Submodular Functions: from Discrete to Continous Domains

Francis Bach

Submodular Functions: from Discrete to Continous Domains

Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

February 23, 2016

Abstract

Submodular set-functions have many applications in combinatorial optimization, as they can be minimized and approximately maximized in polynomial time. A key element in many of the algorithms and analyses is the possibility of extending the submodular set-function to a convex function, which opens up tools from convex optimization. Submodularity goes beyond set-functions and has naturally been considered for problems with multiple labels or for functions defined on continuous domains, where it corresponds essentially to cross second-derivatives being nonpositive. In this paper, we show that most results relating submodularity and convexity for set-functions can be extended to all submodular functions. In particular, (a) we naturally define a continuous extension in a set of probability measures, (b) show that the extension is convex if and only if the original function is submodular, (c) prove that the problem of minimizing a submodular function is equivalent to a typically non-smooth convex optimization problem, and (d) propose another convex optimization problem with better computational properties (e.g., a smooth dual problem). Most of these extensions from the set-function situation are obtained by drawing links with the theory of multi-marginal optimal transport, which provides also a new interpretation of existing results for set-functions. We then provide practical algorithms to minimize generic submodular functions on discrete domains, with associated convergence rates.

1 Introduction

Submodularity has emerged as an important concept in combinatorial optimization, akin to convexity in continuous optimization, with many applications in machine learning, computer vision or signal processing [7, 31, 34, 1]. Most of the literature on submodular functions focuses on *set-functions*, i.e., functions defined on the set of subsets of a given base set. Such functions are classically equivalently obtained as functions defined on the vertices of the hypercube $\{0,1\}^n$, if n is the cardinality of the base set. Throughout the paper, we will make this identification and refer to *set-functions* as functions defined on $\{0,1\}^n$.

Like convex functions, submodular set-functions can be minimized exactly in polynomial time, either through combinatorial algorithms akin to max-flow algorithms [48, 23, 42], or algorithms based on a convex optimization techniques [1, Section 10]. Unlike convex functions, submodular set-functions can also be *maximized* approximately in polynomial time with simple greedy algorithms that come with approximation guarantees [41, 16].

In this paper, we focus primarily on submodular function minimization and links with convexity. In the set-function situation, it is known that submodular function minimization is equivalent to a convex optimization problem, which is obtained by considering a *continuous extension* of the submodular function, from vertices of the hypercube $\{0,1\}^n$ to the full hypercube $[0,1]^n$. This extension, usually referred to as the Choquet integral [10] or the Lovász extension [36], is convex if and only if the original set-function is submodular. Moreover, when the set-function is submodular, minimizing the original set-function or the convex extension is equivalent. Finally, simple and efficient algorithms based on generic convex optimization algorithms may be used for minimization (see, e.g., [1]).

The main goal of this paper is to show that all of these results naturally extend to all submodular functions defined more generally on subsets of $\mathbb{R}^n$. In this paper, we focus on functions defined on subsets of $\mathbb{R}^n$ of the form $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, where each $\mathcal{X}_i$ is a *compact* subset of $\mathbb{R}$. A function $H : \mathcal{X} \to \mathbb{R}$, is then submodular if and only if for all $(x, y) \in \mathcal{X} \times \mathcal{X}$,

$$H(x) + H(y) \geqslant H(\max\{x, y\}) + H(\min\{x, y\}),$$

where the min and max operations are applied component-wise. This extended notion of submodularity has been thoroughly studied [35, 52], in particular in economics [37].

Finite sets. Some of the results on submodular set-functions (which correspond to $\mathcal{X}_i = \{0, 1\}$ for all $i \in \{1, \dots, n\}$) have already been extended, such as the possibility of minimizing discrete functions (i.e., when all $\mathcal{X}_i$'s are finite) in polynomial time. This is done usually by a reduction to the problem of minimizing a submodular function on a ring family [48]. Moreover, particular examples, such as certain cuts with ordered labels [22, 43, 3] lead to min-cut/max-flow reformulations with efficient algorithms. Finally, another special case corresponds to functions defined as sums of local functions, where it is known that the usual linear programming relaxations are tight for submodular functions (see [57, 59] and references therein). Moreover, some of these results extend to continuous Markov random fields [56, 45], but those results depend primarily on the decomposable structure of graphical models, while the focus of our paper is independent of decomposability of functions in several factors.

Infinite sets. While finite sets already lead to interesting applications in computer vision [22, 43, 3], functions defined on products of sub-intervals of $\mathbb{R}$ are particularly interesting. Indeed, when twice-differentiable, a function is submodular if and only if all cross-second-order derivatives are non-positive, i.e., for all $x \in \mathcal{X}$:

$$\frac{\partial^2 H}{\partial x_i \partial x_j}(x) \leqslant 0.$$

In this paper, we provide simple algorithms based solely on function evaluations to minimize all of these functions. This thus opens up a new set of "simple" functions that can be efficiently minimized, which neither is included nor includes convex functions, with potentially many interesting theoretical or algorithmic developments.

In this paper, we make the following contributions, all of them are extensions of the set-function case:

- We propose in Section 3.1 a continuous extension in a set of probability measures and show in Section 3.4 that it is is convex if and only if the function is submodular and that for submodular functions, minimizing the extension, which is a convex optimization problem, is equivalent to minimizing the original function. This is made by drawing links with the theory of optimal transport, which provides simple intuitive proofs (even for submodular set-functions).

– We show in Section 3.5 that minimizing the extension plus a well-defined separable convex function is equivalent to minimizing a series of submodular functions. This may be useful algorithmically because the resulting optimization problem is strongly convex and thus may be easier to solve using duality with Frank-Wolfe methods [2].

– For finite sets, we show in Section 4 a direct link with existing notions for submodular set-functions, such as the base polytope. In the general situation, two polyhedra naturally emerge (instead of one). Moreover, the greedy algorithm to maximize linear functions on these polyhedra are also extended.

– For finite sets, we provide in Section 5 two sets of algorithms for minimizing submodular functions, one based on a non-smooth optimization problem on a set of measures (projected subgradient descent), and one based on smooth functions and "Frank-Wolfe" methods (see, e.g., [24, 2] and references therein). They can be readily applied to all situations by discretizing the sets if continuous. We provide in Section 6 a simple experiment on a one-dimensional signal denoising problem.

In this paper, we assume basic knowledge of convex analysis (see, e.g., [6, 5]), while the relevant notions of submodular analysis (see, e.g., [19, 1]) and optimal transport (see, e.g., [54, 46]) will be rederived as needed.

2 Submodular functions

Throughout this paper, we consider a *continuous function* $H : \mathcal{X} = \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$, defined on the product of n *compact* subsets $\mathcal{X}_i$ of $\mathbb{R}$, and thus equipped with a *total order*. Typically, $\mathcal{X}_i$ will be a finite set such as $\{0, \dots, k_i - 1\}$, where the notion of continuity is vacuous, or an interval (which we refer to as a *continuous domain*).

2.1 Definition

The function H is said to be *submodular* if and only if [35, 52]:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{X}, \;\; H(x) + H(y) \geqslant H(\min\{x, y\}) + H(\max\{x, y\}), \tag{1}$$

where the min and max operations are applied component-wise. An important aspect of submodular functions is that the results we present in this paper only rely on considering sets $\mathcal{X}_i$ with a *total order*, from $\{0, 1\}$ (where this notion is not striking) to sub-intervals of $\mathbb{R}$. For submodular functions defined on more general lattices, see [19, 53].

Like for set-functions, an equivalent definition is that for any $x \in \mathcal{X}$ and (different) basis vectors e_i, e_j and $a_i, a_j \in \mathbb{R}_+$ such that $x_i + a_i \in \mathcal{X}_i$ and $x_j + a_j \in \mathcal{X}_j$, then

$$H(x + a_i e_i) + H(x + a_j e_j) \geqslant H(x) + H(x + a_i e_i + a_j e_j). \tag{2}$$

Moreover, we only need the statement above in the limit of a_i and a_j tending to zero (but different from zero and such that $x_i + a_i \in \mathcal{X}_i$ and $x_j + a_j \in \mathcal{X}_j$): for discrete sets included in $\mathbb{Z}$, then we only need to consider $a_i = a_j = 1$, while for continuous sets, this will lead to second-order derivatives.

Modular functions. We define modular functions as functions H such that both H and $-H$ are submodular. This happens to be equivalent to H being a *separable* function, that is a function which

is a sum of n functions that depend arbitrarily on single variables [52, Theorem 3.3]. These will play the role that linear functions play for convex functions. Note that when each $\mathcal{X}_i$ is a sub-interval of $\mathbb{R}$, this set of functions is much larger than the set of linear functions.

Submodularity-preserving operations. Like for set-functions, the set of submodular functions is a cone, that is, the sum of two submodular functions is submodular and multiplication by a positive scalar preserves submodularity. Moreover, restrictions also preserve submodularity: any function defined by restriction on a product of subsets of $\mathcal{X}_i$ is submodular. This will be useful when discretizing a continuous domain in Section 5.

Moreover, submodularity is invariant by *separable strictly increasing reparameterizations*, that is, if for all $i \in \{1, \dots, n\}$, $\varphi_i : \mathcal{X}_i \to \mathcal{X}_i$ is a strictly increasing bijection, H is submodular, if and only if, $x \mapsto H\big[\varphi_1(x_1), \dots, \varphi_n(x_n)\big]$ is submodular. Note the difference with convex functions which are invariant by *affine* reparameterizations.

Finally, partial minimization does preserve submodularity, that is, if $H : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ is submodular so is $(x_1, \dots, x_k) \mapsto \inf_{x_{k+1}, \dots, x_n} H(x)$ (exact same proof as for set-functions), while maximization does not, that is the pointwise maxima of two submodular functions may not be submodular.

Set of minimizers of submodular functions. Given a submodular function, the set $\mathcal{M}$ of minimizers of H is a sublattice of $\mathcal{X}$, that is, if if $(x, y) \in \mathcal{M} \times \mathcal{M}$, then $\max\{x, y\}$ and $\min\{x, y\}$ are also in $\mathcal{M}$ [52].

Strict submodularity. We define the notion of strict submodularity through a strict inequality in Eq. (1) for any two x and y which are not comparable [52], where, like in the rest of this paper, we consider the partial order on $\mathbb{R}^n$ such that $x \leqslant x'$ if and only if $x_i \leqslant x'_i$ for all $i \in \{1, \dots, n\}$. Similarly, this corresponds to a strict inequality in Eq. (2) as soon as both a_i and a_j are strictly positive. The set of minimizers of a strictly submodular function is a chain, that is the set $\mathcal{M}$ of minimizers is totally ordered [52].

2.2 Examples

In this section, we provide simple examples of submodular functions. We will use as running examples throughout the paper: submodular set-functions defined on $\{0, 1\}^n$ (to show that our new results directly extend the ones for set-functions), modular functions (because they provide a very simple example of the concepts we introduce), and functions that are sums of terms $\varphi_{ij}(x_i - x_j)$ where φ_{ij} is convex (for the link with Wasserstein distances between probability measures [54, 46]).

- **Set-functions**: When each $\mathcal{X}_i$ has exactly two elements, e.g., $\mathcal{X}_i = \{0, 1\}$, we recover exactly submodular set-functions defined on $\{1, \dots, n\}$, with the usual identification of $\{0, 1\}^n$ with the set of subsets of $\{1, \dots, n\}$. Many examples may be found in [1, 19], namely cut functions, entropies, set covers, rank functions of matroids, network flows, etc.
- **Functions on intervals**: When each $\mathcal{X}_i$ is a interval of $\mathbb{R}$ and H is twice differentiable on $\mathcal{X}$, then H is submodular if and only if all cross-second-derivatives are non-negative, i.e.,

$$\forall i \neq j, \forall x \in \mathcal{X}, \ \frac{\partial^2 H}{\partial x_i \partial x_j}(x) \leqslant 0.$$

  This can be shown by letting a_i and a_j tend to zero in Eq. (2). A sufficient condition for strict submodularity is that the cross-order derivatives are strictly negative. As shown in this paper,

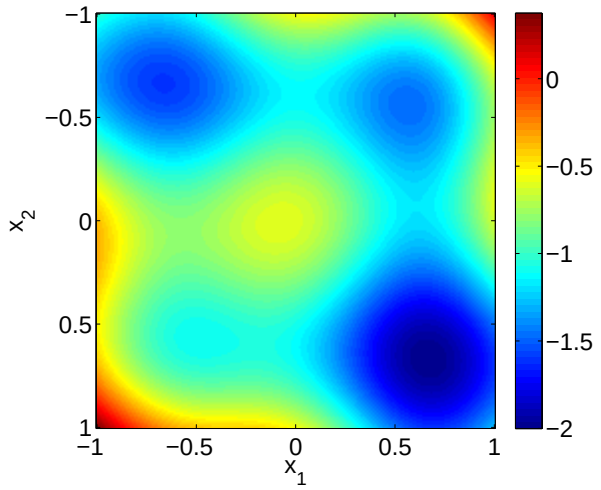


Figure 1: Level sets of the submodular function $(x_1, x_2) \mapsto \frac{7}{20}(x_1 - x_2)^2 - e^{-4(x_1 - \frac{2}{3})^2} - \frac{3}{5} e^{-4(x_1 + \frac{2}{3})^2} - e^{-4(x_2 - \frac{2}{3})^2} - e^{-4(x_2 + \frac{2}{3})^2}$, with several local minima, local maxima and saddle points.

this class of functions can be minimized efficiently while having potentially many local minima and stationary points (see an example in Figure 1).

A quadratic function $x \mapsto x^\top Q x$ is submodular if and only if all off-diagonal elements of Q are non-positive, a class of quadratic functions with interesting behavior, e.g., tightness of semi-definite relaxations [28], which is another instance of the good behavior of such functions.

The class of submodular functions includes functions of the form $\varphi_{ij}(x_i - x_j)$ for $\varphi_{ij} : \mathbb{R} \to \mathbb{R}$ convex, and $x \mapsto g(\sum_{i=1}^n \lambda_i x_i)$ for g concave and $(\lambda_i)_{i=1,\dots,n}$ non-negative weights; this gives examples of functions which are submodular, but convex or concave.

Other examples are $\varphi(\sum_{i=1}^n \lambda_i x_i) - \sum_{i=1}^n \lambda_i \varphi(x_i)$ for φ strictly concave and $\lambda \in \mathbb{R}^n$ in the interior of the simplex, which is non-negative and zero if and only if all x_i are equal.

Moreover, functions of the form $x \mapsto \log\det\big(\sum_{i=1}^n x_i A_i\big)$, where A_i are positive definite matrices and $x \geqslant 0$, are submodular—this extends to other spectral functions [18]. Moreover, if g is the Lovász extension of a submodular set-function, then it is submodular (as a function defined on $\mathbb{R}^n$)—see proof in Appendix A.1. These give examples of functions which are both convex and submodular. Similarly, the multi-linear extension of a submodular set-function [55], defined on $[0,1]^n$, is submodular as soon as the original set-function is submodular (see Appendix A.2), but is not convex in general.

For algorithms, these functions will be approximated on a discrete grid (one separate grid per variable, with a total complexity which is linear in the dimension n), but most of our formulations and convex analysis results extend in the continuous setting with appropriate regularity assumptions.

– **Discrete labels**: in this paper, we will often consider the case where the sets $\mathcal{X}_i$ are all finite. They will serve as approximations of functions defined on intervals. We will still use a functional notation to make the extension to continuous settings explicit. Examples of functions are naturally obtained from restrictions of functions defined on continuous intervals. Moreover, as shown in [52, Theorem 5.2], any Lipschitz-continuous submodular function defined on a product of subsets of $\mathbb{R}^n$ may be extended into a Lipschitz-continuous function on $\mathbb{R}^n$ (with the same constant).

– **Log-supermodular densities**: Submodular functions have also been studied as negative log-densities of probability distributions. These distributions are referred to as "multivariate totally positive of order 2" and classical examples are the multivariate logistic, Gamma and F distributions, as well as characteristic roots of random Wishart matrices (see more examples and additional properties in [27]).

Submodular minimization problems. In this paper, we focus on simple and efficient methods to minimize general submodular functions, based only on function evaluations. Many examples come from signal and image processing, with functions to minimize of the form

$$H(x) = \sum_{i=1}^{n} f_i(x_i) + \sum_{C \in \mathcal{C}} f_C(x_C),$$

where $\mathcal{C}$ is a set of small subsets (often a set of edges) and each f_C is submodular (while each f_i may be arbitrary) [22, 43, 3]. We consider a simple one-dimensional example in Section 6 as an illustration.

Another motivating example is a probabilistic modelling problem where submodularity on continuous domains appears naturally, namely probabilistic models on $\{0,1\}^n$ with log-densities which are negatives of submodular functions [11, 12], that is $\gamma(x) = \frac{1}{Z}\exp(-F(x))$, with F submodular and Z the normalizing constant equal to $Z = \sum_{x \in \{0,1\}^n} \exp(-F(x))$, which is typically hard to compute. In this context, *mean field* inference aims at approximating p by a product of independent distributions $\mu(x) = \prod_{i=1}^n \mu_i(x_i)$, by minimizing the Kullback-Leibler divergence between μ and γ, that is, by minimizing

$$\sum_{x \in \mathcal{X}} \mu(x) \log \frac{\mu(x)}{\gamma(x)} = \sum_{i=1}^{n} \big\{ \mu_i(1) \log \mu_i(1) + \mu_i(0) \log \mu_i(0) \big\} + \sum_{x \in \mathcal{X}} \mu(x) F(x) + Z.$$

The first term in the right-hand side is separable, hence submodular, while the second term is exactly the multi-linear extension of the submodular set-function F, which is itself a submodular function (see [55] and Appendix A.2). This implies that in this context, mean field inference may be done globally with arbitrary precision in polynomial time. Note that in this context, previous work [11] has considered replacing the multi-linear extension by the Lovász extension, which is also submodular but also convex (it then turns out to correspond to a different divergence than the KL divergence between μ and γ).

3 Extension to product probability measures

Our goal is to minimize the function H through a tight convex relaxation. Since all our sets $\mathcal{X}_i$ are subsets of $\mathbb{R}$, we could look for extensions to $\mathbb{R}^n$ directly such as done for certain definitions of discrete convexity [15, 20]; this in fact exactly the approach for functions defined on $\{0,1\}^n$, where one defines extensions on $[0,1]^n$. The view that we advocate in this paper is that $[0,1]$ is in bijection with the set of distributions on $\{0,1\}$ (as the probability of observing 1).

When the sets $\mathcal{X}_i$ have more than two elements, we are going to consider the convex set $\mathcal{P}(\mathcal{X}_i)$ of *Radon probability measures* [44] μ_i on $\mathcal{X}_i$, which is the closure (for the weak topology) of the convex hull of all Dirac measures; for $\mathcal{X}_i = \{0, \dots, k_i - 1\}$, this is essentially a simplex in dimension k_i. In order to get an *extension*, we look for a function defined on the set of *products of probability measures* $\mu \in \mathcal{P}^{\otimes}(\mathcal{X}) = \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$, such that if all μ_i, $i = 1, \dots, n$, are Dirac measures at points $x_i \in \mathcal{X}_i$, then we have a function value equal to $H(x_1, \dots, x_n)$.

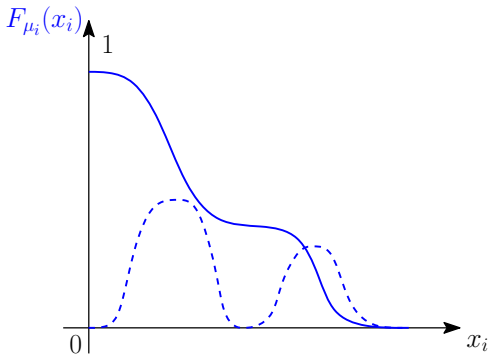


Figure 2: Cumulative function for a "continuous" distribution on the real line, with the corresponding density (with respect to the Lebesgue measure) in dotted.

We will define two types of extensions for all functions, not necessarily submodular, one based on inverse cumulative distribution functions, one based on convex closure. The two will happen to be identical for submodular functions.

3.1 Extension based on inverse cumulative distribution functions

For a probability distribution $\mu_i \in \mathcal{P}(\mathcal{X}_i)$ defined on a totally ordered set $\mathcal{X}_i$, we can define the (reversed) cumulative distribution function $F_{\mu_i} : \mathcal{X}_i \to [0,1]$ as $F_{\mu_i}(x_i) = \mu_i\big(\{y_i \in \mathcal{X}_i, y_i \geqslant x_i\}\big)$. This is a non-increasing left-continuous function from $\mathcal{X}_i$ to $[0,1]$. Such that $F_{\mu_i}(\min \mathcal{X}_i) = 1$ and $F_{\mu_i}(\max \mathcal{X}_i) = \mu_i(\{\max \mathcal{X}_i\})$. See illustrations in Figure 2 and Figure 3 (left).

When $\mathcal{X}_i$ is assumed discrete with k_i elements, it may be exactly represented as a vector in $\mathbb{R}^{k_i - 1}$ elements with non-decreasing components, that is, given μ_i, we define $\mu_i(x_i) + \cdots + \mu_i(k_i - 1) = F_{\mu_i}(x_i)$, for $x_i \in \{1, \dots, k_i - 1\}$. Because the measure μ_i has unit total mass, $F_{\mu_i}(0)$ is always equal to 1 and can thus be omitted to obtain a simpler representation (as done in Section 4). For example, for $k_i = 2$ (and $\mathcal{X}_i = \{0,1\}$), then we simply have $F_{\mu_i}(1) \in [0,1]$ which represents the probability that the associated random variable is equal to 1.

Note that in order to preserve the parallel with submodular set-functions, we choose to deviate from the original definition of the cumulative function by considering the mass of the set $\{y_i \in \mathcal{X}_i, y_i \geqslant x_i\}$ (and not the other direction).

We can define the "inverse" cumulative function from $[0,1]$ to $\mathcal{X}_i$ as

$$F^{-1}_{\mu_i}(t_i) = \sup\{x_i \in \mathcal{X}_i,\ F_{\mu_i}(x_i) \geqslant t_i\}.$$

The function $F^{-1}_{\mu_i}$ is non-increasing and right-continuous, and such that $F^{-1}_{\mu_i}(1) = \min \mathcal{X}_i$ and $F^{-1}_{\mu_i}(0) = \max \mathcal{X}_i$. Moreover, we have $F_{\mu_i}(x_i) \geqslant t_i \Leftrightarrow F^{-1}_{\mu_i}(t_i) \geqslant x_i$. See an illustration in Figure 3 (right). When $\mathcal{X}_i$ is assumed discrete, $F^{-1}_{\mu_i}$ is piecewise constant with steps at every t_i equal to $F_{\mu_i}(x_i)$ for a certain x_i. For $k_i = 2$, we get $F^{-1}_{\mu_i}(t_i) = 1$ if $t_i < \mu_i(1)$ and 0 if $t_i \geqslant \mu_i(1)$. What happens at $t_i = \mu_i(1)$ does not matter because this corresponds to a set of zero Lebesgue measure.

We now define our extension from $\mathcal{X}$ to the set of product probability measures, by considering a single threshold t applied to all n cumulative distribution functions. See an illustration in Figure 4.

Definition 1 (Extension based on cumulative distribution functions) *Let $H : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ be any continuous function. We define the extension $h_{\text{cumulative}}$ of H to $\mathcal{P}^{\otimes}(\mathcal{X}) = \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$ as*

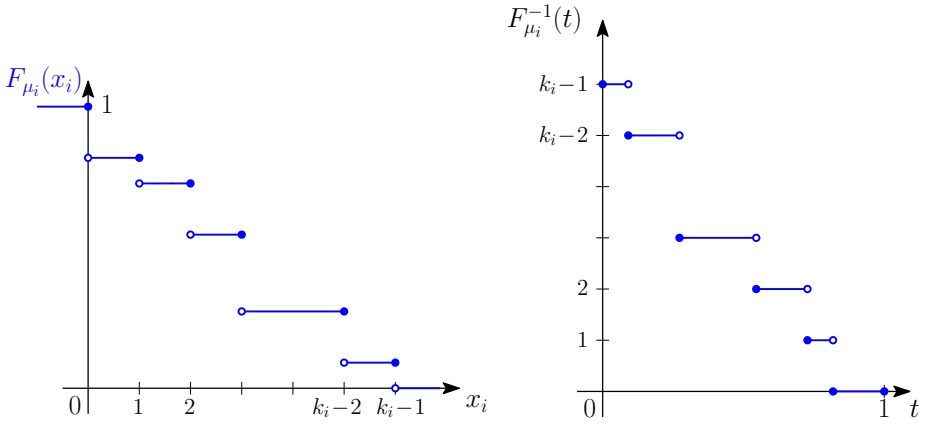


Figure 3: Left: cumulative function for a distribution on the real line supported in the set $\{0, \dots, k_i - 1\}$. Right: inverse cumulative function (which would be the same for the distribution with discrete domain).

follows:

$$\forall \mu \in \prod_{i=1}^{n} \mathcal{P}(\mathcal{X}_i), \quad h_{\text{cumulative}}(\mu_1, \dots, \mu_n) = \int_0^1 H\big[F^{-1}_{\mu_1}(t), \dots, F^{-1}_{\mu_n}(t)\big] dt. \tag{3}$$

If all μ_i, $i = 1, \dots, n$ are Diracs at $x_i \in \mathcal{X}_i$, then for all $t \in (0, 1)$, $F^{-1}_{\mu_i}(t) = x_i$ and we indeed have the extension property (again, what happens for $t = 0$ or $t = 1$ is irrelevant because this is on a set of zero Lebesgue measure). For $\mathcal{X}_i = \{0, 1\}$ for all i, then the extension is defined on $[0, 1]^n$ and is equal to $h_{\text{cumulative}}(\mu) = \int_0^1 H(1_{\{\mu(1) \geqslant t\}}) dt$ and we exactly recover the Choquet integral (i.e., the Lovász extension) for set-functions (see [1, Prop. 3.1]). These properties are summarized in the following proposition.

Proposition 1 (Properties of extension) *For any function $H : \prod_{i=1}^{n} \mathcal{X}_i \to \mathbb{R}$, the extension $h_{\text{cumulative}}$ satistfies the following properties:*

- *If μ is a Dirac at $x \in \mathcal{X}$, then $h_{\text{cumulative}}(\mu) = H(x)$.*
- *If all $\mathcal{X}_i$ are finite, then $h_{\text{cumulative}}$ is piecewise affine.*

Note that the extension is defined on all tuples of measures $\mu = (\mu_1, \dots, \mu_n)$ but it can equivalently be defined through non-increasing functions from $\mathcal{X}_i$ to $[0, 1]$, e.g., the representation in terms of cumulative distribution functions F_{μ_i} defined above. As we will see for discrete domains in Section 4, it may also be defined for all non-increasing functions with no contraints to be in $[0, 1]$. Moreover, this extension can be easily computed, either by sampling, or, when all $\mathcal{X}_i$ are finite, by sorting all values of $F_{\mu_i}(x_i)$, $i \in \{1, \dots, n\}$ and $x_i \in \mathcal{X}_i$ (see Section 4 for details).

Examples. For our three running examples, we may look at the extension. For set-functions, we recover the usual Choquet integral; for modular functions $H(x) = \sum_{i=1}^{n} H_i(x_i)$, then we have $h(\mu) = \sum_{i=1}^{n} \int_{\mathcal{X}_i} H_i(x_i) d\mu_i(x_i)$ which is the expectation of $H(x)$ under the product measure defined by μ. Finally, for the function $\varphi_{ij}(x_i - x_j)$, we obtain a Wasserstein distance between the measures μ_i and μ_j (which is a distance between their cumulative functions) [54]. See more details in Section 3.3.

3.2 Extension based on convex closures

We now describe a second way of extending a function defined on $\mathcal{X}$ to a function defined on $\mathcal{P}^{\otimes}(\mathcal{X}) = \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$, using the concept of *convex closure*. We consider the function g defined on $\mathcal{P}^{\otimes}(\mathcal{X})$ as $g(\mu) = H(x)$ if μ is the Dirac measure δ_x at $x \in \mathcal{X}$, and $+\infty$ otherwise. The following proposition gives an expression for its convex closure, in terms of Kantorovich multi-marginal optimal transport [8, 54], which looks for a joint probability measure on $\mathcal{X}$ with given marginals on each $\mathcal{X}_i$, $i = 1, \dots, n$.

Proposition 2 (Extension by convex closure - duality) *Assume all $\mathcal{X}_i$ are compact subsets of $\mathbb{R}$, for $i \in \{1, \dots, n\}$ and that $H : \mathcal{X} \to \mathbb{R}$ is continuous. The largest lower-semi continuous (for the weak topology on Radon measures) convex function $h : \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i) \to \mathbb{R}$ such that $h(\delta_x) \leqslant H(x)$ for any $x \in \mathcal{X}$ is equal to*

$$h_{\text{closure}}(\mu_1, \dots, \mu_n) = \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x), \tag{4}$$

where the infimum is taken over all probability measures γ on $\mathcal{X}$ such that the i-th marginal γ_i is equal to μ_i. Moreover, the infimum is attained and we have the dual representation:

$$h_{\text{closure}}(\mu) = \sup_v \sum_{i=1}^n \int_{\mathcal{X}_i} v_i(x_i) d\mu_i(x_i) \text{ such that } \forall x \in \mathcal{X}, \ \sum_{i=1}^n v_i(x_i) \leqslant H(x_1, \dots, x_n), \tag{5}$$

over all continuous functions $v_i : \mathcal{X}_i \to \mathbb{R}$, $i = 1, \dots, n$. We denote by $\mathcal{V}(H)$ the set of such potentials $v = (v_1, \dots, v_n)$.

Proof The function h_{closure} can be computed as the Fenchel bi-conjugate of g. The first step is to compute $g^*(v)$ for v in the dual space to $\prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$, that is $v \in \prod_{i=1}^n \mathcal{C}(\mathcal{X}_i)$, with $\mathcal{C}(\mathcal{X}_i)$ the set of continuous functions on $\mathcal{X}_i$. We have, by definition of the Fenchel-Legendre dual function, with $\langle \mu_i, v_i \rangle = \int_{\mathcal{X}_i} v_i(x_i) d\mu_i(x_i)$ the integral of v_i with respect to μ_i:

$$g^*(v) = \sup_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \sum_{i=1}^n \langle \mu_i, v_i \rangle - g(\mu) = \sup_{x \in \mathcal{X}} \sum_{i=1}^n v_i(x_i) - H(x).$$

This supremum is equal to

$$g^*(v) = \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \Big\{ \sum_{i=1}^n v_i(x_i) - H(x) \Big\} d\gamma(x) \tag{6}$$

over all probability measures γ on $\mathcal{X}$. We may then expand using γ_i the i-th marginal of γ on $\mathcal{X}_i$ defined as $\gamma_i(A_i) = \gamma(\{x \in \mathcal{X}, \ x_i \in A_i\})$ for any measurable set $A_i \subset \mathbb{R}$, as follows:

$$g^*(v) = \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{i=1}^n \int_{\mathcal{X}_i} v_i(x_i) d\gamma_i(x_i) - \int_{\mathcal{X}} H(x) d\gamma(x) \right\}.$$

The second step is to compute the bi-dual $g^{**}(\mu) = \sup_{v \in \prod_{i=1}^n \mathcal{C}(\mathcal{X}_i)} \sum_{i=1}^n \langle \mu_i, v_i \rangle - g^*(v)$ for $\mu \in \prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$:

$$\begin{aligned} g^{**}(\mu) &= \sup_{v \in \prod_{i=1}^n \mathcal{C}(\mathcal{X}_i)} \sum_{i=1}^n \langle v_i, \mu_i \rangle - \sup_{\gamma \in \mathcal{P}(\mathcal{X})} \Big\{ \sum_{i=1}^n \sum_{x_i \in \mathcal{X}_i} v_i(x_i) \gamma_i(x_i) - \int_{\mathcal{X}} H(x) d\gamma(x) \Big\} \\ &= \sup_{v \in \prod_{i=1}^n \mathcal{C}(\mathcal{X}_i)} \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^n \int_{\mathcal{X}_i} v_i(x_i) \big(d\gamma_i(x_i) - d\mu_i(x_i) \big) - \int_{\mathcal{X}} H(x) d\gamma(x) \\ &= \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \sup_{w \in \prod_{i=1}^n \mathcal{C}(\mathcal{X}_i)} \sum_{i=1}^n \int_{\mathcal{X}_i} v_i(x_i) \big(d\gamma_i(x_i) - d\mu_i(x_i) \big) - \int_{\mathcal{X}} H(x) d\gamma(x). \end{aligned}$$

In the last equality, we use strong duality which holds here because of the continuity of H and the compactness of all sets $\mathcal{X}_i$, $i = 1, \dots, n$. See for example [54] for details. Note that the infimum in γ is always attained in this type of optimal transport problems.

Thus, by maximizing over each $v_i \in \mathcal{C}(\mathcal{X}_i)$, we get and additional constraint and thus $g^{**}(\mu) = \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x)$ such that $\forall i,\ \gamma_i = \mu_i$. This leads to the desired result. ■

The extension by convex closure h_{closure} has several interesting properties, independently of the submodularity of H, as we now show.

Proposition 3 (Properties of convex closure) *For any continuous function $H : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$, the extension h_{closure} satisfies the following properties:*

(a) If μ is a Dirac at $x \in \mathcal{X}$, then $h_{\text{closure}}(\mu) \leqslant H(x)$.

(b) The function h_{closure} is convex.

(c) Minimizing h_{closure} on $\prod_{i=1}^n \mathcal{P}(\mathcal{X}_i)$ and minimizing H on $\prod_{i=1}^n \mathcal{X}_i$ is equivalent, that is, the two optimal values are equal, and one may find minimizers of one problem given the other one.

Proof Property (a) is obvious from the definition. Note that in general, the inequality may be strict (it will not for submodular functions). Since the objective function and constraint set in Eq. (4) are jointly convex in γ and μ, the infimum with respect to γ is thus convex in μ, which implies property (b). In order to show (c), we note that $\inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h_{\text{closure}}(\mu)$ is trivially less than $\inf_{x \in \mathcal{X}} H(x)$ because of (a), and we consider the sequence of equalities:

$$\begin{aligned} \inf_{x \in \mathcal{X}} H(x) &= \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x) = \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \inf_{\gamma \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} H(x) d\gamma(x) \text{ such that } \forall i,\ \gamma_i = \mu_i \\ &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu). \end{aligned}$$

Moreover, any optimal γ is supported on the (compact) set of minimizers of H on $\mathcal{X}$. Thus any optimal μ is the set of marginals of any distribution γ supported on the minimizers of H. ■

While the convex closure is attractive because it is convex and allows the minimization of H, the key difficulty in general is that h_{closure} cannot be computed in general. These are opposite properties to the extension $h_{\text{cumulative}}$ based on cumulative distribution functions. We now show that the two extensions are equal when H is submodular.

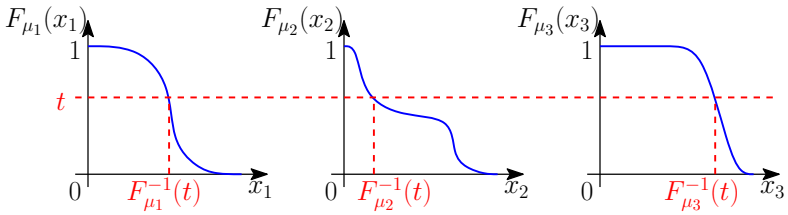


Figure 4: Multi-marginal optimal transport by thresholding inverse cumulative distribution functions: definition of the transport plan γ_{mon}.

3.3 Equivalence between the two extensions through one-dimensional optimal transport

We have seen two possible extensions of $H : \mathcal{X} \to \mathbb{R}$ to $h : \mathcal{P}^{\otimes}(\mathcal{X}) \to \mathbb{R}$. When H is submodular, the two are equal, as a consequence of the following proposition, which is itself obtained directly from the theory of multi-marginal optimal transport between one-dimensional distributions [8].

Proposition 4 (One-dimensional multi-marginal transport) *Let $\mathcal{X}_i$ be a compact subset of $\mathbb{R}$, for $i \in \{1, \dots, n\}$ and H a continuous submodular function defined on $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$. Then the two functions $h_{\text{cumulative}}$ defined in Eq. (3) and h_{closure} defined in Eq. (4) are equal.*

Proof Since [8] does not provide exactly our result, we give a detailed proof here from first principles. In order to prove this equivalence, there are three potential proofs: (a) based on convex duality, by exhibiting primal and dual candidates (this is exactly the traditional proof for submodular set-functions [36, 13], which is cumbersome for continuous domains, but which we will follow in Section 4 for finite sets); (b) based on Hardy-Littlewood's inequalities [35]; or (c) using properties of optimal transport. We consider the third approach (based on the two-marginal proof of [46]) and use four steps:

(1) We define $\gamma_{\text{mon}} \in \mathcal{P}(\mathcal{X})$ as the distribution of $(F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)) \in \mathcal{X}$ for t uniform in $[0, 1]$. See an illustration in Figure 4. The extension $h_{\text{cumulative}}$ corresponds to the distribution γ_{mon} and we thus need to show that γ_{mon} is an optimal distribution. In this context, probability distributions, i.e., elements of $\mathcal{P}(\mathcal{X})$ with given marginals are often referred to as "transport plan", a terminology we now use.

The transport plan γ_{mon} is trivially "monotone" so that two elements of its support are comparable for the partial order $x \leqslant x'$ if all components of x are less than or equal to the corresponding components of x'. Moreover, is it such that $\gamma_{\text{mon}}\Big(\prod_{i=1}^n \big\{y_i \in \mathcal{X}_i, y_i \geqslant x_i\big\}\Big)$ is the Lebesgue measure of the set of $t \in [0, 1]$ such that $F_{\mu_i}^{-1}(t) \geqslant x_i$ for all i, that is such that $F_{\mu_i}(x_i) \leqslant t$ for all i, thus

$$\gamma_{\text{mon}}\Big(\prod_{i=1}^n \big\{y_i \in \mathcal{X}_i, y_i \geqslant x_i\big\}\Big) = \max_{i \in \{1,\dots,n\}} F_{\mu_i}(x_i). \tag{7}$$

(2) We show that if $\gamma \in \mathcal{P}(\mathcal{X})$ is a distribution so that any two elements $x, x' \in \mathcal{X}$ of its support are comparable, then it is equal to γ_{mon}. We simply need to compute the mass of a product of rectangle as in Eq. (7). For $n = 2$ marginals, we consider the 4 possible combinations of the sets $\big\{y_i \in \mathcal{X}_i, y_i \geqslant x_i\big\}$, $i \in \{1, 2\}$ and their complements. because of the comparability assumption, either $\big\{y_1 \in \mathcal{X}_1, y_1 \geqslant x_1\big\} \times \big\{y_2 \in \mathcal{X}_2, y_2 < x_2\big\}$ or $\big\{y_1 \in \mathcal{X}_1, y_1 < x_1\big\} \times \big\{y_2 \in \mathcal{X}_2, y_2 \geqslant x_2\big\}$ is empty (see Figure 5), which implies that the measure of $\big\{y_1 \in \mathcal{X}_1, y_1 \geqslant x_1\big\} \times \big\{y_2 \in \mathcal{X}_2, y_2 \geqslant$

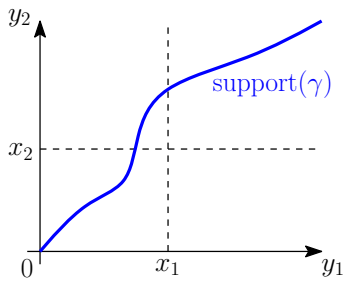


Figure 5: Monotone transport plan γ such that $\big\{y_1 \in \mathcal{X}_1, y_1 \geqslant x_1\big\} \times \big\{y_2 \in \mathcal{X}_2, y_2 < x_2\big\}$ is empty, leading to $\gamma\Big(\big\{y_1 \in \mathcal{X}_1, y_1 \geqslant x_1\big\} \times \big\{y_2 \in \mathcal{X}_2, y_2 \geqslant x_2\big\}\Big) = F_{\mu_1}(x_1)$.

$x_2\big\}$ is either $F_{\mu_1}(x_1)$ or $F_{\mu_2}(x_2)$, and hence larger than the maximum of these two. The fact that it is lower is trivial, hence the result.

(3) If H is strictly submodular, any optimal transport plan $\gamma \in \mathcal{P}(\mathcal{X})$ satisfies the property above of monotone support. Indeed, let us assume that x and x' are two non-comparable elements of the support of γ. From convex duality used in the proof of Proposition 2, in particular, Eq. (6), there exist continuous potentials $v_i : \mathcal{X}_i \to \mathbb{R}$ such that $H(x) = \sum_{i=1}^n v_i(x_i)$ and $H(x') = \sum_{i=1}^n v_i(x'_i)$ (because of complementary slackness applied to any element of the support of γ), while for any $y \in \mathcal{X}$, we simply have $H(y) \geqslant \sum_{i=1}^n v_i(y_i)$. By considering $y = \max\{x, x'\}$ and $y = \min\{x, x'\}$, we obtain: $H(x) + H(x') \leqslant H(\max\{x, x'\}) + H(\min\{x, x'\})$, which is in contradiction with the strict submodularity of H. Thus any optimal plan has to be equal to γ_{mon} for strictly submodular functions.

(4) When H is submodular, by adding $\varepsilon \sum_{i \neq j}(x_i - x_j)^2$, we obtain a strictly submodular function and by letting ε tend to zero, we obtain the desired result.

■

3.4 Relationship between convexity and submodularity

We can now prove our first formal result relating convexity and submodularity, that extends the similar result of [36] from set-functions to all continuous functions. Given the theory of multi-marginal optimal transport outlined above, the proof is straightforward and our result provides an alternative proof even for set-functions; note that while an implicit connection had been made for $n = 2$ through monotonicity properties of optimal assignment problems [50], the link we propose here is novel.

Theorem 1 (Convexity and submodularity) *Assume H is continuous and all $\mathcal{X}_i$'s are compact. The extension $h_{\text{cumulative}}$ defined in Eq. (3) is convex if and only if H is submodular.*

Proof We first assume that H is submodular. As shown in Proposition 4, optimal transport problems on subsets of real numbers with submodular costs are known to have closed-form solutions [8], which leads to the convexity of $h_{\text{cumulative}} = h_{\text{closure}}$.

We now assume that the function $h_{\text{cumulative}}$ is convex. Following the argument of [35] for the related problem of rearrangement inequalities and [36] for submodular set-functions, we consider

two arbitrary elements a and b in $\mathcal{X}$, and the Dirac measures δ_{a_i} and δ_{b_i}. We have, by convexity:

$$h_{\text{cumulative}}\big(\frac{1}{2}\delta_{a_1} + \frac{1}{2}\delta_{b_1}, \dots, \frac{1}{2}\delta_{a_n} + \frac{1}{2}\delta_{b_n}\big) \leqslant \frac{1}{2}h_{\text{cumulative}}(\delta_{a_1}, \dots, \delta_{a_n}) + \frac{1}{2}h_{\text{cumulative}}(\delta_{b_1}, \dots, \delta_{b_n}).$$

The right-hand side is equal to $\frac{1}{2}H(a) + \frac{1}{2}H(b)$, while we can compute the left-hand side by computing $F_{\frac{1}{2}\delta_{a_i}+\frac{1}{2}\delta_{b_i}}(x_i)$, which is equal to 0 if $x_i > \max\{a_i, b_i\}$, 1 if $x_i < \min\{a_i, b_i\}$ and $\frac{1}{2}$ if $x_i \in (\min\{a_i, b_i\}, \max\{a_i, b_i\})$. This implies that $F^{-1}_{\frac{1}{2}\delta_{a_i}+\frac{1}{2}\delta_{b_i}}(t)$ is equal to $\min\{a_i, b_i\}$ if $t > \frac{1}{2}$, and to $\max\{a_i, b_i\}$ if $t < \frac{1}{2}$. Thus the left-hand side of the inequality above is equal to $\frac{1}{2}H(\min\{a, b\}) + \frac{1}{2}H(\max\{a, b\})$. Hence, the submodularity. ■

From now on, we will assume that H is submodular and refer to h as its extension, which is both defined as a convex closure and through cumulative distribution functions. Note that a consequence of Theorem 1 is that for submodular functions, the closure of the sum is the sum of the closures, which is not true in general.

We now show that minimizing the extension is equivalent to minimizing the original function, implying that we may minimize any submodular function as a convex optimization problem:

Theorem 2 (Equivalent minimization problems) *Assume each $\mathcal{X}_i$ is compact, $i = 1, \dots, n$. If H is submodular, then*

$$\inf_{x \in \mathcal{X}} H(x) = \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu). \tag{8}$$

Moreover, μ is a minimizer if and only if $F^{-1}_{\mu}(t)$ is a minimizer of H for almost all $t \in [0, 1]$.

Proof Since h is the convex closure of H, the two infima have to be equal. Indeed, from Proposition 3, we have

$$\inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) = \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \inf_{\gamma \in \Pi(\mu)} \int_{\mathcal{X}} H(x) d\gamma(x),$$

which is the infimum over all probability measures on $\mathcal{X}$ (without any marginal constraints). It is thus achieved at a Dirac measure at any minimizer $x \in \mathcal{X}$ of $H(x)$.

Moreover, given a minimizer μ for the convex problem, we have:

$$\inf_{x \in \mathcal{X}} H(x) = h(\mu) = \int_0^1 H\big[F^{-1}_{\mu_1}(t), \dots, F^{-1}_{\mu_n}(t)\big] dt \geqslant \int_0^1 \inf_{x \in \mathcal{X}} H(x) dt = \inf_{x \in \mathcal{X}} H(x).$$

Thus, for almost all $t \in [0, 1]$, $\big(F^{-1}_{\mu_1}(t), \dots, F^{-1}_{\mu_n}(t)\big) \in \mathcal{X}$ is a minimizer of H. ■

From the proof above, we see that a minimizer of H may be obtained from a minimizer of h by inverting the cumulative distribution for any t. Many properties are known regarding minimizers of submodular functions [52], i.e., if x and y are minimizers of H, so are $\min\{x, y\}$ and $\max\{x, y\}$. As opposed to convex function where imposing strict convexity leads to a unique minimizer, imposing strict submodularity only imposes that the set of minimizers forms a chain.

In practice, given a (potentially approximate) minimizer μ and for discrete domains, we can look at the minimal value of $H\big[F^{-1}_{\mu_1}(t), \dots, F^{-1}_{\mu_n}(t)\big]$ over all $t \in [0, 1]$ by enumerating and sorting all possible values of $F_{\mu_i}(x_i)$ (see Section 4.1). Moreover, we still need a subgradient of h for our optimization algorithms. We will consider these in the simpler situation of finite sets in Section 4.

Dual of submodular function minimization. Algorithms for minimizing submodular set-functions rely on a dual problem which allows to provide optimality certificates. We obtain a similar dual problem in the general situation as we now show:

Proposition 5 (Dual of submodular function minimization) *The problem of minimizing $H(x)$ over $x \in \mathcal{X}$, has the following dual formulation*

$$\inf_{x \in \mathcal{X}} H(x) = \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) = \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^{n} \inf_{x_i \in \mathcal{X}_i} v_i(x_i). \tag{9}$$

Proof We have $\inf_{x \in \mathcal{X}} H(x) = \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu)$ from Theorem 2. Moreover, we may use convex duality [38] like in Prop. 2 to get:

$$\begin{aligned}
\inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^{n} \int_{\mathcal{X}_i} v_i(x_i) d\mu_i(x_i) \\
&= \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^{n} \inf_{\mu_i \in \mathcal{P}(\mathcal{X}_i)} \int_{\mathcal{X}_i} v_i(x_i) d\mu_i(x_i) = \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^{n} \inf_{x_i \in \mathcal{X}_i} v_i(x_i).
\end{aligned}$$

■

It allows to provide certificates of optimality when minimizing H. Note that like for set-functions, checking that a given function v is in $\mathcal{V}(H)$ is difficult, and therefore, most algorithms will rely on convex combinations of outputs of the greedy algorithm presented in Section 4.

3.5 Strongly-convex separable submodular function minimization

In the set-function situation, minimizing the Lovász extension plus a separable convex function has appeared useful in several scenarios [9, 1], in particular because a single of such problems leads to a solution to a continuously parameterized set of submodular minimization problems on $\mathcal{X}$. In our general formulation, the separable convex functions that can be combined are themselves defined through a submodular function and optimal transport.

We choose strongly convex separable costs of the form $\sum_{i=1}^{n} \varphi_i(\mu_i)$, with:

$$\varphi_i(\mu_i) = \int_{\mathcal{X}_i} a_i(x_i, F_{\mu_i}(x_i)) dx_i, \tag{10}$$

where for all $x_i \in \mathcal{X}_i$, $a_i(x_i, \cdot)$ is a differentiable λ-strongly convex function on $[0,1]$. This implies that the function φ_i, as a function of F_{μ_i} is λ-strongly convex (for the L_2-norm on cumulative distribution functions). A key property is that it may be expressed as a transport cost, as we now show.

Proposition 6 (Convex functions of cumulative distributions) *For $a_i : \mathcal{X}_i \times [0,1] \to \mathbb{R}$ convex and differentiable with respect to the second variable, the function $\varphi_i : \mathcal{P}(\mathcal{X}_i) \to \mathbb{R}$ defined in Eq. (10) is equal to*

$$\varphi_i(\mu_i) = \int_0^1 c_i(F_{\mu_i}^{-1}(t), 1-t) dt, \tag{11}$$

for $c_i(z_i, t_i) = \displaystyle\int_{\mathcal{X}_i} a_i(x_i, 0) dx_i + \int_{\mathcal{X}_i} \left(\int_{\mathcal{X}_i \cap (-\infty, z_i]} \frac{\partial a_i}{\partial t_i}(x_i, 1-t_i) dx_i \right)$ *is a submodular cost on $\mathcal{X}_i \times [0,1]$.*

Proof We have the following sequence of equalities:

$$\begin{aligned}
\varphi_i(\mu_i) &= \int_{\mathcal{X}_i} a_i(x_i, F_{\mu_i}(x_i))dx_i \\
&= \int_{\mathcal{X}_i} \left(a_i(x_i, 0) + \int_0^{F_{\mu_i}(x_i)} \frac{\partial a_i}{\partial t_i}(x_i, t_i)dt_i \right) dx_i \\
&= \int_{\mathcal{X}_i} a_i(x_i, 0)dx_i + \int_{\mathcal{X}_i} \left(\int_{\mathcal{X}_i \cap [x_i, +\infty)} \frac{\partial a_i}{\partial t_i}(x_i, F_{\mu_i}(y_i))d\mu_i(y_i) \right) dx_i \\
&\qquad \text{by the change of variable } t_i = F_{\mu_i}(y_i), \\
&= \int_{\mathcal{X}_i} a_i(x_i, 0)dx_i + \int_{\mathcal{X}_i} \left(\int_{\mathcal{X}_i \cap (-\infty, y_i]} \frac{\partial a_i}{\partial t_i}(x_i, F_{\mu_i}(y_i))dx_i \right) d\mu_i(y_i) \text{ by Fubini's theorem,} \\
&= \int_{\mathcal{X}_i} c_i(y_i, 1 - F_{\mu_i}(y_i))d\mu_i(y_i) \text{ by definition of } c_i, \\
&= \int_0^1 c_i(F_{\mu_i}^{-1}(t), 1 - t)dt \text{ by the change of variable } t_i = F_{\mu_i}(y_i).
\end{aligned}$$

Moreover, c_i is indeed a submodular function as, for any $z_i' > z_i$ in $\mathcal{X}_i$, we have $c_i(z_i', t_i) - c_i(z_i, t_i) = \int_{\mathcal{X}_i} \left(\int_{\mathcal{X}_i \cap (z_i, z_i']} \frac{\partial a_i}{\partial t_i}(x_i, 1 - t_i)dx_i \right)$, which is a decreasing function in t_i because a_i is convex. Thus c_i is submodular. It is also strictly submodular if $a_i(x_i, \cdot)$ is strongly convex for all $x_i \in \mathcal{X}_i$. ■

Because c_i is submodular, we have, following Prop. 4, a formulation of φ_i as an optimal transport problem between the measure μ_i on $\mathcal{X}_i$ and the uniform distribution $U[0, 1]$ on $[0, 1]$, as

$$\varphi_i(\mu_i) = \inf_{\gamma_i \in \mathcal{P}(\mathcal{X}_i \times [0,1]} \int_{\mathcal{X}_i \times [0,1]} c_i(x_i, t_i)d\gamma_i(x_i, t_i),$$

such that γ_i has marginals μ_i and the uniform distribution on $[0, 1]$. It is thus always convex (as as the minimum of a jointly convex problem in γ_i and μ_i)—note that it is already convex from the definition in Eq. (11) and the relationship between a_i and c_i in Prop. 6.

We now consider the following problem:

$$\inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) + \sum_{i=1}^n \varphi_i(\mu_i), \tag{12}$$

which is an optimization problem with h, with additional separable transport costs $\varphi_i(\mu_i)$. Given our assumption regarding the strong-convexity of the functions a_i above, this system has a unique solution. We may derive a dual problem using the representation from Eq. (5):

$$\begin{aligned}
\inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) + \sum_{i=1}^n \varphi_i(\mu_i) &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^n \left\{ \int_{\mathcal{X}_i} v_i(x_i)d\mu_i(x_i) + \varphi_i(\mu_i) \right\} \\
&= \sup_{v \in \mathcal{V}(H)} \sum_{i=1}^n \inf_{\mu_i \in \mathcal{P}(\mathcal{X}_i)} \left\{ \int_{\mathcal{X}_i} v_i(x_i)d\mu_i(x_i) + \varphi_i(\mu_i) \right\} \\
&= \sup_{v \in \mathcal{V}(H)} -\sum_{i=1}^n \varphi_i^*(-v_i),
\end{aligned} \tag{13}$$

where we use the Fenchel-dual notation $\varphi_i^*(v_i) = \sup_{\mu_i \in \mathcal{P}(\mathcal{X}_i)} \left\{ \int_{\mathcal{X}_i} v_i(x_i)d\mu_i(x_i) - \varphi_i(\mu_i) \right\}$. The equation above provides a dual problem to Eq. (12). We may also consider a family of submodular

minimization problems, parameterized by $t \in [0,1]$:

$$\min_{x \in \mathcal{X}} \; H(x) + \sum_{i=1}^{n} c_i(x_i, 1-t), \tag{14}$$

with their duals defined from Eq. (9). Note that the two dual problems are defined on the same set $\mathcal{V}(H)$. We can now prove the following theorem relating the two optimization problems.

Theorem 3 (Separable optimization - general case) *Assume H is continuous and submodular and all c_i, $i = 1, \dots, n$ are defined as in Prop. 6. Then:*

(a) If x and x' are minimizers of Eq. (14) for $t > t'$, then $x \leqslant x'$ (for the partial order on $\mathbb{R}^n$), i.e., the solutions of Eq. (14) are non-increasing in $t \in [0,1]$.

(b) Given a primal candidate $\mu \in \mathcal{P}^{\otimes}(\mathcal{X})$ and a dual candidate $v \in \mathcal{V}(H)$, then the duality gap for the problem in Eq. (12) is the integral from $t = 0$ to $t = 1$ of the gaps for the problem in Eq. (14) for the same dual candidate and the primal candidate $(F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)) \in \mathcal{X}$.

(c) Given the unique solution μ of Eq. (12), for all $t \in [0,1]$, $(F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)) \in \mathcal{X}$ is a solution of Eq. (14).

(d) Given any solutions $x^t \in \mathcal{X}$ for all problems in Eq. (14), we may define μ through $F_{\mu_i}(x_i) = \sup\big\{t \in [0,1], \; x_i^t \geqslant x_i\big\}$, for all i and $x_i \in \mathcal{X}_i$, so that μ is the optimal solution of Eq. (12).

Proof The first statement (a) is a direct and classical consequence of the submodularity of c_i [53, Section 2.8]. The main idea is that when we go from t to $t' < t$, then the function difference, i.e., $x_i \mapsto c_i(x_i, t') - c_i(x_i, t)$ is strictly increasing, hence the minimizer has to decrease.

For the second statement (b), we may first re-write the cost function in Eq. (12) as an integral in t, that is, for any $\mu \in \mathcal{P}^{\otimes}(\mathcal{X})$:

$$h(\mu) + \sum_{i=1}^{n} \varphi_i(\mu_i) \quad = \quad \int_0^1 \Big\{ H\big[F_{\mu_1}^{-1}(t), \dots, F_{\mu_n}^{-1}(t)\big] + \sum_{i=1}^{n} c_i\big[F_{\mu_i}^{-1}(t), 1-t\big] \Big\} dt.$$

The gap defined in Prop. 5 for a single submodular minimization problem in Eq. (14) is, for a primal candidate $x \in \mathcal{X}$ and a dual candidate $v \in \mathcal{V}(H)$:

$$H(x) + \sum_{i=1}^{n} c_i\big[x_i, F_{\mu_i}^{-1}(t)\big] - \sum_{i=1}^{n} \min_{y_i \in \mathcal{X}_i} \Big\{ v_i(y_i) + c_i\big[y_i, 1-t\big] \Big\},$$

and its integral with respect to $t \in [0,1]$ for $x_i = F_{\mu_i}^{-1}(t)$, for all $i \in \{1, \dots, n\}$, is equal to

$$h(\mu) + \sum_{i=1}^{n} \varphi_i(\mu_i) - \sum_{i=1}^{n} \int_0^1 \min_{y_i \in \mathcal{X}_i} \Big\{ v_i(y_i) + c_i\big[y_i, 1-t\big] \Big\} dt.$$

Finally, we have, by using the formulation of $\varphi_i(\mu_i)$ through optimal transport, an expression of the elements appearing in the dual problem in Eq. (13):

$$\begin{aligned}
-\varphi_i^*(-v_i) \quad &= \quad \inf_{\mu_i \in \mathcal{P}_i(\mathcal{X}_i)} \Big\{ \int_{\mathcal{X}_i} v_i(x_i) d\mu_i(x_i) + \varphi_i(\mu_i) \Big\} \\
&= \quad \inf_{\mu_i \in \mathcal{P}_i(\mathcal{X}_i)} \inf_{\gamma_i \in \Pi(\mu_i, U[0,1])} \int_{\mathcal{X}_i \times [0,1]} \big[v_i(x_i) + c_i(x_i, 1-t_i) \big] d\gamma_i(x_i, t_i) \\
&= \quad \int_0^1 \inf_{y_i \in \mathcal{X}_i} \big[v_i(y_i) + c_i(y_i, 1-t_i) \big] dt_i,
\end{aligned}$$

because we may for any t_i choose the conditional distribution of x_i given t_i equal to a Dirac at the minimizer y_i of $v_i(y_i) + c_i(y_i, 1 - t_i)$. This implies (b) and thus, by continuity, (c). The statement (d) is proved exactly like for set-functions [1, Prop. 8.3]. ■

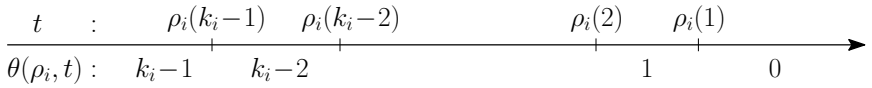
In Section 4.2, we will consider a formulation for finite sets that will exactly recover the set-function case, with the additional concept of base polytopes.

4 Discrete sets

In this section, we consider only finite sets for all $i \in \{1, \dots, n\}$, i.e., $\mathcal{X}_i = \{0, \dots, k_i - 1\}$. For this important subcase, we will extend many of the notions related to submodular set-functions, such as the base polytope and the greedy algorithm to compute its support function. This requires to extend the domain where we compute our extensions, from product of probability measures to products of non-increasing functions. Throughout this section, all measures are characterized by their probability mass functions, thus replacing integrals by sums.

This extension will be done using a specific representation of the measures $\mu_i \in \mathcal{P}(\mathcal{X}_i)$. Indeed, we may represent μ_i through its cumulative distribution function $\rho_i(x_i) = \mu_i(x_i) + \cdots + \mu_i(k_i - 1) = F_{\mu_i}(x_i)$, for $x_i \in \{1, \dots, k_i - 1\}$. Because the measure μ_i has unit total mass, $\rho_i(0)$ is always equal to 1 and can be left out. The only constraint on ρ_i is that is has non-increasing components and that all of them belong to $[0, 1]$. We denote by $[0, 1]_{\downarrow}^{k_i - 1}$ this set, which is in bijection with $\mathcal{P}(\{0, \dots, k_i - 1\})$. Therefore, ρ_i may be seen as a truncated cumulative distribution function equal to a truncation of F_{μ_i}; however, we will next extend its domain and remove the restriction of being between 0 and 1; hence the link with the cumulative distribution function F_{μ_i} is not direct anymore, hence a new notation ρ_i.

We are going to consider the set of non-increasing vectors $\mathbb{R}_{\downarrow}^{k_i - 1}$ (without the constraint that they are between 0 and 1). For any such ρ_i with $k_i - 1$ non-increasing components, the set of real numbers is divided into k_i parts, as shown below. Note that this is simply a rephrasing of the definition of $F_{\mu_i}^{-1}(t)$, as, when $\rho_i \in [0, 1]_{\downarrow}^{k_i - 1}$, we have $\theta(\rho_i, t) = F_{\mu_i}^{-1}(t)$.



This creates a map $\theta(\rho_i, \cdot) : \mathbb{R} \to \{0, \dots, k_i - 1\}$ such that $\theta(\rho_i, t) = k_i - 1$ for $t < \rho_i(k_i - 1)$, $\theta(\rho_i, t) = x_i$ if $t \in (\rho_i(x_i + 1), \rho_i(x_i))$, for $x_i \in \{1, \dots, k_i - 2\}$, and $\theta(\rho_i, t) = 0$ for $t > \rho_i(1)$. What happens at the boundary points is arbitrary and irrelevant.

For example, for $k_i = 2$ (and $\mathcal{X}_i = \{0, 1\}$), then we simply have $\rho_i \in \mathbb{R}$ and $\theta(\rho_i, t) = 1$ for $t < \rho_i$, and 0 if $t > \rho_i$. We can now give an expression of $h(\mu)$ as a function of $\rho \in \prod_{i=1}^{n} [0, 1]_{\downarrow}^{k_i - 1}$, which will be extended to all non-increasing vectors (not constrained to be between 0 and 1). This will then allow us to define extensions of base polytopes.

4.1 Extended extension on all products of non-increasing sequences

We first start by a simple lemma providing an expression of $h(\mu)$ as a function of ρ—note the similarity with the Lovász extension for set-functions [1, Prop. 3.1].

Lemma 1 (Extension as a function of ρ) *For $\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}$, define $h_\downarrow(\rho)$ as*

$$h_\downarrow(\rho) = \int_0^1 H(\theta(\rho_1,t),\dots,\theta(\rho_n,t))dt. \tag{15}$$

If $\mu \in \mathcal{P}^\otimes(\mathcal{X})$ and ρ are linked through $\rho_i(x_i) = F_{\mu_i}(x_i)$ for all i and $x_i \in \{1,\dots,k_i-1\}$, then, $h(\mu) = h_\downarrow(\rho)$.

Proof This is simply a re-writing of the definition of θ, as for almost all $t \in [0,1]$, and all $i \in \{1,\dots,n\}$, we have $\theta(\rho_i,t) = F_{\mu_i}^{-1}(t)$. ■

We can give an alternative formulation for $h_\downarrow(\rho)$ in Eq. (15) for $\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}$, as (with $\max\{\rho\}$ the maximum value of all $\rho_i(x_i)$, and similarly for $\min\{\rho\}$), using that $\theta(\rho_i,t) = \max \mathcal{X}_i = k_i - 1$ for $t < \min\{\rho\}$ and $\theta(\rho_i,t) = \min \mathcal{X}_i = 0$ for $t > \max\{\rho\}$:

$$\begin{aligned} h_\downarrow(\rho) &= \Big(\int_0^{\min\{\rho\}} + \int_{\min\{\rho\}}^{\max\{\rho\}} + \int_{\max\{\rho\}}^1 \Big) H(\theta(\rho_1,t),\dots,\theta(\rho_n,t))dt \\ &= \int_{\min\{\rho\}}^{\max\{\rho\}} H(\theta(\rho_1,t),\dots,\theta(\rho_n,t))dt + \min\{\rho\} H(k_1-1,\dots,k_n-1) + (1-\max\{\rho\})H(0). \end{aligned} \tag{16}$$

The expression in Eq. (16) may be used as a definition of $h_\downarrow(\rho)$ for all $\rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}$. With this definition, the function $\rho \mapsto h_\downarrow(\rho) - H(0)$ is piecewise linear. The following proposition shows that it is convex.

Proposition 7 (Convexity of extended extension) *Assume H is submodular. The function $h_\downarrow$ defined in Eq. (16) is convex on $\prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}$.*

Proof From the definition in Eq. (16), we have that, with $\rho + C$ being defined as adding the constant C to all components of all ρ_i's: $h_\downarrow(\rho + C) = h_\downarrow(\rho) + C\big[H(k_1-1,\dots,k_n-1) - H(0)\big]$. Moreover, $\rho \mapsto h_\downarrow(\rho) - H(0)$ is positively homogeneous. Thus, any set of ρ's in $\prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}$ may be transformed linearly to $\prod_{i=1}^n [0,1]_\downarrow^{k_i-1}$ by subtracting the global minimal value and normalizing by the global range of all $\rho's$. Since $h_\downarrow(\rho)$ coincides with the convex function $h(\mu)$ where μ_i is the probability distribution associated (in a linear way) to ρ_i, we obtain the desired result. ■

Greedy algorithm. We now provide a simple algorithm to compute $h_\downarrow(\rho)$ in Eq. (16), that extends the greedy algorithm for submodular set-functions. We thus now assume that we are given $\rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}$, and we compute $h_\downarrow(\rho)$ without sampling.

We first order all $r = \sum_{i=1}^n k_i - n$ values of $\rho_i(x_i)$ for all $x_i \in \{1,\dots,k_i-1\}$, in decreasing order, breaking ties randomly, except to ensure that all values for $\rho_i(x_i)$ for a given i are in the correct order (such ties may occur when one associated $\mu_i(x_i)$ is equal to zero). See Figure 6 for an example.

We assume that the s-th value is equal to $t(s)$ and corresponds to $\rho_{i(s)}(j(s))$. We have $t(1) = \max\{\rho\}$ and $t(r) = \min\{\rho\}$. For $s \in \{1,\dots,r-1\}$, we define the vector $y(s) \in \mathcal{X}$ so that $y(s)_i$ will be the value of $\theta(\rho_i,t)$ on the open interval (potentially empty) $(t(s+1),t(s))$; note that what happens at break points is still irrelevant. By convention, we define $y(0) = (0,\dots,0)$ and $y(r) = (k_1-1,\dots,k_n-1)$. We then go from $y(s-1)$ to $y(s)$ by increasing the $i(s)$-th component by one. Note that because we have assumed that $(\rho_i(x_i))_{x_i \in \mathcal{X}_i}$ are well-ordered, we always have $y(s)_{i(s)} = j(s)$ and $y(s) = y(s-1) + e_{i(s)}$, where $e_i \in \mathbb{R}^n$ is the i-the canonical basis vector.

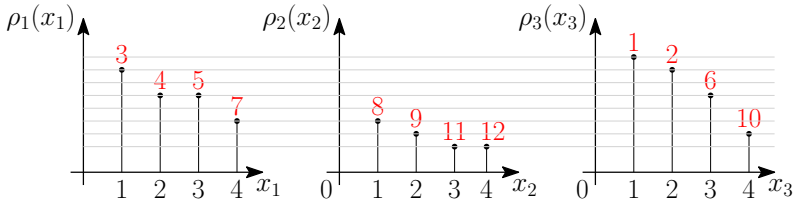


Figure 6: Examples of joint ordering (in red) of all values of $\rho_i(x_i)$ for the greedy algorithm ($n = 3$ and $k_i = 5$ for all i).

We thus get, by cutting the integral from Eq. (16) on $[\min\{\rho\}, \max\{\rho\}] = [t(r), t(1)]$ into pieces:

$$\begin{aligned} h_\downarrow(\rho) &= \min\{\rho\} H(k_1-1,\dots,k_n-1) + (1-\max\{\rho\})H(0) + \sum_{s=1}^{r-1} \int_{t(s+1)}^{t(s)} H\big[y(s)\big]\,dt \\ &= H(0) + t(r)H\big[y(r)\big] - t(1)H\big[y(0)\big] + \sum_{s=1}^{r-1} \big[t(s)-t(s+1)\big]\cdot H\big[y(s)\big] \\ &= H(0) + \sum_{s=1}^{r} t(s)\Big(H\big[y(s)\big] - H\big[y(s-1)\big]\Big). \end{aligned}$$

Since we have ordered all $r = \sum_{i=1}^n k_i - n$ values of $t(s) = \rho_{i(s)}(j(s))$, each $\rho_i(x_i)$ appears exactly once in the sum above. Therefore, $h_\downarrow(\rho)$ is of the form $h_\downarrow(\rho) = H(0) + \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} \rho_i(x_i) w_i(x_i)$, where $w_i(x_i)$ is a difference of two function values of H at arguments that differ from a single canonical basis vector. Moreover, we have $\sum_{i=1}^n \sum_{x_i=1}^{k_i-1} w_i(x_i) = H[y(r)] - H(0) = H(k_1-1,\dots,k_n-1) - H(0)$. We refer to this w as the output of the greedy algorithm (associated with a specific ordering of the values of ρ).

Note that for any ρ, several orderings may be chosen, all leading to the same value $h_\downarrow(\rho)$ but different values for $w_i(x_i)$. The number of such ordering is $n!$ for set-functions (i.e., $k_i = 2$ for all i) and is in general equal to the multinomial coefficient $\frac{\left(\sum_{i=1}^n (k_i-1)\right)!}{\prod_{i=1}^n (k_i-1)!}$. If all k_i, $i = 1,\dots,n$, are equal to k, then we get $\frac{(n(k-1))!}{(k_i-1)!^n}$, which grows very rapidly when n and k grow.

Note that for $k_i = 2$ for all $i \in \{1,\dots,n\}$, then we exactly obtain the greedy algorithm for submodular set-functions. The complexity is $O(r \log r)$ for sorting and r function evaluations, with $r = \sum_{i=1}^n k_i - n$.

4.2 Base polyhedron

We may now provide an extension of the base polyhedron which is usually defined for submodular set-functions. As opposed to set-functions, there are two natural polyhedra $\mathcal{W}(H)$ and $\mathcal{B}(H)$, one defined by linear inequalities, one defined as the convex hull of the outputs of the greedy algorithm [19, 1]. They are equal for set-functions, but not in the general case. The key difference is the monotonicity constraint on each ρ_i, which is only active when $k_i > 2$.

We consider the set $\mathcal{W}(H) \subset \prod_{i=1}^n \mathbb{R}^{k_i-1}$ defined through the inequalities:

$$\forall (x_1,\dots,x_n) \in \mathcal{X}, \qquad \sum_{i=1}^n \sum_{y_i=1}^{x_i} w_i(y_i) \leqslant H(x_1,\dots,x_n) - H(0)$$
$$\sum_{i=1}^n \sum_{y_i=1}^{k_i-1} w_i(y_i) = H(k_1-1,\dots,k_n-1) - H(0).$$

Note that when some $x_i = 0$, then the sum $\sum_{y_i=1}^{x_i} w_i(y_i)$ is equal to zero by convention (alternatively, we can use the convention that $w_i(0) = 0$). When $k_i = 2$ for all $i \in \{1,\dots,n\}$, we obtain exactly the usual base polyhedron (which happens to be bounded) [19]. The following proposition shows that it behaves in a similar way (except that it is not bounded).

Proposition 8 (Support function for $\mathcal{W}(H)$) *Assume H is submodular and $\mathcal{X}_i = \{0,\dots,k_i-1\}$ for $i \in \{1,\dots,n\}$. Then, for any $\rho \in \prod_{i=1}^n \mathbb{R}^{k_i-1}$,*

$$\sup_{w \in \mathcal{W}(H)} \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} w_i(x_i)\rho_i(x_i)$$

is equal to $+\infty$ if $\rho \notin \prod_{i=1}^n \mathbb{R}_{\downarrow}^{k_i-1}$, and to $h_\downarrow(\rho) - H(0)$ otherwise, with an optimal $w \in \mathcal{W}(H)$ obtained from the greedy algorithm for any compatible order.

Proof In this proof, we are going to use a reparameterization of w as follows; we define a vector $v_i \in \mathbb{R}^{\{1,\dots,k_i-1\}}$ as the cumulative sum of w_i, that is, such that $v_i(x_i) = \sum_{y_i=1}^{x_i} w_i(y_i)$ (this is a bijection from w_i to v_i). We then have the constraint that $\sum_{i=1}^n v_i(x_i) \leqslant H(x_1,\dots,x_n) - H(0)$, for all $x \in \mathcal{X}$, with the convention that $v_i(0) = 0$. The extra constraint is that $\sum_{i=1}^n v_i(k_i-1) = H(k_1-1,\dots,k_n-1) - H(0)$. Moreover, we have, with $\mu_i(x_i) = \rho_i(x_i) - \rho_i(x_i+1)$ for $x_i \in \{1,\dots,k_i-2\}$ and $\mu_i(k_i-1) = \rho_i(k_i-1)$, an expression for the linear function we aim to maximize, that is, by Abel's summation formula:

$$\begin{aligned}\sum_{x_i=1}^{k_i-1} w_i(x_i)\rho_i(x_i) &= \sum_{x_i=2}^{k_i-1} \big[v_i(x_i) - v_i(x_i-1)\big]\rho_i(x_i) + v_i(1)\rho_i(1) \\ &= \sum_{x_i=1}^{k_i-2} v_i(x_i)\big[\rho_i(x_i) - \rho_i(x_i+1)\big] + v_i(k_i-1)\rho_i(k_1-1) = \sum_{x_i=1}^{k_i-1} v_i(x_i)\mu_i(x_i).\end{aligned}$$

We first assume that each ρ_i is non-decreasing. We are going to follow the same proof than for set-functions, based on convex duality. First, given the piecewise-linearity of $h_\downarrow - H(0)$ (and hence the homogeneity), and the fact that $h_\downarrow(\rho + C) = h_\downarrow(\rho) + C\big[H(k_1-1,\dots,k_n-1) - H(0)\big]$, we only need to show the result for $\rho_i(x_i) \in [0,1]$ for all i, and $x_i \in \mathcal{X}_i$. Each vector ρ_i is then uniquely associated to a probability measure (with non-negative values because ρ_i is non-increasing) μ_i on $\mathcal{X}_i = \{0,\dots,k_i-1\}$, and, from Lemma 1, $h_\downarrow(\rho) = h(\mu)$.

Using the parameterization in terms of v and μ, we can now consider the Lagrangian, with dual values $\gamma \in \mathbb{R}_+^{\mathcal{X}}$:

$$\mathcal{L}(v,\gamma) = \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} v_i(x_i)\mu_i(x_i) + \sum_{x \in \mathcal{X}} \gamma(x)\bigg(H(x) - H(0) - \sum_{i=1}^n v_i(x_i)\delta(x_i > 0)\bigg).$$

By maximizing with respect to the primal variable v, the dual problem thus becomes the one of minimizing $\sum_{x \in \mathcal{X}} \gamma(x)\big[H(x) - H(0)\big]$ such that

$$\begin{aligned}\forall i \in \{1,\dots,n\},\ \forall x_i \in \{1,\dots,k_i-1\}, &\qquad \gamma_i(x_i) = \mu_i(x_i) \\ \forall x \in \mathcal{X} \backslash \{(k_1-1,\dots,k_n-1)\}, &\qquad \gamma(x) \geqslant 0,\end{aligned}$$

where γ_i is the marginal of γ on the i-th variable, that is, $\gamma_i(x_i) = \sum_{x_j \in \mathcal{X}_j, j \neq i} \gamma(x_1, \ldots, x_n)$.

We now exhibit primal/dual pairs with equal objective values. For the dual variable γ, we consider the solution of the usual optimal transport problem (which happens to satisfy extra constraints, that is, nonnegativity also for $x = 0$ and summing to one), and the dual value is exactly the extension $h(\mu) - H(0)$. For the primal variable, we consider the w parameterization (which is in bijection with v). From the greedy algorithm described in Section 4.1, the vector w satisfies the sum constraint $\sum_{i=1}^n \sum_{y_i=1}^{k_i - 1} w_i(y_i) = H(k_1 - 1, \ldots, k_n - 1) - H(0)$. We simply need to show that for all $x \in \mathcal{X}$, $\sum_{i=1}^n \sum_{z_i=1}^{x_i} w_i(z_i) \leqslant H(x) - H(0)$. We have, using the notation of the greedy algorithm from Section 4.1:

$$
\begin{aligned}
\sum_{i=1}^n \sum_{z_i=1}^{x_i} w_i(z_i) &= \sum_{s=1}^r \big(H[y(s)] - H[y(s-1)]\big)\delta(y(s)_{i(s)} \leqslant x_{i(s)}) \text{ by definition of } i(s) \text{ and } y(s), \\
&= \sum_{a=1}^n \sum_{s, i(s)=a} \big(H[y(s)] - H[y(s-1)]\big)\delta(y(s)_a \leqslant x_a) \text{ by splitting the values of } i(s), \\
&= \sum_{a=1}^n \sum_{s, i(s)=a} \big(H[y(s-1) + e_a] - H[y(s-1)]\big)\delta(y(s)_a \leqslant x_a) \\
&\quad \text{because we go from } y(s-1) \text{ to } y(s) \text{ by incrementing the component } i(s) = a, \\
&\leqslant \sum_{a=1}^n \sum_{s, i(s)=a} \big(H[\min\{y(s-1) + e_a, x\}] - H[\min\{y(s-1), x\}]\big)\delta(y(s-1)_a + 1 \leqslant x_a) \\
&\quad \text{by submodularity and because } y(s)_{i(s)} = y(s-1)_{i(s)} + 1, \\
&= \sum_{a=1}^n \sum_{s, i(s)=a} \big(H[\min\{y(s-1) + e_a, x\}] - H[\min\{y(s-1), x\}]\big) \\
&\quad \text{because the difference in values of } H \text{ is equal to zero for } y(s-1)_a + 1 > x_a, \\
&= \sum_{s=1}^r \big(H[\min\{y(s), x\}] - H[\min\{y(s-1), x\}]\big) \\
&= H[\min\{x, y(r)\}] - H[\min\{x, y(0)\}] = H(x) - H(0).
\end{aligned}
$$

Thus, w is feasible. By construction the primal value is equal to $h(\mu) - H(0)$. We thus have a primal/dual optimal pair and the result is proved for non-decreasing ρ_i's. This notably shows that the polyhedron $\mathcal{W}(H)$ is non-empty.

We can now show that if one ρ_i is not non-decreasing, then the supremum is equal to infinity. In such a case, then there exists $x_i \in \{1, \ldots, k_i - 2\}$ such that $\mu_i(x_i) < 0$. We may then let the corresponding $v_i(x_i)$ tend to $-\infty$. ■

Given the representation of $h_\downarrow(\rho) - H(0)$ as a maximum of linear functions (with an infinite value if some ρ_i does not have non-increasing components), the convex problem of minimizing $H(x)$, which is equivalent to minimizing $h_\downarrow(\rho) - H(0)$ with respect to $\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i - 1}$ has the following dual

problem:

$$\begin{aligned}\min_{\rho\in\prod_{i=1}^n[0,1]_\downarrow^{k_i-1}} h_\downarrow(\rho)-H(0) &= \min_{\rho\in\prod_{i=1}^n[0,1]^{k_i-1}}\max_{w\in\mathcal{W}(H)}\sum_{i=1}^n\sum_{x_i=1}^{k_i-1}w_i(x_i)\rho_i(x_i)\\ &= \max_{w\in\mathcal{W}(H)}\sum_{i=1}^n\min_{\rho_i\in[0,1]^{k_i-1}}\sum_{x_i=1}^{k_i-1}w_i(x_i)\rho_i(x_i)\\ &= \max_{w\in\mathcal{W}(H)}\sum_{i=1}^n\sum_{x_i=1}^{k_i-1}\min\{w_i(x_i),0\}. \qquad (17)\end{aligned}$$

While $\mathcal{W}(H)$ share some properties with the base polytope of a submodular set-function, it is not bounded in general and is not the convex hull of all outputs of the greedy algorithm from Section 4.1.

We now define the base polytope $\mathcal{B}(H)$ as the convex hull of all outputs of the greedy algorithm, when going over all allowed orderings of $\sum_{i=1}^n k_i - n$ elements of ρ that respect the ordering of the individual ρ_i's. In the submodular set-function case, we have $\mathcal{B}(H)=\mathcal{W}(H)$. However, in the general case we only have an inclusion.

Proposition 9 (Properties of the base polytope $\mathcal{B}(H)$) *Assume H is submodular. Then:*

(a) $\mathcal{B}(H)\subset\mathcal{W}(H)$.

(b) For any $\rho\in\prod_{i=1}^n\mathbb{R}_\downarrow^{k_i-1}$, $\max_{w\in\mathcal{B}(H)}\langle w,\rho\rangle=\max_{w\in\mathcal{W}(H)}\langle w,\rho\rangle=h_\downarrow(\rho)-H(0)$, *with a joint maximizer obtained as the output of the greedy algorithm with any particular order.*

(c) $\mathcal{W}(H)=\mathcal{B}(H)+\mathcal{K}$ *with*

$$\mathcal{K}=\left\{w\in\prod_{i=1}^n\mathbb{R}^{k_i-1},\ \forall i\in\{1,\dots,n\},\ \forall x_i\in\{1,\dots,k_i-2\},\ \sum_{y_i=1}^{x_i}w_i(y_i)\leqslant 0 \text{ and } \sum_{y_i=1}^{k_i-1}w_i(y_i)=0\right\}.$$

Proof In the statements above, $\langle w,\rho\rangle$ stands for $\sum_{i=1}^n\sum_{x_i=1}^{k_i-1}w_i(x_i)\rho_i(x_i)$. The statements (a) and (b) were shown in the proof of Prop. 8. Statement (c) is a simple consequence of the fact that the polar cone to $\prod_{i=1}^n\mathbb{R}_\downarrow^{k_i-1}$ is what needs to be added to $\mathcal{B}(H)$ to get to $\mathcal{W}(H)$. In other words, (c) is the equality of two convex sets and is thus equivalent to the equality of their support functions; and we have, for any $\rho_i\in\mathbb{R}^{k_i-1}$, using Abel's summation formula:

$$\langle w_i,\rho_i\rangle=\sum_{x_i=1}^{k_i-1}w_i(x_i)\rho_i(x_i)=\sum_{x_i=1}^{k_i-2}\Big(\sum_{y_i=1}^{x_i}w_i(y_i)\Big)\big[\rho_i(x_i)-\rho_i(x_i+1)\big]+\Big(\sum_{y_i=1}^{x_i-1}w_i(y_i)\Big)\rho_i(k_1-1),$$

from which we see that the supremum of $\langle w,\rho\rangle$ with respect to $\rho\in\prod_{i=1}^n\mathbb{R}_\downarrow^{k_i-1}$ is equal to zero if $w\in\mathcal{K}$ and $+\infty$ otherwise. By convex duality, this implies that the supremum of $\langle w,\rho\rangle$ with respect to $w\in\mathcal{K}$ is equal to zero if ρ has non-increasing components and zero otherwise. We thus have, for any $\rho\in\prod_{i=1}^n\mathbb{R}^{k_i-1}$, $\sup_{w\in\mathcal{B}(H)+\mathcal{K}}\langle w,\rho\rangle=h_\downarrow(\rho)$ if $\rho\in\prod_{i=1}^n\mathbb{R}_\downarrow^{k_i-1}$ and $+\infty$ otherwise. Given Prop. 8, this leads to the desired result. ■

The key difference between $\mathcal{W}(H)$ and $\mathcal{B}(H)$ is that the support function of $\mathcal{B}(H)$ does not include the constraint that the argument should be composed of non-increasing vectors. However, $\mathcal{B}(H)$ is a polytope (i.e., bounded as a convex hull of finitely many points), while $\mathcal{W}(H)$ is not.

To obtain dual problems, we may either choose $\mathcal{W}(H)$ or $\mathcal{B}(H)$ (and then take into account the monotonicity contraints explicitly). For our algorithms in Section 5, we will consider $\mathcal{B}(H)$, while in the section below, we will use $\mathcal{W}(H)$ for the analysis and comparison of several optimization problems. The dual using $\mathcal{W}(H)$ is given in Eq. (17). When using $\mathcal{B}(H)$, the dual problem of minimizing $h_\downarrow(\rho) - H(0)$ with respect to $\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}$ has the following form:

$$
\begin{aligned}
& \min_{\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}} h_\downarrow(\rho) - H(0) \\
= & \min_{\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}} \max_{w \in \mathcal{B}(H)} \sum_{x_i=1}^{k_i-1} w_i(x_i)\rho_i(x_i) = \max_{w \in \mathcal{B}(H)} \sum_{i=1}^n \min_{\rho_i \in [0,1]_\downarrow^{k_i-1}} \sum_{x_i=1}^{k_i-1} w_i(x_i)\rho_i(x_i) \\
= & \max_{w \in \mathcal{B}(H)} \sum_{i=1}^n \min_{\rho_i \in [0,1]_\downarrow^{k_i-1}} \sum_{x_i=1}^{k_i-2} \Big(\sum_{y_i=1}^{x_i} w_i(y_i) \Big) \big[\rho_i(x_i) - \rho_i(x_i+1) \big] + \Big(\sum_{y_i=1}^{x_i-1} w_i(y_i) \Big) \rho_i(k_1-1) \\
= & \max_{w \in \mathcal{B}(H)} \sum_{i=1}^n \min_{\mu_i \in \mathcal{P}(\mathcal{X}_i)} \sum_{x_i=0}^{k_i-1} \mu_i(x_i) \Big(\sum_{y_i=1}^{x_i-1} w_i(y_i) \Big) \\
= & \max_{w \in \mathcal{B}(H)} \sum_{i=1}^n \min_{x_i \in \{0,\dots,k_i-1\}} \sum_{y_i=1}^{x_i} w_i(y_i). \qquad (18)
\end{aligned}
$$

There is thus two dual problems for the submodular function minimization problem, Eq. (17) and Eq. (18). In practice, checking feasibility in the larger set $\mathcal{W}(H)$ is difficult, while checking feasibility in $\mathcal{B}(H)$ will be done by taking convex combinations of outputs of the greedy algorithm.

4.3 Strongly-convex separable submodular function minimization

In this section, we consider the separable optimization problem described in Section 3.5, now in the discrete case, where we will be able to use the polyhedra $\mathcal{W}(H)$ and $\mathcal{B}(H)$ defined above. We thus consider functions $a_{ix_i} : \rho_i(x_i) \mapsto a_{ix_i}\big[\rho_i(x_i)\big]$, for $i \in \{1,\dots,n\}$ and $x_i \in \{1,\dots,k_i-1\}$, which are convex on $\mathbb{R}$. For simplicity, we follow the same assumptions than in [1, Section 8], that is, they are all strictly convex, continuously differentiable and such that the images of their derivatives goes from $-\infty$ to $+\infty$. Their Fenchel conjugates $a^*_{ix_i}$ then have full domains and are differentiable.

We consider the pair of primal/dual problems:

$$
\begin{aligned}
& \min_\rho h_\downarrow(\rho) + \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} a_{ix_i}\big[\rho_i(x_i)\big] \text{ such that } \rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1} \qquad (19) \\
= & \min_{\rho \in \prod_{i=1}^n \mathbb{R}^{k_i-1}} \max_{w \in \mathcal{W}(H)} \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} \Big\{ w_i(x_i)\rho_i(x_i) + a_{ix_i}\big[\rho_i(x_i)\big] \Big\} \text{ by Prop. 8,} \\
= & \max_{w \in \mathcal{W}(H)} \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} -a^*_{ix_i}\big(-w_i(x_i)\big). \qquad (20)
\end{aligned}
$$

Note that we have used $\mathcal{W}(H)$ instead of $\mathcal{B}(H)$ to include automatically the constraints of monotonicity of ρ. If using $\mathcal{B}(H)$, we would get a formulation which is more adapted to optimization algorithms (see Section 5), but not for the theorem below.

The following theorem relates this optimization problem to the following family of submodular minimization problems, for $t \in \mathbb{R}$:

$$\min_{x \in \mathcal{X}} \; H(x) + \sum_{i=1}^{n} \sum_{y_i=1}^{x_i} a'_{iy_i}(t), \tag{21}$$

which is the minimization of the sum of H and a modular function. This is a discrete version of Theorem 3, which directly extends the corresponding results from submodular set-functions.

Theorem 4 (Separable optimization – discrete domains) *Assume H is submodular. Then:*

(a) If x and x' are minimizers of Eq. (21) for $t > t'$, then $x \leqslant x'$, i.e., the solutions of Eq. (21) are non-increasing in $t \in \mathbb{R}$.

(b) Given a primal candidate $\rho \in \prod_{i=1}^{n} \mathbb{R}_{\downarrow}^{k_i - 1}$ and a dual candidate $w \in \mathcal{W}(H)$ for Eq. (19), then the duality gap for the problem in Eq. (19) is the integral from $t = -\infty$ to $t = +\infty$ of the gaps for the problem in Eq. (21) for the same dual candidate and the primal candidate $\theta(\rho, t) \in \mathcal{X}$.

(c) Given the unique solution ρ of Eq. (19), for all $t \in \mathbb{R}$, $\theta(\rho, t) \in \mathcal{X}$ is a solution of Eq. (21).

(d) Given solutions $x^t \in \mathcal{X}$ for all problems in Eq. (21), we may define ρ through $\rho_i(x_i) = \sup \big\{ t \in \mathbb{R}, \; x_i^t \geqslant x_i \big\} = \inf \big\{ t \in \mathbb{R}, \; x_i^t \leqslant x_i \big\}$, for all i and x_i, so that ρ is the optimal solution of Eq. (19).

Proof As for Theorem 3, the first statement (a) is a consequence of submodularity [52]. The main idea is that when we go from t to $t' < t$, then the function difference is increasing by convexity, hence the minimizer has to decrease.

For the second statement (b), we provide expressions for all elements of the gap, following the proof in [1, Prop. 8.5]. For $M > 0$ large enough, we have from Eq. (16):

$$h_{\downarrow}(\rho) - H(0) = \int_{-M}^{+M} H\big[\theta(\rho, t)\big] dt - MH(k_1 - 1, \dots, k_n - 1) - MH(0).$$

Moreover, the integral of the i-th "non-H-dependent" part of the primal objective for the submodular minimization problem in Eq. (21) is equal to, for $i \in \{1, \dots, n\}$ and the primal candidate $x_i = \theta(\rho_i, t)$:

$$\begin{aligned}
& \int_{-M}^{M} \sum_{y_i=1}^{x_i} a'_{iy_i}(t) dt \\
= \; & \int_{-M}^{+M} \Big\{ \sum_{x_i=1}^{k_i - 1} \delta(\theta(\rho_i, t) = x_i) \sum_{y_i=1}^{x_i} a'_{iy_i}(t) \Big\} dt \\
= \; & \int_{-M}^{\rho_i(k_i - 1)} \sum_{y_i=1}^{k_i - 1} a'_{iy_i}(t) dt + \sum_{s_i=2}^{k_i - 1} \int_{\rho_i(s_i)}^{\rho_i(s_i - 1)} \sum_{y_i=1}^{s_i - 1} a'_{iy_i}(t) dt \text{ by definition of } \theta, \\
= \; & \sum_{y_i=1}^{k_i - 1} \big\{ a_{iy_i}(\rho_i(k-1)) - a_{iy_i}(-M) \big\} + \sum_{s_i=2}^{k_i - 1} \sum_{y_i=1}^{s_i - 1} \Big\{ a_{iy_i}(\rho_i(s_i - 1)) - a_{iy_i}(\rho_i(s_i)) \Big\} \\
= \; & - \sum_{y_i=1}^{k_i - 1} a_{iy_i}(-M) + \sum_{x_i=1}^{k_i - 1} a_{ix_i}(\rho_i(x_i)).
\end{aligned}$$

We may now compute for any $i \in \{1,\dots,n\}$ and $x_i \in \{0,\dots,k_i-1\}$, the necessary pieces for the integral of the dual objective values from Eq. (17):

$$\begin{aligned}
& \int_{-M}^{M} \min\{w_i(x_i) + a'_{ix_i}(t), 0\} dt \\
= & \int_{-M}^{(a^*_{ix_i})'(-w_i(x_i))} \big(w_i(x_i) + a'_{ix_i}(t)\big) dt \text{ because } w_i(x_i) + a'_{ix_i}(t) \leqslant 0 \Leftrightarrow t \leqslant (a^*_{ix_i})'(-w_i(x_i)), \\
= & \; a_{ix_i}\big((a^*_{ix_i})'(-w_i(x_i))\big) - a_{ix_i}(-M) + w_i(x_i)\big[M + (a^*_{ix_i})'(-w_i(x_i))\big] \\
= & \; (\psi^*_{iy_i})'(-w_i(x_i))(-w_i(x_i)) - \psi^*_{iy_i}(-w_i(x_i)) - a_{iy_i}(-M) + w_i(x_i)\big[M + (a^*_{ix_i})'(-w_i(x_i))\big] \\
= & \; -a^*_{ix_i}(-w_i(x_i)) - a_{ix_i}(-M) + w_i(x_i) M.
\end{aligned}$$

By putting all pieces together, we obtain the desired result, that is,

$$h_\downarrow(\rho) - H(0) + \sum_{x_i=1}^{k_i-1} a_{ix_i}(\rho_i(x_i)) + \sum_{i=1}^{n} \sum_{x_i=1}^{k_i-1} a^*_{ix_i}(-w_i(x_i))$$

is exactly the integral from $-M$ to M of

$$H\big[\theta(\rho,t)\big] + \sum_{i=1}^{n} \sum_{x_i=1}^{k_i-1} \delta(\theta(\rho_i,t) = x_i) \sum_{y_i=1}^{x_i} a'_{iy_i}(t) - H(0) - \sum_{i=1}^{n} \min\{w_i(x_i) + a'_{ix_i}(t), 0\}.$$

The proofs of statements (c) and (d) follow the same argument as for Theorem 3. ■

In Section 5.2, we will use this result for the functions $\varphi_{ix_i}(t) = \frac{1}{2}t^2$, and from thresholding at $t = 0$ we obtain a solution for the minimization of H, thus directly extending the submodular set-function situation [19].

4.4 Relationship with submodular set-functions and ring families

For finite sets, it is known that one may reformulate the submodular minimization problem in terms of a submodular function minimization problem with added constraints [48]. Given the submodular function H, defined on $\mathcal{X} = \prod_{i=1}^{n}\{0,\dots,k_i-1\}$ we consider a submodular set-function defined on a *ring family* of the set

$$V = \big\{(i,x_i),\ i \in \{1,\dots,n\},\ x_i \in \{1,\dots,k_i-1\}\big\} \subset \{1,\dots,n\} \times \mathbb{R},$$

that is, a family of subsets of V which is invariant by intersection and union. In our context, a member of the ring family is such that if (i,x_i) is in the set, then all (i,y_i), for $y_i \leqslant x_i$ are in the set as well. The cardinality of V is $\sum_{i=1}^{n} k_i - n$; moreover, any member of the ring family is characterized for each i by the largest $x_i \geqslant 1$ such that $(i,x_i) \in V$ (if no (i,x_i) is in V, then we take $x_i = 0$ by convention). This creates a bijection from the ring family to $\mathcal{X}$.

We can identify subsets of V as elements of $\prod_{i=1}^{n}\{0,1\}^{k_i-1}$. It turns out that elements of the ring family as the non-increasing vectors, i.e., $z \in \prod_{i=1}^{n}\{0,1\}^{k_i-1}_{\downarrow}$. Any such element z is also associated uniquely to an element $x \in \mathcal{X} = \prod_{i=1}^{n}\{0,\dots,k_i-1\}$, by taking x_i as the largest $y_i \in \{1,\dots,k_i-1\}$ such that $z_i(y_i) = 1$, and with value zero, if $z_i = 0$. We can thus define a function $H_{\text{ring}}(z)$ equal to $H(x)$ for this uniquely associated x. Then for any two $z, z' \in \prod_{i=1}^{n}\{0,1\}^{k_i-1}_{\downarrow}$ of the ring family with corresponding $x, x' \in \mathcal{X}$, $\min\{z,z'\}$ and $\max\{z,z'\}$ correspond to $\min\{x,x'\}$ and $\max\{x,x'\}$, which implies that $H_{\text{ring}}(z) + H_{\text{ring}}(z') \geqslant H_{\text{ring}}(\max\{z,z'\}) + H_{\text{ring}}(\min\{z,z'\})$, i.e., H_{ring} is submodular

on the ring family, and minimizing $H(x)$ for $x \in \mathcal{X}$, is equivalent to minimizing H_{ring} on the ring family.

There is then a classical reduction to the minimization of a submodular function on $\prod_{i=1}^n \{0,1\}^{k_i-1}$ (without monotonicity constraints) [48, 22, 47], which is a regular submodular set-function minimization problem on a set of size $\sum_{i=1}^n k_i - n$ which we now describe (adapted from [49, Section 49.3]). For a certain $B_i > 0$, $i = 1, \dots, n$, define the function

$$H^{\text{ext}}_{\text{ring}}(z) = H_{\text{ring}}(z^{\downarrow}) + \sum_{i=1}^n B_i \|z_i^{\downarrow} - z_i\|_1 = H_{\text{ring}}(z^{\downarrow}) + \sum_{i=1}^n B_i 1_{k_i-1}^{\top}\big(z_i^{\downarrow} - z_i\big),$$

where $1_{k_i-1} \in \mathbb{R}^{k_i-1}$ is the vector of all ones, and where for $z \in \prod_{i=1}^n \{0,1\}^{k_i-1}$, $z^{\downarrow}$ denotes the smallest element of the ring family containing z; in other words $z_i^{\downarrow}(x_i) = 1$ for all x_i such that there exists $y_i \geqslant x_i$ with $z_i(y_i) = 1$, and zero otherwise. We choose $B_i > 0$ such that for all $x \leqslant y$, $|H(y) - H(x)| \leqslant \sum_{i=1}^n B_i \|x_i - y_i\|_1$, so that $H^{\text{ext}}_{\text{ring}}$ is submodular. Indeed, for any $z, z' \in \prod_{i=1}^n \{0,1\}^{k_i-1}$, we have:

$$\begin{aligned}
& H^{\text{ext}}_{\text{ring}}(z) + H^{\text{ext}}_{\text{ring}}(z') \\
= \;& H_{\text{ring}}(z^{\downarrow}) + H_{\text{ring}}((z')^{\downarrow}) + \sum_{i=1}^n B_i \|z_i^{\downarrow} - z_i\|_1 + \sum_{i=1}^n \|(z')_i^{\downarrow} - z'_i\|_1 \\
\geqslant \;& H_{\text{ring}}(\max\{z^{\downarrow}, (z')^{\downarrow}\}) + H_{\text{ring}}(\min\{z^{\downarrow}, (z')^{\downarrow}\}) + \sum_{i=1}^n B_i \|z_i^{\downarrow} - z_i\|_1 + \sum_{i=1}^n \|(z')_i^{\downarrow} - z'_i\|_1 \\
& \qquad \text{by submodularity of } H_{\text{ring}}, \\
= \;& H_{\text{ring}}(\max\{z^{\downarrow}, (z')^{\downarrow}\}) + H_{\text{ring}}(\min\{z^{\downarrow}, (z')^{\downarrow}\}) \\
& \quad + \sum_{i=1}^n B_i \|\max\{z_i^{\downarrow}, (z')_i^{\downarrow}\} - \max\{z_i, z'_i\}\|_1 + \sum_{i=1}^n B_i \|\min\{z_i^{\downarrow}, (z')_i^{\downarrow}\} - \min\{z_i, z'_i\}\|_1 \\
& \qquad \text{because } z^{\downarrow} \geqslant z, \\
\geqslant \;& H_{\text{ring}}(\max\{z, z'\}^{\downarrow}) + H_{\text{ring}}(\min\{z, z'\}^{\downarrow}) \\
& \quad + \sum_{i=1}^n \|\max\{z, z'\}_i^{\downarrow} - \max\{z, z'\}_i\|_1 + \sum_{i=1}^n B_i \|\min\{z, z'\}_i^{\downarrow} - \min\{z, z'\}_i\|_1 \\
& \qquad \text{because of our choice for } B_i, \\
= \;& H^{\text{ext}}_{\text{ring}}(\min\{z, z'\}) + H^{\text{ext}}_{\text{ring}}(\max\{z, z'\}),
\end{aligned}$$

which shows submodularity. Moreover, for any strictly positive B_i's, any minimizer of $H^{\text{ext}}_{\text{ring}}$ belongs to the ring family, and hence leads to a minimizer of H. Therefore, we have reduced the minimization of H to the minimization of a submodular set-function.

The Lovász extension of $H^{\text{ext}}_{\text{ring}}$ happens to be equal to, for $\nu \in \prod_{i=1}^n [0,1]^{k_i-1}$,

$$h_{\downarrow}(\nu^{\downarrow}) + \sum_{i=1}^n B_i \sum_{x_i=1}^{k_i-2} \big(\nu_i(x_{i+1}) - \nu_i(x_i)\big)_+,$$

where $\nu_i^{\downarrow}$ is the smallest non-increasing vector greater or equal to ν_i. When the B_i's tend to $+\infty$, we recover our convex relaxation from a different point of view. Note that in practice, unless we are in special situations like min-cut/max-flow problems [22] (where we can take $B_i = +\infty$), this strategy adds extra (often unknown) parameters B_i and does not lead to our new interpretations, convex relaxations, and duality certificates, in particular for continuous sets $\mathcal{X}_i$. In particular, when using

preconditioned subgradient descent as described in Section 5.1 to solve the submodular set-function minimization problem, for cases where all B_i are equal and all k_i are equal, we get a complexity bound of $\frac{1}{\sqrt{t}} nk^2 B$ instead of $\frac{1}{\sqrt{t}} nkB$, hence with a worse scaling in the number k of elements of the finite sets, which is due larger Lipschitz-continuity constants.

5 Optimization for discrete sets

In this section, we assume that we are given a submodular function H on $\prod_{i=1}^n \{0, \dots, k_i - 1\}$, which we can query through function values (and usually through the greedy algorithm defined in Section 4.1). We assume that for all $x \in \mathcal{X}$, when $x_i < k_i - 1$, $|H(x + e_i) - H(x)|$ is bounded by B_i (i.e., Lipschitz-continuity).

We present algorithms to minimize H. These can be used in continuous domains by discretizing each $\mathcal{X}_i$. The key is that the overall complexity remains polynomial in n, and the dependence in k_i is weak enough to easily reach high precisions. See the complexity for optimizing functions in continuous domains in the next section.

5.1 Optimizing on measures

We have the first equivalent formulations from Section 4:

$$\min_{\mu \in \prod_{i=1}^n \mathcal{P}(\{0,\dots,k_i-1\})} h(\mu) \quad \Leftrightarrow \quad \min_{\rho \in \prod_{i=1}^n [0,1]_{\downarrow}^{k_i-1}} h_{\downarrow}(\rho).$$

Once we get an approximately optimal solution ρ, we compute $\theta(\rho, t)$ for all $t \in [0, 1]$, and select the one with minimal value (this is a by-product of the greedy algorithm, i.e., with the notation of Section 4.1, $\min_s H\big[y(s)\big]$). If we have a dual candidate $w \in \mathcal{B}(H)$, we can use Eq. (18) to obtain a certificate of optimality.

We consider the *projected subgradient method* on ρ. We may compute a subgradient of $h_{\downarrow}$ in $\mathcal{B}(H)$ by the greedy algorithm, and then use n isotonic regressions to perform the n independent orthogonal projection of $\rho - \gamma h'_{\downarrow}(\rho)$, using the pool-adjacent-violator algorithm [4]. Each iteration has thus complexity $\Big(\sum_{i=1}^n k_i\Big) \log \Big(\sum_{i=1}^n k_i\Big)$, which corresponds to the greedy algorithm (which is here the bottleneck).

In terms of convergence rates, each element $w_i(x_i)$ of a subgradient is in $[-B_i, B_i]$. Moreover, for any two elements ρ, ρ' of $\prod_{i=1}^n [0,1]_{\downarrow}^{k_i-1}$, we have $\|\rho_i - \rho'_i\|_\infty \leqslant 1$. Thus, in ℓ_2-norm, the diameter of the optimization domain is less than $\sqrt{\sum_{i=1}^n k_i}$, and the norm of subgradients less than $\sqrt{\sum_{i=1}^n k_i B_i^2}$. Thus, the distance to optimum (measured in function values) after t steps is less than [51]

$$\frac{1}{\sqrt{t}} \sqrt{\Big(\sum_{i=1}^n k_i\Big)\Big(\sum_{i=1}^n k_i B_i^2\Big)}.$$

Note that by using diagonal preconditioning (see, e.g., [26]), we can replace the bound above by $\frac{1}{\sqrt{t}} \sum_{i=1}^n k_i B_i$. In both cases, if all B_i and k_i are equal, we get a complexity of $O(nk \log(nk))$ per iteration and a convergence rate of $O(nkB/\sqrt{t})$. In practice, we choose the Polyak rule for the step-size, since we have candidates for the dual problem; that is, we use $\gamma = \frac{h_{\downarrow}(\rho) - D}{\|w\|_2^2}$, where D is the best dual value so far, which we can obtain with dual candidates which are the averages of all elements of $\mathcal{B}(H)$ seen so far [40].

Once we have a primal candidate $\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}$ and a dual candidate $w \in \mathcal{B}(H)$, we may compute the minimum value of H along the greedy algorithm, as a primal value, and

$$\min_{\rho \in \prod_{i=1}^n [0,1]_\downarrow^{k_i-1}} \langle w, \rho \rangle = \sum_{i=1}^n \min_{x_i \in \{0,\dots,k_i-1\}} \sum_{y_i=1}^{x_i} w_i(y_i),$$

as the dual value. Note that as for submodular set-functions, the only simple way to certify that $w \in \mathcal{B}(H)$ is to make it a convex combination of outputs of the greedy algorithm.

Finally, note that the projected subgradient descent directly applies when a noisy oracle for H is available, using the usual arguments for stochastic extensions [51], with a similar convergence rate.

Minimizing submodular Lipschitz-continuous functions. We consider a submodular function H defined on $[0,B]^n$ which is G-Lipschitz-continuous with respect to the ℓ_∞-norm, that is $|H(x) - H(x')| \leqslant G\|x - x\|_\infty$. By discretizing $[0,B]$ into k values $\frac{iB}{k}$ for $i \in \{0,\dots,k-1\}$ and minimizing the corresponding submodular discrete function, then we will make an error of at most $\frac{GB}{k}$ on top the bound above, which is equal to $\frac{1}{\sqrt{t}}\sqrt{(nk)(nkB^2G^2/k^2)} = \frac{BGn}{\sqrt{t}}$ after t iterations of cost $O(nk\log(nk))$. Thus to reach a global optimization suboptimality of ε, we may take $\frac{GB}{k} = \frac{\varepsilon}{2}$ and $\frac{BGn}{\sqrt{t}} = \frac{\varepsilon}{2}$, that is, $k = \frac{2GB}{\varepsilon}$ and $t = \big(\frac{2GBn}{\varepsilon}\big)^2$, leading to an overall complexity of $O\big(\big(\frac{2GBn}{\varepsilon}\big)^3 \log\big(\frac{2GBn}{\varepsilon}\big)\big)$.

5.2 Smooth extension and Frank-Wolfe techniques

We consider the minimization of $h_\downarrow(\rho) + \frac{1}{2}\|\rho\|^2$, with $\|\rho\|^2 = \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} |\rho_i(x_i)|^2$. Given an approximate ρ, we compute $\theta(\rho,t)$ for all $t \in \mathbb{R}$ (which can be done by ordering all values of $\rho_i(x_i)$, like in the greedy algorithm).

We have by convex duality:

$$\begin{aligned}\min_{\rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}} h_\downarrow(\rho) + \frac{1}{2}\|\rho\|^2 &= \min_{\rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}} \max_{w \in \mathcal{B}(H)} \langle \rho, w \rangle + \frac{1}{2}\|\rho\|^2 \\ &= \max_{w \in \mathcal{B}(H)} \left\{ \min_{\rho \in \prod_{i=1}^n \mathbb{R}_\downarrow^{k_i-1}} \langle \rho, w \rangle + \frac{1}{2}\|\rho\|^2 \right\}.\end{aligned}$$

We thus need to maximize with respect to w in a compact set of diameter less than $2\sqrt{\sum_{i=1}^n k_i B_i^2}$, a 1-smooth function. Thus, we have that after t steps, the distance to optimum is less than $\frac{2}{t}\sum_{i=1}^n k_i B_i^2$ using Frank-Wolfe techniques, with either line-search of fixed step-sizes [17, 24, 2]. Note that primal candidates may also be obtained with a similar convergence rate; the running-time complexity is the same as for subgradient descent in Section 5.1.

Note that when using a form of line-search, we can use warm-restarts, which could be useful when solving a sequence of related problems, e.g., when discretizing a continuous problem. Finally, more refined versions of Frank-Wolfe algorithms can be used (see [32] and references therein).

Given an approximate minimizer ρ, we can compute all values of $x_i = \theta(\rho_i, t)$ for all $t \in \mathbb{R}$ from the greedy algorithm to get an approximate minimizer of the original submodular function. Given an approximate solution with error ε for the strongly-convex problem, using the exact same argument than for set-functions [1, Prop. 10.5], we get a bound of $2\sqrt{\varepsilon \sum_{i=1}^n k_i}$ for the submodular function minimization problem. Combined with the $O(1/t)$ convergence rate described above for the Frank-Wolfe algorithm, there is no gain in the complexity bounds compared to Section 5.1 (note that a

similar diagonal pre-conditioning can be used); however the empirical performance is significantly better (see Section 6).

6 Experiments

In this section, we consider a basic experiment to illustrate our results. We consider the function defined on $[-1,1]^n$,

$$H(x) = \frac{1}{2}\sum_{i=1}^{n}(x_i - z_i)^2 + \lambda \sum_{i=1}^{n} |x_i|^\alpha + \mu \sum_{i=1}^{n-1}(x_i - x_{i+1})^2.$$

This corresponds to denoising a one-dimensional signal z with the contraint that z is smooth and sparse. See an illustration in Figure 7. The smoothness prior is obtained from the quadratic form (which is submodular and convex), while the sparse prior is only submodular, and not convex. Thus, this is not a convex optimization problem when $\alpha < 1$; however, it can be solved globally to arbitrary precision for any $\alpha > 0$ using the algorithms presented in Section 5. Note that the use of a non-convex sparse prior (i.e., α significantly less than one, e.g., $1/8$ in our experiments) leads to fewer biasing artefacts than the usual ℓ_1-norm [14].

In Figure 8, we show certified duality gaps for the two algorithms from Section 5 on a discretization with 50 grid elements for each of the $n = 50$ variables. We can see that the Frank-Wolfe-based optimization performs better than projected subgradient descent, in particular the "pairwise-Frank-Wolfe" method of [32].

Finally, in Figure 9, we show the estimated values for the vectors ρ, showing that the solution is almost a threshold function for the non-smooth dual problem used in Section 5.1 (left), while it provides more information in the smooth dual case used in Section 5.2 (right), as it solved a series of submodular function minimization problems.

7 Discussion

In this paper, we have shown that a large family of non-convex problems can be solved by a well-defined convex relaxation and efficient algorithms based on simple oracles. This is based on replacing each of the variables x_i by a probability measure and optimizing over the cumulative distributions of these measures. Our algorithms apply to all submodular functions defined on products of subsets of $\mathbb{R}^n$, and hence include continuous domains as well as finite domains. This thus defines a new class of functions that can be minimized in polynomial time.

Several extensions are worth considering:

- **Relationship with convexity for continuous domains**: for functions defined on a product of sub-intervals, there are two notions of "simple" functions, convex and submodular. These two notions are usually disjoint (see Section 2.2 for examples). We study two interesting relationships: (a) the convex closure we define in Section 3.2 for convex functions, and (b) the minimization of the sum of a submodular function and a convex function.

  Given a convex function G defined on a product $\mathcal{X}$ of intervals, its convex closure defined on $\mathcal{P}^{\otimes}(\mathcal{X})$ is defined as:

$$g_{\text{closure}}(\mu) = \inf_{\gamma \in \Pi(\mu)} \int_{\mathcal{X}} G(x) d\gamma(x).$$

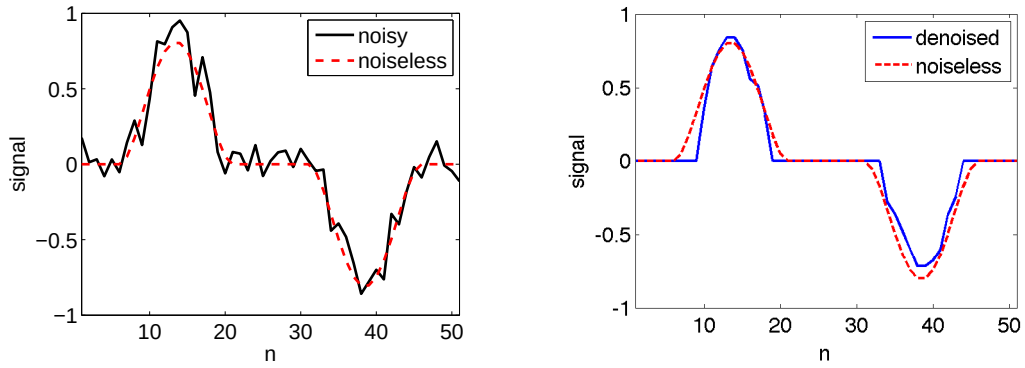


Figure 7: One-dimensional signals: noisy input obtained by adding Gaussian noise to a noiseless signal (left); denoised signal (right).

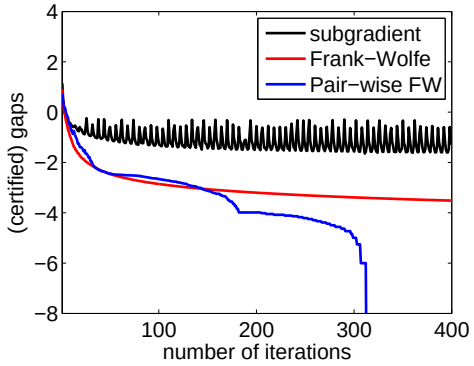


Figure 8: Certified duality gaps: noisy input obtained by adding Gaussian noise to a noiseless signal (left); denoised signal (right).

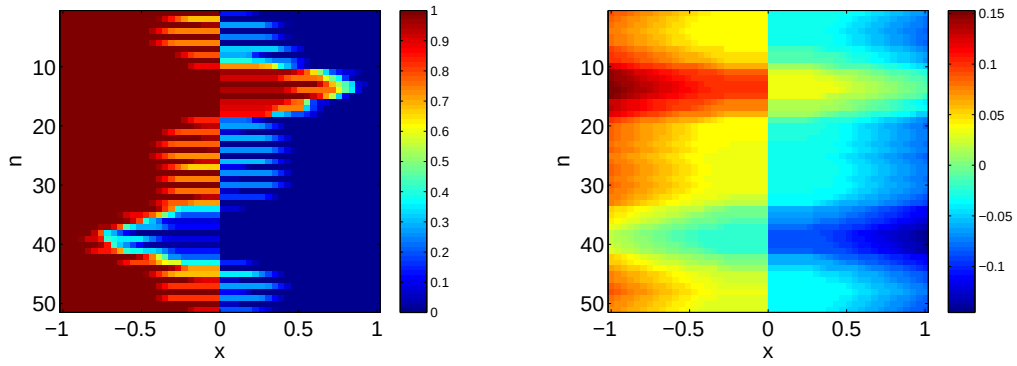


Figure 9: Optimal values of ρ: non-smooth dual problem (left); smooth dual problem (right).

Given Jensen's inequality, it satisfies:

$$g_{\text{closure}}(\mu) \geqslant \inf_{\gamma \in \Pi(\mu)} G\Big(\int_{\mathcal{X}} x d\gamma(x) \Big) = G\Big(\int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) \Big),$$

that is one may obtain a lower bound on the convex closure (note that if perform the convex closure directly on $\mathcal{X}$ and not on $\mathcal{P}\otimes(\mathcal{X})$, we would obtain G). This lower bound on the convex closure may be tight in some situations, for example, when minimizing a convex function G from $\mathcal{X}$ to $\mathbb{R}$. Indeed, we have:

$$\begin{aligned} \inf_{x \in \mathcal{X}} G(x) &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} G_{\text{closure}}(\mu) \\ &\geqslant \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} G\Big(\int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) \Big) = \inf_{x \in \mathcal{X}} G(x), \end{aligned}$$

and thus the convex closure and its relaxation allow an exact minimization, which allows to extend the result of [56, 45] from graphical model-based functions to all convex functions.

Moreover, it is worth considering the two notions of submodularity and convexity *simultaneously*, as many common objective functions are the sums of a convex function and a submodular function (e.g., the negative log-likelihood of a Gaussian vector, where the covariance matrix is parameterized as a linear combination of positive definite matrices [39]). We can thus consider the minimization of $H(x) + G(x)$, where $H : \mathcal{X} \to \mathbb{R}$ is submodular and $G : \mathcal{X} \to \mathbb{R}$ is convex. From the same reasoning as above, we have a natural convex relaxation on our set of measures:

$$\min_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) + G\Big(\int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) \Big). \tag{22}$$

Another relaxation is to replace H by its convex envelope on $\mathcal{X}$ (which is computable, as one can miminizer H plus linear functions), which we can get by bi-conjugation. We have for $z \in \mathbb{R}^n$:

$$H^*(z) = \sup_{x \in \mathcal{X}} x^\top z - H(x) = \sup_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} z^\top \int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) - h(\mu).$$

This implies that for any $x \in \mathcal{X}$:

$$\begin{aligned} H^{**}(x) &= \sup_{z \in \mathbb{R}^n} \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} z^\top x - z^\top \int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) + h(\mu) \\ &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} \sup_{z \in \mathbb{R}^n} z^\top \Big(x - \int_{\mathcal{X}} x \prod_{i=1}^{n} d\mu_i(x_i) \Big) + h(\mu) \\ &= \inf_{\mu \in \mathcal{P}^{\otimes}(\mathcal{X})} h(\mu) \text{ such that } \int_{\mathcal{X}} y \prod_{i=1}^{n} d\mu_i(y_i) = x. \end{aligned}$$

This implies that the relaxation in Eq. (22) is equivalent to the minimization of $H^{**}(x)+G(x)$, which is another natural convex relaxation. The main added benefit of submodularity is that the convex envelope can be computed when H is submodular, whereas typically, it is not possible.

– **Submodular relaxations**: it is possible to write most functions as the difference of two submodular functions H and G, leading to an exact reformulation in terms of minimizing $h(\mu)-$

$g(\mu)$ with respect to the product of measures μ. This problem is however non-convex anymore in general, and we could use majorization-minimization procedures [33]. Alternatively, if a function is a sum of simple functions, we may consider "submodular relaxations" of each of the component (a situation comparable with replacing functions by their convex envelopes). Like in the set-function case, a notion of submodular envelope similar to the convex envelope is not available. However, for any function H, one can always define a convex extension $\tilde{h}(\mu)$ through an optimal transport problem as $h_{\text{closure}}(\mu) = \inf_{\gamma \in \Pi(\mu)} \int_{\mathcal{X}} H(x) d\gamma(x)$. For functions of two variables, this can often be computed in closed form, and by taking the sum of these relaxations, we exactly get the usual linear programming relaxation (see [59] and references therein).

- **Submodular function maximization**: While this paper has focused primarily on minimization, it is worth exploring if algorithms for the maximization of submodular set-functions can be extended to the general case [41, 16], to obtain theoretical guarantees.
- **Divide-and-conquer algorithm**: For submodular set-functions, the separable optimization problem defined in Section 3.5 can be exactly solved by a sequence of at most n submodular optimization problem by a divide-and-conquer procedure [21]. It turns out that a similar procedure extends to general submodular functions on discrete domains (see Appendix B).
- **Minimizing sums of simple submodular functions**: Many submodular functions turn out to be decomposable as the sum of "simple" functions, that is functions for which minimization is particularly simple (see [30] for examples from computer vision). For submodular set-functions, decomposability has been used to derive efficient combinatorial [29] or convex-optimization-based [25] algorithms. These could probably be extended to general submodular functions.
- **Active-set methods**: For submodular set-functions, active-set techniques such as the minimum-norm-point algorithm have the potential to find the exact minimizer in finitely many iterations [58, 19]; they are based on the separable optimization problem from Section 3.5. Given that our extension simply adds inequality constraints, such an approach could easily be extended.
- **Adaptive discretization schemes**: Faced with functions defined on continuous domains, currently the only strategy is to discretize the domains; it would be interesting to study adaptive discretization strategies based on duality gap criteria.

Acknowledgements

The author thanks Marco Cuturi, Stefanie Jegelka, Gabriel Peyré, Suvrit Sra, and Tomas Werner, for helpful discussions and feedback. The main part of this work was carried while visiting the Institut des Hautes Etudes Scientifiques (IHES) in Bures-sur-Yvette, supported by the Schlumberger Chair for mathematical sciences.

References

[1] F. Bach. *Learning with Submodular Functions: A Convex Optimization Perspective*, volume 6 of *Foundations and Trends in Machine Learning*. NOW, 2013.

[2] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[3] E. Bae, J. Yuan, X.-C. Tai, and Y. Boykov. A fast continuous max-flow approach to non-convex multi-labeling problems. In *Efficient Algorithms for Global Optimization Methods in Computer Vision*, pages 134–154. Springer, 2014.

[4] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

[5] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[6] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[8] G. Carlier. On a class of multidimensional optimal transportation problems. *Journal of Convex Analysis*, 10(2):517–530, 2003.

[9] A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

[10] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

[11] J. Djolonga and A. Krause. From MAP to marginals: Variational inference in Bayesian submodular models. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[12] J. Djolonga and A. Krause. Scalable variational inference in log-supermodular models. In *International Conference on Machine Learning (ICML)*, 2015.

[13] J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer, 2003.

[14] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[15] P. Favati and F. Tardella. Convexity in nonlinear integer programming. *Ricerca Operativa*, 53:3–44, 1990.

[16] U. Feige, V. S. Mirrokni, and J. Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

[17] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[18] S. Friedland and S. Gaubert. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 438(10):3872–3884, 2013.

[19] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

[20] S. Fujishige and K. Murota. Notes on l-/m-convex functions and the separation theorems. *Mathematical Programming*, 88(1):129–146, 2000.

[21] H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.

[22] H. Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003.

[23] S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

[24] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[25] S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[26] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 121–148. MIT Press, 2011.

[27] S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

[28] S. Kim and M. Kojima. Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations. *Computational Optimization and Applications*, 26(2):143–154, 2003.

[29] V. Kolmogorov. Minimizing a sum of submodular functions. *Discrete Applied Mathematics*, 160(15):2246–2258, 2012.

[30] N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(3):531–552, 2011.

[31] A. Krause and C. Guestrin. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology*, 2(4), 2011.

[32] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[33] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

[34] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *NAACL/HLT*, 2011.

[35] G. G. Lorentz. An inequality for rearrangements. *American Mathematical Monthly*, 60(3):176–179, 1953.

[36] L. Lovász. Submodular functions and convexity. *Mathematical programming: The state of the art, Bonn*, pages 235–257, 1982.

[37] P. Milgrom and C. Shannon. Monotone comparative statics. *Econometrica: Journal of the Econometric Society*, pages 157–180, 1994.

[38] S. K. Mitter. Convex optimization in infinite dimensional spaces. In *Recent Advances in Learning and Control*, pages 161–179. Springer, 2008.

[39] K. P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[40] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4), February 2009.

[41] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming*, 14(1):265–294, 1978.

[42] J. B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

[43] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. A convex relaxation approach for computing minimal partitions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[44] W. Rudin. *Real and complex analysis*. McGraw-Hill, 1986.

[45] N. Ruozzi. Exactness of approximate MAP inference in continuous MRFs. In *Advances in Neural Information Processing Systems*, 2015.

[46] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

[47] D. Schlesinger and B. Flach. Transforming an arbitrary MinSum problem into a binary one. Technical Report TUD-FI06-01, Dresden University of Technology, Germany, April 2006.

[48] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

[49] A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume 24. Springer Science & Business Media, 2003.

[50] L. S. Shapley. Complements and substitutes in the optimal assignment problem. *Naval Research Logistics Quarterly*, 9(1):45–48, 1962.

[51] N. S. Shor. *Nondifferentiable optimization and polynomial problems*, volume 24. Springer Science & Business Media, 2013.

[52] D. M. Topkis. Minimizing a submodular function on a lattice. *Operations research*, 26(2):305–321, 1978.

[53] D. M. Topkis. *Supermodularity and complementarity*. Princeton University Press, 2011.

[54] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[55] J. Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the fortieth annual ACM symposium on Theory of Computing*, pages 67–74. ACM, 2008.

[56] Y. Wald and A. Globerson. Tightness results for local consistency relaxations in continuous MRFs. *Proc. Uncertainty in Artifical Intelligence (UAI)*, 2014.

[57] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, July 2007.

[58] P. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

[59] S. Živný, T. Werner, and D. Průša. The power of lp relaxation for map inference. In S. Nowozin, P. V. Gehler, J. Jancsary, and C. H. Lampert, editors, *Advanced Structured Prediction*. MIT Press, 2014.

A Proof of miscellaneous results

A.1 Submodularity of the Lovász extension of a submodular set-function

We consider a submodular set-function $G : \{0,1\}^n \to \mathbb{R}$ and its Lovász extension $g : \mathbb{R}^n \to \mathbb{R}$. In order to show the submodularity of g, we simply apply the definition and consider $x \in \mathbb{R}^n$ and two distinct basis vectors e_i and e_j with infinitesimal positive displacements a_i and a_j. If (i,j) belongs to two different level sets of x, then, for a_i and a_j small enough, $g(x+a_ie_i)+g(x+a_je_j)-g(x)-g(x+a_ie_i+a_je_j)$ is equal to zero. If (i,j) belongs to the same level sets, then, by explicitly computing the quantity for $a_i > a_j$ and $a_i < a_j$, it is non-negative (as a consequence of submodularity).

Note that then, the extension has another expression when $\mathcal{X} = [0,1]^n$, as, if $H = g$,

$$\begin{aligned} h(\mu) &= \int_0^1 g(F_{\mu_1}^{-1}(t),\dots,F_{\mu_n}^{-1}(t))dt \\ &= \int_0^1\int_0^1 G(\{i, F_{\mu_i}^{-1}(t) \geqslant z\})dtdz = \int_0^1\int_0^1 G(\{i, F_{\mu_i}(z) \geqslant t\})dtdz \\ &= \int_0^1 g(F_{\mu_1}(z),\dots,F_{\mu_n}(z))dz, \end{aligned}$$

which provides another proof of submodularity since g is convex.

A.2 Submodularity of the multi-linear extension of a submodular set-function

We consider a submodular set-function $G : \{0,1\}^n \to \mathbb{R}$ and its multi-linear extension [55], defined as a function $\tilde{g} : [0,1]^n \to \mathbb{R}$ with

$$\tilde{g}(x_1,\dots,x_n) = \mathbb{E}_{y_i \sim \text{Bernoulli}(x_i), i \in \{1,\dots,n\}} G(y),$$

where all Bernoulli random variables y_i are independent. In order to show submodularity, we only need to consider the case $n = 2$, for which the function $\tilde{g}$ is quadratic in x and the cross-term is non-positive because of the submodularity of G. See more details in [55]. The extension on a product of measures on $[0,1]^n$ does not seem to have the same simple interpretation as for the Lovász extension in Appendix A.1.

B Divide-and-conquer algorithm for separable optimization

We consider the optimization problem studied in Section 4.3:

$$\min_{\rho} h_{\downarrow}(\rho) + \sum_{i=1}^n \sum_{x_i=1}^{k_i-1} a_{ix_i}\big[\rho_i(x_i)\big] \text{ such that } \rho \in \prod_{i=1}^n \mathbb{R}_{\downarrow}^{k_i-1},$$

whose dual problem is the one of maximizing $\sum_{i=1}^n \sum_{x_i=1}^{k_i-1} -a_{ix_i}^*(-w_i(x_i))$ such that $w \in \mathcal{W}(H)$. We assume that all functions a_{ix_i} are strictly convex and differentiable with a Fenchel-conjugate with full domain. From Theorem 4, we know that it is equivalent to a sequence of submodular minimization problems of the form:

$$\min_{x \in \mathcal{X}} H(x) + \sum_{i=1}^n \sum_{y_i=1}^{x_i} a'_{iy_i}(t),$$

i.e., minimizing H plus a modular function. We now show that by solving a sequence of such problems (with added restrictions on the domains of the variables), we recover exactly the solution ρ of the original problem. This algorithm directly extends the one from [21], and we follow the exposition from [1, Section 9.1]. The recursive algorithm is as follows:

(1) Find the unique global maximizer of $-\sum_{i=1}^{n}\sum_{x_i=1}^{k_i-1} a^*_{ix_i}(-w_i(x_i))$ with the single constraint that $\sum_{i=1}^{n}\sum_{x_i=1}^{k_i-1} w_i(x_i) = H(k_1,\dots,k_n) - H(0)$. This can be typically obtained in closed form, or through the following one-dimensional problem (obtained by convex duality), $\min_{t\in\mathbb{R}} \big[H(k_1 - 1,\dots,k_n-1) - H(0)\big]t + \sum_{i=1}^{n}\sum_{x_i=1}^{k_i-1} a_{ix_i}(t)$, with $w_i(x_i)$ then obtained as $w_i(x_i) = -a'_{ix_i}(t)$.

(2) Find an element $y \in \mathcal{X}$ that minimizes $H(y) - \sum_{i=1}^{n}\sum_{z_i=1}^{y_i} w_i(z_i)$. This is a submodular function minimization problem.

(3) If $H(y) - \sum_{i=1}^{n}\sum_{z_i=1}^{y_i} w_i(z_i) = H(0)$, then exit ($w$ is optimal).

(4) Maximize $\sum_{i=1}^{n}\sum_{x_i=1}^{y_i} -a^*_{ix_i}(-w_i(x_i))$ over $w \in \mathcal{W}(H_{x\leqslant y})$, where $H_{x\leqslant y}$ is the restriction of H to $\{x \leqslant y\}$, to obtain $w^{\{x\leqslant y\}}$.

(5) Maximize $\sum_{i=1}^{n}\sum_{x_i=y_i+1}^{k_i-1} -a^*_{ix_i}(-w_i(x_i))$ over $w \in \mathcal{W}(H_{x\geqslant y+1})$, where $H_{x\geqslant y+1}$ is the restriction of H to $\{x \geqslant y+1\}$, to obtain $w^{\{x\geqslant y+1\}}$.

(6) Concatenate the two vectors $w^{\{x\geqslant y+1\}}$ and $w^{\{x\leqslant y\}}$ into w, which is the optimal solution.

The proof of correctness is the same as for set-functions [1, Section 9.1]: if the algorithm stops at step (3), then we indeed have the optimal solution because the optimum on a wider set happens to be in $\mathcal{W}(H)$. Since the optimal w in (1) is such that $w_i(x_i) = -a'_{ix_i}(t)$ for all i and x_i and a single real number t, the problem solved in (2) corresponds to one of the submodular function minimization problems from Theorem 4. From the statement (d) of that theorem, we know that the optimal primal solution ρ will be such that $\rho_i(x_i) \geqslant t$ for $x_i \leqslant y_i$ and $\rho_i(x_i) \leqslant t$ for $x_i \geqslant y_i+1$. Given the expression of $h_\downarrow(\rho)$ obtained from the greedy algorithm, if we impose that $\min_{x_i\leqslant y_i} \rho_i(x_i) \geqslant \max_{x_i\geqslant y_i+1} \rho_i(x_i)$, the function $h_\downarrow(\rho)$ is the sum of two independent terms and the minimization of the primal problem can be done into two separate pieces, which correspond exactly to $h^\downarrow_{x\leqslant y}(\rho^{x\leqslant y})$ and $h^\downarrow_{x\geqslant y+1}(\rho^{x\geqslant y+1})$. If we minimize the two problems separately, like done in steps (4) and (5), we thus only need to check that the decoupled solutions indeed have the correct ordering, that is $\min_{x_i\leqslant y_i} \rho_i(x_i) \geqslant \max_{x_i\geqslant y_i+1} \rho_i(x_i)$, which is true because of Theorem 4 applied to the two decoupled problems with the same value of t.