Subspace Selection via DR-Submodular Maximization on Lattices

So Nakashima,[1,2] **Takanori Maehara**[2]
[1]The University of Tokyo, [2]RIKEN Center for Advanced Intelligence Project
so_nakashima@mist.i.u-tokyo.ac.jp, takanori.maehara@riken.jp

Abstract

The *subspace selection problem* seeks a subspace that maximizes an objective function under some constraint. This problem includes several important machine learning problems such as the principal component analysis and sparse dictionary selection problem. Often, these problems can be (exactly or approximately) solved using greedy algorithms. Here, we are interested in why these problems can be solved by greedy algorithms, and what classes of objective functions and constraints admit this property.

In this study, we focus on the fact that the set of subspaces forms a *lattice*, then formulate the problems as optimization problems on lattices. Then, we introduce a new class of functions on lattices, *directional DR-submodular functions*, to characterize the approximability of problems. We prove that the principal component analysis, sparse dictionary selection problem, and these generalizations are monotone directional DR-submodularity functions. We also prove the "quantum version" of the cut function is a non-monotone directional DR submodular function. Using these results, we propose new solvable feature selection problems (generalized principal component analysis and quantum maximum cut problem), and improve the approximation ratio of the sparse dictionary selection problem in certain instances.

We show that, under several constraints, the directional DR-submodular function maximization problem can be solved efficiently with provable approximation factors.

1 Introduction

Background and motivation *The subspace selection problem* involves seeking a good subspace from data. Mathematically, the problem is formulated as follows. Let $\mathcal{L}$ be a family of subspaces of $\mathbb{R}^d$, $\mathcal{F} \subseteq \mathcal{L}$ be a set of feasible subspaces, and $f\colon \mathcal{L} \to \mathbb{R}$ be an objective function. Then, the task is to solve the following optimization problem.

$$\begin{array}{ll} \text{maximize} & f(X) \\ \text{subject to} & X \in \mathcal{F}. \end{array} \tag{1.1}$$

This problem is a kind of feature selection problem, and contains several important machine learning problems such as the principal component analysis and sparse dictionary selection problem.

In general, the subspace selection problem is a non-convex continuous optimization problem; hence it is hopeless to obtain a provable approximate solution. On the other hand, such solution can be obtained efficiently in some special cases. The most important example is the principal component analysis. Let $\mathcal{L}(\mathbb{R}^d)$ be the set of all the subspaces of $\mathbb{R}^d$, $\mathcal{F}$ be the subspaces with dimension of at most k, and $f\colon \mathcal{L} \to \mathbb{R}$ be the function defined by

$$f(X) = \sum_{i \in I} \|\Pi_X u_i\|^2 \tag{1.2}$$

where $\{u_i\}_{i \in I} \subset \mathbb{R}^d$ is the given data and Π_X is the projection to subspace X. Then, problem (1.1) with these $\mathcal{L}(\mathbb{R}^d)$, $\mathcal{F}$, and f defines the principal component analysis problem. As we know, the greedy algorithm, which iteratively selects a new direction $a_i \in \mathbb{R}^d$ that maximizes the objective function, gives the optimal solution to problem (1.1). Another important problem is the sparse dictionary selection problem. Let $V \subseteq \mathbb{R}^d$ be a set of vectors, called a dictionary. For a subset $S \subseteq V$, we denote by $\mathrm{span}(S)$ the subspace spanned by S. Let $\mathcal{L}(V) = \{\mathrm{span}(S) : S \subseteq V\}$ be the subspaces spanned by a subset of V, and $\mathcal{F}$ be the subspaces spanned by at most k vectors of V. Then, the problem (1.1) with these $\mathcal{L}(V)$, $\mathcal{F}$, and f in (1.2) defines the sparse dictionary selection problem. The problem is in general difficult to solve (Natarajan 1995); however, the greedy-type algorithms, e.g., *orthogonal matching pursuit*, yield provable approximation guarantees depending on the mutual coherence of V.

In this study, we are interested in the following research question:

> Why the principal component analysis and the sparse dictionary selection problem can be solved by the greedy algorithms? What classes of objective functions and constraints have the same property?

Existing approach Several researchers have considered this research question (see Related Work below). One successful approach is employing *submodularity*. Let $V \subseteq \mathbb{R}^d$ be a (possibly infinite) set of vectors. We define $F\colon 2^V \to \mathbb{R}$ by $F(S) = f(\mathrm{span}(S))$. If this function satisfies the submodularity, $F(S) + F(T) \geq F(S \cup T) + F(S \cap T)$, or some its approximation variants, we obtain a provable approximation guarantee of the greedy algorithm (Krause and Cevher 2010;

Das and Kempe 2011; Elenberg et al. 2016; Khanna et al. 2017).

However, this approach has a crucial issue that it cannot capture the structure of vector spaces. Consider three vectors $a = (1, 0)$, $b = (1/\sqrt{2}, 1/\sqrt{2})$, and $c = (0, 1)$ in $\mathbb{R}^2$. Then, we have $\mathrm{span}(\{a, b\}) = \mathrm{span}(\{b, c\}) = \mathrm{span}(\{c, a\})$; therefore, $F(\{a, b\}) = F(\{b, c\}) = F(\{c, a\})$. However, this property (a single subspace is spanned by different bases) is not exploited in the existing approach, which yields underestimation of the approximation factors of the greedy algorithms (see Section 4.2).

Our approach To capture the structure of vector spaces, we employ *Lattice Theory*. A *lattice* $\mathcal{L}$ is a partially ordered set closed under the greatest lower bound (aka., meet, $\wedge$) and the least upper bound (aka., join, $\vee$).

The family of all subspaces of $\mathbb{R}^d$ is called *the vector lattice* $\mathcal{L}(\mathbb{R}^d)$, which forms a lattice whose meet and join operators correspond to the intersection and direct sum of subspaces, respectively. This lattice can capture the structure of vector spaces as mentioned above. Also, the family of subspaces $\mathcal{L}(V)$ spanned by a subset of $V \subseteq \mathbb{R}^d$ forms a lattice.

Our goal is to establish a submodular maximization theory on lattice. Here, the main difficulty is to seek a "nice" definition of submodularity. Usually, the *lattice submodularity* (Topkis 1978), defined by the following inequality, is considered as a natural generalization of set submodularity.

$$f(X) + f(Y) \geq f(X \wedge Y) + f(X \vee Y). \tag{1.3}$$

However, this is too strong that it cannot capture the principal component analysis as shown below.

Example 1. Consider the vector lattice $\mathcal{L}(\mathbb{R}^2)$. Let $X = \mathrm{span}\{(1, 0)\}$ and $Y = \mathrm{span}\{(1, \epsilon)\}$ be subspaces of $\mathbb{R}^2$ where $\epsilon > 0$ is sufficiently small. Let $\{v_i\}_{i \in I} = \{(0, 1)\}$ be the given data, where $|I| = 1$. Then, function (1.2) satisfies $f(X) = 0$, $f(Y) = \epsilon^2/(1 + \epsilon^2)$, $f(X \wedge Y) = 0$, and $f(X \vee Y) = 1$. Therefore, it does not satisfy the lattice submodularity. More seriously, by taking $\epsilon \to 0$, we can see that there is no constants $\alpha > 0$ and $\delta \ll f(X) + f(Y)$ such that $f(X) + f(Y) \geq \alpha(f(X \wedge Y) + f(X \vee Y)) - \delta$. This means that it is hopeless to see this function as an approximated version of a lattice submodular function. □

Another commonly used notion of submodularity is the *diminishing return (DR)-submodularity* (Soma and Yoshida 2015; Bian et al. 2017b; Soma and Yoshida 2017), which is originally introduced in the integer lattice $\mathbb{Z}^V$. A function $f \colon \mathbb{Z}^V \to \mathbb{R}$ is DR-submodular if

$$f(X + e_i) - f(X) \geq f(Y + e_i) - f(Y) \tag{1.4}$$

for all $X \leq Y$ (component wise inequality) and $i \in V$, where e_i is the i-th unit vector. This definition is later extended to distributive lattices (Gottschalk and Peis 2015) and can be extended to general lattices (see Section 3). However, Example 1 above is still crucial, and therefore the objective function of the principal component analysis cannot be an approximated version of a DR-submodular function.

To summarize the above discussion, our main task is to define submodularity on lattices that should satisfy the following two properties:

1. It captures some important practical problems such as the principal component analysis.
2. It admits efficient approximation algorithms on some constraints.

Our contributions In this study, in response to the above two requirements, we make the following contributions:

1. We define *downward DR-submodularity* and *upward DR-submodularity* on *lattices*, which generalize the DR-submodularity on the integer lattice and distributive lattices (Section 3).
2. Our directional DR-submodularities are capable of representing important machine learning problems (Section 4).
   - We show that the objective function of the principal component analysis, and its concave generalization are directional DR-submodular functions. Using this result, we propose the *generalized principal component analysis*, which can limit the contributions from each data; hence, it will provide robust solution.
   - We show that the objective function of the sparse dictionary selection problem is approximately directional DR-submodular; where the approximation ratio depends on the mutual coherence of subspaces. The mutual coherence of subspaces are more robust concept than that of vectors. Thus, using this result, we can improve the approximation ratio of the problem for certain instance.
   - We show that the "quantum version" of the cut function of the graph is (non-monotone) directional DR-submodular function. Using this result, we propose a new feature selection method, which can be used to extract meanings of documents.
3. We propose approximation algorithms for maximizing (1) monotone downward DR-submodular function over height constraint, (2) monotone downward DR-submodular function over knapsack constraint, and (3) non-monotone DR-submodular function (Section 5). These are obtained by generalizing the existing algorithms for maximizing the submodular set functions. Thus, even our directional DR-submodularities are strictly weaker than the strong DR-submodularity, which is a natural generalization of the DR-submodularity on the integer lattice (Definition 3); it is sufficient to admit approximation algorithms.

All the proofs of propositions and theorems are given in Appendix Ain the supplementary material.

Related Work For the principal component analysis, it is well known that the greedy algorithm, which iteratively selects the largest eigenvectors of the correlation matrix, solves the problem (Abdi and Williams 2010).

With regard to the sparse dictionary selection problem, several studies (Gilbert, Muthukrishnan, and Strauss 2003; Tropp et al. 2003; Tropp 2004; Das and Kempe 2008) have analyzed greedy algorithms. In general, the objective function

for the sparse dictionary selection problem is not submodular. Therefore, researchers introduced approximated versions of the submodularity and analyzed the approximation guarantee of algorithms with respect to the parameter.

Krause and Cevher (Krause and Cevher 2010) showed that function (1.2) is an *approximately submodular* function whose additive gap $\delta \geq 0$ depends on the mutual coherence. They also showed that the greedy algorithm gives $(1-1/e, k\delta)$-approximate solution.[1]

Das and Kempe (Das and Kempe 2011) introduced the *submodularity ratio*, which is another measure of submodularity. For the set function maximization problem, the greedy algorithm attains a provable approximation guarantee depending on the submodularity ratio. The approximation ratio of the greedy algorithm is further improved by combining with the curvature (Bian et al. 2017a). Elenberg et al. (Elenberg et al. 2016) showed that, if function $l\colon \mathbb{R}^d \to \mathbb{R}$ has a bounded restricted convexity and a bounded smoothness, the corresponding set function $F(S) := l(0) - \min_{\mathrm{supp}(x)\in S} l(x)$ has a bounded submodularity ratio. Khanna et al. (Khanna et al. 2017) applied the submodularity ratio for the low-rank approximation problem.

It should be emphasized that all the existing studies analyzed the greedy algorithm as a function of a set of vectors (the basis of the subspace), instead of as a function of a subspace. This overlooks the structure of the subspaces causing difficulties as described above.

2 Preliminaries

In this section, we provide standard notion in Lattice Theory; see (Grätzer 2002) for more details.

A *lattice* $(\mathcal{L}, \leq)$ is a partially ordered set (poset) such that, for any $X, Y \in \mathcal{L}$, the least upper bound $X \vee Y := \inf\{Z \in \mathcal{L} : X \leq Z, Y \leq Z\}$ and the greatest lower bound $X \wedge Y := \sup\{Z \in L : Z \leq X, Z \leq Y\}$ uniquely exist. We often say "$\mathcal{L}$ is a lattice" by omitting $\leq$ if the order is clear from the context. In this paper, we assume that the lattice has the smallest element $\perp \in \mathcal{L}$.

A subset $\mathcal{I} \subseteq \mathcal{L}$ is a *lower set* if $Y \in \mathcal{I}$ then any $X \in \mathcal{L}$ with $X \leq Y$ is also $X \in \mathcal{I}$. For $Y \in \mathcal{L}$, the set $\mathcal{I}(Y) = \{X \in \mathcal{L} : X \leq Y\}$ is called the *lower set of* Y.

A sequence $X_1 < \cdots < X_k$ of elements of $\mathcal{L}$ is a *composition series* if there is no $Y \in \mathcal{L}$ such that $X_i < Y < X_{i+1}$ for all i. The length of the longest composition series from $\perp$ to X is referred to as the *height* of X and is denoted by $h(X)$. The height of a lattice is defined by $\sup_{X\in\mathcal{L}} h(X)$. If this value is finite, the lattice has the largest element $\top \in \mathcal{L}$. Note that the height of a lattice can be finite even if the lattice has infinitely many elements. For example, the height of the vector lattice $\mathcal{L}(\mathbb{R}^d)$ is d.

A lattice $\mathcal{L}$ is *distributive* if it satisfies the distributive law: $(X \wedge Y) \vee Z = (X \vee Z) \wedge (Y \vee Z)$. A lattice $\mathcal{L}$ is *modular* if it satisfies the modular law: $X \leq B \Rightarrow X \vee (A \wedge B) = (X \vee A) \wedge B$. Every distributive lattice is modular. On a modular lattice $\mathcal{L}$, all the composition series between $X \in \mathcal{L}$ and $Y \in \mathcal{L}$ have the same length. The lattice is modular if and only if its height function satisfies the modular equality: $h(X) + h(Y) = h(X \vee Y) + h(X \wedge Y)$. Modular lattices often appear with algebraic structures. For example, the set of all subspaces of a vector space forms a modular lattice. Similarly, the set of all normal subgroups of a group forms a modular lattice.

For a lattice $\mathcal{L}$, an element $a \in \mathcal{L} \setminus \{\perp\}$ is *join-irreducible* if there no $X \neq a, Y \neq a$ such that $a = X \vee Y$.[2] We denote by $J(\mathcal{L}) \subseteq \mathcal{L}$ the set of all join-irreducible elements. Any element $X \in \mathcal{L}$ is represented by a join of join-irreducible elements; therefore the structure of $\mathcal{L}$ is specified by the structure of $J(\mathcal{L})$. A join irreducible element $a \in J(\mathcal{L})$ is *admissible* with respect to an element $X \in \mathcal{L}$ if $a \not\leq X$ and any $a' \in \mathcal{L}$ with $a' < a$ satisfies $a' \leq X$. We denote by $\mathrm{adm}(X)$ the set of all admissible elements with respect to X. A set $\mathrm{cl}(a \mid X) = \{a' \in \mathrm{adm}(X) : X \vee a = X \vee a'\}$ is called a *closure* of a at X. See Figures 1 and 2 for the definition of admissible elements and closure. Note that a is admissible with respect to X if and only if the distance from the lower set of X to a is one.

Example 2. In the vector lattice $\mathcal{L}(\mathbb{R}^d)$, each element $X \in \mathcal{L}(\mathbb{R}^d)$ corresponds to a subspace. An element $a \in \mathcal{L}(\mathbb{R}^d)$ is join-irreducible if and only if it has dimension one. A join-irreducible element a is admissible to X if a is not a subspace of X. The closure $\mathrm{cl}(a|X)$ is the one dimensional subspaces contained in $X \vee a$ independent to X.

3 Directional DR-submodular functions on modular lattices

In this section, we introduce new submodularities on lattices. As described in Section 1, our task is to find useful definitions of "submodularities" on lattices; therefore, this is the most important section of this paper.

Recall definition (1.4) of the DR-submodularity on the integer lattice. Then, we can see that $X + e_i = X \vee a$ and $Y + e_i = Y \vee b$ for $a = (X_i + 1)e_i$ and $b = (Y_i + 1)e_i$, where X_i and Y_i are the i-th components of X and Y in, respectively. Here, a and b are join-irreducibles in the integer lattice, $a \in \mathrm{adm}(X)$, $b \in \mathrm{adm}(Y)$, and $a \leq b$. Thus, a natural generalization of the DR-submodularity on lattices may be as follows.

Definition 3 (Strong DR-submodularity)**.** A function $f\colon \mathcal{L} \to \mathbb{R}$ is *strong DR-submodular* if, for all $X, Y \in \mathcal{L}$ with $X \leq Y$ and $a \in \mathrm{adm}(X), b \in \mathrm{adm}(Y)$ with $a \leq b$, the following holds.

$$f(X \vee a) - f(X) \geq f(Y \vee b) - f(Y) \qquad (3.1)$$

This definition generalizes the DR-submodularity on distributive lattices (Gottschalk and Peis 2015) to general lattices using join-irreducibility and admissibility. This is itself non-trivial generalization, and may be useful in some context;

[1] A solution X is an (α, δ)-*approximate solution* if it satisfies $f(X) \geq \alpha \max_{X'\in\mathcal{F}} f(X') - \delta$. If $\delta = 0$ then we simply say that it is an α-approximate solution.

[2] For the set lattice 2^V of a set V, the join-irreducible elements correspond to the singleton sets, $\{a\}$ for $a \in V$. Thus, for clarity, we use upper case letters for general lattice elements (e.g., X or Y) and lower case letters for join-irreducible elements (e.g., a or b).

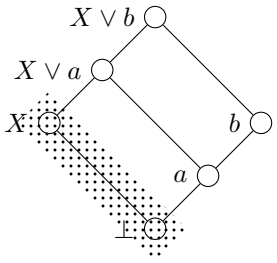

Figure 1: a is admissible with respect to X but b is not because of the existence of a. The shaded area represents the lower set of X.

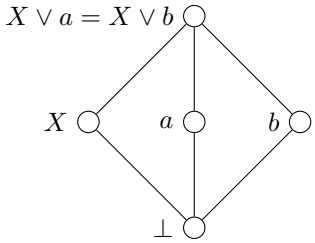

Figure 2: Both a and b are admissible with respect to X, and $X \vee a = X \vee b$. Thus $b \in \mathrm{cl}(a|X)$ and $a \in \mathrm{cl}(b|X)$.

however, in our purpose, it is too strong because it cannot capture the principal component analysis; you can confirm this by Example 1. Thus, here, we do not go into the details of this definition, and seek a weaker concept of DR-submodularity.

To define a weaker concept, we first review why the strong DR-submodularity is too strong. Since $Y \vee b = Y \vee b'$ for all $b' \in \mathrm{cl}(b|Y)$, we have $f(Y \vee b) - f(Y) = f(Y \vee b') - f(Y)$. Then, the strong DR-submodularity (3.1) is equivalent to the following: For all $X, Y \in \mathcal{L}$ with $X \leq Y$, $b \in \mathrm{adm}(Y)$, $b' \in \mathrm{cl}(b|Y)$, and $a \in \mathrm{adm}(X)$ with $a \leq b'$, the following holds.

$$f(Y \vee b) - f(Y) \leq f(X \vee a) - f(X). \tag{3.2}$$

This means that, even if there exists only one bad b', it does not satisfy the strong DR-submodularity.

On the other hand, in the standard analysis of the principal component analysis (Abdi and Williams 2010), we often use the fact that a new direction can be chosen to be orthogonal to the existing subspace; it is a relation between a subspace and "some b'" instead of "all b'". This observation derives the following new definition of the DR-submodularity, which relaxes "for all b'" to "exists b'" as follows.

Definition 4 (Downward DR-submodularity)**.** Let $\mathcal{L}$ be a lattice. A function $f\colon \mathcal{L} \to \mathbb{R}$ is *downward DR-submodular with additive gap* δ if for all $X, Y \in \mathcal{L}$ with $X \leq Y$ and $b \in \mathrm{adm}(Y)$, there exists $b' \in \mathrm{cl}(b|Y)$ such that, for all $a \in \mathrm{adm}(X)$ with $a \leq b'$, the following holds.

$$f(Y \vee b) - f(Y) \leq f(X \vee a) - f(X) + \delta. \tag{3.3}$$

The downward DR-submodularity is obtained by focusing on $Y \vee b$ in the right-hand side of (3.1). Another definition is obtained by focusing on $X \vee a$ in the left-hand side of (3.1) as follows. By renaming $Y \vee b$ to Y, the strong DR-submodularity (3.1) is equivalent to the following: For all $X, Y \in \mathcal{L}$, $a \in \mathrm{adm}(X)$ with $X \vee a \leq Y$, $b \in \mathcal{L}$ with $b \geq a$, and $\mathring{Y} \in \mathcal{L}$ with $Y = \mathring{Y} \vee b$ and $X \leq \mathring{Y}$,

$$f(X \vee a) - f(X) \geq f(Y) - f(\mathring{Y}). \tag{3.4}$$

As same as the downward DR-submodularity, we obtain new definition by relaxing "for all $\mathring{Y}$" to "exists $\mathring{Y}$" as follows.

Definition 5 (Upward DR-submodularity)**.** Let $\mathcal{L}$ be a lattice. $f\colon \mathcal{L} \to \mathbb{R}$ is *upward DR-submodular with additive gap* δ if for all $X, Y \in \mathcal{L}$ and $a \in \mathrm{adm}(X)$ with $X \vee a \leq Y$, $b \in \mathcal{L}$ with $b \geq a$, there exists $\mathring{Y} \in \mathcal{L}$ such that $Y = \mathring{Y} \vee b$ and $X \leq \mathring{Y}$, the following holds.

$$f(X \vee a) - f(X) \geq f(Y) - f(\mathring{Y}) - \delta. \tag{3.5}$$

If a function f is both downward DR-submodular with additive gap δ and upward DR-submodular with additive gap δ, then we say that f is *bidirectional DR-submodular* with additive gap δ. We say *directional DR-submodularity* to refer these new DR-submodularities.

The strong DR-submodularity implies the bidirectional DR-submodularity, because both downward and upward DR-submodularities are relaxations of the strong DR-submodularity. Interestingly, the converse also holds in distributive lattices.

Proposition 6. On a distributive lattice, the strong DR-submodularity, downward DR-submodularity, and upward DR-submodularity are equivalent.

Proof. See Appendix A. □

This implies that the directional DR-submodularities are required to capture the structure of non-distributive lattices such as the vector lattices.

At the cost of generalization, in contrast to the lattice submodularity (1.3) and the strong DR-submodularity (3.1), the downward and upward DR-submodularity are *not* closed under addition, because the elements selected by "exists" in the above definitions can depend on the objective function.

4 Examples

In this section, we present several examples of directional DR-submodular functions to show that our concepts *can* capture several machine learning problems.

4.1 Principal component analysis

Let $\{u_i\}_{i \in I} \subset \mathbb{R}^d$ be the given data. We consider the vector lattice $\mathcal{L}(\mathbb{R}^d)$ of all the subspaces of $\mathbb{R}^d$, and the objective function f defined by (1.2). Then, the following holds.

Proposition 7. The function $f\colon \mathcal{L}(\mathbb{R}^d) \to \mathbb{R}$ defined by (1.2) is a monotone bidirectional DR-submodular function.

Proof. See Appendix A. □

This provides a reason why the principal component analysis is solved by the greedy algorithm from the viewpoint of submodularity (see also Theorem 15 below).

The objective function can be generalized further. Let $\rho_i : \mathbb{R} \to \mathbb{R}$ be a monotone non-decreasing concave function with $\rho_i(0) = 0$ for each $i \in I$. Let

$$f_\rho(X) = \sum_{i \in I} \rho_i(\|\Pi_X u_i\|^2). \tag{4.1}$$

Then, the following holds.

Proposition 8. The function $f_\rho \colon \mathcal{L}(\mathbb{R}^d) \to \mathbb{R}$ defined by (4.1) is a monotone bidirectional DR-submodular function.

Proof. See Appendix A. □

Proposition 8 indicates a new feature selection problem, called the *generalized principal component analysis*, which maximizes the function (4.1) instead of the function (1.2) in (1.1) with the height (dimension) constraint. In this problem, we can ignore the contributions from very large vectors because, if u_i is already well approximated by X, there is less incentive to seek larger subspace for u_i due to the concavity of ρ_i. Hence, it will produce a robust solution.

We verified this idea by conducting numerical experiments on a synthetic dataset; see Appendix C.

Remark 9. Strictly speaking, Proposition 7 does not explain the reason why the principal component analysis can be solved "exactly" using the greedy algorithm, i.e., it only explains the approximate maximization. In Appendix B, we will explain the exact maximization using the concept of the *curvature* suited for the vector lattices.

4.2 Sparse dictionary selection

Let $V \subseteq \mathbb{R}^d$ be a set of vectors called a dictionary. We consider $\mathcal{L}(V) = \{\mathrm{span}(S) : S \subseteq V\}$ of all subspaces spanned by V, which forms a (not necessarily modular) lattice. The height of $X \in \mathcal{L}(V)$ coincides with the dimension of X. Let $\{u_i\}_{i \in I} \subset \mathbb{R}^d$ be the given data. Then the sparse dictionary selection problem is formulated by the maximization problem of f defined by (1.2) on this lattice under the height constraint.

In general, the function f is not a directional DR-submodular function on this lattice. However, we can prove that f is a downward DR-submodular function with a provable additive gap. We introduce the following definition.

Definition 10 (Mutual coherence of lattice)**.** Let $\mathcal{L}$ be a lattice of subspaces. For $\epsilon \geq 0$, the lattice has *mutual coherence* ϵ, if for any $X \in \mathcal{L}$, there exists $X' \in \mathcal{L}$ such that $X \wedge X' = \bot$, $X \vee X' = \top$, and for all unit vectors $u \in X$ and $u' \in X'$, $|\langle u, u' \rangle| \leq \epsilon$. The infimum of such ϵ is called the *mutual coherence of* $\mathcal{L}$, and is denoted by $\mu(\mathcal{L})$.

Our mutual coherence of a lattice is a generalization of the *mutual coherence* of a set of vectors (Donoho and Elad 2003). For a set of unit vectors $V = \{u_1, \ldots, u_N\} \subset \mathbb{R}^d$, its mutual coherence is defined by $\mu(V) = \max_{i \neq j} |\langle u_i, u_j \rangle|$. The mutual coherence of a set of vector is extensively used in compressed sensing to prove the uniqueness of the solution in a sparse recovery problem (Eldar and Kutyniok 2012). Here, we have the following relation between the mutual coherence of a lattice and that of a set of vectors, which is the reason why we named our quantity mutual coherence.

Lemma 11. Let $V = \{u_1, \ldots, u_N\}$ be a set of unit vectors whose mutual coherence is $\mu(V) \leq \epsilon$. Then, the lattice $\mathcal{L}(V)$ generated by the vectors has mutual coherence $\mu(\mathcal{L}(V)) \leq d\epsilon/(1 - d\epsilon)$. □

Proof. See Appendix A. □

This means that if a set of vectors has a small mutual coherence, then the lattice generated by the vectors has a small mutual coherence. Note that the converse does not hold.

Example 12. Consider $V = \{u_1, u_2, u_3\} \subset \mathbb{R}^2$ where $u_1 = (1, 0)^\top$, $u_2 = (1/\sqrt{1+\epsilon^2}, \epsilon/\sqrt{1+\epsilon^2})^\top$, and $u_3 = (0, 1)^\top$ for sufficiently small ϵ. Then the mutual coherence $\mu(V)$ of the vectors is $1/\sqrt{1+\epsilon^2} \approx 1$; however, the mutual coherence $\mu(\mathcal{L})$ of the lattice generated by V is $\epsilon/\sqrt{1+\epsilon^2} \approx \epsilon$.

If a lattice has a small mutual coherence, we can prove that the function f is a monotone downward DR-submodular function with a small additive gap.

Proposition 13. Let $V = \{u_1, \ldots, u_N\} \subseteq \mathbb{R}^d$ be normalized vectors and $\mathcal{L}(V)$ be a lattice generated by V. Suppose that $\mathcal{L}(V)$ forms a modular lattice. Let $\{v_i\}_{i \in I} \subset \mathbb{R}^d$. Then, the function f defined in (4.1) is a downward DR-submodular function with additive gap at most $3\epsilon\rho'(0) \sum_j \|v_j\|^2/(1-\epsilon^2)$ where $\epsilon = \mu(\mathcal{L}(V))$.

Proof. See Appendix A. □

Example 12 and Proposition 13 imply that, even set of vectors V contains some correlated vectors, the function f defined on the lattice can be an approximate submodular function with a small additive gap. On the other hand, if we consider the function defined on the subsets of V, it must have a large additive gap. These implies that (1) there is a situation that the lattice DR-submodularity improves the theoretical approximation guarantee of the greedy algorithm; (2) the lattice formulation is more robust against the contamination of highly correlated vectors. These are strong advantage of considering the lattice instead of the set of vectors.

4.3 Quantum cut

Finally, we present an example of a non-monotone bidirectional DR-submodular function. Let $G = (V, E)$ be a directed graph, and $c : E \to \mathbb{R}_{\geq 0}$ be a weight function. The cut function is then defined by $g(S) = \sum_{(i,j) \in E} c(i, j) 1[i \in S] 1[j \in \bar{S}]$ where $1[i \in S]$ is the indicator function of $i \in S$ and $\bar{S}$ is the complement of S. This is a non-monotone submodular function. Maximizing the cut function has application in feature selection problems with diversity (Lin, Bilmes, and Xie 2009).

We extend the cut function to the "quantum" setting. We say that a lattice of vector spaces $\mathcal{L}$ is *ortho-complementable* if $X \in \mathcal{L}$ then $X^\perp \in \mathcal{L}$ where $X^\perp$ is the orthogonal complement of X. Let $\{u_i\}_{i \in V} \subset \mathbb{R}^d$ be vectors assigned on each vertex. For an ortho-complementable lattice $\mathcal{L}$, the *quantum cut function* $f \colon \mathcal{L} \to \mathbb{R}$ is defined by

$$f(X) = \sum_{(i,j) \in E} c(i, j) \|\Pi_X(u_i)\|^2 \|\Pi_{X^\perp}(v_j)\|^2. \tag{4.2}$$

If $u_i = e_i \in \mathbb{R}^V$ for all i, where e_i is the i-th unit vector, and $\mathcal{L}$ is the lattice of axis-parallel subspaces of $\mathbb{R}^V$, function (4.2) coincides with the original cut function. Moreover, it carries the submodularity.

Proposition 14. The function f defined by (4.2) is a bidirectional DR-submodular function.

Proof. See Appendix A. □

Proposition 14 indicates a new feature selection problem, called *quantum maximum cut problem*, which is the problem of maximizing the function (4.2) as similar to the maximum cut based feature selection as mentioned above.

This problem will be useful in, e.g., natural language processing as follows. Suppose that we have the set of words, which are embedded into latent vector space $\mathbb{R}^d$ (Mikolov et al. 2013). We are given a document, which is a set of words. The task is to summarize the document. Usually, we summarize the document using a subset of words; hence, we can use the maximum cut problem. However, since the "meanings" of the words are encoded in the vector space as subspaces (Kim and de Marneffe 2013), it will be also promising to select a subspace in the latent space.

5 Algorithms

We provide algorithms for maximizing (1) a monotone downward-DR submodular function on the height constraint (Section 5.1), (2) a monotone downward DR-submodular function on knapsack constraint (Section 5.2), and (3) a non-monotone bidirectional DR-submodular function (Section 5.3). Basically, these algorithms are extensions of the algorithms for the set lattice. This indicates that our definitions of directional DR-submodularities are natural and useful.

Below, we always assume that f is normalized, i.e., $f(\bot) = 0$.

5.1 Height constraint

We first consider the height constraint, i.e., $\mathcal{F} = \{X \in \mathcal{L} : h(X) \leq k\}$. This coincides with the cardinality constraint when $\mathcal{L}$ is the set lattice. In general, this constraint is very difficult to analyze because $h(X \vee a) - h(X)$ can be arbitrary large. Thus, we assume that the height function is *p-incremental*, i.e., $h(X \vee a) - h(X) \leq p$ for all X and $a \in \mathrm{adm}(X)$. Note that $p = 1$ if and only if $\mathcal{L}$ is modular.

We show that, as similar to the set lattice (Nemhauser, Wolsey, and Fisher 1978), the greedy algorithm (Algorithm 1) achieves $1 - e^{-1/p}$ approximation for the downward DR-submodular maximization problem over the height constraint.

Theorem 15. Let $\mathcal{L}$ be a lattice whose height function is p-incremental, and $f\colon \mathcal{L} \to \mathbb{R}$ be a downward DR-submodular function with additive gap δ. Then, Algorithm 1 finds $(1 - e^{-\lfloor k/p \rfloor/k}, \delta(1 - e^{-\lfloor k/p \rfloor/k})k)$-approximate solution of the height constrained monotone submodular maximization problem.[3] In particular, on modular lattice with $\delta = 0$, it gives $1 - 1/e$ approximation.

[3] Algorithm 1 requires solving the non-convex optimization problem in Step 3. If we can only obtain an α-approximate solution in Step 3, the approximation ratio of the algorithm reduces to $(1 - e^{-\alpha\lfloor k/p \rfloor/k}, \delta(1 - e^{-\alpha\lfloor k/p \rfloor/k})k)$.

Algorithm 1 Greedy algorithm for monotone height constrained problem.

1: $X = \bot$
2: **for** $i = 1, \ldots, k$ **do**
3: Let $a_i \in \operatorname{argmax}_{a \in \mathrm{adm}(X), X \vee a \in \mathcal{F}} f(X \vee a)$
4: $X \leftarrow X \vee a_i$
5: **end for**
6: **return** X

Algorithm 2 Greedy algorithm for monotone knapsack constrained problem.

1: $X = \bot$
2: **for** $i = 1, 2, \ldots$ **do**
3: Let $a_i \in \operatorname{argmax}_{a \in \mathrm{adm}(X)} (f(X \vee a) - f(X))/(c(X \vee a) - c(X))$
4: **if** $c(X \vee a) \leq B$ **then** $X \leftarrow X \vee a_j$
5: **end for**
6: $a \in \operatorname{argmax}_{a \in \mathrm{adm}(\bot): c(a) \leq B} f(a)$
7: **return** $\operatorname{argmax}\{f(X), f(a)\}$

Proof. See Appendix A. □

This result characterizes the approximate solvability of the (generalized) principal component analysis and the sparse dictionary selection problem since the objective function (1.2) is a monotone downward DR-submodular function (Propositions 7, 8, and 13) and the height constraint coincides with the dimension constraint and the cardinality constraint on these lattices.

5.2 Knapsack constraint

Next, we consider the knapsack constrained problem. A knapsack constraint on a lattice is specified by a nonnegative mod-

Algorithm 3 Double-greedy algorithm for non-monotone unconstrained problem.

1: $A = \bot, B = \top$
2: **while** $A \neq B$ **do**
3: $\mathring{B} \leftarrow \operatorname{argmin}_{\mathring{B}} f(B) - f(\mathring{B})$ where $\mathring{B}$ runs over $A < \mathring{B} < B$ and $h(\mathring{B}) + 1 = h(B)$
4: $a \leftarrow \operatorname{argmax}_{a \in \mathrm{adm}(A), a \leq B} f(A \vee a) - f(A)$
5: **if** $f(A \vee a) - f(A) \geq f(\mathring{B}) - f(B)$ **then** $A \leftarrow A \vee a$ **else** $B \leftarrow \mathring{B}$
6: **end while**
7: **return** A

ular function (cost function) $c\colon \mathcal{L} \to \mathbb{R}_{\geq 0}$ and nonnegative number (budget) $B \in \mathbb{R}$ such that the feasible region is given by $\mathcal{F} = \{X \in \mathcal{L} : c(X) \leq B\}$.

In general, it is NP-hard to obtain a constant factor approximation for a knapsack constrained problem even for a distributive lattice (Gottschalk and Peis 2015). Therefore, we need additional assumptions on the cost function.

We say that a modular function $c\colon \mathcal{L} \to \mathbb{R}$ is *order-consistent* if $c(X \vee a) - c(X) \leq c(Y \vee b) - c(Y)$ for all $X, Y \in \mathcal{L}$, $a \in \mathrm{adm}(X)$, $b \in \mathrm{adm}(Y)$, and $a \leq b$. The height function of a modular lattice is order-consistent, because $c(X \vee a) - c(X) = 1$ for all $X \in \mathcal{L}$ and $a \in \mathrm{adm}(X)$; therefore it generalizes the height function. Moreover, on the set lattice 2^V, any modular function is order-consistent because there is no join-irreducible $a, b \in 2^V$ such that $a < b$ hold; therefore it generalizes the standard knapsack constraint on sets.

For a knapsack constraint with an order-consistent nonnegative modular function, we obtain a provable approximation ratio.

Theorem 16. Let $\mathcal{L}$ be a lattice, $\mathcal{F} = \{X \in \mathcal{L} : c(X) \leq B\}$ be a knapsack constraint where $c\colon \mathcal{L} \to \mathbb{R}_{\geq 0}$ be an order-consistent modular function, $B \in \mathbb{R}_{\geq 0}$, and $f\colon \mathcal{L} \to \mathbb{R}$ be a monotone downward DR-submodular function with additive gap δ. Then, Algorithm 2 gives $((1 - e^{-1})/2, \delta h(X^*)(1 - e^{-1})/2)$ approximation of the knapsack constrained monotone submodular maximization problem.

Proof. See Appendix A. □

In the set lattice, the knapsack constrained problem is solved in approximation ratio of $1 - 1/e$ (Sviridenko 2004). The algorithm first guesses the three largest elements in the optimal solution, then performs the greedy algorithm for the remaining elements. Generalizing this algorithm for general lattices is non-trivial because there are many ways to represent the optimal solution; thus, it is not clear that what should be guessed. Thus, finding an algorithm with better approximation ratio is a future work.

5.3 Non-monotone unconstrained maximization

Finally, we consider the unconstrained non-monotone maximization problem.

In the set lattice and distributive lattices, the double greedy algorithm (Buchbinder et al. 2015; Gottschalk and Peis 2015) achieves the deterministic $1/3$ and randomized $1/2$ approximation ratio for the unconstrained non-monotone submodular maximization problem. We generalize the deterministic version of the double greedy algorithm to lattices. We assume that the lattice has a finite height. This is needed to terminate the algorithm in a finite step. We also assume *both* downward DR-submodularity and upward DR-submodularity, i.e., bidirectional DR-submodularity. Finally, we assume that the lattice is modular. This is needed to analyze the approximation guarantee. Then, we obtain the following result.

Theorem 17. Let $\mathcal{L}$ be a modular lattice of finite height, $\mathcal{F} = \mathcal{L}$, and $f\colon \mathcal{L} \to \mathbb{R}_{\geq 0}$ be non-monotone bidirectional DR-submodular function with additive gap δ. Then, Algorithm 3 gives $(1/3, \delta h(\mathcal{L}))$ approximate solution of the unconstrained non-monotone submodular maximization problem.

Proof. See Appendix A. □

Generalizing the randomized algorithms, which attain the approximation ratio of $1/2$, is non-trivial because lemmas used in the analyses are difficult to generalize for (modular) lattices. Very recently a deterministic optimal $1/2$ approximation algorithm has been proposed (Buchbinder and Feldman 2018); however, it is also non-trivial to generalize this algorithm for lattices. Proposing an algorithm with the optimal approximation ratio is a promising future work.

6 Conclusion

In this paper, we formulated the subspace selection problem as optimization problem over lattices. By introducing new "DR-submodularities" on lattices, named *directional DR-submodularities*, we successfully identified the solvable subspace selection problem in terms of the submodularity. In particular, we found the objective functions of the principal component analysis and sparse dictionary selection problem are directional DR-submodular functions in their lattices. Also, we found a "quantum" version of the cut function is a directional DR-submodular function. These results motivate new feature selection problems, generalized principal component analysis and quantum maximum cut problem, and improves the approximation ratio for sparse dictionary problem in certain instances. We proposed approximation algorithms for maximizing directional DR-submodular functions on several constraints.

There are several interesting future directions. First, in the knapsack constraint problem and the non-monotone unconstrained problem, our results on general lattices have worse approximation ratios than the results on the integer lattice or distributive lattices. Filling these gaps are the most promising future works.

Second, in the set lattices case, a *matroid constraint* is considered as an important setting. Thus, it will promising to establish the lattice counterpart of this constraint; here, the difficulty may appear how to define matroid in general lattice (Barnabei, Nicoletti, and Pezzoli 1998). Related with this direction, in the set lattice case, the *continuous relaxation*-type algorithms play fundamental role in various constraints, including the matroid constraint (Calinescu et al. 2011; Vondrák, Chekuri, and Zenklusen 2011; Hassani, Soltanolkotabi, and Karbasi 2017; Bian et al. 2017a). Generalizing this technique to lattice will connect submodular maximization and the CAT(0) space theory, which is recently discussed in submodular minimization area (Hamada and Hirai 2017).

Finally, it is also an interesting direction to look for machine learning applications of the directional DR-submodular maximization other than the subspace selection problem. The possible candidates include the subgroup selection problem and the subpartition selection problem, which may be naturally formulated as optimization problems on lattices.

References

Abdi, H., and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2(4):433–459.

Barnabei, M.; Nicoletti, G.; and Pezzoli, L. 1998. Matroids on partially ordered sets. *Advances in Applied Mathematics* 21(1):78 – 112.

Bian, A. A.; Buhmann, J. M.; Krause, A.; and Tschiatschek, S. 2017a. Guarantees for greedy maximization of non-submodular functions with applications. In *International Conference on Machine Learning (ICML'17)*.

Bian, A. A.; Mirzasoleiman, B.; Buhmann, J.; and Krause, A. 2017b. Guaranteed Non-convex Optimization: Submodular Maximization over Continuous Domains. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, 111–120.

Buchbinder, N., and Feldman, M. 2018. Deterministic algorithms for submodular maximization problems. *ACM Transactions on Algorithms (TALG)* 14(3):32.

Buchbinder, N.; Feldman, M.; Seffi, J.; and Schwartz, R. 2015. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing* 44(5):1384–1402.

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

Das, A., and Kempe, D. 2008. Algorithms for subset selection in linear regression. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC'08)*, 45–54.

Das, A., and Kempe, D. 2011. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *Proceedings of the 28th International Conference on Machine Learning (ICML'11)* 1057–1064.

Donoho, D. L., and Elad, M. 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences* 100(5):2197–2202.

Eldar, Y. C., and Kutyniok, G. 2012. *Compressed sensing: theory and applications*. Cambridge University Press.

Elenberg, E. R.; Khanna, R.; Dimakis, A. G.; and Negahban, S. 2016. Restricted strong convexity implies weak submodularity. *arXiv preprint arXiv:1612.00804*.

Gilbert, A. C.; Muthukrishnan, S.; and Strauss, M. J. 2003. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the 14th ACM-SIAM Symposium on Discrete algorithms (SODA'03)*, 243–252.

Gottschalk, C., and Peis, B. 2015. Submodular function maximization over distributive and integer lattices. *arXiv preprint arXiv:1505.05423*.

Grätzer, G. 2002. *General lattice theory*. Springer Science & Business Media.

Hamada, M., and Hirai, H. 2017. Maximum vanishing subspace problem, CAT(0)-space relaxation, and block-triangularization of partitioned matrix. *arXiv preprint arXiv:1705.02060*.

Hassani, H.; Soltanolkotabi, M.; and Karbasi, A. 2017. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems (NIPS'17)*, 5843–5853.

Khanna, R.; Elenberg, E. R.; Dimakis, A. G.; Ghosh, J.; and Negahban, S. 2017. On approximation guarantees for greedy low rank optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, 1837–1846.

Kim, J.-K., and de Marneffe, M.-C. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, 1625–1630.

Krause, A., and Cevher, V. 2010. Submodular dictionary selection for sparse representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, 567–574.

Lin, H.; Bilmes, J.; and Xie, S. 2009. Graph-based submodular selection for extractive summarization. In *In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU'09)*, 381–386. IEEE.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS'13)*, 3111–3119.

Natarajan, B. K. 1995. Sparse approximate solutions to linear systems. *SIAM Journal on Computing* 24(2):227–234.

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

Soma, T., and Yoshida, Y. 2015. A generalization of submodular cover via the diminishing return property on the integer lattice. In *Advances in Neural Information Processing Systems (NIPS'15)*, 847–855.

Soma, T., and Yoshida, Y. 2017. Non-monotone dr-submodular function maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'17)*, volume 17, 898–904.

Sviridenko, M. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters* 32(1):41–43.

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

Tropp, J. A.; Gilbert, A. C.; Muthukrishnan, S.; and Strauss, M. J. 2003. Improved sparse approximation over quasiincoherent dictionaries. In *Proceedings of the International Conference on Image Processing (ICIP'03)*, volume 1, I–37. IEEE.

Tropp, J. A. 2004. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory* 50(10):2231–2242.

Vondrák, J.; Chekuri, C.; and Zenklusen, R. 2011. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 783–792.