Effective condition number bounds for convex regularization

Dennis Amelunxen, Martin Lotz and Jake Walvin

***Abstract*—We derive bounds relating Renegar's condition number to quantities that govern the statistical performance of convex regularization in settings that include the ℓ_1-analysis setting. Using results from conic integral geometry, we show that the bounds can be made to depend only on a random projection, or restriction, of the analysis operator to a lower dimensional space, and can still be effective if these operators are ill-conditioned. As an application, we get new bounds for the undersampling phase transition of composite convex regularizers. Key tools in the analysis are Slepian's inequality and the kinematic formula from integral geometry.**

***Index Terms*—Convex regularization, compressed sensing, integral geometry, convex optimization, dimension reduction**

I. Introduction

A well-established approach to solving linear inverse problems with missing information is by means of convex regularization. In one of its manifestations, this approach amounts to solving the minimization problem

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{\Omega}\boldsymbol{x}-\boldsymbol{b}\|_2 \le \varepsilon, \tag{I.1}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{m\times n}$ represents an underdetermined linear operator and $f(\boldsymbol{x})$ is a suitable proper convex function, informed by the application at hand. The typical example is $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, known to promote sparsity, but many other functions have been considered in different settings.

While there are countless algorithms and heuristics to compute or approximate solutions of (I.1) and related problems, the more fundamental question is: when does a solution of (I.1) actually "make sense"? The latter is important because one is usually not interested in a solution of (I.1) per se, but often uses this and related formulations as a proxy for a different, much more intractable problem. The best-known example is the use of the 1-norm to obtain a sparse solution [1], but other popular settings are the total variation norm and its variants for signals with sparse gradient, or the nuclear norm of a matrix when aiming at a low-rank solution.

Regularizers often take the form $f(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$ for a linear map $\boldsymbol{D}$, as in the cosparse recovery setting [2], [3], [4], where $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$ for an analysis operator $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ with possibly $p \ge n$. In this article we present general bounds relating the performance of (I.1) to properties of g and the conditioning of $\boldsymbol{D}$. Moreover, we show that for the analysis we can replace $\boldsymbol{D}$ with a *random projection* applied to $\boldsymbol{D}$, where the target dimension of this projection is independent of the ambient dimension n and only depends on intrinsic properties of the regularizer g.

A. Performance measures for convex regularization

Various parameters have emerged in the study of the performance of problems such as (I.1). Two of the most fundamental ones depend on the *descent cone* $\mathscr{D}(f,\boldsymbol{x}_0)$ of the function f at $\boldsymbol{x}_0$, defined as the convex cone of all directions in which f decreases. These parameters are

- the statistical dimension $\delta(f,\boldsymbol{x}_0) := \delta(\mathscr{D}(f,\boldsymbol{x}_0))$, or equivalently the squared Gaussian width, of the descent cone $\mathscr{D}(f,\boldsymbol{x}_0)$ of f at a solution $\boldsymbol{x}_0$ (cone of direction from $\boldsymbol{x}_0$ in which f decreases), which determines the admissible amount of undersampling m in (I.1) in the noiseless case ($\varepsilon = 0$), in order to uniquely recover a solution $\boldsymbol{x}_0$[1];
- Renegar's condition number $\mathscr{R}_C(\boldsymbol{\Omega})$ of $\boldsymbol{\Omega}$ with respect to the descent cone $C = \mathscr{D}(f,\boldsymbol{x}_0)$ of f at a point $\boldsymbol{x}_0$, which bounds the recovery error $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2$ of a solution $\boldsymbol{x}$ of (I.1).

Before stating the results linking these two parameters, we briefly define them and outline their significance. The statistical dimension of a convex cone is defined as the expected squared length of the projection of a Gaussian vector $\boldsymbol{g}$ onto a cone: $\delta(C) = \mathbb{E}[\|\boldsymbol{\Pi}_C(\boldsymbol{g})\|^2]$ (see Section IV-B for a principled derivation; unless otherwise stated, $\|\cdot\|$ refers to the 2-norm). It has featured as a proxy to the squared Gaussian width in [5], [6] and as the main parameter determining phase transitions in convex optimization [7]. More precisely, let $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\boldsymbol{\Omega} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}_0$. Consider the optimization problem

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{\Omega}\boldsymbol{x} = \boldsymbol{b}, \tag{I.2}$$

which we deem to *succeed* if the solution coincides with $\boldsymbol{x}_0$. In [7, Theorem II] it was shown that for any $\eta \in (0,1)$,

D. Amelunxen was with the Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong

M. Lotz is with the Mathematics Institute, Zeeman Building, University of Warwick, Coventry CV4 7AL, U. K.

J. Walvin was with the School of Mathematics, Alan Turing Building, Univeristy of Manchester, Manchester M13 9PL, U. K.

Manuscript received May 17, 2018; revised September 27, 2019.

[1] Strictly speaking, this is a result for *random* measurement matrices and holds with high probability.

when $\boldsymbol{\Omega}$ has Gaussian entries, then

$$\begin{aligned} &m \geq \delta(f, \boldsymbol{x}_0) + a_\eta \sqrt{n} \\ &\quad \Longrightarrow \quad \text{(I.2) succeeds with probability} \geq 1-\eta; \\ &m \leq \delta(f, \boldsymbol{x}_0) - a_\eta \sqrt{n} \\ &\quad \Longrightarrow \quad \text{(I.2) succeeds with probability} \leq \eta, \end{aligned}$$

with $a_\eta := 4\sqrt{\log(4/\eta)}$. For $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, the relative statistical dimension has been determined precisely by Stojnic [5], and his results match previous derivations by Donoho and Tanner (see [8] and the references). In addition, the statistical dimension / squared Gaussian width also features in the error analysis of the generalized LASSO problem [9], as the minimax mean squared error (MSE) of proximal denoising [10], [11], to study computational and statistical tradeoffs in regularization [12], and in the context of structured regression ([13] and references).

To define Renegar's condition number, first recall the classical condition number of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, defined as the ratio of the operator norm and the smallest singular value. Using the notation $\|\boldsymbol{A}\| := \max_{\boldsymbol{x}\in S^{n-1}} \|\boldsymbol{A}\boldsymbol{x}\|$, $\sigma(\boldsymbol{A}) := \min_{\boldsymbol{x}\in S^{n-1}} \|\boldsymbol{A}\boldsymbol{x}\|$, the classical condition number is given by

$$\kappa(\boldsymbol{A}) = \min\left\{\frac{\|\boldsymbol{A}\|}{\sigma(\boldsymbol{A})}, \frac{\|\boldsymbol{A}\|}{\sigma(\boldsymbol{A}^T)}\right\}.$$

Renegar's condition number arises when replacing the source and target vector spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ with convex cones. Let $C \subseteq \mathbb{R}^n$, $D \subseteq \mathbb{R}^m$ be closed convex cones, and let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Define restricted versions of the norm and the singular value:

$$\|\boldsymbol{A}\|_{C\to D} := \max_{\boldsymbol{x}\in C\cap S^{n-1}} \|\boldsymbol{\Pi}_D(\boldsymbol{A}\boldsymbol{x})\|, \tag{I.3}$$

$$\sigma_{C\to D}(\boldsymbol{A}) := \min_{\boldsymbol{x}\in C\cap S^{n-1}} \|\boldsymbol{\Pi}_D(\boldsymbol{A}\boldsymbol{x})\|, \tag{I.4}$$

where $\boldsymbol{\Pi}_D \colon \mathbb{R}^m \to D$ denotes the orthogonal projection, i.e., $\boldsymbol{\Pi}_D(\boldsymbol{y}) = \arg\min\{\|\boldsymbol{y}-\boldsymbol{z}\| : \boldsymbol{z} \in D\}$.

Renegar's condition number is defined as

$$\mathscr{R}_C(\boldsymbol{A}) := \min\left\{\frac{\|\boldsymbol{A}\|}{\sigma_{C\to\mathbb{R}^m}(\boldsymbol{A})}, \frac{\|\boldsymbol{A}\|}{\sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T)}\right\}. \tag{I.5}$$

In what follows, we simply write $\sigma_C(\boldsymbol{A}) := \sigma_{C\to\mathbb{R}^m}(\boldsymbol{A})$ for the smallest cone-restricted singular value. As mentioned before, Renegar's condition number features implicitly in error bounds solutions of (I.1): if $\boldsymbol{x}_0$ is a feasible point and $\hat{\boldsymbol{x}}$ is a solution of (I.1), then $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\| \leq 2\varepsilon \mathscr{R}_{\mathscr{D}(f,\boldsymbol{x}_0)}(\boldsymbol{\Omega})/\|\boldsymbol{\Omega}\|$ (see, for example, [6]). Renegar's condition number was originally introduced to study the complexity of linear programming [14], see [15] for an analysis of the running time of an interior-point method for the convex feasibility problem in terms of this condition number, and [16] for a discussion and references. In [17], Renegar's condition number is used to study restart schemes for algorithms such as NESTA [18] in the context of compressed sensing.

Unfortunately, computing or even estimating the statistical dimension or condition numbers is notoriously difficult for all but a few examples. For the popular case $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, an effective method of computing $\delta(f, \boldsymbol{x}_0)$ was developed by Stojnic [5], and subsequently generalized in [6], see also [7, Recipe 4.1]. In many practical settings the regularizer f has the form $f(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$ for a matrix $\boldsymbol{D}$, such as in the cosparse or ℓ_1-analysis setting where $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$. Even when it is possible to accurately estimate the statistical dimension (and thus, the permissible undersampling) for a function g, the method may fail for a composite function $g(\boldsymbol{D}\boldsymbol{x})$, due to a lack of certain separability properties [19] (see [20] for recent bounds in the ℓ_1-analysis setting).

B. Main results - deterministic bounds

In this article we derive a characterization of Renegar's condition number associated to a cone as a measure of how much the statistical dimension can change under a linear image of the cone. The first result linking the statistical dimension with Renegar's condition is Theorem A. When using the usual matrix condition number, the upper bound in Equation (I.7) features implicitly in [21], [22] and appears to be folklore.

Theorem A. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone, and $\delta(C)$ the statistical dimension of C. Then for $\boldsymbol{A} \in \mathbb{R}^{p\times n}$,*

$$\delta(\boldsymbol{A}C) \leq \mathscr{R}_C(\boldsymbol{A})^2 \cdot \delta(C), \tag{I.6}$$

where $\mathscr{R}_C(\boldsymbol{A})$ is Renegar's condition number associated to the matrix $\boldsymbol{A}$ and the cone C. If $p \geq n$, $\boldsymbol{A}$ has full rank, and $\kappa(\boldsymbol{A})$ denotes the matrix condition number of $\boldsymbol{A}$, then

$$\frac{\delta(C)}{\kappa(\boldsymbol{A})^2} \leq \delta(\boldsymbol{A}C) \leq \kappa(\boldsymbol{A})^2 \cdot \delta(C). \tag{I.7}$$

Example I.1. Consider the $n \times n$ finite difference matrix

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

This matrix is usually defined with an additional column $(0,\ldots,0,1)^T$, but for simplicity, and to work with a square matrix of full rank, we work with this truncated version. The condition number is known to be of order $\Omega(n)$, making condition bounds using the normal matrix condition number useless. Using Renegar's condition number with respect to a cone, on the other hand, can improve the situation dramatically. Consider, for example, the cone

$$C = \{\boldsymbol{x} \in \mathbb{R}^n : x_1 \geq 0,\ x_i x_{i+1} \leq 0 \text{ for } 1 \leq i < n\}.$$

This cone is the orthant consisting of vectors with alternating signs. The cone-restricted singular value of $\boldsymbol{D}$ is given by

$$\begin{aligned} \sigma_C(\boldsymbol{D})^2 &= \min_{\boldsymbol{x}\in C\cap S^{n-1}} \|\boldsymbol{D}\boldsymbol{x}\|^2 \\ &= \min_{\boldsymbol{x}\in C\cap S^{n-1}} \sum_{i=1}^{n-1} (x_{i+1}-x_i)^2 + x_n^2 \\ &= \min_{\boldsymbol{x}\in C\cap S^{n-1}} 2 - x_1^2 - \sum_{i=1}^{n-1} 2x_i x_{i+1} \geq 1. \end{aligned}$$

Using the same expression for $\|\boldsymbol{Dx}\|^2$, we see that the square of the operator norm is bounded by 4, so that the square of Renegar's condition number with respect to this cone is bounded by 4. If, on the other hand, C is the non-negative orthant, then Renegar's condition number coincides with the normal matrix condition number. Intuitively, Renegar's condition number gives an improvement if the cone C captures a portion of the ellipsoid defined by $\boldsymbol{DD}^T$ that is not too eccentric. Other examples when Renegar's condition number gives significant improvements is for small cones (such as the cone of increasing sequences) or cones contained in linear subspaces of small dimension (such as subdifferential cones of the 1 or ∞ norms).

Theorem A translates into a bound for the statistical dimension of convex regularizers by observing that if $f(\boldsymbol{x}) = g(\boldsymbol{Dx})$ with invertible $\boldsymbol{D}$, then (see Section VI) the descent cone of f at $\boldsymbol{x}_0$ is given by $\mathscr{D}(f, \boldsymbol{x}_0) = \boldsymbol{D}^{-1}\mathscr{D}(g, \boldsymbol{Dx}_0)$. Throughout this paper, we will use $\boldsymbol{A}$ for the transformation matrix in the setting of convex cones, and $\boldsymbol{D}$ for the matrix appearing in a regularizer.

Corollary I.2. *Let $f(\boldsymbol{x}) = g(\boldsymbol{Dx})$, where g is a proper convex function and let $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ be non-singular. Then*

$$\delta(f, \boldsymbol{x}_0) \le \mathscr{R}_{\mathscr{D}(g,\boldsymbol{Dx}_0)}\left(\boldsymbol{D}^{-1}\right) \cdot \delta(g, \boldsymbol{Dx}_0).$$

In particular,

$$\frac{\delta(g, \boldsymbol{Dx}_0)}{\kappa(\boldsymbol{D})^2} \le \delta(f, \boldsymbol{x}_0) \le \kappa(\boldsymbol{D})^2 \cdot \delta(g, \boldsymbol{Dx}_0).$$

Remark I.3. It is interesting to compare the bounds in Corollary I.2 to the condition number bounds for sparse recovery by ℓ_1-minimization from [23]. If $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is invertible, then Problem I.2 with $f(\boldsymbol{x}) = g(\boldsymbol{Dx})$ is mathematically equivalent to

$$\text{minimize} \quad g(\boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{\Omega D}^{-1}\boldsymbol{y} = \boldsymbol{b}. \tag{I.8}$$

In [23], the authors consider measurement matrices $\boldsymbol{\Omega}$ for which the rows $\boldsymbol{\omega}^T$ are sampled according to a distribution with covariance $\mathbb{E}[\boldsymbol{\omega\omega}^T]$. In the isotropic case where the covariance is a multiple of the identity matrix, the measurement ensemble in I.8 is non-isotropic and the covariance matrix has condition number proportional to $\kappa(\boldsymbol{D})^2$. In [23, Theorem 2], a lower bound on the number of measurements needed for recovering a signal is given that involves the condition number of the covariance matrix. The bounds in [23] apply directly to the number of measurements for recovery by ℓ_1-minimization, and under rather general assumptions on the distribution. Moreover, the bounds in [23] rely on the condition number restricted to sparse vectors, while in our case we consider Renegar's condition number with respect to the descent cone. The bounds in Corollary I.2 also apply to any convex regularizer, and their applicability to sparse recovery is via the proxy of the statistical dimension, and thus restricted to situations in which this parameter delivers recovery bounds.

While Renegar's condition number, defined by restricting the smallest singular value to a cone, can improve the bound, computing this condition number is not always practical. Using polarity (IV.4), we get the following version of the bound that ensures that the right-hand side is always bounded by n.

Corollary I.4. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone, and $\delta(C)$ the statistical dimension of C. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be non-singular. Then*

$$\delta(\boldsymbol{A}C) \le \kappa(\boldsymbol{A})^{-2} \cdot \delta(C) + \left(1 - \kappa(\boldsymbol{A})^{-2}\right) \cdot n.$$

If $f(\boldsymbol{x}) = g(\boldsymbol{Dx})$, where g is a proper convex function and $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ with $p \ge n$, then

$$\delta(f, \boldsymbol{x}_0) \le \kappa(\boldsymbol{D})^{-2} \cdot \delta(g, \boldsymbol{Dx}_0) + \left(1 - \kappa(\boldsymbol{D})^{-2}\right) \cdot n. \tag{I.9}$$

The simple proof of Corollary I.4 is given in Section V. One can interpret the upper bounds in Corollary I.4 as interpolating between the statistical dimension of C and the ambient dimension n.

Remark I.5. The restriction to invertible dictionaries $\boldsymbol{D}$ may look limiting at first, but a closer look reveals that it is not necessary when working with the subdifferential cone instead of the descent cone (see Section VI-A for the relevant definitions and background). In fact, given a proper convex function $f(\boldsymbol{x}) = g(\boldsymbol{Dx})$, the statistical dimensions of the descent cone and that of the subdifferential cone are related as

$$\delta(f, \boldsymbol{x}_0) = n - \delta(\text{cone}(\partial f(\boldsymbol{x}_0))).$$

Therefore, lower bounds on the statistical dimension of the subdifferential cone imply upper bounds on the statistical dimension of f. It is well known that $\text{cone}(\partial f(\boldsymbol{x}_0)) = \boldsymbol{D}^T \text{cone}(\partial g(\boldsymbol{Dx}_0))$, and therefore if $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ with $p \le n$, we can apply the lower bound from (I.7). In applications, however, the case $p \ge n$ is of interest. In this case one should note that the subdifferential cone is often contained in a linear subspace of dimension at most n, and by common invariance properties of the statistical dimension (Section IV-B) it is enough to work with the restriction of $\boldsymbol{D}^T$ to this lower dimensional subspace. Proposition I.6 illustrates this idea in the case of the 1-norm.

In the statement of the proposition below, we use the notation $\boldsymbol{A}_I$ for the submatrix of a matrix $\boldsymbol{A}$ with columns indexed by $I \subset [n] = \{1, \dots, n\}$, and denote by $I^c = [n] \setminus I$ the complement of I. The proof is postponed to Section VI.

Proposition I.6. *Let $\boldsymbol{D} \in \mathbb{R}^{p\times n}$, $p \ge n$, be such that all $n \times n$ minors of $\boldsymbol{D}$ have full rank, and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \le n$. Consider the problem*

$$\textit{minimize} \quad \|\boldsymbol{Dx}\|_1 \qquad \textit{subject to} \quad \boldsymbol{\Omega x} = \boldsymbol{b}. \tag{I.10}$$

Let $\boldsymbol{x}_0$ be such that $\boldsymbol{\Omega x}_0 = \boldsymbol{b}$, and such that $\boldsymbol{y}_0 = \boldsymbol{Dx}_0$ is s-sparse with support $I \subset [p]$. Let $\boldsymbol{C} \in \mathbb{R}^{n\times p-s+1}$ be a matrix whose first $p - s$ columns consist of the columns of $\boldsymbol{D}^T$ that are indexed by I^c, and the last column is $\boldsymbol{c}_{p-s+1} = \frac{1}{\sqrt{s}} \sum_{j\in I} \text{sign}((\boldsymbol{y}_0)_j)\boldsymbol{d}_j$, where the vectors $\boldsymbol{d}_j$ denote the columns of $\boldsymbol{D}^T$. Then

$$\delta(\|\boldsymbol{D}\cdot\|_1, \boldsymbol{x}_0) \le \kappa(\boldsymbol{C})^{-2} \cdot \delta(\|\cdot\|_1, \boldsymbol{Dx}_0) + \left(1 - (p/n)\kappa(\boldsymbol{C})^{-2}\right) \cdot n \tag{I.11}$$

In particular, given $\eta \in (0,1)$, Problem (I.10) with Gaussian measurement matrix succeeds with probability $1-\eta$ if

$$m \geq \kappa(\boldsymbol{C})^{-2} \cdot \delta(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0) + \left(1 - (p/n)\kappa(\boldsymbol{C})^{-2}\right) \cdot n + a_\eta \sqrt{n},$$

Example I.7. An illustrative example is the finite difference matrix $\boldsymbol{D}$ of example I.1. The regularizer $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$ is a one-dimensional version of a total variation regularizer, and is used to promote gradient sparsity. The standard method [7, Recipe 4.1] for computing the statistical dimension of the descent cone of f is not easily applicable here, as this regularizer is not separable [19] (in fact, it would require a careful analysis of the structure of the signal with sparse gradient to be recovered). The standard condition number bound Theorem A is also not applicable, as it is known that the condition number satisfies $\kappa(\boldsymbol{D}) \geq \frac{2(n+1)}{\pi}$. Figure 1 plots the upper bound of Proposition I.6 for signals with random support location and sparsity ranging from 1 to 200, and compares it to the actual statistical dimension computed by Monte Carlo simulation. As can be seen in this example, the upper bound is not very useful because of the large condition numbers involved.

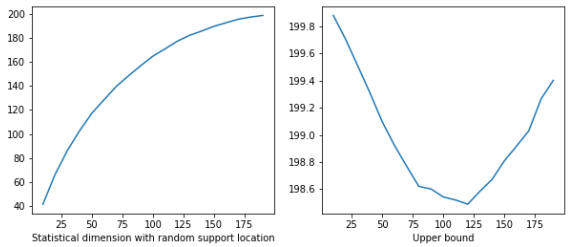

Fig. 1: The statistical dimension of $\|\boldsymbol{D}\cdot\|_1$ for different sparsity levels and the upper bound (I.11).

Remark I.8. It is natural to ask for which dictionaries $\boldsymbol{D}$ Proposition I.6 gives good bounds. This clearly also depends on the support of the signal one wishes to recover. A closer look at the matrix $\boldsymbol{C}$ in the case of the finite difference matrix and for *monotonely increasing* signals shows that $\boldsymbol{C}$ is (up to rows of zeros) itself a finite difference matrix of order $n-s+1$, and the quality of the bounds increases with the size of the support. Another natural example is when $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ is a Gaussian random matrix (that is, a matrix whose entries are independent standard normal distributed random variables). In this case, the invariance properties of Gaussians imply that the matrix $\boldsymbol{C}$ is again a Gaussian matrix in $\mathbb{R}^{n\times p-s+1}$. For such matrices, the condition number is known to be of order $(\sqrt{n}+\sqrt{p-s+1})/(\sqrt{n}-\sqrt{p-s+1})$ with high probability, see for example [24, Theorem 5.32]. In this example we see again that if the support is large, $s \approx p$, then the condition number is close to 1 and the bound becomes useful.

Note that so far we have seen two types of bounds: those based on the upper bound using Renegar's condition number in Theorem A, which improve on the standard condition number by using the cone-restricted smallest singular value, and those based on duality and the lower bound of Theorem A. The latter only work using the standard matrix condition number, but apply to the matrix restricted to the subspace generated by the cone of interest. Both bounds could yield good results for tight frames / well-conditioned matrices, but fail to give useful bounds in cases such as the finite difference matrix, or for redundant dictionaries $\boldsymbol{D}$ for which the statistical dimension $\delta(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0)$ is proportional to the (larger) ambient dimension. In the next section we discuss randomized improvements.

C. Main results - probabilistic bounds

While Corollary I.4 ensures that the upper bound does not become completely trivial, when $\boldsymbol{D}$ is ill-conditioned it still does not give satisfactory results, as seen in Example I.1. The second part, and main contribution, of our work is an improvement of the condition bounds using randomization: using methods from conic integral geometry, we derive a "preconditioned" version of Theorem A. The idea is based on the philosophy that a randomly oriented convex cone C ought to behave roughly like a linear subspace of dimension $\delta(C)$. In that sense, the statistical dimension of a cone C should be approximately invariant under projecting C to a subspace of dimension close to $\delta(C)$. In fact, in Section IV-E we will see that for $n \geq m \gtrapprox \delta(C)$, we have

$$\mathbb{E}_{\boldsymbol{Q}}\left[\delta(\boldsymbol{P}_m \boldsymbol{Q} C)\right] \approx \delta(C),$$

where $\boldsymbol{P}_m$ is the projection on the the first m coordinates and where the expectation is with respect to a random orthogonal matrix $\boldsymbol{Q}$, distributed according to the normalized Haar measure on the orthogonal group. From this it follows that the condition bounds should ideally depend not on the conditioning of $\boldsymbol{D}$ itself, but on a generic projection of $\boldsymbol{D}$ to linear subspace of dimension of order $\delta(C)$. For $m \leq n$ define

$$\overline{\kappa}_m^2(\boldsymbol{A}) := \mathbb{E}_{\boldsymbol{Q}}[\kappa(\boldsymbol{P}_m \boldsymbol{Q}\boldsymbol{A})^2], \quad \overline{\mathscr{R}}_{C,m}^2(\boldsymbol{A}) := \mathbb{E}_{\boldsymbol{Q}}\left[\mathscr{R}_C(\boldsymbol{P}_m \boldsymbol{Q}\boldsymbol{A})^2\right].$$

Theorem B. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone and $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ be a matrix of full rank. Let $\eta \in (0,1)$ and assume that $m \geq \delta(C) + 2\sqrt{\log(2/\eta)m}$. Then*

$$\delta(\boldsymbol{A}C) \leq \overline{\mathscr{R}}_{C,m}^2(\boldsymbol{A}) \cdot \delta(C) + (n-m)\eta.$$

For the matrix condition number,

$$\delta(\boldsymbol{A}C) \leq \overline{\kappa}_m^2(\boldsymbol{A}) \cdot \delta(C) + (n-m)\eta. \tag{I.12}$$

As a consequence of Theorem B we get the following preconditioned version of the previous bounds.

Corollary I.9. *Let $f(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$, where g is a proper convex function and $\boldsymbol{D} \in \mathbb{R}^{n\times n}$ is non-singular. Let $\eta \in (0,1)$ and assume that $m \geq \delta(g, \boldsymbol{D}\boldsymbol{x}_0) + 2\sqrt{\log(2/\eta)m}$. Then*

$$\delta(f, \boldsymbol{x}_0) \leq \overline{\mathscr{R}}_{\mathscr{D}(g,\boldsymbol{D}\boldsymbol{x}_0),m}^2(\boldsymbol{D}^{-1}) \cdot \delta(g, \boldsymbol{D}\boldsymbol{x}_0) + (n-m)\eta \tag{I.13}$$

and

$$\delta(f, \boldsymbol{x}_0) \le \overline{\kappa}_m^2(\boldsymbol{D}^{-1}) \cdot \delta(g, \boldsymbol{D}\boldsymbol{x}_0) + (n-m)\eta.$$

Example I.10. Consider a diagonal matrix $\boldsymbol{\Sigma}$ and the average condition $\overline{\kappa}_m^2(\boldsymbol{\Sigma})$. Intuitively, the average condition measures the expected eccentricity of the projection of an ellipsoid to a random subspace.

Example I.11. Using the finite difference matrix $\boldsymbol{D}$ from Example I.1, note that it is physically not possible, nor do we aim to, locate the precise phase transition for the recovery with $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$ in terms of that of the 1-norm, since the statistical dimension $\delta(f, \boldsymbol{x}_0)$ does not only depend on the sparsity pattern of $\boldsymbol{D}\boldsymbol{x}_0$, but also on the location of the support.

D. Scope and limits of reduction

The condition bounds in Theorem B naturally lead to the question of how to compute or bound the condition number of a random projection of a matrix,

$$\kappa(\boldsymbol{P}_m\boldsymbol{Q}\boldsymbol{A}) \quad \text{or} \quad \mathscr{R}_C(\boldsymbol{P}_m\boldsymbol{Q}\boldsymbol{A})$$

where $\boldsymbol{Q} \in O(n)$ is a random orthogonal matrix. If $m = \lfloor \rho n \rfloor$ with $\rho \in (0,1)$, then in some cases the condition number $\kappa(\boldsymbol{P}_m\boldsymbol{Q}\boldsymbol{A})$ remains bounded with high probability as $n \to \infty$. Below we sketch how such condition numbers can be bounded.

In what follows, let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be fixed and non-singular, and we write $\boldsymbol{Q}_m = \boldsymbol{P}_m\boldsymbol{Q}$ for a random matrix with orthogonal rows, uniformly distributed on the Stiefel manifold. We first reduce to the case of Gaussian matrices, for which tools are readily available. If $\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{1})$ is an $m \times n$ random matrix with Gaussian entries, then $\boldsymbol{Q}_m = (\boldsymbol{G}\boldsymbol{G}^T)^{-1/2}\boldsymbol{G}$ is uniformly distributed on the Stiefel manifold, so that $\mathscr{R}_C(\boldsymbol{Q}_m\boldsymbol{A})$ has the same distribution as $\mathscr{R}_C((\boldsymbol{G}\boldsymbol{G}^T)^{-1/2}\boldsymbol{G}\boldsymbol{A})$. Using Lemma II.7, we can bound (with probability one)

$$\mathscr{R}_C\left((\boldsymbol{G}\boldsymbol{G}^T)^{-1/2}\boldsymbol{G}\boldsymbol{A}\right) \le \kappa\left((\boldsymbol{G}\boldsymbol{G}^T)^{-1/2}\right)\mathscr{R}_C\left(\boldsymbol{G}\boldsymbol{A}\right) = \kappa(\boldsymbol{G})\mathscr{R}_C\left(\boldsymbol{G}\boldsymbol{A}\right),$$

transforming the problem into one in which the orthogonal matrix is replaced with a Gaussian one. The are different ways to estimate such condition numbers, the approach taken here is based on Gordon's inequality. We restrict the analysis to the classical matrix condition number, a more refined analysis using Renegar's condition number is likely to incorporate the Gaussian width of the cone. Moreover, using the invariance of the condition number under transposition, we consider $\kappa(\boldsymbol{A}\boldsymbol{G})$ with a $n \times m$ matrix $\boldsymbol{G}$, $m \le n$. An alternative, suggested by Armin Eftekhari, would be to appeal to the Hanson-Wright inequality [25], [26], or more directly, the Bernstein inequality.

Proposition I.12. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n\times m}$, with $m \le n$. Then*

$$\mathbb{E}[\kappa(\boldsymbol{A}\boldsymbol{G})] \le \frac{\|\boldsymbol{A}\|_F + \sqrt{m}\|\boldsymbol{A}\|_2}{\|\boldsymbol{A}\|_F - \sqrt{m}\|\boldsymbol{A}\|_2} \tag{I.14}$$

whenever $\|\boldsymbol{A}\|_F \ge \sqrt{m}\|\boldsymbol{A}\|_2$.

Using a standard procedure one can show that the singular value and the norm will stay close to their expected values with high probability. More specifically, one can use the above proposition as a basis for a weak average-case analysis of Renegar's condition number for random matrices of the form $\boldsymbol{A}\boldsymbol{G}$, as in [27].

Proof. We will derive the inequalities

$$\|\boldsymbol{A}\|_F - \sqrt{m}\|\boldsymbol{A}\|_2 \le \mathbb{E}[\sigma(\boldsymbol{A}\boldsymbol{G})] \le \mathbb{E}[\|\boldsymbol{A}\boldsymbol{G}\|_2] \le \|\boldsymbol{A}\|_F + \sqrt{m}\|\boldsymbol{A}\|_2.$$

where σ denotes the smallest singular value. We will restrict to showing the lower bound, the upper bound follows similarly by using Slepian's inequality. Without lack of generality assume $\boldsymbol{A} = \boldsymbol{\Sigma}$ is diagonal, with entries $\sigma_1 \ge \cdots \ge \sigma_n$ on the diagonal, and assume $\sigma_1 = 1$. Define the Gaussian processes

$$X_{\boldsymbol{x},\boldsymbol{y}} = \langle \boldsymbol{G}\boldsymbol{x}, \boldsymbol{\Sigma}\boldsymbol{y}\rangle, \quad Y_{\boldsymbol{x},\boldsymbol{y}} = \langle \boldsymbol{g}, \boldsymbol{x}\rangle + \langle \boldsymbol{h}, \boldsymbol{\Sigma}\boldsymbol{y}\rangle,$$

indexed by $\boldsymbol{x} \in S^{m-1}$, $\boldsymbol{y} \in S^{n-1}$, with $\boldsymbol{g} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \mathbb{R}^n$ Gaussian vectors. We get

$$\mathbb{E}[(X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'})^2] = \|\boldsymbol{\Sigma}\boldsymbol{y}\|^2 + \|\boldsymbol{\Sigma}\boldsymbol{y}'\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{x}'\rangle\langle \boldsymbol{\Sigma}\boldsymbol{y}, \boldsymbol{\Sigma}\boldsymbol{y}'\rangle,$$
$$\mathbb{E}[(Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'})^2] = \|\boldsymbol{\Sigma}\boldsymbol{y}\|^2 + \|\boldsymbol{\Sigma}\boldsymbol{y}'\|^2 + 2 - 2\langle \boldsymbol{x}, \boldsymbol{x}'\rangle - 2\langle \boldsymbol{\Sigma}\boldsymbol{y}, \boldsymbol{\Sigma}\boldsymbol{y}'\rangle,$$

so that

$$\begin{aligned}\mathbb{E}[(Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'})^2] - \mathbb{E}[(X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'})^2] &= 2(1 - \langle \boldsymbol{x}, \boldsymbol{x}'\rangle)(1 - \langle \boldsymbol{\Sigma}\boldsymbol{y}, \boldsymbol{\Sigma}\boldsymbol{y}'\rangle) \\ &\ge 0.\end{aligned}$$

This expression is 0 if $\boldsymbol{x} = \boldsymbol{x}'$, and non-negative otherwise, since by assumption $\boldsymbol{\Sigma}$ has largest entry equal to 1. We can therefore apply Gordon's Theorem (see Section [1, 9.2] or [28, Theorem B.1]) to infer an inequality

$$\begin{aligned}\mathbb{E}[\sigma(\boldsymbol{\Sigma}\boldsymbol{G})] &= \mathbb{E}[\min_{\boldsymbol{x}\in S^{m-1}} \max_{\boldsymbol{y}\in S^{n-1}} \langle \boldsymbol{G}\boldsymbol{x}, \boldsymbol{\Sigma}\boldsymbol{y}\rangle] \\ &= \mathbb{E}[\min_{\boldsymbol{x}\in S^{m-1}} \max_{\boldsymbol{y}\in S^{n-1}} X_{\boldsymbol{x},\boldsymbol{y}}] \\ &\ge \mathbb{E}[\min_{\boldsymbol{x}\in S^{m-1}} \max_{\boldsymbol{y}\in S^{n-1}} Y_{\boldsymbol{x},\boldsymbol{y}}] \\ &= \|\boldsymbol{\Sigma}\|_F - \sqrt{m}.\end{aligned}$$

In general, if $\sigma_1 \ne 1$, we replace $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma}/\|\boldsymbol{\Sigma}\|_2 = \boldsymbol{\Sigma}/\|\boldsymbol{A}\|_2$, and obtain the desired bound. □

It would be interesting to characterize those matrices $\boldsymbol{A}$ for which $\kappa(\boldsymbol{P}_m\boldsymbol{Q}\boldsymbol{A}) \approx 1$ using a kind of restricted isometry property, as for example in [29]. We leave a detailed discussion of the probability distribution of $\kappa(\boldsymbol{P}_m\boldsymbol{Q}\boldsymbol{A})$ and its ramifications for another occasion, and instead consider a special case.

Example I.13. Consider again the matrix $\boldsymbol{D}$ from Example I.1. For $\rho \in \{0.1, 0.2, 0.3, 0.4\}$ and n ranging from 1 to 400, $m = \lfloor \rho n \rfloor$, we plot the average condition number $\kappa(\boldsymbol{D}\boldsymbol{G})$, where $\boldsymbol{G} \in \mathbb{R}^{n\times m}$ is a Gaussian random matrix. As n increases, this condition number appears to converge to a constant value. We also plot the condition number $\kappa(\boldsymbol{D}^{-1}\boldsymbol{G})$, where $\boldsymbol{D}^{-1}$ is the upper triangular matrix with non-zero entries -1. The different decay of the singular values leads to condition number that increase with n.

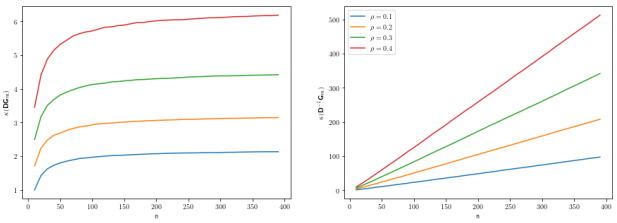

Fig. 2: Condition number $\kappa(\boldsymbol{G}_m\boldsymbol{D})$ for the matrix $\boldsymbol{D}$ from Example I.1, and for its inverse. $\boldsymbol{G}_m$ is the projection to the first $m = \lfloor \rho n \rfloor$ coordinates of a Gaussian $n \times n$ matrix $\boldsymbol{G}$

As we saw in Example I.1, the operator norm of $\boldsymbol{D}$ is bounded by $\|\boldsymbol{\sigma}\|_\infty \le 2$. The Frobenius norm, on the other hand, is easily seen to be $\|\boldsymbol{D}\|_F = \|\sigma\|_2 = \sqrt{2n-1}$. Setting $m = \rho n$, the condition number thus concentrates on a value bounded by

$$\frac{\sqrt{2n-1}+2\sqrt{m}}{\sqrt{2n-1}-2\sqrt{m}} \approx \frac{1+\sqrt{2\rho}}{1-\sqrt{2\rho}},$$

which is sensible if $\rho < 1/2$. We remark that, by construction, the bounds are not sharp, and also do not apply to the inverse $\boldsymbol{D}^{-1}$.

1) A note on applicability: The previous discussion has shown that the condition number bounds need only consider the restricted condition number of a random projection of a matrix, rather than the full matrix condition. However, as the bounds are multiplicative, even small values (for example, 2) lead to bounds for the the statistical dimension of the transformed cone that may not be practical. In addition, the statistical dimension of the reference cone also determines how small the projected dimension m is allowed to become, further limiting the amount of potential reduction in condition. If, for example, C is the descent cone of the ℓ_1-norm, then the resulting bounds can only be used for the descent cones of the ℓ_1-norm at very sparse vectors. The same applies when considering, instead of the difference matrix $\boldsymbol{D}$ and its inverse, diagonal matrices with various forms of decay in the entries (this corresponds to a version of weighted ℓ_1 recovery). In these cases, the expected condition of the randomly projected matrices can be improve dramatically, but still not enough to give non-trivial bounds across all sparsity levels. This limitation is inherent to the notion of condition number: Condition bounds are, by definition, pessimistic. In numerical analysis, they measure the worst case sensitivity of a problem to perturbations in the input. As such, it would be unrealistic to expect condition bounds to be able to accurately locate the statistical dimension of the descent cone of a composite regularizer, unless the matrix $\boldsymbol{D}$ involved is close to orthogonal.

2) A note on distributions: The results presented are based on integral geometry, and as such depend crucially on $\boldsymbol{Q}$ being uniformly distributed in the orthogonal group with the Haar measure. By known universality results [30], the results are likely to carry over to other distributions. In the context of this paper, however, we are neither interested in actually preconditioning the matrices involved, nor are we using them as a model for observation or measurement matrices as is common in compressive sensing. The randomization here is merely a technical tool to improve bounds based on the condition number, and the question of whether this is a "realistic" distribution is of no concern.

E. Organisation of the paper

In Section II we introduce the setting of conically restricted linear operators, the biconic feasibility problem, and Renegar's condition number in some detail. The characterization of this condition number in the generality presented here is new and of independent interest. Section III derives the main condition bound. In Section IV we change the scene and give a brief overview of conic integral geometry, culminating in a proof of Theorem B in Section V. Finally, in Section VI we translate the results to the setting of convex regularizers. Appendix A presents some more details on the biconic feasibility problem, while Appendix B presents a general version of Gordon's inequality. While this version is more general than what is needed in this paper, it may be of independent interest.

II. Conically restricted linear operators

In this section we discuss the restriction of a linear operator to closed convex cones and discuss Renegar's condition number in some detail.

A. Restricted norm and restricted singular value

Before discussing conically restricted operators in more detail, we record the following simple but useful lemma, which generalizes the relation $\ker \boldsymbol{A} = (\operatorname{im} \boldsymbol{A}^T)^\perp$.

Lemma II.1. *Let $D \subseteq \mathbb{R}^m$ be a closed convex cone. Then the polar cone is the inverse image of the origin under the projection map, $D^\circ := \{\boldsymbol{z} \in \mathbb{R}^m : \langle \boldsymbol{y}, \boldsymbol{z}\rangle \le 0 \text{ for all } \boldsymbol{y} \in D\} = \boldsymbol{\Pi}_D^{-1}(\boldsymbol{0})$. Furthermore, if $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, then*

$$\boldsymbol{A}^{-1}(D^\circ) = \left(\boldsymbol{A}^T D\right)^\circ, \tag{II.1}$$

where $\boldsymbol{A}^{-1}(D^\circ) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \in D^\circ\}$ denotes the inverse image of D° under $\boldsymbol{A}$.

Proof. For the first claim, note that $\|\boldsymbol{\Pi}_D(\boldsymbol{z})\| = \max_{\boldsymbol{y}\in D\cap B^m} \langle \boldsymbol{z}, \boldsymbol{y}\rangle$, and $\max_{\boldsymbol{y}\in D\cap B^m} \langle \boldsymbol{z}, \boldsymbol{y}\rangle = 0$ is equivalent to $\langle \boldsymbol{z}, \boldsymbol{y}\rangle \le 0$ for all $\boldsymbol{y} \in D$, i.e., $\boldsymbol{z} \in D^\circ$.

For (II.1), let $\boldsymbol{x} \in \boldsymbol{A}^{-1}(D^\circ)$ and $\boldsymbol{y} \in D$. Then $\langle \boldsymbol{x}, \boldsymbol{A}^T\boldsymbol{y}\rangle = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \le 0$, as $\boldsymbol{A}\boldsymbol{x} \in D^\circ$. Therefore, $\boldsymbol{A}^{-1}(D^\circ) \subseteq (\boldsymbol{A}^T D)^\circ$. On the other hand, if $\boldsymbol{v} \in (\boldsymbol{A}^T D)^\circ$ and $\boldsymbol{y} \in D$, then $\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{y}\rangle = \langle \boldsymbol{v}, \boldsymbol{A}^T\boldsymbol{y}\rangle \le 0$, so that $\boldsymbol{A}\boldsymbol{v} \in D^\circ$ and hence, $(\boldsymbol{A}^T D)^\circ \subseteq \boldsymbol{A}^{-1}(D^\circ)$. □

Recall from (I.3) that for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $C \subseteq \mathbb{R}^n$ and $D \subseteq \mathbb{R}^m$ closed convex cones, the restricted norm and singular value of $\boldsymbol{A}$ are defined by $\|\boldsymbol{A}\|_{C\to D} := \max\{\|\boldsymbol{\Pi}_D(\boldsymbol{A}\boldsymbol{x})\| : \boldsymbol{x} \in C \cap S^{n-1}\}$ and $\sigma_{C\to D}(\boldsymbol{A}) := \min\{\|\boldsymbol{\Pi}_D(\boldsymbol{A}\boldsymbol{x})\| : \boldsymbol{x} \in C \cap S^{n-1}\}$, respectively.

The following proposition provides geometric conditions for the vanishing of the restricted norm or singular value.

Proposition II.2. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $C \subseteq \mathbb{R}^n$ and $D \subseteq \mathbb{R}^m$ be closed convex cones. Then the restricted norm vanishes, $\|\boldsymbol{A}\|_{C\to D} = 0$, if and only if $C \subseteq (\boldsymbol{A}^T D)^\circ$. Furthermore, the restricted singular value vanishes, $\sigma_{C\to D}(\boldsymbol{A}) = 0$, if and only if $C \cap (\boldsymbol{A}^T D)^\circ \neq \{\boldsymbol{0}\}$, which is equivalent to $\boldsymbol{A}C \cap D^\circ \neq \{\boldsymbol{0}\}$ or* $\ker \boldsymbol{A} \cap C \neq \{\boldsymbol{0}\}$.

Proof. Using Lemma II.1 we have $\boldsymbol{\Pi}_D(\boldsymbol{Ax}) = \boldsymbol{0}$ if and only if $\boldsymbol{Ax} \in D^\circ$. This shows $\|\boldsymbol{A}\|_{C\to D} = 0$ if and only if $\boldsymbol{Ax} \in D^\circ$ for all $\boldsymbol{x} \in C \cap S^{n-1}$, or equivalently, $C \subseteq \boldsymbol{A}^{-1}(D^\circ) = (\boldsymbol{A}^T D)^\circ$ by (II.1). The claim about the restricted singular value follows similarly: $\sigma_{C\to D}(\boldsymbol{A}) = 0$ if and only if $\boldsymbol{Ax} \in D^\circ$ for some $\boldsymbol{x} \in C \cap S^{n-1}$, or equivalently, $C \cap \boldsymbol{A}^{-1}(D^\circ) \neq \{\boldsymbol{0}\}$. If $\boldsymbol{x} \in C \cap \boldsymbol{A}^{-1}(D^\circ) \setminus \{\boldsymbol{0}\}$, then either $\boldsymbol{Ax}$ is nonzero or $\boldsymbol{x}$ lies in the kernel of $\boldsymbol{A}$, which shows the second characterization. □

It is easily seen that the restricted norm is symmetric $\|\boldsymbol{A}\|_{C\to D} = \|\boldsymbol{A}^T\|_{D\to C}$,

$$\begin{aligned}\|\boldsymbol{A}\|_{C\to D} &= \max_{\boldsymbol{x}\in C\cap B^m} \max_{\boldsymbol{y}\in D\cap B^n} \langle \boldsymbol{Ax}, \boldsymbol{y}\rangle \\ &= \max_{\boldsymbol{y}\in D\cap B^n} \max_{\boldsymbol{x}\in C\cap B^m} \langle \boldsymbol{A}^T\boldsymbol{y}, \boldsymbol{x}\rangle \\ &= \|\boldsymbol{A}^T\|_{D\to C}.\end{aligned} \tag{II.2}$$

Such a relation does not hold in general for the restricted singular value. In fact, in Section II-B we will see that, unless $C = D = \mathbb{R}^n$, the minimum of $\sigma_{C\to D}(\boldsymbol{A})$ and $\sigma_{D\to C}(-\boldsymbol{A}^T)$ is always zero, if C and D have nonempty interior, cf. (II.5). And if C or D is a linear subspace then $\sigma_{D\to C}(-\boldsymbol{A}^T) = \sigma_{D\to C}(\boldsymbol{A}^T)$.

Remark II.3. In the case $C = \mathbb{R}^n$, $D = \mathbb{R}^m$, with $m \geq n$, one can characterize the smallest singular value of $\boldsymbol{A}$ as the inverse of the norm of the (Moore-Penrose) *pseudoinverse* of $\boldsymbol{A}$:

$$\sigma(\boldsymbol{A}) = \|\boldsymbol{A}^\dagger\|^{-1}.$$

Such a characterization does *not* hold in general for the restricted singular value, i.e., in general one cannot write $\sigma_{C\to D}(\boldsymbol{A})$ as $\|\boldsymbol{A}^\dagger\|^{-1}_{D\to C}$. Consider for example the case $D = \mathbb{R}^m$ and C a circular cone of angle α around some center $\boldsymbol{p} \in S^{n-1}$. Both cones have nonempty interior, but letting α go to zero, it is readily seen that $\sigma_{C\to D}(\boldsymbol{A})$ tends to $\|\boldsymbol{Ap}\|$, while $\|\boldsymbol{A}^\dagger\|_{D\to C}$ tends to $\|\boldsymbol{p}^T\boldsymbol{A}^\dagger\|$, which is in general not equal to $\|\boldsymbol{Ap}\|^{-1}$, unless $\boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_n$.

B. The biconic feasibility problem

The convex feasibility problem in the setting with two nonzero closed convex cones $C \subseteq \mathbb{R}^n$, $D \subseteq \mathbb{R}^m$ is given as:

$$\exists \boldsymbol{x} \in C \setminus \{\boldsymbol{0}\} \quad \text{s.t.} \quad \boldsymbol{Ax} \in D^\circ, \tag{P}$$

$$\exists \boldsymbol{y} \in D \setminus \{\boldsymbol{0}\} \quad \text{s.t.} \quad -\boldsymbol{A}^T\boldsymbol{y} \in C^\circ. \tag{D}$$

Using Lemma II.1 and Proposition II.2 we obtain the following characterizations of the primal feasible matrices $\mathscr{P}(C,D) := \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \text{(P) is feasible}\}$,

$$\begin{aligned}\mathscr{P}(C,D) &\overset{\text{(II.1)}}{=} \left\{\boldsymbol{A} \in \mathbb{R}^{m\times n} : C \cap \left(\boldsymbol{A}^T D\right)^\circ \neq \{\boldsymbol{0}\}\right\} \\ &\overset{\text{[Prop. II.2]}}{=} \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \sigma_{C\to D}(\boldsymbol{A}) = 0\}.\end{aligned} \tag{II.3}$$

By symmetry, we obtain for the dual feasible matrices $\mathscr{D}(C,D) := \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \text{(D) is feasible}\}$,

$$\begin{aligned}\mathscr{D}(C,D) &= \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : D \cap (-\boldsymbol{A}C)^\circ \neq \{\boldsymbol{0}\}\} \\ &= \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \sigma_{D\to C}(-\boldsymbol{A}^T) = 0\}.\end{aligned} \tag{II.4}$$

In fact, we will see that $\sigma_{C\to D}(\boldsymbol{A})$ and $\sigma_{D\to C}(-\boldsymbol{A}^T)$ can be characterized as the distances to $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$, respectively. We defer the proofs for this section to Appendix A.

In the following proposition we collect some general properties of $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$.

Proposition II.4. *Let $C \subseteq \mathbb{R}^n$, $D \subseteq \mathbb{R}^m$ be closed convex cones with nonempty interior. Then*

1) *$\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$ are closed;*
2) *the union of these sets is given by*

$$\mathscr{P}(C,D) \cup \mathscr{D}(C,D) = \begin{cases} \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \det \boldsymbol{A} = 0\} & C = D = \mathbb{R}^n \\ \mathbb{R}^{m\times n} & \text{else;} \end{cases}$$

3) *the intersection $\mathscr{P}(C,D) \cap \mathscr{D}(C,D)$ is nonempty but has Lebesgue measure zero.*

Note that from (2) and the characterizations (II.3) and (II.4) of $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$, respectively, we obtain for every $\boldsymbol{A} \in \mathbb{R}^{m\times n}$: $\min\{\sigma_{C\to D}(\boldsymbol{A}), \sigma_{D\to C}(-\boldsymbol{A}^T)\} = 0$ or, equivalently,

$$\max\left\{\sigma_{C\to D}(\boldsymbol{A}), \sigma_{D\to C}(-\boldsymbol{A}^T)\right\} = \sigma_{C\to D}(\boldsymbol{A}) + \sigma_{D\to C}(-\boldsymbol{A}^T), \tag{II.5}$$

unless $C = D = \mathbb{R}^n$.

In the following we simplify the notation by writing $\mathscr{P}, \mathscr{D}$ instead of $\mathscr{P}(C,D), \mathscr{D}(C,D)$. For the announced interpretation of the restricted singular value as distance to $\mathscr{P}, \mathscr{D}$ we introduce the following notation: for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ define

$$\operatorname{dist}(\boldsymbol{A}, \mathscr{P}) := \min\{\|\boldsymbol{\Delta}\| : \boldsymbol{A} + \boldsymbol{\Delta} \in \mathscr{P}\}, \quad \operatorname{dist}(\boldsymbol{A}, \mathscr{D}) := \min\{\|\boldsymbol{\Delta}\| : \boldsymbol{A} + \boldsymbol{\Delta} \in \mathscr{D}\},$$

where as usual, the norm considered is the operator norm. The proof of the following proposition, given in Appendix A, follows along the lines of similar derivations in the case with a cone and a linear subspace [31].

Proposition II.5. *Let $C \subseteq \mathbb{R}^n$, $D \subseteq \mathbb{R}^m$ nonzero closed convex cones with nonempty interior. Then*

$$\begin{aligned}\operatorname{dist}(\boldsymbol{A}, \mathscr{P}) &= \sigma_{C\to D}(\boldsymbol{A}), \\ \operatorname{dist}(\boldsymbol{A}, \mathscr{D}) &= \sigma_{D\to C}(-\boldsymbol{A}^T).\end{aligned}$$

We finish this section by considering the intersection of $\mathscr{P}$ and $\mathscr{D}$, which we denote by

$$\Sigma(C,D) := \mathscr{P}(C,D) \cap \mathscr{D}(C,D),$$

or simply Σ when the cones are clear from context. This set is usually referred to as the set of *ill-posed inputs*. As shown in Proposition II.4, the set of ill-posed inputs, assuming $C \subseteq \mathbb{R}^n$ and $D \subseteq \mathbb{R}^m$ each have nonempty interior, is a nonempty zero volume set. In the special case $C = \mathbb{R}^n$, $D = \mathbb{R}^m$,

$$\Sigma(\mathbb{R}^n, \mathbb{R}^m) = \{\text{rank deficient matrices in } \mathbb{R}^{m\times n}\}.$$

From (II.5) and Proposition II.5 we obtain, if $(C,D) \neq (\mathbb{R}^n, \mathbb{R}^n)$,

$$\begin{aligned}\operatorname{dist}(\boldsymbol{A}, \Sigma) &= \max\big\{\operatorname{dist}(\boldsymbol{A}, \mathscr{P}), \operatorname{dist}(\boldsymbol{A}, \mathscr{D})\big\} \\ &= \operatorname{dist}(\boldsymbol{A}, \mathscr{P}) + \operatorname{dist}(\boldsymbol{A}, \mathscr{D}).\end{aligned}$$

The inverse distance to ill-posedness forms the heart of Renegar's condition number [32], [14]. We denote

$$\begin{aligned}\mathscr{R}_{C,D}(\boldsymbol{A}) &:= \frac{\|\boldsymbol{A}\|}{\operatorname{dist}(\boldsymbol{A}, \Sigma(C,D))} \\ &= \min\left\{\frac{\|\boldsymbol{A}\|}{\sigma_{C\to D}(\boldsymbol{A})}, \frac{\|\boldsymbol{A}\|}{\sigma_{D\to C}(-\boldsymbol{A}^T)}\right\}.\end{aligned} \tag{II.6}$$

Furthermore, we abbreviate the special case $D = \mathbb{R}^m$, which corresponds to the classical feasibility problem, by the notation

$$\mathscr{R}_C(\boldsymbol{A}) := \mathscr{R}_{C,\mathbb{R}^m}(\boldsymbol{A}). \tag{II.7}$$

Note that the usual matrix condition number is recovered in the case $C = \mathbb{R}^n$, $D = \mathbb{R}^m$,

$$\mathscr{R}_{\mathbb{R}^n}(\boldsymbol{A}) = \mathscr{R}_{\mathbb{R}^n,\mathbb{R}^m}(\boldsymbol{A}) = \kappa(\boldsymbol{A}).$$

Another simple but useful property is the symmetry $\mathscr{R}_{C,D}(\boldsymbol{A}) = \mathscr{R}_{D,C}(-\boldsymbol{A}^T)$. Finally, note that the restricted singular value has the following monotonicity properties

$$\begin{aligned}C \subseteq C' &\Rightarrow \sigma_{C\to D}(\boldsymbol{A}) \geq \sigma_{C'\to D}(\boldsymbol{A}), \\ D \subseteq D' &\Rightarrow \sigma_{C\to D}(\boldsymbol{A}) \leq \sigma_{C\to D'}(\boldsymbol{A}).\end{aligned}$$

This indicates that not necessarily $\mathscr{R}_C(\boldsymbol{A}) \leq \mathscr{R}_{C'}(\boldsymbol{A})$ if $C \subseteq C'$. But in the case $C' = \mathbb{R}^n$ and $m \geq n$ this inequality does hold, which we formulate in the following lemma.

Lemma II.6. *Let $C \subseteq \mathbb{R}^n$ closed convex cone with nonempty interior and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$. Then*

$$\mathscr{R}_C(\boldsymbol{A}) \leq \kappa(\boldsymbol{A}). \tag{II.8}$$

Proof. In the case $C = \mathbb{R}^n$ we have $\mathscr{R}_{\mathbb{R}^n}(\boldsymbol{A}) = \kappa(\boldsymbol{A})$. If $C \neq \mathbb{R}^n$ then $\boldsymbol{A}C \neq \mathbb{R}^m$, as $m \geq n$. It follows that $\mathbb{R}^m \cap (-\boldsymbol{A}C)^\circ \neq \{\boldsymbol{0}\}$, and thus $\sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T) = 0$, cf. (II.4). Hence,

$$\mathscr{R}_C(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|}{\sigma_{C\to\mathbb{R}^m}(\boldsymbol{A})} \leq \frac{\|\boldsymbol{A}\|}{\sigma_{\mathbb{R}^n\to\mathbb{R}^m}(\boldsymbol{A})} = \kappa(\boldsymbol{A}). \qquad \square$$

To conclude this section, we state a useful bound on the condition number of a product of matrices.

Lemma II.7. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \leq n$ and let $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ be nonsingular. Then*

$$\mathscr{R}_C(\boldsymbol{B}\boldsymbol{A}) \leq \kappa(\boldsymbol{B}) \cdot \mathscr{R}_C(\boldsymbol{A}).$$

Proof. We need to bound the numerator from above and the denominator from below in the definition of Regenar's condition number (II.6). For the norms we have $\|\boldsymbol{B}\boldsymbol{A}\| \leq \|\boldsymbol{B}\| \cdot \|\boldsymbol{A}\|$. If $\sigma_C(\boldsymbol{B}\boldsymbol{A}) = \sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T\boldsymbol{B}^T) = 0$, then clearly also $\sigma_C(\boldsymbol{A}) = \sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T) = 0$. Assume that $\sigma_C(\boldsymbol{B}\boldsymbol{A}) \neq 0$, and let $\boldsymbol{x} \in C \cap S^{n-1}$. Since $\boldsymbol{B}$ is non-singular, $\boldsymbol{A}\boldsymbol{x} \neq \boldsymbol{0}$ and set $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}/\|\boldsymbol{A}\boldsymbol{x}\|$. Then

$$\|\boldsymbol{B}\boldsymbol{A}\boldsymbol{x}\| = \|\boldsymbol{B}\boldsymbol{z}\| \cdot \|\boldsymbol{A}\boldsymbol{x}\| \geq \sigma(\boldsymbol{B}) \cdot \sigma_C(\boldsymbol{A}\boldsymbol{x}) \neq 0.$$

If $\sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T\boldsymbol{B}^T) \neq 0$, then if $\boldsymbol{x} \in S^{m-1}$ and $\boldsymbol{z} = \boldsymbol{B}^T\boldsymbol{x}/\|\boldsymbol{B}^T\boldsymbol{x}\|$, then

$$\begin{aligned}\|\boldsymbol{\Pi}_C(\boldsymbol{A}^T\boldsymbol{B}^T\boldsymbol{x})\| &= \|\boldsymbol{\Pi}_C(\boldsymbol{A}^T\boldsymbol{z})\| \cdot \|\boldsymbol{B}^T\boldsymbol{x}\| \\ &\geq \sigma(\boldsymbol{B}) \cdot \sigma_{\mathbb{R}^m\to C}(-\boldsymbol{A}^T) \neq 0.\end{aligned}$$

The condition bound follows. $\square$

III. Linear images of cones

The norm of the projection is a special case of a cone-restricted norm:

$$\|\boldsymbol{\Pi}_C(\boldsymbol{g})\| = \|\boldsymbol{g}\|_{\mathbb{R}_+\to C}, \tag{III.1}$$

where on the right-hand side we interpret $\boldsymbol{g} \in \mathbb{R}^{n\times 1}$ as linear map from $\mathbb{R}$ to $\mathbb{R}^n$. In this section we relate these norms for linear images of convex cones. The upper bound in Theorem III.1 is a special case of a more general bound for moment functionals [28, Proposition 3.9].

Theorem III.1. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone, and $\nu_r(C) := \mathbb{E}[\|\boldsymbol{\Pi}_C(\boldsymbol{g})\|^r]$, with $\boldsymbol{g} \in \mathbb{R}^n$ Gaussian. Then for $\boldsymbol{A} \in \mathbb{R}^{p\times n}$, and $r \geq 1$,*

$$\nu_r(\boldsymbol{A}C) \leq \mathscr{R}_C(\boldsymbol{A})^r \nu_r(C). \tag{III.2}$$

In particular, if $p \geq n$ and $\boldsymbol{A}$ has full rank, then

$$\frac{\delta(C)}{\kappa(\boldsymbol{A})^2} \leq \delta(\boldsymbol{A}C) \leq \kappa(\boldsymbol{A})^2 \delta(C). \tag{III.3}$$

The proof of Theorem III.1 relies on the following auxiliary result, Lemma III.2, and on a generalized form of Slepian's inequality, Theorem III.3.

Lemma III.2. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone and $\boldsymbol{A} \in \mathbb{R}^{p\times n}$. Then*

$$\frac{1}{\|\boldsymbol{A}\|}\boldsymbol{A}(C \cap B^n) \subseteq \boldsymbol{A}C \cap B^p \subseteq \tfrac{1}{\lambda}\boldsymbol{A}(C \cap B^n), \tag{III.4}$$

with $\lambda := \max\{\sigma_{C\to\mathbb{R}^p}(\boldsymbol{A}), \sigma_{\mathbb{R}^p\to C}(-\boldsymbol{A}^T)\}$.

Proof. For the lower inclusion, note that any $\boldsymbol{y} \in \frac{\boldsymbol{A}(C\cap B^n)}{\|\boldsymbol{A}\|}$ can be written as $\boldsymbol{y} = \frac{\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{A}\|}$, with $\boldsymbol{x} \in C \cap B^n$. Since $\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\|$, we have $\boldsymbol{y} \in \operatorname{conv}\left\{\boldsymbol{0}, \frac{\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{A}\boldsymbol{x}\|}\right\} \subset (\boldsymbol{A}C) \cap B^p$. which was to be shown.

For the upper inclusion, let $\lambda_1 := \sigma_{C\to\mathbb{R}^p}(\boldsymbol{A})$, $\lambda_2 := \sigma_{\mathbb{R}^p\to C}(-\boldsymbol{A}^T)$. We show in two steps that $\boldsymbol{A}C \cap B^p \subseteq \frac{1}{\lambda_1}\boldsymbol{A}(C \cap B^n)$ if $\lambda_1 > 0$ and $\boldsymbol{A}C \cap B^p \subseteq \frac{1}{\lambda_2}\boldsymbol{A}(C \cap B^n)$ if $\lambda_2 > 0$.

(1) Let $\lambda_1 > 0$. Since $\boldsymbol{A}C \cap B^p$ as well as $\boldsymbol{A}(C \cap B^n)$ contain the origin, it suffices to show that $\boldsymbol{A}C \cap S^{p-1} \subseteq \frac{1}{\lambda_1}\boldsymbol{A}(C \cap B^n)$. Every element in $\boldsymbol{A}C \cap S^{p-1}$ can be written as $\frac{\boldsymbol{A}\boldsymbol{y}_0}{\|\boldsymbol{A}\boldsymbol{y}_0\|}$ for some $\boldsymbol{y}_0 \in C \cap S^{n-1}$, and since $\sigma_{C\to\mathbb{R}^p}(\boldsymbol{A}) = \min_{\boldsymbol{y}\in C\cap S^{n-1}} \|\boldsymbol{A}\boldsymbol{y}\| \leq \|\boldsymbol{A}\boldsymbol{y}_0\|$, we obtain $\sigma_{C\to\mathbb{R}^p}(\boldsymbol{A})\frac{\boldsymbol{A}\boldsymbol{y}_0}{\|\boldsymbol{A}\boldsymbol{y}_0\|} \in \operatorname{conv}\{\boldsymbol{0}, \boldsymbol{A}\boldsymbol{y}_0\} \subseteq \boldsymbol{A}(C \cap B^n)$. This shows $\boldsymbol{A}C \cap S^{p-1} \subseteq \frac{1}{\lambda_1}\boldsymbol{A}(C \cap B^n)$.

(2) Let $\lambda_2 > 0$. Recall from (II.4) that $\lambda_2 = \sigma_{\mathbb{R}^p \to C}(-\boldsymbol{A}^T) > 0$ only if $(\boldsymbol{A}C)^\circ = \{\boldsymbol{0}\}$, i.e., $\boldsymbol{A}C = \mathbb{R}^p$. Observe that

$$\begin{aligned}\sigma_{\mathbb{R}^p \to C}(-\boldsymbol{A}^T) &= \min_{\boldsymbol{z}\in S^{p-1}} \max_{\boldsymbol{y}\in C\cap B^n} \langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{z}\rangle \\ &= \max\{r \geq 0 : rB^p \subseteq \boldsymbol{A}(C\cap B^n)\}.\end{aligned}$$

This shows $B^p \subseteq \frac{1}{\lambda_2}\boldsymbol{A}(C\cap B^n)$ and thus finishes the proof. □

The following generalization of Slepian's inequality is the special case of a generalized version of Gordon's inequality for Gaussian processes, [28, Theorem B.2], when setting $m = 1$ in that theorem.

Theorem III.3. *Let X_j, Y_j, $j \in \{0,\dots,n\}$, be centered Gaussian random variables, and assume that for all $j,k \geq 0$ we have*

$$\mathbb{E}|X_j - X_k|^2 \leq \mathbb{E}|Y_j - Y_k|^2.$$

Then for any monotonically increasing convex function $f\colon \mathbb{R}_+ \to \mathbb{R}$,

$$\mathbb{E}\max_j f_+(X_j - X_0) \leq \mathbb{E}\max_j f_+(Y_j - Y_0), \tag{III.5}$$

where $f_+(x) := f(x)$, if $x \geq 0$, and $f_+(x) := f(0)$, if $x \leq 0$.

Proof of Theorem III.1. Set $\lambda := \max\{\sigma_{C\to\mathbb{R}^p}(\boldsymbol{A}), \sigma_{\mathbb{R}^p\to C}(-\boldsymbol{A}^T)\}$. For the upper bound, note that by Lemma III.2 we have

$$\mathbb{E}[\|\boldsymbol{\Pi}_{\boldsymbol{A}C}(\boldsymbol{g})\|^r] = \mathbb{E}\left[\max_{\boldsymbol{x}\in\boldsymbol{A}C\cap B^p} \langle \boldsymbol{g}, \boldsymbol{x}\rangle^r\right] \leq \frac{1}{\lambda^r}\mathbb{E}\left[\max_{\boldsymbol{x}\in C\cap B^n} \langle \boldsymbol{g}, \boldsymbol{A}\boldsymbol{x}\rangle^r\right].$$

Let $\boldsymbol{g}$ be a standard Gaussian vector and consider the Gaussian processes $X_{\boldsymbol{x}} = \langle \boldsymbol{g}, \boldsymbol{A}\boldsymbol{x}\rangle$ and $Y_{\boldsymbol{x}} = \langle \boldsymbol{g}, \|\boldsymbol{A}\|\boldsymbol{x}\rangle$, indexed by $\boldsymbol{x} \in C\cap B^n$. For any $\boldsymbol{x},\boldsymbol{y} \in C\cap B^n$ we have

$$\mathbb{E}(X_{\boldsymbol{x}} - X_{\boldsymbol{y}})^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}\|^2 \leq \|\|\boldsymbol{A}\|\boldsymbol{x} - \|\boldsymbol{A}\|\boldsymbol{y}\|^2 = \mathbb{E}(Y_{\boldsymbol{x}} - Y_{\boldsymbol{y}})^2,$$

we get $\mathbb{E}(X_{\boldsymbol{x}} - X_{\boldsymbol{y}})^2 \leq \mathbb{E}(Y_{\boldsymbol{x}} - Y_{\boldsymbol{y}})^2$. From Theorem III.3 we conclude that for any finite subset $S \subset C\cap B^n$ containing the origin,

$$\mathbb{E}[\max_{\boldsymbol{x}\in S} X_{\boldsymbol{x}}^r] \leq \mathbb{E}[\max_{\boldsymbol{x}\in S} Y_{\boldsymbol{x}}^r].$$

By a standard compactness argument (see, e.g., [1, 8.6]), this extends to the whole index set $C\cap B^n$, which yields the inequalities

$$\begin{aligned}\nu_r(\boldsymbol{A}C) &= \mathbb{E}[\|\boldsymbol{\Pi}_{\boldsymbol{A}C}(\boldsymbol{g})\|^r] \\ &\leq \frac{1}{\lambda^r}\mathbb{E}\left[\max_{\boldsymbol{x}\in C\cap B^n} \langle \boldsymbol{g}, \boldsymbol{A}\boldsymbol{x}\rangle^r\right] \\ &\leq \frac{\|\boldsymbol{A}\|^r}{\lambda^r}\mathbb{E}\left[\max_{\boldsymbol{x}\in C\cap B^n} \langle \boldsymbol{g}, \boldsymbol{x}\rangle^r\right] \\ &= \mathscr{R}_C(\boldsymbol{A})^r \nu_r(C).\end{aligned}$$

The upper bound in terms of the usual matrix condition number follows courtesy of (II.8). The lower bound proceeds along the lines, with the roles of $\|\boldsymbol{A}\|$ and λ reversed. More specifically, from Lemma III.2 we get the inequality

$$\mathbb{E}[\|\boldsymbol{\Pi}_{\boldsymbol{A}C}(\boldsymbol{g})\|^r] \geq \frac{1}{\|\boldsymbol{A}\|^r}\mathbb{E}\left[\max_{\boldsymbol{x}\in C\cap B^n} \langle \boldsymbol{g}, \boldsymbol{A}\boldsymbol{x}\rangle^r\right].$$

Define

$$\sigma_{C-C}(\boldsymbol{A}) = \min_{\boldsymbol{z}\in S(C-C)} \|\boldsymbol{A}\boldsymbol{z}\|,$$

where $S(C-C) := \{(\boldsymbol{x}-\boldsymbol{y})/\|\boldsymbol{x}-\boldsymbol{y}\| : \boldsymbol{x}\in C\cap B^n, \boldsymbol{y}\in C\cap B^n, \boldsymbol{x}\neq\boldsymbol{y}\}$. Consider the processes $Y_{\boldsymbol{x}} = \langle \boldsymbol{g}, \boldsymbol{A}\boldsymbol{x}\rangle$ and $X_{\boldsymbol{x}} = \langle \boldsymbol{g}, \sigma_{C-C}(\boldsymbol{A})\boldsymbol{x}\rangle$ indexed by $\boldsymbol{x} \in C\cap B^n$. Then for distinct $\boldsymbol{x},\boldsymbol{y} \in C\cap B^n$,

$$\begin{aligned}\mathbb{E}(X_{\boldsymbol{x}} - X_{\boldsymbol{y}})^2 &= \|\sigma_{C-C}(\boldsymbol{A})\boldsymbol{x} - \sigma_{C-C}(\boldsymbol{A})\boldsymbol{y}\|^2 \\ &\leq \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}\|^2 = \mathbb{E}(Y_{\boldsymbol{x}} - Y_{\boldsymbol{y}})^2.\end{aligned}$$

We can now apply Slepian's inequality as we did for the upper bound, and conclude that

$$\mathbb{E}[\|\boldsymbol{\Pi}_{\boldsymbol{A}C}(\boldsymbol{g})\|^r] \geq \frac{\sigma_{C-C}(\boldsymbol{A})^r}{\|\boldsymbol{A}\|^r}\nu_r(C).$$

To finish the argument, note that we have $\sigma_{C-C}(\boldsymbol{A}) \geq \sigma(\boldsymbol{A})$. □

IV. Conic integral geometry

In this section we use integral geometry to develop the tools needed for deriving a preconditioned bound in Theorem B. A comprehensive treatment of integral geometry can be found in [33], while a self-contained treatment in the setting of polyhedral cones, which uses our language, is given in [34].

A. Intrinsic volumes

The theory of conic integral geometry is based on the *intrinsic volumes* $v_0(C),\dots,v_n(C)$ of a closed convex cone $C \subseteq \mathbb{R}^n$. The intrinsic volumes form a discrete probability distribution on $\{0,\dots,n\}$ that capture statistical properties of the cone C. For a polyhedral cone C and $0 \leq k \leq n$, the intrinsic volumes can be defined as

$$v_k(C) = \mathbb{P}\{\boldsymbol{\Pi}_C(\boldsymbol{g}) \in \operatorname{relint}(F), \quad \dim F = k\},$$

where F is a face of C and relint denotes the relative interior.

Example IV.1. Let $C = L \subseteq \mathbb{R}^n$ be a linear subspace of dimension i. Then

$$v_k(C) = \begin{cases} 1 & \text{if } k = i, \\ 0 & \text{if } k \neq i. \end{cases}$$

Example IV.2. Let $C = \mathbb{R}^n_{\geq 0}$ be the non-negative orthant, i.e., the cone consisting of points with non-negative coordinates. A vector $\boldsymbol{x}$ projects orthogonally to a k-dimensional face of C if and only if exactly k coordinates are non-positive. By symmetry considerations and the invariance of the Gaussian distribution under permutations of the coordinates, it follows that

$$v_k(\mathbb{R}^n_{\geq 0}) = \binom{n}{k} 2^{-n}.$$

For non-polyhedral closed convex cones, the intrinsic volumes can be defined by polyhedral approximation. To avoid having to explicitly take care of upper summation bounds in many formulas, we use the convention that $v_k(C) = 0$ if $C \subseteq \mathbb{R}^n$ and $k > n$ (that this is not just a

convention follows from the fact that intrinsic volumes are "intrinsic", i.e., not dependent on the dimension of the space in which C lives).

The following important properties of the intrinsic volumes, which are easily verified in the setting of polyhedral cones, will be used frequently:

(a) **Orthogonal invariance.** For an orthogonal transformation $\boldsymbol{Q} \in O(n)$,

$$v_k(\boldsymbol{Q}C) = v_k(C);$$

(b) **Polarity.**

$$v_k(C) = v_{n-k}(C^\circ);$$

(c) **Product rule.**

$$v_k(C \times D) = \sum_{i+j=k} v_i(C)\, v_j(D). \tag{IV.1}$$

In particular, if $D = L$ is a linear subspace of dimension j, then $v_{k+j}(C \times L) = v_k(C)$.

(d) **Gauss-Bonnet.**

$$\sum_{k=0}^{n} (-1)^k v_k(C) = \begin{cases} 0 & \text{if } C \text{ is not a linear subspace,} \\ 1 & \text{else.} \end{cases} \tag{IV.2}$$

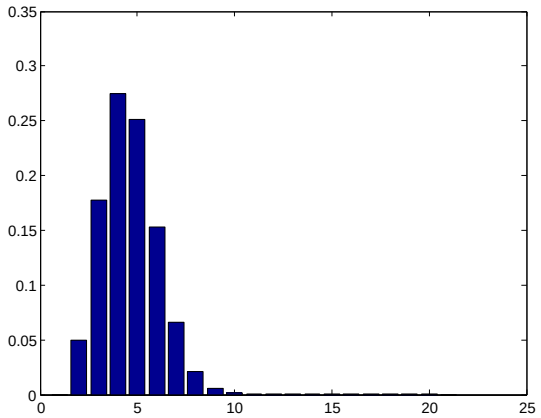

Fig. 3: Intrinsic volumes of the cone $C = \{\boldsymbol{x} : x_1 \le \cdots \le x_n\}$.

B. The statistical dimension

In what follows it will be convenient to work with reparametrizations of the intrinsic volumes, namely the tail and half-tail functionals

$$t_k(C) = \sum_{i \ge 0} v_{k+i}(C), \qquad h_k(C) = 2 \sum_{i \ge 0 \text{ even}} v_{k+i}(C),$$

which are defined for $0 \le k \le n$. Adding (or subtracting) the Gauss-Bonnet relation (IV.2) to the identity $\sum_{i \ge 0} v_i(C) = 1$, we see that $h_0(C) = h_1(C) = 1$ if C is not a linear subspace, so that the sequences $2v_0(C), 2v_2(C), \ldots$ and $2v_1(C), 2v_3(C), \ldots$ are probability distributions in their own right. Moreover, we have the interleaving property

$$t_{i+1}(C) \le h_i(C) \le t_i(C).$$

The intrinsic volumes can be recovered from the half-tail functionals as

$$v_i(C) = \begin{cases} \frac{1}{2}(h_i(C) - h_{i+2}(C)) & \text{for } 0 \le i \le n-2, \\ \frac{1}{2} h_i(C) & \text{else.} \end{cases} \tag{IV.3}$$

An important summary parameter is the *statistical dimension* of a cone C, defined as the expected value of the intrinsic volumes considered as probability distribution:

$$\delta(C) = \sum_{k=0}^{n} k v_k(C) = \frac{1}{2} h_1(C) + \sum_{i \ge 2} h_i(C).$$

The statistical dimension coincides with the expected squared norm of the projection of a Gaussian vector on the cone, $\delta(C) = \mathbb{E}\left[\|\boldsymbol{\Pi}_C(\boldsymbol{g})\|^2\right]$. Moreover, it differs from the squared Gaussian width by at most 1,

$$w^2(C) \le \delta(C) \le w^2(C) + 1,$$

see [7, Proposition 10.2].

The statistical dimension reduces to the usual dimension for linear subspaces, and also extends various properties of the dimension to closed convex cones $C \subseteq \mathbb{R}^n$:

(a) **Orthogonal invariance.** For an orthogonal transformation $\boldsymbol{Q} \in O(n)$,

$$\delta(\boldsymbol{Q}C) = \delta(C);$$

(b) **Complementarity.**

$$\delta(C) + \delta(C^\circ) = n; \tag{IV.4}$$

This generalizes the relation $\dim L + \dim L^\perp = n$ for a linear subspace $L \subseteq \mathbb{R}^n$.

(c) **Additivity.**

$$\delta(C \times D) = \delta(C) + \delta(D).$$

(d) **Monotonicity.**

$$\delta(C) \le \delta(D) \text{ if } C \subseteq D.$$

The analogy with linear subspaces will be taken further when discussing concentration of intrinsic volumes, see Section IV-D.

C. The kinematic formulas

The intrinsic volumes allow to study the properties of random intersections of cones via the *kinematic formulas*. A self-contained proof of these formulas for polyhedral cones is given in [34, Section 5]. In what follows, when we say that $\boldsymbol{Q}$ is drawn uniformly at random from the orthogonal group $O(d)$, we mean that it is drawn from the Haar probability measure ν on $O(n)$. This is the unique regular Borel measure on $O(n)$ that is left and right invariant ($\nu(\boldsymbol{Q}A) = \nu(A\boldsymbol{Q}) = \nu(A)$ for $\boldsymbol{Q} \in O(n)$ and a Borel measurable $A \subseteq O(n)$) and satisfies $\nu(O(n)) = 1$. Moreover, for measurable $f : O(n) \to \mathbb{R}_+$, we write

$$\mathbb{E}_{\boldsymbol{Q} \in O(n)}[f(\boldsymbol{Q})] := \int_{\boldsymbol{Q} \in O(n)} f(\boldsymbol{Q})\ \nu(\mathrm{d}\boldsymbol{Q})$$

for the integral with respect to the Haar probability measure, and we will occasionally omit the subscript

$\boldsymbol{Q} \in O(n)$, or just write $\boldsymbol{Q}$ in the subscript, when there is no ambiguity.

Theorem IV.3 (Kinematic Formula)**.** *Let $C, D \subseteq \mathbb{R}^n$ be polyhedral cones. Then, for $\boldsymbol{Q} \in O(n)$ uniformly at random, and $k > 0$,*

$$\mathbb{E}[v_k(C \cap \boldsymbol{Q}D)] = v_{k+n}(C \times D), \quad \mathbb{E}[v_0(C \cap \boldsymbol{Q}D)] = t_0(C \times D). \tag{IV.5}$$

If $D = L$ is a linear subspace of dimension $n - m$, then

$$\mathbb{E}[v_k(C \cap \boldsymbol{Q}L)] = v_{k+m}(C), \qquad \mathbb{E}[v_0(C \cap \boldsymbol{Q}L)] = \sum_{j=0}^{m} v_j(C). \tag{IV.6}$$

Combining Theorem IV.3 with the Gauss-Bonnet relation (IV.2) yields the so-called *Crofton formulas*, which we formulate in the following corollary. The intersection probabilities are also know as Grassmann angles in the literature (see [34, 2.33] for a discussion and references).

Corollary IV.4. *Let $C, D \subseteq \mathbb{R}^n$ be polyhedral cones such that not both of C and D are linear subspaces, and let $L \subset \mathbb{R}^n$ be a linear subspace of dimension $n - m$. Then, for $\boldsymbol{Q} \in O(n)$ uniformly at random,*

$$\mathbb{P}\{C \cap \boldsymbol{Q}D \neq \boldsymbol{0}\} = h_{n+1}(C \times D), \quad \mathbb{P}\{C \cap \boldsymbol{Q}L \neq \boldsymbol{0}\} = h_{m+1}(C).$$

Applying the polarity relation $(C \cap D)^\circ = C^\circ + D^\circ$ (see [34, Proposition 2.5]) to the kinematic formulas, we obtain a polar version of the kinematic formula, for $k > 0$,

$$\mathbb{E}[v_{n-k}(C + \boldsymbol{Q}D)] = v_{n-k}(C \times D), \quad \mathbb{E}[v_n(C + \boldsymbol{Q}D)] = t_n(C \times D). \tag{IV.7}$$

A convenient consequence of this polar form is a projection formula for intrinsic volumes, due to Glasauer [35]. Let $\boldsymbol{Q} \in O(n)$ uniform at random and $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ a fixed orthogonal projection onto a linear subspace L of dimension m. Then for $0 < k \leq m$,

$$\mathbb{E}[v_{m-k}(\boldsymbol{P}\boldsymbol{Q}C)] = v_{m-k}(C), \quad \mathbb{E}[v_m(\boldsymbol{P}\boldsymbol{Q}C)] = t_m(C). \tag{IV.8}$$

As we will see in Section IV-E, this results holds for *any* full rank $\boldsymbol{T} \in \mathbb{R}^{m \times n}$, instead of just for projections $\boldsymbol{P}$.

Remark IV.5. The astute reader may notice that the projection $\boldsymbol{P}\boldsymbol{Q}C$ does not need to be a closed convex cone. For random $\boldsymbol{Q}$, however, the probability of this happening can be shown to be zero.

D. Concentration of measure

It was shown in [7] (with a more streamlined and improved derivation in [36]), that the intrinsic volumes concentrate sharply around the statistical dimension. For a closed convex cone C, let X_C denote the discrete random variable satisfying

$$\mathbb{P}\{X_C = k\} = v_k(C).$$

The following result is from [36].

Theorem IV.6. *Let $\lambda \geq 0$. Then*

$$\mathbb{P}\{|X_C - \delta(C)| \geq \lambda\} \leq 2 \exp\left(\frac{-\lambda^2/4}{\min\{\delta(C), \delta(C^\circ)\} + \lambda/3}\right).$$

Roughly speaking, the intrinsic volumes of a convex cone in high dimensions approximate those of a linear subspace of dimension $\delta(C)$. The concentration result IV.6, used in conjunction with the kinematic formula, gives rise to an approximate kinematic formula, which in turn underlies the phase transition results from [7]. We will only need the following direct consequence of Theorem IV.6.

Corollary IV.7. *Let $\eta \in (0, 1)$, let C be a closed convex cone, and let $0 \leq m \leq n$. Then*

$$\delta(C) \leq m - a_\eta \sqrt{m} \implies t_m \leq \eta;$$
$$\delta(C) \geq m + a_\eta \sqrt{m} \implies t_m \geq 1 - \eta,$$

with $a_\eta := 2\sqrt{\log(2/\eta)}$.

Applying the above to the statistical dimension, we get the following expression.

Corollary IV.8. *Let $\eta \in (0, 1)$ and assume that $m \geq \delta(C) + a_\eta \sqrt{m}$, with $a_\eta = 2\sqrt{\log(2/\eta)}$. Then*

$$\delta(C) - (n - m)\eta \leq \mathbb{E}_{\boldsymbol{Q}}[\delta(\boldsymbol{P}\boldsymbol{Q}C)] \leq \delta(C).$$

Proof. A direct application of the projection formulas (IV.8) and the definition of the statistical dimension shows that

$$\mathbb{E}_{\boldsymbol{Q}}[\delta(\boldsymbol{P}\boldsymbol{Q}C)] = \delta(C) - \sum_{k=1}^{n-m} k v_{k+m}(C).$$

The bound then follows by bounding the right-hand side in a straight-forward way and applying Corollary IV.7. □

E. The TQC Lemma

The following generalization of the projection formulas (IV.8), first observed by Mike McCoy and Joel Tropp, may at first sight look surprising. While it can be deduced from general integral-geometric considerations (see, for example, [37]), we include a proof because it is illustrative.

Lemma IV.9. *Let $\boldsymbol{T} \in \mathbb{R}^{m \times n}$ be of full rank. Then for $0 \leq k < m$,*

$$\mathbb{E}[v_k(\boldsymbol{T}\boldsymbol{Q}C)] = v_k(C), \qquad \mathbb{E}[v_m(\boldsymbol{T}\boldsymbol{Q}C)] = t_m(C) \tag{IV.9}$$

Proof. In view of (IV.3), it suffices to show (IV.9) for the half-tail functionals h_j instead of the intrinsic volumes v_j. Let $L \subset \mathbb{R}^n$ be a linear subspace of dimension $\dim L = k \leq m$. From Proposition II.2 it follows that

$$\boldsymbol{Q}C \cap \boldsymbol{T}^{-1}L \neq \{\boldsymbol{0}\} \iff \boldsymbol{T}\boldsymbol{Q}C \cap L \neq \{\boldsymbol{0}\} \text{ or } \ker \boldsymbol{T} \cap \boldsymbol{Q}C \neq \{\boldsymbol{0}\},$$

where in this case, as before, $\boldsymbol{T}^{-1}L$ denotes the pre-image of L under $\boldsymbol{T}$. Denoting by $\boldsymbol{P}$ the orthogonal projection onto the complement $(\ker \boldsymbol{T})^\perp$, we thus get

$$\boldsymbol{P}\boldsymbol{Q}C \cap (\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^\perp) \neq \{\boldsymbol{0}\} \iff \boldsymbol{T}\boldsymbol{Q}C \cap L \neq \{\boldsymbol{0}\},$$

and taking probabilities,

$$\mathbb{P}\left\{\boldsymbol{PQ}C \cap (\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^{\perp}) \neq \{\boldsymbol{0}\}\right\} = \mathbb{P}\{\boldsymbol{TQ}C \cap L \neq \{\boldsymbol{0}\}\}. \tag{IV.10}$$

To compute the probability on the left, let $\boldsymbol{Q}_0$ is a random orthogonal transformation of the space $(\ker \boldsymbol{T})^{\perp}$. Restricting to $(\ker \boldsymbol{T})^{\perp}$ as ambient space,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{Q}}\left\{\boldsymbol{PQ}C \cap (\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^{\perp}) \neq \{\boldsymbol{0}\}\right\} \\
&\quad = \mathbb{P}_{\boldsymbol{Q}}\left\{\boldsymbol{PQ}C \cap \boldsymbol{Q}_0(\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^{\perp}) \neq \{\boldsymbol{0}\}\right\} \\
&\quad = \mathbb{E}_{\boldsymbol{Q}_0}\mathbb{P}_{\boldsymbol{Q}}\left\{\boldsymbol{PQ}C \cap \boldsymbol{Q}_0(\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^{\perp}) \neq \{\boldsymbol{0}\}\right\} \\
&\quad \overset{(1)}{=} \mathbb{E}_{\boldsymbol{Q}}\mathbb{P}_{\boldsymbol{Q}_0}\left\{\boldsymbol{PQ}C \cap \boldsymbol{Q}_0(\boldsymbol{T}^{-1}L \cap (\ker \boldsymbol{T})^{\perp}) \neq \{\boldsymbol{0}\}\right\} \\
&\quad \overset{(2)}{=} \mathbb{E}_{\boldsymbol{Q}}[h_{m-k+1}(\boldsymbol{PQ}C)]
\end{aligned}$$

where for (1) we summoned Fubini on the representation of the probability as expectation of an indicator variable and for (2) the Crofton formula IV.4 with $(\ker \boldsymbol{T})^{\perp}$ as ambient space. A similar argument on the right-hand side of (IV.10) shows that

$$\mathbb{P}_{\boldsymbol{Q}}\{\boldsymbol{TQ}C \cap L \neq \{\boldsymbol{0}\}\} = \mathbb{E}_{\boldsymbol{Q}}[h_{m-k+1}(\boldsymbol{TQ}C)].$$

In summary, we have for shown that $\mathbb{E}_{\boldsymbol{Q}}[h_{m-k+1}(\boldsymbol{TQ}C)] = \mathbb{E}_{\boldsymbol{Q}}[h_{m-k+1}(\boldsymbol{PQ}C)]$ for $0 \le k \le m$, and hence also $\mathbb{E}_{\boldsymbol{Q}}[v_i(\boldsymbol{TQ}C)] = \mathbb{E}_{\boldsymbol{Q}}[v_i(\boldsymbol{PQ}C)]$ for $0 \le i \le m$. The claim now follows by applying the projection formula (IV.8). □

As with the case where $\boldsymbol{T}$ is a projection, applying the above to the statistical dimension, we get the following expression.

Corollary IV.10. *Let $\eta \in (0,1)$ and assume that $m \ge \delta(C) + a_\eta\sqrt{m}$, with $a_\eta = 2\sqrt{\log(2/\eta)}$. Then under the conditions of Lemma IV.9, we have*

$$\delta(C) - (n-m)\eta \le \mathbb{E}_{\boldsymbol{Q}}[\delta(\boldsymbol{TQ}C)] \le \delta(C) - \eta.$$

It remains to be seen whether the fact that the main preconditionining results can be formulated with an arbitrary matrix $\boldsymbol{T}$, rather than just a projection $\boldsymbol{P}$, can be of use.

V. Improved condition bounds

In this section we derive the improved condition number bounds on the statistical dimension. We first derive Corollary I.4, restated here as a proposition, which is a simple consequence of the behaviour of the statistical dimension under polarity.

Proposition V.1. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone, and $\delta(C)$ the statistical dimension of C. Then for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ of full rank,*

$$\delta(\boldsymbol{A}C) \le \kappa(\boldsymbol{A})^{-2} \cdot \delta(C) + \left(1 - \kappa(\boldsymbol{A})^{-2}\right) \cdot n.$$

Proof. We have

$$\begin{aligned}
\delta(\boldsymbol{A}C) &\overset{(1)}{=} n - \delta(\boldsymbol{A}^{-T}C^{\circ}) \\
&\overset{(2)}{\le} n - \kappa(\boldsymbol{A})^{-2}\delta(C^{\circ}) \\
&\overset{(3)}{=} n - \kappa(\boldsymbol{A})^{-2}(n - \delta(C)) \\
&= \kappa(\boldsymbol{A})^{-2} \cdot \delta(C) + \left(1 - \kappa(\boldsymbol{A})^{-2}\right) \cdot n,
\end{aligned}$$

where for (1) we used (IV.4) and Lemma II.1, for (2) we used Theorem A, and for (3) we used (IV.4) again. □

We conclude this section by proving Theorem B, which we restate for convenience.

Theorem V.2. *Let $C \subseteq \mathbb{R}^n$ be a closed convex cone and $\boldsymbol{A} \in \mathbb{R}^{p\times n}$ have full rank. Let $\eta \in (0,1)$ and assume that $m \ge \delta(C) + 2\sqrt{\log(2/\eta)m}$. Then*

$$\delta(\boldsymbol{A}C) \le \overline{\mathscr{R}}^2_{C,m}(\boldsymbol{A}) \cdot \delta(C) + (n-m)\eta.$$

For the matrix condition number,

$$\delta(\boldsymbol{A}C) \le \overline{\kappa}^2_m(\boldsymbol{A}) \cdot \delta(C) + (n-m)\eta. \tag{V.1}$$

Proof. The upper bound follows from

$$\begin{aligned}
\delta(\boldsymbol{A}C) &\le \mathbb{E}_{\boldsymbol{Q}}[\delta(\boldsymbol{P}_m\boldsymbol{QA}C)] + (n-m)\eta \\
&\le \mathbb{E}_{\boldsymbol{Q}}\left[\mathscr{R}_C(\boldsymbol{P}_m\boldsymbol{QA})^2\right]\delta(C) + (n-m)\eta,
\end{aligned}$$

where we used Theorem A for the second inequality. The upper bound in terms for the matrix condition number follows as in the proof of Theorem A. □

VI. Applications

In this section we apply the results derived for convex cones to the setting of convex regularizers. To give this application some context, we briefly review some of the theory.

A. Convex regularization, subdifferentials and the descent cone

In practical applications the cones of interest often arise as cones generated by the subgradient of a proper convex function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. The exact form of the general convex regularization problem is

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{\Omega x} = \boldsymbol{b}, \tag{VI.1}$$

while the noisy form is

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{\Omega x} - \boldsymbol{b}\|_2 \le \varepsilon. \tag{VI.2}$$

Interchanging the role of the function f and the residual, we get the *generalized LASSO*

$$\text{minimize} \quad \|\boldsymbol{\Omega x} - \boldsymbol{b}\|_2 \quad \text{subject to} \quad f(\boldsymbol{x}) \le \tau. \tag{VI.3}$$

Finally, we have the Lagrangian form,

$$\text{minimize} \quad \|\boldsymbol{\Omega x} - \boldsymbol{b}\|_2^2 + \lambda f(\boldsymbol{x}). \tag{VI.4}$$

These last three problems are, in fact, equivalent (see [1, Chapter 3] for a concise derivation in the case $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$). The practical problem consists in effectively finding the parameters involved.

The first-order optimality condition states that $\hat{\boldsymbol{x}}$ is a unique solution of (VI.1) if and only if

$$\exists \boldsymbol{y} \neq \boldsymbol{0}\colon \boldsymbol{\Omega}^T\boldsymbol{y} \in \partial f(\hat{\boldsymbol{x}}), \tag{VI.5}$$

where $\partial f(\hat{\boldsymbol{x}})$ denotes the subdifferential of f at $\hat{\boldsymbol{x}}$, i.e., the set

$$\partial f(\hat{\boldsymbol{x}}) = \{\boldsymbol{z} \in \mathbb{R}^n : f(\hat{\boldsymbol{x}} + \boldsymbol{z}) \ge f(\hat{\boldsymbol{x}}) + \langle \boldsymbol{z}, \boldsymbol{x}\rangle\}.$$

If f is differentiable at $\hat{\boldsymbol{x}}$, then of course the subdifferential contains only the gradient of f at $\hat{\boldsymbol{x}}$, and the vector $\boldsymbol{y}$ in (VI.5) consists of the Lagrange multipliers.

Example VI.1. If f is a norm, with dual norm f°, then the subdifferential of f at $\hat{\boldsymbol{x}}$ is

$$\partial f(\hat{\boldsymbol{x}}) = \begin{cases} \{\boldsymbol{z} \in \mathbb{R}^n : f^\circ(\boldsymbol{z}) = 1, \langle \boldsymbol{z}, \hat{\boldsymbol{x}} \rangle = f(\hat{\boldsymbol{x}})\} & \hat{\boldsymbol{x}} \neq \boldsymbol{0} \\ \{\boldsymbol{z} \in \mathbb{R}^n : f^\circ(\boldsymbol{z}) \leq 1\} & \hat{\boldsymbol{x}} = \boldsymbol{0}. \end{cases}$$

Example VI.2. For the ℓ_1-norm at an s-sparse vector $\hat{\boldsymbol{x}}$,

$$\partial \|\hat{\boldsymbol{x}}\|_1 = \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_\infty = 1, \langle \boldsymbol{z}, \hat{\boldsymbol{x}} \rangle = \|\hat{\boldsymbol{x}}\|_1\},$$

or more explicitly,

$$\partial \|\hat{\boldsymbol{x}}\|_1 = \{\boldsymbol{z} \in \mathbb{R}^n : z_i = \text{sign}\,(\hat{x}_i) \text{ if } \hat{x}_i \neq 0,\ z_j \in [-1,1] \text{ if } \hat{x}_j = 0\}. \tag{VI.6}$$

The descent cone of f at $\hat{\boldsymbol{x}}$ is defined as

$$\mathscr{D}(f, \hat{\boldsymbol{x}}) = \bigcup_{\tau > 0} \left\{ \boldsymbol{y} \in \mathbb{R}^n : f(\hat{\boldsymbol{x}} + \tau \boldsymbol{y}) \leq f(\hat{\boldsymbol{x}}) \right\}.$$

The convex cone generated by the subdifferential of f at $\hat{\boldsymbol{x}}$ is the closure of the polar cone of $\mathscr{D}(f, \hat{\boldsymbol{x}})$,

$$\text{cone}\left(\partial f(\hat{\boldsymbol{x}})\right) = \overline{\mathscr{D}(f, \hat{\boldsymbol{x}})^\circ}, \tag{VI.7}$$

Condition (VI.5) is therefore equivalent to

$$\ker \boldsymbol{\Omega} \cap \mathscr{D}(f, \hat{\boldsymbol{x}}) = \{\boldsymbol{0}\},$$

namely, that the kernel of $\boldsymbol{\Omega}$ does not intersect the descent cone nontrivially.

An important class of regularizers are of the form $f(\boldsymbol{x}) := g(\boldsymbol{A}\boldsymbol{x}) + h(\boldsymbol{B}\boldsymbol{x})$, with $\boldsymbol{A}$ and $\boldsymbol{B}$ linear maps. It follows from [38, Theorems 23.8, 23.9] that the subdifferential is

$$\partial f(\boldsymbol{x}) = \boldsymbol{A}^T \partial g(\boldsymbol{A}\boldsymbol{x}) + \boldsymbol{B}^T \partial h(\boldsymbol{B}\boldsymbol{x}).$$

Example VI.3. In the ℓ_1-analysis, or cosparse, model, one considers regularizers of the form $\|\boldsymbol{D}\boldsymbol{x}\|_1$, with $\boldsymbol{D} \in \mathbb{R}^{p \times n}$ with typically $p \geq n$. The interest is on vectors for which $\boldsymbol{D}\boldsymbol{x}_0$ is s-sparse. If $\boldsymbol{D}$ has full rank and $\boldsymbol{x}_0 \neq \boldsymbol{0}$, then $s \geq p - n + 1$, as otherwise $\boldsymbol{D}$ would have a $n \times n$ minor that maps $\boldsymbol{x}_0$ to $\boldsymbol{0}$. The focus in this model has traditionally been on the *cosupport*, i.e., the location of the entries of $\boldsymbol{D}\boldsymbol{x}_0$ that vanish. A typical example would be a shift invariant wavelet transform. The subdifferential of $\|\boldsymbol{D}\cdot\|_1$ is given by $\boldsymbol{D}^T \partial \|\boldsymbol{D}\boldsymbol{x}_0\|_1$. For invertible $\boldsymbol{D}$, combining (VI.7) with Lemma II.1 we get,

$$\mathscr{D}(\|\boldsymbol{D}\cdot\|_1, \boldsymbol{x}_0) = \boldsymbol{D}^{-1} \mathscr{D}(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0). \tag{VI.8}$$

When working with the subdifferential cone rather than the descent cone, we don't need the invertibility requirement.

Example VI.4 (Finite differences)**.** Let $\boldsymbol{x} \in \mathbb{R}^n$ and let

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix} \tag{VI.9}$$

be the discrete finite difference matrix. Thus

$$\boldsymbol{D}\boldsymbol{x} = (x_2 - x_1, x_3 - x_2, \ldots, x_d - x_{d-1}, -x_d)^T.$$

Define $g(\boldsymbol{x}) := f(\boldsymbol{D}\boldsymbol{x})$. Then for a fixed $\hat{\boldsymbol{x}}$, the subdifferential is given by

$$\partial g(\hat{\boldsymbol{x}}) = \boldsymbol{D}^T \partial f(\boldsymbol{D}\hat{\boldsymbol{x}}).$$

In the special case where f is the ℓ_1-norm and $\boldsymbol{D}\hat{\boldsymbol{x}}$ is s-sparse with support $I \subset [n]$,

$$\partial g(\hat{\boldsymbol{x}}) = \{\boldsymbol{D}^T \boldsymbol{z} : \|\boldsymbol{z}\|_\infty = 1, \langle \boldsymbol{z}, \boldsymbol{D}\hat{\boldsymbol{x}} \rangle = \|\boldsymbol{D}\hat{\boldsymbol{x}}\|_1\}.$$

One can think of such a vector $\hat{\boldsymbol{x}}$ as a signal with sparse gradient.

Example VI.5. (Weighted ℓ_1 norm). Let $\boldsymbol{\omega} \in \mathbb{R}^n$ be a vector of weights and define the weighted ℓ_1-norm

$$\|\boldsymbol{x}\|_{\boldsymbol{\omega},1} = \sum_{j=1}^n \omega_j |x_j|.$$

By extension from the ℓ_1 example, we have

$$\begin{aligned} \partial \|\hat{\boldsymbol{x}}\|_{\boldsymbol{\omega},1} &= \{\boldsymbol{z} \in \mathbb{R}^n : z_i = \omega_i \text{ sign } (\hat{x}_i) \text{ if } \hat{x}_i \neq 0,\ z_j \in [-\omega_j, \omega_j] \text{ if } \hat{x}_j = 0\} \\ &= \text{diag}(\boldsymbol{\omega})\ \partial \|\hat{\boldsymbol{x}}\|_1. \end{aligned}$$

This example becomes interesting when considering *weighted* s-sparse vectors, that is, vectors such that

$$\|\boldsymbol{x}\|_{\boldsymbol{\omega},0} = \sum_{x_j \neq 0} \omega_j^2 = s.$$

The use of composite regularizers to recover simultaneously structured models was studied in [39].

B. Performance bounds in convex regularization

As mentioned in the introduction, computing the statistical dimension of convex regularizers is in general a difficult problem, with only few cases allowing for closed-form expressions. Using the condition bounds for the statistical dimension of linear images of convex cones, and translating these to the setting of convex regularizers, we get the corresponding statements in Corollary I.2, which we restate here.

Corollary VI.6. *Let $f(\boldsymbol{x}) = g(\boldsymbol{D}\boldsymbol{x})$, where g is a proper convex function and let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be non-singular. Then*

$$\delta(f, \boldsymbol{x}_0) \leq \mathscr{R}_{\mathscr{D}(g, \boldsymbol{D}\boldsymbol{x}_0)}\left(\boldsymbol{D}^{-1}\right) \cdot \delta(g, \boldsymbol{D}\boldsymbol{x}_0).$$

In particular,

$$\frac{\delta(g, \boldsymbol{D}\boldsymbol{x}_0)}{\kappa(\boldsymbol{D})^2} \leq \delta(f, \boldsymbol{x}_0) \leq \kappa(\boldsymbol{D})^2 \cdot \delta(g, \boldsymbol{D}\boldsymbol{x}_0).$$

Proof. Let $C = \mathscr{D}(g, \boldsymbol{D}\boldsymbol{x}_0)$. Then from (VI.8) we get that

$$\delta(f, \boldsymbol{x}_0) = \delta(\boldsymbol{D}^{-1} C).$$

The claims then follows from Theorem A and Proposition I.4, noting that $\kappa(\boldsymbol{D}^{-1}) = \kappa(\boldsymbol{D})$. □

For convenience, we also recall the statement of Proposition I.6.

Proposition VI.7. *Let $\boldsymbol{D} \in \mathbb{R}^{p\times n}$, $p \geq n$, be such that all $n\times n$ minors of $\boldsymbol{D}$ have full rank, and $\boldsymbol{\Omega} \in \mathbb{R}^{m\times n}$ with $m \leq n$. Consider the problem*

$$\text{minimize} \quad \|\boldsymbol{D}\boldsymbol{x}\|_1 \qquad \text{subject to} \quad \boldsymbol{\Omega}\boldsymbol{x} = \boldsymbol{b}. \tag{VI.10}$$

Let $\boldsymbol{x}_0 \neq \boldsymbol{0}$ be such that $\boldsymbol{\Omega}\boldsymbol{x}_0 = \boldsymbol{b}$, and such that $\boldsymbol{y}_0 = \boldsymbol{D}\boldsymbol{x}_0$ is s-sparse with support $I \subset [p]$. Let $\boldsymbol{C} \in \mathbb{R}^{n\times p-s+1}$ be a matrix whose first $p-s$ columns consist of the columns of $\boldsymbol{D}^T$ that are indexed by I^c, and the last column is $\boldsymbol{c}_{p-s+1} = \frac{1}{\sqrt{s}}\sum_{j\in I}\operatorname{sign}((\boldsymbol{y}_0)_j)\boldsymbol{d}_j$, where the vectors $\boldsymbol{d}_j$ denote the columns of $\boldsymbol{D}^T$. Then

$$\delta(\|\boldsymbol{D}\cdot\|_1, \boldsymbol{x}_0) \leq \kappa(\boldsymbol{C})^{-2}\cdot\delta(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0) + \left(1-(p/n)\kappa(\boldsymbol{C})^{-2}\right)\cdot n$$

In particular, given $\eta \in (0,1)$, Problem (VI.10) with Gaussian measurement matrix succeeds with probability $1-\eta$ if

$$m \geq \kappa(\boldsymbol{C})^{-2}\cdot\delta(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0) + \left(1-(p/n)\kappa(\boldsymbol{C})^{-2}\right)\cdot n + a_\eta\sqrt{n},$$

Proof. Set $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$ with $\boldsymbol{D} \in \mathbb{R}^{p\times n}$ and $p \geq n$. Let $\boldsymbol{x}_0$ be given such that $\boldsymbol{y}_0 = \boldsymbol{D}\boldsymbol{x}_0$ is s-sparse with support I. Assuming that all the $n\times n$ minors of $\boldsymbol{D}$ has rank n and $\boldsymbol{x}_0 \neq \boldsymbol{0}$, $\boldsymbol{y}_0$ has at most $n-1$ zero entries, and the support therefore satisfies $s \geq p-n+1$. As shown in Section VI, the descent cone $\mathscr{D}(f, \boldsymbol{x}_0)$ is polar to the subdifferential cone $\operatorname{cone}(\partial f(\boldsymbol{x}_0))$. Moreover, the statistical dimension satisfies $\delta(C) + \delta(C^\circ) = n$, so that

$$\delta(\mathscr{D}(f, \boldsymbol{x}_0)) = n - \delta(\operatorname{cone}(\partial f(\boldsymbol{x}_0))) = n - \delta(\boldsymbol{D}^T\operatorname{cone}(\partial\|\boldsymbol{y}_0\|_1)).$$

The subdifferential of the 1-norm is given by (see (VI.6))

$$\partial\|\boldsymbol{y}_0\|_1 = \{\boldsymbol{z} \in \mathbb{R}^p : z_i = \operatorname{sign}((\boldsymbol{y}_0)_i) \text{ if } i \in I,\ z_j \in [-1,1] \text{ if } j \notin I\},$$

and we denote by $C := \operatorname{cone}(\partial\|\boldsymbol{y}_0\|_1)$ the cone generated by this subdifferential.

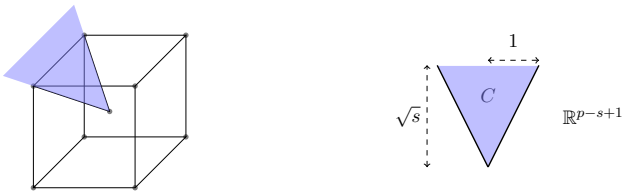

Fig. 4: Cone spanned by $(p-s)$-face of d-dimensional hypercube

It follows that the cone generated by this subdifferential is contained in a subspace L of dimension $\dim L = p-s+1 \leq n$. An orthonormal basis of this subspace is given by the columns of a matrix $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{p-s+1}]$, where for $1 \leq i \leq p-s$, the $\boldsymbol{b}_i$ are the unit vectors $\boldsymbol{e}_j$ for $j \in I^c$ and $\boldsymbol{b}_{p-s+1} = \frac{1}{\sqrt{s}}\sum_{j\in I}\operatorname{sign}((\boldsymbol{y}_0)_j)\boldsymbol{e}_j$. A moment's thought shows that $C = \boldsymbol{B}\tilde{C}$, where $\tilde{C} \subset \mathbb{R}^{p-s+1}$ is the cone in $\mathbb{R}^{p-s+1}$ spanned by vectors of the form $\pm\boldsymbol{e}_i + \sqrt{s}\boldsymbol{e}_{p-s+1}$ for $1 \leq i \leq n-p$ (see Figure 4). By the orthogonal invariance and the embedding invariance of the statistical dimension (see Properties (a) and (c) in Section IV-B), we get $\delta(C) = \delta(\tilde{C})$. With this setup, we have

$$\boldsymbol{D}^T C = \boldsymbol{D}^T\boldsymbol{B}\tilde{C} = \boldsymbol{C}\tilde{C},$$

with the matrix $\boldsymbol{C} := \boldsymbol{D}^T\boldsymbol{B} \in \mathbb{R}^{n\times(p-s+1)}$ is then given as in the statement of the theorem. Applying the bounds from Theorem A we thus get

$$\begin{aligned}\delta(\mathscr{D}(f, \boldsymbol{x}_0)) &= n - \delta(\boldsymbol{D}^T C)\\ &= n - \delta(\boldsymbol{C}\tilde{C})\\ &\leq n - \kappa^{-2}(\boldsymbol{C})\delta(\tilde{C})\\ &= n - \kappa^{-2}(\boldsymbol{C})\delta(C)\\ &= \kappa(\boldsymbol{C})^{-2}\cdot\delta(\|\cdot\|_1, \boldsymbol{D}\boldsymbol{x}_0) + \left(1-(p/n)\kappa(\boldsymbol{C})^{-2}\right)\cdot n,\end{aligned}$$

as was to be shown. □

C. A note on the Stojnic method

A popular method [7, Recipe 4.1], going back to Stojnic [5] and generalized in [6], is to approximate the statistical dimension of the descent cone $\mathscr{D}(f, \boldsymbol{x}_0)$ by the expected value

$$\inf_{\tau\geq 0}\mathbb{E}[\operatorname{dist}^2(\boldsymbol{g}, \tau\cdot\partial f(\boldsymbol{x}))]. \tag{VI.11}$$

This approximation, however, does not work for all regularizers f for two reasons: it my not be tight, and computing the quantity may not be feasible. In [7, Theorem 4.1], the following error bound is derived:

$$0 \leq \inf_{\tau\geq 0}\mathbb{E}[\operatorname{dist}^2(\boldsymbol{g}, \tau\cdot\partial f(\boldsymbol{x}))] - \delta(f, \boldsymbol{x}_0) \leq \frac{2\sup\{\|\boldsymbol{s}\| :\ \boldsymbol{s} \in \partial f(\boldsymbol{x})\}}{f(\boldsymbol{x}/\|\boldsymbol{x}\|)}. \tag{VI.12}$$

In [19], the error (VI.12) was analyzed in the case of TV minimization and it was shown to be bounded, so that the approximation is asymptotically tight. If $f(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$ and assuming that $\boldsymbol{y}_0 = \boldsymbol{D}\boldsymbol{x}_0$ is s-sparse, we can express this bound in terms of the condition number of $\boldsymbol{D}$. First note that the subdifferential of the 1-norm is contained in the unit cube:

$$\partial\|\boldsymbol{y}_0\|_1 \subset \{\boldsymbol{z} :\ \|\boldsymbol{z}\|_\infty \leq 1\}.$$

Using the expression for the subdifferential of g at $\boldsymbol{x}_0$, namely $\partial g(\boldsymbol{x}_0) = \boldsymbol{D}^T\partial\|\boldsymbol{y}_0\|_1$, the error bound (VI.12) translates to

$$\frac{2\sup\{\|\boldsymbol{x}\|_2 : \boldsymbol{x} \in \boldsymbol{D}^T\partial\|\boldsymbol{y}_0\|_1\}}{\|\boldsymbol{y}_0\|_1/\|\boldsymbol{x}_0\|_2} \leq \frac{2}{\|\boldsymbol{y}_0\|_1/\|\boldsymbol{x}_0\|_2}\sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{D}^T\boldsymbol{x}\|.$$

Using the norm inequality $\|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty$, we get the bound

$$\sup_{\|\boldsymbol{x}\|_\infty\leq 1}\|\boldsymbol{D}^T\boldsymbol{x}\|_2 \leq \sqrt{n}\sup_{\|\boldsymbol{x}\|_2\leq 1}\|\boldsymbol{D}^T x\|_2 = \sqrt{n}\|\boldsymbol{D}\|_2.$$

On the other hand, by the norm inequality $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1$ we have that

$$\frac{\|\boldsymbol{y}_0\|_1}{\|\boldsymbol{x}_0\|_2} = \frac{\|\boldsymbol{D}\boldsymbol{x}_0\|_1}{\|\boldsymbol{x}_0\|_2} \geq \frac{\|\boldsymbol{D}\boldsymbol{x}_0\|_2}{\|\boldsymbol{x}_0\|_2} \geq \sigma(\boldsymbol{D}).$$

We therefore get the condition bound

$$0 \leq \inf_{\tau\geq 0}\mathbb{E}[\operatorname{dist}^2(\boldsymbol{g}, \tau\cdot\partial f(\boldsymbol{x}))] - \delta(f, \boldsymbol{x}_0) \leq \sqrt{n}\kappa(\boldsymbol{D}).$$

From this we see that we can guarantee good bounds on the relative statistical dimension $\delta(f, \boldsymbol{x}_0)/n$ if the condition

number of $\boldsymbol{D}$ is small. The bound can actually be improved when considering that we only need to maximize and minimize over certain subspaces in the definition of the singular values.

While this bound is not sharp (the derivation makes use of norm inequalities), it is enlightening as it gives sufficient conditions for the applicability of Bound (VI.12) in terms of the condition number of $\boldsymbol{A}$. It remains to be seen whether randomized preconditioning can be incorporated into this bound, and therefore whether this approach can lead to bounds that would rival those derived in [19].

Appendix A
The biconic feasibility problem - proofs

In this appendix we provide the proofs for Section II-B. Recall that for $C \subseteq \mathbb{R}^n$, $D \subseteq \mathbb{R}^m$ closed convex cones, the biconic feasibility problem is given by

$$\exists \boldsymbol{x} \in C \setminus \{\boldsymbol{0}\} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \in D^\circ, \tag{P}$$

$$\exists \boldsymbol{y} \in D \setminus \{\boldsymbol{0}\} \quad \text{s.t.} \quad -\boldsymbol{A}^T\boldsymbol{y} \in C^\circ, \tag{D}$$

and the sets of primal feasible and dual feasible instances can be characterized by

$$\begin{aligned}\mathscr{P}(C,D) &= \left\{\boldsymbol{A} \in \mathbb{R}^{m\times n} : C \cap \left(\boldsymbol{A}^T D\right)^\circ \neq \{\boldsymbol{0}\}\right\} \\ &= \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \sigma_{C\to D}(\boldsymbol{A}) = 0\}, \\ \mathscr{D}(C,D) &= \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : D \cap (-\boldsymbol{A}C)^\circ \neq \{\boldsymbol{0}\}\} \\ &= \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \sigma_{D\to C}(-\boldsymbol{A}^T) = 0\},\end{aligned}$$

respectively, cf. (II.3)/(II.4). The proof of Proposition II.4 uses the following generalization of Farkas' Lemma.

Lemma A.1. *Let $C, \tilde{C} \subseteq \mathbb{R}^n$ be closed convex cones with* $\operatorname{int}(C) \neq \emptyset$. *Then*

$$\operatorname{int}(C) \cap \tilde{C} = \emptyset \iff C^\circ \cap (-\tilde{C}^\circ) \neq \{\boldsymbol{0}\}. \tag{A.1}$$

Proof. If $\operatorname{int}(C) \cap \tilde{C} = \emptyset$, then there exists a separating hyperplane $H = \boldsymbol{v}^\perp$, $\boldsymbol{v} \neq \boldsymbol{0}$, so that $\langle \boldsymbol{v}, \boldsymbol{x}\rangle \leq 0$ for all $\boldsymbol{x} \in C$ and $\langle \boldsymbol{v}, \boldsymbol{y}\rangle \geq 0$ for all $\boldsymbol{y} \in \tilde{C}$. But this means $\boldsymbol{v} \in C^\circ \cap (-\tilde{C}^\circ)$. On the other hand, if $\boldsymbol{x} \in \operatorname{int}(C) \cap \tilde{C}$ then only in the case $C = \mathbb{R}^n$, for which the claim is trivial, can $\boldsymbol{x} = \boldsymbol{0}$. If $\boldsymbol{x} \neq \boldsymbol{0}$, then $C^\circ \setminus \{\boldsymbol{0}\}$ lies in the open half-space $\{\boldsymbol{v} : \langle \boldsymbol{v}, \boldsymbol{x}\rangle < 0\}$ and $-\tilde{C}^\circ$ lies in the closed half-space $\{\boldsymbol{v} : \langle \boldsymbol{v}, \boldsymbol{x}\rangle \geq 0\}$, and thus $C^\circ \cap (-\tilde{C}^\circ) = \{\boldsymbol{0}\}$. □

For the proof of the third claim in Proposition II.4 we also need the following well-known convex geometric lemma; a proof can be found, for example, in [33, proof of Thm. 6.5.6]. We say that two cones $C, D \subseteq \mathbb{R}^n$, with $\operatorname{int}(C) \neq \emptyset$, *touch* if $C \cap D \neq \{\boldsymbol{0}\}$ but $\operatorname{int}(C) \cap D = \emptyset$.

Lemma A.2. *Let $C, D \subseteq \mathbb{R}^n$ closed convex cones with* $\operatorname{int}(C) \neq \emptyset$. *If $\boldsymbol{Q} \in O(n)$ uniformly at random, then the randomly rotated cone $\boldsymbol{Q}D$ almost surely does not touch C.*

Proof of Proposition II.4. (1) The sets $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$ are closed as they are preimages of the closed set $\{0\}$ under continuous functions, c.f. (II.3)/(II.4). Indeed, for any $\boldsymbol{x}$, the function $\boldsymbol{A} \mapsto \|\boldsymbol{\Pi}_D(\boldsymbol{A}\boldsymbol{x})\|$ is continuous, and as a minimum of such functions over the compact set $C \cap S^{m-1}$, it follows that $\sigma_{C\to D}(\boldsymbol{A})$ is continuous. Hence, $\mathscr{P}(C,D) = \{\boldsymbol{A} \in \mathbb{R}^{n\times m} : \sigma_{C\to D}(\boldsymbol{A}) = 0\}$ is closed. The same argument applies to $\mathscr{D}(C,D)$.

(2) For the claim about the union of the sets $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$ we first consider the case $C \neq \mathbb{R}^n$, so that $\boldsymbol{0} \notin \operatorname{int}(C)$. Using the generalized Farkas' Lemma A.1, we obtain

$$\begin{aligned}\boldsymbol{A} \notin \mathscr{P}(C,D) &\iff C \cap \left(\boldsymbol{A}^T D\right)^\circ = \{\boldsymbol{0}\} \\ &\Rightarrow \operatorname{int}(C) \cap \left(\boldsymbol{A}^T D\right)^\circ = \emptyset \\ &\overset{(A.1)}{\Longrightarrow} C^\circ \cap (-\boldsymbol{A}^T D) \neq \{\boldsymbol{0}\} \Rightarrow \boldsymbol{A} \in \mathscr{D}(C,D).\end{aligned}$$

This shows $\mathscr{P}(C,D) \cup \mathscr{D}(C,D) = \mathbb{R}^{n\times m}$. For $D \neq \mathbb{R}^n$ the argument is the same. For $C = \mathbb{R}^n$ and $D = \mathbb{R}^m$:

$$\begin{aligned}\mathscr{P}(\mathbb{R}^n,\mathbb{R}^m) &= \left\{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \ker \boldsymbol{A} \neq \{\boldsymbol{0}\}\right\} \\ &= \begin{cases}\{\text{rank deficient matrices}\} & \text{if } n \leq m \\ \mathbb{R}^{m\times n} & \text{if } n > m,\end{cases} \\ \mathscr{D}(\mathbb{R}^n,\mathbb{R}^m) &= \left\{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \ker \boldsymbol{A}^T \neq \{\boldsymbol{0}\}\right\} \\ &= \begin{cases}\mathbb{R}^{m\times n} & \text{if } n < m \\ \{\text{rank deficient matrices}\} & \text{if } n \geq m.\end{cases}\end{aligned}$$

In particular,this shows $\mathscr{P}(\mathbb{R}^n,\mathbb{R}^n) \cup \mathscr{D}(\mathbb{R}^n,\mathbb{R}^n) = \{\text{rank deficient matrices}\}$.

(3) If $(C,D) = (\mathbb{R}^n,\mathbb{R}^m)$ then by the characterization above $\Sigma(\mathbb{R}^n,\mathbb{R}^m)$ consists of the rank deficient matrices, which is a nonempty set. If $(C,D) \neq (\mathbb{R}^n,\mathbb{R}^n)$, then the union of the closed sets $\mathscr{P}(C,D)$ and $\mathscr{D}(C,D)$ equals $\mathbb{R}^{m\times n}$, which is an irreducible topological space, so that their intersection $\Sigma(C,D) = \mathscr{P}(C,D) \cap \mathscr{D}(C,D)$ must be nonempty.

As for the claim about the Lebesgue measure of $\Sigma(C,D)$, we may use the symmetry between (P) and (D) to assume without loss of generality $m \leq n$. If $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has full rank, then $\boldsymbol{A}C$ has nonempty interior and from Proposition II.2 and Farkas' Lemma,

$$\begin{aligned}\sigma_{C\to D}(\boldsymbol{A}) = 0 &\iff C \cap (\boldsymbol{A}^T D)^\circ \neq \{\boldsymbol{0}\} \\ &\iff \boldsymbol{A}C \cap D^\circ \neq \{\boldsymbol{0}\} \text{ or } \ker \boldsymbol{A} \cap C \neq \{\boldsymbol{0}\}, \\ \sigma_{D\to C}(-\boldsymbol{A}^T) = 0 &\iff D \cap (-\boldsymbol{A}C)^\circ \neq \{\boldsymbol{0}\} \\ &\overset{(A.1)}{\iff} D^\circ \cap \operatorname{int}(\boldsymbol{A}C) = \emptyset.\end{aligned}$$

Note that if $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ for some $\boldsymbol{x} \in \operatorname{int}(C)$, then $\boldsymbol{A}$, being a continuous surjection, maps an open neighborhood of $\boldsymbol{x}$ to an open neighborhood of the origin, so that $\boldsymbol{A}C = \mathbb{R}^m$. Hence, $D \cap (-\boldsymbol{A}C)^\circ \neq \{\boldsymbol{0}\}$ implies $\ker \boldsymbol{A} \cap \operatorname{int}(C) = \emptyset$, since otherwise $\boldsymbol{A}C = \mathbb{R}^m$, i.e., $(\boldsymbol{A}C)^\circ = \{\boldsymbol{0}\}$.

If $\boldsymbol{A} \in \Sigma(C,D)$, i.e., $\sigma_{C\to D}(\boldsymbol{A}) = \sigma_{D\to C}(-\boldsymbol{A}^T) = 0$, and if $\boldsymbol{A}$ has full rank, then $\boldsymbol{A}C \cap D^\circ \neq \{\boldsymbol{0}\}$ implies that D° touches $\boldsymbol{A}C$, while $\ker \boldsymbol{A} \cap C \neq \{\boldsymbol{0}\}$ implies that $\ker \boldsymbol{A}$ touches C. Hence, if $\boldsymbol{A} = \boldsymbol{G}$ Gaussian, then $\boldsymbol{G}$ has almost surely full rank, and Lemma A.2 implies that both touching events have zero probability, so that almost surely $\boldsymbol{G} \notin \Sigma(C,D)$. □

We next provide the proof for the characterization of the restricted singular values as distances to the primal and dual feasible sets. From now on we use again the short-hand notation $\mathscr{P} := \mathscr{P}(C,D)$ and $\mathscr{D} := \mathscr{D}(C,D)$.

Proof of Proposition II.5. By symmetry, it suffices to show that $\mathrm{dist}(\boldsymbol{A},\mathscr{P}) = \sigma_{C\to D}(\boldsymbol{A})$. If $\boldsymbol{A}\in\mathscr{P}$ then $\mathrm{dist}(\boldsymbol{A},\mathscr{P}) = 0 = \sigma_{C\to D}(\boldsymbol{A})$, so assume that $\boldsymbol{A}\notin\mathscr{P}$. Let $\boldsymbol{\Delta A}\in\mathbb{R}^{m\times n}$ such that $\boldsymbol{A}+\boldsymbol{\Delta A}\in\mathscr{P}$ and $\mathrm{dist}(\boldsymbol{A},\mathscr{P}) = \|\boldsymbol{\Delta A}\|$. Since $\boldsymbol{A}+\boldsymbol{\Delta A}\in\mathscr{P}$, there exists $\boldsymbol{x}_0\in C\cap S^{n-1}$ such that $\boldsymbol{w}_0 := (\boldsymbol{A}+\boldsymbol{\Delta A})\boldsymbol{x}_0\in D^\circ$. For all $\boldsymbol{y}\in D$

$$0\geq\langle\boldsymbol{w}_0,\boldsymbol{y}\rangle = \langle(\boldsymbol{A}+\boldsymbol{\Delta A})\boldsymbol{x}_0,\boldsymbol{y}\rangle = \langle\boldsymbol{A}\boldsymbol{x}_0,\boldsymbol{y}\rangle - \langle-\boldsymbol{\Delta A}\boldsymbol{x}_0,\boldsymbol{y}\rangle.$$

If $\boldsymbol{y}_0\in B^m\cap D$ is such that $\|\Pi_D(\boldsymbol{A}\boldsymbol{x}_0)\| = \langle\boldsymbol{A}\boldsymbol{x}_0,\boldsymbol{y}_0\rangle$, then

$$\begin{aligned}\mathrm{dist}(\boldsymbol{A},\mathscr{P}) = \|\boldsymbol{\Delta A}\| &\geq \|\boldsymbol{\Delta A}\boldsymbol{x}_0\| \geq \|\Pi_D(-\boldsymbol{\Delta A}\boldsymbol{x}_0)\|\\ &= \max_{\boldsymbol{y}\in B^m\cap D}\langle-\boldsymbol{\Delta A}\boldsymbol{x}_0,\boldsymbol{y}\rangle\\ &\geq \langle-\boldsymbol{\Delta A}\boldsymbol{x}_0,\boldsymbol{y}_0\rangle \geq \langle\boldsymbol{A}\boldsymbol{x}_0,\boldsymbol{y}_0\rangle = \|\Pi_D(\boldsymbol{A}\boldsymbol{x}_0)\|\\ &\geq \min_{\boldsymbol{x}\in C\cap S^{n-1}}\|\Pi_D(\boldsymbol{A}\boldsymbol{x})\| = \sigma_{C\to D}(\boldsymbol{A}).\end{aligned}$$

For the reverse inequality $\mathrm{dist}(\boldsymbol{A},\mathscr{P})\leq\sigma_{C\to D}(\boldsymbol{A})$ we need to construct a perturbation $\boldsymbol{\Delta A}$ such that $\boldsymbol{A}+\boldsymbol{\Delta A}\in\mathscr{P}$ and $\|\boldsymbol{\Delta A}\|\leq\sigma_{C\to D}(\boldsymbol{A})$. Let $\boldsymbol{x}_0\in C\cap S^{n-1}$ and $\boldsymbol{y}_0\in D\cap B^m$ such that

$$\sigma_{C\to D}(\boldsymbol{A}) = \min_{\boldsymbol{x}\in C\cap S^{n-1}}\max_{\boldsymbol{y}\in D\cap B^m}\langle\boldsymbol{A}\boldsymbol{x},\boldsymbol{y}\rangle = \langle\boldsymbol{A}\boldsymbol{x}_0,\boldsymbol{y}_0\rangle.$$

Since $\boldsymbol{A}\notin\mathscr{P}$ we have $\sigma_{C\to D}(\boldsymbol{A})>0$, which implies $\|\boldsymbol{y}_0\| = 1$, i.e., $\boldsymbol{y}_0\in D\cap S^{m-1}$. We define

$$\boldsymbol{\Delta A} := -\boldsymbol{y}_0\boldsymbol{y}_0^T\boldsymbol{A}.$$

Note that

$$\|\boldsymbol{\Delta A}\| = \|\boldsymbol{A}^T\boldsymbol{y}_0\|\leq\langle\boldsymbol{A}^T\boldsymbol{y}_0,\boldsymbol{x}_0\rangle = \sigma_{C\to D}(\boldsymbol{A}).$$

Furthermore,

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{\Delta A})\boldsymbol{x}_0 &= \boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{y}_0\boldsymbol{y}_0^T\boldsymbol{A}\boldsymbol{x}_0\\ &= \boldsymbol{A}\boldsymbol{x}_0 - \langle\boldsymbol{A}\boldsymbol{x}_0,\boldsymbol{y}_0\rangle\boldsymbol{y}_0\\ &= \boldsymbol{A}\boldsymbol{x}_0 - \Pi_D(\boldsymbol{A}\boldsymbol{x}_0) = \Pi_{D^\circ}(\boldsymbol{A}\boldsymbol{x}_0).\end{aligned}$$

So $\boldsymbol{x}_0\in C\setminus\{\boldsymbol{0}\}$ and $(\boldsymbol{A}+\boldsymbol{\Delta A})\boldsymbol{x}_0\in D^\circ$, which shows that $\boldsymbol{A}+\boldsymbol{\Delta A}\in\mathscr{P}$, and hence $\mathrm{dist}(\boldsymbol{A},\mathscr{P})\leq\|\boldsymbol{\Delta A}\|\leq\sigma_{C\to D}(\boldsymbol{A})$. □

ACKNOWLEDGMENT

The authors would like to thank Mike McCoy and Joel Tropp for fruitful discussions on integral geometry, and in particular for suggesting the TQC Lemma, and Armin Eftekhari for helpful discussions on random projections. I would also like to thank the anonymous referees for valuable feedback and suggestions.

REFERENCES

[1] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, ser. Applied and Numerical Harmonic Analysis. Basel: Birkhäuser, 2013, vol. 336.

[2] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse problems*, vol. 23, no. 3, p. 947, 2007.

[3] E. Candes, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[4] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30–56, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.acha.2012.03.006

[5] M. Stojnic, "Various thresholds for ℓ_1-optimization in compressed sensing," *preprint*, 2009, arXiv:0907.3666.

[6] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10208-012-9135-7

[7] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: phase transitions in convex programs with random data," *Information and Inference*, vol. 3, no. 3, pp. 224–294, 2014.

[8] D. L. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, 2009. [Online]. Available: http://dx.doi.org/10.1090/S0894-0347-08-00600-0

[9] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized lasso: A precise analysis," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, 2013, pp. 1002–1009.

[10] D. L. Donoho, I. Johnstone, and A. Montanari, "Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3396–3433, June 2013.

[11] S. Oymak and B. Hassibi, "Sharp MSE bounds for proximal denoising," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 965–1029, 2016.

[12] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. E1181–E1190, 2013.

[13] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth, "Isotonic regression in general dimensions," *arXiv preprint arXiv:1708.09468*, 2017.

[14] J. Renegar, "Incorporating condition measures into the complexity theory of linear programming," *SIAM J. Optim.*, vol. 5, no. 3, pp. 506–524, 1995.

[15] J. C. Vera, J. C. Rivera, J. Peña, and Y. Hui, "A primal-dual symmetric relaxation for homogeneous conic systems," *J. Complexity*, vol. 23, no. 2, pp. 245–261, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jco.2007.01.002

[16] P. Bürgisser and F. Cucker, *Condition: The geometry of numerical algorithms*, ser. Grundlehren der Mathematischen Wissenschaften. Springer Verlag, 2013, no. 349.

[17] V. Roulet, N. Boumal, and A. d'Aspremont, "Computational complexity versus statistical performance on sparse recovery problems," *arXiv preprint arXiv:1506.03295*, 2015.

[18] S. Becker, J. Bobin, and E. Candès, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[19] B. Zhang, W. Xu, J.-F. Cai, and L. Lai, "Precise phase transition of total variation minimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4518–4522.

[20] M. Genzel, G. Kutyniok, and M. März, "ℓ_1-analysis minimization and generalized (co-) sparsity: When does recovery succeed?" *arXiv preprint arXiv:1710.04952*, 2017.

[21] M. Kabanava, H. Rauhut, and H. Zhang, "Robust analysis ℓ_1-recovery from gaussian measurements and total variation minimization," *European Journal of Applied Mathematics*, vol. 26, no. 06, pp. 917–929, 2015.

[22] M. Kabanava and H. Rauhut, "Analysis ℓ_1-recovery with frames and gaussian measurements," *Acta Applicandae Mathematicae*, vol. 140, no. 1, pp. 173–195, 2015.

[23] R. Kueng and D. Gross, "Ripless compressed sensing from anisotropic measurements," *Linear Algebra and its Applications*, vol. 441, pp. 110–123, 2014.

[24] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed sensing*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge: Cambridge University Press, 2012, pp. xii+544, theory and applications.

[25] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-gaussian concentration," *Electron. Commun. Probab*, vol. 18, no. 82, pp. 1–9, 2013.

[26] A. Eftekhari, 2017, private communication.

[27] D. Amelunxen and M. Lotz, "Average-case complexity without the black swans," *Journal of Complexity*, vol. 41, pp. 82–101, 2017.

[28] ——, "Gordon's inequality and condition numbers in conic optimization," *arXiv preprint arXiv:1408.3016*, 2014.

[29] S. Oymak, B. Recht, and M. Soltanolkotabi, "Isometric sketching of any set via the restricted isometry property," *Information and Inference: A Journal of the IMA*, 2015.
[30] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," *Information and Inference: A Journal of the IMA*, 2015.
[31] A. Belloni and R. M. Freund, "A geometric analysis of Renegar's condition number, and its interplay with conic curvature," *Math. Program.*, vol. 119, no. 1, Ser. A, pp. 95–107, 2009.
[32] J. Renegar, "Some perturbation theory for linear programming," *Math. Programming*, vol. 65, no. 1, Ser. A, pp. 73–91, 1994.
[33] R. Schneider and W. Weil, *Stochastic and integral geometry*, ser. Probability and its Applications (New York). Berlin: Springer-Verlag, 2008. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-78859-1
[34] D. Amelunxen and M. Lotz, "Intrinsic volumes of polyhedral cones: a combinatorial perspective," *Discrete & Computational Geometry*, vol. 58, no. 2, pp. 371–409, 2017.
[35] S. Glasauer, "Integralgeometrie konvexer Körper im sphärischen Raum," 1995, thesis, Univ. Freiburg i. Br.
[36] M. B. McCoy and J. A. Tropp, "From Steiner formulas for cones to concentration of intrinsic volumes," *Discrete & Computational Geometry*, vol. 51, no. 4, pp. 926–963, 2014.
[37] D. Amelunxen, "Measures on polyhedral cones: characterizations and kinematic formulas," *arXiv preprint arXiv:1412.1569*, 2014.
[38] R. T. Rockafellar, *Convex analysis*, ser. Princeton Mathematical Series, No. 28. Princeton, N.J.: Princeton University Press, 1970.
[39] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

Dennis Amelunxen was Assistant Professor at the Mathematics Department of City University of Hong Kong. Dennis Amelunxen received his PhD from the University of Paderborn, Germany, in 2011. Before joining City University in 2014, he worked as a postdoctoral fellow at Cornell University, USA, and at The University of Manchester, UK.

Martin Lotz is Associate Professor of Mathematics at the University of Warwick. Prior to joining Warwick, Martin Lotz was a Lecturer in Numerical Analysis at the University of Manchester and held research positions at the City University of Hong Kong, at the University of Oxford, and at the University of Edinburgh, supported by a Leverhulme Trust and Seggie Brown Fellowship. Martin Lotz received his undergraduate degree from the ETH Zürich, and his PhD at the University of Paderborn, with a thesis on Algebraic Complexity Theory.

Jake Walvin completed his PhD in Mathematics at the University of Manchester in 2019.