7-2016

Geometric Inference for General High-Dimensional Linear Inverse Problems

Tony Cai
University of Pennsylvania

Tengyuan Liang
University of Pennsylvania

Alexander Rakhlin
University of Pennsylvania

Recommended Citation

Cai, T., Liang, T., & Rakhlin, A. (2016). Geometric Inference for General High-Dimensional Linear Inverse Problems. *The Annals of Statistics, 44* (4), 1536-1563. http://dx.doi.org/10.1214/15-AOS1426

Geometric Inference for General High-Dimensional Linear Inverse Problems

Abstract
This paper presents a unified geometric framework for the statistical analysis of a general ill-posed linear inverse model which includes as special cases noisy compressed sensing, sign vector recovery, trace regression, orthogonal matrix estimation and noisy matrix completion. We propose computationally feasible convex programs for statistical inference including estimation, confidence intervals and hypothesis testing. A theoretical framework is developed to characterize the local estimation rate of convergence and to provide statistical inference guarantees. Our results are built based on the local conic geometry and duality. The difficulty of statistical inference is captured by the geometric characterization of the local tangent cone through the Gaussian width and Sudakov estimate.

Disciplines
Physical Sciences and Mathematics

Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

GEOMETRIC INFERENCE FOR GENERAL HIGH-DIMENSIONAL LINEAR INVERSE PROBLEMS

By T. Tony Cai[*,‡], Tengyuan Liang[‡] and Alexander Rakhlin[†,‡]

University of Pennsylvania[‡]

This paper presents a unified geometric framework for the statistical analysis of a general ill-posed linear inverse model which includes as special cases noisy compressed sensing, sign vector recovery, trace regression, orthogonal matrix estimation, and noisy matrix completion. We propose computationally feasible convex programs for statistical inference including estimation, confidence intervals and hypothesis testing. A theoretical framework is developed to characterize the local estimation rate of convergence and to provide statistical inference guarantees. Our results are built based on the local conic geometry and duality. The difficulty of statistical inference is captured by the geometric characterization of the local tangent cone through the Gaussian width and Sudakov estimate.

1. Introduction. Driven by a wide range of applications, high-dimensional linear inverse problems such as noisy compressed sensing, sign vector recovery, trace regression, orthogonal matrix estimation, and noisy matrix completion have drawn significant recent interest in several fields, including statistics, applied mathematics, computer science, and electrical engineering. These problems are often studied in a case-by-case fashion, with the main focus on estimation. Although similarities in the technical analyses have been suggested heuristically, a general unified theory for statistical inference including estimation, confidence intervals and hypothesis testing is still yet to be developed.

In this paper, we consider a general linear inverse model

$$Y = \mathcal{X}(M) + Z \tag{1.1}$$

where $M \in \mathbb{R}^p$ is the vectorized version of the parameter of interest, $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ is a linear operator (matrix in $\mathbb{R}^{n\times p}$), and $Z \in \mathbb{R}^n$ is a noise

[*]The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

[†]Alexander Rakhlin gratefully acknowledges the support of NSF under grant CAREER DMS-0954737.

MSC 2010 subject classifications: 62Jxx

Keywords and phrases: linear inverse problems, high-dimensional statistics, statistical inference, convex relaxation, conic geometry, geometric functional analysis

vector. We observe $(\mathcal{X}, Y)$ and wish to recover the unknown parameter M. A particular focus is on the high-dimensional setting where the ambient dimension p of the parameter M is much larger than the sample size n, i.e., the dimension of Y. In such a setting, the parameter of interest M is commonly assumed to have, with respect to a given atom set $\mathcal{A}$, a certain low complexity structure which captures the true dimension of the statistical estimation problem. A number of high-dimensional inference problems actively studied in the recent literature can be seen as special cases of this general linear inverse model.

High Dimension Linear Regression/Noisy Compressed Sensing. In high-dimensional linear regression, one observes (X, Y) with

$$Y = XM + Z, \tag{1.2}$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ with $p \gg n$, $M \in \mathbb{R}^p$ is a sparse signal, and $Z \in \mathbb{R}^n$ is a noise vector. The goal is to recover the unknown sparse signal of interest $M \in \mathbb{R}^p$ based on the observation (X, Y) through an efficient algorithm. Many estimation methods including ℓ_1-regularized procedures such as the Lasso and Dantzig Selector have been developed and analyzed. See, for example, [41, 15, 2, 4] and the references therein. Confidence intervals and hypothesis testing for high-dimensional linear regression have also been actively studied in the last few years. A common approach is to first construct a de-biased Lasso or de-biased scaled-Lasso estimator and then make inference based on the asymptotic normality of low-dimensional functionals of the de-biased estimator. See, for example, [3, 48, 44, 23].

Trace Regression. Accurate recovery of a low-rank matrix based on a small number of linear measurements has a wide range of applications and has drawn much recent attention in several fields. See, for example, [37, 26, 38, 27, 13]. In trace regression, one observes (X_i, Y_i), $i = 1, ..., n$ with

$$Y_i = \mathsf{Tr}(X_i^T M) + Z_i, \tag{1.3}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{p_1\times p_2}$ are measurement matrices, and Z_i are noise. The goal is to recover the unknown matrix $M \in \mathbb{R}^{p_1\times p_2}$ which is assumed to be of low rank. Here the dimension of the parameter M is $p \equiv p_1 p_2 \gg n$. A number of constrained and penalized nuclear minimization methods have been introduced and studied in both the noiseless and noisy settings. See the aforementioned references for further details.

Sign Vector Recovery. The setting of sign vector recovery is similar to the one for the high-dimensional regression except the signal of interest is a sign vector. More specifically, one observes (X, Y) with

$$Y = XM + Z \tag{1.4}$$

where $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}$, $M \in \{+1, -1\}^p$ is a sign vector, and $Z \in \mathbb{R}^n$ is a noise vector. The goal is to recover the unknown sign signal M. Exhaustive search over the parameter set is computationally prohibitive. The noiseless case of (1.4), known as the generalized multi-knapsack problem [25, 31], can be solved through an integer program which is known to be computationally difficult even for checking the uniqueness of the solution, see [36, 43].

Orthogonal Matrix Recovery. In some applications the matrix of interest in trace regression is known to be an orthogonal/rotation matrix [40, 21]. More specifically, in orthogonal matrix recovery, we observe (X_i, Y_i), $i = 1, \ldots, n$ as in the trace regression model (1.3) where $X_i \in \mathbb{R}^{m\times m}$ are measurement matrices and $M \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. The goal is to recover the unknown M using an efficient algorithm. Computational difficulties come in because of the non-convex constraint.

Other high-dimensional inference problems that are closely connected to the structured linear inverse model (1.1) include Matrix Completion [12, 17, 9], sparse and low rank decomposition in robust principal component analysis [11], and sparse noise and sparse parameter in demixing problem [1], to name a few. We will discuss the connections in Section 3.5.5.

There are several fundamental questions for this general class of high-dimensional linear inverse problems:

> **Statistical Questions:** How well can the parameter M be estimated? What is the intrinsic difficulty of the estimation problem? How to provide inference guarantees for M, i.e., confidence intervals and hypothesis testing, in general?
>
> **Computational Questions:** Are there computationally efficient (polynomial time complexity) algorithms that are also sharp in terms of statistical estimation and inference?

1.1. *High-Dimensional Linear Inverse Problems.* Linear inverse problems have been well studied in the classical setting where the parameter of interest lies in a convex set. See, for example, [42], [33], and [24]. In particular, for estimation of a linear functional over a convex parameter space, [18] developed an elegant geometric characterization of the minimax theory in terms of the modulus of continuity. However, the theory relies critically on the convexity assumption of the parameter space. As shown in [7, 8], the behavior of the functional estimation and confidence interval problems is significantly different even when the parameter space is the union of two convex sets. For the high-dimensional linear inverse problems considered in the present paper, the parameter space is highly non-convex and the theory and techniques developed in the classical setting are not readily applicable.

For high-dimensional linear inverse problems such as those mentioned earlier, the parameter space has low-complexity and exhaustive search often leads to the optimal solution in terms of statistical accuracy. However, it is computationally prohibitive and requires the prior knowledge of the true low complexity. In recent years, relaxing the problem to a convex program such as ℓ_1 or nuclear norm minimization and then solving it with optimization techniques has proven to be a powerful approach in individual cases.

Unified approaches to signal recovery recently appeared both in the applied mathematics literature [16, 1, 34] and in the statistics literature [32]. [34] studied the generalized LASSO problem through conic geometry with a simple bound in terms of the ℓ_2 norm of the noise vector (which may not vanish to 0 as sample size n increases). [16] introduced the notion of atomic norm to define a low complexity structure and showed that Gaussian width captures the minimum sample size required to ensure recovery. [1] studied the phase transition for the convex algorithms for a wide range of problems. These suggest that the geometry of the local tangent cone determines the minimum number of samples to ensure successful recovery in the noiseless or deterministic noise settings. [32] studied the regularized M-estimation with a decomposable norm penalty in the additive Gaussian noise setting.

Another line of research is focused on a detailed analysis of the Empirical Risk Minimization (ERM) [28]. The analysis is based on the empirical processes theory, with a proper localized rather than global analysis. In addition to convexity, the ERM requires the prior knowledge on the size of the bounded parameter set of interest. This knowledge is not needed for the algorithm we propose in the present paper.

Compared to estimation, there is a paucity of methods and theoretical results for confidence intervals and hypothesis testing for these linear inverse models. Specifically for high-dimensional linear regression, [3] studied a bias correction method based on ridge estimation, while [48] proposed bias correction via score vector using scaled Lasso as the initial estimator. [44, 23] focused on de-sparsifying Lasso by constructing a near inverse of the Gram matrix [5]; the first paper uses nodewise Lasso, while the other uses ℓ_∞ constrained quadratic programing, with similar theoretical guarantees. To the best of our knowledge, a unified treatment of inference procedures for general high-dimensional linear inverse models is yet to be developed.

1.2. *Geometric Characterization of Linear Inverse Problems.* We take a geometric perspective in studying the model (1.1). The parameter M inherits certain low complexity structure with respect to a given atom set in a high-dimensional space, thus introducing computationally difficult non-convex

constraints. However, proper convex relaxation based on the atom structure provides a computationally feasible solution. For point estimation, we are interested in how the local convex geometry around the true parameter affects the estimation procedure and the intrinsic estimation difficulty. For inference, we develop general procedures induced by the convex geometry, addressing inferential questions such as confidence intervals and hypothesis testing. We are also interested in the sample size condition induced by the local convex geometry for valid inference guarantees. This local geometry plays a key role in our analysis.

Complexity measures such as Gaussian width and Rademacher complexity are well studied in the empirical processes theory [29, 39], and are known to capture the difficulty of the estimation problem. Covering/Packing entropy and volume ratio [46, 45, 30] are also widely used in geometric functional analysis to measure the complexity. In this paper, we will show how these geometric quantities affect the computationally efficient estimation/inference procedure, as well as the intrinsic difficulties.

1.3. *Our Contributions.* The main result can be summarized as follows:

Unified convex algorithms. We propose a general computationally feasible convex program that provides near optimal rate of convergence simultaneously for a collection of high-dimensional linear inverse problems. We also study a general efficient convex program that leads to statistical inference for linear contrasts of M, such as confidence intervals and hypothesis testing. The point estimation and statistical inference are adaptive in the sense that the difficulty (rate of convergence, conditions on sample size, etc.) automatically adapts to the low complexity structure of the true parameter.

Local geometric theory. A unified theoretical framework is provided for analyzing high-dimensional linear inverse problems based on the local conic geometry and duality. Local geometric complexities govern the difficulty of statistical inference for the linear inverse problems.

Specifically, on the local tangent cone $T_{\mathcal{A}}(M)$ (defined in (2.4)), geometric quantities such as the Gaussian width $w(B_2^p \cap T_{\mathcal{A}}(M))$ and Sudakov minoration estimate $e(B_2^p \cap T_{\mathcal{A}}(M))$ (both defined in Section 2.2; B_2^p denotes unit Euclidean ball in $\mathbb{R}^p$) capture the rate of convergence. In terms of the upper bound, with overwhelming probability, if $n \gtrsim w^2(B_2^p \cap T_{\mathcal{A}}(M))$, the estimation error under ℓ_2 norm for our algorithm is

$$\sigma \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}},$$

where $\gamma_{\mathcal{A}}(M)$ is the local asphericity ratio defined in (2.11). A minimax lower bound for estimation over the local tangent cone $T_{\mathcal{A}}(M)$ is

$$\sigma \frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}}.$$

For statistical inference, we establish valid asymptotic normality for any linear functional $\langle v, M\rangle$ (with $\|v\|_{\ell_1}$ bounded) of the parameter M under the condition

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

which can be compared to the condition for point estimation consistency

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0.$$

There is a critical difference on the sufficient conditions between valid inference and estimation consistency — more stringent condition on sample size n is required for inference beyond estimation. Intuitively, statistical inference is purely geometrized by Gaussian width and Sudakov minoration estimate.

1.4. *Organization of the Paper.* The rest of the paper is structured as follows. In Section 2, after notation, definitions, and basic convex geometry are reviewed, we formally present convex programs for recovering the parameter M, and for providing inference guarantees for M. The properties of the proposed procedures are then studied in Section 3 under the Gaussian setting, where a geometric theory is developed, along with the minimax lower bound, as well as the confidence intervals and hypothesis testing. Applications to particular high-dimensional estimation problems are caculated in Section 3.5. Section 4 extends the geometric theory beyond the Gaussian case. Further discussions appear in Section 5, and the proofs of the main results are given in Section 6 and Supplement [6].

2. Preliminaries and Algorithms. Let us first review notation and definitions that will be used in the rest of the paper. We use $\|\cdot\|_{\ell_q}$ to denote the ℓ_q norm of a vector or induced norm of a matrix, and use B_2^p to denote the unit Euclidean ball in $\mathbb{R}^p$. For a matrix M, denote by $\|M\|_F$, $\|M\|_*$, and $\|M\|$ the Frobenius norm, nuclear norm, and spectral norm of M respectively. When there is no confusion, we also denote $\|M\|_F = \|M\|_{\ell_2}$ for a matrix M. For a vector $V \in \mathbb{R}^p$, denote its transpose by V^*. The inner product on vectors is defined as usual $\langle V_1, V_2\rangle = V_1^* V_2$. For matrices

$\langle M_1, M_2\rangle = \mathsf{Tr}(M_1^* M_2) = \mathsf{Vec}(M_1)^* \mathsf{Vec}(M_2)$, where $\mathsf{Vec}(M) \in \mathbb{R}^{pq}$ denotes the vectorized version of matrix $M \in \mathbb{R}^{p\times q}$. $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ denotes a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^n$. Following the notation above, $M^* \in \mathbb{R}^{q\times p}$ is the adjoint (transpose) matrix of M and $\mathcal{X}^* : \mathbb{R}^n \to \mathbb{R}^p$ is the adjoint operator of $\mathcal{X}$ such that $\langle \mathcal{X}(V_1), V_2\rangle = \langle V_1, \mathcal{X}^*(V_2)\rangle$.

For a convex compact set K in a metric space with the metric d, the ϵ-entropy for a convex compact set K with respect to the metric d is denoted in the following way: ϵ-packing entropy $\log \mathcal{M}(K, \epsilon, d)$ is the logarithm of the cardinality of the largest ϵ-packing set. Similarly, ϵ-covering entropy $\log \mathcal{N}(K, \epsilon, d)$ is the log-cardinality of the smallest ϵ-covering set with respect to metric d. A well known result is $\mathcal{M}(K, 2\epsilon, d) \le \mathcal{N}(K, \epsilon, d) \le \mathcal{M}(K, \epsilon, d)$. When the metric d is the usual Euclidean distance, we will omit d in $\mathcal{M}(K, \epsilon, d)$ and $\mathcal{N}(K, \epsilon, d)$ and simply write $\mathcal{M}(K, \epsilon)$ and $\mathcal{N}(K, \epsilon)$.

For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we denote $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ if there exist constants c_0, C_0 such that $\frac{a_n}{b_n} \ge c_0$ and $\frac{a_n}{b_n} \le C_0$ respectively, for all n. We write $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. Throughout the paper, c, C denote constants that may vary from place to place.

2.1. *Basic Convex Geometry.* The notion of low complexity is based on a collection of basic atoms. We denote the collection of these basic atoms as an atom set $\mathcal{A}$, either countable or uncountable. A parameter M is of complexity k in terms of the atoms in $\mathcal{A}$ if M can be expressed as a linear combination of at most k atoms in $\mathcal{A}$, i.e., there exists a decomposition

$$M = \sum_{a\in\mathcal{A}} c_a(M)\cdot a, \text{ where } \sum_{a\in\mathcal{A}} 1_{\{c_a(M)\neq 0\}} \le k.$$

In convex geometry [35], the Minkowski functional (gauge) of a symmetric convex body K is defined as

$$\|x\|_K = \inf\{t > 0 : x \in tK\}.$$

Let $\mathcal{A}$ be a collection of atoms that is a compact subset of $\mathbb{R}^p$. Without loss of generality, assume $\mathcal{A}$ is contained inside ℓ_∞ ball. We assume that the elements of $\mathcal{A}$ are extreme points of the convex hull $\mathsf{conv}(\mathcal{A})$ (in the sense that for any $x \in \mathbb{R}^p$, $\sup\{\langle x, a\rangle : a \in \mathcal{A}\} = \sup\{\langle x, a\rangle : a \in \mathsf{conv}(\mathcal{A})\}$). The atomic norm $\|x\|_{\mathcal{A}}$ for any $x \in \mathbb{R}^p$ is defined as the gauge of $\mathsf{conv}(\mathcal{A})$:

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 : x \in t\ \mathsf{conv}(\mathcal{A})\}.$$

As noted in [16], the atomic norm can also be written as

$$\|x\|_{\mathcal{A}} = \inf\left\{\sum_{a\in\mathcal{A}} c_a : x = \sum_{a\in\mathcal{A}} c_a\cdot a,\ c_a \ge 0\right\}. \tag{2.1}$$

The dual norm of this atomic norm is defined in the following way (since the atoms in $\mathcal{A}$ are the extreme points of $\mathsf{conv}(\mathcal{A})$),

$$\|x\|_{\mathcal{A}}^* = \sup\{\langle x, a\rangle : a \in \mathcal{A}\} = \sup\{\langle x, a\rangle : \|a\|_{\mathcal{A}} \leq 1\}. \tag{2.2}$$

We have the following ("Cauchy-Schwarz") symmetric relation for the norm and its dual

$$\langle x, y\rangle \leq \|x\|_{\mathcal{A}}^* \|y\|_{\mathcal{A}}. \tag{2.3}$$

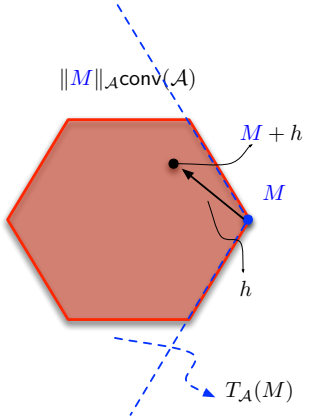

FIG 1. *Tangent cone: general illustration in 2D. The red shaped area is the scaled convex hull of atom set. The blue dashed line forms the tangent cone at M. Black arrow denotes the possible directions inside the cone.*

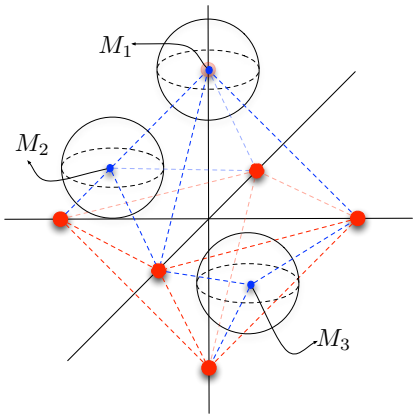

FIG 2. *Tangent cone illustration in 3D for sparse regression. For three possible locations $M_i, 1 \leq i \leq 3$, the tangent cone are different, with cones becoming more complex as i increases.*

It is clear that the unit ball with respect to the atomic norm $\|\cdot\|_{\mathcal{A}}$ is the convex hull of the set of atoms $\mathcal{A}$. The **tangent cone** at x with respect to the scaled unit ball $\|x\|_{\mathcal{A}}\ \mathsf{conv}(\mathcal{A})$ is defined to be

$$T_{\mathcal{A}}(x) = \mathsf{cone}\left\{h : \|x + h\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\right\}. \tag{2.4}$$

Also known as a recession cone, $T_{\mathcal{A}}(x)$ is the collection of directions where the atomic norm becomes smaller. The "size" of the tangent cone at the true parameter M will affect the difficulty of the recovery problem. We focus on the cone intersected with the unit ball $B_2^p \cap T_{\mathcal{A}}(M)$ in analyzing the complexity of the cone. See Figure 1 for an intuitive illustration.

It is helpful to look at the atom set, atomic norm and tangent cone geometry in a few examples to better illustrate the general model and notion of low complexity.

EXAMPLE 1. For sparse signal recovery in high-dimensional linear regression, the atom set consists of the unit basis vectors $\{\pm e_i\}$, the atomic norm is the vector ℓ_1 norm, and its dual norm is the vector ℓ_∞ norm. The convex hull $\mathsf{conv}(\mathcal{A})$ is called the cross-polytope. Figure 2 illustrates this tangent cone for 3D ℓ_1 norm ball for 3 different cases $T_{\mathcal{A}}(M_i), 1 \leq i \leq 3$. The "angle" or "complexity" of the local tangent cone determines the difficulty of recovery. Previous work showed that the algebraic characterization (sparsity) of the parameter space drives the global rate, and we are arguing that the geometric characterization through the local tangent cone provides an intuitive and refined local approach.

EXAMPLE 2. In trace regression and matrix completion, the goal is to recover low rank matrices. In such settings, the atom set consists of the rank one matrices (matrix manifold) $\mathcal{A} = \{uv^* : \|u\|_{\ell_2} = 1,\ \|v\|_{\ell_2} = 1\}$ and the atomic norm is the nuclear norm and the dual norm is the spectral norm. The convex hull $\mathsf{conv}(\mathcal{A})$ is called the nuclear norm ball of matrices. The position of the true parameter on the scaled nuclear norm ball determines the geometry of the local tangent cone, thus affecting the estimation difficulty.

EXAMPLE 3. In integer programming, one would like to recover the sign vectors whose entries take on values ± 1. The atom set is all sign vectors (cardinality 2^p) and the convex hull $\mathsf{conv}(\mathcal{A})$ is the hypercube. Tangent cones for each parameter have the same structure in this case.

EXAMPLE 4. In orthogonal matrix recovery, the matrix of interest is constrained to be orthogonal. In this case, the atom set is all orthogonal matrices and the convex hull $\mathsf{conv}(\mathcal{A})$ is the spectral norm ball. Similar to sign vector recovery, the local tangent cones for each orthogonal matrix share similar geometric property.

2.2. *Gaussian Width, Sudakov Estimate, and Other Geometric Quantities.* We first introduce two complexity measures, the Gaussian width and Sudakov estimate.

DEFINITION 1 (Gaussian Width). *For a compact set $K \in \mathbb{R}^p$, the Gaussian width is defined as*

$$w(K) := \mathbb{E}_g \left[\sup_{v \in K} \langle g, v \rangle \right]. \tag{2.5}$$

where $g \sim N(0, I_p)$ is the standard multivariate Gaussian vector.

Gaussian width quantifies the probability that a randomly oriented subspace misses a convex subset. It was used in Gordon's analysis [20], and was shown recently to play a crucial rule in linear inverse problems in various noiseless or deterministic noise settings, see, for example, [16, 1]. Explicit upper bounds on the Gaussian width for different convex sets have been given in [16, 1]. For example, if $M \in \mathbb{R}^p$ is a s-sparse vector, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{s \log p/s}$. When $M \in \mathbb{R}^{p\times q}$ is a rank-r matrix, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{r(p+q-r)}$. For sign vector in $\mathbb{R}^p$, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{p}$, while for orthogonal matrix in $\mathbb{R}^{m\times m}$, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{m(m-1)}$. See Section 3.4 propositions 3.10-3.14 in [16] for detailed calculations. The Gaussian width as a complexity measure of the local tangent cone will be used in the upper bound analysis in Sections 3 and 4.

Definition 2 (Sudakov Minoration Estimate). *The Sudakov estimate of a compact set $K \in \mathbb{R}^p$ is defined as*

$$e(K) := \sup_{\epsilon} \ \epsilon\sqrt{\log \mathcal{N}(K,\epsilon)}. \tag{2.6}$$

where $\mathcal{N}(K,\epsilon)$ denotes the ϵ-covering number of set K with respect to the Euclidean norm.

Sudakov estimate has been used in the literature as a measure of complexity for a general functional class that nearly matches (from below) the expected supremum of a gaussian process. By balancing the cardinality of the covering set at scale ϵ and the covering radius ϵ, the estimate maximizes

$$\epsilon\sqrt{\log \mathcal{N}(B_2^p \cap T_{\mathcal{A}}(M),\epsilon)},$$

thus determining the complexity of the cone $T_{\mathcal{A}}(M)$. Sudakov estimate as a complexity measure of the local tangent cone is useful for the minimax lower bound analysis.

The following well-known result [19, 29] establishes a relation between the Gaussian width $w(\cdot)$ and Sudakov estimate $e(\cdot)$:

Lemma 1 (Sudakov Minoration and Dudley Entropy Integral). *For any compact subset $K \subseteq \mathbb{R}^p$, there exist a universal constant $c > 0$ such that*

$$c \cdot e(K) \le w(K) \le 24 \int_0^\infty \sqrt{\log \mathcal{N}(K,\epsilon)}d\epsilon. \tag{2.7}$$

In the literature, another complexity measure—volume ratio—has also been used to characterize the minimax lower bounds [30]. Volume ratio has

been studied in [35] and [45]. For a convex set $K \in \mathbb{R}^p$, volume ratio used in the present paper is defined as follows.

DEFINITION 3 (Volume Ratio). *The volume ratio is defined as*

$$v(K) := \sqrt{p}\left(\frac{\mathsf{vol}(K)}{\mathsf{vol}(B_2^p)}\right)^{\frac{1}{p}}. \tag{2.8}$$

The recovery difficulty of the linear inverse problem also depends on other geometric quantities defined on the local tangent cone $T_{\mathcal{A}}(M)$: the local isometry constants $\phi_{\mathcal{A}}(M, \mathcal{X})$ and $\psi_{\mathcal{A}}(M, \mathcal{X})$ and the local asphericity ratio $\gamma_{\mathcal{A}}(M)$. The **local isometry constants** are defined for the local tangent cone at the true parameter M as

$$\phi_{\mathcal{A}}(M, \mathcal{X}) := \inf\left\{\frac{\|\mathcal{X}(h)\|_{\ell_2}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\} \tag{2.9}$$

$$\psi_{\mathcal{A}}(M, \mathcal{X}) := \sup\left\{\frac{\|\mathcal{X}(h)\|_{\ell_2}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\}. \tag{2.10}$$

The local isometry constants measure how well the linear operator preserves the ℓ_2 norm within the local tangent cone. Intuitively, the larger the ψ or the smaller the ϕ is, the harder the recovery is. We will see later that the local isometry constants are determined by the Gaussian width under the Gaussian ensemble design.

The **local asphericity ratio** is defined as

$$\gamma_{\mathcal{A}}(M) := \sup\left\{\frac{\|h\|_{\mathcal{A}}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\} \tag{2.11}$$

and measures how extreme the atomic norm is relative to the ℓ_2 norm within the local tangent cone.

2.3. *Point Estimation via Convex Relaxation.* We now return to the linear inverse model (1.1) in the high-dimensional setting. Suppose we observe $(\mathcal{X},\ Y)$ as in (1.1) where the parameter of interest M is assumed to have low complexity with respect to a given atom set $\mathcal{A}$. The low complexity of M introduces a non-convex constraint, which leads to serious computational difficulties if solved directly. Convex relaxation is an effective and natural approach in such a setting. In most interesting cases, the atom set is not too rich in the sense that $\mathsf{conv}(\mathcal{A}) \subset B_2^p$. For such cases, we propose a generic convex constrained minimization procedure induced by the atomic norm and the corresponding dual norm to estimate M:

$$\hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} :\ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \leq \lambda\} \tag{2.12}$$

where λ is a localization radius (tuning parameter) that depends on the sample size, noise level, and geometry of the atom set $\mathcal{A}$. An explicit formula for λ is given in (3.1) in the case of Gaussian noise. The atomic norm minimization (2.12) is a convex relaxation of the low complexity structure, and λ specifies the localization scale based on the noise. This generic convex program utilizes the duality and recovers the low complexity structure adaptively. The Dantzig selector for high-dimensional sparse regression [15] and the constrained nuclear norm minimization [13] for trace regression are particular examples of (2.12). The properties of the estimator $\hat{M}$ will be investigated in Sections 3 and 4.

In cases where the atomic norm ball is rich, i.e. $\mathsf{conv}(\mathcal{A}) \not\subset B_2^p$, a slightly stronger program

$$\hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} : \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \le \lambda, \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\ell_2} \le \mu\} \tag{2.13}$$

with λ, μ as tuning parameters will yield optimal guarantees. The analysis of (2.13) is essentially the same as (2.12). For conciseness, we will present the main result for the interesting case (2.12). We remark that the atomic dual norm constraint is crucial for attaining optimal behavior unless $\mathsf{conv}(\mathcal{A}) \supset B_2^p$. For instance, the convex program in [16] with only the ℓ_2 constraint will lead to a suboptimal estimator.

2.4. *Statistical Inference via Feasibility of Convex Program.* In the high-dimensional setting, p-values as well as confidence intervals are important inferential questions beyond point estimation. In this section we will show how to perform statistical inference for the linear inverse model (1.1). Let $M \in \mathbb{R}^p$ be the vectorized parameter of interest, and $\{e_i, 1 \le i \le p\}$ are the corresponding basis vectors. Consider the following convex feasibility problem for matrix $\Omega \in \mathbb{R}^{p\times p}$, where each row $\Omega_{i\cdot}$ satisfies

$$\|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall 1 \le i \le p. \tag{2.14}$$

Here η is some tuning parameter that depends on the sample size and geometry of the atom set $\mathcal{A}$. One can also solve a stronger version of the above convex program for $\eta \in \mathbb{R}, \Omega \in \mathbb{R}^{p\times p}$ simultaneously:

$$(\Omega, \eta_n) = \arg\min_{\Omega,\eta} \{\eta : \ \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall 1 \le i \le p\}. \tag{2.15}$$

Built upon the constrained minimization estimator $\hat{M}$ in (2.12) and feasible matrix Ω in (2.15), the de-biased estimator for inference on parameter

M is defined as

$$\tilde{M} := \hat{M} + \Omega \mathcal{X}^*(Y - \mathcal{X}(\hat{M})). \tag{2.16}$$

We will establish the asymptotic normality for linear contrast $\langle v, M \rangle$, where $v \in \mathbb{R}^p, \|v\|_{\ell_1} \leq \rho$, ρ does not grow with $n, p(n)$, and construct confidence intervals and hypothesis tests based on the asymptotic normality result. In the case of high-dimensional linear regression, de-biased estimators has been investigated in [3, 48, 44, 23]. The convex feasibility program we proposed here can be viewed as a unified treatment for general linear inverse models. We will show that under some conditions on the sample size and the local tangent cone, asymptotic confidence intervals and hypothesis tests are valid for linear contrast $\langle v, M \rangle$ which include as a special case the individual coordinates of M.

3. Local Geometric Theory: Gaussian Setting. We establish in this section a general theory of geometric inference in the Gaussian setting where the noise vector Z is Gaussian and the linear operator $\mathcal{X}$ is the Gaussian ensemble design (Definition 4). In analyzing Model 1.1, without loss of generality, we can scale $\mathcal{X}, Z$ simultaneously such that column ℓ_2 norm does not grow with n. In the stochastic noise setting, the noise $Z_i, 1 \leq i \leq n$ is scaled correspondingly to noise level $\sigma/\sqrt{n}$.

DEFINITION 4 (Gaussian Ensemble Design). *Let $\mathcal{X} \in \mathbb{R}^{n \times p}$ be the matrix form of the linear operator $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$. $\mathcal{X}$ is Gaussian ensemble if each element is an i.i.d Gaussian random variable with mean 0 and variance $\frac{1}{n}$.*

Our analysis is quite different from the case by case global analysis of the Dantzig selector, Lasso and nuclear norm minimization. We show a stronger result which adapts to the local tangent cone geometry. All the analyses in our theory are non-asymptotic, and the constants are explicit. Another advantage is that the local analysis yields robustness for a given parameter (with near but not exact low complexity), as the convergence rate is captured by the geometry of the associated local tangent cone at a given M. Later in Section 4 we will show how to extend the theory to a more general setting.

3.1. *Local Geometric Upper Bound.* For the upper bound analysis, we need to choose a suitable localization radius λ (in the convex program (2.12)) to guarantee that the true parameter M is in the feasible set with high probability. In the case of Gaussian noise the tuning parameter is chosen as

$$\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n) = \frac{\sigma}{\sqrt{n}} \left\{ w(\mathcal{X}\mathcal{A}) + \delta \cdot \sup_{v \in \mathcal{A}} \|\mathcal{X}v\|_{\ell_2} \right\} \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}) \tag{3.1}$$

where $\mathcal{X}T$ is the image of the set T under the linear operator $\mathcal{X}$, and $\delta > 0$ can be chosen arbitrarily according to the probability of success we would like to attain (δ is commonly chosen at order $\sqrt{\log p}$). $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$ is a global parameter that depends on the linear operator $\mathcal{X}$ and the atom set $\mathcal{A}$, but, importantly, not on the complexity of M. The following theorem geometrizes the local rate of convergence in the Gaussian case.

THEOREM 1 (Gaussian Ensemble: Convergence Rate). *Suppose we observe $(\mathcal{X}, Y)$ as in (1.1) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Let $\hat{M}$ be the solution of (2.12) with λ chosen as in (3.1). Let $0 < c < 1$ be a constant. For any $\delta > 0$, if*

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c},$$

then with probability at least $1 - 3\exp(-\delta^2/2)$,

$$\|\hat{M} - M\|_{\mathcal{A}} \leq \gamma_{\mathcal{A}}(M) \cdot \|\hat{M} - M\|_{\ell_2}, \quad \text{and further we have}$$
$$\|\hat{M} - M\|_{\ell_2} \leq \frac{1}{1-c} \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq \frac{2\sigma}{(1-c)^2} \cdot \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}}.$$

Theorem 1 gives bounds for the estimation error under both the ℓ_2 norm loss and the atomic norm loss, as well as for the in sample prediction error. The upper bounds are determined by the geometric quantities $w(\mathcal{X}\mathcal{A}), \gamma_{\mathcal{A}}(M)$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$. Take, for example, the estimation error under the ℓ_2 loss. Given any $\epsilon > 0$, the smallest sample size n to ensure the recovery error $\|\hat{M} - M\|_{\ell_2} \leq \epsilon$ with probability at least $1 - 3\exp(-\delta^2/2)$ is

$$n \geq \max\left\{ \frac{4\sigma^2}{(1-c)^4} \cdot \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\epsilon^2}, \ \frac{4w^2(B_2^p \cap T_{\mathcal{A}}(M))}{c^2} \right\}.$$

That is, the minimum sample size for guaranteed statistical accuracy is driven by two geometric terms $w(\mathcal{X}\mathcal{A})\gamma_{\mathcal{A}}(M)$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$. We will see in Section 3.5 that these two rates match in a range of specific high-dimensional estimation problems.

The proof of Theorem 1 (and Theorem 4 in Section 4) relies on the following two key lemmas.

LEMMA 2 (Choice of Tuning Parameter). *Consider the linear inverse model (1.1) with $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. For any $\delta > 0$, with probability at least $1 - \exp(-\delta^2/2)$ on the σ-field of Z (conditional on $\mathcal{X}$),*

$$\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \leq \frac{\sigma}{\sqrt{n}} \left\{ w(\mathcal{X}\mathcal{A}) + \delta \cdot \sup_{v \in \mathcal{A}} \|\mathcal{X}v\|_{\ell_2} \right\}. \tag{3.2}$$

This lemma is proved in Section 6. The particular value of $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$ for a range of examples will be calculated in Section 3.5.

The next lemma addresses the local behavior of the linear operator $\mathcal{X}$ around the true parameter M under the Gaussian ensemble design. We call a linear operator *locally near-isometric* if the local isometry constants are uniformly bounded. The following lemma tells us that in the most widely used Gaussian ensemble case, the local isometry constants are guaranteed to be bounded, given the sample size n is at least of order $[w(B_2^p \cap T_{\mathcal{A}}(M))]^2$. Hence, the difficulty of the problem is captured by the Gaussian width.

LEMMA 3 (Local Isometry Bound for Gaussian Ensemble). *Assume the linear operator $\mathcal{X}$ is the Gaussian ensemble design. Let $0 < c < 1$ be a constant. For any $\delta > 0$, if*

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c},$$

then with probability at least $1 - 2\exp(-\delta^2/2)$, the local isometry constants are around 1 with

$$\phi_{\mathcal{A}}(M, \mathcal{X}) \geq 1 - c \quad \textit{and} \quad \psi_{\mathcal{A}}(M, \mathcal{X}) \leq 1 + c.$$

3.2. *Local Geometric Inference: Confidence Intervals and Hypothesis Testing.* For statistical inference on the general linear inverse model, we would like to choose the smallest η in (2.14) to ensure that, under the Gaussian ensemble design, the feasibility set for (2.14) is non-empty with high probability. The following theorem establishes geometric inference for Model (1.1).

THEOREM 2 (Geometric Inference). *Suppose we observe $(\mathcal{X}, Y)$ as in (1.1) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Let $\hat{M} \in \mathbb{R}^p, \Omega \in \mathbb{R}^{p \times p}$ be the solution of (2.12) and (2.14) , and let $\tilde{M} \in \mathbb{R}^p$ be the de-biased estimator as in (2.16). Assume $p \geq n \gtrsim w^2(B_2^p \cap T_{\mathcal{A}}(M))$. If the tuning parameters λ, η are chosen with*

$$\lambda \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}), \quad \eta \asymp \frac{1}{\sqrt{n}} w(\mathcal{X}\mathcal{A}),$$

convex programs (2.12) and (2.14) have non-empty feasibility set for Ω with high probability.

The following decomposition

$$\tilde{M} - M = \Delta + \frac{\sigma}{\sqrt{n}} \Omega \mathcal{X}^* W \tag{3.3}$$

holds, where $W \sim N(0, I_n)$ *is the standard Gaussian vector with*

$$\Omega\mathcal{X}^*W \sim N(0, \Omega\mathcal{X}^*\mathcal{X}\Omega^*)$$

and $\Delta \in \mathbb{R}^p$ *satisfies* $\|\Delta\|_{\ell_\infty} \precsim \gamma_{\mathcal{A}}^2(M) \cdot \lambda\eta \asymp \sigma\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{n}$. *Suppose* $(n, p(n))$ *as a sequence satisfies*

$$\limsup_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

then for any $v \in \mathbb{R}^p, \|v\|_{\ell_1} \le \rho$ *with* ρ *finite, we have the asymptotic normality for the functional* $\langle v, \tilde{M}\rangle$,

$$\frac{\sqrt{n}}{\sigma}\left(\langle v, \tilde{M}\rangle - \langle v, M\rangle\right) = \sqrt{v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v} \cdot Z_0 + o_p(1) \tag{3.4}$$

where $Z_0 \sim N(0,1)$ *and* $\lim_{n,p(n)\to\infty} o_p(1) = 0$ *means convergence in probability.*

It follows from Theorem 2 that a valid asymptotic $(1-\alpha)$-level confidence intervals for $M_i, 1 \le i \le p$ (when v is taken as e_i in Theorem 2) is

$$\left[\tilde{M}_i + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}},\quad \tilde{M}_i + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}}\right]. \tag{3.5}$$

If we are interested in a linear contrast $\langle v, M\rangle = v_0$, $\|v\|_{\ell_1} \le \rho$ with ρ fixed, consider the hypothesis testing problem

$$H_0 : \sum_{i=1}^p v_i M_i = v_0 \quad \text{v.s.} \quad H_\alpha : \sum_{i=1}^p v_i M_i \ne v_0.$$

The test statistic is $\frac{\sqrt{n}\left(\langle v,\tilde{M}\rangle - v_0\right)}{\sigma(v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)^{1/2}}$ and under the null, it follows an asymptotic standard normal distribution as $n \to \infty$. Similarly, the p-value is of the form $2 - 2\Phi^{-1}\left(\left|\frac{\sqrt{n}\left(\langle v,\tilde{M}\rangle - v_0\right)}{\sigma(v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)^{1/2}}\right|\right)$ as $n \to \infty$.

Note the asymptotic normality holds for any finite linear contrast, and the asymptotic variance nearly achieves the Fisher information lower bound, as Ω is an estimate of the inverse of $\mathcal{X}^*\mathcal{X}$. For fixed dimension inference, Fisher information lower bound is asymptotically optimal.

REMARK 1. Note that the condition required for asymptotic normality and valid confidence intervals,

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

is stronger than the one for estimation consistency of the parameter M under the ℓ_2 norm,

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0.$$

For inference, we do require stronger condition in order to learn the order of the bias of the estimate. In the case when $n > p$ and the Gaussian ensemble design, $\mathcal{X}^*\mathcal{X}$ is non-singular with high probability. With the choice of $\Omega = (\mathcal{X}^*\mathcal{X})^{-1}$ and $\eta = 0$, for any $i \in [p]$, the following holds non-asymptotically,

$$\sqrt{n}(\tilde{M}_i - M_i) \sim N(0, \sigma^2[(\mathcal{X}^*\mathcal{X})^{-1}]_{ii}).$$

3.3. *Extension: Correlated Design.* The results in Section 3.1 and 3.2 can be extended beyond Gaussian ensemble (where $\mathbb{E}[\mathcal{X}^*\mathcal{X}] = I$) to Gaussian design with known covariance matrix Σ (where $\mathbb{E}[\mathcal{X}^*\mathcal{X}] = \Sigma$). Consider the following slightly modified point estimation and inference procedure (with tuning parameter λ, η)

$$\begin{aligned} &\text{Point Estimation via } \hat{M} \quad \hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} : \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \le \lambda\} \\ &\text{Inference via } \tilde{M} \quad \Omega : \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - \Sigma^{\frac{1}{2}} e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall 1 \le i \le p \\ &\qquad\qquad\qquad\quad \tilde{M} := \hat{M} + \Sigma^{-\frac{1}{2}}\Omega\mathcal{X}^*(Y - \mathcal{X}(\hat{M})) \end{aligned} \tag{3.6}$$

where $\Omega \in \mathbb{R}^{p\times p}$ is an solution to the convex feasibility problem (3.6). Then the following Corollary holds.

COROLLARY 1. *Suppose we observe $(\mathcal{X}, Y)$ as in (1.1), where the Gaussian design $\mathcal{X}$ has covariance Σ and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Consider the convex programs for estimation $\hat{M}$ and inference $\tilde{M}$ with the tuning parameters chosen as*

$$\lambda \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}), \quad \eta \asymp \frac{1}{\sqrt{n}} w(\mathcal{X}\mathcal{A}).$$

Under the condition $n \gtrsim w(B_2^p \cap \Sigma^{\frac{1}{2}} \circ T_{\mathcal{A}}(M))$, $\hat{M}$ satisfies

$$\|\hat{M} - M\|_{\ell_2} \precsim \sigma \frac{\gamma_{\mathcal{A}}(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}}, \quad \|\hat{M} - M\|_{\mathcal{A}} \precsim \sigma \frac{\gamma_{\mathcal{A}}^2(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}}.$$

Suppose $(n, p(n))$ *as a sequence satisfies*

$$\limsup_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

then for any $v \in \mathbb{R}^p, \|v\|_{\ell_1} \le \rho$ *with* ρ *finite, we have the asymptotic normality for the functional* $\langle \Sigma^{\frac{1}{2}} v, \tilde{M}\rangle$,

$$\frac{\sqrt{n}}{\sigma}\left(\langle \Sigma^{\frac{1}{2}} v, \tilde{M}\rangle - \langle \Sigma^{\frac{1}{2}} v, M\rangle\right) = \sqrt{v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v}\cdot Z_0 + o_p(1)$$

where $Z_0 \sim N(0,1)$ *and* $\lim_{n,p(n)\to\infty} o_p(1) = 0$ *means convergence in probability.*

3.4. *Minimax Lower Bound for Local Tangent Cone.* As seen in Section 3.1 and 3.2, the local tangent cone plays an important role in the upper bound analysis. In this section, we are interested in restricting the parameter space to the local tangent cone and seeing how the geometry of the cone affects the minimax lower bound.

THEOREM 3 (Lower bound Based on Local Tangent Cone). *Suppose we observe* $(\mathcal{X}, Y)$ *as in* (1.1) *with the Gaussian ensemble design and* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. *Let* M *be the true parameter of interest. Let* $0 < c < 1$ *be a constant. For any* $\delta > 0$, *if* $n \ge \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c}$. *Then with probability at least* $1 - 2\exp(-\delta^2/2)$,

$$\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{\ell_2}^2 \ge \frac{c_0\sigma^2}{(1+c)^2} \cdot \left(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}}\right)^2$$

for some universal constant $c_0 > 0$. *Here* $\mathbb{E}_{\cdot|\mathcal{X}}$ *stands for the conditional expectation given the design matrix* $\mathcal{X}$, *and the probability statement is with respect to the distribution of* $\mathcal{X}$ *under the Gaussian ensemble design.*

Recall Theorem 1, the local upper bound is basically determined by $\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})$, which in many examples in Section 3.5 is of the rate $w^2(B_2^p \cap T_{\mathcal{A}}(M))$. The general relationship between these two quantities is given in Lemma 4 below, which is proved in Supplement [6] Section A.

LEMMA 4. *For any atom set* $\mathcal{A}$, *we have the following relation*

$$\gamma_{\mathcal{A}}(M)w(\mathcal{A}) \ge w(B_2^p \cap T_{\mathcal{A}}(M))$$

where $w(\cdot)$ *is the Gaussian width and* $\gamma_{\mathcal{A}}(M)$ *is defined in* (2.11).

From Theorem 3, the minimax lower bound for estimation over the local tangent cone is determined by the Sudakov estimate $e^2(B_2^p \cap T_{\mathcal{A}}(M))$. It follows directly from Lemma 1 that there exists a universal constant $c > 0$ such that $c \cdot e(B_2^p \cap T_{\mathcal{A}}(M)) \leq w(B_2^p \cap T_{\mathcal{A}}(M)) \leq 24 \int_0^\infty \sqrt{\log \mathcal{N}(B_2^p \cap T_{\mathcal{A}}(M), \epsilon)} d\epsilon$. Thus under the Gaussian setting, both in terms of the upper bound and lower bound, geometric complexity measures govern the difficulty of the estimation problem, through closely related quantities: Gaussian width and Sudakov estimate.

3.5. *Application of the Geometric Approach.* In this section we apply the general theory under the Gaussian setting to some of the actively studied high-dimensional problems mentioned in Section 1 to illustrate the wide applicability of the theory. The detailed proofs are deferred to Supplement [6] Section B.

3.5.1. *High-Dimensional Linear Regression.* We begin by considering the high-dimensional linear regression model (1.2) under the assumption that the true parameter $M \in \mathbb{R}^p$ is sparse, say $\|M\|_{l_0} = s$. Our general theory applying to the ℓ_1 minimization recovers the optimality results as in Dantzig selector and Lasso. In this case, it can be shown that $\gamma_{\mathcal{A}}(M) w(\mathcal{A})$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$ are of the same rate $\sqrt{s \log p}$. See Supplement [6] Section B for the detailed calculations. The asphericity ratio $\gamma_{\mathcal{A}}(M) \leq 2\sqrt{s}$ reflects the sparsity of M through the local tangent cone and the Gaussian width $w(\mathcal{X}\mathcal{A}) \asymp \sqrt{\log p}$. The following corollary follows from the geometric analysis of the high-dimensional regression model.

COROLLARY 2. *Consider the linear regression model (1.2). Assume that $\mathcal{X} \in \mathbb{R}^{n \times p}$ is the Gaussian ensemble design and the parameter of interest $M \in \mathbb{R}^p$ is of sparsity s. Let $\hat{M}$ be the solution to the constrained ℓ_1 minimization (2.12) with $\lambda = C_1 \sigma \sqrt{\frac{\log p}{n}}$. If $n \geq C_2 s \log p$, then*

$$\|\hat{M} - M\|_{\ell_2} \precsim \sigma \sqrt{\frac{s \log p}{n}}, \ \|\hat{M} - M\|_{\ell_1} \precsim \sigma s \sqrt{\frac{\log p}{n}}, \ \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \precsim \sigma \sqrt{\frac{s \log p}{n}}.$$

with high probability, where $C_1, C_2 > 0$ are some universal constants.

For ℓ_2 norm consistency of the estimation for M, we require $\lim_{n,p(n)\to\infty} \frac{s \log p}{n} = 0$. However, for valid inferential guarantee, the de-biased Dantzig selector type estimator $\tilde{M}$ satisfies asymptotic normality under the condition $\lim_{n,p(n)\to\infty} \frac{s \log p}{\sqrt{n}} = 0$ through Theorem 2. Under this condition, the confidence

interval given in (3.5) has asymptotic coverage probability of $(1-\alpha)$ and its expected length is at the parametric rate $\frac{1}{\sqrt{n}}$. Furthermore, the confidence intervals do not depend on the specific value of s. Results in Section 3.2 and 3.3 recover the best known result on confidence intervals as in [48, 44, 23]. Our result is a generic procedure that compensates for the bias introduced by the point estimation convex program. All these procedures are driven by local geometry.

3.5.2. *Low Rank Matrix Recovery.* We now consider the recovery of low-rank matrices under the trace regression model (1.3). The geometric theory leads to the optimal recovery results for nuclear norm minimization and penalized trace regression in the existing literature.

Assume the true parameter $M \in \mathbb{R}^{p\times q}$ has rank r. Let us examine the behavior of $\phi_{\mathcal{A}}(M,\mathcal{X})$, $\gamma_{\mathcal{A}}(M)$, and $\lambda_{\mathcal{A}}(\mathcal{X},\sigma,n)$. Detailed calculations given in Supplement [6] Section B show that in this case $\gamma_{\mathcal{A}}(M)w(\mathcal{A})$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$ are of the same order $\sqrt{r(p+q)}$. The asphericity ratio $\gamma_{\mathcal{A}}(M) \le 2\sqrt{2r}$ characterizes the low rank structure and the Gaussian width $w(\mathcal{X}\mathcal{A}) \asymp \sqrt{p+q}$. We have the following corollary for low rank matrix recovery.

COROLLARY 3. *Consider the trace regression model (1.3). Assume that $\mathcal{X} \in \mathbb{R}^{n\times pq}$ is the Gaussian ensemble design and the true parameter $M \in \mathbb{R}^{p\times q}$ is of rank r. Let $\hat{M}$ be the solution to the constrained nuclear norm minimization (2.12) with $\lambda = C_1\sigma\sqrt{\frac{p+q}{n}}$. If $n \ge C_2 r(p+q)$, then for some universal constants $C_1, C_2 > 0$, with high probability,*

$$\|\hat{M}-M\|_F \precsim \sigma\sqrt{\frac{r(p+q)}{n}},\ \|\hat{M}-M\|_* \precsim \sigma r\sqrt{\frac{p+q}{n}},\ \|\mathcal{X}(\hat{M}-M)\|_{\ell_2} \precsim \sigma\sqrt{\frac{r(p+q)}{n}}.$$

For point estimation consistency under the Frobenius norm loss, the condition is $\lim_{n,p(n),q(n)\to\infty} \frac{\sqrt{r(p+q)}}{\sqrt{n}} = 0$. For statistical inference, Theorem 2 requires $\lim_{n,p(n),q(n)\to\infty} \frac{r(p+q)}{\sqrt{n}} = 0$, which is essentially $n \gtrsim pq$ (sample size is larger than the dimension) for $r = 1$. This phenomenon happens when the Gaussian width complexity of the rank-1 matrices is large, i.e., the atom set is too rich. We remark that in practice, convex program (2.15) can still be used for constructing confidence intervals and performing hypothesis testing. However, it is harder to provide sharp upper bound theoretically for the approximation error η in (2.15), for any given r, p, q.

3.5.3. *Sign Vector Recovery.* We turn to the sign vector recovery model (1.4) where the parameter of interest $M \in \{+1,-1\}^p$ is a sign vector. The

convex hull of the atom set is then the ℓ_∞ norm ball. Applying the general theory to the constrained ℓ_∞ norm minimization (2.13) leads to the optimal rates of convergence for the sign vector recovery. The calculations given in Supplement [6] Section B show that the asphericity ratio $\gamma_{\mathcal{A}}(M) \leq 1$ and the Gaussian width $w(\mathcal{X}B_2^p) \asymp \sqrt{p}$. Geometric theory when applied to sign vector recovery shows the following Corollary.

COROLLARY 4. *Consider the model (1.4) where the true parameter $M \in \{+1, -1\}^p$ is a sign vector. Assume that $\mathcal{X} \in \mathbb{R}^{n\times p}$ is the Gaussian ensemble design. Let $\hat{M}$ be the solution to the convex program (2.13) with $\lambda = C_1\sigma\frac{p}{\sqrt{n}}$ and $\mu = C_1\sigma\sqrt{\frac{p}{n}}$. If $n \geq C_2 p$, then for some universal constant $C > 0$, with high probability,*

$$\|\hat{M} - M\|_{\ell_2}, \|\hat{M} - M\|_{\ell_\infty}, \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq C \cdot \sigma\sqrt{\frac{p}{n}}.$$

3.5.4. *Orthogonal Matrix Recovery.* We now treat orthogonal matrix recovery using the spectral norm minimization. Please see Example 4 in Section 2.1 for details. Consider the same model as in trace regression, but the parameter of interest $M \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. One can show that $w(B_2^p \cap T_{\mathcal{A}}(M))$ is of order $\sqrt{m^2}$ and $\gamma_{\mathcal{A}}(M) \leq 1$. Applying the geometric analysis to the constrained spectral norm minimization (2.13) yields the following.

COROLLARY 5. *Consider the orthogonal matrix recovery model (1.3). Assume that $\mathcal{X} \in \mathbb{R}^{n\times m^2}$ is the Gaussian ensemble matrix and the true parameter $M \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. Let $\hat{M}$ be the solution to the program (2.13) with $\lambda = C_1\sigma\sqrt{\frac{m^3}{n}}$ and $\mu = C_1\sigma\sqrt{\frac{m^2}{n}}$. If $n \geq C_2 m^2$, then, with high probability,*

$$\|\hat{M} - M\|_F, \|\hat{M} - M\|, \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq C \cdot \sigma\sqrt{\frac{m^2}{n}},$$

where $C > 0$ is some universal constant.

3.5.5. *Other examples.* Other examples that can be formalized under the framework of the linear inverse model include permutation matrix recovery [22], sparse plus low rank matrix recovery [11] and matrix completion [14]. The convex relaxation of permutation matrix is double stochastic matrix; the atomic norm corresponding to sparse plus low rank atom set is the infimal convolution of the ℓ_1 norm and nuclear norm; for matrix completion, the

design matrix can be viewed as a diagonal matrix with diagonal elements being independent Bernoulli random variables. See Section 5 for a discussion on further examples.

4. Local Geometric Theory: General Setting. We have developed in the last section a local geometric theory for the linear inverse model in the Gaussian setting. The Gaussian assumption on the design and noise enables us to carry out concrete and more specific calculations as seen in the examples given in Section 3.5, but the distributional assumption is not essential. In this section we extend this theory to the general setting.

4.1. *General Local Upper Bound.* We shall consider a fixed design matrix $\mathcal{X}$ (in the case of random design, results we will establish are conditional on the design) and condition on the event that the noise is controlled $\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \le \lambda_n$. We have seen in Lemma 2 of Section 3.1 how to choose λ_n to make this happen with overwhelming probability under Gaussian noise.

THEOREM 4 (Geometrizing Local Convergence). *Suppose we observe* $(\mathcal{X}, Y)$ *as in* (1.1). *Condition on the event that the noise vector* Z *satisfies, for some given choice of localization radius* λ_n, $\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \le \lambda_n$. *Let* $\hat{M}$ *be the solution to the convex program* (2.12) *with* λ_n *being the tuning parameter. Then the geometric quantities defined on the local tangent cone capture the local convergence rate for* $\hat{M}$*:*

$$\|\hat{M} - M\|_{\mathcal{A}} \le \gamma_{\mathcal{A}}(M)\|\hat{M} - M\|_{\ell_2}, \quad \text{and further}$$
$$\|\hat{M} - M\|_{\ell_2} \le \frac{1}{\phi_{\mathcal{A}}(M, \mathcal{X})}\|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \le \frac{2\gamma_{\mathcal{A}}(M)\lambda_n}{\phi_{\mathcal{A}}^2(M, \mathcal{X})}$$

with the local asphericity ratio $\gamma_{\mathcal{A}}(M)$ *defined in* (2.11) *and the local lower isometry constant* $\phi_{\mathcal{A}}(M, \mathcal{X})$ *defined in* (2.9).

Theorem 4 does not require distributional assumptions on the noise, nor does it impose conditions on the design matrix. Theorem 1 can be viewed as a special case where the local isometry constant $\phi_{\mathcal{A}}(M, \mathcal{X})$ and the local radius λ_n are calculated explicitly under the Gaussian assumption. Theorem 4 is proved in Section 6 in a general form, which analyzes convex programs (2.12) and (2.13) simultaneously.

4.2. *General Geometric Inference.* Geometric inference can also be extended to other fixed designs when Z is Gaussian. We can modify the convex

feasibility program (2.14) into the following stronger form

$$(\Omega, \eta_n) = \arg\min_{\Omega,\eta} \{\eta : \quad \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall 1 \le i \le p\}. \tag{4.1}$$

Then the following theorem holds (proof is analogous to Theorem 2).

THEOREM 5 (Geometric Inference). *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (1.1) *with* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. *Let* $\hat{M}$ *be the solution to the convex program* (2.12). *Denote* Ω *and* η_n *as the optimal solution to the convex program* (4.1), *and* $\tilde{M}$ *as the de-biased estimator. The following decomposition*

$$\tilde{M} - M = \Delta + \frac{\sigma}{\sqrt{n}} \Omega \mathcal{X}^* W \tag{4.2}$$

holds, where $W \sim N(0, I_n)$ *is the standard Gaussian vector and*

$$\Omega\mathcal{X}^*W \sim N(0, \Omega\mathcal{X}^*\mathcal{X}\Omega^*).$$

Here the bias part $\Delta \in \mathbb{R}^p$ *satisfies, with high probability,*

$$\|\Delta\|_{\ell_\infty} \le \frac{2 \cdot \gamma_{\mathcal{A}}^2(M)}{\phi_{\mathcal{A}}(M, \mathcal{X})} \cdot \lambda_n \eta_n,$$

provided we choose λ_n *as in Lemma 2.*

4.3. *General Local Minimax Lower Bound.* The lower bound given in the Gaussian case can also be extended to the general setting where the class of noise distributions contains the Gaussian distributions. We aim to geometrize the intrinsic difficulty of the estimation problem in a unified manner. We first present a general result for a convex cone T in the parameter space, which illustrates how the Sudakov estimate, volume ratio and the design matrix affect the minimax lower bound.

THEOREM 6. *Let* $T \in \mathbb{R}^p$ *be a compact convex cone. The minimax lower bound for the linear inverse model* (1.1), *if restricted to the cone* T, *is*

$$\inf_{\hat{M}} \sup_{M \in T} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M\|_{\ell_2}^2 \ge \frac{c_0 \sigma^2}{\psi^2} \cdot \left(\frac{e(B_2^p \cap T)}{\sqrt{n}} \vee \frac{v(B_2^p \cap T)}{\sqrt{n}} \right)^2$$

where $\hat{M}$ *is any measurable estimator,* $\psi = \sup_{v \in B_2^p \cap T} \|\mathcal{X}(v)\|_{\ell_2}$ *and* c_0 *is a universal constant. Here* $\mathbb{E}_{\cdot|\mathcal{X}}$ *is conditioned on the design matrix.* $e(\cdot)$ *and* $v(\cdot)$ *denote the Sudakov estimate* (2.6) *and volume ratio* (2.8). *Then*

$$\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{\ell_2}^2 \ge \frac{c_0 \sigma^2}{\psi_{\mathcal{A}}^2(M, \mathcal{X})} \cdot \left(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \vee \frac{v(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \right)^2.$$

Theorem 6 gives minimax lower bounds in terms of the Sudakov estimate and volume ratio. In the Gaussian setting, Lemma 3 shows that the local upper isometry constant satisfies $\psi_{\mathcal{A}}(M, \mathcal{X}) \leq 1+c$ with probability at least $1 - 2\exp(-\delta^2/2)$, as long as

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c}.$$

We remark that $\psi_{\mathcal{A}}(M, \mathcal{X})$ can be bounded under more general design matrix $\mathcal{X}$. However, under the Gaussian design (even correlated design), the minimum sample size n to ensure that $\psi_{\mathcal{A}}(M, \mathcal{X})$ is upper bounded, is directly determined by Gaussian width of the tangent cone.

The geometric complexity of the lower bound provided by Theorem 6 matches $w(B_2^p \cap T_{\mathcal{A}}(M))$ if Sudakov minoration of Lemma 1 can be reversed on the tangent cone, in the sense that $w(B_2^p \cap T_{\mathcal{A}}(M)) \leq C \cdot e(B_2^p \cap T_{\mathcal{A}}(M))$. Further, recalling Urysohn's inequality we have $v(B_2^p \cap T_{\mathcal{A}}(M)) \leq w(B_2^p \cap T_{\mathcal{A}}(M))$. Hence, if the reverse Urysohn's inequality $w(B_2^p \cap T_{\mathcal{A}}(M)) \leq C \cdot v(B_2^p \cap T_{\mathcal{A}}(M))$ holds for the local tangent cone, the obtained rate is, again, of the order $w(B_2^p \cap T_{\mathcal{A}}(M))$.

5. Discussion. This paper presents a unified geometric characterization of the local estimation rates of convergence as well as statistical inference for high-dimensional linear inverse problems. Exploring other interesting applications that can be subsumed under the general framework is an interesting future research direction.

For statistical inference, both independent Gaussian design and correlated Gaussian design with known covariance Σ are considered. The case of unknown Σ is an interesting problem for future work.

The lower bound constructed in the current paper can be contrasted with the lower bounds in [47, 10]. Both the above two papers consider specifically the minimax lower bound for high-dimensional linear regression. We focus on a more generic perspective – lower bounds in Theorem 6 hold in general for arbitrary star-shaped body T, which includes $\ell_p, 0 \leq p \leq \infty$, balls and cones as special cases.

6. Proofs. The proofs of the main results are divided into several parts. For the upper bound of point estimation, we will first prove Theorem 4 and then two lemmas, Lemma 3 and Lemma 2 (these two Lemmas are included in Supplement [6] Section A). Theorem 1 is then easy to prove. As for the statistical inference, Theorem 2 is proved based on Theorem 1. For the lower bound of point estimation, Theorem 3 is a direct result combining Lemma

3 and Theorem 6, which is proved in Supplement [6] Section A. Proofs of Corollaries are deferred to Supplement [6] Section B.

PROOF OF THEOREM 4. We will prove a stronger version of the Theorem, analyzing both (2.12) and (2.13). The proof is clean and in a general fashion, following directly from the assumptions of the theorem and the definitions:

$$\begin{aligned} \|\mathcal{X}^*(Y-\mathcal{X}M)\|_{\mathcal{A}}^* \le \lambda, \|\mathcal{X}^*(Y-\mathcal{X}M)\|_{\ell_2} \le \mu \quad & \text{Assumption of the Theorem} \\ \|\mathcal{X}^*(Y-\mathcal{X}\hat{M})\|_{\mathcal{A}}^* \le \lambda, \|\mathcal{X}^*(Y-\mathcal{X}\hat{M})\|_{\ell_2} \le \mu \quad & \text{Constraint in program} \\ \|\hat{M}\|_{\mathcal{A}} \le \|M\|_{\mathcal{A}} \quad & \text{Definition of minimizer} \end{aligned}$$

Thus we have

(6.1)
$$\|\mathcal{X}^*\mathcal{X}(\hat{M}-M)\|_{\mathcal{A}}^* \le 2\lambda, \|\mathcal{X}^*\mathcal{X}(\hat{M}-M)\|_{\ell_2} \le 2\mu \quad \text{and} \quad \hat{M}-M \in T_{\mathcal{A}}(M).$$

The first equation is due to triangle inequality and second one due to Tangent cone definition. Define $H = \hat{M} - M \in T_{\mathcal{A}}(M)$. According to the "Cauchy-Schwarz" (2.3) relation between atomic norm and its dual,

$$\|\mathcal{X}(H)\|_{\ell_2}^2 = \langle \mathcal{X}(H), \mathcal{X}(H)\rangle = \langle \mathcal{X}^*\mathcal{X}(H), H\rangle \le \|\mathcal{X}^*\mathcal{X}(H)\|_{\mathcal{A}}^* \|H\|_{\mathcal{A}}.$$

Using the earlier result $\|\mathcal{X}^*\mathcal{X}(H)\|_{\mathcal{A}}^* \le 2\lambda$, as well as the following two equations for any $H \in T_{\mathcal{A}}(M)$

$$\begin{aligned} \phi_{\mathcal{A}}(M,\mathcal{X})\|H\|_{\ell_2} \le \|\mathcal{X}(H)\|_{\ell_2} \quad & \text{local isometry constant} \\ \|H\|_{\mathcal{A}} \le \gamma_{\mathcal{A}}(M)\|H\|_{\ell_2} \quad & \text{local asphericity ratio} \end{aligned}$$

we get the following self-bounding relationship

$$\begin{aligned} \phi_{\mathcal{A}}^2(M,\mathcal{X})\|H\|_{\ell_2}^2 \le \|\mathcal{X}(H)\|_{\ell_2}^2 \le 2\lambda\|H\|_{\mathcal{A}} \le 2\lambda\gamma_{\mathcal{A}}(M)\|H\|_{\ell_2}, \\ \phi_{\mathcal{A}}^2(M,\mathcal{X})\|H\|_{\ell_2}^2 \le \|\mathcal{X}(H)\|_{\ell_2}^2 \le 2\mu\|H\|_{\ell_2}. \end{aligned}$$

Thus $\|H\|_{\ell_2} \le \frac{2}{\phi_{\mathcal{A}}^2(M,\mathcal{X})} \min\{\gamma_{\mathcal{A}}(M)\lambda, \mu\}$. The proof is then completed by simple algebra. Note here under the Gaussian setting, we can plug in $\lambda \asymp w(\mathcal{X}\mathcal{A})/\sqrt{n}$ and $\mu \asymp w(\mathcal{X}B_2^p)/\sqrt{n}$ using Lemma 2. □

PROOF OF THEOREM 1. Theorem 1 is a special case of Theorem 4 under Gaussian setting, combining with Lemma 3 and Lemma 2. All we need to show is a good control of λ_n and $\phi_{\mathcal{A}}(M,\mathcal{X})$ with probability at least $1-3\exp(-\delta^2/2)$ under Gaussian ensemble and Gaussian noise. We bound

λ_n with probability at least $1-\exp(-\delta^2/2)$ via Lemma 2. For $\phi_{\mathcal{A}}(M,\mathcal{X})$, we can lower bound by $1-c$ with probability at least $1-2\exp(-\delta^2/2)$. Let's define good event to be when

$$\lambda_n \le \frac{\sigma}{\sqrt{n}}\left\{\mathbb{E}_g\left[\sup_{v\in\mathcal{A}}\langle g,\mathcal{X}v\rangle\right]+\delta\cdot\sup_{v\in\mathcal{A}}\|\mathcal{X}v\|_{\ell_2}\right\}$$

and $1-c\le\phi_{\mathcal{A}}(M,\mathcal{X})\le\psi_{\mathcal{A}}(M,\mathcal{X})\le 1+c$ both hold. It is easy to see this good event holds with probability $1-3\exp(-\delta^2/2)$. Thus all we need to prove is $\max_{z\in\mathcal{A}}\|\mathcal{X}z\|\le 1+c$ under the good event.

According to Lemma 3, equation (A.3)'s calculation, $\max_{z\in\mathcal{A}}\|\mathcal{X}z\|/\|z\|\le 1+c$ is satisfied under the condition $n\ge\frac{1}{c^2}[w(B_2^p\cap\mathcal{A})+\delta]^2$. As we know for any M, the unit atomic norm ball $\mathsf{conv}(\mathcal{A})$ is contained in $2B_2^p$ and $T_{\mathcal{A}}(M)$, which means $B_2^p\cap\mathcal{A}\subset 2B_2^p\cap T_{\mathcal{A}}(M)$, thus $w(B_2^p\cap\mathcal{A})\le 2w(B_2^p\cap T_{\mathcal{A}}(M))$ (monotonic property of Gaussian width). So we have for any M, if $n\ge\frac{4}{c^2}[w(B_2^p\cap T_{\mathcal{A}}(M))+\delta]^2\vee\frac{1}{c}$. we have the following two bounds with probability at least $1-2\exp(-\delta^2/2)$

$$\max_{z\in\mathcal{A}}\|\mathcal{X}z\|\le 1+c$$
$$1-c\le\phi_{\mathcal{A}}(M,\mathcal{X})\le\psi_{\mathcal{A}}(M,\mathcal{X})\le 1+c. \tag{6.2}$$

Now plugging (6.2) into the expression of Lemma 2, together with Lemma 3, Theorem 4 reduces to Theorem 1. □

Proof of Theorem 2. We first prove that, with high probability, the convex program (2.14) is indeed feasible with $\Omega=I_n$. Equivalently we establish that, with high probability, for any $1\le i\le p$, $\|\mathcal{X}^*\mathcal{X}e_i-e_i\|_{\mathcal{A}}^*\le\eta$ for some proper choice of η. Here $\mathcal{X}\in\mathbb{R}^{n\times p}$, and the entries $\mathcal{X}_{ij}\overset{iid}{\sim}N(0,1/n)$. Denote $g=\sqrt{n}\mathcal{X}_{\cdot i}$ as a scaling version of the i-th column of $\mathcal{X}$, $g\sim N(0,I_n)$ and $g'\sim N(0,I_n)$ being an independent copy. Below $O_p(\cdot)$ denotes the

asymptotic order in probability. We have, for all $1 \le i \le p$ uniformly,

$$\begin{aligned}
\|\mathcal{X}^*\mathcal{X}e_i - e_i\|_{\mathcal{A}}^* &= \sup_{v\in\mathcal{A}}\langle \mathcal{X}^*\mathcal{X}e_i - e_i, v\rangle = \sup_{v\in\mathcal{A}}\langle \mathcal{X}^*g - e_i, v\rangle/\sqrt{n} \\
&\le \sup_{v\in\mathcal{A}}\langle \mathcal{X}_{(-i)}^* g, v\rangle/\sqrt{n} + \sup_{v\in\mathcal{A}}\left(\frac{1}{n}\sum_{j=1}^n g_j^2 - 1\right) v_i \\
&\overset{w.h.p}{\precsim} \frac{w(\mathcal{X}_{(-i)}\mathcal{A})}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \quad \text{invoking Lemma 2} \\
&\le \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \frac{\mathbb{E}_{g'}\sup_{v\in\mathcal{A}}\sum_{k=1}^n g_k'\mathcal{X}_{ki}(-v_i)}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \\
&\le \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \frac{\sqrt{\mathbb{E}_{g'}(\sum_{k=1}^n g_k'\mathcal{X}_{ki})^2 \cdot \sup_{v\in\mathcal{A}} v_i^2}}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \\
&\le \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \sqrt{\frac{1 + O_p(\sqrt{\log p/n})}{n}} + O_p(\sqrt{\log p/n}) \qquad (6.3)
\end{aligned}$$

where $\mathcal{X}_{(-i)}$ is the linear operator setting i-th column to be all zeros. We applied Lemma 2 in establishing the above bounds.

For the de-biased estimate $\tilde{M}$, we have $\tilde{M} = \hat{M} + \Omega\mathcal{X}^*(Y - \mathcal{X}(\hat{M}))$ and $\tilde{M} - M = (\Omega\mathcal{X}^*\mathcal{X} - I_p)(M - \hat{M}) + \Omega\mathcal{X}^*Z := \Delta + \frac{\sigma}{\sqrt{n}}\Omega\mathcal{X}^*W$. Then for any $1 \le i \le p$, from the Cauchy-Schwartz relationship (2.3),

(6.4)

$$|\Delta_i| = |\langle \mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i, M - \hat{M}\rangle| \le \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \|M - \hat{M}\|_{\mathcal{A}} \le \sigma\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{n}.$$

The last line invokes the consistency result in Theorem 1, $\|\hat{M} - M\|_{\mathcal{A}} \precsim \sigma\frac{\gamma_{\mathcal{A}}^2(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}}$. Thus we have $\|\Delta\|_{\ell_\infty} \precsim \sigma\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{n}$. For any linear contrast $\|v\|_{\ell_1} \le \rho$, we have $\frac{\sqrt{n}}{\sigma}v^*(\tilde{M} - M) = v^*\Omega\mathcal{X}^*W + \frac{\sqrt{n}}{\sigma}v^*\Delta$,

$$\limsup_{n,p(n)\to\infty}\frac{\sqrt{n}}{\sigma}v^*\Delta \le \limsup_{n,p(n)\to\infty}\frac{\sqrt{n}}{\sigma}\|v\|_{\ell_1}\|\Delta\|_{\ell_\infty} \le \rho\cdot\limsup_{n,p(n)\to\infty}\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

and $v^*\Omega\mathcal{X}^*W \sim N(0, v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)$. □

Proof of Theorem 3. Theorem 3 is a special case of Theorem 6, combining with Lemma 3 (both in Supplement [6] Section A). Plug in the general convex cone T by local tangent cone $T_{\mathcal{A}}(M)$, then upper bound $\psi_{\mathcal{A}}(M, \mathcal{X}) \le 1 + c$ with high probability via Lemma 3. □

SUPPLEMENTARY MATERIAL

Supplement to: "Geometric Inference for General High-Dimensional Linear Inverse Problems"

(doi: COMPLETED BY THE TYPESETTER; .pdf). Due to space constraints, we have relegated remaining proofs to the Supplement [6], where details of proof for Lemma 2-4, Theorem 6 and Corollary 1-5 are included.

References.

[1] Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2013). Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*.

[2] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

[3] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

[5] Cai, T., Liu, W., and Luo, X. (2011). A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

[6] Cai, T. T., Liang, T., and Rakhlin, A. (2014). Supplement to "geometric inference for general high-dimensional linear inverse problems". Technical report.

[7] Cai, T. T. and Low, M. G. (2004a). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32:1805–1840.

[8] Cai, T. T. and Low, M. G. (2004b). Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of Statistic*, 32(2):552–576.

[9] Cai, T. T. and Zhou, W. (2013). Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*.

[10] Candes, E. J. and Davenport, M. A. (2013). How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323.

[11] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

[12] Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

[13] Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359.

[14] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

[15] Candès, E. J. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351.

[16] Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.

[17] Chatterjee, S. (2012). Matrix estimation by universal singular value thresholding. *arXiv preprint arXiv:1212.1247*.

[18] Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270.

[19] Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

[20] Gordon, Y. (1988). *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer.

[21] Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes problems*, volume 3. Oxford University Press Oxford.

[22] Jagabathula, S. and Shah, D. (2011). Inferring rankings using constrained sensing. *Information Theory, IEEE Transactions on*, 57(11):7288–7306.

[23] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

[24] Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *The Annals of Statistics*, pages 251–280.

[25] Khuri, S., Bäck, T., and Heitkötter, J. (1994). The zero/one multiple knapsack problem and genetic algorithms. In *Proceedings of the 1994 ACM symposium on Applied computing*, pages 188–193. ACM.

[26] Koltchinskii, V. (2011). Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6):2936–2973.

[27] Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

[28] Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*.

[29] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer.

[30] Ma, Z. and Wu, Y. (2013). Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *arXiv preprint arXiv:1306.3609*.

[31] Mangasarian, O. L. and Recht, B. (2011). Probability of unique integer solution to a system of linear equations. *European Journal of Operational Research*, 214(1):27–30.

[32] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

[33] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518.

[34] Oymak, S., Thrampoulidis, C., and Hassibi, B. (2013). Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*.

[35] Pisier, G. (1999). *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press.

[36] Prokopyev, O. A., Huang, H.-X., and Pardalos, P. M. (2005). On complexity of unconstrained hyperbolic 0–1 programming problems. *Operations Research Letters*, 33(3):312–318.

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

[38] Rohde, A., Tsybakov, A. B., et al. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

[39] Talagrand, M. (1996). Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103.

[40] Ten Berge, J. M. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276.

[41] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[42] Tikhonov, A. and Arsenin, V. Y. (1977). *Methods for solving ill-posed problems*. John Wiley and Sons, Inc.

[43] Valiant, L. G. and Vazirani, V. V. (1986). Np is as easy as detecting unique solutions. *Theoretical Computer Science*, 47:85–93.

[44] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

[45] Vershynin, R. (2011). Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal. umich. edu/˜ romanv/papers/GFA-book/GFA-book. pdf*.

[46] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599.

[47] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540.

[48] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
USA
E-mail: tcai@wharton.upenn.edu
E-mail: tengyuan@wharton.upenn.edu
E-mail: rakhlin@wharton.upenn.edu