From Sets to Multisets: Provable Variational Inference for Probabilistic Integer Submodular Models

Aytunc Sahin [1]   **Yatao Bian** [2]   **Joachim M. Buhmann** [1]   **Andreas Krause** [1]

Abstract

Submodular functions have been studied extensively in machine learning and data mining. In particular, the optimization of submodular functions over the *integer lattice* (integer submodular functions) has recently attracted much interest, because this domain relates naturally to many practical problem settings, such as multilabel graph cut, budget allocation and revenue maximization with discrete assignments. In contrast, the use of these functions for probabilistic modeling has received surprisingly little attention so far. In this work, we firstly propose the Generalized Multilinear Extension, a continuous DR-submodular extension for integer submodular functions. We study central properties of this extension and formulate a new probabilistic model which is defined through *integer* submodular functions. Then, we introduce a block-coordinate ascent algorithm to perform approximate inference for this class of models and finally, we demonstrate its effectiveness and viability on several real-world social connection graph datasets with integer submodular objectives.

1. Introduction

Submodular functions have many applications in machine learning, which include data summarization (Tschiatschek et al., 2014; Lin & Bilmes, 2012), sensor placement (Krause et al., 2006) and computer vision (Boykov et al., 1999). They are defined on the subsets of the ground set $\mathcal{V}$ which contains n elements. Alternatively, they can be described as a pseudo-Boolean function which is defined on vertices of the unit hypercube $\{0, 1\}^n$. A crucial benefit of submodular functions is their attractive properties from the optimiza-

tion perspective. Using the celebrated greedy algorithm of Nemhauser et al. (1978), one can get a constant factor approximation guarantee for maximizing a submodular function subject to cardinality constraints. However, for more complex constraints, such as matroid constraints, the best guarantees can be obtained via a continuous relaxation of the problem. The *multilinear extension* (Calinescu et al., 2007) is a continuous extension of a submodular function to the full hypercube $[0, 1]^n$ and one can get the best approximation guarantee for matroid constraints by optimizing it using the continuous greedy algorithm (Calinescu et al., 2011).

Recently, *Probabilistic Submodular Models (PSMs)* (Djolonga & Krause, 2014; Tschiatschek et al., 2016; Gotovos et al., 2015) were introduced as Gibbs distributions over subsets parameterized via a submodular energy function. These distributions can represent expressive probabilistic models by capturing complex interactions among elements. Since the partition function of these models requires a sum over all subsets, it is *intractable* for large ground sets. One way to approximate this probability distribution is to use *variational inference*. When a fully factorized Bernoulli distribution is chosen as a variational distribution, the lower bound for the log-partition function – the *Evidence Lower Bound (ELBO)* – can be decomposed into two parts: The multilinear extension and the entropy of the variational distribution (Bian et al., 2019).

Though originally associated with set functions, submodularity can be generalized to *integer* (Soma et al., 2014; Soma & Yoshida, 2018) and *continuous* domains (Bach, 2019; Staib & Jegelka, 2017; Bian et al., 2017b; Hassani et al., 2017). Submodular functions over the integer lattice have applications in optimal budget allocation (Soma et al., 2014), sensor placement with discrete energy levels (Soma & Yoshida, 2015) and influence maximization with partial incentives (Demaine et al., 2014). Submodular optimization over continuous domains captures a rich class of tractable non-convex optimization problems, and arises in various applications, such as revenue maximization with continuous assignments (Bian et al., 2017b; 2020), robust budget allocation (Staib & Jegelka, 2017) and provable mean-field inference for PSMs (Bian et al., 2019). However, the family of *probabilistic submodular models defined through*

---

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland [2]Tencent AI Lab, Shenzhen, China 518057. Correspondence to: Aytunc Sahin <aytunc.sahin@inf.ethz.ch>, Yatao Bian <yatao.bian@gmail.com>.

integer and continuous submodular functions has been little explored, and it is, for example, not clear how to apply variational inference in a principled way.

Our main contributions are summarized as follows:

1. We present the *Generalized Multilinear Extension* (GME) for integer submodular functions and continuous submodular functions. The relation of GME to other continuous extensions is described in Figure 1. Though being non-convex, GME is DR-submodular in general, which allows to use guaranteed DR-submodular maximization techniques to conduct provable variational inference.

2. We introduce *Probabilistic Integer Submodular Models*, which are characterized through integer submodular functions. We show that using an appropriate variational distribution we can obtain a DR-submodular ELBO for variational inference. This ELBO is obtained as an expectation w.r.t. the same variational distribution through which the GME is defined.

3. To optimize this DR-submodular ELBO, we develop an efficient block coordinate ascent algorithm and show its effectiveness on real-world graph datasets.

		Optimization	
		Minimize	Maximize
Domain	Set	Lovasz Ext. (Lovász, 1983)	Multilinear Ext. (Calinescu et al., 2007)
	Integer	Extended Ext. (Bach, 2019)	Generalized ME **This work**

Figure 1. Continuous extensions for submodular functions depending on the function domain and optimization type.

2. Problem Setting

We consider submodular functions defined on the 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 $\mathbb{R}$ (Bach, 2019). A function f is *submodular* if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$ we have

$$f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \vee \mathbf{y}) + f(\mathbf{x} \wedge \mathbf{y}), \quad (1)$$

where $\vee$ and $\wedge$ are component-wise maximum and minimum functions respectively. From now on, a submodular function f will be referred as

- *set submodular* if $\mathcal{X}_i = \{0, 1\}$

- *integer submodular* if $\mathcal{X}_i = \{0, 1, \ldots, k-1\}$

- *continuous submodular* if $\mathcal{X}_i = [a, b]$

For ease of notation, we assume that $\mathcal{X}_i$ is same along each dimension, i.e. k does not depend on i. Our proofs will work with different $\mathcal{X}_i$ as well, i.e. using k_i instead of k. When f is continuous submodular and twice differentiable, we have

$$\forall i \neq j, \forall \mathbf{x} \in \mathcal{X}, \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \leq 0. \quad (2)$$

A function f has *diminishing returns* property if $\forall A \subseteq B \subseteq \mathcal{V}, e \notin B$ we have $f(A+e) - f(A) \geq f(B+e) - f(B)$. This property plays a significant role for the maximization of submodular functions. For set functions, this diminishing return property is equivalent to submodularity. However, for integer functions, submodularity is a weaker condition than the diminishing returns property. An integer function is called *diminishing returns submodular* (*DR-submodular*) if $\forall \mathbf{x} \leq \mathbf{y} \in \mathcal{X}$ and $i \in \mathcal{V}$, we have $f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) \geq f(\mathbf{y} + \mathbf{e}_i) - f(\mathbf{y})$ where $\mathbf{e}_i$ is the i-th unit vector. When f is continuous submodular and twice differentiable, DR-submodularity is equivalent to

$$\forall i, j, \forall \mathbf{x} \in \mathcal{X}, \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \leq 0. \quad (3)$$

Thus, submodularity only imposes constraints on *interactions* of coordinates, whereas DR-submodularity also requires that f is *concave along each coordinate* (which of course does not require that f is globally concave).

We will now discuss an equivalence between the integer vectors and classical *multisets*. With this connection, we have the following benefits: *i)* By borrowing well-established concepts, such as cardinality, intersection and set difference from multiset theory, we can define the corresponding concepts for integer vectors in a principled way. *ii)* The notions from multiset theory will ease the proof in the following part.

We follow the exposition from Blizard et al. (1991) and Syropoulos (2000) to define basic notions of multisets. A *multiset* is a natural generalization of a set, where elements can be contained *repeatedly*, i.e., it may contain a finite number of copies of a particular element. In classical sets, in contrast, distinct elements can only occur once. The number of times an element occurs in a multiset is called the *multiplicity* $\mu(i)$ of the element i.

Formally, a multiset $\mathcal{M}_\mathcal{V}$ is defined as a pair $\langle \mathcal{V}, \mu \rangle$, where $\mathcal{V}$ is the support and $\mu : \mathcal{V} \to \mathbb{N}$ is a function defining multiplicity for each element. Any ordinary set $\mathcal{V}$ can be viewed as a multiset $\langle \mathcal{V}, \chi_\mathcal{V} \rangle$ where $\chi_\mathcal{V}$ is the characteristic function of the set $\mathcal{V}$. A *submultiset* is a generalization of a subset. For each element in the support, you can choose a multiplicity to create a submultiset. Finally, we define the *powerset* of a multiset $\mathcal{P}(\mathcal{M}_\mathcal{V})$ as the set of submultisets of $\mathcal{M}_\mathcal{V}$ and the cardinality of the powerset is defined by

$$|\mathcal{P}(\mathcal{M}_\mathcal{V})| = \prod_{i \in \mathcal{V}} (\mu(i) + 1).$$

Note that this concept is consistent with the ordinary set theory where the multiplicity for each element is 1, i.e. $\mu(i) = 1$ and we recover $|\mathcal{P}(\mathcal{V})| = 2^n$.

Given the definitions of multisets, it is clear that one can represent any multiset with an integer vector, by viewing the multiplicity of the multiset μ as the argument $\mathbf{x}$ of the integer vectors. With this connection, we can transfer several important notions from multisets to integer vectors, such as the notion of a subset, set intersection, set union and set difference.

Now we define the notion of the subset of a multiset and the operations between multisets. Suppose $\mathcal{A} = \langle \mathcal{V}, \mu \rangle$ and $\mathcal{B} = \langle \mathcal{V}, \nu \rangle$ are two multisets and the corresponding integer vectors are $\mathbf{a}$ and $\mathbf{b}$. Then $\mathcal{A}$ is a *sub-multiset* of $\mathcal{B}$, $\mathcal{A} \subseteq \mathcal{B}$ if for all $e \in \mathcal{V}$ we have $\mu(e) \leq \nu(e)$ (or if $\mathbf{a} \leq \mathbf{b}$ for integer vectors). The *set difference* of $\mathcal{A}$ from $\mathcal{B}$, $\mathcal{A} \setminus \mathcal{B}$, is the multiset $\mathcal{C} = \langle \mathcal{V}, \xi \rangle$ where for all $e \in \mathcal{V}$ we have $\xi(e) = \max(\mu(e) - \nu(e), 0)$ (or $(\mathbf{a} - \mathbf{b}) \vee 0$ for integer vectors). The union and intersection of two multisets are similarly defined where $\xi(e) = \max(\mu(e), \nu(e))$ and $\xi(e) = \min(\mu(e), \nu(e))$ respectively (or $\mathbf{a} \vee \mathbf{b}$ and $\mathbf{a} \wedge \mathbf{b}$ for integer vectors).

3. The Generalized Multilinear Extension

Let f be an integer submodular function of n variables and $\boldsymbol{\rho}_i \in \mathbb{R}_+^{k-1}$ be the marginals of a k dimensional categorical distribution. Let $\boldsymbol{\rho} := [\boldsymbol{\rho}_1; ...; \boldsymbol{\rho}_n] \in \mathbb{R}_+^{n \times (k-1)}$. The generalized multilinear extension F is defined on the space of product of categorical distributions and can be written as:

$$F(\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_n) = \mathbb{E}_{R(\boldsymbol{\rho}) \sim \boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_n}[f(R(\boldsymbol{\rho}))]. \quad (4)$$

Hereby $R(\boldsymbol{\rho})$ is a random integer vector where each dimension is sampled from an independent categorical distribution. In other words, each $\boldsymbol{\rho}_i$ lives in the $k-1$ dimensional simplex Δ^{k-1} and $\boldsymbol{\rho}$ is the concatenation of all $\boldsymbol{\rho}_i$ vectors. We define the simplex Δ^{k-1} as

$$\Delta^{k-1} := \{\boldsymbol{\rho}_i \in \mathbb{R}^{k-1} : \rho_{i,1} + \ldots + \rho_{i,k-1} \leq 1,$$
$$\rho_{ij} \geq 0, j = 1, \ldots, k-1\}.$$

We define the union of n simplexes as Δ_n^{k-1}, and naturally $\boldsymbol{\rho} \in \Delta_n^{k-1}$. We need to compute the sum of k^n elements to compute the expectation in Equation (4). Note that when $k = 2$, this extension corresponds to the multilinear extension (Calinescu et al., 2011) of a set function. Here is an example with $n = 2$ and $k = 3$. In this case we have two categorical distributions which take three different values.

$F([\boldsymbol{\rho}_1; \boldsymbol{\rho}_2]) = F(\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}) =$
$f(0,0)(1 - \rho_{11} - \rho_{12})(1 - \rho_{21} - \rho_{22}) + f(2,2)\rho_{12}\rho_{22} +$
$f(1,0)\rho_{11}(1 - \rho_{21} - \rho_{22}) + f(0,1)(1 - \rho_{11} - \rho_{12})\rho_{21} +$
$f(2,0)\rho_{12}(1 - \rho_{21} - \rho_{22}) + f(0,2)(1 - \rho_{11} - \rho_{12})\rho_{22} +$
$f(1,1)\rho_{11}\rho_{21} + f(1,2)\rho_{11}\rho_{22} + f(2,1)\rho_{12}\rho_{21},$

where we have the following constraints

$$\rho_{11} + \rho_{12} \leq 1, \ \rho_{21} + \rho_{22} \leq 1, \ \rho_{ij} \geq 0, \ i,j = 1,2.$$

Therefore, maximizing the GME requires solving a constrained optimization problem with linear constraints. Now, we are going to answer the following questions:

1. *What are the properties of this GME when the integer function is submodular?*

2. *Does it have similar properties to the multilinear extension which makes it suitable for optimization purposes?*

This generalized multilinear extension in Equation (4) has the following properties:

Proposition 1. *Let F be the generalized multilinear extension of an integer submodular function f. Then we have*

1. *If f is monotone then $\frac{\partial F}{\partial \rho_{ij}} \geq 0$ for all $i \in \mathcal{V}$ and $j \in \{1, \ldots, k-1\}$;*

2. *F is DR-submodular (even if f is not), i.e., $\frac{\partial^2 F}{\partial \rho_{ij} \partial \rho_{kl}} \leq 0$ for all i, j, k, l.*

Proof. Given $\boldsymbol{\rho} \in \Delta_n^{k-1}$, let $\mathcal{R}_{\mathcal{V}}$ be a random multiset where elements appear independently with probabilities $\boldsymbol{\rho}_i$. Since F is multilinear, the partial derivative can be written as a difference of two GME. For the ease of notation, union and set difference for the multisets will be denoted as plus and minus.

$$\frac{\partial F}{\partial \rho_{ij}} = F(\boldsymbol{\rho}_1, \boldsymbol{\rho}_i = \mathbf{e}_j, \boldsymbol{\rho}_n) - F(\boldsymbol{\rho}_1, \boldsymbol{\rho}_i = \mathbf{0}, \boldsymbol{\rho}_n)$$
$$= \mathbb{E}_{\mathcal{R}_{\mathcal{V}} \sim \boldsymbol{\rho}}[f(\mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j)] - \mathbb{E}_{\mathcal{R}_{\mathcal{V}} \sim \boldsymbol{\rho}}[f(\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j)]$$
$$\geq 0,$$

since f is monotone and $\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j \subseteq \mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j$ where $\mathcal{E}_i^j$ is the multiset where ith element has multiplicity j.

For the second statement, we first observe that $\frac{\partial^2 F}{\partial \rho_{ij} \partial \rho_{kl}} = 0$ when $i = k$ for all j, l since F is multilinear.

$$\frac{\partial^2 F}{\partial \rho_{ij} \partial \rho_{kl}} = \frac{\partial}{\partial \rho_{ij}} \left(F(\boldsymbol{\rho}_k = \mathbf{e}_l, \boldsymbol{\rho}_n) - F(\boldsymbol{\rho}_k = \mathbf{0}, \boldsymbol{\rho}_n) \right)$$
$$= F(\boldsymbol{\rho}_i = \mathbf{e}_j, \boldsymbol{\rho}_k = \mathbf{e}_l) + F(\boldsymbol{\rho}_i = \mathbf{0}, \boldsymbol{\rho}_k = \mathbf{0})$$
$$- F(\boldsymbol{\rho}_i = \mathbf{e}_j, \boldsymbol{\rho}_k = \mathbf{0}) - F(\boldsymbol{\rho}_i = \mathbf{0}, \boldsymbol{\rho}_k = \mathbf{e}_l)$$
$$= \mathbb{E}[f(\mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j + \mathcal{E}_k^l)] + \mathbb{E}[f(\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j - \mathcal{E}_k^l)]$$
$$- \mathbb{E}[f(\mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j - \mathcal{E}_k^l)] - f(\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j + \mathcal{E}_k^l)$$
$$\leq 0,$$

since $f(\mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j + \mathcal{E}_k^l) + f(\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j - \mathcal{E}_k^l) \leq f(\mathcal{R}_{\mathcal{V}} + \mathcal{E}_i^j - \mathcal{E}_k^l) + f(\mathcal{R}_{\mathcal{V}} - \mathcal{E}_i^j + \mathcal{E}_k^l)$ because of the integer submodularity of f. $\quad \square$

Some remarks about Proposition 1 are in order:

1. We *do not* require f to be DR-submodular in order to get a DR-submodular continuous extension. Even if f is only integer submodular we still get a DR-submodular extension.

2. This generalized multilinear extension can also be applied when f is continuous submodular: By discretizing the interval $[a, b]$ where the function is defined, firstly we get an integer submodular function. Then applying the generalized multilinear extension, we get a DR-submodular extension.

3.1. On the Relation with Other Continuous Extensions

To our knowledge, there are two other continuous extensions (Soma & Yoshida, 2018; Bach, 2019) that can be used for integer submodular functions.

Soma & Yoshida (2018) define a continuous extension in order to maximize an integer submodular function subject to a polymatroid constraint. For $\mathbf{x} \in \mathbb{R}^{\mathcal{V}}$, let $\lfloor \mathbf{x} \rfloor$ denote the vector obtained by rounding down each entry of $\mathbf{x}$. For $a \in \mathbb{R}$, let $\langle a \rangle$ denote the fractional part of a, namely $\langle a \rangle = a - \lfloor a \rfloor$. For $\mathbf{x} \in \mathbb{R}$, $\mathcal{D}(\mathbf{x})$ is defined as the distribution from which $\overline{\mathbf{x}}$ is sampled such that $\overline{\mathbf{x}}(i) = \lfloor \mathbf{x}(i) \rfloor$ with probability $1 - \langle \mathbf{x}(i) \rangle$ and $\overline{\mathbf{x}}(i) = \lceil \mathbf{x}(i) \rceil$ with probability $\langle x(i) \rangle$ for each $i \in \mathcal{V}$. Then the continuous extension $F_S : \mathbb{R}_+^{\mathcal{V}} \to \mathbb{R}_+$ is defined as

$$F_S(\mathbf{x}) = \mathbb{E}_{\overline{\mathbf{x}} \sim \mathcal{D}(\mathbf{x})}[f(\overline{\mathbf{x}})]$$
$$= \sum_{S \subseteq \mathcal{V}} f(\lfloor \mathbf{x} \rfloor + \mathbf{e}_S) \prod_{i \in S} \langle \mathbf{x}(i) \rangle \prod_{i \notin S} (1 - \langle \mathbf{x}(i) \rangle).$$

In this case, F is obtained by gluing the multilinear extension of f restricted to each hypercube. However, this extension is non-smooth and it is not clear how it can be used in variational inference for PSMs with integer submodular functions. On the other hand, the proposed GME is smooth and corresponds to the one part of the ELBO for variational inference for PSMs with integer submodular functions as we show below.

Another continuous extension, which is suited for submodular function *minimization* is presented in Bach (2019). It is originally presented for continuous submodular functions but can also be used for integer submodular functions by discretizing the continuous interval. The extension $f_{\text{cumulative}}$ is defined on the set of products of Radon probability measures $\mu \in \prod_{i=1}^n P(\mathcal{X}_i)$ as:

$$f_{\text{cumulative}}(\mu_1, \dots, \mu_n) = \int_0^1 f(\theta_1(t), \dots, \theta_n(t)) dt,$$

where $\theta_i(t) = \sup\{x_i \in \mathcal{X}_i, F_{\mu_i}(x_i) \geq t\}$ with $F_{\mu_i}(x_i)$ denoting the cumulative distribution function of μ_i. This extension is convex when the function f is submodular.

Our result is consistent with previous reduction techniques for DR-submodular functions (Ene & Nguyen, 2016) which depends logarithmically on k. In their setting, they first apply their reduction to get a ground set of size $n \log k$, then use the classical multilinear extension which requires the sum over $2^{n \log k} = k^n$ values. On the other hand, our GME is based on the ground set of size nk but requires the sum over k^n values. If the classical multilinear extension was used, 2^{kn} values would be summed up. In short, our GME requires the same number of function evaluation as the classical multilinear extension applied to the reduced ground set.

4. Variational Inference for Probabilistic Integer Submodular Models

Now, we consider distributions over multisets of $\mathcal{V}$, namely $p(\mathbf{x}) = \frac{1}{Z} e^{f(\mathbf{x})}$, where $\mathbf{x}$ is defined on the integer lattice $\mathbf{x} \in \{0, \dots, k-1\}^{\mathcal{V}}$ and $f(\mathbf{x})$ is submodular. We call such distributions *Probabilistic Integer Submodular Models* (PISMs) and they generalize the existing Probabilistic Submodular Models which are defined on the vertices of the unit hypercube. PISMs can generalize various models in practice. For example, the classical Potts model (Li, 2009) with repulsive pairwise terms is used in relation to the q-coloring of the vertices of a graph (Aldous et al., 2013). With PISMs, we can model higher order interactions and capture more complicated structure.

In our case, w.l.o.g, we assume that each element of $\mathcal{V}$ has k levels, i.e., in multiset notation it is defined as $\{\underbrace{e_1, \dots, e_1}_{k-1 \text{ times}}, \dots, \underbrace{e_n, \dots, e_n}_{k-1 \text{ times}}\}$. Since not having an element is also defined as a level, the multiplicity of each element is $k - 1$ with $|\mathcal{P}(\mathcal{M}_{\mathcal{V}})| = k^n$. The normalization constant

$$Z = \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} e^{f(\mathbf{x})} \tag{5}$$

is also called the partition function and involves a summation over k^n terms. When n is large, computing this sum becomes intractable and variational inference (Blei et al., 2017) is one of the widely used methods to approximate the probability density $p(\mathbf{x})$. Let $q(\mathbf{x}, \boldsymbol{\rho})$ be a variational distribution with its variational parameter $\boldsymbol{\rho}$. When we minimize the KL divergence between $q(\mathbf{x}, \boldsymbol{\rho})$ and $p(\mathbf{x})$ we have

$$\text{KL}(q(\mathbf{x}, \boldsymbol{\rho}) \parallel p(\mathbf{x})) = \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} q(\mathbf{x}, \boldsymbol{\rho}) \log \frac{q(\mathbf{x}, \boldsymbol{\rho})}{p(\mathbf{x})}$$

$$= \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} q(\mathbf{x}, \boldsymbol{\rho}) \log q(\mathbf{x}, \boldsymbol{\rho})$$

$$- \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} q(\mathbf{x}, \boldsymbol{\rho}) f(\mathbf{x}) + \log Z.$$

Since KL divergence is always non-negative, we get an expression which gives a lower bound for the log partition function. This term is called *Evidence Lower Bound (ELBO)* and can be written as:

$$\text{ELBO}(\boldsymbol{\rho}) = \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} q(\mathbf{x}, \boldsymbol{\rho}) f(\mathbf{x})$$
$$- \sum_{\mathbf{x} \in \mathcal{P}(\mathcal{M}_{\mathcal{V}})} q(\mathbf{x}, \boldsymbol{\rho}) \log q(\mathbf{x}, \boldsymbol{\rho}) \leq \log Z.$$

Usually, the ELBO is non-convex and difficult to optimize if we do not make additional assumptions. One of the most popular assumptions is the mean-field assumption which uses a fully-factorized variational distribution. If assuming that our variational distribution belongs to the mean-field family, i.e., $q(\mathbf{x}, \boldsymbol{\rho}) = \prod_i q_i(x_i, \boldsymbol{\rho}_i)$, where each q_i is a categorical distribution with parameter $\boldsymbol{\rho}_i$, the ELBO objective becomes a sum of the GME of f and the entropy of categorical distributions. Moreover, the entropy part of ELBO becomes separable and we can write the mean-field ELBO as:

$$\text{ELBO}_{\text{MF}}(\boldsymbol{\rho}) = F(\boldsymbol{\rho}) + \sum_{i \in \mathcal{V}} H(\boldsymbol{\rho}_i),$$

where H is the entropy of one dimensional categorical distribution. The following proposition establishes the DR-submodularity of $\text{ELBO}_{\text{MF}}(\boldsymbol{\rho})$, all the omitted proofs in the sequel are deferred to Appendix A.

Proposition 2. *$\text{ELBO}_{MF}(\boldsymbol{\rho})$ is DR-submodular in $\boldsymbol{\rho}$ as long as the energy function f is integer submodular.*

Since $\text{ELBO}_{\text{MF}}(\boldsymbol{\rho})$ is DR-submodular we can use various DR-submodular maximization algorithms (Mokhtari et al., 2017; Hassani et al., 2017; Bian et al., 2017a) to optimize it.

5. Extension to Continuous Variables

It is also relevant to consider *continuous* PSMs, defined using a *continuous DR-submodular* function $f(\mathbf{x})$ where the domain is the hypercube $[a, b]^n$. Here, the partition function becomes $Z = \int \exp(f(\mathbf{x})) d\mathbf{x}$. For large values of n, this integral becomes intractable and we will use variational inference to approximate $p(\mathbf{x})$. Let $q(\mathbf{x}, \boldsymbol{\theta})$ be a variational distribution with its variational parameter $\boldsymbol{\theta}$. Following the same logic in the previous section, the ELBO becomes

$$\text{ELBO}(\boldsymbol{\theta}) = \underbrace{\int q(\mathbf{x}, \boldsymbol{\theta}) f(\mathbf{x}) d\mathbf{x}}_{L(\boldsymbol{\theta})} - \int q(\mathbf{x}, \boldsymbol{\theta}) \log q(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x}.$$

We again make a mean-field assumption about the variational distribution $q(\mathbf{x}, \boldsymbol{\theta})$. $\text{ELBO}(\boldsymbol{\theta})$ is still highly non-convex in this circumstance. In order to provably maximize the $\text{ELBO}(\boldsymbol{\theta})$, we ideally would want it to have nice structure, such as DR-submodularity w.r.t. $\boldsymbol{\theta}$.

This amounts to choosing some family of variational distributions $q(\mathbf{x}, \boldsymbol{\theta})$ such that $\text{ELBO}(\boldsymbol{\theta})$ is still tractable. Then the following important question arises:

For which variational family is the ELBO objective DR-submodular w.r.t. to the variational parameter $\boldsymbol{\theta}$?

We consider firstly the first part of the ELBO ($L(\boldsymbol{\theta})$) and give a sufficient condition such that it becomes DR-submodular in Proposition 3. Because of the mean-field assumption, for *simple* variational distributions, the entropy part becomes separable and concave. Therefore, we obtain a DR-submodular ELBO.

Proposition 3. *If $q(x, \theta)$ is a probability density function such that $q_x''(x, \theta) = q_\theta''(x, \theta)$, then we have*

$$L''(\theta) = \int_{-\infty}^{+\infty} q_\theta''(x, \theta) f(x) dx = \int_{-\infty}^{+\infty} q(x, \theta) f''(x) dx \quad (6)$$

for sufficiently smooth f. This also holds in higher dimensions under the mean-field assumption.

Now, we look at various probability distributions and for distributions such as Gaussian and Laplace, this property holds. For example, if we choose a Gaussian with fixed variance, where the variational parameter is μ, we have $q(x, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$ with $\frac{\partial^2 q}{\partial \mu^2} = \frac{\partial^2 q}{\partial x^2}$ which satisfies the above proposition. This implies that first part of the ELBO is DR-submodular. Since the differentiable entropy part is $\sum_{i \in \mathcal{V}} \log(\sigma_i \sqrt{2\pi e})$ and does not depend on μ, ELBO is also DR-submodular with respect to the variational parameter μ. In this case, since there is no entropy regularizer, ELBO puts all the mass on the maximizer of f, which is not desirable.

One way to solve this issue is to discretize the interval and use the categorical distribution similar as that in Section 4. Then the ELBO objective becomes DR-submodular. This discretization approach is also used in Bach (2019) to minimize the convex extension of submodular functions.

6. Algorithms for Approximate Mean-Field Inference

Approximate inference amounts to solving the following non-convex optimization problem:

$$\max_{\boldsymbol{\rho} \in \Delta_n^{k-1}} \text{ELBO}_{\text{MF}}(\boldsymbol{\rho}) = F(\boldsymbol{\rho}) + H(\boldsymbol{\rho}), \quad (7)$$

where $H(\boldsymbol{\rho})$ is the entropy of the n categorical distributions. The optimization problem is one instance of the constrained non-monotone DR-submodular maximization problem, for which several guaranteed algorithms exist. For example, the `Shrunken FW` algorithm has a $1/e$ approximation

guarantee and the `Two-Phase FW` algorithm has a $1/4$ approximation guarantee (Bian et al., 2017a). In addition to these algorithms, we present a block coordinate ascent algorithm, abbreviated as `Block CA`, which tries to optimize the marginals for one categorical random variable in each iteration.

6.1. The Block Coordinate Ascent Algorithm

We summarize the main steps in Algorithm 1, where $\rho^{k-1}|_{i_k}(\boldsymbol{\xi})$ means changing the marginals of the i_k-th random variable to be $\boldsymbol{\xi}$ while keeping all other marginals fixed. Algorithm 1 starts by initializing ρ^0 to be some point in the probability simplex. Then it runs for K iterations. In iteration k, it optimizes the marginals of random variable i_k while keeping all other marginals fixed, which is illustrated in line 4. For the ELBO objective used in this work, one can verify that it has a closed form solution, as shown by the following proposition.

Algorithm 1 The `Block CA` Algorithm

Input: $\max_{\rho \in \Delta_n^{k-1}} \mathrm{ELBO_{MF}}(\rho)$, # iterations K

1   Initialize $\rho^0 \in \Delta_n^{k-1}$, $k \leftarrow 1$
2   **while** $k \leq K$ **do**
3   let i_k be the random variable being operated
4      $\boldsymbol{\xi}_{i_k} \leftarrow \arg\max_{\boldsymbol{\xi}} f(\rho^{k-1}|_{i_k}(\boldsymbol{\xi}))$
5      $\rho^k \leftarrow \rho^{k-1}|_{i_k}(\boldsymbol{\xi}_{i_k})$
6      $k \leftarrow k + 1$

Output: ρ^K

Proposition 4. *When maximizing ELBO(ρ) in Equation* (7), *if one only optimizes for ρ_i while keeping all other marginals fixed, we have the following closed form solution (let $\nabla_{ij} := \nabla_{\rho_{ij}} F(\rho)$ for notational simplicity):*

$$\rho_{ij} = \frac{\exp(\nabla_{ij})}{1 + \sum_{j'} \exp(\nabla_{ij'})}, \forall j \in \{1, ..., k-1\}. \quad (8)$$

Since the Block Coordinate Ascent algorithm can only *increase* the ELBO in each iteration, if we initialize it using the solutions of `Shrunken FW` and `Two-Phase FW` and therefore still retain their respective approximation guarantees.

7. Experiments

We now illustrate the efficacy of the proposed generalized multilinear extension and algorithms on two classes of applications: Revenue maximization with discrete assignments and facility location with discrete levels. Both of these two applications have integer submodular objectives. With a probabilistic treatment using the proposed PISMs, we are

able to i) provide refined probabilistic solutions using variational inference techniques, instead of the discrete solutions without a probabilistic treatment; ii) conduct parameter learning through, for example, the maximum likelihood framework, given that we have enough training data of the corresponding objectives. Due to limit of training data, we are only going to illustrate the first benefit in this section.

We used the oblivious step sizes $2/(t+2)$ for the `Two-Phase FW` algorithm where t is the iteration index. Below, by "one epoch" we mean one full gradient evaluation of the objective. Notice that in general it costs $\mathcal{O}(k^n)$ time to evaluate $F(\rho)$ precisely. In practice, we use Monte Carlo sampling to estimate $F(\rho)$ by sampling the integer function values according to the categorical distribution. Thanks to the Hoeffding bound (Hoeffding, 1963), one can easily show that the estimated function value can be arbitrarily close to true $F(\rho)$ with polynomial number of samples.

7.1. Revenue Maximization with Discrete Assignments

Given a social connection graph with nodes describing n users and edges describing their connection strength, the viral marketing strategy suggests to choose a small subset of buyers to give them some product for free, to trigger a cascade of further adoptions through "word-of-mouth" effects, in order to maximize the total revenue (Hartline et al., 2008). One model with "discrete" product assignments was used by (Soma & Yoshida, 2017) and Dürr et al. (2019), which is motivated by the fact that giving a user more free product, the user will be more likely to advocate the product. It can be viewed as a simplified variant of the Influence-and-Exploit (IE) strategy of Hartline et al. (2008).

In the *influence* stage, for each of the user i, if giving him x_i units of products for free, the user becomes an advocate of the product with probability $1 - q^{x_i}$ (independently from other users), where $q \in (0, 1)$ is a parameter. In the *exploit* stage: suppose that a set S of users advocate the product while the complement set $\mathcal{V} \setminus S$ of users do not. Now the revenue comes from the users in $\mathcal{V} \setminus S$, and they will be influenced by the advocates with probability proportional to the edge weights. So the expected revenue is a function $f : \mathbb{Z}_+^{\mathcal{V}} \to \mathbb{R}_+$:

$$f(\mathbf{x}) = \mathbb{E}_S \left[\sum_{i \in S} \sum_{j \in \mathcal{V} \setminus S} W_{ij} \right]$$
$$= \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V} \setminus \{i\}} W_{ij}(1 - q^{x_i})q^{x_j}, \quad (9)$$

where W is the adjacency matrix of the social connection graph.

Experimental results. For variational inference of the PISMs which are defined through the above revenue objec-
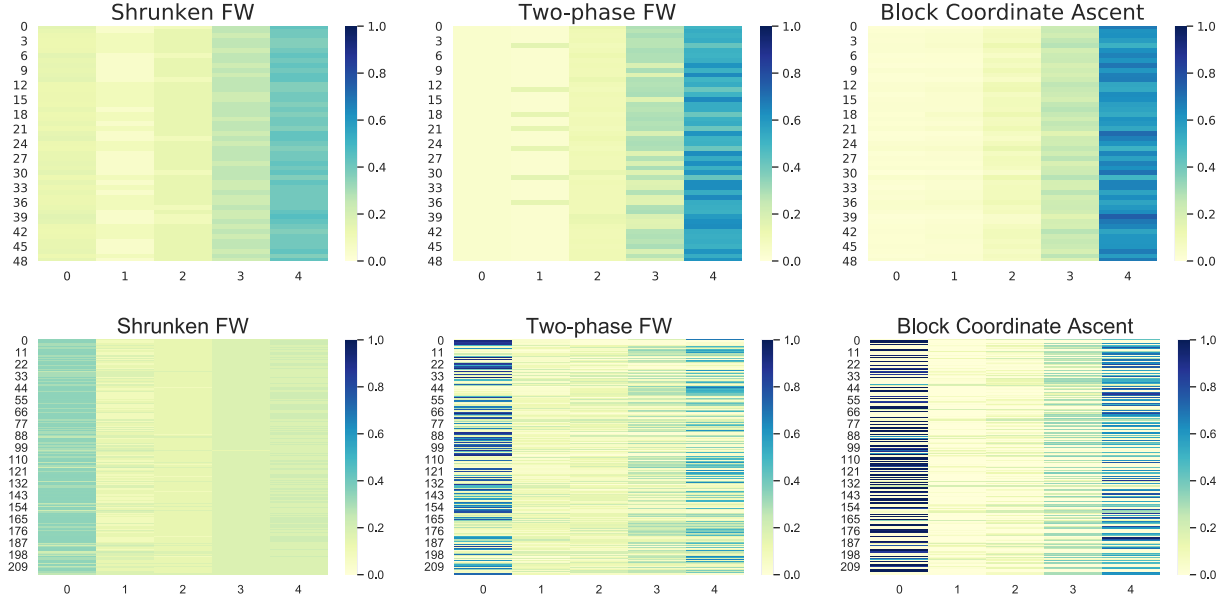
Figure 2. Marginals of variational distribution for two different models: First row corresponds to the model with facility location function with 50 nodes and 5 levels. Second row corresponds to the model with revenue maximization function on the real world graph with 217 nodes and 5 levels. Each row corresponds to elements in the ground set and each column corresponds to their levels. When a box is darker, there is more probability mass assigned to that level. In both models, one can observe that the marginals of `Shrunken FW` are smoother than the marginals of `Block CA`.
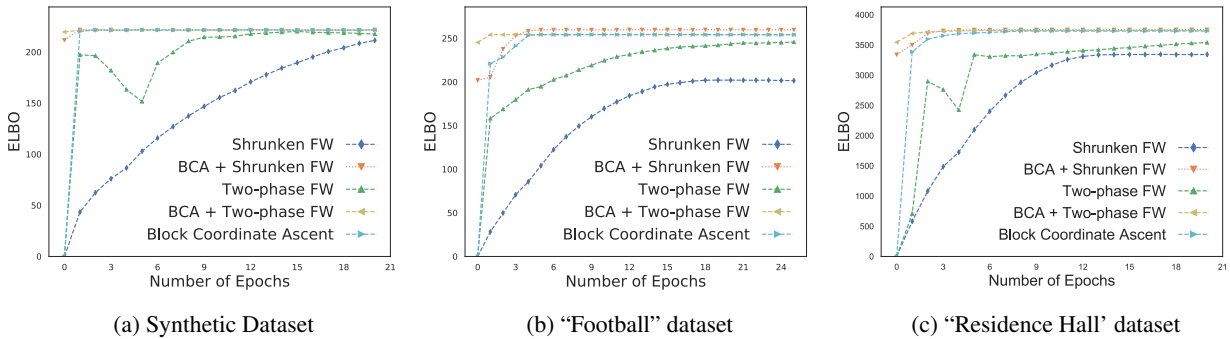

Figure 3. ELBO trajectories with different datasets. In all cases, `Block CA` achieves highest ELBO value and fastest convergence. `Shrunken FW` is the slowest to converge and obtains the lowest ELBO value. `Two-Phase FW` obtains faster convergence than `Shrunken FW`.

tive, we used real-world graphs from the Konect network collection (Kunegis, 2013)[1], for example, the network of "Residence Hall" ($n = 217$, #edges = 2,672) and from Pajek dataset[2], for example, the network of "Football" ($n = 35$, #edges = 118). Full information on the datasets and experimental results are deferred to Appendix B due to space limit.

Figure 3b shows the trajectories of the `Shrunken FW`,

---

[1]http://konect.uni-koblenz.de/networks
[2]http://vlado.fmf.uni-lj.si/pub/networks/data/

`Two-Phase FW` and `Block CA` which is initialized by the output of the `Shrunken FW` and `Two-Phase FW` for the "Football" dataset. One can see that `Block CA` type algorithms converge the fastest, `Two-Phase FW` follows, `Shrunken FW` is the last. `Block CA` which is initialized by the result of the `Shrunken FW` attains the highest ELBO value, followed by `Block CA` initialized by `Two-Phase FW`. They both outperform vanilla `Block CA` initialized by 0. `Shrunken FW` and `Two-Phase FW` converge to the almost same value, both significantly lower than `Block CA` variants. The first line of Figure 2

presents the marginals $\boldsymbol{\rho}$ of the variational distribution. We can see that `Block CA` tends to produce sharp marginals, while `Shrunken FW` and `Two-Phase FW` provide smoother marginals. This is consistent with the intuition that Frank-Wolfe style algorithms usually generate solutions that lie in the interior of the constraints. The output solution of the Frank-Wolfe algorithm is a convex combination of the extreme points of the constraint set (Jaggi, 2013), so it usually lies in the interior of the constraints. In terms of the runtime, we measured the time spent for one epoch for each of the methods. One epoch of `Shrunken FW` takes 50.2 seconds, followed by `Block CA` (53.3 seconds) and `Two-Phase FW` (75.4 seconds). We note that `Block CA` seems to converge at iterations 5-6, whereas `Two-Phase FW` and `Shrunken FW` require much more epochs to converge. Running `Block CA` for a few epochs on top of the `Shrunken FW` result, which is obtained by running a fewer number of epochs, outperforms running `Shrunken FW` for more epochs. Also, Figure 3c shows the trajectories of `Shrunken FW`, `Two-Phase FW` and `Block CA` which is initialized by the output of the `Shrunken FW` and `Two-Phase FW` for the "Residence Hall" dataset. Here similar to the Figure 3b, we can see that `Block CA` type algorithms converge the fastest, `Two-Phase FW` follows, `Shrunken FW` is the last.

7.2. Facility Location with Multiple Levels

We use the facility location function with different levels of facilities from Soma et al. (2014). In classical submodular facility location setting, we decide to either open a facility or not. Here, we have $|\mathcal{V}|$ facilities and we would like to decide on the level of the facility we want to open in order to serve m number of customers. The level of the facilities is represented as integers in $\{0, \ldots, k-1\}$. If a facility j of level c_j is opened, then it provides utility $w_{ij}(c_j)$ to the customer i. We choose $w_{ij} : \{0, \ldots, k-1\} \to \mathbb{R}$ as a monotone function and each customer chooses the facility with highest value. When there are multiple facilities with levels c_j, the total value provided to the customers is given by $\sum_{i=1}^{m} \max_{j \in \mathcal{V}} w_{ij}(c_j)$. This function is submodular and monotone over the integer lattice.

Experimental results. In this experiment, we use an integer facility location function with $n = 50$ facilities and customers and $k = 5$ levels. First, we generate matrices L_i where each element follows a standard normal distribution for $i \in \{1, \ldots, k\}$. Then we compute the matrix $\frac{1}{n} L_i L_i^T$ and take their absolute values. For each level we sum up the previous utility values in order to get a monotone function. We run `Shrunken FW`, `Two-Phase FW` and variants of `Block CA` for 20 epochs and the variants of `Block CA` achieve similar ELBO values whereas `Shrunken FW` and `Two-Phase FW` obtain a lower

ELBO value in Figure 3a. In the second row of Figure 2 we present a heatmap of the probabilities for the variational distribution $\boldsymbol{\rho}$. The x-axis represents different levels and the y-axis represents different facilities. If we look at the marginals of the `Shrunken FW`, we see that there is a gradual increase in the probabilities assigned to different levels. On the other hand, `Block CA` assigns very high probability to the highest level and assigns lower probability to the remaining levels. This behavior can be related to the nature of the Frank-Wolfe style algorithms. Since `Shrunken FW` uses convex combination of the vertices of the constraint set, the marginals are smoother than the `Block CA`.

8. Related Work

The first systematic study for probabilistic models defined through submodular *set* functions is presented by Djolonga & Krause (2014). Djolonga et al. (2016) develop an efficient approximate inference algorithm for more general models with both submodular and supermodular functions. Gotovos et al. (2015) analyze MCMC sampling to perform approximate inference in PSMs and Gotovos et al. (2018) introduce a new sampling strategy for accelerating mixing in PSMs.

Integer and continuous submodular optimization problems attract considerable attention recently. Gottschalk & Peis (2015) present a deterministic algorithm to maximize non-monotone submodular functions on a bounded lattice; Soma et al. (2014) consider maximizing monotone integer submodular functions with a knapsack constraint; and Soma & Yoshida (2017) study the integer submodular cover problem with applications on sensor placement with discrete energy levels. Qian et al. (2018) consider submodular maximization problems subject to size constraints while relaxing the submodularity assumption. Bian et al. (2017b; 2020) characterize the notion of continuous submodularity, present continuous submodularity composition rules and propose an optimal algorithm for monotone DR-submodular maximization. Hassani et al. (2017) show that the projected gradient ascent algorithm achieves a 1/2 approximation for maximizing monotone DR-submodular functions. Bian et al. (2017a) present the local-global relation and guaranteed algorithms for non-monotone DR-submodular maximization. Recently, Bian et al. (2019); Niazadeh et al. (2018) propose optimal algorithms for non-monotone DR-submodular maximization with a box constraint. Continuous submodular maximization is also well studied in the stochastic setting (Hassani et al., 2017; Mokhtari et al., 2018) and online setting (Chen et al., 2018).

9. Conclusion

We considered the problem of variational inference in Probabilistic Integer Submodular Models and introduced a novel

continuous extension for integer submodular functions. It can be viewed as an expectation under fully-factorized marginals and we proved that it is DR-submodular even if the integer function is just submodular. Then we introduced Probabilistic Integer Submodular Models and showed that ELBO becomes DR-submodular if we use a fully-factorized categorical distribution as the variational distribution. Finally, we presented an efficient block coordinate ascent algorithm to optimize the ELBO and showed the effectiveness of our method in real world graph mining applications.

References

Aldous, D., Grimmett, G. R., Howard, C. D., Martinelli, F., Steele, J. M., and Saloff-Coste, L. *Probability on discrete structures*, volume 110. Springer Science & Business Media, 2013.

Bach, F. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, pp. 419–459, 2019.

Bian, A., Levy, K. Y., Krause, A., and Buhmann, J. M. Continuous dr-submodular maximization: Structure and algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 486–496, 2017a.

Bian, A. A., Mirzasoleiman, B., Buhmann, J. M., and Krause, A. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 111–120, 2017b.

Bian, Y., Buhmann, J. M., and Krause, A. Continuous submodular function maximization. *arXiv preprint arXiv:2006.13474*, 2020.

Bian, Y. A., Buhmann, J. M., and Krause, A. Optimal continuous DR-submodular maximization and applications to provable mean field inference. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 644–653, 2019.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Blizard, W. D. et al. The development of multiset theory. *Modern logic*, 1(4):319–352, 1991.

Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pp. 377–384. IEEE, 1999.

Calinescu, G., Chekuri, C., Pál, M., and Vondrák, J. Maximizing a submodular set function subject to a matroid constraint. In *International Conference on Integer Programming and Combinatorial Optimization*, pp. 182–196. Springer, 2007.

Calinescu, G., Chekuri, C., Pál, M., and Vondrák, J. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM J. Comput.*, 40(6):1740–1766, 2011.

Chen, L., Hassani, H., and Karbasi, A. Online continuous submodular maximization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1896–1905, 2018.

Demaine, E. D., Hajiaghayi, M., Mahini, H., Malec, D. L., Raghavan, S., Sawant, A., and Zadimoghadam, M. How to influence people with partial incentives. In *Proceedings of the 23rd international conference on World wide web*, pp. 937–948. ACM, 2014.

Djolonga, J. and Krause, A. From map to marginals: Variational inference in bayesian submodular models. In *Advances in Neural Information Processing Systems*, pp. 244–252, 2014.

Djolonga, J., Tschiatschek, S., and Krause, A. Variational inference in mixed probabilistic submodular models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1767–1775, 2016.

Dürr, C., Thang, N. K., Srivastav, A., and Tible, L. Nonmonotone dr-submodular maximization: Approximation and regret guarantees. *arXiv preprint arXiv:1905.09595*, 2019.

Ene, A. and Nguyen, H. L. A reduction for optimizing lattice submodular functions with diminishing returns. *CoRR*, abs/1606.08362, 2016. URL http://arxiv.org/abs/1606.08362.

Gotovos, A., Hassani, H., and Krause, A. Sampling from probabilistic submodular models. In *Advances in Neural Information Processing Systems*, pp. 1945–1953, 2015.

Gotovos, A., Hassani, H., Krause, A., and Jegelka, S. Discrete sampling using semigradient-based product mixtures. *arXiv preprint arXiv:1807.01808*, 2018.

Gottschalk, C. and Peis, B. Submodular function maximization on the bounded integer lattice. In *International Workshop on Approximation and Online Algorithms*, pp. 133–144. Springer, 2015.

Hartline, J., Mirrokni, V., and Sundararajan, M. Optimal marketing strategies over social networks. In *Proceedings*

of the 17th international conference on World Wide Web, pp. 189–198. ACM, 2008.

Hassani, H., Soltanolkotabi, M., and Karbasi, A. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5837–5847, 2017.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML13, pp. I427I435, 2013.

Krause, A., Guestrin, C., Gupta, A., and Kleinberg, J. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *Proceedings of the 5th International Conference on Information Processing in Sensor Networks*, IPSN '06, pp. 2–10, 2006.

Kunegis, J. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1343–1350. ACM, 2013.

Li, S. Z. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

Lin, H. and Bilmes, J. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pp. 479–490, 2012.

Lovász, L. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pp. 235–257. Springer, 1983.

Mokhtari, A., Hassani, H., and Karbasi, A. Conditional gradient method for stochastic submodular maximization: Closing the gap. *arXiv preprint arXiv:1711.01660*, 2017.

Mokhtari, A., Hassani, H., and Karbasi, A. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *arXiv preprint arXiv:1804.09554*, 2018.

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

Niazadeh, R., Roughgarden, T., and Wang, J. Optimal algorithms for continuous non-monotone submodular and dr-submodular maximization. In *Advances in Neural Information Processing Systems*, pp. 9594–9604, 2018.

Qian, C., Zhang, Y., Tang, K., and Yao, X. On multiset selection with size constraints. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Soma, T. and Yoshida, Y. A generalization of submodular cover via the diminishing return property on the integer lattice. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 847–855, 2015.

Soma, T. and Yoshida, Y. Non-monotone dr-submodular function maximization. In *AAAI*, volume 17, pp. 898–904, 2017.

Soma, T. and Yoshida, Y. Maximizing monotone submodular functions over the integer lattice. *Mathematical Programming*, 172(1-2):539–563, 2018.

Soma, T., Kakimura, N., Inaba, K., and Kawarabayashi, K.-i. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 351–359, 2014.

Staib, M. and Jegelka, S. Robust budget allocation via continuous submodular functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Syropoulos, A. Mathematics of multisets. In *Workshop on Membrane Computing*, pp. 347–358. Springer, 2000.

Tschiatschek, S., Iyer, R. K., Wei, H., and Bilmes, J. A. Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, pp. 1413–1421, 2014.

Tschiatschek, S., Djolonga, J., and Krause, A. Learning probabilistic submodular diversity models via noise contrastive estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.