The Nonconvex Geometry of Linear Inverse Problems

Armin Eftekhari and Peyman Mohajerin Esfahani

Abstract. The gauge function, closely related to the atomic norm, measures the complexity of a statistical model, and has found broad applications in machine learning and statistical signal processing. In a high-dimensional learning problem, the gauge function attempts to safeguard against overfitting by promoting a sparse (concise) representation within the learning alphabet.

In this work, within the context of linear inverse problems, we pinpoint the source of its success, but also argue that the applicability of the gauge function is inherently limited by its convexity, and showcase several learning problems where the classical gauge function theory fails. We then introduce a new notion of statistical complexity, gauge$_p$ function, which overcomes the limitations of the gauge function. The gauge$_p$ function is a simple generalization of the gauge function that can tightly control the sparsity of a statistical model within the learning alphabet and, perhaps surprisingly, draws further inspiration from the Burer-Monteiro factorization in computational mathematics.

We also propose a new learning machine, with the building block of gauge$_p$ function, and arm this machine with a number of statistical guarantees. The potential of the proposed gauge$_p$ function theory is then studied for two stylized applications. Finally, we discuss the computational aspects and, in particular, suggest a tractable numerical algorithm for implementing the new learning machine.

1. Introduction

While data is abundant, information is often sparse, and can be characterized mathematically using a small number of atoms, drawn from an alphabet $\mathcal{A} \subset \mathbb{R}^d$. Concretely, an r-sparse model $x^\sharp$ is specified as $x^\sharp := \sum_{i=1}^r c_i^\sharp A_i^\sharp$, for nonnegative coefficients $\{c_i^\sharp\}_{i=1}^r$ and atoms $\{A_i^\sharp\}_{i=1}^r \subset \mathcal{A}$. Complexity of the model $x^\sharp$ is often measured by its (convex) gauge function $\mathcal{G}_\mathcal{A}$ [1, 2, 3], to be defined later. As a safeguard against overfitting, the gauge function has become a mainstay in linear inverse problems, a large class of learning problems with diverse applications in statistical signal processing and machine learning. More specifically, to discover the true model $x^\sharp$ or its atoms $\{A_i^\sharp\}_{i=1}^r$, the classical gauge function theory studies the (convex) learning machine

$$\min_x \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } \mathcal{G}_\mathcal{A}(x) \leq \gamma, \tag{1.1}$$

or, alternatively, its basis pursuit or lasso reformulations. Above, $\mathcal{L}$ is a linear operator and the vector $y \approx \mathcal{L}(x^\sharp)$ collects m (possibly inexact) observations about the true model $x^\sharp$. A certificate of correctness for the output of the machine (1.1) is at the heart of the classical gauge function theory, and this certificate can be constructed, for example, when $\mathcal{L}$ is a generic linear operator and we have access to sufficiently many observations [1, Corollary 3.3.1]. The literature of the gauge function

Date: January 7, 2021.

The authors are with the Department of Mathematics and Mathematical Statistics, Umea University, Sweden, (Armin.Eftekhari@umu.se), and the Delft Center for Systems and Control, Delft University of Technology, Netherlands (P.MohajerinEsfahani@tudelft.nl).

features numerous successful applications in different areas including statistics [4, 5, 6] and signal processing [7, 8, 9, 10, 11], to name a few. In all these success stories, the gauge function successfully captures the underlying geometry of the learning alphabet, which can be computed efficiently.

The applicability of the gauge function is, however, inherently limited by its convexity. Indeed, there is anecdotal and numerical evidence suggesting that the gauge function is incapable of capturing the geometric details of many learning alphabets, e.g., see the study [12, 13] for sparse principal component analysis (PCA) and [14, 15] in the context of super-resolution. Motivating by these examples, we opt to develop a theoretical foundation along with some basic computational tools for a nonconvex counterpart of the gauge function.

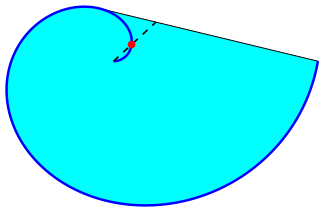


FIGURE 1. A toy example of learning with manifold models that visualizes the limitations of the convex gauge function.

Contributions. Our main objective is to develop a generalized theory, dubbed the gauge$_p$ function theory, addressing the statistical limitations of the classical gauge function theory. More specifically, the following summarizes the contributions of this study:

(i) This work mathematically pinpoints the limitations of the classical gauge function theory (Proposition 2.7 and Observation 2.8), which we then support by concrete examples where the gauge function fails to enforce a sparse representation within the learning alphabet (Examples 2.14—2.16).

(ii) This work proposes and studies the gauge$_p$ function, a simple generalization of the classical gauge function, as a new notion for statistical complexity that can tightly control the sparsity level of a model within the learning alphabet (Proposition 3.4).

(iii) The gauge$_p$ function motivates a new learning machine, for which we develop statistical guarantees that parallel those of the classical gauge function theory (Theorem 3.23). The new theory is showcased with two stylized applications of manifold models and sparse PCA.

(iv) This work also studies the computational aspects of implementing the new learning machine and proposes a tractable algorithm (Proposition 5.2).

Further details and related literature. We provide a section-by-section overview that, whenever necessary, is punctuated with few bibliographic notes borrowed from the corresponding sections.

Section 2.1 mathematically pinpoints the source of success and failure of the classical gauge function theory, which we then corroborate with several examples in Sections 2.2 and 2.3, including Chebyshev systems, group sparsity and manifold models; see also the toy example in Figure 1 where the (blue) curve represents the learning alphabet $\mathcal{A}$, and a level set of the corresponding gauge function $\mathcal{G}_{\mathcal{A}}$ is filled with cyan. The gauge function evidently loses considerable geometric details about the alphabet $\mathcal{A}$. The details can be found in Section 2.3.

Sections 3.1 and 3.2 propose and study the gauge$_p$ function, denoted by $\mathcal{G}_{\mathcal{A},p}$, as a new notion of statistical complexity (Definition 3.3 and Proposition 3.4). The gauge$_p$ function $\mathcal{G}_{\mathcal{A},p}$ generalizes the classical gauge function $\mathcal{G}_{\mathcal{A}}$ and can tightly control the sparsity level of a model within the learning alphabet $\mathcal{A}$. Gauge$_p$ function draws further inspiration from the idea of Burer-Monteiro

factorization [16]. In the success stories of the classical theory (Section 2.2), gauge and gauge$_p$ functions nearly coincide. In contrast, whenever the classical theory fails (Section 2.3), gauge$_p$ function behaves more favourably compared to the classical gauge function, as detailed in Section 3.2. Motivated by this observation, Section 3.3 introduces the learning machine

$$\min_x \|\mathcal{L}(x) - y\|_2^2 \quad \text{subject to} \quad \mathcal{G}_{\mathcal{A},p}(x) \leq \gamma, \tag{1.2}$$

in which the gauge$_p$ function plays the role of regularizer in place of the classical gauge function, see (gauge$_p$). As p varies, the new machine interpolates between two extremes: The classical convex machine (1.1) on one end, and the learning machine

$$\min_x \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } x \text{ has an } r\text{-sparse decomposition in } \mathcal{A}, \tag{1.3}$$

on the other end (Remark 3.13). Above, recall that r is the sparsity level of the true model $x^\sharp$ within the alphabet $\mathcal{A}$. Viewed differently, the new machine (1.2) extends the Burer-Monteiro idea to any alphabet in the sense that (1.2) coincides with the widely-used Burer-Monteiro factorization when the learning alphabet $\mathcal{A}$ is the set of unit-norm rank-1 matrices (Remark 3.12). Implementing the new machine (1.2) often requires solving a nonconvex optimization problem.

Section 3.4 develops some statistical guarantees for the new machine (1.2). In particular, Lemma 3.14 therein introduces a family of certificates for verifying the correctness of the output of the machine (1.2), analogous to Lemma 2.6 for the convex machine (1.1). When $\mathcal{L}$ is a generic linear operator, p is small and m is sufficiently large, we also develop a probabilistic approach to construct these certificates, as detailed in Theorem 3.23, loosely analogous to [1, Corollary 3.3.1] for the convex machine (gauge). The proof technique for Theorem 3.23 appears to be new in this context and might be of independent interest. More specifically, instead of a single certificate, the proof of Theorem 3.23 constructs a family of certificates that *jointly* certify the learning outcome. As a sanity check, we also establish in Proposition 3.16 that the classical gauge function theory is, in a certain nontrivial sense, a special case of the new gauge$_p$ function theory.

In Section 4, we showcase the new theory with two stylized applications, namely, manifold-like models [17] and sparse PCA [18]. Both applications span highly active research areas and it is not our intention to improve over the state of art for these applications, but rather to merely convince the reader that the new machine (1.2) merits further investigation and research.

As mentioned in item (ii) above, implementing the new machine (1.2) often requires solving a nonconvex optimization problem. For certain learning alphabets, such as the one in matrix sensing [19, Chapter 5] or [20, Section 2.1], the landscape of the optimization problem (1.2) is benign (for small p) in the sense that the optimization problem does not have any spurious stationary points and, consequently, problem (1.2) can be solved efficiently [21].

For certain other alphabets, such as smooth manifolds [22], the optimization landscape of (1.2) might in general contain spurious stationary points which could trap first- or second-order optimization algorithms, such as gradient descent. Nevertheless, problem (1.2) is smooth and can be solved efficiently to (near) stationarity, rather than global optimality. This compromise is common in machine learning: As an example, empirical risk minimization is known to be intractable for neural

networks in general, and instead the practitioners seek local (rather than global) optimality by means of first- or second-order optimization algorithms [23, Chapter 20].

For yet other learning alphabets, such as the one in sparse regression [4], the problem (1.2) might be NP-hard in the worst case. Nevertheless, not all is lost here and we draw inspiration from recent developments in mixed-integer programming [24, 25]. Indeed, after decades of research, modern mixed-integer optimization algorithms that directly solve problem (3.18) for sparse regression can now outperform convex heuristics in speed and scalability, and without incurring the well-documented bias of the shrinkage methods. More specifically, inspired by [24], we develop in Section 5 a tractable optimization algorithm to numerically solve the new problem (1.2) when the alphabet is finite and, consequently, problem (1.2) is NP-hard.

To summarize, motivated by the limitations of the classical gauge function theory, this work studies a new learning machine for solving linear inverse problems. The gauge$_p$ function theory, introduced in this work, is far from complete and this study raises several research questions, which require further investigation. For example, both of the applications in Section 4 present interesting opportunities for more in-depth future research. Moreover, this first work is largely focused on the statistical aspects of the new theory and, beyond the preliminary results presented in Section 5, considerable effort is required to better understand the computational aspects of the new learning machine.

Notation. Throughout this study, we adopt the notation from [26] to denote by $\mathrm{lin}(\cdot)$, $\mathrm{aff}(\cdot)$, $\mathrm{cone}(\cdot)$, and $\mathrm{conv}(\cdot)$, the linear, affine, conic, and convex hulls of a set, respectively. A cone is a positive homogeneous subset of a vector space. For a convex set $\mathcal{C}$, its tangent cone at $x \in \mathcal{C}$ is $\mathrm{cone}(\mathcal{C} - x)$, where the subtraction is in the Minkowski's sense. We use $\|x\|_p$ to denote the ℓ_p-norm of a vector $x \in \mathbb{R}^n$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, its convex conjugate is defined as $f^*(z) := \sup_x \langle x, z\rangle - f(x)$. Given a linear operator $\mathcal{L} : \mathbb{X} \to \mathbb{Y}$, defined on a pair of vector spaces $\mathbb{X}$ and $\mathbb{Y}$, the corresponding adjoint operator is denoted by $\mathcal{L}^*$, i.e., $\langle \mathcal{L}(x), y\rangle = \langle x, \mathcal{L}^*(y)\rangle$ for all $(x, y) \in \mathbb{X} \times \mathbb{Y}$. When the spaces are equipped with the norms $(\mathbb{X}, \|\cdot\|_{\mathbb{X}}), (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$, the induced operator norm is denoted by $\|\mathcal{L}\|_{\mathrm{op}} := \sup_{x\in\mathbb{X}} \|\mathcal{L}(x)\|_{\mathbb{Y}}/\|x\|_{\mathbb{X}}$. We also use the notation $[l] := \{1, \cdots, l\}$ for an integer l. Throughout, we always use the convention that $0/0 = 0$.

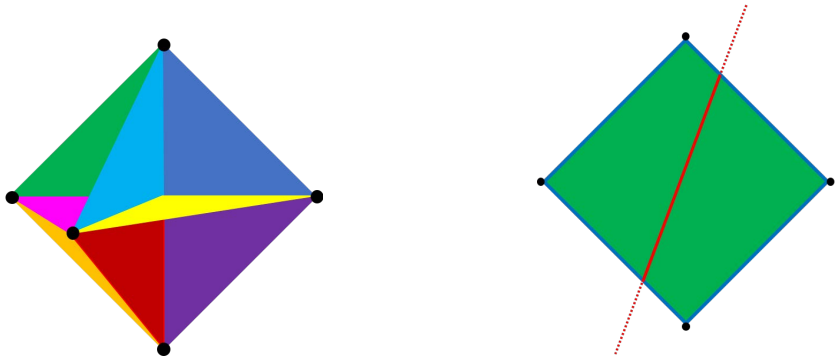
2. Classical (Convex) Gauge Function Theory

As an informal outline of this section, below we will first review the gauge function theory in Section 2.1, and identify a sufficient condition for its success, when given access to unlimited observations. We then feature a few of the success stories of the gauge function theory in Section 2.2, where this key condition is met. We lastly discuss the statistical limitations of the gauge function theory in Section 2.3, where we present several learning problems for which the above key condition is not met. We later aim to address these statistical limitations in Section 3.

2.1. A Geometric Perspective

To review the gauge function theory, this section takes a somewhat different perspective, which appears to be new, to the best of our knowledge. The different geometric perspective of this section

will later help us generalize the classical theory in Section 3. More specifically, this section contains two main results. The first result of this section, Lemma 2.6 below, reviews the standard dual certificate that guarantees successful learning. The second result of this section, Proposition 2.7 below, delineates when this dual certificate exists, *if* the learning machine had access to unlimited data. We now begin with a few definitions. The notion of slice below, visualized in Figure 2a, appears to be new even though it has implicitly appeared before, see for example [27].



(a) Two dimensional slices

(b) Exposed versus hidden faces

FIGURE 2. Pictorial visualization: Figure 2a depicts several two-dimensional slices of the alphabet $\{\pm e_i\}_{i=1}^3$ in different colors where e_i is the i^{th} canonical vector. In Figure 2b, the black dots and the blue line segments are exposed faces while the solid red line segment is a hidden face. The extreme points coincide with the black dots.

Definition 2.1 (Slice)**.** *For an alphabet $\mathcal{A} \subset \mathbb{R}^d$, an integer r and atoms $\{A_i\}_{i=1}^r \subset \mathcal{A}$, the corresponding slice of* $\operatorname{conv}(\mathcal{A})$ *is the set* $\operatorname{conv}(\{A_i\}_{i=1}^r \cup \{0\})$*. We also let* $\operatorname{slice}_r(\mathcal{A})$ *denote the set of all slices of* $\operatorname{conv}(\mathcal{A})$ *formed by at most r atoms.*

Definition 2.1 allows us to rewrite the r-sparse model $x^\sharp = \sum_{i=1}^r c_i^\sharp A_i^\sharp$ in Section 1 as

$$x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp), \qquad \mathcal{S}^\sharp \in \operatorname{slice}_r(\mathcal{A}), \tag{2.1}$$

where the slice $\mathcal{S}^\sharp$ in (2.1) is formed by the atoms $\{A_i^\sharp\}_{i=1}^r$. In convex learning, the complexity of a model, such as $x^\sharp$, is commonly measured by its gauge function, reviewed below [1, 3].

Definition 2.2 (Gauge function)**.** *For an alphabet $\mathcal{A} \subset \mathbb{R}^d$, the gauge function $\mathcal{G}_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}$ is*

$$\begin{aligned} \mathcal{G}_{\mathcal{A}}(x) &:= \inf\left\{t \,:\, x/t \in \operatorname{conv}(\mathcal{A}),\, t \geq 0\right\} \\ &= \inf\left\{\sum_{i=1}^l c_i \,:\, x = \sum_{i=1}^l c_i A_i,\, c_i \geq 0,\; A_i \in \mathcal{A},\, i \in [l]\right\}, \end{aligned} \tag{2.2}$$

with the convention that $0/0 = 0$. Above, $[l] := \{1, \cdots, l\}$ and the infimum is taken over $l, \{c_i\}_i$ and $\{A_i\}_i$.

Let us next collect some classes of alphabets upon which the gauge function theory will be developed.

Assumption 2.3 (Alphabet regularity)**.** *The following assumptions are in order:*

(i) (Origin:) The alphabet $\mathcal{A}$ contains the origin, i.e., $0 \in \mathcal{A}$.
(ii) (Symmetry:) The alphabet $\mathcal{A}$ is symmetric, i.e., $\mathcal{A} = -\mathcal{A}$.
(iii) (Boundedness:) The alphabet $\mathcal{A}$ is bounded, i.e., $\sup_{A\in\mathcal{A}} \|A\|_2 < \infty$.
(iv) (Unit sphere:) The alphabet $\mathcal{A}$ belongs to the unit sphere, i.e., $\|A\|_2 = 1$ for every $A \in \mathcal{A}$.

Under Assumption 2.3(ii), the gauge function $\mathcal{G}_{\mathcal{A}}$ is in fact a norm for $\mathbb{R}^d$ [3], and the unit ball of this norm is $\operatorname{conv}(\mathcal{A})$, i.e.,

$$\operatorname{conv}(\mathcal{A}) = \{x : \mathcal{G}_{\mathcal{A}}(x) \leq 1\}. \tag{2.3}$$

Moreover, the dual norm corresponding to $\mathcal{G}_{\mathcal{A}}$ is denoted by $\mathcal{D}_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}$, and defined as

$$\mathcal{D}_{\mathcal{A}}(z) := \sup\big\{\langle z, x\rangle : \mathcal{G}_{\mathcal{A}}(x) \leq 1\big\} = \sup\big\{\langle z, A\rangle : A \in \mathcal{A}\big\}. \tag{2.4}$$

As a device to control the statistical complexity of learning, the gauge function has found broad applications in statistical signal processing and machine learning. We are particularly interested in linear inverse problems [1], which unify a wide range learning problems, a few of which will be showcased throughout this work. More specifically, for a linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ and an integer r, consider the (exact) model

$$y := \mathcal{L}(x^\sharp), \qquad x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp),\ \mathcal{S}^\sharp \in \operatorname{slice}_r(\mathcal{A}), \tag{exact}$$

where $\operatorname{slice}_r(\mathcal{A})$ was defined in Definition 2.1. In statistical inference or signal processing, for example, $\mathcal{L}$ is the measurement operator and y is the vector of observations [28, 29]. Given y, in order to learn $x^\sharp$ or its sparse decomposition in the alphabet $\mathcal{A}$, consider the learning machine

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } \mathcal{G}_{\mathcal{A}}(x) \leq \mathcal{G}_{\mathcal{A}}(x^\sharp). \tag{gauge}$$

What follows in this section also holds true for the basis pursuit reformulation of the machine (gauge), in which the objective and constraints are swapped and the knowledge of $\mathcal{G}_{\mathcal{A}}(x^\sharp)$ is not required [30]. To study the problem (gauge), let us recall two basic concepts from convex geometry, visualized in Figure 2b, see [26, Definitions 2.6, 3.1].

Definition 2.4 (Extreme point)**.** *An extreme point of a closed convex set $\mathcal{C}$ is a point in $\mathcal{C}$ that cannot be written as a convex combination of other points in $\mathcal{C}$. Let also the set $\operatorname{ext}(\mathcal{C})$ collect all the extreme points of $\mathcal{C}$.*

Definition 2.5 (Face)**.** *For a closed convex set $\mathcal{C}$, the subset $\mathcal{F} \subset \mathcal{C}$ is a face of $\mathcal{C}$ if there exists a hyperplane $\mathcal{H}$ such that $\mathcal{F} = \mathcal{C} \cap \mathcal{H}$. Dimension of a face $\mathcal{F}$ is the dimension of the affine hull of $\mathcal{F}$, i.e., $\dim(\mathcal{F}) = \dim(\operatorname{aff}(\mathcal{F}))$. Moreover, we say that $\mathcal{F}$ is an exposed face of $\mathcal{C}$ if one of the two halfspaces formed by $\mathcal{H}$ contains $\mathcal{C}$. A face $\mathcal{F}$ is hidden if it is not exposed. Lastly, for an integer r, we let $\operatorname{face}_r(\mathcal{A})$ denote the set of all faces of $\operatorname{conv}(\mathcal{A})$ with dimension at most r.*

For an alphabet $\mathcal{A}$, a simple inclusion that we will use frequently in this work is that

$$\operatorname{ext}(\operatorname{conv}(\mathcal{A})) \subseteq \mathcal{A}, \tag{2.5}$$

which states that the extreme points of the convex hull of a set belong to that set. Note also that an exposed 0-dimensional face of a convex set $\mathcal{C}$ is simply an extreme point of $\mathcal{C}$. Equipped with the above definitions, the following result exemplifies learning with the gauge function, in effect stating that the machine (gauge) successfully learns the model $x^\sharp$, provided that a certain certificate of correctness exists. The next lemma is in essence a standard result, see for example [31, Lemma 2.1], though it has not appeared in the literature from the geometric perspective adopted in this section, to the best of our knowledge.

Lemma 2.6 (Dual certificate)**.** *Consider the model $x^\sharp$ in (exact). Suppose that Assumptions 2.3(ii) and (iii) are met. If $x^\sharp = 0$, then the machine (gauge) correctly returns 0. Otherwise, let $\mathcal{F}^\sharp$ be an exposed face of $\operatorname{conv}(\mathcal{A})$ such that $x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \mathcal{F}^\sharp$. Then the machine (gauge) returns $x^\sharp$ if the following holds:*

(i) The linear operator $\mathcal{L}$ in (exact) is injective when restricted to the subspace $\operatorname{lin}(\mathcal{F}^\sharp)$, i.e.,

$$x \in \operatorname{lin}(\mathcal{F}^\sharp) \quad \text{and} \quad \mathcal{L}(x) = 0 \qquad \Longleftrightarrow \qquad x = 0\,.$$

(ii) The face $\mathcal{F}^\sharp$ has a support vector within the range of $\mathcal{L}^*$, where $\mathcal{L}^*$ is the adjoint of the operator $\mathcal{L}$, i.e., there exists $Q \in \operatorname{range}(\mathcal{L}^*)$ such that

$$\langle Q, x - x' \rangle < 0, \qquad \forall x \in \operatorname{conv}(\mathcal{A}) - \mathcal{F}^\sharp,\ \forall x' \in \mathcal{F}^\sharp\,. \tag{2.6}$$

Lemma 2.6 also immediately extends to the basis pursuit formulation of the problem (gauge), in which the objective and constraint are swapped [30]. Designing the dual certificate Q in Lemma 2.6 is perhaps more of an art, often starting with the construction of a dual pre-certificate in $\operatorname{range}(\mathcal{L}^*)$, for which the assertion (2.6) is then verified, see for example [32].

A key fact is that, for a fixed m, constructing the dual certificate in Lemma 2.6 becomes increasingly more difficult as $\dim(\mathcal{F}^\sharp)$ increases. For instance, to respect condition (i) in Lemma 2.6, it is necessary that

$$m \geq \dim(\mathcal{F}^\sharp). \tag{2.7}$$

This fact suggests that the machine (gauge) might fail even when the model $x^\sharp$ has a very sparse decomposition in the alphabet $\mathcal{A}$ and regardless of the choice of the operator $\mathcal{L}$. That is, the machine (gauge) might fail *regardless* of the observations fed to the machine (gauge). To formalize this discussion, we continue with a simple result that identifies a sufficient condition for the success of the machine (gauge), when the data is unlimited.

Proposition 2.7 (Successful convex learning)**.** *Consider the model $x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp)$ in (exact) and let $\mathcal{F}^\sharp$ be an exposed face of $\operatorname{conv}(\mathcal{A})$ such that $x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \mathcal{F}^\sharp$. Suppose also that $\mathcal{S}^\sharp = \operatorname{conv}(\mathcal{F}^\sharp \cup \{0\})$. Lastly suppose that Assumption 2.3(iii) is met. Then there exists a linear operator $\mathcal{L}$ such that the machine (gauge) returns $x^\sharp$ and its r-sparse decomposition in the slice $\mathcal{S}^\sharp$.*

Proposition 2.7 is proved by studying a loss-less linear operator $\mathcal{L}$, particularly the identity map. In words, Proposition 2.7 very loosely establishes that certain alphabets are learnable [23, Chapter 3], in the sense that the machine (gauge) eventually succeeds in recovering $x^\sharp$ when given access to sufficiently many observations. Put differently, again informally speaking, when Proposition 2.7 is

in force, $\mathrm{conv}(\mathcal{A})$ successfully captures the geometric details of the alphabet $\mathcal{A}$, unlike Figure 1. It is worth noting that the tightness of convex relaxation for various alphabets has been an important line of research for decades [33, 34, 35, 13]. Crucially, Proposition 2.7 also leads to the following negative observation.

Observation 2.8 (Failure of convex learning)**.** *For an alphabet $\mathcal{A}$ and the model $x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp)$ in (exact), the machine (gauge) might fail to return $x^\sharp$ or an r-sparse decomposition of $x^\sharp$ in the alphabet $\mathcal{A}$, if the slice $\mathcal{S}^\sharp$ does not contain an exposed face of $\mathrm{conv}(\mathcal{A})$, regardless of the choice of the linear operator $\mathcal{L}$ in (exact).*

The above simple observation is central to our work, and also trivially extends to the basis pursuit formulation of the problem (gauge). In line with Observation 2.8, the statistical failure of the machine (gauge) is visualized with a toy example in Figure 1: In this figure, $\mathcal{L}$ is the identity operator in (exact) and the 1-sparse model $x^\sharp$ is represented by the red dot. Then the machine (gauge) fails to find *any* 1-sparse decomposition of $x^\sharp$, regardless of the number of observations fed to the machine. See also Section 2.3 for more details about this toy example.

We complement Proposition 2.7 and Observation 2.8 by listing several successful and failed applications of the gauge function theory in Sections 2.2 and 2.3, respectively, with examples from Chebyshev systems, sparse PCA, group sparsity, and more.

2.2. Success Stories

We have so far reviewed the gauge function theory and identified in Proposition 2.7 a sufficient condition for the success of the machine (gauge), when data is unlimited. This section reviews a few of the successful applications of the gauge function theory which fulfill that sufficient condition. The reader familiar with sparse regression and low-rank matrix completion might find it easy to browse through this section and then continue with Section 2.3.

Example 2.9 (Sparsity)**.** *This example applies Proposition 2.7 to the alphabet*

$$\mathcal{A} := \{\pm e_i\}_{i=1}^d \subset \mathbb{R}^d, \tag{2.8}$$

which is central to sparse signal processing and high-dimensional statistical inference [11, 4, 36]. Here, $e_i \in \mathbb{R}^d$ is the i^{th} canonical vector, with its i^{th} entry equal to one and the remaining entries equal to zero. Note that sparsity in this example means a concise representation within the alphabet $\mathcal{A}$ and equivalently a small number of nonzero entries. For the alphabet above, it is not difficult to verify from (2.2) that the corresponding gauge function is the ℓ_1-norm, i.e.,

$$\mathcal{G}_{\mathcal{A}} = \|\cdot\|_1, \tag{2.9}$$

and that $\mathrm{conv}(\mathcal{A}) = \{x : \|x\|_1 \le 1\} \subset \mathbb{R}^d$ is the cross polytope. Consider a slice $\mathcal{S}$ of $\mathrm{conv}(\mathcal{A})$ formed by the atoms $\{A_i\}_{i=1}^r \subset \mathcal{A}$, which do not include any antipodal pairs of atoms. In view of Definition 2.1, this slice can be compactly represented as

$$\mathcal{S} := \mathrm{conv}(\{A_i\}_{i=1}^r \cup \{0\}). \tag{2.10}$$

After setting $Q := \sum_{j=1}^{r} A_j$*, note that*

$$\langle Q, A_i \rangle = \sum_{j=1}^{r} \langle A_j, A_i \rangle = \langle A_i, A_i \rangle = 1, \tag{2.11}$$

for every $i \in [r]$*. Moreover, we have that*

$$\langle Q, A \rangle = \sum_{j=1}^{r} \langle A_j, A \rangle < 1, \tag{2.12}$$

for any atom $A \in \mathcal{A}$ *not listed in* $\{A_i\}_{i=1}^r$*. From (2.11) and (2.12), we observe that the atoms* $\{A_i\}_{i=1}^r$ *form an* $(r-1)$*-dimensional exposed face* $\mathcal{F}$ *of* $\operatorname{conv}(\mathcal{A})$ *with a support vector* Q*. Combined with (2.10), it follows that* $\mathcal{S} = \operatorname{conv}(\mathcal{F} \cup \{0\})$*. Since any antipodal pair of atoms is redundant when representing a model, we conclude that every nontrivial* r*-dimensional slice* $\mathcal{S}$ *of* $\operatorname{conv}(\mathcal{A})$ *can be identified with an* $(r-1)$*-dimensional exposed face* $\mathcal{F}$ *of* $\operatorname{conv}(\mathcal{A})$ *in the sense that* $\mathcal{S} = \operatorname{conv}(\mathcal{F} \cup \{0\})$*. Proposition 2.7 is thus in force, predicting that the machine (gauge) succeeds for the choice of alphabet* $\mathcal{A}$ *in this example, when given access to sufficiently many observations. This finding agrees with the well-established wisdom about sparse regression and subset selection. For example, it is well-known that the machine (gauge) successfully learns the true model* $x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp)$ *in (exact) and its* r*-sparse decomposition in the slice* $\mathcal{S}^\sharp$*, provided that* $m = \widetilde{\Omega}(r)$ *and* $\mathcal{L}$ *is a generic linear operator [11]. Here and elsewhere,* $\widetilde{\Omega}$ *hides logarithmic factors for simplicity.*

Example 2.10 (Low-rankness)**.** *For integers* d_1 *and* d_2*, consider the alphabet*

$$\mathcal{A} := \{uv^\top : \|u\|_2 = \|v\|_2 = 1\} \subset \mathbb{R}^{d_1 \times d_2}, \tag{2.13}$$

which is central to matrix factorization, with applications in collaborative filtering, anomaly detection and beyond [37]. This example applies Proposition 2.7 to the above alphabet $\mathcal{A}$*, which can also be identified with an alphabet in* $\mathbb{R}^d$ *with* $d = d_1 d_2$*. For this alphabet, it is again not difficult to verify that the corresponding gauge function is the nuclear norm, i.e.,*

$$\mathcal{G}_{\mathcal{A}} = \|\cdot\|_*, \tag{2.14}$$

and that $\operatorname{conv}(\mathcal{A}) = \{x : \|x\|_* \leq 1\}$ *is the unit ball for the nuclear norm. Recall that the nuclear norm of a matrix is the sum of its singular values [38]. For a model* $x^\sharp \in \mathbb{R}^{d_1 \times d_2}$ *such that* $\operatorname{rank}(x^\sharp) \leq r$*, let*

$$x^\sharp \overset{\text{SVD}}{=} \sum_{i=1}^{r} c_i^\sharp u_i^\sharp (v_i^\sharp)^\top, \tag{2.15}$$

be its singular value decomposition (SVD), where $\{c_i^\sharp\}_{i=1}^r$ *are the leading* r *singular values of* $x^\sharp$*, some of which might be zero. Let us form the matrix* $U^\sharp \in \mathbb{R}^{d_1 \times r}$ *by concatenating the (left) singular vectors* $\{u_i^\sharp\}_{i=1}^r$*, and we also form* $V^\sharp \in \mathbb{R}^{d_2 \times r}$ *similarly. Consider the slice* $\mathcal{S}^\sharp$ *formed by the atoms* $\{U^\sharp \alpha\alpha^\top (V^\sharp)^\top\}_{\|\alpha\|_2 = 1} \subset \mathcal{A}$*. In view of Definition 2.1, we can represent this slice compactly as*

$$\mathcal{S}^\sharp := \operatorname{conv}\left(\{U^\sharp \alpha\alpha^\top (V^\sharp)^\top\}_{\|\alpha\|_2 = 1} \cup \{0\}\right). \tag{2.16}$$

In particular, note that $x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp)$, *see* (2.15). *After setting* $Q := U^\sharp (V^\sharp)^\top$, *we distinguish three cases. First, note that*

$$\langle Q, U^\sharp \alpha\alpha^\top (V^\sharp)^\top \rangle = \langle U^\sharp (V^\sharp)^\top, U^\sharp \alpha\alpha^\top (V^\sharp)^\top \rangle = \|\alpha\|_2^2 = 1, \tag{2.17}$$

for every unit-norm vector $\alpha \in \mathbb{R}^r$. *Second, note that*

$$\langle Q, U^\sharp \alpha\beta^\top (V^\sharp)^\top \rangle = \langle U^\sharp (V^\sharp)^\top, U^\sharp \alpha\beta^\top (V^\sharp)^\top \rangle = \langle \alpha, \beta \rangle < 1, \tag{2.18}$$

for every distinct pair of unit-norm vectors $\alpha, \beta \in \mathbb{R}^r$. *Third, for any pair of unit-norm vectors* u *and* v *where one or both vectors are not in the range of* $U^\sharp$ *and* $V^\sharp$, *respectively, at least one of the two inequalities* $\|u^\top U^\sharp\|_2 < 1$, $\|v^\top V^\sharp\|_2 < 1$ *holds. Consequently, for any such pair of vectors* u *and* v, *it holds that*

$$\langle Q, uv^\top \rangle \leq \|u^\top U^\sharp\|_2 \cdot \|v^\top V^\sharp\|_2 < 1, \tag{2.19}$$

where we used the Cauchy-Schwarz's inequality above. From (2.17)-(2.19), we observe that the atoms $\{U^\sharp \alpha\alpha^\top (V^\sharp)^\top\}_{\|\alpha\|_2=1}$ *form an exposed face* $\mathcal{F}^\sharp$ *of* $\mathrm{conv}(\mathcal{A})$ *with a support vector* Q. *In view of* (2.16), *it immediately follows that* $\mathcal{S}^\sharp = \mathrm{conv}(\mathcal{F}^\sharp \cup \{0\})$. *By counting the number of variables, we also see that* $\dim(\mathcal{F}^\sharp) \leq r^2$, *see* (2.16). *Proposition 2.7 with* r^2 *thus predicts that there exists a linear operator* $\mathcal{L}$ *such that the machine* (gauge) *returns* $x^\sharp$ *and its* r^2*-sparse decomposition in the slice* $\mathcal{S}^\sharp$. *But note that the* r*-sparse decomposition in* (2.15) *is the unique decomposition of* $x^\sharp$ *in the slice* $\mathcal{S}^\sharp$, *because* $U^\sharp$ *and* $V^\sharp$ *have orthonormal columns by construction. We can therefore strengthen the prediction of Proposition 2.7 to read that there exists a linear operator* $\mathcal{L}$ *such that the machine* (gauge) *returns* $x^\sharp$ *and its* r*-sparse decomposition in the slice* $\mathcal{S}^\sharp$. *Indeed, it is well-known that the machine* (gauge) *successfully learns the hidden model* $x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp)$ *and its* r*-sparse decomposition* (2.15) *in the slice* $\mathcal{S}^\sharp$, *provided that* $m = \widetilde{\Omega}(rd)$ *and* $\mathcal{L}$ *is a generic linear operator, see for example* [37].

Before presenting the next example, let us recall the definition of a Chebyshev system [39].

Definition 2.11 (Chebyshev system)**.** *Real-valued and continuous functions* $\{\phi_j\}_{j=1}^d$ *form a Chebyshev system on an interval* $I \subset \mathbb{R}$ *if the* $m \times m$ *matrix* $[\phi_j(\tau_l)]_{l,j=1}^m$ *is nonsingular for every increasing sequence* $\{\tau_l\}_{l=1}^m \subset I$.

Monomials on a closed interval of the real line form a Chebyshev system and, in fact, the notion of Chebyshev system generalizes and preserves the key properties of monomials, see [39, 40, 14] for more examples of Chebyshev systems and their applications in classical approximation theory and modern signal processing.

Example 2.12 (Chebyshev system)**.** *Suppose that* $\{\phi_j\}_{j=1}^d$ *form a Chebyshev system on the interval* $I \subset \mathbb{R}$ *and let us define the map* $\Phi : I \to \mathbb{R}^d$ *as*

$$\Phi(t) = [\phi_1(t), \cdots, \phi_d(t)]^\top. \tag{2.20}$$

Consider also the alphabet

$$\mathcal{A} := \{\Phi(t) : t \in I\} \subset \mathbb{R}^d, \tag{2.21}$$

which arises, for instance, in signal and image super-resolution [40, 14, 41]. *This example applies Proposition 2.7 to the above alphabet. Consider an integer* r *and* $\{t_i\}_{i=1}^r \subset I$. *(In the context of super-resolution,* $\{t_i\}_{i=1}^r$ *are the point sources.) Let* $\mathcal{S}^\sharp$ *be the slice of* $\mathrm{conv}(\mathcal{A})$ *formed by* $\{\Phi(t_i)\}_{i=1}^r \subset \mathcal{A}$, *i.e.,*

$$\mathcal{S}^\sharp := \mathrm{conv}\left(\{\Phi(t_i)\}_{i=1}^r \cup \{0\}\right). \tag{2.22}$$

A classical result in approximation theory guarantees that there exists a vector $Q \in \mathbb{R}^d$ *such that*

$$\langle Q, \Phi(t_i)\rangle = 0, \quad \forall i \in [r], \qquad \langle Q, \Phi(t)\rangle > 0, \quad \forall t \in I - \{t_i\}_{i=1}^r, \tag{2.23}$$

provided that $d \geq 2r+1$, *see for example* [39, Theorem 5.1]. *We might therefore interpret* $\langle Q, \Phi(t)\rangle$ *as a nonnegative polynomial of* t *on the interval* I *that vanishes precisely at* $\{t_i\}_{i=1}^r$. *In view of* (2.23), *we find that* $\mathcal{F}^\sharp = \mathrm{conv}(\{\Phi(t_i)\}_{i=1}^r)$ *is an* $(r-1)$*-dimensional face of* $\mathrm{conv}(\mathcal{A})$ *with a support vector* Q. *After revisiting* (2.22), *we observe that* $\mathcal{S}^\sharp = \mathrm{conv}(\mathcal{F}^\sharp \cup \{0\})$. *In words, similar to the previous two examples, Proposition 2.7 is in force. As before, this finding agrees with our understanding of Chebyshev systems. For the alphabet in this example, it is indeed known that the machine* (gauge) *can successfully learn the true model* $x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp)$ *in* (exact) *and its decomposition in the slice* $\mathcal{S}^\sharp$, *see* [40, 14].

2.3. **Failures**

Despite the success stories featured in Section 2.2, there are many linear inverse problems for which the gauge function theory fails, in line with Observation 2.8. To highlight the statistical failures of the machine (gauge), we focus in this section on structured data factorization, i.e., the special case of the linear inverse problem in (exact) and (gauge), where $\mathcal{L}$ is the identity operator. More specifically, given the r-sparse model $x^\sharp$ in (exact), we search for a minimal decomposition of $x^\sharp$ that achieves $\mathcal{G}_{\mathcal{A}}(x^\sharp)$ by solving the optimization problem

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) = \inf\left\{\sum_{i=1}^l c_i : x^\sharp = \sum_{i=1}^l c_i A_i,\ c_i \geq 0,\ A_i \in \mathcal{A},\ \forall i \in [l]\right\}, \qquad (\text{see } (2.2)) \tag{2.24}$$

where the infimum above is taken over the integer l, coefficients$\{c_i\}_i$ and the atoms $\{A_i\}_i$. Let us list two examples of structured data factorization:

(1) **(Sparse PCA)** The objective here is to decompose a data matrix into a small number of rank-1 and sparse components. More specifically, suppose that the rows and columns of the data matrix $x^\sharp \in \mathbb{R}^{d_1 \times d_2}$ correspond to the samples and features, respectively. In general, the leading principal components of $x^\sharp$ are not sparse, which renders them difficult to interpret. For example, it is often difficult to single out the key features in a data matrix from its leading principal components. In contrast, for an integer r, sparse PCA in effect models the data matrix as

$$\begin{aligned} &x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp), \qquad \mathcal{S}^\sharp \in \mathrm{slice}_r(\mathcal{A}), \\ &\mathcal{A} := \{uv^\top : \|u\|_2 = \|v\|_2 = 1,\ \|v\|_0 \leq k\} \subset \mathbb{R}^{d_1 \times d_2}, \end{aligned} \tag{2.25}$$

where $\|v\|_0$ denotes the number of nonzero entries of v, inspired by [12] and [18, Equation 3.12]. The above interpretation of sparse PCA in (2.25), from the viewpoint of the gauge function theory,

has been documented before, see for example [13]. Note that k is the sparsity level of vector v, i.e., the (maximum) number of its nonzero entries, not to be confused with the sparsity level r of the model $x^\sharp$, as a combination of (at most) r atoms from the alphabet $\mathcal{A}$. Note also that $\mathcal{A}$ may be identified with an alphabet in $\mathbb{R}^d$ with $d = d_1 d_2$. One may also revise the definition of $\mathcal{A}$ by including $\|v\|_\infty = O(1/\sqrt{k})$ to ensure that the atoms that make up $x^\sharp$ are diffuse on their support. Importantly, note that the model (2.25) is exact, i.e., without any noise. This restriction is for the sake of simplicity of the toy examples presented in this section. Sparse PCA with an inexact model will be studied in Section 4.

(2) **(Group sparsity)** Here, the objective is to decompose the model $x^\sharp$ into a small number of vectors with known supports [42, 43]. To be concrete, for a factor $C > 0$ and a collection of index sets $\Omega \subset 2^{[d]}$, the model considered in group sparsity [2] is

$$\begin{aligned} &x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp), \qquad \mathcal{S}^\sharp \in \mathrm{slice}_r(\mathcal{A}), \\ &\mathcal{A} := \{u : \|u\|_2 = 1,\ \|u\|_\infty \leq C,\ \mathrm{supp}(u) \in \Omega\} \subset \mathbb{R}^d, \end{aligned} \tag{2.26}$$

where the above bound on ℓ_∞-norm ensures that the atoms are diffuse on their supports. Recall that the set $\mathrm{supp}(u) \subset [d]$ denotes the support of u, i.e., the index set over which u is nonzero.

Above, we listed two examples of structured data factorization. In order to better motivate the failure of the gauge function theory in this context, let us first study a special case of structured data factorization. More concretely, with $r = 1$, suppose that the 1-sparse model in (exact) is specified as

$$x^\sharp := c^\sharp A^\sharp, \qquad c^\sharp > 0,\ A^\sharp \in \mathcal{A}. \tag{2.27}$$

For the 1-sparse model (2.27), the machine (2.24) is closely related to another learning machine for structured data factorization, introduced in the following result. In words, the result below connects the so-called analysis and synthesis approaches to structured data factorization. These two approaches are common, for example, in sparse PCA and dictionary learning [12, 44], and more broadly in signal processing [45].

Proposition 2.13 (Synthesis vs. analysis)**.** *Suppose that Assumption 2.3(iv) is met. For the model $x^\sharp$ in (2.27), it then holds that*

$$\sup_{A \in \mathcal{A}} \langle x^\sharp, A \rangle = \mathcal{D}_{\mathcal{A}}(x^\sharp) = \mathcal{G}_{\mathcal{A}}(x^\sharp), \tag{2.28}$$

and the machines (2.24) and (2.28) both return $A^\sharp$.

In view of Proposition 2.13, one may solve problem (2.24), or equivalently problem (2.28), in order to recover the atom $A^\sharp$ in (2.27). Indeed, problem (2.28) is the optimization problem widely studied in the sparse PCA literature, for instance as the starting point of the convex relaxation in [44, 46].

A key drawback of the machine (2.28) is that it does not naturally lend itself to generalization beyond the 1-sparse model (2.27). In the context of sparse PCA, a common alternative is *deflation* [47], which involves the potentially unstable process of solving a sequence of optimization problems similar to (2.28), through which the numerical errors might accumulate or amplify. The machine (2.24)

addresses this drawback as it naturally generalizes (2.28) beyond the 1-sparse model (2.27). Unfortunately, however, for an r-sparse model (exact) with $r \geq 2$, the machine (2.24) in general fails to find an r-sparse decomposition of $x^\sharp$, as predicted by Observation 2.8. This failure of the machine (2.24) is demonstrated below with a few toy examples.

Example 2.14 (Manifold models)**.** *Our first example in this section demonstrates the failure of the gauge function theory for manifold manifolds. In a multitude of problems, the alphabet $\mathcal{A}$ is naturally an embedded submanifold of the Euclidean space* [22, 48]*. As a toy example, here we consider the alphabet*

$$\mathcal{A} := \{(t\cos(\pi t), t\sin(\pi t)) : t \in [0,2]\} \subset \mathbb{R}^2, \tag{2.29}$$

which forms an incomplete spiral in $\mathbb{R}^2$, i.e., a one-dimensional manifold with boundary visualized in Figure 1. For the above alphabet, it is important to note that the inclusion reviewed in (2.5) is strict, i.e.,

$$\operatorname{ext}(\operatorname{conv}(\mathcal{A})) \subset \mathcal{A}. \tag{2.30}$$

In particular, the 1-sparse model

$$x^\sharp := A^\sharp = \frac{1}{4\sqrt{2}}(1,1) \in \mathcal{A}, \tag{2.31}$$

which is represented with the red dot in Figure 1, is not an extreme point of $\operatorname{conv}(\mathcal{A})$*, i.e., $x^\sharp$ above belongs to the right-hand side but* not *to the left-hand side of (2.30). Recalling Definition 2.1, we can write that $x^\sharp \in \mathcal{S}^\sharp$, where the slice $\mathcal{S}^\sharp$ is simply the line segment connecting $A^\sharp$ to the origin. A visual inspection of Figure 1 immediately reveals that $\mathcal{S}^\sharp$ does* not *contain an exposed face of* $\operatorname{conv}(\mathcal{A})$ *and, as Observation 2.8 suggests, the machine (2.24) fails to find the (trivial) 1-sparse decomposition $x^\sharp = A^\sharp$. To verify this claim, note that the model $x^\sharp$ in (2.31) has the alternative decomposition*

$$x^\sharp = \frac{A_1}{8\sqrt{2}} + \frac{A_2}{2\sqrt{2}}, \qquad A_1 = \frac{1}{2}(0,1) \in \mathcal{A}, \qquad A_2 = 2(1,0) \in \mathcal{A}. \tag{2.32}$$

By comparing the two alternative representations of $x^\sharp$ above, we find that

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) \leq \min\left(1, \frac{1}{8\sqrt{2}} + \frac{1}{2\sqrt{2}}\right) = \frac{5}{8\sqrt{2}} < 1. \qquad \textit{(see (2.2))} \tag{2.33}$$

In fact, a visual inspection of Figure 1 reveals that the minimal decomposition of $x^\sharp$ that achieves $\mathcal{G}_{\mathcal{A}}(x^\sharp)$ in (2.24) is not *1-sparse. That is, the machine (2.24) fails to learn any 1-sparse decomposition for the model $x^\sharp$.*

Example 2.15 (Sparse PCA)**.** *To see another failure of the gauge function theory, let us revisit sparse PCA, introduced earlier in this section, see (2.25). To be concrete, for $d_1 = d_2 = 3$ and $k = 2$, consider the 2-sparse model*

$$x^\sharp := \frac{A_1^\sharp}{2} + \frac{A_2^\sharp}{2} \in \mathbb{R}^{3\times 3}, \tag{2.34}$$

where the atoms $A_1^\sharp, A_2^\sharp \in \mathcal{A}$ *are specified as*

$$A_1^\sharp := u_1^\sharp (v_1^\sharp)^\top = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^\top \cdot \begin{bmatrix} 0.3122 & 0.95 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3},$$
$$A_2^\sharp := u_2^\sharp (v_2^\sharp)^\top = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 0.95 & 0.3122 \end{bmatrix} \in \mathbb{R}^{3\times 3}. \tag{2.35}$$

Recalling Definition 2.1, let $\mathcal{S}^\sharp$ *be the two-dimensional slice formed by* $\{A_1^\sharp, A_2^\sharp\}$*, so that* $x^\sharp \in \mathcal{S}^\sharp$*. On the other hand, note that* $x^\sharp$ *above also has the alternative decomposition*

$$\begin{aligned} x^\sharp &= c_1 u_1 v_1^\top + c_2 u_2 v_2^\top + c_3 u_3 v_3^\top \qquad (u_1 v_1^\top, u_2 v_2^\top, u_3 v_3^\top \in \mathcal{A}) \\ &= 0.1561 \cdot \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^\top \cdot \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + 0.6717 \cdot \begin{bmatrix} 0.9082 & -0.0918 & 0.4082 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \\ &\quad + 0.1561 \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}. \end{aligned} \tag{2.36}$$

By comparing the two alternative representations of $x^\sharp$ *in* (2.34) *and* (2.36)*, we observe that*

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) \le \min\left(\frac{1}{2} + \frac{1}{2}, 0.1561 + 0.6717 + 0.1561\right) = 0.984 < 1, \qquad \textit{(see (2.2))} \tag{2.37}$$

and thus the 2-sparse decomposition of $x^\sharp$ *in* (2.34) *is not minimal. In fact, we may verify that the machine* (2.24) *fails to find any 2-sparse decomposition for the model* $x^\sharp$*. The failure of the gauge function theory for this toy example in sparse PCA agrees with Observation 2.8. Indeed, from (2.3) and (2.37), we find that* $x^\sharp$ *belongs to the interior of* $\mathrm{conv}(\mathcal{A})$*. Consequently, any face of* $\mathrm{conv}(\mathcal{A})$ *that passes through* $\{A_1^\sharp, A_2^\sharp\}$ *is a hidden face of* $\mathrm{conv}(\mathcal{A})$*, thanks to Definition 2.5 and the fact that* $\mathrm{conv}(\mathcal{A})$ *is convex body (a convex set with nonempty interior). In particular, the slice* $\mathcal{S}^\sharp$ *does not contain an exposed face of* $\mathrm{conv}(\mathcal{A})$*, in agreement with Observation 2.8. We also remark that the failure of machine* (2.24) *in Example 2.15 will persist even after imposing an incoherence requirement on the alphabet* $\mathcal{A}$*, i.e., after including* $\|v\|_\infty = O(1/\sqrt{k})$ *in* (2.25)*. Indeed, the true atoms* $\{A_1^\sharp, A_2^\sharp\}$ *in* (2.35) *are already sufficiently diffuse on their support. To close this example, we note that the potential failure of the gauge function theory, in the context of sparse PCA, has also been documented in* [12].

Example 2.16 (Group sparsity)**.** *For the last failed application of the gauge function theory in this section, let us revisit group sparsity, introduced earlier in this section. As an example of the model* (2.26) *with* $d = 3$*, consider the collection of index sets*

$$\Omega := \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}\} \subset 2^{[3]}, \tag{2.38}$$

and the alphabet

$$\mathcal{A} := \left\{ u : \|u\|_2 = 1,\ \|u\|_\infty \le \|u\|_0^{-\frac{1}{3}},\ \mathrm{supp}(u) \in \Omega \right\}, \tag{2.39}$$

where the bound on ℓ_∞*-norm above ensures that the atoms are diffuse on their support. With this alphabet, consider the model*

$$x^\sharp := \frac{A_1^\sharp}{2} + \frac{A_2^\sharp}{2} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} - \frac{\sqrt{7}}{8} & \frac{3}{8} \end{bmatrix}^\top, \tag{2.40}$$

where the atoms $\{A_1^\sharp, A_2^\sharp\} \subset \mathcal{A}$ *are specified as*

$$A_1^\sharp := \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top, \qquad A_2^\sharp := \begin{bmatrix} 0 & -\frac{\sqrt{7}}{4} & \frac{3}{4} \end{bmatrix}^\top. \tag{2.41}$$

Evidently, the model $x^\sharp$ *in* (2.40) *has the alternative decomposition*

$$x^\sharp = \frac{A_1}{2\sqrt{2}} + \left(\frac{1}{2\sqrt{2}} - \frac{\sqrt{7}}{8} \right) A_2 + \frac{3}{8} A_3, \tag{2.42}$$

where $\{A_i\}_{i=1}^3 \subset \mathcal{A}$ *are the three canonical vectors in* $\mathbb{R}^3$*. By comparing the two alternative representations of* $x^\sharp$ *in (2.40) and (2.42), we find that*

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) \leq \min \left(\frac{1}{2} + \frac{1}{2}, \frac{1}{2\sqrt{2}} + \left| \frac{1}{2\sqrt{2}} - \frac{\sqrt{7}}{8} \right| + \frac{3}{8} \right) = 0.7514 < 1, \qquad \textit{(see (2.2))} \tag{2.43}$$

and thus the 2*-sparse decomposition in* (2.40) *is not minimal. In fact, the machine* (2.24) *fails to find any* 2*-sparse decomposition for* $x^\sharp$*, as verified in Section 3.2. The failure of the gauge function theory in this toy example agrees with Observation 2.8. Indeed, similar to Example 2.15, we can verify here that the slice* $\mathcal{S}^\sharp$*, formed by* $\{A_1^\sharp, A_2^\sharp\}$ *in* (2.41)*, does not contain an exposed face of* $\operatorname{conv}(\mathcal{A})$*.*

While all learning problems discussed in this section fall under the broad umbrella of structured data factorization, it is not difficult to find other problems for which the classical gauge function theory fails. In the interest of space, with minimal details, Figure 3 shows another example in the context of super-resolution [49, 50]. The (blue) curve is the superposition of two Gaussian waves; the red bars show the centers and amplitudes of the two waves. (The red bars are scaled to fit.) The values of the blue curve are then observed at 40 random locations on the interval $[0, 1]$ and stored in a vector y. The learning alphabet here is comprised of Gaussian waves centered on the grid. The centers and amplitudes of the waves, as estimated by the convex machine (gauge), are shown by black bars, see [51] for the details. (The black bars are also scaled to fit.) The resounding failure of the convex machine here in learning the location of the red bars is an example of super-resolution below the diffraction limit [40]. We will revisit this super-resolution example in Section 5.

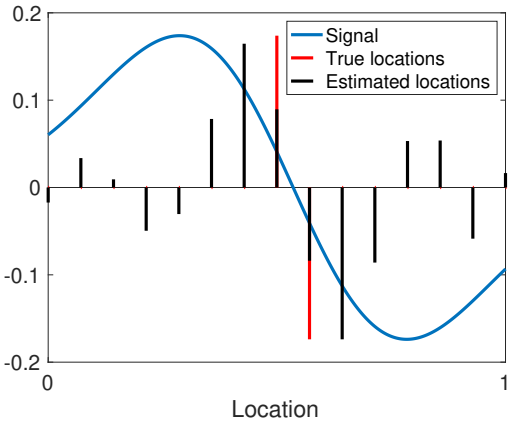


FIGURE 3. Failure of the (convex) gauge function theory in the context of super-resolution, detailed in the last paragraph of Section 2.

3. Gauge$_p$ Function Theory and Main Results

In Section 2, we reviewed the gauge function theory and also identified its limitations. In particular, Observation 2.8 and the negative toy examples in Section 2.3 highlighted the statistical limitations of the gauge function theory, which we aim to overcome in this section by introducing a generalized theory, dubbed gauge$_p$ function theory, as the main contribution of this work.

Below, first in Section 3.1, we will first introduce the gauge$_p$ function, which is the central object of the new theory, and then compare it with the classical gauge function in Section 3.2. The gauge$_p$ function is a simple generalization of the gauge function that can tightly control the sparsity of a model within the learning alphabet. The gauge$_p$ function also draws inspiration from the Burer-Monteiro factorization, which is a successful idea with computational (rather than statistical) origins in semi-definite programming [16]. We finally introduce the new learning machine at the heart of the gauge$_p$ function theory in Section 3.3, and describe its statistical guarantees in Section 3.4. This new learning machine uses, as the regularizer, the gauge$_p$ function instead of the classical gauge function. Aside from the statistical benefits, the computational strengths of the proposed machine are discussed later in Section 5.

3.1. Gauge$_p$ Function

Central to our generalized theory is a new notion of complexity for statistical models, dubbed gauge$_p$ function, which generalizes the gauge function in Definition 2.2, and can tightly control the sparsity level of a model within the learning alphabet. Before defining the gauge$_p$ function, we begin below with a few geometric concepts. To be specific, let us first introduce a geometric object which, as we will see shortly, generalizes the notion of the convex hull of a set.

Definition 3.1 ($\mathrm{conv}_p(\mathcal{A})$)**.** *For an alphabet $\mathcal{A}$ and integer p, we define*

$$\mathrm{conv}_p(\mathcal{A}) := \bigcup_{\mathcal{S}\in \mathrm{slice}_p(\mathcal{A})} \mathcal{S}, \tag{3.1}$$

to be the union of all slices of $\mathrm{conv}(\mathcal{A})$ *formed by at most p atoms. The notation* $\mathrm{slice}_p(\mathcal{A})$ *above was introduced in Definition 2.1.*

For example, for the alphabet $\mathcal{A} := \{\pm e_i\}_{i=1}^3$ in Figure 2a, $\mathrm{conv}_2(\mathcal{A})$ is the union of all colored triangles, some of which are hidden from the view. We next record a simple but key property of $\mathrm{conv}_p(\mathcal{A})$.

Proposition 3.2 (Nested hulls)**.** *Suppose that Assumption 2.3(i) or (ii) is met. Then it holds that*

$$\bigcup_{0\le\tau\le 1} \tau\mathcal{A} = \mathrm{conv}_1(\mathcal{A}) \subset \mathrm{conv}_2(\mathcal{A}) \subset \cdots \subset \mathrm{conv}_{d+1}(\mathcal{A}) = \mathrm{conv}_{d+2}(\mathcal{A}) = \cdots = \mathrm{conv}(\mathcal{A}), \tag{3.2}$$

where $\tau\mathcal{A} = \{\tau A : A \in \mathcal{A}\}$.

In words, the sets $\{\mathrm{conv}_p(\mathcal{A})\}_p$ provide a nested sequence of approximations to the alphabet $\mathcal{A}$. That is, as p decreases, $\mathrm{conv}_p(\mathcal{A})$ becomes an increasingly finer approximation to the alphabet $\mathcal{A} \subset \mathbb{R}^d$. Note also that the sets $\{\mathrm{conv}_p(\mathcal{A})\}_{p=1}^d$ in (3.2) might be nonconvex. In contrast, the sets $\{\mathrm{conv}_p(\mathcal{A})\}_{p\ge d+1}$ are convex and, in fact, coincide with $\mathrm{conv}(\mathcal{A})$. Indeed, the identities in (3.2) follow from an application of the Carathéodory theorem [26, Theorem 2.3]. To each set $\mathrm{conv}_p(\mathcal{A})$ above, we associate a gauge$_p$ function in analogy with (2.2), as formalized below.

Definition 3.3 (Gauge$_p$ function)**.** *For an alphabet $\mathcal{A} \subset \mathbb{R}^d$ and integer p, the corresponding gauge$_p$ function $\mathcal{G}_{\mathcal{A},p} : \mathbb{R}^d \to \mathbb{R}$ is defined as*

$$\mathcal{G}_{\mathcal{A},p}(x) := \inf \{t : x/t \in \mathrm{conv}_p(\mathcal{A}),\, t \geq 0\}. \tag{3.3}$$

Gauge$_p$ function generalizes the notion of gauge function in Definition 2.2 in the sense that $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ for every $p \geq d+1$. In contrast, when $p \leq d$, the gauge$_p$ function tightly controls the sparsity level of a model in the sense that $\mathcal{G}_p(x)$ is finite only when x has a p-sparse decomposition in the alphabet $\mathcal{A}$. This and other basic properties of the gauge$_p$ functions are collected below, and it is straightforward to prove them using Proposition 3.2 and Definition 3.3.

Proposition 3.4 (Properties of gauge$_p$ functions)**.** *Consider the gauge$_p$ function in Definition 3.3 for an integer p and an alphabet $\mathcal{A} \subset \mathbb{R}^d$. Suppose that Assumptions 2.3 (i) and (iii) are met. Then the following statements are true:*

(i) The gauge$_p$ function has the equivalent definition

$$\mathcal{G}_{\mathcal{A},p}(x) := \inf \left\{ \sum_{i=1}^{p} c_i \,:\, x = \sum_{i=1}^{p} c_i A_i,\, c_i \geq 0,\, A_i \in \mathcal{A},\, \forall i \in [p] \right\}. \tag{3.4}$$

(ii) $\mathcal{G}_{\mathcal{A},p}(x) = 0$ if and only if $x = 0$.

(iii) The convex conjugate of $\mathcal{G}_{\mathcal{A},p}$, denoted here by $\mathcal{G}^*_{\mathcal{A},p}$, is the convex indicator function for the unit ball of the norm $\mathcal{D}_{\mathcal{A}}$ in (2.4). That is,

$$\mathcal{G}^*_{\mathcal{A},p}(z) := \begin{cases} 0 & \mathcal{D}_{\mathcal{A}}(z) \leq 1 \\ \infty & \text{otherwise.} \end{cases} \tag{3.5}$$

(iv) The convex envelope of $\mathcal{G}_{\mathcal{A},p}$ is the gauge function $\mathcal{G}_{\mathcal{A}}$ in (2.2), i.e., $\mathcal{G}^{**}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$.

(v) If x does not admit a p-sparse decomposition in the alphabet $\mathcal{A}$, then $\mathcal{G}_{\mathcal{A},p}(x) = \infty$.

(vi) If $\mathcal{G}_{\mathcal{A},p}(x) < \infty$, then any minimal decomposition of x in the alphabet $\mathcal{A}$ that achieves $\mathcal{G}_{\mathcal{A},p}(x)$ in (3.4) is p-sparse.

(vii) The gauge$_p$ functions have the nested property

$$\mathcal{G}_{\mathcal{A},1}(x) \geq \mathcal{G}_{\mathcal{A},2}(x) \geq \cdots \geq \mathcal{G}_{\mathcal{A},d+1}(x) = \mathcal{G}_{\mathcal{A},d+2}(x) = \cdots = \mathcal{G}_{\mathcal{A}}(x), \tag{3.6}$$

where $\{\mathcal{G}_{\mathcal{A},p}\}_{p=1}^{d}$ above may be nonconvex functions.

Let us take a moment to parse Proposition 3.4. In view of Proposition 3.4(vii), the gauge$_p$ function coincides with the gauge function for $p \geq d+1$. Moreover, when $p \leq d$, the gauge$_p$ function still preserves certain properties of the gauge function, see Proposition 3.4(ii)-(iv).

Crucially, for $p \leq d$, the gauge$_p$ function directly controls the sparsity level, as articulated in Proposition 3.4(v) and (vi), and thus addresses the key shortcoming of the classical gauge function, i.e., the lack of sparsity. Indeed, in the negative examples in Section 2.3, the gauge function failed to enforce sparsity (its minimal decomposition failed to be sparse at all), whereas any minimal decomposition of x that achieves $\mathcal{G}_{\mathcal{A},p}(x)$ is p-sparse by Proposition 3.4(vi). Put differently, in the negative examples highlighted of Section 2.3, some of the inequalities in (3.6) were strict. We will

further investigate these key differences in Section 3.2 by comparing the gauge and gauge$_p$ functions for several learning alphabets.

It is worth pointing out a natural interpretation of the chain of inequalities in (3.6): In view of [52, Definition 2.2] or [53], $\mathcal{G}_{\mathcal{A},1}$ is the "most nonconvex" among the gauge$_p$ functions, followed by $\mathcal{G}_{\mathcal{A},2}$ and so on. Here the nonconvexity of the gauge$_p$ functions is measured by the distance from their shared convex envelope $\mathcal{G}_{\mathcal{A}}$. Another interesting perspective is offered by the the approximate Carathéodory theorem of Maurey [52, 54] which can, in principle, quantify just how well the sets $\text{conv}_p(\mathcal{A})$ approximate $\text{conv}(\mathcal{A})$. We will not further investigate this connection here.

3.2. Why Gauge$_p$ Function?

We earlier reviewed the gauge function in Definition 2.2, as the notion of statistical complexity at the heart of the classical gauge function theory. We then introduced the gauge$_p$ function in Definition 3.3 as a new device for measuring the complexity of statistical models. To develop a better understanding of these concepts, we next compare the gauge and gauge$_p$ functions for several learning alphabets. Our discussion in this section is limited to the statistical benefits of the gauge$_p$ function and we defer the computational strengths of gauge$_p$ functions to Section 5.

The synopsis of this section is as follows: In the first two examples below, which correspond to the success stories of the gauge function theory in Section 2.2, we will discover that gauge and gauge$_p$ functions are closely related. In contrast, in the third and fourth examples below, the difference between the gauge and gauge$_p$ functions helps explain the statistical failures of the gauge function theory in Section 2.3.

Example 3.5 (Sparsity, continued)**.** *Given the alphabet of Example 2.9, a simple calculation shows*

$$\mathcal{G}_{\mathcal{A},p}(x) := \begin{cases} \mathcal{G}_{\mathcal{A}}(x) = \|x\|_1 & \|x\|_0 \leq p \\ \infty & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\|x\|_0$ is the number of nonzero entries of x, and we recall from (2.9) that $\mathcal{G}_{\mathcal{A}} = \|\cdot\|_1$. In particular, it follows from (3.7) that $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ when $p \geq d$. On the other hand, if $\|x\|_0 > p$, then note that $\mathcal{G}_{\mathcal{A},p}(x) = \infty$. This property of the gauge$_p$ function will later enable us to limit the learning machine to sufficiently sparse models. We also note that, in general, the relation between gauge and gauge$_p$ functions is more complex than the one in (3.7), as we will see in other examples in this section. It is worth noting that, if one replaces the ℓ_1-norm in (3.7) with the ℓ_2-norm, then the convex envelope of the resulting function would coincide with the k-support norm [55], an alternative to elastic net regularization [56]. We close by noting that, for this choice of the alphabet, the relation $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ for $p \geq d$ is an improvement over the conservative but more general result in (3.6).

Example 3.6 (Low-rankness, continued)**.** *Given the alphabet of Example 2.10, it is not difficult to verify that*

$$\mathcal{G}_{\mathcal{A},p}(x) := \begin{cases} \mathcal{G}_{\mathcal{A}}(x) = \|x\|_* & \textit{rank}(x) \leq p \\ \infty & \textit{otherwise,} \end{cases}$$

where we recall from (2.14) *that* $\mathcal{G}_{\mathcal{A}} = \|\cdot\|_*$ *is the nuclear norm. In particular,* $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ *for* $p \geq \min(d_1, d_2)$*. Similar to the previous example, it is worth noting that, if one replaces the nuclear norm in* (3.7) *with the Frobenius norm, then the convex envelope of the resulting function would coincide with the so-called spectral* k*-support norm* [57].

In both Examples 2.9 and 2.10, which correspond to the success stories of the gauge function theory in Section 2.2, observe that the gauge$_p$ function of a p-sparse model coincides with the corresponding gauge function. However, in general, recall from (3.6) that $\mathcal{G}_{\mathcal{A},p}(x) \geq \mathcal{G}_{\mathcal{A}}(x)$ for an alphabet $\mathcal{A}$ and model x. As we will see below, this inequality might be strict, which precisely explains the failures of the classical gauge function theory in the toy examples presented in Section 2.3. To see this, let us continue with the group sparsity example below.

Example 3.7 (Group sparsity, continued)**.** *Given the alphabet of group sparsity in* (2.26) *in Example 2.16, the gauge and gauge*$_p$ *functions have key differences which we highlight in this example. For this alphabet, note that the gauge function becomes*

$$\mathcal{G}_{\mathcal{A}}(x) = \inf\left\{\sum_{i=1}^{l} \|u_i\|_2 : x = \sum_{i=1}^{l} u_i,\ \|u_i\|_\infty \leq C,\ \mathrm{supp}(u_i) \in \Omega,\ \forall i \in [l]\right\}, \tag{3.8}$$

where $\Omega \subset 2^{[d]}$ *is a collection of index sets in* $[d]$ *(cf.* (2.2)*). If* Ω *only contains disjoint subsets of* $[d]$*, then* (3.8) *reduces to*

$$\mathcal{G}_{\mathcal{A}}(x) := \sum_{I \in \Omega} \|x_I\|_2, \tag{3.9}$$

where x_I *is the restriction of* x *to the index set* $I \subset [d]$*. Likewise, the corresponding gauge*$_p$ *function for the group sparsity alphabet in* (2.26) *is*

$$\mathcal{G}_{\mathcal{A},p}(x) := \inf\left\{\sum_{i=1}^{p} \|u_i\|_2 : x = \sum_{i=1}^{p} u_i,\ \|u_i\|_\infty \leq C,\ supp(u_i) \in \Omega,\ \forall i \in [p]\right\}. \quad \textit{(see (3.4))} \tag{3.10}$$

If Ω *only contains disjoint subsets of* $[d]$*, then* (3.10) *simplifies to*

$$\mathcal{G}_{\mathcal{A},p}(x) := \begin{cases} \mathcal{G}_{\mathcal{A}}(x) & \text{if } x \text{ is supported on at most } p \text{ index sets in } \Omega \\ \infty & \text{otherwise.} \end{cases} \tag{3.11}$$

In particular, when Ω *only contains disjoint subsets of* $[d]$*, the gauge*$_p$ *function of a* p*-sparse model coincides with its gauge function, similar to Examples 2.9 and 2.10 above.*

However, if Ω *contains overlapping subsets of* $[d]$*, then* (3.11) *does not hold in general. For instance, for the* 2*-sparse model* $x^\sharp$ *in* (2.40)*, for which the classical gauge function theory failed in Section 2.3, it is not difficult to verify that*

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) \leq 0.7514 < 0.9179 \leq \mathcal{G}_{\mathcal{A},2}(x^\sharp). \tag{3.12}$$

It is important to note the strict inequality above in contrast to the equality in (3.11)*. As a byproduct,* (3.12) *also posits that the machine* (2.24) *fails to find any* 2*-sparse decomposition of* $x^\sharp$ *within the alphabet, thus explaining the failure of the classical gauge function theory for group sparsity in*

Section 2.3. To see why (3.12) holds, observe that the inequality on the far-left of (3.12) was indeed established in (2.43). To verify the far-right inequality in (3.12), for nonnegative coefficients c_1, c_2, suppose that $x^\sharp = c_1 A_1 + c_2 A_2$, where the atoms $A_1, A_2 \in \mathcal{A}$ are supported on the index sets $\{1,2\} \in \Omega$ and $\{2,3\} \in \Omega$, respectively. In particular, it follows from (2.40) that

$$\frac{1}{2\sqrt{2}} = c_1 A_1(1), \qquad \frac{3}{8} = c_2 A_2(3), \tag{3.13}$$

where $A_1(1)$ is the first coordinate of the vector A_1 and $A_2(3)$ is defined similarly. Recalling from (2.39) that the atoms satisfy $\|A_1\|_\infty \leq 2^{-1/3}$ and $\|A_2\|_\infty \leq 2^{-1/3}$, it follows from (3.13) that

$$c_1 + c_2 \geq \frac{2^{\frac{1}{3}}}{2\sqrt{2}} + \frac{2^{\frac{1}{3}}3}{8} = 0.9179, \tag{3.14}$$

which immediately establishes the far-right inequality in (3.12), after recalling Definition 3.3.

We now present another example to highlight the differences between gauge and gauge$_p$ functions.

Example 3.8 (Manifold models, continued)**.** *When the alphabet $\mathcal{A}$ is an embedded submanifold, we saw in Section 2.3 that the gauge function $\mathcal{G}_\mathcal{A}$ often loses vital geometric details about the manifold $\mathcal{A}$. In contrast, $\mathcal{G}_{\mathcal{A},1}$ captures far more information about the manifold $\mathcal{A}$. More specifically, note that $\mathcal{G}_{\mathcal{A},1}(x^\sharp) = \infty$ unless the model $x^\sharp$ has a 1-sparse decomposition in the alphabet $\mathcal{A}$, i.e., $\mathcal{G}_{\mathcal{A},1}(x^\sharp) = \infty$ unless $x^\sharp/t \in \mathcal{A}$ for some $t \geq 0$, see Proposition 3.4(v). As an example, for the 1-sparse model $x^\sharp$ in (2.31), it is easy to verify that*

$$\mathcal{G}_\mathcal{A}(x^\sharp) \leq \frac{5}{8\sqrt{2}} < 1 = \mathcal{G}_{\mathcal{A},1}(x^\sharp), \tag{3.15}$$

and the strict inequality above means that the machine (gauge) fails to find any 1-sparse decomposition of $x^\sharp$, which in turn explains the failure of the classical gauge function theory in Section 2.3. Indeed, to see why (3.15) holds, note that the far-left inequality in (3.15) was established in (2.33) and the identity in (3.15) is evident from a visual inspection of Figure 1.

One can also verify that the strict inequality between gauge and gauge$_p$ functions also holds in Example 2.15 about sparse PCA, thus explaining the failure of the gauge function in that example.

Observation 3.9 (Gauge vs. gauge$_p$ functions)**.** *To summarize this section so far, the gauge$_p$ function in Definition 3.3 is a simple generalization of the gauge function that better controls the sparsity of a model within the learning alphabet, compared to the classical gauge function. In every failed example of the gauge function theory above, we observed that $\mathcal{G}_\mathcal{A}(x^\sharp) < \mathcal{G}_{\mathcal{A},r}(x^\sharp)$ and, consequently, any minimal decomposition of $x^\sharp$ in the alphabet $\mathcal{A}$ that achieves $\mathcal{G}_\mathcal{A}(x^\sharp)$ cannot be r-sparse. In contrast, any minimal decomposition of $x^\sharp$ that achieves $\mathcal{G}_{\mathcal{A},r}(x^\sharp)$ is r-sparse by definition, see Proposition 3.4(vi). Motivated by this observation, we next introduce a new learning machine that replaces the gauge function in (gauge) with a gauge$_p$ function.*

3.3. A New Learning Machine

The gauge function theory, reviewed in Section 2.1, is a theory for (convex) statistical learning that uses the gauge function to promote sparsity within the learning alphabet. However, in various linear inverse problems, the gauge function fails to enforce sparsity and, consequently, the gauge function theory fails, as suggested by Observation 2.8 and several examples in Section 2.3.

To overcome the above statistical limitations of the gauge function theory, this section introduces and motivates a new learning machine that hinges on the gauge$_p$ function in Definition 3.3 as its key building block. The new learning machine is a simple generalization of the convex machine (gauge) which allows us to tightly control the sparsity of the learning outcome. We will later study the statistical guarantees and the computational aspects of the new machine in Sections 3.4 and 5, respectively. To begin, consider the (inexact) model

$$y := \mathcal{L}(x^\sharp) + e, \qquad x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp),\ \mathcal{S}^\sharp \in \operatorname{slice}_r(\mathcal{A}), \qquad \|e\|_2 \leq \epsilon, \tag{inexact}$$

where ϵ controls the inexactness of the model. In signal processing, for example, ϵ quantifies the noise level in our measurements y. In statistical inference, e in the model (inexact) is also assigned a probability distribution [58, 29], but we avoid this additional layer of complexity here. In particular, when the distribution assigned to e is sufficiently light-tailed, we can always work with the model (inexact) for a sufficiently large ϵ.

To learn the model $x^\sharp$ in (inexact) or its sparse decomposition in the alphabet $\mathcal{A}$, we introduce the machine

$$\min_x \|\mathcal{L}(x) - y\|_2^2 \quad \text{subject to} \quad \mathcal{G}_{\mathcal{A},p}(x) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \tag{gauge$_p$}$$

which uses the gauge$_p$ function $\mathcal{G}_{\mathcal{A},p}$ in (3.3) and, as detailed below, generalizes the machine (gauge).

Remark 3.10 (A generalization of (gauge))**.** *The machine* (gauge$_p$) *reduces to the convex machine* (gauge) *for* $p \geq d+1$*, in view of Proposition 3.4(vii). On the other hand, when* $p \leq d$*, the machine* (gauge$_p$) *only searches over* p*-sparse models by virtue of Proposition 3.4(v), and thus always returns a* p*-sparse solution by design. This immediately rectifies the key failure of the gauge function theory in Observation 2.8 and Section 2.3, i.e., the lack of sparsity.*

The computational aspects of the new learning machine are discussed later in Section 5, where we provide a tractable numerical scheme for solving the optimization problem (gauge$_p$). A short remark follows below about the alternative formulations of the problem (gauge$_p$).

Example 3.11 (Examples 2.14, 2.15, 2.16, continued)**.** *In these negative toy examples, we saw in Section 2.3 that the convex machine* (gauge) *failed to find any* r*-sparse decomposition of the model* $x^\sharp$ *in the learning alphabet* $\mathcal{A}$*. In contrast, it is not difficult to verify that the new machine* (gauge$_p$) *with* $p = r$ *and* $\mathcal{L} = \mathrm{id}$ *successfully finds the (unique)* r*-sparse decomposition of* $x^\sharp$ *in each example. In general, the machine* (gauge$_p$) *always returns a* p*-sparse solution within the alphabet* $\mathcal{A}$*, see Proposition 3.4(vi). Here,* id *stands for the identity operator.*

In addition to generalizing the convex machine (gauge), the new machine (gauge$_p$) can also be interpreted as a natural extension of the Burer-Monteiro idea [16]:

Remark 3.12 (Burer-Monteiro factorization as inspiration)**.** *An important predecessor of the proposed machine* (gauge$_p$) *appears in the context of matrix factorization. To explain their connection, for simplicity, consider the optimization problem*

$$\min_x \ \|\mathcal{L}(x)-y\|_2^2 \ \ \textit{subject to} \ \operatorname{trace}(x)\le\gamma^2 \ \ \textit{and} \ x\in\mathbb{R}^{d_1\times d_1} \ \textit{is positive semi-definite}, \tag{3.16}$$

where $\mathcal{L}$ is a linear operator and $\gamma\ge 0$. Because x is positive semi-definite above, $\operatorname{trace}(x)$ *coincides with the nuclear norm of x and problem* (3.16) *is therefore a variant of the (convex) learning machines widely studied in matrix completion and sensing* [37]*. Note that the computational cost of solving problem* (3.16) *by directly manipulating x grows rapidly as the dimension d_1 grows. With this computational motivation, it is common to instead solve the Burer-Monteiro factorization of problem* (3.16)*. More specifically, for an integer $p\le d$, the factorized version of problem* (3.16) *is*

$$\min_{u\in\mathbb{R}^{d\times p}} \ \|\mathcal{L}(uu^\top)-y\|_2^2 \ \ \textit{subject to} \ \|u\|_{\mathrm{F}}\le\gamma, \tag{3.17}$$

where $\|\cdot\|_{\mathrm{F}}$ stands for the Frobenius norm and we used the fact that $\operatorname{trace}(uu^\top)=\|u\|_{\mathrm{F}}^2$*. When p is sufficiently small, problem* (3.17) *can offer substantial savings in computational speed and storage, compared to a direct implementation of the problem* (3.16)*. This idea has been successfully utilized in matrix-valued learning problems* [16, 19, 59, 60, 15, 38]*.*

It is not difficult to verify that the factorized problem (3.17) *coincides with problem* (gauge$_p$) *for the choice of alphabet $\mathcal{A}:=\{uu^\top:\|u\|_2=1\}$. In this sense, the proposed machine* (gauge$_p$) *extends the successful idea of Burer-Monteiro factorization to any learning alphabet.*

While the motivation behind the Burer-Monteiro factorization is purely computational [16]*, it is important to emphasize our statistical ambitions in this work: As we saw earlier in Examples 2.14, 2.15 and 2.16, the proposed machine* (gauge$_p$) *can also offer statistical improvements compared to the machine* (gauge)*. In the next section, we will develop a unified statistical theory for the new learning machine in order to quantify these improvements.*

We close this remark by adding that it is often possible to remove the constraints in problem (3.17)*, provided that $p=\widetilde{O}(\mathit{rank}(x^\sharp))$. Here, $x^\sharp$ is the hidden model that satisfies $\mathcal{L}(x^\sharp)\approx y$. In this regime, known as the* thin *Burer-Monteiro factorization, a generic operator $u\to\mathcal{L}(uu^\top)$ is often injective, thus obviating the need for regularization via $\|u\|_{\mathrm{F}}$, see for instance* [20, Section 2.1]*.*

It is natural to ask why we have opted for the new machine (gauge$_p$) instead of an alternative that more directly enforces sparsity. The next remark answers this question.

Remark 3.13 (Sparsity)**.** *This remark compares the new machine* (gauge$_p$) *with an alternative that more directly enforces sparsity within the learning alphabet. For simplicity, suppose that $\epsilon=0$ and $\mathcal{L}(x^\sharp)=y$ in* (inexact)*. As we saw in Remark 3.10, the new machine* (gauge$_p$) *finds a p-sparse model $\widehat{x}$ with no prediction error ($\mathcal{L}(\widehat{x})=y$) such that $\mathcal{G}_{\mathcal{A},p}(\widehat{x})\le\mathcal{G}_{\mathcal{A},p}(x^\sharp)$. (Recall that $\mathcal{G}_{\mathcal{A},p}$ might differ from $\mathcal{G}_{\mathcal{A}}$ even for p-sparse models, as we saw in Section 3.2.) It is natural to ask why we have opted here for the machine* (gauge$_p$)*, instead of the learning machine*

$$\min_x \ \|\mathcal{L}(x)-y\|_2^2 \ \ \textit{subject to} \ x \ \textit{has a p-sparse decomposition in the alphabet} \ \mathcal{A}, \tag{3.18}$$

which more directly penalizes the sparsity level. The machine (3.18) *appears, for instance, in the context of best subset selection* [61]*, for which the choice of alphabet* $\mathcal{A}$ *was specified Example 2.9. We have opted for the new machine* (gauge_p) *in this work in part because it naturally interpolates between two extremes: On the one hand, when* $p \geq d+1$*, the machine* (gauge_p) *reduces to the convex machine* (gauge)*, see Remark 3.10. On the other hand, when* $p = r$*, the machine* (gauge_p) *coincides with* (3.18) *also with* $p = r$*, provided that* $\mathcal{L}$ *is an injective map when restricted to* r*-sparse models. (This last assumption is reasonable because* $x^\sharp$ *would not be identifiable otherwise.) Recall from* (inexact) *that* r *denotes the sparsity level of the true model* $x^\sharp$*.*

As p *decreases from one extreme to the other, the machine* (gauge_p) *unlocks a range of new statistical and computational trade-offs. Informally speaking, as* p *decreases, the statistical accuracy of the new machine* (gauge_p) *improves and we will later quantify this improvement in Remark 3.15. The computational trade-offs are more complex and deferred to Section 5.*

Another advantage of the machine (gauge_p) *over the alternative* (3.18) *is that the former is regularized with the* gauge_p *function. Indeed, if* p *is relatively large, then the operator* $\mathcal{L}$ *might not be injective when restricted to* p*-sparse models and, consequently, the machine* (3.18) *would fail to identify the true model* $x^\sharp$*. One last reason to opt for the new machine* (gauge_p) *is that it readily extends the Burer-Monteiro idea to any learning alphabet, as detailed earlier in Remark 3.12.*

3.4. **Statistical Guarantees**

In Section 3.3, we introduced the machine (gauge_p), as a generalization of the convex machine (gauge) which, at the same time, draws inspiration from the Burer-Monteiro idea. This section will develop some statistical guarantees for the new learning machine, as parts of a generalized theory for the machine (gauge). The first result of this section, Lemma 3.14 below, provides certificates of correctness for the new machine (gauge_p). The second result of this section, Theorem 3.23 below, shows that these certificates exist for a generic operator $\mathcal{L}$ and a sufficiently small p. The proof of the second result is by construction, as outlined in Section 3.4.2, and appears to be new to be the best of our knowledge, which might be of independent interest. We begin now with the first result of this section which posits that the machine (gauge_p) succeeds if certain certificates of correctness exist, reminiscent of Lemma 2.6 in the context of the classical gauge function theory.

Lemma 3.14 (Correctness certificates, exact model)**.** *With* $\epsilon = 0$*, consider the model* $x^\sharp$ *in* (inexact) *and suppose that the alphabet* $\mathcal{A}$ *satisfies Assumptions 2.3(i) and (iii). If* $x^\sharp = 0$*, then the machine* (gauge_p) *correctly returns* 0*. Otherwise, set* $p \geq r$*, where* r *is the sparsity level of* $x^\sharp$ *in* (inexact)*. Then the machine* (gauge_p) *returns* $x^\sharp$ *and a* p*-sparse decomposition of* $x^\sharp$ *in* $\mathcal{A}$*, provided that for every slice* $\mathcal{S} \in \text{slice}_p(\mathcal{A})$*, (one of) the following holds:*

(i) If $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}$*, the linear map* $\mathcal{L}$ *restricted to the slice* $\mathcal{S}$ *is injective.*

(ii) If $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \notin \mathcal{S}$*, the point* $x^\sharp/\mathcal{G}_{p,\mathcal{A}}(x^\sharp)$ *and the slice* $\mathcal{S}$ *are separated along* $\text{range}(\mathcal{L}^*)$*, i.e., there exists a correctness certificate* $Q_\mathcal{S} \in \text{range}(\mathcal{L}^*)$ *such that*

$$\left\langle Q_\mathcal{S}, x - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right\rangle < 0, \qquad \forall x \in \mathcal{S}. \tag{3.19}$$

The proof of Lemma 3.14 largely mirrors that of Lemma 2.6 for the convex machine (gauge) (which justifies our perhaps unusual perspective in the review of the gauge function theory in Section 2.1). In Lemma 3.14, the correctness certificates for the machine (gauge_p) replace the dual certificate for the convex machine (gauge). The remark below compares these two types of certificates.

Remark 3.15 (Convex versus nonconvex learning)**.** *From Lemma 2.6, recall that the convex machine* (gauge) *succeeds if* $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp)$ *is separated along* $\text{range}(\mathcal{L}^*)$ *from the rest of* $\text{conv}(\mathcal{A})$*. In contrast, when* $p \leq d$*, the machine* (gauge_p) *succeeds under the weaker requirement that* $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp)$ *is separated along* $\text{range}(\mathcal{L}^*)$ *from each slice of* $\text{conv}(\mathcal{A})$ *formed by at most* p *atoms. Moreover, when* $p \geq d+1$*, the machine* (gauge_p) *reduces to the convex machine* (gauge) *by Proposition 3.4(vii), and a dual certificate in Lemma 2.6 roughly qualifies as a correctness certificate for Lemma 3.14. Indeed, the classical gauge function theory is a special case of the generalized theory in this section, as detailed in Section 3.4.1.*

In certain scenarios, we can in fact construct the correctness certificates prescribed in Lemma 3.14 and guarantee the success of the new machine (gauge_p), as detailed later in Section 3.4.2.

To close this section, we also note that it is not difficult to extend Lemma 3.14 to account for an inexact model, i.e., the case where $\epsilon > 0$ in (inexact). However, we avoid this added layer of complexity in the interest of readability.

3.4.1. *Gauge function theory as a special case*

In this section, as a sanity check, we show that the (convex) gauge function theory, reviewed in Section 2.1, is, in a certain (nontrivial) sense, a special case of the general theory developed so far in Section 3. To enunciate this connection, for simplicity, consider the exact model

$$y := \mathcal{L}(x^\sharp), \qquad x^\sharp \in \text{cone}(\mathcal{S}^\sharp),\ \mathcal{S}^\sharp \in \text{slice}_r(\mathcal{A}), \tag{3.20}$$

and recall the two learning machines

$$\min_x\ \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } \mathcal{G}_\mathcal{A}(x) \leq \mathcal{G}_\mathcal{A}(x^\sharp), \qquad (\text{see (gauge)}) \tag{3.21}$$

$$\min_x\ \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } \mathcal{G}_{\mathcal{A},p}(x) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \qquad (\text{see } (\text{gauge}_p)) \tag{3.22}$$

for an integer $p \geq r$, where r is the sparsity level of the model $x^\sharp$. First, the nested property of the gauge_p functions in (3.6) ensures that the two machines (3.21) and (3.22) are equivalent when $p \geq d+1$. More generally (and less trivially), when the convex machine (3.21) succeeds in learning $x^\sharp$ and its sparse decomposition in the alphabet $\mathcal{A}$, then so does the nonconvex machine (3.22), as formalized below.

Proposition 3.16. *Suppose that the assumptions in Lemma 2.6 are met. Consider the model* $x^\sharp \in \text{cone}(\mathcal{S}^\sharp)$ *in* (3.20)*. If* $x^\sharp = 0$*, then both machines* (3.21) *and* (3.22) *correctly return* 0*. Otherwise, let* $\mathcal{F}^\sharp$ *be an exposed face of* $\text{conv}(\mathcal{A})$ *such that* $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp) \in \mathcal{F}^\sharp$*, and suppose that* $\mathcal{S}^\sharp = \text{conv}(\mathcal{F}^\sharp \cup \{0\})$*. Then the machines* (3.21) *and* (3.22) *with* $p \geq r$ *both return* $x^\sharp$ *and its* r*-sparse decomposition in the slice* $\mathcal{S}^\sharp$*.*

The proof of Proposition 3.16 uses the dual certificate Q of the convex machine (3.21) in Lemma 2.6 as a correctness certificate for the new machine (3.22). We can also rephrase Proposition 3.16 differently: As soon as the requirement $\mathcal{S}^\sharp = \operatorname{conv}(\mathcal{F}^\sharp \cup \{0\})$ is met, both machines (3.21) and (3.22) succeed, provided that the dual certificate for the convex machine exists. It is important to note that the new machine can also succeed in scenarios where the convex machine fails. Indeed, recall that the requirement $\mathcal{S}^\sharp = \operatorname{conv}(\mathcal{F}^\sharp \cup \{0\})$ was violated in all negative examples of Section 2.3 in which the convex machine (3.21) failed and yet the new machine (gauge_p) succeeded, as we saw in Section 3.3.

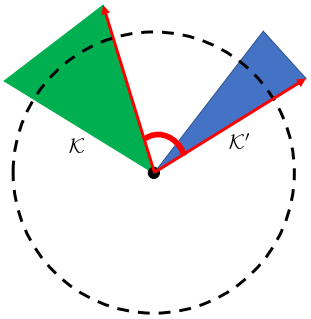


FIGURE 4. The angle between two cones, see Definition 3.18.

3.4.2. *Existence of the correctness certificates*

As was the case with Lemma 2.6 in the classical gauge function theory, establishing the existence of the certificates prescribed in Lemma 3.14 is a problem-specific task. Nevertheless, this section provides a somewhat general recipe for constructing these correctness certificates. To that end, we begin with a few definitions.

Definition 3.17 (Angle of a cone)**.** *The angle $\angle\mathcal{K}$ of a closed cone $\mathcal{K} \subset \mathbb{R}^d$ satisfies*

$$\cos(\angle\mathcal{K}) := \max_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \min_{u' \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle, \tag{3.23}$$

where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$.

As a simple example, the positive orthant in $\mathbb{R}^2$, which is a closed cone, has the angle of $\pi/4$.

Definition 3.18 (Angle between two cones)**.** *The angle between two closed cones $\mathcal{K}, \mathcal{K}' \subset \mathbb{R}^d$ satisfies*

$$\begin{aligned} \cos(\angle[\mathcal{K}, \mathcal{K}']) &:= \min\left(\min_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \max_{u' \in \mathcal{K}' \cap \mathbb{S}^{d-1}} \langle u, u' \rangle, \min_{u' \in \mathcal{K}' \cap \mathbb{S}^{d-1}} \max_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle \right) \\ &= 1 - \frac{1}{2} \left(\operatorname{dist}_{\mathrm{H}}(\mathcal{K} \cap \mathbb{S}^{d-1}, \mathcal{K}' \cap \mathbb{S}^{d-1}) \right)^2, \end{aligned} \tag{3.24}$$

where $\operatorname{dist}_{\mathrm{H}}$ denotes the (Euclidean) Hausdorff distance between two sets [62].

In words, the angle between two closed cones is the Hausdorff distance of their intersections with the unit sphere. For example, in Figure 4, the angle between the (partly drawn) blue and green cones equals the angle formed by the red arrows. It is worth noting that $\angle[\mathcal{K}, \mathcal{K}']$ coincides with the (largest) principal angle between two subspaces in the special case when $\mathcal{K}$ and $\mathcal{K}'$ are two subspaces of $\mathbb{R}^d$ [63]. Next, recall that the metric entropy of a set is a measure for how large or complex that set is [64].

Definition 3.19 (Metric entropy)**.** *The set $\operatorname{cover}(\mathcal{I}, \operatorname{dist}, \delta)$ is a δ-cover for the set $\mathcal{I}$ with respect to the metric* dist*, provided that for every $x \in \mathcal{I}$, there exists $x' \in \operatorname{cover}(\mathcal{I}, \operatorname{dist}, \delta)$ such that $\operatorname{dist}(x, x') \leq \delta$. The metric entropy of the set $\mathcal{I}$, denoted here by $\operatorname{entropy}(\mathcal{I}, \operatorname{dist}, \delta)$, is the (natural) logarithm of the size of its minimal cover.*

In linear inverse problems, it is not uncommon for the linear operator to be random to some degree, see for instance [11, 37, 9]. For our purposes, we quantify the randomness of the operator $\mathcal{L}$ as follows.

Definition 3.20 (Probabilistic near-isometry)**.** *For $\delta' \in [0,1)$, a random linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ is a δ'-near-isometry if*

$$(1-\delta')\|u\|_2 \leq \|\mathcal{L}(u)\|_2 \leq (1+\delta')\|u\|_2, \tag{3.25}$$

for every u in an (arbitrary but fixed) subspace $\mathcal{U} \subset \mathbb{R}^d$ with $m \geq C\dim(\mathcal{U})/\delta'^2$ and except with a probability of at most $\exp(-C'\delta'^2 m)$ over the choice of $\mathcal{L}$. Here, C, C' are universal constants.

For example, a matrix populated by independent Gaussian random variables, which may be identified with a random linear operator, is a near-isometry if the matrix is sufficiently flat and properly scaled, see Appendix B.4. Random matrix theory [65] offers far more general statements with broad applications in statistical inference and signal processing [11]. Finally, let us define the notion of critical angle below, which loosely takes on the role played in convex learning by the Gaussian width of the tangent cone, defined below or in [1].

Definition 3.21 (Critical angle)**.** *For an integer p, alphabet $\mathcal{A}$ and model $x^\sharp$, let*

$$\operatorname{slice}_{x^\sharp,p}(\mathcal{A}) := \left\{\mathcal{S} \in \operatorname{slice}_p(\mathcal{A}) : x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \notin \mathcal{S}\right\}, \tag{3.26}$$

collect all slices of $\operatorname{conv}(\mathcal{A})$, formed by at most p atoms, that do not contain the point $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp)$, see Definitions 2.1 and 3.3. The critical angle of the alphabet $\mathcal{A}$ with respect to the model $x^\sharp$ is then defined as

$$\theta_{x^\sharp,p}(\mathcal{A}) := \sup\left\{\angle\operatorname{cone}\left(\mathcal{S} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right) : \mathcal{S} \in \operatorname{slice}_{x^\sharp,p}(\mathcal{A})\right\}, \tag{3.27}$$

with the conventions that $0/0 = 0$, and $\theta_{x^\sharp,p}(\mathcal{A}) = 0$ if $\operatorname{slice}_{x^\sharp,p}(\mathcal{A})$ is empty.

As we will see shortly, the smaller the critical angle is, the easier it is to construct the correctness certificates prescribed in Lemma 3.14. In this sense, the critical angle has an analogue in convex learning. Precisely, the next result posits that, when $p \geq d+1$, i.e., when the machine (gauge$_p$) coincides with the convex machine (gauge), the critical angle is bounded by twice the angle of the corresponding tangent cone, defined below. Indeed, in the classical gauge function theory [1], the complexity of learning is quantified with a related quantity, i.e., the Gaussian width of the tangent cone.

Proposition 3.22 (Critical angle and the tangent cone)**.** *For an alphabet $\mathcal{A} \subset \mathbb{R}^d$, integer $p \geq d+1$ and model $x^\sharp \in \mathbb{R}^d$, the critical angle of the alphabet $\mathcal{A}$ with respect to the model $x^\sharp$ is bounded by twice the angle of the corresponding tangent cone, i.e., $\theta_{x^\sharp,p}(\mathcal{A}) \leq 2 \cdot \angle\operatorname{cone}(\mathcal{A} - x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp))$.*

Equipped with Definitions 3.18-3.21, we now present the last main result of this section. Informally speaking, Theorem 3.23 below states that, with a generic operator $\mathcal{L}$, the new machine (gauge$_p$) is successful when we have access to sufficiently many observations (m is large enough).

Theorem 3.23 (Exact recovery)**.** *Suppose that Assumptions 2.3(i) and (iii) on the alphabet $\mathcal{A}$ are met, and consider the model $x^\sharp$ in (inexact) with $\epsilon = 0$. Let us equip $\mathrm{slice}_p(\mathcal{A})$ in (3.1) with the pseudo-metric*[1] *that assigns the distance*

$$\mathrm{dist}_p(\mathcal{S}, \mathcal{S}') := \sqrt{2 - 2\cos\left(\angle\left[\mathrm{cone}\left(\mathcal{S} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right), \mathrm{cone}\left(\mathcal{S}' - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right)\right]\right)}, \tag{3.28}$$

to every two slices $\mathcal{S}, \mathcal{S}' \in \mathrm{slice}_p(\mathcal{A})$, with the convention that $0/0 = 0$. Lastly, for $\delta \in [0, 1)$, suppose that the random linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ is a δ-near-isometry in the sense of Definition 3.20. If $\mathcal{G}_{\mathcal{A},p}(x^\sharp) = 0$, then the machine (gauge$_p$) correctly returns 0.

Otherwise, set $p \geq r$, where r is the sparsity level of $x^\sharp$ in (inexact). Then the machine (gauge$_p$) returns $x^\sharp$ and a p-sparse decomposition of $x^\sharp$ in $\mathcal{A}$, provided that

$$m \geq \frac{Cp}{\delta^2} + \frac{2}{C'\delta^2} \cdot \mathrm{entropy}\left(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}_p, \frac{\delta}{\max(\|\mathcal{L}\|_{\mathrm{op}}^2, 1)}\right), \quad \delta \leq \frac{\cos(\theta_{x^\sharp,p}(\mathcal{A}))}{2}, \tag{3.29}$$

and except with a probability of at most $\exp(-C'\delta^2 m/2)$, for universal constants C, C'. Above, $\|\mathcal{L}\|_{\mathrm{op}}$ is the operator norm of $\mathcal{L}$.

The proof technique of Theorem 3.23 appears to be new in this context and might be of independent interest. More specifically, the proof relies on a covering argument, where we form a fine cover for all relevant slices of $\mathrm{conv}(\mathcal{A})$ (with respect to the metric dist_p in (3.28)) and then explicitly construct a correctness certificate for each slice, with high probability over the choice of the random operator $\mathcal{L}$. The failure probabilities are added up via a union bound. We then show that these certificates actually qualify as optimality certificates for *all* relevant slices of $\mathrm{conv}(\mathcal{A})$, even those not present in the fine cover. One key technical challenge in the proof is to handle the slices nearby the true model $x^\sharp$.

As detailed in the next remark, we may consider Theorem 3.23 as a statistical guarantee for the new machine (gauge$_p$) that extends the analogous guarantee in [1, Corollary 3.3.1] for the convex machine (gauge).

Remark 3.24 (Complexity of learning)**.** *When applying Theorem 3.23, we are primarily interested in the regime $p = O(r)$. In this regime, $\mathrm{slice}_p(\mathcal{A})$ is a collection of polytopes with dimension of at most $p = O(r)$ and, loosely speaking, the right-hand side of (3.29) is $\widetilde{O}(r)$. Theorem 3.23 then predicts that, as soon as $m = \widetilde{\Omega}(r)$, the machine (gauge$_p$) successfully recovers the model $x^\sharp$ and a p-sparse decomposition of $x^\sharp$ in the alphabet $\mathcal{A}$. In this sense, Theorem 3.23 is a statistical guarantee for the machine (gauge$_p$) that extends the analogous guarantee in* [1, Corollary 3.3.1] *for the convex machine (gauge). For completeness,* [1, Corollary 3.3.1] *is reviewed in Appendix B.8.*

We can informally compare these two statistical guarantees by comparing their respective lower bounds on the number of observations m. In view of (2.7), the convex machine (gauge) requires $m \geq \dim(\mathcal{F}^\sharp)$ observations to recover the model $x^\sharp$, where $\mathcal{F}^\sharp$ is an exposed face of $\mathcal{A} \subset \mathbb{R}^d$ passing through $x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp)$. In all negative toy examples of Section 2.3, we had $\dim(\mathcal{F}^\sharp) = d$. Consequently,

[1] For a pseudo-metric, $\mathrm{dist}(x, x') = 0$ does not imply that $x = x'$.

the convex machine (gauge) *requires* $m \geq d$ *observations to recover* $x^\sharp$*, in contrast to the potentially much smaller* $m = \widetilde{\Omega}(r)$ *observations needed by the new machine* (gauge$_p$)*.*

Note also that, within the regime $p \geq d+1$*, the machine* (gauge$_p$) *coincides with the convex machine* (gauge) *and we refer the reader to Remark 3.15. Beyond the informal discussion above, the entropy number in Theorem 3.23 should be estimated on a case-by-case basis by taking into account the geometry of the learning alphabet* $\mathcal{A}$*.*

4. Stylized Applications of the Gauge$_p$ Function Theory

In Section 3.3, we introduced the new machine (gauge$_p$). The building block of new machine is the gauge$_p$ function, which was introduced and studied in Sections 3.1 and 3.2. Moreover, in Section 3.4, the machine (gauge$_p$) was equipped by a few statistical guarantees, namely, Lemma 3.14 and Theorem 3.23.

Without being exhaustive, this section showcases the potential of the new learning machine by applying it to two representative problems. As mentioned in the introduction, each problem below represents a highly active research area and it is not our intention to improve over the state of the art in these problems, but rather we only hope to convince the reader about the potential of the new machine (gauge$_p$) and that it merits further investigation in the future. In particular, in the interest of brevity, we have also removed two similar applications (phase retrieval and shallow neural networks) from the manuscript.

4.1. Manifold-Like Models

The classical gauge function theory often fails to learn manifold models, as highlighted in Section 2.3 with a toy example, despite the importance of manifold models in signal processing and machine learning [22, 17, 66]. In this section, we consider a slightly more general family of models and show that the new machine (gauge$_p$) succeeds in learning them. To be specific, suppose that the alphabet $\mathcal{A}$ is an arbitrary subset of $\mathbb{R}^d$, and consider the 1-sparse model

$$y := \mathcal{L}(A^\sharp), \qquad A^\sharp \in \mathcal{A}, \tag{4.1}$$

where $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator and, for simplicity, we have not accounted for any inexactness in the model, such as measurement noise. Note that the model (4.1) is a manifold model when $\mathcal{A}$ is an embedded submanifold of $\mathbb{R}^d$. For this reason, we refer to (4.1) as a manifold-like model. Such manifold-like models are common in learning with nonlinear constraints, see [67, 68] for examples in the context of learning with a generative adversarial network as the prior. To recover the atom $A^\sharp$ in (4.1), we may apply the machine (gauge$_p$) for any $p \geq 1$. In particular, for $p = 1$, the machine (gauge$_p$) reduces to

$$\min_{c,A} \ \|c\mathcal{L}(A) - y\|_2^2 \text{ subject to } 0 \leq c \leq 1, \ A \in \mathcal{A}, \tag{4.2}$$

which is closely related to the machines proposed and numerically tested in [22, Equation 12] and [17, Equation 20] with one image patch. A corollary of Theorem 3.23 for $r = p = 1$, presented below, ensures that the machine (4.2) successfully learns the true atom $A^\sharp$ from the vector of observations y,

when $\mathcal{L}$ is a generic linear operator and we have access to sufficiently many observations. In the corollary below, for tidiness, we assume that $A^\sharp$ has unit norm. This assumption is without any loss of generality, since we can always scale the alphabet $\mathcal{A}$ to enforce it. In the corollary below, also for tidiness, instead of directly using the critical angle in Definition 3.21, we will make use of the quantity

$$\theta'_{A^\sharp}(\mathcal{A}) := \inf_{A \in \mathcal{A}} \angle[A - A^\sharp, A^\sharp], \tag{4.3}$$

which is closely related to the critical angle of the alphabet $\mathcal{A}$ with respect to the atom $A^\sharp$. Indeed, in the proof of the corollary, we show that

$$\theta_{A^\sharp,1}(\mathcal{A}) \leq \frac{1}{2}(\pi - \theta'_{A^\sharp}(\mathcal{A})), \tag{4.4}$$

where $\theta_{A^\sharp,1}(\mathcal{A})$ denotes the critical angle of $\mathcal{A}$ with respect to the atom $A^\sharp$. Let us now state the corollary.

Corollary 4.1 (Learning with manifold-like models)**.** *Consider the model* (4.1) *and assume without loss of generality that* $\|A^\sharp\|_2 = 1$*. Suppose that Assumptions 2.3(i) and (iii) on the alphabet* $\mathcal{A}$ *are met, and that* $\theta'_{A^\sharp}(\mathcal{A}) > 0$*, see* (4.3)*. Let us equip* $\mathcal{A}$ *with the pseudo-metric that assigns the distance*

$$\mathrm{dist}_{A^\sharp}(A, A') := \left\| \frac{A - A^\sharp}{\|A - A^\sharp\|_2} - \frac{A' - A^\sharp}{\|A' - A^\sharp\|_2} \right\|_2, \tag{4.5}$$

to every two atoms $A, A' \in \mathcal{A}$*, with the convention that* $0/0 = 0$*. Lastly, for* $\delta \in [0, 1)$*, suppose that the linear operator* $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ *in* (4.1) *is a* δ*-near-isometry in the sense of Definition 3.20. Then the machine* (4.2) *returns* $A^\sharp$*, provided that*

$$m \geq \frac{C}{\delta^2} + \frac{2}{C'\delta^2} \cdot \mathrm{entropy}\left(\mathcal{A}, \mathrm{dist}_{A^\sharp}, \frac{\delta}{\max(\|\mathcal{L}\|_{\mathrm{op}}^2, 1)}\right), \quad \delta \leq \frac{\sin(\theta'_{A^\sharp}(\mathcal{A}))}{2}, \tag{4.6}$$

and except with a probability of at most $\exp(-C'\delta^2 m/2)$*, for universal constants* C, C'*.*

As a corollary of Theorem 3.23, the proof of Corollary 4.1 simply estimates the entropy number of the set $\mathrm{slice}_1(\mathcal{A}) = \bigcup_{0 \leq \tau \leq 1} \tau\mathcal{A}$ with respect to the metric dist_1 in Theorem 3.23. (The identity in the last sentence follows from Proposition 3.2.) One can better quantify the right-hand side of (4.6) on a case-by-case basis, depending on the alphabet $\mathcal{A}$ at hand. This direction of research is discussed in the remark below.

Remark 4.2 (Complexity of manifold models)**.** *In Corollary 4.1, suppose in addition that* $\mathcal{A}$ *is an embedded submanifold of* $\mathbb{R}^d$*. Informally, one might expect the machine* (4.2) *to succeed with* $m = \widetilde{\Omega}(\dim(\mathcal{A}))$ *observations* [22]*, where* $\dim(\mathcal{A})$ *denotes the dimension of the manifold* $\mathcal{A}$*. Formally, in order to estimate the entropy number in* (4.6) *in terms of the geometric attributes of the alphabet* $\mathcal{A}$ *(such as volume or curvature), we might have to impose additional geometric regularities on* $\mathcal{A}$*, particularly sufficient smoothness near the target atom* $A^\sharp$*. Identifying these regularities and a careful estimation of the entropy number in* (4.6) *appears to be nontrivial and is left as a future research direction.*

The final remark of this section revisits the toy example in Figure 1.

Remark 4.3 (Critical angle)**.** *For an alphabet $\mathcal{A} \subset \mathbb{R}^d$ and an atom $A^\sharp \in \mathcal{A}$, recall from (4.4) that $\theta'_{A^\sharp}(\mathcal{A})$ in (4.3) replaces, for tidiness, the critical angle in Definition 3.21. Note that the angle $\theta'_{A^\sharp}(\mathcal{A})$ plays a key role in Corollary 4.1. More specifically, Corollary 4.1 demands that $\theta'_{A^\sharp}(\mathcal{A}) > 0$, i.e., it requires that $\angle[A - A^\sharp, A^\sharp] > 0$ for every atom $A \in \mathcal{A}$, see (4.3). For an informal comparison, it is not difficult to verify that the convex machine (gauge) requires $m = \Omega(d)$ observations to recover $A^\sharp$ if $\angle[A - A^\sharp, A^\sharp] \leq \pi/2$ for every atom $A \in \mathcal{A}$. In Example 2.14, recall that $\mathcal{A}$ was a spiral and $A^\sharp$ was specified in (2.31), see the red dot in Figure 1. For that example, we can verify that*

$$\theta'_{A^\sharp/\|A^\sharp\|_2}(\mathcal{A}/\|A^\sharp\|_2) \approx 51^\circ,$$

where we have rescaled the alphabet $\mathcal{A}$ so that the model $A^\sharp/\|A^\sharp\|_2$ has unit norm, to conveniently match the prescription of Corollary 4.1.

To summarize, the preliminary result in Corollary 4.1 indicates the potential of the new machine for learning manifold-like models. The important problem of computing the entropy number of $\mathcal{A}$ in terms of its geometric attributes (such as volume and curvature) remains as a nontrivial future research question.

4.2. **Sparse Principal Component Analysis**

With a negative toy example, we saw in Section 2.3 that the classical gauge function theory might fail in general for the sparse PCA problem. More specifically, we saw that the gauge function might fail to promote sparsity, and that the minimal decomposition which achieves the gauge function value might not have a sparse decomposition within the learning alphabet. In contrast, the generalized theory developed in Section 3 immediately rectifies this problem since, by design, any minimal decomposition that achieves the gauge$_p$ function value always has a p-sparse decomposition, see Proposition 3.4(vi).

The remainder of this section is devoted to the popular spiked covariance model for sparse PCA [44, 46, 69, 70], where we will show that the proposed learning machine achieves the information-theoretic performance limit of sparse PCA, and naturally generalizes beyond the spiked covariance model. More specifically, for sparsity level k, consider the alphabet

$$\mathcal{A} := \{uu^\top : \|u\|_2 = 1,\ \|u\|_0 \leq k\} \subset \mathbb{R}^{d_1 \times d_1}, \tag{4.7}$$

where $\|u\|_0$ is the number of nonzero entries of the vector u. As we will see shortly, the sparsity level of u (number of its nonzero entries) should not be confused with the sparsity of a statistical model (number of atoms of the alphabet $\mathcal{A}$ present in a model). For $\theta \in [0, 1)$, consider also a Gaussian random vector in $\mathbb{R}^{d_1}$ with zero mean and the covariance matrix $\Sigma \in \mathbb{R}^{d_1 \times d_1}$, where the covariance matrix is specified as

$$\Sigma := A^\sharp + \theta I_{d_1}, \qquad A^\sharp = u^\sharp (u^\sharp)^\top \in \mathcal{A}, \tag{4.8}$$

where $I_{d_1} \in \mathbb{R}^{d_1 \times d_1}$ is the identity matrix. Above, $A^\sharp$ is the "spike" in the spiked covariance model. Instead of the covariance matrix Σ, we have access to the sample covariance matrix

$$y := \frac{1}{n} \sum_{i=1}^{n} z_i z_i^\top, \tag{4.9}$$

formed by the samples $\{z_i\}_{i=1}^n \subset \mathbb{R}^{d_1}$, drawn independently from the distribution normal$(0, \Sigma)$. The objective of sparse PCA is to identify the spike in Σ, i.e., to identify the atom $A^\sharp$ in (4.8), given the sample covariance matrix y. In view of the model (inexact), here our inexact 1-sparse model is

$$y := A^\sharp + e, \qquad A^\sharp \in \mathcal{A},\ e := \frac{1}{n} \sum_{i=1}^{n} z_i z_i^\top - A^\sharp. \tag{4.10}$$

with the alphabet $\mathcal{A}$ defined in (4.7). For $p = 1$, the proposed machine in (gauge$_p$) is equivalent to

$$\min_{c,A} \|y - cA\|_{\mathrm{F}}^2 \text{ subject to } 0 \le c \le \mathcal{G}_{\mathcal{A},1}(A^\sharp) = 1 \text{ and } A \in \mathcal{A}, \tag{4.11}$$

where the identity $\mathcal{G}_{\mathcal{A},1}(A^\sharp) = 1$ above follows from (3.4) and the fact that $A^\sharp$ is an extreme point of conv$(\mathcal{A})$, which is in turn true because the alphabet $\mathcal{A}$ is a subset of the unit sphere in $\mathbb{R}^{d_1 \times d_1}$, see (4.7). Since we are only interested in recovering the atom $A^\sharp$, and not its coefficient $c^\sharp$, it suffices to consider the optimization over A within (4.11), i.e.,

$$\max_{A \in \mathcal{A}} \langle y, A \rangle = \max_{\|u\|_0 \le k} u^\top y u. \qquad \text{(see (4.7))} \tag{4.12}$$

The problem on the right-hand side above is often the starting point of sparse PCA algorithms and, in particular, the well-known convex relaxation proposed by [44]. This relaxation and, a fortiori, the machine (4.11) achieve the information-theoretic performance limit of sparse PCA [46, Theorems 2 and 3].

Moreover, for $p > 1$, the machine (gauge$_p$) naturally generalizes beyond the spiked covariance model, i.e., when Σ in (4.8) might have multiple spikes. Indeed, as discussed in Section 2.3, for $p > 1$, the common alternative of deflation [47] might be numerically unstable. Yet another alternative to deflation is to search for a subspace with sparse basis vectors, which all together forgoes the individual sparse components in favour of identifying a sparse subspace [71]. In view of the shortcomings of these alternatives, an important future research question might be to study the statistical performance of the machine (gauge$_p$) for $p > 1$ and with the learning alphabet of sparse PCA specified in (4.7).

5. Computational Aspects and a Tractable Numerical Scheme

For certain alphabets, the optimization landscape of the new machine (gauge$_p$) does not have any spurious stationary points, and (gauge$_p$) is therefore amenable to a variety of optimization algorithms. This in turn leads to substantial computational and storage savings, when p is sufficiently small, in comparison to a direct implementation of the convex machine (gauge). A prominent example is Remark 3.12 and the well-known Burer-Monteiro factorization for certain matrix- or tensor-valued learning problems, as documented in, for instance [19, 59] or [20, Section 2.1].

For certain other alphabets, such as smooth manifolds [22, 72] or shallow neural networks [73], the optimization landscape of (gauge$_p$) might in general contain spurious stationary points which could potentially trap first- or second-order optimization algorithms, such as gradient descent. Nevertheless, problem (gauge$_p$) can be reformulated as a smooth nonconvex optimization problem and solved efficiently to (near) stationarity, rather than global optimality, with a variety of algorithms, including the gradient descent [74]. This compromise (between optimality and tractability) is common in machine learning: As an example, empirical risk minimization is known to be NP-hard for neural networks in general and the practitioners instead seek local (rather than global) optimality by means of first- or second-order optimization algorithms [23, Chapter 20].

Yet for many other alphabets, such as the one in Example 2.14 (sparse regression [75]), the problem (gauge$_p$) is NP-hard in general for $p < d$ and, moreover, neither of the above approaches in this section are directly applicable. There are, however, compelling reasons to remain optimistic for such alphabets. For example, after decades of research, modern mixed-integer optimization algorithms that directly solve problem (3.18) for sparse regression are now competitive with convex heuristics in speed and scalability [24, 25].

5.1. Computational Approach

In view of these recent developments in mixed-integer programming, we provide below a tractable numerical scheme for solving problem (gauge$_p$) when the learning alphabet is finite. Let us now turn to the details. When the learning alphabet $\mathcal{A}$ is finite, the following lemma offers an exact reformulation of problem (gauge$_p$) as a mixed integer quadratic programming (MIQP). MIQP is in general an NP-hard problem [76], which comes at no surprise since the original problem of sparse recovery of the model (exact) is also known to be hard [75]. Nonetheless, the lemma below allows us to deploy the rich literature of computational mathematical programming dedicated to MIQP, see [77, 78] and the references therein.

Lemma 5.1 (MIQP reformulation)**.** *Suppose that the alphabet $\mathcal{A}$ is finite with the cardinality $|\mathcal{A}| < \infty$. If $M > 0$ is sufficiently large, the machine (gauge$_p$) is equivalent to the MIQP optimization program*

$$\min_{c,s} \left\{ \Big\| \sum_{i=1}^{|\mathcal{A}|} c_i \mathcal{L}(A_i) - y \Big\|_2^2 \ : \ \sum_{i=1}^{|\mathcal{A}|} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp),\ |c_i| \leq M s_i,\ \sum_{i=1}^{|\mathcal{A}|} s_i = p,\ c_i \geq 0,\ s_i \in \{0,1\} \right\}. \quad (5.1)$$

The MIQP reformulation in Lemma 5.1 builds on the so called "big-M" technique where M is only required to be a sufficiently large constant. It is well known that the choice of M typically has a significant impact on the performance of cutting plane algorithms for convex integer optimization [79]. To address this issue, inspired by the recent work of [24], we utilize a dual reformulation of the optimization program Lemma 5.1 in order to propose a good starting point (warm start) for the branch-and-bound algorithms. This program is a slight generalization of the one proposed in [24] where the linear constraints such as $c_i \geq 0$ and $\sum_{i=1}^{|\mathcal{A}|} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp)$ can also systematically be handled.

Proposition 5.2 (Tractable algorithm)**.** *Let us define the matrices*

$$A = \begin{bmatrix} A_1 & A_2 & \dots & A_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{d\times|\mathcal{A}|}, \quad C = \begin{bmatrix} -I_{|\mathcal{A}|} \\ \mathbb{1}_{|\mathcal{A}|}^\top \end{bmatrix}, \quad g = \begin{bmatrix} 0_{|\mathcal{A}|} \\ \mathcal{G}_{\mathcal{A},p}(x^\sharp) \end{bmatrix},$$

where $|\mathcal{A}|$ *is the size of the finite set* $\mathcal{A}$*. In addition,* $I_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ *is the identity matrix,* $\mathbb{1}_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$ *is a vector of all ones, and* $0_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|}$ *is a vector of zeros. The optimal value of the machine* (gauge$_p$) *coincides with the minimax program*

$$\min_{\substack{S\subset[|\mathcal{A}|]\\ |S|\le p}} \max_{\mu\ge 0,\lambda} -\frac{1}{2}\|\lambda\|_2^2 - \langle g,\mu\rangle - \frac{\gamma}{2}\sum_{i\in S}\left((\mathcal{L}(A))^\top\lambda - C^\top\mu\right)_i^2, \tag{5.2}$$

where γ *is a sufficiently large constant,* $\mathcal{L}(\mathcal{A}) = [\mathcal{L}(A_1)\,\mathcal{L}(A_2)\,\cdots\,\mathcal{L}(A_{|\mathcal{A}|})] \in \mathbb{R}^{m\times|\mathcal{A}|}$*, and the symbol* $(u)_i$ *returns the* i^{th} *coordinate of a vector* u*. Moreover, the indices of the ones in the optimal binary variable* s *in* (5.1) *coincide with the optimal subset* $S \subset [|\mathcal{A}|] = \{1,\cdots,|\mathcal{A}|\}$ *in* (5.2)*. Lastly, the solution of* (5.2) *is the fixed-point of the algorithm*

$$\begin{bmatrix} \lambda_{k+1} \\ \mu_{k+1} \end{bmatrix} = \arg\max_{\mu\ge 0,\,\lambda} -\frac{1}{2}\|\lambda\|_2^2 - \langle g,\mu\rangle - \frac{\gamma}{2}\sum_{i\in S_k}\left((\mathcal{L}(A))^\top\lambda - C^\top\mu\right)_i^2 \tag{5.3a}$$

$$S_{k+1} = \arg\max_{S\subset[|\mathcal{A}|],|S|\le p}\sum_{i\in S}\left((\mathcal{L}(A))^\top\lambda_k - C^\top\mu_k\right)_i^2. \tag{5.3b}$$

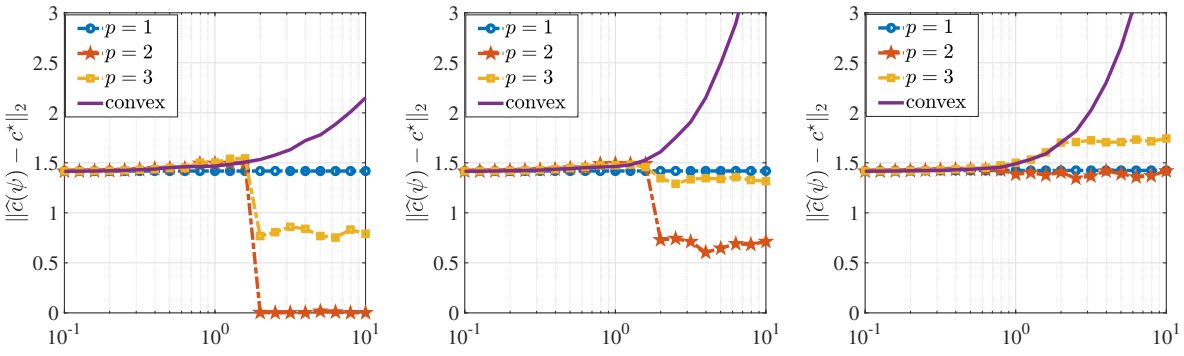
We emphasize that in (5.3a) the objective function is concave and quadratic jointly in the variable $(\lambda,\,\mu)$ where the number of the summands is the sparsity level $|S| = p$. Moreover, and more importantly, the set-valued optimization (5.3b) admits an (almost) analytic solution as it suffices to select only the first p coordinates $i \le |\mathcal{A}|$ where $\left((\mathcal{L}(A))^\top\lambda_k - C^\top\mu_k\right)_i^2$ is maximized. The complexity of this step reduces to a sorting algorithm. Therefore, the algorithm (5.3) is indeed computationally a highly tractable implementation of the machine (gauge$_p$), which may merit a more comprehensive numerical investigation in the future. If the iteration of (5.3) converges, we solve the the learning machine (gauge$_p$). However, it often happens that the algorithm (5.3) oscillates between several discrete solutions S. We note that these candidates can then be chosen as "warm start" for the MIQP problem (5.1) à la [24].

5.2. **Numerical Example**

As a numerical example of the MIQP formulation above, let us revisit super-resolution in Figure 3, where the alphabet is comprised of 20 translated copies of a Gaussian wave, i.e.,

$$\mathcal{A} := \{A_\theta\}_{\theta\in\Theta}, \qquad A_\theta(\vartheta) = e^{-\frac{(\vartheta-\theta)^2}{0.35^2}}, \qquad \vartheta\in\mathbb{R}, \tag{5.4}$$

and $\Theta = \{\theta_i\}_{i=1}^{20}$ is the uniform grid of size 20 over the observational interval $\vartheta \in [0,1]$. We consider the model (inexact) with $x^\sharp := c_{10}^\sharp A_{\theta_{10}} + c_{11}^\sharp A_{\theta_{11}}$. The operator $\mathcal{L}$ in (inexact) evaluates and stores its input function at 40 random locations on the interval $[0,1]$. We consider three different noise levels, namely, when e is a vector of zero-mean and independent Gaussian random variables with variances of $\sigma_e^2 \in \{10^{-6}, 10^{-4}, 10^{-2}\}$. For various values of a constant ψ and the integer $p \in \{1,2,3,20\}$, we



(a) $\sigma_e = 0.001$ (b) $\sigma_e = 0.01$ (c) $\sigma_e = 0.1$

FIGURE 5. Performance of the nonconvex machine ((gauge$_p$) for $p \in \{1, 2, 3\}$) versus the convex counterpart ((gauge) or equivalently (gauge$_p$) for $p = |\mathcal{A}| = 20$).

then solve the problem (5.1) for different values of $\psi = \mathcal{G}_{\mathcal{A},p}(x^\sharp)$, and show the error between the recovered and true coefficient vectors ($\|\widehat{c}(\psi) - c^\sharp\|_2$) in Figure 5. Each plot is obtained by averaging the recovery errors over 200 independent experiments with different realizations of the noise vector. For $p \in \{1, 2, 3\}$, we used MOSEK with the interface of YALMIP [80] to solve (5.1) and implement the machine (gauge$_p$). For $p = |\mathcal{A}| = 20$, the problem (5.1) is a convex program. Indeed, for the choice of $p = 20$, the machine (gauge$_p$) coincides with the convex machine (gauge).

In Figure 5, the sharp transition in the plots is explained by the fact that the true model $x^\sharp$ is not feasible for small values of ψ. Moreover, the poor performance of the machine (gauge$_p$) for $p = 1$ is explained by the fact that $x^\sharp$, with sparsity level of two, is never feasible for problem (5.1) with $p = 1$. However, for $p \in \{2, 3\}$, the proposed machine (gauge$_p$) considerably outperforms the convex machine (gauge).

ACKNOWLEDGEMENTS

The authors are grateful to Gongguo Tang, Michael Wakin and Konstantinos Zygalakis for helpful discussions and their valuable feedback.

APPENDIX A. TECHNICAL DETAILS OF SECTION 2

A.1. Proof of Lemma 2.6

Let $\widehat{x}$ be a minimizer of problem (gauge). Suppose that $\mathcal{G}_{\mathcal{A}}(x^\sharp) = 0$. By feasibility of $\widehat{x}$ for problem (gauge), it holds that $\mathcal{G}_{\mathcal{A}}(\widehat{x}) \leq \mathcal{G}_{\mathcal{A}}(x^\sharp) = 0$. By Assumption 2.3(ii), $\mathcal{A}$ is symmetric and $\mathcal{G}_{\mathcal{A}}$ is thus a norm in $\mathbb{R}^d$. Because $\mathcal{G}_{\mathcal{A}}$ is a norm, $\mathcal{G}_{\mathcal{A}}(\widehat{x}) = \mathcal{G}_{\mathcal{A}}(x^\sharp) = 0$ implies that $\widehat{x} = x^\sharp = 0$. We thus assume that $\mathcal{G}_{\mathcal{A}}(x^\sharp) > 0$ from now on. By definition of the gauge function in (2.2), we have

$$\widehat{x}/\mathcal{G}_{\mathcal{A}}(\widehat{x}) \in \operatorname{conv}(\mathcal{A}),$$

with the convention that $0/0 = 0$. Since $\mathcal{A}$ is symmetric by Assumption 2.3(ii), it also holds that

$$t \cdot \widehat{x}/\mathcal{G}_{\mathcal{A}}(\widehat{x}) \in \operatorname{conv}(\mathcal{A}), \qquad \forall t \in [-1, 1], \tag{A.1}$$

i.e., the line segment connecting $\pm\widehat{x}/\mathcal{G}_{\mathcal{A}}(\widehat{x})$ also belongs to $\operatorname{conv}(\mathcal{A})$. Moreover, by feasibility of $\widehat{x}$ in problem (gauge), we have that

$$\mathcal{G}_{\mathcal{A}}(\widehat{x}) \leq \mathcal{G}_{\mathcal{A}}(x^\sharp).$$

In view of the above relation, for the choice of $t = \mathcal{G}_{\mathcal{A}}(\widehat{x})/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in [0, 1]$, (A.1) reduces to

$$\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \operatorname{conv}(\mathcal{A}). \tag{A.2}$$

For future reference, note also that the feasibility of $x^\sharp$ and optimality of $\widehat{x}$ in problem (gauge) imply that $\mathcal{L}(x^\sharp) = \mathcal{L}(\widehat{x}) = y$ and, consequently,

$$\mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A}}(x^\sharp)}\right) = \mathcal{L}(\widehat{x} - x^\sharp) = 0. \qquad (\mathcal{G}_{\mathcal{A}}(x^\sharp) > 0) \tag{A.3}$$

We now consider two cases:

(1) Suppose that $\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \mathcal{F}^\sharp$. Then we find that

$$\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} \in \mathcal{F}^\sharp - \mathcal{F}^\sharp \subset \operatorname{lin}(\mathcal{F}^\sharp). \tag{A.4}$$

Consequently, (A.3), (A.4) and the injectivity of $\mathcal{L}$ on $\operatorname{lin}(\mathcal{F}^\sharp)$ together imply that $\widehat{x} = x^\sharp$.

(2) Suppose that $\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^\sharp) \notin \mathcal{F}^\sharp$. We can therefore strengthen (A.2) as

$$\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \operatorname{conv}(\mathcal{A}) - \mathcal{F}^\sharp. \tag{A.5}$$

By assumption of Lemma 2.6, there exists a certificate $Q = \mathcal{L}^*(q)$ that satisfies (2.6). Recalling (A.3), we then write that

$$\begin{aligned}
0 &= \langle q, \mathcal{L}(\widehat{x} - x^\sharp)\rangle \qquad \text{(see (A.3))} \\
&= \langle \mathcal{L}^*(q), \widehat{x} - x^\sharp\rangle = \langle Q, \widehat{x} - x^\sharp\rangle = \mathcal{G}_{\mathcal{A}}(x^\sharp)\left\langle Q, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A}}(x^\sharp)}\right\rangle \quad (\mathcal{G}_{\mathcal{A}}(x^\sharp) > 0) \\
&= \left\langle Q, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A}}(x^\sharp)}\right\rangle < 0
\end{aligned}$$

where above we used the assumption that $\mathcal{G}(x^\sharp) > 0$ as well as (2.6) and (A.5). To avoid the above contradiction, it must hold that $\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^\sharp) \in \mathcal{F}^\sharp$ which again implies that $\widehat{x} = x^\sharp$.

This completes the proof of Lemma 2.6.

A.2. **Proof of Proposition 2.7**

Suppose that $m = d$ and that $\mathcal{L}$ is the identity operator. Then problem (gauge) reduces to computing a minimal decomposition of $x^\sharp$ that achieves $\mathcal{G}_{\mathcal{A}}(x^\sharp)$. We next record a simple result that completes the proof of Proposition 2.7.

Lemma A.1. *Suppose that Assumption 2.3(iii) on the alphabet $\mathcal{A} \subset \mathbb{R}^d$ is fulfilled. Consider model $x^\sharp \in \text{cone}(\mathcal{S}^\sharp)$ in (exact) and let $\mathcal{F}^\sharp$ be an exposed face of $\text{conv}(\mathcal{A})$ such that $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp) \in \mathcal{F}^\sharp$. Suppose also that $\mathcal{S}^\sharp = \text{conv}(\mathcal{F}^\sharp \cup \{0\})$. Then $x^\sharp$ has a minimal decomposition in $\mathcal{F}^\sharp \subset \mathcal{S}^\sharp$ that achieves the gauge function value $\mathcal{G}_\mathcal{A}(x^\sharp)$, see (2.2).*

Proof. Note that $x^\sharp \in \text{cone}(\mathcal{F}^\sharp) = \text{cone}(\mathcal{S}^\sharp)$ by assumption. Using the definition of the gauge function in (2.2), it follows that

$$\mathcal{G}_\mathcal{A}(x^\sharp) = \inf\left\{t : x^\sharp/t \in \text{conv}(\mathcal{A})\right\} = \inf\left\{t : x^\sharp/t \in \mathcal{F}^\sharp\right\}.$$

Consequently, $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp) \in \mathcal{F}^\sharp \subset \mathcal{S}^\sharp$ is the minimal decomposition of $x^\sharp$ that achieves the value of the gauge function $\mathcal{G}_\mathcal{A}(x^\sharp)$. This completes the proof of Lemma A.1. □

A.3. Proof of Proposition 2.13

Recall from (2.27) that $x^\sharp = c^\sharp A^\sharp$. We first use the definition of the gauge function in (2.2) to write that

$$\begin{aligned}\mathcal{G}_\mathcal{A}(x^\sharp) &= c^\sharp \cdot \mathcal{G}_\mathcal{A}(A^\sharp) \qquad \text{(positive homogenity)} \\ &= c^\sharp \cdot \inf\left\{\sum_{i=1}^{l} c_i : A^\sharp = \sum_{i=1}^{l} c_i A_i,\ c_i \geq 0,\ A_i \in \mathcal{A},\ \forall i \in [l]\right\} \leq c^\sharp. \end{aligned} \tag{A.6}$$

On the other hand, note that from (2.4) we have

$$\begin{aligned}\mathcal{G}_\mathcal{A}(x^\sharp) &= \sup\left\{\langle x^\sharp, z\rangle : \mathcal{D}_\mathcal{A}(z) \leq 1\right\} = \sup\left\{\langle x^\sharp, z\rangle : \langle A, z\rangle \leq 1,\ \forall A \in \mathcal{A}\right\} \\ &\geq \langle x^\sharp, A^\sharp\rangle = c^\sharp\langle A^\sharp, A^\sharp\rangle = c^\sharp, \end{aligned} \tag{A.7}$$

where the third line above holds because $\mathcal{D}_\mathcal{A}(A^\sharp) \leq 1$, as a consequence of Assumption 2.3(iv). Note that the supremum in (A.7) is achieved for $A^\sharp$. By combining (A.6) and (A.7), we find that

$$c^\sharp = \mathcal{G}_\mathcal{A}(x^\sharp) = \langle x^\sharp, A^\sharp\rangle. \tag{A.8}$$

In view of (A.8), the Holder's inequality allows us to strengthen $\mathcal{D}_\mathcal{A}(A^\sharp) \leq 1$ to $\mathcal{D}_\mathcal{A}(A^\sharp) = 1$, which in turn implies that

$$\begin{aligned}\mathcal{D}_\mathcal{A}(x^\sharp) &= c^\sharp \cdot \mathcal{D}_\mathcal{A}(A^\sharp) \qquad \text{(positive homogenity)} \\ &= c^\sharp \qquad (\mathcal{D}_\mathcal{A}(A^\sharp) = 1) \\ &= c^\sharp \sup_{A \in \mathcal{A}} \langle A^\sharp, A\rangle, \qquad \text{(Assumption 2.3(iv))} \end{aligned} \tag{A.9}$$

and the supremum above is achieved for $A^\sharp$. This completes the proof of Proposition 2.13.

Appendix B. Technical Details of Section 3

B.1. Proof of Proposition 3.2

The nested property of $\{\text{conv}_p(\mathcal{A})\}_p$ in (3.2) is evident from its definition in (3.1). To show the far-left identity in (3.2), we use (3.1) for $r = 1$ to write that

$$\begin{aligned}\text{conv}_1(\mathcal{A}) &= \bigcup_{A\in\mathcal{A}} \text{conv}(\{A, 0\}) \qquad \text{(see (3.1))} \\ &= \bigcup_{A\in\mathcal{A}} \bigcup_{0\le\tau\le 1} \tau A = \bigcup_{0\le\tau\le 1} \bigcup_{A\in\mathcal{A}} \tau A = \bigcup_{0\le\tau\le 1} \tau\mathcal{A}. \end{aligned} \tag{B.1}$$

To show the far-right identity in (3.2), recall that every point $\text{conv}(\mathcal{A}) \subset \mathbb{R}^d$ can be expressed as a convex combination of at most $d+1$ atoms in the alphabet $\mathcal{A}$, by Carathéodory theorem [26]. We now use both (3.1) and the Carathéodory theorem to write that

$$\begin{aligned}\text{conv}_{d+1}(\mathcal{A}) &= \bigcup_{\{A_i\}_{i=1}^{d+1}\subset\mathcal{A}} \text{conv}(\{A_i\}_{i=1}^{d} \cup \{0\}) = \bigcup_{\{A_i\}_{i=1}^{d+1}\subset\mathcal{A}} \bigcup_{0\le\tau\le 1} \tau\cdot\text{conv}(\{A_i\}_{i=1}^{d+1}) \\ &= \bigcup_{0\le\tau\le 1} \bigcup_{\{A_i\}_{i=1}^{d+1}\subset\mathcal{A}} \tau\cdot\text{conv}(\{A_i\}_{i=1}^{d+1}) = \bigcup_{0\le\tau\le 1} \tau\cdot\text{conv}(\mathcal{A}) = \text{conv}(\mathcal{A}\cup\{0\}) = \text{conv}(\mathcal{A}),\end{aligned}$$

where the last line holds because $0 \in \mathcal{A}$ by Assumption 2.3(i). This establishes (3.2) and completes the proof of Proposition 3.2.

B.2. Proof of Proposition 3.4

To prove (3.4), we use the expression for $\text{conv}_p(\mathcal{A})$ in (3.1) to rewrite the definition of gauge_p function in (3.3) as

$$\begin{aligned}\mathcal{G}_{\mathcal{A},p}(x) &= \inf\left\{t : x/t \in \bigcup_{\mathcal{S}\in\text{slice}_r(\mathcal{A})} \mathcal{S},\, t\ge 0\right\} \qquad \text{(see (3.1,3.3))} \\ &= \inf\left\{t : x = \sum_{i=1}^{p} c_i A_i,\, \sum_{i=1}^{p} c_i \le t,\, c_i\ge 0, A_i\in\mathcal{A},\, \forall i\in[p]\right\} \\ &= \inf\left\{\sum_{i=1}^{p} c_i : x = \sum_{i=1}^{p} c_i A_i,\, c_i\ge 0, A_i\in\mathcal{A},\, \forall i\in[p]\right\},\end{aligned}$$

which proves (3.4). To show Proposition 3.4(ii), suppose that $\mathcal{G}_{\mathcal{A},p}(x) = 0$ which implies by definition in (3.3) that $x/t \in \text{conv}_p(\mathcal{A})$ for every $t > 0$. Since the alphabet $\mathcal{A}$ and, consequently, $\text{conv}_p(\mathcal{A})$ in (3.1) are both bounded by Assumption 2.3(iii), we conclude that $x = 0$.

To prove Proposition 3.4(iii), we begin by writing down the convex conjugate of $\mathcal{G}_{\mathcal{A},p}$ as

$$\begin{aligned}\mathcal{G}^*_{\mathcal{A},p}(z) &= \sup_x \langle x, z\rangle - \mathcal{G}_{\mathcal{A},p}(x) \\ &= \sup\left\{\langle x, z\rangle - \sum_{i=1}^{p} c_i : x = \sum_{i=1}^{p} c_i A_i,\, c_i\ge 0,\, A_i\in\mathcal{A},\, \forall i\in[p]\right\} \qquad \text{(see (3.4))}\end{aligned}$$

$$= \sup\left\{\sum_{i=1}^{p} c_i(\langle A_i, z\rangle - 1) : c_i \geq 0,\ A_i \in \mathcal{A},\ \forall i \in [p]\right\} = \begin{cases} 0 & \text{if } \sup_{A\in\mathcal{A}} \langle A, z\rangle \leq 1 \\ \infty & \text{if } \sup_{A\in\mathcal{A}} \langle A, z\rangle > 1 \end{cases}$$
$$= \text{indicator}_{\text{ball}(\mathcal{D}_{\mathcal{A}})}(z),$$

where $\text{ball}(\mathcal{D}_{\mathcal{A}})$ is the unit ball for the dual norm of the gauge function, i.e., $\mathcal{D}_{\mathcal{A}}$ in (2.4). It also immediately follows that

$$\mathcal{G}_{\mathcal{A},p}^{**} = (\text{indicator}_{\text{ball}(\mathcal{D}_{\mathcal{A}})})^* = \mathcal{G}_{\mathcal{A}},$$

which proves Proposition 3.4(iv). Proposition 3.4(v) and (vi) trivially follow from the definition of the gauge_p function in (3.4). Lastly, the nested property of the gauge_p functions in (3.6) follows immediately from (3.4). The identity on the far-right of (3.6) follows by combining the far-right identity in (3.2) with (3.3). This completes the proof of Proposition 3.4.

B.3. **Proof of Lemma 3.14**

Let $\widehat{x}$ be a minimizer of problem (gauge_p). Suppose that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) = 0$. By feasibility of $\widehat{x}$ in problem (gauge_p), it holds that $\mathcal{G}_{\mathcal{A},p}(\widehat{x}) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp) = 0$ and, consequently, $\widehat{x} = x^\sharp = 0$ by Proposition 3.4(ii). We thus assume that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0$ from now on.

By definition of the gauge_p function in (3.3), it holds that

$$x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \text{conv}_p(\mathcal{A}).$$

Again by definition of the gauge_p function and using also the definition of $\text{conv}_p(\mathcal{A})$ in (3.1), there exists a slice $\mathcal{S} \in \text{slice}_p(\mathcal{A})$ such that

$$\widehat{x}/\mathcal{G}_{\mathcal{A},p}(\widehat{x}) \in \mathcal{S} \subset \text{conv}_p(\mathcal{A}). \tag{B.2}$$

Moreover, since the slice $\mathcal{S}$ is a convex set containing the origin, see Definition 2.1, it follows from (B.2) that

$$t \cdot \widehat{x}/\mathcal{G}_{\mathcal{A},p}(\widehat{x}) \in \mathcal{S}, \qquad \forall t \in [0,1]. \tag{B.3}$$

By feasibility of $\widehat{x}$ in problem (gauge_p), we have that $\mathcal{G}_{\mathcal{A},p}(\widehat{x}) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp)$, and we can thus take $t = \mathcal{G}_{\mathcal{A},p}(\widehat{x})/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in [0,1]$ in (B.3) to find that

$$\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}\,. \tag{B.4}$$

For future reference, note also that the feasibility of $x^\sharp$ and optimality of $\widehat{x}$ in problem (gauge) implies that $\mathcal{L}(x^\sharp) = \mathcal{L}(\widehat{x}) = y$ and, consequently,

$$\mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right) = \mathcal{L}(\widehat{x} - x^\sharp) = 0. \qquad (\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0) \tag{B.5}$$

We can now proceed to the body of the proof by considering two cases:

(1) Suppose that $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}$. Then it follows from (B.2) that

$$\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \in \mathcal{S} - \mathcal{S} \subset \text{lin}(\mathcal{S}). \qquad (\text{see (B.2)}) \tag{B.6}$$

Then, (B.5,B.6) together with the injectivity of $\mathcal{L}$ on $\text{lin}(\mathcal{S})$ imply that $\widehat{x} = x^\sharp$.

(2) Suppose that $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \notin \mathcal{S}$. It therefore exists a certificate $Q_\mathcal{S} = \mathcal{L}^*(q_\mathcal{S})$ that satisfies (3.19). With this in mind and after recalling (B.5), we write that

$$\begin{aligned} 0 &= \left\langle q_\mathcal{S}, \mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right)\right\rangle = \left\langle \mathcal{L}^*(q_\mathcal{S}), \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right\rangle \\ &= \left\langle Q_\mathcal{S}, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right\rangle < 0 \end{aligned}$$

which leads to a contradiction. We conclude that $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}$ and, consequently, $\widehat{x} = x^\sharp$.

This completes the proof of Lemma 3.14.

B.4. **Example of a Near-Isometry Operator**

Let $G \in \mathbb{R}^{m\times d}$ be a standard random Gaussian matrix, i.e., the entries of G are independent Gaussian random variables with zero mean and unit variance. Consider a (linear) subspace $\mathcal{U} \subset \mathbb{R}^d$ and let the $d \times \dim(\mathcal{U})$ matrix U be an orthonormal basis for the span of this subspace. We then write that

$$\sup_{u\in\mathcal{U}} \frac{\|Gu\|_2}{\|u\|_2} = \sup_v \frac{\|GUv\|_2}{\|v\|_2} = \sigma_{\max}(GU) =: \sigma_{\max}(G'), \tag{B.7}$$

where we set $G' = GU$ for short, and $\sigma_{\max}(G')$ is the largest singular value of G'. Likewise,

$$\inf_{u\in\mathcal{U}} \frac{\|Gu\|_2}{\|u\|_2} = \inf_v \frac{\|GUv\|_2}{\|v\|_2} = \sigma_{\min}(GU) = \sigma_{\min}(G').$$

Note that the $m \times \dim(\mathcal{U})$ matrix $G' = GU$ too is a standard random Gaussian matrix because $U^\top U = I_{\dim(\mathcal{U})}$ by construction. The largest and smallest singular values of a standard random Gaussian matrix are well-known [65, Corollary 5.35]. In particular, it holds that

$$(1-\delta')\sqrt{m} \le \sigma_{\min}(G') \le \sigma_{\max}(G') \le (1+\delta')\sqrt{m}, \tag{B.8}$$

provided that $m \ge C\dim(\mathcal{U})/\delta'^2$ and except with a probability of at most $\exp(-C'\delta'^2 m)$, for universal constants C, C'. By combining (B.7-B.8) for the linear operator $\mathcal{L} = G/\sqrt{m}$, we arrive at

$$(1-\delta')\|u\|_2 \le \|\mathcal{L}(u)\|_2 \le (1+\delta')\|u\|_2, \qquad \forall u \in \mathcal{U},$$

provided that $m \ge C\dim(\mathcal{U})/\delta'^2$ and except with a probability of at most $\exp(-C'\delta'^2 m)$. The random linear operator $\mathcal{L}$ constructed above thus a δ'-near-isometry.

B.5. **Proof of Proposition 3.22**

Since $p \ge d+1$ by assumption, recall from (3.6) that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) = \mathcal{G}_\mathcal{A}(x^\sharp)$. In particular, $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) = x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp) =: \widetilde{x}$. For a slice $\mathcal{S} \in \mathrm{slice}_p(\mathcal{A})$, note that

$$\begin{aligned} \mathcal{S} &\subset \mathrm{conv}_p(\mathcal{A}) \qquad \text{(see (3.1))} \\ &= \mathrm{conv}(\mathcal{A}), \qquad \text{(see (3.2))} \end{aligned} \tag{B.9}$$

where the identity above holds by (3.2) and because $p \ge d+1$. Let us assume that $\widetilde{x} \notin \mathcal{S}$. An immediate implication of (B.9) is that $\mathcal{S} - \widetilde{x} \subset \mathrm{conv}(\mathcal{A}) - \widetilde{x}$ and, consequently,

$$\mathrm{cone}(\mathcal{S} - \widetilde{x}) \subset \mathrm{cone}(\mathrm{conv}(\mathcal{A}) - \widetilde{x}) = \mathrm{cone}(\mathrm{conv}(\mathcal{A} - \widetilde{x})) = \mathrm{cone}(\mathcal{A} - \widetilde{x}),$$

and, in turn, $\angle\text{cone}(\mathcal{S}-\widetilde{x}) \le 2\angle\text{cone}(\mathcal{A}-\widetilde{x})$, where we invoked Lemma B.2 below to obtain the last inequality. Using this last inequality and (3.27), we arrive at

$$\theta_{x^\sharp,p}(\mathcal{A}) = \sup\left\{\angle\text{cone}\left(\mathcal{S}-\widetilde{x}\right) : \mathcal{S} \in \text{slice}_{x^\sharp,p}(\mathcal{A})\right\} \le 2\angle\text{cone}\left(\mathcal{A}-\widetilde{x}\right),$$

which completes the proof of Proposition 3.22.

To prove Lemma B.2 below, in addition to (3.23), we first introduce two other notions of angle for a closed cone $\mathcal{K} \subset \mathbb{R}^d$, i.e.,

$$\begin{aligned}\cos(\phi(\mathcal{K})) &:= \max_{u\in\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}\cap\mathbb{S}^{d-1}} \langle u, u'\rangle,\\ \cos(\psi(\mathcal{K})) &:= \min_{u\in\mathcal{K}\cap\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}\cap\mathbb{S}^{d-1}} \langle u, u'\rangle.\end{aligned} \tag{B.10}$$

These quantities are related as follows.

Lemma B.1. *For a closed cone $\mathcal{K}$, it holds that*

$$\frac{\psi(\mathcal{K})}{2} \le \phi(\mathcal{K}) \le \angle\mathcal{K} \le \psi(\mathcal{K}). \tag{B.11}$$

Proof. Note that

$$\begin{aligned}\cos(\psi(\mathcal{K})) &= \min_{u\in\mathcal{K}\cap\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}\cap\mathbb{S}^{d-1}} \langle u, u'\rangle \qquad \text{(see (B.10))}\\ &\le \max_{u\in\mathcal{K}\cap\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}\cap\mathbb{S}^{d-1}} \langle u, u'\rangle = \cos(\angle\mathcal{K}) \qquad \text{(see (3.23))}\\ &\le \max_{u\in\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}\cap\mathbb{S}^{d-1}} \langle u, u'\rangle = \cos(\phi(\mathcal{K})),\end{aligned}$$

which establishes the last two inequalities in (B.11). On the other hand, note that $\phi(\mathcal{K})$ is the angle of the smallest spherical cap that contains $\mathcal{K}\cap\mathbb{S}^{d-1}$, whereas $\psi(\mathcal{K})$ is the largest pairwise angle in $\mathcal{K}$. The two quantities are thus related as

$$\psi(\mathcal{K}) \le 2\phi(\mathcal{K}),$$

which proves the remaining inequality in (B.11), and completes the proof of Lemma B.1. □

We next prove a weak inclusion result for cones.

Lemma B.2. *For closed cones $\mathcal{K}_1 \subset \mathcal{K}_2$, it holds that $\angle\mathcal{K}_1 \le 2\angle\mathcal{K}_2$.*

Proof. Note that by (B.10) and (B.11) we have

$$\begin{aligned}\cos(\angle\mathcal{K}_1) \ge \cos(\psi(\mathcal{K}_1)) &= \min_{u\in\mathcal{K}_1\cap\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}_1\cap\mathbb{S}^{d-1}} \langle u, u'\rangle\\ &\ge \min_{u\in\mathcal{K}_2\cap\mathbb{S}^{d-1}} \min_{u'\in\mathcal{K}_2\cap\mathbb{S}^{d-1}} \langle u, u'\rangle = \cos(\psi(\mathcal{K}_2)) \ge \cos(2\angle\mathcal{K}_2),\end{aligned}$$

which completes the proof of Lemma B.2. □

In general, Lemma B.2 cannot be improved. For example, consider the example $\mathcal{K}_1 = \text{cone}(\{e_1, e_2\})$ and $\mathcal{K}_2 = \text{cone}(\{e_1, e_2, e_1+e_2\})$. It is easy to verify that $\angle\mathcal{K}_1 = \pi/2$ whereas $\angle\mathcal{K}_2 = \pi/4$, and thus the inequality in Lemma B.2 holds with equality. However, we strongly suspect that Lemma B.2 could be improved to $\angle\mathcal{K}_1 \le \angle\mathcal{K}_2$ when $\mathcal{K}_1$ is a closed convex cone.

B.6. **Proof of Theorem 3.23**

Throughout the proof, we frequently use the shorthand

$$\widetilde{x} := \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}, \qquad u_x := \frac{x - \widetilde{x}}{\|x - \widetilde{x}\|_2}, \tag{B.12}$$

for $x \neq \widetilde{x}$. As in Lemma 3.14, we assumed above without loss of generality that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0$. The proof of Theorem 3.23 relies on the following technical result, which is proved similar to [81, Lemma 2.1].

Lemma B.3. *For $\delta' \in [0,1)$, suppose that the random linear operator $\mathcal{L}$ is a δ'-near-isometry. Then, for a slice $\mathcal{S}' \in \mathrm{slice}_p(\mathcal{A})$, it holds that*

$$\frac{\langle Q_{\mathcal{S}'}, u_{x'} \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} \leq -\cos(\angle \mathrm{cone}(\mathcal{S}' - \widetilde{x})) + \delta', \qquad \forall x' \in \mathcal{S}', \tag{B.13}$$

provided that $m \geq Cp/\delta'^2$, and except with a probability of $\exp(-C'\delta'^2 m)$, where

$$Q_{\mathcal{S}'} = \mathcal{L}^*(\mathcal{L}(\widetilde{x} - x_{\mathcal{S}'})), \tag{B.14}$$

and $x_{\mathcal{S}'}$ above is selected such that

$$\cos(\angle \mathrm{cone}(\mathcal{S}' - \widetilde{x})) = \min_{x' \in \mathcal{S}'} \left\langle u_{x_{\mathcal{S}'}}, u_{x'} \right\rangle. \tag{B.15}$$

Before proving Lemma B.3 in Section B.7, we first complete the proof of Theorem 3.23. Recall that $\mathrm{slice}_p(\mathcal{A})$ in (3.1) denotes the set of all slices of $\mathrm{conv}(\mathcal{A})$ formed by at most p atoms. For a resolution $\delta > 0$, let $\mathrm{cover}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ denote a minimal δ-cover for $\mathrm{slice}_p(\mathcal{A})$ with respect to the pseudo-metric dist, specified as

$$\mathrm{dist}_p(\mathcal{S}, \mathcal{S}') = \sqrt{2 - 2\cos(\angle[\mathrm{cone}(\mathcal{S} - \widetilde{x}), \mathrm{cone}(\mathcal{S}' - \widetilde{x})])}, \qquad \forall \mathcal{S}, \mathcal{S}' \in \mathrm{slice}_p(\mathcal{A}). \tag{B.16}$$

Indeed, dist above is a pseudo-metric, as it coincides with the Hausdorff distance between the intersection of the cones on the right-hand side above and the unit sphere in $\mathbb{R}^d$, see (3.24). For $\delta' \in [0,1)$, suppose that the random linear operator $\mathcal{L}$ is a δ'-near-isometry, see (3.25). Consequently, by applying the union bound to all slices in the cover, we find that

$$\begin{aligned} &(1-\delta')\|u\|_2 \leq \|\mathcal{L}(u)\|_2 \leq (1+\delta')\|u\|_2, \\ &\qquad \forall u \in \mathrm{lin}(\mathcal{S}' - \widetilde{x}), \qquad \forall \mathcal{S}' \in \mathrm{cover}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta), \end{aligned} \tag{B.17}$$

provided that $m \geq Cp/\delta'^2$ and except with a probability of at most $\exp(-C'\delta'^2 m + \mathrm{entropy}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta'))$, where we used the fact that $\mathrm{cover}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ is a minimal cover by construction and thus has the size of $\exp(\mathrm{entropy}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta'))$.

Consider an arbitrary slice $\mathcal{S} \in \mathrm{slice}_p(\mathcal{A})$. By Definition 3.19, there exists another slice $\mathcal{S}' \in \mathrm{cover}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ such that $\mathrm{dist}_p(\mathcal{S}, \mathcal{S}') \leq \delta$, which implies that

$$\cos(\angle[\mathrm{cone}(\mathcal{S} - \widetilde{x}), \mathrm{cone}(\mathcal{S}' - \widetilde{x})]) \geq 1 - \frac{\delta^2}{2}. \qquad \text{(see (B.16))} \tag{B.18}$$

Consider also an arbitrary point $x \in \mathcal{S}$ and the corresponding unit-norm vector u_x, see (B.12). By (B.18) and after recalling the definition of angle between cones in (3.24), there exists a unit-norm vector $u' \in \text{cone}(\mathcal{S}' - \widetilde{x})$ such that

$$\langle u_x, u' \rangle \geq 1 - \frac{\delta^2}{2}, \tag{B.19}$$

which also immediately implies that

$$\begin{aligned} \|u_x - u'\|_2^2 &= \|u_x\|_2^2 + \|u'\|_2^2 - 2\langle u_x, u' \rangle \\ &\leq 2 - 2\left(1 - \frac{\delta^2}{2}\right) \leq \delta^2, \qquad \text{(see (B.12), (B.19))} \end{aligned} \tag{B.20}$$

which is also easy to arrive at from the Hausdorff distance interpretation discussed earlier. We next distinguish two cases:

(1) Suppose that $\widetilde{x} \in \mathcal{S}$. Without loss of generality, we may assume that the ray passing through $\widetilde{x}$ belongs to the interior of $\text{cone}(\mathcal{S})$. (Indeed, otherwise there exists a lower-dimensional slice, to the interior of which the ray passing through $\widetilde{x}$ belongs.) Consequently, from definition of $\widetilde{x}$ in (B.12), it follows that

$$\text{cone}(\mathcal{S} - \widetilde{x}) \cup -\text{cone}(\mathcal{S} - \widetilde{x}) = \text{lin}(\mathcal{S} - \widetilde{x}). \tag{B.21}$$

Moreover, note that

$$\text{lin}(\mathcal{S} - \widetilde{x}) = \text{lin}(\mathcal{S}), \tag{B.22}$$

where the above identity holds because $0 \in \mathcal{S}$ by the definition of slice in Definition 2.1. On the other hand, we observe that

$$\begin{aligned} \|\mathcal{L}(u_x)\|_2 &\geq \|\mathcal{L}(u')\|_2 - \|\mathcal{L}(u_x - u')\|_2 \qquad \text{(triangle inequality)} \\ &\geq 1 - \delta' - \|\mathcal{L}\|_{\text{op}} \cdot \|u_x - u'\|_2 \qquad \text{(see (B.17))} \\ &= 1 - \delta' - \|\mathcal{L}\|_{\text{op}} \cdot \delta > 0, \qquad \text{(see (B.20))} \end{aligned} \tag{B.23}$$

where the last line above holds if $\delta' + \|\mathcal{L}\|_{\text{op}} \cdot \delta < 1$. Since the choice of the point $x \in \mathcal{S}$ in (B.23) was arbitrary, we conclude that $\mathcal{L}$ is an injective operator when restricted to $\text{cone}(\mathcal{S} - \widetilde{x})$ and, consequently, also when restricted to $-\text{cone}(\mathcal{S} - \widetilde{x})$. In view of (B.21), $\mathcal{L}$ is also injective when restricted to $\text{lin}(\mathcal{S} - \widetilde{x})$ and, by (B.22), when restricted to $\text{lin}(\mathcal{S})$.

(2) Suppose that $\widetilde{x} \notin \mathcal{S}$. Recall $Q_{\mathcal{S}'}$ defined in (B.14) and note that

$$\begin{aligned} \frac{\langle Q_{\mathcal{S}'}, u_x \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} &= \frac{\langle \mathcal{L}^*(\mathcal{L}(\widetilde{x} - x_{\mathcal{S}'})), u_x \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} \qquad \text{(see (B.14))} \\ &= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u_x) \rangle \qquad \text{(see (B.12))} \\ &= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u') \rangle + \langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u' - u_x) \rangle \\ &= \frac{\langle Q_{\mathcal{S}'}, u' \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} + \langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u' - u_x) \rangle \\ &\leq \frac{\langle Q_{\mathcal{S}'}, u' \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} + \|\mathcal{L}\|_{\text{op}}^2 \cdot \|u_{x_{\mathcal{S}'}}\|_2 \cdot \|u' - u_x\|_2 \\ &\leq -\cos(\text{cone}(\mathcal{S}' - \widetilde{x})) + \delta' + \|\mathcal{L}\|_{\text{op}}^2 \cdot \delta, \qquad \text{(see (B.13), (B.20)} \end{aligned}$$

$$\leq -\inf_{\mathcal{S}''\in \mathrm{slice}_{x^\sharp,p}(\mathcal{A})} \cos(\mathrm{cone}(\mathcal{S}''-\widetilde{x})) + \delta' + \|\mathcal{L}\|_{\mathrm{op}}^2\delta$$
$$= -\cos(\theta_{x^\sharp,p}(\mathcal{A})) + \delta' + \|\mathcal{L}\|_{\mathrm{op}}^2\delta < 0, \qquad \text{(see (3.27))}$$

where the last line above holds if

$$\delta' + \|\mathcal{L}\|_{\mathrm{op}}^2\delta < \cos(\theta_{x^\sharp,p}(\mathcal{A})). \tag{B.24}$$

Since the choice of the point $x \in \mathcal{S}$ above was arbitrary, we thus established that there exists $Q_{\mathcal{S}} \in \mathrm{range}(\mathcal{L}^*)$ such that

$$\langle Q_{\mathcal{S}}, x-\widetilde{x}\rangle < 0, \qquad \forall x \in \mathcal{S}.$$

We may now invoke Lemma 3.14 to conclude that $x^\sharp$ is the unique minimizer of problem (gauge$_p$). This completes the proof of Theorem 3.23.

B.7. Proof of Lemma B.3

For $x' \in \mathcal{S}'$, we write that

$$\begin{aligned}
&\frac{\langle Q_{\mathcal{S}'}, x'-\widetilde{x}\rangle}{\|x_{\mathcal{S}'}-\widetilde{x}\|_2\|x'-\widetilde{x}\|_2} \\
&= \frac{\langle \mathcal{L}^*(\mathcal{L}(\widetilde{x}-x_{\mathcal{S}'})), x'-\widetilde{x}\rangle}{\|x_{\mathcal{S}'}-\widetilde{x}\|_2\|x'-\widetilde{x}\|_2} \qquad \text{(see (B.14))} \\
&= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u_{x'})\rangle \qquad \text{(see (B.12))} \\
&= \frac{-1}{4}(\|\mathcal{L}(u_{x_{\mathcal{S}'}}+u_{x'})\|_2^2 - \|\mathcal{L}(u_{x_{\mathcal{S}'}}-u_{x'})\|_2^2) \qquad \text{(parallelogram identity)} \\
&\leq -\frac{1}{4}((1-\delta')\|u_{x_{\mathcal{S}'}}+u_x\|_2^2 - (1+\delta')\|u_{x_{\mathcal{S}'}}-u_{x'}\|_2^2) \qquad (\mathcal{L} \text{ is a } \delta'\text{-near-isometry, see (3.25)}) \\
&= -\frac{1}{4}(\|u_{x_{\mathcal{S}'}}+u_{x'}\|_2^2 - \|u_{x_{\mathcal{S}'}}-u_{x'}\|_2^2) + \frac{\delta'}{4}(\|u_{x_{\mathcal{S}'}}+u_{x'}\|_2^2 + \|u_{x_{\mathcal{S}'}}-u_{x'}\|_2^2) \\
&= -\langle u_{x_{\mathcal{S}'}}, u_{x'}\rangle + \frac{\delta'}{2}(\|u_{x_{\mathcal{S}'}}\|_2^2 + \|u_{x'}\|_2^2) \\
&= -\langle u_{x_{\mathcal{S}'}}, u_{x'}\rangle + \delta' \qquad \text{(see (B.12))} \\
&\leq -\min_{u_{x''}\in\mathcal{S}'} \langle u_{x_{\mathcal{S}'}}, u_{x''}\rangle + \delta' \\
&= -\max_{u_{x'}\in\mathcal{S}'}\min_{u_{x''}\in\mathcal{S}'} \langle u_{x'}, u_{x''}\rangle + \delta' \qquad \text{(see (3.23,B.15))} \\
&= -\cos(\angle\mathrm{cone}(\mathcal{S}'-\widetilde{x})) + \delta', \qquad \text{(see (3.23))}
\end{aligned} \tag{B.25}$$

which completes the proof of Lemma B.3.

B.8. Corollary 3.3.1 in [1]

For completeness, below we review Corollary 3.3.1 in [1], adapted to our notation.

Corollary B.4. *Suppose that the alphabet $\mathcal{A} \subset \mathbb{R}^d$ is a compact set and that (2.5) holds with equality, i.e., $\mathrm{ext}(\mathrm{conv}(\mathcal{A})) = \mathcal{A}$. Consider the model $x^\sharp \in \mathbb{R}^d$ in (exact) and let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ be*

the linear map associated with the $m \times d$ Gaussian random matrix, populated with independent and zero-mean normal random variables with the variance of $1/m$. Then the learning machine

$$\min_x \ \mathcal{G}_{\mathcal{A}}(x) \ \text{subject to} \ y = \mathcal{L}(x),$$

returns $x^\sharp$, provided that $m \geq w(\Omega)^2 + 1$, and except with a probability of at most $\exp(-C(\sqrt{m} - w(\Omega))^2)$, for a universal constant C. Above, we set $\Omega = \operatorname{cone}(\mathcal{A} - x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp)) \cap \mathbb{S}^{d-1}$ to be the intersection of the unit sphere with the tangent cone of $\operatorname{conv}(\mathcal{A})$ at $x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp)$, and $w(\Omega)$ is the Gaussian width of Ω, i.e.,

$$w(\Omega) = \mathbb{E}_g \left[\sup_{x \in \Omega} \langle g, x \rangle \right],$$

where $g \in \mathbb{R}^d$ is a standard Gaussian random vector, i.e., populated with independent, zero-mean and unit-variance normal random variables.

B.9. **Proof of Proposition 3.16**

Since Lemma 2.6 is in force, the machine (3.21) returns $x^\sharp$. Moreover, by invoking Lemma A.1, we find that the decomposition of $x^\sharp$ in the slice $\mathcal{S}^\sharp$ is minimal for the gauge function $\mathcal{G}_{\mathcal{A}}$. To complete the proof of Proposition 3.16, it remains to establish the same claim for the machine (3.22).

As in the proof of Lemma 3.14, we assume without loss of generality that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0$. Recall from model (3.20) that $x^\sharp \in \operatorname{cone}(\mathcal{S}^\sharp)$. From the definition of the gauge$_p$ function in (3.3), it then follows that

$$x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}^\sharp = \operatorname{conv}(\mathcal{F}^\sharp \cup \{0\}), \tag{B.26}$$

where the identity above follows from the assumption made in Lemma 2.6. In fact, we may apply Lemma A.1 to strengthen (B.26) as

$$x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{F}^\sharp. \tag{B.27}$$

Recall also from (B.4) that

$$\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \operatorname{conv}_p(\mathcal{A}) \subset \operatorname{conv}(\mathcal{A}), \tag{B.28}$$

where the inclusion above follows from (3.2). By optimality of $\widehat{x}$ in problem (3.22), it also holds that $\mathcal{L}(\widehat{x}) = \mathcal{L}(x^\sharp) = y$ and, consequently,

$$\mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right) = \mathcal{L}(\widehat{x} - x^\sharp) = 0. \qquad (\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0) \tag{B.29}$$

We again consider two cases:

(1) Suppose that $\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{F}^\sharp$. Then it follows from (B.27) that

$$\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \in \mathcal{F}^\sharp - \mathcal{F}^\sharp \subset \operatorname{lin}(\mathcal{F}^\sharp). \tag{B.30}$$

By combining (B.29) and (B.30), and using the injectivity of $\mathcal{L}$ on $\operatorname{lin}(\mathcal{F}^\sharp)$, we conclude that $\widehat{x} = x^\sharp$.

(2) Suppose that $\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \notin \mathcal{F}^\sharp$. In combination with (B.28), we then find that

$$\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathrm{conv}(\mathcal{A}) - \mathcal{F}^\sharp. \tag{B.31}$$

Since Lemma 2.6 is in force, there exists a dual certificate $Q = \mathcal{L}^*(q)$ that satisfies (2.6). With this in mind and after recalling (B.29), we write that

$$\begin{aligned} 0 &= \left\langle q, \mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right)\right\rangle \qquad \text{(see (B.29))} \\ &= \left\langle Q, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right\rangle \qquad (Q = \mathcal{L}^*(q)) \\ &< 0, \qquad \text{(see (2.6) and (B.31))} \end{aligned}$$

which is a contradiction. We conclude that $\widehat{x}/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{F}^\sharp$ and, consequently, $\widehat{x} = x^\sharp$.

We have thus far established that machine (3.22) returns $x^\sharp$. We now complete the proof of Proposition 3.16 by presenting a version of Lemma A.1 for the gauge$_p$ function.

Lemma B.5. *Suppose that Assumption 2.3(iii) on the alphabet $\mathcal{A} \subset \mathbb{R}^d$ is fulfilled. Consider the model $x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp)$ in (exact) and let $\mathcal{F}^\sharp$ be an exposed face of $\mathrm{conv}(\mathcal{A})$ such that $x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{F}^\sharp$. Suppose also that $\mathcal{S}^\sharp = \mathrm{conv}(\mathcal{F}^\sharp \cup \{0\})$. Then $x^\sharp$ has a minimal decomposition in $\mathcal{F}^\sharp \subset \mathcal{S}^\sharp$ that achieves the gauge$_p$ function value $\mathcal{G}_{\mathcal{A},p}(x^\sharp)$, see (3.4).*

Proof. Note that $x^\sharp \in \mathrm{cone}(\mathcal{F}^\sharp) = \mathrm{cone}(\mathcal{S}^\sharp)$ by assumption. Using the definition of the gauge$_p$ function in (3.4), it follows that

$$\begin{aligned} \mathcal{G}_{\mathcal{A},p}(x^\sharp) &= \inf\left\{t : x^\sharp/t \in \mathrm{conv}_p(\mathcal{A})\right\} = \inf\left\{t : x^\sharp/t \in \mathcal{S}^\sharp\right\} = \inf\left\{t : x^\sharp/t \in \mathrm{conv}(\mathcal{F}^\sharp \cup \{0\})\right\} \\ &= \inf\left\{t : x^\sharp/t \in t'\mathcal{F}^\sharp,\, t' \in [0,1]\right\} = \inf\left\{t : x^\sharp/t \in \mathcal{F}^\sharp\right\}, \end{aligned}$$

where the third line above uses the assumption that $\mathcal{S}^\sharp = \mathrm{conv}(\mathcal{F}^\sharp \cup \{0\})$. Consequently, $x^\sharp \in \mathcal{F}^\sharp \subset \mathcal{S}^\sharp$ is a minimal decomposition of $x^\sharp$ that achieves the gauge$_p$ function value $\mathcal{G}_{\mathcal{A},p}(x^\sharp)$. This completes the proof of Lemma B.5. □

Appendix C. Technical Details of Section 4

C.1. Proof of Corollary 4.1

First note that

$$\mathcal{G}_{\mathcal{A},1}(A^\sharp) = 1. \tag{C.1}$$

Indeed, otherwise if $\mathcal{G}_{\mathcal{A},1}(A^\sharp) < 1$, then there would exist an atom $A' \in \mathcal{A}$ aligned with $A^\sharp$ with $\|A'\|_2 > \|A^\sharp\|_2 = 1$, which contradicts the assumption that $\angle[A - A^\sharp, A^\sharp] \geq \theta_0 > 0$ for every atom $A \in \mathcal{A}$. Moreover, for atoms $A, A' \in \mathcal{A}$, consider the corresponding one-dimensional slices $\mathcal{S}, \mathcal{S}' \in \mathrm{slice}_1(\mathcal{A})$, specified as

$$\mathcal{S} = \bigcup_{0 \leq \tau \leq 1} \tau A, \qquad \mathcal{S}' = \bigcup_{0 \leq \tau \leq 1} \tau A',$$

which are simply the line segments connecting A and A' to the origin. With $p = 1$, the distance in (3.28) is

$$\begin{aligned}\mathrm{dist}_1(\mathcal{S},\mathcal{S}') &= \sqrt{2-2\cos\left(\angle\left[\mathrm{cone}(\mathcal{S}-A^\sharp),\mathrm{cone}(\mathcal{S}'-A^\sharp)\right]\right)} \qquad \text{(see (3.28,C.1))}\\ &= \mathrm{dist}_{\mathrm{H}}\left(\mathrm{cone}\left(\mathcal{S}-A^\sharp\right)\cap\mathbb{S}^{d-1},\mathrm{cone}(\mathcal{S}'-A^\sharp)\cap\mathbb{S}^{d-1}\right) \qquad \text{(see (3.24))}\\ &\le \left\|\frac{A-A^\sharp}{\|A-A^\sharp\|_2}-\frac{A'-A^\sharp}{\|A'-A^\sharp\|_2}\right\| =: \mathrm{dist}_{A^\sharp}(A,A'). \end{aligned} \tag{C.2}$$

with the convention that $0/0 = 0$. Since each one-dimensional slice can be identified with its corresponding atom, it immediately follows that

$$\mathrm{entropy}(\mathrm{slice}_1(\mathcal{A}),\mathrm{dist}_1,\delta) = \mathrm{entropy}(\mathcal{A},\mathrm{dist}_{A^\sharp},\delta).$$

On the other hand, the critical angle in (3.27) satisfies

$$\begin{aligned}\theta_{1,A^\sharp} &= \sup\left\{\frac{1}{2}\angle[A-A^\sharp,-A^\sharp] : A\in\mathcal{A}-\{A^\sharp\}\right\} \qquad \text{(see (3.27))}\\ &= \sup\left\{\frac{\pi}{2}-\frac{1}{2}\angle[A-A^\sharp,A^\sharp] : A\in\mathcal{A}-\{A^\sharp\}\right\} \le \frac{\pi-\theta_0}{2}, \end{aligned} \tag{C.3}$$

where the last line above uses the assumption on the alphabet $\mathcal{A}$. In view of (C.2) and (C.3), we may now invoke Theorem 3.23 to complete the proof of Corollary 4.1.

Appendix D. Technical Details of Section 5

D.1. Proof of Lemma 5.1

It is straightforward to see that the continuous variables $c_i = 0$ if and only if the binary variables $s_i = 0$. Therefore, the constraint $\sum_{i=1}^{l} s_i = p$ of the binary variables directly imposes the required p-sparsity condition on c.

D.2. Proof of Proposition 5.2

First, observe that the machine (gauge$_p$) (or equivalently the MIQP reformulation (5.1)) can be rewritten as

$$\min_{c}\left\{\|\mathcal{L}(A)c-y\|_2^2 \;:\; c\ge 0,\quad \mathbb{1}^\top c\le \mathcal{G}_{\mathcal{A},p}(x^\sharp),\quad \|c\|_0\le p\right\}.$$

Above, as usual, $\|c\|_0$ denotes the number of nonzero entries of the vector c. One can encode the nonzero elements of the vector c as a subset $S\subset[|\mathcal{A}|]$. That is, we can introduce a new vector c_S such that $(c_S)_i = (c)_i$, when $i\in S$, otherwise $[c_S]_i = 0$. In this way, the above optimization program can be rewritten as

$$\min_{\substack{S\subset[|\mathcal{A}|]\\ |S|\le p}}\ \min_{c_S,z}\left\{\|z\|_2^2 \;:\; z=\mathcal{L}(A)c_S-y,\quad Cc_S\le g\right\}, \tag{D.1}$$

where the matrix C and the vector g were defined in the proposition. Note that the inner optimization program in (D.1) is indeed a convex quadratic programming. This observation allows us to claim two things: (i) We can add an additional term $\|c_S\|_2^2/\gamma$ in the objective function where

for all sufficiently large γ the optimal solution does not change. Indeed, the objective value of the convex inner problem does not change by adding the constraint $\|c_S\|_2 \leq \epsilon$ for a sufficiently large ϵ. By convexity of the inner problem, this is equivalent to adding the penalty term $\|c_S\|_2^2/\gamma$ for a sufficiently large γ. (ii) Thanks to the convexity, we can dualize the linear constraints and arrive at the equivalent optimization program

$$\min_{\substack{S \subset [|\mathcal{A}|] \\ |S| \leq p}} \max_{\mu \geq 0, \lambda} \min_{c_S, z} \Big(\|z\|_2^2 + \frac{1}{\gamma}\|c_S\|_2^2 + \lambda^\top \big(\mathcal{L}(A)c_S - y - z\big) + \mu^\top \big(Cc_S - g\big) \Big).$$

Note that the most inner minimization above is an unconstrained convex quadratic program. Computing the analytical solution for the variables (c_S, z) yields the desired program (5.2). With regards to the algorithm described through the dynamics (5.3), first observe that the relation (5.3b) is the same as the maximizer of the objective function (5.2) when the set is fixed to S_k. Note further that the fixed point of (5.3) is indeed a saddle-point equilibrium for the zero-sum game between the player S and (μ, λ). Therefore, the equilibrium $S^\star$ is in fact also a "policy security", i.e., the pair is the solution to the minimax program (5.2) and its dual when the order of the minimization and maximization operators are changed [82, Proposition 4.2].

References

[1] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[2] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[3] R.T. Rockafellar. *Convex Analysis: (PMS-28)*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 2015.

[4] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[5] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[6] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[7] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

[8] Emmanuel Candes and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

[9] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

[10] Parikshit Shah, Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Linear system identification via atomic norm regularization. In *2012 IEEE 51st IEEE conference on decision and control (CDC)*, pages 6265–6270. IEEE, 2012.

[11] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[12] Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

[13] Emile Richard, Guillaume R Obozinski, and Jean-Philippe Vert. Tight convex relaxations for sparse matrix factorization. In *Advances in neural information processing systems*, pages 3284–3292, 2014.

[14] Geoffrey Schiebinger, Elina Robeva, and Benjamin Recht. Superresolution without separation. *Information and Inference: A Journal of the IMA*, 7(1):1–30, 2018.

[15] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. A super-resolution framework for tensor decomposition. *arXiv preprint arXiv:1602.08614*, 2016.

[16] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[17] Gabriel Peyré. Manifold models for signals and images. *Computer vision and image understanding*, 113(2):249–260, 2009.

[18] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[19] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[20] Armin Eftekhari and Konstantinos Zygalakis. Implicit regularization in matrix sensing: A geometric view leads to stronger results. *arXiv preprint arXiv:2008.12091*, 2020.

[21] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[22] Armin Eftekhari and Michael B Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.

[23] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[24] Dimitris Bertsimas, Bart Van Parys, et al. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, 48(1):300–323, 2020.

[25] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, pages 813–852, 2016.

[26] A. Barvinok. *A Course in Convexity*. Graduate studies in mathematics. American Mathematical Society, 2002.

[27] Yingjie Bi and Ao Tang. Refined shapley-folkman lemma and its application in duality gap estimation. *arXiv preprint arXiv:1610.05416*, 2016.

[28] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[29] A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[30] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[31] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[32] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang. Overcomplete tensor decomposition via convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 53–56. IEEE, 2015.

[33] Robin Vujanic, Peyman Mohajerin Esfahani, Paul J Goulart, Sébastien Mariéthoz, and Manfred Morari. A decomposition method for large scale MILPs, with performance guarantees and a power system application. *Automatica*, 67:144–156, 2016.

[34] Alexander Barvinok. Convexity of the image of a quadratic map via the relative entropy distance. *Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry*, 55(2):577–593, 2014.

[35] Jean-Pierre Aubin and Ivar Ekeland. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3):225–245, 1976.

[36] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier Science, 2008.

[37] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[38] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[39] William J Studden. *Tchebycheff systems: with applications in analysis and statistics*. Wiley, 1966.

[40] Armin Eftekhari, Jared Tanner, Andrew Thompson, Bogdan Toader, and Hemant Tyagi. Sparse non-negative super-resolution—simplified and stabilised. *Applied and Computational Harmonic Analysis*, 2019.

[41] Armin Eftekhari, Tamir Bendory, and Gongguo Tang. Stable super-resolution of images: A theoretical study. *arXiv preprint arXiv:1805.09513*, 2018.

[42] Junzhou Huang, Tong Zhang, et al. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[43] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

[44] Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48, 2005.

[45] Emmanuel J Candes, Yonina C Eldar, Deanna Needell, and Paige Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[46] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.

[47] Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

[48] NICOLAS Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 2020.

[49] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[50] Armin Eftekhari and Michael B Wakin. Greed is super: A fast algorithm for super-resolution. *arXiv preprint arXiv:1511.03385*, 2015.

[51] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[52] Thomas Kerdreux, Igor Colin, and Alexandre d'Aspremont. An approximate shapley-folkman theorem. *arXiv preprint arXiv:1712.08559*, 2017.

[53] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[54] Gilles Pisier. Remarks on an unpublished result 'e of b. maurey. *S 'e minaire Functional analysis (called "Maurey-Schwartz")*, pages 1–12, 1981.

[55] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems*, 25:1457–1465, 2012.

[56] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[57] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. Spectral k-support norm regularization. In *Advances in neural information processing systems*, pages 3644–3652, 2014.

[58] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[59] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic guarantees for burer-monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

[60] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[61] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[62] R.T. Rockafellar, M. Wets, and R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009.

[63] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[64] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer Berlin Heidelberg, 2013.

[65] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[66] Mark A Iwen, Felix Krahmer, Sara Krause-Solberg, and Johannes Maly. On recovery guarantees for one-bit compressed sensing on manifolds. *arXiv preprint arXiv:1807.06490*, 2018.

[67] Fabian Latorre Gómez, Armin Eftekhari, and Volkan Cevher. Fast and provable admm for learning with generative priors. *arXiv preprint arXiv:1907.03343*, 2019.

[68] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[69] Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[70] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

[71] Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[72] Parikshit Shah and Venkat Chandrasekaran. Iterative projections for signal identification on manifolds: Global recovery guarantees. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 760–767. IEEE, 2011.

[73] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441. PMLR, 2018.

[74] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[75] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[76] Christos H Papadimitriou. On the complexity of integer programming. *Journal of the ACM (JACM)*, 28(4):765–768, 1981.

[77] Daniel Bienstock. Computational study of a family of mixed-integer quadratic programming problems. *Mathematical programming*, 74(2):121–140, 1996.

[78] Alberto Del Pia, Santanu S Dey, and Marco Molinaro. Mixed-integer quadratic programming is in np. *Mathematical Programming*, 162(1-2):225–240, 2017.

[79] Big-M and convex hulls. https://yalmip.github.io/tutorial/bigmandconvexhulls/.

[80] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[81] Emmanuel J Candes et al. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[82] Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.