Submodular-Bregman and the Lovász-Bregman Divergences with Applications: Extended Version

Rishabh Iyer
Department of Electrical Engineering
University of Washington
rkiyer@u.washington.edu

Jeff Bilmes
Department of Electrical Engineering
University of Washington
bilmes@uw.edu

Abstract

We introduce a class of discrete divergences on sets (equivalently binary vectors) that we call the submodular-Bregman divergences. We consider two kinds of submodular Bregman divergence, defined either from tight modular upper or tight modular lower bounds of a submodular function. We show that the properties of these divergences are analogous to the (standard continuous) Bregman divergence. We demonstrate how the submodular Bregman divergences generalize many useful divergences, including the weighted Hamming distance, squared weighted Hamming, weighted precision, recall, conditional mutual information, and a generalized KL-divergence on sets. We also show that the generalized Bregman divergence on the Lovász extension of a submodular function, which we call the Lovász-Bregman divergence, is a continuous extension of a submodular Bregman divergence. We point out a number of applications of the submodular Bregman and the Lovász Bregman divergences, and in particular show that a proximal algorithm defined through the submodular Bregman divergence provides a framework for many mirror-descent style algorithms related to submodular function optimization. We also show that a generalization of the k-means algorithm using the Lovász Bregman divergence is natural in clustering scenarios where ordering is important. A unique property of this algorithm is that computing the mean ordering is extremely efficient unlike other order based distance measures. Finally we provide a clustering framework for the submodular Bregman, and we derive fast algorithms for clustering sets of binary vectors (equivalently sets of sets).

1 Introduction

The Bregman divergence first appeared in the context of relaxation techniques in convex programming ([4]), and has found numerous applications as a general framework in clustering ([2]), proximal minimization ([5]) and online learning ([31]). Many of these applications are due to the nice properties of the Bregman divergence, and the fact that they are parameterized by a single convex function. They also generalize a large class of divergences on vectors. Recently Bregman divergences have also been defined between matrices ([29, 7]) and between functions ([10]).

In this paper we define a class of divergences between sets, where each divergence is parameterized by a submodular function. This can alternatively and equivalently be seen as a divergence between binary vectors in the same way that submodular functions are special cases of pseudo-Boolean functions [3]. We call this the class of *submodular Bregman divergences* (or just *submodular Bregman*) ,and in the following sections show how its properties are related to the (classical continuous) Bregman divergence. We show an interesting mathematical property of the submodular Bregman, namely that they can be defined based on either a tight modular (linear) upper bound or *alternatively* a tight modular lower bound, unlike the traditional (continuous) Bregman definable only via a tight linear lower bound.

Let V refer to a finite ground set $\{1, 2, \ldots, |V|\}$. A set function $f : 2^V \to \mathbb{R}$ is submodular if $\forall S, T \subseteq V$, $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. Submodular functions have attractive properties that make their exact or approximate optimization efficient and often practical. They naturally arise in many problems in machine learning, computer vision, economics, operations research, etc. Submodularity can be seen as a discrete counterpart to convexity and concavity ([22]) and often the problems are closely related ([1]).The link between convexity and submodularity is seen via the Lovász extension ([8, 22]) of the submodular function. Indeed, as we shall see in this paper, the connections between submodularity and convexity and concavity will help us formulate certain discrete divergences that are analogous to the Bregman divergence. We in fact show a direct connection between a submodular Bregman and a generalized Bregman divergence defined through the Lovász extension. Exploiting many of these relationships then gives us clustering algorithms for the submodular Bregman and the Lovász Bregman divergences. Further background on submodular functions may be found in the text [11].

An outline of the paper follows. We first define the different types of submodular Bregman in Section 2. We also define the Lovász Bregman divergence, and show its relation to a version of the submodular Bregman. Then in Section 3, we prove a number of properties of the submodular Bregman and show how they are related to the Bregman divergence. Finally in Section 4, we provide applications in the context of clustering and proximal methods. In particular, we show how the proximal framework of the submodular Bregman generalizes a number of mirror-descent style approximate submodular optimization algorithms. We also consider generalizations of the k-means algorithm using the Lovász Bregman divergence, and show how they can be used in clustering applications where ordering or ranking is important. We also provide an efficient class of clustering algorithms on sets of binary vectors via the submodular Bregman.

2 The Bregman and Submodular Bregman divergences

Notation: We use ϕ to refer to a convex function, f to refer to a submodular function, and $\hat{f}$ as f's Lovász extension. Lowercase characters x, y will refer to continuous vectors, while upper case characters X, Y, S will refer to sets. We will also refer to the characteristic vectors of a set X as $1_X \in \{0,1\}^V$. Note that the characteristic vector of a set X, 1_X is such that $1_X(j) = I(j \in X)$, where $I(\cdot)$ is the standard indicator function. We will refer to the ground set as V, and the cardinality of the ground set as $n = |V|$. The (regular continuous) Bregman divergence will be expressed as d_ϕ while we refer to the upper bound submodular Bregman using d^f and the lower bound submodular Bregman using d_f. A divergence on vectors and sets is formally defined as follows: Given a domain of vectors $\mathbb{S}$, a function $d : \mathbb{S} \times \mathbb{S} \to \mathbb{R}_+$ is called a *divergence* if $\forall x, y \in \mathbb{S}$, $d(x, y) \geq 0$ and $\forall x \in \mathbb{S}$, $d(x, x) = 0$. Similarly we can define the notion of a divergence on sets Given a lattice of sets $\mathcal{L}$ (recall, $\mathcal{L}$ is a lattice if $\forall X, Y \in \mathcal{L}, X \cup Y, X \cap Y \in \mathcal{L}$), a function $d : \mathcal{L} \times \mathcal{L} \to \mathbb{R}_+$ is called a divergence if $\forall X, Y \in \mathcal{L}, d(X, Y) \geq 0$ and $\forall X \in \mathcal{L}, d(X, X) = 0$. For simplicity, we consider mostly the Boolean lattice $\mathcal{L} = 2^V$ but generalizations are possible as well [11].

2.1 Bregman and Generalized Bregman divergences

The Taylor series approximation of a twice differentiable convex function provides a natural way of generating a (regular continuous) Bregman divergence ([4]). In particular the first order Taylor series approximation of a convex function is a lower bound on the function, and is linear in x for a given y and hence given a twice differentiable convex function $\phi(x)$, we can define a divergence $d_\phi : \mathbb{S} \times \mathbb{S} \to \mathbb{R}_+$ as:

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle. \tag{1}$$

For non-differentiable convex functions, we can extend equation (1) to define the generalized Bregman divergence [14, 20]. Define a subgradient map $\mathcal{H}_\phi$, which for every vector y, gives a subgradient $\mathcal{H}_\phi(y) = h_y \in \partial\phi(y)$ [14], where $\partial\phi(y)$ is the subdifferential of ϕ at y.

$$d_\phi^{\mathcal{H}_\phi}(x, y) = \phi(x) - \phi(y) - \langle \mathcal{H}_\phi(y), x - y \rangle, \forall x, y \in \mathbb{S}. \tag{2}$$

When ϕ is differentiable, then $\partial\phi(x) = \{\nabla\phi(x)\}$ and $\mathcal{H}_\phi(y) = \nabla\phi(y)$. More generally, there may be multiple distinct subgradients in the subdifferential, hence the generalized Bregman divergence is parameterized both by ϕ and the subgradient-map $\mathcal{H}_\phi$. The generalized Bregman divergences have also been defined in terms of "extreme" subgradients [28, 20].

$$d_\phi^\sharp(x, y) = \phi(x) - \phi(y) - \sigma_{\partial\phi(y)}(x - y) \quad \text{and} \quad d_\phi^\natural(x, y) = \phi(x) - \phi(y) + \sigma_{\partial\phi(y)}(y - x), \tag{3}$$

where, for a convex set C, $\sigma_C(.) \triangleq \sup_{x \in C}\langle ., x\rangle$. Then notice that we have: $\sigma_{\partial\phi(y)}(x-y) \geq \langle h_y, x-y\rangle \geq -\sigma_{\partial\phi(y)}(y-x), \forall h_y \in \partial\phi(y)$. This then implies that: $d^{\sharp}_{\phi}(x,y) \leq d^{\mathcal{H}_\phi}_{\phi}(x,y) \leq d^{\natural}_{\phi}(x,y), \forall \mathcal{H}_\phi$ which justifies their being called the extreme generalized Bregman divergences [14].

2.2 The Submodular Bregman divergences

In a similar spirit, we define a submodular Bregman divergence parameterized by a submodular function and defined as the difference between the function and its modular (sometimes called linear) bounds. Surprisingly, any submodular function has both a tight upper and lower modular bound ([17]), unlike strict convexity where only a tight first-order lower bound exists. Hence, we define two distinct forms of submodular Bregman parameterized by a submodular function and in terms of either its tight upper or tight lower bounds.

2.2.1 Lower bound form of the Submodular Bregman

Given a submodular function f, the submodular polymatroid $\mathcal{P}_f$, the corresponding base polytope $\mathcal{B}_f$ and the subdifferential $\partial_f(Y)$ (at a set Y) for a submodular function f [11] are respectively:

$$\mathcal{P}_f = \{x : x(S) \leq f(S), \forall S \subseteq V\}, \qquad \mathcal{B}_f = \mathcal{P}_f \cap \{x : x(V) = f(V)\}, \text{ and} \tag{4}$$

$$\partial_f(Y) = \{y \in \mathbb{R}^V : \forall X \subseteq V, f(Y) - y(Y) \leq f(X) - y(X)\}. \tag{5}$$

Note that here $y(S) = \sum_{j \in S} y(j)$ is a modular function. In a manner similar to the generalized Bregman divergence ([14]), we define a discrete subgradient map for a submodular function $\mathcal{H}_f$, which for every set Y, picks a subgradient $\mathcal{H}_f(Y) = h_Y \in \partial_f(Y)$. Then, given a submodular function f and a subgradient-map $\mathcal{H}_f$, the generalized lower bound submodular Bregman – which we shall henceforth call $d^{\mathcal{H}_f}_f$, is defined as:

$$d^{\mathcal{H}_f}_f(X,Y) = f(X) - f(Y) - h_Y(X) + h_Y(Y) = f(X) - f(Y) - \langle \mathcal{H}_f(Y), 1_X - 1_Y\rangle). \tag{6}$$

This form of submodular Bregman is parameterized both by the submodular function f and the subgradient map $\mathcal{H}_f$. In the sequel, we shall consider some examples below of the generalized lower bound submodular Bregman divergence, by instantiating the submodular function f and the subgradient map $\mathcal{H}_f$.

The subdifferential corresponding to a submodular function is an unbounded polyhedron [11], with a large number of possible subgradients. Its extreme points, however, are easy to find and characterize using the greedy algorithm [8]. Thus, we define a subclass of $d^{\mathcal{H}_f}_f$ with $\mathcal{H}_f$ chosen so that it picks an extreme points of $\partial_f(Y)$, which we will call the permutation based lower bound submodular Bregman, henceforth referred to with d^{Σ}_f. The extreme points of $\partial_f(Y)$ can be obtained via a greedy algorithm ([8, 11]) as follows:

Lemma 2.1. *([11], Theorem 6.11) A point y is an extreme point of $\partial_f(Y)$, if and only if there exists a chain $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_n$ with $Y = S_j$ for some j, such that $y(S_i) - y(S_{i-1}) = f(S_i) - f(S_{i-1})$.*

Let σ be a permutation of V and define $S_i = \{\sigma(1), \sigma(2), \dots, \sigma(i)\}$ as its corresponding chain. We define $\mathbf{\Sigma}_Y$ as the set of permutations σ_Y such that their corresponding chains contain Y, meaning $S_{|Y|} = Y$. Then we can define a subgradient h_{Y,σ_Y} (which is an extreme point of $\partial_f(Y)$) where:

$$\forall \sigma_Y \in \mathbf{\Sigma}_Y, \;\; h_{Y,\sigma_Y}(\sigma_Y(i)) = \begin{cases} f(S_1) & \text{if } i = 1 \\ f(S_i) - f(S_{i-1}) & \text{otherwise} \end{cases}. \tag{7}$$

In the above, $h_{Y,\sigma_Y}(Y) = f(Y)$. Hence define $\mathcal{H}^{\Sigma}_f$ as a subgradient map which picks a subgradient h_{Y,σ_Y}, for some $\Sigma(Y) = \sigma_Y \in \mathbf{\Sigma}_Y$. Here we treat Σ as a permutation operator which, for a given set Y, produces a permutation $\sigma_Y \in \mathbf{\Sigma}_Y$. Hence the above, directly provides us with a subclass, which we call the permutation based lower bound submodular Bregman and we can rewrite Eqn. (6), with the above subgradient as

$$d^{\Sigma}_f(X,Y) = f(X) - h_{Y,\sigma_Y}(X) = f(X) - \langle \mathcal{H}^{\Sigma}_f(Y), 1_X\rangle. \tag{8}$$

As can readily be seen, the d_f^{Σ} are special cases of the $d_f^{\mathcal{H}_f}$.

Similar to the extreme generalized Bregman divergence above, we can define forms of the "extreme" lower bound submodular Bregman divergences $d_f^{\sharp}(X, Y)$ and $d_f^{\natural}(X, Y)$, which provide bounds on the forms of the lower bound submodular Bregman. In order to obtain these extreme forms of submodular Bregman, we would need to compute $\max_{h \in \partial_f(Y)} \langle h, 1_X - 1_Y \rangle$ and $\min_{h \in \partial_f(Y)} \langle h, 1_X - 1_Y \rangle$. Both these expressions are linear programs over the submodular subdifferential. As we shall show below, this can be obtained easily. In order to show this, we invoke the following theorem from [11].

Lemma 2.2. *[11] For any submodular function f, $\partial_f(Y) = \partial_{f^Y}(Y) \times \partial_{f_Y}(\emptyset)$, where $f^Y(X) = f(X), \forall X \in [\emptyset, Y]$ and $f_Y(Z) = f(Z \cup Y) - f(Y), \forall Z \in [Y, V] \backslash Y$. Further define $f_{\#}^Y(X) = f(Y) - f(Y \backslash X)$. Then $\partial_f(Y) = \mathcal{P}_{f_{\#}^Y} \times \mathcal{P}_{f_Y}$, where $\mathcal{P}_{f_{\#}^Y}$ is a supermodular polyhedron, corresponding to the supermodular function $f_{\#}^Y$ on $[\emptyset, Y]$ and $\mathcal{P}_{f_Y}$ is a submodular polyhedron of f_Y on $[Y, V] \backslash Y$.*

In other words, the submodular subdifferential is an inner product between a submodular polyhedron and a supermodular polyhedron.

We are now in a position to show that similar to the submodular polyhedron, a linear program over the submodular sub-differential can be solved efficiently in certain cases. Though this result follows directly from the results above, we could not find this result in the literature. Hence we explicitely prove it here. We introduce the notion of a base sub-differential $\partial_f^{\mathcal{B}}(Y) = \partial_f(Y) \cap \mathcal{B}_f$, which is similar to the base polytope. Define $\partial_f^{\mathcal{B}}(Y) = \{y : y(X) \leq f(X), \forall X \subseteq V, y(Y) = f(Y), y(V) = f(V)\}$. Then we have the following facts about $\partial_f^{\mathcal{B}}(Y)$:

Lemma 2.3. *For a submodular function f,*

$$\partial_f^{\mathcal{B}}(Y) = \mathcal{B}_{f^Y} \times \mathcal{B}_{f_Y} \tag{9}$$

where $\mathcal{B}_{f^Y}$ is the base polytope of f^Y on $[\emptyset, Y]$ and $\mathcal{B}_{f_Y}$ is the base polytope of f_Y on $[Y, V] \backslash Y$. Further $\partial_f^{\mathcal{B}}(Y)$ is the convex combinations of the extreme points of $\partial_f(Y)$. In other words, $\partial_f^{\mathcal{B}}(Y) = conv(h_{\sigma_Y}^f, \forall \sigma_Y \in \mathbf{\Sigma}_Y)$.

The proof of the above Lemma is in Appendix A.1 Then we show the following crucial theorem.

Theorem 2.1. *Given a vector $w \in \mathbb{R}^n$ and a set Y, consider a permutation $\sigma \in \mathbf{\Sigma}_Y$ such that $w(\sigma(1)) \geq w(\sigma(2) \cdots \geq w(\sigma(|Y|))$ and $w(\sigma(|Y| + 1)) \geq \cdots \geq w(\sigma(n))$. Define $s^* \in \mathbb{R}^n$ such that $s^*(\sigma(i)) = f(S_i) - f(S_{i-1})$, for $i = [1, 2, \cdots, n]$, with $S_i = [\sigma_1, \cdots, \sigma_i]$. Then $\operatorname{argmax}_{s \in \partial_f^{\mathcal{B}}(Y)} s^\top w = s^*$. Further if additionally w is such that $w(i) \leq 0, \forall i \in Y$ and $w(i) \geq 0, \forall i \notin Y$, then $\operatorname{argmax}_{s \in \partial_f(Y)} s^\top w = s^*$.*

The proof of this theorem is in Appendix A.2. Notice that this theorem is very analogous to the greedy algorithm for the submodular polyhedron. In particular we can use exactly the same procedure, except that we individually order the elements inside Y and those outside Y, based on w.

Now define the "extreme lower bound submodular Bregman as follows:

$$d_f^{\sharp}(X, Y) = f(X) - f(Y) - \sigma_{\partial_f(Y)}(1_X - 1_Y) \tag{10}$$

The theorem above, shall play a significant role in showing that the above expression can be obtained easily for a submodular function.

Theorem 2.2. *For a submodular function f, $d_f^{\sharp}(X, Y) = f(X) + f(Y) - f(X \cap Y) - f(X \cup Y)$.*

The proof of this theorem is in Appendix A.3. Unfortunately an analogous expression for the other extreme form of the lower bound submodular Bregman will be unbounded (if maximized over the entire sub-differential). This can be verified as follows. Notice that the subdifferential is an unbounded polyhedron and correspondingly $d_f^{\natural}(X, Y)$ defined on the polyhedron is unbounded above. Let $h_Y \in \partial_f(Y) \cap \mathcal{B}_f$. Then we have that $d_f^{\mathcal{H}_f}(X, Y) = f(X) - h_Y(X)$. We then can define $h'_Y = h_Y - c1_{\bar{Y}}$ (where $\bar{Y}$ is the complement of Y) for any constant $c \geq 0$. Then with respect to the subgradient h'_Y, $d_f^{\mathcal{H}'_f}(X, Y) = f(X) - h'_Y(X) - f(Y) + h'_Y(Y) = d_f^{\mathcal{H}_f}(X, Y) + c|\bar{Y} \cap X|$.

This can be unbounded above. Hence we define $d_f^{\natural}(X,Y)$ over $\partial_f(Y) \cap \mathcal{B}_f = \partial_f^{\mathcal{B}}(Y)$. Hence define the other extreme generalized submodular Bregman as:

$$d_f^{\natural}(X,Y) = f(X) - f(Y) + \sigma_{\partial_f^{\mathcal{B}}(Y)}(1_Y - 1_X) \tag{11}$$

Then $d_f^{\natural}(X,Y)$ also has a nice representation, as we show below:

Theorem 2.3. *For a submodular function f, $d_f^{\natural}(X,Y) = f(X) - f(Y) + f(Y \backslash X) - f^{\sharp}(X \backslash Y)$, where $f^{\sharp}(A) = f(V) - f(V \backslash A)$*

The proof of this theorem is in Appendix A.4

Finally we relate the different forms of lower bound submodular Bregmans in the Lemma below:

Corollary 2.3.1. *For every $h_Y \in \partial_f(Y) \cap \mathcal{B}_f$, $d_f^{\sharp}(X,Y) \leq d_f^{\mathcal{H}_f}(X,Y) \leq d_f^{\natural}(X,Y)$. Similarly for every permutation map Σ, $d_f^{\sharp}(X,Y) \leq d_f^{\Sigma}(X,Y) \leq d_f^{\natural}(X,Y)$.*

The above corollary shows that the extreme submodular Bregman divergences give bounds for $d_f^{\mathcal{H}_f}$ and d_f^{Σ}. Further we see that $d_f^{\sharp}$ is exactly the divergence which defines the submodularity of f. Also notice that this is unlike the generalized Bregman divergences, where the "extreme" forms may not be easy to obtain in general [14]. It is easy to check that $d_f^{\sharp}(X,Y) = 0$ whenever $X \subseteq Y$ and $Y \subseteq X$. This is not surprising since in these cases the minimum value of $d_f^{\mathcal{H}_f}$ and d_f^{Σ} over $\partial_f(Y)$ is zero. Further it is possible to show just from the definition of submodularity (independently of the above theorem) that $d_f^{\sharp}(X,Y) \leq d_f^{\natural}(X,Y)$.

We show below three examples of the lower bound submodular Bregman divergence. Few more examples are shown in table 1.

Hamming and weighted Hamming distance: Recall that given binary vectors $1_X, 1_Y$, we can define the hamming distance as $d_H(X,Y) = \sum_{j=1}^n |1_X(j) - 1_Y(j)| = |X \backslash Y| + |Y \backslash X|$. Similarly for a weight vector $w \in \mathbb{R}_+^n$, $d_H^w(X,Y) = \sum_{j=1}^n w_j |1_X(j) - 1_Y(j)| = w(X \backslash Y) + w(Y \backslash X)$. This is the weighted Hamming distance. Both these distance measures are special cases of the lower bound submodular Bregman.

Let $f(X) = w(X)$ and $\mathcal{H}_f(Y) = 2.w \odot 1_Y$.

$$\begin{aligned} d_f^{\mathcal{H}_f}(X,Y) =& f(X) - f(Y) - h_Y(X) + h_Y(Y) \\ =& w(X) - w(Y) - 2w(X \cup Y) + 2w(Y) \\ =& w(X|) + w(Y) - 2w(X \cup Y) \\ =& w(X \backslash Y) + w(Y \backslash X) = d_H^w(X,Y) \end{aligned} \tag{12}$$

Substituting $w = \mathbf{1}$, gives us the Hamming distance.

Recall and weighted Recall: For sets X, Y we can define the recall divergence (note that recall is a similarity and its inverse is a distance measure) as $d_R(X,Y) = 1 - \frac{|X \cap Y|}{|Y|}$. Similarly we can define a weighted recall as: $d_R^w(X,Y) = 1 - \frac{w(X \cap Y)}{w(Y)}$, for a weight vector $w \in \mathbb{R}_+^n$.

Let $f(X) = 1$ and $\mathcal{H}(Y) = \frac{w \odot 1_Y}{w(Y)}$.

$$\begin{aligned} d_f^{\mathcal{H}_f}(X,Y) &= f(X) - f(Y) - h_Y(X) + h_Y(Y) \\ &= 1 - 1 - 2\frac{w(X \cup Y)}{w(Y)} + 1 \\ &= 1 - \frac{w(X \cap Y)}{w(Y)} = d_R^w(X,Y) \end{aligned} \tag{13}$$

Again with $w = \mathbf{1}$, we get back Recall.

Table 1: Instances of the $d_f^{\mathcal{H}_f}$

Name	Type	$d(X,Y)$	$f(X)$	h_Y
Hamming	$d_f^{\mathcal{H}_f}(X,Y)$	$\lvert X\backslash Y\rvert + \lvert Y\backslash X\rvert$	$\lvert X\rvert$	$2 \cdot 1_Y$
Weighted Hamming	$d_f^{\mathcal{H}_f}(X,Y)$	$w(X\backslash Y) + w(Y\backslash X)$	$w(X)$	$2 \cdot w \odot 1_Y$
Recall	$d_f^{\mathcal{H}_f}(X,Y)$	$1 - \frac{\lvert X\cap Y\rvert}{\lvert Y\rvert}$	1	$\frac{1_Y}{\lvert Y\rvert}$
Weighted Recall	$d_f^{\mathcal{H}_f}(X,Y)$	$1 - \frac{w(X\cap Y)}{w(Y)}$	1	$\frac{w\odot 1_Y}{w(Y)}$
$d(X,Y) = \text{AER}(Y,X;Y)$	$d_f^{\mathcal{H}_f}(X,Y)$	$1 - \frac{\lvert Y\rvert + \lvert Y\cap X\rvert}{2\lvert Y\rvert}$	$\frac{1}{2}$	$\frac{1_Y}{2\lvert Y\rvert}$
Cond. Mutual Information	$d_f^\sharp(X\cup C, Y\cup C)$	$I(\mathcal{X}_X; \mathcal{X}_Y \mid \mathcal{X}_C)$, when $X\cap Y = \emptyset$	$H(\mathcal{X}_X)$	-

Conditional Mutual Information as a special case of $d_f^\sharp(X,Y)$: Define a set function divergence $d_I(A,B) = I(X_A; X_B|X_C)$ for sets $A, B : A\cap B = \emptyset$ and a given set C. Then we have the famous equality:

$$I(\mathcal{X}_{X\backslash Y}; \mathcal{X}_{Y\backslash X}|\mathcal{X}_{X\cap Y}) = H(\mathcal{X}_X) + H(\mathcal{X}_Y) - H(\mathcal{X}_{X\cap Y}) - H(\mathcal{X}_{X\cup Y})$$
$$= d_f^\sharp(X,Y) \tag{14}$$

with the submodular function $f(X) = H(\mathcal{X}_X)$.

This is interesting since conditional mutual information $I(\mathcal{X}_{X\backslash Y}; \mathcal{X}_{Y\backslash X}|\mathcal{X}_{X\cap Y})$ can be seen as a special case of the lower bound submodular Bregman divergence.

2.2.2 The upper bound submodular Bregman

For submodular f, [26] established the following properties:

$$f(Y) \leq f(X) - \sum_{j\in X\backslash Y} f(j|X-\{j\}) + \sum_{j\in Y\backslash X} f(j|X\cap Y) \tag{15}$$

$$\text{and} \qquad f(Y) \leq f(X) - \sum_{j\in X\backslash Y} f(j|X\cup Y-\{j\}) + \sum_{j\in Y\backslash X} f(j|X), \tag{16}$$

where $f(j|X) = f(X\cup j) - f(X)$ is the gain of element j in the context of set X. In [26], it is shown that these in fact characterize submodular functions, in that a function f is a submodular function if and only if it satisfies the above bounds. Then we define two divergences, which we call the Nemhauser divergences:

$$d_\sharp^f(X,Y) \triangleq f(X) - \sum_{j\in X\backslash Y} f(j|X-\{j\}) + \sum_{j\in Y\backslash X} f(j|X\cap Y) - f(Y) \tag{17}$$

$$d_\natural^f(X,Y) \triangleq f(X) - \sum_{j\in X\backslash Y} f(j|X\cup Y-\{j\}) + \sum_{j\in Y\backslash X} f(j|X) - f(Y), \tag{18}$$

Notice that $d_\sharp^f(X,Y)$ and $d_\natural^f(X,Y)$ are valid divergences if and only if f is submodular. Similar to the approach in ([17]), we can relax the Nemhauser divergences to obtain three modular upper bound submodular Bregmans as:

$$d_1^f(X,Y) \triangleq f(X) - \sum_{j\in X\backslash Y} f(j|X-\{j\}) + \sum_{j\in Y\backslash X} f(j|\emptyset) - f(Y), \tag{19}$$

$$d_2^f(X,Y) \triangleq f(X) - \sum_{j\in X\backslash Y} f(j|V-\{j\}) + \sum_{j\in Y\backslash X} f(j|X) - f(Y). \tag{20}$$

$$d_3^f(X,Y) \triangleq f(X) - \sum_{j\in X\backslash Y} f(j|V-\{j\}) + \sum_{j\in Y\backslash X} f(j|\emptyset) - f(Y). \tag{21}$$

We call these the Nemhauser based upper-bound submodular Bregmans of, respectively, type-I, II and III. Henceforth, we shall represent andrefer to them as d_1^f, d_2^f and d_3^f and when referring to them collectively, we will use $d_{1:3}^f$. The Nemhauser divergences are analogous to the extreme divergences of the generalized Bregman divergences since they bound the Nemhauser based submodular Bregmans. Its not hard to observe th following fact:

Lemma 2.4. *Given a submodular function f, $d^f_3(X, Y) \geq d^f_1(X, Y) \geq d^f_\sharp(X, Y)$. Similarly $d^f_3(X, Y) \geq d^f_2(X, Y) \geq d^f_\natural(X, Y)$*

Similar to the generalized lower bound submodular Bregman $d_f^{\mathcal{H}_f}$, we define a generalized upper bound submodular Bregman divergence $d^f_{\mathcal{G}^f}$ in terms of any supergradient of f. Interestingly for a submodular function, we can define a superdifferential $\partial^f(X)$ at X as follows:

$$\partial^f(X) = \{x \in \mathbb{R}^V : \forall Y \subseteq V, f(X) - x(X) \geq f(Y) - x(Y)\}. \tag{22}$$

Similar to the subgradient map, we can define $\mathcal{G}^f$ as the supergradient map, which picks a supergradient from $\mathcal{G}^f(X) = g_X \in \partial^f(X)$. Given a supergradient at X, $\mathcal{G}^f(X) = g_X \in \partial^f(X)$, we can define a divergence $d^f_{\mathcal{G}^f}$, as:

$$d^f_{\mathcal{G}^f}(X, Y) = f(X) - f(Y) - g_X(X) - g_X(Y) = f(X) - f(Y) - \langle \mathcal{G}^f(X), 1_X - 1_Y \rangle \tag{23}$$

In fact, it can be shown that all three forms of $d^f_{1:3}$ are actually special cases of $d^f_{\mathcal{G}^f}$, in that they form specific supergradient maps.

Define three supergradients as follows:

$$g^1_X(j) = \begin{cases} f(j|X - j) & \text{if } j \in X \\ f(j|\emptyset) & \text{if } j \notin X \end{cases} \tag{24}$$

$$g^2_X(j) = \begin{cases} f(j|V - j) & \text{if } j \in X \\ f(j|X) & \text{if } j \notin X \end{cases} \tag{25}$$

$$g^3_X(j) = \begin{cases} f(j|V - j) & \text{if } j \in X \\ f(j|\emptyset) & \text{if } j \notin X \end{cases} \tag{26}$$

Denote the super-gradient corresponding maps $\mathcal{G}^f_1, \mathcal{G}^f_2$ and $\mathcal{G}^f_3$. Then we have the following theorem. (proof in Appendix B.1).

Theorem 2.4. *For a submodular function f, $g^1_X, g^2_X, g^3_X \in \partial^f(X)$. Correspondingly the divergences d^f_1, d^f_2 and d^f_3 are special cases of $d^f_{\mathcal{G}^f}$ with g_X being g^1_X, g^2_X and g^3_X respectively.*

Note that a convex function does not have a supergradient and hence an analogous expression does not exist for the standard Bregman divergence. However this can be seen to be akin to a form of a concave Bregman divergence (which is actually identical to the class of standard Bregman divergences, since for every concave function g, $-g$ is convex.) $d^f_{\mathcal{G}^f}$ also subsumes an interesting class of divergences for any submodular function representable as concave over modular. Consider any decomposable submodular function [27] f, representable as: $f(X) = \sum_i \lambda_i h_i(m_i(X))$, where the h_is are (not necessarily smooth) concave functions and the m_is are vectors in $\mathbb{R}^n$. Let h'_i be any supergradient of h_i. Then we define $g^{cm}_X = \sum_i \lambda_i h'_i(m_i(X)) m_i$. Further we can define a divergence defined for a concave over modular function as:

$$d^f_{cm}(X, Y) = \sum_i \lambda_i(h_i(m_i(X)) - h_i(m_i(Y)) - h_i(m_i(X))(m_i(X) - m_i(Y)) \tag{27}$$

Then we have the following Lemma. (proof in Appendix B.2).

Lemma 2.5. *Let $f(X) = \sum_i \lambda_i h_i(m_i(X))$, be a decomposable submodular function, where h_i's are concave functions and m_i's are modular. Then $g^{cm}_X \in \partial^f(X)$ and correspondingly d^f_{cm} is a special case of $d^f_{\mathcal{G}^f}$ with $g_X = g^{cm}_X$.*

We now consider below a number of examples of the upper bound submodular Bregman divergence. A complete list of examples is in Table 2.

Hamming and weighted Hamming distance: The Hamming and weighted Hamming can be shown to be special cases of the upper bound submodular Bregman as well.

Let $f(X) = -w(X)$ and $h_Y = -2.w \odot 1_Y$.

$$\begin{aligned}
d^f_{\mathcal{G}^f}(X, Y) =& f(X) - f(Y) - h_Y(X) + h_Y(Y) \\
=& - w(X) + w(Y) - 2w(X \cup Y) + 2w(X) \\
=& w(X|) + w(Y) - 2w(X \cup Y) \\
=& w(X \backslash Y) + w(Y \backslash X) = d^w_H(X, Y)
\end{aligned} \tag{28}$$

Again substituting $w =$**1**, gives us the Hamming distance

Precision and weighted Precision: Recall that the lower bound submodular Bregman gives recall and weighted recall based divergence. With the upper bound submodular Bregman, we can get precision and weighted precision. Given sets X, Y we can define the precision divergence as $d_P(X, Y) = 1 - \frac{|X \cap Y|}{|X|}$. Similarly we can define a weighted precision as: $d^w_P X, Y) = 1 - \frac{w(X \cap Y)}{w(Y)}$, for a weight vector $w \in \mathbb{R}^n_+$.

Let $f(X) = -1$ and $g_X = -\frac{w \odot 1_Y}{w(X)}$.

$$\begin{aligned}
d^f_{\mathcal{G}^f}(X, Y) &= f(X) - f(Y) - h_Y(X) + h_Y(Y) \\
&= -1 + 1 - 2\frac{w(X \cup Y)}{w(X)} + 1 \\
&= 1 - \frac{w(X \cap Y)}{w(X)} = d^w_P(X, Y)
\end{aligned} \tag{29}$$

Again substituting $w =$**1**, gives us the Precision.

Generalized KL like divergence on sets: The upper bound submodular Bregman also generalizes a divergence which looks very much like a generalized KL divergence on sets. Given sets X, Y we define a divergence $d_{KL}(X, Y) = |Y| \log \frac{|Y|}{|X|} - |Y| + |X|$. Similarly we can define a weighted generalized KL divergence as: $d^w_{KL} X, Y) = w(Y) \log \frac{w(Y)}{w(X)} - w(Y) + w(X)$, for a weight vector $w \in \mathbb{R}^n_+$.

Let $f(X) = -w(X) \log w(X)$ and $g_X = -(1 + \log w(X))w$. Then we have that:

$$\begin{aligned}
d^f_{\mathcal{G}^f}(X, Y) &= f(X) - f(Y) - g_X(X) + g_X(Y) \\
&= -w(X) \log w(X) + w(Y) \log w(Y) + w(X)(1 + \log w(X)) - (1 + \log w(X))w(Y) \\
&= w(Y) \log \frac{w(Y)}{w(X)} - w(Y) + w(X) = d^w_{KL}(X, Y)
\end{aligned} \tag{30}$$

Again substituting $w =$**1**, we get a cardinality based distance measure $d_{KL}(X, Y)$.

We next consider a viewpoint that will make the $d^f_{1:3}$ look more like divergences. We show this in the following theorem (proof in Appendix B.3).

Theorem 2.5. *Given sets X and Y, define the series $X = X_0 \subseteq X_1 \subseteq X_2 \cdots \subseteq X_k = X \cup Y$, and $Y = Y_0 \subseteq Y_1 \subseteq Y_2 \cdots \subseteq Y_l = X \cup Y$. Define $X_j \backslash X = [x_1, \cdots x_j]$ and $Y_j \backslash Y = [y_1, \cdots y_j]$, then $d^f_2(X, Y) = \sum_{j=1}^k \big[f(x_j|(X) - f(x_j|X_{j-1})\big] + \sum_{j=1}^l \big[f(y_j|Y_{j-1}) - f(y_j|V - y_j)\big]$. Similarly define the series $X = X_0 \supseteq X_1 \supseteq X_2 \cdots \supseteq X_k = X \cap Y$, and $Y = Y_0 \supseteq Y_1 \supseteq Y_2 \cdots \supseteq Y_l = X \cap Y$. Define $X \backslash X_j = [x_1, \cdots x_j]$ and $Y \backslash Y_j = [y_1, \cdots y_j]$, then $d^f_1(X, Y) = \sum_{j=1}^k \big[f(x_j|\emptyset) - f(x_j|X_{j-1})\big] + \sum_{j=1}^l \big[f(y_j|Y_{j-1}) - f(y_j|Y - y_j)\big]$. Finally $d^f_3(X, Y) = \sum_{j=1}^k \big[f(x_j|\emptyset) - f(x_j|X_{j-1})\big] + \big[f(y_j|Y_{j-1}) - f(y_j|V - y_j)\big]$.*

2.3 The Lovász Bregman divergence

The Lovász extension ([22]) offers a natural connection between submodularity and convexity. The Lovász extension is a non-smooth convex function, and hence we can define a generalized Bregman

Table 2: Instances of $d^f_{\mathcal{G}^f}$

Name	$d^f_{\mathcal{G}^f}(X,Y)$	$f(X)$	g_X
Hamming	$\lvert X\backslash Y\rvert + \lvert Y\backslash X\rvert$	$\lvert X\rvert$	$2 \cdot 1_Y$
Weighted Hamming	$w(X\backslash Y) + w(Y\backslash X)$	$-w(X)$	$-2 \cdot w \odot 1_X$
Precision	$1 - \frac{\lvert X\cap Y\rvert}{\lvert X\rvert}$	-1	$-\frac{1_X}{\lvert X\rvert}$
Weighted Precision	$1 - \frac{w(X\cap Y)}{w(X)}$	-1	$-\frac{w\odot 1_X}{\lvert X\rvert}$
Itakura-Saito like	$\frac{\lvert Y\rvert}{\lvert X\rvert} - log\frac{\lvert Y\rvert}{\lvert X\rvert} - 1$	$log\lvert X\rvert$	$\frac{1}{\lvert X\rvert}$
Weighted Itakura-Saito like	$\frac{w(Y)}{w(X)} - \log \frac{w(Y)}{w(X)} - 1$	$\log w(X)$	$\frac{w}{w(X)}$
Generalized KL like divergence	$\lvert Y\rvert \log \frac{\lvert Y\rvert}{\lvert X\rvert} - \lvert Y\rvert + \lvert X\rvert$	$-\lvert X\rvert \log \lvert X\rvert$	$-(1 + \log \lvert X\rvert)1$
Weight. Gen. KL divergence	$w(Y) \log \frac{w(Y)}{w(X)} - w(Y) + w(X)$	$-w(X) \log w(X)$	$-w(1 + \log w(X))$
-	$e^{\lvert Y\rvert} - e^{\lvert X\rvert} - e^{\lvert X\rvert}(\lvert Y\rvert - \lvert X\rvert)$	$-e^{\lvert X\rvert}$	$1e^{\lvert X\rvert}$
Cut based - 1	$2\lvert\tau(X\backslash Y)\rvert + 2\lvert\tau(Y\backslash X)\rvert + 2\lvert\tau(X \cap Y, \bar{X} \cap Y)\rvert$	$\lvert\delta(X)\rvert$	Eqn. (47)
Cut based - 2	$2\lvert\tau(X\backslash Y)\rvert + 2\lvert\tau(Y\backslash X)\rvert + 2\lvert\tau(\bar{X} \cap \bar{Y}, X \cap \bar{Y})\rvert$	$\lvert\delta(X)\rvert$	Eqn. (49)

divergence ([14, 20]) which has a number of properties and applications analogous to the Bregman divergence. Recall that the generalized Bregman divergence corresponding to a convex function ϕ is parameterized by the choice of the subgradient map $\mathcal{H}_\phi$. The Lovász extension of a submodular function has a very interesting set of subgradients, which have a particularly nice structure in that there is a very simple way of obtaining them [8].

For simplicity, we define the Lovász Bregman divergence on vectors $x, y \in [0,1]^n$. Then given a vector y, define a permutation σ_y such that $y[\sigma_y(1)] \geq y[\sigma_y(2)] \geq \cdots \geq y[\sigma_y(n)]$ and define $Y_k = \{\sigma_y(1), \cdots, \sigma_y(k)\}$. The Lovász extension ([8, 22]) is defined as: $\hat{f}(y) = \sum_{k=1}^n y[\sigma_y(k)] f(\sigma_y(k)|Y_{k-1})$. For each point y, we can define a subdifferential $\partial\hat{f}(y)$, which has a particularly nice form [11]: for any point $y \in [0,1]^n$, $\partial\hat{f}(y) = \cap\{\partial_f(Y_i)|i = 1, 2\cdots, n\}$. This naturally defines a generalized Bregman divergence $d_{\hat{f}}^{\mathcal{H}_{\hat{f}}}$ of the Lovász extension, parameterized by a subgradient map $\mathcal{H}_{\hat{f}}$, which we can define as:

$$d_{\hat{f}}^{\mathcal{H}_{\hat{f}}}(x,y) = \hat{f}(x) - \hat{f}(y) - \langle h_y, x - y\rangle, \text{ for some } h_y = H_{\hat{f}}(y) \in \partial\hat{f}(y). \tag{31}$$

We can also define specific subgradients of $\hat{f}$ at y as h_{y,σ_y}, with $h_{y,\sigma_y}(\sigma_y(k)) = f(Y_k) - f(Y_{k-1}), \forall k$ [22]. These subgradients are really the extreme points of the submodular polyhedron. Then define the Lovász Bregman divergence $d_{\hat{f}}$ as the Bregman divergence of $\hat{f}$ and the subgradient h_{y,σ_y} ,which can be obtained as follows:

$$\langle y, h_{y,\sigma_y}\rangle = \sum_i y[\sigma_y(i)] h_{y,\sigma_y}[\sigma_y(i)] = \sum_i y[\sigma_y(i)](f(Y_i) - f(Y_{i-1})) = \hat{f}(y) \tag{32}$$

Thus we have that:

$$d_{\hat{f}}(x,y) = \hat{f}(x) - \langle h_{y,\sigma_y}, x\rangle = \langle x, h_{x,\sigma_x} - h_{y,\sigma_y}\rangle \tag{33}$$

Note that if the vector y is totally ordered (no two elements are equal to each other), the subgradient of $\hat{f}$ and the corresponding permutation σ_y at y will actually be unique. When the vector is not totally ordered, we can consider σ_y as a permutation operator which defines a valid and consistent total ordering for every vector y, and we can then define the Bregman divergence in terms of it. Note also that the points with no total ordering in the interior of the hypercube is of measure zero. Hence for simplicity we just refer to the Lovász Bregman divergence as $d_{\hat{f}}$. The Lovász Bregman divergence is closely related to the lower bound submodular Bregman, as we show below.

Lemma 2.6. *The Lovász Bregman divergences are an extension of the lower bound submodular Bregman, over the interior of the hypercube. Further the Lovász Bregman divergence can be expressed as $d_{\hat{f}}(x,y) = \langle x, h_{x,\sigma_x} - h_{y,\sigma_y}\rangle$, and hence depends only x, the permutation σ_x and the permutation of $y(\sigma_y)$, but is independent of the values of y.*

Further this relationship can be made more precise as we show in the theorem below. The proof of this theorem is in Appendix C.

Theorem 2.6. *The lower bound submodular Bregman is closely related to the Lovász Bregman divergence in the following ways (Assume $X_0 \subseteq X_1 \subseteq \cdots \subseteq X_n$ and $Y_0 \subseteq Y_1 \subseteq \cdots \subseteq Y_n$ are the chains corresponding to x and y):*

- *If x and y are vertices of the hypercube, then d_f^{Σ} is exactly $\mathrm{d}_{\hat{f}}$, if the chosen subgradients of both are the same.*
- *If x is a point in the interior of the hypercube and $y = 1_Y$ is a vertex of the hypercube, then $\mathrm{d}_{\hat{f}}(x, y) = (1 - x(\sigma_X(n)))\mathrm{d}_f^{\mathcal{H}_f}(X_0, Y) + \sum_{i=1}^{n-1}(x(\sigma_X(i)) - x(\sigma_X(i+1))\mathrm{d}_f^{\mathcal{H}_f}(X_i, Y) + x(\sigma_X(n))\mathrm{d}_f^{\mathcal{H}_f}(X_n, Y)$, for $\mathrm{d}_f^{\mathcal{H}_f}$, and $\mathrm{d}_{\hat{f}}$, as long as the chosen subgradient of $\mathrm{d}_{\hat{f}}$ (at $y = 1_Y$) is $\mathcal{H}_f(Y)$.*
- *Let y be a point in the interior of the hypercube and $x = 1_X$ is a vertex of the hypercube. Then $\mathrm{d}_{\hat{f}}(x, y) = (1 - y(\sigma_Y(n)))\mathrm{d}_f^{\Sigma}(X, Y_0) + \sum_{i=1}^{n-1}(y(\sigma_Y(i)) - y(\sigma_Y(i+1))\mathrm{d}_f^{\Sigma}(X, Y_i) + y(\sigma_Y(n))\mathrm{d}_f^{\Sigma}(X, Y_n)$, where the permutation map Σ satisfies $\Sigma(Y_i) = \sigma_y, \forall i$.*
- *Finally, when both x and y are vectors within the hypercube, $\mathrm{d}_{\hat{f}}(x, y)$ is a convex combination of $\mathrm{d}_f^{\Sigma}(X_i, Y_j)$, where X_i and Y_j belong to the chain of sets corresponding to vectors x and y and the permutation map Σ satisfies $\Sigma(Y_i) = \sigma_y, \forall i$.*

2.4 Extension of the upper bound submodular Bregman

In this section, we introduce a partial extention of the upper bound submodular Bregman in terms of the Lovász extention. This is unlike the Lovász Bregman divergence, since the concave extention of a submodular function is NP hard to compute [30]. However as we show below, we define $d^{\hat{f}}(x, y)$ as a partial extension of the generalized upper bound submodular Bregman to the interior of the hypercube through the Lovász extension. We however restrict x be a vertex of the hypercube, while y is allowed to be a point in the interior of the hypercube, i.e $y \in [0, 1]^n$. Then we have

$$d^{\hat{f}}_{\mathcal{G}^f}(1_X, y) = \hat{f}(1_X) - \hat{f}(y) - \langle \mathcal{G}^f(X), 1_X - y \rangle \tag{34}$$

For simplicity we just represent this as $d^{\hat{f}}$. It is evident that $d^{\hat{f}}(1_X, 1_Y) = d^f_{\mathcal{G}^f}(X, Y)$. However we can relate $d^{\hat{f}}(1_X, y)$ to $d^f_{\mathcal{G}^f}$, for any $y \in [0, 1]^n$ (proof in Appendix C).

Theorem 2.7. *$d^{\hat{f}}(1_X, y)$ is a valid divergence from $\{0, 1\}^n \times [0, 1]^n \to \mathbb{R}_+$. Further $d^{\hat{f}}(1_X, y) = (1 - y(\sigma(n)))d^f_{\mathcal{G}^f}(X, \emptyset) + \sum_{i=1}^{n-1}(y(\sigma(i)) - y(\sigma(i+1))d^f_{\mathcal{G}^f}(X, Y_i) + y(\sigma(n))d^f_{\mathcal{G}^f}(X, Y_n)$, as long as the chosen supergradinet of $d^{\hat{f}}$ is $\mathcal{G}^f(X)$.*

We shall use this extention in providing effecient algorithms for clustering the submodular Bregman divergences.

3 Properties of the submodular Bregman and Lovász Bregman divergences

In this section, we investigate some of the properties of the submodular Bregman and Lovász Bregman divergences which make these divergences interesting, both from theorotical and practical viewpoints.

3.1 The submodular Bregman

In the following, we list a number of interestig properties of the submodular Bregman divergences. A highlight of these properties is as follows. All forms of the submodular Bregman divergences are non-negative, and hence they are valid divergences. Also the lower bound submodular Bregman is submodular in X for a given Y, while the upper bound submodular Bregman is supermodular in Y for a given X. A direct consequence of this is that problems involving optimization in X or Y (for example in finding the discrete representatives in a discrete k-means like application which we consider in the later part of this paper), can be performed either exactly or approximately in polynomial time. In addition to these the forms of the submodular Bregman divergence also satisfy interesting properties like a characterization of equivalence classes, a form of set separation, a generalized triangle inequality over sets and a form of both Fenchel and submodular duality. Finally the generalized submodular Bregman divergence has an interesting alternate characterization, which shows that they can potentially subsume a large number of discrete divergences. In particular, a divergence d is of the form $d_f^{\mathcal{H}_f}$ *iff* for any sets $A, B \subseteq V$, the set function $f_A(X) = d(X, A)$ is submodular in X and the set function $d(X, A) - d(X, B)$ is modular in X. Similarly a divergence d is of the form $d_{\mathcal{G}^f}^f$ *iff*, for any set $A, B \subseteq V$, the set function $f_A(Y) = d(A, Y)$ is supermodular in Y and the set function $d(A, Y) - d(B, Y)$ is modular in Y. These facts show that the generalized Bregman divergences are potentially a very large class of divergences while Tables 2 and 1 provide just a few of them.

We now list and prove some interesting properties about the various forms of the submodular Bregman divergences, and compare it with the corresponding property of its continuous counterpart. We have already seen some close correspondences while defining them, but now we make the relations more formal. Note that any property true for $d_f^{\mathcal{H}_f}$ will also be obeyed by the d_f^{Σ}, and any property is true for $d_{\mathcal{G}^f}^f$ will also be true for any of the special cases $d_{1:3}^f$ and d_{cm}^f .

1) Nonnegativity: All forms of submodular Bregmans are divergences, in that $d(X, Y) \geq 0, \forall X, Y \subseteq V$ and $d(X, X) = 0$. The property "$d(X, Y) = 0$ iff $X = Y$" we refer to as the "iff non-negativity property." In general, the submodular Bregman do not satisfy this property. However in certain cases they do satisfy this property:

Theorem 3.1. *Given a strictly submodular function f, $d_f^{\mathcal{H}_f}$ (and $d_{\mathcal{G}^f}^f$) satisfy the iff non-negativity property if the subgradient h_Y (respectively supergradient g_X) lie in the strict interior of the subdifferential (respectively superdifferential). Correspondingly, for a strictly submodular functions f, both d_{cm}^f and d_3^f satisfy the iff non-negativity property.*

The above theorem can directly be verified from the definitions of the different forms of the submodular Bregman divergences.

2) Submodularity and convexity: The Bregman divergence (and the generalized Bregman divergence) is convex in x for fixed y, but not necessarily convex in y for fixed x. Similarly the Lovász Bregman divergence is convex in x for a given y. Correspondingly, $d_f^{\mathcal{H}_f}(X, Y)$ is submodular in X for fixed Y and $d_{\mathcal{G}^f}^f(X, Y)$ is supermodular in Y under fixed X. Further, d_1^f and d_2^f can be naturally expressed as a difference between submodular functions in X under certain conditions on the function f (stated and proved in Appendix-D.1) and $d_3^f(X, Y)$ is submodular in X for a given Y and supermodular in Y for a given X.

3) Linearity: The Bregman divergence is a linear operator in ϕ. In the theorem below, we give similar properties for the some forms of the submodular Bregman (proof in Appendix D.2).

Theorem 3.2. *For a submodular function f, d_f^{Σ}, $d_{1:3}^f$, d_{cm}^f and $d_{\hat{f}}$ are linear operators in f.*

4) Equivalence classes: The Bregman divergence of functions which differ only in an affine term are equal. A similar property holds for certain forms of the submodular Bregman, as shown in the following (proof in Appendix D.2).

Theorem 3.3. *Let $m(X)$ be a modular function. Then d_f^{Σ} satisfies: $d_f^{\Sigma}(X, Y) = d_{f+m}^{\Sigma}(X, Y)$. Similarly $d_{1:3}^f$ and d_{cm}^f , satisfy the property that: $d_{1:3}^f(X, Y) = d_{1:3}^{f+m}(X, Y)$ and $d_{cm}^f(X, Y) = d_{cm}^{f+m}(X, Y)$.*

Thus, we need use only polymatroidal rank functions f within d_f^{Σ}, $d_{1:3}^f$, d_{cm}^f and $d_{\hat{f}}$[6].

5) Set Separation: The Bregman divergence has the property of linear separation — the set of points x equidistant to two fixed points μ_1 and μ_2 (i.e., $\{x : d_\phi(x, \mu_1) = d_\phi(x, \mu_2)\}$) comprise a hyperplane. The theorem below shows that the upper and lower bound submodular Bregmans have similar properties (the proofs follow immediately from the definitions).

Theorem 3.4. *A set X which is equidistant to two sets Y_1 and Y_2 (for the generalized lower bound case $d_f^{\mathcal{H}_f}(X, Y_1) = d_f^{\mathcal{H}_f}(X, Y_2)$ and generalized upper bound cases $d_{\mathcal{G}^f}^f(Y_1, X) = d_{\mathcal{G}^f}^f(Y_2, X)$), must satisfy an equation of the type $m(X) = c(Y_1, Y_2)$, where $m : 2^V \to \mathbb{R}$ is a modular function and $c(Y_1, Y_2)$ is a constant dependent on Y_1 and Y_2.*

While the classical Bregman divergence has this property only when x is the first argument, the different forms of submodular Bregmans have it for alternative arguments due to their complementary nature.

6) Generalized Triangle Inequality: The Bregman divergence doesn't follow the triangle inequality in general. However the following generalized Pythagorean theorem holds:

$$d_\phi(x_1, x_3) = d_\phi(x_1, x_2) + d_\phi(x_2, x_3) - \langle x_1 - x_2, \nabla\phi(x_3) - \nabla\phi(x_2)\rangle \tag{35}$$

We next derive similar relationships for the submodular Bregman (proof in Appendix D.3).

Theorem 3.5. *The generalized triangle inequality for $d_f^{\mathcal{H}_f}$ is:*

$$d_f^{\mathcal{H}_f}(X_1, X_3) = d_f^{\mathcal{H}_f}(X_1, X_2) + d_f^{\mathcal{H}_f}(X_2, X_3) - \langle 1_{X_1} - 1_{X_2}, h_{X_3} - h_{X_2}\rangle \tag{36}$$

Similarly we can give a generalized triangle inequality for $d_{\mathcal{G}^f}^f$ as:

$$d_{\mathcal{G}^f}^f(X_1, X_3) = d_{\mathcal{G}^f}^f(X_1, X_2) + d_{\mathcal{G}^f}^f(X_2, X_3) - \langle 1_{X_2} - 1_{X_3}, g_{X_1} - g_{X_2}\rangle \tag{37}$$

Interestingly, the $d_{1:3}^f$ in certain cases satisfy the triangle inequality (proof also in Appendix D.3).

Theorem 3.6. *For sets X, Y, Z, if $X \subseteq Y$, then we have: $d_1^f(X, Y) + d_1^f(Y, Z) \geq d_1^f(X, Z)$. Similarly if $Y \subseteq X$, then $d_2^f(X, Y) + d_2^f(Y, Z) \geq d_2^f(X, Z)$. Further d_3^f always satisfies the triangle inequality, in that $\forall X, Y, Z \subseteq V, d_3^f(X, Y) + d_3^f(Y, Z) \geq d_3^f(X, Z)$.*

7) Fenchel conjugate divergence: $d_f^{\mathcal{H}_f}$ also enjoys a nice duality relationship with respect to the Fenchel conjugate of a submodular function [11]. The Fenchel conjugate of a submodular function is a convex function defined as: $f^*(y) = \max\{y(X) - f(X) \mid X \subseteq V\}, (y \in \mathbb{R}^V)$. Noting that f^* is convex, we can define $\partial_2 f^*(y)$ as the binary subdifferential [11] of f^* at y defined as $\partial_2 f^*(y) = \{Y \subseteq E \mid \forall x \in \mathbb{R}^V, x(Y) - y(Y) \leq f^*(x) - f^*(y)\}$. f^* may not in general be a smooth convex function, but we can still define a generalized Bregman divergence with respect to f^*, for binary subgradient $Y \in \partial_2 f^*(y)$, as:

$$d_{f^*,Y}(x, y) = f^*(x) - f^*(y) - x(Y) + y(Y), \text{ for } Y \in \partial_2 f^*(y) \tag{38}$$

An interesting result from [11] states that $x \in \partial_f(X)$ iff $X \in \partial_2 f^*(x)$. This yields the following theorem (proof in Appendix-D.4).

Theorem 3.7. *Let $h_X \in \partial_f(X)$, $h_Y \in \partial_f(Y)$ be two vectors, with h_Y being the subgradient at Y defining $d_f^{\mathcal{H}_f}(X, Y)$. Then $d_f^{\mathcal{H}_f}(X, Y) = d_{f^*}(h_Y, h_X)$.*

8) Submodular Dual divergence The $d_{1:3}^f$ have an interesting property related to submodular duality. A dual function of a submodular function [11] is defined as $f^d(X) = f(V) - f(V - X)$. It is known that this dual is a supermodular function, and hence we define a submodular dual as: $f^{\#}(X) = -f^d(X)$. Then we have the following theorem.

Theorem 3.8. *For a submodular function f, $d_1^{f^{\#}}(X, Y) = d_2^f(V - X, V - Y)$. Similarly, $d_2^{f^{\#}}(X, Y) = d_1^f(V - X, V - Y)$ and $d_3^{f^{\#}}(X, Y) = d_3^f(V - X, V - Y)$.*

The proof is in Appendix-D.4. This property is interesting, since it connects the (Nemhauser) upper bound based submodular Bregmans.

9) Necessary and sufficient conditions: The necessary and sufficient conditions for a divergence d to represent a Bregman divergence is that for any vector a, the function $\phi_a(x) = d(x, a)$ is strictly convex and differentiable, and $d(x, y) = d_{\phi_a}(x, y)$. We can similarly define necessary and sufficient conditions for a divergence d to represent the submodular Bregman, which we state in the form of the following theorem (proof in Appendix D.5).

Theorem 3.9. *The following are the necessary and sufficient conditions for a divergence d to be an instance of a submodular Bregman.*

(a): *A divergence d is of the form $d_f^{\mathcal{H}_f}$* iff *for any sets $A, B \subseteq V$, the set function $f_A(X) = d(X, A)$ is submodular in X and the set function $d(X, A) - d(X, B)$ is modular in X.*

(b): *A divergence* d *is of the form* $\mathrm{d}^f_{\mathcal{G}^f}$ iff, *for any set* $A, B \subseteq V$, *the set function* $f_A(Y) = \mathrm{d}(A, Y)$ *is supermodular in* Y *and the set function* $\mathrm{d}(A, Y) - \mathrm{d}(B, Y)$ *is modular in* Y.

(c): *A divergence* d *is of the form* d^{Σ}_f iff, *for any set* A, *the function* $f_A(X) = \mathrm{d}(X, A)$, *is submodular in* X, *and for every* Y, *there exists a permutation* σ *of* V *such that* $\mathrm{d}(X, Y) = \mathrm{d}_{f_A, \sigma}(X, Y)$.

(d): *A divergence* d *is of the form* $\mathrm{d}^f_{1:3}$ iff, *for any set* A, *the function* $f_A(Y) = \mathrm{d}(A, Y)$ *is supermodular in* Y, *and* $\mathrm{d}(X, Y) = \mathrm{d}^{-f_A}_{1:3}(X, Y)$, *for one of the three Nemhauser based upper bound submodular Bregman.*

3.2 The Lovász Bregman divergence

The Lovász Bregman divergence also has a number of very interesting properties. In particular, notice that the Lovász Bregman divergences are also non-negative (and hence valid divergences). In addition, they are convex in their first argument for a given second argument. Additionally they also satisfy the property of linearity, i.e given submodular functions $f_1, f_2, d_{f_1 \hat{+} f_2}(x, y) = d_{\hat{f}_1}(x, y) + d_{\hat{f}_2}(x, y)$. Further, since it is a form of generalized Bregman divergence, a lot of the properties of generalized Bregman divergences extend to the Lovász Bregman divergence as well [28, 14].

In addition, the Lovász Bregman divergences satisfy a number of other interesting properties. Notable amongst these is the fact that it has an interesting property related to permutations.

Theorem 3.10. *Given a submodular function whose polyhedron contains all possible extreme points (e.g.,* $f(X) = \sqrt{|X|}$*),* $d_{\hat{f}}(x, y) = 0$ *if and only if* $\sigma_x = \sigma_y$*.*

Proof. The proof of this theorem follows from standard notions of the submodular polyhedron, and the definition of the Lovász Bregman divergence. Recall Eqn. (39), and it follows that $x \neq 0$, $d_{\hat{f}}(x, y) = 0$, iff $h_{x,\sigma_x} = h_{y,\sigma_y}$. Further from the definition of h and the fact that the function f has all possible extreme points, corresponding to every permutation σ we have a unique extreme point. Hence proved. □

Hence the Lovász Bregman divergence can be seen as a divergence between the permutations. While a number of distance measures capture the notion of a distance amongst orderings [19], the Lovász Bregman divergences has a unique feature not present in these distance measures. The Lovász Bregman divergences not only capture the distance between σ_x and σ_y, but also weighs it with the value of x, thus giving preference to the values and not just the orderings. Hence it can be seen as a divergence between a score x and a permutation σ_y, and hence we shall also represent it as $d_{\hat{f}}(x, y) = d_{\hat{f}}(x||\sigma_y) = \langle x, h_{x,\sigma_x} - h_{x,\sigma_y} \rangle$. Correspondingly, given a collection of scores, it also measures how confident the scores are about the ordering. For example given two scores x and y with the same orderings such that the values of x are nearly equal (low confidence), while the values of y have large differences, the distance to any other permutation will be more for y than x. This property intuitively desirable in a permutation based divergence. Finally, as we shall see the Lovász Bregman divergences are easily amenable to k-means style alternating minimization algorithms for clustering ranked data, a process that is typically difficult using other permutation-based distances.

4 Applications

In this section, we show the utility of the submodular Bregman and Lovász Bregman divergences by considering some practical applications in machine learning and optimization. The first application is that of proximal algorithms which generalize several mirror descent algorithms. As a second application, we motivate the use of the Lovász Bregman divergence as a natural choice in clustering where the order is important. Finally we provide a clustering framework for the submodular Bregman, and we derive fast algorithms for clustering sets of binary vectors (equivalently sets of sets).

4.1 A proximal framework for the submodular Bregman divergence

The Bregman divergence has some nice properties related to a proximal method. In particular ([5]), let ψ be a convex function that is hard to optimize, but suppose the function $\psi(x) + \lambda d_\phi(x, y)$ is easy to optimize for a given fixed y. Then a proximal algorithm, which starts with a particular x^0 and updates at every iteration $x^{t+1} = \text{argmax}_x \psi(x) + \lambda d_\phi(x, x^t)$, is bound to converge to the global minima. In fact the standard proximal procedures used ϕ as the quadratic function which gives the Euclidean distance.

We define a similar framework for the submodular Bregmans. Consider a set function F, and an underlying combinatorial constraint $\mathcal{S}$. Optimizing this set function may not be easy — e.g., if $\mathcal{S}$ is the constraint that X be a graph-cut, then this optimization problem is NP hard even if F is submodular ([17]). Consider now a divergence $d(X, Y)$ that can be either an upper or lower bound submodular Bregman. Note, the combinatorial constraints $\mathcal{S}$ are the discrete analogs of the convex set projection in the proximal method. We offer a proximal minimization algorithm (Algorithm 1) in a spirit similar to [5]. We have the following theorem which guarantees that the solution is monotonically decreasing over iterations.

Algorithm 1: Proximal Minimization Algorithm

```
$X^0 = \emptyset$
while until convergence do
    $X^{t+1} := \text{argmin}_{X \in \mathcal{S}} F(X) + \lambda d(X, X^t)$
    $t \leftarrow t + 1$
```

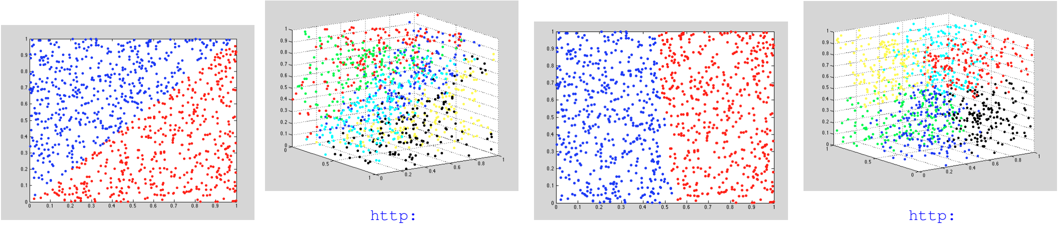
Theorem 4.1. *Consider the proximal minimization algorithm. Then the values of $F(X^t)$ are non-increasing with t (i.e., $F(X^0) \geq F(X^1) \geq F(X^2) \geq \dots$). Further since F is a set function, it is finite, and hence the algorithm is guaranteed to reach a local minimum.*

Proof. Observe from the algorithm that: $F(X^{v+1}) + d(X^{v+1}, X^v) \leq F(X^v) \Rightarrow F(X^{v+1}) \leq F(X^v)$, since $d(X_1^{v+1}, X^v) \geq 0$. Hence we have that $F(X^{v+1}) \leq F(X^v)$. □

Interestingly, a number of approximate optimization problems considered in the past turn out to be special cases of the proximal framework. We analyze this in detail in [15], and hence provide only a summary of the results below:

Minimizing the difference between submodular (DS) functions: Consider the case where $F(X) = f(X) - g(X)$ is a difference between two submodular functions f and g. This problem is known to be NP hard and even NP hard to approximate [25, 13]. However there are a number of heuristic algorithms which have been shown to perform well in practice [25, 13]. Consider first: $d(X, X^t) = d_g^{\Sigma_t}(X, X^t)$ (for some appropriate schedule Σ_t of permutations), with $\lambda = 1$ and $\mathcal{S} = 2^V$. Correspondingly at every iteration we have: $X^{t+1} \in \text{argmin}_X \left(f(X) - g(X) + d_{g,\sigma_{X^t}}(X, X^t)\right) = \text{argmin}_X f(X) - h^g_{X^t}(X)$, where $h^g_{X^t,\sigma_{X^t}}(X)$ refers to the modular lower bound of g at X^t. Thus at every iteration we minimize a submodular function, a process which is the submodular-supermodular (sub-sup) procedure ([25]). Moreover, we can define $d(X, X^t) = d^f_{1:3}(X^t, X)$, again with $\lambda = 1$ and $\mathcal{S} = 2^V$. Then at every iteration, we have: $X^{t+1} \in \text{argmin}_X \Big(f(X) - g(X) + d^f_{1:3}(X^t, X)\Big) = \text{argmin}_X m^f_{X^t}(X) - g(X)$, where $m^f_{X^t}(X)$ refers to one of the modular upper bounds of g at X^v. Thus every iteration is a submodular function maximization, which is exactly the supermodular-submodular (sup-sub) procedure [13]. Finally defining $d(X, X^t) = d^f_{1:3}(X^t, X) + d_g^{\Sigma_t}(X, X^t)$, we get the modular-modular (mod-mod) procedure [13]. Further, the sup-sub and mod-mod procedures can be used with more complicated constraints like cardinality, matroid and knapsack constraints while the mod-mod algorithm can be extended with even combinatorial constraints like the family of cuts, spanning trees, shortest paths, covers, matchings, etc. [13]

Submodular function minimization: Algorithm 1 also generalizes a number of approximate submodular minimization algorithms. If F is a submodular function and the underlying constraints $\mathcal{S}$ represent the family of cuts, then we obtain the cooperative cut problem ([17], [16]) and one of the algorithms developed in ([17]) is a special case of Algorithm 1 with $F = f\ \lambda = 1$, $d(X, X^t) = d^f_2(X^t, X)$, and $\mathcal{S}$ representing the family of cuts. In this case at every iteration is a standard graph-cut problem which is relatively easy. If $\mathcal{S} = 2^V$ above, we get a form of the approximate submodular minimization algorithm suggested for arbitrary (non-graph representable) submodular functions ([18]). The proximal minimization algorithm also generalizes three submodular function minimization algorithms IMA-I, II and III, described in detail in [15] again with $\lambda = 1, \mathcal{S} = 2^V$ and $d(X, X^t) = d^f_1(X^t, X), d^f_2(X^t, X)$ and $d(X, X^t) = d^f_3(X^t, X)$ respectively. These algorithms are similar to the greedy algorithm for submodular maximization [26]. Interestingly these algorithms provide bounds to the lattice of minimizers of the submodular functions. It is known [1] that the sets $A = \{j : f(j|\emptyset) < 0\}, B = \{j : f(j|V - \{j\}) > 0$ are such that, for every minimizer X^*, $A \subseteq X^* \subseteq B$. Thus the lattice formed with A and B defined as the join and meet respectively,



http://youtu.be/kfEnLOmvEVc

http://youtu.be/IqRhemUg14I

Figure 1: Results of k-means clustering using the Lovász Bregman divergence (two plots on the left) and the Euclidean distance (two plots on the right). URLs above link to videos.

gives a bound on the minimizers, and we can restrict the submodular minimization algorithms to this lattice. However using $d = d_3^f$ as a regularizer (which is IMA-III) and starting with $X^0 = \emptyset$ and $X^0 = V$, we get the sets A and B [15] respectively from Algorithm 1. With versions of algorithm 1 with $d = d_1^f$ and $d = d_2^f$, and starting respectively from $X^0 = \emptyset$ and $X^0 = V$, we get sets that provide a tighter bound on the lattice of minimizers than the one obtained with A and B. Further these algorithms also provide improved bounds in the context of monotone submodular minimization subject to combinatorial constraints. In particular, these algorithms provide bounds which are better than $\frac{1}{\nu}$, where ν is a parameter related to the curvature of the submodular function. Hence when the parameter ν is a constant, these bounds are constant factor guarantees, which contrasts the $O(n)$ bounds for most of these problems. As an example, for monotone submodular minimization subject to a spanning tree, perfect matching or edge cover constraints, we obtain an improved bound of $O(\frac{n}{1+(n-1)\nu})$, which matches the lower bound for these problems for the class of submodular functions with curvature ν. For a more elaborate and detailed discussion related to this, refer to [15].

Submodular function maximization: If f is a submodular function, then using $d(X, X^v) = d_f^{\Sigma_v}(X, X^v)$ forms an iterative algorithm for maximizing the modular lower bound of a submodular function. This algorithm then generalizes a algorithms number of unconstrained submodular maximization and constrained submodular maximization, in that by an appropriate schedule of Σ_v we can obtain these algorithms. Notable amongst them is a $\frac{1}{2}$ approximate algorithm and a $1 - \frac{1}{e}$ approximation algorithm for unconstrained and cardinality constrained submodular maximization respectively. Further, similar to the minimization case, we get improved curvature related bounds for monotone submodular maximization subject to cardinality and matroid constraints. For a complete list of algorithms (and results) generalized by this and a much detailed description, refer to [15].

We point out that the proximal framework provided above, is very broad and can be used for a vast class of optimization problems. In this section however, we have only considered a few special cases of this framework. Notice that akin to the proximal framework of [5], this framework will be useful only if the function $F(X) + \lambda d(X, X^t)$ for an appropriate choice of a submodular Bregman, is easier than minimizing F directly. In the above, we considered special cases, which make $F(X) + \lambda d(X, X^t)$ either modular, submodular or supermodular, thus making every step of the algorithm optimal (or approximately optimal). On the other hand, there could be other cases as well. For example, the performance of most approximate algorithms for constrained submodular minimization or maximization, depend on the curvature ν of a submodular function [15]. In particular, if F is monotone but too highly curved (for example a matroid rank function), it might be the case that adding the regularizer, preserves the monotonicity (and submodularity, which is possible if f is chosen to be an appropriate monotone submodular function), improves the curvature at every iteration, thus improving the guarantees of every inner iteration. These are some interesting open questions, which could be investigated in future work.

4.2 Clustering framework with the Lovász Bregman divergence

In this section we investigate a clustering framework similar to [2], using the Lovász Bregman divergence and show how this is natural for a number of applications. Recall that the Lovász Bregman divergence can be written as:

$$d_{\hat{f}}(x, y) = \hat{f}(x) - \langle h_{y,\sigma_y}, x \rangle = \langle x, h_{x,\sigma_x} - h_{y,\sigma_y} \rangle. \tag{39}$$

Recall also that the Lovász Bregman divergence in some sense measures the distance between the ordering of the vectors and can be seen as a form of the "sorting" distance. We define the clustering problem as given a set of vectors, find a clustering into subsets of vectors with similar orderings. For example, given a set of voters and their corresponding ranked preferences, we might want to find subsets of voters who mostly agree. Let $\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$ represent a set of m vectors, such that $\forall i, x_i \in [0, 1]^n$. We first consider the problem of finding the representative of these vectors. Given a set of vectors $\mathcal{X}$ and a Lovász Bregman divergence $d_{\hat{f}}$, a natural choice of a representative (in this case a permutation) is the point with minimum average distance, or in other words:

$$\sigma = \operatorname*{argmin}_{\sigma'} \sum_{i=1}^{n} d_{\hat{f}}(x_i || \sigma') \tag{40}$$

Interestingly for the Lovász Bregman divergence this problem is easy and the representative permutation is exactly the permutation of the arithmetic mean of $\mathcal{X}$

Theorem 4.2. *Given a submodular function f, the Lovász Bregman representative $\operatorname{argmin}_{\sigma'} \sum_{i=1}^{n} d_{\hat{f}}(x_i || \sigma')$ is exactly $\sigma = \sigma_\mu, \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$*

The proof of the above theorem is in Appendix E.1 It may not suffice to encode $\mathcal{X}$ using a single representative, and hence we partition $\mathcal{X}$ into disjoint blocks $\mathcal{C} = \{C_1, \cdots, C_k\}$ with each block having its own Lovász Bregman representative, with the set of representatives given by $\mathcal{M} = \{\sigma_1, \sigma_2, \cdots, \sigma_k\}$. Then we define an objective, which captures this idea of clustering vectors into subsets of similar orderings:

$$\min_{\mathcal{M}, \mathcal{C}} \sum_{j=1}^{k} \sum_{x_i \in C_j} d_{\hat{f}}(x_i || \sigma_j) \tag{41}$$

Consider then a k-means like alternating algorithm [21, 23]. It has two stages, often called the *assignment* and the *re-estimation* step. In the assignment stage, for every point x_i we choose its cluster membership C_j such that $j = \operatorname{argmin}_l d_{\hat{f}}(x_i || \sigma_l)$. The re-estimation step involves finding the representatives for every cluster C_j, which is exactly the permutation of the mean of the vectors in C_j.

Algorithm 2: The Lovász Bregman k-means algorithm

```
Given a set of sets $\mathcal{X}$, find a clustering $\mathcal{C}$ and set of permutations $\mathcal{M}$.
Initialize $\mathcal{M}^0$ as a particular choice of initial permutations.
$t = 0$
while not converged do
    $t \leftarrow t + 1$
    The assignment step:
    $\forall i = 1, 2, \cdots, m$, assign $x_i$ to a cluster $C_j^t$ such that $j = \operatorname{argmin}_l d(x_i || \sigma_l^{t-1})$.
    The re-estimation step:
    For the clustering $\mathcal{C}^t$ obtained above, find the representative (or mean) $\sigma_j^t$ for each $C_j^t$.
```

Algorithm 4 is very similar to the k-means algorithm. Further it is obvious that the performance of this algorithm depends on the choice of the the initial permutations. We do not consider any schemes of choosing the initial permutations, however similar to the standard k-means, it is possible to provide hueristics for this.

We remark here that a number of distance measures capture the notion of orderings, like the bubble-sort distance [19], etc. However for these distance measures, finding the representative may not be easy. The Lovász Bregman divergence naturally captures the notion of distance between orderings of vectors and yet, the problem of finding the representative in this case is very easy. Then similar to [2], we can show that:

Lemma 4.1. *A k-means clustering algorithm defined above will monotonically decrease the objective of equation (43) at every iteration.*

This theorem can be proved in a manner similar to [2]. Further It is interesting to analyze the results of this algorithm at convergence. We have the following theorem.

Lemma 4.2. *At the convergence of Algorithm 2, we are guaranteed to converge to a local minimum in the sense that the loss function cannot be improved by either the assignment step, or by changing the means (permutations) of any existing clusters.*

To demonstrate the utility of our clustering framework, we show some results in 2 and 3 dimensions (Fig. 1), where we compare our framework to a k-means algorithm using the euclidean distance. We use the submodular function $f(X) = \sqrt{w(X)}$, for an arbitrary vector w ensuring unique base extreme points. The results clearly show that the Lovász Bregman divergence clusters the data based on the orderings of the vectors.

4.3 A clustering framework for the submodular Bregman

It has been shown ([2]) that the Bregman divergence possesses very interesting properties regarding clustering in the k-means framework. It is the only class of continuous divergences which has the property that the point having the minimum average distance to a set of points $x_1, x_2, \cdots, x_m$ is simply the mean of these points. The previous section, shows yet another application of this through the Lovász Bregman divergences.

In this section however, we extend the clustering framework of [2] to cluster sets of binary vectors (or equivalently sets), with the centroids themselves being binary vectors (sets). Clearly, we cannot naïvely use the continuous clustering framework for such vectors. Hence the underlying problem is to find the set (binary vector) that has the minimum average distance to a set of sets. This can be useful in machine learning applications where objects (which might be structured and/or variable length such as strings, trees, or graphs) are well-represented by a fixed size set of binary features. We motivate this by an example of the Hamming distance clustering.

Example 4.1. *In this context, we want to cluster a set of sets, using the Hamming distance. We can use a k-means like algorithm for this, and the Hamming representative has a particularly nice form. Given a set of sets $X_1, X_2, \cdots, X_n$, the hamming representative (we formally define the notion of a discrete representative in the following section), is X_H which is an integer rounding of $\frac{\sum_{i=1}^n 1_{X_i}}{n}$ (We show this in Appendix E.2).*

In this paper, we consider a clustering framework using the submodular Bregmans as the class of discrete divergences (notice that this subsumes the Hamming distance clustering discussed above).

4.3.1 Finding the submodular Bregman representatives

Let $\mathcal{X} = \{X_1, X_2, \cdots, X_m\}$ represent a set of m sets (or m binary vectors) that we want to cluster. Then we can define two problems, which we call the left and right means problems respectively. Let S_l and S_r represent the left and right *Bregman representatives* respectively — then we have, for a divergence d:

$$S_l = \operatorname*{argmin}_{S \subseteq V} \sum_{i=1}^{n} d(S, X_i), \qquad S_r = \operatorname*{argmin}_{S \subseteq V} \sum_{i=1}^{n} d(X_i, S) \tag{42}$$

Both the above are set function minimization problems that in general can be intractable. However when d is an instance of the submodular Bregman divergence, some very interesting properties exist. It is evident from the definitions, for example, that for $d = d_f^{\mathcal{H}_f}$, the left means problem is a submodular minimization problem, and correspondingly for $d = d_{\mathcal{G}^f}^f$, the right means problem is a submodular maximization problem. Hence there are a number of fast algorithms to approximately (or sometimes exactly) solve them [9, 18, 12, 24].

The other two problems (the right mean with $d = d_f^{\mathcal{H}_f}$, and left mean with $d = d_{\mathcal{G}^f}^f$) can also be approximately solved, thanks to the structure of the submodular Bregmans and their connection to the Lovász extension (Sections 2.3 and 2.4). As shown in the theorem below, the right means problem with $d = d_f^{\mathcal{H}_f}$, is exactly equivalent to solving $X_r = \operatorname{argmin}_S d_{\hat{f}}(\mu, 1_S)$, where μ is the *continuous* mean of $1_{X_1}, 1_{X_2}, \cdots, 1_{X_n}$. Similarly the left means problem with $d = d_{\mathcal{G}^f}^f$, is equivalent to solving $X_r = \operatorname{argmin}_S d^{\hat{f}}(1_S, \mu)$. Thus, we need to find the vertex of the hypercube that is closest to the continuous mean μ (in the generalized Bregman divergence sense), and correspondingly we would expect that rounding μ would give the optimal mean.

Algorithm 3: Generalized rounding procedure

```
1. Define $L(S) = \sum_{i=1}^{n} \frac{1}{n} d(S, X_i)$ and $R(S) = \sum_{i=1}^{n} \frac{1}{n} d(X_i, S)$
2. Sort the elements of $\mu$ (the continuous mean of $1_{X_1}, 1_{X_2}, \cdots, 1_{X_n}$) to obtain the chain of sets $U_1 \subseteq U_2 \subseteq \cdots \subseteq U_n$
3. Assign $S_l = U_k$, where $k = \text{argmin}_j L(U_j)$ (for the left mean) and $k = \text{argmin}_j R(U_j)$ (for the right mean).
```

Consider the procedure described in Algorithm 3 for obtaining the means. Observe that this procedure is identical to rounding the continuous mean at different thresholds and picking the best one — that is, picking any of the sets U_i corresponds to thresholding the mean μ at $\mu(\sigma(i))$ (i.e., setting all elements less than it to 0, and the ones above it to 1). We formally show this in Theorem 4.3 (proof in Appendix E.2).

Theorem 4.3. *We have that:*

- *The left means problem for* d $=\mathrm{d}^{f}_{\mathcal{G}^f}$*, is equivalent to solving* $X_l = \text{argmin}_X\, \mathrm{d}^{\hat{f}}(1_X, \mu) = \text{argmin}_X \sum_{i=1}^{m} \lambda_i \mathrm{d}^{f}_{\mathcal{G}^f}(X, U_i)$*, where* λ_i *are constants and* U_i*'s are as defined in Algorithm 3.*
- *The right means problem for the lower bound submodular Bregman is equivalent to solving* $X_r = \mathit{argmin}_X \mathrm{d}_{\hat{f}}(\mu, 1_X) = \text{argmin}_X \sum_{i=1}^{m} \lambda_i \mathrm{d}_{f}^{\mathcal{H}_f}(U_i, X)$*, where* λ_i *are constants and* U_i*'s are as defined in Algorithm 3.*
  For the Hamming distance, Algorithm 3 gives the exact representatives.

The above theorem suggests the utility of the rounding procedure 3 for these problems. Notice that the problem of finding the means is equivalent to minimizing the weighted sum of the divergences to the chain of sets corresponding to the continuous means. Hence intuitively if the divergence is not too curved, one would expect the minimizer to be one of the sets in the chain (which corresponds to rounding the continuous mean at different points). In particular, notice that this is the case for a modular divergence, like the Hamming or Recall/ Precision etc. Finally however we point out that we do not have any theorotical guarantees for general forms of submodular Bregman, and it would be interesting if either constant factor or log factor guarantees could be provided for this.

Hence we see, similar to [2], finding the representative of a given set is computationally efficient and salable (very low polynomial time complexity, and essentially $O(n)$ in some cases). Importantly, note that it is the structure of the submodular Bregman divergences that give these nice properties, and the generalized rounding algorithm will not work for arbitrary discrete divergences.

We formally define the clustering problem in the framework of [2]. Let $\mathcal{X} = \{X_1, X_2, \cdots, X_m\}$ represent a set of m sets (or m binary vectors) that we want to cluster.

4.3.2 The clustering framework

We first consider the problem of finding the submodular Bregman representatives. Since the submodular Bregman is not in general symmetric, we have two problems above, which we have called the left mean (L) and right mean (R) problems respectively. The sets that minimize the expressions above are the submodular Bregman representatives and are named S_l and S_r respectively. Hence we generalize equation (42). Thanks to the nice structure of the submodular Bregman, we can efficiently find the submodular Bregman representatives as we show in Section 4.3 and Theorem 4.3.

It may not suffice to encode $\mathcal{X}$ using a single representative, and hence we partition $\mathcal{X}$ into disjoint blocks $\mathcal{C} = \{C_1, \cdots, C_k\}$ with each block having its own submodular Bregman representative. Let d be a discrete divergence (which in this framework, we assume is a form of submodular Bregman) and $\mathcal{M} = \{M_1, \cdots, M_k\}$ be the set of representatives. Then we define the clustering objective as:

$$\min_{\mathcal{M},\mathcal{C}} \sum_{j=1}^{k} \sum_{X_i \in C_j} d(X_i, M_j) \tag{43}$$

Here and in the below, we have explicitly defined the right-means problem, but a similar expression and equations can be given for the left means as well.

Consider now a k-means like alternating algorithm [21, 23]. It will typically have two stages, often called the *assignment* and the *re-estimation* step. In the assignment step, for every point X_i we choose its cluster membership C_j such that $j = \text{argmin}_l\, d(X_i, M_l)$. The re-estimation step involves finding the representatives for every cluster C_j (by solving equation (42)). This immediately yields Algorithm 4.

Algorithm 4: The submodular Bregman k-means algorithm

```
Given a set of sets $\mathcal{X}$, find a clustering $\mathcal{C}$ and set of centroids $\mathcal{M}$ that attempts to minimize Eqn. (43)
Initialize $\mathcal{M}^0$ to be random centroids.
$t = 0$
while not converged do
    $t \leftarrow t + 1$
    The assignment step:
    $\forall i = 1, 2, \cdots, m$, assign $X_i$ to a cluster $C_j^t$ such that $j = \text{argmin}_l\, d(X_i, M_l^{t-1})$.
    The re-estimation step:
    For the clustering $\mathcal{C}^t$ obtained above, find the representative $M_j^t$ for each $C_j^t$ using equation (42).
    For each $j$, pick the best out of $M_j^t$ and $M_j^{t-1}$
```

Algorithm 4 is very similar to the k-means algorithm, except that at every stage, for each $j = 1, 2, \cdots, k$, we pick the best representative amongst M_j^t and M_j^{t-1}. By “best”, we mean the one that has a lesser average distance to the sets in C_j^t. Note that the reason for this is that, in all except one case (finding the left mean of a lower-bound submodular Bregman function, which is a submodular minimization problem that can be solved exactly in polynomial time) the representatives of the submodular Bregman are approximate means (e.g., the right mean problem of the upper bound submodular Bregman divergence is a submodular maximization problem which can only be solved approximately in polynomial time, and also the right (respectively left) mean problem of the lower (respectively upper) bound submodular Bregman function which is also only approximate). Hence by picking the best of M_j^t and M_j^{t-1}, we are guaranteed to not increase objective (43) at every iteration. Further, if M_j^{t-1} is a better representative, then the preceding mean is a better representative of the current clustering and hence we choose it.

Then similar to [2], we can show that:

Lemma 4.3. *A k-means clustering algorithm defined above will monotonically decrease the objective of equation (43) at every iteration.*

This theorem can be proved in a manner similar to [2]. The assignment step clearly reduces the objective. Further, the way we have defined the re-estimation step, again the objective can only reduce at every iteration. Since this is a discrete problem with a finite number of clusters k, the number of distinct clusterings are finite, and hence we are guaranteed to converge.

It is interesting to analyze the results of this algorithm at convergence. We have the following theorem.

Theorem 4.4. *The following is true at convergence of Algorithm 4:*

- *For the left means problem, with the lower bound submodular Bregman, we are guaranteed to converge to a local minimum in the sense that the loss function cannot be improved by either the assignment step, or by changing the means of any existing clusters.*

- *For the right means problem, with the upper bound submodular Bregman, if we use a local search technique for non-monotone submodular maximization [9], then we are guaranteed to converge to a local minimum, in the sense that the loss function cannot improve by either the assignment step, or by taking any subset or superset of the means of the existing clusters.*

- *For the right means problem of the lower bound submodular Bregman, and the left means problem, of the upper bound submodular Bregman, we are guaranteed to converge to a local minimum, in the sense that the loss function cannot improve by either the assignment step or by any other rounding of the means of the current clustering.*

Proof. For the proof of this theorem, observe first that since submodular minimization can exactly in polytime, the means computed at every step are the global minima, and hence for the current clustering no other means can improve the solution. Now for the right means problem of the upper bound submodular Bregman, since it is a submodular maximization problem, the local search solution of [9] converges to a local maxima. Further the local maxima has the property that no subset of superset of the optimal set can be better than it. Hence the means obtained at convergence will have the property that no subset of superset of them will be better. Finally observe that the generalized clustering algorithm gives the best rounding of the continuous mean and hence we are guaranteed that if the assignment step does not change the clustering, then the means obtained will be the best rounding of the means of the current clusterings. □

5 Conclusions

In this paper, we introduced a submodular Bregman divergence, characterized by a submodular function. Using both the upper and lower bounds, we defined two forms of the submodular Bregman, and also analyzed their relation with the Bregman divergence. We also introduced the Lovász Bregman divergences, as a form of the generalized Bregman divergence on the Lovász extention. We showed how the submodular Bregman divergences generalize many useful distance measures like the hamming, precision and recall, and showed how the Lovász Bregman divergences provide a natural framework for clustering vectors based on their ordering. We Finally we showed some useful applications of these, in the context of clustering and as a proximal operator, which provides a framework for submodular optimization.

This new notion of the submodular Bregman divergences are indeed very exciting, and we hope that they can be found to be useful in many more application contexts in machine learning, in addition to the ones we consider. Further the Lovász Bregman divergences also provide a very exciting class of continuous divergences, which enrich the already existing utilities of the Bregman divergences in Machine Learning.

Acknowledgments: We thank Stefanie Jegelka, Karthik Narayanan, Andrew Guillory, Hui Lin, John Halloran and the rest of the submodular group at UW-EE for discussions. This material is based upon work supported by the National Science Foundation under Grant No. (IIS-1162606), and is also supported by a Google, a Microsoft, and an Intel research award.

References

[1] F. Bach. Learning with Submodular functions: A convex Optimization Perspective. *Arxiv*, 2011.

[2] A. Banerjee, S. Meregu, I. S. Dhilon, and J. Ghosh. Clustering with Bregman divergences. *JMLR*, 6:1705–1749, 2005.

[3] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Applied Math.*, 123(1–3):155 – 225, 2002.

[4] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math and Math Physics*, 7, 1967.

[5] Y. Censor and S. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press, USA, 1997.

[6] W. H. Cunningham. Decomposition of submodular functions. *Combinatorica*, 3(1):53–68, 1983.

[7] I. Dhillon and J. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

[8] J. Edmonds. Submodular functions, matroids and certain polyhedra. *Combinatorial structures and their Applications*, 1970.

[9] U. Feige, V. Mirrokni, and J. Vondrák. Maximizing non-monotone submodular functions. *SIAM J. COMPUT.*, 40(4):1133–1155, 2007.

[10] B. Frigyik, S. Srivastava, and M. Gupta. Functional Bregman divergence. In *In ISIT*, pages 1681–1685. IEEE, 2008.

[11] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier Science, 2005.

[12] S. Fujishige and S. Isotani. A submodular function minimization algorithm based on the minimum-norm base. *Pacific Journal of Optimization*, 7:3–17, 2011.

[13] R. Iyer and J. Bilmes. Algorithms for approximate minimization of the difference between submodular functions, with applications. *In UAI*, 2012.

[14] R. Iyer and J. Bilmes. A unified theory on the generalized Bregman divergences. *Manuscript*, 2012.

[15] R. Iyer, S. Jegelka, and J. Bilmes. Mirror descent like algorithms for submodular optimization. *NIPS Workshop on Discrete Optimization in Machine Learning (DISCML)*, 2012.

[16] S. Jegelka and J. Bilmes. Cooperative cuts: Graph cuts with submodular edge weights. Technical report, Technical Report TR-189, Max Planck Institute for Biological Cybernetics, 2010.

[17] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[18] S. Jegelka, H. Lin, and J. Bilmes. On fast approximate submodular minimization. *In NIPS*, 2011.

[19] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[20] K. C. Kiwiel. Free-steering relaxation methods for problems with strictly convex costs and linear constraints. *Mathematics of Operations Research*, 22(2):326–349, 1997.

[21] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on IT*, 28(2):129–137, 1982.

[22] L. Lovász. Submodular functions and convexity. *Mathematical Programming*, 1983.

[23] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on math. stats and probability*, volume 1, pages 281–297. California, USA, 1967.

[24] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, pages 234–243, 1978.

[25] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *UAI*, 2005.

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

[27] P. Stobbe and A. Krause. Efficient minimization of decomposable submodular functions. In *NIPS*, 2010.

[28] M. Telgarsky and S. Dasgupta. Agglomerative Bregman clustering. *In ICML*, 2012.

[29] K. Tsuda, G. Ratsch, and M. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *JMLR*, 6(1):995, 2006.

[30] J. Vondrák. *Submodularity in combinatorial optimization*. PhD thesis, Charles University, 2007.

[31] M. K. Warmuth. Online learning and Bregman divergences. *Tutorial at the Machine Learning Summer School*, 2006.

A Proofs related to the lower bound submodular Bregman

A.1 Proof of Lemma 2.3

Proof. To prove this theorem, observe that $\partial_f(Y) = \mathcal{P}_{f_{\#}^Y} \times \mathcal{P}_{f_Y}$. However since we have that $\forall y \in \partial_f^{\mathcal{B}}(Y), y(Y) = f(Y)$, we have that the vectors in $\mathcal{P}_{f_{\#}^Y}$ actually belong to the base polytope of $f_{\#}^Y$. Further the base polytope of $f_{\#}^Y$ is exactly the base polytope of f^Y. A similar argument proves that the other part is the base polytope of f_Y. Lastly observe that the base polytope is a closed polytope and hence it is the convex combinations of its extreme points. Hence each part in the direct product is a convex combination of their respective extreme points, and hence the base sub-differential is a convex combination of the extreme points of the sub-differential. □

A.2 Proof of theorem 2.1

Proof. Observe that from the earlier result, the submodular sub-differential is a direct product of a dual submodular polyhedron and a submodular polyhedron. Hence the linear program actually can be broken up into two smaller linear programs as follows:

$$max_{s\in\partial_f(Y)}s^\top w = \max_{s_1\in P_{f^Y_{\#}}} s_1^\top w_1 + \max_{s_2\in\mathcal{P}_{f_Y}} s_2^\top w_2 \tag{44}$$

where w_1 is a vector formed of elements of w inside Y, and w_2 is a vector formed of elements outside Y. Then the results from Fujishige [11] directly show that we obtain the maximizer if we order the elements using the greedy algorithm, to order the elements both within and outside the set Y. For the maximizers over $\partial_f(Y)$ to be bounded however, we require that the elements within Y be non-positive, while the elements outside Y be non-negative. Since the base subdifferential is bounded, the linear program over the base subdifferential will always remain bounded. □

A.3 Proof of theorem 2.2

Proof. Observe that $\sigma_{\partial_f(Y)\cap\mathcal{P}_f}(1_X - 1_Y) = \max_{y\in\partial_f(Y)\cap\mathcal{P}_f}\langle y, 1_X - 1_Y\rangle$. We then invoke the theorem above, to order the elements in $1_X - 1_Y$ and find a permutation $S_i = [\sigma(1), \sigma(2), \cdots, \sigma(i)]$, with $S_{|Y|} = Y$. In addition $\{1_X - 1_Y\}(j) \leq 0, j \in Y$ and $\{1_X - 1_Y\}(j) \geq 0, j \notin Y$, and hence the maximization over the subdifferential will be bounded.

We require the orderings then to ensure that $\{1_X - 1_Y\}(\sigma(1) \geq \{1_X - 1_Y\}(\sigma(2)) \cdots \{1_X - 1_Y\}(\sigma(|Y|))$ and similarly for elements for those outside Y, $\{1_X - 1_Y\}(\sigma(|Y| + 1)) \geq \{1_X - 1_Y\}(\sigma(|Y| + 2)) \cdots \{1_X - 1_Y\}(\sigma(n))$.

Let h represent the maximizer. Then the ordering of h, should contain the elements in $X \cap Y$ first since those correspond to the maximum value of $1_X - 1_Y$ inside Y (Notice that for $j \in Y, 1_Y(j) = 1$. The next set of elements would be the elements in $Y\backslash X$. Similarly we can show that for the elements outside Y (i.e $j \notin Y$), we order the elements in $X\backslash Y$ first followed by the rest of the elements. Hence $h(X \cap Y) = f(X \cap Y), h(Y) = f(Y)$ and $h(X \cup Y) = f(X \cup Y)$. Then simple algebra reveals that:

$$\begin{aligned}\sigma_{\partial_f(Y)}(1_X - 1_Y) &= \langle h, 1_X - 1_Y\rangle \\ &= h(X) - h(Y) \\ &= h(X \cap Y) + h(X\backslash Y) - h(Y) \\ &= f(X \cap Y) + f(X \cup Y) - f(Y) - f(Y)\end{aligned} \tag{45}$$

Substituting this back, we get $d^\sharp_f(X, Y) = f(X) + f(Y) - f(X \cap Y) - f(X \cup Y)$. □

A.4 Proof of theorem 2.3

Proof. Again this follows from the greedy algorithm corresponding to the subdifferential. Notice that $\sigma_{\partial_f(Y)^\mathcal{B}}(1_Y - 1_X) = \max_{h\in\partial_f(Y)^\mathcal{B}}(1_Y - 1_X)$.

Let h represent the maximizer. Then the ordering of h, should contain the elements in $Y\backslash X$ first since those correspond to the maximum value of $1_Y - 1_X$ inside Y. The next set of elements would be the elements in $X \cap Y$. Similarly we can show that for the elements outside Y (i.e $j \notin Y$), we order the elements which are not in $X \cup Y$ first followed by the rest of the elements. Hence $h(Y\backslash X) = f(Y\backslash X), h(Y) = f(Y)$ and $h(X\backslash Y) = f(V) - f(V\backslash\{X\backslash Y\}) = f^\sharp(X\backslash Y)$. Then simple algebra reveals that:

$$\begin{aligned}\sigma_{\partial_f(Y)\cap\mathcal{B}_f}(1_Y - 1_X) &= \langle h, 1_Y - 1_X\rangle \\ &= h(Y) - h(X) \\ &= -h(X \cap Y) - h(X\backslash Y) + h(Y) \\ &= -h(Y) + h(Y\backslash X) - h(X\backslash Y) + h(Y) \\ &= f(Y\backslash X) - f^\sharp(X\backslash X)\end{aligned} \tag{46}$$

Substituting this back, we get the expression for $d^\natural_f$. □

B Proofs related to the upper bound submodular Bregman

B.1 Proof of Theorem 2.4

Proof. Consider first d_1^f. Now define a vector $g_X^1 \in \mathbb{R}^V$, such that:

$$g_X^1(j) = \begin{cases} f(j|X - j) & \text{if } j \in X \\ f(j|\emptyset) & \text{if } j \notin X \end{cases} \tag{47}$$

Now for any $Y \subseteq V$ we have:

$$\begin{aligned} g_X^1(Y) - g_X^1(X) &= g_X^1(Y \backslash X) + g_X^1(X \cap Y) - g_X^1(X) \\ &= \sum_{j \in Y \backslash X} f(j|\emptyset) + \sum_{j \in X \cap Y} f(j|X - j) - \sum_{j \in X} f(j|X - j) \\ &= \sum_{j \in Y \backslash X} f(j|\emptyset) - \sum_{j \in X \backslash Y} f(j|X - j) \end{aligned} \tag{48}$$

and we obtain terms belonging to the submodular Bregman of the first type. Hence we have that d_1^f, is a special case of $d_{\mathcal{G}^f}^f$. We can also similarly define a g_X^2 for the submodular Bregman of the second type, as:

$$g_X^2(j) = \begin{cases} f(j|V - j) & \text{if } j \in X \\ f(j|X) & \text{if } j \notin X \end{cases} \tag{49}$$

It is not hard to see that from $g_X^2(Y) - g_X^2(X)$ we get terms in the submodular Bregman of the second type, and from g_X^3, we get the submodular Bregman of third type and hence d_2^f d_3^f are also special cases of $d_{\mathcal{G}^f}^f$. □

B.2 Proof of Lemma 2.5

Proof. We define $f(X) = \sum_i \lambda_i h_i(m_i(X))$ with $\lambda_i \geq 0, \forall i$. Given that h_i's are concave functions, we have from the definition of concavity ($\forall i$):

$$h_i(x) - h_i(y) - h_i'(x)(x - y) \geq 0 \tag{50}$$

Where h_i' here represents a supergradient of the concave function at x. There should be no confusion to the reader, however, that if h_i's are differentiable at x, then the supergradient is exactly the derivative. Now consider evaluating this function at $x = m_i(X)$ and $y = m_i(Y)$. Further let m_i be the vector corresponding to $m_i(X)$, or in other words $m_i(X) = \langle m_i, 1_X \rangle$. Then we have that:

$$\begin{aligned} & h_i(m_i(X)) - h_i(m_i(Y)) - h_i'(m_i(X))(m_i(X) - m_i(Y)) \geq 0 \\ \Rightarrow\; & h_i(m_i(X)) - h_i(m_i(Y)) - h_i'(m_i(X))(\langle m_i, 1_X \rangle - \langle m_i, 1_Y \rangle) \geq 0 \\ \Rightarrow\; & h_i(m_i(X)) - h_i(m_i(Y)) - h_i'(m_i(X))\langle m_i, 1_X - 1_Y \rangle \geq 0 \\ \Rightarrow\; & h_i(m_i(X)) - h_i(m_i(Y)) - \langle h_i'(m_i(X))m_i, 1_X - 1_Y \rangle \geq 0 \end{aligned} \tag{51}$$

Note that $h_i(m(X))$ is just a scalar. Hence we have that the supergradient of $g_i(X) \triangleq h_i(m_i(X))$ (seen as a submodular function) at X is $h_i'(m_i(X))m_i$. We can then easily show the result, by multiplying the expression above by λ_i, and summing over i.

$$\begin{aligned} \sum_i \lambda_i \Big(h_i(m_i(X)) - h_i(m_i(Y)) - \langle h_i'(m_i(X))m_i, 1_X - 1_Y \rangle \Big) \geq 0 \\ \sum_i \lambda_i h_i(m_i(X)) - \sum_i \lambda_i h_i(m_i(Y)) - \langle \sum_i \lambda_i h_i'(m_i(X))m_i, 1_X - 1_Y \rangle \geq 0 \\ f(X) - f(Y) - \left\langle \sum_i \lambda_i h_i'(m_i(X))m_i, 1_X - 1_Y \right\rangle \geq 0 \end{aligned} \tag{52}$$

Hence $\sum_i \lambda_i h_i'(m_i(X))m_i$ is actually then a supergradient of $f(X)$ at X. Hence proved. □

B.3 Proof of Theorem 2.5

Proof. Consider for now d_2^f:

$$\begin{aligned} d_2^f(X, Y) &= f(X) - \sum_{j \in X \backslash Y} f(j | V - \{j\}) + \sum_{j \in Y \backslash X} f(j | X) - f(Y) \\ &= \sum_{j \in Y \backslash X} f(j|X) - f(X \cup Y) + f(X) + f(X \cup Y) - f(Y) - \sum_{j \in X \backslash Y} f(j | V - \{j\}) \\ &= \sum_{j=1}^{k} \Big[f(x_j | X) - f(x_j | X_{j-1}) \Big] + \sum_{j=1}^{l} \Big[f(y_j | Y_{j-1}) - f(y_j | V - y_j) \Big] \end{aligned} \quad (53)$$

Note that the two sums are over elements, respectively, in $X \backslash Y$ and $Y \backslash X$ and that the terms within each of the sums is non-negative. The submodular Bregman seen in this form now seems like a distance measure, since would expect that (like the Hamming distance $d_H(X, Y) = |X \backslash Y| + |Y \backslash X|$) the distance would be larger if $|X \backslash Y|$ and $|Y \backslash X|$ is larger. Analogously d_1^f can also be written as:

$$d_1^f(X, Y) = \sum_{j=1}^{k} \Big[f(x_j | \emptyset) - f(x_j | X_{j-1}) \Big] + \sum_{j=1}^{l} \Big[f(y_j | Y_{j-1}) - f(y_j | Y - y_j) \Big] \quad (54)$$

The bounds now, directly follow from the above equations. The case of d_3^f is analogous to the above, and we leave it to the reader. □

C Proofs related to the continuous extentions of the submodular Bregman divergences

Proof of Theorem 2.6

Proof. To show the first part, we argue that if $x = 1_X$ and $y = 1_Y$ are vertices of the hypercube corresponding to sets X and Y, then the subgradient defined on the Lovász extension is actually the same as the set of modular lower bounds, corresponding to the lower bound submodular Bregman. In other words if Y is the set corresponding to the vertex 1_Y, then $h_Y = h_{1_Y}$. To see this, observe that when 1_Y is ordered following a permutation π, all the permutations of sets will involve the ones first followed by the zeros. Clearly 1_Y has $|Y|$ ones and hence every such permutation π of V will be such that $W_{|Y|} = Y$, and hence $h_{Y,\pi} = h_{y,\pi}$. This fact seen with equation (33) and (6) shows the third item of the theorem. Another probably simpler way to see this fact is that since the subdifferentials corresponding to h_y and h_Y are identical, and hence their extreme points are identical.

We then show the second item as follows. Note that this is true for any generalized lower bound Bregman divergence, and any generalized Bregman divergence of the Lovász extension (any subgradient), as long as the chosen subgradient of $\hat{f}$ at 1_Y and that of f at Y are the same. Now observe that from the Lovász extension that we have: $\hat{f}(x) = x(\sigma(n))f(X_n) + \sum_{i=1}^{n-1}(x(\sigma(i)) - x(\sigma(i+1))f(X_i) = \sum_{i=1}^{n} \lambda_i f(X_i)$. Further also notice that $x = \sum_{i=1}^{n} \lambda_i 1_{X_i}$ and $\sum_{i=1}^{n} \lambda_i = x(\sigma(1))$. Thus we have (using that $h_{1_Y} = h_Y$)

$$\begin{aligned} d_{\hat{f}}(x, y) &= \hat{f}(x) - \hat{f}(y) - \langle h_Y, x - y \rangle \\ &= \sum_i \lambda_i f(X_i) - \hat{f}(y) - \langle h_Y, \sum_i \lambda_i 1_{X_i} - y \rangle \\ &= \sum_i \lambda_i f(X_i) + (\sum_i \lambda_i + 1 - x(\sigma(1)))f(Y) - \langle h_Y, (\sum_i \lambda_i + 1 - x(\sigma(1)))1_Y - \sum_i \lambda_i 1_{X_i} \rangle \\ &= \sum_i \lambda_i d_f^{\mathcal{H}_f}(X_i, Y) + (1 - y(\sigma(1)))d_f^{\mathcal{H}_f}(X_0, Y) \end{aligned} \quad (55)$$

Finally note that since $\lambda_i \geq 0$ and $x(\sigma(1)) \leq 1$ and thus we have the Lovász extension of the Bregman divergence, when x is a continuous vector within the hypercube and y is a vertex of the hypercube, is a convex combination of the submodular Bregman between Y and the chain of sets, corresponding to x.

We now show the third statement of the theorem. Observe that from the Lovász extension, we have: $\hat{f}(y) = y(\sigma(n))f(Y_n) + \sum_{i=1}^{n-1}(y(\sigma(i)) - y(\sigma(i+1))f(Y_i) = \sum_{i=1}^{n} \lambda_i f(Y_i)$. Further also notice that $y = \sum_{i=1}^{n} \lambda_i Y_i$ and $\sum_{i=1}^{n} \lambda_i = y(\sigma(1))$. First we define now a permutation for every set Y_i. Notice that for every i, the permutation σ is a valid permutation for Y_i since each Y_i appears in that chain. Hence let, $h_{Y_i,\sigma} = h_{y,\sigma}$. In other words, corresponding to every set in the chain, we define the subgradient corresponding to the same permutation σ. Further let $x = 1_X$. Thus we have:

$$\begin{aligned}
d_{\hat{f},\sigma}(x, y) &= \hat{f}(x) - \hat{f}(y) - \langle h_{y,\sigma}, x - y\rangle \\
&= \hat{f}(1_X) - \sum_i \lambda_i f(Y_i) - \langle h_{y,\sigma}, 1_X - \sum_i \lambda_i 1_{Y_i}\rangle \\
&= (\sum_i \lambda_i + 1 - y(\sigma(1)))f(X) - \sum_i \lambda_i f(Y_i) - \langle h_{y,\sigma}, (\sum_i \lambda_i + 1 - y(\sigma(1)))1_X - \sum_i \lambda_i 1_{Y_i}\rangle \\
&= \sum_i \lambda_i (f(X) - f(Y_i) - \langle h_{Y_i,\sigma}, 1_X - 1_{Y_i}\rangle) + (1 - y(\sigma(1)))(f(X) - \langle h_{Y_0,\sigma}, 1_X\rangle) \\
&= \sum_i \lambda_i f(X) - h_{Y_i,\sigma}(X) + (1 - y(\sigma(1)))f(X) - h_{Y_0,\sigma}(X) \\
&= \sum_i \lambda_i d_f^{\mathcal{H}_f}(X, Y_i) + (1 - y(\sigma(1)))d_f^{\mathcal{H}_f}(X, Y_0) \qquad (56)
\end{aligned}$$

Finally the last statement of the theorem directly follows from the above two statements. □

This is the proof of Theorem 2.7.

Proof. First observe that from the Lovász extension, we have: $\hat{f}(y) = y(\sigma(n))f(Y_n) + \sum_{i=1}^{n-1}(y(\sigma(i)) - y(\sigma(i+1))f(Y_i) = \sum_{i=1}^{n} \lambda_i f(Y_i)$. Further also notice that $y = \sum_{i=1}^{n} \lambda_i 1_{Y_i}$ and $\sum_{i=1}^{n} \lambda_i = y(\sigma(1))$. Thus we have:

$$\begin{aligned}
d^{\hat{f}}(x, y) &= \hat{f}(x) - \sum_i \lambda_i f(Y_i) - \langle g_X, x - \sum_i \lambda_i 1_{Y_i}\rangle \\
&= (\sum_i \lambda_i + 1 - y(\sigma(1)))\hat{f}(1_X) - \sum_i \lambda_i f(Y_i) - \langle g_X, (\sum_i \lambda_i + 1 - y(\sigma(1)))1_X - \sum_i \lambda_i 1_{Y_i}\rangle \\
&= \sum_i \lambda_i (f(X) - f(Y_i) - \langle g_X, 1_X - 1_{Y_i}\rangle) + (1 - y(\sigma(1)))(f(X) - \langle g_X, 1_X\rangle) \\
&= \sum_i \lambda_i d_{\mathcal{G}^f}^f(X, Y_i) + (1 - y(\sigma(1)))d_{\mathcal{G}^f}^f(X, Y_0) \qquad (57)
\end{aligned}$$

□

D Proofs related to the properties of the submodular Bregman divergences

D.1 Submodularity of the Upper bound submodular Bregman in the first argument

Theorem D.1. *In the below, let $m_Y(X)$ be a given modular function in X parameterized by a fixed Y.*

*For a fixed Y and if f is monotone **non-increasing** submodular function, the submodular Bregman of the first type $d_f^1(X, Y)$ can be expressed as a difference between two submodular functions in X as follows:*

$$d_f^1(X, Y) = \Big(f(X) + \sum_{j \in X \setminus Y} f(X - j) + m_Y(X)\Big) - \Big(\sum_{j \in X \setminus Y} f(X)\Big) \qquad (58)$$

Similarly, for fixed Y and if f is monotone ***non-decreasing*** *submodular function, then the submodular Bregman of the second type $d^2_f(X,Y)$ can be expressed as a difference between two submodular functions in X as follows:*

$$d^2_f(X,Y) = \Big(f(X) + \sum_{j \in Y \backslash X} f(X+j) + m_Y(X)\Big) - \Big(\sum_{j \in Y \backslash X} f(X)\Big) \tag{59}$$

Proof. Recall the expression for the upper bound submodular Bregman of the first type:

$$d^f_1(X,Y) = f(X) - \sum_{j \in X \backslash Y} f(j|X - \{j\}) + \sum_{j \in Y \backslash X} f(j|\emptyset) - f(Y) \tag{60}$$

and if we fix Y, and letting $m_Y : V \to \mathbb{R}$ be a modular function in X, we get

$$= f(X) - \sum_{j \in X \backslash Y} f(X) + \sum_{j \in X \backslash Y} f(X-j) + m_Y(X) \tag{61}$$

We now show that $\sum_{j \in X \backslash Y} f(X)$ and $\sum_{j \in X \backslash Y} f(X-j)$ are both submodular. Observe first that if f is non-increasing, then $g(X) = \sum_{j \in X \backslash Y} f(X)$ is submodular. Again consider $X \subseteq Z \subseteq V$ and $e \notin Z$. Then we have:

$$\begin{aligned} g(e|X) &= g(X \cup e) - g(X) \\ &= \sum_{j \in X \cup e \backslash Y} f(X \cup e) - \sum_{j \in X \backslash Y} f(X) \\ &= \sum_{j \in X \backslash Y} (f(e|X)) + I(e \notin Y) f(X \cup e) \end{aligned} \tag{62}$$

Now note that $X \backslash Y \subseteq Z \backslash Y$, and $f(e|X) \geq f(e|Z)$ since f is submodular. Both, however, are non-positive since f is non-increasing. Hence we have $\sum_{j \in X \backslash Y} f(e|X) \geq \sum_{j \in X \backslash Y} f(e|Z) \geq \sum_{j \in Z \backslash Y} f(e|Z)$. Finally note that again as f is non-increasing $f(X \cup e) \geq f(Z \cup e)$ and hence $g(e|X) \geq g(e|Z)$.

The proof that $\sum_{j \in X \backslash Y} f(X - j)$ is submodular in X is similar to the above except replacing X by $X - j$. Note that all the same steps will follow. The claim is proved.

The proof of the second part of the theorem for the submodular Bregman of the second type is very similar so we leave it for the reader. □

D.2 Proofs of theorems 3.2 and Theorem 3.3

Proof of Theorem 3.2

Proof. Observe the way we have defined the permutation based lower bound Bregman divergence. For a given Y, $\partial_f(Y)$ is a linear operator in f [11]. Hence the extreme points are linear operators of f and correspondingly h_Y is a linear operator of f. From this it directly follows from the definition that $d^{\mathcal{H}_f}_f(X,Y)$ is a linear operator in f.

For the modular upper bound based submodular Bregman, we see that the gain of a submodular function (i.e., $f(j|.)$) is a linear operator in f, and hence $f(j|X)$, $f(j|E-j)$, $f(j|\emptyset)$ and $f(j|X-\{j\})$ are all linear operators of f and hence the upper bound submodular Bregmans are linear operators in f. □

Proof of Theorem 3.3

Proof. Consider first the permutation based lower bound Bregman divergence d^{Σ}_f. Notice that $d^{\Sigma}_m(X,Y) = m(X) - h_{Y,\sigma}(X)$, where h_Y is the modular lower bound of $m(X)$ in Eqn. (7). However for any permutation σ, we have $\sum_{j \in X} h_Y(j) = \sum_{j \in X} m(j) = m(X)$, and correspondingly

$d_m(X,Y) = 0, \forall X, Y$. Hence $d_{f+m}(X,Y) = d_f^{\mathcal{H}_f}(X,Y)$ using the linearity of $d_f^{\mathcal{H}_f}$ given by Theorem 3.2.

In a similar manner we can show this for both of the upper bound Bregman divergences. Consider for example the first type for $f = m$. Then we have $f(X) - \sum_{j \in X \backslash Y} f(j|X-j) + \sum_{j \in Y \backslash X} f(j|\emptyset) - f(Y) = m(X) - m(X \backslash Y) + m(Y \backslash X) - m(Y) = 0$. A similar argument can be provided for the second type and hence again from Theorem 3.2's linearity, we have that $d^{f+m}(X,Y) = d_{1:3}^f(X,Y)$ □

D.3 Proof of Theorem 3.5 and Theorem 3.6

Proof of Theorem 3.5.

Proof. Consider first the generalized lower bound submodular Bregman. Given three sets X_1, X_2 and X_3, we can write:

$$\begin{aligned} d_f^{\mathcal{H}_f}(X_1,X_2) + d_f^{\mathcal{H}_f}(X_2,X_3) &= f(X_1) - f(X_2) - \langle h_{X_2}, 1_{X_1} - 1_{X_2} \rangle \\ &\quad + f(X_2) - f(X_3) - \langle h_{X_3}, 1_{X_2} - 1_{X_3} \rangle \\ &= f(X_1) - f(X_3) - \langle h_{X_3}, 1_{X_1} - 1_{X_3} \rangle \\ &\quad - \langle h_{X_2}, 1_{X_1} - 1_{X_2} \rangle + \langle h_{X_3}, 1_{X_1} - 1_{X_2} \rangle \\ &= d_f^{\mathcal{H}_f}(X_1,X_3) + \langle h_{X_3} - h_{X_2}, 1_{X_1} - 1_{X_2} \rangle \end{aligned} \tag{63}$$

To show the second part, we have the generalized upper bound submodular Bregman. (Eqn. (23)).

$$\begin{aligned} d_{\mathcal{G}^f}^f(X_1,X_2) + d_{\mathcal{G}^f}^f(X_2,X_3) &= f(X_1) - f(X_2) - \langle g_{X_1}, 1_{X_1} - 1_{X_2} \rangle \\ &\quad + f(X_2) - f(X_3) - \langle g_{X_2}, 1_{X_2} - 1_{X_3} \rangle \\ &= f(X_1) - f(X_3) - \langle g_{X_1}, 1_{X_1} - 1_{X_3} \rangle \\ &\quad - \langle g_{X_2}, 1_{X_2} - 1_{X_3} \rangle + \langle g_{X_1}, 1_{X_2} - 1_{X_3} \rangle \\ &= d_{\mathcal{G}^f}^f(X_1,X_3) + \langle g_{X_1} - g_{X_2}, 1_{X_2} - 1_{X_3} \rangle \end{aligned} \tag{64}$$

as desired. □

Proof of Theorem 3.6.

Proof. First define: $g = X \cap Y \cap Z$, $f = Y \cap Z \backslash g$, $e = X \cap Z \backslash g$, $d = X \cap Y \backslash g$, $a = X \setminus \{d, e, g\}$, $b = Y \setminus \{d, g, f\}$, and $c = Z \setminus \{e, g, f\}$. Consider now:

$$\begin{aligned} d_2^f(X,Y) + d_2^f(Y,Z) \quad &= \quad f(X) - f(Y) + \sum_{j \in Y \backslash X} f(j|X) - \sum_{j \in X \backslash Y} f(j|E-j) + f(Y) - f(Z) \\ &\quad + \sum_{j \in Z \backslash Y} f(j|Y) - \sum_{j \in Y \backslash Z} f(j|E-j) \\ &= \quad f(X) - f(Z) + \sum_{j \in Y \backslash X} f(j|X) + \sum_{j \in Z \backslash Y} f(j|Y) - \sum_{j \in Y \backslash Z} f(j|E-j) - \sum_{j \in X \backslash Y} f(j|E-j) \end{aligned}$$

Now we consider $d_2^f(X,Y) + d_2^f(Y,Z) - d_2^f(X,Z) = \sum_{j \in Y \backslash X} f(j|X) + \sum_{j \in Z \backslash Y} f(j|Y) - \sum_{j \in Z \backslash X} f(j|X) - \sum_{j \in Y \backslash Z} f(j|E-j) - \sum_{j \in X \backslash Y} f(j|E-j) + \sum_{j \in X \backslash Z} f(j|E-j) = \sum_{j \in b} f(j|X) + \sum_{j \in e} f(j|Y) + \sum_{j \in c}(f(j|Y) - f(j|X)) - \sum_{j \in b \cup e} f(j|E - \{j\})$. The last step can be verified from the Venn-diagram for sets X, Y, Z, and using the expressions for a, b, c, d, e, f and g. Hence for the triangle inequality, we need:

$$\sum_{j \in b \cup e} f(j|E - \{j\}) \leq \sum_{j \in b} f(j|X) + \sum_{j \in e} f(j|Y) + \sum_{j \in c}(f(j|Y) - f(j|X)) \tag{65}$$

For this to be true we require that $f(j|Y) \geq f(j|X), \forall j \in c$, which will be true if either $Y \subseteq X$ or $c = \emptyset$. Also we get an analogous expression from the triangle inequality for the upper bound submodular Bregman of second type:

$$\sum_{j \in b \cup e} f(j|\emptyset) \geq \sum_{j \in b} f(j|X) + \sum_{j \in e} f(j|Y) + \sum_{j \in d} (f(j|Y - \{j\}) - f(j|X - \{j\})) \qquad (66)$$

Thus here we need $f(j|Y - \{j\}) \leq f(j|X - \{j\}), \forall j \in d$, which will follow if $X \subseteq Y$ or $d = \emptyset$.

The proof of the triangle inequality of d_3^f follows in lines very similar to the above, and we leave it to the reader. □

D.4 Proofs of Theorem 3.7 and Theorem 3.8

Proof of Theorem 3.7

Proof. First observe that $h_Y \in \partial_f(Y)$ and hence $Y \in \partial_2 f(h_Y)$. Thus, directly from the definition, we have $d_{f^*}(h_Y, h_X) = f^*(h_Y) - f^*(h_X) - h_Y(X) + h_X(X)$. Further, also from [11], we have, $f^*(h_Y) = h_Y(Y) - f(Y)$ and $f^*(h_X) = h_X(X) - f(X)$. Therefore, we have: $d_{f^*}(h_Y, h_X) = h_Y(Y) - f(Y) - h_X(X) + f(X) - h_Y(X) + h_X(X) = f(X) - f(Y) - h_Y(X) + h_Y(Y) = d_f^{\mathcal{H}_f}(X, Y)$, as desired. □

Proof of Theorem 3.8

Proof. Consider the first expression and we start for the dual function, with the submodular Bregman of the first type. $d_1^{f^\#}(X, Y) = f^\#(X) - \sum_{j \in X \backslash Y} f^\#(j|X - \{j\}) + \sum_{j \in Y \backslash X} f^\#(j|\emptyset) - f^\#(Y) = f(V - X) - f(V) + \sum_{j \in X \backslash Y} f(j|V - X) - \sum_{j \in Y \backslash X} f(j|V - \{j\}) + f(V) - f(V - Y) = d_2^f(V - X, V - Y)$. We have used here the fact that $f^\#(j|X - \{j\}) = f^\#(X) - f^\#(X - \{j\}) = -f(V) + f(V - X) + f(V) - f(V - X + \{j\}) = -f(j|V - X)$. Similarly we can show that $f^\#(j|\emptyset) = f(j|V - \{j\})$. Note that we started with the submodular Bregman of the first type, for X and Y of the dual function $f^\#$, but obtained the submodular Bregman of the second type for $V - X, V - Y$ of f. The other expressions can be shown in a similar fashion. □

D.5 Proof of Theorem 3.9

Proof of Theorem 3.9

Proof. We first show the necessary and sufficient conditions of the generalized lower bound submodular Bregman. Notice that a generalized lower bound submodular Bregman satisfies both of the given properties (of being submodular in the first argument given the other, and that for given sets A, B, the difference $d(X, A) - d(X, B)$ is modular in X). Hence one side is direct. To show that any divergence satisfying these properties is a lower bound submodular Bregman, define $f(X) = d(X, \emptyset)$. Clearly f is a submodular function, and correspondingly, define $h_\emptyset = 0$. Now consider any $d(X, A)$. From the second property we know that for a modular function m, and any set function g, $d(X, A) - d(X, \emptyset) = m_A(X) + g(A) \Rightarrow d(X, A) = d(X, \emptyset) + m_A(X) + g(A)$. Now since $d(A, A) = 0$, we have $d(A, \emptyset) + m_A(A) + g(A) = f(A) + m_A(A) + g(A) = 0$. Hence we have $g(A) = -m_A(A) - f(A)$. Then we have that $d(X, A) = d(X, \emptyset) + m_A(X) + g(A) = f(X) + m_A(X) - m_A(A) - f(A)$. This is a lower bound submodular Bregman, using $h_A = -m_A$ for all sets A. Further we can see that since d is a valid divergence, h_A is a subgradient. The statement is proved. The necessary and sufficient conditions for the generalized upper bound based submodular Bregman can be showed exactly using the same approach like above.

Now consider the necessary and sufficient conditions of the permutation based lower bound submodular Bregman. Observe that the function $f_A(X)$ is submodular in X, and $f_A(X)$ and $f(X)$ differ in only a modular term and hence the corresponding submodular Bregman are identical. To show that a divergence satisfying these properties is a lower bound submodular Bregman, observe that d is directly a permutation based lower bound submodular Bregman characterized by a permutation σ and function f_A. Hence proved.

Finally, for the Nemhauser based upper bound submodular Bregman, observe that first the function f_A is supermodular. Further we also have: $f_A(X) = d^f_{1:3}(A, Y) = f(A) - f(X) + \text{modular}(X) + c$. Hence $f(X)$ and $-f_A(X)$ differ in only a modular term and hence the submodular Bregman associated with $f(X)$ and $-f_A(X)$ are identical. Clearly a function satisfying these properties is indeed a Nemhauser based upper bound submodular Bregman parameterized by a submodular function $-f_A(X)$. The statement is now proved. □

E Proofs in the Applications

E.1 Proof of theorem 4.2

Proof. The proof of this is direct from the definition. Define $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$.

$$\begin{aligned}
\mu &= \operatorname*{argmin}_{x \in [01]^n} \sum_{i=1}^n d_{\hat{f}}(x_i, x) \\
&= \operatorname*{argmin}_{x \in [01]^n} \sum_{i=1}^n \hat{f}(x_i) - \langle h_{x,\sigma_x}, x_i \rangle \\
&= \operatorname*{argmin}_{x \in [01]^n} \sum_{i=1}^n \hat{f}(x_i) - \langle h_{x,\sigma_x}, x_i \rangle \\
&= \operatorname*{argmin}_{x \in [01]^n} \sum_{i=1}^n \hat{f}(x_i) - \hat{f}(\hat{\mu}) + \hat{f}(m\hat{u}) - \langle h_{x,\sigma_x}, x_i \rangle \\
&= \operatorname*{argmin}_{x \in [01]^n} \sum_{i=1}^n \hat{f}(x_i) - \hat{f}(\hat{\mu}) + d_{\hat{f},\sigma_x}(\hat{\mu}, x) \qquad (67)
\end{aligned}$$

Hence $\mu = \hat{\mu}$ is a minimizer of equation (40). Correspondingly $\sigma = \sigma_\mu$ is the Lovász Bregman permutation representative. □

E.2 Proof of theorem-4.3

Proof of item 1:

Proof. Recall that we can define a continuous extension of the upper bound submodular Bregman $d^{\hat{f}}(1_X, y)$ from Eqn. (34), for y being a point in the interior of the hypercube. Define $\mu = \frac{1}{n}\sum_{i=1}^n 1_{X_i}$ as the continuous mean. Then we have:

$$\begin{aligned}
\text{argmin}_X \sum_{i=1}^n d^f_{\mathcal{G}^f}(X, X_i) &= \text{argmin}_X \sum_{i=1}^n d^{\hat{f}}(1_X, 1_{X_i}) \\
&= \text{argmin}_X \sum_{i=1}^n \Big(\hat{f}(1_X) - \hat{f}(1_{X_i}) - \langle g_X, 1_X - 1_{X_i} \rangle \Big) \\
&= \text{argmin}_X \sum_{i=1}^n \Big(\hat{f}(1_X) - \hat{f}(\mu) - \langle g_X, 1_X - \mu \rangle + \hat{f}(\mu) - \hat{f}(1_{X_i}) \Big) \\
&= \text{argmin}_X \hat{f}(1_X) - \hat{f}(\mu) - \langle g_X, 1_X - \mu \rangle = \text{argmin}_X d^{\hat{f}}(1_X, \mu)
\end{aligned}$$

This gives a nice relation that we need to find a hypercube vertex 1_X that is "closest" to the continuous mean μ through a divergence $d_{\hat{f}}$. It seems that rounding the continuous mean μ should provide us with the set X. We use this intuition and provide a generalized rounding procedure as follows. Recall then from Theorem 2.7, we have:

$$d^{\hat{f}}(1_X, \mu) = (1 - \mu(\sigma(1)))d^f_{\mathcal{G}^f}(X, \emptyset) + \sum_{i=1}^{n-1} (\mu(\sigma(i)) - \mu(\sigma(i+1))d^f_{\mathcal{G}^f}(X, U_i) + \mu(\sigma(n))d^f_{\mathcal{G}^f}(X, U_n)$$

where the sets U_i are obtained from ordering the elements of μ in decreasing order, on the basis of a permutation σ such that $U_i = [\sigma(1), \cdots, \sigma(i)]$.

Now from the intuition of vectors, we can see that the problem of finding X is now equivalent to finding the minimum of a weighted sum of $d_f^{\mathcal{H}_f}(X, U_i)$ with the sets U_i being a chain of sets. Thus we would expect the set X to be one of the sets U_i to minimize this sum.

□

Proof of item-2 and clustering using hamming distance:

This theorem actually follows from the relation of the generalized submodular Bregman and the generalized Bregman divergence through the Lovász extension from Theorem 2.6. Now observe that:

$$\begin{aligned}
\text{argmin}_X \sum_{i=1}^n d_f^{\mathcal{H}_f}(X_i, X) &= \text{argmin}_X \sum_{i=1}^n d_{\hat{f}}(1_{X_i}, 1_X) \\
&= \text{argmin}_X \sum_{i=1}^n \Big(\hat{f}(1_{X_i}) - \hat{f}(1_X) - \langle h_X, 1_{X_i} - 1_X \rangle \Big) \\
&= \text{argmin}_X \sum_{i=1}^n \Big(\hat{f}(1_X) - \hat{f}(\mu) - \langle h_X, 1_X - \mu \rangle + \hat{f}(\mu) - \hat{f}(1_{X_i}) \Big) \\
&= \text{argmin}_X \hat{f}(1_X) - \hat{f}(\mu) - \langle h_X, 1_X - \mu \rangle = \text{argmin}_X d_{\hat{f}}(1_X, \mu)
\end{aligned}$$

Thus again we have now a relation that the mean set X is the closest point in terms of the generalized Bregman divergence $\hat{f}$ of the continuous mean μ, and hence we approximate this problem by rounding the continuous mean.

Now consider clustering using the Hamming distance. We have:

$$\begin{aligned}
\underset{X}{\text{argmin}} \sum_i d_H(X, X_i) &= \underset{X}{\text{argmin}} \sum_i |X| + |X_i| - 2|X \cap X_i| \\
&= \underset{X}{\text{argmin}}\, n|X| - \sum_i 2|X \cap X_i| \\
&= \underset{X}{\text{argmin}} \langle 1_X, 1_V \rangle - 2\langle 1_X, \mu \rangle \\
&= \underset{X}{\text{argmin}} \langle 1_X, 1_V - 21_{X_i} \rangle
\end{aligned} \tag{68}$$

Clearly if $1 - 2\mu(j) \geq 0$, we have that $1_X(j) = 0$ and vice-verse. Thus we have $1_X(j) = 0$, if $\mu(j) \leq 0.5$, and hence X is equivalent to rounding the continuous mean μ at 0.5. Hence the generalized rounding procedure will give an exact result.