Submodular Analysis, Duality and Optimization

Yao-Liang Yu
yaoliang@cs.ualberta.ca
Dept. of Computing Science
University of Alberta

December 14, 2015

This note is intended to present some fundamental results about submodular functions and their applications in discrete optimization (with special interest in machine learning applications). Most results are taken from various sources, with some occasional improvements.

The algorithm section is expected to go through a large update soon.

Page 38 – 43 are not available currently, due to conflict with some ongoing work.

Contents

1 Distributive Lattice

Alert 1.1: Skipping this section

If one is willing to restrict himself to the full domain 2^Ω, the power set of Ω, then this section can be skipped without much harm. Nevertheless, it is recommended to read at least Theorem 1.1 so that one understands how to deal with general (distributive) lattices.

Let $(\mathfrak{L}, \leq)$ be a partially ordered set. For any pair $x, y \in \mathfrak{L}$, its least upper bound (or the *join* operator w.r.t. the order $\leq$), if exists, is denoted as $x \vee y$, and its greatest lower bound (the *meet* operator), if exists, is similarly denoted as $x \wedge y$. When all pairs have least upper bound (supremum) and greatest lower bound (infimum), we call $(\mathfrak{L}, \leq)$ a lattice—the central domain for us. We will focus on distributive lattices—those enjoy the distributive law: for all $x, y, z \in \mathfrak{L}$,

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$$
$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z).$$

For the product set of ordered sets, we equip it with the natural pointwise order. It is a (distributive) lattice if the factors are.

Example 1.1: Not all lattices are distributive

Let $\mathfrak{L} = \{\bot, \top, x, y, z\}$ where $\top$ is largest, $\bot$ is smallest, and $\{x, y, z\}$ are not directly comparable. Then $x \vee (y \wedge z) = x \vee \bot = x$ while $(x \vee y) \wedge (x \vee z) = \top \wedge \top = \top$.

Remark 1.1: Lattice operators characterize the order

It is clear that the lattice operators $\wedge$ and $\vee$ are completely determined by the underlying order, with the following properties:

- Idempotent: $\forall x \in \mathfrak{L}, x \wedge x = x, x \vee x = x$;
- Symmetric: $\forall x, y \in \mathfrak{L}, x \wedge y = y \wedge x, x \vee y = y \vee x$;
- Absorptive: $\forall x, y \in \mathfrak{L}, (x \wedge y) \vee x = x, (x \vee y) \wedge x = x$;
- Associative: $\forall x, y, z \in \mathfrak{L}, (x \wedge y) \wedge z = x \wedge (y \wedge z), (x \vee y) \vee z = x \vee (y \vee z)$.

On the other hand, given two operators $\wedge$ and $\vee$ on some set $\mathfrak{L}$ with the above four properties, we can define an order on $\mathfrak{L}$: $x \leq y \iff x \wedge y = x$ (or $x \leq y \iff x \vee y = y$), and the lattice operators associated with the defined order are exactly the $\wedge$ and $\vee$ that we begin with.

Perhaps the most important example for a distributive lattice is the power set, denoted as 2^Ω, of a nonempty ground set Ω, ordered by the set inclusion. A bit surprisingly, the converse is also true. Recall that two lattices $\mathfrak{L}_1$ and $\mathfrak{L}_2$ are isomorphic if there exists some bijective function $f : \mathfrak{L}_1 \to \mathfrak{L}_2$ such that $f(x \wedge y) = f(x) \wedge f(y)$ and $f(x \vee y) = f(x) \vee f(y)$. We say $\mathfrak{L}' \subseteq \mathfrak{L}$ a sublattice if $\mathfrak{L}'$ is itself a lattice with the inherited lattice operators. Note that it is possible for $\mathfrak{L}'$ to have different lattice operators than $\mathfrak{L}$, in which case we call $\mathfrak{L}'$ a lattice subspace.

Example 1.2: Not all lattice subspaces are sublattices

Let $\mathfrak{L} = C([0, 1])$ be the set of continuous functions on the interval $[0, 1]$, equipped with the pointwise order. It is clearly a lattice. Take $\mathfrak{L}'$ to be all affine functions. Again $\mathfrak{L}'$ is a lattice, but its own lattice operators are different from those of $\mathfrak{L}$.

Theorem 1.1: [Birkhoff, 1948, p. 140]

Any distributive lattice is isomorphic to a sublattice of 2^Ω for some ground set Ω.

Proof: Proof will be added. ■

More definitions. We call a lattice $\mathfrak{L}$ *bounded* if it has a largest element $\top$ and a smallest element $\bot$; *complete* if any subset has infimum and supremum; *complemented* if for any $X \subseteq \mathfrak{L}$ there exists $Y \subseteq \mathfrak{L}$ such that $X \vee Y = \top, X \wedge Y = \bot$; finite if the cardinality $|\mathfrak{L}| < \infty$. Moreover, the product set is bounded, complete, complemented if the factors are, and finite if we only have finitely many factors which themselves are finite. A bounded, complemented, distributive lattice is called a Boolean algebra.

Proposition 1.1: Finite lattices are nice

Any finite lattice is bounded and complete.

Remark 1.2: A special ordered set

We consider decomposing a complete sublattice $\mathfrak{L} \subseteq 2^\Omega$ with $\emptyset, \Omega \in \mathfrak{L}$. Let $\mathcal{I}_{[x]} := \bigcap_{x \in X \in \mathfrak{L}} X$ be the smallest element in $\mathfrak{L}$ that contains $x \in \Omega$. $\mathcal{I}_{[x]}$ is well defined since $\Omega \in \mathfrak{L}$ and $\mathfrak{L}$ is complete. Clearly, $y \in \mathcal{I}_{[x]} \iff \mathcal{I}_{[y]} \subseteq \mathcal{I}_{[x]}$. Define the equivalence class $[x] = \{y : \mathcal{I}_{[y]} = \mathcal{I}_{[x]}\}$. Then $\mathcal{P} = \{[x] : x \in \Omega\}$ is a partition of Ω, and we order its elements by $[x] \preceq [y] \iff \mathcal{I}_{[x]} \subseteq \mathcal{I}_{[y]}$. Note that the cardinality of $\mathcal{P}$ may be strictly smaller than that of Ω. The resulting ordered set $(\mathcal{P}, \preceq)$ will freely appear many times in our later development.

Alert 1.2: Notation

Following set theory, when $\mathcal{I} = \{A_j : j \in J\}$ is a set of sets, we use $\bigcup \mathcal{I}$ as a shorthand for $\bigcup_{j \in J} A_j$.

Recall that for an ordered set $(\mathbb{O}, \leq)$, $\mathcal{I} \subseteq \mathbb{O}$ is called an (lower) ideal if $x \preceq y \in \mathcal{I} \implies x \in \mathcal{I}$, i.e., an ideal contains all of its dominated elements. Similarly $\mathcal{I}$ is called an upper ideal if $\mathcal{I} \ni x \preceq y \implies y \in \mathcal{I}$. We verify that $\mathcal{I} \subseteq \mathbb{O}$ is a lower ideal iff $\mathbb{O} \setminus \mathcal{I}$ is an upper ideal. Besides, all ideals themselves, under the inclusion order, form a lattice whose join and meet operators are simply the set union and intersection, respectively. An ideal in the form of $\{x \in \mathbb{O} : x \preceq y\}$ for some $y \in \mathbb{O}$ is called the principal ideal and denoted as $\mathcal{I}_y$. The collection of all principal ideals form a lattice subspace (not necessarily sublattice!) of the set of all ideals, in fact, it is isomorphic to the original ordered set $\mathbb{O}$ under the identification $x \mapsto \mathcal{I}_x$. Every ideal in a finite ordered set is a union of principal ideals, and an ideal remains to be an ideal after removing any of its maximal elements (which always exists for a finite set).

Theorem 1.2: Distributive lattices correspond to ideals

Let $\mathfrak{L} \subseteq 2^\Omega$ be a complete distributive sublattice that contains $\emptyset, \Omega$. Consider the ordered set $(\mathcal{P}, \preceq)$ constructed in Remark 1.2. Then for each ideal $\mathcal{I} \subseteq \mathcal{P}$, $\bigcup \mathcal{I} \in \mathfrak{L}$. Conversely, for any $X \in \mathfrak{L}$, $\mathcal{I} := \{I \in \mathcal{P} : I \subseteq X\}$ is an ideal of $\mathcal{P}$ that forms a partition of X.

Proof: Suppose $\mathcal{I}$ is an ideal in $\mathcal{P}$. Fix $x \in X := \bigcup \mathcal{I}$. Thus $[x] \in \mathcal{I}$. For each $y \in \mathcal{I}_{[x]}$, $\mathcal{I}_{[y]} \subseteq \mathcal{I}_{[x]} \implies [y] \preceq [x] \implies [y] \in \mathcal{I}$ since $\mathcal{I}$ is an ideal in $(\mathcal{P}, \preceq)$. Therefore $y \in X$ and consequently $X = \bigcup_{x \in X} \mathcal{I}_{[x]} \in \mathfrak{L}$ due to the completeness of $\mathfrak{L}$.

Conversely, let $X \in \mathfrak{L}$. Then for any $I \in \mathcal{P}$, either $I \subseteq X$ or $I \cap X = \emptyset$. Since $\bigcup \mathcal{P} = \Omega$, $\mathcal{I} := \{I \in \mathcal{P} : I \subseteq X\}$ forms a partition of X. Clearly, if $\mathcal{P} \ni J \preceq I \in \mathcal{I}$, then $J \subseteq X$ hence $J \in \mathcal{I}$, meaning that $\mathcal{I}$ thus defined is indeed an ideal of $(\mathcal{P}, \preceq)$. ■

In other words, any complete distributive lattice $(\mathfrak{L}, \leq)$ consists of merely the union of each ideal of a potentially different set $\mathcal{P}$ equipped with a potentially different order $\preceq$.

Theorem 1.3: Maximal increasing sequence determines the partition

Let $\mathfrak{L} \subseteq 2^\Omega$ be a finite distributive sublattice that contains $\emptyset, \Omega$. Let

$$\emptyset = S_0 \subset S_1 \subset \cdots \subset S_k = \Omega$$

be any *maximal* increasing sequence in $\mathfrak{L}$. Then

$$\mathcal{P} = \{S_i \setminus S_{i-1}, i = 1, \dots, k\}. \tag{1}$$

In particular, all maximal increasing sequences are equally long.

Proof: According to Theorem 1.2, each S_i is the union of an ideal in $(\mathcal{P}, \preceq)$, i.e., $S_i = \bigcup \mathcal{I}_i, \mathcal{I}_i = \{I_1, \dots, I_{j_i}\} \subseteq 2^{\mathcal{P}}$. Let I be a maximal element in $\mathcal{I}_i \setminus \mathcal{I}_{i-1}$ (w.r.t. the order $\preceq$). Clearly I is also maximal in $\mathcal{I}_i$ since $S_i \supset S_{i-1}$. Let $\mathcal{I} = \mathcal{I}_i \setminus \{I\}$. By the maximality of I, $\mathcal{I}$ is again an ideal and $\mathcal{I} \supseteq \mathcal{I}_{i-1}$. Thus by the maximality of $\{S_i\}$ we must have $\mathcal{I} = \mathcal{I}_{i-1}$, i.e., $I = S_i \setminus S_{i-1} \in \mathcal{P}$. This proves that $\{S_i \setminus S_{i-1}, i = 1, \dots, k\} \subseteq \mathcal{P}$. Since $S_0 = \emptyset, S_k = \Omega$ and $|S_i \setminus S_{i-1}| = 1$ we must have $k = |\mathcal{P}|$, i.e., the equality in (1). ■

From the proof it is clear that for any $j > i$ we cannot have $S_j \setminus S_{j-1} \preceq S_i \setminus S_{i-1}$.

Therefore we can deduce the set $\mathcal{P}$ from any maximal increasing sequence in $\mathfrak{L}$. For the extreme case where $\mathcal{P} = \Omega$, we say the lattice $\mathfrak{L}$ is simple. Pleasantly, the simple lattice $(\mathfrak{L} \subseteq 2^{\Omega}, \leq)$ is just the collection of all ideals of the ordered set $(\Omega, \preceq)$. On the other hand, the collection of all ideals of a finite set $(\Omega, \preceq)$, equipped with the set inclusion order, is a simple lattice: Successively adding minimal elements one by one in the remaining ground set we arrive at a maximal increasing sequence. Besides, the lattice $\mathfrak{L}$ (simple or not) is a Boolean algebra iff the order $\preceq$ is trivial, i.e., no two elements are comparable.

Remark 1.3: "Simplification"

Let $F : \mathfrak{L} \to \mathbb{R}$ be a function defined on the finite distributive lattice $\mathfrak{L}$. If $\mathfrak{L}$ is *not* simple, then we construct the simple lattice $\tilde{\mathfrak{L}}$ that is consisted of all ideals (without taking union) of the ordered set $(\mathcal{P}, \preceq)$. Define $\tilde{F} : \tilde{\mathfrak{L}} \to \mathbb{R}$ by $\tilde{F}(\mathcal{I}) = F(\bigcup \mathcal{I})$. Conveniently, many good properties of F, such as monotonicity or submodularity defined below, transfer to $\tilde{F}$. Note that the ground set for $\tilde{\mathfrak{L}}$ is $\mathcal{P}$, potentially a collection of subsets of Ω. By simplification, a Boolean algebra can be taken simple.

There is a beautiful theory on vector lattices (those compatible with a linear structure), Banach lattices (also compatible with a norm), and Boolean algebras, but we shall not need any of such knowledge.

2 Submodular Functions

Thanks to Theorem 1.1 and Proposition 1.1, we know any finite distributive lattice is a complete sublattice of 2^{Ω} for some ground set Ω, with $\emptyset, \Omega \in \mathfrak{L}$. Thus from now on, the following convention should be kept in mind.

Alert 2.1: Notation

Our domain $\mathfrak{L}$ is a sublattice of 2^{Ω} for some nonempty finite ground set Ω, with always $\emptyset, \Omega \in \mathfrak{L}$. We use $\mathbf{1}_X$ to denote the characteristic function of the set X, i.e., $\mathbf{1}_X(x) = 1$ if $x \in X$, otherwise $\mathbf{1}_X(x) = 0$. The shorthand $\mathbf{1} := \mathbf{1}_{\Omega}$ (all ones) is also adopted whenever the ground set Ω is clear from context. All empty sums, unless stated otherwise, are understood to take value 0. As usual, $\mathbb{R}$ denotes the real line (although many results extend immediately to any totally ordered vector space) and $\{\mathbf{e}_i\}_{i \in \Omega}$ denotes the canonical basis in $\mathbb{R}^{\Omega}$.

We start with the definition of submodularity.

Definition 2.1: Submodular function

The set function $F : \mathfrak{L} \to \mathbb{R}$ is called submodular if

$$\forall X \in \mathfrak{L},\ \forall Y \in \mathfrak{L},\quad F(X \cup Y) + F(X \cap Y) \leq F(X) + F(Y). \tag{2}$$

Similarly, we can define supermodular functions by reversing the inequality in (2). Clearly, F is supermodular iff $-F$ is submodular, which allows us to focus exclusively on submodularity.

Alert 2.2: Domain

To verify submodularity, one must first check whether or not the domain is a lattice.

Proposition 2.1: Minimizers constitute a sublattice

For a submodular function $F : \mathfrak{L} \to \mathbb{R}$, its minimizing set $\left\{ Y \in \mathfrak{L} : F(Y) = \min_{X \in \mathfrak{L}} F(X) \right\}$ is a sublattice of the domain $\mathfrak{L}$. ■

Let us illustrate the ubiquity of submodularity through a few examples, each of which can be verified directly from Definition 2.1.

Example 2.1: Maximal element is submodular

Consider a weight function $\mathbf{w} : \Omega \to \mathbb{R}$ and define for any $A \subseteq \Omega$

$$F(A) = \begin{cases} \max_{a \in A} w_a, & A \neq \emptyset \\ c, & A = \emptyset \end{cases} . \tag{3}$$

As long as $c \leq \min_{a \in \Omega} w_a$, F is easily verified to be submodular.

Example 2.2: Entropy is (increasing) submodular

Consider a collection of random variables $X_i, i \in \Omega := \{1, \ldots, n\}$. For any nonempty subset $A \subseteq \Omega$, let us denote X_A as the set $\{X_i : i \in A\}$ and define

$$E(A) := \mathsf{H}(X_A) := \mathsf{E}(-\log \mathsf{p}(X_A)), \quad E(\emptyset) \leq 0. \tag{4}$$

The submodularity of E follows from the nonnegativity of the conditional mutual information:

$$\begin{aligned} E(A) + E(B) - E(A \cup B) - E(A \cap B) &= \mathsf{E}\left(\log \frac{\mathsf{p}(X_{A \cup B}) \mathsf{p}(X_{A \cap B})}{\mathsf{p}(X_A) \mathsf{p}(X_B)} \right) \\ &= \mathsf{E}\left(\log \frac{\mathsf{p}(X_A, X_B \mid X_{A \cap B})}{\mathsf{p}(X_A \mid X_{A \cap B}) \mathsf{p}(X_B \mid X_{A \cap B})} \right) \\ &= \mathsf{I}(X_A; X_B \mid X_{A \cap B}) \geq 0. \end{aligned}$$

Note also that E is increasing since the conditional entropy is nonnegative as well.

Example 2.3: Graph cut is (strictly) submodular

Let $\mathcal{G} = (V, c)$ be a complete directed graph, with vertices V, and the capacity function $c : V \times V \to \mathbb{R}_+ \cup \{\infty\}$. The graph cut for any subset $X \subseteq V$ is defined as

$$\mathsf{Cut}(X) = \sum_{u \in X} \sum_{v \in V \setminus X} c(u, v). \tag{5}$$

Note that $\mathsf{Cut}(X) = \mathsf{Cut}(V - X)$ if $c(u, v) = c(v, u)$ for all $u, v \in V$. Simple algebra yields

$$\begin{aligned} \mathsf{Cut}(X) &= \sum_{X \cap Y} \sum_{Y \setminus X} + \sum_{X \cap Y} \sum_{V \setminus (X \cup Y)} + \sum_{X \setminus Y} \sum_{V \setminus X} \\ \mathsf{Cut}(Y) &= \sum_{Y \cap X} \sum_{X \setminus Y} + \sum_{Y \cap X} \sum_{V \setminus (Y \cup X)} + \sum_{Y \setminus X} \sum_{V \setminus Y} \\ \mathsf{Cut}(X \cap Y) &= \sum_{X \cap Y} \sum_{Y \setminus X} + \sum_{X \cap Y} \sum_{V \setminus (X \cup Y)} + \sum_{X \cap Y} \sum_{X \setminus Y} \end{aligned}$$

$$\mathsf{Cut}(X \cup Y) = \sum_{X \cup Y} \sum_{V \setminus (X \cup Y)}$$

therefore

$$\begin{aligned}\mathsf{Cut}(X) + \mathsf{Cut}(Y) - \mathsf{Cut}(X \cap Y) - \mathsf{Cut}(X \cup Y) &= \sum_{X \setminus Y} \sum_{V \setminus X} + \sum_{X \cap Y} \sum_{V \setminus (X \cup Y)} + \sum_{Y \setminus X} \sum_{V \setminus Y} - \sum_{X \cup Y} \sum_{V \setminus (X \cup Y)} \\ &\geq \sum_{X \setminus Y} \sum_{V \setminus (X \cup Y)} + \sum_{X \cap Y} \sum_{V \setminus (X \cup Y)} + \sum_{Y \setminus X} \sum_{V \setminus (X \cup Y)} - \sum_{X \cup Y} \sum_{V \setminus (X \cup Y)} \\ &= 0,\end{aligned}$$

meaning that the cut function, when restricted to the lattice $\{X : \mathsf{Cut}(X) < \infty\}$, is submodular. Observe that the nonnegativity of the capacity is needed in deriving the inequality.

If the weights are bounded away from 0, we have

$$\min_{X \not\subseteq Y, Y \not\subseteq X} \mathsf{Cut}(X) + \mathsf{Cut}(Y) - \mathsf{Cut}(X \cap Y) - \mathsf{Cut}(X \cup Y) \geq 2 \cdot \min_{u,v} c(u,v) > 0.$$

Such functions will be called *strictly* submodular.

Remark 2.1: Hardness of submodular maximization

Example 2.3 immediately implies that there exists some constant $0 \leq \alpha < 1$ up to which submodular maximization cannot be approximated, modulus certain complexity conjectures. On the other hand, as we will discuss later on, the tractability of submodular minimization implies, in particular, the tractability of min-cut.

Definition 2.2: Modular function

The function $M : \mathfrak{L} \to \mathbb{R}$ is called modular if $\forall X \in \mathfrak{L}, \forall Y \in \mathfrak{L}$,

$$M(X \cup Y) + M(X \cap Y) = M(X) + M(Y). \tag{6}$$

Clearly, M is modular iff it is both submodular and supermodular. If $\mathfrak{L} = 2^\Omega$, inducting from (6) we obtain

$$M(X) = M(\emptyset) + \sum_{\omega \in X} \Big(M(\{\omega\}) - M(\emptyset) \Big). \tag{7}$$

In other words, the modular function is determined by its values on singletons (and the empty set). On the other hand, one easily verifies that (7) indeed defines a modular function on 2^Ω. Moreover, assuming $M(\emptyset) = 0$ and denote $\mathbf{q}(X)$ as $\sum_{\omega \in X} q_\omega$ for any $\mathbf{q} \in \mathbb{R}^\Omega$, we can identify any modular function M with a vector $\mathbf{q} \in \mathbb{R}^\Omega$, in a way that $M(\{\omega\}) = \mathbf{q}(\{\omega\})$, and consequently $M(X) = \mathbf{q}(X)$. Later on we will see that this conclusion generalizes to any simple lattice, and modular functions are exactly those corresponding to linear functions.

Example 2.4: Cardinality is modular

The cardinality of a set $X \subseteq \Omega$ is simply the number of elements in X. It is elementary to verify that indeed the cardinality function is modular. We can extend the definition of cardinality to any real vector $x \in \mathbb{R}^\Omega$:

$$|x| := |\mathrm{Supp}(x)|, \tag{8}$$

where

$$\mathrm{Supp}(x) := \{i \in \Omega : x_i \neq 0\} \tag{9}$$

is the support of x. In essence, the cardinality of a real vector is the number of its nonzero components. If we treat the vector space $\mathbb{R}^\Omega$ as a lattice equipped with the pointwise order, then the cardinality

function defined in (8) is indeed modular.

Example 2.5: log det **is submodular**

Let $\Omega = \{1, \ldots, n\}$ and fix $X \in \mathbb{S}^n_{++}$. Denote X_A as the submatrix $[X]_{i,j}$ where $(i,j) \in A \times A$. Then

$$F(A) := \log\det(X_A) \tag{10}$$

is submodular. Indeed, the (differential) entropy of a multivariate Gaussian vector $\mathbf{x}$ with covariance matrix $X \in \mathbb{S}^n_{++}$ is known as

$$\mathsf{H}_g(X) = \frac{1}{2}\log[(2\pi)^n \det(X)] = \frac{1}{2}\log(2\pi)\cdot n + \frac{1}{2}\log\det X,$$

hence $F(A) = 2\mathsf{H}_g(X_A) - \log(2\pi)\cdot|A|$, but Example 2.2 showed that entropy is submodular while Example 2.4 showed that cardinality is modular.

The following proposition will be convenient in determining submodularity.

Proposition 2.2: Determining submodularity

Let $\mathfrak{L} \subseteq 2^\Omega$ be a *simple* sublattice. The following are equivalent:

(I). $F : \mathfrak{L} \to \mathbb{R}$ is submodular;

(II). $\forall \mathfrak{L} \ni X \subseteq Y \in \mathfrak{L}, \forall S \subseteq \Omega \setminus Y$ such that $X \cup S \in \mathfrak{L}, Y \cup S \in \mathfrak{L}$, we have $F(X \cup S) - F(X) \geq F(Y \cup S) - F(Y)$;

(III). $\forall \mathfrak{L} \ni X \subseteq Y \in \mathfrak{L}, \forall \omega \in \Omega \setminus Y$ such that $X \cup \{\omega\} \in \mathfrak{L}, Y \cup \{\omega\} \in \mathfrak{L}$, we have $F(X \cup \{\omega\}) - F(X) \geq F(Y \cup \{\omega\}) - F(Y)$;

(IV). $\forall x, y \in \Omega, \forall Z \in \mathfrak{L}$ such that $Z \cup \{x\} \in \mathfrak{L}, Z \cup \{y\} \in \mathfrak{L}$, we have $F(Z \cup \{x\}) - F(Z) \geq F(Z \cup \{x,y\}) - F(Z \cup \{y\})$.

Proof: (I) $\Rightarrow$ (II) $\Rightarrow$ (III) $\Rightarrow$ (IV) is clear.
(IV) $\Rightarrow$ (I): Let $X, Y \in \mathfrak{L}$, we want to prove

$$F(X) + F(Y) \geq F(X \cup Y) + F(X \cap Y). \tag{11}$$

If the set difference $|X \Delta Y| \leq 2$, then either $X \subseteq Y$ or $Y \subseteq X$ or $X \setminus Y = \{x\}, Y \setminus X = \{y\}$. (11) holds trivially for the first two cases. For the last case, taking $Z = X \cap Y$ and applying (IV) we have again (11). We perform induction on $|X \Delta Y|$. Assume w.l.o.g. $|X \setminus Y| \geq 2$. Take any *maximal* element in $X \setminus Y$, say ω. Since $\mathfrak{L}$ is simple, we know $\omega \not\preceq z$ for any $z \in X \cap Y$ for otherwise $\omega \in Y$ as Y is an ideal. Thus ω is a maximal element of X hence $X \setminus \{\omega\} \in \mathfrak{L}$. By the induction hypothesis

$$F(X \cup Y) - F(X) \leq F((X \setminus \{\omega\}) \cup Y) - F(X \setminus \{\omega\}) \leq F(Y) - F(X \cap Y).$$

Rearranging completes the induction hence our proof. ■

Intuitively the quantity $F(X \cup \{\omega\}) - F(X)$ denotes the "gain" of adding the element ω into the set X, and the above proposition states that submodular functions have diminishing gain (as the set X gets bigger).

Thanks to Proposition 2.2, we can present (and verify) more examples about submodularity.

Example 2.6: Rank is submodular

Let us consider a fixed matrix $A \in \mathbb{R}^{m \times n}$. Define $\Omega := \{1, \ldots, n\}$ and the rank function R on $X \subseteq \Omega$ as the number of linearly independent columns of A, indexed by the elements of X. Using

Proposition 2.2 we easily verify that R is submodular: First note that $R^j(X) := R(X \cup \{j\}) - R(X) \in \{0, 1\}$ for any $X \subseteq \Omega$ and $j \in \Omega$. Take $j \notin Y \supset X$,

$$R^j(X) = 0 \implies R^j(Y) = 0,$$

hence $R^j(X) \geq R^j(Y)$, proving the submodulartiy of R.

Alert 2.3: Rank vs. cardinality

The (matrix) rank function is usually treated as a matrix generalization of the cardinality function. But compare Example 2.4 and Example 2.6.

Example 2.7: Union coverage is submodular

Let $\Omega = \{A_1, \ldots, A_n\}$ be a collection of (possibly overlapping) sets and $\mu : 2^{\cup\Omega} \to \mathbb{R}$ be some increasing set function that is additive over disjoint sets (an outer measure). The coverage function on 2^Ω, defined as

$$\mathbf{U}(X) := \mu(\cup X) \tag{12}$$

is submodular: Let $X \subset Y \subseteq \Omega \setminus \{A_j\}$, then

$$\begin{aligned}
\mathbf{U}(X \cup \{A_j\}) - \mathbf{U}(X) &= \Big[\mu\big((\cup X) \cap A_j\big) + \mu\big((\cup X) \setminus A_j\big) + \mu\big(A_j \setminus (\cup X)\big)\Big] - \Big[\mu\big((\cup X) \cap A_j\big) + \mu\big((\cup X) \setminus A_j\big)\Big] \\
&= \mu\big(A_j \setminus (\cup X)\big) \\
&\geq \mu\big(A_j \setminus (\cup Y)\big) \\
&= \mathbf{U}(Y \cup \{A_j\}) - \mathbf{U}(Y).
\end{aligned}$$

Example 2.8: Intersection coverage is supermodular

Similar as Example 2.7, consider the intersection

$$\mathbf{I}(X) := \mu(\cap X), \tag{13}$$

which is supermodular:

$$\mathbf{I}(X \cup \{A_j\}) - \mathbf{I}(X) = -\mu\big((\cap X) \setminus A_j\big) \leq -\mu\big((\cap Y) \setminus A_j\big) = \mathbf{I}(Y \cup \{A_j\}) - \mathbf{I}(Y).$$

The next theorem about composition is occasionally useful.

Theorem 2.1: Concave composing monotonic modular is submodular

Let $|\Omega| \geq 3$, $M : 2^\Omega \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. $F := g \circ M$ is (strictly) submodular for every monotone modular M iff g is (strictly) concave.

Proof: Notice that a real-valued function g is concave iff $\forall x < y < z < w$,

$$\frac{g(y) - g(x)}{y - x} \geq \frac{g(z) - g(w)}{z - w}. \tag{14}$$

$\Leftarrow$: Let g be concave (hence satisfy (14)) and M be increasing modular. For any $X \subset Y \subseteq 2^\Omega$, we verify (III) in Proposition 2.2 for any $\omega \in \Omega \setminus Y$:

$$\begin{aligned}
F(X \cup \{\omega\}) - F(X) &= g\Big(M(X) + M(\{\omega\}) - M(\emptyset)\Big) - g(M(X)) \\
&\geq g\Big(M(Y) + M(\{\omega\}) - M(\emptyset)\Big) - g(M(Y))
\end{aligned} \tag{15}$$

$$= F(Y \cup \{\omega\}) - F(Y).$$

If F is decreasing, simply consider $-F$ and $g(-\cdot)$, which is again modular and concave, respectively.

$\Rightarrow$: If g is not concave, (14) is violated at some $x < y = z < w$. Based on this observation we can easily construct a monotonic modular function M that does not satisfy (15). ■

By (7) a modular function M on 2^Ω is increasing iff

$$M(\emptyset) \leq \min_{\omega \in \Omega} M(\{\omega\}). \tag{16}$$

A more general result is in Proposition 4.1 below. Note that this theorem does *not* generalize to simple lattices (or σ-algebras): take $\mathfrak{L} = \{\emptyset, \{1\}, \{1,2\}, \{1,2,3\}\}$, then any function defined on $\mathfrak{L}$ is trivially submodular.

Our last examples of submodularity require an important definition.

Definition 2.3: Base polyhedron and subbase polyhedron

For any set function (not necessarily submodular) F, with $F(\emptyset) = 0$, we associate it with the subbase polyhedron

$$\mathbf{P}_F := \bigcap_{X \in \mathfrak{L}} \{\mathbf{q} \in \mathbb{R}^\Omega : \mathbf{q}(X) \leq F(X)\}, \tag{17}$$

and the base polyhedron

$$\mathbf{B}_F := \mathbf{P}_F \cap \{\mathbf{q} \in \mathbb{R}^\Omega : \mathbf{q}(\Omega) = F(\Omega)\}. \tag{18}$$

Note that $\mathbf{P}_F$ is unbounded with nonempty interior while $\mathbf{B}_F$ can be empty, although we shall see this cannot happen if F is submodular. By definition $\mathbf{P}_F$ and $\mathbf{B}_F$ are closed and convex (which has nothing to do with submodularity). For supermodular functions we reverse the inequality in (17).

Alert 2.4: Centering

Whenever we are referring to the subbase or base polyhedron, it is implicitly assumed that the underlying function satisfies $F(\emptyset) = 0$. This is merely for convenience and we can always achieve it by subtracting $F(\emptyset)$, without affecting properties such as submodularity or monotonicity.

Theorem 2.2: Pointed base polyhedron

The (sub)base polyhedron (of any function), if nonempty, is pointed (i.e. with extreme points) iff $\mathfrak{L}$ is simple.

Proof: A closed convex set is pointed iff it does not contain a line. In our setting, it is further equivalent to require

$$\mathbf{0} = \bigcap_{X \in \mathfrak{L}} \{\mathbf{q} \in \mathbb{R}^\Omega : \mathbf{q}(X) = 0\}.$$

Using Theorem 1.3 the above is satisfied iff

$$\mathbf{0} = \bigcap_{\mathcal{X} \in \mathcal{P}} \{\mathbf{q} \in \mathbb{R}^\Omega : \mathbf{q}(\mathcal{X}) = 0\}.$$

This is possible iff $|\mathcal{X}| \equiv 1$, i.e., $\mathfrak{L}$ is simple. ■

The next result about lattices is useful at times.

Proposition 2.3: Coupling

For any $\emptyset \neq X \notin \mathfrak{L}$, there exist $x \in X$ and $y \notin X$ such that $x \in Y \in \mathfrak{L} \implies y \in Y$.

Proof: $X \notin \mathfrak{L}$ implies X is not the union of an ideal in $(\mathcal{P}, \preceq)$, i.e., there exist $J \preceq I \in \mathcal{P}$ such that $X \cap J = \emptyset, I \subseteq X$. Pick arbitrary $x \in I$ and $y \in J$. ■

Theorem 2.3: Bounded base polyhedron

The base polyhedron (of any function), if nonempty, is bounded iff $\mathfrak{L}$ is a simple Boolean algebra, i.e. $\mathfrak{L} = 2^\Omega$.

Proof: If $\mathfrak{L} = 2^\Omega$, then $\mathbf{B}_F$ is contained in $F(\Omega) - \sum_{y \neq x} F(\{y\}) \leq \mathbf{q}(\{x\}) \leq F(\{x\})$, clearly bounded. On the other hand, if $\mathfrak{L} \subset 2^\Omega$, by Proposition 2.3 there exists $x, y \in \Omega$ so that $x \in Y \in \mathfrak{L} \implies y \in Y$. Take an arbitrary $\mathbf{b} \in \mathbf{B}_F$, the half line $\{\mathbf{b} + \alpha(\mathbf{e}_x - \mathbf{e}_y) : \alpha \geq 0\} \subseteq \mathbf{B}_F$, showing the unboundedness of the base polyhedron $\mathbf{B}_F$. ■

Now we can proceed with more examples of submodularity.

Example 2.9: Permutahedron

Consider the concave function $g(x) = (n + \frac{1}{2})x - \frac{1}{2}x^2$ and the cardinality function $|\cdot| : 2^\Omega \to \mathbb{N}$, where $\Omega := \{1, \ldots, n\}$. By Theorem 2.1 we know the composition

$$F(X) := g(|X|) = \sum_{i=1}^{|X|} (n - i + 1) \tag{19}$$

is submodular. Note that the extreme points of the base polyhedron $\mathbf{B}_F$ are exactly all permutations of Ω.

Example 2.10: Majorization

Take an arbitrary $\mathbf{w} \in \mathbb{R}^\Omega$, where $\Omega := \{1, \ldots, n\}$. Define for all $X \subseteq \Omega$,

$$F_{\mathbf{w}}(X) := \sum_{i=1}^{|X|} w_i, \tag{20}$$

where for the empty sum we take $F_{\mathbf{w}}(\emptyset) = 0$. By Proposition 2.2, $F_{\mathbf{w}}$ is submodular iff for all $X \subseteq Y$, $w_{|X|+1} \geq w_{|Y|+1}$, i.e., $\mathbf{w}$ is ordered decreasingly: $w_1 \geq \cdots \geq w_n$, in which case $\mathbf{b} \in \mathbb{R}^\Omega$ is majorized by $\mathbf{w}$ iff $\mathbf{b} \in \mathbf{B}_{F_{\mathbf{w}}}$, and $\mathbf{p}$ is weakly majorized by $\mathbf{w}$ iff $\mathbf{p} \in \mathbf{P}_{F_{\mathbf{w}}}$ (whereas the if parts hold even when $\mathbf{w}$ is not ordered).

Note that $F_{\mathbf{w}}$ depends only on the cardinality of its input. In fact, any such function F (that vanishes at the empty set) is easily seen to be in the form of (20).

There are many other important submodular functions, but we shall contend ourselves with what we have discussed, for the time being.

We end this section with an important definition.

Proposition 2.4: Saturation sublattice

For any subbase $\mathbf{p} \in \mathbf{P}_F$ of a submodular function $F : \mathfrak{L} \to \mathbb{R}$,

$$\mathsf{S}_{\mathbf{p}} := \{X \in \mathfrak{L} : \mathbf{p}(X) = F(X)\}, \tag{21}$$

if nonempty, is a sublattice of $\mathfrak{L}$.

Proof: Simply note that $\mathsf{S}_{\mathbf{p}}$ is the minimizing set of the nonnegative submodular function $F(X) - \mathbf{p}(X)$.

Apply Proposition 2.1. ■

Definition 2.4: Capacity function

For any subbase $\mathbf{p} \in \mathbf{P}_F$ of an arbitrary function $F : \mathfrak{L} \to \mathbb{R}$ and for any $x, y \in \Omega$,

$$\mathsf{c}(\mathbf{p}, x, y) := \min\{F(X) - \mathbf{p}(X) : x \in X \in \mathfrak{L}, y \notin X\}, \tag{22}$$

and

$$\mathsf{c}(\mathbf{p}, x) := \min\{F(X) - \mathbf{p}(X) : x \in X \in \mathfrak{L}\}. \tag{23}$$

Clearly $\mathsf{c}(\mathbf{p}, x, y) \geq \mathsf{c}(\mathbf{p}, x) \geq 0$. The importance of the capacity function is that it determines how much we can add to $\mathbf{p}$ before it leaves the subbase polyhedron $\mathbf{P}_F$:

- $\mathbf{p} + \alpha \mathbf{e}_x \in \mathbf{P}_F$ for all $0 \leq \alpha \leq \mathsf{c}(\mathbf{p}, x)$;
- $\mathbf{p} + \alpha(\mathbf{e}_x - \mathbf{e}_y) \in \mathbf{P}_F$ for all $0 \leq \alpha \leq \mathsf{c}(\mathbf{p}, x, y)$.

Let $\mathbf{p} \in \mathbf{P}_F$ and F be submodular, the unique *maximal* element in $\mathsf{S}_\mathbf{p}$ (see (21)) coincides with

$$\mathsf{sat}(\mathbf{p}) := \{x \in \Omega : \mathsf{c}(\mathbf{p}, x) = 0\} \in \mathfrak{L}. \tag{24}$$

Define the dependent set $\mathsf{dep}(\mathbf{p}, x) \in \mathfrak{L}$ to be the unique *smallest* element in $\mathsf{S}_\mathbf{p}$ that contains x. We verify that

$$\mathsf{c}(\mathbf{p}, x, y) = 0 \iff x \in \mathsf{sat}(\mathbf{p}) \text{ and } y \notin \mathsf{dep}(\mathbf{p}, x). \tag{25}$$

Proposition 2.5: Bases are maximal

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular and $\mathbf{b} \in \mathbf{P}_F$. Then $\mathbf{b} \in \mathbf{B}_F$ iff $\mathbf{b}$ is maximal in $\mathbf{P}_F$ (under the pointwise order). In particular, $\mathbf{B}_F \neq \emptyset$.

Proof: Clearly, any $\mathbf{b} \in \mathbf{B}_F$ is maximal in $\mathbf{P}_F$. On the other hand, suppose $\mathbf{b}$ is maximal in $\mathbf{P}_F$, then $\mathsf{c}(\mathbf{b}, x) = 0$ for any $x \in \Omega$ for otherwise $\mathbf{b} + \alpha \mathbf{e}_x$ with some small positive α will contradict the maximality of $\mathbf{b}$. Thus $\mathsf{sat}(\mathbf{b}) = \Omega$ and $\mathbf{b}(\Omega) = \mathbf{b}(\mathsf{sat}(\mathbf{b})) = F(\mathsf{sat}(\mathbf{b})) = F(\Omega)$. ■

3 Basic Properties

We study basic properties of submodular functions in this section, starting with trivialities.

Theorem 3.1: Submodular functions form convex cone

The set of submodular functions is a convex cone, i.e., if $F_i : \mathfrak{L} \to \mathbb{R}$ are submodular, so is $\lambda_1 F_1 + \lambda_2 F_2$ for any $\lambda_i \geq 0$ where $i \in \{1, 2\}$. ■

Therefore the set of submodular functions on the same domain $\mathfrak{L}$ induces a pre-order on $\mathbb{R}^\mathfrak{L}$, the vector space of all real-valued set functions on $\mathfrak{L}$. In fact, this pre-order is generating.

Theorem 3.2: Submodular-Supermodular decomposition

Any function $G : \mathfrak{L} \to \mathbb{R}$ can be written as $F_1 - F_2$, with F_1 and F_2 being submodular on $\mathfrak{L}$.

Proof: Define

$$\delta_G := \min_{X \in \mathfrak{L}, Y \in \mathfrak{L}, X \not\subseteq Y, Y \not\subseteq X} G(X) + G(Y) - G(X \cup Y) - G(X \cap Y),$$

and let H be a strictly submodular function on $\mathfrak{L}$ with $\delta_H > |\delta_G|$ (the existence of such a function is discussed in Example 2.3). If $\delta_G \geq 0$, then let $F_1 = G$ and $F_2 = 0$; otherwise let $F_1 = H$ and $F_2 = H - G$. ■

Of course, the decomposition need not be unique. It remains a tricky question to define certain *minimal* decompositions that are useful in applications.

Denote $\mathfrak{L}_1 \wedge \mathfrak{L}_2$ as the largest sublattice contained in both $\mathfrak{L}_1$ and $\mathfrak{L}_2$. In fact, $\mathfrak{L}_1 \wedge \mathfrak{L}_2 = \mathfrak{L}_1 \cap \mathfrak{L}_2$. Similarly, denote $\mathfrak{L}_1 \vee \mathfrak{L}_2$ as the smallest lattice that contains $\mathfrak{L}_1 \cup \mathfrak{L}_2$. It is easily seen that $\mathfrak{L}_1 \vee \mathfrak{L}_2$ can be obtained by taking all *disjoint* unions of elements in $\mathfrak{L}_1 \cup \mathfrak{L}_2$.

Theorem 3.3: Summation

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ and $G : \mathfrak{L}_2 \to \mathbb{R}$ be centered submodular, then $F + G : \mathfrak{L}_1 \wedge \mathfrak{L}_2 \to \mathbb{R}$ with $X \mapsto F(X) + G(X)$ is submodular, and

$$\mathbf{P}_{F+G} = \mathbf{P}_F + \mathbf{P}_G. \tag{26}$$

Proof: Clearly LHS $\supseteq$ RHS. For the converse, note that $\mathbf{q} \in$ LHS iff $\forall X \in \mathfrak{L}_1 \wedge \mathfrak{L}_2$, $\mathbf{q}(X) \leq F(X) + G(X)$, i.e., $\mathbf{q}(X) - F(X) \leq G(X)$. According to Theorem 4.9 below, there exists $\mathbf{p} \in \mathbb{R}^\Omega$ such that $\mathbf{q} - F \leq \mathbf{p}$ on $\mathfrak{L}_1$ and $\mathbf{p} \leq G$ on $\mathfrak{L}_2$. So we have found $\mathbf{q} - \mathbf{p} \in \mathbf{P}_F$ and $\mathbf{p} \in \mathbf{P}_G$. ■

This result is a bit surprising because we do *not*, although we could, restrict F (or G) on the RHS of (26) to $\mathfrak{L}_1 \wedge \mathfrak{L}_2$.

Theorem 3.4: Basic set operations preserve submodularity

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular. Then for any $S \subseteq \Omega$

$$\begin{aligned} U_S(X) &:= F(X \cup S) \text{ defined on } \{Z \subseteq \Omega : Z \cup S \in \mathfrak{L}\} \\ I_S(X) &:= F(X \cap S) \text{ defined on } \{Z \subseteq \Omega : Z \cap S \in \mathfrak{L}\} \\ C_S(X) &:= F(S \setminus X) \text{ defined on } \{Z \subseteq \Omega : S \setminus Z \in \mathfrak{L}\} \end{aligned}$$

are all submodular. ■

However, some caution must also be taken (especially when treating submodularity as *discrete convexity*).

Example 3.1: Pointwise maximum/minimum does not preserve submodularity

Consider the modular functions defined on 2^Ω with $\Omega := \{\omega_1, \omega_2, \omega_3\}$:

$$\begin{aligned} M_1(\emptyset) = 0, M_1(\{\omega_1\}) = 1, M_1(\{\omega_2\}) = -1, M_1(\{\omega_3\}) = 1; \\ M_2(\emptyset) = 0, M_2(\{\omega_1\}) = 2, M_1(\{\omega_2\}) = -2, M_1(\{\omega_3\}) = 2. \end{aligned}$$

Taking $X = \{\omega_1, \omega_2\}$ and $Y = \{\omega_2, \omega_3\}$ we verify that $M_1 \vee M_2$ is not submodular, while taking $X = \{\omega_1\}$ and $Y = \{\omega_2\}$ we see that $M_1 \wedge M_2$ is not submodular either.

The disappointment brought by Example 3.1 is that we cannot meaningfully talk about the *tightest* submodular envelop of a general set function. Nevertheless, we do have the following theorem. Note also that for $F_1 \vee F_2 : \mathfrak{L}_1 \wedge \mathfrak{L}_2 \to \mathbb{R}$,

$$\mathbf{P}_{F_1 \vee F_2} \supseteq \text{conv}\{\mathbf{P}_{F_1} \cup \mathbf{P}_{F_2}\}. \tag{27}$$

Theorem 3.5: Monotonic minimum preserves submodularity

Let $F, G : \mathfrak{L} \to \mathbb{R}$ be submodular. If $F - G$ is monotonic, then $F \wedge G$ is submodular.

Proof: Fix two arbitrary sets $X, Y \in \mathfrak{L}$. Denote $H = F \wedge G$. If H agrees with, say F on $\{X, Y\}$, then

$$H(X) + H(Y) = F(X) + F(Y) \geq F(X \cup Y) + F(X \cap Y) \geq H(X \cup Y) + H(X \cap Y).$$

Otherwise assume that $H(X) = F(X), H(Y) = G(Y)$ and that $F - G$ is increasing, hence

$$\begin{aligned} H(X) + H(Y) = F(X) + G(Y) &\geq F(X \cup Y) - F(Y) + F(X \cap Y) + G(Y) \\ &\geq G(X \cup Y) - G(Y) + F(X \cap Y) + G(Y) \\ &\geq H(X \cup Y) + H(X \cap Y). \end{aligned}$$

The case when $F - G$ is decreasing is proved similarly. ■

Corollary 3.1: Monotone saturation preserves submodularity

Let $F : \mathfrak{L} \to \mathbb{R}$ be monotone, then F is submodular iff $\forall c \in \mathbb{R}$, $F \wedge c$ is submodular.

Proof: $\Rightarrow$: Follows from Theorem 3.5 since $F - c$ is monotone.
$\Leftarrow$: Take $c \geq \max_{X \in \mathfrak{L}} F(X)$. ■

Example 3.2: Monotonicity is not essential in Corollary 3.1

Let $\Omega := \{\omega_1, \omega_2\}$ and define the non-monotone submodular function $F(\emptyset) = 0, F(\{\omega_1\}) = -1, F(\{\omega_2\}) = 1, F(\Omega) = -100$. We notice that $F \wedge c$ is submodular $\forall c \in \mathbb{R}$. On the other hand, changing $F(\Omega)$ to 0 we observe that F becomes modular but $\forall -1 < c < 1$, $F \wedge c$ is not submodular.

Definition 3.1: (Infimal) convolution

For any function $F : \mathfrak{L}_1 \to \mathbb{R}$ and $G : \mathfrak{L}_2 \to \mathbb{R}$, we define their convolution $F \boxdot G : \mathfrak{L}_1 \vee \mathfrak{L}_2 \to \mathbb{R}$ as

$$(F \boxdot G)(X) = \min\{F(Y) + G(Z) - F(\emptyset) - G(\emptyset) : Y \in \mathfrak{L}_1, Z \in \mathfrak{L}_2, Y \cap Z = \emptyset, Y \cup Z = X\}. \quad (28)$$

Proposition 3.1: Convolution is commutative and associative

$F \boxdot G = G \boxdot F$, $(F \boxdot G) \boxdot H = F \boxdot (G \boxdot H)$, and $(F - F(\emptyset)) \boxdot (G - G(\emptyset)) = F \boxdot G$. If $F \leq G$ are centered submodular on the same lattice, then $F \boxdot G = F$.

Therefore when discussing the convolution we may assume the functions are centered. Clearly $F \boxdot G \leq F - F(\emptyset)$ on $\mathfrak{L}_1$ and similarly $F \boxdot G \leq G - G(\emptyset)$ on $\mathfrak{L}_2$, namely the convolution always brings down the function values. Thus $\mathbf{P}_{F \boxdot G} \subseteq \mathbf{P}_F \cap \mathbf{P}_G$. The converse is also true.

Theorem 3.6: Polyhedron of the convolution is intersection of polyhedra

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ and $G : \mathfrak{L}_2 \to \mathbb{R}$ be centered, then

$$\mathbf{P}_{F \boxdot G} = \mathbf{P}_F \cap \mathbf{P}_G. \quad (29)$$

Proof: We need only prove LHS $\supseteq$ RHS. Indeed, by definition, each defining inequality of LHS is simply a summation of the defining inequalities of RHS. ■

Quite unfortunately, submodularity in general is not preserved under convolution.

Example 3.3: Convolution of submodular functions need not be submodular

Let $\mathfrak{L}_1 = \mathfrak{L}_2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. Let $F = \{0,3,3,3,3,4,4,4\}, G = \{0,3,3,2,4,3,4,4\}$. Clearly both F and G are submodular. But $F \boxdot G = \{0,3,3,2,3,3,4,4\}$, which is not submodular: $(F \boxdot G)(\{1,2\}) + (F \boxdot G)(\{1,3\}) = 3+3 < 3+4 = (F \boxdot G)(\{1\}) + (F \boxdot G)(\{1,2,3\})$.

However, things are different when one of the functions is modular.

Theorem 3.7: Submodular convolving modular is submodular

The convolution of a submodular functon $F : \mathfrak{L} \to \mathbb{R}$ and a modular function $M : 2^\Omega \to \mathbb{R}$, $F \boxdot M : 2^\Omega \to \mathbb{R}$, remains submodular.

Proof: Let $X, Y \in 2^\Omega$ and $X_1, Y_1 \in \mathfrak{L}, X_2, Y_2 \in 2^\Omega$ so that $X_1 \cup X_2 = X, X_1 \cap X_2 = \emptyset, Y_1 \cup Y_2 = Y, Y_1 \cap Y_2 = \emptyset$, and

$$\begin{aligned}(F \boxdot M)(X) + (F \boxdot M)(Y) &= F(X_1) + M(X_2) + F(Y_1) + M(Y_2) \\ &\geq F(X_1 \cup Y_1) + F(X_1 \cap Y_1) + M(X_2 \cup Y_2) + M(X_2 \cap Y_2) \\ &= F(X_1 \cup Y_1) + F(X_1 \cap Y_1) + M((X \cup Y) - (X_1 \cup Y_1)) \\ &\qquad + M((X \cap Y) - (X_1 \cap Y_1)) \\ &\geq (F \boxdot M)(X \cup Y) + (F \boxdot M)(X \cap Y),\end{aligned}$$

where the second equality follows from the modularity of M. Note that we need M to be defined on the whole Boolean algebra 2^Ω since we have little control of where, say $(X \cup Y) - (X_1 \cup Y_1)$ sits in. ■

Consider the simple lattice $\mathfrak{L} = \{\emptyset, \{1\}, \{1,2\}, \{1,3\}, \{1,2,3\}\}$, define M as the modular function $\{0,0,1,1,2\}$ and F as the submodular function $\{0,0,3,0,2\}$. Then $F \boxdot M = \{0,0,1,0,2\}$ is not submodular.

Clearly, $F \boxdot M$ is the largest centered submodular function that is majorized by both F and M.

Corollary 3.2: Monotonization

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular, then for all $X \subseteq \Omega$,

$$F^{\downarrow}(X) := \min_{\mathfrak{L} \ni Y \subseteq X} F(Y) \tag{30}$$

is monotonically decreasing and submodular.

Proof: $F^{\downarrow}(X) = (F \boxdot \mathbf{0})(X) + F(\emptyset)$. Apply Theorem 3.7. ■

The next theorem provides a decomposition rule for submodular functions.

Theorem 3.8: Restriction and contraction are submodular

If $F : \mathfrak{L} \to \mathbb{R}$ is submodular, then $\forall S \in \mathfrak{L}$,

$$\mathfrak{L}_S := \{X \in \mathfrak{L} : X \subseteq S\}, \quad F_S(Z) := F(Z), \tag{31}$$

$$\mathfrak{L}^S := \{X \setminus S : X \in \mathfrak{L}\}, \quad F^S(Z) := F(Z \cup S) - F(S) \tag{32}$$

are both submodular, called restriction and contraction of F w.r.t. the set S, respectively. ■

For any vector $\mathbf{q} \in \mathbb{R}^\Omega$, we denote $\mathbf{q}_S$ the subvector in $\mathbb{R}^S$ that restricts to the components in S, and $\mathbf{q}^S$ the subvector in $\mathbb{R}^{\Omega \setminus S}$ that restricts to the complement of S.

Theorem 3.9: (Sub)base decomposition

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular and $S \in \mathfrak{L}$. The direct sum $\mathbf{p} := \mathbf{p}_S \oplus \mathbf{p}^S \in \mathbf{P}_F$ if $\mathbf{p}_S \in \mathbf{P}_{F_S}$ and

$\mathbf{p}^S \in \mathbf{P}_{F^S}$. Moreover, the converse is also true if $\mathbf{p}_S$ is additionally maximal, and any two of the following imply the third:

- $\mathbf{p}$ is maximal in $\mathbf{P}_F$;
- $\mathbf{p}_S$ is maximal in $\mathbf{P}_{F_S}$;
- $\mathbf{p}^S$ is maximal in $\mathbf{P}_{F^S}$.

Proof: First assume $\mathbf{p}_S \in \mathbf{P}_{F_S}$ and $\mathbf{p}^S \in \mathbf{P}_{F^S}$. For any $X \in \mathfrak{L}$,

$$\begin{aligned}\mathbf{p}(X) &= \mathbf{p}(X \cap S) + \mathbf{p}(X \setminus S) = \mathbf{p}_S(X \cap S) + \mathbf{p}^S(X \setminus S) \le F_S(X \cap S) + F^S(X \setminus S) \\ &= F(X \cap S) + F(X \cup S) - F(S) \le F(X).\end{aligned}$$

Hence $\mathbf{p} \in \mathbf{P}_F$.

Conversely, let $\mathbf{p} \in \mathbf{P}_F$ and $\mathbf{p}_S(S) = F_S(S) = F(S)$. For $X \in \mathfrak{L}_S$, $\mathbf{p}_S(X) = \mathbf{p}(X) \le F(X) = F_S(X)$ while for $Y \in \mathfrak{L}^S$, $\mathbf{p}^S(Y) = \mathbf{p}(Y \cup S) - \mathbf{p}(S) \le F(Y \cup S) - F(S) = F^S(Y)$.

Moreover, if $\mathbf{p}_S$ and $\mathbf{p}^S$ are maximal, then $\mathbf{p}(\Omega) = \mathbf{p}_S(S) + \mathbf{p}^S(\Omega \setminus S) = F(S) + F(\Omega) - F(S) = F(\Omega)$, proving the maximality of $\mathbf{p}$; if $\mathbf{p}^S$ and $\mathbf{p}$ are maximal, then $\mathbf{p}_S(S) = \mathbf{p}(\Omega) - \mathbf{p}^S(\Omega \setminus S) = F(\Omega) - F(\Omega) + F(S) = F(S)$, proving the maximality of $\mathbf{p}_S$; finally if $\mathbf{p}_S$ and $\mathbf{p}$ are maximal, then $\mathbf{p}^S(\Omega \setminus S) = \mathbf{p}(\Omega) - \mathbf{p}_S(S) = F(\Omega) - F(S) = F^S(\Omega \setminus S)$, proving the maximality of $\mathbf{p}^S$. ■

Remark 3.1: Decomposition and divide and conquer

Under our definition, for any $A, B \in \mathfrak{L}$, it makes sense to talk about $(F_A)^B$ only when $B \subseteq A$, and similarly $(F^B)_A$ only when $A \cap B = \emptyset$. Therefore the notation F_A^B has a unique meaning.

Moreover, for any $A, B \in \mathfrak{L}$, if $A \subseteq B$, $(F_B)_A = F_A$ and $(F^A)^B = F^B$. For an increasing sequence $A_1 \subset A_2 \subset \cdots \subset A_k$ in $\mathfrak{L}$, we have the decomposition (in the sense of Theorem 3.9) $F = F_{A_k} \oplus F^{A_k}$, and of course $F_{A_k} = (F_{A_k})_{A_{k-1}} \oplus (F_{A_k})^{A_{k-1}} = F_{A_{k-1}} \oplus F_{A_k}^{A_{k-1}}$. Recursively this gives

$$F = F_{A_1} \oplus F_{A_2}^{A_1} \oplus F_{A_3}^{A_2} \cdots \oplus F_{A_k}^{A_{k-1}} \oplus F^{A_k}. \tag{33}$$

The importance of the above decomposition is self-evident: divide and conquer now comes into play.

Proposition 3.2: Attainability

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular. For any $X \in \mathfrak{L}$, there exists a (sub)base $\mathbf{b}$ of F such that $\mathbf{b}(X) = F(X)$; while for any $X \notin \mathfrak{L}$ and any $K > 0$, there exists a subbase $\mathbf{p}$ such that $\mathbf{p}(X) \ge K$.

Proof: Suppose $X \in \mathfrak{L}$. Take a base of F_X and a (sub)base of F^X, their direct sum is a (sub)base of F and satisfy our requirement, according to Theorem 3.9.

If $X \notin \mathfrak{L}$, by Proposition 2.3 we can find $x \in X, y \notin X$ so that $x \in Y \in \mathfrak{L} \implies y \in Y$. Thus for any subbase $\mathbf{p}$, $\mathbf{p} + \alpha(\mathbf{e}_x - \mathbf{e}_y)$ for α large enough satisfies our requirement. ■

Suppose we have a submodular function $F : \mathfrak{L} \to \mathbb{R}$. Of course F is determined by its values on all sets in $\mathfrak{L}$, but is it possible to identify a smaller subset of $\mathfrak{L}$, namely a "base" or "subbase", that nevertheless would still allow us to recover F? The answer is a pleasant yes, but definitions first.

Definition 3.2: Intersecting family

A family of subsets $\mathfrak{D}$ of Ω is called intersecting if for all $X, Y \in \mathfrak{D}$, $X \cap Y \neq \emptyset \implies X \cap Y \in \mathfrak{D}, X \cup Y \in \mathfrak{D}$. The function $F : \mathfrak{D} \to \mathbb{R}$ is called intersecting submodular if for all $X, Y \in \mathfrak{D}$, $X \cap Y \neq \emptyset$ we have

$$F(X) + F(Y) \ge F(X \cap Y) + F(X \cup Y).$$

Clearly, an intersecting family $\mathfrak{D}$ need not be a lattice since for $X, Y \in \mathfrak{D}$, if $X \cap Y = \emptyset$, then we may not have $X \cup Y \in \mathfrak{D}$. We fix this issue by enlarging $\mathfrak{D}$ with all unions of its elements. Denote this enlargement as $\breve{\mathfrak{D}}$. Since $\mathfrak{D}$ is intersecting, we easily verify that $\breve{\mathfrak{D}}$ consists of all *disjoint* unions of elements in $\mathfrak{D}$. Note that we allow taking an empty union, which results in the empty set, and an empty sum is taken to be 0. Recall that a partition is named *proper* if none of its members is empty.

Theorem 3.10: Extending an intersecting submodular function

Let $\mathfrak{D}$ be an intersecting family and $F : \mathfrak{D} \to \mathbb{R}$ be intersecting submodular. Then $\breve{\mathfrak{D}}$ is a lattice. If for each $X \in \breve{\mathfrak{D}}$,

$$\breve{F}(X) := \min \left\{ \sum_{A \in \mathcal{A}} F(A) : \mathcal{A} \subseteq \mathfrak{D} \text{ is a proper partition of } X \right\}, \tag{34}$$

then $\breve{F}$ is submodular (on $\breve{\mathfrak{D}}$) and $\breve{F}(\emptyset) = 0$.

Proof: It is clear that $\breve{\mathfrak{D}}$ is a lattice. Also, $\emptyset \in \breve{\mathfrak{D}}$ and $\breve{F}(\emptyset) = 0$. By definition, each $X \in \breve{\mathfrak{D}}$ is a *disjoint* union of elements in $\mathfrak{D}$, therefore $\breve{F}$ is well-defined. Let $X, Y \in \breve{\mathfrak{D}}$ and $\mathcal{A}, \mathcal{B} \subseteq \mathfrak{D}$ be partitions of X and Y, respectively, so that $\breve{F}(X) = \sum_{A \in \mathcal{A}} F(A)$ and $\breve{F}(Y) = \sum_{B \in \mathcal{B}} F(B)$. For any $A, B \in \mathcal{C} := \mathcal{A} \cup \mathcal{B}$, if $A \cap B \neq \emptyset$ and $A \not\subseteq B \not\subseteq A$, we replace A with $A \cap B$ and B with $A \cup B$. Since

$$|A| \cdot |\Omega \setminus A| + |B| \cdot |\Omega \setminus B| > |A \cap B| \cdot |\Omega \setminus (A \cap B)| + |A \cup B| \cdot |\Omega \setminus (A \cup B)|,$$

after a finite number of these steps, we must have either $A \cap B = \emptyset$ or $A \subseteq B$ or $B \subseteq A$, i.e., a laminar system. All maximal elements in $\mathcal{C}$ are disjoint and their union is $X \cup Y$, i.e. a partition, while the remaining elements form a partition for $X \cap Y$ (since during our re-shuffling we never change the number of total elements). Since F is intersecting submodular, we have

$$\breve{F}(X) + \breve{F}(Y) = \sum_{A \in \mathcal{A}} F(A) + \sum_{B \in \mathcal{B}} F(B) \geq \sum_{C \in \mathcal{C}} F(C) \geq \breve{F}(X \cap Y) + \breve{F}(X \cup Y).$$

■

Very pleasantly, if $F(\emptyset) \geq 0$, by construction we have

$$\mathbf{P}_F = \mathbf{P}_{\breve{F}}, \tag{35}$$

and $\breve{F}$ is integer valued if F is.

Frequently we will need to center a submodular function F so that $F(\emptyset) = 0$. This can be achieved by simply subtracting the constant function $F(\emptyset)$ without affecting submodularity. It turns out that there is another fancier way.

Corollary 3.3: Dilworth truncation

For any submodular function $F : \mathfrak{L} \to \mathbb{R}$, its Dilworth truncation

$$\breve{F}(X) = \min \left\{ \sum_{A \in \mathcal{A}} F(A) : \mathcal{A} \subseteq \mathfrak{L} \text{ is a proper partition of } X \right\} \tag{36}$$

is the unique maximal submodular function that vanishes at $\emptyset$ and that is majorized by F.

Recall that we can generate a topology by taking (arbitrary) unions of a basis; Theorem 3.10 is similar in spirit. Quite naturally, one wonders if we can generate a submodular function from a subbase, perhaps by taking intersections first and then unions?

Definition 3.3: Crossing family

A family of subsets $\mathfrak{D}$ of Ω is called crossing if for all $X, Y \in \mathfrak{D}$, $X \cap Y \neq \emptyset$ and $X \cup Y \neq \Omega \implies X \cap Y \in \mathfrak{D}, X \cup Y \in \mathfrak{D}$. The function $F : \mathfrak{D} \to \mathbb{R}$ is called crossing submodular if for all $X, Y \in \mathfrak{D}$,

$X \cap Y \neq \emptyset, X \cup Y \neq \Omega$ we have

$$F(X) + F(Y) \geq F(X \cap Y) + F(X \cup Y).$$

As suggested, we enlarge $\mathfrak{D}$ by taking all intersections of its elements, denoted as $\hat{\mathfrak{D}}$. We expect $\hat{\mathfrak{D}}$ to be intersecting. Note that we allow taking empty intersections, which result in the full set Ω. To understand the structure of $\hat{\mathfrak{D}}$, recall that a family of sets $\mathcal{A} \subseteq 2^\Omega$ is called a (proper) copartition of Ω if $\{X \setminus A : A \in \mathcal{A}\}$ form a (proper) partition of Ω. For each $x \in \Omega$, define $\mathfrak{D}_x := \{X \in \mathfrak{D} : x \in X\}$, then $\mathfrak{D} = \sum_{x\in\Omega} \mathfrak{D}_x$ is a *disjoint* union. So we need only take intersections within each $\mathfrak{D}_x$. Moreover, since any two sets in $\mathfrak{D}_x$ intersect and $\mathfrak{D}$ is crossing, we need only take intersections of any two sets in $\mathfrak{D}_x$ whose union is the full set Ω. In other words, each set in $\hat{\mathfrak{D}}_x$ admits a copartition whose elements come from $\mathfrak{D}_x$.

Theorem 3.11: Extending a crossing submodular function

Let $\mathfrak{D}$ be a crossing family and $F : \mathfrak{D} \to \mathbb{R}$ be crossing submodular. Then $\hat{\mathfrak{D}}$ is intersecting. If for each $X \in \hat{\mathfrak{D}}$,

$$\hat{F}(X) := \min \left\{ \sum_{A\in\mathcal{A}} F(A) : \mathcal{A} \subseteq \mathfrak{D} \text{ is a proper copartition of } \Omega \setminus X \right\}, \tag{37}$$

then $\hat{F}$ is intersecting submodular (on $\hat{\mathfrak{D}}$) and $\hat{F}(\Omega) = 0$.

Proof: For each $x \in \Omega$ define $\mathsf{E}_x := \{\Omega \setminus X : x \in X \in \mathfrak{D}\}$. Fix any $\Omega \setminus Y, \Omega \setminus Z \in \mathsf{E}$ where $x \in Y \in \mathfrak{D}, x \in Z \in \mathfrak{D}$. If $(\Omega \setminus Y) \cap (\Omega \setminus Z) \neq \emptyset$, meaning $Y \cup Z \neq \Omega$, then $x \in Y \cap Z \in \mathfrak{D}$ since $\mathfrak{D}$ is crossing. Thus $(\Omega \setminus X) \cup (\Omega \setminus Y) = \Omega \setminus (X \cap Y) \in \mathsf{E}_x$ and similarly $(\Omega \setminus X) \cap (\Omega \setminus Y) = \Omega \setminus (X \cup Y) \in \mathsf{E}_x$, i.e., E_x is intersecting. Therefore $\check{\mathsf{E}}_x$ is a lattice, and so is $\hat{\mathfrak{D}}_x = \Omega \setminus \check{\mathsf{E}}_x$. Since $\hat{\mathfrak{D}} = \sum_{x\in\Omega} \hat{\mathfrak{D}}_x$ is a *disjoint* union, it is intersecting.

To see that $\hat{F}$ is intersecting submodular, it suffices to consider its restriction $\hat{F}_x$ to the lattice $\hat{\mathfrak{D}}_x$. Define $G : \mathsf{E}_x \to \mathbb{R}$ by $G(X) = F(\Omega \setminus X)$ for each $X \in \mathsf{E}_x$. Since F is crossing submodular, G is intersecting submodular. By Theorem 3.10 $\check{G}$ is submodular on $\check{\mathsf{E}}_x$. As it can be verified that $\hat{F}_x(X) = \check{G}(\Omega \setminus X)$, we know $\hat{F}_x$ is submodular. ■

Pleasantly, if $F(\Omega) = 0$, $\mathbf{b}(X) \leq \hat{F}(X) \iff \mathbf{b}(X) \leq \sum_{A\in\mathcal{A}} F(A) \iff \mathbf{b}(\Omega \setminus X) \geq -\sum_{A\in\mathcal{A}} F(A) \iff \sum_{A\in\mathcal{A}} \mathbf{b}(\Omega \setminus A) \geq -\sum_{A\in\mathcal{A}} F(A) \iff \sum_{A\in\mathcal{A}} \mathbf{b}(A) \leq \sum_{A\in\mathcal{A}} F(A)$, thus

$$\mathbf{B}_F = \mathbf{B}_{\hat{F}}. \tag{38}$$

More generally, for any given $F(\Omega)$, we translate back and forth to get a similar result. Clearly $\hat{F}$ is integer valued if F is. On the other hand, if $F(\Omega)$ is not given, we may not maintain even the subbase polyhedron: Consider $\Omega = \{1, 2, 3\}, \mathfrak{L} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ and $F = \{1, 1, 1\}$. Its subbase polyhedron is not even integral, with $(1/2, 1/2, 1/2)$ being a vertex.

4 Greedy Algorithm

We consider the following maximization problem in this section:

$$\sigma^\flat_F(\mathbf{w}) := \max_{\mathbf{b}\in\mathbf{B}_F} \langle \mathbf{b}, \mathbf{w} \rangle, \tag{39}$$

namely the support function of the base polyhedron of the centered function $F : \mathfrak{L} \to \mathbb{R}$. Here and throughout we use $\langle \cdot, \cdot \rangle$ for the inner product of the underlying space.

Remark 4.1: "Simplification" in the primal

If $\mathfrak{L}$ is not simple, then $\mathcal{P} \neq \Omega$. By considering $\mathbf{b} + \alpha(\mathbf{e}_x - \mathbf{e}_y)$, we know (39) is finite only when $w_x = w_y$ for all $x, y \in \mathcal{X} \in \mathcal{P}$. Therefore we can take the simplification of F, that is, $\tilde{F}$ constructed

in Remark 1.3, and consider (at least in theory) the equivalent problem:

$$\max_{\tilde{\mathbf{b}}\in\mathbf{B}_{\tilde{F}}} \left\langle \tilde{\mathbf{b}}, \tilde{\mathbf{w}} \right\rangle,$$

where for all $\mathcal{X} \in \mathcal{P}$, $\tilde{w}(\{\mathcal{X}\}) = w(\{x\})$ for any $x \in \mathcal{X}$, and $\tilde{b}(\{\mathcal{X}\}) = \sum_{x\in\mathcal{X}} b(\{x\})$.

To motivate the next result, recall from Theorem 2.3 that the base polyhedron is bounded iff $\mathfrak{L}$ is a simple Boolean algebra, thus for a simple $\mathfrak{L}$ we still need some assumption on the "weight" vector $\mathbf{w}$ so that (39) admits a maximizer. This is achieved by a beautiful result that characterizes the polar cone of the isotonic cone, defined as follows.

The ordered set $(\Omega = \mathcal{P}, \preceq)$ induces the (anti)isotonic (convex) cone

$$\mathcal{K} := \{\mathbf{w} \in \mathbb{R}^\Omega : w_x \geq w_y,\ \forall x \preceq y\}, \tag{40}$$

which is generating as it clearly has nonempty interior. Our goal is to characterize its polar cone: $\mathcal{K}^\circ := \{\mathbf{z} \in \mathbb{R}^\Omega : \langle \mathbf{w}, \mathbf{z}\rangle \leq 0, \forall \mathbf{w} \in \mathcal{K}\}$. Recall that a subset Y of Ω is called an ideal if $\forall x \preceq y \in Y \implies x \in Y$.

Theorem 4.1: Characterizing the polar cone

$$\mathcal{K} = \{\mathbf{w} \in \mathbb{R}^\Omega : \langle \mathbf{w}, \mathbf{e}_y - \mathbf{e}_x\rangle \leq 0, \forall y \text{ that is an immediate successor of } x\} \tag{41}$$

$$\mathcal{K}^\circ = \{\mathbf{z} \in \mathbb{R}^\Omega : \mathbf{z}(\Omega) = 0, \mathbf{z}(X) \leq 0, \forall \text{ proper ideal } X \text{ in } (\Omega, \preceq)\} =: 0^+\mathbf{B}_F \tag{42}$$

Proof: The first equality is standard. Thus $\mathcal{K}^\circ$ consists of all conic combinations of the vectors $\mathbf{e}_y - \mathbf{e}_x$, each of which clearly belongs to the right-hand side, meaning LHS $\subseteq$ RHS in (42). For the other direction, we slightly abuse the notation to let

$$w_1 > w_2 > \cdots > w_k \tag{43}$$

be the distinct elements of $\mathbf{w} \in \mathcal{K}$. Define for each $i \in \{1, \ldots, k\}$,

$$A_i := \{x \in \Omega : w_x \geq w_i\}. \tag{44}$$

Thanks to the monotonicity of $\mathbf{w}$, $A_k = \Omega$ and all other A_i's are proper ideals. Let $A_0 = \emptyset$, then for any fixed $\mathbf{z} \in$ RHS,

$$\langle \mathbf{w}, \mathbf{z}\rangle = \sum_{x\in\Omega} w_x z_x = \sum_{i=1}^{k} w_i \sum_{x\in A_i\setminus A_{i-1}} z_x = \sum_{i=1}^{k} w_i(\mathbf{z}(A_i) - \mathbf{z}(A_{i-1})) = \sum_{i=1}^{k-1}(w_i - w_{i+1})\mathbf{z}(A_i) \leq 0,$$

where we have used the fact that $\mathbf{z}(\Omega) = 0$ and $\mathbf{z}(A_i) \leq 0$. Therefore RHS $\subseteq$ LHS too. ■

Note that (41) is a minimal characterization while (42) may contain redundant constraints. Indeed, we need only consider all principal ideals (since any ideal is a union of principal ideals when $|\Omega| < \infty$).

Now we are ready to discuss when (39) has a maximizer.

Theorem 4.2: Existence of maximizer: Base polyhedron

Let $\mathfrak{L}$ be simple, then problem (39) has a solution iff $\mathbf{B}_F \neq \emptyset$ and $\mathbf{w} : \Omega \to \mathbb{R}$ is monotonically decreasing from $(\Omega, \preceq)$ to $\mathbb{R}$ (equipped with its usual order), namely $\mathbf{w} \in \mathcal{K}$.

Proof: Problem (39) is bounded from below iff $\mathbf{w}$ is in the polar of the recession cone of the base polyhedron, that is $\langle \mathbf{b}, \mathbf{w}\rangle \leq 0$ for all $\{\mathbf{b} \in \mathbb{R}^\Omega : \mathbf{b}(\Omega) = 0, \mathbf{b}(X) \leq 0$ for all proper ideals X in $(\Omega, \preceq)\}$. By Theorem 4.1 $\mathbf{w} \in (\mathcal{K}^\circ)^\circ = \mathcal{K}$, where the isotonic cone $\mathcal{K}$ is defined in (40). ■

Note that Theorem 4.2 is mostly about the underlying lattice: we have *not* used any information about the function F except the nonemptyness of its base polyhedron. In particular, when $\mathfrak{L}$ is a simple Boolean algebra, the ordered set $(\Omega, \preceq)$ is trivial, hence no assumption on the weight vector $\mathbf{w}$ is needed, matching the conclusion we have drawn from Theorem 2.3.

Before continuing, let us point out an easy observation about the support function of the subbase polyhedron:

$$\sigma_F(\mathbf{w}) := \max_{\mathbf{p}\in\mathbf{P}_F} \langle \mathbf{p}, \mathbf{w}\rangle \,. \tag{45}$$

Theorem 4.3: Existence of maximizer: Subbase polyhedron

Let $\mathfrak{L}$ be simple, (45) has a maximizer iff $\mathbf{w} \in \mathcal{K} \cap \mathbb{R}^\Omega_+$, whose polar is $\mathcal{K}^\circ + \mathbb{R}^\Omega_- = 0^+\mathbf{P}_F$. If F is submodular, (45) reduces to (39), and for any $t \le \min_i w_i$,

$$\sigma^\flat_F(\mathbf{w}) = t \cdot F(\Omega) + \sigma_F(\mathbf{w} - t\mathbf{1}). \tag{46}$$

Proof: (45) has a maximizer iff $\mathbf{w} \in (0^+\mathbf{P}_F)^\circ$, with the recession cone $0^+\mathbf{P}_F := \{\mathbf{z} \in \mathbb{R}^\Omega : \mathbf{z}(X) \le 0, \text{ for all ideal } X \text{ in } (\Omega, \preceq)\}$. Thus all we need is to prove $\mathcal{K}^\circ + \mathbb{R}^\Omega_- = 0^+\mathbf{P}_F$. Thanks to Theorem 4.1, $\mathsf{LHS} \subseteq \mathsf{RHS}$ is clear. Fix any $\mathbf{w} \in \mathcal{K} \cap \mathbb{R}^\Omega_+$, a similar argument as in the proof of Theorem 4.1 yields $\langle \mathbf{w}, \mathbf{z}\rangle \le 0$ for all $\mathbf{z} \in 0^+\mathbf{P}_F$. Therefore $\mathcal{K} \cap \mathbb{R}^\Omega_+ \subseteq (0^+\mathbf{P}_F)^\circ \implies \mathcal{K}^\circ + \mathbb{R}^\Omega_- \supseteq 0^+\mathbf{P}_F$.

If F is submodular, w.l.o.g. we assume $F(\emptyset) \ge 0$. By Proposition 2.5 $\mathbf{p} \in \mathbf{P}_F \iff \exists \mathbf{b} \in \mathbf{B}_F$ such that $\mathbf{p} \le \mathbf{b}$. The rest follows from Theorem 4.2. ■

Edmonds [1970] showed that for submodular functions there is a simple greedy algorithm that computes the support functions (39) and (45).

Algorithm 4.1: Greedy algorithm for (39)

- Find a maximal increasing sequence $\emptyset = S_0 \subset S_1 \ldots \subset S_n = \Omega$ that contains each A_i in (44);
- Set $b^\star(S_i \setminus S_{i-1}) = F(S_i) - F(S_{i-1})$.

The algorithm is greedy in the sense that for any $\mathbf{b} \in \mathbf{B}_F$, by maximality $S_i \setminus S_{i-1} \in \mathfrak{L}$, hence $b(S_i \setminus S_{i-1}) \le F(S_i \setminus S_{i-1}) \le F(S_i) - F(S_{i-1}) + F(\emptyset)$—we go for the upper bound (but neglect the constant term $F(\emptyset)$).

Theorem 4.4: Correctness of Algorithm 4.1

Let $\mathfrak{L}$ be simple and $\mathbf{w}$ be monotonically decreasing from $(\Omega, \preceq)$ to $\mathbb{R}$. Algorithm 4.1 always returns a maximizer of (39) iff F is submodular and $F(\emptyset) \ge 0$.

Proof: Let $\mathbf{b}^\star$ be the solution returned by Algorithm 4.1 and $\mathbf{b} \in \mathbf{B}_F$ be arbitrary. Then

$$\begin{aligned}
\langle \mathbf{b} - \mathbf{b}^\star, \mathbf{w}\rangle &= \sum_{x\in\Omega} w_x(b_x - b^\star_x) = \sum_{i=1}^{k} w_i \sum_{x\in A_i\setminus A_{i-1}} (b_x - b^\star_x) \\
&= \sum_{i=1}^{k} w_i\big(\mathbf{b}(A_i) - \mathbf{b}(A_{i-1}) - \mathbf{b}^\star(A_i) + \mathbf{b}^\star(A_{i-1})\big) \\
&= \sum_{i=1}^{k} w_i\big(\mathbf{b}(A_i) - \mathbf{b}(A_{i-1}) - F(A_i) + F(A_{i-1})\big) \\
&= \sum_{i=1}^{k-1} (w_i - w_{i+1})\big(\mathbf{b}(A_i) - F(A_i)\big) \le 0.
\end{aligned} \tag{47}$$

Therefore we need only prove that $\mathbf{b}^\star \in \mathbf{B}_F$ iff F is submodular.

The correctness of Algorithm 4.1 implies $\mathbf{b}^\star \in \mathbf{B}_F$. For any incomparable $X, Y \in \mathfrak{L}$, let $\{S_i\}_{i=0}^n$

be a maximal increasing sequence that contains $X \cap Y$, $X \cup Y$ and of course $\{A_i\}_{i=1}^k$. Thus

$$F(X) + F(Y) \geq \mathbf{b}^\star(X) + \mathbf{b}^\star(Y) = \mathbf{b}^\star(X \cap Y) + \mathbf{b}^\star(X \cup Y) \geq F(X \cap Y) + F(X \cup Y).$$

For the reverse direction, we note that $F_{A_i}^{A_{i-1}}$ is defined on $\{X \setminus A_{i-1} : \mathfrak{L} \ni X \subseteq A_i\}$ and sends $Z = X \setminus A_{i-1}$ to $F(Z \cup A_{i-1}) - F(A_{i-1})$, see Theorem 3.8 and Remark 3.1. We claim that $\mathbf{b}^\star_{A_i \setminus A_{i-1}}$ is a base of $F_{A_i}^{A_{i-1}}$. Indeed, let $S_{j_0} = A_{i-1}$, $S_{j_{m+1}} = A_i$, and $Z = \{z_{j_1}, \ldots, z_{j_m}\}$ with $z_{j_t} = S_{j_t} \setminus S_{j_t - 1}, t = 1, \ldots, m$. Since $\mathfrak{L}$ is simple and $\{S_j\}_{j=0}^n$ is maximal, by Theorem 1.3 $|S_j \setminus S_{j-1}| = 1$ for all $1 \leq j \leq n$. Thanks to the submodularity of F, for all $m \geq t \geq 1$,

$$F(S_{j_0} \cup \{z_{j_1}, \ldots, z_{j_t}\}) + F(S_{j_t - 1}) \geq F(S_{j_0} \cup \{z_{j_1}, \ldots, z_{j_{t-1}}\}) + F(S_{j_t}).$$

Here notice that $S_{j_0} \cup \{z_{j_1}, \ldots, z_{j_t}\} \in \mathfrak{L}$ implies $S_{j_0} \cup \{z_{j_1}, \ldots, z_{j_{t-1}}\} \in \mathfrak{L}$. Thus summing from $t = m$ to $t = 1$ we obtain

$$F(S_{j_0} \cup \{z_{j_1}, \ldots, z_{j_m}\}) - F(S_{j_0}) \geq \sum_{t=1}^m F(S_{j_t}) - F(S_{j_t - 1}),$$

that is, $F(A_{i-1} \cup Z) - F(A_{i-1}) \geq \mathbf{b}^\star(Z)$. By definition $\mathbf{b}^\star(A_i \setminus A_{i-1}) = F(A_i) - F(A_{i-1})$. So we have proved that $\mathbf{b}^\star_{A_i \setminus A_{i-1}}$ is a base of $F_{A_i}^{A_{i-1}}$. Since $1 \leq i \leq n$ is arbitrary, we know from Remark 3.1 that $\mathbf{b}^\star \in \mathbf{B}_F$. ■

Theorem 4.5: Characterizing the maximizer by base decomposition

If $\mathfrak{L}$ is simple, $F : \mathfrak{L} \to \mathbb{R}$ is submodular, and $\mathbf{w}$ is monotonically decreasing from $(\Omega, \preceq)$ to $\mathbb{R}$, then the maximizers of (39) are given by

$$\mathbf{B}_{F_{A_1}} \oplus \mathbf{B}_{F_{A_2}^{A_1}} \oplus \cdots \oplus \mathbf{B}_{F_{A_k}^{A_{k-1}}} \oplus \mathbf{B}_{F^{A_k}}, \tag{48}$$

where the sets $\{A_i\}_{i=1}^k$ are defined in (44).

Proof: It is clear from the inequality (47) that any vector in the form of (48) is a maximizer for (39). On the other hand, if $\mathbf{b}$ is a maximizer, then for all $x \in A_i$ we have $\mathsf{dep}(\mathbf{b}, x) \subseteq A_i$, for otherwise $\mathbf{b} + \alpha(\mathbf{e}_x - \mathbf{e}_y)$ with any $y \in \mathsf{dep}(\mathbf{b}, x) - A_i$ and small positive α gives a larger objective than $\mathbf{b}$ (note that $w_x > w_y$). Thus $A_i = \bigcup_{x \in A_i} \mathsf{dep}(\mathbf{b}, x)$. By definition $\mathbf{b}(\mathsf{dep}(\mathbf{b}, x)) = F(\mathsf{dep}(\mathbf{b}, x))$ hence by Proposition 2.4 $\mathbf{b}(A_i) = f(A_i)$. Due to feasibility $\mathbf{b}$ is a base of F, and applying Remark 3.1 completes our proof. ■

Corollary 4.1: Uniqueness of maximizer

Let $\mathfrak{L}$ be simple and F be submodular. (39) has a unique solution iff $F(\emptyset) \geq 0$ and $\mathbf{w} \in \mathcal{K}$ is 1-1.

Proof: If $\mathbf{w}$ is not 1-1, then one of $\mathsf{B}_{F_{A_i}^{A_{i-1}}}, i = 1, \ldots, k$, will have multiple bases. ■

Corollary 4.2: Extreme points of base polyhedron

Let $\mathfrak{L}$ be simple and F be submodular. $\mathbf{b} \in \mathsf{B}_F$ is an extreme point iff $F(\emptyset) \geq 0$ and there is a maximal increasing sequence $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_n = \Omega$ in $\mathfrak{L}$ so that $b(S_i \setminus S_{i-1}) = F(S_i) - F(S_{i-1})$ for all $i = 1, \ldots, n$.

Proof: If $\mathbf{b}$ is an extreme point, then any of its supporting hyperplanes will give us the desired maximal increasing sequence. On the other hand, given the maximal sequence, we construct a 1-1 $\mathbf{w}$ so that its associated sets $A_i := \{x \in \Omega : w_x \geq w_i\} = S_i$. According to Corollary 4.1 $\mathbf{b}$ is a unique solution hence an extreme point. ■

Theorem 4.6: Characterizing the maximizer by local optimality

If $\mathfrak{L}$ is simple, $F : \mathfrak{L} \to \mathbb{R}$ is submodular with $F(\emptyset) \geq 0$, and $\mathbf{w}$ is monotonically decreasing from $(\Omega, \preceq)$ to $\mathbb{R}$, then $\mathbf{b} \in \mathbf{B}_F$ is a maximizer of (39) iff for all $x, y \in \Omega$, $y \in \mathsf{dep}(\mathbf{b}, x) \implies w_y \geq w_x$.

Proof: We have seen the necessity in the proof of Theorem 4.5: Simply consider $\mathbf{b} + \alpha(\mathbf{e}_x - \mathbf{e}_y)$. For the sufficiency, note that the given condition implies that for the set A_i defined in (44), we have $A_i = \bigcup_{x \in A_i} \mathsf{dep}(\mathbf{b}, x)$ hence $\mathbf{b}(A_i) = F(A_i)$ by Proposition 2.4. Appeal to Theorem 4.5. ■

Recall that the tangent cone at the point $\mathbf{b}$ in a set $C \subseteq \mathbb{R}^\Omega$ is defined as $\mathrm{cl}\left(\bigcup_{\lambda>0} \lambda(C - \{\mathbf{b}\})\right)$ while the normal cone is defined as $\{\mathbf{z} \in \mathbb{R}^\Omega : \langle \mathbf{z}, \mathbf{w} - \mathbf{b} \rangle \leq 0, \forall \mathbf{w} \in C\}$. Clearly the tangent cone and the normal cone are polar to each other.

Corollary 4.3: Tangent cone of base polyhedron

Let $\mathfrak{L}$ be simple and F be submodular with $F(\emptyset) \geq 0$, for any $\mathbf{b} \in \mathbf{B}_F$, its tangent cone is the conic combinations of $\mathbf{1}_x - \mathbf{1}_y$ for all $x \in \Omega, y \in \mathsf{dep}(\mathbf{b}, x)$.

Proof: From convex analysis we know that $\mathbf{b}$ is optimal for (39) iff $\mathbf{w}$ is in the normal cone of $\mathbf{b}$, while from Theorem 4.6 we know $\mathbf{b}$ is optimal iff $\mathbf{w}$ is in the polar of the cone spanned by $\mathbf{1}_x - \mathbf{1}_y$ for all $x \in \Omega, y \in \mathsf{dep}(\mathbf{b}, x)$. ■

To discuss the monotonicity of a submodular function F defined on a simple lattice $\mathfrak{L}$, let us define for each $x \in \Omega$, $\mathcal{I}_x := \bigcap\{X \in \mathfrak{L} : x \in X\}$, i.e., the (union of the) principal ideal of $[x]$ in $(\mathcal{P}, \preceq)$, and let

$$\forall x \in \Omega,\ u_x := F(\mathcal{I}_x) - F(\mathcal{I}_x \setminus \{x\}). \tag{49}$$

Note that x is a maximal element in $\mathcal{I}_x$, hence $\mathcal{I}_x \setminus \{x\}$ is an ideal therefore belongs to $\mathfrak{L}$. Since F is submodular, for all $X \in \mathfrak{L}$ so that $x \in X$ and $X \setminus \{x\} \in \mathfrak{L}$, we have

$$F(X) - F(X \setminus \{x\}) \leq F(\mathcal{I}_x) - F(\mathcal{I}_x \setminus \{x\}) = u_x.$$

Thanks to Corollary 4.2, $\mathbf{u}$ is the least upper bound of all extreme points, hence all points, in $\mathbf{B}_F$. In particular, for all $X \in \mathfrak{L}$, $F(X) \leq \mathbf{u}(X)$ since there exists a base $\mathbf{b}$ with $\mathbf{b}(X) = F(X)$, see Proposition 3.2. Similarly, we define $\mathcal{I}^x := \bigcup\{X \in \mathfrak{L} : x \notin X\}$ and

$$\forall x \in \Omega,\ \ell_x := F(\mathcal{I}^x \cup \{x\}) - F(\mathcal{I}^x), \tag{50}$$

i.e., $\boldsymbol{\ell}$ is the greatest upper bound of all extreme points, hence all points, in $\mathbf{B}_F$.

Proposition 4.1: Characterizing monotonicity

Let $\mathfrak{L}$ be simple and $F : \mathfrak{L} \to \mathbb{R}$ be submodular, then F is (strictly) increasing iff $\boldsymbol{\ell}$ is (strictly) positive; (strictly) decreasing iff $\mathbf{u}$ is (strictly) negative.

Proof: We prove the if part for the decreasing case; others are similar. Let $X \subset Y$ both in $\mathfrak{L}$ and $\omega \in \Omega$ be a maximal element in $Y \setminus X$. Since X is an ideal, ω is in fact a maximal element of Y, thus $Y \setminus \{\omega\}$ is an ideal hence in $\mathfrak{L}$. Then $F(Y) - F(Y \setminus \{\omega\}) \leq u_\omega \leq 0$. Continue removing elements until we reach $Y = X$. ■

Proposition 4.2: Characterizing modularity

Let $\mathfrak{L}$ be simple and $F : \mathfrak{L} \to \mathbb{R}$ be centered submodular, then F is modular iff $\mathbf{B}_F$ has one and only one extreme point iff there exists a base $\mathbf{b} \in \mathbf{B}_F$ such that $\mathbf{b} = \boldsymbol{\ell}$ or $\mathbf{b} = \mathbf{u}$.

Proof: If $\mathbf{B}_F$ has exactly one extreme point $\mathbf{b}$, then for all $X \in \mathfrak{L}$, $F(X) = \sigma_F^\flat(\mathbf{1}_X) = \mathbf{b}(X)$, a modular function. On the other hand, if F is modular, any $\mathbf{b} \in \mathbf{B}_F$ coincides with $\boldsymbol{\ell}$ (or $\mathbf{u}$) by definition. ■

In fact, each centered modular function M can be identified with its unique extreme base $\mathbf{b}$ so that $M(X) = \mathbf{b}(X)$ for all $X \in \mathfrak{L}$.

Remark 4.2: Monotone-Modular decomposition

Here we show how to decompose a submodular function into the difference of two increasing submodular functions. First observe that the map $F \mapsto \boldsymbol{\ell}$, see (50), is clearly linear. Therefore, for all X in a simple lattice $\mathfrak{L}$,

$$F(X) \quad = \quad \underbrace{F(X) - M(X)}_{\text{increasing submodular}} \quad + \quad \underbrace{M(X)}_{\text{decreasing modular}} \quad , \tag{51}$$

where we pick any modular function M whose (unique) extreme base dominates the lower bound $\boldsymbol{\ell}$ of F. It is obvious that we can make both parts *strictly* monotonic. Additionally, if F is centered and integer valued, then the decomposition can be made of two polymatroid functions, that is, monotonically increasing centered submodular functions which take integral values.

Besides, if we are interested in minimizing F, then we know the minimizer may *not* be written as the union of two *disjoint* sets $X, Y \in \mathfrak{L}$ such that, say $\ell_y \geq 0$ for all $y \in Y$, after all $\ell_y \geq 0$ intuitively means that F is increasing along the "coordinate" y. Indeed, following the proof of Proposition 4.1 to remove each element in Y we can only decrease F. Note that it may well happen that the minimizer still consists of some coordinate y with $\ell_y \geq 0$, as long as we do not have a partition of the minimizer in $\mathfrak{L}$: take $\mathfrak{L} = \{\emptyset, \{1\}, \{1,2\}, \{1,2,3\}\}$ and consider $F = \{2,1,2,0\}$.

Remark 4.3: The greedy Algorithm 4.1 is primal-dual

The reasoning in the proof of Theorem 4.4 actually shows that the greedy Algorithm 4.1 is primal-dual. Indeed, consider the dual problem of (39):

$$\min_{\boldsymbol{\lambda}} \quad \sum_{X \in \mathfrak{L}} \lambda_X F(X) \tag{52}$$

$$\text{s.t.} \quad \forall \mathfrak{L} \ni X \neq \Omega, \quad \lambda_X \geq 0, \tag{53}$$

$$\forall x \in \Omega, \quad w_x = \sum_{\mathfrak{L} \ni X \ni x} \lambda_X. \tag{54}$$

Let us set for each A_i in (44),

$$\lambda^\star_{A_i} = w_i - w_{i+1}, \tag{55}$$

with the understanding that $w_{k+1} = 0$, and for all other $X \in \mathfrak{L}$, $\lambda^\star_X = 0$. Clearly (53) is satisfied since $w_i > w_{i+1}$ by definition. Moreover, for some j, $w_x = w_j = \sum_{i \geq j}(w_i - w_{i+1}) = \sum_{\mathfrak{L} \ni X \ni x} \lambda^\star_X$, satisfying (54). Finally,

$$\langle \mathbf{b}^\star, \mathbf{w} \rangle = \sum_{i=1}^{k} w_i (\mathbf{b}^\star(A_i) - \mathbf{b}^\star(A_{i-1})) = \sum_{i=1}^{k} (w_i - w_{i+1}) F(A_i) = \sum_{X \in \mathfrak{L}} \boldsymbol{\lambda}^\star_X F(X). \tag{56}$$

Therefore by linear programming duality $\boldsymbol{\lambda}^\star$ defined in (55) is indeed a minimizer of the dual problem (52). In particular, if $\mathbf{w}$ is integral, so is the dual minimizer $\boldsymbol{\lambda}^\star$.

Remark 4.4: "Simplification" in the dual

The simplification we mentioned in Remark 4.1 extends to the dual problem (52)-(54). In fact, the dual problem of $\max\{\langle \mathbf{w}, \mathbf{b} \rangle : \mathbf{b} \in \mathbf{B}_F\}$, when finite, is exactly the dual of $\max\{\langle \tilde{\mathbf{w}}, \tilde{\mathbf{b}} \rangle : \tilde{\mathbf{b}} \in \mathbf{B}_{\tilde{F}}\}$, where $\tilde{F}$ is the simplification constructed in Remark 4.1.

The above primal-dual result is in fact way more far-reaching than it appears to be. We need some background first. Recall that the linear system $\mathfrak{P} := \{\mathbf{z} \in \mathbb{R}^\Omega : A_1\mathbf{z} \le \mathbf{s}_1, A_2\mathbf{z} = \mathbf{s}_2\}$ is called box totally dual integral (box-TDI) if for any $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^\Omega$ and any integral $\mathbf{c} \in \mathbb{Z}^\Omega$, the dual problem of $\max\{\langle \mathbf{c}, \mathbf{z}\rangle : \mathbf{z} \in \mathfrak{P}, \boldsymbol{\ell} \le \mathbf{z} \le \mathbf{u}\}$ always has an integral solution (whenever there exists one). The matrix A is called totally unimodular (TUM) if the determinants of all its submatrices are either 0 or ± 1. We will thoroughly discuss the many consequences of box-TDI and TUM later. For now, the following two propositions, also due to Edmonds [1970], are enough.

Proposition 4.3: TUM support in the dual implies box-TDI

Consider the linear system $\mathfrak{P} := \{\mathbf{z} \in \mathbb{R}^\Omega : A_1\mathbf{z} \le \mathbf{s}_1, A_2\mathbf{z} = \mathbf{s}_2\}$ for some matrix $A = [A_1, A_2]$ and vector $\mathbf{s} = [\mathbf{s}_1; \mathbf{s}_2]$. If for any $\mathbf{w} \in \mathbb{R}^\Omega$, the problem $\max\{\langle \mathbf{z}, \mathbf{w}\rangle : \mathbf{z} \in \mathfrak{P}\}$ has an optimal dual minimizer $\boldsymbol{\lambda}^\star = [\boldsymbol{\lambda}_1^\star; \boldsymbol{\lambda}_2^\star]$ such that the rows of A corresponding to the support of $\boldsymbol{\lambda}^\star$ form a totally unimodular submatrix, then the system $\mathfrak{P}$ is box totally dual integral.

Proof: Fix any $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^\Omega$, $\mathbf{c} \in \mathbb{Z}^\Omega$ and consider the primal problem

$$\max\{\langle \mathbf{z}, \mathbf{c}\rangle : \mathbf{z} \in \mathfrak{P}, \boldsymbol{\ell} \le \mathbf{z} \le \mathbf{u}\},$$

whose dual is

$$\min\{\langle \boldsymbol{\lambda}_1, \mathbf{s}_1\rangle + \langle \boldsymbol{\lambda}_2, \mathbf{s}_2\rangle - \langle \boldsymbol{\alpha}, \boldsymbol{\ell}\rangle + \langle \boldsymbol{\beta}, \mathbf{u}\rangle : A_1^\top \boldsymbol{\lambda}_1 + A_2^\top \boldsymbol{\lambda}_2 - \boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{c}, \boldsymbol{\lambda}_1, \boldsymbol{\alpha}, \boldsymbol{\beta} \ge \mathbf{0}\}. \tag{57}$$

Fix any minimizer $\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star \ge \mathbf{0}$, (57) is the dual problem of

$$\max\{\langle \mathbf{z}, \mathbf{c} + \boldsymbol{\alpha}^\star - \boldsymbol{\beta}^\star\rangle - \langle \boldsymbol{\alpha}^\star, \boldsymbol{\ell}\rangle + \langle \boldsymbol{\beta}^\star, \mathbf{u}\rangle : \mathbf{z} \in \mathfrak{P}\}.$$

Therefore by assumption the minimizer $\boldsymbol{\lambda}^\star = [\boldsymbol{\lambda}_1^\star; \boldsymbol{\lambda}_2^\star]$ in (57) can be chosen so that the rows of $A = [A_1, A_2]$ corresponding to the support of $\boldsymbol{\lambda}^\star$ form a TUM submatrix. Call the submatrix $\tilde{A} = [\tilde{A}_1, \tilde{A}_2]$ and denote $\tilde{\boldsymbol{\lambda}}_i$ as the restriction of $\boldsymbol{\lambda}_i$ to the support of $\boldsymbol{\lambda}_i^\star$, then $(\boldsymbol{\alpha}^\star, \beta^\star, \tilde{\boldsymbol{\lambda}}_1^\star, \tilde{\boldsymbol{\lambda}}_2^\star)$ is feasible for

$$\min\{\left\langle \tilde{\boldsymbol{\lambda}}_1, \tilde{\mathbf{s}}_1\right\rangle + \left\langle \tilde{\boldsymbol{\lambda}}_2, \tilde{\mathbf{s}}_2\right\rangle - \langle \boldsymbol{\alpha}, \boldsymbol{\ell}\rangle + \langle \boldsymbol{\beta}, \mathbf{u}\rangle : \tilde{A}_1^\top \tilde{\boldsymbol{\lambda}}_1 + \tilde{A}_2^\top \tilde{\boldsymbol{\lambda}}_2 - \boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{c}, \tilde{\boldsymbol{\lambda}}_1, \boldsymbol{\alpha}, \boldsymbol{\beta} \ge \mathbf{0}\}. \tag{58}$$

Clearly the constraints in (58) are TUM, thus it has an integral minimizer, which, after padding zero (in $\tilde{\boldsymbol{\lambda}}$), must also be optimal for (57) (since it has a smaller objective than $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, \boldsymbol{\lambda}^\star)$, a minimizer of (57)). Thus (57) has an integral minimizer, meaning that $\mathfrak{P}$ is box-TDI. ■

The proof, although straightforward, is a very useful technique and will re-appear a few times later. We remark that box-TDI is a property of the system parameterized by A and $\mathbf{s}$. It is *not* entirely, although close to, a property of the underlying polyhedron.

A collection of subsets of Ω , say $\mathcal{F}$, is called a laminar system if for all $X \in \mathcal{F}, Y \in \mathcal{F}$ we have either $X \subseteq Y$ or $Y \subseteq X$ or $X \cap Y = \emptyset$.

Proposition 4.4: Incidence of two laminar systems is TUM

Let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ be the union of two laminar systems $\mathcal{F}_1, \mathcal{F}_2$, the incidence matrix $A \in \mathbb{R}^{\Omega \times \mathcal{F}}$, defined as $A_{\omega,X} = 1$ if $\omega \in X$ and 0 otherwise, is totally unimodular.

Proof: Let B be any square submatrix of A. Note that a subset of a laminar system is clearly laminar. Pick any column of B that is indexed by the set X in, say $\mathcal{F}_1$. For all other columns of B that are indexed also by sets in $\mathcal{F}_1$, we replace them with the set difference by X. This does not change the determinant of B and the system remains laminar. Repeatedly, we can assume the sets in $\mathcal{F}_1$ and $\mathcal{F}_2$ are pairwise disjoint, respectively. Therefore each row of B contains at most two 1's. If there exists a row with all 0's, then $\det(B) = 0$, while if there exists a row with a singe 1, we perform induction. Finally if every row has two 1's, then they must each come from $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Hence the sum of the columns indexed by $\mathcal{F}_1$ equals the sum of the columns indexed by $\mathcal{F}_2$, again $\det(B) = 0$. ■

This proposition is clearly false for more than two laminar systems: Take $\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$ whose determinant is -2.

Theorem 4.7: (Sub)base system is box-TDI

The (sub)base system of a submodular function is box totally dual integral.

Proof: Thanks to Remark 4.4, we can assume the underlying lattice is simple. According to Remark 4.3, the dual problem (55) admits a minimizer $\boldsymbol{\lambda}^\star$ whose support is an increasing sequence in $\mathfrak{L}$, in particular, a laminar system. As a simple consequence of Proposition 4.4, we verify that the rows corresponding to the support form a TUM submatrix. Apply Proposition 4.3.

Thanks to Theorem 4.3, the same conclusion for the subbase polyhedron is clear. ■

Theorem 4.7 can be further improved, after all we have not used the full power of Proposition 4.4. Note that we do not discuss the intersection of two base systems since it can easily be empty, unless the functions coincide at the ground set Ω, in which case Theorem 4.8 and Corollary 4.4 continue to hold with essentially the same proof.

Theorem 4.8: Intersection of two subbase systems is box-TDI

The intersection of the subbase systems of two submodular/supermodular functions (not necessarily defined on the same lattice) is box totally dual integral.

Proof: We prove the case for two submodular functions; the other cases are completely analogous (recall that the subbase system for a supermodular function is defined by reversing the inequalities, without taking negation of the set).

The dual problem is

$$\min\left\{\sum_{i=1}^{2}\sum_{X\in\mathfrak{L}_i}\boldsymbol{\lambda}_i(X)F_i(X) : \forall X\in\mathfrak{L}_i, \boldsymbol{\lambda}_i(X)\geq 0, \forall x\in\Omega,\ \sum_{i=1}^{2}\sum_{x\in X\in\mathfrak{L}_i}\boldsymbol{\lambda}_i(X)=w_x\right\}. \tag{59}$$

Fix any minimizer $\boldsymbol{\lambda}_i^\star$ of (59), and let $\alpha_i=\sum_{x\in X\in\mathfrak{L}_i}\boldsymbol{\lambda}_i^\star(X)$. Consider the restricted subproblem for $i=1,2$:

$$\min\left\{\sum_{X\in\mathfrak{L}_i}\boldsymbol{\lambda}_i(X)F_i(X) : \forall X\in\mathfrak{L}_i, \boldsymbol{\lambda}_i(X)\geq 0, \forall x\in\Omega,\ \sum_{x\in X\in\mathfrak{L}_i}\boldsymbol{\lambda}_i(X)=\alpha_{3-i}\right\}. \tag{60}$$

As before, we assume w.l.o.g. that the lattices $\mathfrak{L}_i$ are simple. Since F_i is submodular, from Remark 4.3 we can choose a minimizer $\tilde{\boldsymbol{\lambda}}_i$ for (60) whose support is an increasing sequence in $\mathfrak{L}_i$. Clearly $(\tilde{\boldsymbol{\lambda}}_1,\tilde{\boldsymbol{\lambda}}_2)$ is also optimal for the dual problem (59), proving the existence of a dual minimizer whose support is the union of two increasing sequences. Apply Proposition 4.3 and Proposition 4.4. ■

Corollary 4.4: Min-max duality

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ and $G : \mathfrak{L}_2 \to \mathbb{R}$ be centered submodular, then for any $X \in \mathfrak{L}_1 \vee \mathfrak{L}_2$ we have

$$\max\{\mathbf{p}(X) : \mathbf{p} \in \mathbf{P}_F \cap \mathbf{P}_G\} = (F \boxdot G)(X). \tag{61}$$

Moreover, if F and G are integer valued, then we can choose an integral maximizer on the LHS.

Proof: Clearly LHS is equal to the dual problem (59), with $\mathbf{w} = \mathbf{1}_X$. According to Theorem 4.8, we have an integral minimizer for the dual, but this can only happen when $\boldsymbol{\lambda}_1(Y) = 1$ and $\boldsymbol{\lambda}_2(X \setminus Y) = 1$ for some $Y \in \mathfrak{L}_1, X \setminus Y \in \mathfrak{L}_2, Y \subseteq X$ and $\boldsymbol{\lambda}_i$ vanish on all other sets.

When F, G are integer valued, the box-TDI of the system $\mathbf{P}_F \cap \mathbf{P}_G$ implies it is integral. ■

The result is not entirely trivial since $F \boxdot G$ need not be submodular.

We need a handy construction that sends a submodular function to a supermodular one, and vice versa.

Definition 4.1: Negation

For any centered function $F : \mathfrak{L} \to \mathbb{R}$, we associate it with the negated function $F^\neg : \bar{\mathfrak{L}} \to \mathbb{R}$, where

$$\bar{\mathfrak{L}} := \{\Omega \setminus X : X \in \mathfrak{L}\}, \tag{62}$$
$$F^\neg(Z) := F(\Omega) - F(\Omega \setminus Z). \tag{63}$$

One easily verifies that indeed F is submodular iff $F^\neg$ is supermodular, and F is supermodular iff $F^\neg$ is submodular.

Remark 4.5: Reversed order

As shown in Theorem 1.2, $\mathfrak{L}$ consists of all lower ideals of the underlying ordered set $(\mathcal{P}, \preceq)$. Not surprisingly, $\bar{\mathfrak{L}}$ consists of exactly all upper ideals of the same ordered set. Or we can treat $\bar{\mathfrak{L}}$ as the set of all lower ideals of the reversed ordered set $(\mathcal{P}, \tilde{\preceq})$, i.e., $\mathcal{X} \tilde{\preceq} \mathcal{Y} \iff \mathcal{X} \succeq \mathcal{Y}$. Therefore, if $\mathcal{K}$ is the isotonic cone of $\mathfrak{L}$, then the isotonic cone of $\tilde{\mathfrak{L}}$ is simply $-\mathcal{K}$.

Proposition 4.5: Double negation cancels

For any centered $F : \mathfrak{L} \to \mathbb{R}$, $\mathbf{B}_F = \mathbf{B}_{F^\neg}$ and $(F^\neg)^\neg = F$.

Proof: One first verifies that $\bar{\bar{\mathfrak{L}}} = \mathfrak{L}$, hence for $X \in \mathfrak{L}$, $(F^\neg)^\neg(X) = F^\neg(\Omega) - F^\neg(\Omega \setminus X) = F(\Omega) - F(\emptyset) - F(\Omega) + F(X) = F(X)$ since we assume $F(\emptyset) = 0$.

For any $\mathbf{b} \in \mathbf{B}_F$ and $X \in \mathfrak{L}, Z = \Omega \setminus X$, $\mathbf{b}(X) = \mathbf{b}(\Omega \setminus Z) = \mathbf{b}(\Omega) - \mathbf{b}(Z) \leq F(X) = F(\Omega) - F^\neg(Z)$. Since $F(\Omega) = \mathbf{b}(\Omega)$, we have $F^\neg(Z) \leq \mathbf{b}(Z)$. ■

The next theorem, due to Frank [1982], resembles a classic result in convex analysis, and the real difference lies in its integral part.

Theorem 4.9: Discrete sandwich theorem

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ be submodular and $G : \mathfrak{L}_2 \to \mathbb{R}$ be supermodular. If for all $X \in \mathfrak{L}_1 \wedge \mathfrak{L}_2$, $G(X) \leq F(X)$, then there exists $\mathbf{p}^\star \in \mathbb{R}^\Omega$ so that for all $X \in \mathfrak{L}_2$, $G(X) \leq \mathbf{p}^\star(X)$ while for all $X \in \mathfrak{L}_1$, $\mathbf{p}^\star(X) \leq F(X)$. Moreover, if F and G are integer valued, we can choose $\mathbf{p}^\star \in \mathbb{Z}^\Omega$.

Proof: By translation we can assume w.l.o.g. $F(\emptyset) = 0$ and $G(\emptyset) \leq 0$. Redefine $G(\emptyset) = 0$, and note that this does not ruin the supermodularity of G nor affect the theorem.

By Corollary 4.4 we have

$$\begin{aligned} \max\{\mathbf{p}(\Omega) : \mathbf{p} \in \mathbf{P}_F \cap \mathbf{P}_{G^\neg}\} &= \min\{F(Y) + G^\neg(\Omega \setminus Y) : Y \in \mathfrak{L}_1, Y \subseteq \Omega, \Omega \setminus Y \in \bar{\mathfrak{L}}_2\} \\ &= \min\{F(Y) + G(\Omega) - G(Y) : Y \in \mathfrak{L}_1 \cap \mathfrak{L}_2\} \\ &= G(\Omega), \end{aligned} \tag{64}$$

since $F \geq G$ on $\mathfrak{L}_1 \wedge \mathfrak{L}_2$ and they are centered. Any maximizer for the LHS, say $\mathbf{p}^\star$ satisfies our requirement. Indeed, by feasibility, for all $X \in \mathfrak{L}_1$, $\mathbf{p}^\star(X) \leq F(X)$ and for all $X \in \mathfrak{L}_2$, $\mathbf{p}^\star(\Omega \setminus X) \leq G^\neg(\Omega \setminus X) = G(\Omega) - G(X)$, that is $\mathbf{p}^\star(X) \geq G(X)$ since $\mathbf{p}^\star(\Omega) = G(\Omega)$.

When F, G are integer valued, we can choose a integral $\mathbf{p}^\star$, according to Corollary 4.4. ■

Of course, we could have assumed $G \geq F$ on $\mathfrak{L}_1 \wedge \mathfrak{L}_2$ and get a similar result.

5 The Lovász Extension

Motivated by the greedy Algorithm 4.1, in this section we *extend* a set function $F : \mathfrak{L} \to \mathbb{R}$ to $f : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K} \subseteq \mathbb{R}^\Omega$ is the (anti)isotonic cone defined in (40). The result, generally known as the Lovász extension, appeared in Lovász [1982], but a more general theory was developed before by Choquet [1954]. The following convention will be adopted.

Remark 5.1: Centering and simplification

If $\mathfrak{L}$ is not simple, we first identify $(F, \mathfrak{L})$ with its "simplification" $(\tilde{F}, \tilde{\mathfrak{L}})$, which is defined in Remark 4.1. In case $F(\emptyset) \neq 0$, we first translate $F - F(\emptyset)$, compute the Lovász extension, and finally add $F(\emptyset)$ back to the extension.

For any $\mathbf{w} \in \mathcal{K}$, as before we identify its unique elements as $w_1 > w_2 > \cdots > w_k$ and define $A_i := \{x \in \Omega : w_x \geq w_i\}$. Then we have the decomposition

$$\mathbf{w} = \sum_{i=1}^{k} \lambda_i \cdot \mathbf{1}_{A_i}, \tag{65}$$

where, with the understanding $w_{k+1} = 0$,

$$\lambda_i := w_i - w_{i+1} \tag{66}$$

is uniquely determined, thanks to the monotonicity of $\mathbf{w}$. Next, we define the Lovász extension of F as

$$f(\mathbf{w}) = \sum_{i=1}^{k} \lambda_i \cdot F(A_i) = \sum_{i=1}^{k} (w_i - w_{i+1}) \cdot F(A_i) = \sum_{i=1}^{k} w_i \cdot \big(F(A_i) - F(A_{i-1})\big), \tag{67}$$

with the understanding that $F(A_0) = F(\emptyset) = 0$. Clearly, f is a positively homogeneous real-valued function defined on the isotonic cone $\mathcal{K}$. More importantly, f is an "extension" of F in the following sense.

Theorem 5.1: Lovász extension is an extension

Consider $F : \mathfrak{L} \to \mathbb{R}$ and f its Lovász extension. For all $X \in \mathfrak{L}$, $f(\mathbf{1}_X) = F(X)$.

Proof: We first show that $\mathbf{1}_X \in \mathcal{K}$ for all $X \in \mathfrak{L}$. Indeed, fix any $y \preceq x$. If $x \in X$, then $y \in X$ since each X in the simple lattice $\mathfrak{L}$ is an ideal. Thus $\mathbf{1}_X(y) \geq \mathbf{1}_X(x)$, which is also true when $x \notin X$. Finally, observing that for $\mathbf{w} = \mathbf{1}_X$, $\lambda_i = 0, \forall i < k$, $\lambda_k = 1$, and $A_k = X$ completes the proof. ■

Theorem 5.2: Linearity

The map that sends $F : \mathfrak{L} \to \mathbb{R}$ to its Lovász extension $f : \mathcal{K} :\to \mathbb{R}$ is linear, 1-1 but not onto the space of positively homogeneous functions (unless $|\Omega| = 1$).

Proof: The linearity is clear while the 1-1 property follows from Theorem 5.1. Lastly, the space of all set functions from $\mathfrak{L}$ to $\mathbb{R}$ is of finite dimension while the space of all positively homogeneous $f : \mathcal{K} \to \mathbb{R}$ is of infinite dimension (whenever $|\Omega| \geq 2$). ■

An explicit counterexample: take $f(\mathbf{w}) = \max\{w_1, w_2 + w_3\}$. It is not clear to us what is the Lovász extension of the convolution.

Similarly, the map that sends a submodular function to its Lovász extension is *not* onto the space of all sublinear functions. A complete characterization of the Lovász extension will be given in Corollary 6.2 below. For now, let us observe that we do have a nice 1-1 correspondence for modular functions. In the following we call the function $\ell : \mathcal{K} \to \mathbb{R}$ linear if for all $\alpha, \beta \geq 0$, $\mathbf{w}, \mathbf{z} \in \mathcal{K}$, we have $\ell(\alpha\mathbf{w} + \beta\mathbf{z}) = \alpha\ell(\mathbf{w}) + \beta\ell(\mathbf{z})$.

Theorem 5.3: Modularity corresponds to linearity

The Lovász extension of any centered modular function $M : \mathfrak{L} \to \mathbb{R}$, is linear. Conversely, any linear function on $\mathcal{K}$ is the Lovász extension of such a modular function M.

Proof: Clearly if M is modular, it is both submodular and supermodular, hence by Theorem 5.5 below its Lovász extension is both convex and concave, i.e., linear. Conversely, a linear function $\ell : \mathcal{K} \to \mathbb{R}$ is uniquely determined by its values on $\{\mathbf{1}_X : X \in \mathfrak{L}\}$ through (65). For any $X \in \mathfrak{L}$, define $M(X) := \ell(\mathbf{1}_X)$, we easily verify that M is modular and ℓ is its Lovász extension. ■

As a sanity check, we note that the space of linear functions on $\mathcal{K}$ is of finite dimension, hence the possibility of Theorem 5.3. Besides, we can easily extend ℓ from the cone $\mathcal{K}$ to the linear subspace $\mathcal{K} - \mathcal{K} = \mathbb{R}^\Omega$ (recall that $\mathcal{K}$ is generating), thus the Lovász extension of a modular function is given by $\langle \cdot, \mathbf{z} \rangle$ for some $\mathbf{z} \in \mathbb{R}^\Omega$ being the unique extreme base (of the base polyhedron). The identification by $\mathbf{z}$ is clearly unique. Interestingly, this immediately implies that the space of all centered modular functions defined on any (finite distributive) simple lattice is isomorphic to $\mathbb{R}^\Omega$—a result which does not seem to be entirely trivial (although it is indeed trivial for Boolean algebras, see (7)).

The next a few results should convince us the importance of the Lovász extension.

Theorem 5.4: Lovász extension preserves the minimum

Let $F : \mathfrak{L} \to \mathbb{R}$ be any function and f its Lovász extension, then

$$\min_{X \in \mathfrak{L}} F(X) \quad = \quad \min_{\mathbf{w} \in C} f(\mathbf{w}), \tag{68}$$

where $C = \mathcal{K} \cap [0,1]^\Omega$ is the convex hull of $\{\mathbf{1}_X : X \in \mathfrak{L}\}$.

Proof: Following Remark 5.1 we assume w.l.o.g. that F is centered. Thanks to Theorem 5.1, we clearly have $\mathsf{LHS} \geq \mathsf{RHS}$ in (68). On the other hand,

$$\mathsf{RHS} = \min_{\mathbf{w} \in C} \sum_{i=1}^{k} (w_i - w_{i+1}) F(A_i) \geq \min_{\mathbf{w} \in C} \sum_{i=1}^{k} (w_i - w_{i+1}) \cdot \mathsf{LHS} = \min_{\mathbf{w} \in C} w_1 \cdot \mathsf{LHS} \geq \mathsf{LHS},$$

since $\mathsf{LHS} \leq 0$. To see C is the convex hull of $\{\mathbf{1}_X : X \in \mathfrak{L}\}$, first note that C clearly is bigger. However, any $\mathbf{w} \in C$ can be written uniquely as in (65), where λ_i is given in (66) and satisfies $\lambda_i \geq 0, \sum_i \lambda_i = w_1 \leq 1$. Since $\emptyset \in \mathfrak{L}$, $\mathbf{w}$ is in the convex hull of $\{\mathbf{1}_X : X \in \mathfrak{L}\}$. ■

Theorem 5.5: Submodularity is equivalent to convexity

Let $F : \mathfrak{L} \to \mathbb{R}$ be centered, then F is submodular iff its Lovász extension f is convex iff the support function $\sigma^\flat_F = f$.

Proof: Firstly, $f = \sigma^\flat_F$ implies f is convex, since the support function is always convex. Secondly, if f is convex, then it is subadditive due to its built-in positive homogeneity. Thus for all $X, Y \in \mathfrak{L}$, using Theorem 5.1,

$$F(X) + F(Y) = f(\mathbf{1}_X) + f(\mathbf{1}_Y) \geq f(\mathbf{1}_X + \mathbf{1}_Y) = f(\mathbf{1}_{X \cup Y} + \mathbf{1}_{X \cap Y}) = F(X \cup Y) + F(X \cap Y),$$

where the last equality follows from (67). Lastly, if F is submodular, according to Remark 4.3 we know $f = \sigma^\flat_F$. ■

Corollary 5.1: Lovász extension is d.c.

The Lovász extension of any $F : \mathfrak{L} \to \mathbb{R}$ is the difference of two positively homogeneous convex functions.

Proof: By Theorem 3.2 any set function F can be written as the difference of two submodular functions, and the rest follows from Theorem 5.2 and Theorem 5.5. ■

Of course, the decomposition need not be unique, even for a submodular F.

Remark 5.2: Algorithmic consequence of the Lovász extension

Thanks to Theorem 5.4, we can turn the minimization of a submodular function into the minimization of its Lovász extension, which, by Theorem 5.5, is a convex program hence can be efficiently solved. Moreover, the definition of Lovász extension in (65) and (67) allows us to recover a minimizer of the original submodular function: any set $A_i, i = 0, \dots, k-1$ in the decomposition must be optimal.

6 The Choquet Integral

It turns out that the Lovász extension is a very special case of a much more general theory—the non-additive integration theory originated from Choquet [1954].

For this section only, Ω is an *arbitrary* nonempty set and $\mathfrak{L} \subseteq 2^\Omega$ always includes the empty set $\emptyset$ and the full set Ω. At first we consider a monotonically *increasing* set function $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ is the extended real line. With the additional assumption that $F(\emptyset) = 0$, such an increasing set function F will be called a capacity. A function $g : \Omega \to \overline{\mathbb{R}}$ is called $\mathfrak{L}$-measurable if for all $\gamma \in \mathbb{R}$ the upper level set $\{x \in \Omega : g(x) \geq \gamma\} \in \mathfrak{L}$. Not surprisingly, all of our results still hold if we consider instead the strict upper level sets $\{x \in \Omega : g(x) > \gamma\}$. Note that the two definitions coincide when $\mathfrak{L}$ is a σ-algebra. From now on we use the short hand $[\![g \geq \gamma]\!]$ for the upper level set. It is an easy exercise to verify that if g is $\mathfrak{L}$-measurable, then so are $g \wedge \mu$, $g \vee \mu$, $(g - \mu)_+$ and $\lambda g + \mu$ for all $\mu \in \mathbb{R}, \lambda \in \mathbb{R}_+$. Our initial goal is to build an integration theory for any $\mathfrak{L}$ and any capacity F. As usual, we start with nonnegative functions, in particular, step functions.

Definition 6.1: Choquet integral for nonnegative function

Fix any capacity $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$. For any $\mathfrak{L}$-measurable function $g : \Omega \to \overline{\mathbb{R}}_+$, its Choquet integral w.r.t. F is defined as

$$\int g \, \mathrm{d}F := \int_0^\infty F([\![g \geq \gamma]\!]) \, \mathrm{d}\gamma, \tag{69}$$

where the right integral is of the Riemann type. Since the integrand on the RHS is a decreasing function (of γ), the Choquet integral is well-defined and always nonnegative.

If F is a measure (on a σ-algebra $\mathfrak{L}$), then the Choquet integral coincides with the Lebesgue integral, since the RHS in (69) is nothing but the familiar "integrating the tail". The next result is immediate.

Proposition 6.1: Increasing and positively homogeneous

Let $g, h : \Omega \to \overline{\mathbb{R}}_+$ be $\mathfrak{L}$-measurable and $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be a capacity. Then $h \geq g \implies \int h \, \mathrm{d}F \geq \int g \, \mathrm{d}F$, and for all $\lambda \geq 0$, $\int (\lambda g) \, \mathrm{d}F = \lambda \cdot \int g \, \mathrm{d}F$. Moreover, if $\bar{F}$ is an extension of F, we have $\int g \, \mathrm{d}F = \int g \, \mathrm{d}\bar{F}$.

Abstract as it is, let us compute the Choquet integral for step functions, whose range is a finite set. We use $\mathcal{M} = \mathcal{M}(\mathfrak{L})$ and $\mathcal{S} = \mathcal{S}(\mathfrak{L})$ to denote $\mathfrak{L}$-measurable functions and step functions, respectively, with $(\cdot)_+$ signaling the nonnegative ones. For any $s \in \mathcal{S}_+$, we identify it with its range $\{0 \leq w_k < \cdots < w_1\}$ and the increasing sequence $A_i := [\![s \geq w_i]\!]$, $i = 1, \dots, k$. Clearly $s = \sum_{i=1}^k (w_i - w_{i+1}) \mathbf{1}_{A_i}$, where $w_{k+1} := 0$. We verify that $\int s \, \mathrm{d}F = \sum_{i=1}^k (w_i - w_{i+1}) F(A_i)$. Recall that this is exactly how we defined the Lovász extension in Section 5.

We can extend the centered set function $F : \mathfrak{L} \to \overline{\mathbb{R}}$ to the power set 2^Ω in the following way:

$$\forall \, X \subseteq \Omega, \quad F_\star(X) = \sup\{F(Y) : \mathfrak{L} \ni Y \subseteq X\}, \tag{70}$$

which is always increasing and nonnegative. Clearly any function is 2^Ω-measurable, hence $\int g\,\mathrm{d}F_\star$ is well-defined for all functions $g : \Omega \to \overline{\mathbb{R}}_+$. The next three results are fundamental.

Theorem 6.1: Regularity

Fix any capacity $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$. For all $g : \Omega \to \overline{\mathbb{R}}_+$ we have

$$\int g\,\mathrm{d}F_\star = \sup\left\{\int s\,\mathrm{d}F : g \geq s \in \mathcal{S}_+(\mathfrak{L})\right\}. \tag{71}$$

Proof: Due to monotonicity, it is clear that LHS $\geq$ RHS. For the converse, suppose first that $F_\star([\![g \geq \gamma]\!]) = \infty$ for some γ. Then for each $\alpha \geq 0$, there exists $\mathfrak{L} \ni A \subseteq [\![g \geq \gamma]\!]$ so that $\gamma F(A) \geq \alpha$. Clearly $\gamma \mathbf{1}_A \in \mathcal{S}_+$, thus RHS $\geq \int \gamma \mathbf{1}_A\,\mathrm{d}F = \gamma F(A) \geq \alpha$. Since α is arbitrary, we have RHS $= \infty$ and there is nothing to prove.

Now fix $0 < a < b < \infty$ and $\epsilon > 0$. By the definition of Riemann integral, there exists a subdivision $a = t_{k+1} < t_k < \ldots < t_1 = b$ such that

$$\int_a^b F_\star([\![g \geq \gamma]\!])\,\mathrm{d}\gamma \leq \epsilon + \sum_{i=1}^k (t_i - t_{i+1}) F_\star([\![g \geq t_i]\!]),$$

and for each $1 \leq i \leq k$, there exists a chain $\mathfrak{L} \ni A_i \subseteq [\![g \geq t_i]\!]$ such that $F_\star([\![g \geq t_i]\!]) \leq F(A_i) + \epsilon/(b-a)$. Note that we can assume w.l.o.g. that $F_\star([\![g \geq \gamma]\!]) < \infty$ for all $\gamma \in \mathbb{R}$. Therefore

$$\int_a^b F_\star([\![g \geq \gamma]\!])\,\mathrm{d}\gamma \leq 2\epsilon + \sum_{i=1}^k (t_i - t_{i+1}) F(A_i) = 2\epsilon + \int \overbrace{\sum_{i=1}^k (t_i - t_{i+1}) \mathbf{1}_{A_i}}^{\leq g}\,\mathrm{d}F.$$

Taking limits we are done. ∎

Note that $F_\star$ extends F due to the monotonicity of the latter. Of course, if g is $\mathfrak{L}$-measurable, we can replace $F_\star$ in (71) with F, thanks to the observation made in Proposition 6.1.

Recall that a set function $F : \mathfrak{L} \to \overline{\mathbb{R}}$ is called continuous from below if $\mathfrak{L} \ni A_i \uparrow A \in \mathfrak{L}$ and $F(A_i) > -\infty \implies \lim_i F(A_i) \uparrow F(A)$, and similarly it is called continuous from above if $\mathfrak{L} \ni A_i \downarrow A \in \mathfrak{L}$ and $F(A_i) < \infty \implies \lim_i F(A_i) \downarrow F(A)$.

Theorem 6.2: Monotone convergence

Let the capacity $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be continuous from below (above, resp.) and $\mathcal{M}_+(\mathfrak{L}) \ni g_n \uparrow (\downarrow$, resp.$)\ g \in \mathcal{M}_+(\mathfrak{L})$ (with $\int g_n\,\mathrm{d}F < \infty$, resp.). Then

$$\lim_n \int g_n\,\mathrm{d}F = \int g\,\mathrm{d}F. \tag{72}$$

Proof: Simply note that, for example, $g_n \downarrow g \implies [\![g_n \geq \gamma]\!] \downarrow [\![g \geq \gamma]\!]$ hence $F([\![g_n \geq \gamma]\!]) \downarrow F([\![g \geq \gamma]\!])$, thanks to the continuity of F and the fact that $\int g_n\,\mathrm{d}F < \infty \implies \gamma F([\![g \geq \gamma]\!]) < \infty$. Apply the usual monotone convergence theorem for Riemann's integral. ∎

It is clear that we can replace all sequences with nets.

When the capacity F is submodular (on a lattice $\mathfrak{L}$), resorting to the definition we have the inequality

$$\int (g \wedge h)\,\mathrm{d}F + \int (g \vee h)\,\mathrm{d}F \leq \int g\,\mathrm{d}F + \int h\,\mathrm{d}F, \tag{73}$$

which is reversed if F is supermodular.

Proposition 6.2: Choquet [1954, Theorem 54.1]

Let $\mathcal{K}$ be a lattice convex cone and $\Gamma : \mathcal{K} \to \mathbb{R}$ be positively homogeneous. Then

$$\forall\, g, h \in \mathcal{K}, \Gamma(g \vee h) + \Gamma(g \wedge h) \leq \Gamma(g) + \Gamma(h) \implies \Gamma(g+h) \leq \Gamma(g) + \Gamma(h). \tag{74}$$

Proof: This result is not entirely true. Repair it later (see Marinacci and Montrucchio [2008] and König [2003]). ■

For the Choquet integral, we shall provide a direct proof.

Theorem 6.3: Subadditive = submodular, Choquet [1954, Theorem 54.2]

Let $\mathfrak{L}$ be a lattice and $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be a capacity, then the Choquet integral is subadditive (superadditive) iff F is submodular (supermodular).

Proof: Suppose first that the Choquet integral is subadditive. Let $A, B \in \mathfrak{L}$, then

$$F(A\cap B)+F(A\cup B) = \int (\mathbf{1}_{A\cap B}+\mathbf{1}_{A\cup B})\,\mathrm{d}F = \int (\mathbf{1}_A+\mathbf{1}_B)\,\mathrm{d}F \leq \int \mathbf{1}_A\,\mathrm{d}F+\int \mathbf{1}_B\,\mathrm{d}F = F(A)+F(B).$$

Conversely, suppose F is submodular. Clearly we can assume w.l.o.g. that $\int g\,\mathrm{d}F < \infty$ and $\int h\,\mathrm{d}F < \infty$. Consequently (assuming $g + h \in \mathcal{M}_+$)

$$\begin{aligned}\int (g+h)\,\mathrm{d}F &= \int_0^\infty F([\![g+h \geq \gamma]\!])\,\mathrm{d}\gamma \leq \int_0^\infty F\Big([\![g \geq \gamma/2]\!] \cup [\![h \geq \gamma/2]\!]\Big)\,\mathrm{d}\gamma \\ &\leq \int_0^\infty F([\![g \geq \gamma/2]\!])\,\mathrm{d}\gamma + \int_0^\infty F([\![h \geq \gamma/2]\!])\,\mathrm{d}\gamma < \infty.\end{aligned}$$

Note that for any $A \in \mathfrak{L}$ and $g \in \mathcal{M}_+$, $g + \mathbf{1}_A \in \mathcal{M}_+$ since $\mathfrak{L}$ is a lattice. Consider first the step function $s \in \mathcal{S}_+$ with $s_{\max}$ its maximal value. Then

$$\begin{aligned}\int (s+\mathbf{1}_A)\,\mathrm{d}F &= \int_0^1 F(A \cup [\![s \geq \gamma]\!])\,\mathrm{d}\gamma + \int_1^{s_{\max}} F\Big([\![s \geq \gamma]\!] \cup ([\![s \geq \gamma - 1]\!] \cap A)\Big)\,\mathrm{d}\gamma \\ &\leq \int_0^1 \Big(F(A) + F([\![s \geq \gamma]\!]) - F([\![s \geq \gamma]\!] \cap A)\Big)\,\mathrm{d}\gamma + \int_1^{s_{\max}} \Big(F([\![s \geq \gamma]\!]) \\ &\qquad + F([\![s \geq \gamma - 1]\!] \cap A) - F([\![s \geq \gamma]\!] \cap A)\Big)\,\mathrm{d}\gamma \\ &\leq F(A) + \int_0^{s_{\max}} F([\![s \geq \gamma]\!])\,\mathrm{d}\gamma = \int s\,\mathrm{d}F + \int \mathbf{1}_A\,\mathrm{d}F.\end{aligned}$$

Now let $g \in \mathcal{M}_+$. According to Theorem 6.1 we can find $s \in \mathcal{S}_+$ so that $\int (g + \mathbf{1}_A)\,\mathrm{d}F \leq \epsilon + \int s\,\mathrm{d}F$ and $s \leq g + \mathbf{1}_A$. Clearly $t := s \vee \mathbf{1}_A \leq g + \mathbf{1}_A$ hence $0 \leq t - \mathbf{1}_A \leq g$. Moreover $t - \mathbf{1}_A \in \mathcal{S}_+$. Thus

$$\int (g+\mathbf{1}_A)\,\mathrm{d}F \leq \epsilon + \int s\,\mathrm{d}F \leq \epsilon + \int t\,\mathrm{d}F \leq \epsilon + \int (t-\mathbf{1}_A)\,\mathrm{d}F + \int \mathbf{1}_A\,\mathrm{d}F \leq \epsilon + \int g\,\mathrm{d}F + \int \mathbf{1}_A\,\mathrm{d}F.$$

Iterating the argument for each component of a step function we can replace $\mathbf{1}_A$ in the above inequality with any step function s. Finally applying a similar approximation we can replace the step function s with any $h \in \mathcal{M}_+$ (such that $g + h \in \mathcal{M}_+$). ■

In general, the Choquet integral is *not* even subadditive, which is not too surprising, after all we are dealing with a general capacity F. When F is a charge (namely, finitely additive) on an algebra $\mathfrak{L}$, it is modular, thus we gain additivity for the Choquet integral. When F is a measure (namely, countably additive) on a σ-algebra, the Choquet integral reduces to the Lebesgue integral, which, of course, is additive.

The next notion is extremely important.

Definition 6.2: Comonotone

Two functions $g, h : \Omega \to \overline{\mathbb{R}}$ are called comonotone iff there are no pairs $x, y \in \Omega$ such that

$$g(x) < g(y) \text{ and } h(x) > h(y). \tag{75}$$

When g, h are real valued, the above definition can be written as for all $x, y \in \Omega$,

$$\big(g(x) - g(y)\big) \cdot \big(h(x) - h(y)\big) \geq 0,$$

hence the name comonotone.

Recall that a subset of an ordered set is a chain if it is totally ordered, i.e., all pairs are comparable.

Theorem 6.4: Characterizing comonotonicity, Denneberg [1994]

Let $g, h : \Omega \to \overline{\mathbb{R}}$ be $\mathfrak{L}$-measurable. The following are equivalent:

(I). g, h are comonotone;

(II). The set $\{[\![g \geq \gamma]\!] : \gamma \in \mathbb{R}\} \cup \{[\![h \geq \gamma]\!] : \gamma \in \mathbb{R}\}$ is a chain;

(III). The set $\{\big(g(x), h(x)\big) : x \in \Omega\} \subseteq \mathbb{R}^2$ is a chain;

If g, h are real valued, then we have the additional equivalence:

(IV). There exists $f : \Omega \to \mathbb{R}$ and increasing functions $\mathsf{i}, \mathsf{j} : \mathbb{R} \to \mathbb{R}$ such that $g = \mathsf{i} \circ f, h = \mathsf{j} \circ f$;

(V). There exists continuous increasing functions $\mathsf{i}, \mathsf{j} : \mathbb{R} \to \mathbb{R}$ such that $\mathsf{i} + \mathsf{j} = \mathsf{Id}$ and $g = \mathsf{i} \circ (g + h), h = \mathsf{j} \circ (g + h)$.

Proof: (I) $\iff$ (II) is easily proved from considering its contrapositive. The implications (I) $\iff$ (III), (V) $\implies$ (IV) $\implies$ (I) are clear. Only (I) $\implies$ (V) is left.

Note first that for $t = (g + h)(z)$ with some $z \in \Omega$, we can decompose it into $t = p + q$ with $p \in g(\Omega)$ and $q \in h(\Omega)$. From comonotonicity, we know p and q are uniquely determined by t. Define $\mathsf{i}(t) = p$ and $\mathsf{j}(t) = q$ we satisfy the identity $\mathsf{i} + \mathsf{j} = \mathsf{Id}$. By comonotonicity again, we know both i and j are increasing, which in turn implies their continuity: for example, for $\delta \geq 0$, $\mathsf{i}(t) \leq \mathsf{i}(t + \delta) = t + \delta - \mathsf{j}(t + \delta) \leq t + \delta - \mathsf{j}(t) = \mathsf{i}(t) + \delta$. So we have constructed i and j on $(g+h)(\Omega)$. Extend to the closure $\mathrm{cl}((g+h)(\Omega))$ by continuity. For $\mathbb{R} - \mathrm{cl}((g+h)(\Omega))$, we apply linear interpolation in each of its connected components. The extensions maintain the identity $\mathsf{i} + \mathsf{j} = \mathsf{Id}$, as well as the monotonicity and continuity. ■

This theorem is very handy in verifying comonotonicity (for real valued functions). For instance, from (IV) it is clear that $g^+ := g \vee 0$ and $-g^- := g \wedge 0$ are comonotone (which also pleasantly fits (V)). Similarly using (IV), if g and h are comonotone, so is g and $g + h$.

Theorem 6.5: Comonotone additivity, Bassanezi and Greco [1984, Theorem 2.1]

Let g, h and $g + h$ belong to $\mathcal{M}_+$. Then

$$\text{for all capacity } F : \mathfrak{L} \to \overline{\mathbb{R}}_+, \quad \int (g + h)\, \mathrm{d}F = \int g\, \mathrm{d}F + \int h\, \mathrm{d}F \iff g, h \text{ comonotone}.$$

Proof: $\Longleftarrow$: Let $\mathring{\mathfrak{L}} := \{[\![g \geq \gamma]\!] : \gamma \in \mathbb{R}\} \cup \{[\![h \geq \gamma]\!] : \gamma \in \mathbb{R}\}$. Since f, g are comonotone, by Theorem 6.4, $\mathring{\mathfrak{L}}$ is a chain hence trivially a lattice and the restriction of F to $\mathring{\mathfrak{L}}$, $\mathring{F}$, is modular. Thus using Theorem 6.3 we have

$$\int g\, \mathrm{d}F + \int h\, \mathrm{d}F = \int g\, \mathrm{d}\mathring{F} + \int h\, \mathrm{d}\mathring{F} = \int (g + h)\, \mathrm{d}\mathring{F} = \int (g + h)\, \mathrm{d}F,$$

where the last equality follows from Proposition 6.1.

$\Longrightarrow$: Suppose g and h are not comonotone, implying the existence of $\alpha \neq \beta$ such that there exists $x \in [\![g \geq \alpha]\!] - [\![h \geq \beta]\!]$ and $x \neq y \in [\![h \geq \beta]\!] - [\![g \geq \alpha]\!]$. Define the capacity $F : \mathfrak{L} \to \{0, 1\}$, $X \mapsto \mathbf{1}_{\{x,y\} \subseteq X}$. In fact, F is supermodular. An easy calculation reveals that $\int g \, \mathrm{d}F < \alpha, \int h \, \mathrm{d}F < \beta$ while $\int (g + h) \, \mathrm{d}F \geq \alpha + \beta$. ■

As an application, consider $g = \sum_{i=1}^{k} w_i \mathbf{1}_{A_i}$ for a chain $A_1 \subset \cdots \subset A_k$. Clearly the functions $\mathbf{1}_{A_i}$ are comonotone hence immediately we have $\int g \, \mathrm{d}F = \sum_{i=1}^{k} \int w_i \mathbf{1}_{A_i} \, \mathrm{d}F = \sum_{i=1}^{k} w_i F(A_i)$.

Alert 6.1: Non-existence interpretation

From now on the integrals we consider might not exist in pathological cases, due to the possibility of $\infty - \infty$. Instead of tediously repeating the existence assumption each time, we adopt the convention that if something does not make sense, the related result shall not be interpreted.

It is time to consider the integral of a real valued function. It turns out there are two different ways to proceed. One is familiar: we split g into $g_+ := g \vee 0$ and $g_- = g_+ - g$. This yields the symmetric integral of Šipoš [1979]:

$$\cdot\!\!\int\!\!\cdot \, g \, \mathrm{d}F := \int g_+ \, \mathrm{d}F - \int g_- \, \mathrm{d}F = \int_0^\infty F([\![g \geq \gamma]\!]) \, \mathrm{d}\gamma - \int_0^\infty F([\![-g \geq \gamma]\!]) \, \mathrm{d}\gamma. \tag{76}$$

Note that we now need $-g$ to be $\mathfrak{L}$-measurable too. Let $s = \sum_{i=1}^{k} (a_i - a_{i+1}) \mathbf{1}_{A_i} + \sum_{j=1}^{m} (b_j - b_{j+1}) \mathbf{1}_{B_j}$ where $0 = a_{k+1} < a_k < \cdots < a_1$, $b_1 < \cdots < b_m < b_{m+1} = 0$, and $A_i = [\![s \geq a_i]\!]$, $B_j = [\![s \leq b_j]\!]$, then

$$\cdot\!\!\int\!\!\cdot \, s \, \mathrm{d}F = \sum_{i=1}^{k} (a_i - a_{i+1}) F(A_i) + \sum_{j=1}^{m} (b_j - b_{j+1}) F(B_j). \tag{77}$$

The following result about the symmetric integral is useful to know.

Proposition 6.3: Basic properties of the symmetric integral

Let $\pm g, \pm h \in \mathcal{M}$ and $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ a capacity. Then

- $g \geq h \implies \cdot\!\int\!\cdot \, g \, \mathrm{d}F \geq \cdot\!\int\!\cdot \, h \, \mathrm{d}F$;
- $\forall \mu \in \mathbb{R}$, $\cdot\!\int\!\cdot \, (\mu g) \, \mathrm{d}F = \mu \cdot \cdot\!\int\!\cdot \, g \, \mathrm{d}F$;
- $\cdot\!\int\!\cdot \, g \, \mathrm{d}F = \sup\{\cdot\!\int\!\cdot \, s \, \mathrm{d}F : g \geq s \in \mathcal{S}\}$;
- $\forall \lambda \geq 0$, $\cdot\!\int\!\cdot \, g \, \mathrm{d}F = \cdot\!\int\!\cdot \, (g \wedge \lambda) \, \mathrm{d}F + \cdot\!\int\!\cdot \, (g - \lambda)_+ \, \mathrm{d}F$.

Proof: The first two claims are clear while the third follows from the definition and Theorem 6.1. We prove the last one:

$$\begin{aligned} \mathsf{RHS} &= \int (g \wedge \lambda)_+ \, \mathrm{d}F - \int (g \wedge \lambda)_- \, \mathrm{d}F + \int (g - \lambda)_+ \, \mathrm{d}F \\ &= \int \Big((g \wedge \lambda)_+ + (g - \lambda)_+ \Big) \, \mathrm{d}F - \int g_- \, \mathrm{d}F \\ &= \int g_+ \, \mathrm{d}F - \int g_- \, \mathrm{d}F = \mathsf{LHS}, \end{aligned}$$

where the second equality follows from Theorem 6.5 and the nonnegativity of λ. ■

However, the symmetric integral does not maintain comonotone additivity. Indeed, consider a bounded function g with $\sup_x |g(x)| \leq g_{\max}$. If we had comonotone additivity, then

$$\int g \, \mathrm{d}F = \int (g + g_{\max} \mathbf{1}) \, \mathrm{d}F - g_{\max} \int \mathbf{1} \, \mathrm{d}F = \int_0^{g_{\max}} F([\![g \geq \gamma]\!]) \, \mathrm{d}\gamma + \int_{-g_{\max}}^{0} (F([\![g \geq \gamma]\!] - F(\Omega)) \, \mathrm{d}\gamma,$$

since any function is comonotone with the constant function $\mathbf{1}$. Thus we are motivated to extend the Choquet integral in a different way.

Definition 6.3: Choquet integral w.r.t. capacity

Fix any capacity $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$. For any $\mathfrak{L}$-measurable $g : \Omega \to \overline{\mathbb{R}}$, its Choquet integral w.r.t. F is defined as

$$\int g \, \mathrm{d}F := \int_0^\infty F([\![g \geq \gamma]\!]) \, \mathrm{d}\gamma + \int_{-\infty}^0 \Big(F([\![g \geq \gamma]\!]) - F(\Omega) \Big) \, \mathrm{d}\gamma, \tag{78}$$

where the right integrals are of the Riemann type, and are well-defined respectively due to monotonicity. Unless we have $\infty - \infty$ on the RHS, the Choquet integral always exists in $\overline{\mathbb{R}}$.

Again, let us compute the Choquet integral for step functions $s = \sum_{i=1}^k (w_i - w_{i+1}) \mathbf{1}_{A_i}$ where $w_k < \cdots < w_1$, $w_{k+1} := 0$ and $A_i = [\![s \geq w_i]\!]$. An easy calculation gives

$$\int s \, \mathrm{d}F = \sum_{i=1}^k (w_i - w_{i+1}) F(A_i). \tag{79}$$

Note the difference with the symmetric integral in (77).

We remind that two measurable functions agree almost everywhere do not necessarily yield the same Choquet integral: Take $\Omega = \{1, 2\}$ and $\mathfrak{L} = 2^\Omega$. Set $F = \{0, 0, 0, 1\}$ and consider $g = [0, 1], h = [2, 1]$. Clearly $F([\![g \neq h]\!]) = F(\{1\}) = 0$ but $\int g \, \mathrm{d}F = 0, \int h \, \mathrm{d}F = 1$. We need a stronger notion of "almost everywhere".

Definition 6.4: s-null set

A set $N \subseteq \Omega$ is called s-null w.r.t. the capacity $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ if for all $A, B \in \mathfrak{L}$, $A \subseteq B \cup N \implies F(A) \leq F(B)$.

Note that an s-null set N need not be $\mathfrak{L}$-measurable, however, when it does, we have $F(N) = 0$ since the empty set is s-null. Also, a subset of an s-null set is s-null.

Proposition 6.4: Choquet integals agree up to an s-null set

Let $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be a capacity. If two $\mathfrak{L}$-measurable functions g and h agree up to an s-null set N, then $\int g \, \mathrm{d}F = \int h \, \mathrm{d}F$.

Proof: The definition of s-null set allows us to restrict both the set function F and measurable functions g, h to $\mathring{\mathfrak{L}} := \{A \cap (\Omega \setminus N) : A \in \mathfrak{L}\}$ and $\Omega \setminus N$, respectively, without changing the Choquet integral. ■

Next, we remove the increasing assumption on the set function F, which requires a new concept.

Definition 6.5: Bounded (chain) variation

We define the chain variation of a *centered* set function $F : \mathfrak{L} \to \overline{\mathbb{R}}$ at $X \in \mathfrak{L}$ as

$$|F|(X) := \sup \left\{ \sum_i |F(A_i) - F(A_{i-1})| : \emptyset = A_0 \subset A_1 \subset \cdots \subset A_k = X \right\}. \tag{80}$$

Clearly $|F|$ is a capacity on the same domain $\mathfrak{L}$, and $|F(X)| \leq |F|(X)$. The set function F is said to have bounded variation if $\|F\| := |F|(\Omega) < \infty$.

Denote $\mathsf{BV}(\mathfrak{L})$ as the set of all centered set functions on $\mathfrak{L} \subseteq 2^\Omega$ that have bounded variation.

Proposition 6.5: $\mathsf{BV}(\mathfrak{L})$ is Banach

The space $\mathsf{BV}(\mathfrak{L})$, equipped with the chain variation norm $\|\cdot\| := |\cdot|(\Omega)$, is Banach.

Proof: Clearly, $\|F + G\| \leq \|F\| + \|G\|$, therefore $\mathsf{BV}(\mathfrak{L})$ is a vector space and $\|\cdot\|$ is a norm on it.

The completeness of the norm is proved as usual. ■

Theorem 6.6: Monotone decomposition, Aumann and Shapley [1974]

$F \in \mathsf{BV}(\mathfrak{L})$ iff $F = F_1 - F_2$ where F_1, F_2 are capacities. Moreover, for $F \in \mathsf{BV}(\mathfrak{L})$,

$$\|F\| = \min\{F_1(\Omega) + F_2(\Omega) : F = F_1 - F_2, F_1, F_2 \text{ are capacities } \}. \tag{81}$$

Proof: As usual, define

$$F_+(X) = \sup\left\{\sum_i (F(A_i) - F(A_{i-1}))_+ : \emptyset = A_0 \subset A_1 \subset \cdots \subset A_k = X\right\} \tag{82}$$

$$F_- = F_+ - F. \tag{83}$$

We easily verify F_+ and F_- are increasing hence the first claim is clear.

For the second claim, note first that $\|F\| \leq \|F_1\| + \|F_2\| = F_1(\Omega) + F_2(\Omega)$. On the other hand, $F_+(\Omega) + F_-(\Omega) \geq |F|(\Omega)$ is clear. Let $\emptyset = A_0 \subset \cdots \subset A_k = \Omega$ be a chain such that F_+ attains its norm on it (up to some $\epsilon \geq 0$). Divide the chain into two subchains: $S_1 = \{(A_{i-1}, A_i) : F(A_{i-1}) \leq F(A_i)\}$ and S_2 the rest. Clearly F_+ still attains its norm on S_1 hence F_- attains its (negated) norm on S_2 too, since $\sum_{S_1}(F(A_i) - F(A_{i-1})) + \sum_{S_2}(F(A_i) - F(A_{i-1})) = \sum_i (F(A_i) - F(A_{i-1})) = F(\Omega) = F_+(\Omega) - F_-(\Omega)$. Thus $F_+(\Omega) + F_-(\Omega) = \sum_{S_1}(F(A_i) - F(A_{i-1})) - \sum_{S_2}(F(A_i) - F(A_{i-1})) = \sum_i |F(A_i) - F(A_{i-1})| \leq |F|(\Omega)$. ■

We actually proved that $\|F\| = F_+(\Omega) + F_-(\Omega) = \|F_+\| + \|F_-\|$, from which we can also show that the norm is submultiplicative: $\|FG\| \leq \|F\| \cdot \|G\|$.

Theorem 6.6 allows us to extend the Choquet integral in a linear way.

Definition 6.6: Choquet integral for bounded variation set function

Fix any centered set function $F \in \mathsf{BV}(\mathfrak{L})$. For any $\mathfrak{L}$-measurable $g : \Omega \to \overline{\mathbb{R}}$, its Choquet integral w.r.t. F is defined as (see (82) and (83) for the definitions of F_+ and F_-)

$$\int g \,\mathrm{d}F := \int g \,\mathrm{d}F_+ - \int g \,\mathrm{d}F_- \tag{84}$$

$$= \int_0^\infty F([\![g \geq \gamma]\!]\,\mathrm{d}\gamma + \int_{-\infty}^0 \Big(F([\![g \geq \gamma]\!] - F(\Omega)\Big)\,\mathrm{d}\gamma, \tag{85}$$

where the right integrals in (84) are defined in Definition 6.3. Again, when we encounter $\infty - \infty$, the Choquet integral does not exist. Clearly, the Choquet integral is linear in the set function F. Moreover, the integral is even w.r.t. g iff F is symmetric, i.e., for all $A \in \mathfrak{L}$, $\Omega \setminus A \in \mathfrak{L}$ and $F(A) = F(\Omega \setminus A)$. (introduce the complement capacity and firm this claim up)

Of course we could extend the symmetric integral in a similar way, but that is of less interest to us. Judging from the end result in (85), it might appear that we could just directly drop the monotone assumption on F without going through the notion of bounded variation. This is of course inappropriate as how do we know the integrals in (85) are sensible at all?

We summarize some useful results about the general Choquet integral.

Theorem 6.7: Basic properties of the general Choquet integral

Let $F, G : \mathfrak{L} \to \mathbb{R}$ be centered and of bounded variation. For any $\mathfrak{L}$-measurable $g, h : \Omega \to \overline{\mathbb{R}}$,

(I). for all $\alpha, \beta \in \mathbb{R}$, $\int g \,\mathrm{d}(\alpha F + \beta G) = \alpha \int g \,\mathrm{d}F + \beta \int g \,\mathrm{d}G$;

(II). for all $\lambda \in \mathbb{R}_+$, $\int (\lambda g)\,\mathrm{d}F = \lambda \int g \,\mathrm{d}F$;

(III). f, g are comonotone iff $\int(g+h)\,\mathrm{d}F = \int g\,\mathrm{d}F + \int h\,\mathrm{d}F$ for all F;

If F is additionally increasing, i.e. a capacity, then we also have

(IV). $g \geq h \implies \int g\,\mathrm{d}F \geq \int h\,\mathrm{d}F$;

(V). for all $g : \Omega \to \overline{\mathbb{R}}$, $\int g\,\mathrm{d}F_\star = \sup\{\int s\,\mathrm{d}F : g \geq s \in \mathcal{S}(\mathfrak{L})\}$;

(VI). provided that $\mathfrak{L}$ is a lattice, F is submodular iff for all $\mathfrak{L}$-measurable g and h, $\int(g+h)\,\mathrm{d}F \leq \int g\,\mathrm{d}F + \int h\,\mathrm{d}F$. A similar claim for a supermodular F also holds.

Proof: (V), (VI), and (III) can be proved (sequentially) as Theorem 6.1, Theorem 6.3 and Theorem 6.5, respectively. The rest are obvious. ■

Comparing Proposition 6.3 and the current theorem we see that, despite their coincidence for nonnegative functions and the many similarities, the symmetric integral enjoys complete homogeneity but only partial comonotone additivity (see the last bullet point of Proposition 6.3) while the Choquet integral enjoys full comonotone additivity but only positively homogeneity. Both reduces to the usual Lebesgue integral when we have a (signed) measure.

We now come to a very natural question: What kind of functionals $I : \mathbb{R}^\Omega \to \mathbb{R}$ can be represented by some set function F as the Choquet integral $I(\cdot) = \int \cdot\,\mathrm{d}F$? The next representation theorem, usually attributed to Greco [1982], gives a satisfactory answer. Recall that a (nonempty) class of functions $\mathcal{G} \subseteq \mathbb{R}^\Omega$ (equipped with the pointwise order) is Stonean if $g \in \mathcal{G} \implies \forall \lambda \in \mathbb{R}_+, \lambda g, g \wedge \lambda, (g-\lambda)_+ \in \mathcal{G}$, in particular $\mathbf{0} \in \mathcal{G}$. A functional $\Gamma : \mathcal{G} \to \mathbb{R}$ is monotonically increasing if $g \geq h \implies \Gamma(g) \geq \Gamma(h)$ and comonotone additive if $\Gamma(g+h) = \Gamma(g) + \Gamma(h)$ for all comonotone $g, h \in \mathcal{G}$. We use $\mathfrak{M} = \mathfrak{M}(\mathcal{G})$ and $\mathfrak{L} = \mathfrak{L}(\mathcal{G})$ to denote the minimal set system $\{[\![g \geq \gamma]\!] : g \in \mathcal{G}, \gamma \in \mathbb{R}\}$ and the minimal lattice under which $\mathcal{G}$ is measurable, respectively. For any $h : \Omega \to \overline{\mathbb{R}}$, define

$$\Gamma^\star(h) := \inf\{\Gamma(g) : h \leq g \in \mathcal{G}\} \tag{86}$$
$$\Gamma_\star(h) := \sup\{\Gamma(g) : h \geq g \in \mathcal{G}\}, \tag{87}$$

which are clearly increasing functionals on $\overline{\mathbb{R}}^\Omega$. Similarly, for $A \in \mathfrak{L}$, we let

$$F^{\max}(A) := \Gamma^\star(\mathbf{1}_A) = \inf\{\Gamma(g) : \mathbf{1}_A \leq g \in \mathcal{G}\} \tag{88}$$
$$F^{\min}(A) := \Gamma_\star(\mathbf{1}_A) = \sup\{\Gamma(g) : \mathbf{1}_A \geq g \in \mathcal{G}\}, \tag{89}$$

which are centered (if $\mathbf{0} \in \mathcal{G}$), increasing and potentially infinite.

Theorem 6.8: Representing functionals as Choquet integral

Let $\mathcal{G} \subseteq \mathbb{R}^\Omega$ be a class of real valued functions on Ω that is nonnegative and Stonean; $\Gamma : \mathcal{G} \to \mathbb{R}_+$ be a functional that is increasing and comonotone additive, then

(I). Γ is represented by some capacity $F : \mathfrak{M} \to \bar{\mathbb{R}}_+$ iff Γ is truncation friendly:

- $\forall g \in \mathcal{G}, \Gamma(g \wedge t) \downarrow 0$ as $t \downarrow 0$, and $\Gamma(g \wedge t) \uparrow \Gamma(g)$ as $t \uparrow \infty$;

Also, F can only take ∞ at the full set Ω, which is avoided if there exists $g \in \mathcal{G}, \lambda \geq 0$ such that $\lambda \cdot g \geq \mathbf{1}_\Omega$;

(II). under (I), a centered F (not necessarily increasing) represents Γ iff $F^{\min} \leq F \leq F^{\max}$ on $\mathfrak{M}$;

(III). if F representing Γ is continuous from above (below), then $F = F^{\max}$ ($F = F^{\min}$) on $\mathfrak{M}$;

(IV). under (I), if $\mathcal{G}$ is convex and Γ is superadditive, then for all $h : \Omega \to \overline{\mathbb{R}}$, $\Gamma_\star(h) = \int h\,\mathrm{d}(F^{\min})_\star$ (see (70) for the definition of $F_\star$).

Proof: (I): Assume first that $F : \mathfrak{M} \to \mathbb{R}_+$ represents Γ. Thus for $t > 0$,

$$\Gamma(g \wedge t) = \int (g \wedge t)\, \mathrm{d}F = \int_0^\infty F([\![g \wedge t \geq \gamma]\!])\, \mathrm{d}\gamma = \int_0^t F([\![g \geq \gamma]\!])\, \mathrm{d}\gamma,$$

meaning I is truncation friendly. Conversely, for $g \in \mathcal{G}$ and $0 < \gamma < \lambda$, we have

$$\mathbf{1}_{[\![g \geq \lambda]\!]} \leq \frac{1}{\lambda - \gamma}\big((g \wedge \lambda) - (g \wedge \gamma)\big) \leq \mathbf{1}_{[\![g \geq \gamma]\!]},$$

whereas $(g \wedge \lambda) - (g \wedge \gamma) = (g \wedge \lambda) - (g \wedge \lambda) \wedge \gamma = (g \wedge \lambda - \gamma)_+ \in \mathcal{G}$ since $\mathcal{G}$ is Stonean. By comonotone additivity,

$$\Gamma\Big(\frac{1}{\lambda - \gamma}\big((g \wedge \lambda) - (g \wedge \gamma)\big)\Big) = \frac{1}{\lambda - \gamma}\Big(\Gamma(g \wedge \lambda) - \Gamma(g \wedge \gamma)\Big),$$

whence follows from the definitions (88)-(89) that

$$F^{\max}([\![g \geq \lambda]\!]) \leq \frac{1}{\lambda - \gamma}\Big(\Gamma(g \wedge \lambda) - \Gamma(g \wedge \gamma)\Big) \leq F^{\min}([\![g \geq \gamma]\!]).$$

Fix the subdivision $0 < a = t_{k+1} < \cdots < t_1 = b < \infty$ and apply the previous inequality:

$$\sum_{i=1}^k (t_i - t_{i+1}) F^{\max}([\![g \geq t_i]\!]) \leq \Gamma(g \wedge b) - \Gamma(g \wedge a) \leq \sum_{i=1}^k (t_i - t_{i+1}) F^{\min}([\![g \geq t_{i+1}]\!]).$$

Taking limits, applying the truncation property of Γ, and using $F^{\max} \geq F^{\min}$ on $\mathfrak{M}$ (which is implied by the monotonicity of Γ) we obtain

$$\Gamma(g) = \int_0^\infty F^{\max}([\![g \geq \gamma]\!])\, \mathrm{d}\gamma = \int_0^\infty F^{\min}([\![g \geq \gamma]\!])\, \mathrm{d}\gamma,$$

i.e., Γ is represented by both $F^{\max}$ and $F^{\min}$, hence also any F satisfying $0 \leq F^{\min} \leq F \leq F^{\max}$. Note that since $\mathcal{G}$ is Stonean, for all $A \in \mathfrak{M}$, $F^{\min}(A) \leq F^{\max}(A) < \infty$ except perhaps $A = \Omega$, which is further avoided if there exists $g \in \mathcal{G}, \lambda \geq 0$ such that $\lambda \cdot g \geq \mathbf{1}_\Omega$.

(II): Assume F represents Γ. Consider any $A \in \mathfrak{M}$ and any $g \in \mathcal{G}$ with $\mathbf{1}_A \geq g$. Note that $\mathbf{1}_A$ need not in $\mathcal{G}$. If $1 \geq \gamma \geq 0$ we have $[\![g \geq \gamma]\!] \subseteq A$ while if $\gamma > 1$ we have $[\![g \geq \gamma]\!] = \emptyset$. Thus

$$F(A) \geq \int_0^1 F([\![g \geq \gamma]\!])\, \mathrm{d}\gamma + \int_1^\infty F([\![g \geq \gamma]\!])\, \mathrm{d}\gamma = \Gamma(g) \implies F(A) \geq F^{\min}(A).$$

Similarly we prove $F(A) \leq F^{\max}(A)$.

(III): By (II) $F \leq F^{\max}$. Fix $g \in \mathcal{G}$ and put $p(\gamma) = F([\![g \geq \gamma]\!]), q(\gamma) = F^{\max}([\![g \geq \gamma]\!])$ which are both decreasing functions. The continuity assumption implies the left continuity of p. But $\int_0^t p(\gamma)\, \mathrm{d}\gamma = \int_0^t q(\gamma)\, \mathrm{d}\gamma = \Gamma(g \wedge t)$, hence $p = q$ almost everywhere. Since p is left continuous, $p \geq q$.

(IV): For all $\mathcal{G} \ni g \leq h$, $\Gamma(g) = \int g\, \mathrm{d}F^{\min} = \int g\, \mathrm{d}F^{\min}_\star \leq \int h\, \mathrm{d}F^{\min}_\star$, hence $\Gamma_\star(h) \leq \int h\, \mathrm{d}F^{\min}_\star$. On the other hand, by Theorem 6.1, $\int h\, \mathrm{d}F^{\min}_\star = \sup\left\{\int s\, \mathrm{d}F^{\min} : h \geq s \in \mathcal{S}_+(\mathfrak{L})\right\}$. Take an arbitrary $s = \sum_{i=1}^k a_i \mathbf{1}_{A_i}$ for some $A_1 \subset \cdots \subset A_k$ so that $h \geq s \in \mathcal{S}_+(\mathfrak{L})$, then

$$\int s\, \mathrm{d}F^{\min} = \sum_{i=1}^k a_i F^{\min}(A_i) = \sum_{i=1}^k a_i \Gamma_\star(\mathbf{1}_{A_i}) \leq \Gamma_\star\left(\sum_{i=1}^k a_i \mathbf{1}_{A_i}\right) = \Gamma_\star(s) \leq \Gamma_\star(h),$$

where the superadditivity of $\Gamma_\star$ follows from that of Γ and the convexity of $\mathcal{G}$. Thus we have the other inequality $\Gamma_\star(h) \geq \int h\, \mathrm{d}F^{\min}_\star$. ∎

It is tempted to prove a similar claim as in (IV) when Γ is subadditive. However, this does not seem plausible simply because in Theorem 6.1 we can only approximate from below, not above.

Note that comonotone additivity and monotonicity together imply positive homogeneity: the former

The model function for $\mathfrak{d}$ is the decreasing distribution function. Let g be an $\mathfrak{L}$-measurable function and $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be an increasing set function, we define $\mathfrak{d}(t) = F([\![g \geq t]\!])$ and its inverse $\mathfrak{q} := \mathfrak{d}^\dagger$ is known as the quantile function.

Note that If t is a continuity point of $\mathfrak{d}^\dagger$, then $\mathfrak{d}(s) \geq t \implies \mathfrak{d}^\dagger(t) = \mathfrak{d}^\dagger_{\max}(t) \geq s \implies (\mathfrak{d}^\dagger)^\dagger_{\max}(s) \geq t$, meaning $\mathfrak{d} \overset{\text{e.c.}}{\leq} (\mathfrak{d}^\dagger)^\dagger$. Similarly, if t is a continuity point of $\mathfrak{d}^\dagger$, then $\mathfrak{d}(s) \leq t \implies \mathfrak{d}^\dagger(t) = \mathfrak{d}^\dagger_{\min}(t) \leq s \implies (\mathfrak{d}^\dagger)^\dagger_{\min}(s) \leq t$, meaning $\mathfrak{d} \overset{\text{e.c.}}{\geq} (\mathfrak{d}^\dagger)^\dagger$. Therefore $(\mathfrak{d}^\dagger)^\dagger \overset{\text{e.c.}}{=} \mathfrak{d}$.

Proposition 6.6: Quantile retains expectation

Let $\mathfrak{d} : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be decreasing and $\mathfrak{d}^\dagger$ its inverse.

$$\int_0^\infty \mathfrak{d}(t)\,\mathrm{d}t = \int_0^\infty \mathfrak{d}^\dagger(t)\,\mathrm{d}t. \tag{104}$$

Proof: $\mathfrak{d}$, as a decreasing function, is Lebesgue measurable. Denote μ the Legesgue measure on the real line and apply the "integrating the tail" trick:

$$\mathsf{LHS} = \int_0^\infty \mu(\{s : \mathfrak{d}(s) \geq t\})\,\mathrm{d}t \leq \int_0^\infty \mu(\{s : \mathfrak{d}^\dagger(t) \geq s\})\,\mathrm{d}t = \int_0^\infty \mathfrak{d}^\dagger(t)\,\mathrm{d}t = \mathsf{RHS}. \qquad \blacksquare$$

Intuitively, this result says that switching the xy-axis does not change the (Riemann) integral of a nonnegative decreasing function.

Theorem 6.14: Change of formula

Let $\mathfrak{d} : [0, b] \to \overline{\mathbb{R}}$ be decreasing, where $0 < b < \infty$.

$$\int_0^b \mathfrak{d}(t)\,\mathrm{d}t = \int_0^\infty \mathfrak{d}^\dagger(t)\,\mathrm{d}t + \int_{-\infty}^0 (\mathfrak{d}^\dagger(t) - b)\,\mathrm{d}t. \tag{105}$$

Proof: Due to monotonicity, we can split the LHS into two parts $\int_0^a \mathfrak{d}(t)\,\mathrm{d}t + \int_a^b \mathfrak{d}(t)\,\mathrm{d}t$ so that $\mathfrak{d} \geq 0$ on the first term and $\mathfrak{d} \leq 0$ one the other term. Apply Proposition 6.6 on the first term directly and on the second term after an appropriate translation and sign change. $\blacksquare$

Thus we could define the Choquet integral of g w.r.t. the capacity F as

$$\int g\,\mathrm{d}F := \int_0^{F(\Omega)} \mathfrak{d}_g^\dagger(t)\,\mathrm{d}t, \quad \text{where} \quad \mathfrak{d}_g(t) := F([\![g \geq t]\!]). \tag{106}$$

At least when $F(\Omega) < \infty$, the above definition coincides with our previous Definition 6.3.

Proposition 6.7: Composition rule

Let $F : \mathfrak{L} \to \overline{\mathbb{R}}_+$ be an increasing set function, $g : \Omega \to \mathbb{R}$ be $\mathfrak{L}$-measurable, $\mathfrak{i} : \mathbb{R} \to \overline{\mathbb{R}}$ be increasing, and define $\mathfrak{d}_g(t) := F([\![g \geq t]\!])$. If $\mathfrak{d}_g$ has no common discontinuity point with $\mathfrak{i}$, then

$$\mathfrak{d}^\dagger_{\mathfrak{i}\circ g} \overset{\text{e.c.}}{=} \mathfrak{i} \circ \mathfrak{d}^\dagger_g. \tag{107}$$

Proof: For convenience we take $\mathfrak{d}^\dagger = \mathfrak{d}^\dagger_{\max}$, which clearly does not affect the result. Let $\mathfrak{i}^\dagger(s) := \inf\{t : \mathfrak{i}(t) \geq s\}$. We claim that

$$\mathfrak{d}^\dagger_{\mathfrak{i}\circ g}(t) \overset{\text{e.c.}}{=} \sup\{s : F([\![g \geq \mathfrak{i}^\dagger(s)]\!]) > t\} \leq \mathfrak{i}(\mathfrak{d}^\dagger_g(t)).$$

Indeed, $\mathfrak{d}_g(\mathfrak{i}^\dagger(s)) > t \implies \mathfrak{d}^\dagger_g(t) \geq \mathfrak{i}^\dagger(s)$. If $\mathfrak{d}^\dagger_g(t) > \mathfrak{i}^\dagger(s)$, then $\mathfrak{i}(\mathfrak{d}^\dagger_g(t)) \geq s$; while if $\mathfrak{d}^\dagger_g(t) = \mathfrak{i}^\dagger(s)$ we know $\mathfrak{d}_g$ is discontinuous at $\mathfrak{d}^\dagger_g(t)$ (otherwise $\mathfrak{d}_g(\mathfrak{d}^\dagger_g(t)) = \mathfrak{d}_g(\mathfrak{i}^\dagger(s)) = t$), then $\mathfrak{i}$ is continuous at $\mathfrak{d}^\dagger_g(t)$ and $\mathfrak{i}(\mathfrak{d}^\dagger_g(t)) = \mathfrak{i}(\mathfrak{i}^\dagger(s)) = s$.

On the other hand,

$$\mathfrak{i}(\mathfrak{d}_g^\dagger(t)) = \mathfrak{i}(\sup\{s : \mathfrak{d}_g(s) \geq t\}) = \sup\{\mathfrak{i}(s) : \mathfrak{d}_g(s) \geq t\}.$$

Indeed, due to monotonicity the set $\{s : \mathfrak{d}_g(s) \geq t\}$ is an interval whose right most endpoint a may or may not be present. We need only consider when a is not present, in which case $\mathfrak{d}_g$ is discontinuous at a. Thus by assumption $\mathfrak{i}$ is continuous at a, whence follows that both sides above are equal to $\mathfrak{i}(a)$. Therefore we have the other direction

$$\mathfrak{i}(\mathfrak{d}_g^\dagger(t)) = \sup\{\mathfrak{i}(s) : F([\![g \geq s]\!] \geq t)\} \leq \sup\{r : F([\![\mathfrak{i}(g) \geq r]\!] \geq t)\} = \mathfrak{d}_{\mathfrak{i}\circ g}^\dagger(t)$$

by taking $r = \mathfrak{i}(s)$. ■

We have an alternative proof of the comonotone additivity of the Choquet integral, under the equivalent definition in (106).

Theorem 6.15: Comonotone additivity, revisited

Let $F : \mathfrak{L} \to \mathbb{R}$ be centered and of bounded variation. If $g, h \in \mathcal{M}(\mathfrak{L})$ are comonotone and real valued, then $\int (g+h)\,\mathrm{d}F = \int g\,\mathrm{d}F + \int h\,\mathrm{d}F$.

Proof: We assume w.l.o.g. that F is increasing. According to Theorem 6.4, $g = \mathfrak{i}(g+h), h = \mathfrak{j}(g+h)$ for some continuous increasing functions $\mathfrak{i}, \mathfrak{j} : \mathbb{R} \to \mathbb{R}$ with $\mathfrak{i} + \mathfrak{j} = \mathsf{Id}$. Using Proposition 6.7 and its notation,

$$\begin{aligned}\int g\,\mathrm{d}F + \int h\,\mathrm{d}F &= \int_0^{F(\Omega)} \left(\mathfrak{d}_{\mathfrak{i}\circ(g+h)}^\dagger(t) + \mathfrak{d}_{\mathfrak{j}\circ(g+h)}^\dagger(t)\right) \mathrm{d}t = \int_0^{F(\Omega)} \left((\mathfrak{i}\circ\mathfrak{d}_{g+h}^\dagger)(t) + (\mathfrak{j}\circ\mathfrak{d}_{g+h}^\dagger)(t)\right) \mathrm{d}t \\ &= \int_0^{F(\Omega)} \mathfrak{d}_{g+h}^\dagger(t)\,\mathrm{d}t = \int (g+h)\,\mathrm{d}F.\end{aligned}$$

■

7 Duality

In this section we develop a duality theory for submodular functions by mimicking that for convex functions. Most of our results are taken from Fujishige [2005]. An important definition first.

Definition 7.1: Conjugate function

For any function $F : \mathfrak{L} \to \mathbb{R}$, define its conjugate $f_C^* : \mathbb{R}^\Omega \to \mathbb{R} \cup \{\infty\}$ as

$$f_C^*(\mathbf{z}) := \max_{X \in \mathfrak{L}} \mathbf{z}(X) - F(X) \tag{108}$$

$$= \mathbf{z}(\Omega) - (F \boxdot \mathbf{z})(\Omega). \tag{109}$$

In some cases, such as a supermodular F, we change the max to min.

Theorem 7.1: Conjugate is supermodular

The conjugate f_C^* is always convex and supermodular on the distributive lattice $(\mathbb{R}^\Omega, \leq)$.

In fact, as the notation suggests, f_C^* is the Fenchel conjugate of the Lovász extension, restricted to some convex set C.

Theorem 7.2: Justifying the conjugate

Consider $F : \mathfrak{L} \to \mathbb{R}$, then

$$f_C^*(\mathbf{z}) = \max_{\mathbf{w} \in C} \langle \mathbf{w}, \mathbf{z} \rangle - f(\mathbf{w}), \tag{110}$$

where f is the Lovász extension of F and $C = \mathcal{K} \cap [0,1]^\Omega$ is the convex hull of $\{\mathbf{1}_X : X \in \mathfrak{L}\}$.

Proof: By its definition in (108), we have

$$f_C^*(\mathbf{z}) = \max_{X \in \mathfrak{L}} \mathbf{z}(X) - F(X) = -\min_{X \in \mathfrak{L}} F(X) - \mathbf{z}(X) = -\min_{\mathbf{w} \in C} f(\mathbf{w}) - \langle \mathbf{w}, \mathbf{z} \rangle = \max_{\mathbf{w} \in C} \langle \mathbf{w}, \mathbf{z} \rangle - f(\mathbf{w}),$$

where the second last equality follows from Theorem 5.4 and from identifying the modular function $\mathbf{z}(\cdot)$ with its Lovász extension $\langle \cdot, \mathbf{z} \rangle$. ■

Corollary 7.1: Double conjugation

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular, then

$$(f_C^*)^*(\mathbf{w}) := \sup_{\mathbf{z} \in \mathbb{R}^\Omega} \langle \mathbf{w}, \mathbf{z} \rangle - f_C^*(\mathbf{z}) = f(\mathbf{w}) + \iota_C(\mathbf{w}), \tag{111}$$

where $\iota_C(\mathbf{w}) = 0$ if $\mathbf{w} \in C$ and ∞ otherwise. In particular, for all $X \in \mathfrak{L}$,

$$(f_C^*)^*(\mathbf{1}_X) = \sup_{\mathbf{z} \in \mathbb{R}^\Omega} \mathbf{z}(X) - f_C^*(\mathbf{z}) = F(X). \tag{112}$$

Proof: Following Remark 5.1 we assume w.l.o.g. that F is centered. When F is submodular, the Lovász extension f coincides with the support function, hence it is closed and convex. Then (111) follows from the classic result in convex analysis while (112) follows from Theorem 5.1. ■

Proposition 7.1: Conjugate of supermodular

Let $G : \mathfrak{L} \to \mathbb{R}$ be supermodular, then its conjugate $g_C^* : \mathbb{R}^\Omega \to \mathbb{R} \cup \{-\infty\}$ satisfies

$$g_C^*(\mathbf{z}) := \min_{X \in \mathfrak{L}} \{\mathbf{z}(X) - G(X)\} = \mathbf{z}(\Omega) - G(\Omega) - (g_{\tilde{C}}^\neg)^*(\mathbf{z}), \tag{113}$$

where $C = \mathcal{K} \cap [0,1]^\Omega$, $\tilde{C} = (-\mathcal{K}) \cap [0,1]^\Omega$, and $(g_{\tilde{C}}^\neg)^*$ is the conjugate of $G^\neg$ (see Definition 4.1).

Proof: Clearly $G^\neg : \bar{\mathfrak{L}} \to \mathbb{R}$ is submodular and the isotonic cone of $\bar{\mathfrak{L}}$, as discussed in Remark 4.5, is $-\mathcal{K}$. By the definition of $G^\neg$ and Theorem 7.2,

$$\begin{aligned} g_C^*(\mathbf{z}) &= \min_{X \in \mathfrak{L}} \mathbf{z}(X) - G(X) \\ &= \mathbf{z}(\Omega) - G(\Omega) - \max_{X \in \mathfrak{L}} \mathbf{z}(\Omega - X) - (G(\Omega) - G(X)) \\ &= \mathbf{z}(\Omega) - G(\Omega) - (g_{\tilde{C}}^\neg)^*(\mathbf{z}), \end{aligned}$$

■

The Lovász extension nicely bridges convex functions in the continuous domain and submodular functions in the discrete domain. Let us illustrate this point with another duality result.

Theorem 7.3: Discrete Fenchel duality

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ be submodular and $G : \mathfrak{L}_2 \to \mathbb{R}$ be supermodular, then for their conjugate functions $f_{C_1}^*$ and $g_{C_2}^*$ (which we use min instead of max in Definition 7.1), we have

$$\min\{F(X) - G(X) : X \in \mathfrak{L}_1 \wedge \mathfrak{L}_2\} = \max\{g_{C_2}^*(\mathbf{z}) - f_{C_1}^*(\mathbf{z}) : \mathbf{z} \in \mathbb{R}^\Omega\}. \tag{114}$$

Moreover, if F, G are integer valued, the maximizer of RHS can be chosen in $\mathbb{Z}^\Omega$.

Proof: Through translation we may assume w.l.o.g. that $G(\Omega) = 0$. By Corollary 4.4 and Proposi-

tion 7.1, we have (after restricting to $\mathfrak{L}_1 \wedge \mathfrak{L}_2$)

$$\begin{aligned}\mathsf{LHS} &= \min\{F(X) + G^\neg(\Omega - X) : X \in \mathfrak{L}_1 \wedge \mathfrak{L}_2\} = (F \boxdot G^\neg)(\Omega) = \max\{\mathbf{p}(\Omega) : \mathbf{p} \in \mathbf{P}_F \cap \mathbf{P}_{G^\neg}\} \\ \mathsf{RHS} &= \max\{(G^\neg \boxdot \mathbf{z})(\Omega) + (F \boxdot \mathbf{z})(\Omega) - \mathbf{z}(\Omega) : \mathbf{z} \in \mathbb{R}^\Omega\} \\ &= \max\{\mathbf{s}(\Omega) + \mathbf{t}(\Omega) - \mathbf{z}(\Omega) : \mathbf{s} \in \mathbf{P}_F, \mathbf{s} \le \mathbf{z}, \mathbf{t} \in \mathbf{P}_{G^\neg}, \mathbf{t} \le \mathbf{z}\} \\ &= \max\{\mathbf{s}(\Omega) \wedge \mathbf{t}(\Omega) : \mathbf{s} \in \mathbf{P}_F, \mathbf{t} \in \mathbf{P}_{G^\neg}\},\end{aligned}$$

where in the last equality we used the fact that $\mathbf{z} = \mathbf{s} \vee \mathbf{t}$ and $\mathbf{s} + \mathbf{t} = \mathbf{s} \vee \mathbf{t} + \mathbf{s} \wedge \mathbf{t}$. It is clear now that by taking $\mathbf{p} = \mathbf{s} = \mathbf{t} = \mathbf{s} \wedge \mathbf{t} = \mathbf{z}$ we have $\mathsf{LHS} = \mathsf{RHS}$. When F, G are integer valued, $\mathbf{p}$ can be taken integral, thanks to Corollary 4.4, hence also $\mathbf{z}$ on the RHS. ■

Definition 7.2: Subgradient

$\mathbf{w} \in \mathbb{R}^\Omega$ is a subgradient of $F : \mathfrak{L} \to \mathbb{R}$ at $X \in \mathfrak{L}$ if for all $Y \in \mathfrak{L}$,

$$F(Y) \ge F(X) + \mathbf{w}(Y - X). \tag{115}$$

The subdifferential $\partial F(X)$ consists of all subgradients of F at $X \in \mathfrak{L}$. Clearly the subdifferential is a polyhedron.

Theorem 7.4: Duality

Let $F : \mathfrak{L} \to \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^\Omega$, and $X \in \mathfrak{L}$, then

$$\mathbf{w} \in \partial F(X) \iff f_C^*(\mathbf{w}) + F(X) = \mathbf{w}(X) \iff \mathbf{w} \in \partial f_C(\mathbf{1}_X). \tag{116}$$

If F is submodular, we can add the equivalence $\mathbf{1}_X \in \partial f_C^*(\mathbf{w})$.

Theorem 7.5: Optimality condition

$X \in \mathfrak{L}$ minimizes $F : \mathfrak{L} \to \mathbb{R}$ iff $\mathbf{0} \in \partial F(X)$.

Proposition 7.2: Irrelevance of disjoint sets

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular and $X \in \mathfrak{L}$, then $\mathbf{w} \in \mathbb{R}^\Omega$ belongs to the subdifferential $\partial F(X)$ iff (115) holds for all $\mathfrak{L} \ni Y \subseteq X$ and $\mathfrak{L} \ni Y \supseteq X$.

Proof: We need only prove the sufficiency. Fix any $Z \in \mathfrak{L}$, then

$$\begin{aligned}\mathbf{w}(X \cup Z) - \mathbf{w}(X) &\le F(X \cup Z) - F(X), \\ \mathbf{w}(X \cap Z) - \mathbf{w}(X) &\le F(X \cap Z) - F(X).\end{aligned}$$

Adding the inequalities and applying the submodularity of F completes the proof. ■

Theorem 7.6: Decomposing the subdifferential

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular and $X \in \mathfrak{L}$, then

$$\partial F(X) = \partial F_X(X) \otimes \partial F^X(\emptyset), \tag{117}$$

where F_X and F^X are defined in Theorem 3.8.

Proof: By definition, $\mathbf{p} \in \partial F_X(X)$ and $\mathbf{q} \in \partial F^X(\emptyset)$ iff for all $\mathfrak{L} \ni Y \subseteq X$ and $\mathfrak{L} \ni Z \supseteq X$,

$$\begin{aligned}\mathbf{p}(Y) - \mathbf{p}(X) &\leq F_X(Y) - F_X(X) = F(Y) - F(X),\\ \mathbf{q}(Z - X) &\leq F^X(Z - X) - F^X(\emptyset) = F(Z) - F(X).\end{aligned}$$

Apply Proposition 7.2. ■

Theorem 7.7: Divide and conquer

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular, then $X \in \mathfrak{L}$ minimizes F iff X minimizes F_X and $\emptyset$ minimizes F^X.

Proof: Apply Theorem 7.5. ■

Clearly we can iterate the argument in the spirit of Remark 3.1.

Proposition 7.3: Subdifferential of extremal sets

Let $F : \mathfrak{L} \to \mathbb{R}$ be centered, then

$$\partial F(\emptyset) = \mathbf{P}_F \tag{118}$$
$$\partial F(\Omega) = \mathbf{P}_{F^\neg}. \tag{119}$$

If F is additionally submodular, then for all $X \in \mathfrak{L}$, $\partial F(X) \cap \mathbf{B}_F \neq \emptyset$.

Proof: We only prove the last claim. If F is submodular, by Proposition 3.2 there exists some base $\mathbf{b}$ such that $\mathbf{b}(X) = F(X)$. Verify (115). ■

Theorem 7.8: Subdifferential is positively homogeneous

For any function $F : \mathfrak{L} \to \mathbb{R}$ and $\lambda > 0$, $\partial(\lambda F)(X) = \lambda \cdot \partial F(X)$ for all $X \in \mathfrak{L}$.

Theorem 7.9: Subdifferential is additive

Let $F : \mathfrak{L}_1 \to \mathbb{R}$ and $G : \mathfrak{L}_2 \to \mathbb{R}$ be submodular. For all $X \in \mathfrak{L} := \mathfrak{L}_1 \wedge \mathfrak{L}_2$,

$$\partial(F + G)(X) = \partial F(X) + \partial G(X). \tag{120}$$

Proof: According to Theorem 7.6 and Proposition 7.3,

$$\begin{aligned}\partial(F+G)(X) &= \partial(F+G)_X(X) \otimes \partial(F+G)^X(\emptyset)\\ &= \partial(F_X + G_X)(X) \otimes \partial(F^X + G^X)(\emptyset)\\ &= \mathbf{P}_{(F_X+G_X)^\neg} \otimes \mathbf{P}_{F^X+G^X}\\ &= \mathbf{P}_{F_X^\neg+G_X^\neg} \otimes \mathbf{P}_{F^X+G^X}\\ &= \left(\mathbf{P}_{F_X^\neg} + \mathbf{P}_{G_X^\neg}\right) \otimes \left(\mathbf{P}_{F^X} + \mathbf{P}_{G^X}\right)\\ &= \left(\partial(F_X)(X) + \partial(G_X)(X)\right) \otimes \left(\partial(F^X)(\emptyset) + \partial(G^X)(\emptyset)\right)\\ &= \left(\partial(F_X)(X) \otimes \partial(F^X)(\emptyset)\right) + \left(\partial(G_X)(X) \otimes \partial(G^X)(\emptyset)\right)\\ &= \partial F(X) + \partial G(X).\end{aligned}$$

■

Again, the surprising part of this result is that we do *not*, although we could, restrict F or G to $\mathfrak{L}$.

Theorem 7.10: Summation is dual to convolution

Let F and G be submodular on $\mathfrak{L}$, then

$$(F+G)^* = (f_C + g_C)^* = f_C^* \boxdot g_C^*, \tag{121}$$

where recall that $C := \mathcal{K} \cap [0,1]^\Omega$ is the convex hull of $\{\mathbf{1}_X : X \in \mathfrak{L}\}$.

Proof: The first equality is always true due to Theorem 7.2 and Theorem 5.2. In case of submodularity, the Lovász extensions are convex, thanks to Theorem 5.5. ■

Theorem 7.11: Extreme points of subdifferential

Let $\mathfrak{L}$ be simple and $F : \mathfrak{L} \to \mathbb{R}$ be submodular. For any $X \in \mathfrak{L}$, $\mathbf{w} \in \mathbb{R}^\Omega$ is an extreme point of $\partial F(X)$ iff there exists a maximal increasing sequence

$$\emptyset = S_0 \subset S_1 \subset \cdots \subset S_n = \Omega$$

in $\mathfrak{L}$ that contains X such that for all $i \in \{1, \dots, n\}$, $w(S_i \setminus S_{i-1}) = F(S_i) - F(S_{i-1})$.

Proof: Following Remark 5.1 we assume w.l.o.g. that F is centered. By Proposition 7.3 we have $\partial F(\emptyset) = \mathbf{P}_F$. Thanks to submodularity, the current theorem for $X = \emptyset$ is true by Proposition 2.5 and Corollary 4.2. Similarly for $X = \Omega$. Finally apply Theorem 7.6. ■

Theorem 7.12: Intersection of subdifferential

Consider $F : \mathfrak{L} \to \mathbb{R}$. Every $\mathbf{w} \in C$ can be written as $\mathbf{w} = \sum_i \lambda_i \mathbf{1}_{A_i}$ for some $A_i \in \mathfrak{L}$, $\lambda_i > 0$ and $\sum_i \lambda_i = 1$, and

$$\partial f_C(\mathbf{w}) = \bigcap_i \partial F(A_i). \tag{122}$$

Proof: By Theorem 7.4, $\mathbf{z} \in \partial f_C(\mathbf{w})$ iff

$$\begin{aligned} \langle \mathbf{z}, \mathbf{w} \rangle = f_C(\mathbf{w}) + f_C^*(\mathbf{z}) \quad &\iff \quad \sum_i \lambda_i \langle \mathbf{z}, \mathbf{1}_{A_i} \rangle = \sum_i \lambda_i \Big(F(A_i) + f_C^*(\mathbf{z}) \Big) \\ &\iff \quad \forall\, i, \quad \langle \mathbf{z}, \mathbf{1}_{A_i} \rangle = F(A_i) + f_C^*(\mathbf{z}). \end{aligned}$$

Apply Theorem 7.4 once more. ■

Definition 7.3: Normal cone

Following convex analysis, we define the normal cone of a lattice $\mathfrak{L}$ at the set $X \in \mathfrak{L}$ as

$$\mathcal{N}_\mathfrak{L}(X) := \{\mathbf{w} \in \mathbb{R}^\Omega : \langle \mathbf{w}, \mathbf{1}_Y - \mathbf{1}_X \rangle \leq 0, \forall\, Y \in \mathfrak{L}\}.$$

Theorem 7.13: Constrained optimality

Let $F : \mathfrak{L} \to \mathbb{R}$ be submodular and $\mathfrak{L}_0 \subseteq \mathfrak{L}$ a sublattice, then $X \in \mathfrak{L}_0$ minimizes F on $\mathfrak{L}_0$ iff

$$\mathbf{0} \in \partial F(X) + \mathcal{N}_{\mathfrak{L}_0}(X). \tag{123}$$

Proof: The sufficiency is immediate from the definitions. On the other hand, if X minimizes F on $\mathfrak{L}_0$, we consider the constant $F(X)$ as a supermodular function on the sublattice $\mathfrak{L}_0$. Apply the sandwich Theorem 4.9, we have some $\mathbf{p} \in \mathbb{R}^\Omega$ so that $F \geq \mathbf{p}$ on $\mathfrak{L}$ while $\mathbf{p} \geq F(X)$ on $\mathfrak{L}_0$. Necessarily $F(X) = \mathbf{p}(X)$. Thus for all $Z \in \mathfrak{L}$, $F(Z) \geq F(X) + \mathbf{p}(Z) - \mathbf{p}(X)$ while for all $Y \in \mathfrak{L}_0$, $\mathbf{p}(X) - \mathbf{p}(Y) \leq F(X) - F(X) = 0$. Clearly $\mathbf{p}$ satisfies (123). ■

Usually the sublattice $\mathfrak{L}_0$ is given explicitly as the constraint set

$$\mathfrak{L}_0 := \{X \in \mathfrak{L} : F_i(X) = \alpha_i, i = 1, \ldots, m\} \tag{124}$$

where $F_i : \mathfrak{L} \to \mathbb{R}$ are submodular and $\alpha_i = \min\{F_i(X) : X \in \mathfrak{L}\}$. Thanks to Proposition 2.1, $\mathfrak{L}_0$ thus defined is indeed a sublattice of $\mathfrak{L}$. Very interestingly, we can also develop a theory of Lagrangian duality for the (discrete) constrained minimization problem

$$\min_{X \in \mathfrak{L}_0} F(X). \tag{125}$$

First recall that the Lagrangian is defined as

$$L(X, \boldsymbol{\mu}) := F(X) + \sum_{i=1}^{m} \mu_i (F_i(X) - \alpha_i), \tag{126}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m_+$ is called a Lagrangian multiplier if

$$\min_{X \in \mathfrak{L}} L(X, \boldsymbol{\mu}) = \min_{X \in \mathfrak{L}_0} F(X). \tag{127}$$

Note that the constraint $F_i(X) = \alpha_i$ is equivalent to $F_i(X) \leq \alpha_i$ since α_i is the minimum of F_i. That partly explains why we take $\boldsymbol{\mu}$ to be nonnegative, but the true reason will become clear only later.

Theorem 7.14: Existence of a Lagrangian multiplier

There always exists a Lagrangian multiplier for problem (125), where $\mathfrak{L}_0$ is given by (124) and $\alpha_i = \min_{X \in \mathfrak{L}} F_i(X)$.

Proof: Let

$$\gamma := \min_{i=1,\ldots,m} \min_{X \in \mathfrak{L} : F_i(X) \neq \alpha_i} F_i(X) - \alpha_i.$$

Clearly $\gamma > 0$, due to the minimality of α_i. Let $\beta = \min_{X \in \mathfrak{L}} F(X)$. If $X \notin \mathfrak{L}_0$, then

$$L(X, \boldsymbol{\mu}) \geq \beta + \gamma \mu_i,$$

for some $1 \leq i \leq m$. Therefore if,

$$\min_{i=1,\ldots,m} \mu_i \geq \frac{1}{\gamma} \left(\min_{X \in \mathfrak{L}_0} F(X) - \beta \right),$$

we must have the equality in (127). ■

No assumption on submodularity is needed here, thanks to the inherent finite structure. Note that our proof is semi-constructive, as it requires knowledge of $\min_{X \in \mathfrak{L}} F(X)$ and $\min_{X \in \mathfrak{L}_0} F(X)$ while the latter is our true interest! Also clear from the proof is that any upper bound of a Lagrangian multiplier remains a Lagrangian multiplier.

Given a Lagrangian multiplier, clearly any minimizer of the RHS in (127) also minimizes its LHS. Conversely, any minimizer of the LHS lying in $\mathfrak{L}_0$ minimizes the RHS too. Recall that $(X^\star, \boldsymbol{\mu}^\star)$ is called a saddle-point of the Lagrangian iff

$$\sup_{\boldsymbol{\mu} \in \mathbb{R}^m_+} L(X^\star, \boldsymbol{\mu}) \leq L(X^\star, \boldsymbol{\mu}^\star) \leq \min_{X \in \mathfrak{L}} L(X, \boldsymbol{\mu}^\star). \tag{128}$$

Note that the inequalities obviously imply equalities.

Theorem 7.15: Saddle-point and optimality

Let $F, F_i : \mathfrak{L} \to \mathbb{R}, i = 1, \ldots, m$ be submodular, then $X^\star \in \mathfrak{L}_0$ minimizes F on $\mathfrak{L}_0$ and $\boldsymbol{\mu}^\star \in \mathbb{R}^m_+$ is a Lagrangian multiplier iff $(X^\star \in \mathfrak{L}, \boldsymbol{\mu}^\star \in \mathbb{R}^m_+)$ is a saddle-point of the Lagrangian iff $X^\star \in \mathfrak{L}_0, \boldsymbol{\mu} \in \mathbb{R}^m_+$ and

$$\mathbf{0} \in \partial F(X^\star) + \sum_{i=1}^m \mu_i^\star \partial F_i(X^\star), \tag{129}$$

where the subdifferentials are taken on $\mathfrak{L}$.

Proof: If $(X^\star, \boldsymbol{\mu}^\star)$ is a saddle-point of the Lagrangian, then (128) holds, hence we have

$$\sup_{\boldsymbol{\mu} \in \mathbb{R}^m_+} L(X^\star, \boldsymbol{\mu}) = L(X^\star, \boldsymbol{\mu}^\star) = \min_{X \in \mathfrak{L}} L(X, \boldsymbol{\mu}^\star), \tag{130}$$

from which we immediately know $X^\star \in \mathfrak{L}_0$ for otherwise the LHS would be unbounded from above. On the other hand, clearly RHS $\leq \min_{X \in \mathfrak{L}_0} F(X)$. Thus $X^\star$ minimizes F on $\mathfrak{L}_0$ and $\boldsymbol{\mu}^\star$ is a Lagrangian multiplier. The converse is obvious, hence we have proved the first equivalence.

If $X^\star$ minimizes F on $\mathfrak{L}_0$ and $\boldsymbol{\mu}^\star$ is a Lagrangian multiplier, then applying Theorem 7.5, Theorem 7.8 and Theorem 7.9 on the RHS of (130) yields (129). Note that we need the Lagrangian multiplier $\boldsymbol{\mu}$ to be nonnegative in order to apply Theorem 7.8. For the converse, note that (129) implies $X^\star$ is optimal for the RHS of (130), whence follows that $X^\star$ minimizes F on $\mathfrak{L}_0$ and $\boldsymbol{\mu}$ is a Lagrangian multiplier. ■

Theorem 7.16: Characterizing the Lagrangian multipliers

Let $F, F_i : \mathfrak{L} \to \mathbb{R}, i = 1, \ldots, m$ be submodular. The Lagrangian multipliers for (125) are given by

$$\operatorname*{argmax}_{\boldsymbol{\mu} \in \mathbb{R}^m_+} \left\{ g(\boldsymbol{\mu}) := \min_{X \in \mathfrak{L}} L(X, \boldsymbol{\mu}) \right\}. \tag{131}$$

Proof: Theorem 7.14 proves the existence of a Lagrangian multiplier $\boldsymbol{\mu}^\star$. Let $X^\star$ be any minimizer of F on $\mathfrak{L}_0$. For any $\hat{\boldsymbol{\mu}}$ that maximizes (131), according to (130),

$$\left\{ \min_{X \in \mathfrak{L}} L(X, \hat{\boldsymbol{\mu}}) \right\} = g(\hat{\boldsymbol{\mu}}) \leq \left\{ \sup_{\boldsymbol{\mu} \in \mathbb{R}^m_+} L(X^\star, \boldsymbol{\mu}) \right\} = L(X^\star, \boldsymbol{\mu}^\star) = \min_{X \in \mathfrak{L}_0} F(X),$$

verifying that $\hat{\boldsymbol{\mu}}$ is a Lagrangian multiplier, since LHS $\leq$ RHS is clear due to nonnegativity. By (128), any Lagrangian multiplier clearly belongs to (131). ■

Since the function g by definition is increasing w.r.t. $\boldsymbol{\mu}$, we see again that any upper bound of a Lagrangian multiplier remains a Lagrangian multiplier.

Remark 7.1: Optimizing the dual

Thanks to Theorem 7.16 and Theorem 7.14, to optimize the constrained problem (125), we can first optimize the dual problem (131) to find a Lagrangian multiplier, based on which we minimize the "unconstrained" Lagrangian w.r.t. $X \in \mathfrak{L}$. If the minimizer $X^\star$ happens to be in $\mathfrak{L}_0$, we have found a minimizer to the primal problem (125). Happily, the dual problem (131) is to maximize a bounded (from above) continuous concave function g on $\mathbb{R}^m_+$.

8 Algorithms

We consider the separable minimization problem first:

$$\min_{\mathbf{b}\in\mathbf{B}_F} \sum_{x\in\Omega} \varphi_x(b_x), \tag{132}$$

where $F : \mathfrak{L} \to \mathbb{R}$ is a centered submodular function with $\mathbf{B}_F$ its base polyhedron, and each φ_x is a univariate (finite-valued) convex function.

Theorem 8.1: Optimality for separable convex minimization

$\mathbf{b} \in \mathbf{B}_F$ is optimal for (132) iff for all $x \in \Omega, y \in \mathsf{dep}(\mathbf{b}, x)$ we have

$$\varphi_x^+(b_x) \geq \varphi_y^-(b_y), \tag{133}$$

where φ^- and φ^+ denote the left and right derivative of φ, respectively.

Proof: Clearly $\mathbf{b}$ is optimal iff $\mathbf{0} \in \partial\boldsymbol{\varphi}(\mathbf{b}) + \mathcal{N}_{\mathbf{B}_F}(\mathbf{b})$, namely, according to Corollary 4.3,

$$\forall\, x \in \Omega, \forall\, y \in \mathsf{dep}(\mathbf{b}, x), \quad \langle -\partial\boldsymbol{\varphi}(\mathbf{b}), \mathbf{1}_x - \mathbf{1}_y \rangle \leq 0,$$

which is (133). ■

Take $\varphi_x(b_x) = w_x b_x$ we recover Theorem 4.6.

The next result is important for developing a decomposition algorithm.

Theorem 8.2: Optimality certificate

$\mathbf{b} \in \mathbf{B}_F$ is optimal for (132) iff there exists a chain $\emptyset = A_0 \subset A_1 \subset \cdots \subset A_k = \Omega$ in $\mathfrak{L}$ such that

(I). $\forall i, \mathbf{b}(A_i) = F(A_i)$;

(II). $\forall i \geq j$, $\forall x \in A_i \setminus A_{i-1}, \forall y \in A_j \setminus A_{j-1}$, $\varphi_x^+(b_x) \geq \varphi_y^-(b_y)$.

Proof: $\Longleftarrow$: We verify Theorem 8.1. Let $x \in \Omega, y \in \mathsf{dep}(\mathbf{b}, x)$. Thanks to (I), we know $x \in A_i \setminus A_{i-1}$, $y \in A_j \setminus A_{j-1}$ for some $i \geq j$. Thus (II) implies (133).

$\Longrightarrow$: Consider the saturation lattice $\mathsf{S}_\mathbf{b} := \{X \in \mathfrak{L} : \mathbf{b}(X) = F(X)\}$. Following Remark 1.2 we construct the ordered set $\mathcal{P}(\mathsf{S}_\mathbf{b}) = \{[X_1], \ldots, [X_k]\}$. For each i, define $\mathfrak{g}_i^- = \max\{\varphi_y^-(b_y) : y \in [X_i]\}$ and similarly $\mathfrak{g}_i^+ = \min\{\varphi_x^+(b_x) : x \in [X_i]\}$. Note that $x \in [X_i], y \in [X_j]$ with $[X_j] \preceq [X_i]$ implies $y \in \mathsf{dep}(\mathbf{b}, x)$ hence by Theorem 8.1 $\varphi_x^+(b_x) \geq \varphi_y^-(b_y)$, thus $\mathfrak{g}_i^+ \geq \mathfrak{g}_j^-$. Finally define $\mathfrak{g}_i^\uparrow := \max\{\mathfrak{g}_j^- : [X_j] \preceq [X_i]\}$. Clearly $\mathfrak{g}_i^- \leq \mathfrak{g}_i^\uparrow \leq \mathfrak{g}_i^+$, and $[X_j] \preceq [X_i] \implies \mathfrak{g}_j^\uparrow \leq \mathfrak{g}_i^\uparrow$. Assume w.l.o.g. that $\mathfrak{g}_1^\uparrow \leq \ldots \leq \mathfrak{g}_k^\uparrow$ and let $A_i = \bigcup_{j\leq i}[X_j]$. (I) is clearly met since $[X_j] \in \mathsf{S}_\mathbf{b}$. For (II), consider $x \in [X_i], y \in [X_j]$ with $i \geq j$, we have $\varphi_x^+(b_x) \geq \mathfrak{g}_i^+ \geq \mathfrak{g}_i^\uparrow \geq \mathfrak{g}_j^\uparrow \geq \mathfrak{g}_j^- \geq \varphi_y^-(b_y)$, as required. ■

In the following algorithm, we assume for each $x \in \Omega$, φ_x is super-coercive, i.e., $\lim_{|b|\to\infty} \varphi_x(b)/|b| \to \infty$, which guarantees that the Fenchel conjugate φ_x^* is finite hence subdifferentiable everywhere.

Algorithm 8.1: Decomposition algorithm

(I). Find $\eta \in \mathbb{R}$ such that $F(\Omega) \in \sum_{x\in\Omega} \partial\varphi_x^*(\eta)$;

(II). Find $\mathbf{b} \in \mathbf{B}_F$ such that for all $x, y \in \Omega$:

a) $\varphi_x^+(b_x) < \eta$ and $\varphi_y^-(b_y) > \eta$ imply $y \notin \mathsf{dep}(\mathbf{b}, x)$;

b) $\varphi_x^+(b_x) < \eta$, $\varphi_y^-(b_y) = \eta$ and $y \in \mathsf{dep}(\mathbf{b}, x)$ imply $b_y = \inf \partial\varphi_y^*(\eta)$;

c) $\varphi_x^+(b_x) = \eta$, $\varphi_y^-(b_y) > \eta$ and $y \in \mathsf{dep}(\mathbf{b}, x)$ imply $b_x = \sup \partial\varphi_x^*(\eta)$;

(III). Let $\Omega_+ = \{\}$, $\Omega_- = \{\}$ and $\Omega_0 = \Omega \setminus \Omega_+ \setminus \Omega_-$. Set $\mathbf{b}^\star(\Omega_0) = \mathbf{b}(\Omega_0)$ and repeat the algorithm for the subproblems:

Needless to say that the current algorithm generalizes the greedy Algorithm 4.1.

Let us explain Algorithm 8.1: According to Theorem 8.1, we need to satisfy the optimality condition (133), hence (II)a collects those pairs which could potentially violate this optimality condition while (II)b and (II)c are needed merely for technical reasons.

Proposition 8.1: Algorithm 8.1 is well-defined and terminates

Algorithm 8.1 is well-defined and terminates in finite steps.

Proof: (I): Equivalently, we may consider the problem

$$\min_{\mathbf{q}\in\mathbb{R}^\Omega} \sum_{x\in\Omega} \varphi_x(q_x) \qquad \text{s.t.} \qquad \mathbf{q}(\Omega) = F(\Omega), \tag{134}$$

whose Fenchel dual is

$$\max_{\eta\in\mathbb{R}} \eta \cdot F(\Omega) - \sum_{x\in\Omega} \varphi_x^*(\eta). \tag{135}$$

Therefore (I) is nothing but the optimality condition of (135). Since φ_x by assumption is finite everywhere, (135) admits a maximizer.

(II): We provide an implementation for (II). Note first that $\varphi_x^+(b_x) < \eta \iff b_x < \inf \partial\varphi_x^*(\eta)$ and similarly $\varphi_y^-(b_y) > \eta \iff b_y > \sup \partial\varphi_y^*(\eta)$. Start with any base $\mathbf{b} \in \mathbf{B}_F$ and we perform the following procedure:

1). find $x \in \Omega$ so that $b_x < \inf \partial\varphi_x^*(\eta)$; if no such x exists, $\mathbf{b}$ is optimal;

2). if there exists $y \in \mathsf{dep}(\mathbf{b}, x)$ such that $b_y > \sup \partial\varphi_y^*(\eta)$, update $\mathbf{b} \leftarrow \mathbf{b} + \alpha(\mathbf{e}_x - \mathbf{e}_y)$ with

$$0 \le \alpha = \min\{\mathsf{c}(\mathbf{b}, x, y),\ \sup \partial\varphi_x^*(\eta) - b_x\ ,\ b_y - \inf \partial\varphi_y^*(\eta)\ \};$$

if there is no such y, remove $\mathsf{dep}(\mathbf{b}, x)$ from later considerations and go to 1);

3). if $\alpha = \sup \partial\varphi_x^*(\eta) - b_x$, remove x from later considerations and go to 1); otherwise repeat 2).

Some explanations: If we find no x in Step 1), then $\mathbf{b} \in \mathbf{B}_F$ is optimal for the *relaxation* (134) hence also optimal for (132). Step 2) makes one of (II)a, (II)b and (II)c happen for the pair (x, y). We need to argue that the subsequent updates never ruin previous updates. Indeed, note first that if $\alpha = \sup \partial\varphi_x^*(\eta) - b_x$ (resp. $\alpha = b_y - \inf \partial\varphi_y^*(\eta)$), then b_x (resp. b_y), which meets (II)c (resp. (II)b), is never updated again. Then we prove that step 2) is repeated for each x at most $|\Omega|$ times: We need only consider $\alpha = \mathsf{c}(\mathbf{b}, x, y)$, after the immediate update of $\mathbf{b}$, $y \notin \mathsf{dep}(\mathbf{b}, x)$ and in subsequent repetitions of step 2) (for the same x), $\mathsf{dep}(\mathbf{b}, x)$ can only shrink. Finally observe that we either remove x from our consideration because $b_x = \sup \partial\varphi_x^*(\eta)$ in which case (II)c is satisfied, or we remove $\mathsf{dep}(\mathbf{b}, x)$ in which case any element in $\mathsf{dep}(\mathbf{b}, x)$ is never updated anymore: For all $x \neq x' \in \mathsf{dep}(\mathbf{b}, x)$, either $b_{x'} = \inf \partial\varphi_{x'}^*(\eta)$ (which clearly remains intact) or $b_{x'} < \inf \partial\varphi_{x'}^*(\eta)$ (but $\mathsf{dep}(\mathbf{b}, x') \subseteq \mathsf{dep}(\mathbf{b}, x)$ thus step (II)b is never executed for x').

(III): ■

Proposition 8.2: Algorithm 8.1 is correct

Fujishige [2005] minimum-point algorithm

The problem of maximizing a symmetric non-monotone submodular function subject to no constraints admits a 1/2 approximation algorithm [Feige et al., 2011]. Computing the maximum cut of a graph is a special case of this problem. The more general problem of maximizing an arbitrary non-monotone submodular function subject to no constraints also admits a 1/2 approximation algorithm [Buchbinder et al., 2012]. The problem of maximizing a monotone submodular function subject to a cardinality constraint admits a 1 - 1/e approximation algorithm [Nemhauser et al., 1978]. The maximum coverage problem is a special case of this problem. The more general problem of maximizing a monotone submodular function subject to a matroid constraint also admits a 1 - 1/e approximation algorithm.

9 Graph Theorems and Algorithms

We collect here some of the most important graph algorithms in combinatorial optimization. Recall that a graph is defined as the pair $(V; E)$ where V is the set of vertices and E is the set of edges. If the graph is directed, we will use A to denote the arcs (instead of E for the edges).

Definition 9.1: Walk, Path, Cycle and Circuit

A walk in a graph is simply a collection of vertices and edges (arcs) $v_0, e_1, v_1, \ldots, e_k, v_k$ where $\{v_i : 0 \le i \le k\}$ are vertices and $\{e_i : 1 \le i \le k\}$ are edges (arcs) connecting v_{i-1} and v_i. A path is simply a walk whose vertices (hence edges/arcs) are all distinct. A cycle is a walk with $v_0 = v_k$ and a circuit is a cycle whose vertices, except the first one, are all distinct.

Alert 9.1: Simple Graph

Note that when we say graph we allow it to have multiple edges/arcs between two vertices. If not the case, we will use explicitly the phrase *simple* graph.

Algorithm 9.1: Graph Scanning

Let $\mathcal{G} = (V; E)$ be a graph and s be a vertex. The following procedure scans all vertices and edges/arcs reachable from s.

Start with $Q = \{s\}, R = \{s\}, T = \emptyset$.

(I). If $Q = \emptyset$ stop; otherwise choose v from Q.

(II). Choose $w \in V - R$ with $vw \in E$ and set $Q \leftarrow Q \cup \{w\}, R \leftarrow R \cup \{w\}, T \leftarrow T \cup vw$; If there is no such w, $Q \leftarrow Q - \{v\}$.

(III). Repeat.

When the algorithm stops, R is the set of vertices reachable from s and T is the set of edges/arcs reachable from s. The overall complexity is $O(|V| + |E|)$ since each vertex/arc is scanned at most once.

Algorithm 9.2: Breadth First Search

This is a specialization of Algorithm 9.1, where the query set Q follows first in first out.

Fix a vertex s and denote V_i as the set of vertices that has i unit distance to s. Clearly $V_0 = \{s\}$, and $V_{i+1} = \{v \in V - \bigcup_{j=1}^{i} V_i : \exists u \in V_i \text{ such that } uv \in E\}$. The partition $V = \sum_i V_i$ can be performed in time $O(|V| + |E|)$.

Clearly we can use breadth first search to find a shortest path from the vertex s to any other vertex t: Simply check $j := \min\{i : t \in V_i\}$. The V_i's also indicate all the distances from s to other vertices.

Algorithm 9.3: Depth First Search

This is another specialization of Algorithm 9.1, where the query set Q follows first in last out.

Given the vertex s we can also perform a depth first search by recursion, using the scan operator:

$\mathsf{scan}(v)$: Delete all edges/arcs pointing to v. For each $u \in V$ such that $vu \in E$, $\mathsf{scan}(u)$.

We order the vertices by

$$v_i \succ v_j : \mathsf{scan}(v_i) \text{ finishes earlier than } \mathsf{scan}(v_j),$$

which is a strict partial ordering on all vertices that are reachable by s. Moreover, this order enjoys the property that if $v_i \succ v_j$ and v_j is reachable by v_i, then v_i must be reachable by v_j too, i.e., there exists a circuit in the graph. It is easy to count that the complexity of determining this ordering is $O(m)$, where m is the number of edges/arcs reachable by s.

By adding an extra vertex t to the graph and connecting it to all vertices, $\mathsf{scan}(t)$ will give us an ordering of all vertices. Clearly the complexity is $O(|V| + |E|)$.

Algorithm 9.4: Strong Components

To find the strong components in a digraph, we first order the vertices such that $v_1 \succ \ldots \succ v_n$. Scan the largest vertex v_1 and let V_1 be the set of vertices reachable by it. Due to the property of the ordering, V_1 is exactly the component containing v_1. Delete from the graph the component V_1 and repeat the procedure (no need to reorder the remaining vertices). The overall complexity is dominated by ordering, that is $O(|V| + |A|)$.

Definition 9.2: Eulerian Cycle

An Eulerian cycle in a (di)graph is a cycle which transverses each edge/arc exactly once. It is well-known that an undirected graph has an Eulerian cycle iff each vertex has even degree, while a directed graph has an Eulerian cycle iff for each vertex its in-degree and out-degree agree.

Algorithm 9.5: Eulerian Cycle

Choose a non-isolated vertex, make a walk as long as possible so that no edge/arc is transversed twice. If there is an Eulerian cycle we must end in the starting vertex again. Delete the transversed edges/arcs and repeat the procedure. Since each edge is transversed at most once, the complexity is $O(|V| + |E|)$.

Algorithm 9.6: Dijkstra's Shortest Path

Given a digraph $\mathcal{G} = (V; A)$ and a nonnegative length function $\ell : A \to \mathbb{R}_+$, we want to find a shortest path from the vertex s to another vertex t.

Start with the distance vector $d \in \mathbb{R}_+^V$ with $d(s) = 0$ and $d(v) = \infty, \forall v \neq s$. Set $U = V$ and repeat the following: Find $u \in U$ such that $d(u)$ is minimal over $u \in U$. For each $uv \in A$ set $d(v) = d(v) \wedge (d(u) + \ell(uv))$. Delete u from U.

Clearly the complexity is at most $O(|V|^2)$.

Using induction it is clear that the distance vector d maintained by the algorithm is always an upper bound of the true distance dist in the graph. In each iteration, we claim that $d(u) = \mathsf{dist}(u)$. Use induction: $d(s) = 0 = \mathsf{dist}(s)$. Suppose $d(u) > \mathsf{dist}(u)$. Let $s = v_0, \ldots, v_k = u$ be a shortest $s - u$ path and $i \geq 1$ be the smallest index such that $v_i \in U$. Then the fact that v_{i-1} has been chosen previously leads to $d(v_i) \leq d(v_{i-1}) + \ell(v_{i-1}v_i) = \mathsf{dist}(v_{i-1}) + \ell(v_{i-1}v_i) = \mathsf{dist}(v_i) \leq \mathsf{dist}(u) < d(u)$, contradicting our choice of $u \in U$. Note that the nonnegativity of the length function is used to derive the second inequality.

By using a fancier data structure (the Fibonacci heap), Dijkstra's algorithm can be accelerated to

$O(|A| + |V| \log |V|)$,i.e., almost linear time.

Algorithm 9.7: Bellman-Ford Shortest Path

This is an instance of dynamic programming. Note first that it is NP-hard to find the shortest path between two vertices in a digraph, when the length function is arbitrary (reduction from Hamiltonian path). However, if there is no negative-length directed circuit that is reachable by s, the following algorithm will do.

Fix $s \in V$ and denote $d_k \in \mathbb{R}^V$ as the distance vector where $d_k(v)$ is the minimum length of $s - v$ paths transversing at most k edges. Clearly $d_0(s) = 0$ and $d_0(v) = \infty, \forall v \neq s$. Given d_k, d_{k+1} is given as $d_{k+1}(v) = d_k(v) \wedge (\min_{uv \in A} d_k(u) + \ell(uv))$.

The complexity is easily seen to be $O(|V||A|)$. Moreover, if there exists a negative-length directed circuit that is reachable by s, by comparing the distance vectors $d_{|V|}$ with $d_{|V|-1}$, we can detect its existence: If $d_{|V|} = d_{|V|-1}$, denote the circuit reachable by s as $v_0, v_1, \ldots, v_k = v_0$, then $d_{|V|}(v_i) \leq d_{|V|-1}(v_{i-1}) + \ell(v_{i-1}v_i)$ for all $1 \leq i \leq k$. Summing up we get that the circuit must have nonnegative length. Conversely, if $\exists t$ such that $d_{|V|}(t) < d_{|V|-1}(t)$ then there exists a shortest $s - t$ path which transverses $|V|$ edges, i.e., there exists a circuit. Of course, the circuit can only have negative length.

Definition 9.3: Potential

Let $\mathcal{G} := (V; A)$ be a digraph with the length function $\ell : A \to \mathbb{R}$. The function $p : V \to \mathbb{R}$ is called a potential for $\mathcal{G}$ if

$$\forall uv \in A, \ell(uv) \geq p(v) - p(u).$$

Add a new vertex s to the graph and compute the shortest distance of each vertex to it. The resulting distance function is trivially a potential for the graph. Conversely, given a potential p, we can change the length function to $\ell(uv) - p(v) + p(u)$ to make it nonnegative, without affecting the relative lengths of paths between any pair of vertices. When a potential can be found cheaply, we will apply the previous trick so that the cheaper Dijkstra's algorithm can be employed to find the shortest path.

Algorithm 9.8: Floyd-Warshall Shortest Path

To compute the shortest distances between all pairs of vertices in a digraph (with arbitrary length function ℓ but without negative-length circuits), we proceed as follows. Arbitrarily order the vertices as $v_1, \ldots, v_n$. Denote $D_k \in \mathbb{R}^{V \times V}$ as the distance matrix where $D_k(s, t)$ is the minimum length of $s - t$ paths using only vertices $\{s, v_1, \ldots, v_k, t\}$. Define $D_0(s, t) = \ell(s, t)$, if $st \in A$ otherwise $D_0(s, t) = \infty$. Given D_k, $D_{k+1}(s, t) = D_k(s, t) \wedge (D_k(s, v_{k+1}) + D_k(v_{k+1}, t))$. This is yet another instance of dynamic programming.

The complexity is $O(|V|^3)$. However, we can easily get a faster algorithm: Using Bellman-Ford algorithm to compute a potential and then employ Dijkstra's algorithm for each vertex. This gives us $O(|V|(|A| + |V| \log |V|)$.

Using the Floyd-Marshall algorithm, it is easy to find a minimum length circuit (provided that there is no negative-length circuit): Simply solve $\min_{s \in V} \min_{t \in V} D_n(s, t) + D_n(t, s)$, where $n := |V|$.

Algorithm 9.9: Karp's Minimum Average-Length Directed Circuit

Given a digraph $\mathcal{G} = (V; A)$ (with arbitrary length function ℓ), we want to find a directed circuit whose average length is minimal. Define $\forall v \in V$, $d_0(v) = 0$, and $d_k(v)$ as the minimum length of walks ending in v with *exactly* k arcs. That is

$$d_{k+1}(v) = \min_{uv \in A} d_k(u) + \ell(uv).$$

Set $d_{k+1}(v) = \infty$ if no such walk exists. Denote $n := |V|$.

We prove that the minimum average length of *cycles* is equal to

$$\min_{v\in V} \max_{0\le k\le n-1: d_k(v)<\infty} \frac{d_n(v) - d_k(v)}{n-k}. \tag{136}$$

Note first that if there is no cycle (i.e., the minimum average length is ∞), then $d_n \equiv \infty$ hence the formula is correct. By subtracting some constant from each arc, we can assume w.l.o.g. the minimum average length of cycles is 0. Let the minimum in (136) be attained by u. It is easy to see that the minimum is nonnegative: If $d_n(u) = \infty$ this is trivially true, otherwise the walk that leads to $d_n(u)$ must be composed of a walk with k arcs ending at u and a nonnegative-length cycle C with $n-k$ arcs.

Conversely, let $C := \{v_0, v_1, \dots, v_t = v_0\}$ be a zero length cycle. Then $\arg\min_r d_r(v_0)$ can be chosen in the interval $n-t \le r < n$ (since $d_n(v_0)$, containing a nonnegative length cycle, can be safely ignored). Fix such an r and split C into $P := \{v_0, \dots, v_{n-r}\}$ and $Q := \{v_{n-r}, \dots, v_t = v_0\}$. We have $d_k(v_{n-r}) + \ell(Q) \ge d_{k+(t-(n-r))}(v_0) \ge d_r(v_0) \ge d_n(v_{n-r}) - \ell(P)$, hence $d_n(v_{n-r}) - d_k(v_{n-r}) \le \ell(C) = 0$.

Algorithmically, we compute the distance vector d_n in time $O(|V||A|)$; find the minimizer $u \in V$ which achieves the minimum in (136); and finally any circuit in the walk that leads to $d_n(u)$ is the one we are looking for: By subtracting the circuit we get an upper bound on $d_k(u)$ for some k, which in turn is an upper bound on $d_n(u)$ due to the minimality of u. The overall complexity is dominated by the first step hence $O(|V||A|)$.

Recall that for a collection of subsets $\mathcal{Q} = \{\emptyset \ne P_i \subseteq \Omega : i \in I\}$, $p_i \in \Omega, i \in I$ is called a system of representation for $\mathcal{Q}$ if there exists a bijection $\pi : I \to I$ such that $p_i \in P_{\pi(i)}$. Note that we might have $p_i = p_j$ for some $i, j \in I$, otherwise it will be called a system of distinct representation.

Theorem 9.1: Hall-Rado Theorem

Let $F : 2^\Omega \to \mathbb{R}$ be a polymatroid function and $\mathcal{Q} = \{\emptyset \ne P_i \subseteq \Omega : i \in I\}$. Fix $d \in \mathbb{N}$, then $\mathcal{Q}$ has a system of representation $p_i, i \in \mathcal{I}$ with $F(\bigcup_{i\in J}\{p_i\}) \ge |J| + d, \forall J \subseteq I$ iff

$$F\left(\bigcup_{i\in J} P_i\right) \ge |J| + d, \forall J \subseteq I. \tag{137}$$

Proof: $\Rightarrow$: Trivial, since $\bigcup_{i\in J}\{p_i\} \subseteq \bigcup_{i\in J} P_i$.

$\Leftarrow$: If $|P_i| = 1, \forall i \in I$, then the implication is trivial. Otherwise let, say, $|P_{i_1}| > 1$, hence there exist $x_1, x_2 \in P_{i_1}$ and $x_1 \ne x_2$. We claim that $\exists k \in \{1, 2\}$ such that the system $\{\hat{P}_{i_1} := P_{i_1} - \{x_k\}, \hat{P}_i := P_i, i_1 \ne i \in I\}$ satisfies (137). Indeed, suppose not, then there exist $J_1 \cup J_2 \not\ni i_1$ such that

$$F\big((P_{i_1} - \{x_k\}) \cup P(J_k)\big) \le |J_k| + d,$$

where we introduce the notation $P(J) := \bigcup_{i\in J} P_i$. But since F is polymatroid, we have

$$\begin{aligned} F\Big((P_{i_1} - \{x_1\}) \cup P(J_1)\Big) + F\Big((P_{i_1} - \{x_2\}) \cup P(J_2)\Big) &\ge F\Big(P(\{i_1\} \cup J_1 \cup J_2)\Big) + F\Big(P(J_1 \cap J_2)\Big) \\ &\ge |J_1 \cup J_2| + 1 + |J_1 \cap J_2| + 2d \\ &= |J_1| + |J_2| + 1 + 2d, \end{aligned}$$

contradiction. Therefore we can keep deleting elements so that $|P_i| = 1, \forall i \in I$. ■

The case with $d = 0$ is mostly interesting. If we let $F = |\cdot|$ then the system of representation is forced to be distinct, and (137) becomes a sufficient and necessary condition for the existence of transversals, usually known as Hall's Marriage Theorem (marrying p_i to $P_{\pi(i)}$, one husband and one wife!). If we let F be the rank function of some matroid, then (137) becomes a sufficient and necessary condition for the existence of transversals satisfying $\{p_i : i \in J\} \in \mathcal{I}, \forall J \subseteq I$.

We add a few other theorems that each is related to Hall's and König's theorems.

Theorem 9.2: Birkhoff Theorem

Let $S \in \mathbb{R}_+^{n\times n}$ be a doubly stochastic matrix (i.e., each row and column sums to 1). Then S can be written as a convex combination of permutation matrices.

Proof: For each $i \in \Omega := \{1, \dots, n\}$, define $P_i := \{j \in \Omega : S_{ij} > 0\}$, which is non-empty due to the assumption $\sum_j S_{ij} = 1$. Consider any subset $J \subseteq \Omega$. Denote $P(J)$ as $\bigcup_{j\in J} P_j$, we have

$$|P(J)| = \sum_{i=1}^{n} \sum_{j\in P(J)} S_{ij} \geq \sum_{i\in J} \sum_{j\in P_i} S_{ij} = |J|,$$

where the first equality is due to $\sum_i S_{ij} = 1$ and the last equality follows from the definition of P_i. Apply Hado's theorem we get a transversal from Ω to $\{P_i\}_{i\in\Omega}$. The transversal can be represented as a permutation matrix whose multiple subtracted from S will zero out one more element in S. Iterating the above procedure until S becomes 0. ■

It is not clear if Hall's theorem (or any of the above equivalents) follows *directly* from Birkhoff's Theorem, although for a regular (i.e., each node has the same degree) bipartite graph, the incidence matrix is doubly stochastic (after padding zeros) hence Hall's theorem does follow from Birkhoff's in this special case.

Theorem 9.3: König's Theorem

Let $\mathcal{G} = (U, V; E)$ be a bipartite graph, then

$$\min_{C \text{ is a vertex cover}} |C| \quad = \quad \max_{M \text{ is a matching}} |M|. \tag{138}$$

Proof: The inequality $\geq$ is easily seen to be true for any graph. Now let C be a minimum cover with $C \cap U = X, C \cap V = Y$. For each $x \in X$, define $V_x := \{v \in V - Y : xv \in E\}$. The minimality of C allows us to apply Hall's theorem to $\{V_x\}_{x\in X}$ hence we can find a matching M_x that connects X to $V - Y$. Similarly another matching M_y that connects Y to $U - X$ exists. Apparently $M_x \cup M_y$ is a matching that has the same size as the minimum cover C. ■

Conversely, Hall's theorem follows from König's theorem: Build the bipartite graph $\mathcal{G} := (U, V; E)$ where U is the disjoint union of $\{p_i\}$, V is the collection of $\{P_i\}$, and $uv \in E$ iff $u \in v$. Let C be a minimum cover and $X = U - C$. All the edges between X and V, whose number denoted as k, must be covered by nodes in V, collectively denoted as Y. Hence $k \geq |\bigcup_{i\in Y} P_i| \geq |Y| = |X|$. Therefore $|C| \geq |U|$, and by König's theorem a transversal exists.

Algorithm 9.10: Bipartite Maximum-Size Matching/Minimum-Size Vertex Cover

To get a maximum size matching in a bipartite graph $\mathcal{G}_b = (U, V; E)$, start with a *maximal* matching M. Let U_M denote the vertices in U that are *not* covered by M. Similarly define V_M. Build a *digraph* $\mathcal{G}_M = (U, V; D)$ where the edges in M oriented from V to U while the rest edges in E oriented from U to V. Find a (directed) path from U_M to V_M and augment it with M (by taking their symmetric difference). Repeat until this no augmenting path. The overall complexity is clearly $O(\nu \cdot |E|)$, where ν is the size of a maximum matching.

A better algorithm is to add a source s and point it to U, add a sink t and point V to it, and orient edges in E from U to V. Using Algorithm 9.13 below to find the maximum number of vertex disjoint $s-t$ paths, which naturally lead to a maximum matching. The overall complexity is brought down to $O(\nu^{1/2} \cdot |E|)$ (König's theorem equates ν with the minimum size of a vertex cover).

After getting a maximum-size matching M, we can construct a minimum-size vertex cover as follows: As before build the digraph $\mathcal{G}_M$. Denote R_M as the set of vertices in $\mathcal{G}_M$ that can be

reached from some vertex in U_M. By the maximality of M we know $R_M \cap V_M = \emptyset$. Moreover there is no edge connecting $U \cap R_M$ and $V - R_M$ (for otherwise the latter will be reachable). Hence $C := (U - R_M) \cup (V \cap R_M)$ is a vertex cover whose vertices are covered by the matching M. However, the vertices on the same edge in M cannot both be present in C (because then both vertices are reachable, violating the definition of C). Therefore $|C| \leq |M|$, i.e., C is a minimum-size vertex cover. The complexity is dominated by finding the maximum matching, hence can be bounded by $O(\tau^{1/2} \cdot |E|)$, where τ is the size of a minimum vertex cover.

Note that the above construction does not generalize to the weighted setting, although minimum-weight vertex cover in a bipartite graph can be solved using the linear programming relaxation.

Algorithm 9.11: Bipartite Maximum-Weight Matching

To find a maximum-weight matching in any bipartite graph, we can use the Hungarian method: Start with the empty matching M in the bipartite graph $\mathcal{G}_b = (U, V; E)$. Let U_M denote the vertices in U that are *not* covered by M. Similarly define V_M. As before build a digraph $\mathcal{G}_M = (U, V; D)$ where the edges in M oriented from V to U with length equal to the weight while the rest edges in E oriented from U to V with length equal to the *negation* of the weight. Among all the paths (if any) in $\mathcal{G}_M$ that start from U_M and end in V_M, pick a shortest one and augment M with it. Iterate the procedure. Using induction it can easily be shown that M is always a maximum-weight matching among all matchings with size $|M|$. We terminate when the shortest path we find has positive length. If we want to find a maximum-weight *perfect* matching, we stop until all vertices are covered. The complexity can be bounded as $O(\nu \cdot |V||E|)$ if we use the Bellman-Ford algorithm (cf. Algorithm 9.7) to find shortest paths, where ν is the minimum size of a maximum-weight matching.

A refined algorithm can maintain a potential (cf. Definition 9.3) so that the faster Dijkstra's algorithm (cf. Algorithm 9.6) can be applied to find shortest paths. Indeed, denote R_M as the set of vertices in $\mathcal{G}_M$ that can be reached from some vertex in U_M. Initially, when $M = \emptyset$, we can set the potential p as $p(v) := \max_{e \in E} w_e$ if $v \in U$ and $p(v) = 0$ otherwise. This gives us a potential for the graph $\mathcal{G}_M(R_M)$, the restriction of the graph $\mathcal{G}_M$ to the vertex set R_M. Then for each $v \in R_M$ we calculate $\hat{p}(v) = \mathsf{dist}(U_M, v)$ using Dijkstra's algorithm. We find a shortest $U_M - V_M$ path P and obtain the bigger matching $\hat{M} = M \Delta P$. We claim that $\hat{p}$ is a potential for $\mathcal{G}_{\hat{M}}(R_{\hat{M}})$: Clearly $U_{\hat{M}} \subseteq U_M$. Moreover $R_{\hat{M}} \subseteq R_M$, for otherwise there exists a path starting from $U_{\hat{M}}$ hence U_M and ending in $R_{\hat{M}}$ but not R_M. Each arc in $\mathcal{G}_M$ does not leave R_M therefore there must be an arc in P that leaves R_M, which impossible by the definition of R_M. So $\hat{p}$, defined on R_M, also covers $R_{\hat{M}}$. Now take an arc $uv \in \mathcal{G}_{\hat{M}}(R_{\hat{M}})$. Either $uv \in \mathcal{G}_M$ hence $\ell(uv) \geq \hat{p}(v) - \hat{p}(u)$ due to the definiton of $\hat{p}$, or $vu \in P$, the augmenting path, hence also $\hat{p}(v) - \hat{p}(u) = -\ell(vu) = \ell(uv)$ since P is shortest. In conclusion, $\hat{p}$ is a potential for $\mathcal{G}_{\hat{M}}(R_{\hat{M}})$. The overall complexity is thus $O(\nu \cdot (|E| + |V| \log |V|))$, where ν is the minimum size of a maximum-weight matching.

Theorem 9.4: Gallai's Theorem

For any graph without isolated vertices the sum of the maximum size of an independent set and the minimum size of a vertex cover, equals the sum of the maximum size of a matching and the minimum size of an edge cover, equals the number of vertices.

Proof: Consider a matching M, augmenting it with edges that cover vertices not covered by M gives us an edge cover whose size is $|V| - 2|M| + |M| = |V| - |M|$. Conversely, for any edge cover C, find a maximum matching M in it. Denote the vertices in V that are not covered by M as X. Then $|X| \leq |C| - |M|$, since every vertex is covered by C while no two vertices connected in C are missed by M. Therefore $|X| = |V| - 2|M| \leq |C| - |M|$. Finally note that a vertex set is independent iff its complement is a vertex cover. ■

It is clear from the proof that if we can find a maximum matching then we can easily grow from it a minimum edge cover. Also if we can find a maximum independent set in the complement graph then we get a minimum vertex cover in the original graph, and vice versa (although it is unlikely to have polynomial-time algorithms for them).

Algorithm 9.12: Minimum-Weight Edge Cover

We reduce the minimum-weight edge cover problem (in any graph without isolated vertices) to minimum-weight perfect matching: Simply build a copy of the graph, and for each vertex v, connect it to its twin sister v' with weight $2w_v = \min_{e \ni v} w(e)$. Find a minimum-weight *perfect* matching (which exists by construction) in the twin-graph, retain edges within one copy of the graph and replace the crossing edge with a lightest edge that resides in the chosen copy of the twin-graph. Since the matching is perfect, we get an edge cover whose weight is half of the matching. Conversely, any edge cover induces a perfect matching in the twin-graph whose weight is at most twice bigger (provided that the weights are nonnegative). The complexity is $O(|V| \cdot (|E| + |V| \log |V|))$.

Theorem 9.5: Dilworth's Theorem

Let $(P, \preceq)$ be a partially ordered set. The minimal number m of disjoint chains covering P equals the maximal size M of antichains.

Proof: Clearly $m \geq M$. We prove König $\Rightarrow$ Dilworth: Build the bipartite graph $\mathcal{G} = (P, P; E)$ where $ab \in E$ iff $a \preceq b$. A chain of length n in $(P, \preceq)$ corresponds to a matching with size $n-1$ in $\mathcal{G}$. Therefore a covering of disjoint j chains in P corresponds to a matching with size $|P| - j$, and vice versa. Now consider a minimum covering of m disjoint chains, which corresponds to a maximum matching in $\mathcal{G}$ with size $|P| - m$. By König's theorem, there exists a minimum vertex cover in $\mathcal{G}$ with size $|P| - m$. The rest of the vertexes that are not in the minimum vertex cover form a antichain with size m. ■

Conversely, we show Dilworth $\Rightarrow$ Hall: Let $P_i \subseteq \Omega, i \in I$ satisfy (137) (with $d = 0$, $F =$ card). Build the partially ordered set $(P := \bigcup_{i \in I} \{P_i\} \cup \Omega, \preceq)$ where $\omega \preceq P_i$ iff $\omega \in P_i$ while all other pairs are not comparable. Let $S \cup \{P_i\}_{i \in J}$ be a maximum antichain where $S \subseteq \Omega, J \subseteq I$. By Dilworth's theorem, P can be partitioned into $|S| + |J|$ disjoint chains. In particular, $|J|$ of these chains cover $\{P_i\}_{i \in J}$ hence also cover $\bigcup_{i \in J} P_i$ (because other chains cover S hence cannot cover elements in P_i). Therefore $|J| \geq |\bigcup_{i \in J} P_i|$ (each chain can contain at most one element of Ω) while by assumption $|\bigcup_{i \in J} P_i| \geq |J|$. So we find from these J chains a system of distinct representation for $P_i, i \in J$. On the other hand $\{P_i\}_{i \in I - J}$ are covered by the remaining $|S|$ chains, each containing one element in S. We then easily find a system of distinct representation for $P_i, i \in I - J$. Combing the results completes the proof of Hall's theorem.

Theorem 9.6: Menger's Theorem

The maximum number m of vertex-disjoint $S - T$ paths equals the minimum size n of $S - T$ separating *vertex* sets (possibly containing vertices in S, T).

Proof: Clearly we have $n \geq m$. Consider the digraph $\mathcal{G}_d := (V, D)$ with $S, T \subseteq V$. For each node in $S \cap T$, we create a "twins", one responsible for outward arcs (assigned to S) and one for inward arcs (assigned to T). Add also arcs in both directions between the twins. One can verify that by doing so we do not change $n - m$, hence we can assume $S \cap T = \emptyset$.

For each $v \in V - S$ introduce v^t and for each $v \in V - T$ introduce v^s. Construct the bipartite graph $\mathcal{G}_b = (V^s, V^t; E)$, where for $u^s \in V^s, v^t \in V^t$, $uv \in E$ iff $u = v$ or $uv \in D$. By construction $N := \{v^s v^t : v \in V - (S \cup T)\}$ is a matching in $\mathcal{G}_b$. Take a *maximum* matching M in $\mathcal{G}_b$ with size μ and consider $N \Delta M$, the symmetric difference. Pick a component K of $N \Delta M$, there are three possibilities:

(I). $|K \cap M| < |K \cap N|$. However, by replacing $K \cap M$ with $K \cap N$ in M we would have got a larger matching, contradicting the maximality of M;

(II). $|K \cap M| = |K \cap N|$. Similar as above, by (repeatedly) replacing $K \cap M$ with $K \cap N$ in M we can assume this case does not happen either;

(III). $|K \cap M| > |K \cap N|$. Due to our construction of the special matching N, we know it must be

true that $|K \cap M| = |K \cap N| + 1$ and this component is an $S^s - T^t$ path.

Of course the components of $N \Delta M$ are disconnected. So we have $|M| - |N| = \mu - |V - (S \cup T)|$ vertex-disjoint $S^s - T^t$ paths in $\mathcal{G}_b$, which easily translate to the same number of vertex-disjoint $S - T$ paths in $\mathcal{G}_d$.

Now let $X \subseteq V - T$ and $Y \subseteq V - S$ be such that $C := X^s \cup Y^t$ is a minimum vertex cover in $\mathcal{G}_b$. Denote $U := (X \cap S) \cup (Y \cap T) \cup (X \cap Y)$. Let $P := (v_0, v_1, \ldots, v_k)$ be an $S - T$ path in $\mathcal{G}_d$. W.l.o.g., we assume only $v_0 \in S$ and only $v_k \in T$. The path P induces the path $Q := (v_0^s, v_1^t, v_1^s, \ldots, v_{k-1}^t, v_{k-1}^s, v_k^t)$ in $\mathcal{G}_b$. Since Q has $2k - 1$ edges, the vertex cover C must intersect it in at least k vertices: either $v_0^s \in C$ hence $v_0 \in X \cap S \subseteq U$, or $v_k^t \in C$ hence $v_k \in Y \cap T \subseteq U$, or $v_i^s, v_i^t \in C$ for some $1 \leq i \leq k-1$ hence $v_i \in S \cap T \subseteq U$. In any case the vertex set U intersects each $S - T$ path in $\mathcal{G}_d$, i.e., U is $S - T$ separating. The size of U is

$$\begin{aligned} |U| &= |X \cap S| + |Y \cap T| + |X \cap Y| \\ &= |X \cap S| + |Y \cap T| + |X| + |Y| - |X \cup Y| \\ &= |X| + |Y| - |V - (S \cup T)| \\ &= |C| - |V - (S \cup T)| \\ &= \mu - |V - (S \cup T)|. \end{aligned}$$

The third equality follows from $(X \cup Y) - (S \cup T) = V - (S \cup T)$ since $\forall v \in V - (S \cup T), v^s v^t \in E$ and C is a vertex cover while the last equality follows from König's theorem. Therefore $n \leq m$ thus $m = n$. ■

Two variations of Menger's theorem are also useful.

The maximum number of internally vertex-disjoint paths connecting two distinct non-adjacent vertices s and t equals the minimum size of $s - t$ vertex cuts (not containing s, t).

Proof: Let $S = \{v \in V : sv \in D\}, T = \{v \in V : vt \in D\}$ and delete s, t. ■

The maximum number of arc-disjoint paths connecting two distinct non-adjacent vertices s and t equals the minimum size of $s - t$ arc cuts.

Proof: Consider the line graph and define S and T similarly as above. ■

Algorithm 9.13: Maximum Collection of Arc Disjoint Paths

Given a digraph $\mathcal{G} := (V; A)$ and vertices s and t, we want to find a maximum number of arc disjoint $s - t$ paths. Denote $\mathsf{dist}_{\mathcal{G}}(s, t)$ as the minimum length of an $s - t$ path in $\mathcal{G}$. The idea is to *reverse* a *blocking* collection of arc disjoint $s - t$ paths at a time so that $\mathsf{dist}_{\mathcal{G}}(s, t)$ increases.

Start with $i = 0$ and $\mathcal{G}_0 = \mathcal{G}$.

First we use breadth first search (cf. Algorithm 9.2) to identify the arcs A_i^{st} that appear in some $s - t$ path in $\mathcal{G}_i$: Simply compute the distance d_s to s and the distance d_t to t for each vertex, then arc $uv \in A_i^{st}$ iff $uv \in A(\mathcal{G}_i)$ and $d_s(u) + d_t(v) + 1 = \mathsf{dist}_{\mathcal{G}_i}(s, t)$. Notice that the arcs in A_i^{st} form a DAG.

Next we recursively find a blocking collection of $s - t$ paths in the DAG $\tilde{\mathcal{G}}_i := (V, A_i^{st})$: Use depth first search (cf. Algorithm 9.3) to scan s and stop immediately when t is reached. Delete all arcs that have been scanned (for every arc other than the ones in the found $s - t$ path must have finished scanning, therefore if it is contained in another $s - t$ path, it is reachable by t, i.e. there is a circuit, contradicting the DAG property). Repeat until there is no $s - t$ path. Using induction we see that the paths we find are indeed blocking, i.e., after reversing them in $\tilde{\mathcal{G}}_i$, there is no other $s - t$ path. The complexity for this step is $O(|A_i^{st}|)$.

Reverse the arcs in the blocking collection of $s - t$ paths in $\mathcal{G}_i$, call the new graph $\mathcal{G}_{i+1}$. We show that $\mathsf{dist}_{\mathcal{G}_{i+1}}(s, t) > \mathsf{dist}_{\mathcal{G}_i}(s, t)$. We need the graph $\hat{\mathcal{G}}_i := (V; A(\mathcal{G}_i) \cup (A_i^{st})^{-1})$, obtained by *adding* the reversed arcs in A_i^{st} to the graph $\mathcal{G}_i$. It is not hard to see that $\mathsf{dist}_{\mathcal{G}_i}(s, t) = \mathsf{dist}_{\hat{\mathcal{G}}_i}(s, t)$, hence the reversed arcs in A_{st} cannot appear in any $s - t$ path in $\hat{\mathcal{G}}_i$ (due to the DAG nature of arcs appearing

in shortest paths). Since $\mathcal{G}_{i+1} \subseteq \hat{\mathcal{G}}_i$, we have $\mathsf{dist}_{\mathcal{G}_{i+1}}(s,t) \geq \mathsf{dist}_{\hat{\mathcal{G}}_i}(s,t)$. Let P be an $s-t$ path in $\mathcal{G}_{i+1}$ with length $\mathsf{dist}_{\hat{\mathcal{G}}_i}(s,t)$. Then P is arc disjoint from the (reversed) blocking collection we found in $\tilde{\mathcal{G}}_i$ (otherwise contradicting the DAG property of A_i^{st}). But P is also an $s-t$ path in $\hat{\mathcal{G}}_i$ hence in $\mathcal{G}_i$ hence also in $\tilde{\mathcal{G}}_i$, contradicting to the blocking property. Thus we have strict inequality.

Increment i and repeat until there is no $s-t$ path in the graph, say, $\mathcal{G}_k$. We show how to extract a minimum $s-t$ cut first. Recall that an $s-t$ cut is defined as the set of arcs pointing from $U \ni s$ to $(V-U) \ni t$ for some vertex set $U \subseteq V$. Each time when we reverse an $s-t$ path, the size of *any* $s-t$ cut decreases by 1. At the end of iteration, there is no $s-t$ path left hence by Menger's theorem there exists a zero $s-t$ cut. Therefore the minimum size of an $s-t$ cut equals the times we reverse the $s-t$ paths. Moreover, the set of vertices reachable by s in the final graph $\mathcal{G}_k$ consists of a minimum $s-t$ cut.

Finally we show that the arcs reversed in the final graph $\mathcal{G}_k$ consists of a maximum collection of arc disjoint $s-t$ paths in the graph $\mathcal{G}$. Let R_j be the arcs remaining reversed in $\mathcal{G}$ when we reverse the j-th $s-t$ path. Using induction we can easily show that the graph $(V; R_j)$ with j direct arcs from s to t added is Eulerian. Therefore we can find as many $s-t$ arc disjoint paths in $\mathcal{G}_k$ as the times we reverse $s-t$ paths in $\mathcal{G}$. But this number of arc disjoint paths coincides with the minimum size of an $s-t$ cut we found above, hence by Menger's theorem it is maximum. In fact we have proved something stronger: Each time we reverse an $s-t$ path, we add to our repository one more arc disjoint $s-t$ path.

Note that we choose to reverse arcs instead of deleting them, because the graph in general can have multiple arcs between vertices. The overall complexity of the above procedure is $O(k \cdot |A|)$, where $k \leq \mathsf{dist}_{\mathcal{G}}(s,t))$ is the terminating index. Clearly an easy upper bound is $O(|V| \cdot |A|)$. On the other hand, let $p := \lfloor |A|^{1/2} \rfloor$, then each $s-t$ path in $\mathcal{G}_p$ has length at least $p+1 \geq |A|^{1/2}$, i.e., there are at most $|A|/(p+1) \approx |A|^{1/2}$ arc disjoint $s-t$ paths in $\mathcal{G}_p$. Therefore there are at most $|A|^{1/2}$ iterations after $\mathcal{G}_p$ since each iteration afterwards increases the cut in $\mathcal{G}_p$ by at least 1 while by Menger's theorem the minimum size of a cut equals the maximum number of arc disjoint paths. Hence $k \leq 2 \cdot |A|^{1/2}$, i.e., the complexity of the procedure is also bounded by $O(|A|^{3/2})$. Furthermore, if $\mathcal{G}$ is *simple*, we can tighten the analysis. Let $p = \lfloor |V|^{2/3} \rfloor$. Denote U_i as the set of vertices in $\mathcal{G}_p$ that has i-unit distance to s. Clearly $\sum_{i=1}^{p} |U_i| + |U_{i+1}| \leq 2 \cdot |V|$ hence exists some j such that $|U_j| + |U_{j+1}| \leq 2 \cdot |V|^{1/3}$, i.e., $|U_j| \leq |V|^{2/3}$ for some j. Since the graph is simple, no two arc disjoint paths can share vertices. Then $\mathcal{G}_p$ has at most $|V|^{2/3}$ arc disjoint paths, therefore the number of iterations after $\mathcal{G}_p$ is at most $|V|^{2/3}$, i.e., the complexity is bounded by $O(|V|^{2/3}|A|)$ for simple graphs.

To find a maximum collection of vertex disjoint $s-t$ paths, we split each vertex v into v_1 and v_2 where v_1v_2 is added to the arc set while any $uv \in A$ is replaced by u_2v_1. A maximum collection of arc disjoint $s_2 - t_1$ paths in the new graph suffices for our purpose. However, we can tighten the complexity analysis a bit by exploiting the special structure of the new graph: Denote τ as the size of a minimum vertex cover C in $\mathcal{G}$. Set $p = \lfloor \tau^{1/2} \rfloor$. Each vertex disjoint $s_2 - t_1$ path in $\mathcal{G}_p$ has length at least $p+1$ hence contains at least $p/2$ vertices in C, therefore there are at most $2\tau^{1/2}$ vertex disjoint paths in $\mathcal{G}_p$. It follows that the maximum number of iterations after $\mathcal{G}_p$ is at most $2\tau^{1/2}$, i.e., the complexity is bounded by $O(\tau^{1/2}|A|) \leq O(|V|^{1/2}|A|)$.

Theorem 9.7: Ford-Fulkerson Theorem

For any network (a digraph with a single source and a single sink), the minimum cut equals the maximum flow.

Proof: Let us assume the capacities are all integral (the general case will be handled in **??**). Replace each arc with capacity c to c arcs, all pointing to the original direction. Clearly by construction, the maximum number of $s-t$ arc-disjoint paths equals the (value of the) maximum flow while the minimum number of arcs separating $s-t$ equals the (value of the) minimum cut. The theorem follows from (the edge version of) Menger's theorem. ■

Conversely, Ford-Fulkerson theorem easily implies König's theorem: Let $\mathcal{G} = (U, V; E)$ be a bipartite graph. Add a source vertex s to U and a sink vertex t to V. Put infinite (or a sufficiently

large number) capacity on each edge between U and V and put unit capacity from the source to U and similarly unit capacity from V to the sink t. Clearly, under our construction a maximum matching corresponds to a maximum flow. On the other hand, given a minimum cut $C = C_s \cup C_t$ with value c (note that $c \leq |U| \wedge |V|$ hence is finite), we define $U_s := U \cap C_s$ and $V_s := V \cap C_s$. Then we verify that $(U - U_s) \cup V_s$ form a vertex cover for $\mathcal{G}$ with value c, since a minimum cut cannot afford any edge between $(U_s, V - V_s)$ and also $(U - U_s, V_s)$. Similarly, given a minimum vertex cover, we can find a cut with the same value. Therefore (the integral version of) Ford-Fulkerson theorem implies König's theorem.

Algorithm 9.14: Edmonds-Karp Algorithm for Max-Flow

The algorithm is extremely simple: Start with any flow, build the residual graph, find a shortest path, augment the flow and repeat. Note that if we find an arbitrary path instead the algorithm might not be correct for irrational capacities although it remains correct for rational capacities at the price of exponential slow down (in the extreme case) [**?**, page 152 & figure 10.1].

To see the correctness: Let f be the current flow, $\mathcal{G}_f$ be the residual graph. Denote $A(\mathcal{G}_f)$ as the set of arcs in $\mathcal{G}_f$ which are on at least one shortest path in $\mathcal{G}_f$. Reverse the arcs in $A(\mathcal{G}_f)$ and add them to $\mathcal{G}_f$, call it $\hat{\mathcal{G}}_f$. Note that we have $A(\mathcal{G}_f) = A(\hat{\mathcal{G}}_f)$ since the added arcs in $\hat{\mathcal{G}}_f$ clearly cannot on any shortest path. Consequently $\ell(\mathcal{G}_f) = \ell(\hat{\mathcal{G}}_f)$, i.e., the length of the shortest paths does not change. After augmenting f with a shortest path P, we get a bigger flow g whose residual graph $\mathcal{G}_g$ is a subgraph of $\hat{\mathcal{G}}_f$, hence $\ell(\mathcal{G}_g) \geq \ell(\hat{\mathcal{G}}_f)$. If we actually have equality, then $A(\mathcal{G}_g) \subseteq A(\hat{\mathcal{G}}_f) = A(\mathcal{G}_f)$. The inclusion is in fact proper since at least one arc in P is no longer in $\mathcal{G}_g$ after the augmentation. In summary, after each augmentation, we either increase the length of the shortest path of the residual graph by one or we lose one edge in $A(\mathcal{G}_f)$. Therefore the algorithm will terminate after at most $|V| \cdot |A|$ steps.

The overall complexity of the Edmonds-Karp algorithm is $O(|V| \cdot |A|^2)$, and the best algorithm so far achieves $O(|V| \cdot |E|)$ [**?**].

Algorithm 9.15: Tardos' Minimum Cost Circulation

TO BE ADDED.

Theorem 9.8: Tutte-Berge Formula

For an arbitrary graph $\mathcal{G} = (V, E)$, the maximum size $m(\mathcal{G})$ of its matchings satisfies

$$m(\mathcal{G}) = \min_{U \subseteq V} \frac{1}{2} \left[|U| + |V| - o(\mathcal{G} - U) \right], \tag{139}$$

where $\mathcal{G} - U$ is the graph with all nodes in U removed and $o(\mathcal{G})$ denotes the number of components of $\mathcal{G}$ that have odd number of vertices.

Proof: Clearly $\forall U \subseteq V$, $m \leq |U| + m(\mathcal{G} - U) \leq |U| + \frac{1}{2}[|V - U| - o(\mathcal{G} - U)] = \frac{1}{2}[|U| + |V| - o(\mathcal{G} - U)]$.

For the other direction, we do induction on $|V|$. The case $V = \emptyset$ trivially holds and we assume $\mathcal{G}$ is connected (otherwise apply the induction hypothesis to each component). If there exists a vertex v that is contained in every matching, then we can delete the vertex and apply the induction hypothesis. If we prove by contradiction that there exists a maximum matching that misses at most one node, then $m(\mathcal{G}) \geq \frac{1}{2}[|V| - o(\mathcal{G})]$ while the r.h.s. of (139) is at most $\frac{1}{2}[|V| - o(\mathcal{G})]$ (by setting $U = \emptyset$), hence the theorem will be proved.

Indeed, suppose every maximum matching misses at least two distinct nodes. Let

$$(u, v) = \arg \min_{M, p \notin M, q \notin N, p \neq q} \operatorname{dist}(p, q),$$

where M ranges over all maximum matchings while $\operatorname{dist}(p, q)$ denotes the distance between node p and q in $\mathcal{G}$. Note that $\operatorname{dist}(u, v) \geq 2$ for otherwise we can augment M to get a bigger matching.

Choose a distinct third node w on the shortest path from u to v and select among maximum matchings which do not contain w the one intersects M *most*, dented as N. By the minimality of (u, v) (in the sense of dist), N must contain both u and v. Therefore there exists $z \neq w$ that is covered by M but not N (for they have the same size). Let zx be the edge contained in M. Then there exists an edge xy contained in N (for otherwise we can augment N with zx). But replacing N by $N - xy + zx$ increases the intersection of N with M, contradiction to the maximality of N. ■

Algorithm 9.16: Maximum-Size Matching

We describe Edmonds' maximum-size matching algorithm for any graph.

Theorem 9.9: Brualdi Formula

Theorem 9.10: Edmonds-Galai Decomposition

Algorithm 9.17: Maximum-Weight Matching

We describe Edmonds' maximum-size matching algorithm for any graph.

10 Matroid

Definition 10.1: Matroid by Independent Sets

Let Ω be a nonempty finite set (called the ground set) and call the pair $\mathfrak{M} = (\Omega, \mathcal{I})$ matroid, where $\mathcal{I} \subseteq 2^\Omega$ define the *independent sets*:

(I). (Non-empty) $\emptyset \in \mathcal{I}$;

(II). (Inheritable) $J \in \mathcal{I} \implies \mathcal{I} \ni I \subseteq J$;

(III). (Augmentable) If $I \in \mathcal{I}, J \in \mathcal{I}$ and $|J| > |I|$, then $\exists \omega \in J - I$ such that $I \cup \{\omega\} \in \mathcal{I}$.

The temptation to assume $\Omega = \cup\mathcal{I}$ must be resisted for good reasons (cf. Definition 10.11 below). The last property of $\mathcal{I}$ allows us to consider *maximal* independent sets, which will be called basis. It should be clear that all bases are equipotent, and their common cardinality is called the rank of the matroid. (Do not confuse the rank of the matroid with the rank function that is defined below.)

Remark 10.1: Weakening the Augmentable Property

For a nonempty, inheritable collection of sets $\mathcal{I}$, the augmentable property can be weakened to

(III'). $I \in \mathcal{I}, J \in \mathcal{I}, |I - J| = 1, |J - I| = 2 \implies \exists \omega \in J - I$ s.t. $I \cup \{\omega\} \in \mathcal{I}$.

Proof: Take $I \in \mathcal{I}, J \in \mathcal{I}, |J| > |I|$ and we show (III') $\implies$ (III) in Definition 10.1.
Induction on $|I - J|$:

- $|I - J| = 0$: Trivial.
- $|I - J| = 1$: Take a subset of J and apply (III');
- $|I - J| = k + 1$: Consider $I - \{i\}$ where $i \in I - J$. Note that $|J| > |I| > |I - \{i\}|$ while $|I - \{i\} - J| = k$ hence we can apply the induction hypothesis to conclude that $\exists j \in J - I$ such that $(I - \{i\}) \cup \{j\} \in \mathcal{I}$. Now consider $(I - \{i\}) \cup \{j\}$ and J, we have $|(I - \{i\}) \cup \{j\} - J| = k$ while $|J| > |(I - \{i\}) \cup \{j\}|$. Therefore we can apply the induction hypothesis again to conclude

that $\exists j' \in J - I - \{j\}$ such that $(I - \{i\}) \cup \{j, j'\} \in \mathcal{I}$. Finally consider $(I - \{i\}) \cup \{j, j'\}$ and I, apply (III') we get either $I \cup \{j\} \in \mathcal{I}$ or $I \cup \{j'\} \in \mathcal{I}$, finishing the induction.

■

As will be seen, (III') is usually more convenient than (III) to check.

Definition 10.2: Polymatroid and Rank Function

We call any positive integer valued, nondecreasing, submodular function F^0 with $F^0(\emptyset) = 0$ polymatroid function. Any polymatroid function satisfying the "1-Lipschitz" property (denote $X \Delta Y$ as $(X - Y) \cup (Y - X)$, the symmetric difference)

$$| F^0(X) - F^0(Y) | \leq |X \Delta Y| \tag{140}$$

is called the rank function, denoted usually as R^0.

Every matroid $\mathfrak{M} = (\Omega, \mathcal{I})$ is equipped with a rank function defined as

$$R^0_{\mathfrak{M}}(X) := \max\{|I| : I \subseteq X \text{ and } I \in \mathcal{I}\}. \tag{141}$$

One need only apply **??** to verify the submodularity of (141): Let $X \subseteq Y \subseteq \Omega - \omega$. Note that $(R^0_{\mathfrak{M}})^{\omega}(\cdot) \in \{0, 1\}$. We show that $(R^0_{\mathfrak{M}})^{\omega}(Y) = 1 \implies (R^0_{\mathfrak{M}})^{\omega}(X) = 1$. Let $\mathcal{I} \ni I_X \subseteq X$ such that $R^0(X) = |I_X|$ and similarly define I_Y such that $R^0(Y \cup \{\omega\}) = |I_Y \cup \{\omega\}| > |I_X|$. By the augmentable property of independent sets we can build independent set $I_X \cup J$ where $J \subseteq I_Y \cup \{\omega\}$ so that $|I_X \cup J| = |I_Y \cup \{\omega\}|$. We claim $\omega \in J$ for otherwise $I_X \cup J \subseteq Y$ hence $R^0(Y) = |I_Y \cup \{\omega\}| = R^0(Y) + 1$, contradiction. Therefore $I_X \cup \{\omega\}$ is independent due to the inheritable property, i.e., $(R^0_{\mathfrak{M}})^{\omega}(X) = 1$.

Conversely, we can also define a matroid from a given rank function.

Definition 10.3: Matroid by Rank Function

Given a rank function R^0 on Ω, we associate it with the matroid $\mathfrak{M} = (\Omega, \mathcal{I})$, where $\mathcal{I} \ni I \subseteq \Omega$ iff $R^0(I) = |I|$. Such a definition indeed yields a matroid in the sense of Definition 10.1:

(I). $R^0(\emptyset) = 0 = |\emptyset|$, hence $\emptyset \in \mathcal{I}$;

(II). Let $I \subseteq J \in \mathcal{I}$. From the 1-Lipschitz property of R^0 we have $R^0(I) - R^0(\emptyset) = R^0(I) \leq |I \Delta \emptyset| = |I|$. Similarly, $R^0(J) - R^0(I) = |J| - R^0(I) \leq |J \Delta I| = |J - I|$ hence $R^0(I) \geq |J| - |J - I| = |I|$. Therefore $R^0(I) = |I|$, i.e., $I \in \mathcal{I}$;

(III'). We verify (III') in Remark 10.1 instead of (III) in Definition 10.1. Let $I \in \mathcal{I}, J \in \mathcal{I}$ such that $I - J = \{i\}$ while $J - I = \{j_1, j_2\}$. Suppose $\forall k \in \{1, 2\}, I \cup \{j_k\} \notin \mathcal{I}, i.e., R^0(I \cup \{j_k\}) = R^0(I)$. Then $R^0(J) \leq R^0(I \cup \{j_1, j_2\}) \leq R^0(I \cup \{j_1\}) + R^0(I \cup \{j_2\}) - R^0(I) = R^0(I) = R^0(J) - 1$, contradiction.

Finally we show that the rank function (141) $R_0^{\mathfrak{M}}$ defined for the constructed matroid coincides with the rank function R_0 we start with: Note first $R^0_{\mathfrak{M}}(X) := \max\{|I| : I \subseteq X \text{ and } |I| = R^0(I)\} \leq R^0(X)$, since R^0 is nondecreasing. The converse can be shown by induction: $R^0(\emptyset) = R^0_{\mathfrak{M}}(\emptyset) = 0$. Let $X \neq \emptyset$ and consider $X - \{x\}$ for $x \in X$. By the induction hypothesis, there exists $\mathcal{I} \ni I \subseteq X - \{x\}$ such that $R^0_{\mathfrak{M}}(X - \{x\}) = |I| = R^0(X - \{x\})$. If $R^0(X) = R^0(X - \{x\}) = R^0_{\mathfrak{M}}(X - \{x\}) \leq R^0_{\mathfrak{M}}(X)$ then we are done, otherwise assume $R^0(X) = R^0(X - \{x\}) + 1$. We claim $I \cup \{x\} \in \mathcal{I}$ since $R^0(I \cup \{x\}) \geq R^0(I) + R^0(X) - R^0(X - \{x\}) = |I| + 1$. Therefore $R^0(X) \leq R^0(X - \{x\}) + R^0(\{x\}) \leq |I| + 1 \leq R^0_{\mathfrak{M}}(X)$, finishing the induction.

More generally, we call a set Ω, equipped with a polymatroid function P^0, a polymatroid.

As mentioned before, a maximal independent set is called *basis* while a minimal dependent set is called *circuit*. We can also define a matroid by specifying its bases or circuits.

Definition 10.4: Matroid by Bases

Let $\mathcal{B} \subseteq 2^{\Omega}$ be a collection of bases, which satisfies

(I). (Non-empty) $\mathcal{B} \neq \emptyset$;

(II). (Exchangeable) For any $B_1, B_2 \in \mathcal{B}$, for any $x \in B_1$ there exists $y \in B_2$ such that $(B_1 - \{x\}) \cup \{y\} \in \mathcal{B}$.

(II'). For any $B_1, B_2 \in \mathcal{B}$, for any $x \in B_1$ there exists $y \in B_2$ such that $(B_2 - \{y\}) \cup \{x\} \in \mathcal{B}$.

The independent sets are precisely those contained in some basis.

We only prove (II) $\Rightarrow$ (II') (the other direction is completely analogous): Take $B_1, B_2 \in \mathcal{B}$ and fix $x \in B_1 - B_2$ (the interesting case). For each $z \in B_1 - B_2 - \{x\}$, by (II) there exists $w \in B_2 - B_1$ such that $(B_1 - \{z\}) \cup \{w\} \in \mathcal{B}$. Repeating the replacement until $B_1 - B_2 - \{x\} = \emptyset$.

Definition 10.5: Matroid by Circuits

Let $\mathcal{C} \subseteq 2^{\Omega}$ be a collection of circuits, which satisfies

(I). (Non-containing) $\forall C_1, C_2 \in \mathcal{C}$, $C_1 \subseteq C_2 \implies C_1 = C_2$;

(II). $\forall C_1, C_2 \in \mathcal{C}$ with $C_1 \neq C_2$, $\forall x \in C_1 \cap C_2, \exists C \in \mathcal{C}$ such that $C \subseteq (C_1 \cup C_2) - \{x\}$;

(II'). If $C_1, C_2 \in \mathcal{C}, x \in C_1 \cap C_2, y \in C_1 - C_2$, then $\exists C \in \mathcal{C}$ such that $y \in C \subseteq (C_1 \cup C_2) - \{x\}$.

Circuits are precisely "bases" for *dependent* sets. Define the independent sets to be those containing no elements in $\mathcal{C}$.

We show first that (II) indeed guarantees the new definition of independent sets satisfies Definition 10.1. We verify only (III') in Remark 10.1. Let $I \in \mathcal{I}, J \in \mathcal{I}, I - J = \{i\}, J - I = \{j_1, j_2\}$, and suppose to the contrary $I \cup \{j_k\} \notin \mathcal{I}$ for $k \in \{1, 2\}$. Then $I \cup \{j_k\} \subseteq J \cup \{i\} \notin \mathcal{I}$,i.e., $\exists C_1 \in \mathcal{C}$ such that $C_1 \subseteq J \cup \{i\}$. Let $C_2 \in \mathcal{C}$ be another circuit contained in $J \cup \{i\}$. We must have $i \in C_1 \cap C_2$ (otherwise $C_k \subseteq J$ contradicting $J \in \mathcal{I}$). If $C_1 \neq C_2$, by (II) we can construct $C \in \mathcal{C}$ such that $C \subseteq J$, i.e., contradicting $J \in \mathcal{I}$. Therefore $C_1 = C_2$. Moreover $I \in \mathcal{I}$ implies C_1 is not contained in I hence exists, say, $j_1 \in C_1 \cap (J - I)$. It follows from the uniqueness of C_1 that $\mathcal{I} \ni (J \cup \{i\}) - \{j_1\} = I \cup \{j_2\}$.

Next we prove that (II') is satisfied for the collection of circuits of any matroid (defined through Definition 10.1). Consider the submatroid $\mathfrak{M}_s = (C_1 \cup C_2, \mathcal{I}_s)$ (cf. Definition 10.6 below). Obviously $C_1 - \{y\}$ does not contain any (sub)circuit (otherwise contradicting (I)) hence exists some basis $y \notin B_1 \supseteq C_1 - \{y\}$, and similarly exists another basis $x \notin B_2 \supseteq C_2 - \{x\}$. If $y \in B_2$, then by the exchangeability of bases, $\exists z \in B_1 - B_2$ such that $(B_2 - \{y\}) \cup \{z\}$ is still a basis. Note that $z \notin \{x, y\}$ for otherwise $B_2 \supseteq C_2$ (as $y \notin C_2$). Replace B_2 with $(B_2 - \{y\}) \cup \{z\}$ we can assume $\{x, y\} \cap B_2 = \emptyset$. Therefore $B_2 \cup \{y\} \notin \mathcal{I}$, i.e., $\exists C \in \mathcal{C}_s$ such that $C \subseteq B_2 \cup \{y\}$. Apparently $y \in C$ (for otherwise $C \subseteq B_2$) and $x \notin C$ (for $x \notin B_2$).

Needless to say that (II') $\implies$ (II). Another consequence of (II) is that for any $I \in \mathcal{I}$, if $I \cup \{j\} \notin \mathcal{I}$ then there exists a unique circuit in $I \cup \{j\}$ (since all circuits must contain j).

Definition 10.6: Submatroid

Given a matroid $\mathfrak{M} = (\Omega, \mathcal{I})$ and any set $S \subseteq \Omega$ we can define a submatroid $\mathfrak{M}_s = (S, \mathcal{I}_s)$, where $\mathcal{I}_s := \{I \in \mathcal{I} : I \subseteq S\}$. It is easy to verify that indeed $\mathcal{I}_s$ satisfies Definition 10.1. The rank function for $\mathfrak{M}_s$ is simply the one for $\mathfrak{M}$ restricted to subsets of S, and the circuits of $\mathfrak{M}_s$ are precisely those of $\mathfrak{M}$ which are contained in S.

Definition 10.7: Matroid by Closure Operator

Fix a matroid $\mathfrak{M} = (\Omega, \mathcal{I})$ with its rank function $R^0_{\mathfrak{M}}$ defined in (141), we define its closure operator

$\mathrm{cl}: 2^\Omega \to 2^\Omega$ by

$$\mathrm{cl}(X) := \Big\{\omega \in \Omega : R^0_{\mathfrak{M}}(X) = R^0_{\mathfrak{M}}(X \cup \{\omega\})\Big\}. \tag{142}$$

Sets satisfy $\mathrm{cl}(X) = X$ are called flats (closed sets) while sets satisfy $\mathrm{cl}(X) = \Omega$ are called spanning (dense). We note first that $R^0_{\mathfrak{M}}(\mathrm{cl}(X)) = R^0_{\mathfrak{M}}(X)$ hence $\mathrm{cl}(\mathrm{cl}(X)) = \mathrm{cl}(X)$: Let $I \in \mathcal{I}, I \subseteq X$ satisfy $|I| = R^0_{\mathfrak{M}}(X)$, then if $R^0_{\mathfrak{M}}(\mathrm{cl}(X)) > R^0_{\mathfrak{M}}(X)$ there must exists $\omega \in \mathrm{cl}(X) - X$ such that $I \cup \{\omega\} \in \mathcal{I}$, contradicting $R^0_{\mathfrak{M}}(X \cup \{\omega\}) = R^0_{\mathfrak{M}}(X)$. Similarly we can prove $X \subseteq Y \implies \mathrm{cl}(X) \subseteq \mathrm{cl}(Y)$, hence consequently $\mathrm{cl}(X \cap Y) \subseteq \mathrm{cl}(X) \cap \mathrm{cl}(Y)$.

One easily verifies that the closure operator satisfies

(I). $X \subseteq \mathrm{cl}(X)$;

(II). $X \subseteq \mathrm{cl}(Y) \implies \mathrm{cl}(X) \subseteq \mathrm{cl}(Y)$;

(III). $\forall X \subseteq \Omega, \omega \in \Omega : \omega' \in \mathrm{cl}(X \cup \{\omega\}) - \mathrm{cl}(X) \implies \omega \in \mathrm{cl}(X \cup \{\omega'\}) - \mathrm{cl}(X)$.

We prove conversely the closure operator defines a matroid by specifying

$$\mathcal{I} := \Big\{I \subseteq \Omega : \forall \omega \in I, \omega \notin \mathrm{cl}(I - \{\omega\})\Big\}.$$

We first prove that

$$I \in \mathcal{I} \implies \mathrm{cl}(I) = I \cup \{\omega : I \cup \{\omega\} \notin \mathcal{I}\}. \tag{143}$$

If $\omega \in \mathrm{cl}(I) - I$, then $I \cup \{\omega\} \notin \mathcal{I}$. Conversely if $I \cup \{\omega\} \notin \mathcal{I}$ then $\exists \omega' \in I \cup \{\omega\}$ such that $\omega' \in \mathrm{cl}((I \cup \{\omega\}) - \omega')$. If $\omega = \omega'$ then $\omega \in \mathrm{cl}(I)$; otherwise $\omega' \in I$, since $I \in \mathcal{I}$ we have $\omega' \notin \mathrm{cl}(I - \{\omega'\})$ hence by (III) (where $X = I - \{\omega'\}$) we get again $\omega \in \mathrm{cl}(I)$.

Now we can prove (III') in Remark 10.1 (the other two are easy). Let $I \in \mathcal{I}, J \in \mathcal{I}, I - J = \{i\}, J - I = \{j_1, j_2\}$. Assume that $I \cup \{j_1\} \notin \mathcal{I}$, i.e., $(J \cup \{i\}) - \{j_2\} \notin \mathcal{I}$. Apply (143) we know $i \in \mathrm{cl}(J - \{j_2\})$ (for $J - \{j_2\} \in \mathcal{I}$). Therefore $I \subseteq \mathrm{cl}(J - \{j_2\})$ and consequently $\mathrm{cl}(I) \subseteq \mathrm{cl}(J - \{j_2\})$. Since $J \in \mathcal{I}$ we have $j_2 \notin \mathrm{cl}(J - \{j_2\})$ hence $j_2 \notin \mathrm{cl}(I)$. Apply (143) once more we obtain $I \cup \{j_2\} \in \mathcal{I}$.

Finally we show that the closure operator $\mathrm{cl}(\cdot)$ we start with coincides with the closure operator $\mathrm{cl}_{\mathfrak{M}}(\cdot)$ for the constructed matroid. Consider an arbitrary $X \subseteq \Omega$ and let $I \in \mathcal{I}, I \subseteq X$ satisfy $R^0_{\mathfrak{M}}(X) = |I|$. Then $\mathrm{cl}_{\mathfrak{M}}(X) = I \cup \{\omega : I \cup \{\omega\} \notin \mathcal{I}\}$, which by (143) is also equal to $\mathrm{cl}(I) \subseteq \mathrm{cl}(X)$. Let $\omega \in X - I$, then $I \cup \{\omega\} \notin \mathcal{I}$ (by our choice of I). Hence $\omega \in \mathrm{cl}(I)$ due to (143), i.e., $X \in \mathrm{cl}(I)$. Therefore $\mathrm{cl}(X) \subseteq \mathrm{cl}(I)$, implying $\mathrm{cl}_{\mathfrak{M}}(X) = \mathrm{cl}(I) = \mathrm{cl}(X)$.

Definition 10.8: Matroid by Flats

Recall that flats are closed sets $F : \mathrm{cl}(F) = F$, which collectively, denoted as $\mathcal{F}$, satisfy:

(I). $\Omega \in \mathcal{F}$;

(II). $F_1 \in \mathcal{F}, F_2 \in \mathcal{F} \implies F_1 \cap F_2 \in \mathcal{F}$;

(III). $\forall F \in \mathcal{F}, \forall \omega \in \Omega$, denote F_s as the *smallest* flat containing $F \cup \{\omega\}$, then there is no $F' \in \mathcal{F}$ with $F \subset F' \subset F_s$.

It is easy to verify that *bona fide* flats do satisfy (I) and (II). Note also that due to (II), it is meaningful to talk about *smallest* containing flats in (III). To show (III), assume such an F' exists. Then $\exists \omega' \in F'$ such that $\omega' \notin \mathrm{cl}(F)$. By assumption $\omega \in F_s - F'$ and $\omega \notin \mathrm{cl}(F \cup \{\omega'\})$ (due to the minimality of F_s). By property (III) of the closure operator we have $\omega' \notin \mathrm{cl}(F \cup \{\omega\}) = F_s \supset F'$, contradiction.

Conversely, a collection of sets satisfy the above three properties defines a matroid, through the closure operator:

$$\mathrm{cl}(X) := \bigcap_{F \in \mathcal{F}, F \supseteq X} F.$$

We need only prove the last property of the closure operator. Let $X \subseteq \Omega, \omega \in \Omega, \omega' \in \mathrm{cl}(X \cup \{\omega\}) - \mathrm{cl}(X)$. Then $\mathrm{cl}(X) \subset \mathrm{cl}(X \cup \{\omega'\}) \subseteq \mathrm{cl}(X \cup \{\omega\})$, therefore by property (III) of the flats we have $\mathrm{cl}(X \cup \{\omega'\}) = \mathrm{cl}(X \cup \{\omega\})$, i.e., $\omega \in \mathrm{cl}(X \cup \{\omega'\})$.

We can also specify the independent sets through flats (and vice versa):

$$\mathcal{I} = \{I \subseteq \Omega : \forall \omega \in I, \exists F \in \mathcal{F} \text{ such that } \omega \notin F \text{ and } I - \{\omega\} \subseteq F\}.$$

Indeed, if I is independent and $\omega \in I$, let $F = \mathrm{cl}(I - \{\omega\})$, then $R^0_{\mathfrak{M}}(F \cup \{\omega\}) \geq R^0_{\mathfrak{M}}(I) = R^0_{\mathfrak{M}}(I - \{\omega\}) + 1 = R^0_{\mathfrak{M}}(F) + 1$, meaning $\omega \notin F$. Conversely, if I is not independent, then $\exists \omega \in I$ such that $\omega \in \mathrm{cl}(I - \{\omega\})$, hence all flats containing $\mathrm{cl}(I - \{\omega\})$ must also contain ω.

Remark 10.2: Polynomial-time Implications

The following polynomial-time implications are easy to check:

$$\text{Closure} \Leftrightarrow \text{Independence} \Leftrightarrow \text{Rank}$$
$$\Downarrow$$
$$\{\text{Basis, Circuit, Flat}\}$$

Definition 10.9: Dual Matroid

Given a matroid $\mathfrak{M} = (\Omega, \mathcal{I})$ we can define its dual

$$\mathfrak{M}^* := (\Omega, \mathcal{I}^*) \quad \text{where} \quad \mathcal{I}^* := \{I \subseteq \Omega : \mathrm{cl}_{\mathfrak{M}}(\Omega - I) = \Omega\}, \tag{144}$$

i.e., the complement of spanning sets. Equivalently we can define

$$\mathcal{B}^* := \{B \subseteq \Omega : \Omega - B \in \mathcal{B}\}, \tag{145}$$

or

$$R^0_{\mathfrak{M}^*}(X) = |X| - R^0_{\mathfrak{M}}(\Omega) + R^0_{\mathfrak{M}}(\Omega - X). \tag{146}$$

The equivalence can be easily established by noting that (146) indeed is a rank function, whose independent sets are given by precisely (144), whose bases are precisely (145).

To justify the name *dual* matroid, note simply from (145) that

$$\mathfrak{M} = (\mathfrak{M}^*)^*. \tag{147}$$

Definition 10.10: Deletion, Contraction and Truncation

Let $\mathfrak{M} = (\Omega, \mathcal{I})$ be a matroid and $\omega \in \Omega$, then $\mathfrak{M}\backslash\{\omega\}$ is the submatroid with ground set $\Omega - \{\omega\}$, i.e., deleting ω. The contraction $\mathfrak{M}/\{\omega\} := (\mathfrak{M}^*\backslash\{\omega\})^*$, and the truncation $\mathfrak{M}_k := (\Omega, \mathcal{I}_k)$, where $\mathcal{I}_k := \{I \in \mathcal{I} : |I| \leq k\}$. It is easily verified that deletion is commutative, and as a consequence of the deliberate definition, contraction is also commutative. Therefore we can extend the definitions of deletion and contraction from singletons to subsets by successive application to each element (or simply revise the definition directly). It follows from (146) that $\forall S \subseteq \Omega$

$$\forall X \subseteq \Omega - S, R^0_{\mathfrak{M}/S}(X) = R^0_{\mathfrak{M}}(X \cup S) - R^0_{\mathfrak{M}}(S), \tag{148}$$

which implies further that deletion commutes with contraction (by verifying the equality of the rank functions). Any matroid arising from deletion and contraction of $\mathfrak{M}$ is called a minor of $\mathfrak{M}$.

Checking again the rank functions we verify the duality between deletion and contraction:

$$(\mathfrak{M}\backslash S)^* = \mathfrak{M}^*/S, \tag{149}$$
$$(\mathfrak{M}/S)^* = \mathfrak{M}^*\backslash S. \tag{150}$$

Definition 10.11: Matroid Isomorphism

We call two matroids $\mathfrak{M}_i = (\Omega_i, \mathcal{I}_i), i \in \{1,2\}$ isomorphic if there exists a bijection $f : \Omega_1 \to \Omega_2$ such that $I_1 \in \mathcal{I}_1 \iff f(I_1) \in \mathcal{I}_2$. Necessarily isomorphic matroids satisfy $|\Omega_1| = |\Omega_2|$.

Now it is time to see some examples.

Example 10.1: Uninteresting Matroids

$\mathfrak{M}_t = (\Omega, \{\emptyset\})$ is called the trivial matroid while $\mathfrak{M}_f = (\Omega, 2^\Omega)$ is called the free matroid. Clearly the two are dual to each other.

Example 10.2: k-Uniform Matroid

Denote $n = |\Omega| \geq k$. Define the uniform matroid

$$\mathcal{U}_{k,n} := (\Omega, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq \Omega : |I| \leq k\}. \tag{151}$$

Obviously the bases $\mathcal{B} = \{I \subseteq \Omega : |I| = k\}$ and the rank function $R^0(X) = |X| \wedge k$. It is exactly the truncation of the free matroid.

Example 10.3: Linear Matroid

This example is extremely important due to its historical significance and many deep questions related to it. Let Ω be the *disjoint* union of $\{a_i\}$ where $i \in \{1, \ldots, n\}, a_i \in V$ and V is an arbitrary vector space over some field $\mathbb{F}$. A subset is claimed independent iff its elements are linearly independent over $\mathbb{F}$. One should not have difficulty in verifying Definition 10.1. From the matroid structure point of view, we can take $V = \mathbb{F}^r$ where r is the rank of the matroid. Hence a linear matroid can be simply represented as a matrix $A \in \mathbb{F}^{r\times n}$, which is easily seen to be isomorphic to $[\mathbf{I}_r, B]$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix, whereas the dual matroid can be shown to be isomorphic to $[B^\top, \mathbf{I}_{n-r}]$, hence is also linear, consequently so is the minor. Surprisingly, it is still an open problem to characterize linear matroids up to *isomorphism*.

The whole theory on matroids starts from Hassler Whitney's (also independently Takeo Nakasawa's) efforts in abstracting our example above. The power of abstraction is again witnessed!

Example 10.4: Matching Matroid

Let $\mathcal{G} = (V, E)$ be a given graph and $\Omega \subseteq V$. Define

$$\mathfrak{M}_M = (\Omega, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq \Omega : I \text{ is covered by some matching in } \mathcal{G}\}. \tag{152}$$

It is easy to verify 3' in Remark 10.1, and $\mathfrak{M}_M$ so defined is called the matching matroid.

Example 10.5: Transversal Matroid

Let $\mathcal{Q} \subseteq 2^\Omega$. Define the bipartite graph $\mathcal{G} := (\Omega, \mathcal{Q}; E)$ where $\forall \omega \in \Omega, Q \in \mathcal{Q}, E(\omega, Q) = 1$ iff $\omega \in Q$. Define

$$\mathfrak{M}_T := (\Omega, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq \Omega : \exists \text{1-1 map } \pi : I \to \mathcal{Q} \text{ such that } \forall i \in I, i \in \pi(i)\}, \tag{153}$$

i.e., $\mathcal{I}$ is the collection of all partial transversals of $\mathcal{Q}$ (w.r.t. Ω). Clearly, $\mathfrak{M}_T$ is a special matching matroid, called the transversal matroid.

Using Theorem 9.3, the rank function of $\mathfrak{M}_T$ equals the minimum size of vertex covers in the bipartite graph, which is in the following form (S is the part in X *not* chosen by the vertex cover and $\mathcal{P}$ is the part in $\mathcal{Q}$ *not* chosen by the vertex cover):

$$R^0_{\mathfrak{M}_T}(X) = \min_{S\subseteq X} |X - S| + |\{Q \in \mathcal{Q} : Q \cap S \neq \emptyset\}| \tag{154}$$

$$= \min_{\mathcal{P}\subseteq\mathcal{Q}} |(\cup\mathcal{P}) \cap X| + |\mathcal{Q}| - |\mathcal{P}|. \tag{155}$$

Transversal matroids are linearly representable in all but finitely many finite fields [].

Remark 10.3: Matching Matroid = Transversal Matroid

Example 10.6: Partition Matroid

Let $\Omega = \sum_{i=1}^n P_i$ be a partition and $\{k_1, \ldots, k_n\}$ be given natural numbers. Define

$$\mathfrak{M}_P := (\Omega, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq \Omega : |I \cap P_i| \leq k_i\}. \tag{156}$$

By setting $\mathcal{Q}$ to be the disjoint union of P_i, each with k_i copies, we see that $\mathfrak{M}_P$ is a special transversal matroid, called the partition matroid. Note that the k-Uniform matroid $\mathcal{U}_{k,n}$ is a special partition matroid (with $P_1 = \Omega$ and $k_1 = k$). The rank function is easily seen to be

$$R^0_{\mathfrak{M}_P}(X) = \sum_{i=1}^n |X \cap P_i| \wedge k_i.$$

Example 10.7: Algebraic Matroid

Let $\mathbb{E}$ be a field extension of the field $\mathbb{F}$ and Ω be a finite subset of $\mathbb{E}$. Define

$$\mathfrak{M}_A := (\Omega, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq \Omega : I \text{ is algebraically independent over } \mathbb{F}\} \tag{157}$$

To verify 3' in Remark 10.1, let $\mathcal{I} \ni I = \{e_n\} \cup K, \mathcal{I} \ni J = K \cup \{e_1, e_2\}, K = \{e_3, \ldots, e_{n-1}\}$. Assume to the contrary $\mathcal{I} \cup \{e_i\} \notin \mathcal{I}, i \in \{1, 2\}$. Then there exist nonzero polynomials $p_i \in \mathbb{F}[x, x_3, \ldots, x_n]$ such that $p(e_i, e_3, \ldots, e_n) = 0, i \in \{1, 2\}$. W.l.o.g. assume p_1 and p_2 are irreducible. Since $J \in \mathcal{I}$, p_1 and p_2 are relatively prime. Define $\mathbb{L} := \mathbb{F}(x_1, \ldots, x_{n-1})$. So $p_i \in \mathbb{L}[x_n]$. Let r be the g.c.d. of p_1 and p_2 in $\mathbb{L}[x_n]$. As p_i are relatively prime, we know $r \in \mathbb{L}[x_n]$ hence $r \in \mathbb{F}[x_1, \ldots, x_{n-1}]$. Now $r = \sum_{i=1}^2 \alpha_i p_i$ for some $\alpha_i \in \mathbb{L}[x_n]$. So $r(e_1, \ldots, e_{n-1}) = 0$, contradicting $I \in \mathcal{I}$.

Linear matroids are algebraic.

[?] gives an example for algebraic nonlinear matroids and another example for non-algebraic matroids. Compared to linear matroids, much less is known about algebraic matroids. In particular, it is not known if the dual of an algebraic matroid is algebraic. However, algebraic matroids over any field $\mathbb{F}$ are closed under taking minors.

As it turns out, greedy algorithms are surprisingly effective in discrete optimization. Let us illustrate this with an example. Given a matroid $\mathfrak{M} = (\Omega, \mathcal{B})$ and a weight vector $w \in \mathbb{R}^\Omega$, we want to find a basis $B \in \mathcal{B}$ so that it has the maximum weight $w(B) := \sum_{i\in B} w_i$. Given an independent set $\mathcal{I} \ni I \notin \mathcal{B}$, let us call the set

$$\emptyset \neq A^I \subseteq A^I_{\max} := \{i \in \Omega - I : I \cup \{i\} \in \mathcal{I}\} \tag{158}$$

admissible iff for all **maximum-weight bases** $B \supseteq I$ the following basis exchange property holds:

$$\forall i \in A^I \ \exists j \in B \cap A^I \text{ s.t. } (B - \{j\}) \cup \{i\} \in \mathcal{B}. \tag{159}$$

Notice that $A^I_{\max}$ itself is admissible in a strong sense: For any basis $B \supseteq I$ we have $B \cap A^I_{\max} \neq \emptyset$, and for any $i \in A^I_{\max}$, $\mathcal{I} \ni I \cup \{i\} \subseteq B \cup \{i\} \notin \mathcal{I}$ hence $\exists j \in B \cap A^I_{\max}$ such that $(B - \{j\}) \cup \{i\} \in \mathcal{B}$.

The greedy algorithm, shown in Algorithm 1, only requires an oracle for testing basis and another oracle for constructing the admissible set A^I. If we set $A^I = A^I_{\max}$ then both oracles reduce to testing independence. An apparent dual version of Algorithm 1 exists (which finds a minimum-weight basis in the dual matroid).

Algorithm 1: The greedy algorithm for finding a maximum-weight basis

$I = \emptyset$.
while I *is not a basis* **do**
  $i^* \leftarrow \arg\max_{i \in A^I} w_i$
  $I \leftarrow I \cup \{i^*\}$

We now prove that the greedy algorithm *works*, and more importantly, we show that it is precisely the matroid structure that makes the greedy algorithm work. Note that for $A^I = A^I_{\max}$, one easily verifies that the weights picked by the greedy algorithm is non-increasing, and moreover, the k-th weight is always bigger than the k-th largest weight of any basis (for otherwise the greedy algorithm would have picked differently in the k-th stage).

Theorem 10.1: Correctness of the Greedy Algorithm

The greedy algorithm is correct for any matroid. Conversely, given any collection of sets $\mathcal{I}$ satisfying 1 and 2 in Definition 10.1, if the greedy algorithm, equipped with the admissible set $A^I_{\max}$, always returns a maximum-weight element in $\mathcal{I}$ for any nonnegative weight, then $\mathcal{I}$ defines a matroid.

Proof: We show by induction that the independent set I maintained by the greedy algorithm is always contained in some maximum-weight basis. Initially this is vacuously true. Suppose at some intermediate stage $I \in \mathcal{I}$ is contained in some maximum-weight basis B. Suppose the newly added i^* does not belong to B (otherwise we are done). But by the admissibility of A^I there exists $j \in B \cap A^I$ such that $I \cup \{i^*\} \subseteq (B - \{j\}) \cup \{i^*\} \in \mathcal{B}$, which is also a maximum-weight basis due to the optimality of i^*.

Conversely, given two sets $I, J \in \mathcal{I}$ with $|J| > |I| = k$. Assume $I \cup \{j\} \notin \mathcal{I}$ for every $j \in J - I$. Define

$$w_i = \begin{cases} k+1, & i \in I \\ k+2, & i \in J - I \\ 0, & \text{otherwise} \end{cases}.$$

The greedy algorithm will pick all elements in I and then pick elements not in $I \cup J$, hence it returns a set in $\mathcal{I}$ with weight $k(k+1)$. On the other hand, had we take all elements in J we would have a set in $\mathcal{I}$ with weight bigger than $(k+1)(k+2) > k(k+1)$, contradicting the optimality of the greedy algorithm. ■

We observe that for a *nonnegative* weight, finding the maximum-weight independent set is the same as finding the maximum-weight basis, while for an arbitrary weight, finding the maximum-weight independent set reduces to the subproblem operating on the submatroid with all negatively weighted elements deleted.

Another observation about the greedy Algorithm 1 (using $A^I_{\max}$) is that every element that is *not* chosen is dominated in some circuit (i.e., having the smallest weight in that circuit): Indeed, suppose j is not chosen, then there exists the smallest number k such that $\{i_1, \ldots, i_k\}$, chosen sequentially by the greedy algorithm, is dependent when augmenting j. If $k = 0$ there is nothing to prove, otherwise the fact that the greedy algorithm chose w_k implies that $w_k \geq w_j$. On the other hand, if the element j is dominated in some circuit C, then it can be safely discarded: Assume j is contained in some maximum-weight basis B. Since $C - \{j\}$ is independent, it is contained in some basis B'. By the enhanced basis exchange property (cf. Remark 10.4) we know $(B - \{j\}) \cup \{i\}$ and $(B' - \{i\}) \cup \{j\}$ are again bases. But then we must have $i \in C - \{j\}$, i.e., $(B - \{j\}) \cup \{i\}$ is again a maximum-weight basis (due to the minimality of j in C), hence j can be deleted. Note that the maximum-weight basis in the matroid and the minimum-weight basis in the dual matroid can be chosen to complement each other. Under this duality we know the

dominating element in some circuit of the dual matroid can be safely picked and then contracted. This variation of Algorithm 1 is summarized in Algorithm 2, while yet another similar variation is summarized in Algorithm 3. The slight difference is on how to avoid picking the same circuit.

Whether or not it is easier to work with the matroid or its dual, testing independence or constructing circuits, performing deletion or contraction, are of course dependent on the particular problem at hand.

Algorithm 2: Variation 1 of the greedy Algorithm 1

$I = \emptyset, J = \emptyset$.
repeat
  Perform Procedure a or a'.
until *Convergence*;
begin Procedure a
  if $\exists C \in \mathcal{C}_{\mathfrak{M}}$ **then**
    $\omega \leftarrow \arg\min_{c \in C} w_c$
    $J \leftarrow J \cup \{\omega\}$
    $\mathfrak{M} \leftarrow \mathfrak{M} \backslash \{\omega\}$
  else
    $I \leftarrow \Omega_{\mathfrak{M}}$
    return
begin Procedure a'
  if $\exists C^* \in \mathcal{C}_{\mathfrak{M}^*}$ **then**
    $\omega^* \leftarrow \arg\max_{c \in C^*} w_c$
    $I \leftarrow I \cup \{\omega^*\}$
    $\mathfrak{M} \leftarrow \mathfrak{M}/\{\omega^*\}$.
  else
    $J \leftarrow \Omega_{\mathfrak{M}}$
    return

Definition 10.12: Regular Matroid

A matroid is called regular iff it can be linearly represented in every field, or equivalently, iff it can be represented as a TUM over $\mathbb{R}$ [**?**, Theorem 6.6.3, page 205]. Clearly the dual, hence minor, of a regular matroid is regular.

Example 10.8: Graphic Matroid

Let $\mathcal{G} = (V, E)$ be a graph. Define

$$\mathfrak{M}_G := (E, \mathcal{I}) \quad \text{where} \quad \mathcal{I} := \{I \subseteq E : I \text{ contains no cycle } \}. \tag{160}$$

One easily verifies Remark 10.1 hence $\mathfrak{M}_G$ is a matroid, whose independent sets are exactly forests of the graph $\mathcal{G}$. The rank function for the graphic matroid is given by

$$R^0_{\mathfrak{M}_G}(X) = |V| - \kappa(V, X),$$

where $\kappa(V, X)$ is the number of connected components in the subgraph $\mathcal{G}_s := (V, X)$. Note that if the graph $\mathcal{G}$ is connected then the bases of $\mathfrak{M}_G$ are exactly spanning trees.

Graphic matroids are regular. Indeed, construct the matrix $A \in \mathbb{R}^{|V| \times |E|}$, where $A_{v,e} = 1$ iff the edge $e \in E$ leaves the node $v \in V$, $A_{v,e} = -1$ iff the edge e enters the node v, and $A_{v,e} = 0$ iff e does not meet v. It is easy to see that $X \subseteq E$ contains no cycle iff A_X is linearly independent. Therefore graphic matroids are linearly representable in $\mathbb{R}$ by a TUM A, i.e., they are regular.

Consider a connected graph $\mathcal{G}$ and apply the greedy Algorithm 1 to it. If we choose $A^I = A^I_{\max}$ we end up with Kruskal's algorithm, while if we choose A^I to be the set of all edges in $E - I$ that

Algorithm 3: Variation 2 of the greedy Algorithm 1

$I = \emptyset, J = \emptyset, K = \Omega$.
repeat
  Either Procedure b or b'.
until *convergence*;
begin Procedure b
  if $\exists C \in \mathcal{C}_{\mathfrak{M}}$ *and* $C \subseteq I \cup K$ **then**
    $\omega \leftarrow \arg\min_{c \in C} w_c$
    $J \leftarrow J \cup \{\omega\}$
    $K \leftarrow K - \{\omega\}$.
  else
    $I \leftarrow I \cup K$
    return
begin Procedure b'
  if $\exists C^* \in \mathcal{C}_{\mathfrak{M}^*}$ *and* $C \subseteq J \cup K$ **then**
    $\omega^* \leftarrow \arg\max_{c \in C^*} w_c$
    $I \leftarrow I \cup \{\omega^*\}$
    $K \leftarrow K - \{\omega^*\}$.
  else
    $J \leftarrow J \cup K$
    return

are connected to I we get Prim's algorithm, and finally if we choose A^I to be edges in $E - I$ that connect different components in the subgraph (V, I) we recover Borůvka's algorithm (assuming all weights are different).

Definition 10.13: Matroid Intersection

For two (or more) matroids, we can define their intersection by intersecting their ground sets and independent classes, respectively.

Note that the intersection need not satisfy (III) in Definition 10.1, as shown in the following example: Take $\Omega_1 = \Omega_2 = \{1,2,3\}, \mathcal{I}_1 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}\}, \mathcal{I}_2 = \{\emptyset, \{1\}, \{1,2\}\}, \mathcal{I}_1 \cap \mathcal{I}_2 = \{\emptyset, \{1\}, \{1,2\}\}$.

Definition 10.14: Matroid Union and Direct Sum

We can define the union (by taking unions of the ground sets and independent sets, respectively) and the direct sum (by taking disjoint unions of the ground sets and independent sets, respectively) of two (or more) matroids. It is trivial to verify that we indeed end up with *bona fide* matroids.

Remark 10.4: Strengthening the Basis Exchangeability

The basis exchangeability properties (II) and (II') in Definition 10.4 can be unified and strenghthened as:

(II'). Let $B_1, B_2 \in \mathcal{B}$, for any partition $B_1 = X_1 + Y_1$ there exists a partition $B_2 = X_2 + Y_2$ such that $X_1 \cup Y_2 \in \mathcal{B}, X_2 \cup Y_1 \in \mathcal{B}$.

11 Integer Polyhedra

Our main reference in this section is Schrijver [1986].

Definition 11.1: Polyhedron and Polytope

A polyhedron is a point set in, say $\mathbb{R}^n$, defined by finitely many linear inequalities $A\mathbf{x} \leq \mathbf{b}$ where $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. When $A, \mathbf{b}$ can be chosen with rational entries, we call the polyhedron rational. By definition, polyhedra are closed and convex. A bounded (in the sense of any norm) polyhedron is called polytope.

Let us describe the (nonempty) polyhedron $\mathfrak{P}$ as the set of linear inequalities $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^\top \mathbf{x} \leq b_i, i \in I\}$. Denote $I^= := \{i \in I : \forall \mathbf{x} \in \mathfrak{P}, \mathbf{a}_i^\top \mathbf{x} = b_i\}$ and denote $I^\leq := \{i \in I : \exists \mathbf{x} \in \mathfrak{P}, \mathbf{a}_i^\top \mathbf{x} < b_i\}$. Clearly $I = I^= \cup I^\leq$ and $\forall i \in I^\leq, (\mathbf{a}_i^\top, b_i)$ is linearly independent of $\{(\mathbf{a}_j^\top, b_j) : j \in I^=\}$. Collectively we will also use the notation $A^\leq, A^=, \mathbf{b}^\leq, \mathbf{b}^=$. A point $\mathbf{x} \in \mathfrak{P}$ is called an inner point if $A^\leq \mathbf{x} < \mathbf{b}^\leq$.

Proposition 11.1: Inner Point Exists

Every nonempty polyhedron has an inner point.

Proof: For $i \in I^\leq$, by definition each inequality $\mathbf{a}_i^\top \mathbf{x} \leq b_i$ admits a point $\mathbf{x}_i \in \mathfrak{P}$ such that $\mathbf{a}_i^\top \mathbf{x}_i < b_i$. Take the convex combination of $\mathbf{x}_i$'s. ■

Theorem 11.1: Dimension Formula

For any polyhedron $\mathfrak{P} \subseteq \mathbb{R}^n$, $\dim(\mathfrak{P}) + \operatorname{rank}(A^=, \mathbf{b}^=) = n$.

Proof: Note first that the affine space $H := \{\mathbf{x} \in \mathbb{R}^n : A^=\mathbf{x} = \mathbf{b}^=\}$ satisfies $\dim(H) = n - \operatorname{rank}(A^=, \mathbf{b}^=)$. Center H at an inner point of $\mathfrak{P}$ we see that $\dim(\mathfrak{P}) \geq n - \operatorname{rank}(A^=, \mathbf{b}^=)$. The other direction $n - \operatorname{rank}(A^=, \mathbf{b}^=) \geq \dim(\mathfrak{P})$ is apparent. ■

Definition 11.2: Face and Facet

The inequality $\mathbf{c}^\top \mathbf{x} \leq d$ is called a valid inequality for the polyhedron $\mathfrak{P}$ if it is satisfied by all points in $\mathfrak{P}$. A face $\mathfrak{F}$ of $\mathfrak{P}$ is defined as $\mathfrak{F} := \{\mathbf{x} \in \mathfrak{P} : \mathbf{c}^\top \mathbf{x} = d\}$, where $\mathbf{c}^\top \mathbf{x} \leq d$ is a valid inequality. And we say the inequality $\mathbf{c}^\top \mathbf{x} \leq d$ represents the face $\mathfrak{F}$. A maximal proper face is called a facet.

Proposition 11.2: Description of Faces

For any face $\mathfrak{F}$ of the polyhedron $\mathfrak{P}$, there exists $I^= \subseteq J \subseteq I$ such that $\mathfrak{F} = \{\mathbf{x} \in \mathfrak{P} : \mathbf{a}_j^\top \mathbf{x} = b_j, \forall j \in J\}$.

Proof: By definition $\mathfrak{F} = \arg\max_{A\mathbf{x}\leq\mathbf{b}} \mathbf{c}^\top \mathbf{x}$. Consider the dual problem $\mathbf{y}^* \in \arg\min_{\mathbf{y}\geq\mathbf{0}, A\mathbf{y}=\mathbf{c}} \mathbf{b}^\top \mathbf{y}$ and let $I^* := \{i \in I : \mathbf{y}^* > 0\}$. It is easy to see that the polyhedron $\mathfrak{F}^* := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^\top \mathbf{x} = b_i, \forall i \in I^*, \mathbf{a}_j^\top \mathbf{x} \leq b_j, \forall j \in I \backslash I^*\}$ coincides with the face $\mathfrak{F}$. ■

It follows immediately that a face of a polyhedron is a polyhedron, and a face of a face is still a face. Moreover, there is only finitely many faces.

Alert 11.1: Face of Convex Sets

The definition we gave for the face of a polyhedron can be generalized to arbitrary convex sets, and in that case it is usually called the exposed face. However, in general, an exposed face of an exposed face (of a convex set) need not be an exposed face (think about the convex hull of a donut).

Proposition 11.3: Dimension of Facets

Each facet $\mathfrak{F}$ is represented by the inequality $\mathbf{a}_i^\top \mathbf{x} \leq b_i$ for some $i \in I^{\leq}$, hence $\dim(\mathfrak{F}) = \dim(\mathfrak{P}) - 1$.

Proof: Since facets by definition are maximal and proper, it follows from Proposition 11.2 that they are represented by exactly one inequality. The dimension formula follows from Theorem 11.1. ■

Proposition 11.4: Necessity of Facets

For each facet $\mathfrak{F}$ of $\mathfrak{P}$, one of its representing inequalities is necessary for describing $\mathfrak{P}$.

Proof: Let $\mathbf{a}_r^\top \mathbf{x} \leq b_r$ be some representing inequality for the facet $\mathfrak{F}$ and $\mathfrak{P}_{\mathfrak{F}}$ be the polyhedron after removing all representing inequalities of the facet $\mathfrak{F}$. We prove that $\mathfrak{P}_{\mathfrak{F}} \backslash \mathfrak{P} \neq \emptyset$. Clearly $(\mathbf{a}_r^\top, b_r)$ is linearly independent of $(A^=, \mathbf{b}^=)$. Moreover, since $\mathbf{a}_r^\top \mathbf{x} \leq b_r$ represents a facet we must have $\mathbf{a}_r^\top$ linearly independent of $A^=$. Therefore there exists $\mathbf{y} \in \mathbb{R}^n$ such that $A^= \mathbf{y} = \mathbf{0}, \mathbf{a}_r^\top \mathbf{y} > 0$. Take an inner point $\mathbf{x}$ of $\mathfrak{F}$, then for some small $\epsilon > 0$ we have $\mathbf{x} + \epsilon \mathbf{y} \in \mathfrak{P}_{\mathfrak{F}}$ while $\mathbf{x} + \epsilon \mathbf{y} \notin \mathfrak{P}$ (since $\mathbf{a}_r^\top \mathbf{x} = b_r$). ■

Proposition 11.5: Sufficiency of Facets

Every inequality $\mathbf{a}_r^\top \mathbf{x} \leq b_r$ for some $r \in I^{\leq}$ that represents a face $\mathfrak{F}$ of $\mathfrak{P}$ with $\dim(\mathfrak{F}) < \dim(\mathfrak{P}) - 1$ is irrelevant for describing $\mathfrak{P}$.

Proof: Suppose $\mathbf{a}_r^\top \mathbf{x} \leq b_r$ is not irrelevant, then $\exists \mathbf{x}$ such that $A^= \mathbf{x} = \mathbf{b}^=, \mathbf{a}_i^\top \mathbf{x} \leq b^i, i \in I^{\leq} \backslash \{r\}$ and $\mathbf{a}_r^\top \mathbf{x} > b_r$. Take an inner point $\mathbf{y}$ of $\mathfrak{P}$, then there is some $\mathbf{z}$ on the line between $\mathbf{x}$ and $\mathbf{y}$ such that $A^= \mathbf{z} = \mathbf{b}^=, \mathbf{a}_i^\top \mathbf{z} \leq b^i, i \in I^{\leq} \backslash \{r\}$ and $\mathbf{a}_r^\top \mathbf{z} = b_r$, i.e., $\mathbf{z} \in \mathfrak{F}$. Therefore $\dim(\mathfrak{F}) \geq n - \operatorname{rank}(A^=, \mathbf{b}^=; \mathbf{a}_r^\top, b_r) = \dim(\mathfrak{P}) - 1$, contradiction. ■

Combining Proposition 11.4 and Proposition 11.5, we get

Theorem 11.2: Minimal Description of a Polyhedron

Any polyhedron $\mathfrak{P} \subseteq \mathbb{R}^n$ admits a minimal description $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^\top \mathbf{x} = b_i, i = 1, \ldots, n - \dim(\mathfrak{P}), \mathbf{a}_j^\top \mathbf{x} \leq b_j, j = n - \dim(\mathfrak{P}) + 1, \ldots, n - \dim(\mathfrak{P}) + t\}$, where t is the number of facets of $\mathfrak{P}$. In particular, $(A^=, \mathbf{b}^=)$ is unique up to nonsingular linear transformation while $(A^{\leq}, \mathbf{b}^{\leq})$ is unique up to positive scalar multiplication and addition of the row space of $(A^=, \mathbf{b}^=)$.

Proof: The existence of a minimal description is clear. If there exist two different minimal descriptions of the same polyhedron $\mathfrak{P} = \mathfrak{P}_1 = \mathfrak{P}_2$, then also $\mathfrak{P} = \mathfrak{P}_1 \cap \mathfrak{P}_2$, whence the uniqueness claim follows. ■

Proposition 11.6: Facet Characterization

Consider the polyhedron $\mathfrak{P}$ with equality set $(A^=, \mathbf{b}^=)$, then the proper face $\mathfrak{F} := \{\mathbf{x} \in \mathfrak{P} : \mathbf{c}^\top \mathbf{x} = d\}$ is a facet iff $\forall x \in \mathfrak{F}, \mathbf{e}^\top \mathbf{x} = f \implies (\mathbf{e}^\top, f) = \operatorname{Range}(A^=, \mathbf{b}^=; \mathbf{c}^\top, d)$

Proof: $\Rightarrow$: Follows from Theorem 11.2.

$\Leftarrow$: Since $\mathfrak{F}$ is proper, $(\mathbf{c}^\top, d)$ is linearly independent of $(A^=, \mathbf{b}^=)$. It follows that $\operatorname{rank}(A^=, \mathbf{b}^=; \mathbf{c}^\top, d) = n - \dim(\mathfrak{P}) + 1$. On the other hand, the system $\mathbf{e}^\top \mathbf{x} = f, \forall \mathbf{x} \in \mathfrak{F}$, with $(\mathbf{e}; f)$ indeterminate, has solution space whose dimension equals $n - \dim(\mathfrak{F})$. Therefore $\dim(\mathfrak{F}) = \dim(\mathfrak{P}) - 1$, i.e., $\mathfrak{F}$ is a facet. ■

Example 11.1: Permutahedron

Let X be the set of all permutations of $\{1,\dots,n\}$, we claim its convex hull is

$$\mathfrak{P} = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_i x_i = \binom{n}{2}, \sum_{i\in S} x_i \geq \binom{|S|+1}{2}, \forall S \subset \{1,\dots,n\} \right\}. \tag{161}$$

Indeed, it is easy to see that $\text{conv}(X) \subseteq \mathfrak{P}$ (for $X \subseteq \mathfrak{P}$). For the reverse inclusion, we verify two things: 1). $\dim(\text{conv}(X)) = \dim(\mathfrak{P})$, which is clear and ensures us that we did not miss any equality constraint in $\text{conv}(X)$; 2). each facet of $\text{conv}(X)$ is represented by some inequality in $\mathfrak{P}$. Once 2) is verified, together with $\text{conv}(X) \subseteq \mathfrak{P}$ we know that $\text{conv}(X) = \text{conv}(X) \cap \mathfrak{P}$ hence by Theorem 11.2 it follows that both the equality constraints in $\text{conv}(X)$ and $\mathfrak{P}$ are equivalent and therefore $\mathfrak{P} \subseteq \text{conv}(X)$.

To verify 2), let $\mathfrak{F}$ be a proper nonempty face of $\text{conv}(X)$, represented by some linear inequality $\mathbf{a}^\top \mathbf{x} \leq b$. Consider $Y = \text{argmax}\{\mathbf{a}^\top \mathbf{x} : \mathbf{x} \in X\}$ and let S be the indexes of the smallest entry in $\mathbf{a}$. Note that $|S| \neq n$ for otherwise the face $\mathfrak{F} = \text{conv}(X)$ will not be proper. By an exchange argument we know any permutation $\mathbf{y} \in Y$ must satisfy $y_i \in \{1,\dots,|S|\}$ for any $i \in S$, i.e., it will satisfy the inequality $\sum_{i\in S} x_i \leq \binom{|S|+1}{2}$ with equality. Since $|S| \neq n$, we have found an inequality in $\mathfrak{P}$ that represents the face $\mathfrak{F}$.

Proposition 11.7: Minimal Face

Any minimal face of the polyhedron $\mathfrak{P}$ has the form $\{\mathbf{a}_j^\top \mathbf{x} = b_j, j \in J\}$ for some $I^= \subseteq J \subseteq I$. In particular, if $\text{rank}(A^=; A^\leq) = n - k$, then $\mathfrak{P}$ has a minimal face with dimension k.

Proof: Take a minimal face which by Proposition 11.2 can be written as $\mathfrak{F} := \{\mathbf{a}_i^\top \mathbf{x} \leq b_i, i \in J', \mathbf{a}_j^\top \mathbf{x} = b_j, j \in J\}$. If there exists $r \in J', \mathbf{x}_r \in \mathfrak{F}$ such that $\mathbf{a}_r^\top \mathbf{x}_r < b_r$, then take an outside point $\mathbf{y}$ which satisfies $\mathbf{a}_i^\top \mathbf{y} \leq b_i, \forall i \in J\backslash\{r\}$, $\mathbf{a}_r^\top \mathbf{y} > b_r$ and $\mathbf{a}_i^\top \mathbf{y} = b_i, \forall i \in J$. There exists a point $\mathbf{z}$ on the line between $\mathbf{x}_r$ and $\mathbf{y}$ such that $\mathbf{a}_i^\top \mathbf{z} \leq b_i, i \in J' - \{r\}$ and $\mathbf{a}_j^\top \mathbf{z} = b_j, j \in J \cup \{r\}$, i.e., $\mathfrak{F}$ contains a smaller face, contradiction. The last claim follows from the fact that $\mathfrak{P}$ has at least one minimal face, and the number of (linearly independent) equality constraints that define any face cannot exceed the rank of A. ■

Proposition 11.8: Extreme Point and Ray

$\mathbf{p}$ is an extreme point of the polyhedron $\mathfrak{P}$ iff it is a zero-dimensional face of $\mathfrak{P}$; $\mathbf{r}$ is an extreme Ray iff $\text{cone}(\mathbf{r})$ is a one-dimensional face of the associated polyhedral cone $\mathfrak{P}0+ := \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \leq \mathbf{0}\}$.

Proof: We only prove the extreme direction case.

$\Rightarrow$: If $\mathbf{r}$ is a one-dimensional face, then for some A' with $\text{rank}(A') = n - 1$ we have $A'\mathbf{r} = \mathbf{0}$. Let $\mathbf{r}_1, \mathbf{r}_2 \in \mathfrak{P}0+$ satisfy $\mathbf{r} = \frac{\mathbf{r}_1+\mathbf{r}_2}{2}$, then $A'\mathbf{r}_1 \leq 0, A'\mathbf{r}_2 \leq 0$ while $A'\mathbf{r}_1 + A'\mathbf{r}_2 = 0$ hence $A'\mathbf{r}_1 = A'\mathbf{r}_2 = 0$. But $\text{rank}(A') = n - 1$, therefore we must have $\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r}$.

$\Leftarrow$: Consider the minimal face containing $\mathbf{r}$, whose dimension is at least 2. We can find an inner point $\mathbf{r}'$ in the same minimal face such that $\mathbf{r}' \neq \text{cone}(\mathbf{r})$, hence $\mathbf{r} = \frac{1}{2}(\mathbf{r} + \epsilon\mathbf{r}') + \frac{1}{2}(\mathbf{r} - \epsilon\mathbf{r}')$ for some small $\epsilon > 0$. ■

Theorem 11.3: Minkowski-Weyl Theorem

Polyhedral sets are precisely those in the form

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \sum_i \lambda_i \mathbf{p}_i + \sum_j \mu_j \mathbf{r}_j + \sum_k \nu_k \boldsymbol{\ell}_k, \sum_i \lambda_i = 1, \lambda_i \geq 0, \mu_j \geq 0\}. \tag{162}$$

In particular, $\boldsymbol{\ell}_k$ can be chosen as some basis of the lineality space L of the polyhedron $\mathfrak{P}$, and $\mathbf{p}_i, \mathbf{r}_j$ can be chosen as the extreme points and extreme rays of $\mathfrak{P} \cap L^\perp$, respectively.

Proof: Consider the polyhedron $\mathfrak{P}$ with lineality space L, then $\mathfrak{P} = L + (\mathfrak{P} \cap L^\perp)$. We only need to consider the set $\mathfrak{P}' := \mathfrak{P} \cap L^\perp$ which contains no lines. Let $\mathfrak{Q} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \sum_i \lambda_i \mathbf{p}_i + \sum_j \mu_j \mathbf{r}_j, \sum_i \lambda_i = 1, \lambda_i \geq 0, \mu_j \geq 0\}$ with $\mathbf{p}_i, \mathbf{r}_j$ being the extreme points and rays of $\mathfrak{P}'$. Clearly $\mathfrak{Q} \subseteq \mathfrak{P}'$. Suppose there exists $\mathbf{z} \in \mathfrak{P}' \backslash \mathfrak{Q}$, i.e., the system

$$\sum_i \lambda_i \mathbf{p}_i + \sum_j \mu_j \mathbf{r}_j = \mathbf{z}, \quad \sum_i \lambda_i = 1, \quad \lambda_i \geq 0, \quad \mu_j \geq 0$$

has no solution. By Farkas' lemma, there exists $(\mathbf{y}^\top, c)$ such that $\mathbf{p}_i^\top \mathbf{y} - c \geq 0, \forall i$, $\mathbf{r}_j^\top \mathbf{y} \geq 0, \forall j$ and $\mathbf{z}^\top \mathbf{y} - c < 0$. Consider the linear program $\min_{\mathbf{x} \in \mathfrak{P}'} \mathbf{x}^\top \mathbf{y}$, if the optimal value is finite then it is attained by some extreme point (since $\mathfrak{P}'$ contains no lines hence any face of it contains no lines either). Therefore the optima value is lower bounded by c, contradicting to $\mathbf{z}^\top \mathbf{y} < c, \mathbf{z} \in \mathfrak{P}'$. On the other hand, if the linear program is unbounded (from below), then it is unbounded on some extreme ray (since $\mathfrak{P}'$ contains no lines), i.e., for some $\mathbf{r}_k, \mathbf{r}_k^\top \mathbf{y} < 0$, again contradiction.

Conversely, if any set is written in the claimed form, it is the projection of a polyhedron, hence is itself a polyhedron. ■

To explicitly find the representing vectors $\boldsymbol{\ell}_k, \mathbf{p}_i, \mathbf{r}_j$ for the polyhedron $\mathfrak{P} = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \leq \mathbf{b}\}$, we first find the basis for the lineality space by solving $A\mathbf{x} = \mathbf{0}$. Then consider $\mathfrak{P}' = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \leq \mathbf{b}, \boldsymbol{\ell}_k^\top \mathbf{x} = 0, \forall k\}$ and find its zero-dimensional faces (by solving $A'\mathbf{x} = \mathbf{b}', \boldsymbol{\ell}_k^\top \mathbf{x} = 0, \forall k$). Lastly consider $\mathfrak{P}'0+$ and find its one-dimensional faces (by solving $A''\mathbf{x} = \mathbf{0}, \boldsymbol{\ell}_k^\top \mathbf{x} = 0, \forall k$). Clearly, if $A, \mathbf{b}$ are rational, all representing vectors $\boldsymbol{\ell}_k, \mathbf{p}_i, \mathbf{r}_j$ can be chosen rational too.

Definition 11.3: Blocking and Antiblocking Polar

Let $\mathfrak{P} = \{\mathbf{x} \in \mathbb{R}^n_+ : A\mathbf{x} \geq 1\}$ where $A \in \mathbb{R}^{m \times n}_+$ with nonzero rows, define its blocking polar as

$$\mathfrak{P}^b := \{\mathbf{y} \in \mathbb{R}^n_+ : \mathbf{y}^\top \mathbf{x} \geq 1, \forall \mathbf{x} \in \mathfrak{P}\}. \tag{163}$$

Similarly let $\mathfrak{Q} = \{\mathbf{x} \in \mathbb{R}^n_+ : A\mathbf{x} \leq 1\}$ where $A \in \mathbb{R}^{m \times n}_+$ with nonzero rows, define its antiblocking polar as

$$\mathfrak{Q}^a := \{\mathbf{y} \in \mathbb{R}^n_+ : \mathbf{y}^\top \mathbf{x} \leq 1, \forall \mathbf{x} \in \mathfrak{Q}\}. \tag{164}$$

Proposition 11.9: (Anti)blocking Polar is Self-dual

Let $\mathbf{p}_i, \mathbf{q}_j$ be extreme points of $\mathfrak{P}$ and $\mathfrak{Q}$ respectively, then

$$\mathfrak{P}^b = \{\mathbf{y} \in \mathbb{R}^n_+ : \mathbf{y}^\top \mathbf{p}_i \geq 1\}, \tag{165}$$
$$(\mathfrak{P}^b)^b = \mathfrak{P}, \tag{166}$$
$$\mathfrak{Q}^a = \{\mathbf{y} \in \mathbb{R}^n_+ : \mathbf{y}^\top \mathbf{q}_j \leq 1\}, \tag{167}$$
$$(\mathfrak{Q}^a)^a = \mathfrak{Q}. \tag{168}$$

Proof: Note that $\mathfrak{P} = \text{conv}(\{\mathbf{p}_i\}) + \mathbb{R}^n_+$ (for $A \in \mathbb{R}^{m \times n}_+$). Since $\mathfrak{P}^b \subseteq \mathbb{R}^n_+$, (165) follows. Clearly $\mathfrak{P} \subseteq (\mathfrak{P}^a)^a$ while on the other hand $\mathbf{a}_i \in \mathfrak{P}^a$ hence also $(\mathfrak{P}^a)^a \subseteq \mathfrak{P}$. ■

For $A \in \mathbb{R}^{m \times n}_+$ and $B \in \mathbb{R}^{r \times n}_+$ with nonzero rows, they are called a blocking pair if $\{\mathbf{x} \in \mathbb{R}^n_+ : A\mathbf{x} \geq 1\}$ are $\{\mathbf{y} \in \mathbb{R}^n_+ : B\mathbf{y} \geq 1\}$ are blocking polar to each other. Define similarly the antiblocking pair.

Theorem 11.4: Min-Max Duality for (Anti)Blocking Pair

$A \in \mathbb{R}_+^{m\times n}$ and $B \in \mathbb{R}_+^{r\times n}$ with nonzero rows are a blocking pair iff $\forall \mathbf{z} \in \mathbb{R}_+^n$,

$$\max\{\mathbf{1}^\top \mathbf{y} : A^\top \mathbf{y} \le \mathbf{z}, \mathbf{y} \in \mathbb{R}_+^m\} = \min_{1\le j\le r} \mathbf{z}^\top \mathbf{b}_j, \tag{169}$$

an antiblocking pair iff $\forall \mathbf{z} \in \mathbb{R}_+^n$,

$$\min\{\mathbf{1}^\top \mathbf{y} : A^\top \mathbf{y} \ge \mathbf{z}, \mathbf{y} \in \mathbb{R}_+^m\} = \max_{1\le j\le r} \mathbf{z}^\top \mathbf{b}_j. \tag{170}$$

Proof: Straightforward linear programming duality. ■

Proposition 11.10: When does $A\mathbf{x} = \mathbf{b}$ Have an Integral Solution?

Fix $A \in \mathbb{Q}^{m\times n}$ and $b \in \mathbb{Q}^m$. The system $A\mathbf{x} = \mathbf{b}$ has an integral solution iff $\forall \mathbf{y} \in \mathbb{Q}^m, A^\top \mathbf{y} \in \mathbb{Z}^n \implies \mathbf{b}^\top \mathbf{y} \in \mathbb{Z}$.

Remark 11.1: Finding an Integral Solution

We can also decide if $A\mathbf{x} = \mathbf{b}$ has an integral solution in polynomial time: Use Gaussian elimination to decide if $A\mathbf{x} = \mathbf{b}$ has any solution, if so, find a maximum number of linearly independent rows of A, denoted as A'. Employ (integral) elementary column transformations to reduce A' to its Hermite canonical form, i.e., find nonsingular integral U such that $A'U = \begin{bmatrix} B & \mathbf{0} \end{bmatrix}$ where B is lower triangular sharing the same rank with A. Then $A\mathbf{x} = \mathbf{b}$ has an integral solution iff $U \begin{bmatrix} B^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix}$ is an integral solution (note that $B^{-1}\mathbf{b} = B^{-1}A'\mathbf{y} = \begin{bmatrix} \mathsf{Id} & \mathbf{0} \end{bmatrix} U^{-1}\mathbf{y}$, where U^{-1} is also integral).

The next result tells us when the A matrix will always admit an integral solution for all integral $\mathbf{b}$.

Proposition 11.11: Pseudo-Unimodular (PUM)

Let $A \in \mathbb{Z}^{m\times n}$ with $\text{rank}(A) = r$. The following are equivalent:

(I). A can be converted into $\begin{bmatrix} \mathsf{Id}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ using (integral) elementary operations;

(II). the g.c.d. of the order r subdeterminants of A is 1;

(III). $\forall \mathbf{b} \in \mathbb{Z}^n$ the system $A\mathbf{x} = \mathbf{b}$ either has no solution or has an integral solution;

(IV). $\exists A^\dagger \in \mathbb{Z}^{n\times m}$ such that $AA^\dagger A = A$.

An integral matrix satisfying any of the above is called pseudo-unimodular (PUM). Moreover, if $r = m$, i.e., A is of full row rank, we have an additional equivalence

(V). $A^\top \mathbf{y} \in \mathbb{Z}^n \implies \mathbf{y} \in \mathbb{Z}^m$.

Proof: Note that if we interchange two columns or multiply some column by -1 or add some column to another column, we do not change the g.c.d. of the subdeterminants: each transformation is invertible and each subdeterminant after the transformation is an integral combination of the subdeterminants of the original matrix. Therefore we can convert A to $\begin{bmatrix} B & \mathbf{0} \\ A' & \mathbf{0} \end{bmatrix}$, where $B \in \mathbb{Z}_+^{r\times r}$ is lower triangular with positive entries on the diagonal. Similarly, using elementary row transformations (from right to

left) we can further convert A to $\begin{bmatrix} D & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, where $D \in \mathbb{Z}_+^{r\times r}$ is diagonal. From here it is clear that (I), (II) and (III) are all equivalent.

(III) $\implies$ (IV): We can, without loss of generality, assume $A = \begin{bmatrix} B \\ CB \end{bmatrix}$ where $B \in \mathbb{Z}^{r\times n}$ is of full row rank. The linear system $AX = \begin{bmatrix} \mathsf{Id} \\ C \end{bmatrix}$ has solutions hence by (III) has an integral solution $X^* \in \mathbb{Z}^{n\times r}$. Now one easily verifies that $A^\dagger := \begin{bmatrix} X^* & \mathbf{0}_{n,m-r} \end{bmatrix}$ satisfies $AA^\dagger A = A$.

(IV) $\implies$ (III): Suppose the linear system has a solution $\mathbf{z}$ (not necessarily integral), then $AA^\dagger A\mathbf{z} = A\mathbf{z} = \mathbf{b}$, i.e., $A^\dagger A\mathbf{z} = A^\dagger\mathbf{b}$ is an integral solution.

(IV) $\implies$ (V): Suppose A is of full row rank, then $AA^\dagger$ is nonsingular hence $\forall\mathbf{y}\exists\mathbf{w}$ such that $\mathbf{y} = (AA^\dagger)^\top\mathbf{w}$. Therefore if $A^\top\mathbf{y} = A^\top(AA^\dagger)^\top\mathbf{w} = (AA^\dagger A)^\top\mathbf{w} = A^\top\mathbf{w}$ is integral, so is $(A^\dagger)^\top A^\top\mathbf{w} = \mathbf{y}$.

Finally (V) $\implies$ (I) can be proved using again elementary transformations. ■

From the proof it is clear that (integral) elementary transformations preserve PUM. Note that the equivalence (V) is false if A is not of full row rank, for example $A = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$.

Definition 11.4: Unimodular (UM)

Matrix $A \in \mathbb{Z}^{m\times n}$ with $\operatorname{rank}(A) = r$ is called unimodular (UM) if for each submatrix B consisting of r linearly independent columns of A, the g.c.d. of the order r subdeterminants of B is 1.

Unimodularity is strictly stronger than pseudo-unimodularity: $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix}$ is not unimodular but apparently satisfies (II) in Proposition 11.11.

Remark 11.2: Operations Preserving Unimodularity

The following operations clearly preserve unimodularity:

- deleting columns;
- subtracting one row from another;
- permuting, replicating or negating columns/rows;
- adding zero rows/columns;
- taking direct products.

The first property is the main motivation to introduce unimodularity, as it is not possessed by pseudo-unimodular matrices.

Note that $\begin{bmatrix} \mathsf{Id} \\ A \end{bmatrix}$ is always unimodular while $\begin{bmatrix} \mathsf{Id} & A \end{bmatrix}$ need not be: take $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$. Also, subtracting one column from another does not preserve unimodularity.

We will use the notion of integer polyhedron to characterize unimodularity. Some basic properties of integral hull and integer polyhedra can be found in [Korte and Vygen, 2012, Chapter 5].

Definition 11.5: Integral Hull and Integer Polyhedra

The integral hull of the polyhedron $\mathfrak{P}$ is the convex hull of all integral points in $\mathfrak{P}$, denoted as $\mathfrak{P}_I$. Clearly $\mathfrak{P}_I \subseteq \mathfrak{P}$ and we call $\mathfrak{P}$ integral if $\mathfrak{P}_I = \mathfrak{P}$.

Recall that a column submatrix is called a basis if its rank is maximal.

Theorem 11.5: Characterizing Unimodularity

Let $A \in \mathbb{Z}^{m\times n}$. The following are equivalent:

(I). A is unimodular;

(II). $\forall \boldsymbol{\ell}, \mathbf{u} \in \mathbb{Z}^n_{\pm\infty}, \forall \mathbf{b} \in \mathbb{Z}^m$, the polyhedron $\mathfrak{P} := \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{b}, \boldsymbol{\ell} \le \mathbf{x} \le \mathbf{u}\}$ is integral;

(III). $\exists \boldsymbol{\ell}, \mathbf{u} \in \mathbb{Z}^n_{\pm\infty}, \mathbf{u} \ge \mathbf{1} + \boldsymbol{\ell}, |\boldsymbol{\ell}| \wedge |\mathbf{u}| < \infty, \forall \mathbf{b} \in \mathbb{Z}^m$, the polyhedron $\mathfrak{P}$ is integral;

(IV). $\forall \mathbf{c} \in \mathbb{Z}^n$ the polyhedron $\mathfrak{Q} := \{\mathbf{y} \in \mathbb{R}^m : A^\top \mathbf{y} \ge \mathbf{c}\}$ is integral;

(V). $\exists$ a basis B that is UM and $\exists$ a unique (T)UM matrix C such that $BC = A$;

(VI). $\forall$ basis B, it is UM and $\exists$ a unique (T)UM matrix C such that $BC = A$.

Proof: (I) $\implies$ (II): Let A be unimodular and $\mathfrak{F}$ be a minimal face of $\mathfrak{P}$, which is determined by $A'\mathbf{x}' = \mathbf{c}$ where A' is some column submatrix of A (cf. Proposition 11.7) and $\mathbf{c} \in \mathbb{Z}^m$. From Remark 11.2 we know A' is unimodular hence it follows from Proposition 11.11 that $\mathfrak{F}$ has an integral point. Since $\mathfrak{F}$ is chosen arbitrarily, $\mathfrak{P}$ is integral.

(I) $\implies$ (IV): Let $\mathfrak{F}$ be a minimal face of $\mathfrak{Q}$, then $\mathfrak{F} = \{\mathbf{y} \in \mathbb{R}^m : B^\top \mathbf{y} = \mathbf{c}'\}$, where B consists of linearly independent columns of A, therefore is unimodular. Proposition 11.11 implies that $\mathfrak{F}$ has an integral point hence $\mathfrak{Q}$ is integral.

(III) $\implies$ (I): We prove the contrapositive. Suppose A is not unimodular, w.l.o.g. let $A = \begin{bmatrix} B & C \end{bmatrix}$ with $\operatorname{rank}(A) = r$, $B \in \mathbb{Z}^{m\times r}$ has full column rank and the order r subdeterminants of B have g.c.d. greater than 1. Therefore by Proposition 11.11 $\exists \mathbf{y}' \notin \mathbb{Z}^r$ such that $B\mathbf{y}' \in \mathbb{Z}^m$. If necessary adding integers so that $\boldsymbol{\ell}_{1:r} \le \mathbf{y}' \le \mathbf{u}_{1:r}$. Set $\mathbf{y} = \begin{bmatrix} \mathbf{y}' \\ \mathbf{y}'' \end{bmatrix}$ where $\mathbf{y}''_i$ equals the finite entry in $\{\ell_{r+i}, u_{r+i}\}$. Consider the polyhedron $\{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = A\mathbf{y}, \boldsymbol{\ell} \le \mathbf{x} \le \mathbf{u}\}$, it has a non-integral extreme point $\{B\mathbf{x}_{1:r} = B\mathbf{y}', \mathbf{x}_{r+1:n} = \mathbf{y}''\}$, contradiction.

(IV) $\implies$ (I): We prove the contrapositive. Suppose A is not unimodular, then by definition $\exists B \in \mathbb{Z}^{m\times r}$ consisting of r linearly independent columns of A such that the order r subdeterminants of B have g.c.d. greater than 1. Therefore by Proposition 11.11 $\exists \mathbf{c} \in \mathbb{Z}^r$ such that the system $B^\top \mathbf{y} = \mathbf{c}$ has no integral solution, although some real-valued solution $\mathbf{y}'$ does exist (since B is of full column rank). Consider the polyhedron $\mathfrak{Q} := \{\mathbf{y} : A^\top \mathbf{y} \ge \lfloor A^\top \mathbf{y}' \rfloor\}$. Its face $\mathfrak{F} := \{\mathbf{y} \in \mathfrak{Q} : B^\top \mathbf{y} = \mathbf{c}\}$ is nonempty ($\mathbf{y}' \in \mathfrak{F}$) and contains no integral points, contradicting the fact that $\mathfrak{Q}$ is integral.

Finally we turn to the equivalence between (I), (V) and (VI). From Remark 11.2 we know that (integral) elementary *row* transformations preserve unimodularity. Moreover, the claims in (V) and (VI) are robust w.r.t. (integral) elementary *row* transformations as well. Therefore we may assume $A = \begin{bmatrix} U & D \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, where U is a nonsingular matrix sharing the same rank with A. Take $B = \begin{bmatrix} U \\ \mathbf{0} \end{bmatrix}$, then $C = \begin{bmatrix} \mathsf{Id} & U^{-1}D \end{bmatrix}$. Under (V), B is UM and C is UM (equivalently TUM, see (I) and (II) in Theorem 11.6 below). Checking the determinants verifies that $BC = A$ is also UM, hence proves that (V) $\implies$ (I). On the other hand, if A is UM, then $U = \mathsf{Id}$ and any basis of $\begin{bmatrix} U & D \end{bmatrix}$ must have determinant ± 1. Therefore w.l.o.g. $B = \begin{bmatrix} \mathsf{Id} \\ \mathbf{0} \end{bmatrix}$ is UM and $C = \begin{bmatrix} \mathsf{Id} & D \end{bmatrix}$ is UM (equivalently TUM), proving (I) $\implies$ (VI). ■

From the proof of (I) $\implies$ (II) and (I) $\implies$ (IV) it is clear that the column inheritability of unimodularity is the key (as compared to pseudo-unimodularity). On the other hand, pseudo-unimodularity is sufficient for proving $\neg$ (I) $\implies$ $\neg$ (III) and $\neg$ (I) $\implies$ $\neg$ (IV).

Remark 11.3: Testing Unimodularity

The UM factorization shown in (VI) of Theorem 11.5 can be employed to reduce testing UM to testing TUM (defined below): Find a basis B in A (using Gaussian elimination); use (integral) elementary column transformations to convert B to its Hermite canonical form, i.e., find nonsingular

UM matrix U such that $B^\top U = \begin{bmatrix} \mathsf{Id} & \mathbf{0} \end{bmatrix}$; finally check if $U^\top A = U^\top BC = \begin{bmatrix} C \\ \mathbf{0} \end{bmatrix}$ is TUM.

Definition 11.6: Totally Unimodular (TUM)

$A \in \{1, 0, -1\}^{m \times n}$ is totally unimodular (TUM) if all its subdeterminants belong to $\{1, 0, -1\}$.

There is an important property of TUM that we would like to mention: Recall that the columns of $A \in \mathbb{F}^{m \times n}$ are linearly independent (over the scalar field $\mathbb{F}$) iff there exists a submatrix $A' \in \mathbb{F}^{n \times n}$ with $\det_{\mathbb{F}}(A') \neq 0$. Now let $A \in \{1, 0, -1\}^{m \times n}$, if its columns are linearly independent over $\mathsf{GF}(2)$, i.e., the scalar field with two elements 0 and 1, then they are also linearly independent over $\mathbb{R}$, since for any square submatrix A', $\det_{\mathbb{R}}(A') \equiv \det_{\mathsf{GF}(2)}(A') \mod 2$. If A is furthermore TUM, then the reverse implication is also true because $\det_{\mathbb{R}}(A') = \{1, 0, -1\}$.

Remark 11.4: Operations Preserving TUM

The following operations, which can be verified using (V) in Theorem 11.6 below, preserve TUM:

- transposing;
- permuting, negating, replicating or deleting columns/rows;
- adding a row/column with at most one nonzero entry, being ± 1;
- pivoting, i.e., replacing $\begin{bmatrix} 1 & \mathbf{c}^\top \\ \mathbf{b} & D \end{bmatrix}$ by either $\begin{bmatrix} 1 & \mathbf{c}^\top \\ \mathbf{0} & D - \mathbf{b}\mathbf{c}^\top \end{bmatrix}$, $\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{b} & D - \mathbf{b}\mathbf{c}^\top \end{bmatrix}$, or $\begin{bmatrix} -1 & \mathbf{c}^\top \\ \mathbf{b} & D - \mathbf{b}\mathbf{c}^\top \end{bmatrix}$;
- 1-sum: $A \oplus_1 B := \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & B \end{bmatrix}$;
- 2-sum: $\begin{bmatrix} A & \mathbf{a} \end{bmatrix} \oplus_2 \begin{bmatrix} \mathbf{b}^\top \\ B \end{bmatrix} := \begin{bmatrix} A & \mathbf{a}\mathbf{b}^\top \\ \mathbf{0} & B \end{bmatrix}$;
- 3-sum: $\begin{bmatrix} A & \mathbf{a} & \mathbf{a} \\ \mathbf{c}^\top & 0 & 1 \end{bmatrix} \oplus_3 \begin{bmatrix} 1 & 0 & \mathbf{b}^\top \\ \mathbf{d} & \mathbf{d} & B \end{bmatrix} := \begin{bmatrix} A & \mathbf{a}\mathbf{b}^\top \\ \mathbf{d}\mathbf{c}^\top & B \end{bmatrix}$.

A beautiful result of Seymour shows that every totally unimodular matrix arises from these operations on the so-called network matrices and two special totally unimodular matrices:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & -1 \\ -1 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 & -1 \\ -1 & 0 & 0 & -1 & 1 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{171}$$

Moreover, it implies a polynomial-time algorithm for determining total unimodularity!

Example 11.2: 2-sum or 3-sum do not preserve UM

$\begin{bmatrix} 1 & 1 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ are both UM while their 2-sum $\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ is not UM. Similarly, $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ are both UM while their 3-sum $\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$ is not UM. Also, the last two pivoting rules do not preserve UM: take $\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$.

Proposition 11.12

Let $A \in \mathbb{R}^{m \times n}$ be of full row rank. The following are equivalent:

(I). $\forall$ basis B, the matrix $B^{-1}A$ is integral;

(II). $\forall$ basis B, the matrix $B^{-1}A$ is (T)UM;

(III). $\exists$ basis B such that the matrix $B^{-1}A$ is (T)UM.

Proof: Simply note that all claims are robust w.r.t. left multiplying A with any nonsingular matrix, hence we can assume w.l.o.g. that $A = \begin{bmatrix} \mathsf{Id}_m & C \end{bmatrix}$ (after permuting the columns if necessary). Then all claims are easily seen to be equivalent as requiring C to be TUM. ■

Theorem 11.6: Characterizing TUM

Let $A \in \{1, 0, -1\}^{m \times n}$, the following are equivalent:

(I). A is TUM;

(II). $\begin{bmatrix} \mathsf{Id} & A \end{bmatrix}$ is (T)UM;

(III). $\forall \boldsymbol{\ell}, \mathbf{u} \in \mathbb{Z}^n_{\pm\infty}, \forall \mathbf{c}, \mathbf{d} \in \mathbb{Z}^m_{\pm\infty}$, the polyhedron $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{c} \leq A\mathbf{x} \leq \mathbf{d}, \boldsymbol{\ell} \leq \mathbf{x} \leq \mathbf{u}\}$ is integral;

(IV). $\exists \boldsymbol{\ell}, \mathbf{u} \in \mathbb{Z}^n_{\pm\infty}, \mathbf{u} \geq \mathbf{1} + \boldsymbol{\ell}, |\mathbf{u}| \wedge |\boldsymbol{\ell}| < \infty, \exists \mathbf{c}, \mathbf{d} \in \mathbb{Z}^n_{\pm\infty}, \mathbf{d} \geq \mathbf{1} + \mathbf{c}, |\mathbf{c}| \wedge |\mathbf{d}| < \infty, \forall \mathbf{b} \in \mathbb{Z}^m$ the polyhedron $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{c} \leq A\mathbf{x} - \mathbf{b} \leq \mathbf{d}, \boldsymbol{\ell} \leq \mathbf{x} \leq \mathbf{u}\}$ is integral;

(V). $\forall S \subseteq \{1, \ldots, n\}$, $\exists S_1 + S_2 = S$ such that $\sum_{j \in S_1} A_{:,j} - \sum_{j \in S_2} A_{:,j} \in \{1, 0, -1\}^m$;

(VI). $\forall$ nonsingular submatrix B of A, $\exists i$ such that $|\{j : B_{i,j} \neq 0\}|$ is odd;

(VII). $\forall$ square, hence also rectangular, submatrix B of A, if $\sum_i B_{i,:}$ and $\sum_j B_{:,j}$ are even vectors then $\sum_{ij} B_{ij}$ is divisible by 4;

(VIII). $\forall$ nonsingular submatrix B of A, $|\det(B)| \neq 2$;

(IX). $\forall$ integral $\mathbf{b}, \mathbf{y} \geq 0$ and $\forall k \in \mathbb{N}$ such that $A\mathbf{y} \leq k\mathbf{b}$, $\exists$ integral $\mathbf{x}_i \in \{\mathbf{x} \geq \mathbf{0} : A\mathbf{x} \leq \mathbf{b}\}, i = 1, \ldots, k$ such that $\mathbf{y} = \sum_{i=1}^k \mathbf{x}_i$;

(X). $\forall$ nonsingular submatrix B of A and the g.c.d. of the entries in $\mathbf{y}^\top B$ for $\mathbf{y} = \mathbf{1}$, hence also for all $\{1, 0, -1\}$-valued $\mathbf{y}$, is 1.

Proof: (I) $\iff$ (II) is obvious while (II) $\implies$ (III) $\implies$ (IV) $\implies$ (II) follows from Theorem 11.5.

(III) $\implies$ (V): Consider the polyhedron $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}_S, \lfloor \frac{1}{2} \cdot A\mathbf{1}_S \rfloor \leq A\mathbf{x} \leq \lceil \frac{1}{2} \cdot A\mathbf{1}_S \rceil\}$. Clearly it is not empty and since A is unimodular it is integral by (III). Let $\mathbf{y}$ be an extreme point, which is integral, then $\mathbf{1}_S - 2\mathbf{y}$ gives the desired partition.

(V) $\implies$ (VI): By (V) there exists a $\{\pm 1\}$-valued vector $\mathbf{z}$ such that $B\mathbf{z}$ is $\{1, 0, -1\}$-valued. If $\forall i, |\{j : B_{i,j} \neq 0\}|$ is even, then $B|\mathbf{z}|$ is an even vector, therefore $B\mathbf{z} = 0$. But $\mathbf{z} \neq \mathbf{0}$, contradicting to the non-singularity of B.

(V) $\implies$ (VII): Since $\sum_j B_{:,j}$ is an even vector, by (V) we can partition the columns of B into two classes whose column sums coincide. Since $\sum_i B_{i,:}$ is an even vector, it follows that $\sum_{ij} B_{ij}$ is divisible by 4.

(VIII) $\implies$ (I): We prove the contrapositive. Suppose A is not TUM, then $\exists$ square submatrix $B \in \mathbb{R}^{t \times t}$ with $|\det(B)| \geq 2$. We will exhibit a further square submatrix of B that has determinant either 2 or -2. Consider the matrix $C := \begin{bmatrix} B & \mathsf{Id}_t \end{bmatrix}$. Iteratively adding or subtracting rows from or to other rows and multiplication of columns by -1, we convert C to C' so that 1). C' remains $\{1, 0, -1\}$-valued; 2). Id_t remains a column submatrix of C' (not necessarily in the same positions as in C); 3). C' contains among its first t columns as many columns of Id_t as possible. Note that the subdeterminants of the first t columns of C' remain the same as those of C (up to sign changes).

Permuting if necessary we may assume $C' = \begin{bmatrix} \mathsf{Id}_k & B_1 & \mathbf{0}_{k,t-k} \\ \mathbf{0}_{t-k,k} & B_2 & \mathsf{Id}_{t-k} \end{bmatrix}$. Clearly $k < t$ since otherwise $|\det(\mathsf{Id}_k)| = |\det(B)| > 2$. Therefore $\exists k+1 \le i,j \le t$ with $C'_{ij} = 1$. By the maximality of C' there must then exist a submatrix $\begin{bmatrix} 1 & \pm 1 \\ \pm 1 & -1 \end{bmatrix}$ in the first t columns. This submatrix has determinant -2.

(VI) or (VII) $\implies$ (VIII): We use induction. Suppose the theorem is true for all proper submatrices of A. Suppose (VIII) is false, then $\det_{\mathbb{R}}(A) = \pm 2 \equiv 0 \mod 2$, therefore A is singular over $\mathsf{GF}(2)$ (since A is $\{1,0,-1\}$-valued). On the other hand, any proper subcolumn matrix of A is linearly independent over $\mathbb{R}$ (since $\det_{\mathbb{R}}(A) \neq 0$) hence also linearly independent over $\mathsf{GF}(2)$ (since any proper submatrix of A is TUM by the induction hypothesis). Therefore we must have $A\mathbf{1} \equiv \mathbf{0} \mod 2$, which contradicts (VI).

Similarly one can show that $A^\top \mathbf{1} \equiv \mathbf{0} \mod 2$. Partition A as $\begin{bmatrix} \alpha & \mathbf{b}^\top \\ \mathbf{c} & A' \end{bmatrix}$. By the induction hypothesis, the submatrix $B := \begin{bmatrix} \mathbf{c} & A' \end{bmatrix}$ is TUM hence by (V) $\exists \mathbf{x} \in \{1,-1\}^n$ such that $B\mathbf{x} \in \{1,0,-1\}^{m-1}$. However $B\mathbf{1} \equiv 0 \mod 2$ hence we conclude that $B\mathbf{y} = 0$. The equality $A \begin{bmatrix} \mathbf{x} & C \end{bmatrix} = \begin{bmatrix} \beta & \mathbf{b}^\top \\ \mathbf{0} & A' \end{bmatrix}$, where $C := \begin{bmatrix} \mathbf{0} \\ \mathsf{Id}_{n-1} \end{bmatrix}$, tells us that $|\beta| = |\det(A)| = 2$ because $\det(A') = \pm 1$. Moreover, $\mathbf{1}^\top A(1-\mathbf{x}) \equiv 0 \mod 4$ due to the fact $1 - \mathbf{x} \equiv 0 \mod 2$ and the previously established result $A^\top \mathbf{1} \equiv \mathbf{0} \mod 2$. Finally note that $\mathbf{1}^\top A\mathbf{x} = \beta \equiv 2 \mod 4$ hence $\mathbf{1}^\top A\mathbf{1} \equiv 2 \mod 4$, contradicting to (VII).

(III) $\implies$ (IX): We use induction. $k = 1$ holds trivially. Let $k \ge 2$ and consider the polyhedron $\{\mathbf{x} : \mathbf{0} \le \mathbf{x} \le \mathbf{y}, A\mathbf{y} - k\mathbf{b} + \mathbf{b} \le A\mathbf{x} \le \mathbf{b}\}$, which is nonempty (containing $k^{-1}\mathbf{y}$) hence integral. Take any extreme point $\mathbf{x}_k$ and note that $\mathbf{y}' := \mathbf{y} - \mathbf{x}_k \ge 0$ and $A\mathbf{y}' \le (k-1)\mathbf{b}$.

(IX) $\implies$ (IV): Let $\mathbf{x}_0$ be an arbitrary extreme point of $\mathfrak{P} := \{\mathbf{x} \ge \mathbf{0} : A\mathbf{x} \le \mathbf{b}\}$. Take $k \in \mathbb{N}$ so that $\mathbf{y} := k\mathbf{x}_0$ is integral. Since $\mathbf{y} \ge \mathbf{0}, A\mathbf{y} \le k\mathbf{b}$, by (IX) $k\mathbf{x}_0 = \sum_{i=1}^k \mathbf{x}_i$, with $\mathbf{x}_i$ being integral vectors in $\mathfrak{P}$. Clearly we must then have $k = 1$ hence $\mathbf{x}_0$ is integral.

(I) $\implies$ (X): Clearly B^{-1} is integral. Let k be the g.c.d. of the entries in $\mathbf{y}^\top B$, then $k^{-1}\mathbf{y}^\top = (k^{-1}\mathbf{y}^\top B)B^{-1}$ is integral. Since $\mathbf{y}$ is $\{1,0,-1\}$-valued, we must have $k = 1$.

(X) $\implies$ (VI): Simply take $\mathbf{y} = \mathbf{1}$. ■

Definition 11.7: (Box) Totally Dual Integral (TDI)

The system $A\mathbf{x} \le \mathbf{b}$ is called box totally dual integral iff $\forall \boldsymbol{\ell}, \mathbf{u} \in \mathbb{F}^n, \forall \mathbf{c} \in \mathbb{Z}^n$, the dual problem of $\max\{\mathbf{c}^\top \mathbf{x} : A\mathbf{x} \le \mathbf{b}, \boldsymbol{\ell} \le \mathbf{x} \le \mathbf{u}\}$ has an integral solution (whenever there exists one). The system is simply called totally dual integral if $\boldsymbol{\ell} = -\infty, \mathbf{u} = \infty$.

Proposition 11.13: Box-TDI Implies TDI

Box-TDI $\forall \boldsymbol{\ell}, \mathbf{u} \in \mathbb{F}^n$ implies Box-TDI $\forall \boldsymbol{\ell}, \mathbf{u} \in \mathbb{F}^n_{\pm\infty}$.

Proof: Let $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{F}^n_{\pm\infty}$ and assume the dual of $\max\{\mathbf{c}^\top \mathbf{x} : A\mathbf{x} \le \mathbf{b}, \boldsymbol{\ell} \le \mathbf{x} \le \mathbf{u}\}$ has a minimizer. By the LP duality the primal problem has a maximizer, say $\mathbf{x}^*$. Consider $\max\{\mathbf{c}^\top \mathbf{x} : A\mathbf{x} \le \mathbf{b}, \mathbf{x}^* - \mathbf{1}_{\boldsymbol{\ell}=-\infty} \le \mathbf{x} \le \mathbf{x}^* + \mathbf{1}_{\mathbf{u}=\infty}\}$, of which $\mathbf{x}^*$ is still a maximizer, hence by assumption its dual has an integral solution. By KKT conditions, this dual integral solution (after dropping zeros if necessary) remains optimal for the dual of the original LP. ■

Theorem 11.7: (Box) TDI Implies Primal Integrality

If the system $A\mathbf{x} \le \mathbf{b}, A \in \mathbb{Q}^{m\times n}, \mathbf{b} \in \mathbb{Z}^m$ is (box) TDI, then the underlying polyhedron is integral.

Proof: By Proposition 11.13, we only need to consider the TDI case. By definition the dual LP always has an integral solution hence the primal LP always has integral value, whence the integrality of the polyhedron. ■

Theorem 11.8: Characterization of Box-TDI

A rational polyhedron $\mathfrak{P} \subseteq \mathbb{R}^n$ is box-TDI iff $\forall \mathbf{c} \in \mathbb{Q}^n, \exists \mathbf{d} \in \mathbb{Z}^n$ such that $\lfloor \mathbf{c} \rfloor \leq \mathbf{d} \leq \lceil \mathbf{c} \rceil$ and every maximizer of $\max\{\mathbf{c}^\top \mathbf{x} : \mathbf{x} \in \mathfrak{P}\}$ also maximizes $\max\{\mathbf{d}^\top \mathbf{x} : \mathbf{x} \in \mathfrak{P}\}$.

Proof: $\Rightarrow$: We use induction to prove that $\forall k \leq n$ there exists $\mathbf{d} \in \mathbb{Z}^n$ such that $d_i \leq \lceil c_i \rceil$, $d_i = c_i$ if $c_i \in \mathbb{Z}$, $d_i \geq \lfloor c_i \rfloor$ for all $i \leq k$, and the claim in Proposition 11.13 holds.

Let $A\mathbf{x} \leq \mathbf{b}$ be a box-TDI representation of $\mathfrak{P}$, and let $\mathbf{x}^* \in \mathbb{F}^n$ be an arbitrary point in the relative interior of the optimum set of $\max\{\mathbf{c}^\top \mathbf{x} : \mathbf{x} \in \mathfrak{P}\}$. Let $k = 0$ and consider

$$\max\{\mathbf{x}^\top \lceil \mathbf{c} \rceil : A\mathbf{x} \leq \mathbf{b}, \mathbf{x}_i \leq \mathbf{x}_i^* \text{ if } c_i \notin \mathbb{Z}\}.$$

Clearly, $\mathbf{x}^*$ remains feasible and because of $(\lceil \mathbf{c} \rceil - \mathbf{c})^\top (\mathbf{x}^* - \mathbf{x}) \geq 0$ it is actually optimal. By box-TDI, the dual LP has an integral optimum $\mathbf{y}, \mathbf{z} \geq 0$ such that $A^\top \mathbf{y} + \mathbf{z} = \lceil \mathbf{c} \rceil, \mathbf{b}^\top \mathbf{y} + \mathbf{z}^\top \mathbf{x}^* = \lceil \mathbf{c} \rceil^\top \mathbf{x}^*$ and $\mathbf{z}_i = 0$ if $\mathbf{c}_i \in \mathbb{Z}$. Take $\mathbf{d} := \lceil \mathbf{c} \rceil - \mathbf{z}$, then using $\mathbf{y}$ as a dual certificate we verify that $\mathbf{x}^*$ is optimal for $\max\{\mathbf{x}^\top \mathbf{d} : \mathbf{x} \in \mathfrak{P}\}$. Since $\mathbf{x}^*$ is chosen in the relative interior, we know all maximizers of $\max\{\mathbf{c}^\top \mathbf{x} : \mathbf{x} \in \mathfrak{P}\}$ remain optimal for $\max\{\mathbf{x}^\top \mathbf{d} : \mathbf{x} \in \mathfrak{P}\}$. This completes the proof for $k = 0$.

Now we prove for $k + 1$. By the induction hypothesis we know $\exists \mathbf{d} \in \mathbb{Z}^n$ such that $d_i \leq \lceil c_i \rceil$, $d_i = c_i$ if $c_i \in \mathbb{Z}$, $d_i \geq \lfloor c_i \rfloor$ for all $i \leq k$, and the claim in Proposition 11.13 holds. Suppose $d_{k+1} < \lfloor c_{k+1} \rfloor$. Consider the convex combination $\mathbf{f}$ of $\mathbf{d}$ and $\mathbf{c}$ such that $f_{k+1} = \lfloor c_{k+1} \rfloor$. Note that $\mathbf{x}^*$ remains optimal for $\max\{\mathbf{f}^\top \mathbf{x} : \mathbf{x} \in \mathfrak{P}\}$, therefore by the induction hypothesis, $\exists \mathbf{e} \in \mathbb{Z}^n$ such that $e_i \leq \lceil f_i \rceil$, $e_i = f_i$ if $f_i \in \mathbb{Z}$, $e_i \geq \lfloor f_i \rfloor$ for all $i \leq k$, and the claim in Proposition 11.13 holds. Clearly, $e_i \leq \lceil f_i \rceil \leq \lceil c_i \rceil$; if $c_i \in \mathbb{Z}$, then $c_i = d_i = f_i = e_i$; $e_{k+1} = f_{k+1} = \lfloor c_{k+1} \rfloor$; and finally $e_i \geq \lfloor f_i \rfloor \geq \lfloor c_i \rfloor$ for $i \leq k$. Thus we have proved the case for $k + 1$. Let $k = n$ completes the proof.

$\Leftarrow$: We know every rational polyhedron admits a TDI representation [Korte and Vygen, 2012, Theorem 5.17, page 110]. Let $A\mathbf{x} \leq \mathbf{b}$ be such a TDI representation. We need only prove that $A\mathbf{x} \leq \mathbf{b}$ is box-TDI. Consider $\max\{\mathbf{e}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}, \boldsymbol{\ell} \leq \mathbf{x} \leq \mathbf{u}\}$ with some maximizer $\mathbf{x}^*$. By LP duality, there exist $\mathbf{y}, \mathbf{z}, \mathbf{w} \geq 0$ such that

$$\begin{aligned} A^\top \mathbf{y} + \mathbf{z} - \mathbf{w} &= \mathbf{e} \\ \mathbf{b}^\top \mathbf{y} + \mathbf{u}^\top \mathbf{z} - \boldsymbol{\ell}^\top \mathbf{w} &= \mathbf{e}^\top \mathbf{x}^* \\ z_i w_i &= 0, \forall i. \end{aligned}$$

Define $\mathbf{c} = A^\top \mathbf{y} = \mathbf{e} - \mathbf{z} + \mathbf{w}$. By assumption $\exists \mathbf{d} \in \mathbb{Z}^n$ such that $\lfloor \mathbf{c} \rfloor \leq \mathbf{d} \leq \lceil \mathbf{c} \rceil$ and such that each maximizer of $\max\{\mathbf{c}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$ also maximizes $\max\{\mathbf{d}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$. Using $\mathbf{y}$ as a dual certificate we verify that $\mathbf{x}^*$ is optimal for $\max\{\mathbf{c}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$ hence also optimal for $\max\{\mathbf{d}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}\}$. Since $A\mathbf{x} \leq \mathbf{b}$ is TDI, there exists $\mathbf{v} \in \mathbb{Z}_+^m$ such that $A^\top \mathbf{v} = \mathbf{d}$ and $\mathbf{b}^\top \mathbf{v} = \mathbf{d}^\top \mathbf{x}^*$. Define $\mathbf{g} = (\mathbf{e} - \mathbf{d})_+$ and $\mathbf{h} = (\mathbf{d} - \mathbf{e})_+$. Note that $x_i^* \leq u_i \implies z_i = 0 \implies c_i \geq e_i \implies d_i \geq e_i \implies g_i = 0$ hence $\mathbf{g}^\top (\mathbf{u} - \mathbf{x}^*) = 0$. Similarly $\mathbf{h}^\top (\mathbf{x}^* - \boldsymbol{\ell}) = 0$. Therefore

$$\begin{aligned} A^\top \mathbf{v} + \mathbf{g} - \mathbf{h} &= A^\top \mathbf{v} + \mathbf{e} - \mathbf{d} = \mathbf{e} \\ \mathbf{b}^\top \mathbf{v} + \mathbf{u}^\top \mathbf{g} - \boldsymbol{\ell}^\top \mathbf{h} &= \mathbf{d}^\top \mathbf{x}^* + \mathbf{u}^\top \mathbf{g} - \boldsymbol{\ell}^\top \mathbf{h} = \mathbf{d}^\top \mathbf{x}^* + \mathbf{g}^\top \mathbf{x}^* - \mathbf{h}^\top \mathbf{x}^* = \mathbf{e}^\top \mathbf{x}^*, \end{aligned}$$

i.e., $(\mathbf{v}, \mathbf{g}, \mathbf{h})$ is an integral solution of the dual of $\max\{\mathbf{e}^\top \mathbf{x} : A\mathbf{x} \leq \mathbf{b}, \boldsymbol{\ell} \leq \mathbf{x} \leq \mathbf{u}\}$ (for integral $\mathbf{e}$). ■

It follows from the proof that any TDI representation of a box-TDI polyhedron is actually box-TDI.

Next, we discuss the "robustness" of TDI under various operations. Clearly multiplying both sides by an arbitrary scalar does not preserve TDI, however, division is a different story:

Proposition 11.14

If $A\mathbf{x} \leq \mathbf{b}$ is (box)TDI then $k^{-1} A \leq \alpha \mathbf{b}$ is also (box)TDI for all $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}_+$.

Even more interestingly, we have

Theorem 11.9

For each rational system $A\mathbf{x} \leq \mathbf{b}$ there exists $k \in \mathbb{N}$ such that $(k^{-1}A)\mathbf{x} \leq (k^{-1}\mathbf{b})$ is TDI.

Proof: As mentioned before, there is a rational TDI system $C\mathbf{x} \leq \mathbf{d}$ that yields the same polyhedron as $A\mathbf{x} \leq \mathbf{b}$. For any row $\mathbf{c}$ of C, $\min\{\mathbf{b}^\top\mathbf{y} : \mathbf{y} \geq 0, A^\top\mathbf{y} = \mathbf{c}\}$ has a rational minimizer, whose components has k_c the least common multiple of their denominators. Take k as the least common multiple of all k_c.

Consider $\mathbf{e} \in \mathbb{Z}^n$ with $\delta = \max\{\mathbf{e}^\top\mathbf{x} : A\mathbf{x} \leq \mathbf{b}\} = \min\{\mathbf{b}^\top\mathbf{y} : \mathbf{y} \geq 0, A^\top\mathbf{y} = \mathbf{e}\} < \infty$. Since $C\mathbf{x} \leq \mathbf{d}$ is TDI, $(\mathbf{e}^\top, \delta)$ is a nonnegative integral combination of $(\mathbf{c}^\top, d)$, which itself is a nonnegative integral combination of $(k^{-1}\mathbf{a}^\top, k^{-1}b)$. Therefore $\exists \mathbf{z} \in k^{-1}\mathbb{Z}^n_+$ such that $\mathbf{b}^\top\mathbf{z} = \delta, A^\top\mathbf{z} = \mathbf{e}$. This easily implies the TDI of $\{(k^{-1}A)\mathbf{x} \leq (k^{-1}\mathbf{b})\}$. ■

Proposition 11.15

If $A\mathbf{x} \leq \mathbf{b}$ is (box)TDI, then $\forall \mathbf{w} \in \mathbb{R}^n$, $A\mathbf{x} \leq \mathbf{b} - A\mathbf{w}$ is (box)TDI.

Proposition 11.16

Let $A_i\mathbf{x} \leq \mathbf{b}_i, i \in \{1, 2\}$ represent the same polyhedron. If each inequality in $A_1\mathbf{x} \leq \mathbf{b}_1$ is a nonnegative integral combination of the inequalities in $A_2\mathbf{x} \leq \mathbf{b}_2$, then (box)TDI of the former implies (box)TDI of the latter.

Proposition 11.17: Adding/Removing Slacks

Consider the rational system $A\mathbf{x} \leq \mathbf{b}$. Let $\mathbf{a} \in \mathbb{Z}^n, \beta \in \mathbb{Q}$. The system $A\mathbf{x} \leq \mathbf{b}, \mathbf{a}^\top\mathbf{x} \leq \beta$ is (box)TDI iff $A\mathbf{x} \leq \mathbf{b}, \mathbf{a}^\top\mathbf{x} + \eta = \beta, \eta \geq 0$ is (box)TDI (with η an added slack variable).

Proposition 11.18: Intersection and Projection

Intersection and projection to a coordinate hyperplane preserves box-TDI.

Proof: Consider intersection to, say $x_1 = \alpha$. It amounts to adding $\alpha \leq x_1 \leq \alpha$, hence maintains box-TDI. For projection, apply Theorem 11.8 (by padding zero to $\mathbf{c}$). ■

Proposition 11.19: Repetition

If $A\mathbf{x} \leq \mathbf{b}$ is (box)TDI, then $\mathbf{a}x_0 + A\mathbf{x} \leq \mathbf{b}$ is also (box)TDI, where $\mathbf{a}$ is the first column of A and x_0 is a new variable.

Proof: The claim for the TDI case follows from the definition. Apply Theorem 11.8 for box-TDI. ■

The next result shows that box-TDI polyhedra are quite special.

Proposition 11.20

The box-TDI polyhedron admits the representation $A\mathbf{x} \leq \mathbf{b}$ for some $\{1, 0, -1\}$-valued matrix A.

Proof: Let $C\mathbf{x} \leq \mathbf{d}$ be some box-TDI system representing the box-TDI polyhedron $\mathfrak{P}$. Note that $\mathbf{p} \in \mathfrak{P}$ iff

$$\max\{-\mathbf{1}^\top\mathbf{z} + \mathbf{1}^\top\mathbf{w} : C\mathbf{x} + C\mathbf{w} + C\mathbf{z} \leq \mathbf{d}, \mathbf{x} = \mathbf{p}, \mathbf{z} \geq 0, \mathbf{w} \leq 0\} \geq 0. \tag{172}$$

By Proposition 11.19, $C\mathbf{x} + C\mathbf{w} + C\mathbf{z} \leq \mathbf{d}$ is box-TDI, hence (172) is equivalent to

$$\min\{\mathbf{d}^\top \mathbf{y} - \mathbf{y}^\top C\mathbf{p} : \mathbf{y} \geq 0, -\mathbf{1} \leq C^\top \mathbf{y} \leq \mathbf{1},\ \mathbf{y} \text{ integral},\ C^\top \mathbf{y} \text{ integral}\} \geq 0.$$

Therefore

$$\mathfrak{P} = \{\mathbf{x} : (\mathbf{y}^\top C)\mathbf{x} \leq \mathbf{d}^\top \mathbf{y},\ \text{for all } \mathbf{y} \geq 0, -\mathbf{1} \leq C^\top \mathbf{y} \leq \mathbf{1},\ \mathbf{y} \text{ integral},\ C^\top \mathbf{y} \text{ integral}\}.$$

Note that there are finitely many $C^\top \mathbf{y}$ satisfying the above while the right hand side $\mathbf{d}^\top \mathbf{y}$ is attainable by some $\mathbf{y}$ satisfying above. In summary, $\mathfrak{P}$ is defined by some $\{1, 0, -1\}$-valued matrix A. ■

Proposition 11.21: Domination

If $\mathfrak{P}$ is box-TDI, then its dominant $\mathfrak{Q} := \{\mathbf{z} : \mathbf{z} \geq \mathbf{x} \text{ for some } \mathbf{x} \in \mathfrak{P}\}$ is also box-TDI.

Proof: This also follows easily from Theorem 11.8. ■

Proposition 11.22: Substitution

If $\mathfrak{P}$ is box-TDI, then its dominant $\mathfrak{Q} := \{\mathbf{z} : \mathbf{z} \geq \mathbf{x} \text{ for some } \mathbf{x} \in \mathfrak{P}\}$ is also box-TDI.

Proof: This also follows easily from Theorem 11.8. ■

Proposition 11.23: Schur Complement

If $\mathfrak{P}$ is box-TDI, then its dominant $\mathfrak{Q} := \{\mathbf{z} : \mathbf{z} \geq \mathbf{x} \text{ for some } \mathbf{x} \in \mathfrak{P}\}$ is also box-TDI.

Proof: This also follows easily from Theorem 11.8. ■

Example 11.3: Operations Not Preserving TDI

The system $A = \begin{bmatrix} 1 & 5 \\ 1 & 6 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is TDI, however, $A\mathbf{x} - A\mathbf{y} \leq \mathbf{b}, \mathbf{x} \geq 0, \mathbf{y} \geq 0$ is not TDI.

The systems $A\mathbf{x} \leq \mathbf{b}_i$ with $A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}, \mathbf{b}_1 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 0 \end{bmatrix}, \mathbf{b}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ are both TDI, however, $A\mathbf{x} \leq \mathbf{b}_1 + \mathbf{b}_2$ is not TDI.

Theorem 11.10: Carathéodory's Theorem

(173)

References

Robert J. Aumann and Lloyd S. Shapley. *Values of non-atomic games*. Princeton University Press, 1974.

Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

Rodney C. Bassanezi and Gabriele H. Greco. Sull'additività dell'integrale. *Rendiconti del Seminario Matematico della Università di Padova*, 72:249–275, 1984.

Garrett Birkhoff. *Lattice Theory*. AMS, 2nd edition, 1948.

Niv Buchbinder, Moran Feldman, Joseph (Seffi) Naor, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *IEEE* 53^{rd} *Annual Symposium on Foundations of Computer Science*, pages 649–658, 2012.

Gustave Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

Claude Dellacherie. Quelques commentaires sur les prolongements de capacités. *Séminaire de Probabilités (Strasbourg)*, 5:77–81, 1971.

Dieter Denneberg. *Non-Additive Measure and Integral*. Kluwer Academic, 1994.

Jack Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial structures and their applications*, pages 69–87. 1970.

Uriel Feige, Vahab Mirrokni, and Jan Vondràk. Maximizing non-monotone submodular functions. *SIAM Journal of Computing*, 40(4):1133–1153, 2011.

András Frank. An algorithm for submodular functions on graphs. In *Bonn Workshop on Combinatorial Optimization*, volume 66, pages 97–120. North-Holland, 1982.

Satoru Fujishige. *Submodular Functions and Optimization*. Elsevier, 2nd edition, 2005.

Gabriele H. Greco. Sulla rappresentazione di funzionali mediante integrali. *Rendiconti del Seminario Matematico della Università di Padova*, 66:21–42, 1982.

Heinz König. The (sub/super) additivity assertion of choquet. *Studia Mathematica*, 157:171–197, 2003.

Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, 5th edition, 2012.

László Lovász. Submodular functions and convexity. In *Mathematical programming: the state of the art*, pages 235–257. 1982.

Massimo Marinacci and Luigi Montrucchio. On concavity and supermodularity. *Journal of Mathematical Analysis and Applications*, 344:642–654, 2008.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, 14:265–294, 1978.

David Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97(2):255–261, 1986.

Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1986.

Lloyd S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1(1):11–26, 1971.

Ján Šipoš. Integral representations of non-linear functionals. *Mathematica Slovaca*, 29(4):333–345, 1979.

Yaoliang Yu. On decomposing the proximal map. In *Advances in Neural Information Processing Systems*, 2013.