differential entropies. One application is for Gaussian random variables, leading to the submodularity of the function defined through $F(A) = \log \det Q_{AA}$, for some positive definite matrix $Q \in \mathbb{R}^{p \times p}$ (see further related examples in Section 10.6).

10.6 Spectral functions of submatrices

Given a positive semidefinite matrix $Q \in \mathbb{R}^{p \times p}$ and a real-valued function h from $\mathbb{R}_+ \rightarrow \mathbb{R}$, one may define $\text{tr}[h(Q)]$ as $\sum_{i=1}^p h(\lambda_i)$ where $\lambda_1, \dots, \lambda_p$ are the (nonnegative) eigenvalues of Q [37]. We can thus define the function $F(A) = \text{tr} h(Q_{AA})$ for $A \subset V$.

The concavity of h is not sufficient for submodularity (as can be seen by generating random examples with $h(\lambda) = \lambda/(\lambda + 1)$).

We know however that the functions $h(\lambda) = \log(\lambda + t)$ for $t \geq 0$ lead to submodular functions; thus, since for $p \in (0, 1)$, $\lambda^p = \frac{p \sin p\pi}{\pi} \int_0^\infty \log(1 + \lambda/t) t^{p-1} dt$ (see, e.g., [38]), $h(\lambda) = \lambda^p$ for $p \in (0, 1]$ are positive linear combinations of functions that lead to non-decreasing submodular set-functions. We thus obtain a non-decreasing submodular function. Applications may be found in [4].

This can be generalized to functions of the singular values of $X(A, B)$ where X is a rectangular matrix, by considering the fact that singular values of a matrix X are related to the eigenvalues of $\begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix}$ (see, e.g., [39]).

10.7 Best subset selection

Following [40], we consider p random variables (covariates) X_1, \dots, X_p , and a random response Y with unit variance, i.e., $\text{var}(Y) = 1$. We consider predicting Y linearly from X . We consider $F(A) = \text{var}(Y | X_A)$. The function F is a non-increasing function.

A variable X_j is a suppressor for variable X_i , if $|\text{Corr}(Y, X_i | X_j)| > |\text{Corr}(Y, X_i)|$. Following [40], we assume that there are no suppressor variables given any set A , i.e., we assume that for all $A \subset V$, $i, j \notin A$,

$$|\text{Corr}(Y, X_i | X_j, X_A)| \leq |\text{Corr}(Y, X_i | X_A)|,$$

We then have:

$$\begin{aligned} \text{var}(Y | X_A, X_k) - \text{var}(Y | X_A) &= -\text{Corr}(Y, X_k | X_A)^2, \\ \text{var}(Y | X_A, X_j, X_k) - \text{var}(Y | X_A, X_j) &= -\text{Corr}(Y, X_k | X_A, X_j)^2. \end{aligned}$$

This implies that F is supermodular. Note however that the condition on suppressors is rather strong.

10.8 Matroids

Given a set V , we consider a family \mathcal{I} of subsets of V such that (a) $\emptyset \in \mathcal{I}$, (b) $I_1 \subset I_2 \in \mathcal{I} \Rightarrow I_1 \in \mathcal{I}$, and (c) for all $I_1, I_2 \in \mathcal{I}$, $|I_1| < |I_2| \Rightarrow \exists k \in I_2 \setminus I_1, I_1 \cup \{k\} \in \mathcal{I}$. The pair (V, \mathcal{I}) is then referred to as a matroid, with \mathcal{I} its family of independent sets. Then the rank function of the matroid, defined as $\rho(A) = \max_{I \subset A, I \in \mathcal{I}} |I|$, is submodular.

The classical example is the *graphic matroid*; it corresponds to V being an edge set of a certain graph, and \mathcal{I} being the set of subsets of edges which do not contain any cycle. The rank function $\rho(A)$ is then equal to p minus the number of connected components of the subgraph induced by A .

The other one is the *linear matroid*. Given a matrix M with p columns, then a set I is independent if and only if the set of columns indexed by I is independent. The rank function $\rho(A)$ is then the rank of the columns indexed by A (this is also an instance of functions from Section 10.6).

Acknowledgements

This tutorial was partially supported by grants from the Agence Nationale de la Recherche (MGA Project) and from the European Research Council (SIERRA Project). The author would like to thank Rodolphe Jenatton, Armand Joulin, Julien Mairal and Guillaume Obozinski for discussions related to submodular functions.

References

- [1] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.
- [2] Y. Kawahara, K. Nagano, K. Tsuda, and J.A. Bilmes. Submodularity cuts and applications. In *Adv. NIPS 22*, 2009.
- [3] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proc. ICML*, 2010.
- [4] F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, 2010.
- [5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.
- [6] D.S. Hochbaum. An efficient algorithm for image segmentation, Markov random fields and related problems. *Journal of the ACM (JACM)*, 48(4):686–701, 2001.
- [7] M. Queyranne and A. Schulz. Scheduling unit jobs with compatible release dates on parallel machines with nonstationary speeds. *Integer Programming and Combinatorial Optimization*, 920:307–320, 1995.
- [8] H. Narayanan. *Submodular Functions and Electrical Networks*. North-Holland, 2009. Second edition.

- [9] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [10] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.
- [11] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.
- [12] A. Toshev. Submodular function minimization. Technical report, University of Pennsylvania, 2010. Written Preliminary Examination.
- [13] A. Krause and C. Guestrin. Beyond convexity: Submodularity in machine learning, 2008. Tutorial at ICML.
- [14] L. Lovász. Submodular functions and convexity. *Mathematical programming: the state of the art, Bonn*, pages 235–257, 1982.
- [15] G. Choquet. Theory of capacities. *Ann. Inst. Fourier*, 5:131–295, 1954.
- [16] A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International journal of computer vision*, 84(3):288–307, 2009.
- [17] P.L. Combettes and J.C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing. New York: Springer-Verlag, 2010.
- [18] N. Megiddo. Optimal flows in networks with multiple sources and sinks. *Mathematical Programming*, 7(1):97–107, 1974.
- [19] G. Gallo, M.D. Grigoriadis, and R.E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.
- [20] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions–i. *Mathematical Programming*, 14(1):265–294, 1978.
- [21] P. Wolfe. Finding the nearest point in a polytope. *Math. Progr.*, 11(1):128–149, 1976.
- [22] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.
- [23] S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.
- [24] J.B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.
- [25] M. Queyranne. Minimizing symmetric submodular functions. *Mathematical Programming*, 82(1):3–12, 1998.
- [26] H. Nagamochi and T. Ibaraki. A note on minimizing submodular functions. *Information Processing Letters*, 67(5):239–244, 1998.

- [27] K. Nagano. A strongly polynomial algorithm for line search in submodular polyhedra. *Discrete Optimization*, 4(3-4):349–359, 2007.
- [28] S. Fujishige. Lexicographically optimal base of a polymatroid with respect to a weight vector. *Mathematics of Operations Research*, 5(2):186–196, 1980.
- [29] H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.
- [30] J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer, 2003.
- [31] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1989.
- [32] S. Foldes and P. L. Hammer. Submodularity, supermodularity, and higher-order monotonicities of pseudo-Boolean functions. *Mathematics of Operations Research*, 30(2):453–461, 2005.
- [33] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proc. ICML*, 2010.
- [34] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.
- [35] D.S. Hochbaum and S.P. Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical Programming*, 69(1):269–309, 1995.
- [36] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- [37] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge Univ. Press, 1990.
- [38] T. Ando. Concavity of certain maps on positive definite matrices and applications to hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.
- [39] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
- [40] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the 40th annual ACM symposium on Theory of computing*. ACM, 2008.