Foundations and Trends® in Optimization

Atomic Decomposition via Polar Alignment

The Geometry of Structured Optimization

Suggested Citation: Zhenan Fan, Halyun Jeong, Yifan Sun and Michael P. Friedlander (2020), "Atomic Decomposition via Polar Alignment", Foundations and Trends® in Optimization: Vol. 3, No. 4, pp 280–366. DOI: 10.1561/2400000028.

Zhenan Fan
University of British Columbia
Canada
zhenanf@cs.ubc.ca

Halyun Jeong
University of British Columbia
Canada
clatar1@gmail.com

Yifan Sun
Stony Brook University
USA
yifan.sun@stonybrook.edu

Michael P. Friedlander
University of British Columbia
Canada
michael.friedlander@ubc.ca

now
the essence of knowledge
Boston — Delft

Contents

Atomic Decomposition via Polar Alignment

Zhenan Fan[1], Halyun Jeong[2], Yifan Sun[3] and Michael P. Friedlander[4]

[1] *University of British Columbia, Canada; zhenanf@cs.ubc.ca*
[2] *University of British Columbia, Canada; clatar1@gmail.com*
[3] *Stony Brook University, USA; yifan.sun@stonybrook.edu*
[4] *University of British Columbia, Canada; michael.friedlander@ubc.ca*

ABSTRACT

Structured optimization uses a prescribed set of atoms to assemble a solution that fits a model to data. Polarity, which extends the familiar notion of orthogonality from linear sets to general convex sets, plays a special role in a simple and geometric form of convex duality. This duality correspondence yields a general notion of alignment that leads to an intuitive and complete description of how atoms participate in the final decomposition of the solution. The resulting geometric perspective leads to variations of existing algorithms effective for large-scale problems. We illustrate these ideas with many examples, including applications in matrix completion and morphological component analysis for the separation of mixtures of signals.

Zhenan Fan, Halyun Jeong, Yifan Sun and Michael P. Friedlander (2020), "Atomic Decomposition via Polar Alignment", Foundations and Trends® in Optimization: Vol. 3, No. 4, pp 280–366. DOI: 10.1561/2400000028.

1

Introduction

Convex optimization provides a valuable computational framework that renders many problems tractable because of the range of powerful algorithms that can be brought to the task. The key is that a certain mathematical structure—i.e., convexity of the functions and sets defining the problem—lays open an enormous range of theoretical and algorithmic tools that lend themselves astonishingly well to computation. There are limits, however, to the scalability of general-purpose algorithms for convex optimization. As has been recognized in the optimization and related communities for at least the past decade, significant efficiencies can be gained by acknowledging the latent structure in the solution itself, coupled with the overarching structure provided by convexity.

Structured optimization proceeds along these lines by using a prescribed set of atoms from which to assemble an optimal solution. In effect, the atoms selected to participate in forming a solution decompose the model into simpler parts, which offers opportunities for algorithmic efficiency in solving the optimization problem. From a modeling point of view, the particular atoms that constitute the computed solution often represent key explanatory components of a model. An atomic decomposition thus provides a description of the most informative features of a solution—in other words, a kind of generalized principal component analysis.

Our purpose with this monograph is to describe the rich convex geometry that underlies atomic decomposition. The path we follow builds on the duality inherent in convex cones: every convex cone is paired uniquely with another cone that is polar to it. The extreme rays of each cone in this pair are in some sense *aligned*. Brought into the context of atomic decomposition, this notion of alignment through the polar operation provides a theoretical framework that can be harnessed to identify the atoms that participate in a decomposition. This approach facilitates certain algorithmic design patterns that promote computational efficiency, as we demonstrate with concrete examples. Similar computational economies accrue within reduced-space active-set methods for optimization problems with inequality constraints, such as implemented by the MINOS software package [1].

Early work in structured optimization focused on problem formulations meant to produce sparse solution vectors, i.e., a solution with relatively few non-zero elements. Compressed sensing [2]–[4] and model selection [5], [6], with their many applications in signal processing and statistics, helped to establish sparse optimization as an important class of problems with a range of specialized algorithms. Generalizations that accommodated different notions of sparsity soon followed, including matrix problems with low-rank solutions (sparsity in the vector of singular values), fused index pairs (sparsity in terms of the norms of subgroups of variables), and sparsity in specialized dictionaries, such as mass spectrographs of simple molecules used to represent structures of more complicated molecules [7, Section 6.3.1].

Nonsmooth regularization functions that promote sparsity, such as the 1-norm for sparse vectors, or the nuclear norm for low-rank matrices, are key features of these formulations. Gauge functions, which significantly generalize the notion of a norm, were recognized as flexible regularization functions that promote a broad range of sparse structures. By defining a set of atoms from which to build a solution, an almost arbitrary set of solution structures can be considered. The gauge function to this set can be incorporated into a convex optimization problem in order to obtain a solution with the desired structure. The convex analysis of gauges and support functions, which are their dual counterparts, is rich in geometry and rife with opportunity for efficient algorithm implementations for high-dimensional problems. Our purpose with this monograph is to expose the basic elements of this theory and its many connections to sparse and structured optimization. To make it accessible to researchers who are not specialists in convex analysis, we chose a largely self-contained treatment and make a few modest assumptions that greatly simplify the derivations.

1.1 Applications and Prior Work

One of the main implications of our approach is its usefulness in adapting dual optimization methods for discovering atomic decompositions. With the tools of polar alignment, a dual optimization method can be interpreted as solving for an aligning dual vector z that exposes the support of a primal solution x. If the number of exposed atoms is small, a solution x of the primal problem can be obtained from a reduced problem defined over the exposed support, but without the nonsmooth atomic regularization. The resulting reduced problem is often computationally much cheaper [8] and better conditioned [9]. Alternatively, two-metric methods can be designed to act differently on a primal iterate's suspected support [10]. In many applications, such as feature selection, knowing the optimal support may itself be sufficient. As we illustrate through various examples, there are several important cases where the dual aligning vector z can be computed directly.

Machine Learning. The regularized optimization problems described in Section 5 frequently appear in applications of machine learning for the purpose of model complexity reduction. The most popular tools are the vector 1-norm in feature selection [5], its group-norm variant [11], and the nuclear norm in matrix completion [12]. Many other sparsity-promoting regularizers, however, appear in practice [13]. Although unconstrained formulations are most popular, particularly when the proximal operator is computationally convenient [14], the gauge-constrained formulation is frequently used and solved via the conditional gradient method [15]–[17]. Popular dual methods, which iterate over a dual variable $z^{(k)}$ but maintain the corresponding primal variable $x^{(k)}$ only implicitly, include bundle methods [18] and dual averaging [19], [20].

Linear Conic Optimization. Conic programs are a cornerstone of convex optimization. The nonnegative cone, the second-order cone and the semidefinite cone respectively, give rise to linear, second-order, and semidefinite programs. These problem classes capture an enormous range of important models, and can be solved efficiently by a variety of algorithms, including interior methods [21]–[23]. Conic programs and their associated solvers are key ingredients for general purpose optimization software packages such as YALMIP [24] and CVX [25]. The alignment conditions for these specific cones have been exploited in dual methods, such as in the spectral bundle method for large-scale semidefinite programming [26]. Example 3.6 demonstrates this alignment principle in the context of conic optimization.

Gauge Optimization. The class of gauge optimization problems, as defined by Freund's 1987 seminal work [27], can be simply stated: find the element of a convex set that is minimal with respect to a gauge function. These conceptually simple problems appear in a remarkable array of applications, and include parts of sparse optimization and all of conic optimization [28, Example 1.3]. This class of optimization problems admits a duality relationship different from classical Lagrange duality, and is founded on the polar inequality. In this context, the polar inequality provides an analogue to weak duality, well-known in Lagrange duality, which guarantees that any feasible primal value provides an upper bound for any feasible dual value. In the gauge optimization context, a primal-dual pair (x, z) is optimal if and only if the polar inequality holds as an equation, which under Definition 2.4 implies that x and z are aligned. The connection between polar alignment and optimality is discussed further in Subsection 5.2.

Two-Stage Methods. In sparse optimization, two-stage methods first identify the primal variable support, and then solve the problem over a reduced support [29], [30]. If the support is sparse enough, the second problem may be computationally much cheaper because it can allow for faster Newton-like methods. The atomic alignment principles we describe in Section 4 give a general recipe for extracting primal variable support from a computed dual variable, which at optimality is aligned with the primal variable; see Section 5. This property forms the basis for our approach to morphological component analysis, described in Subsection 7.4.

Method Interpretability. The connection between sparsity and alignment points to a likely "aligning behavior" in many of the most effective methods for sparse optimization [31]. Indeed, we show in Section 6 that this is true for a range of methods, including proximal gradient, conditional gradient, and cutting-plane methods. Surprisingly, we also find hints of aligning behavior in seemingly unrelated methods, such as augmented Lagrangian and bundle methods. The alignment point of view thus offers greater interpretability of commonly used methods in many modern optimization applications.

1.2 Basic Definitions and Notation

We work with n-vectors in $\mathbb{R}^n$ and p-by-n matrices in $\mathbb{R}^{p\times n}$. The restriction to real-valued vectors and matrices considerably simplifies our development, though many of the ideas set forth in this monograph extend to more general functional spaces, as described by Zălinescu [32] and Bauschke and Combettes [33].

Vectors are always denoted by lower-case letters; matrices by capital letters. A vector norm $\|x\|$ always refers to the 2-norm, unless otherwise specified. Matrix norms always refer to the Schatten norm, e.g., if $(s_1, s_2, \ldots)$ are the singular values of X, then

$$\|X\|_1 = \sum_i s_i, \quad \|X\|_2 = \Big(\sum_i s_i^2\Big)^{1/2}, \quad \text{and} \quad \|X\|_\infty = \max_i s_i.$$

Let e_i denote the ith canonical unit vector, i.e., the vector of all zeros except a single 1 in the ith position. The dot product of two n-vectors x and z is $\langle x, z\rangle = \sum_j x_j z_j$. The dot product of two p-by-n matrices

X and Z is the trace inner product $\langle X, Z\rangle = \mathrm{tr}(X^TZ) = \sum_{ij} X_{ij}Z_{ij}$. The adjoint F^* of any linear map F is the unique linear map that satisfies the relationship $\langle Fx, z\rangle = \langle x, F^*z\rangle$ for all x and z. Thus, for the linear map F: $\mathbb{R}^n \to \mathbb{R}^m$, the product of the adjoint and an m-vector y is $F^*y = \sum_{i=1}^m y_i(Fe_i)$. For the linear map $\mathcal{F}$: $\mathbb{R}^{p\times n} \to \mathbb{R}^m$, the forward and adjoint maps take the form

$$\mathcal{F}X = \begin{bmatrix} \langle F_1, X\rangle \\ \vdots \\ \langle F_m, X\rangle \end{bmatrix} \quad \text{and} \quad \mathcal{F}^*y = \sum_{i=1}^m y_iF_i, \tag{1.1}$$

where each $F_1, \dots, F_m$ is a p-by-n matrix. The notation $X \succeq 0$ indicates that X is symmetric positive definite.

Throughout the monograph, we use the symbol $\mathcal{C}$ to denote a convex set in $\mathbb{R}^n$. The convex hull of any set $\mathcal{D}$ in $\mathbb{R}^n$ contains all weighted averages of the elements of the set, denoted

$$\mathrm{conv}\,\mathcal{D} = \left\{ \sum_{i=1}^m \alpha_i x_i \;\middle|\; x_i \in \mathcal{D},\ \alpha_i \geq 0,\ \sum_{i=1}^m \alpha_i = 1 \right\},$$

for some positive integer m. Define the conic extension of $\mathcal{D}$ by

$$\mathrm{cone}\,\mathcal{D} = \{\alpha d \mid d \in \mathcal{D},\ \alpha \geq 0\}.$$

The closure, boundary and relative interior, respectively, of $\mathcal{D}$ denoted $\mathrm{cl}\,\mathcal{D}$, $\mathrm{bnd}\,\mathcal{D}$ and $\mathrm{ri}\,\mathcal{D}$. The indicator to $\mathcal{D}$ is the function

$$\delta_{\mathcal{D}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{D}; \\ +\infty & \text{otherwise.} \end{cases}$$

The normal cone to the set $\mathcal{C}$ at $x \in \mathcal{C}$ is defined as

$$\mathcal{N}_{\mathcal{C}}(x) = \{d \mid \langle d, u - x\rangle \leq 0 \text{ for all } u \in \mathcal{C}\}.$$

The Euclidean projection onto the set $\mathcal{C}$ is denoted

$$\mathrm{proj}_{\mathcal{C}}(x) = \operatorname*{arg\,min}_{u\in\mathcal{C}} \|x - u\|_2,$$

which defines the distance of a point to the set $\mathcal{C}$, denoted by

$$\mathrm{dist}_{\mathcal{C}}(x) = \|x - \mathrm{proj}_{\mathcal{C}}(x)\|_2.$$

Let $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any function. The domain is denoted $\operatorname{dom} f = \{x \mid f(x) < +\infty\}$, and the convex conjugate is denoted

$$f^*(z) = \sup_{x \in \mathbb{R}^n} \{\langle x, z\rangle - f(x)\}\,.$$

2

Atomic Decomposition

The atomic decomposition of a vector $x \in \mathbb{R}^n$ with respect to an atomic set $\mathcal{A} \subset \mathbb{R}^n$ is given by the weighted superposition

$$x = \sum_{a \in \mathcal{A}} c_a a, \quad c_a \geq 0 \ \forall a \in \mathcal{A}. \tag{2.1}$$

Each coefficient c_a in the atomic decomposition measures the contribution of the corresponding atom a toward the representation of x. Intuitively, an atomic decomposition reveals structural information implicit in a vector, with large coefficients in the decomposition indicating the more significant structures. Within the context of an optimization problem, the atomic decomposition reveals structural elements most significant in the minimization process. In the simplest case, the atoms $\mathcal{A}$ may be formed from the collection of signed canonical unit vectors $\{\pm e_1, \ldots, \pm e_n\}$, which leads to the atomic decomposition

$$x = \sum_{j=1}^{n} c_j a_j, \quad c_j := |x_j|, \quad a_j := (\operatorname{sgn} x_j) \cdot e_j.$$

Trivially, the most significant atoms thus correspond to the variables x_j in the vector $x = (x_1, \ldots, x_n)$ with the largest magnitude. The definition given by (2.1) allows for decompositions with respect to arbitrary atomic sets, including atomic sets that are uncountably infinite.

This generic model for atomic decompositions was promoted by Chen *et al.* [2], [3] in the context of sparse signal decomposition, and more recently, by Chandrasekaran *et al.* [34], who were concerned with obtaining sparse solutions to linear inverse problems.

We are particularly interested in the question of determining which of the atoms in $\mathcal{A}$ are essential to the atomic decomposition of x, and conversely, which atoms can be safely ignored.

2.1 Gauge Functions Reveal the Atomic Support

The gauge function to the atomic set $\mathcal{A}$, defined by

$$\gamma_{\mathcal{A}}(x) = \inf_{c_a} \left\{ \sum_{a \in \mathcal{A}} c_a \;\middle|\; x = \sum_{a \in \mathcal{A}} c_a a,\ c_a \geq 0\ \forall a \in \mathcal{A} \right\}, \tag{2.2}$$

helps to define the answer to the question above. This function returns the minimal sum of weights over all valid atomic decompositions of x with respect to the set $\mathcal{A}$. The value $\gamma_{\mathcal{A}}(x) = \infty$ indicates that there doesn't exist a valid atomic decomposition for x. (For example, an atomic set composed of nonnegative vectors cannot decompose a vector x that contains negative entries.)

In the framework outlined by Chandrasekaran *et al.* [34], the gauge function $\gamma_{\mathcal{A}}$ defines the objective of a convex optimization problem suitable for recovering a signal from a small number of partial observations. In that context, the number of atoms needed to decompose the signal determines the number of observations needed to reconstruct the signal.

The significant atoms—those that *support* the vector x—are those that contribute positively in forming the minimal sum. We are thus led to the following definition.

Definition 2.1 (Atomic Support)**.** A subset of atoms $\mathcal{S}_{\mathcal{A}}(x) \subset \mathcal{A}$ is a *support set* for x with respect to $\mathcal{A}$ if every atom $a \in \mathcal{S}_{\mathcal{A}}(x)$ has a strictly positive coefficient c_a in the atomic decomposition (2.1). That is,

$$\gamma_{\mathcal{A}}(x) = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a, \qquad x = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a a, \quad \text{and} \quad c_a > 0\ \forall a \in \mathcal{S}_{\mathcal{A}}(x). \tag{2.3}$$

The set $\mathbf{supp}_{\mathcal{A}}(x)$ is defined as the set of all support sets. Thus, any set in $\mathbf{supp}_{\mathcal{A}}(x)$ is a valid support set. □

2.2 Polar Inequality

How do we identify the support of a vector x with respect to an arbitrary atomic set $\mathcal{A}$? The direct approach requires us to solve the minimum-weight problem defined by the gauge (2.2), and determine a valid decomposition from the positive elements of the computed solution. Thus, to identify all possible atomic support sets, we need to compute all possible solutions to (2.2). As we will demonstrate, however, a complete description of all possible solution sets can be obtained using the concept of *polar alignment*, which we define in this subsection. Our approach is based on a certain duality correspondence particular to gauge functions and to convex cones that are implicit in their definition. We describe in Section 3 this correspondence and its relationship to atomic decompositions. Here we give only the basic elements needed to define the notion of alignment.

Throughout the monograph, let

$$\widehat{\mathcal{A}} := \operatorname{cl}\operatorname{conv}(\mathcal{A} \cup \{0\})$$

denote the closed convex hull of the atomic set $\mathcal{A}$ adjoined with the origin. The Minkowski functional to the atomic set provides an equivalent expression for the definition (2.2) of the gauge:

$$\gamma_{\mathcal{A}}(x) = \inf\{\lambda \geq 0 \mid x \in \lambda\widehat{\mathcal{A}}\}\,; \tag{2.4}$$

cf. Proposition 4.1. This characterization reveals two important properties. First, the gauge function is blind to non-convexity of an atomic set. (A finite set of atoms, for instance, is always non-convex.) Second, it generalizes the standard notion of a norm because it encompasses all convex functions that are nonnegative, vanish at the origin, and are positively homogeneous, i.e.,

$$\gamma_{\mathcal{A}}(\alpha x) = \alpha\gamma_{\mathcal{A}}(x) \quad \forall\alpha \geq 0. \tag{2.5}$$

The support function

$$\sigma_{\mathcal{A}}(z) = \sup\{\langle a, z\rangle \mid a \in \mathcal{A} \cup \{0\}\} \tag{2.6}$$

to the set $\mathcal{A}$ shares these same properties, and is related to the gauge function through the polar inequality, described by the next result.

Proposition 2.2 (Polar Inequality)**.** For all pairs of vectors $(x, z) \in \operatorname{dom}\gamma_{\mathcal{A}} \times \operatorname{dom}\sigma_{\mathcal{A}}$,

$$\langle x, z\rangle \leq \gamma_{\mathcal{A}}(x) \cdot \sigma_{\mathcal{A}}(z). \tag{2.7}$$

Proof. The proof for the polar inequality in Rockafellar [35, Section 15] relies on the polarity of cones. Here we provide an elementary proof that only depends on the provided definitions of gauge and support functions. First, consider the case $\gamma_{\mathcal{A}}(x) > 0$. Let $\widehat{x} = x/\gamma_{\mathcal{A}}(x)$. Thus, $\widehat{x} \in \widehat{\mathcal{A}}$, and

$$\langle x, z\rangle = \gamma_{\mathcal{A}}(x) \cdot \langle \widehat{x}, z\rangle \leq \gamma_{\mathcal{A}}(x) \cdot \sigma_{\mathcal{A}}(z),$$

where the inequality follows from the maximality property of the support function; see Figure 2.1. Next, consider the case $\gamma_{\mathcal{A}}(x) = 0$, and proceed by contradiction. Suppose $\langle x, z\rangle > 0$. Because $\gamma_{\mathcal{A}}(x) = 0$ implies $\lambda x \in \widehat{\mathcal{A}}$ for all $\lambda > 0$, it follows from the positive homogeneity of $\sigma_{\mathcal{A}}$ that $\sigma_{\mathcal{A}}(z) = \infty$. This contradicts the assumption that $z \in \operatorname{dom}\sigma_{\mathcal{A}}$. Thus,

$$\langle x, z\rangle \leq 0 = \gamma_{\mathcal{A}}(x) \cdot \sigma_{\mathcal{A}}(z). \qquad \square$$

Example 2.3 (Norms)**.** When $\mathcal{A} = \{x \mid \|x\| \leq 1\}$ is the unit level set to any norm, $\|\cdot\|\colon \mathbb{R}^n \to \mathbb{R}_+$, then

$$\gamma_{\mathcal{A}}(x) = \|x\| \quad \text{and} \quad \sigma_{\mathcal{A}}(z) = \|z\|_d,$$

where $\|\cdot\|_d$ is the dual norm. The polar inequality then reduces to the standard inequality between inner products and dual pairs of norms:

$$\langle x, z\rangle \leq \|x\| \cdot \|z\|_d. \qquad \square$$

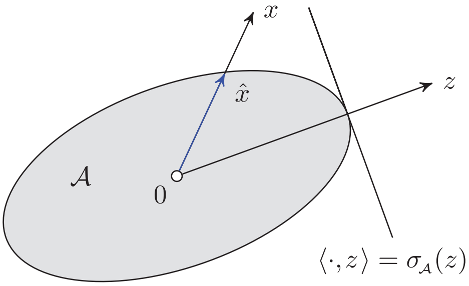


Figure 2.1: An illustration of the proof of the polar inequality given by Proposition 2.2. The inner-product of the scaled vector $\widehat{x} = x/\gamma_{\mathcal{A}}(x)$ with z always lies in the left halfspace defined by $\sigma_{\mathcal{A}}(z)$, i.e., $\langle \widehat{x}, z\rangle \leq \sigma_{\mathcal{A}}(z)$.

2.3 Alignment and Support Identification

The polar inequality motivates the following generalized definition of aligned pairs of vectors.

Definition 2.4 (Alignment). A pair $(x, z) \in \mathbb{R}^n \times \mathbb{R}^n$ is *aligned* with respect to the atomic set $\mathcal{A}$, i.e., x and z are $\mathcal{A}$-aligned, if the polar inequality (2.7) holds as an equation.

This general notion of alignment follows from the special case where $\mathcal{A} = \{x \mid \|x\|_2 \leq 1\}$ is the unit 2-norm ball. In this case, it follows from Example 2.3 that the polar inequality reduces the Cauchy-Schwartz inequality

$$\langle x, z\rangle \leq \|x\|_2 \cdot \|z\|_2.$$

This inequality holds as an equation if and only if x and z are aligned in the usual sense: there exists a nonnegative scalar α such that $x = \alpha z$. The more general notion of alignment captures other important special cases, including the Hölder inequality, which is a special case of (2.7) in which $\mathcal{A}$ is the unit p-norm ball, with $p \in [1, \infty]$.

A rich convex geometry underlies this general notion of alignment, and plays a role in identifying the atoms important for the decomposition (2.2). Suppose that a vector z is $\mathcal{A}$-aligned with x. As we demonstrate in Proposition 4.5, all atoms $a \in \mathcal{A}$ in the atomic support of x must also be $\mathcal{A}$-aligned with z, i.e.,

$$\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z) := \{a \in \mathcal{A} \cup \{0\} \mid \langle a, z\rangle = \sigma_{\mathcal{A}}(z)\}\,. \tag{2.8}$$

To see that $\mathcal{E}_{\mathcal{A}}(z)$ indeed contains all the atoms that are $\mathcal{A}$-aligned with z, note that any atom $a \in \mathcal{E}_{\mathcal{A}}(z)$ necessarily has unit gauge value, i.e., $\gamma_{\mathcal{A}}(a) = 1$, which follows from (2.4) and Proposition 2.2. Thus the condition $\langle a, z\rangle = \sigma_{\mathcal{A}}(z)$ implies that a is $\mathcal{A}$-aligned with z. Figure 2.2 presents a visualization of this concept. The atoms in $\mathcal{E}_{\mathcal{A}}(z)$ are said to be *exposed* by the vector z. As we show in Subsection 3.2, this set of atoms is contained in the face of $\widehat{\mathcal{A}}$ exposed by the vector z.

2.4 Examples

There are many varieties of atomic sets and recognizable convex regularizers used to obtain sparse decompositions. Chandrasekaran *et al.* [34]

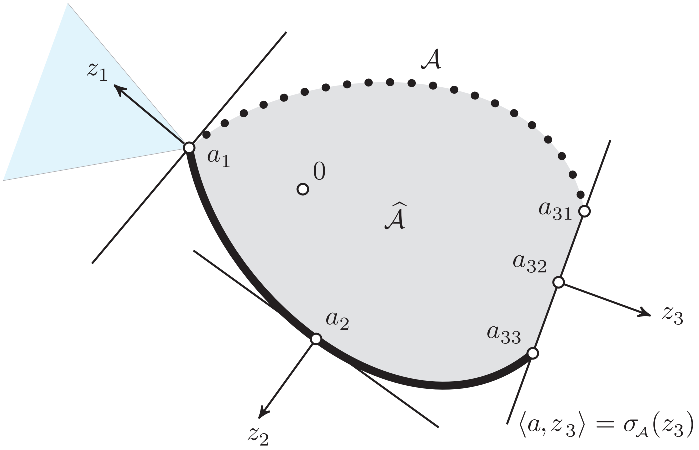


Figure 2.2: The set of atoms in the set $\mathcal{A}$ generally (but not necessarily) defines the boundary of the convex hull $\widehat{\mathcal{A}}$. The set of exposed atoms $\mathcal{E}_{\mathcal{A}}(z)$ are contained within the supporting hyperplane $\{a \mid \langle a, z\rangle = \sigma_{\mathcal{A}}(z)\}$ normal to z. The atom a_1 is exposed by z_1 and all other directions that lie in the shaded cone; the atom a_2 is exposed by the unique direction along z_2; and the set of atoms $\{a_{3i}\}_{i=1}^{3}$ are exposed by the unique direction along z_3.

and Jaggi [17] both give extensive lists of atoms and the norms that they induce, as well as their applications in practice. Here we provide several simple examples that illustrate the variety of ways in which vectors can be aligned.

Example 2.5 (One Norm). Let

$$\mathcal{A} = \{\pm e_1, \dots, \pm e_n\}$$

be the signed standard basis vectors. The gauge to this atomic set induces the 1-norm, which is the canonical example of a sparsifying convex penalty. The corresponding support function is the dual ∞-norm:

$$\gamma_{\mathcal{A}}(x) = \|x\|_1 \quad \text{and} \quad \sigma_{\mathcal{A}}(z) = \|z\|_\infty.$$

The polar inequality (2.7) reduces to Hölder's inequality for these norms—i.e., $\langle x, z\rangle \le \|x\|_1 \cdot \|z\|_\infty$. As is well known, this holds with equality—and thus x and z are $\mathcal{A}$-aligned—if and only if

$$x_i \neq 0 \quad \Longrightarrow \quad \operatorname{sgn}(x_i) z_i = \max_j |z_j| \quad \forall i = 1:n.$$

Alignment of the pair (x, z) with respect to the atomic set $\mathcal{A}$ is hence equivalent to the statement that $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$, where the atomic support for x and the atoms exposed by z, respectively, are given the the sets

$$\begin{aligned}\mathcal{S}_{\mathcal{A}}(x) &= \{\operatorname{sgn}(x_i) \cdot e_i \mid x_i \neq 0\}\,,\\ \mathcal{E}_{\mathcal{A}}(z) &= \{\operatorname{sgn}(z_i) \cdot e_i \mid |z_i| = \max_j |z_j|\}\,.\end{aligned}$$

The inclusion $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$ also characterizes an optimality condition. For example, consider the LASSO [6] problem

$$\operatorname*{minimize}_x \ \ \tfrac{1}{2}\|Ax - b\|_2^2 \ \ \text{subject to} \ \ \|x\|_1 \leq \tau,$$

where τ is a positive parameter. It's straightforward to verify that x is optimal for this problem if and only if $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$ where $z := A^T(b - Ax)$ is the negative gradient of the objective. Section 5 describes in detail the connection between optimality and alignment. □

Example 2.6 (Nuclear Norm)**.** The nuclear norm, or Schatten 1-norm, of a matrix is the spectral analog to the vector 1-norm. The nuclear norm and its dual spectral norm can be obtained via the atomic set

$$\mathcal{A} = \{uv^T \mid \|u\|_2 = \|v\|_2 = 1\}$$

of normalized n-by-m rank-1 matrices. Let X and Z both be m-by-n matrices with singular values $c_1 \geq \cdots \geq c_{m\wedge n} \geq 0$ and $s_1 \geq \cdots \geq s_{m\wedge n} \geq 0$, where $m \wedge n := \min\{m, n\}$. The corresponding gauge for X is the nuclear norm

$$\gamma_{\mathcal{A}}(X) = \|X\|_1 := \sum_{i=1}^{m\wedge n} c_i,$$

and the support function for Z is the Schatten 2-norm

$$\sigma_{\mathcal{A}}(Z) = \|Z\|_\infty := \max_{i=1:m\wedge n} s_i.$$

The atomic description of these functions is consistent with the notion that the nuclear norm is a convex function that promotes low rank (e.g., sparsity with respect to rank-1 matrices) [12]. The alignment condition $\langle X, Z\rangle = \|X\|_1 \cdot \|Z\|_\infty$ holds when X and Z have a simultaneously ordered singular value decomposition (SVD). Suppose, then, that X has rank r and that the largest singular value of Z has multiplicity d. If the SVDs of X and Z are

$$X = \sum_{i=1}^{r} c_i u_i v_i^T \quad \text{and} \quad Z = \sum_{i=1}^{m \wedge n} s_i u_i v_i^T,$$

then the atomic support of X is

$$\mathcal{S}_{\mathcal{A}}(X) = \{u_1 v_1^T, \dots, u_r v_r^T\},$$

and the set of atoms exposed by Z is

$$\mathcal{E}_{\mathcal{A}}(Z) = \{u_1 v_1^T, \dots, u_d v_d^T\}.$$

The inclusion (2.8), which identifies the support as a subset of the exposed atoms, implies $d \geq r$. Thus, the singular vectors of Z corresponding to the d singular values $s_1, \dots, s_d$ contain the singular vectors of X. Note that this can also be proven as a consequence of von Neumann's trace inequality [36], [37]. Friedlander and Macêdo [38] use this property for the construction of space-efficient dual methods for low-rank semidefinite optimization. □

Example 2.7 (Linear Subspaces)**.** Suppose that the set of atoms $\mathcal{A}$ contains all the elements of a linear subspace $\mathcal{L}$. In this case, the gauge $\gamma_{\mathcal{L}}(x)$ is finite only if x is in $\mathcal{L}$, and similarly, the support function $\sigma_{\mathcal{L}}(z)$ is finite only if z is in its orthogonal complement $\mathcal{L}^\perp$. In particular, because $\mathcal{L}$ and $\mathcal{L}^\perp$ are cones,

$$\gamma_{\mathcal{L}}(x) = \delta_{\mathcal{L}}(x) \quad \text{and} \quad \sigma_{\mathcal{L}}(z) = \delta_{\mathcal{L}^\perp}(z).$$

The respective domains of the gauge and support functions are thus $\mathcal{L}$ and $\mathcal{L}^\perp$. It follows that, under the atomic set $\mathcal{L}$, the vectors x and z are $\mathcal{L}$-aligned if and only if $x \in \mathcal{L}$ and $z \in \mathcal{L}^\perp$. Thus, the aligned vectors are orthogonal. □

3

Alignment with Respect to General Convex Sets

The alignment principles we develop depend on basic notions of convex sets and their supporting hyperplanes. Gauge and support functions are central because they furnish a complete and convenient calculus for manipulating and interpreting atomic sets. The following blanket assumption, which holds throughout the monograph, ensures a desirable symmetry between a set and its polar, as explained in Subsection 3.1. This assumption considerably simplifies our analysis and fortunately holds for many of the most important and relevant examples.

Assumption 3.1 (Origin Containment)**.** The set $\mathcal{C} \subset \mathbb{R}^n$ is closed convex and contains the origin.

3.1 Polarity

Our notion of alignment is based on the polarity of convex sets. Polarity is most intuitive in the context of convex cones, which are convex sets closed under positive scaling: the set $\mathcal{K}$ is a convex cone if $\alpha\mathcal{K} \subset \mathcal{K}$ for all positive α and $\mathcal{K} + \mathcal{K} \subset \mathcal{K}$. Its polar

$$\mathcal{K}^\circ = \{z \mid \langle x, z\rangle \le 0 \ \forall x \in \mathcal{K}\} \tag{3.1}$$

is also a convex cone, and its vectors make an oblique angle (i.e., a nonpositive inner product) with every vector in $\mathcal{K}$. The definition of the polar operation for general convex set $\mathcal{C}$ is similar, except that the 0 bound is replaced with a 1:

$$\mathcal{C}^\circ = \{z \mid \langle x, z\rangle \le 1 \text{ for all } x \in \mathcal{C}\}\,. \tag{3.2}$$

One way to connect the two polarity definitions (3.1) and (3.2) is by "lifting" the set $\mathcal{C}$ and its polar $\mathcal{C}^\circ$ and embedding them into slices of the 1 and -1 level sets of the opposing cones in $\mathbb{R}^{n+1}$:

$$\mathcal{K}_\mathcal{C} := \operatorname{cone}(\mathcal{C} \times \{1\}) \quad \text{and} \quad \mathcal{K}_\mathcal{C}^\circ := \operatorname{cone}(\mathcal{C}^\circ \times \{-1\}).$$

Then for any nonzero $(n+1)$-vectors $\bar{x} \in \mathcal{K}_\mathcal{C}$ and $\bar{z} \in \mathcal{K}_\mathcal{C}^\circ$, there exist positive scalars α_x and α_z, and vectors $x \in \mathcal{C}$ and $z \in \mathcal{C}^\circ$, such that

$$\begin{aligned} \langle \bar{x}, \bar{z}\rangle &= \left\langle \alpha_x \begin{pmatrix} x \\ 1 \end{pmatrix}, \alpha_z \begin{pmatrix} z \\ -1 \end{pmatrix} \right\rangle \\ &= (\alpha_x\, \alpha_z) \cdot (\langle x, z\rangle - 1) \\ &\le 0, \end{aligned} \tag{3.3}$$

where the last inequality follows from the polar definition in (3.2). The last inequality in (3.3) confirms that the cones $\mathcal{K}_\mathcal{C}$ and $\mathcal{K}_\mathcal{C}^\circ$ are polar to each other under definition (3.1).

The blanket Assumption 3.1, which asserts $\mathcal{C}$ is closed and contains the origin, yields a special symmetry because then the polar $\mathcal{C}^\circ$ also contains the origin and $\mathcal{C}^{\circ\circ} = \mathcal{C}$ [35, Theorem 14.5]. This is one of the reasons why we define $\widehat{\mathcal{A}} = \operatorname{conv}(\mathcal{A} \cup \{0\})$ to include the origin.

The pair of polar sets $\mathcal{C}$ and $\mathcal{C}^\circ$ can be said to generate the corresponding gauge and support functions $\gamma_\mathcal{C}$ and $\sigma_\mathcal{C}$, as we show below. Because the gauge and support functions are positively homogeneous (2.5), the epigraphs for these functions are convex cones. Moreover, it is straightforward to verify from the Minkowski characterization of the gauge (2.4) and the definition of the polar that the unit level sets for $\gamma_\mathcal{C}$ and $\sigma_\mathcal{C}$ are the sets that define them:

$$\mathcal{C} = \{x \mid \gamma_\mathcal{C}(x) \le 1\} \quad \text{and} \quad \mathcal{C}^\circ = \{z \mid \sigma_\mathcal{C}(z) \le 1\}\,. \tag{3.4}$$

It thus follows that

$$\operatorname{epi} \gamma_\mathcal{C} = \operatorname{cone}(\mathcal{C} \times \{1\}) \quad \text{and} \quad \operatorname{epi} \sigma_\mathcal{C} = \operatorname{cone}(\mathcal{C}^\circ \times \{1\}). \tag{3.5}$$

Figure 3.1 shows a visualization of the epigraph of the gauge to $\mathcal{C}$.

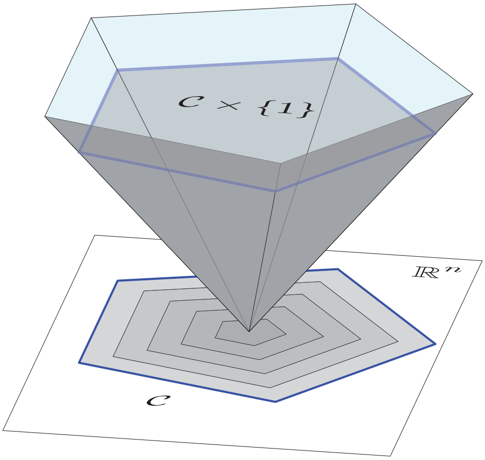


Figure 3.1: The epigraph of the gauge $\gamma_{\mathcal{C}}$ is the cone in $\mathbb{R}^n \times \mathbb{R}$ generated by the set $\mathcal{C} \subset \mathbb{R}^n$; see (3.5).

The recession cone (also known as the asymptotic cone) of a set $\mathcal{C}$ contains the set of directions in which the set is unbounded:

$$\operatorname{rec} \mathcal{C} := \{d \mid x + \lambda d \in \mathcal{C} \text{ for every } \lambda \geq 0 \text{ and } x \in \mathcal{C}\}. \tag{3.6}$$

See Figure 3.2 for an illustration. Vectors in the recession cone can also be thought of as "horizon points" of $\mathcal{C}$ [35, p. 60]. With respect to the gauge and support functions to the set $\mathcal{C}$, vectors $u \in \operatorname{rec} \mathcal{C}$ have the property that $\gamma_{\mathcal{C}}(u) = 0$ and $\sigma_{\mathcal{C}}(u) = +\infty$; see Proposition 3.2. We must therefore be prepared to consider cases where these functions can take on infinite values. Far from being a nuisance, this property is useful in modelling important cases in optimization.

The following proposition collects standard results regarding gauge and support functions and establishes the polarity correspondence between these two functions. The proofs of these claims can be found in standard texts, notably Rockafellar [35] and Hiriart-Urruty and Lemaréchal [39]. Those proofs typically rely on properties of conjugate functions. Because our overall theoretical development doesn't require conjugacy, we provide self-contained proofs that depend only on properties of closed convex sets.

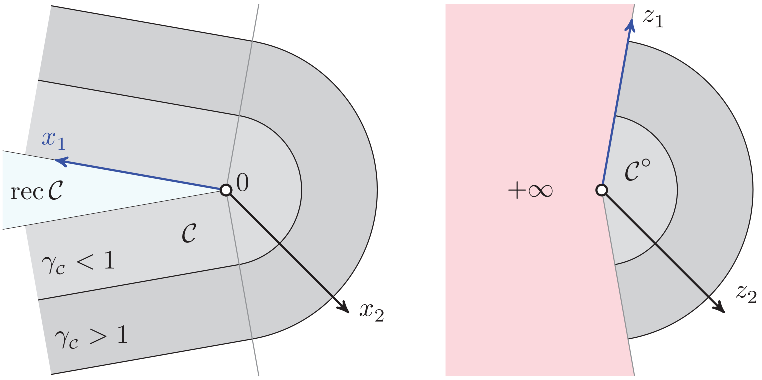

x_1
$\operatorname{rec}\mathcal{C}$
0
$\mathcal{C}$
$\gamma_{\mathcal{C}} < 1$
$\gamma_{\mathcal{C}} > 1$
x_2
z_1
$+\infty$
$\mathcal{C}^\circ$
z_2

Figure 3.2: The contours of the gauge function of $\mathcal{C}$ (left) and of $\mathcal{C}^\circ$ (right). All vectors x in the recession cone of $\mathcal{C}$ have gauge value $\gamma_{\mathcal{C}}(x) = 0$. A vector x_1 can only be $\mathcal{C}$-aligned with another vector z_1 if they are orthogonal to each and each is an extreme ray, respectively, of $\operatorname{rec}\mathcal{C}$ and $\operatorname{dom}\gamma_{\mathcal{C}^\circ} = (\operatorname{rec}\mathcal{C})^\circ$. Each of the pairs (x_1, z_1) and (x_2, z_2) are $\mathcal{C}$-aligned.

Proposition 3.2 (Properties of Gauges and Support Functions)**.** Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed convex set that contains the origin, and $\mathcal{D} \subset \mathbb{R}^n$ be an arbitrary set. The following statements hold.

(a) (Closure and convex hull) $\gamma_{\mathcal{D}} = \gamma_{\operatorname{cl}\operatorname{conv}\mathcal{D}}$ and $\sigma_{\mathcal{D}} = \sigma_{\operatorname{cl}\operatorname{conv}\mathcal{D}}$.

(b) (Polarity and conjugacy) $\gamma_{\mathcal{C}^\circ} = \sigma_{\mathcal{C}} = \delta^*_{\mathcal{C}}$.

(c) (Linear transformation) For a linear map M with adjoint M^*,

$$\gamma_{M^{-1}\mathcal{C}}(x) = \gamma_{\mathcal{C}}(Mx) \quad \text{and} \quad \sigma_{M\mathcal{C}}(z) = \sigma_{\mathcal{C}}(M^*z),$$

where we interpret the image of $\mathcal{C}$ under a linear map M and its inverse as the sets

$$\begin{aligned} M\mathcal{C} &= \{Mx \mid x \in \mathcal{C}\}, \\ M^{-1}\mathcal{C} &= \{x \mid Mx \in \mathcal{C}\}. \end{aligned}$$

(d) (Scaling) $\alpha\gamma_{\mathcal{C}} = \gamma_{\frac{1}{\alpha}\mathcal{C}}$ and $\alpha\sigma_{\mathcal{C}} = \sigma_{\alpha\mathcal{C}}$ for all $\alpha > 0$.

(e) (Bijection) $\mathcal{C} = \{x \in \mathbb{R}^n \mid \langle x, z\rangle \le \sigma_{\mathcal{C}}(z) \text{ for all } z \in \mathbb{R}^n\}$.

(f) (Domains) $\operatorname{dom}\gamma_{\mathcal{C}} = \operatorname{cone}\mathcal{C}$ and $\operatorname{dom}\sigma_{\mathcal{C}} = (\operatorname{rec}\mathcal{C})^\circ$.

(g) (Subdifferential) $\partial\sigma_{\mathcal{C}}(z) = \operatorname{conv}\{x \in \mathcal{C} \mid \langle x, z\rangle = \sigma_{\mathcal{C}}(z)\}$.

(h) (Recession cones) $\gamma_{\mathcal{C}}(x) = 0$ if and only if $x \in \operatorname{rec}\mathcal{C}$.

(i) (Duality correspondence) For all $x \in \mathcal{C}$,

$$x \in \partial\sigma_{\mathcal{C}}(z) \text{ if and only if } z \in \mathcal{N}_{\mathcal{C}}(x).$$

Proof.

(a) The stated property for the gauge function follows immediately from the Minkowski-functional description (2.4). Now consider the support function. Because $\mathcal{D} \subseteq \operatorname{cl}\operatorname{conv}\mathcal{D}$, it follows that $\sigma_{\mathcal{D}}(z) \le \sigma_{\operatorname{cl}\operatorname{conv}\mathcal{D}}(z)$ for all z. Hence it's sufficient to prove that $\sigma_{\operatorname{cl}\operatorname{conv}\mathcal{D}}(z) \le \sigma_{\mathcal{D}}(z)$ for all z. Fix any $d \in \operatorname{cl}\operatorname{conv}\mathcal{D}$ and choose an arbitrary sequence $\{d_k\}_{n=1}^{\infty} \subset \operatorname{conv}\mathcal{D}$ such that $d_k \to d$. Each element of the sequence $\{d_k\}$ is a convex combination of points in $\mathcal{D}$, and so it follows that $\langle d_k, z\rangle \le \sigma_{\mathcal{D}}(z)$ for all k and z. Since $d_k \to d$ and $\langle d_k, z\rangle \le \sigma_{\mathcal{D}}(z)$ for all n, it follows that $\langle d, z\rangle \le \sigma_{\mathcal{D}}(z)$. But d is arbitrary, and so we can conclude that $\sigma_{\operatorname{cl}\operatorname{conv}\mathcal{D}}(z) \le \sigma_{\mathcal{D}}(z)$.

(b) First, we show $\gamma_{\mathcal{C}^\circ} = \sigma_{\mathcal{C}}$. The gauge to $\mathcal{C}^\circ$ (cf. (2.4)) can be expressed as

$$\gamma_{\mathcal{C}^\circ}(x) = \inf\left\{\lambda > 0 \mid \lambda^{-1}x \in \mathcal{C}^\circ\right\}.$$

Thus, from the definition of the polar set (3.2),

$$\begin{aligned}\gamma_{\mathcal{C}^\circ}(x) &= \inf\left\{\lambda > 0 \mid \langle \lambda^{-1}x, y\rangle \le 1,\ \forall y \in \mathcal{C}\right\}\\ &= \left[\sup\left\{\mu > 0 \mid \langle \mu x, y\rangle \le 1,\ \forall y \in \mathcal{C}\right\}\right]^{-1}\\ &= \left[\sup\left\{\mu > 0 \mid \langle x, y\rangle \le \mu^{-1},\ \forall y \in \mathcal{C}\right\}\right]^{-1}\\ &= \sup_{y\in\mathcal{C}} \langle x, y\rangle\\ &= \sigma_{\mathcal{C}}(x).\end{aligned}$$

Next, we show $\sigma_{\mathcal{C}} = \delta^*_{\mathcal{C}}$. By the definition of conjugate function,

$$\begin{aligned}\delta^*_{\mathcal{C}}(x) &= \sup_{z\in\mathbb{R}^n} \{\langle x, z\rangle - \delta_{\mathcal{C}}(z)\}\\ &= \sup_{z\in\mathcal{C}} \langle x, z\rangle\\ &= \sigma_{\mathcal{C}}(x).\end{aligned}$$

(c) From the Minkowski functional expression for the gauge function,

$$\begin{aligned}\gamma_{\mathcal{C}}(Mx) &= \inf\{\lambda \mid Mx \in \lambda C\}\\ &= \inf\{\lambda \mid x \in \lambda M^{-1}C\}\\ &= \gamma_{M^{-1}C}(x).\end{aligned}$$

Also, from the definition of the adjoint of a linear map,

$$\begin{aligned}\sigma_{M\mathcal{C}}(z) &= \sup\{\langle Mx, z\rangle \mid x \in \mathcal{C}\}\\ &= \sup\{\langle x, M^*z\rangle \mid x \in \mathcal{C}\}\\ &= \sigma_{\mathcal{C}}(M^*z).\end{aligned}$$

(d) By defining $M = \alpha$, the proof follows directly from Proposition 3.2(c).

(e) Let $\mathcal{D} = \{x \in \mathbb{R}^n \mid \langle x, z\rangle \le \sigma_{\mathcal{C}}(z) \text{ for all } z \in \mathbb{R}^n\}$. By the definition of support function, it can be easily shown that $\mathcal{C} \subseteq \mathcal{D}$. So we only need to prove that $\mathcal{D} \subseteq \mathcal{C}$. Assume there is some $x \in \mathcal{D}$ such that $x \notin \mathcal{C}$. Then by the separating hyperplane theorem, there exists $s \in \mathbb{R}^n$ such that

$$\langle s, x\rangle > \sup\{\langle s, y\rangle \mid y \in \mathcal{C}\} = \sigma_{\mathcal{C}}(s).$$

This leads to a contradiction. We therefore conclude that $\mathcal{C} = \mathcal{D}$.

(f) It follows from the definition of the domain that $\operatorname{dom}\gamma_{\mathcal{C}} = \operatorname{cone}\mathcal{C}$. Thus we only need to show that $\operatorname{dom}\sigma_{\mathcal{C}} = (\operatorname{rec}\mathcal{C})^\circ$. First we show that $\operatorname{dom}\sigma_{\mathcal{C}} \subseteq (\operatorname{rec}\mathcal{C})^\circ$. For any $x \in \operatorname{dom}\sigma_{\mathcal{C}}$, the support $\sigma_{\mathcal{C}}(x)$ is finite. Thus for any $d \in \operatorname{rec}\mathcal{C}$,

$$\langle c + \lambda d, x\rangle < \infty, \quad \forall c \in \mathcal{C}, \lambda \ge 0;$$

see (3.6). It follows that $\langle d, x\rangle \le 0$, and thus $x \in (\operatorname{rec}\mathcal{C})^\circ$. For the other direction, instead we will show that $(\operatorname{dom}\sigma_{\mathcal{C}})^\circ \subseteq \operatorname{rec}\mathcal{C}$. Assume $x \in (\operatorname{dom}\sigma_{\mathcal{C}})^\circ$, then for any $c \in \mathcal{C}$, $\lambda \ge 0$, $y \in \operatorname{dom}\sigma_{\mathcal{C}}$,

$$\langle c + \lambda x, y\rangle = \langle c, y\rangle + \lambda\langle x, y\rangle \le \langle c, y\rangle \le \sigma_{\mathcal{C}}(y).$$

Because $\mathcal{C}$ is a closed convex set, we can conclude that $c + \lambda x \in \mathcal{C}$, for all $c \in \mathcal{C}$ and $\lambda \ge 0$ by Proposition 3.2(e). Therefore, $x \in \operatorname{rec}\mathcal{C}$ by (3.6).

(g) Let $\mathcal{D} = \operatorname{conv}\{x \in \mathcal{C} \mid \langle x, z\rangle = \sigma_{\mathcal{C}}(z)\}$. First, we show that $\mathcal{D} \subseteq \partial\sigma_{\mathcal{C}}(z)$. Assume $x \in \mathcal{D}$. Then for any $w \in \mathbb{R}^n$,

$$\sigma_{\mathcal{C}}(w) \ge \langle x, w\rangle = \sigma_{\mathcal{C}}(z) + \langle x, w - z\rangle.$$

Thus, $x \in \partial\sigma_{\mathcal{C}}(z)$. Next, we prove that $\partial\sigma_{\mathcal{C}}(z) \subseteq \mathcal{D}$. Assume $x \in \partial\sigma_{\mathcal{C}}(z)$, then

$$\sigma_{\mathcal{C}}(w) \ge \sigma_{\mathcal{C}}(z) + \langle x, w - z\rangle \quad \forall w \in \mathbb{R}^n. \tag{3.7}$$

By the subadditivity of support functions,

$$\sigma_{\mathcal{C}}(z) + \sigma_{\mathcal{C}}(w - z) \ge \sigma_{\mathcal{C}}(w) \quad \forall w \in \mathbb{R}^n. \tag{3.8}$$

It then follows from (3.7) and (3.8) that $\sigma_{\mathcal{C}}(v) \ge \langle x, v\rangle$ for all v. By Proposition 3.2(e), we thus conclude that $x \in \mathcal{C}$. Now let $w = 0$ in (3.7), it follows that $\langle x, z\rangle \ge \sigma_{\mathcal{C}}(z)$. Therefore, it follows that $\langle x, z\rangle = \sigma_{\mathcal{C}}(z)$ and thus $x \in \mathcal{D}$.

(h) First, assume $\gamma_{\mathcal{C}}(x) = 0$. Then for any $\widehat{x} \in \mathcal{C}$ and $\lambda \ge 0$,

$$\gamma_{\mathcal{C}}(\widehat{x} + \lambda x) \le \gamma_{\mathcal{C}}(\widehat{x}) + \lambda\gamma_{\mathcal{C}}(x) = \gamma_{\mathcal{C}}(\widehat{x}).$$

It follows that $\widehat{x} + \lambda x \in \mathcal{C}$ and therefore $x \in \operatorname{rec}\mathcal{C}$. Next, assume $x \in \operatorname{rec}\mathcal{C}$. Then by the definition of recession cone, we have $\lambda x \in \mathcal{C}$ for all $\lambda \ge 0$, which implies $\gamma_{\mathcal{C}}(x) = 0$.

(i) Let $x \in \mathcal{C}$ and $z \in \mathbb{R}^n$, then by Proposition 3.2(g) we know that $x \in \partial\sigma_{\mathcal{C}}(z)$ if and only if

$$\langle x, z\rangle \ge \langle u, z\rangle \quad \text{for all } u \in \mathcal{C}.$$

Therefore, from the definition of normal cone we know that this holds if and only if $z \in \mathcal{N}_{\mathcal{C}}(x)$. □

3.2 Exposed Faces

A face $\mathcal{F}_\mathcal{C}$ of a convex set $\mathcal{C}$ is a subset with the property that for all elements x_1 and x_2 both in $\mathcal{C}$, and for all $\theta \in (0,1)$,

$$\theta x_1 + (1-\theta)x_2 \in \mathcal{F}_\mathcal{C} \quad \Longleftrightarrow \quad x_1 \in \mathcal{F}_\mathcal{C} \quad \text{and} \quad x_2 \in \mathcal{F}_\mathcal{C}.$$

Note that the face must itself be convex. A face $\mathcal{F}_\mathcal{C}(d)$ is *exposed* by a direction $d \in \mathbb{R}^n$ if the face is contained in the supporting hyperplane with normal d:

$$\mathcal{F}_\mathcal{C}(d) = \{c \in \mathcal{C} \mid \langle c, d\rangle = \sigma_\mathcal{C}(d)\} = \partial\sigma_\mathcal{C}(d), \tag{3.9}$$

where the second equality follows from Proposition 3.2(g). The elements of the exposed face $\mathcal{F}_\mathcal{C}(d)$ are thus precisely those elements of $\mathcal{C}$ that achieve the supremum for $\sigma_\mathcal{C}(d)$.

In Section 4 we will consider atomic sets that are not convex. In that case, the exposed face of the convex hull of those atoms coincides with the convex hull of the exposed atoms. In particular, if $\mathcal{A} = \{a_i\}_{i\in\mathcal{I}}$ is any collection of atoms, then

$$\mathcal{F}_\mathcal{A}(d) = \operatorname{conv}\mathcal{E}_\mathcal{A}(d).$$

The face of a set is exposed by the direction of a vector, regardless of its magnitude. In particular, it follows from positive homogeneity of the support function $\sigma_\mathcal{C}$ to the set $\mathcal{C}$ that

$$\mathcal{F}_{\alpha\mathcal{C}}(d) = \alpha\mathcal{F}_\mathcal{C}(d) \quad \text{and} \quad \mathcal{F}_\mathcal{C}(\alpha d) = \mathcal{F}_\mathcal{C}(d) \quad \forall\alpha > 0. \tag{3.10}$$

For nonpolyhedral sets, it's possible that some faces may not be exposed [35, p. 163].

3.3 Alignment Characterization

The alignment condition specified by Definition 2.4 rests on the tightness of the polar inequality (2.7). In this subsection we tie the alignment condition and the polar inequality to a geometric concept based on exposed faces. This geometric vantage illuminates an intuitive notion of the dual relationship between a pair of aligned vectors. We proceed in two steps. The first step characterizes the alignment for vectors normalized to unit length, as defined by the gauge to a set and its polar; see Proposition 3.3. The second step generalizes the result by removing the normalization assumption; see Corollary 3.4.

Proposition 3.3 (Normalized Alignment). Any pair of vectors $(x, z) \in \mathcal{C} \times \mathcal{C}^\circ$ is $\mathcal{C}$-aligned if any of the following equivalent conditions holds:

(a) $\langle x, z \rangle = 1$,

(b) $x \in \operatorname{bnd} \mathcal{C}$ and $z \in \mathcal{F}_{\mathcal{C}^\circ}(x)$,

(c) $x \in \mathcal{F}_{\mathcal{C}}(z)$ and $z \in \operatorname{bnd} \mathcal{C}^\circ$.

Proof. Suppose (a) holds. By definition (3.2) of the polar set $\mathcal{C}^\circ$,

$$\sigma_{\mathcal{C}^\circ}(x) = \sup \{ \langle x, u \rangle \mid u \in \mathcal{C}^\circ \} \leq 1 \quad \forall x \in \mathcal{C}.$$

Then (a) implies that z achieves the supremum above, and so by (3.9), this holds if and only if $z \in \mathcal{F}_{\mathcal{C}^\circ}(x)$ and $x \in \operatorname{bnd} \mathcal{C}$. Thus (b) holds. The fact that (b) implies (a) follows by simply reversing this chain of arguments.

To prove that (a) is equivalent to (c), we only need to use the assumption that $\mathcal{C}$ is closed and contains the origin, and hence that $\mathcal{C} = \mathcal{C}^{\circ\circ}$ [35, Theorem 14.5]. This allows us to reuse the arguments above by exchanging the roles of x and z, and $\mathcal{C}$ and $\mathcal{C}^\circ$. □

The following corollary characterizes the general alignment condition without assuming that the vector pair (x, z) is normalized.

Corollary 3.4 (Alignment). Any pair of vectors $(x, z) \in \operatorname{cone} \mathcal{C} \times \operatorname{cone} \mathcal{C}^\circ$ is $\mathcal{C}$-aligned if any of the following equivalent conditions holds:

(a) $\langle x, z \rangle = \gamma_{\mathcal{C}}(x) \cdot \sigma_{\mathcal{C}}(z)$,

(b) $z \in \operatorname{cone} \mathcal{F}_{\mathcal{C}^\circ}(x) + \operatorname{rec} \mathcal{C}^\circ$,

(c) $x \in \operatorname{cone} \mathcal{F}_{\mathcal{C}}(z) + \operatorname{rec} \mathcal{C}$.

Proof. First suppose that $\gamma_{\mathcal{C}}(x)$ and $\sigma_{\mathcal{C}}(z)$ are positive. Then the equivalence of the statements follows by applying Proposition 3.3 to the normalized pair of vectors $\widehat{x} := x/\gamma_{\mathcal{C}}(x)$ and $\widehat{z} := z/\sigma_{\mathcal{C}}(z)$. In that case, (a) follows immediately after multiplying $\langle \widehat{x}, \widehat{z} \rangle = 1$ by the quantity $\gamma_{\mathcal{C}}(x) \cdot \sigma_{\mathcal{C}}(z)$. Parts (b) and (c) follow from the fact that for any convex set $\mathcal{D}$ and any vector $d \in \mathbb{R}^n$, $\mathcal{F}_{\mathcal{D}}(d) = \mathcal{F}_{\mathcal{D}}(\alpha d)$ for any positive scalar α; see (3.10).

We now show equivalence of the statements in the case where $\gamma_{\mathcal{C}}(x) = 0$. By Proposition 3.2(h), this holds if and only if $x \in \operatorname{rec} \mathcal{C}$, but not in $\mathcal{F}_{\mathcal{C}}(z)$. Thus (c) holds. But because $\sigma_{\mathcal{C}}(z)$ is finite, x and z together satisfy $\langle x, z \rangle = 0$. Thus, (a) holds. To show that (b) holds, note that $\sigma_{\mathcal{C}^\circ}(x) = \gamma_{\mathcal{C}}(x) = 0$, and so by (3.9),

$$\operatorname{cone} \mathcal{F}_{\mathcal{C}^\circ}(x) = \{u \mid \langle x, u \rangle = 0\},$$

which certainly contains z. Thus, (b) holds. The case with $\sigma_{\mathcal{C}}(z) = 0$ follows using the same symmetric argument used in the proof of Proposition 3.3. □

Corollary 3.4 dispenses with the normalization requirement and allows for one of the vectors of the aligned pair to lie in the recession cone of $\mathcal{C}$ or its polar $\mathcal{C}^\circ$. In that case, the alignment condition in Corollary 3.4(a) reduces to an orthogonality condition, i.e., $\langle x, z \rangle = 0$. But if $x \in \operatorname{rec} \mathcal{C}$, this implies that $z \in (\operatorname{rec} \mathcal{C})^\circ$. In other words, x and z are extreme rays of their respective recession cones. Figure 3.2 illustrates the geometry of this situation.

Example 3.5 (Alignment for Cones)**.** Suppose that $\mathcal{K}$ is a cone, and that the pair of vectors (x, z) is $\mathcal{K}$-aligned. Because a cone is its own recession cone, i.e., $\operatorname{rec} \mathcal{K} = \mathcal{K}$, Corollary 3.4 asserts

$$\langle x, z \rangle = 0 \quad \Longleftrightarrow \quad x \in \mathcal{K} \quad \Longleftrightarrow \quad z \in \mathcal{K}^\circ.$$

This assertion effectively generalizes Example 2.7, which made the same claim for linear subspaces.

Thus, for convex cones we see that alignment is equivalent to orthogonality. This principle applies to general convex sets $\mathcal{C}$ using the lifting technique described in Subsection 3.1. Take any pair of vectors $(x, z) \in \mathcal{C} \times \mathcal{C}^\circ$ satisfying $\langle x, z \rangle = 1$, which implies that they are $\mathcal{C}$-aligned by Proposition 3.3. Then

$$\bar{x} := (x, 1) \in \mathcal{K}_{\mathcal{C}} \quad \text{and} \quad \bar{z} := (z, -1) \in \mathcal{K}_{\mathcal{C}}^\circ,$$

and

$$\langle \bar{x}, \bar{z} \rangle = \langle x, z \rangle - 1 = 0.$$

This last equation coincides with tightness of the inequality (3.3), which characterizes polarity of cones. □

The next example shows how the alignment property is connected to complementarity in conic programming [40, Section 5.3.6]. Section 5 explores a more general connection between alignment and optimality in convex optimization.

Example 3.6 (Alignment as Optimality in Conic Optimization). Consider the pair of dual linear conic optimization problems

$$\begin{array}{llll} \underset{x}{\text{minimize}} & \langle c, x\rangle & \qquad \underset{y,\,z}{\text{maximize}} & \langle b, y\rangle \\ \text{subject to} & Fx = b,\ x \in \mathcal{K}, & \qquad \text{subject to} & F^T y - z = c,\ z \in \mathcal{K}^\circ, \end{array}$$

where $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, $(b, c) \in \mathbb{R}^m \times \mathbb{R}^n$ are arbitrary vectors, and $\mathcal{K}^\circ$ is the polar cone of $\mathcal{K}$.

The feasible triple (x, y, z) is optimal if strong duality holds, i.e.,

$$0 = \langle c, x\rangle - \langle b, y\rangle = \langle F^T y - z, x\rangle - \langle Fx, y\rangle = \langle x, z\rangle.$$

But because $x \in \mathcal{K}$ and $z \in \mathcal{K}^\circ$, it follows from Example 3.5 that x and z are $\mathcal{K}$-aligned. □

3.4 Alignment as Conic Orthogonal Decomposition

The Moreau decomposition for convex cones asserts that any vector can be orthogonally decomposed as the projection onto a cone and its polar [39, Theorem 3.2.5]. In other words, for any vector x, the conditions

$$x = x_1 + x_2, \quad x_1 \in \mathcal{K},\ \ x_2 \in \mathcal{K}^\circ,\ \ \langle x_1, x_2\rangle = 0$$

hold if and only if

$$x_1 = \mathrm{proj}_{\mathcal{K}}(x) \quad \text{and} \quad x_2 = \mathrm{proj}_{\mathcal{K}^\circ}(x).$$

This conic polar decomposition generalizes the classical notion of decomposition by orthogonal subspaces, and sheds light on the relationship between vectors aligned with respect to any convex set $\mathcal{C}$.

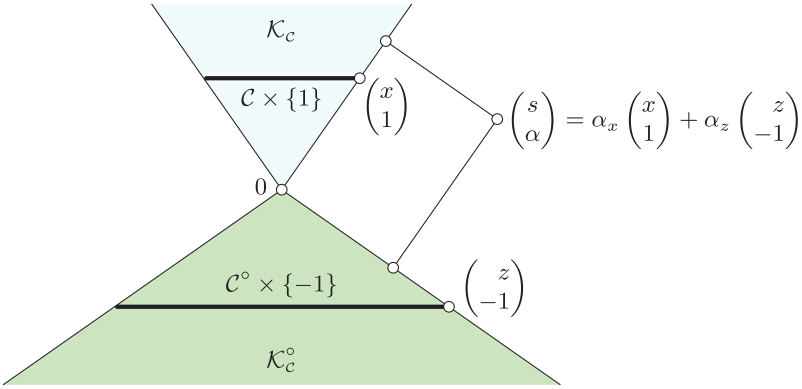

$\mathcal{K}_{\mathcal{C}}$
$\mathcal{C} \times \{1\}$
$\begin{pmatrix} x \\ 1 \end{pmatrix}$
$\begin{pmatrix} s \\ \alpha \end{pmatrix} = \alpha_x \begin{pmatrix} x \\ 1 \end{pmatrix} + \alpha_z \begin{pmatrix} z \\ -1 \end{pmatrix}$
0
$\mathcal{C}^\circ \times \{-1\}$
$\begin{pmatrix} z \\ -1 \end{pmatrix}$
$\mathcal{K}_{\mathcal{C}}^\circ$

Figure 3.3: Any vector $(s, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ can be decomposed into orthogonal components in the cones generated by a convex set $\mathcal{C} \subset \mathbb{R}^n$ and its polar. The components of the decomposition $s = \alpha_x x + \alpha_z z$ are $\mathcal{C}$-aligned.

Every element, respectively, in $\mathcal{K}_{\mathcal{C}}$ and $\mathcal{K}_{\mathcal{C}}^\circ$ is a nonnegative multiple of $(x, 1)$ and $(z, 1)$ for some vectors $x \in \mathcal{C}$ and $z \in \mathcal{C}^\circ$. Thus, for any vector $(s, \alpha) \in \mathbb{R}^n \times \mathbb{R}$, Moreau's decomposition implies unique nonnegative scalars α_x and α_z such that

$$\begin{aligned}(s, \alpha) &= \operatorname{proj}_{\mathcal{K}_{\mathcal{C}}}(s, \alpha) + \operatorname{proj}_{\mathcal{K}_{\mathcal{C}}^\circ}(s, \alpha) \\ &= \alpha_x(x, 1) + \alpha_z(z, -1).\end{aligned}$$

Because the elements of the decomposition are orthogonal,

$$(\alpha_x \alpha_z) \cdot (\langle x, z \rangle - 1) = 0.$$

Moreover, the vectors $\widehat{x} := \alpha_x x$ and $\widehat{z} := \alpha_z z$, respectively, have gauge and support values

$$\alpha_x = \gamma_{\mathcal{C}}(\widehat{x}) \quad \text{and} \quad \alpha_z = \sigma_{\mathcal{C}}(\widehat{z}).$$

Thus, the pair $(\widehat{x}, \widehat{z})$ is $\mathcal{C}$-aligned because $\langle \widehat{x}, \widehat{z} \rangle = \alpha_x \alpha_z$. Figure 3.3 illustrates the geometry of this decomposition.

4

Alignment with Respect to Atomic Sets

Section 3 describes properties of gauges and support functions generated by general convex sets. In this section, we study the properties of these functions and their interpretations that are particular to atomic sets $\mathcal{A} \subset \mathbb{R}^n$.

4.1 Atomic Decomposition

Section 2 described two different characterizations of the gauge function, given in terms of a minimal conic decomposition of the atoms (2.2), and as Minkowski functional, which gives the infimal dilation of the atomic sets (2.4). The former "sum form" of the gauge function is useful because it provides an interpretation of all gauge functions as weighted 1-norm specialized to a particular atomic set. This suggests that gauges are the natural promoters of atomic sparsity. The next elementary result establishes the equivalence between the two characterizations.

Proposition 4.1 (Gauge Equivalence). For any set $\mathcal{A} \subset \mathbb{R}^n$,

$$\begin{aligned}\gamma_{\mathcal{A}}(x) &= \inf_{c_a} \left\{ \sum_{a \in \mathcal{A}} c_a \;\middle|\; x = \sum_{a \in \mathcal{A}} c_a a,\ c_a \geq 0\ \forall a \in \mathcal{A} \right\} \\ &= \inf \left\{ \lambda \geq 0 \mid x \in \lambda \widehat{\mathcal{A}} \right\}.\end{aligned}$$

Proof. Take any $x \in \operatorname{cone}\widehat{\mathcal{A}}$, since otherwise the sets above are empty, and by convention, both expressions have infinite value. Then, by the definition of $\widehat{\mathcal{A}}$,

$$\begin{aligned}\gamma_{\mathcal{A}}(x) &= \inf_{\lambda} \{\lambda \geq 0 \mid x \in \lambda\widehat{\mathcal{A}}\} \\ &= \inf_{\lambda,\, \bar{c}_a} \left\{ \lambda \geq 0 \;\middle|\; x = \lambda \sum_{a \in \mathcal{A}} \bar{c}_a a,\ \sum_{a \in \mathcal{A}} \bar{c}_a = 1,\ \bar{c}_a \geq 0\ \forall a \in \mathcal{A} \right\} \\ &= \inf_{\lambda,\, c_a} \left\{ \lambda \;\middle|\; x = \sum_{a \in \mathcal{A}} c_a a,\ \sum_{a \in \mathcal{A}} c_a = \lambda,\ c_a \geq 0\ \forall a \in \mathcal{A} \right\},\end{aligned}$$

which, after eliminating λ, yields the required equivalence. □

Some atomic sets, such as the set of rank-1 outer products used to define the nuclear-norm ball (cf. Example 2.6), may be uncountably infinite. Even in that case, however, whenever x admits a valid atomic decomposition with respect to the atoms in $\mathcal{A}$, the gauge value, and thus the sum $\sum_{a \in \mathcal{A}} c_a$, necessarily has finite value.

Proposition 4.2 (Finite Support). For any n-vector x that has a valid atomic decomposition with respect to the set $\mathcal{A} \subset \mathbb{R}^n$, a finite atomic support set $\mathcal{S}_{\mathcal{A}}(x) \in \mathbf{supp}_{\mathcal{A}}(x)$ always exists.

Proof. If $\gamma_{\mathcal{A}}(x) = 0$, the assertion is trivially true, since the empty set is the only element of $\mathbf{supp}_{\mathcal{A}}(x)$. Now suppose $\gamma_{\mathcal{A}}(x) > 0$, and define the normalized vector $\widehat{x} = x/\gamma_{\mathcal{A}}(x)$. Then $\widehat{x} \in \widehat{\mathcal{A}}$, and $\gamma_{\mathcal{A}}(x) = 1$. By Carathéory's Theorem [35, Theorem 17.1], there exists a finite convex decomposition of $\widehat{x}$ in terms of at most $n+1$ atoms in $\mathcal{A}$. That is, there exists a set $\mathcal{S} \subset \mathcal{A}$ with $n+1$ elements such that

$$\widehat{x} = \sum_{a \in \mathcal{S}} \widehat{c}_a a, \quad \sum_{a \in \mathcal{S}} \widehat{c}_a = 1, \quad \widehat{c}_a > 0,\ \forall a \in \mathcal{S}.$$

Taking $c_a = \gamma_{\mathcal{A}}(x) \cdot \widehat{c}_a$ for each $a \in \mathcal{S}$ gives a solution to the equations in (2.3), showing that $\mathcal{S} \in \mathbf{supp}_{\mathcal{A}}(x)$. □

The support may not be unique, even if it's minimal.

Example 4.3 (Non-Uniqueness). Consider the atomic set

$$\mathcal{A} = \{(\pm 1, \pm 1, 1)\} \subset \mathbb{R}^3.$$

The point $x = (0, 0, 2)$ can be expressed in at least three different ways,

$$\begin{aligned} x &= (1,1,1) + (-1,-1,1) \\ &= (1,-1,1) + (-1,1,1) \\ &= \frac{1}{2}[(1,1,1) + (-1,-1,1) + (1,-1,1) + (-1,1,1)], \end{aligned}$$

all of which give the same gauge value $\gamma_{\mathcal{A}}(x) = 2$. In this case, the set elements of $\mathbf{supp}_{\mathcal{A}}(x)$ are

$$\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\}, \left\{ \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \right\}, \quad \text{and}$$

$$\frac{1}{2} \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \right\}.$$

Any element of $\mathbf{supp}_{\mathcal{A}}(x)$ is a valid support set of x with respect to the atomic set $\mathcal{A}$. However, for functions commonly used to promote sparsity, often the support set is always unique. □

Proposition 4.1 establishes that the gauge value $\gamma_{\mathcal{A}}(x)$ of a vector x yields an atomic decomposition whose coefficient sum is minimal. If another vector v can be atomically decomposed as a subset of the atoms from x, then the support of v is a subset of the support of x, i.e., $\mathcal{S}_{\mathcal{A}}(v) \subset \mathcal{S}_{\mathcal{A}}(x)$. This is established in the following proposition.

Proposition 4.4 (Same Support Sets). Take any n-vector x and atomic set $\mathcal{A} \subset \mathbb{R}^n$ such that $\gamma_{\mathcal{A}}(x)$ is positive and finite. Then a vector v that has a valid atomic decomposition in terms of any support $\mathcal{S}_{\mathcal{A}}(x)$, i.e., there exists coefficients c_a such that

$$v = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a a, \quad c_a \geq 0, \tag{4.1}$$

must have the gauge value

$$\gamma_{\mathcal{A}}(v) = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a.$$

Proof. Suppose, by way of contradiction, that there exists a atomic decomposition of v with respect to $\mathcal{A}$ that isn't given by (4.1), i.e.,

$$v = \sum_{a\in\mathcal{A}} c'_a a, \qquad c'_a \geq 0, \qquad \sum_{a\in\mathcal{A}} c'_a < \sum_{a\in\mathcal{S}_\mathcal{A}(x)} c_a.$$

Because $\mathcal{S}_\mathcal{A}(x)$ is the support set of x, there exist positive coefficients $\widehat{c}_a$ where

$$x = \sum_{a\in\mathcal{S}_\mathcal{A}(x)} \widehat{c}_a a, \quad \gamma_\mathcal{A}(x) = \sum_{a\in\mathcal{S}_\mathcal{A}(x)} \widehat{c}_a.$$

But a valid decomposition of x is

$$x = \beta v + x - \beta v = \beta \sum_{a\in\mathcal{A}} c'_a a + \sum_{a\in\mathcal{S}_\mathcal{A}(x)} (\widehat{c}_a - \beta c_a) a,$$

where we pick $\beta = [\min_{a\in\mathcal{S}_\mathcal{A}(x)} \widehat{c}_a]/[\max_{a\in\mathcal{S}_\mathcal{A}(x)} c_a]$ to guarantee that all the coefficients are nonnegative. Then by the definition of a gauge,

$$\sum_{a\in\mathcal{S}_\mathcal{A}(x)} \widehat{c}_a \leq \beta \sum_{a\in\mathcal{A}} c'_a + \sum_{a\in\mathcal{S}_\mathcal{A}(x)} (\widehat{c}_a - \beta c_a),$$

which holds if and only if

$$\sum_{a\in\mathcal{A}} c'_a \geq \sum_{a\in\mathcal{S}_\mathcal{A}(x)} c_a.$$

This implies that the decomposition of v with respect to $\mathcal{S}_\mathcal{A}(x)$ is in fact the *minimal* decomposition of v with respect to $\mathcal{A}$, and the sum of the coefficients indeed giving its gauge value; cf. Definition 2.1. □

Proposition 4.5 (Support Identification). For any set $\mathcal{A} \subset \mathbb{R}^n$, the pair of n-vectors (x, z) is $\mathcal{A}$-aligned if and only if $\mathcal{S}_\mathcal{A}(x) \subseteq \mathcal{E}_\mathcal{A}(z)$ for all $\mathcal{S}_\mathcal{A}(x) \in \mathbf{supp}_\mathcal{A}(x)$.

Proof. First, we show that if x and z are $\mathcal{A}$-aligned, then $\mathcal{S}_\mathcal{A}(x) \subseteq \mathcal{E}_\mathcal{A}(z)$ for all $\mathcal{S}_\mathcal{A}(x) \in \mathbf{supp}_\mathcal{A}(x)$. Because x and z are $\mathcal{A}$-aligned,

$$\langle x, z\rangle = \gamma_\mathcal{A}(x) \cdot \sigma_\mathcal{A}(z). \tag{4.2}$$

Now suppose that $\gamma_\mathcal{A}(x) > 0$. Then all support sets $\mathcal{S}_\mathcal{A}(x) \in \mathbf{supp}_\mathcal{A}(x)$ are nonempty. Suppose that $a \in \mathcal{S}_\mathcal{A}(x)$ but $a \notin \mathcal{E}_\mathcal{A}(z)$. We show that this leads to a contradiction. By definition, $a \notin \mathcal{E}_{\mathcal{A}}(z)$ implies that

$$\langle a, z\rangle < \sigma_{\mathcal{A}}(z). \tag{4.3}$$

Define $v = x - c_a a$, which is the vector that results from deleting the atom a from the support of x. Then by Proposition 4.4,

$$\gamma_{\mathcal{A}}(v) = \gamma_{\mathcal{A}}(x) - c_a. \tag{4.4}$$

Thus,

$$\begin{aligned}\langle x, z\rangle &\overset{\text{(i)}}{=} \langle v, z\rangle + c_a\langle a, z\rangle \\ &\overset{\text{(ii)}}{<} \gamma_{\mathcal{A}}(v)\sigma_{\mathcal{A}}(z) + c_a\sigma_{\mathcal{A}}(z) \\ &= (\gamma_{\mathcal{A}}(v) + c_a)\sigma_{\mathcal{A}}(z) \\ &\overset{\text{(iii)}}{=} \gamma_{\mathcal{A}}(x)\cdot\sigma_{\mathcal{A}}(z),\end{aligned}$$

where (i) follows by construction ($x = v + c_a a$); (ii) follows from the polar inequality (2.7) and (4.3); and (iii) follows from (4.4). But this contradicts (4.2), and therefore $a \in \mathcal{S}_{\mathcal{A}}(x)$ implies $a \in \mathcal{E}_{\mathcal{A}}(z)$, i.e., $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$.

Now assume $\gamma_{\mathcal{A}}(x) = 0$. Then $x \in \operatorname{rec}\widehat{\mathcal{A}}$ and $\mathbf{supp}_{\mathcal{A}}(x)$ contains only the empty set. Because the empty set is also a subset of $\mathcal{E}_{\mathcal{A}}(z)$ for any z, the statement is trivially true.

Next, we show that if $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$ for all $\mathcal{S}_{\mathcal{A}}(x) \in \mathbf{supp}_{\mathcal{A}}(x)$, then x and z are $\mathcal{A}$-aligned. By the definition of support set (2.3), we can assume that

$$\gamma_{\mathcal{A}}(x) = \sum_{a\in\mathcal{S}_{\mathcal{A}}(x)} c_a, \quad x = \sum_{a\in\mathcal{S}_{\mathcal{A}}(x)} c_a a, \quad c_a > 0\ \forall a \in \mathcal{S}_{\mathcal{A}}(x).$$

Then by Corollary 3.4, we only need to show that $\langle x, z\rangle = \gamma_{\mathcal{A}}(x)\cdot\sigma_{\mathcal{A}}(z)$. Indeed,

$$\begin{aligned}\langle x, z\rangle &= \sum_{a\in\mathcal{S}_{\mathcal{A}}(x)} c_a\langle a, z\rangle \\ &\overset{\text{(i)}}{=} \Bigg(\sum_{a\in\mathcal{S}_{\mathcal{A}}(x)} c_a\Bigg)\sigma_{\mathcal{A}}(z) = \gamma_{\mathcal{A}}(x)\cdot\sigma_{\mathcal{A}}(z),\end{aligned}$$

where (i) follows from the assumption that $\mathcal{S}_{\mathcal{A}}(x) \subseteq \mathcal{E}_{\mathcal{A}}(z)$. □

4.2 Examples

The general alignment result described by Corollary 3.4 includes the possibility that aligned vectors may contain elements from the recession cone of the atomic set. Elements in the recession cone may be interpreted as directions, rather than just points in the set. The presence of a nontrivial recession cone must be considered in practice, and is exhibited, for example, by all seminorms: these are nonnegative functions that behave like norms with the exception that they may vanish at nonzero points and are not necessarily symmetric. The next example describes a common atomic set composed by points and directions.

Example 4.6 (Total Variation). The anisotropic total-variation norm of an n-vector x is defined as

$$\|x\|_{\mathrm{TV}} = \sum_{i=2}^{n} |x_i - x_{i-1}| = \|Dx\|_1, \qquad D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

The bi-diagonal matrix D has a 1-dimensional nullspace spanned by the constant vector of all ones e, and so $De = 0$. Thus the TV norm is a seminorm and the generating atomic set must include a direction of recession, given by the range of e. Interestingly, the atomic set that induces this norm isn't unique: for any matrix $A = [a_1, \dots, a_{n-1}]$ where $DA = I$, the corresponding TV norm is the gauge with respect to the atoms

$$\mathcal{A} = \{\pm a_1, \dots, \pm a_{n-1}\} + \mathrm{cone}(\pm e).$$

To see this, write

$$x = \sum_{i=1}^{n-1} c_i a_i + c_e e = Ac + c_e e$$

for some scalars $c_1, \dots, c_{n-1}$ and c_e. (The scalars are not restricted to be nonnegative because the set of atoms includes vectors with both positive and negative signs.) Note that the $n-1$ vectors a_i span null(e), so the above decomposition always exists, with unique values for c_i

and c_e. The solution to (2.3) thus determines the unique decomposition

$$x = \sum_{i=1}^{n-1} \underbrace{(s_i c_i)}_{c_a} \cdot \underbrace{(s_i a_i)}_{a} + c_e e, \quad s_i = \operatorname{sgn}(c_i),$$

where $(s_i c_i)$ are the coefficients for the atoms $(s_i a_i) \in \mathcal{A}$, and c_e is the coefficient for the recession direction e. Then

$$\|Dx\|_1 = \|DAc\|_1 = \|c\|_1 = \gamma_{\mathcal{A}}(x).$$

If $x \in \operatorname{cone}(\pm e)$, then $x \in \operatorname{rec} \widehat{\mathcal{A}}$ and thus $\gamma_{\mathcal{A}}(x) = 0$.

To see that the atomic set isn't unique, note that $DA = I$ for any matrix of the form $A = B + es^T$, where

$$B = [b_1, \ldots, b_{n-1}] := \begin{bmatrix} 1 & 1 & \ldots & 1 & 1 \\ 0 & 1 & \ldots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix}$$

and $s \in \mathbb{R}^{n-1}$ is an arbitrary vector. However, the gauge function with respect to the atomic set formed by the columns of B and e is well defined. Specifically, note that the range of the matrix $[B \ e]$ spans all of $\mathbb{R}^n$. Thus the decomposition

$$x = Bc + c_e e \tag{4.5}$$

uniquely defines the vector c and the scalar c_e, and $\gamma_{\mathcal{A}}(x) = \|Dx\|_1 = \|c\|_1$, as before.

The support function for this set of atoms is

$$\begin{aligned} \sigma_{\mathcal{A}}(z) &= \sup \{\langle x, z\rangle \mid x = Bc + c_e e, \ \|c\|_\infty \le 1\} \\ &= \sup \{\langle c, B^T z\rangle + c_e \langle e, z\rangle \mid \|c\|_\infty \le 1\}. \end{aligned}$$

Note that if $z \notin \operatorname{null}(e)$, then $\sigma_{\mathcal{A}}(z)$ clearly unbounded because c_e isn't constrained. This confirms the fact that the domain of $\sigma_{\mathcal{A}}$ is $(\operatorname{rec} \widehat{\mathcal{A}})^\circ = \operatorname{null}(e)$, as shown by Proposition 3.2(f). Corollary 3.4 asserts that if z is $\mathcal{A}$-aligned with x, then it exposes all of the atoms that contribute non-trivially towards the decomposition (4.5). In particular, $\mathcal{S}_{\mathcal{A}}(x) \subset \mathcal{E}_{\mathcal{A}}(z)$, where one such decomposition gives, respectively, the support of x and the atoms exposed by z:

$$\mathcal{S}_{\mathcal{A}}(x) = \{\, \operatorname{sgn}((Dx)_i) b_i \mid x_i \neq 0 \,\},$$
$$\mathcal{E}_{\mathcal{A}}(z) = \Big\{ \operatorname{sgn}(\langle b_i, z\rangle) b_i | \max_j \langle b_j, z\rangle = |\langle b_i, z\rangle| \Big\}.$$

Note that these alignment conditions do not depend on the specific choice of the representation $\mathcal{A}$, and are defined only with respect to the columns of B, which are fixed. □

Group norms arise in applications where the nonzero entries of a vector are concentrated in patterns across the vector. Applications include source localization, functional magnetic resonance imaging, and others [11], [41], [42]. One interesting feature of group norms is that they are not polyhedral.

Example 4.7 (Group Norms). Consider the ℓ subsets $g_i \subseteq \{1 : n\}$ such $\cup_{i=1}^{\ell} g_i = \{1 : n\}$. Define the *group norm* with respect to the groups $\mathcal{G} = \{g_1, \ldots, g_\ell\}$ as the solution of the convex optimization problem

$$\|x\|_{\mathcal{G}} = \min_{y_i} \left\{ \sum_{i=1}^{\ell} \|y_i\|_2 \;\middle|\; x = \sum_{i=1}^{\ell} P_{g_i} y_i \right\}, \tag{4.6}$$

where the linear operator $P_{\mathcal{I}}\colon \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^n$ scatters the elements of a vector into an n vector at positions indexed by $\mathcal{I}$, i.e., $\{(P_{\mathcal{I}} y)_i\}_{i \in \mathcal{I}} = y$, and $(P_{\mathcal{I}} y)_k = 0$ for any $k \notin \mathcal{I}$. This norm is induced by the atomic set

$$\mathcal{A} = \{ P_{g_i} s_i \mid s_i \in \mathbb{R}^{|g_i|},\ \|s_i\|_2 = 1,\ i = 1 : \ell \},$$

which yields the decomposition

$$x = \sum_{i=1}^{\ell} c_i (P_{g_i} s_i), \tag{4.7}$$

where c_i and $(P_{g_i} s_i)$ are, respectively, the coefficients and atoms of the decomposition.

If the sets in $\mathcal{G}$ form a partition of $\{1 : n\}$ then the (non-overlapping) group norm is simply

$$\|x\|_{\mathcal{G}} = \sum_{i=1}^{\ell} \|x_{g_i}\|_2.$$

A common example is the matrix (1, 2) norm, which is the sum of the Euclidean norms of the columns of a matrix [43]. In the non-overlapping group case, the support set is unique, and for all $i = 1 : \ell$, the coefficients and atoms of the decomposition (4.7) are given by

$$c_i = \|x_{g_i}\|_2 \quad \text{and} \quad (P_{g_i} s_i) \quad \text{with } s_i = (c_i)^{-1} x_{g_i}.$$

More generally, the support sets g_i may overlap, and thus the gauge value of x must be obtained as the solution of the convex optimization problem (4.6).

The conditions under which a vector z is $\mathcal{A}$-aligned with x is similar to the 1-norm case. We first decompose by each group g_i:

$$\begin{aligned}\sup_{x\in\mathcal{A}} \langle x, z\rangle &\overset{\text{(i)}}{=} \max_{i=1\,:\,\ell}\ \sup\{\langle s_i, z_{g_i}\rangle \mid \|s_i\|_2 \le 1,\ s_i \in \mathbb{R}^{|g_i|}\}\\ &\overset{\text{(ii)}}{=} \max_{i=1\,:\,\ell}\ \|z_{g_i}\|_2,\end{aligned}$$

where (i) follows from applying the supremum to each atom in $\mathcal{A}$ and (ii) follows from the definition of the 2-norm. That is to say, x is $\mathcal{A}$-aligned with z if the decomposition (4.7) has $\mathcal{S}_{\mathcal{A}}(x) \subset \mathcal{E}_{\mathcal{A}}(z)$, where

$$\mathcal{S}_{\mathcal{A}}(x) = \{P_{g_i} y_i / \|y_i\|_2 \mid \|y_i\|_2 > 0\} \ \text{ with } y \text{ as the solution to (4.6)},$$

and

$$\mathcal{E}_{\mathcal{A}}(z) = \{z_{g_i}/\|z_{g_i}\|_2 \mid \|z_{g_i}\|_2 = \max_j\ \|z_{g_i}\|_2\}. \qquad \square$$

Bogdan *et al.* [44] proposed the ordered weighted 1-norm (OWL) as a statistical tool for promoting models with low false discovery rates over certain design matrices. Zeng and Figueiredo [13] derive the atomic set for this norm, and show that it generalizes the octagonal shrinkage and clustering algorithm for regression (OSCAR) [45], which has been shown to have good sparse clustering properties.

Example 4.8 (OWL Norm). Consider a nonnegative vector $w \in \mathbb{R}^n$ with elements ordered as $w_1 \ge \cdots \ge w_n \ge 0$. The OWL norm with respect to w is defined as

$$\|x\|_w = \sum_{i=1}^{n} w_i |x_{[i]}|,$$

where $x_{[i]}$ is the ith-largest component of x in magnitude. This norm is induced by the atomic set

$$\mathcal{A} = \{Qb_i \mid Q \in \mathcal{P}_n,\ \ i = 1 : n\},$$

where $\mathcal{P}_n$ is the set of all $n \times n$ signed permutation matrices and the vectors

$$b_i := (\underbrace{\tau_i, \dots, \tau_i}_{i \text{ entries}}, 0, \dots, 0) \quad \text{with } \tau_i := \left(\sum_{j=1}^{i} w_j\right)^{-1}.$$

To see this, first derive the support function with respect to $\mathcal{A}$:

$$\begin{aligned}\sigma_{\mathcal{A}}(z) &= \max_{i=1:n} \ \max_{Q \in \mathcal{P}_n} \ \langle z, Qb_i \rangle \\ &= \max_{i=1:n} \ \tau_i \left(\sum_{j=1}^{i} |z_{[j]}|\right).\end{aligned}$$

Use Proposition 3.2(b) together with (3.4) to obtain an expression for the gauge function to the atomic set $\mathcal{A}$:

$$\begin{aligned}\gamma_{\mathcal{A}}(x) &= \sup_z \{\langle x, z \rangle \mid \sigma_{\mathcal{A}}(z) \le 1\} \\ &= \sup_z \left\{ \langle x, z \rangle \;\middle|\; \sum_{j=1}^{i} |z_{[j]}| \le \sum_{j=1}^{i} w_j \text{ for all } i = 1 : n \right\} \\ &= \sum_{i=1}^{n} w_i |x_{[i]}|.\end{aligned}$$

Now consider the atomic decomposition of x with respect to the atomic set $\mathcal{A}$. Let $\widehat{x} = [\,|x_{[1]}|, \dots, |x_{[n]}|\,]^T$ denote the absolute ordered version of x. Then there exists $Q_x \in \mathcal{P}_n$ such that $x = Q_x \widehat{x}$. Let

$$c = B^{-1}\widehat{x} \quad \text{with } B = [b_1, \dots, b_n],$$

where B^{-1} exists because B is upper-triangular with strictly positive entries. It is straightforward to determine the components of the vector c, which are all nonnegative: for all $i = 1 : n$,

$$c_i = \begin{cases} (|x_{[i]}| - |x_{[i+1]}|)/\tau_i & \text{if } i \in \{1 : n-1\}, \\ |x_{[n]}|/\tau_n & \text{if } i = n. \end{cases} \tag{4.8}$$

It then follows that

$$x = Q_x Bc = \sum_{i=1}^{n} c_i \cdot Q_x b_i. \tag{4.9}$$

Next, we verify that the decomposition (4.9) provides a valid atomic support set for x with respect to $\mathcal{A}$. Indeed,

$$\begin{aligned}\gamma_{\mathcal{A}}(x) &= \sum_{i=1}^{n} w_i |x_{[i]}| \\ &= \frac{1}{\tau_1}|x_{[1]}| + \sum_{i=2}^{n} \left(\frac{1}{\tau_i} - \frac{1}{\tau_{i-1}} \right) |x_{[i]}| \\ &= \sum_{i=1}^{n} c_i.\end{aligned}$$

Then by Definition 2.1, we confirm that one valid support set for x with respect to $\mathcal{A}$ is given by

$$\mathcal{S}_{\mathcal{A}}(x) = \{ Q_x b_i \mid i = 1 : n \text{ and } c_i > 0 \},$$

where each c_i is defined by (4.8). Use Proposition 4.5 to determine the essential atoms exposed by z:

$$\mathcal{E}_{\mathcal{A}}(z) = \left\{ Qb_i \;\middle|\; i = \operatorname*{arg\,max}_{i=1\,:\,n} \tau_i g_i(\widehat{z}),\; Q \in \mathcal{P}_n,\; g_i(Q^T z) = g_i(\widehat{z}) \right\},$$

where the vector $\widehat{z} = (|z_{[1]}|, \ldots, |z_{[n]}|)$ and $g_i(z) = \sum_{j=1}^{i} z_j$. □

The next two examples are for gauges that encourage sparsity (i.e., low-rank) for matrices.

Example 4.9 (Semidefinite Matrix Trace Norm). An important gauge function is generated by the spectrahedron

$$\mathcal{A} = \{ uu^T \mid u \in \mathbb{R}^n,\; \|u\|_2 = 1 \},$$

which is a subset of the nuclear-norm ball that only includes symmetric rank-1 matrices. As with the nuclear-norm, this gauge encourages sparsity with respect to the set of rank-1 matrices—i.e., low-rank—and only admits positive semidefinite matrices.

We first derive the support function with respect to $\mathcal{A}$:

$$\begin{aligned}\sigma_{\mathcal{A}}(Z) &= \sup_{X \in \widehat{\mathcal{A}}} \langle X, Z \rangle \\ &= \max\left\{0, \sup_{\|u\|_2=1} \langle u, Zu \rangle\right\} \\ &= \max\{0, \lambda_{\max}(Z)\},\end{aligned}$$

which vanishes only if Z is negative semidefinite, and otherwise is achieved when u is a maximal eigenvector of Z. Let $X = U\Lambda U^T$ be the eigenvalue decomposition of X. Use Proposition 3.2(b) together with (3.4) to obtain an expression for the gauge to this atomic set:

$$\begin{aligned}\gamma_{\mathcal{A}}(X) &= \sup\{\langle X, Z\rangle \mid \lambda_{\max}(Z) \le 1\} \\ &= \sup\{\langle U\Lambda U^T, Z\rangle \mid \lambda_{\max}(Z) \le 1\} \\ &= \sup\{\langle \operatorname{diag}(\Lambda), \operatorname{diag}(U^T Z U)\rangle \mid \lambda_{\max}(Z) \le 1\}, \\ &= \operatorname{tr}(\Lambda) + \delta_{\succeq 0}(X),\end{aligned}$$

where the last equality holds because the supremum is achieved by $Z = UU^T$. The indicator $\delta_{\succeq 0}$ on the semidefinite cone arises because indefinite matrices cannot be atomically decomposed with respect to the atomic set $\mathcal{A}$. Moreover, it follows that the nontrivial eigenvectors provide an atomic support set for X, i.e.,

$$\{u_1 u_1^T, \ldots, u_r u_r^T\} \subseteq \mathcal{S}_{\mathcal{A}}(X),$$

where r is the rank of X.

This support isn't unique, however, and in fact the set of supports of X is very large. To see this, consider any valid atomic decomposition

$$X = c_1 v_1 v_1^T + \cdots + c_k v_k v_k^T = VCV^T,$$

where c_i and v_i, respectively, are the ith diagonal entry of the diagonal matrix C and ith column of the matrix V. Then

$$\begin{aligned}\operatorname{tr}(X) &= \operatorname{tr}(VCV^T) \\ &= \operatorname{tr}(CV^T V) \\ &= \langle \operatorname{diag}(C), \operatorname{diag}(V^T V)\rangle = \sum_{i=1}^{k} c_i,\end{aligned}$$

where the last equality follows from the fact that each $v_i v_i^T$ is in $\mathcal{A}$ and thus has unit norm. Therefore any atomic decomposition of X yields the same gauge value, which is the trace of X. Specifically, the support of X with respect to the spectrahedron $\mathcal{A}$ can be characterized as

$$\mathcal{S}_{\mathcal{A}}(X) = \{v_1 v_1^T, \dots, v_k v_k^T \mid \|v_i\|_2 = 1,\ \operatorname{range}(V) = \operatorname{range}(X)\}.$$

Because we don't impose orthonormality among the vectors v_i, this set isn't unique.

According to Proposition 4.5, the exposed atoms are given by the eigenvectors corresponding to the maximal eigenvalue of Z, including all of their convex combinations:

$$\mathcal{E}_{\mathcal{A}}(Z) = \operatorname{conv}\{uu^T \mid u^T Z u = \lambda_{\max}(Z)\}.$$

This set coincides with the exposed face $\mathcal{F}_{\mathcal{A}}(Z)$; cf. (3.10). □

An important challenge in recommender systems is incorporating *side information*, e.g., exogenous information about a user or product that may enhance recommendation quality. With the growth of social media platforms, one way to incorporate side information is to promote similarity of recommendations according to friendship networks. In such applications, a commonly used regularization term is $g(U) = \operatorname{tr}(U^T L U)$, where $L \in \mathbb{R}^{n\times n}$ is the positive semidefinite *graph Laplacian matrix* and $U \in \mathbb{R}^{n\times r}$ contains *node embeddings*, which represent the user archetypes [46]–[48]. The next example provides a gauge function that contains g as a special case. Let $X = UU^T$.

Example 4.10 (Weighted Trace Norm for Semidefinite Matrices)**.** We describe a generalization of the trace norm for positive semidefinite matrices, which was covered by Example 4.9. The weighted trace norm is given by the function

$$\gamma(X) = \langle L, X\rangle + \delta_{\succeq 0}(X),$$

where L is positive semidefinite. Write the decomposition of L as

$$L = [V \ \ \bar{V}] \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^T \\ \bar{V}^T \end{bmatrix} = V\Lambda V^T,$$

where Λ is diagonal with strictly positive elements and V and $\bar{V}$, respectively, span the range and nullspace of L.

We claim that γ is the gauge to the atomic set

$$\mathcal{A} = \{rr^T \mid r = Vp,\ p^T\Lambda p = 1\} + \{ss^T \mid s = \bar{V}q \text{ for all } q\}\,, \tag{4.10}$$

which we establish by showing $X \in \mathcal{A}$ implies $\gamma(X) = 1$, and vice versa.

Take any element $X \in \mathcal{A}$, and observe

$$\gamma(X) = \langle L, X\rangle = \langle L, Vpp^TV^T\rangle = p^TV^TLVp = p^T\Lambda p = 1.$$

Conversely, take any X such that $\gamma(X) = 1$. Then, X is positive semidefinite. The orthogonal decomposition of X onto the range and nullspace of L is given by

$$X = VV^TXVV^T + \bar{V}\bar{V}^TX\bar{V}\bar{V}^T.$$

Then,

$$1 = \gamma(X) = \langle L, X\rangle = \langle L, VV^TXVV^T\rangle = \langle \Lambda, V^TXV\rangle,$$

which implies that $V^TXV \in \operatorname{conv}\{pp^T \mid p^T\Lambda p = 1\}$. Therefore, X is in the convex hull of $\mathcal{A}$. The second set in the sum (4.10) is in the nullspace of L and thus can be ignored. This establishes the claim, and also provides an expression for the support set to X:

$$\mathcal{S}_{\mathcal{A}}(X) = \{(Vp_i)(Vp_i)^T \mid p_i^T\Lambda p_i = 1,\ \operatorname{range}(V[p_1 \dots p_k]) = \operatorname{range}(X)\}\,.$$

The minimal set of vectors needed to complete the support is equal to the rank of X.

We now derive the support function with respect to the atomic set $\mathcal{A}$. Because the nullspace of L characterizes a recession direction for γ, the domain of $\sigma_{\mathcal{A}}$ must be restricted to $\operatorname{null}(\bar{V}^T)$. Thus, for $Z \in \operatorname{dom}\sigma_{\mathcal{A}}$,

$$\begin{aligned}
\sigma_{\mathcal{A}}(Z) &= \sup_X \{\langle X, Z\rangle \mid X \in \widehat{\mathcal{A}}\} \\
&= \sup_{p,q} \{\langle Vpp^TV^T + \bar{V}qq^T\bar{V}^T, Z\rangle \mid p^T\Lambda p \le 1, \forall q\} \\
&= \sup_u \{\langle V^TZV, \Lambda^{-1/2}uu^T\Lambda^{-1/2}\rangle \mid u^Tu \le 1\} \\
&\qquad + \sup_q \{\langle qq^T, \bar{V}^TZ\bar{V}\rangle\} \\
&= \sup_u \{\langle \Lambda^{-1/2}V^TZV\Lambda^{-1/2}, uu^T\rangle \mid u^Tu \le 1\} \\
&= \max\{0,\ \lambda_{\max}(\Lambda^{-1/2}V^TZV\Lambda^{-1/2})\},
\end{aligned}$$

where the fourth equality follows from the assumption that $Z \in \operatorname{dom} \sigma_{\mathcal{A}}$. It's evident from this derivation that if $\bar{V}^T Z \bar{V} \neq 0$, then $\sigma_{\mathcal{A}}(Z) = \infty$.

We recognize that the expression inside the supremum is the generalized eigenvalue of the pencil (Z, L), so that for all $Z \in \operatorname{dom} \sigma_{\mathcal{A}}$

$$\sigma_{\mathcal{A}}(Z) = \max \{0,\, \lambda_{\max}(Z, L)\}.$$

Hence, the exposed atoms are given by the maximal generalized eigenvectors and their convex combinations:

$$\mathcal{E}_{\mathcal{A}}(Z) = \operatorname{conv} \{uu^T \mid \langle u, Zu \rangle = \lambda_{\max}(Z, L) \cdot \langle u, Lu \rangle\}. \qquad \square$$

The nuclear (Example 2.6), trace (Example 4.9), and weighted (Example 4.10) trace norms are examples of gauges generated by continuous atomic sets. We give another example of a continuous atomic set built from sinuoidal functions that vary continuously in their frequencies. The corresponding regularization function features widely in applications of super-resolution imaging. The derivation in this example largely follows Chi and Da Costa [49].

Example 4.11 (Continuous Sinusoidal Dictionary). Consider a signal $x \in \mathbb{C}^n$ that can be expressed as a superposition of complex sinusoids:

$$x = \sum_{k=1}^{r} c_k s(\tau_k), \tag{4.11}$$

where r is the number of spikes, the nonnegative coefficients c_k denote the complex amplitudes, the parameters $\tau_k \in [0, 1)$ denote the delays of the spikes, and the n-vector

$$s(\tau) = (1, \exp(i2\pi\tau), \ldots, \exp(i2\pi(n-1)\tau))$$

represents the complex sinusoids. Now consider the atomic set

$$\mathcal{A} = \{\exp(i\phi)s(\tau) \mid \tau \in [0, 1),\, \phi \in [0, 2\pi)\}.$$

Use (4.11) to reparameterize x as a decomposition of atoms in $\mathcal{A}$:

$$x = \sum_{k=1}^{r} |c_k| \exp(i\phi_k) s(\tau_k),$$

where $\phi_k \in [0, 2\pi)$ satisfies $|c_k| \exp(i\phi_k) = c_k$ for all k. It follows that the support set of x with respect to $\mathcal{A}$ is given by

$$\mathcal{S}_{\mathcal{A}}(x) = \{\exp(i\phi_k)s(\tau_k) \mid |c_k| > 0,\ k = 1 : r\} .$$

Next we show the corresponding computable gauge and support functions. The support function with respect to $\mathcal{A}$ at $z \in \mathbb{C}^n$ is given by

$$\begin{aligned}\sigma_{\mathcal{A}}(z) &= \sup\{\mathrm{Re}\langle \exp(i\phi)s(\tau), z\rangle \mid \tau \in [0, 1),\ \phi \in [0, 2\pi)\}\\ &= \sup\{|\langle s(\tau), z\rangle| \mid \tau \in [0, 1)\}.\end{aligned}$$

Given $z \in \mathbb{C}^n$, the complex trigonometric polynomial

$$p_z(\tau) = \langle s(\tau), z\rangle,$$

has coefficients z. Evaluating the support function $\sigma_{\mathcal{A}}(z)$ is equivalent to finding the maximum modulus of p_z on $[0, 1)$. McLean and Woerdeman [50] describe a stable root-finding algorithm that can be used to compute the maximizing values.

Use Proposition 3.2(b) and (3.4) to derive the gauge function

$$\begin{aligned}\gamma_{\mathcal{A}}(x) &= \sup_{z\in\mathbb{C}^n} \{\mathrm{Re}\langle x, z\rangle \mid \sigma_{\mathcal{A}}(z) \le 1\}\\ &= \sup_{z\in\mathbb{C}^n} \{\mathrm{Re}\langle x, z\rangle \mid |\langle s(\tau), z\rangle| \le 1\ \forall \tau \in [0, 1)\}\\ &= \sup_{z\in\mathbb{C}^n} \{\mathrm{Re}\langle x, z\rangle \mid s(\tau)^H zz^H s(\tau) \le 1\ \forall \tau \in [0, 1)\}\\ &= \sup\Bigg\{\mathrm{Re}\langle x, z\rangle \Bigg|\ z \in \mathbb{C}^n,\ H \in \mathbb{C}^{n\times n},\ H \succeq zz^H,\\ &\qquad\qquad \mathrm{tr}(H) = 1,\ \sum_{j=1}^{n-k} H_{j,j+k} = 0,\ k = 1 : n - 1\Bigg\}.\end{aligned}$$

The last equality follows from the bounded real lemma [51, Lemma 4.23].

From the definition of the support function $\sigma_{\mathcal{A}}$, we also conclude that the set of atoms exposed by the vector z is given by

$$\mathcal{E}_{\mathcal{A}}(z) = \{\exp(i\phi)s(\tau) \mid \mathrm{Re}\langle \exp(i\phi)s(\tau), z\rangle = \sigma_{\mathcal{A}}(z)\} . \qquad \square$$

5

Alignment as Optimality

A pair of vectors aligned with respect to an atomic set inform each other about their respective atomic supports. If the two vectors are related through a gradient map of a convex function, then the alignment condition can be interpreted as an optimality condition for a constrained or regularized optimization problem. The alignment condition can also be interpreted as providing an optimality certificate for the problem of finding minimum gauge elements of a convex set. This subsection describes both perspectives.

5.1 Regularized Smooth Problems

Consider the three related convex optimization problems

$$\underset{x}{\text{minimize}} \quad f(x) + \rho\gamma_{\mathcal{C}}(x), \tag{5.1a}$$

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to } \gamma_{\mathcal{C}}(x) \le \alpha, \tag{5.1b}$$

$$\underset{x}{\text{minimize}} \quad \gamma_{\mathcal{C}}(x) \quad \text{subject to } \ f(x) \le \tau, \tag{5.1c}$$

where the parameters ρ and α are positive, and $\tau > \inf f$. Note that the constraint $\gamma_{\mathcal{C}}(x) \le \alpha$ is equivalent to the constraint $x \in \alpha\mathcal{C}$, which follows from Proposition 3.2(d). Assumption 3.1 on $\mathcal{C}$ continues to hold throughout.

Theorem 5.1 (Optimality). Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and $\mathcal{C} \subset \mathbb{R}^n$. Assume that at the respective solutions for all three problems, the gauge values are positive and finite, and that for problems (5.1b) and (5.1c), the constraints hold with equality. For (5.1c), let $\mathcal{P} = \{x \mid f(x) \le \tau\}$ and assume that $\operatorname{ri}(\mathcal{P}) \cap \operatorname{ri}(\operatorname{dom}\gamma_{\mathcal{C}}) \neq \emptyset$. Then x^* is optimal if and only if it's $\mathcal{C}$-aligned with $z^* := -\nabla f(x^*)$, and

$$\begin{aligned} \sigma_{\mathcal{C}}(z^*) &= \rho \quad && \text{for problem (5.1a);} \\ \gamma_{\mathcal{C}}(x^*) &= \alpha \quad && \text{for problem (5.1b);} \\ f(x^*) &= \tau \quad && \text{for problem (5.1c).} \end{aligned}$$

Proof. First consider the unconstrained problem (5.1a). A vector x^* is a solution if and only if

$$0 \in \nabla f(x^*) + \rho \partial \gamma_{\mathcal{C}}(x^*).$$

Equivalently,

$$\rho^{-1} z^* \in \partial \gamma_{\mathcal{C}}(x^*) = \partial \sigma_{\mathcal{C}^\circ}(x^*) \equiv \mathcal{F}_{\mathcal{C}^\circ}(x^*).$$

Then by Corollary 3.4 and the requirement that $\rho^{-1} z^* \in \operatorname{bnd} \mathcal{C}^\circ$, this condition is equivalent to the $\mathcal{C}$-alignment of the pair (x^*, z^*) and $\sigma_{\mathcal{C}}(z^*) = \rho$.

Next, consider the gauge constrained problem (5.1b). Because the constraint is equivalent to $\alpha^{-1} x \in \mathcal{C}$, a feasible vector x^* is optimal if and only if

$$0 \in \nabla f(x^*) + \partial \delta_{\mathcal{C}}(\alpha^{-1} x^*) \quad \text{i.e., } z^* \in \partial \delta_{\mathcal{C}}(\alpha^{-1} x^*),$$

where $\delta_{\mathcal{C}}$ is the indicator function for set $\mathcal{C}$. By Rockafellar [35, Theorem 23.5] and Proposition 3.2(b), this is equivalent to $x^* \in \alpha \partial \sigma_{\mathcal{C}}(z^*)$. Thus by Corollary 3.4 and the theorem hypothesis, this condition equivalent to the $\mathcal{C}$-alignment of the pair (x^*, z^*) and $\gamma_{\mathcal{C}}(x^*) = \alpha$. Note that by Proposition 3.2(i), we know that this condition is equivalent to $z^* \in \mathcal{N}_{\mathcal{C}}(x^*/\alpha)$, which is the standard optimality condition for (5.1b).

Finally, consider the level-constrained problem (5.1c). Because of the hypothesis on the nonempty intersection between the relative interiors of $\mathcal{P}$ and $\operatorname{dom}\gamma_{\mathcal{C}}$, we can use Rockafellar [35, Theorem 23.8] to conclude that

$$0 \in \partial(\gamma_{\mathcal{C}}(x^*) + \delta_{\mathcal{P}}(x^*)) = \partial\gamma_{\mathcal{C}}(x^*) + \partial\delta_{\mathcal{P}}(x^*),$$

which holds if and only if x^* is optimal. The assumption that $f(x^*) = \tau$ together with Hiriart-Urruty and Lemaréchal [39, Theorem D.1.3.5] guarantees that there exists a positive scalar λ such that

$$0 \in \partial\gamma_{\mathcal{C}}(x^*) + \lambda\nabla f(x^*) \quad \text{i.e., } z^* \in \operatorname{cone}\partial\gamma_{\mathcal{C}}(x^*).$$

Then by Corollary 3.4, it's equivalent to say that the pair (x^*, z^*) is $\mathcal{C}$-aligned. □

5.1.1 Objective Value Bound

With only slightly more effort, Theorem 5.1 implies that the residual in the satisfaction of the polar inequality can be used to bound the difference between the objective value $f(x)$ and the optimal value $f(x^*)$ of any solution x^*.

Let α^* be an upper bound on the gauge value $\gamma_{\mathcal{C}}(x^*)$ of any optimal solution x^*, and define

$$g_{\mathcal{C}}(x) = \alpha^*\sigma_{\mathcal{C}}(z_x) - \langle x, z_x\rangle,$$

as the residual in the polar inequality, where

$$z_x := -\nabla f(x).$$

Although a bound α^* isn't generally available, a notable exception is for problems of the form (5.1b), where feasibility implies that $\gamma_{\mathcal{C}}(x^*) \le \alpha$, and in that case we may simply take $\alpha^* = \alpha$. To see how g_c provides the bound on the optimal value of f, note that

$$\begin{aligned} f(x^*) &\ge f(x) + \langle x^* - x, \nabla f(x)\rangle \\ &\ge f(x) + \min_{a\in\alpha^*\mathcal{C}} \langle a - x, \nabla f(x)\rangle \\ &= f(x) + \langle x, z\rangle - \alpha^*\sigma_{\mathcal{C}}(z), \end{aligned}$$

where the first inequality follows from the subgradient inequality. Rearranging terms and using the definition of g_c, we obtain the bound

$$g_c(x) \ge f(x) - f(x^*) \quad \forall x.$$

Jaggi [17] and Ndiaye *et al.* [52] derive a similar bound in the context of the conditional gradient method applied to (5.1b).

5.2 Gauge Optimization

Consider the problem of finding the minimum-gauge element of a convex set. This broad class of problems, originally proposed by Freund [27], generalizes a range of problems, including specialized classes such as least-norm solutions to linear systems and convex conic programming; see Aravkin *et al.* [53] and Friedlander *et al.* [28]. These conceptually simple problems have a special duality characterized by the following pair of dual problems:

$$\begin{array}{llcll} \underset{x}{\text{minimize}} & \gamma_{\mathcal{C}}(x) & \qquad & \underset{z}{\text{minimize}} & \sigma_{\mathcal{C}}(z) \\ \text{subject to} & x \in \mathcal{D}, & & \text{subject to} & z \in \mathcal{D}', \end{array} \tag{5.2}$$

where $\mathcal{D} \subset \mathbb{R}^n \backslash \{0\}$ is any closed convex set and

$$\mathcal{D}' := \{z \mid \langle x, z\rangle \geq 1 \ \forall x \in \mathcal{D}\}$$

is its antipolar. The following results describes the alignment correspondence between primal and dual solutions.

Proposition 5.1 (Polar Duality). A pair of primal-dual feasible vectors $(x, z) \in \mathcal{D} \times \mathcal{D}'$ is primal-dual optimal for (5.2) if and only if they are $\mathcal{C}$-aligned and $\langle x, z\rangle = 1$.

Proof. First, assume that the pair (x, z) is primal-dual optimal for (5.2). Then by strong duality [28, Corollary 5.2],

$$1 = \langle x, z\rangle = \gamma_{\mathcal{C}}(x) \cdot \sigma_{\mathcal{C}}(z). \tag{5.3}$$

We prove the other direction by contradiction. Assume (x, z) are $\mathcal{C}$-aligned and $\langle x, z\rangle = 1$ and suppose there exists $\widehat{x} \in C$ such that $\gamma_{\mathcal{C}}(\widehat{x}) < \gamma_{\mathcal{C}}(x)$. It then follows that

$$\langle \widehat{x}, z\rangle \overset{\text{(i)}}{\geq} 1 \overset{\text{(ii)}}{=} \gamma_{\mathcal{C}}(x) \cdot \sigma_{\mathcal{C}}(z) > \gamma_{\mathcal{C}}(\widehat{x}) \cdot \sigma_{\mathcal{C}}(z),$$

where the inequality (i) follows from the definition of the antipolar $\mathcal{D}'$, and the equality (ii) follows from (5.3). This violates the polar gauge inequality, and thus leads to a contradiction. □

Example 5.2 (Phase Retrieval)**.** The phase retrieval problem aims to recover a complex signal $x^\natural \in \mathbb{C}^n$ from magnitude-only measurements

$$b_i = |\langle x^\natural, f_i\rangle|^2 \quad \text{for } i = 1 : m,$$

where $f_i \in \mathbb{C}^n$, $i = 1 : m$, are the measurement vectors. This problem has many applications, including X-ray crystallography [54], optical imaging [55], and more [56]. Here we consider the PhaseLift formulation proposed by Candès *et al.* [57]. By lifting the signal $x^\natural$, the measurements can be expressed equivalently as

$$b_i = \langle x^\natural (x^\natural)^*, f_i f_i^*\rangle = \langle X^\natural, F_i\rangle \quad \text{for } i = 1 : m,$$

where $X^\natural := x^\natural(x^\natural)^*$ is the lifted signal and each $F_i := f_i f_i^*$ is a lifted measurement matrix. Candès *et al.* [57] show that $m = \mathcal{O}(n \log n)$ measurements are sufficient to recover $X^\natural$ with high probability by solving the convex semidefinite problem

$$\underset{X\in\mathbb{C}^{n\times n}}{\text{minimize}} \;\; \mathrm{tr}(X) \;\; \text{subject to} \;\; \mathcal{F}X = b, \;\; X \succeq 0, \tag{5.4}$$

where the linear operator $\mathcal{F}\colon \mathbb{C}^{n\times n} \to \mathbb{R}^m$ is defined by (1.1). Move the semidefinite constraint on X into the redefined objective $\mathrm{tr}(X)+\delta_{\succeq 0}(X)$, which is a gauge function, as shown by Example 4.9. It then follows from (5.2) that the gauge dual of problem (5.4) is given by

$$\underset{y\in\mathbb{R}^m}{\text{minimize}} \;\; \max\{0, \lambda_{\max}(F^* y)\} \;\; \text{subject to} \;\; \langle b, y\rangle \geq 1. \tag{5.5}$$

Compared with the primal problem (5.4), which has dimension n^2, the gauge dual problem (5.5) has a much smaller dimension on the order of $n \log n$, and only a single linear constraint. Thus, it's more efficient to instead solve the dual problem to obtain an optimal dual solution y^*, which then can be used to expose the primal support. In particular, use Proposition 5.1 and Example 4.9 to deduce that the primal solution must have the form $X = USU^T$, where the columns of $U \in \mathbb{R}^{n\times r}$ form a basis for the r-dimensional maximal eigenspace of $Z = \mathcal{F}^* y^*$, and the r-by-r positive semidefinite matrix S is the solution of the reduced primal problem

$$\underset{S\in\mathbb{C}^{r\times r}}{\text{minimize}} \;\; \mathrm{tr}(S) \;\; \text{subject to} \;\; \mathcal{F}(USU^T) = b, \;\; S \succeq 0.$$

In practice, r is usually very small, and in some cases, $r = 1$. Thus, the reduced primal problem will be easy to solve. This approach, including the noisy phase retrieval problem, is developed by Friedlander and Macêdo [38]. $\square$

6

Alignment in Optimization Methods

We show in Section 5 how the alignment between a vector and a gradient signals the vector's optimality for gauge-constrained and gauge-regularized optimization problems. In this section, we interpret many known methods for solving these problem, including proximal-gradient and conditional-gradient methods, in terms of alignment. In particular, we show that for many of these methods, each iterate achieves an approximate alignment between the iterates $x^{(k)}$ and $z^{(k)} := -\nabla f(x^{(k)})$. Moreover, we also show that the alignment condition can be used to prove the equivalence between seemingly unrelated primal and dual methods, such as the augmented Lagrangian and bundle methods.

6.1 Proximal Gradient and Mirror Descent Methods

Proximal-gradient (PG) methods [14], [58] are widely used for problems such as (5.1a), which involve the sum of a smooth and a nonsmooth function. Because the regularization parameter ρ in (5.1a) can be absorbed into the atomic set $\mathcal{A}$, as shown in Proposition 3.2(d), there is no loss in generality in assuming that $\rho = 1$. We thus restrict our

Algorithm 6.1: Proximal-gradient method for problem (6.1).

```
0 Input: $x^{(0)} \in \mathbb{R}^n$, $f^{(0)} = f(x^{(0)})$, $\lambda^{(0)} > 0$,  $\beta \in (0,1)$
1 for $k = 0, 1, 2, \ldots$ do
2     $z^{(k)} = -\nabla f(x^{(k)})$
3     $x^{(k+1)} = \mathrm{prox}_{\lambda^{(k)}\gamma_{\mathcal{A}}}(x^{(k)} + \lambda^{(k)} z^{(k)})$
4     $f^{(k+1)} = f(x^{(k+1)})$
5     $\epsilon^{(k)} = (x^{(k)} - x^{(k+1)})/\lambda^{(k)}$
6     if $f^{(k+1)} \leq f^{(k)} + \lambda^{(k)}\langle z^{(k)}, \epsilon^{(k)}\rangle + \lambda^{(k)}/2\|\epsilon^{(k)}\|_2^2$ then break
7     $\lambda^{(k+1)} = \beta\lambda^{(k)}$
8 return $x^{(k+1)}$
```

attention to the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) + \gamma_{\mathcal{A}}(x). \tag{6.1}$$

The iterates of the basic PG method are summarized in Algorithm 6.1, as drawn from Parikh and Boyd [14, Section 4.2].

For any convex function g, the proximal map, used in Line 3 of the algorithm, is defined by

$$\mathrm{prox}_{\lambda g}(u) = \underset{x\in\mathbb{R}^n}{\arg\min} \; g(x) + \frac{1}{2\lambda}\|x - u\|_2^2.$$

Applied to the function $g := \gamma_{\mathcal{A}}$, it's straightforward to instead express this line of the algorithm as

$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{\arg\min} \; \gamma_{\mathcal{A}}(x) - \langle z^{(k)}, x\rangle + \frac{1}{2\lambda^{(k)}}\|x - x^{(k)}\|^2. \tag{6.2}$$

Interpretation as Alignment. Theorem 5.1 asserts that any optimal point x^* must be $\mathcal{A}$-aligned with the negative gradient $z^* := -\nabla f(x^*)$, which implies

$$\langle x^*, z^*\rangle = \gamma_{\mathcal{A}}(x^*) \cdot \sigma_{\mathcal{A}}(z^*).$$

Again apply Theorem 5.1, but this time to (6.2). Rearrange terms to deduce that the updated iterate $x^{(k+1)}$ is aligned with $z^{(k)} + \epsilon^{(k)}$, i.e.,

$$\langle x^{(k+1)}, z^{(k)} + \epsilon^{(k)}\rangle = \gamma_{\mathcal{A}}(x^{(k+1)}) \cdot \sigma_{\mathcal{A}}(z^{(k)} + \epsilon^{(k)}), \tag{6.3}$$

where the convergence residual $\epsilon^{(k)}$ is computed in Line 5 of Algorithm 6.1. Thus, each iterate $x^{(k+1)}$ is $\mathcal{A}$-aligned with the corrected gradient $z^{(k)} + \epsilon^{(k)}$.

A similar phenomenon occurs when applying the mirror descent method [59], [60] to the same problem (6.1). The mirror descent method includes many popular algorithm variations, including weighted majority [61] and boosting [62]. The method generalizes the proximal-gradient iteration (6.2) to

$$x^{(k+1)} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ \gamma_{\mathcal{A}}(x) - \langle z^{(k)}, x\rangle + \frac{1}{\lambda^{(k)}} D_\phi(x, x^{(k)}), \tag{6.4}$$

where the *Bregman function* ϕ: $\mathbb{R}^n \to \mathbb{R}$ is continuously-differentiable and strictly convex, and induces the *Bregman divergence*

$$D_\phi(x, x^{(k)}) = \phi(x) - \phi(x^{(k)}) - \langle \nabla\phi(x^{(k)}), x - x^{(k)}\rangle.$$

The mirror descent iteration (6.4) is equivalent to the classical proximal gradient iteration (6.2) with the Bregman function $\phi = \frac{1}{2}\|\cdot\|_2^2$. Similar to the classical case, the generalized proximal map (6.4) generates iterates $x^{(k+1)}$ and $z^{(k)} + \epsilon^{(k)}$ that are $\mathcal{A}$-aligned and satisfy (6.3), except that the residual quantity

$$\epsilon^{(k)} := \frac{1}{\lambda^{(k)}}(\nabla\phi(x^{(k)}) - \nabla\phi(x^{(k+1)}))$$

is defined with respect to the gradient map of Bregman function.

6.2 Conditional Gradient Method

Conditional gradient (CG) methods [15]–[17] naturally exhibit the atomic alignment property in several ways. Here we describe an important property related to alignment useful for implementing memory efficient variations of this class of methods.

In its simplest form, the CG method applies to problems formulated as (5.1b), which feature a smooth objective function. Because here we wish to make explicit the atomic set in the constraint, we instead express that problem in the equivalent form

$$\operatorname*{minimize}_{x\in\widehat{\mathcal{A}}} \ \ f(x). \tag{6.5}$$

Algorithm 6.2: Conditional gradient method for (6.5).

```
0 Input: $x^{(0)} \in \mathcal{A}$, $\epsilon > 0$
1 for $k = 0, 1, 2, \ldots$ do
2     $z^{(k)} = -\nabla f(x^{(k)})$
3     $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$
4     if $\langle a^{(k)} - x^{(k)}, z^{(k)} \rangle < \epsilon$ then break      [break if optimal]
5     $x^{(k+1)} = \theta^{(k)} a^{(k)} + (1 - \theta^{(k)}) x^{(k)}, \quad \theta^{(k)} \in (0, 1)$
6 return $x^{(k)}$
```

We adopt throughout this subsection the simplifying assumption that the atomic set $\mathcal{A}$ is compact, which implies that every direction exposes a well-defined face of the closed convex hull $\widehat{\mathcal{A}}$. Algorithm 6.2 summarizes the iterates of the basic CG method.

The linear minimization oracle (LMO) in Line 3 selects an atom or a convex combination of atoms from the set $\mathcal{A}$ exposed by the current negative gradient $z^{(k)} := -\nabla f(x^{(k)})$. In the language of atomic alignment, the LMO step selects an atom $a^{(k)}$ that is $\mathcal{A}$-aligned with $z^{(k)}$. In particular, observe

$$\langle a^{(k)}, z^{(k)} \rangle = \sigma_{\mathcal{A}}(z^{(k)}) = \gamma_{\mathcal{A}}(a^{(k)}) \cdot \sigma_{\mathcal{A}}(z^{(k)}),$$

where the second equality follows from $a^{(k)} \in \mathcal{A}$—i.e., $\gamma_{\mathcal{A}}(a^{(k)}) = 1$.

Line 5 of Algorithm 6.2 merges the selected element $a^{(k)}$ with the collection of atoms exposed at previous iterations, and thus the latest iterate $x^{(k)}$ represents an aggregate of these atoms. Various choices for the steplength $\theta^{(k)}$ exist, including a steplength derived from a linesearch on the objective function f (which requires additional evaluations of the function to ensure sufficient decrease) and a decaying steplength that follows a predetermined schedule.

We express the merge step at iteration k recursively as

$$x^{(k)} = \sum_{i=0}^{k} \widehat{\theta}^{(i)} a^{(i)}, \quad \widehat{\theta}^{(i)} := \theta^{(i)} \prod_{j=0}^{i} (1 - \theta^{(j)}). \tag{6.6}$$

This expression makes explicit the one-atom-at-a-time construction of the current iterate $x^{(k)}$, each taken from a face exposed by the negative gradients. Thus, the latest iterate lies in the convex hull of the faces exposed up to iteration k:

$$x^{(k)} \in \sum_{i=1}^{k} \widehat{\theta}^{(i)} \mathcal{F}_{\mathcal{A}}(z^{(i)}).$$

In an idealized, perfectly greedy run of the algorithm, the sequence of exposed faces $\mathcal{F}_{\mathcal{A}}(z^{(k)})$ are expanding, i.e., $\mathcal{F}_{\mathcal{A}}(z^{(k)}) \subseteq \mathcal{F}_{\mathcal{A}}(z^{(k+1)})$, and converge to an optimal face $\mathcal{F}_{\mathcal{A}}(z^*)$, where $z^* := -\nabla f(x^*)$. In practice, however, we do not expect such efficiency, and the algorithm may inadvertently collect many atoms not at all related to the optimal face. Thus, the computed decomposition (6.6) at any iteration may contain atoms $a^{(k)}$ not in the optimal support $\mathcal{S}_{\mathcal{A}}(x^*)$. In applications such as matrix-completion, described in Example 6.2 below, the cost of storing intermediate atoms $a^{(k)}$—say, as singular pairs $(u^{(k)}, v^{(k)})$—can be prohibitively expensive for large problems. Various modifications of the basic CG method aim to compress or trim the collected atoms to reduce unnecessary storage [63].

The recent appeal of these methods lies with the computational efficiency of the linear minimization oracle (Line 3 of Algorithm 6.2) for many important atomic sets, especially those where the associated projection or proximal operations are not computationally feasible. The next example illustrates the point.

Example 6.1 (Euclidean Projection Onto the Nuclear-Norm Ball). Suppose the n-by-m matrix X has the singular-value decomposition

$$X = U\,\mathrm{Diag}(\{\sigma_i(X)\}_{i=1}^{m\wedge n})V^T.$$

The projection of X onto the unit nuclear-norm ball $\mathcal{A} := \{Z \mid \|Z\|_* \leq 1\}$ is given by the matrix

$$\mathrm{proj}_{\mathcal{A}}(X) = U\bar{\Sigma}V^T,$$

where the diagonal matrix

$$\bar{\Sigma} = \mathrm{Diag}(\min\{1, \sigma_i(X)\}_{i=1}^{m\wedge n})$$

contains the singular values of X truncated to unit value. Thus, the projection operation requires computing all of the singular triples of X

up to unit value. In contrast, the linear minimization oracle in Line 3 of Algorithm 6.2 requires only computing one of the maximal singular triples of the negative gradient (a matrix, in this case). For this reason, the CG method often features in applications of matrix completion [64]–[66], which are characterized typically by high-dimensional data. □

6.3 Constrained Least-Squares

For problems with a least-squares objective function, the alignment principle provides a simple device that can be used to reduce storage requirements for the CG method. Instead of storing intermediate primal iterates and atoms, we store only reduced versions of these quantities. We illustrate this approach with the constrained least-squares problem

$$\underset{x \in \widehat{\mathcal{A}}}{\text{minimize}} \quad \tfrac{1}{2}\|Mx - b\|_2^2, \tag{6.7}$$

where $M\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator with $m < n$, b is an m-vector, and $\widehat{\mathcal{A}}$ coincides with the constraint of the generic constrained problem (6.5). The approach requires storage of order $\mathcal{O}(m)$, and thus is most effective when $m \ll n$.

Algorithm 6.3 describes a dual version of the standard CG method shown in Algorithm 6.2. It's similar to the approach implemented by Yurtsever *et al.* [67], who maintain a low-memory random sketch of the primal iterate $x^{(k)}$. Algorithm 6.3, however, forgoes direct reference to the primal iterate during the CG iteration loop, and instead tracks a sequence of m-vectors that satisfy the relations

$$\begin{aligned} r^{(k)} &= b - Mx^{(k)}, & p^{(k)} &= Ma^{(k)}, \\ z^{(k)} &= M^* r^{(k)} = -\nabla f(x^{(k)}), & q^{(k)} &= Mx^{(k)}, \\ \Delta r^{(k)} &= M(a^{(k)} - x^{(k)}). \end{aligned}$$

The negative-gradient vector $z^{(k)} \to z^* = -\nabla f(x^*)$. The corresponding approximation $x^{(k)}$ to the primal solution x^* is subsequently recovered in Line 11. The justification for this step is based on the $\mathcal{A}$-alignment between x^* and z^*, as spelled out by Theorem 5.1. In some applications,

Algorithm 6.3: Dual conditional gradient for the constrained least-squares problem (6.7).

```
0  Input: $M$, $b$, $\epsilon > 0$
1  $r^{(0)} = b$;  $q^{(0)} = 0$
2  for $k = 0, 1, 2, \dots$ do
3      $z^{(k)} = M^* r^{(k)}$                                  [$z^{(k)} \equiv -\nabla f(x^{(k)})$]
4      $p^{(k)} \in M\mathcal{F}_{\mathcal{A}}(z^{(k)})$        [$p^{(k)} \equiv Ma^{(k)}$,  $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$]
5      $\Delta r^{(k)} = p^{(k)} - q^{(k)}$                     [$\Delta r^{(k)} \equiv M(a^{(k)} - x^{(k)})$]
6      $\rho^{(k)} = \langle \Delta r^{(k)}, r^{(k)} \rangle$   [optimality gap]
7      if $\rho^{(k)} < \epsilon$ then break                     [break if optimal]
8      $\theta^{(k)} = \min\{1,\ \rho^{(k)}/\|\Delta r^{(k)}\|_2^2\}$   [exact linesearch]
9      $r^{(k+1)} = r^{(k)} - \theta^{(k)} \Delta r^{(k)}$      [$r^{(k+1)} \equiv b - Mx^{(k+1)}$]
10     $q^{(k+1)} = q^{(k)} + \theta^{(k)} \Delta r^{(k)}$      [$q^{(k+1)} \equiv Mx^{(k+1)}$]
11 $x^{(k)} \in \arg\min_x \{\frac{1}{2}\|Mx - b\|_2^2 \mid \mathcal{S}_{\mathcal{A}}(x) \subseteq \tau\mathcal{E}_{\mathcal{A}}(z^{(k)})\}$
12 return $x^{(k)}$
```

however, it may be sufficient to skip this step and instead return the set of exposed atoms $\mathcal{E}_{\mathcal{A}}(z^{(k)})$.

The optimality test on Line 7 is equivalent to the optimality test in Algorithm 6.2 because

$$\begin{aligned}\langle \Delta r^{(k)}, r^{(k)} \rangle &= \langle p^{(k)} - q^{(k)}, r^{(k)} \rangle \\ &= \langle M(a^{(k)} - x^{(k)}), r^{(k)} \rangle \\ &= \langle a^{(k)} - x^{(k)}, M^* r^{(k)} \rangle = \langle a^{(k)} - x^{(k)}, z^{(k)} \rangle.\end{aligned}$$

The linesearch parameter $\theta^{(k)}$ is an exact minimizer of the equivalent objective function $\|r^{(k)} - \theta \Delta r^{(k)}\|_2$ over $\theta \in [0, 1]$.

The following example describes an application of Algorithm 6.3 to solve the matrix-completion problem.

Example 6.2 (Matrix Completion and Delayed atom Generation). Consider the low-rank matrix completion problem

$$\underset{X \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \tfrac{1}{2}\|\Omega \circ X - B\|_F^2 \quad \text{subject to} \quad \|X\|_* \le \tau. \tag{6.8}$$

This problem appears in recommender systems [68], where the (i, j)th element of the sparse matrix B records the rating score given by user i

for product j. Ratings are observed only for a subset of user-product pairs indexed by the binary mask

$$\Omega_{ij} = \begin{cases} 1 & \text{if user } i \text{ has rated product } j; \\ 0 & \text{otherwise.} \end{cases}$$

The goal is to predict the unseen ratings, captured in the dense unknown matrix X. A structural low-rank assumption is used to capture an "archetype" phenomenon—users who often like the same movies serve as good predictors for each other, and movies that are liked by the same users probably are also similar. Therefore, we consider each user as a sparse linear combination of archetypal individuals (and similarly with products), where the inner product of their feature vectors give the same prediction rating. The nuclear-norm constraint on X is a common approach for encouraging low-rank solutions [12].

Algorithm 6.4 describes a specialization of the dual CG method (Algorithm 6.3) for the matrix-completion problem (6.8). Most of the computational cost for an implementation of this specialization is represented in Line 4, which computes a maximal singular triple of the negative gradient

$$Z^{(k)} \equiv -\nabla f(X^{(k)}) = \Omega \circ (B - X^{(k)}),$$

represented as a sparse matrix indexed by Ω. Thus, the atoms exposed by $Z^{(k)}$ are rank-1 outer products in the set

$$\mathcal{E}_{\mathcal{A}}(Z^{(k)}) = \{uv^T \in \mathcal{A} \mid \langle u, Z^{(k)}v \rangle = \sigma_{\max}(Z^{(k)})\},$$

where $\mathcal{A}$ is the unit nuclear norm ball; cf. Example 2.6. Algorithm 6.4 does not store these pairs or their aggregate in a primal iteration matrix $X^{(k)}$. Instead, the algorithm computes only a sparse matrix of the form $\Omega \circ (uv^T)$, which requires minimal storage. The algorithm returns the primal iterate in the factored form $X^{(k)} := USV^T$.

Table 6.1 lists the results of applying the primal and dual CG algorithm variants on a set of random matrix-completion problems. For varying problem sizes with $m = n$, we generate the binary mask Ω with 10% nonzeros, and generate the observation matrix

$$B = \Omega \circ (UV^T + 0.1 \cdot N),$$

Table 6.1: Performance of the primal and dual variants of conditional gradient for the matrix-completion problem (Example 6.2) after 10 iterations of Algorithm 6.2 (primal CG) and Algorithm 6.4 (dual CG). Estimated rank of final solution is computed as the smallest number of singular values that account for 90% of its Frobenious norm. Time is measured in seconds

Size	**Primal CG**			**Dual CG**		
$m = n$	**Residual**	**Rank**	**Time**	**Residual**	**Rank**	**Time**
100	10.3	6	0.0	10.3	1	0.0
250	25.2	6	0.1	25.2	1	0.1
1,000	100.4	6	1.3	100.4	1	0.3
5,000	501.7	6	48.3	501.7	1	11.4
10,000	998.9	6	242.9	998.9	1	63.3

Algorithm 6.4: Specialization of the dual conditional gradient method (Algorithm 6.3) for the matrix-completion problem (6.8).

```
Input: $\Omega$, $B$, $\ell$
$R^{(k)} = \Omega \circ B;\ Q^{(k)} = 0$
for $k = 1, 2, \ldots$ do
    $Z^{(k)} = \Omega \circ R^{(k)}$
    $(u, v) = \texttt{svds(}Z^{(k)}\texttt{, 1)}$            [expose atom $A^{(k)} \equiv \tau u v^T$]
    $\Delta R^{(k)} = \Omega \circ (\tau u v^T) - Q^{(k)}$      [$\Omega \circ (\tau u v^T) \equiv \tau (u_i v_i)_{(ij) \in \Omega}$]
    $\rho^{(k)} = \langle \Delta R^{(k)}, R^{(k)} \rangle$                [optimality gap]
    if $\rho^{(k)} < \epsilon$ then break                 [break if optimal]
    $\theta^{(k)} = \min\{1,\ \rho^{(k)} / \|\Delta R^{(k)}\|_F^2\}$       [exact linesearch]
    $R^{(k+1)} = R^{(k)} - \theta^{(k)} \Delta R^{(k)}$          [$R^{(k+1)} \equiv \Omega \circ (B - X^{(k)})$]
    $Q^{(k+1)} = Q^{(k)} + \theta^{(k)} \Delta R^{(k)}$          [$Q^{(k+1)} \equiv \Omega \circ X^{(k)}$]
$(U, V, \Sigma) = \texttt{svds(}Z^{(k)}\texttt{, }\ell\texttt{)}$                [top $\ell$ singular vectors]
$S \in \arg\min_S \left\{ \frac{1}{2} \|\Omega \circ (U S V^T - B)\|_2^2 \mid \mathrm{tr}(S) \le \tau,\ S \succeq 0 \right\}$
return $(U, S, V)$
```

where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$, and N are generated i.i.d. from a standard Gaussian distribution. The "true rank" r was set to **round**$(m/100)$. Interestingly, the leading singular value of the final computed dual solution estimate $Z^{(k)}$ was always isolated (i.e., multiplicity 1), which

Algorithm 6.5: Fully-corrective conditional gradient method for problem (6.5).

```
Input: $x^{(0)} \in \mathcal{A}$, $\epsilon > 0$,  $\mathcal{A}^{(0)} = \{0\}$
for $k = 0, 1, 2, \dots$ do
    $z^{(k)} = -\nabla f(x^{(k)})$
    $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$
    if $\langle a^{(k)} - x^{(k)}, z^{(k)} \rangle < \epsilon$ then break
    $\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \cup \{a^{(k)}\}$        [optionally compress $\mathcal{A}^{(k+1)}$]
    $x^{(k+1)} = \arg\min_x \{f(x) \mid x \in \operatorname{conv} \mathcal{A}^{(k+1)}\}$
return $x^{(k)}$
```

made trivial the primal-recovery phase in Line 12 of Algorithm 6.4. The residual values between the two variants are the same, confirming that they recover solutions of similar quality. The dual variant, however, is significantly faster because it doesn't need to manipulate storage for a primal iterate $X^{(k)}$. □

6.4 Simplicial Conditional Gradient and Cutting Planes

The simplicial version of the CG method maintains a collection of ν atoms exposed through to iteration k, and replaces the 1-dimensional linesearch in Line 5 of Algorithm 6.2 with a ν-dimensional linesearch over the convex hull of the collected atoms. (The term *simplicial* refers to the simplex in which the atom weights lie.) When $\nu = k$ (i.e., the collection contains a complete history of exposed atoms), the method is also known as the fully-corrective conditional gradient (FC-CG) method [17], [69]. Although there are no generic guarantees that this algorithm performs better than the basic CG method, for some problems an exact minimizer of the original problem can be obtained after a small number of iterations. One relevant example is orthogonal matching pursuit [70], which can be interpreted as a special case of the FC-CG method. The iterates of the FC-CG variant for problem (6.5) are summarized in Algorithm 6.5.

Algorithm 6.6: Cutting plane method for problem (6.9).

```
0 Input: $z^{(0)} \in \mathbb{R}^n$, $\epsilon > 0$,  $\mathcal{A}^{(0)} = \{0\}$
1 for $k = 0, 1, 2, \ldots$ do
2     $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$
3     if $\langle a^{(k)}, z^{(k)} \rangle - \sigma_{\mathcal{A}^{(k)}}(z^{(k)}) < \epsilon$ then break
4     $\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \cup \{a^{(k)}\}$        [optionally compress $\mathcal{A}^{(k+1)}$]
5     $z^{(k+1)} = \arg\min_z \{f^*(-z) + \sigma_{\mathcal{A}^{(k+1)}}(z)\}$
6 return $z^{(k)}$
```

Bertsekas and Yu [71] show that the FC-CG method and Kelley's cutting plane (CP) method [72] are dual variations of the same algorithm. Here we show the equivalence of the two methods by the atomic alignment property. We begin with the Fenchel–Rockafellar dual problem of (6.5)

$$\underset{z \in \mathbb{R}^n}{\text{minimize}} \quad f^*(-z) + \sigma_{\mathcal{A}}(z). \tag{6.9}$$

The iterates of the CP method applied to problem (6.9) are summarized by Algorithm 6.6. The version we describe is a *partial* CP method that builds a cutting-plane model for the function $\sigma_{\mathcal{A}}$, but uses the conjugate function f^* directly [73]. The key step in Algorithm 6.6 is Line 5, where $\sigma_{\mathcal{A}^{(k)}}$ can be viewed as a sublinear minorant for $\sigma_{\mathcal{A}}$ built from a collection of k atoms in $\mathcal{A}$. Proposition 6.3 shows that the iteration pair $(x^{(k)}, z^{(k)})$ generated by Algorithms 6.5 and 6.6 are $\mathcal{A}^{(k)}$-aligned and identical. It also follows that stopping conditions in both algorithms are equivalent, and thus that both algorithms are equivalent.

Proposition 6.3 (Equivalence Between FC-CG and CP Methods)**.** Assume that $f \colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex, differentiable and finite everywhere. Suppose that Algorithms 6.5 and 6.6, respectively, are initialized with the iterates $x^{(0)}$ and $z^{(0)} := -\nabla f(x^{(0)})$. If the atom-selection strategy from the exposed faces $\mathcal{F}_{\mathcal{A}}(z^{(k)})$ is deterministic, then the iterates $x^{(k)}$ and $z^{(k)}$ generated by the algorithms are $\mathcal{A}^{(k)}$-aligned, and $z^{(k)} = -\nabla f(x^{(k)})$ for all $k \geq 1$.

Proof. It is sufficient to prove that for all $k \geq 1$, $x^{(k)}$ and $z^{(k)}$ can be obtained as the solutions to the saddle-point problem

$$\min_x \max_z \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z). \tag{6.10}$$

First, we verify $x^{(k)}$:

$$\begin{aligned} x^{(k)} &= \arg\min_{x \in \operatorname{conv} \mathcal{A}^{(k)}} \ f(x) \\ &= \arg\min_x \ f(x) + \delta_{\operatorname{conv} \mathcal{A}^{(k)}}(x) \\ &= \arg\min_x \ f(x) + \sigma^*_{\mathcal{A}^{(k)}}(x) \\ &= \arg\min_x \ f(x) + \max_z \left[\langle x, z\rangle - \sigma_{\mathcal{A}^{(k)}}(z)\right] \\ &= \arg\min_x \max_z \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z). \end{aligned} \tag{6.11}$$

Next, we verify $z^{(k)}$:

$$\begin{aligned} z^{(k)} &= \arg\max_z \ -f^*(-z) - \sigma_{\mathcal{A}^{(k)}}(z) \\ &= \arg\max_z \ \min_x \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z). \end{aligned} \tag{6.12}$$

Note that Line 6 in Algorithm 6.5 and Line 5 in Algorithm 6.6 are Fenchel–Rockafellar duals to each other and problem (6.10) is the associated saddle-point problem. Since f is finite everywhere, by [35, Theorem 31.1] we know that the strong duality holds and it follows that the $\min_x$ and $\max_z$ in problem (6.10) can be exchanged. So $z^{(k)}$ is indeed the maximizer in the saddle-point problem (6.10).

Since f is strongly convex and differentiable, it follows that both $x^{(k)}$ and $z^{(k)}$ are unique. Differentiate the inner minimization problem in (6.12) with respect to x to obtain

$$z^{(k)} = -\nabla f(x^{(k)}).$$

Subdifferentiate the inner maximization in (6.11) with respect to z to obtain

$$x^{(k)} \in \mathcal{F}_{\mathcal{A}^{(k)}}(z^{(k)}),$$

which implies that $x^{(k)}$ and $z^{(k)}$ are $\mathcal{A}^{(k)}$ aligned by Corollary 3.4. □

6.5 Connections to the Augmented Lagrangian Method

In the previous subsection, we show the equivalence between FC-CG method and CP method through alignment. In this subsection, we generalize this result to augmented Lagrangian fully-corrective conditional gradient (AL-FC-CG) method and proximal bundle (PB) method. To the best of our knowledge, we are the first to talk about such equivalence.

Consider Line 6 in Algorithm 6.5, which can be equivalently expressed as

$$\underset{x\in\mathbb{R}^n, u\in\mathbb{R}^n}{\text{minimize}} \quad f(x) + \delta_{\operatorname{conv}\mathcal{A}^{(k)}}(u) \ \text{ subject to } \ x = u. \tag{6.13}$$

The augmented Lagrangian methods, originate from Hestenes [74], replace the constraint $x = u$ with a penalty function that promotes the feasibility. Specifically, given an augmented Lagrangian parameter $\lambda > 0$ and a Lagrange multiplier $z \in \mathbb{R}^n$, (6.13) will become

$$\begin{aligned} x^{(k+1)} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \min_{u\in\mathbb{R}^n} f(x) + \delta_{\operatorname{conv}\mathcal{A}^{(k)}}(u) + \langle z, x-u\rangle + \frac{\lambda}{2}\|x-u\|_2^2 \\ &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ f(x) + \min_{u\in\operatorname{conv}\mathcal{A}^{(k)}} \left(\langle z, x-u\rangle + \frac{\lambda}{2}\|x-u\|_2^2\right) \\ &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ f(x) + \frac{\lambda}{2} \min_{u\in\operatorname{conv}\mathcal{A}^{(k)}} \left\|x + \frac{1}{\lambda}z - u\right\|_2^2 \\ &= \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \ f(x) + \frac{\lambda}{2}\operatorname{dist}^2_{\mathcal{A}^{(k)}}\left(x + \frac{1}{\lambda}z\right). \end{aligned} \tag{6.14}$$

The detailed AL-FC-CG method for (6.5) is shown in Algorithm 6.7.

PB method was first introduced by Kiwiel [75] as a stabilization of CP method. The detailed PB method for (6.9) is shown in Algorithm 6.8. Similar to Algorithm 6.6, here we only build CP model for $\sigma_{\mathcal{A}}$ and f^* is unchanged. The main improvement is Line 5 in Algorithm 6.8, where the difference is the addition of a quadratic penalty function. The next proposition shows the equivalence between Algorithm 6.7 and Algorithm 6.8.

Algorithm 6.7: Augmented Lagrangian fully-corrective conditional gradient method for problem (6.5).

```
Input: $x^{(0)} \in \mathcal{A}$, $\epsilon > 0$, $\mathcal{A}^{(0)} = \{0\}$, $\lambda > 0$
for $k = 0, 1, 2, \dots$ do
    $z^{(k)} = -\nabla f(x^{(k)})$
    $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$
    if $\langle a^{(k)} - x^{(k)}, z^{(k)} \rangle < \epsilon$ then break
    $\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \cup \{a^{(k)}\}$        [optionally compress $\mathcal{A}^{(k+1)}$]
    $x^{(k+1)} = \arg\min_x \{f(x) + \frac{\lambda}{2} \operatorname{dist}^2_{\mathcal{A}^{(k+1)}}(x + \frac{1}{\lambda} z^{(k)})\}$
return $x^{(k)}$
```

Algorithm 6.8: Proximal bundle method for problem (6.9).

```
Input: $z^{(0)} \in \mathbb{R}^n$, $\epsilon > 0$, $\mathcal{A}^{(0)} = \{0\}$, $\lambda > 0$
for $k = 0, 1, 2, \dots$ do
    $a^{(k)} \in \mathcal{F}_{\mathcal{A}}(z^{(k)})$
    if $\langle a^{(k)}, z^{(k)} \rangle - \sigma_{\mathcal{A}^{(k)}}(z^{(k)}) < \epsilon$ then break
    $\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} \cup \{a^{(k)}\}$        [optionally compress $\mathcal{A}^{(k+1)}$]
    $z^{(k+1)} = \arg\min_z \{f^*(-z) + \sigma_{\mathcal{A}^{(k+1)}}(z) + \frac{1}{2\lambda}\|z - z^{(k)}\|_2^2\}$
return $z^{(k)}$
```

Proposition 6.4 (Equivalence Between AL-FC-CG Method and PB Method)**.** Assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is strongly convex, differentiable and finite everywhere. Suppose that Algorithms 6.7 and 6.8, respectively, are initialized with the iterates $x^{(0)}$ and $z^{(0)} := -\nabla f(x^{(0)})$. If the atom-selection strategy from the exposed faces $\mathcal{F}_{\mathcal{A}}(z^{(k)})$ is deterministic, then the iterates $x^{(k)} + \epsilon^{(k)}$ and $z^{(k)}$ generated by the algorithms are $\mathcal{A}^{(k)}$-aligned with

$$\epsilon^{(k)} = \frac{1}{\lambda}(z^{(k-1)} - z^{(k)}),$$

and $z^{(k)} = -\nabla f(x^{(k)})$ for all $k \geq 1$.

Proof. It is sufficient to prove that for all $k \geq 1$, $x^{(k)}$ and $z^{(k)}$ can be obtained as the solutions to the saddle-point problem

$$\min_x \max_z \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z) - \frac{1}{2\lambda}\|z - z^{(k)}\|^2. \tag{6.15}$$

First, we verify $x^{(k)}$:

$$\begin{aligned} x^{(k)} &= \arg\min_x \ f(x) + \frac{\lambda}{2}\operatorname{dist}^2_{\mathcal{A}^{(k)}}\left(x + \frac{1}{\lambda}z^{(k-1)}\right) \\ &= \arg\min_x \ f(x) + \min_{u \in \operatorname{conv}\mathcal{A}^{(k)}}\left(\langle z^{(k-1)}, x - u\rangle + \frac{\lambda}{2}\|x - u\|^2\right) \\ &= \arg\min_x \ f(x) + \delta_{\operatorname{conv}\mathcal{A}^{(k)}} \square \left(\langle z^{(k-1)}, \cdot\rangle + \frac{\lambda}{2}\|\cdot\|^2\right)(x) \\ &= \arg\min_x \ f(x) + \left(\sigma_{\mathcal{A}^{(k)}}(\cdot) + \frac{1}{2\lambda}\|\cdot - z^{(k-1)}\|^2\right)^*(x) \\ &= \arg\min_x \ f(x) + \max_z \left(\langle x, z\rangle - \sigma_{\mathcal{A}^{(k)}}(z) - \frac{1}{2\lambda}\|z - z^{(k-1)}\|^2\right) \\ &= \arg\min_x \max_z \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z) - \frac{1}{2\lambda}\|z - z^{(k-1)}\|^2, \end{aligned} \tag{6.16}$$

where the second equality follows from (6.14) and the fourth equality follows from Rockafellar [35, Theorem 16.4]. Next, we check $z^{(k)}$:

$$\begin{aligned} z^{(k)} &= \arg\max_z \ -f^*(-z) - \sigma_{\mathcal{A}^{(k)}}(z) - \frac{1}{2\lambda}\|z - z^{(k-1)}\|^2 \\ &= \arg\max_z \min_x \ \langle x, z\rangle + f(x) - \sigma_{\mathcal{A}^{(k)}}(z) - \frac{1}{2\lambda}\|z - z^{(k-1)}\|^2. \end{aligned} \tag{6.17}$$

Note that Line 6 in Algorithm 6.7 and Line 5 in Algorithm 6.8 are Fenchel–Rockafellar duals to each other and problem (6.15) is the associated saddle-point problem. Since f is finite everywhere, by [35, Theorem 31.1] we know that the strong duality holds and it follows that the $\min_x$ and $\max_z$ in problem (6.15) can be exchanged. So $z^{(k)}$ is indeed the maximizer in the saddle-point problem (6.15).

Since f is strongly convex and differentiable, it follows that both $x^{(k)}$ and $z^{(k)}$ are unique. Differentiate the inner minimization problem in (6.17) with respect to x to obtain

$$z^{(k)} = -\nabla f(x^{(k)}).$$

Subdifferentiate the inner maximization in (6.16) with respect to z to obtain

$$x^{(k)} + \epsilon^{(k)} \in \mathcal{F}_{\mathcal{A}^{(k)}}(z^{(k)}),$$

which implies that $x^{(k)} + \epsilon^{(k)}$ and $z^{(k)}$ are $\mathcal{A}^{(k)}$ aligned by Corollary 3.4. □

7

Alignment in Convolution of Atomic Sets

The theory of atomic decomposition and polar alignment developed thus far is tied to a single atomic set $\mathcal{A}$. Correspondingly, the regularized optimization problems considered in Section 5 involve only a single gauge regularization function $\gamma_{\mathcal{A}}$ meant to encourage minimizers that have sparse atomic decompositions with respect to $\mathcal{A}$. Richer atomic decompositions and regularized formulations, however, may be constructed by combining different atomic sets. We describe in this section the theory of polar convolution, which mixes any number of atomic sets to form computable regularization functions useful for defining complex decompositions with respect to multiple atomic sets.

Consider the atomic decomposition of an n-vector x with respect to two atomic sets $\mathcal{A}_1$ and $\mathcal{A}_2$:

$$x = \sum_{a_1 \in \mathcal{A}_1} c_{a_1} a_1 + \sum_{a_2 \in \mathcal{A}_2} c_{a_2} a_2 \quad c_{a_i} \geq 0 \ \forall a_i. \tag{7.1}$$

(It's convenient to restrict our discussion to two atomic sets, but all of the results in this section readily extend to three or more sets.) This decomposition appears often in models for separating a mixture of structurally different signals, each one well represented by atoms from one of the base atomic sets $\mathcal{A}_1$ or $\mathcal{A}_2$ [76]–[81].

The common approach to constructing sparse representations using multiple atomic sets is to form a regularization function from the sum of corresponding gauges, i.e., $\gamma_{\mathcal{A}_1} + \gamma_{\mathcal{A}_2}$. However, the alignment principles that connect atomic decompositions and regularization, including their algorithmic implications, don't apply to these constructions.

Thus, we focus on an alternative and less-often used approach that forms an aggregate atomic set from the vector sum

$$\mathcal{A}_1 + \mathcal{A}_2 = \{a_1 + a_2 \mid a_1 \in \mathcal{A}_1,\ a_2 \in \mathcal{A}_2\}$$

of the base atomic sets. This construction directly mirrors the desired decomposition in (7.1). As we show in this section, the sum of atomic sets corresponds to the *polar convolution* of the corresponding gauge functions [82]. This connection to the theory of polar convolution allows us to extend properties of alignment to aggregate atomic sets, and also to extend the dual algorithms discussed in Section 6. We show how these extensions lead to practical approaches for optimization formulations that arise in demixing applications.

7.1 Atomic Sums

One important application of the alignment principles that we discuss in this section is in the analysis of the various demixing problems

$$\underset{x_1,\, x_2}{\text{minimize}} \quad f(x_1 + x_2) + \rho \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2)\}, \tag{7.2a}$$

$$\underset{x_1,\, x_2}{\text{minimize}} \quad f(x_1 + x_2) \ \text{ subj to } \ \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2)\} \le \alpha, \tag{7.2b}$$

$$\underset{x_1,\, x_2}{\text{minimize}} \quad \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2)\} \ \text{ subj to } \ f(x_1 + x_2) \le \tau. \tag{7.2c}$$

Compared with the formulations in (5.1), these formulations aim to decompose a solution as $x = x_1 + x_2$, where each component x_i has a sparse structure with respect to the atomic set $\mathcal{A}_i$. The regularization function

$$\max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2)\}$$

replaces the single regularizer $\gamma_{\mathcal{C}}$, used in the generic problems (5.1). The central result of this subsection is a corollary to Theorem 5.1 that reveals the optimal alignment property that each of the computed components x_i holds with respect to $\mathcal{A}_i$.

Corollary 7.1 (Optimality and Atomic Sums). Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and sets $\mathcal{A}_i \subset \mathbb{R}^n$, $i = 1, 2$, contain the origin in its interior. Assume that at the respective solutions, the gauge values are positive in all three problems and the constraints in (7.2b) and (7.2c) hold with equality. Then the vectors x_1^* and x_2^* are optimal if and only if x_i^* is $\mathcal{A}_i$-aligned with $z^* := -\nabla f(x_1^* + x_2^*)$ for $i = 1, 2$, and

$$\begin{aligned} \sigma_{\mathcal{A}_1}(z^*) + \sigma_{\mathcal{A}_2}(z^*) &= \rho \quad \text{for problem (7.2a);} \\ \gamma_{\mathcal{A}_1}(x_1^*) = \gamma_{\mathcal{A}_2}(x_2^*) &= \alpha \quad \text{for problem (7.2b);} \\ f(x_1^* + x_2^*) &= \tau \quad \text{for problem (7.2c).} \end{aligned}$$

Before we can establish the proof of this result, we first require several tools from polar convolution and its connection to the sum of atomic sets. The proof of Corollary 7.1 is given in Subsection 7.3.1.

7.2 Polar Convolution

The polar convolution of two gauges functions $\gamma_{\mathcal{A}_1}$ and $\gamma_{\mathcal{A}_2}$ is defined by the function

$$(\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2})(x) = \inf_{x_1, x_2} \max \{\gamma_{\mathcal{A}_1}(x_1), \gamma_{\mathcal{A}_2}(x_2) \mid x = x_1 + x_2\}. \tag{7.3}$$

This operation first appears in Rockafellar [35, Theorem 5.8] for general convex functions, and is subsequently analyzed by Seeger and Volle [83]. When specialized to gauge functions, as shown in (7.3), this convolution operation is tightly connected to the polarity operation applied to the defining atomic sets. In that case, Friedlander *et al.* [82] refer to the operation as *polar convolution.*

The next result shows that the gauge values are necessarily equal at a solution of the infimum defined in (7.3).

Lemma 7.2 (Balancing in Polar Convolution). Suppose that the sets $\mathcal{A}_1$ and $\mathcal{A}_2$ contain the origin in their interiors and that the infimum in (7.3) is achieved at (x_1^*, x_2^*). Then

$$\gamma_{\mathcal{A}_1}(x_1^*) = \gamma_{\mathcal{A}_2}(x_2^*).$$

Proof. We proceed by contradiction. Without loss of generality, assume that

$$\gamma_{\mathcal{A}_1}(x_1^*) > \gamma_{\mathcal{A}_2}(x_2^*) = \gamma_{\mathcal{A}_2}(x - x_1^*).$$

Then

$$\begin{aligned} 0 &\in \partial \max \{\gamma_{\mathcal{A}_1}(x_1^*),\ \gamma_{\mathcal{A}_2}(x - x_1^*)\} \\ &= \partial \gamma_{\mathcal{A}_1}(x_1^*) \\ &= \partial \sigma_{\mathcal{A}_1^\circ}(x_1^*), \end{aligned}$$

where the inclusion follows from the optimality of x_1^* for (7.3), the first equality follows from Hiriart-Urruty and Lemaréchal [39, Theorem D.4.4.2], and the second equality follows from Proposition 3.2(b). Thus, $\sigma_{\mathcal{A}_1^\circ}(x_1^*) = 0$. As a consequence,

$$\gamma_{\mathcal{A}_2}(x_2^*) < \gamma_{\mathcal{A}_1}(x_1^*) = \sigma_{\mathcal{A}_1^\circ}(x_1^*) = 0,$$

which contradicts the fact the gauge function is non-negative. Therefore, we must have

$$\gamma_{\mathcal{A}_1}(x_1^*) = \gamma_{\mathcal{A}_2}(x_2^*).$$ □

The next result demonstrates that the polar convolution of gauges is the functional counterpart to the vector sum of the underlying sets.

Proposition 7.3 (Polar Convolution of Gauges)**.** Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be nonempty closed convex sets that contain the origin. If at least one set contains the origin in its interior, then the polar convolution of the gauges $\gamma_{\mathcal{A}_1}$ and $\gamma_{\mathcal{A}_2}$ is the gauge

$$\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2} = \gamma_{\mathcal{A}_1 + \mathcal{A}_2}.$$

Proof. The hypothesis that one of the sets $\mathcal{A}_1$ or $\mathcal{A}_2$ contains the origin implies that the corresponding gauge (say, $\gamma_{\mathcal{A}_1}$) is finite and therefore continuous. Thus,

$$\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2} = (\gamma_{\mathcal{A}_1^\circ} + \gamma_{\mathcal{A}_1^\circ})^\circ = \gamma_{\mathcal{A}_1 + \mathcal{A}_2},$$

where the first equality follows from [82, Lemma 3.3] and the continuity of $\gamma_{\mathcal{A}_1}$, and the second equality follows from [82, Lemma 3.4]. □

For the remainder of this subsection, we assume that any gauge $\gamma_{\mathcal{A}_i}$ are continuous, which holds if the origin is contained in the interior of its generating set $\mathcal{A}_i$.

7.3 Alignment to the Sum of Sets

The polar convolution operation, which mixes atoms via the sum of sets, has the appealing property that it explicitly decomposes a vector as a sum of elements, each belonging to one of the atomic sets. In particular, evaluating the polar convolution

$$\begin{aligned}(\gamma_{\mathcal{A}_1} \diamond \gamma_{\mathcal{A}_2})(x) &= \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) \\ &= \inf_{x_1,\, x_2} \max\left\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2) \mid x = x_1 + x_2\right\}\end{aligned}$$

at a point x implicitly generates a decomposition

$$x = \sum_{a \in \mathcal{A}_1 + \mathcal{A}_2} c_a a = \sum_{\substack{a_1 \in \mathcal{A}_1 \\ a_2 \in \mathcal{A}_2}} c_a (a_1 + a_2) = x_1 + x_2,$$

where each $x_i \in \operatorname{cone} \mathcal{A}_i$, i.e., x_i has a valid atomic decomposition with respect to $\mathcal{A}_i$. The vectors x_i in this decomposition exhibit an alignment property with their corresponding atomic sets $\mathcal{A}_i$, as described by the next result.

Theorem 7.1 (Alignment in Polar Convolution)**.** Suppose that the pair of n-vectors (x, z) is $(\mathcal{A}_1 + \mathcal{A}_2)$-aligned and

$$0 < \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) = \gamma_{\mathcal{A}_1}(x_1) = \gamma_{\mathcal{A}_2}(x_2), \quad \text{where } x = x_1 + x_2.$$

Then the pair (x_i, z) is $\mathcal{A}_i$-aligned for $i = 1, 2$.

Proof. Because x and z are $(\mathcal{A}_1 + \mathcal{A}_2)$-aligned,

$$\gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) \cdot \sigma_{\mathcal{A}_1+\mathcal{A}_2}(z) = \langle x, z \rangle = \langle x_1, z \rangle + \langle x_2, z \rangle.$$

Use the fact that $\sigma_{\mathcal{A}_1+\mathcal{A}_2} = \sigma_{\mathcal{A}_1} + \sigma_{\mathcal{A}_2}$ and rearrange terms to deduce

$$\sigma_{\mathcal{A}_1}(z) + \sigma_{\mathcal{A}_2}(z) = \left\langle \frac{x_1}{\gamma_{\mathcal{A}_1}(x_1)}, z \right\rangle + \left\langle \frac{x_2}{\gamma_{\mathcal{A}_1}(x_2)}, z \right\rangle. \tag{7.4}$$

Because $x_i / \gamma_{\mathcal{A}_i}(x_i) \in \mathcal{A}_i$ it follows that

$$\sigma_{\mathcal{A}_i}(z) \geq \left\langle \frac{x_i}{\gamma_{\mathcal{A}_i}(x_i)}, z \right\rangle, \quad i = 1, 2.$$

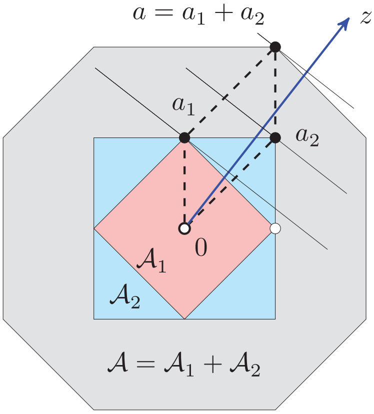


Figure 7.1: Illustration of the polar alignment principle for atomic sums, as described by Theorem 7.1 (alignment in polar convolution). The vector z simultaneously exposes atoms, indicated by black dots, in the atomic sets $\mathcal{A}_1$ and $\mathcal{A}_2$, and also in the sum of atomic sets $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$.

Therefore, the equality (7.4) implies that

$$\gamma_{\mathcal{A}_i}(x_i) \cdot \sigma_{\mathcal{A}_i}(z) = \langle x_i, z \rangle, \quad i = 1, 2,$$

which establish, respectively, that each (x_i, z) is $\mathcal{A}_i$-aligned. □

Figure 7.1 illustrates this result.

7.3.1 Proof of Corollary 7.1

The first step in the proof is to establish that the regularized optimization problems in (7.2) are equivalent, respectively, with the problems

$$\underset{x}{\text{minimize}} \quad f(x) + \rho\gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) \tag{7.5a}$$

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to } \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) \le \alpha, \tag{7.5b}$$

$$\underset{x}{\text{minimize}} \quad \gamma_{\mathcal{A}_1+\mathcal{A}_2}(x) \quad \text{subject to} \quad f(x) \le \tau. \tag{7.5c}$$

We establish the equivalence for (7.5a); the equivalence for (7.5b) and (7.5c) follows the same line of reasoning. Observe that

$$\begin{aligned}
&\inf_{x_1,\, x_2} \{f(x_1 + x_2) + \rho \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x_2)\}\} \\
&\quad = \inf_{x,\, x_1} \{f(x) + \rho \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x - x_1)\}\} \\
&\quad = \inf_{x} \{f(x) + \rho \inf_{x_1} \max\{\gamma_{\mathcal{A}_1}(x_1),\, \gamma_{\mathcal{A}_2}(x - x_1)\}\} \\
&\quad = \inf_{x} \{f(x) + \rho \gamma_{\mathcal{A}_1 + \mathcal{A}_2}(x)\},
\end{aligned}$$

where the last equality follows from the definition of polar convolution (7.3) and Proposition 7.3.

Next, use Theorem 5.1 to establish that a point x^* is a solution to one of the three problems (7.5) if and only if x^* is $(\mathcal{A}_1 + \mathcal{A}_2)$-aligned with $z^* := -\nabla f(x^*)$. The equivalence of the formulations (7.5) and (7.2) implies that $x^* = x_1^* + x_2^*$, where x_1^* and x_2^* are optimal for (7.2). Moreover, optimality of x_1^* and x_2^* implies that $\gamma_{\mathcal{A}_1+\mathcal{A}_2}(x^*) = \gamma_{\mathcal{A}_1}(x_1^*) = \gamma_{\mathcal{A}_2}(x_2^*)$. Thus, Theorem 7.1 applies in this case and each pair (x_i^*, z^*) is $\mathcal{A}_i$-aligned. □

7.4 Morphological Component Analysis

We show how the alignment principle can be used as part of a demixing application in signal separation known as morphological component analysis [81]. Our discussion focuses on demixing using the constrained formulation (7.5b), but can be easily extended to the other regularized formulations shown in (7.5).

Suppose that x^* is the solution of (7.5b). Then $x^* = x_1 + x_2$ for some $x_i \in \alpha\mathcal{A}_i$. We recover the constituent components x_i using two stages. In the first stage, we apply the conditional gradient method (Algorithm 6.2) to (7.5b) with $\mathcal{A} := \mathcal{A}_1 + \mathcal{A}_2$ to obtain the negative gradient $z^* = -\nabla f(x^*)$. (The primal iterate $x^{(k)}$ doesn't need to be stored). The key to the efficient application of this method is to recognize that the exposed face of the sum of sets is equal to the sum of exposed faces, i.e.,

$$\mathcal{F}_{\mathcal{A}_1+\mathcal{A}_2}(z) = \mathcal{F}_{\mathcal{A}_1}(z) + \mathcal{F}_{\mathcal{A}_2}(z).$$

Thus, Step 3 in the CG method can be implemented using separate procedures available for exposing a face in each of the atomic sets $\mathcal{A}_i$.

In the second stage, we use the vector z^* obtained in the first stage to expose the atoms in each component x_i. Theorem 7.1 asserts that each x_i is $\mathcal{A}_i$-aligned with the negative gradient $z^* := -\nabla f(x^*)$, and therefore exposes the atoms in $\mathcal{A}_i$ that supports x_i. Thus, each component x_i can be recovered as the solution of the reduced optimization problem

$$\underset{x_1,\,x_2}{\text{minimize}} \quad f(x_1 + x_2) \quad \text{subject to} \quad \gamma_{\mathcal{E}_{\mathcal{A}_i}(z^*)}(x_i) \le \alpha, \qquad i = 1, 2.$$

The underlying assumption, of course, is that the exposed face $\mathcal{F}_{\mathcal{A}_i}(z^*)$ containing the relevant atoms has small dimension, since otherwise this problem could be as expensive as the original problem. A variety of algorithms can be applied to solve this reduced problem.

Although our discussion above considered only two atomic sets, the analysis extends easily to any number of atomic sets. The example below illustrates how to use the alignment principles to separate a mixture of three signals.

Example 7.4 (Separating Background from Foreground in a Noisy Image). We give a concrete example from morphological component analysis that illustrates how this approach can be used in practice to separate background and foreground from a noisy image. Suppose that the m-vector

$$b = x_s + x_\ell + \epsilon$$

encodes a 2-dimensional image composed of a sparse component x_s, a low-rank component x_ℓ, and structured noise ϵ. The ability to decouple b into these three components rests on their incoherence [76], [79], [80]. Because our aim here is only to illustrate the polar-alignment property, we make the simplifying assumption that the noise ϵ is sparse in the Fourier basis, which is known to be incoherent with sparsity and low-rank. Based on these assumptions, we choose $\mathcal{A}_1$ to be the unit 1-norm ball (Example 2.5), $\mathcal{A}_2$ to be the nuclear-norm ball (Example 2.6), and $\mathcal{A}_3 = D^T\mathcal{A}_1$, where D is the discrete cosine transform. Use Proposition 3.2(c) to deduce that the gauge that corresponds to $\mathcal{A}_3$ is the transformed 1-norm:

$$\gamma_{D^T\mathcal{A}_1}(v) = \gamma_{\mathcal{A}_1}(Dv) = \|Dv\|_1.$$

We follow the approach outlined in Subsection 7.4. For the first stage, we apply the dual CG method to the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \frac{1}{2}\|x-b\|_2^2 \\ \text{subject to} & \gamma_{\mathcal{A}_1+\mathcal{A}_2+\mathcal{A}_3}(x) \leq \tau, \end{array}$$

to obtain the negative gradient $z^* := b - x^*$ (without storing the primal iterates $x^{(k)}$ or solution x^*). In the second stage, the primal solution x^* is recovered by solving the problem

$$\begin{array}{ll} \underset{c_a^{(1)},c_a^{(2)},c_a^{(3)}}{\text{minimize}} & \frac{1}{2}\|x-b\|_2^2 \\ \text{with} & x = \sum_{i=1,2,3} \sum_{a^{(i)} \in \mathcal{E}_{\mathcal{A}_i}(z^*)} c_a^{(i)} a^{(i)} \end{array}$$

over the coefficients $c_a^{(i)}$. Because in this case the atomic sets are centrosymmetric, we may ignore the nonnegativity requirements of the coefficients.

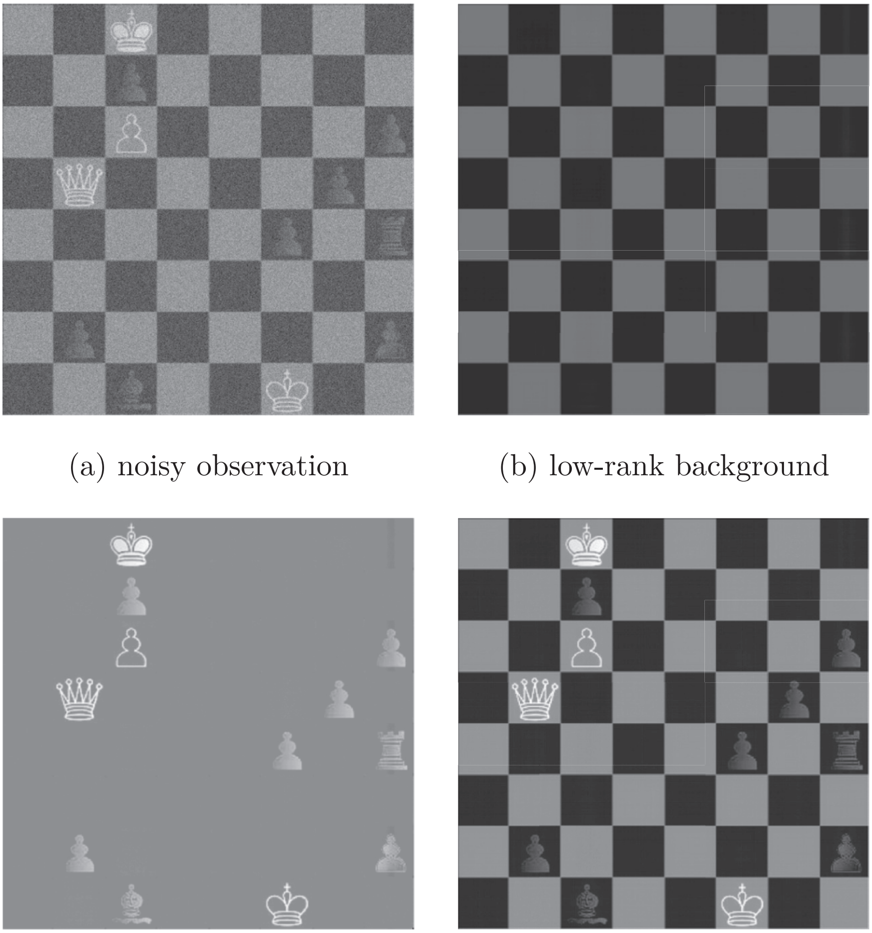
The first panel in Figure 7.2 shows a noisy 500-by-500 pixel image of a chess board. The remaining panels show the separated images obtained after 2000 iterations of the CG algorithm as described above. □

7.5 Atomic Unions and Sum Convolution

The infimal sum convolution between two gauges $\gamma_{\mathcal{A}_1}$ and $\gamma_{\mathcal{A}_2}$ is defined through the optimization problem

$$(\gamma_{\mathcal{A}_1} \Box \gamma_{\mathcal{A}_2})(x) = \inf_{x_1,x_2} \left\{ \gamma_{\mathcal{A}_1}(x_1) + \gamma_{\mathcal{A}_2}(x_2) \mid x = x_1 + x_2 \right\}.$$

Although here we define this operation only for gauges, it can be applied to any two convex functions and always results in another convex function [35, Theorem 5.4]. Normally the operation is simply called *infimal convolution*, but here we use the term *sum convolution* to distinguish it from the polar convolution operation (i.e., infimal max convolution) that we use in Subsection 7.1.



(a) noisy observation

(b) low-rank background

(c) sparse foreground

(d) reconstructed image

Figure 7.2: Morphological component analysis via polar convolution is used to denoise and separate foreground from background in an image. The chess-board image in panel (a) has been corrupted with noise. Panels (b) and (c) show the extracted low-rank and sparse parts of the image, which are assembled in panel (d) as the final reconstruction of the observed image (a). See Example 7.4.

Proposition 7.5 (Sum Convolution of Gauges)**.** Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be nonempty closed convex sets that contain the origin. The sum convolution of the gauges $\gamma_{\mathcal{A}_1}$ and $\gamma_{\mathcal{A}_2}$ is the gauge

$$\gamma_{\mathcal{A}_1} \square \gamma_{\mathcal{A}_2} = \gamma_{\mathcal{A}_1 \cup \mathcal{A}_2}.$$

Proof. Use Proposition 4.1 to derive the following equivalent expressions:

$$\begin{aligned}
(\gamma_{\mathcal{A}_1} \square \gamma_{\mathcal{A}_2})(x) &= \inf_w \Bigg\{ \inf_{c_a \geq 0} \Bigg\{ \sum_{a \in \mathcal{A}_1} c_a \;\Bigg|\; w = \sum_{a \in \mathcal{A}_1} c_a a \Bigg\} \\
&\qquad + \inf_{c_a \geq 0} \Bigg\{ \sum_{a \in \mathcal{A}_2} c_a \;\Bigg|\; x - w = \sum_{a \in \mathcal{A}_2} c_a a \Bigg\} \Bigg\} \\
&= \inf_{c_a \geq 0,\, w} \Bigg\{ \sum_{a \in \mathcal{A}_1 \cup \mathcal{A}_2} c_a \;\Bigg|\; w = \sum_{a \in \mathcal{A}_1} c_a a, \; x - w = \sum_{a \in \mathcal{A}_2} c_a a \Bigg\} \\
&= \inf_{c_a \geq 0} \Bigg\{ \sum_{a \in \mathcal{A}_1 \cup \mathcal{A}_2} c_a \;\Bigg|\; x = \sum_{a \in \mathcal{A}_1 \cup \mathcal{A}_2} c_a a \Bigg\} \\
&= \gamma_{\mathcal{A}_1 \cup \mathcal{A}_2}(x),
\end{aligned}$$

which establishes the claim. □

8

Conclusions

The theory of polar alignment and its relationship with atomic decompositions offers a rich grammar with which to reason about structured optimization. Of course, the underlying ideas are not entirely new and many of the conclusions can be derived using standard arguments from Lagrange multiplier theory. However, we have found that the theory of polarity and alignment offer a clarifying viewpoint and a powerful suite of tools. Indeed, concepts such as active sets and supports, which are intuitive for polyhedral constraints and vectors, easily extend to more abstract settings when we adopt the vocabulary of alignment, exposed faces, and the machinery of gauges and support functions.

Further research opportunities remain. For example, most (if not all) of the ideas we have presented could be generalized to the infinite-dimensional setting, which would accommodate more general decompositions. Also, other standard algorithms, such as splitting and bundle methods [73], seem to exhibit properties that can easily be explained using the language of polar alignment.

Acknowledgments

The last author (MPF) is indebted to James Burke at the University of Washington, who over years of collaboration and discussion, and through his mastery of convex analysis and optimization, helped to influence many of the ideas in this manuscript.

References

[1] B. A. Murtagh and M. A. Saunders, *MINOS 5.5 User's Guide.* Systems Optimization Laboratory. Tech. Rep. 83-20R, Department of Management Science and Engineering, Stanford University, Stanford, CA, 1983.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[4] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Royal Stat. Soc. Ser. B*, pp. 267–288, 1996.

[6] R. Tibshirani, “The lasso method for variable selection in the Cox model,” *Stat. Med.*, vol. 16, no. 4, pp. 385–395, 1997.

[7] L. Vandenberghe, “The CVXOPT linear and quadratic cone program solvers”, 2010. [Online]. Available: http://www.seas.ucla.edu/~vandenbe/publications/coneprog.pdf.

[8] R. M. Freund, P. Grigas, and R. Mazumder, "An extended Frank–Wolfe method with 'in-face' directions, and its application to low-rank matrix completion," *SIAM J. Optim.*, vol. 27, no. 1, pp. 319–346, 2017.

[9] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, May 2012.

[10] E. M. Gafni and D. P. Bertsekas, "Two-metric projection methods for constrained optimization," *SIAM J. Control Optim.*, vol. 22, no. 6, pp. 936–964, 1984.

[11] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Inter. Conf. Mach. Learning (ICML 2009)*, ACM, pp. 433–440, 2009.

[12] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[13] X. Zeng and M. A. Figueiredo, "The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms," *arXiv:1409.4271*, 2014.

[14] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[15] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics (NRL)*, vol. 3, no. 1–2, pp. 95–110, 1956.

[16] J. C. Dunn and S. Harshbarger, "Conditional gradient algorithms with open loop step size rules," *J. Math. Anal. Appl.*, vol. 62, no. 2, pp. 432–444, 1978.

[17] M. Jaggi, "Revisiting Frank–Wolfe: Projection-free sparse convex optimization," in *Inter. Conf. Mach. Learning (ICML 2013)*, pp. 427–435, 2013.

[18] C. Lemarechal, J.-J. Strodiot, and A. Bihain, "On a bundle algorithm for nonsmooth optimization," in *Nonlinear Programming 4*, Elsevier, 1981, pp. 245–282.

[19] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *J. Mach. Learn. Res.*, vol. 11, pp. 2543–2596, Oct. 2010.

[20] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Automat. Control*, vol. 57, no. 3, pp. 592–606, 2012.

[21] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, ACM, pp. 302–311, 1984.

[22] Y. E. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, vol. 13, ser. Stud. Appl. Math. Philadelphia: Society of Industrial and Applied Mathematics, 1994.

[23] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, ser. MPS/SIAM Series on Optimization. Philadelphia: Society of Industrial and Applied Mathematics, 2001.

[24] J. Lofberg, "Yalmip: A toolbox for modeling and optimization in matlab," in *2004 IEEE International Conference on Robotics and Automation*, IEEE, pp. 284–289, 2004.

[25] M. Grant and S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming (Web Page and Software)*, http://cvxr.com, 2009.

[26] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming," *SIAM J. Optim.*, vol. 10, no. 3, pp. 673–696, 2000.

[27] R. M. Freund, "Dual gauge programs, with applications to quadratic programming and the minimum-norm problem," *Math. Program.*, vol. 38, no. 1, pp. 47–67, 1987.

[28] M. P. Friedlander, I. Macêdo, and T. K. Pong, "Gauge optimization and duality," *SIAM J. Optim.*, vol. 24, no. 4, pp. 1999–2022, 2014.

[29] M. Kocvara and J. Zowe, "An iterative two-step algorithm for linear complementarity problems," *Numerische Mathematik*, vol. 68, no. 1, pp. 95–106, 1994.

[30] A. Cristofari, M. De Santis, S. Lucidi, and F. Rinaldi, "A two-stage active-set algorithm for bound-constrained optimization," *J. Optim. Theory Appl.*, vol. 172, no. 2, pp. 369–401, 2017.

[31] W. Hare and A. S. Lewis, “Identifying active constraints via partial smoothness and prox-regularity,” *J. Convex Anal.*, vol. 11, no. 2, pp. 251–266, 2004.

[32] C. Zălinescu, *Convex Analysis in General Vector Spaces.* World scientific, 2002.

[33] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408. Springer, 2011.

[34] V. Chandrasekaran, B. Recht, P. Parrilo, and A. S. Willsky, “The convex geometry of linear inverse problems,” English, *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[35] R. T. Rockafellar, *Convex Analysis.* Princeton: Princeton University Press, 1970.

[36] J. von Neumann, “Some matrix inequalities and metrization of matric-space,” in *Univ. Tomsk. Rev.* Ser. Collected Works, vol. IV, Oxford: Pergamon, 1962, pp. 205–218.

[37] A. S. Lewis, “The convex analysis of unitarily invariant matrix functions,” *J. Convex Anal.*, vol. 2, no. 1, pp. 173–183, 1995.

[38] M. P. Friedlander and I. Macêdo, “Low-rank spectral optimization via gauge duality,” *SIAM J. Sci. Comput.*, vol. 38, no. 3, A1616–A1638, 2016.

[39] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis.* New York, NY: Springer, 2001.

[40] D. P. Bertsekas, *Convex Optimization Theory.* Athena Scientific, 2009.

[41] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” *IEEE Trans. Sig. Proc.*, vol. 53, no. 7, pp. 2477–2488, 2005.

[42] I. F. Gorodnitsky, J. S. George, and B. D. Rao, “Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm,” *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, 1995.

[43] C. Ding, D. Zhou, X. He, and H. Zha, “R 1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization,” in *Inter. Conf. Mach. Learning (ICML 2006)*, ACM, pp. 281–288, 2006.

[44] M. Bogdan, E. v. d. Berg, W. Su, and E. Candès, "Statistical estimation and testing via the sorted L1 norm," *arXiv:1310.1969*, 2013.

[45] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar," *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.

[46] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Advances in Neural Information Processing Systems (NIPS 2015)*, pp. 2107–2115, 2015.

[47] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems (NIPS 2002)*, pp. 585–591, 2002.

[48] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[49] Y. Chi and M. F. Da Costa, "Harnessing sparsity over the continuum: Atomic norm minimization for superresolution," *IEEE Sig. Proc. Mag.*, vol. 37, no. 2, pp. 39–57, 2020.

[50] J. W. McLean and H. J. Woerdeman, "Spectral factorizations and sums of squares representations via semidefinite programming," *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 3, pp. 646–655, 2002.

[51] B. Dumitrescu, *Positive Trigonometric Polynomials and Signal Processing Applications*, vol. 103. Springer, 2007.

[52] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap safe screening rules for sparse-group lasso," in *Advances in Neural Information Processing Systems (NIPS 2016)*, pp. 388–396, 2016.

[53] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. MacPhee, "Foundations of gauge and perspective duality," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2406–2434, 2018.

[54] R. W. Harrison, "Phase problem in crystallography," *J. Opt. Soc. Am. A*, vol. 10, pp. 1046–1055, 1993.

[55] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: A contemporary overview," *IEEE Sig. Proc. Mag.*, vol. 32, no. 3, pp. 87–109, 2015.

[56] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imag. Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[57] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[58] M. Teboulle, "Convergence of proximal-like algorithms," *SIAM J. Optim.*, vol. 7, 1997.

[59] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization.* New York: Wiley, 1983.

[60] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Oper. Res. Lett.*, vol. 31, no. 3, pp. 167–175, 2003.

[61] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inform. and Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[62] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 771–780, p. 1612, 1999.

[63] N. Rao, P. Shah, and S. Wright, "Forward–backward greedy algorithms for atomic norm regularization," *IEEE Trans. Sig. Proc.*, vol. 63, no. 21, pp. 5798–5811, 2015.

[64] M. Jaggi and M. Sulovsk, "A simple algorithm for nuclear norm regularized problems," in *Inter. Conf. Mach. Learning (ICML 2010)*, pp. 471–478, 2010.

[65] S. Shalev-Shwartz, A. Gonen, and O. Shamir, "Large-scale convex minimization with a low-rank constraint," *arXiv:1106.1622*, 2011.

[66] K. Lee and Y. Bresler, "Efficient and guaranteed rank minimization by atomic decomposition," in *IEEE International Symposium on Information Theory (ISIT 2009)*, IEEE, pp. 314–318, 2009.

[67] A. Yurtsever, M. Udell, J. A. Tropp, and V. Cevher, "Sketchy decisions: Convex low-rank matrix optimization with optimal storage," in *International Conference on Artificial Intelligence and Statistics (AISTATS 2017)*, pp. 1188–1196, 2017.

[68] R. M. Bell and Y. Koren, "Lessons from the Netflix prize challenge," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 75–79, 2007.

[69] C. A. Holloway, "An extension of the Frank and Wolfe method of feasible irections," *Math. Program.*, vol. 6, no. 1, pp. 14–27, 1974.

[70] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[71] D. P. Bertsekas and H. Yu, "A unifying polyhedral approximation framework for convex optimization," *SIAM J. Optim.*, vol. 21, no. 1, pp. 333–360, 2011.

[72] J. E. Kelley Jr, "The cutting-plane method for solving convex programs," *J. Soc. Indust. Appl. Math.*, vol. 8, no. 4, pp. 703–712, 1960.

[73] Z. Fan, Y. Sun, and M. P. Friedlander, "Bundle methods for dual atomic pursuit," in *Asilomar Conference on Signals, Systems, and Computers (ACSSC 2019)*, IEEE, pp. 264–270, 2019.

[74] M. R. Hestenes, "Multiplier and gradient methods," *J. Optim. Theory Appl.*, vol. 4, no. 5, pp. 303–320, 1969.

[75] K. C. Kiwiel, "Proximity control in bundle methods for convex nondifferentiable minimization," *Math. Program.*, vol. 46, no. 1–3, pp. 105–122, 1990.

[76] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems (NIPS 2009)*, pp. 2080–2088, 2009.

[77] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. Assoc. Comput. Mach.*, vol. 58, no. 3, p. 11, 2011.

[78] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *IMA Inform. Inference*, vol. 2, no. 1, pp. 32–68, 2013.

[79] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex demixing, with applications," *Found. Comput. Math.*, vol. 14, no. 3, pp. 503–567, 2014.

[80] S. Oymak and J. A. Tropp, “Universality laws for randomized dimension reduction, with applications,” *IMA Inform. Inference*, vol. 7, no. 3, pp. 337–446, 2017.

[81] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, “Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit,” *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.

[82] M. P. Friedlander, I. Macêdo, and T. K. Pong, “Polar convolution,” *SIAM J. Optim.*, vol. 29, no. 4, pp. 1366–1391, 2019.

[83] A. Seeger and M. Volle, “On a convolution operation obtained by adding level sets: Classical and new results,” *RAIRO Recherche Opérationnelle*, vol. 29, pp. 131–154, 1995.