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. Serving 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

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

Above, $\mathcal{L} \colon \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator and the vector y typically in the form of $\mathcal{L}(x^\sharp)$ stores m (possibly inexact) observations of the true model $x^\sharp$. Alternatively, as briefly discussed later, one can consider the basis pursuit or lasso reformulations of the problem (1.1).

A certificate of correctness for the output of the machine (1.1) is at the heart of the classical gauge function theory. 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 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.
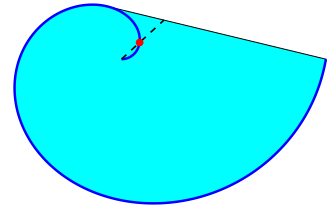

FIGURE 1. As a toy example, the (blue) curve represents the learning alphabet $\mathcal{A}$. A level set of the corresponding gauge function $\mathcal{G}_\mathcal{A}$ is filled with cyan. The (convex) gauge function evidently loses the geometric details of the alphabet $\mathcal{A}$.

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. For example, see [12, 13] for sparse principal component analysis (PCA) and see [14, 15] in the context of super-resolution. Motivated by these examples, in this work we will develop a theoretical foundation for a nonconvex counterpart of the gauge function, along with some basic computational tools.

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

(i) 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).

(ii) 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 3.20). The new theory is showcased with two stylized applications to manifold models and sparse PCA.

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

Additional details. We now provide a section-by-section overview of this work, punctuated by a few bibliographic notes: Section 2 reviews the classical gauge function theory. Several successful and failed applications of this theory are highlighted in Section 2 and in the appendices.

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. 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, gauge and gauge$_p$ functions nearly coincide. In contrast, whenever the classical theory fails, 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

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

in which the new gauge$_p$ function plays the role of regularizer in place of the classical gauge function in (1.1). Section 3.3 elucidates that, as p varies, the new machine interpolates between two extremes:

- the classical convex machine (1.1); and
- the ℓ_0-pursuit: $\min_x \|\mathcal{L}(x) - y\|_2^2$ subject to x has an r-sparse decomposition in $\mathcal{A}$.

We recall that r is the sparsity level of the true model $x^\sharp$ within the alphabet $\mathcal{A}$. Moreover, the new machine (1.2) extends the Burer-Monteiro idea to any alphabet in the following sense: As detailed in Section 3.3, the new machine (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. We also note that 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 3.11 therein introduces a family of certificates for verifying the correctness of the solutions of the optimization problem (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 3.20, loosely analogous to [1, Corollary 3.3.1] for the convex machine (gauge). The proof technique for Theorem 3.20 appears to be new in this context and might be of independent interest. More specifically, instead of a single certificate, the proof of Theorem 3.20 constructs a family of certificates that *jointly* certify the learning outcome.

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.

Computational aspects. 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 a sufficiently small p. That is, the optimization problem does not have any spurious stationary points when p is small. For such alphabets, 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) 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. Instead the practitioners often 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 the problem (ℓ_0-pursuit) 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 5.1 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.

Kurzgesagt. 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, this first work is largely focused on the statistical aspects of the new theory. Beyond the preliminary results presented in Section 5.1, more effort is required to better understand the computational aspects of the new 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

In this section, we first review the classical gauge function theory. We then highlight both successful and failed applications of the theory in order to motivate the main contribution of this work, which is a generalized gauge function theory that addresses the statistical limitations of the classical theory.

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.

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].

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 $\mathrm{conv}(\mathcal{A})$ is defined to be the set $\mathrm{conv}(\{A_i\}_{i=1}^r \cup \{0\})$. We also let $\mathrm{slice}_r(\mathcal{A})$ be the set of all slices of $\mathrm{conv}(\mathcal{A})$ formed by at most r atoms. Note that $\mathrm{slice}_r(\mathcal{A})$ is a set of sets.*

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

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

where the slice $\mathcal{S}^\sharp$ in (2.1) is formed by the atoms $\{A_i^\sharp\}_{i=1}^r$. That is, $\mathcal{S}^\sharp = \mathrm{conv}(\{A_i^\sharp\}_{i=1}^r \cup \{0\})$. The representation in (2.1) is particularly helpful when we are more interested in the atoms $\{A_i^\sharp\}_{i=1}^r$ rather than their coefficients $\{c_i^\sharp\}_{i=1}^r$. In convex statistical learning, the complexity of a model, such as $x^\sharp$, is commonly measured by its gauge function [1, 3].
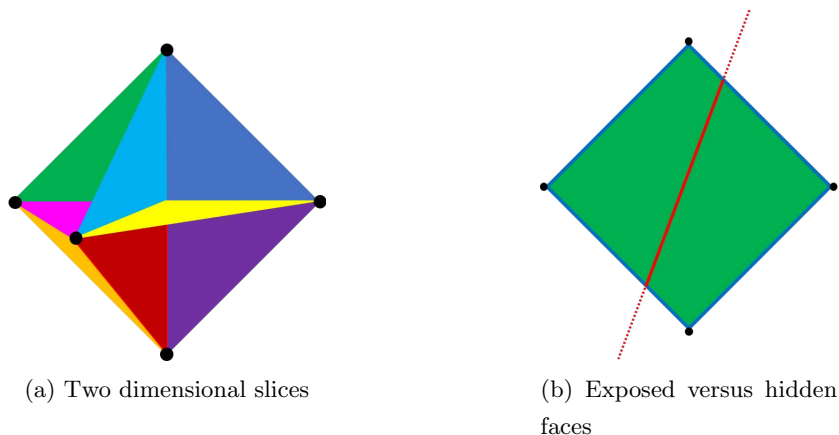
(a) Two dimensional slices

(b) Exposed versus hidden faces

FIGURE 2. Figure 2a depicts several two-dimensional slices of the alphabet $\{\pm e_i\}_{i=1}^3$ in different colors where e_i is the ith 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 2.2 (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*

$$\mathcal{G}_\mathcal{A}(x) := \inf\left\{ t \, : \, x/t \in \text{conv}(\mathcal{A}), \, t \geq 0 \right\} \tag{2.2}$$

$$= \inf\left\{ \sum_{i=1}^l c_i \, : \, x = \sum_{i=1}^l c_i A_i, \, c_i \geq 0, \, A_i \in \mathcal{A}, \, i \in [l] \right\},$$

with the convention that $0/0 = 0$. Above, $[l] := \{1, \cdots, l\}$. Lastly, the second infimum above is taken over l and $\{c_i\}_{i=1}^l$ and $\{A_i\}_{i=1}^l$.

Let us next collect some standard assumptions on learning alphabets. Throughout this work, each technical result is centered around an alphabet that satisfies a subset of the assumptions below.

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 the convex hull of $\mathcal{A}$, i.e.,

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

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

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

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 and we will encounter a few of them 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) setup

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

where $\mathrm{slice}_r(\mathcal{A})$ was defined in Definition 2.1. For example, in statistical inference or signal processing, $\mathcal{L}$ is the measurement operator and y is the vector of observations [28, 29]. (exact) is called an exact setup because it does not account for any noise or numerical inaccuracy that might distort $\mathcal{L}(x^\sharp)$. Given y, in order to learn $x^\sharp$ or its sparse decomposition in the alphabet $\mathcal{A}$, consider the learning machine

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \text{ subject to } \mathcal{G}_\mathcal{A}(x) \le \mathcal{G}_\mathcal{A}(x^\sharp). \tag{2.5}$$

Even though the gauge function appears in the constraints of the problem (2.5), what follows in this section also holds true for the basis pursuit reformulation of (2.5), in which the objective and constraints are swapped and the exact knowledge of $\mathcal{G}_\mathcal{A}(x^\sharp)$ is not required [30]. It seems more convenient for us to work with (2.5), compared to its basis pursuit formulation. Using the definition of gauge function in (2.2), we can also reformulate the above machine as

$$\inf \left\{ \left\| \sum_{i=1}^{l} c_i \mathcal{L}(A_i) - y \right\|_2^2 : \sum_{i=1}^{l} c_i \le \mathcal{G}_\mathcal{A}(x^\sharp), \ c_i \ge 0, \ A_i \in \mathcal{A}, \ i \in [l] \right\}, \tag{gauge}$$

where the infimum is over l and $\{c_i\}_{i=1}^l$ and $\{A_i\}_{i=1}^l$. It is the above reformulation of (2.5) that we will often work with in this paper. To study (gauge), let us recall two basic concepts from convex geometry, both visualized in Figure 2b. See [26, Definitions 2.6 and 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 $\mathrm{ext}(\mathcal{C})$ 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., $\dim(\mathcal{F}) = \dim(\mathrm{aff}(\mathcal{F}))$. 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 $\mathrm{face}_r(\mathcal{A})$ denote the set of all faces of $\mathrm{conv}(\mathcal{A})$ with dimension at most r.*

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

$$\mathrm{ext}(\mathrm{conv}(\mathcal{A})) \subseteq \mathcal{A}, \tag{2.6}$$

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 two definitions, the following lemma exemplifies learning with the gauge function. In effect, the lemma below states 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 (CERTIFICATE OF CORRECTNESS). *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. That is, $x^\sharp = \sum_{i=1}^{\widehat{l}} \widehat{c}_i \widehat{A}_i = 0$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^{\widehat{l}}$ is a solution of the optimization problem* (gauge).

Otherwise, let $\mathcal{F}^\sharp$ be an exposed face of $\mathrm{conv}(\mathcal{A})$ *such that $x^\sharp / \mathcal{G}_\mathcal{A}(x^\sharp) \in \mathcal{F}^\sharp$. Suppose also that the following holds:*

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

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

(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 $Q \in \mathrm{range}(\mathcal{L}^*)$ such that*

$$\langle Q, x - x' \rangle < 0, \qquad \forall x \in \mathrm{conv}(\mathcal{A}) - \mathcal{F}^\sharp, \ \forall x' \in \mathcal{F}^\sharp\,. \tag{2.7}$$

Then the machine (gauge) *successfully returns $x^\sharp$. That is, $x^\sharp = \sum_{i=1}^{\widehat{l}} \widehat{c}_i \widehat{A}_i$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^{\widehat{l}}$ is a solution of the optimization problem* (gauge).

Lemma 2.6 also immediately extends to the basis pursuit formulation of the problem (gauge), in which the objective and constraint are swapped [30].

The correctness certificate Q in Lemma 2.6 can be successfully designed for a variety of linear inverse problems, e.g., in compressive sensing [11] and low-rank matrix completion [32]. Often the starting point is the construction of a pre-certificate within $\mathrm{range}(\mathcal{L}^*)$, for which the assertion (2.7) is then verified, see for example [33].

Despite these success stories, there are many linear inverse problems for which the classical gauge function theory fails. To highlight the statistical failures of the machine (gauge), we focus in this section on structured data factorization, i.e., the particular linear inverse problem in (gauge), for which $\mathcal{L}$ is the identity operator. In this special case, it is not difficult to verify that solving (gauge) is equivalent to finding a minimal decomposition of $x^\sharp$ that achieves $\mathcal{G}_\mathcal{A}(x^\sharp)$. That is, when $\mathcal{L}$ is the identity operator, (gauge) reduces to the optimization problem

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

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

Example 2.7 (MANIFOLD MODELS). *In a multitude of problems, the alphabet $\mathcal{A}$ is naturally an embedded submanifold of the Euclidean space [22, 34]. Our first example showcases the failure of the gauge function theory for manifold manifolds. As a toy example, here we consider the alphabet*

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

which forms a spiral in $\mathbb{R}^2$. That is, $\mathcal{A}$ is the one-dimensional manifold with boundary visualized in Figure 1. For the above alphabet, it is important to note that the inclusion in (2.6) is strict, i.e.,

$$\mathrm{ext}(\mathrm{conv}(\mathcal{A})) \subset \mathcal{A}. \tag{2.8}$$

In particular, the 1-sparse model

$$x^\sharp := A^\sharp = \frac{1}{4\sqrt{2}}(1,1) \in \mathcal{A}, \tag{2.9}$$

is not an extreme point of $\mathrm{conv}(\mathcal{A})$, *i.e.,* $x^\sharp$ *above belongs to the right-hand side but not to the left-hand side of* (2.8). *The model* $x^\sharp$ *in* (2.9) *is represented with a red dot in Figure 1. We note that the model* $x^\sharp$ *in* (2.9) *has the alternative decomposition*

$$x^\sharp = \frac{A_1}{8\sqrt{2}} + \frac{A_2}{2\sqrt{2}}, \qquad A_1 = \frac{1}{2}(0,1) \in \mathcal{A}, \qquad A_2 = 2(1,0) \in \mathcal{A}. \tag{2.10}$$

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

$$\mathcal{G}_\mathcal{A}(x^\sharp) \le \min\left(1, \frac{1}{8\sqrt{2}} + \frac{1}{2\sqrt{2}}\right) = \frac{5}{8\sqrt{2}} < 1. \qquad \text{(see (2.2))} \tag{2.11}$$

That is, (2.9) *is not the minimal decomposition of* $x^\sharp$ *that achieves* $\mathcal{G}_\mathcal{A}(x^\sharp)$ *in* (gauge : $\mathcal{L} = \mathrm{id}$). *In fact, a visual inspection of Figure 1 reveals that the minimal decomposition of* $x^\sharp$ *that achieves* $\mathcal{G}_\mathcal{A}(x^\sharp)$ *is not 1-sparse. We conclude that the machine* (gauge : $\mathcal{L} = \mathrm{id}$) *fails to learn any 1-sparse decomposition for the model* $x^\sharp$.

As as another failed application of the gauge function theory, we turn to sparse PCA. Here the objective 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 samples and features, respectively. In general, the leading principal components of $x^\sharp$ are not sparse, which renders them difficult to interpret. That is, it is often not possible 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 $x^\sharp$ as

$$x^\sharp \in \mathrm{cone}(\mathcal{S}^\sharp), \qquad \mathcal{S}^\sharp \in \mathrm{slice}_r(\mathcal{A}),$$
$$\mathcal{A} := \{uv^\top : \|u\|_2 = \|v\|_2 = 1, \|v\|_0 \le k\} \subset \mathbb{R}^{d_1 \times d_2}, \qquad \text{(sparse PCA)}$$

where $\|v\|_0$ denotes the number of nonzero entries of v. In words, the data matrix $x^\sharp$ is a conic combination of r atoms from the alphabet $\mathcal{A}$. The representation in (sparse PCA) is closely related to [12, 13] and [18, Equation 3.12]. Note that k is the sparsity level of vector v (i.e., the number of its nonzero entries) and should not to be confused with the sparsity level r of the model $x^\sharp$ (i.e., the number of atoms from the alphabet $\mathcal{A}$ that make up $x^\sharp$). 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 (sparse PCA) is exact , namely, we have complete access to the matrix $x^\sharp$. Often, $x^\sharp$ is distorted by noise and we will study this general case in Section 4.

Example 2.8 (SPARSE PCA). *In this example, we consider the model* (sparse PCA) *with* $d_1 = d_2 = 3$ *and* $k = 2$. *To be specific, we consider the 2-sparse model*

$$x^\sharp := \frac{A_1^\sharp}{2} + \frac{A_2^\sharp}{2} \in \mathbb{R}^{3\times 3}, \tag{2.12}$$

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

$$A_1^\sharp := u_1^\sharp (v_1^\sharp)^\top = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^\top \cdot \begin{bmatrix} 0.3122 & 0.95 & 0 \end{bmatrix} \in \mathbb{R}^{3\times3},$$

$$A_2^\sharp := u_2^\sharp (v_2^\sharp)^\top = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 0.95 & 0.3122 \end{bmatrix} \in \mathbb{R}^{3\times3}. \tag{2.13}$$

Recalling Definition 2.1, let $\mathcal{S}^\sharp$ be the two-dimensional slice formed by $\{A_1^\sharp, A_2^\sharp\}$. That is, $\mathcal{S}^\sharp = \mathrm{conv}(\{A_1^\sharp, A_2^\sharp, 0\})$. In particular, in view of (2.12), note that $x^\sharp \in \mathcal{S}^\sharp$. On the other hand, note also that $x^\sharp$ above has the alternative decomposition

$$x^\sharp = c_1 u_1 v_1^\top + c_2 u_2 v_2^\top + c_3 u_3 v_3^\top \qquad (u_1 v_1^\top, u_2 v_2^\top, u_3 v_3^\top \in \mathcal{A})$$

$$= 0.1561 \cdot \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}^\top \cdot \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + 0.6717 \cdot \begin{bmatrix} 0.9082 & -0.0918 & 0.4082 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

$$+ 0.1561 \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top \cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}. \tag{2.14}$$

By comparing the two alternative representations of $x^\sharp$ in (2.12) and (2.14), we observe that

$$\mathcal{G}_\mathcal{A}(x^\sharp) \leq \min\left(\frac{1}{2} + \frac{1}{2}, 0.1561 + 0.6717 + 0.1561 \right) = 0.984 < 1, \qquad \text{(see (2.2))} \tag{2.15}$$

and thus the 2-sparse decomposition of $x^\sharp$ in (2.12) is not minimal, i.e., the decomposition in (2.12) does not achieve $\mathcal{G}_\mathcal{A}(x^\sharp)$. In fact, we may verify that the machine (gauge : $\mathcal{L} = \mathrm{id}$) fails to find any 2-sparse decomposition for the model $x^\sharp$.

We also remark that the failure of (gauge : $\mathcal{L} = \mathrm{id}$) in this example persists even after imposing an incoherence requirement on the alphabet $\mathcal{A}$, i.e., after including $\|v\|_\infty = O(1/\sqrt{k})$ in (sparse PCA). Indeed, the true atoms $\{A_1^\sharp, A_2^\sharp\}$ in (2.13) 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].

In both Examples 2.7 and 2.8, it is easy to verify that $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp)$ belongs to a high- (rather than low-) dimensional exposed face of $\mathrm{conv}(\mathcal{A})$. While both Examples 2.7 and 2.8 fall under the broad umbrella of structured data factorization, it is also not difficult to find other examples for which the classical gauge function theory fails beyond structured data factorization, e.g., see Section 5.2.

3. Gauge$_p$ Function Theory and Main Results

In Section 2, we reviewed the gauge function theory and highlighted its statistical limitations. This section aims to overcome some of the limitations of the classical theory by introducing a generalized gauge function theory, dubbed gauge$_p$ function theory. This generalized theory is the main contribution of this study.

Below, in Section 3.1, we will first introduce the gauge$_p$ function, which is the central object of the new theory. We then compare the gauge$_p$ function with the classical gauge function in Section 3.2. In short, 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. We finally introduce the new learning machine at the heart of the gauge$_p$ function theory in Section 3.3, and list 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 its statistical benefits, the computational aspects of the proposed machine are discussed later in Section 5.1.

3.1. **Gauge$_p$ Function**

Central to the generalized theory is a new notion of complexity for statistical models, dubbed gauge$_p$ function, which generalizes the gauge function in Definition 2.2. Gauge$_p$ function 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 necessary 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 (conv$_p(\mathcal{A})$)**.** *For an alphabet $\mathcal{A}$ and integer p, we define*

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

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

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

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

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

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

In words, the sets $\{\text{conv}_p(\mathcal{A})\}_p$ provide a nested sequence of approximations to the alphabet $\mathcal{A} \subset \mathbb{R}^d$. That is, as p decreases, $\text{conv}_p(\mathcal{A})$ becomes an increasingly better approximation to $\mathcal{A}$. In view of (3.2), we note that the sets $\{\text{conv}_p(\mathcal{A})\}_{p=1}^d$ in (3.2) might be nonconvex. In contrast, the sets $\{\text{conv}_p(\mathcal{A})\}_{p \geq d+1}$ are convex and, in fact, all coincide with $\text{conv}(\mathcal{A})$. Indeed, the identities in (3.2) follow from an application of the Carathéodory theorem [26, Theorem 2.3].

To each set $\text{conv}_p(\mathcal{A})$ above, we associate a gauge$_p$ function in analogy with (2.2).

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*

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

Gauge$_p$ function generalizes the notion of gauge function in Definition 2.2 in the sense that $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ when $p > d$, see (2.2) and (3.3). In contrast, when $p \leq d$, the gauge$_p$ function tightly controls the sparsity level of a model. That is, $\mathcal{G}_p(x)$ is finite only when x has a p-sparse decomposition in the alphabet $\mathcal{A}$. This and other basic properties of 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 both 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) If x does not admit a p-sparse decomposition in the alphabet $\mathcal{A}$, then $\mathcal{G}_{\mathcal{A},p}(x) = \infty$.

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

(v) The convex conjugate of $\mathcal{G}_{\mathcal{A},p}$, denoted here by $\mathcal{G}^*_{\mathcal{A},p}$, is specified as

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

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

(vii) The gauge$_p$ functions satisfy 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 the above result: Proposition 3.4(vii) posits that the gauge$_p$ function coincides with the gauge function for $p > d$. Moreover, even when $p \leq d$, the gauge$_p$ function still preserves certain properties of the gauge function, see Proposition 3.4(ii)-(vi). Crucially, when $p \leq d$, the gauge$_p$ function directly controls the sparsity level, as articulated in Proposition 3.4(iii) and (iv). This property of the gauge$_p$ function remedies the key shortcoming of the classical gauge function concerning the lack of sparsity of the outcome. Indeed, in the negative examples of Section 2, the classical gauge function failed to enforce sparsity, i.e., its minimal decomposition failed to be sparse at all. In contrast, any minimal decomposition of x that achieves $\mathcal{G}_{\mathcal{A},p}(x)$ is p-sparse by Proposition 3.4(iv). Put differently, in the negative examples highlighted of Section 2, at least one of the inequalities in (3.6) was strict. We will further investigate these key differences in Section 3.2 by comparing the gauge and gauge$_p$ functions for several learning alphabets.

Let us also describe a natural interpretation of (3.6): In view of [35, Definition 2.2] or [36], $\mathcal{G}_{\mathcal{A},1}$ is the "most nonconvex" among the gauge$_p$ functions, followed by $\mathcal{G}_{\mathcal{A},2}$, then $\mathcal{G}_{\mathcal{A},3}$ and so on. Here, the nonconvexity of gauge$_p$ functions is measured by their distance from their shared convex envelope $\mathcal{G}_{\mathcal{A}}$, see Proposition 3.4(vi). Another interesting perspective is offered by the the approximate Carathéodory theorem of Maurey [35, 37]. In principle, this theorem can quantify just how well the sets $\text{conv}_p(\mathcal{A})$ approximate $\text{conv}(\mathcal{A})$. However, we will not further pursue these connections.

3.2. **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 a few learning alphabets. Our discussion in this section is limited to the statistical benefits of gauge$_p$ functions and we defer their computational aspects to Section 5.1.

Example 3.5 (SPARSITY). *Consider the alphabet*

$$\mathcal{A} := \{\pm e_i\}_{i=1}^d \subset \mathbb{R}^d, \tag{3.7}$$

which is central to sparse signal processing and high-dimensional statistical inference [4, 11, 38]. *Here, $e_i \in \mathbb{R}^d$ is the ith canonical vector, with its ith entry equal to one and the remaining entries equal to zero. For this alphabet, a simple calculation shows that*

$$\mathcal{G}_{\mathcal{A},p}(x) := \begin{cases} \mathcal{G}_{\mathcal{A}}(x) = \|x\|_1 & \|x\|_0 \le p \\ \infty & otherwise, \end{cases} \tag{3.8}$$

where $\|x\|_0$ is the number of nonzero entries of x. Above, $\|\cdot\|_1$ stands for the ℓ_1-norm. In particular, it follows from (3.8) that $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ when $p \ge d$. 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 outcome to sufficiently sparse models. If one replaces the ℓ_1-norm in (3.8) with the ℓ_2-norm, then the convex envelope of the resulting function would coincide with the k-support norm [39], *an alternative to elastic net regularization* [40]. *We close by noting that, for this choice of the alphabet $\mathcal{A}$, the relation $\mathcal{G}_{\mathcal{A},p} = \mathcal{G}_{\mathcal{A}}$ for $p \ge d$ is an improvement over the conservative but general result in (3.6).*

In Example 3.5, 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 $\mathcal{G}_{\mathcal{A},p}(x) \ge \mathcal{G}_{\mathcal{A}}(x)$ for an arbitrary 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 Section 2. To see this, let us continue below with the manifold example.

Example 3.6 (MANIFOLD MODELS, CONTINUED). *When the alphabet $\mathcal{A}$ is an embedded submanifold, we saw in Section 2 that the gauge function $\mathcal{G}_{\mathcal{A}}$ might lose 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, the gauge$_1$ value of a model is infinity unless that model is 1-sparse, see Proposition 3.4(iii). That is, $\mathcal{G}_{\mathcal{A},1}(x^\sharp) = \infty$, unless $x^\sharp = tA$ for some $t \ge 0$ and atom $A \in \mathcal{A}$. As an example, for the 1-sparse model $x^\sharp$ in (2.9), it is easy to verify that*

$$\mathcal{G}_{\mathcal{A}}(x^\sharp) \le \frac{5}{8\sqrt{2}} < 1 = \mathcal{G}_{\mathcal{A},1}(x^\sharp). \tag{3.9}$$

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. To see

why (3.9) *holds, note that the far-left inequality in* (3.9) *was established in* (2.11) *and the identity in* (3.9) *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 2.8 about sparse PCA, thus explaining the failure of the gauge function in that example.

> The gauge$_p$ function in Definition 3.3 is a simple generalization of the classical gauge function that better controls the sparsity of a model within the learning alphabet. In every failed example of the classical gauge function theory, we observed that $\mathcal{G}_\mathcal{A}(x^\sharp) < \mathcal{G}_{\mathcal{A},r}(x^\sharp)$. Consequently, in these examples, 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 the result in Proposition 3.4(iv). Motivated by this encouraging 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

As reviewed in Section 2, the classical gauge function theory is a theory for (convex) statistical learning that leverages the gauge function to promote sparsity within the learning alphabet. However, the classical gauge function might fail to enforce sparsity and, in such cases, the gauge function theory fails too. To overcome the statistical limitations of the classical gauge function theory, this section introduces a new learning machine, the main building block of which is the gauge$_p$ function in Definition 3.3. As we will see shortly, 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.

To begin, consider the (inexact) setup

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

where $y \in \mathbb{R}^m$ is the available measurement, and ϵ reflects the inexactness of the setup. In signal processing, for example, ϵ quantifies the noise level in our measurements y. In statistical inference, e in the setup (inexact) is also assigned a probability distribution [41, 29], but we avoid this additional layer of complexity here. In particular, when the distribution assigned to e is light-tailed, it often suffices to take ϵ large enough and then work with the setup (inexact).

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

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \quad \text{subject to} \quad \mathcal{G}_{\mathcal{A},p}(x) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \tag{3.10}$$

which uses the gauge$_p$ function $\mathcal{G}_{\mathcal{A},p}$ defined in (3.3). Using the equivalent definition of gauge$_p$ function in (3.4), we can also reformulate the above learning machine as

$$\inf \left\{ \left\| \sum_{i=1}^p c_i \mathcal{L}(A_i) - y \right\|_2^2 : \sum_{i=1}^p c_i \leq \mathcal{G}_\mathcal{A}(x^\sharp), \ c_i \geq 0, \ A_i \in \mathcal{A}, \ i \in [p] \right\}, \qquad \text{(gauge}_p)$$

where the infimum is over $\{c_i\}_{i=1}^p$ and $\{A_i\}_{i=1}^p$. It is the above reformulation of (3.10) that we will primarily work with in this paper. In particular, if $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of (gauge$_p$), then the learning outcome of (gauge$_p$) is the p-sparse model $\widehat{x} = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i$.

The computational aspects of the new learning machine are discussed later in Section 5.1, where we provide a tractable numerical scheme for solving the optimization problem (gauge$_p$). As detailed below, the proposed machine (gauge$_p$) generalizes both the classical machine (gauge) and the familiar ℓ_0-pursuit approach.

Remark 3.7 (A GENERALIZATION OF (gauge)). *The proposed machine* (gauge$_p$) *reduces to the classical convex machine* (gauge) *for $p \geq d + 1$. Indeed, this claim follows from comparing the equivalent formulations* (2.5) *and* (3.10), *and then invoking Proposition 3.4(vii). On the other hand, when $p \leq d$, the machine* (gauge$_p$) *only searches over p-sparse models and thus always returns a p-sparse solution by design. More specifically, the learning output of* (gauge$_p$) *is the p-sparse model $\widehat{x} = \sum_{i=1}^{p} \widehat{c}_i \widehat{A}_i$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^{p}$ is a solution of* (gauge$_p$). *This guaranteed sparsity of the learning outcome immediately rectifies the key failure of the classical gauge function theory, i.e., the lack of sparsity that we observed in Examples 2.7 and 2.8.*

Suppose that $\epsilon = 0$ in (inexact), i.e., for simplicity, consider the exact setup $\mathcal{L}(x^\sharp) = y$. Recall the familiar ℓ_0-pursuit,

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \ \text{ subject to } x \text{ having a } p\text{-sparse decomposition in the alphabet } \mathcal{A}, \quad (\ell_0\text{-pursuit})$$

which finds a p-sparse model $\widetilde{x}$ that satisfies $\mathcal{L}(\widetilde{x}) = y$. For instance, in the context of Example 2.7, the machine (ℓ_0-pursuit) reduces to the familiar optimization problem in [24]:

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \ \text{ subject to } \|x\|_0 \leq p. \tag{3.11}$$

Remark 3.8 (A GENERALIZATION OF THE ℓ_0-PURSUIT). *The new machine also generalizes the well-known ℓ_0-pursuit: For $\epsilon = 0$, recall from Remark 3.7 that the proposed machine* (gauge$_p$) *finds a p-sparse model $\widehat{x}$ that satisfies $\mathcal{L}(\widehat{x}) = y$ and $\mathcal{G}_{\mathcal{A},p}(\widehat{x}) \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp)$. When $p = r$, it not difficult to verify that the new machine* (gauge$_p$) *reduces to* (ℓ_0-pursuit), *provided that $\mathcal{L}$ is an injective map when its domain is restricted to r-sparse models. This restricted injectivity assumption is reasonable because the true model $x^\sharp$ would not be identifiable otherwise. (Recall from* (inexact) *that r denotes the sparsity level of the true model $x^\sharp$.) When $p > r$, however, a key advantage of the new machine* (gauge$_p$) *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. Consequently, searching over p-sparse models might fail to recover the true model $x^\sharp$.*

We summarize this subsection so far in the following note:

> The proposed machine (gauge$_p$) generalizes both of the familiar machines (gauge) and (ℓ_0-pursuit). That is, (gauge$_p$) interpolates two extremes:
>
> (i) In one extreme, when $p \geq d + 1$, the new machine (gauge$_p$) reduces to the classical convex machine (gauge), see Remark 3.7.
>
> (ii) In the other extreme, when $p = r$, the new machine (gauge$_p$) coincides with (ℓ_0-pursuit), provided that $\mathcal{L}$ is an injective map when restricted to r-sparse models. See Remark 3.8. Recall from (inexact) that r denotes the sparsity level of the true model $x^\sharp$.
>
> As we decrease the value of p from $d + 1$ towards r, the machine (gauge$_p$) unlocks a range of potentially new statistical and computational trade-offs. Informally speaking, as p decreases, the statistical accuracy of the new machine (gauge$_p$) improves and we will quantify this improvement later in this section. However, the computational trade-offs of (gauge$_p$) are more complex and we defer their discussion to Section 5.1.

In addition to generalizing (gauge) and (ℓ_0-pursuit), the new machine (gauge$_p$) can also be interpreted as a natural extension of the Burer-Monteiro idea [16]:

Remark 3.9 (GENERALIZATION OF THE BURER-MONTEIRO FACTORIZATION). *A predecessor of the proposed machine* (gauge$_p$) *appears in the context of matrix factorization. To explain their connection, for simplicity, consider the optimization problem*

$$\min_x \ \|\mathcal{L}(x) - y\|_2^2 \ \text{ subject to } \ \mathrm{trace}(x) \leq \gamma^2 \ \text{ and } \ x \in \mathbb{R}^{d \times d} \text{ is positive semi-definite,} \qquad (3.12)$$

where $\mathcal{L} : \mathbb{R}^{d \times d} \to \mathbb{R}^m$ is a linear operator and $\gamma \geq 0$. Because x is a positive semi-definite matrix above, $\mathrm{trace}(x)$ coincides with the nuclear norm of x and the problem (3.12) is therefore a variant of the (convex) learning machines that are widely studied in matrix completion and sensing [32]. Note that the direct computational cost of solving problem (3.12) grows rapidly as the dimension d_1 grows. Instead, it is common to solve the Burer-Monteiro factorization of problem (3.12). More specifically, for an integer $p \leq d$, the factorized version of problem (3.12) is

$$\min_{u \in \mathbb{R}^{d \times p}} \ \|\mathcal{L}(uu^\top) - y\|_2^2 \ \text{ subject to } \ \|u\|_{\mathrm{F}} \leq \gamma, \qquad (3.13)$$

where $\| \cdot \|_{\mathrm{F}}$ stands for the Frobenius norm. Above, we also used the fact that $\mathrm{trace}(uu^\top) = \|u\|_{\mathrm{F}}^2$. When p is sufficiently small, solving the problem (3.13) can offer substantial savings in computational speed and storage, compared to a direct implementation of the problem (3.12). This idea has been extensively studied for matrix-valued learning problems [16, 19, 42, 43, 44, 15].

It is not difficult to verify that the factorized problem (3.13) coincides with (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.*

It is worth noting that the Burer-Monteiro factorization is primarily concerned with those values of p for which (3.12) and (3.13) have the same minimizers. In contrast, as to be seen shortly, we are often interested in those values of p for which the two problems have different minimizers.

Let us also add that it is often possible to remove the constraints in problem (3.13), provided that $p \lesssim \mathrm{rank}(x^\sharp)$. Here, $x^\sharp$ is the hidden model that satisfies $\mathcal{L}(x^\sharp) \approx y$ and $\lesssim$ hides any constant factors. In this regime, which is 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 with $\|u\|_F$, see for instance [20]. Finally, let us revisit the failed applications of the classical gauge function theory in Section 2.

Example 3.10 (EXAMPLES 2.7 AND 2.8, REVISITED). *In these negative toy examples, we saw in Section 2 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$_p$) successfully finds the (unique) r-sparse decomposition of $x^\sharp$ in each example ($p = r$ and $\mathcal{L} = \mathrm{id}$). Here, id stands for the identity operator. In general, the machine (gauge$_p$) always returns a p-sparse model, see Proposition 3.4(iv).*

3.4. Statistical Guarantees

In Section 3.3, we introduced the machine (gauge$_p$) as a generalization of the convex machine (gauge). This section will develop some statistical guarantees for this new learning machine as part of a generalized gauge function theory. The first result of this section, Lemma 3.11 below, provides certificates of correctness for the new machine (gauge$_p$). Lemma 3.13 later extends Lemma 3.11 to account for noise. Lastly, Theorem 3.20 shows that these certificates exist in certain generic learning problems.

Let us start with a lemma that posits that the machine (gauge$_p$) succeeds if certain certificates of correctness exist. This result is reminiscent of Lemma 2.6 in the classical gauge function theory.

Lemma 3.11 (CORRECTNESS CERTIFICATES, EXACT SETUP). *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. That is, $x^\sharp = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i = 0$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge$_p$).*

Otherwise, suppose that $p \geq r$, where r is the sparsity level of $x^\sharp$ in (inexact). For every slice $\mathcal{S} \in \mathrm{slice}_p(\mathcal{A})$, suppose also that (one of) the following holds:

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

(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 $\mathrm{range}(\mathcal{L}^*)$, i.e., there exists a correctness certificate $Q_{\mathcal{S}} \in \mathrm{range}(\mathcal{L}^*)$ such that

$$\left\langle Q_{\mathcal{S}}, x - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right\rangle < 0, \qquad \forall x \in \mathcal{S}. \tag{3.14}$$

Then the machine (gauge$_p$) returns $x^\sharp$ and a p-sparse decomposition of $x^\sharp$ in $\mathcal{A}$. That is, $x^\sharp = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge).

The proof of Lemma 3.11 largely mirrors that of Lemma 2.6 for the convex machine (gauge). This symmetry justifies our unusual perspective when we reviewed the classical theory in Section 2. In Lemma 3.11, the correctness certificates for the machine (gauge$_p$) replace the correctness certificate for the convex machine (gauge). The remark below compares the two types of certificates.

Remark 3.12 (CONVEX VERSUS NONCONVEX LEARNING). *When $p > d$, the new machine (gauge$_p$) reduces to the convex machine (gauge), see Remark 3.7. From Lemma 2.6, recall that the convex machine (gauge) succeeds if $x^\sharp / \mathcal{G}_\mathcal{A}(x^\sharp)$ is separated along $\mathrm{range}(\mathcal{L}^*)$ from the rest of $\mathrm{conv}(\mathcal{A})$. However, when $p \leq d$, the new machine (gauge$_p$) succeeds under the weaker requirement that $x^\sharp / \mathcal{G}_{\mathcal{A},p}(x^\sharp)$ is separated along $\mathrm{range}(\mathcal{L}^*)$ from every slice of $\mathrm{conv}(\mathcal{A})$ formed by at most p atoms.*

For a family of generic learning problems, we will later construct the correctness certificates in Lemma 3.11. This will, in turn, guarantee the success of the new machine (gauge$_p$). It is not difficult to extend Lemma 3.11 to account for an inexact setup, i.e., the case where $\epsilon > 0$ in (inexact).

Lemma 3.13 (CORRECTNESS CERTIFICATES, INEXACT SETUP). *Consider the model $x^\sharp$ in (inexact) and suppose that the alphabet $\mathcal{A}$ satisfies Assumptions 2.3(i) and (iii). If $\mathcal{G}_{\mathcal{A},p}(x^\sharp) = 0$, then the machine (gauge$_p$) correctly returns 0. That is, $x^\sharp = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i = 0$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge$_p$). Otherwise, suppose that $p \geq r$, where r is the sparsity level of $x^\sharp$ in (inexact). For every slice $\mathcal{S} \in \mathrm{slice}_p(\mathcal{A})$, suppose also that (one of) the following holds:*

(i) *If $x^\sharp / \mathcal{G}_{\mathcal{A},p}(x^\sharp) \in \mathcal{S}$, the linear map $\mathcal{L}$ is injective when restricted to the slice $\mathcal{S}$ and its smallest singular value is at least $\sigma > 0$.*

(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 well-separated along $\mathrm{range}(\mathcal{L}^*)$, i.e., there exists a correctness certificate $Q_\mathcal{S} \in \mathbb{R}^d$ and a vector $q_\mathcal{S} \in \mathbb{R}^m$ such that*

$$Q_\mathcal{S} = \mathcal{L}^*(q_\mathcal{S}),$$

$$\|q_\mathcal{S}\|_2 \leq q,$$

$$\left\langle Q_\mathcal{S}, x - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right\rangle < -\gamma \left\| x - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right\|_2, \qquad \forall x \in \mathcal{S}. \tag{3.15}$$

Finally, let the p-sparse model $\widehat{x}$ be a learning outcome of the machine (gauge$_p$). That is, $\widehat{x} = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge$_p$). Then it holds that

$$\|\widehat{x} - x^\sharp\|_2 \leq \max\left(\frac{2\epsilon}{\sigma}, \frac{2\epsilon q}{\gamma} \right). \tag{3.16}$$

Finding the certificates prescribed in Lemma 3.11 is a problem-specific task, similar to the classical gauge function theory in Section 2. Nevertheless, this section provides a somewhat general recipe for constructing these correctness certificates. To that end, we begin with a few definitions.

Definition 3.14 (ANGLE OF A CONE). *The angle $\angle \mathcal{K} \in [0, \pi]$ of a closed cone $\mathcal{K} \subset \mathbb{R}^d$ satisfies*

$$\cos(\angle \mathcal{K}) := \max_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \min_{u' \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle, \tag{3.17}$$

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

As an example, the positive orthant in $\mathbb{R}^2$ has the angle $\pi/4$. That is, $\angle \mathbb{R}_+^2 = \pi/4$. We also recall the Hausdorff metric below [45]. This concept is visualized in Figure 3.

Definition 3.15 (HAUSDORFF DISTANCE). *The Hausdorff distance of two sets $\mathcal{S}$ and $\mathcal{S}'$ is*

$$\mathrm{dist}_H(\mathcal{S}, \mathcal{S}') := \max\left(\max_{s \in \mathcal{S}} \min_{s' \in \mathcal{S}'} \|s - s'\|_2, \ \max_{s' \in \mathcal{S}'} \min_{s \in \mathcal{S}} \|s - s'\|_2 \right) \tag{3.18}$$
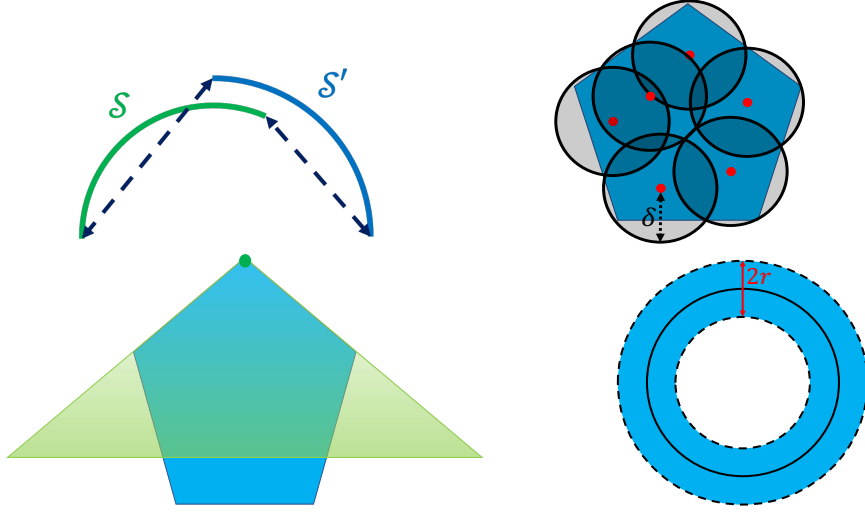
FIGURE 3. This figure visualizes some of the geometric concepts in this work: In the top left panel, the Hausdorff distance between the blue and green arcs equals the length of the longer dashed line segment, see Definition 3.15. The (red) dots in the top right panel form a δ-net for the (blue) trapezoid, see Definition 3.16. The (green) triangle in the bottom left panel is a section of the tangent cone of the (blue) trapezoid at the (green) dot, see Proposition 3.19. Lastly, in the bottom right panel, the reach of the circle in solid black is its radius, see Definition 4.1. The blue annulus shows the normal bundle of radius r.

Next, recall that the metric entropy of a set is a measure for how large or complex that set is [46].

Definition 3.16 (METRIC ENTROPY). *Consider a metric space with distance* dist. *In this space, consider a set $\mathcal{I}$ and its subset $\mathcal{I}'$ that satisfy the following: For every $x \in \mathcal{I}$, there exists $x' \in \mathcal{I}'$ such that $\mathrm{dist}(x, x') \leq \delta$. The set $\mathcal{I}'$ is called a δ-net for the set $\mathcal{I}$ with respect to the metric* dist. *There are often many δ-nets for a set $\mathcal{I}$. In particular, let $\mathrm{net}(\mathcal{I}, \mathrm{dist}, \delta)$ denote a minimal δ-net for $\mathcal{I}$, i.e., the one with the smallest size. The logarithm of the size of this net is called the metric entropy of the set $\mathcal{I}$, denoted by $\mathrm{entropy}(\mathcal{I}, \mathrm{dist}, \delta)$. That is, $\mathrm{entropy}(\mathcal{I}, \mathrm{dist}, \delta) := \log[|\mathrm{net}(\mathcal{I}, \mathrm{dist}, \delta)|]$.*

As an example, the red dots in Figure 3 form a δ-net for the (blue) trapezoid. In linear inverse problems, it is not uncommon for the linear operator to be random, see for instance [11, 32, 9]. For our purposes, we quantify the randomness of the operator $\mathcal{L}$ as follows.

Definition 3.17 (PROBABILISTIC RESTRICTED INJECTIVITY). *The random linear operator $\mathcal{L}$: $\mathbb{R}^d \to \mathbb{R}^m$ satisfies the probabilistic δ-restricted isometry property if $\mathcal{L}$ is a near-isometry when restricting its domain to a low-dimensional subspace with high-probability. More specifically, for $\delta \in [0, 1)$, we say that $\mathcal{L}$ satisfies the probabilistic δ-RIP if an arbitrary subspace $\mathcal{U} \subset \mathbb{R}^d$ with $\dim(\mathcal{U}) \leq Cm\delta^2$ satisfies*

$$(1 - \delta)\|u\|_2 \leq \|\mathcal{L}(u)\|_2 \leq (1 + \delta)\|u\|_2, \qquad \forall u \in \mathcal{U}, \tag{3.19}$$

except with a probability of at most $\exp(-C'\delta^2 m)$. Here, C, C' are universal constants and the probability is over the choice of the operator $\mathcal{L}$.

For example, consider a matrix populated by independent Gaussian random variables, which can be identified with a random linear operator. This matrix satisfies the probabilistic RIP if it is sufficiently flat and properly scaled, see Appendix B.4. Random matrix theory [47] offers similar statements with broad applications in statistical inference and signal processing [11].

Note also that the two-sided nature of (3.19) does not restrict the generality of Definition 3.17. Indeed, consider an operator $\mathcal{L}'$ that satisfies $0 < a\|u\|_2 \leq \|\mathcal{L}'(u)\|_2$ instead of (3.19). Then we can verify that $\mathcal{L} = \sqrt{2/(a^2 + \|\mathcal{L}'\|^2)}\mathcal{L}'$ satisfies the probabilistic δ-RIP with $\delta = \sqrt{(\|\mathcal{L}'\|^2 - a^2)/(\|\mathcal{L}'\|^2 + a^2)}$, where $\|\mathcal{L}'\|$ is the operator norm of $\mathcal{L}'$. Finally, let us define the notion of critical angle below. As we will see shortly, the critical angle quantifies how difficult it is to construct the correctness certificates prescribed in Lemma 3.11.

Definition 3.18 (CRITICAL ANGLE). *For an integer p, alphabet $\mathcal{A}$ and model $x^\sharp$, we define*

$$\text{slice}_{x^\sharp,p}(\mathcal{A}) := \left\{ \mathcal{S} \in \text{slice}_p(\mathcal{A}) : x^\sharp/\mathcal{G}_{\mathcal{A},p}(x^\sharp) \notin \mathcal{S} \right\}. \tag{3.20}$$

In words, $\text{slice}_{x^\sharp,p}(\mathcal{A})$ collects all the slices of $\text{conv}(\mathcal{A})$ that are formed by at most p atoms and 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 $x^\sharp$, denoted by $\theta_{x^\sharp,p}(\mathcal{A}) \in [0,\pi]$, is then defined as

$$\theta_{x^\sharp,p} = \theta_{x^\sharp,p}(\mathcal{A}) := \sup \left\{ \angle\text{cone}\left(\mathcal{S} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} \right) : \mathcal{S} \in \text{slice}_{x^\sharp,p}(\mathcal{A}) \right\}, \tag{3.21}$$

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

To help visualize this new notion, the next result relates the critical angle to a familiar quantity in convex statistical learning: Tangent cones of the set $\text{conv}(\mathcal{A})$.

Proposition 3.19 (EXAMPLE FOR CRITICAL ANGLE). *Consider an alphabet $\mathcal{A} \subset \mathbb{R}^d$, an integer $p \geq d+1$, and a 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.,*

$$\theta_{x^\sharp,p}(\mathcal{A}) \leq 2 \cdot \angle\text{cone}\left(\mathcal{A} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A}}(x^\sharp)} \right).$$

Here, $\text{cone}(\mathcal{A} - x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp))$ is the tangent cone of $\text{conv}(\mathcal{A})$ at the point $x^\sharp/\mathcal{G}_{\mathcal{A}}(x^\sharp)$ [45]. The angle of this tangent cone, which appears above, should be calculated according to Definition 3.14.

As we will see shortly, the smaller the critical angle is, the easier it is to construct the correctness certificates prescribed in Lemma 3.11. Equipped with Definitions 3.14-3.18, we now present the last main result of this section. Informally speaking, the result below states that the new learning machine (gauge$_p$) succeeds when the linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ in (inexact) is generic and m is large enough. For simplicity, the result below is limited to the exact setup, i.e., below we set $\epsilon = 0$ in (inexact).

Theorem 3.20 (EXACT RECOVERY). *Consider the model $x^\sharp$ in (inexact) with $\epsilon = 0$, and assume that the alphabet $\mathcal{A}$ satisfies Assumptions 2.3(i) and (iii). Assume also that the corresponding*

critical angle in Definition 3.18 satisfies $\theta_{x^\sharp,p}(\mathcal{A}) < \frac{\pi}{2}$. Let us equip $\mathrm{slice}_p(\mathcal{A})$ in (3.1) with the pseudo-metric[1] that assigns the distance

$$\mathrm{dist}_p(\mathcal{S},\mathcal{S}') := \mathrm{dist}_H\left(\mathrm{cone}\left(\mathcal{S} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right) \cap \mathbb{S}^{d-1}, \mathrm{cone}\left(\mathcal{S}' - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right) \cap \mathbb{S}^{d-1}\right), \quad (3.22)$$

to every pair of slices $\mathcal{S}, \mathcal{S}' \in \mathrm{slice}_p(\mathcal{A})$, with the convention that $0/0 = 0$. Above, $\mathbb{S}^{d-1}$ is the unit sphere and dist_H denotes the (Euclidean) Hausdorff distance between two sets [45]. Lastly, for $\delta \in [0,1)$, suppose that the random linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ satisfies the probabilistic δ-RIP, see Definition 3.17. If $\mathcal{G}_{\mathcal{A},p}(x^\sharp) = 0$, then the machine (gauge_p) correctly returns 0. That is, $x^\sharp = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i = 0$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge_p). Otherwise, suppose that $p \geq r$, where r is the sparsity level of $x^\sharp$ in (inexact). Suppose also that

$$m \geq \frac{1}{\left[\cos(\theta_{x^\sharp,p})\right]^2}\left(Cp + C'\mathrm{entropy}\left(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}_p, \frac{\cos(\theta_{x^\sharp,p})}{2\max(\|\mathcal{L}\|_{\mathrm{op}}^2, 1)}\right)\right). \quad (3.23)$$

Then the machine (gauge_p) returns $x^\sharp$ and a p-sparse decomposition of $x^\sharp$ in $\mathcal{A}$, except with a probability of at most $\exp(-[\cos(\theta_{x^\sharp,p})]^2 m/C')$. That is, with high probability, it holds that $x^\sharp = \sum_{i=1}^p \widehat{c}_i \widehat{A}_i$, where $\{\widehat{c}_i, \widehat{A}_i\}_{i=1}^p$ is a solution of the optimization problem (gauge). Here, C, C' are universal constants and $\|\mathcal{L}\|_{\mathrm{op}}$ is the operator norm of $\mathcal{L}$.

The proof technique of Theorem 3.20 appears to be new in this context. More specifically, the proof relies on a covering argument, where we form a fine net for all relevant slices of $\mathrm{conv}(\mathcal{A})$, with respect to the metric dist_p in (3.22). We 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. To complete the proof, we finally show that the constructed certificates qualify as optimality certificates for *all* relevant slices of $\mathrm{conv}(\mathcal{A})$, even those not present in the net.

As discussed below, we may consider Theorem 3.20 as a statistical guarantee for the new machine (gauge_p) that is analogous to a similar guarantee for the convex machine (gauge). The regime of interest in Theorem 3.20 is

$$p \lesssim r, \quad \text{and} \quad 0 \leq \theta_{x^\sharp,p} < \frac{\pi}{2},$$

$$\text{and} \quad \mathrm{entropy}\left(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}_p, \frac{\cos(\theta_{x^\sharp,p})}{2\max(\|\mathcal{L}\|_{\mathrm{op}}^2, 1)}\right) \lesssim r, \quad (3.24)$$

in which, for brevity, the symbol $\lesssim$ suppresses any factors that might depend on the alphabet $\mathcal{A}$. An example of the regime (3.24) is given in Section 4.1 with $r = p = 1$. In that example, the factors suppressed in (3.24) are given explicitly and their dependence on $\mathcal{A}$ is through elementary geometric attributes, e.g., intrinsic dimension and volume.

In the regime (3.24), Theorem 3.20 predicts that the machine (gauge_p) successfully recovers the true model $x^\sharp$ and returns a p-sparse decomposition of $x^\sharp$ in the alphabet $\mathcal{A}$, provided that $m \gtrsim r$. That is, in the regime (3.24), solving the optimization problem (gauge_p) with high probability recovers $x^\sharp$ and finds a p-sparse decomposition of $x^\sharp$ if $m \gtrsim r$. Recall that m is the number of observations and r is the sparsity level of $x^\sharp$ within the alphabet $\mathcal{A}$. In contrast, it is not difficult to verify

---

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

that $m \geq \dim(\mathcal{F}^\sharp)$ is necessary (rather than sufficient) for the classical convex machine (gauge) to recover the model $x^\sharp$. Here, $\mathcal{F}^\sharp$ is an exposed face of $\mathrm{conv}(\mathcal{A}) \subset \mathbb{R}^d$ that passes through $x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp)$, see Definition 2.5. In particular, in the negative toy examples of Section 2, one can verify that $\dim(\mathcal{F}^\sharp) = d$ which, in turn, necessitates $m \geq d$.

The informal discussion above highlights the potential benefits of the new machine in the regime (3.24). Beyond this discussion, the entropy number and the critical angle in Theorem 3.20 should be calculated on a case-by-case basis by taking into account the geometry of the learning alphabet $\mathcal{A}$. One such case is presented in the next section.

Finally, in the regime $p \geq d + 1$, the machine (gauge$_p$) reduces to the convex machine (gauge), see Remark 3.12. In particular, our bound in (3.23) is too conservative in this regime and reads as $m \gtrsim d$. Instead, with high probability, a classical result guarantees that the machine (gauge) recovers $x^\sharp$ if $m \geq w(\Omega)^2 + 1$, e.g., see [1, Corollary 3.3.1]. Here, $w(\Omega)$ is the Gaussian width of the set Ω. In turn, Ω is the intersection of the corresponding tangent cone of $\mathcal{A}$ with the unit sphere, i.e., $\Omega := \mathrm{cone}(\mathcal{A} - x^\sharp/\mathcal{G}_\mathcal{A}(x^\sharp)) \cap \mathbb{S}^{d-1}$. For completeness, this result is reviewed in Appendix B.8. Through the celebrated Dudley's inequality, the Gaussian width of the above set Ω relates to its entropy number with respect to the Euclidean metric [1]. This might be contrasted with (3.23) which involves the entropy number of the set slice$_p(\mathcal{A})$ with respect to the metric dist$_p$.

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

In Section 3.3, we studied the new machine (gauge$_p$). Without being exhaustive, this section applies the new learning machine to two representative problems to showcase its potential.

4.1. Manifold-Like Models

Despite the importance of manifold models in signal processing and machine learning [22, 17, 48], the classical gauge function theory might fail to learn manifold models, as highlighted in Section 2 with a toy example. In this section, we consider a slightly more general family of models and show that the new machine (gauge$_p$) succeeds in learning them from limited observations.

Suppose that the alphabet $\mathcal{A}$ is an arbitrary subset of $\mathbb{R}^d$, and consider the 1-sparse setup

$$y := \mathcal{L}(A^\sharp) \in \mathbb{R}^m, \qquad A^\sharp \in \mathcal{A}, \qquad \text{(manifold-like)}$$

where $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator. For simplicity, we have not accounted for measurement noise in the setup above. The equation (manifold-like) is common in learning with nonlinear constraints, e.g., when using a generative adversarial network as the prior [49, 50]. In particular, when $\mathcal{A}$ is an embedded submanifold of $\mathbb{R}^d$, then (manifold-like) reduces to the well-known manifold model [22].

Since the alphabet $\mathcal{A}$ in (manifold-like) can be arbitrary, the 1-sparsity of the (manifold-like) setup does not reduce its generality. Indeed, consider another alphabet $\mathcal{A}' \subset \mathbb{R}^d$. For an integer r, let us set $\mathcal{A} := \mathrm{conv}_r(\mathcal{A}')$, according to Definition 3.1. Then, every 1-sparse model in the alphabet $\mathcal{A}$ corresponds to an r-sparse model of the alphabet $\mathcal{A}'$. On the other hand, as we will see shortly, the 1-sparsity of (manifold-like) will simplify the presentation of the main result in this section.

To recover the atom $A^\sharp$ in (manifold-like), we may implement the machine (gauge$_p$) for any $p \geq 1$. In particular, the choice of $p = 1$ leads us to consider the learning machine

$$\min_{c,A} \ \|c\mathcal{L}(A) - y\|_2^2 \text{ subject to } 0 \leq c \leq 1 \text{ and } A \in \mathcal{A}, \qquad \text{(gauge$_1$: manifold-like)}$$

which is closely related to those numerically studied in [22, Equation 12] and [17, Equation 20].

In the remainder of this section, we will limit ourselves to the special of case of (manifold-like) in which the alphabet $\mathcal{A}$ is a compact embedded submanifold of $\mathbb{R}^d$ [51]. For example, the set $\{x : h(x) = 0\}$ is a k-dimensional embedded submanifold of $\mathbb{R}^d$ if the Jacobian of $h : \mathbb{R}^d \to \mathbb{R}^k$ is full-rank everywhere [52]. This restriction does not considerably reduce the generality of our results below. Indeed, if a bounded alphabet $\mathcal{A}' \subset \mathbb{R}^d$ is not an embedded submanifold, one can always replace the alphabet $\mathcal{A}'$ with a new alphabet $\mathcal{A} \subset \mathbb{R}^d$ such that $\mathcal{A}$ is a compact embedded submanifold and $\mathcal{A}' \subset \mathcal{A}$. For example, a (potentially conservative) choice for $\mathcal{A}$ is a sufficiently large (closed) Euclidean ball that contains $\mathcal{A}'$.

The main result of this section is a corollary of Theorem 3.20 for the special case of $r = p = 1$, presented below. This corollary predicts that the machine (gauge$_1$: manifold-like) successfully recovers the true atom $A^\sharp$ from the vector of observations $y = \mathcal{L}(A^\sharp) \in \mathbb{R}^m$, provided that $\mathcal{L}$ is a generic linear operator and m is sufficiently large. In the corollary below, for tidiness, we assume that $A^\sharp$ has unit norm. We make this assumption without any loss of generality because it can always be enforced by scaling the alphabet $\mathcal{A}$. Again for tidiness, we also introduce a new parameter: In the corollary below, instead of directly using the critical angle $\theta_{A^\sharp,1}(\mathcal{A})$ in Definition 3.18, we will make use of another angle, which is defined as

$$\theta'_{A^\sharp,1}(\mathcal{A}) := \inf \left\{ \angle[A - A^\sharp, A^\sharp] : A \in \mathcal{A} - \{A^\sharp\} \right\} \in [0, \pi]. \qquad (4.1)$$

As shown in the proof of the corollary, the two angles $\theta_{A^\sharp,1}(\mathcal{A})$ and $\theta'_{A^\sharp,1}(\mathcal{A})$ are closely related:

$$\theta_{A^\sharp,1}(\mathcal{A}) \leq \frac{\pi}{2} - \frac{\theta'_{A^\sharp,1}(\mathcal{A})}{2}. \qquad (4.2)$$

In the corollary below, we will also make use of reach of $\mathcal{A}$, a geometric attribute of the manifold $\mathcal{A}$ that is reviewed below [53]. This elementary property, rooted in geometric measure theory, has become somewhat popular in the analysis of manifold models for signal processing [54, 55, 56].

Definition 4.1 (REACH). *Suppose that $\mathcal{A}$ is a compact embedded submanifold of $\mathbb{R}^d$. The reach of $\mathcal{A}$, denoted by $\mathrm{reach}(\mathcal{A})$, is the largest number r that satisfies the following: The open normal bundle of $\mathcal{A}$ of radius r is embedded in $\mathbb{R}^d$ for all $r < \mathrm{reach}(\mathcal{A})$. Recall that the normal bundle of radius r is the set of all normal vectors to the manifold $\mathcal{A}$ of length at most r [51].*

It is not difficult to verify that the reach of a circle is its radius, see Figure 3. As another example, consider the so-called moment curve $t \to [1, \cdots, e^{i2\pi(d-1)t}]$, which can be embedded in $\mathbb{R}^{2d}$. The reach of the moment curve is known to be proportional to $\sqrt{d}$, see [22, Section 2.2.2]. Let us now state the main result of this section.

Corollary 4.2 (MANIFOLD-LIKE MODELS). *Suppose that Assumptions 2.3(i) and (iii) on the alphabet $\mathcal{A}$ are met. Consider the setup (manifold-like) and assume without loss of generality that*

$\|A^{\sharp}\|_2 = 1$. *For an integer k, suppose also that $\mathcal{A}$ is a compact k-dimensional embedded submanifold of $\mathbb{R}^d$. Let $\mathrm{vol}_k(\mathcal{A})$ and $\mathrm{reach}(\mathcal{A}) > 0$ denote the k-dimensional volume of $\mathcal{A}$ and its reach, respectively. We also make the mild technical assumption that $\mathrm{vol}_k(\mathcal{A}) \cdot \mathrm{reach}(\mathcal{A})^k \geq (10.5/\sqrt{k})^k$. Lastly, assume that $\theta'_{A^{\sharp},1}(\mathcal{A}) > 0$, see (4.1). For $\delta \in [0,1)$, suppose that the linear operator $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ in (manifold-like) satisfies the probabilistic δ-RIP, see Definition 3.17. Then the machine (gauge$_1$: manifold-like) returns $A^{\sharp}$ if*

$$m \geq \frac{Ck}{\left[\sin(\theta'_{A^{\sharp},1}(\mathcal{A}))\right]^2} \log\left[\frac{\max(\|\mathcal{L}\|_{\mathrm{op}}^2, 1)}{\sin(\theta'_{A^{\sharp},1}(\mathcal{A}))} (\mathrm{vol}_k(\mathcal{A}))^{\frac{1}{k}} \mathrm{reach}(\mathcal{A})\right], \tag{4.3}$$

except with a probability of at most $\exp\left(-C'[\sin(\theta'_{A^{\sharp},1})]^2 m\right)$ where C, C' are universal constants. That is, with high probability, the pair $(1, A^{\sharp})$ is the unique solution of the optimization problem (gauge$_1$: manifold-like), provided that m is sufficiently large.

The proof of Corollary 4.2 estimates the entropy number on the right-hand side of (3.23) in Theorem 3.20. Note that the number m of observations depends logarithmically on the volume and reach of the manifold $\mathcal{A}$. Moreover, if we ignore the logarithmic term, the number of observations in (4.3) is linear in the dimension k of the manifold $\mathcal{A}$. Corollary 4.2, which specializes Theorem 3.20 to manifolds, is in the same vein as [22, Theorem 4]. To apply Corollary 4.2, one only needs to have access to (estimates of) four geometric attributes of the compact manifold $\mathcal{A}$, namely, its dimension, volume, reach, and critical angle. Except for perhaps the critical angle, these might be considered widely-studied attributes of a manifold $\mathcal{A}$. The final remark of this section revisits the toy example in Figure 1 and computes its critical angle.

Remark 4.3 (CRITICAL ANGLE). *Consider an alphabet $\mathcal{A} \subset \mathbb{R}^d$ and an atom $A^{\sharp} \in \mathcal{A}$. Recall from (4.1) that, for tidiness in this section, we earlier replaced the critical angle in Definition 3.18 with the angle $\theta'_{A^{\sharp},1}(\mathcal{A})$. This new angle $\theta'_{A^{\sharp},1}(\mathcal{A})$ evidently plays a key role in Corollary 4.2. More specifically, to recover the atom $A^{\sharp}$, Corollary 4.2 requires that $\theta'_{A^{\sharp},1}(\mathcal{A}) \neq 0$. That is, Corollary 4.2 requires that $\angle[A - A^{\sharp}, A^{\sharp}] \neq 0$ for every atom $A \in \mathcal{A}$, see (4.1). In Example 2.7, recall that $\mathcal{A}$ was a (spiral) and the atom $A^{\sharp}$ was specified in (2.9), represented by the red dot in Figure 1. For that example, one can verify that $\theta'_{A^{\sharp},1}(\mathcal{A}) \approx 51°$. In contrast, unless we assume that $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ is injective, it is not difficult to carefully construct an alphabet $\mathcal{A}$ for which the convex machine (gauge) would fail to recover the atom $A^{\sharp}$ and $\angle[A_i - A^{\sharp}, A^{\sharp}] \leq \pi/2$ for some atoms $\{A_i\}_i \subset \mathcal{A}$.*

4.2. Sparse Principal Component Analysis

With a negative toy example, we saw in Section 2 that the classical gauge function theory might fail for the sparse PCA problem. This happens because the gauge function might fail to promote sparsity. That is, the minimal decomposition that achieves the gauge function value might not be sparse in the corresponding learning alphabet, see (sparse PCA). In contrast, the generalized theory, developed in Section 3, immediately rectifies this issue. More specifically, by design, any minimal decomposition that achieves the gauge$_p$ function value is always p-sparse, see Proposition 3.4(iv). This observation is precisely the improvement offered by the generalized gauge function theory.

In the remainder of this section, we will show that the new learning machine asymptotically approaches the information-theoretic performance limit of sparse PCA for the spiked covariance model [57, 58, 59, 60]. We will also show that the new machine generalizes beyond the spiked covariance model. More specifically, for sparsity level k and dimension d, consider the alphabet

$$\mathcal{A} := \{uu^\top : \|u\|_2 = 1, \|u\|_0 \leq k, u \in \mathbb{R}^d\} \subset \mathbb{R}^{d \times d}, \tag{4.4}$$

where $\|u\|_0$ is the number of nonzero entries of the vector u. Throughout this section, it is important not to confuse the sparsity level of the vector u (number of its nonzero entries) with the sparsity of a statistical model (number of atoms of the alphabet $\mathcal{A}$ that are present in the model). For $\theta \in [0, 1)$, consider also a Gaussian random vector of length d, with zero mean and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. This covariance matrix is specified as

$$\Sigma := A^\sharp + \theta I_d, \qquad A^\sharp = u^\sharp (u^\sharp)^\top \in \mathcal{A}, \tag{4.5}$$

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

$$y := \frac{1}{n} \sum_{i=1}^n z_i z_i^\top, \tag{4.6}$$

formed by the samples $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$, drawn independently from the distribution normal$(0, \Sigma)$.

Given the sample covariance matrix y, the objective of sparse PCA is to identify the spike in the covariance matrix Σ, i.e., our objective is to identify the atom $A^\sharp$ in (4.5). In view of (inexact), our inexact 1-sparse setup is defined as

$$y := \mathcal{L}(A^\sharp) + e, \qquad A^\sharp \in \mathcal{A}, \quad \text{where} \quad \mathcal{L} = \text{id}, e := \frac{1}{n} \sum_{i=1}^n z_i z_i^\top - A^\sharp. \tag{spike}$$

The alphabet $\mathcal{A}$ is specified as in (4.4). Above, the operator id denotes the identity operator. To recover the spike $A^\sharp$, we may apply the machine (gauge_p) for any $p \geq 1$. In particular, the choice of $p = 1$ leads us to consider the learning machine

$$\min_{c,A} \|y - cA\|_F^2 \text{ subject to } 0 \leq c \leq 1 \text{ and } A \in \mathcal{A}. \tag{4.7}$$

Since we are only interested in recovering the atom $A^\sharp$, and not its amplitude $c^\sharp$, it suffices to consider the optimization over A within (4.7), which reads as

$$\max_{A \in \mathcal{A}} \langle y, A \rangle = \max_{u \in \mathbb{R}^d} \left\{ u^\top y u : \|u\|_2 = 1, \|u\|_0 \leq k \right\}, \qquad \text{(see (4.4))} \qquad (\text{gauge}_1 : \text{SPCA})$$

The above optimization problem is the starting point of the well-known convex relaxation proposed by [57]. To see this connection, note that $(\text{gauge}_1 : \text{SPCA})$ implies that

$$\max_{A \in \mathcal{A}} \langle y, A \rangle \leq \max \left\{ \langle y, A \rangle : \text{trace}(A) = 1, \|A\|_0 \leq k^2, A \in \mathbf{S}_+^d \right\}, \tag{4.8}$$

where $\mathbf{S}_+^d = \{A \in \mathbb{R}^{d \times d} : A = A^\top, A \succeq 0\}$ is the cone of positive semi-definite (PSD) matrices, and $\|A\|_0$ denotes the number of nonzero entries of A. We can obtain a convex relaxation of the right-hand side above by replacing $\|A\|_0$ with the ℓ_1-norm of the matrix A, which is $\|A\|_1 := \sum_{i,j} |A_{i,j}|$.

By doing so, we obtain the convex relaxation

$$\max\left\{\langle y, A\rangle - \lambda\|A\|_1 : \text{trace}(A) = 1, A \in \mathbf{S}_+^d\right\},\tag{4.9}$$

with $\lambda > 0$, which is precisely the optimization problem studied in [57]. As detailed in the result below, under mild assumptions, the machine (gauge$_1$: SPCA) provably discovers the spike $A^\sharp$ in the spiked covariance model. The proof of the result below is standard in the context of empirical processes and the result itself is in the same vein as [58, Proposition 1].

Proposition 4.4 (SPIKED COVARINCE MODEL). *Consider the spiked covariance setup in* (spike). *The machine* (gauge$_1$: SPCA) *asymptotically returns the spike* $A^\sharp$ *in* (spike). *More specifically, consider a sequence* $\{k_l, d_l, n_l\}_l$ *such that* $\lim_{l\to\infty} n_l = \infty$. *Suppose that*

$$\lim_{l\to\infty} \frac{k_l \log d_l}{n_l} = 0.\tag{4.10}$$

Then, in the limit of $l \to \infty$, *solving the optimization problem* $\max_{u\in\mathbb{R}^d}\left\{u^\top yu : \|u\|_2 = 1, \|u\|_0 \le k\right\}$ *returns a vector* $u^\sharp \in \mathbb{R}^d$ *such that* $u^\sharp(u^\sharp)^\top = A^\sharp$, *with a probability that approaches one.*

For the sake of comparison, recall from [58, Theorem 3] that it is impossible for *any* method to discover the spike $A^\sharp$ if

$$\frac{k_l \log(d_l - k_l)}{n_l} > \frac{1}{\theta + \theta^2},$$

where θ is the noise level, see (4.5). Moving on, we have so far focused on the spiked covariance setup, i.e., the covariance matrix Σ in (4.5) contains only one spike. When Σ in (4.5) contains multiple spikes, the common alternative of deflation [61] 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 [62]. However, as we will see below, the proposed learning machine naturally generalizes to multiple spikes. More specifically, instead of (4.5), suppose that the covariance matrix Σ is specified as

$$\Sigma := x^\sharp + \theta I_d, \qquad x^\sharp := \sum_{i=1}^r c_i^\sharp A_i^\sharp, \qquad c_i^\sharp \ge 0, A_i^\sharp \in \mathcal{A}, i \le r,\tag{4.11}$$

where r is the number of atoms (spikes) present in Σ. In view of (inexact), this time our inexact r-sparse setup is

$$y := \mathcal{L}(x^\sharp) + e, \qquad \mathcal{L} = \text{id}, e := \frac{1}{n}\sum_{i=1}^n z_i z_i^\top - x^\sharp.\tag{multiple spikes}$$

To recover the model $x^\sharp$ and/or the spikes $\{A_i^\sharp\}_{i=1}^r$, we may apply the machine (gauge$_p$) for any $p \ge r$. In particular, the choice of $p \ge r$ leads us to consider the learning machine

$$\min\left\{\left\|y - \sum_{i=1}^p c_i A_i\right\|_F^2 : \sum_{i=1}^p c_i \le \text{trace}(x^\sharp), c_i \ge 0, A_i \in \mathcal{A}\right\}.\tag{gauge$_p$: SPCA}$$

Note that $\text{trace}(x^\sharp)$ in (gauge$_p$: SPCA) might not be known advance. However, in principle, guarantees for (gauge$_p$: SPCA) can be transferred to its basis pursuit reformulation in which the objective

function and constraint of (gauge$_p$: SPCA) are swapped, as discussed in Section 2. The following result provides the sufficient conditions for (gauge$_p$: SPCA) to successfully recover the spikes.

Proposition 4.5 (GENERALIZED SPIKED COVARIANCE SETUP). *Consider the generalized spiked covariance setup in* (multiple spikes). *The machine* (gauge$_p$: SPCA) *asymptotically returns the spikes* $\{A_i^\sharp\}_{i=1}^r$ *in* (multiple spikes), *provided that* $p < \mathrm{spark}(\mathcal{A}) - r$. *More specifically, consider the same sequence* $\{k_l, d_l, n_l\}_l$ *as in Proposition 4.4. Then, in the limit of* $l \to \infty$, *solving the optimization problem* (gauge$_p$: SPCA) *successfully returns* $\{c_i^\sharp, A_i^\sharp\}_{i=1}^r$ *with probability that approaches one, provided that* $p < \mathrm{spark}(\mathcal{A}) - r$. *Recall that* $\mathrm{spark}(\mathcal{A})$ *is the smallest number of atoms in* $\mathcal{A} \subset \mathbb{R}^{d \times d}$ *that form a linearly dependent subset of* $\mathbb{R}^{d \times d}$ [63].

For the sparse PCA alphabet in (4.4), we are not aware of any estimates for $\mathrm{spark}(\mathcal{A})$ and it appears to be nontrivial to obtain one. Nevertheless, Proposition 4.5 posits that the new machine (gauge$_p$: SPCA) succeeds when p is sufficiently small. In contrast, the corresponding convex machine, which coincides with (gauge$_p$: SPCA) for $p \geq d(d+1)/2 + 1$, might fail. This last claim about the value of p follows from Remark 3.7 and the fact that dimension of $\mathbf{S}_+^d$ is $d(d+1)/2$.

5. COMPUTATIONAL ASPECTS AND A TRACTABLE NUMERICAL SCHEME

This section discusses the computational aspects of solving the optimization problem (gauge$_p$). As discussed earlier, this problem might be nonconvex, particularly for small values of p. We identify three classes of alphabets:

(i) For certain alphabets, the optimization landscape of the new machine (gauge$_p$) does not have any spurious stationary points and (gauge$_p$) is amenable to a variety of standard optimization algorithms. A prominent example was discussed in Remark 3.9, i.e., the well-known Burer-Monteiro factorization for certain matrix- or tensor-valued learning problems, see [19, 42, 20].

(ii) For certain other alphabets, such as smooth manifolds [22, 64] or shallow neural networks [65], 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 then solved efficiently to stationarity (rather than global optimality) with a variety of first- or second-order algorithms [66]. 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 [23, Chapter 20].

(iii) Yet for many other alphabets, such as the one in Example 2.7 (sparse regression), the problem (gauge$_p$) is NP-hard in general for $p < d$ [67]. Moreover, the second approach above is not 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 the problem (ℓ_0-pursuit) for sparse regression are now competitive with convex heuristics in speed and scalability [24, 25, 68]. We will pursue this direction in the next section.

5.1. **Tractable Numerical Scheme**

In this section, we provide a tractable numerical scheme for solving the problem (gauge_p) for a finite learning alphabet. If the alphabet is infinite, it is sometimes possible to discretize it and apply the algorithm in this section, e.g., in super-resolution [69]. The algorithm in this section builds on the recent developments in mixed-integer programming [24, 25, 68]. To begin, when the learning alphabet $\mathcal{A}$ is finite, the following lemma offers an exact reformulation of the problem (gauge_p) as a mixed integer quadratic programming (MIQP). MIQP is in general an NP-hard problem [70], which comes at no surprise since the original problem of sparse recovery of the setup (exact) is also known to be hard [67]. Nonetheless, the lemma below allows us to deploy the rich literature of computational mathematical programming dedicated to MIQP, see [71, 72].

Lemma 5.1 (MIQP REFORMULATION). *Suppose that the alphabet $\mathcal{A} \subset \mathbb{R}^d$ is finite and denote its size by $|\mathcal{A}| < \infty$. If $M > 0$ is sufficiently large, the machine (gauge_p) is equivalent to the MIQP optimization problem*

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

The MIQP reformulation in Lemma 5.1 leverages the so-called "big-M" technique in which M is only required to be a sufficiently large constant. It is well known that the choice of M has a significant impact on the performance of cutting plane algorithms for convex integer optimization [73].

To address this issue, inspired by the recent work of [24], we next provide a dual reformulation of the optimization problem in Lemma 5.1. This dual reformulation supplies a good starting point (warm start) for branch-and-bound algorithms. This reformulation is a slight generalization of the one proposed in [24] which also handles the linear constraints such as $c_i \geq 0$ and $\sum_{i=1}^{|\mathcal{A}|} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp)$. Below, we will use the notation $[n] := \{1, \cdots, n\}$ for an integer n.

Proposition 5.2 (TRACTABLE ALGORITHM). *Let us define the matrices*

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

where $|\mathcal{A}|$ is the size of the finite set $\mathcal{A}$. Above, $I_{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$ 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 optimization problem (5.1) coincides with the optimal value of the minimax problem

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

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 notation $(u)_i$ returns the ith coordinate of the vector u. Moreover, an optimal vector $s \in \{0,1\}^{|\mathcal{A}|}$ for the problem (5.1) corresponds to an optimal set $S \subset [|\mathcal{A}|]$ in (5.2). Lastly, any fixed-point of the

algorithm below is a solution of (5.2).

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

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

We emphasize that the objective function in (5.3a) is concave and quadratic jointly in the variables (λ, μ), and the number of the summands in (5.3a) 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 $i \leq |\mathcal{A}|$ for which $\left((\mathcal{L}(A))^\top \lambda_k - C^\top \mu_k \right)_i^2$ 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 Examples**

We now investigate two numerical examples that support the theoretical findings of this work.

Example 1: sparse PCA

Our first numerical example is a simplified version of sparse PCA in Section 4.2. More specifically, here the alphabet $\mathcal{A}$ of size 35 is sub-sampled from the set

$$\left\{ uu^\top : \|u\|_2 = 1, \|u\|_0 \leq 2, u \in \mathbb{R}^4 \right\} \subset \mathbb{R}^{4 \times 4},$$

and the model $x^\sharp$ in this example is a combination of three atoms from $\mathcal{A}$, i.e., $x^\sharp = \frac{1}{3} A_1^\sharp + \frac{1}{3} A_2^\sharp + \frac{1}{3} A_3^\sharp$. The setup is $y = x^\sharp + e$, where e is the white Gaussian noise with the standard deviation $\sigma_e \in \{0.01, 0.05, 0.1\}$. For $p \in \{1, 2, 3, 16\}$ and various values of ψ, we numerically solve the program

$$\min_{c,s} \left\{ \left\| \sum_{i=1}^{|\mathcal{A}|} c_i A_i - y \right\|_2^2 : \sum_{i=1}^{|\mathcal{A}|} c_i \leq \psi, \ |c_i| \leq M s_i, \ \sum_{i=1}^{|\mathcal{A}|} s_i = p, \ c_i \geq 0, \ s_i \in \{0,1\} \right\}. \tag{5.4}$$

Note that the above optimization problem is an instantiation of the program (5.1) in which the observation operator is the identity matrix ($\mathcal{L} = \mathrm{id}$) and the the ground-truth gauge value $\mathcal{G}_{\mathcal{A},p}(x^\sharp)$ is approximated by different values ψ. The vector of coefficients that minimize the above problem is denoted by $\widehat{c}_p(\psi)$. Also note that for $p = 16$, the problem (5.4) coincides with the convex machine (gauge), see Remark 3.7.

The numerical results are reported in Figure 4 for different noise levels. Each plot shows the average error across 100 independent experiments with fresh realizations of the noise vector. For $p \in \{1, 2, 3\}$, we solved (5.4) using the solver MOSEK with the interface of YALMIP [74]. The sharp transition in the plots can be explained by the fact that the true model $x^\sharp$ is not feasible in (5.4) for small

(a) $\sigma_e = 0.01$ (b) $\sigma_e = 0.05$ (c) $\sigma_e = 0.1$
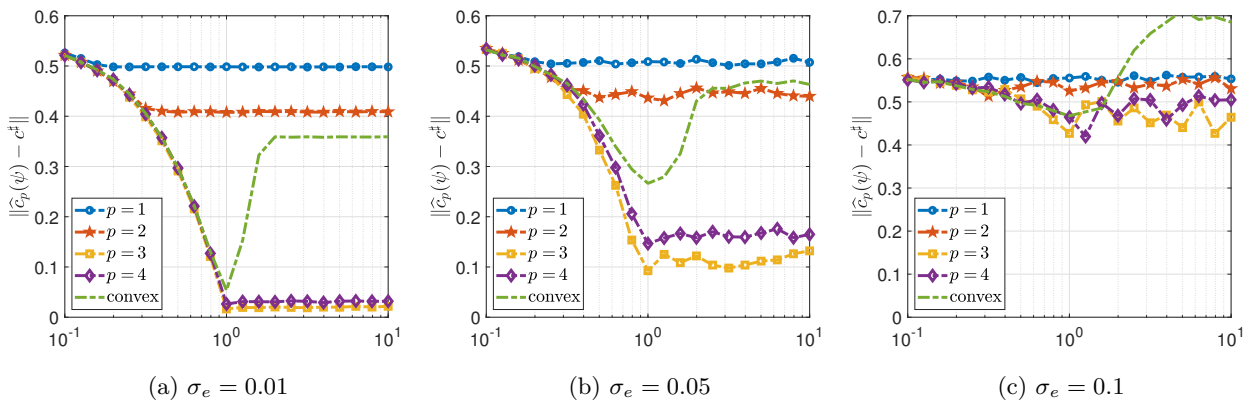
FIGURE 4. Sparse PCA with the sparsity level $\|c^\sharp\|_0 = 3$: performance of the non-convex machine (gauge$_p$) for $p \in \{1, 2, 3, 4\}$ versus the convex counterpart (gauge) (or equivalently (gauge$_p$) for $p = |\mathcal{A}| = 16$)

values of ψ. Moreover, the poor performance for $p = 1$ is explained by the fact that $x^\sharp$, with the sparsity level of two, is never feasible for problem (5.4) with $p = 1$. However, for $p \in \{2, 3\}$, the proposed machine (5.4) considerably outperforms the convex machine (gauge).

Example 2: Super-resolution

Our second numerical example showcases the failure of the classical theory in the context of super-resolution below the diffraction limit. We take the alphabet to be $\mathcal{A} := \{A_\theta\}_{\theta \in \Theta}$, where $A_\theta(\cdot) = \exp(-(\cdot - \theta)^2/0.35^2)$ and $\Theta = \{\theta_i\}_{i=1}^{20}$ is the uniform grid over the interval $[0, 1]$. In words, our alphabet is comprised of twenty (scaled) Gaussian waves centered on a uniform grid over $[0, 1]$. In this example, we consider the two-sparse model $x^\sharp := A_{\theta_{10}} - A_{\theta_{11}}$. This model is shown in blue in Figure 5. In words, the blue curve in Figure 5 represents the superposition of two (scaled) Gaussian waves and the two red bars show the centers and amplitudes of these two waves. The red bars are scaled to fit in the figure.

The (exact) values of the blue curve are then observed at $m = 40$ random locations on the interval $[0, 1]$ and then stored in a vector $y \in \mathbb{R}^{40}$. That is, we set $e = 0$ and the corresponding linear operator $\mathcal{L}$ in (inexact) evaluates and stores the values of its input function at 40 random locations on the interval $[0, 1]$.

We then estimated the true centers and amplitudes (in red) by solving the convex quadratic problem (gauge) with YALMIP in MATLAB [74]. The estimated centers and amplitudes are shown with the black bars in Figure 5. The black bars are also scaled
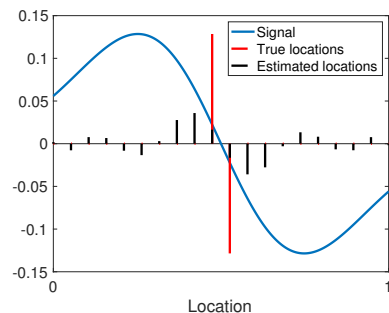

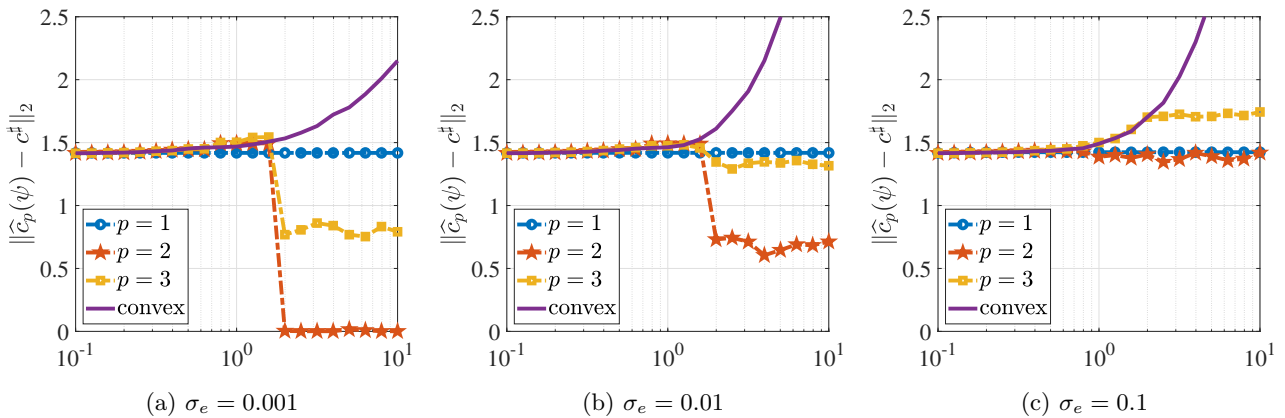FIGURE 5. Failure of the classical gauge function theory in super-resolution, see Section 5.2.

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

FIGURE 6. Super-resolution with the sparsity level $\|c^\sharp\|_0 = 2$: performance of the nonconvex machine (gauge$_p$) for $p \in \{1, 2, 3\}$ versus the convex counterpart (gauge) (or equivalently (gauge$_p$) for $p = |\mathcal{A}| = 20$)

to fit in the figure. The resounding failure of the convex machine in Figure 5 in learning the location of the red bars is an example that visualizes the difficulty of super-resolution below the diffraction limit [75]. Motivated by this observation, we now apply the framework developed in this paper. We let the noise e in (inexact) be a vector of zero-mean and independent Gaussian random variables with standard deviation $\sigma_e \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. For various values of ψ and integer $p \in \{1, 2, 3, 20\}$, we then numerically solve the problem

$$\min_{c,s} \left\{ \left\| \sum_{i=1}^{|\mathcal{A}|} c_i \mathcal{L}(A_i) - y \right\|_2^2 \ : \ \sum_{i=1}^{|\mathcal{A}|} c_i \leq \psi, \ |c_i| \leq M s_i, \ \sum_{i=1}^{|\mathcal{A}|} s_i = p, \ c_i \geq 0, \ s_i \in \{0, 1\} \right\}, \quad (5.5)$$

and collect the optimal coefficients in the vector $\widehat{c}$. The above problem should be compared with (5.1), which was an equivalent reformulation of the proposed learning machine (gauge$_p$).

Each plot in Figure (6) is obtained by averaging the recovery errors over 200 independent experiments with fresh realizations of the noise vector. For $p \in \{1, 2, 3\}$, we solved (5.5) using MOSEK with the interface of YALMIP [74]. For $p = |\mathcal{A}| = 20$, however, the problem (5.5) is a convex program. Indeed, for the choice of $p = 20$, the problem (5.5) coincides with the convex machine (gauge), see Remark 3.7. As in Figure 4, the sharp transitions in Figure 6 can be explained by the fact that the true model $x^\sharp$ is not feasible in (5.5) for small values of ψ. Similarly, the poor performance for $p = 1$ is explained by the fact that $x^\sharp$, with the sparsity level of two, is never feasible for problem (5.5) with $p = 1$. Lastly, for $p \in \{2, 3\}$, the proposed machine (5.5) considerably outperforms the convex machine (gauge).

ACKNOWLEDGEMENTS

Appendix A. Technical Details of Section 2

A.1. **Group Sparsity**

This section presents a third toy example for which the classical gauge function theory fails. Here, the objective is to decompose the model $x^\sharp$ into a small number of vectors with known supports [76, 77]. 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

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

$$\mathcal{A} := \{u : \|u\|_2 = 1, \|u\|_\infty \leq C, \text{supp}(u) \in \Omega\} \subset \mathbb{R}^d, \tag{A.1}$$

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

Example A.1 (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* (A.1) *with $d = 3$, consider the collection of index sets*

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

and the alphabet

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

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

$$x^\sharp := \frac{A_1^\sharp}{2} + \frac{A_2^\sharp}{2} = \left[\begin{array}{ccc} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} - \frac{\sqrt{7}}{8} & \frac{3}{8} \end{array} \right]^\top, \tag{A.4}$$

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

$$A_1^\sharp := \left[\begin{array}{ccc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{array} \right]^\top, \qquad A_2^\sharp := \left[\begin{array}{ccc} 0 & -\frac{\sqrt{7}}{4} & \frac{3}{4} \end{array} \right]^\top. \tag{A.5}$$

Evidently, the model $x^\sharp$ in (A.4) *has the alternative decomposition*

$$x^\sharp = \frac{A_1}{2\sqrt{2}} + \left(\frac{1}{2\sqrt{2}} - \frac{\sqrt{7}}{8} \right) A_2 + \frac{3}{8} A_3, \tag{A.6}$$

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 (A.4) *and* (A.6)*, we find that*

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

and thus the 2-sparse decomposition in (A.4) *is not minimal. In fact, it is not difficult to verify that the machine* (gauge : $\mathcal{L} = \text{id}$) *fails to find any 2-sparse decomposition for $x^\sharp$.*

A.2. 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

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

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

$$t \cdot \widehat{x}/\mathcal{G}_{\mathcal{A}}(\widehat{x}) \in \mathrm{conv}(\mathcal{A}), \qquad \forall t \in [-1, 1], \tag{A.8}$$

i.e., the line segment connecting $\pm\widehat{x}/\mathcal{G}_{\mathcal{A}}(\widehat{x})$ also belongs to $\mathrm{conv}(\mathcal{A})$. 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.8) reduces to

$$\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^{\sharp}) \in \mathrm{conv}(\mathcal{A}). \tag{A.9}$$

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,

$$\mathcal{L}\left(\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} - \frac{x^{\sharp}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})}\right) = \mathcal{L}(\widehat{x} - x^{\sharp}) = 0. \qquad (\mathcal{G}_{\mathcal{A}}(x^{\sharp}) > 0) \tag{A.10}$$

We now consider two cases:

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

$$\frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} - \frac{x^{\sharp}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} \in \mathcal{F}^{\sharp} - \mathcal{F}^{\sharp} \subset \mathrm{lin}(\mathcal{F}^{\sharp}). \tag{A.11}$$

Consequently, (A.10), (A.11) and the injectivity of $\mathcal{L}$ on $\mathrm{lin}(\mathcal{F}^{\sharp})$ 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.9) as

$$\widehat{x}/\mathcal{G}_{\mathcal{A}}(x^{\sharp}) \in \mathrm{conv}(\mathcal{A}) - \mathcal{F}^{\sharp}. \tag{A.12}$$

By assumption of Lemma 2.6, there exists a certificate $Q = \mathcal{L}^*(q)$ that satisfies (2.7). Recalling (A.10), we then write that

$$0 = \langle q, \mathcal{L}(\widehat{x} - x^{\sharp}) \rangle \qquad (\text{see (A.10)})$$

$$= \langle \mathcal{L}^*(q), \widehat{x} - x^{\sharp} \rangle = \langle Q, \widehat{x} - x^{\sharp} \rangle = \mathcal{G}_{\mathcal{A}}(x^{\sharp}) \left\langle Q, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} - \frac{x^{\sharp}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} \right\rangle \quad (\mathcal{G}_{\mathcal{A}}(x^{\sharp}) > 0)$$

$$= \left\langle Q, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} - \frac{x^{\sharp}}{\mathcal{G}_{\mathcal{A}}(x^{\sharp})} \right\rangle < 0$$

where above we used the assumption that $\mathcal{G}(x^{\sharp}) > 0$ as well as (2.7) and (A.12). 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.

APPENDIX B. TECHNICAL DETAILS OF SECTION 3

B.1. Proof of Proposition 3.2

The nested property of $\{\mathrm{conv}_p(\mathcal{A})\}_p$ 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

$$
\mathrm{conv}_1(\mathcal{A}) = \bigcup_{A \in \mathcal{A}} \mathrm{conv}(\{A, 0\}) \qquad (\text{see (3.1)})
$$

$$
= \bigcup_{A \in \mathcal{A}} \bigcup_{0 \le \tau \le 1} \tau A = \bigcup_{0 \le \tau \le 1} \bigcup_{A \in \mathcal{A}} \tau A = \bigcup_{0 \le \tau \le 1} \tau \mathcal{A}. \tag{B.1}
$$

To show the far-right identity in (3.2), recall that every point $\mathrm{conv}(\mathcal{A}) \subset \mathbb{R}^d$ 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

$$
\mathrm{conv}_{d+1}(\mathcal{A}) = \bigcup_{\{A_i\}_{i=1}^{d+1} \subset \mathcal{A}} \mathrm{conv}(\{A_i\}_{i=1}^{d} \cup \{0\}) = \bigcup_{\{A_i\}_{i=1}^{d+1} \subset \mathcal{A}} \bigcup_{0 \le \tau \le 1} \tau \cdot \mathrm{conv}(\{A_i\}_{i=1}^{d+1})
$$

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

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 $\mathrm{conv}_p(\mathcal{A})$ in (3.1) to rewrite the definition of gauge_p function in (3.3) as

$$
\mathcal{G}_{\mathcal{A},p}(x) = \inf \left\{ t : x/t \in \bigcup_{\mathcal{S} \in \mathrm{slice}_r(\mathcal{A})} \mathcal{S}, \, t \ge 0 \right\} \qquad (\text{see (3.1,3.3)})
$$

$$
= \inf \left\{ t : x = \sum_{i=1}^{p} c_i A_i, \, \sum_{i=1}^{p} c_i \le t, \, c_i \ge 0, \, A_i \in \mathcal{A}, \, \forall i \in [p] \right\}
$$

$$
= \inf \left\{ \sum_{i=1}^{p} c_i : x = \sum_{i=1}^{p} c_i A_i, \, c_i \ge 0, \, A_i \in \mathcal{A}, \, \forall i \in [p] \right\},
$$

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 $x/t \in \mathrm{conv}_p(\mathcal{A})$ for every $t > 0$. Since the alphabet $\mathcal{A}$ and, consequently, $\mathrm{conv}_p(\mathcal{A})$ in (3.1) are both bounded by Assumption 2.3(iii), we conclude that $x = 0$.

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

$$
\mathcal{G}_{\mathcal{A},p}^*(z) = \sup_x \langle x, z \rangle - \mathcal{G}_{\mathcal{A},p}(x)
$$

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

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

$$= \text{indicator}_{\text{ball}(\mathcal{D}_\mathcal{A})}(z),$$

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(vi). Proposition 3.4(iii) and (iv) 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 3.11**

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) and (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 (3.14). With this in mind and after recalling (B.5), we write that

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

$$= \left\langle Q_\mathcal{S}, \frac{\widehat{x}}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)} - \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}\right\rangle < 0$$

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 3.11.

B.4. **Example of an Operator That Satisfies the RIP**

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 [47, Corollary 5.35]. In particular, it holds that

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

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

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

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

B.5. **Proof of Proposition 3.19**

Since $p \geq d+1$ 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 $\mathcal{S} \in \text{slice}_p(\mathcal{A})$, note that

$$\mathcal{S} \subset \text{conv}_p(\mathcal{A}) \qquad \text{(see (3.1))}$$

$$= \text{conv}(\mathcal{A}), \qquad \text{(see (3.2))} \tag{B.9}$$

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

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

and, in turn, $\angle\operatorname{cone}(\mathcal{S} - \widetilde{x}) \leq 2\angle\operatorname{cone}(\mathcal{A} - \widetilde{x})$, where we invoked Lemma B.2 below to obtain the last inequality. Using this last inequality and (3.21), we arrive at

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

which completes the proof of Proposition 3.19.

To prove Lemma B.2 below, in addition to (3.17), we first introduce two other notions of angle for a closed cone $\mathcal{K} \subset \mathbb{R}^d$, i.e.,

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

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

These quantities are related as follows.

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

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

Proof. Note that

$$\cos(\psi(\mathcal{K})) = \min_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \min_{u' \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle \qquad (\text{see (B.10)})$$

$$\leq \max_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \min_{u' \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle = \cos(\angle\mathcal{K}) \qquad (\text{see (3.17)})$$

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

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

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

which proves the remaining inequality in (B.11), and completes the proof of Lemma B.1. $\qquad\square$

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 $\angle\mathcal{K}_1 \leq 2\angle\mathcal{K}_2$.*

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

$$\cos(\angle\mathcal{K}_1) \geq \cos(\psi(\mathcal{K}_1)) = \min_{u \in \mathcal{K}_1 \cap \mathbb{S}^{d-1}} \min_{u' \in \mathcal{K}_1 \cap \mathbb{S}^{d-1}} \langle u, u' \rangle$$

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

which completes the proof of Lemma B.2. $\qquad\square$
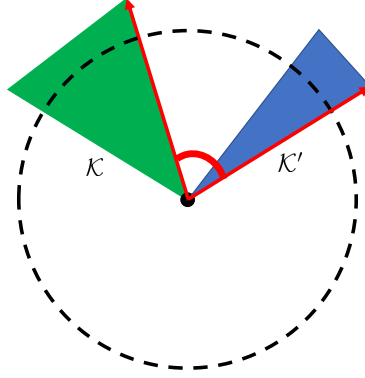
FIGURE 7. The angle between the green and blue cones equals the angle formed by the red arrows, see Definition B.3.

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 $\angle \mathcal{K}_1 \leq \angle \mathcal{K}_2$ when $\mathcal{K}_1$ is a closed convex cone.

B.6. Proof of Theorem 3.20

Introducing the concept of angle between two cones is beneficial for this proof.

Definition B.3 (ANGLE BETWEEN TWO CONES). *The angle between two closed cones $\mathcal{K}, \mathcal{K}' \subset \mathbb{R}^d$, denoted by $\angle[\mathcal{K}, \mathcal{K}'] \in [0, \pi]$, satisfies*

$$\cos(\angle[\mathcal{K}, \mathcal{K}']) := \min \left(\min_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \max_{u' \in \mathcal{K}' \cap \mathbb{S}^{d-1}} \langle u, u' \rangle, \min_{u' \in \mathcal{K}' \cap \mathbb{S}^{d-1}} \max_{u \in \mathcal{K} \cap \mathbb{S}^{d-1}} \langle u, u' \rangle \right)$$
$$= 1 - \frac{1}{2} \left(\text{dist}_H(\mathcal{K} \cap \mathbb{S}^{d-1}, \mathcal{K}' \cap \mathbb{S}^{d-1}) \right)^2, \tag{B.12}$$

where dist_H *denotes the (Euclidean) Hausdorff distance between two sets* [45].

In words, the angle between two closed cones is the Hausdorff distance of their intersections with the unit sphere. For example, Figure 7 shows sections of a blue cone and a green cone in $\mathbb{R}^2$. The angle between the blue and green cones equals the angle formed by the red arrows in the figure.

When $\mathcal{K}$ and $\mathcal{K}'$ are two (linear) subspaces of $\mathbb{R}^d$ [78], $\angle[\mathcal{K}, \mathcal{K}']$ coincides with the (largest) principal angle between the two subspaces. Throughout the proof, we will frequently use the shorthand

$$\widetilde{x} := \frac{x^\sharp}{\mathcal{G}_{\mathcal{A},p}(x^\sharp)}, \qquad u_x := \frac{x - \widetilde{x}}{\|x - \widetilde{x}\|_2}, \tag{B.13}$$

for $x \neq \widetilde{x}$. As in Lemma 3.11, we assumed above without loss of generality that $\mathcal{G}_{\mathcal{A},p}(x^\sharp) > 0$. The proof of Theorem 3.20 relies on the following technical result, which is similar to [79, Lemma 2.1].

Lemma B.4. *For $\delta' \in [0, 1)$, suppose that the random linear operator $\mathcal{L}$ satisfies the probabilistic δ'-RIP. Then, for a slice $\mathcal{S}' \in \text{slice}_p(\mathcal{A})$, it holds that*

$$\frac{\langle Q_{\mathcal{S}'}, u_{x'} \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} \leq -\cos(\angle \text{cone}(\mathcal{S}' - \widetilde{x})) + \delta', \qquad \forall x' \in \mathcal{S}', \tag{B.14}$$

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

$$Q_{\mathcal{S}'} := \mathcal{L}^*(\mathcal{L}(\widetilde{x} - x_{\mathcal{S}'})), \tag{B.15}$$

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

$$\cos(\angle\mathrm{cone}(\mathcal{S}' - \widetilde{x})) = \min_{x' \in \mathcal{S}'} \left\langle u_{x_{\mathcal{S}'}}, u_{x'} \right\rangle. \tag{B.16}$$

Before proving Lemma B.4 in the next appendix, let us first complete the proof of Theorem 3.20. 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 $\mathrm{net}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ denote a minimal δ-net for $\mathrm{slice}_p(\mathcal{A})$ with respect to the pseudo-metric dist, specified as

$$\mathrm{dist}_p(\mathcal{S}, \mathcal{S}') := \sqrt{2 - 2\cos(\angle[\mathrm{cone}(\mathcal{S} - \widetilde{x}), \mathrm{cone}(\mathcal{S}' - \widetilde{x})])}, \qquad \forall \mathcal{S}, \mathcal{S}' \in \mathrm{slice}_p(\mathcal{A}). \tag{B.17}$$

Indeed, dist above is a pseudo-metric because 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 (B.12). For $\delta' \in [0, 1)$, suppose that the random linear operator $\mathcal{L}$ satisfies the probabilistic δ'-RIP, see (3.19). Consequently, by applying the union bound to all slices in $\mathrm{net}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$, we find that

$$(1 - \delta')\|u\|_2 \leq \|\mathcal{L}(u)\|_2 \leq (1 + \delta')\|u\|_2,$$
$$\forall u \in \mathrm{lin}(\mathcal{S}' - \widetilde{x}), \qquad \forall \mathcal{S}' \in \mathrm{net}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta), \tag{B.18}$$

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 $\mathrm{net}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ is a minimal net by construction. The size of this net is therefore $\exp(\mathrm{entropy}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta))$ by Definition 3.16.

Consider an arbitrary slice $\mathcal{S} \in \mathrm{slice}_p(\mathcal{A})$. By Definition 3.16, there exists another slice $\mathcal{S}' \in \mathrm{net}(\mathrm{slice}_p(\mathcal{A}), \mathrm{dist}, \delta)$ such that $\mathrm{dist}_p(\mathcal{S}, \mathcal{S}') \leq \delta$. After recalling (B.17), this observation leads to

$$\cos(\angle[\mathrm{cone}(\mathcal{S} - \widetilde{x}), \mathrm{cone}(\mathcal{S}' - \widetilde{x})]) \geq 1 - \frac{\delta^2}{2}. \qquad \text{(see (B.17))} \tag{B.19}$$

Consider also an arbitrary point $x \in \mathcal{S}$ and the corresponding unit-norm vector u_x, see (B.13). By (B.19) and after recalling the definition of angle between cones in (B.12), there exists a unit-norm vector $u' \in \mathrm{cone}(\mathcal{S}' - \widetilde{x})$ such that

$$\langle u_x, u' \rangle \geq 1 - \frac{\delta^2}{2}, \tag{B.20}$$

which also immediately implies that

$$\|u_x - u'\|_2^2 = \|u_x\|_2^2 + \|u'\|_2^2 - 2\langle u_x, u' \rangle$$
$$\leq 2 - 2\left(1 - \frac{\delta^2}{2}\right) = \delta^2, \qquad \text{(see (B.13), (B.20))} \tag{B.21}$$

Alternatively, it is also easy to arrive at the above conclusion from the Hausdorff distance interpretation of dist_p that we 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 $\mathrm{cone}(\mathcal{S})$. (Indeed, otherwise there exists a lower-dimensional slice, to the interior of which the ray passing through $\widetilde{x}$ would belong.) Consequently, we can use the definition of $\widetilde{x}$ in (B.13) to write that

$$\mathrm{cone}(\mathcal{S} - \widetilde{x}) \cup -\mathrm{cone}(\mathcal{S} - \widetilde{x}) = \mathrm{lin}(\mathcal{S} - \widetilde{x}). \tag{B.22}$$

Moreover, note that

$$\mathrm{lin}(\mathcal{S} - \widetilde{x}) = \mathrm{lin}(\mathcal{S}), \tag{B.23}$$

because the definition of slice in Definition 2.1 implies that $0 \in \mathcal{S}$. On the other hand, note that

$$
\begin{aligned}
\|\mathcal{L}(u_x)\|_2 &\geq \|\mathcal{L}(u')\|_2 - \|\mathcal{L}(u_x - u')\|_2 && \text{(triangle inequality)} \\
&\geq 1 - \delta' - \|\mathcal{L}\|_{\mathrm{op}} \cdot \|u_x - u'\|_2 && \text{(see (B.18))} \\
&= 1 - \delta' - \|\mathcal{L}\|_{\mathrm{op}} \cdot \delta > 0, && \text{(see (B.21))}
\end{aligned}
\tag{B.24}
$$

where the last line above holds if $\delta' + \|\mathcal{L}\|_{\mathrm{op}} \cdot \delta < 1$. Since the choice of the point $x \in \mathcal{S}$ in (B.24) was arbitrary, we conclude that $\mathcal{L}$ is an injective operator when restricted to $\mathrm{cone}(\mathcal{S} - \widetilde{x})$ and, consequently, also when restricted to $-\mathrm{cone}(\mathcal{S} - \widetilde{x})$. In view of (B.22), $\mathcal{L}$ is also injective when restricted to $\mathrm{lin}(\mathcal{S} - \widetilde{x})$ and, by (B.23), when restricted to $\mathrm{lin}(\mathcal{S})$.

(2) Suppose that $\widetilde{x} \notin \mathcal{S}$. Recall $Q_{\mathcal{S}'}$ from (B.15) and note that

$$
\begin{aligned}
\frac{\langle Q_{\mathcal{S}'}, u_x \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} &= \frac{\langle \mathcal{L}^*(\mathcal{L}(\widetilde{x} - x_{\mathcal{S}'})), u_x \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} && \text{(see (B.15))} \\
&= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u_x) \rangle && \text{(see (B.13))} \\
&= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u') \rangle + \langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u' - u_x) \rangle \\
&= \frac{\langle Q_{\mathcal{S}'}, u' \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} + \langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u' - u_x) \rangle \\
&\leq \frac{\langle Q_{\mathcal{S}'}, u' \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2} + \|\mathcal{L}\|_{\mathrm{op}}^2 \cdot \|u_{x_{\mathcal{S}'}}\|_2 \cdot \|u' - u_x\|_2 \\
&\leq -\cos(\mathrm{cone}(\mathcal{S}' - \widetilde{x})) + \delta' + \|\mathcal{L}\|_{\mathrm{op}}^2 \delta && \text{(see (B.14) and (B.21))} \\
&\leq - \inf_{\mathcal{S}'' \in \mathrm{slice}_{x^{\sharp},p}(\mathcal{A})} \cos(\mathrm{cone}(\mathcal{S}'' - \widetilde{x})) + \delta' + \|\mathcal{L}\|_{\mathrm{op}}^2 \delta \\
&= -\cos(\theta_{x^{\sharp},p}(\mathcal{A})) + \delta' + \|\mathcal{L}\|_{\mathrm{op}}^2 \delta < 0, && \text{(see (3.21))}
\end{aligned}
$$

where the last line above holds if

$$\delta' + \|\mathcal{L}\|_{\mathrm{op}}^2 \delta < \cos(\theta_{x^{\sharp},p}(\mathcal{A})). \tag{B.25}$$

Since the choice of the point $x \in \mathcal{S}$ above was arbitrary, we arrive at the following: 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 3.11 to conclude that $x^{\sharp}$ is the unique minimizer of problem (gauge$_p$). This completes the proof of Theorem 3.20 after choosing $\delta' = \cos(\theta_{x^{\sharp},p}(\mathcal{A}))/2$.

B.7. **Proof of Lemma B.4**

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

$$\frac{\langle Q_{\mathcal{S}'}, x' - \widetilde{x} \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2 \|x' - \widetilde{x}\|_2}$$

$$= \frac{\langle \mathcal{L}^*(\mathcal{L}(\widetilde{x} - x_{\mathcal{S}'})), x' - \widetilde{x} \rangle}{\|x_{\mathcal{S}'} - \widetilde{x}\|_2 \|x' - \widetilde{x}\|_2} \qquad \text{(see (B.15))}$$

$$= -\langle \mathcal{L}(u_{x_{\mathcal{S}'}}), \mathcal{L}(u_{x'}) \rangle \qquad \text{(see (B.13))}$$

$$= \frac{-1}{4}(\|\mathcal{L}(u_{x_{\mathcal{S}'}} + u_{x'})\|_2^2 - \|\mathcal{L}(u_{x_{\mathcal{S}'}} - u_{x'})\|_2^2) \qquad \text{(parallelogram identity)}$$

$$\leq -\frac{1}{4}((1 - \delta')\|u_{x_{\mathcal{S}'}} + u_x\|_2^2 - \|\mathcal{L}\|_{\text{op}}\|u_{x_{\mathcal{S}'}} - u_{x'}\|_2^2) \qquad (\mathcal{L} \text{ satisfies the probabilistic RIP, see (3.19)})$$

$$= -\frac{1}{4}(\|u_{x_{\mathcal{S}'}} + u_{x'}\|_2^2 - \|u_{x_{\mathcal{S}'}} - u_{x'}\|_2^2) + \frac{\delta'}{4}\|u_{x_{\mathcal{S}'}} + u_{x'}\|_2^2 + \frac{\|\mathcal{L}\|_{\text{op}} - 1}{4}\|u_{x_{\mathcal{S}'}} - u_{x'}\|_2^2$$

$$\leq -\frac{1}{4}(\|u_{x_{\mathcal{S}'}} + u_{x'}\|_2^2 - \|u_{x_{\mathcal{S}'}} - u_{x'}\|_2^2) + \frac{1}{4}\max(\delta, \|\mathcal{L}\|_{\text{op}} - 1)\left(\|u_{x_{\mathcal{S}'}} + u_{x'}\|_2^2 + \|u_{x_{\mathcal{S}'}} - u_{x'}\|_2^2\right)$$

$$= -\langle u_{x_{\mathcal{S}'}}, u_{x'} \rangle + \frac{1}{2}\max(\delta, \|\mathcal{L}\|_{\text{op}} - 1)\left(\|u_{x_{\mathcal{S}'}}\|_2^2 + \|u_{x'}\|_2^2\right)$$

$$= -\langle u_{x_{\mathcal{S}'}}, u_{x'} \rangle + \max(\delta', \|\mathcal{L}\|_{\text{op}} - 1) \qquad \text{(see (B.13))}$$

$$\leq -\min_{u_{x''} \in \mathcal{S}'} \langle u_{x_{\mathcal{S}'}}, u_{x''} \rangle + \max(\delta', \|\mathcal{L}\|_{\text{op}} - 1)$$

$$= -\max_{u_{x'} \in \mathcal{S}'} \min_{u_{x''} \in \mathcal{S}'} \langle u_{x'}, u_{x''} \rangle + \max(\delta', \|\mathcal{L}\|_{\text{op}} - 1) \qquad \text{(see (3.17) and (B.16))}$$

$$= -\cos(\angle\text{cone}(\mathcal{S}' - \widetilde{x})) + \max(\delta', \|\mathcal{L}\|_{\text{op}} - 1), \qquad \text{(see (3.17))} \tag{B.26}$$

which completes the proof of Lemma B.4.

B.8. **Review of Corollary 3.3.1 in [1]**

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

Corollary 3.3.1 in [1]. *Suppose that the alphabet $\mathcal{A} \subset \mathbb{R}^d$ is a compact set and that (2.6) holds with equality, i.e., $\text{ext}(\text{conv}(\mathcal{A})) = \mathcal{A}$. 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 $\exp(-C(\sqrt{m} - w(\Omega))^2)$. Here, C is a universal constant. Above, we set $\Omega := \text{cone}(\mathcal{A} - x^{\sharp}/\mathcal{G}_{\mathcal{A}}(x^{\sharp})) \cap \mathbb{S}^{d-1}$. In words, Ω is the intersection of the unit sphere with the tangent cone of $\text{conv}(\mathcal{A})$ at $x^{\sharp}/\mathcal{G}_{\mathcal{A}}(x^{\sharp})$. Moreover, $w(\Omega)$ 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.

APPENDIX C. TECHNICAL DETAILS OF SECTION 4

C.1. **Proof of Corollary 4.2**

First note that

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

We can establish (C.1) by way of contradiction: Recall the definition of gauge$_p$ function in (3.4). It is easy to see that $\mathcal{G}_{\mathcal{A},1}(A^\sharp) \le 1$. In particular, if $\mathcal{G}_{\mathcal{A},1}(A^\sharp) < 1$, then there exists an atom $A' \in \mathcal{A}$ aligned with $A^\sharp$ such that $\|A'\|_2 > \|A^\sharp\|_2 = 1$. This contradicts the assumption that $\angle[A' - A^\sharp, A^\sharp] > 0$ and establishes (C.1).

For atoms $A, A' \in \mathcal{A}$, consider the corresponding one-dimensional slices $\mathcal{S}, \mathcal{S}' \in \text{slice}_1(\mathcal{A})$, which are specified as

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

Note that $\mathcal{S}$ and $\mathcal{S}'$ are simply the line segments connecting A and A' to the origin, respectively. With $p = 1$, the distance in (3.22) between these two slices is

$$
\begin{aligned}
\text{dist}_1(\mathcal{S}, \mathcal{S}') &= \sqrt{2 - 2\cos\left(\angle\left[\text{cone}(\mathcal{S} - A^\sharp), \text{cone}(\mathcal{S}' - A^\sharp)\right]\right)} \qquad \text{(see (3.22,C.1))} \\
&= \text{dist}_\text{H}\left(\text{cone}(\mathcal{S} - A^\sharp) \cap \mathbb{S}^{d-1}, \text{cone}(\mathcal{S}' - A^\sharp) \cap \mathbb{S}^{d-1}\right) \qquad \text{(see (B.12))} \\
&\le \left\|\frac{A - A^\sharp}{\|A - A^\sharp\|_2} - \frac{A' - A^\sharp}{\|A' - A^\sharp\|_2}\right\|_2,
\end{aligned}
\tag{C.2}
$$

with the convention that $0/0 = -A^\sharp$. In the last line above, we used the fact that the arguments of dist_H are two arcs on the unit sphere. The first arc passes through $-A^\sharp$ and $(A - A^\sharp)/\|A - A^\sharp\|_2$. The second arc passes through $-A^\sharp$ and $(A' - A^\sharp)/\|A' - A^\sharp\|_2$. The Hausdorff distance between these two arcs is bounded by the distance of their end points, see the last line of (C.2). Since each one-dimensional slice can be identified with its corresponding atom, it follows that

$$\text{entropy}(\text{slice}_1(\mathcal{A}), \text{dist}_1, \delta) = \text{entropy}(U_{A^\sharp}(\mathcal{A}), \|\cdot\|_2, \delta),$$

$$\text{where} \quad U_{A^\sharp}(\mathcal{A}) = \left\{\frac{A - A^\sharp}{\|A - A^\sharp\|_2} : A \in \mathcal{A}\right\},$$

for every $\delta > 0$ and with the convention that $0/0 = -A^\sharp$. We can now invoke Lemma 15 from [22] to find that

$$\text{entropy}(U_{A^\sharp}(\mathcal{A}), \|\cdot\|_2, \delta) \le C'' k \log\left(\delta^{-1}(\text{vol}_k(\mathcal{A}))^{\frac{1}{k}} \text{reach}(\mathcal{A})\right), \tag{C.3}$$

for a universal constant C'' and every $\delta > 0$.

On the other hand, by definition, the critical angle in (3.21) satisfies

$$\theta_{1,A^\sharp}(\mathcal{A}) = \sup\left\{\frac{1}{2}\angle[A - A^\sharp, -A^\sharp] : A \in \mathcal{A} - \{A^\sharp\}\right\} \qquad \text{(see (3.21))}$$

$$= \sup \left\{ \frac{\pi}{2} - \frac{1}{2} \angle [A - A^\sharp, A^\sharp] : A \in \mathcal{A} - \{A^\sharp\} \right\}$$

$$= \frac{\pi}{2} - \frac{\theta'_{A^\sharp,1}(\mathcal{A})}{2} \qquad \text{(see (4.1))}$$

$$< \frac{\pi}{2}, \tag{C.4}$$

where the last line above uses the assumption on the alphabet $\mathcal{A}$. In view of (C.3) and (C.4), we may now invoke Theorem 3.20 to complete the proof of Corollary 4.2.

C.2. **Proof of Proposition 4.4**

Let

$$\mathcal{U} := \{u \in \mathbb{R}^d : \|u\|_2 = 1, \|u\|_0 \leq k\}, \tag{C.5}$$

for short. For a fixed vector $u \in \mathcal{U}$, note that $\mathbb{E}[u^\top z_i] = 0$ and that

$$\mathbb{E}[(u^\top z_i)^2] = u^\top \mathbb{E}[z_i z_i^\top] u = u^\top \Sigma u$$

$$= u^\top A^\sharp u + \theta \|u\|_2^2 = (u^\top u^\sharp)^2 + \theta, \qquad \text{(see (4.5))} \tag{C.6}$$

for every $i \leq n$. In the second line above, we used the fact that any vector $u \in \mathcal{U}$ has unit ℓ_2-norm. We are particularly interested in the deviation of the random variable $(u^\top z_i)^2$ from its expectation.

Throughout, C is a universal constant, the value of which might change in every appearance. Recall from [80, Definition 2.7] the notion of sub-exponential norm of a random variable, which we denote by $\|\cdot\|_{\psi_1}$. In particular, the sub-exponential norm of the random variable $(u^\top z_i)^2 - (u^\top u^\sharp)^2 - \theta$ can be calculated as

$$\left\| (u^\top z_i)^2 - (u^\top u^\sharp)^2 - \theta \right\|_{\psi_1} \leq C \left\| (u^\top z_i)^2 \right\|_{\psi_1} \qquad \text{[80, Lemma 2.6.8]}$$

$$= C \|u^\top z_i\|_{\psi_2}^2 \qquad \text{[80, Lemma 2.7.6]}$$

$$\leq C \mathbb{E}[(u^\top z_i)^2], \qquad \text{[80, Example 2.5.8]} \tag{C.7}$$

where $\|\cdot\|_{\psi_2}$ returns the sub-Gaussian norm of a random variable [80, Definition 2.5.6]. In view of (C.6), we can revisit (C.7) and write that

$$\left\| (u^\top z_i)^2 - (u^\top u^\sharp)^2 - \theta \right\|_{\psi_1} \leq C(u^\top u^\sharp)^2 + C\theta \leq C, \tag{C.8}$$

where the last inequality above uses the fact that $u, u^\sharp \in \mathcal{U}$ are unit-norm vectors and that $\theta < 1$. Because $\{z_i\}_{i=1}^n$ are independent random variables, we can now apply the Bernstein inequality [80, Corollary 2.8.3] and find that

$$\left| u^\top y u - (u^\top u^\sharp)^2 - \theta \right| = \left| \frac{1}{n} \sum_{i=1}^n (u^\top z_i)^2 - (u^\top u^\sharp)^2 - \theta \right| \leq \delta, \qquad \text{(see (4.6))} \tag{C.9}$$

except with the probability of at most $\exp(-C \min(\delta^2, \delta)n)$ and for every $\delta > 0$. Here, we used the fact that $\theta < 1$ to simplify the failure probability.

The remainder of the proof is a standard covering argument. For $\epsilon > 0$ to be set later, let $\mathcal{U}_\epsilon$ denote a minimal ϵ-net for $\mathcal{U}$, with respect to the ℓ_2-norm, see Definition 3.16 or [80, Definition

4.2.1]. By construction, $\log|\mathcal{U}_\epsilon| = \text{entropy}(\mathcal{U}, \epsilon)$, where $|\cdot|$ returns the size of a finite set and the right-hand side denotes the entropy number of $\mathcal{U}$ at resolution ϵ. Using the definition of $\mathcal{U}$ in (C.5), it is not difficult to calculate that

$$\log|\mathcal{U}_\epsilon| = \text{entropy}(\mathcal{U}, \epsilon) \leq Ck\left(\log d + \log(1/\epsilon)\right). \qquad \text{(see [80, Corollary 4.2.13])} \qquad \text{(C.10)}$$

Applying the union bound to (C.9) and using (C.10), we find that

$$\max_{u \in \mathcal{U}_\epsilon}\left|u^\top yu - (u^\top u^\sharp)^2 - \theta\right| \leq \delta, \qquad\qquad \text{(C.11)}$$

except with a probability of at most $\exp(Ck\log d - Ck\log(1/\epsilon) - C\min(\delta^2, \delta)n)$.

Next, consider an arbitrary $u \in \mathcal{U}$ and choose $u_\epsilon \in \mathcal{U}_\epsilon$ such that $\|u - u_\epsilon\|_2 \leq \epsilon$. Such a point u_ϵ is guaranteed to exist by construction of the ϵ-net $\mathcal{U}_\epsilon$. Using the reverse triangle inequality, we then write that

$$\left|u^\top yu - (u^\top u^\sharp)^2 - \theta\right| - \left|u_\epsilon^\top yu_\epsilon - (u_\epsilon^\top u^\sharp)^2 - \theta\right|$$
$$\leq \left|\left(u^\top yu - u_\epsilon^\top yu_\epsilon\right) - \left((u^\top u^\sharp)^2 - (u_\epsilon^\top u^\sharp)^2\right)\right| \qquad \text{(reverse triangle inequality)}$$
$$\leq \left|u^\top yu - u_\epsilon^\top yu_\epsilon\right| + \left|(u^\top u^\sharp)^2 - (u_\epsilon^\top u^\sharp)^2\right| \qquad \text{(triangle inequality)}$$
$$= \frac{1}{n}\left|\sum_{i=1}^{n}(u^\top z_i)^2 - (u_\epsilon^\top z_i)^2\right| + \left|(u^\top u^\sharp)^2 - (u_\epsilon^\top u^\sharp)^2\right| \qquad \text{(see (4.6))}$$
$$\leq \|u - u_\epsilon\|_2(\|u\|_2 + \|u_\epsilon\|_2)\frac{1}{n}\sum_{i=1}^{n}\|z_i\|_2^2 + \|u - u_\epsilon\|_2(\|u\|_2 + \|u_\epsilon\|_2)\|u^\sharp\|_2^2, \qquad \text{(C.12)}$$

where the last line above uses the Cauchy-Schwarz's inequality multiple times. By construction, u and u_ϵ satisfy $\|u\|_2 = \|u_\epsilon\|_2 = \|u^\sharp\|_2 = 1$ and $\|u - u_\epsilon\|_2 \leq \epsilon$. With this in mind, we bound the last line above

$$\left|u^\top yu - (u^\top u^\sharp)^2 - \theta\right| - \left|u_\epsilon^\top yu_\epsilon - (u_\epsilon^\top u^\sharp)^2 - \theta\right| \leq \frac{2\epsilon}{n}\sum_{i=1}^{n}\|z_i\|_2^2 + 2\epsilon. \qquad \text{(C.13)}$$

Recall that $z_i \sim \text{normal}(0, \Sigma)$, which allows us to write that $z_i \stackrel{\text{dist.}}{=} \Sigma^{\frac{1}{2}}z_i'$ for a standard Gaussian random variable $z_i' \sim \text{normal}(0, I_d)$. Because $\{z_i\}_{i=1}^{n}$ are statistically independent, then so are the new random variables $\{z_i'\}_{i=1}^{n}$. We now revisit (C.13) and write its right-hand side as

$$\left|u^\top yu - (u^\top u^\sharp)^2 - \theta\right| - \left|u_\epsilon^\top yu_\epsilon - (u_\epsilon^\top u^\sharp)^2 - \theta\right| \leq \frac{2\epsilon}{n}\sum_{i=1}^{n}\|\Sigma^{\frac{1}{2}}z_i'\|_2^2 + 2\epsilon$$
$$\leq \frac{2\epsilon\|\Sigma\|}{n}\sum_{i=1}^{n}\|z_i'\|_2^2 + 2\epsilon$$
$$\leq \frac{4\epsilon}{n}\sum_{i=1}^{n}\|z_i'\|_2^2 + 2\epsilon, \qquad \text{(C.14)}$$

where the last line above follows because $\|\Sigma\| = 1 + \theta < 2$, see (4.5). Note that $\sum_{i=1}^{n}\|z_i'\|_2^2$ is a Chi-square random variable of degree nd because $\{z_i'\}_{i=1}^{n}$ are independent standard Gaussian random

vectors. The tail probability of a Chi-square random variable is well-known and we therefore have that

$$\left| u^\top y u - (u^\top u^\sharp)^2 - \theta \right| - \left| u_\epsilon^\top y u_\epsilon - (u_\epsilon^\top u^\sharp)^2 - \theta \right| \le C\epsilon d, \tag{C.15}$$

except with a probability of at most $\exp(-Cnd)$. Together, (C.11) and (C.15) imply that

$$\max_{u \in \mathcal{U}} \left| u^\top y u - (u^\top u^\sharp)^2 - \theta \right| \le \max_{u \in \mathcal{U}_\epsilon} \left| u^\top y u - (u^\top u^\sharp)^2 - \theta \right| + C\epsilon d$$

$$\le \delta + C\epsilon d, \tag{C.16}$$

except with a probability of at most

$$\exp\left(Ck \log d + Ck \log(1/\epsilon) - C \min(\delta^2, \delta)n \right) + \exp(-Cnd). \tag{C.17}$$

Consider a sequence $\{k_l, d_l, n_l\}_l$ such that $\lim_{l \to \infty} n_l = \infty$. Consider also the sequence $\{\epsilon_l\}_l$ with $\epsilon_l = 1/d_l^2$. As $l \to \infty$, in view of (C.16) and (C.17), there exists a sequence $\{\delta_l\}_l$ such that

$$\max_{u \in \mathcal{U}} \left| u^\top y u - (u^\top u^\sharp)^2 - \theta \right| \to 0, \tag{C.18}$$

with a probability that converges to one and provided that

$$\lim_{l \to \infty} k_l \log(d_l)/n_l = 0.$$

Lastly, we note that $u^\sharp$ is the unique maximizer of the function $u \to (u^\top u^\sharp)^2 + \theta$. This completes the proof of Proposition 4.4.

C.3. Proof of Proposition 4.5

Consider nonnegative coefficients $\{c_i\}_{i=1}^p$ and atoms $\{A_i\}_{i=1}^p \subset \mathcal{A}$ that are feasible for the optimization problem (gauge$_p$), i.e., $\sum_{i=1}^p c_i \le \mathcal{G}_{\mathcal{A},p}(x^\sharp)$. Note that

$$\left\| y - \sum_{i=1}^p c_i A_i \right\|_{\mathrm{F}}^2 = \|y\|_{\mathrm{F}}^2 - 2 \sum_{i=1}^p c_i \langle y, A_i \rangle + \left\| \sum_{i=1}^p c_i A_i \right\|_{\mathrm{F}}^2,$$

where only the second and third components depend on $\{c_i, A_i\}_i$. Consequently, the problem (gauge$_p$) has the same solutions as

$$\min \left\{ -2 \sum_{i=1}^p c_i \langle y, A_i \rangle + \left\| \sum_{i=1}^p c_i A_i \right\|_{\mathrm{F}}^2 : \sum_{i=1}^p c_i \le \mathcal{G}_{\mathcal{A},p}(x^\sharp), \, c_i \ge 0, \, A_i \in \mathcal{A} \right\}. \tag{C.19}$$

Since y in (4.6) is random, the only random term above is $\sum_{i=1}^p c_i \langle y, A_i \rangle$. We will focus on this random term first. For every $i \le p$, $A_i \in \mathcal{A}$ implies that there exists $u_i \in \mathcal{U}$ such that $A_i = u_i u_i^\top$. Recall that the sets $\mathcal{A}$ and $\mathcal{U}$ were defined in (gauge$_p$: SPCA) and (C.5), respectively. We can now rewrite the only random term in (C.19) as

$$\sum_{i=1}^p c_i \langle y, A_i \rangle = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^n c_i \left\langle z_j z_j^\top, u_i u_i^\top \right\rangle = \sum_{i,j} c_i \langle u_i, z_j \rangle^2. \quad \text{(see (4.6))} \tag{C.20}$$

Recall from (C.6) that

$$\mathbb{E}[\langle y, A_i \rangle] = \mathbb{E}[\langle u_i, z_j \rangle^2] = \langle \Sigma, A_i \rangle = u_i^\top \Sigma u_i \quad \text{(see (C.6))}$$

$$= u_i^\top \left(\sum_{j=1}^r c_j^\sharp u_j^\sharp (u_j^\sharp)^\top + \theta I \right) u_i \qquad \text{(see (multiple spikes))}$$

$$= \sum_{j=1}^r c_j^\sharp (u_i^\top u_j^\sharp)^2 + \theta \sum_{j=1}^r c_j^\sharp \|u_i\|_2^2$$

$$= \sum_{j=1}^r c_j^\sharp (u_i^\top u_j^\sharp)^2 + \theta \sum_{j=1}^r c_j^\sharp, \qquad \text{(see (C.5))} \tag{C.21}$$

where the last line follows because $u_i \in \mathcal{U}$ is a unit-length vector for every i. On the other hand, following the same steps as in the proof of Proposition 4.4, it is easy to verify that

$$\max_{u \in \mathcal{U}} \left| u^\top y u - \sum_{j=1}^r c_j^\sharp (u^\top u_j^\sharp)^2 - \theta \sum_{j=1}^r c_j^\sharp \right| \le \delta + C\epsilon d, \tag{C.22}$$

except with a probability of at most

$$\exp \left(Ck \log d + Ck \log(1/\epsilon) - C \min(\delta^2, \delta)n \right) + \exp(-Cnd). \tag{C.23}$$

Only this time, the factor C in (C.22) and (C.23) may depend on $\mathcal{G}_p(x^\sharp)$. With (C.22) and (C.23) at hand, we now write that

$$\left| \sum_{i=1}^p c_i \langle y, A_i \rangle - \sum_{i=1}^p c_i \langle \Sigma, A_i \rangle \right| = \left| \sum_{i=1}^p c_i \langle y, A_i \rangle - \sum_{i=1}^p c_i \mathbb{E}[\langle y, A_i \rangle] \right| \qquad \text{(see the first line of (C.21))}$$

$$\le \sum_{i=1}^p c_i \cdot \max_{A \in \mathcal{A}} |\langle y, A \rangle - \mathbb{E}[\langle y, A \rangle]| \qquad \text{(Holder's inequality)}$$

$$= \sum_{i=1}^p c_i \max_{u \in \mathcal{U}} \left| u^\top y u - \sum_{j=1}^r c_j^\sharp (u^\top u_j^\sharp)^2 - \theta \sum_{j=1}^r c_j^\sharp \right| \qquad \text{(see (C.21))}$$

$$\le \sum_{i=1}^p c_i \left(\delta + C\epsilon d \right) \qquad \text{(see (C.22))}$$

$$\le \mathcal{G}_{\mathcal{A},p}(x^\sharp) \left(\delta + C\epsilon d \right), \qquad \text{(feasibility of } \{c_i\}_{i=1}^p) \tag{C.24}$$

except with the failure probability specified in (C.23). With the same argument as in the proof of Proposition 4.4, we find that (C.19) has asymptotically the same minimizers as

$$\min \left\{ -2 \sum_{i=1}^p c_i \langle \Sigma, A_i \rangle + \left\| \sum_{i=1}^p c_i A_i \right\|_{\mathrm{F}}^2 : \sum_{i=1}^p c_i \le \mathcal{G}_{\mathcal{A},p}(x^\sharp),\, c_i \ge 0,\, A_i \in \mathcal{A} \right\}, \tag{C.25}$$

provided that (C.1) holds. We next focus on the deterministic optimization problem (C.25). Recall from (multiple spikes) that

$$\Sigma = x^\sharp + \theta I = \sum_{i=1}^r c_i^\sharp A_i^\sharp + \theta I.$$

Substituting for Σ in (C.25), we find that

$$\min \left\{ -2 \left\langle \sum_{j=1}^r c_j^\sharp A_j^\sharp, \sum_{i=1}^p c_i A_i \right\rangle - 2\theta \sum_{i=1}^p c_i \operatorname{trace}(A_i) + \left\| \sum_{i=1}^p c_i A_i \right\|_{\mathrm{F}}^2 : \sum_{i=1}^p c_i \le \mathcal{G}_p(x^\sharp),\, c_i \ge 0,\, A_i \in \mathcal{A} \right\}$$

$$= \min \left\{ -2 \left\langle \sum_{j=1}^{r} c_j^\sharp A_j^\sharp, \sum_{i=1}^{p} c_i A_i \right\rangle - 2\theta \sum_{i=1}^{p} c_i + \left\| \sum_{i=1}^{p} c_i A_i \right\|_F^2 : \sum_{i=1}^{p} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \, c_i \geq 0, \, A_i \in \mathcal{A} \right\}$$

$$\geq \min \left\{ -2 \left\langle \sum_{j=1}^{r} c_j^\sharp A_j^\sharp, \sum_{i=1}^{p} c_i A_i \right\rangle - 2\theta \mathcal{G}_p(x^\sharp) + \left\| \sum_{i=1}^{p} c_i A_i \right\|_F^2 : \sum_{i=1}^{p} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \, c_i \geq 0, \, A_i \in \mathcal{A} \right\},$$

$$(C.26)$$

where the second line above uses the fact that $A_i \in \mathcal{A}$ satisfies $\text{trace}(A_i) = 1$. (We will later show that the relaxation in the last line above is, in fact, tight.) In the last line above, note also that we can remove the term $-2\theta \mathcal{G}_{\mathcal{A},p}(x^\sharp)$ without changing the minimizers. That is, instead of the optimization problem in the last line above, we can solve

$$\min \left\{ -2 \left\langle \sum_{j=1}^{r} c_j^\sharp A_j^\sharp, \sum_{i=1}^{p} c_i A_i \right\rangle + \left\| \sum_{i=1}^{p} c_i A_i \right\|_F^2 : \sum_{i=1}^{p} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \, c_i \geq 0, \, A_i \in \mathcal{A} \right\}. \quad (C.27)$$

We can add and subtract $\| \sum_{j=1}^{r} c_j^\sharp A_j^\sharp \|_F^2$ to the objective function above. By doing so, we observe that (C.27) has, in turn, the same minimizers as

$$\min \left\{ \left\| \sum_{i=1}^{r} c_i^\sharp A_i^\sharp - \sum_{i=1}^{p} c_i A_i \right\|_F^2 : \sum_{i=1}^{p} c_i \leq \mathcal{G}_{\mathcal{A},p}(x^\sharp), \, c_i \geq 0, \, A_i \in \mathcal{A} \right\}, \quad (C.28)$$

Recall from (multiple spikes) that $x^\sharp = \sum_{i=1}^{r} c_i^\sharp A_i^\sharp$. Let $x^\sharp = \sum_{i=1}^{p} c_i A_i$ be a different decomposition of $x^\sharp$, where $c_i \geq 0$ and $A_i \in \mathcal{A}$ for every $i \leq p$. If we take the trace of both sides of the last identity, we find that

$$\sum_{i=1}^{p} c_i = \sum_{i=1}^{p} c_i \, \text{trace}(A_i) = \text{trace}(x^\sharp) = \sum_{i=1}^{r} c_i^\sharp \, \text{trace}(A_i^\sharp) = \sum_{i=1}^{r} c_i^\sharp, \quad (C.29)$$

where we also used the fact that every atom $A \in \mathcal{A}$ satisfies $\text{trace}(A) = 1$. After recalling the definition of the gauge$_p$ function in (3.4), note that (C.29) implies

$$\mathcal{G}_{\mathcal{A},p}(x^\sharp) = \text{trace}(x^\sharp). \quad (C.30)$$

It follows from (C.30) that $\{c_i, A_i\}_{i=1}^{p}$ is feasible for the problem (C.28). Note also that

$$\sum_{i=1}^{r} c_i^\sharp A_i^\sharp - \sum_{i=1}^{p} c_i A_i = x^\sharp - x^\sharp = 0. \quad (C.31)$$

That is, $\{A_i^\sharp\}_{i=1}^{r} \cup \{A_i\}_{i=1}^{p}$ are linearly dependent. Recalling the definition of spark, we conclude that $r + p \geq \text{spark}(\mathcal{A})$. In other words, if we take $p < \text{spark}(\mathcal{A}) - r$, then $x^\sharp = \sum_{i=1}^{r} c_i^\sharp A_i^\sharp$ is the unique p-sparse decomposition of $x^\sharp$ in the alphabet $\mathcal{A}$. Consequently, $\{c_i^\sharp, A_i^\sharp\}_{i=1}^{r}$ is the unique solution of (C.28). In fact, as we saw earlier, (C.28) has the same minimizers as the problem in the last line of (C.26). Therefore, $\{c_i^\sharp, A_i^\sharp\}_{i=1}^{r}$ is also the unique solution of the problem in the last line of (C.26). Recall from (C.29) and (C.30) that $\sum_{i=1}^{r} c_i^\sharp = \mathcal{G}_p(x^\sharp)$. Therefore, for the choice of $\{c_i^\sharp, A_i^n\}_{i=1}^{r}$, the objective function in the second and third lines of (C.26) coincide. That is, the relaxation in (C.26) is tight and we can replace the inequality in (C.26) with equality. This completes the proof of Proposition 4.5.

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

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

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

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

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

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

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 [81, 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. 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] 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.

[33] 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.

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

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

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

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

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

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

[40] 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.

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

[42] 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.

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

[44] 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.

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

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

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

[48] 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.

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

[50] 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.

[51] John M Lee. *Introduction to Riemannian manifolds*. Springer, 2018.

[52] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[53] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[54] Richard G Baraniuk and Michael B Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[55] Mark A Davenport, Marco F Duarte, Michael B Wakin, Jason N Laska, Dharmpal Takhar, Kevin F Kelly, and Richard G Baraniuk. The smashed filter for compressive classification and target recognition. In *Computational Imaging V*, volume 6498, page 64980H. International Society for Optics and Photonics, 2007.

[56] Mark A Davenport, Chinmay Hegde, Marco F Duarte, and Richard G Baraniuk. Joint manifolds for data fusion. *IEEE Transactions on Image Processing*, 19(10):2580–2594, 2010.

[57] 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.

[58] 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.

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

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

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

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

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

[64] 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.

[65] 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.

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

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

[68] Robbie Vreugdenhil, Viet Anh Nguyen, Armin Eftekhari, and Peyman Mohajerin Esfahani. Principal component hierarchy for sparse quadratic programs. In *International Conference on Machine Learning*, pages 10607–10616, 2021.

[69] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2013.

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

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

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

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

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

[75] 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.

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

[77] 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.

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

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

[80] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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