On the well-spread property and its relation to linear regression

Hongjie Chen HONGJIE.CHEN@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Tommaso d'Orsi TOMMASO.DORSI@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the robust linear regression model $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, where an adversary oblivious to the design $X \in \mathbb{R}^{n\times d}$ may choose $\boldsymbol{\eta}$ to corrupt all but a (possibly vanishing) fraction of the observations $\boldsymbol{y}$ in an arbitrary way. Recent work d'Orsi et al. (2021a,b) has introduced efficient algorithms for consistent recovery of the parameter vector. These algorithms crucially rely on the design matrix being well-spread (a matrix is well-spread if its column span is far from any sparse vector).

In this paper, we show that there exists a family of design matrices lacking well-spreadness such that consistent recovery of the parameter vector in the above robust linear regression model is information-theoretically impossible.

We further investigate the average-case time complexity of certifying well-spreadness of random matrices. We show that it is possible to efficiently certify whether a given n-by-d Gaussian matrix is well-spread if the number of observations is quadratic in the ambient dimension. We complement this result by showing rigorous evidence —in the form of a lower bound against low-degree polynomials— of the computational hardness of this same certification problem when the number of observations is $o(d^2)$.

Keywords: spread subspaces, robust optimization, oblivious outliers, regression, information-theoretic bounds, average-case analysis

1. Introduction

For a subspace $V \subseteq \mathbb{R}^n$, the well-spreadness property describes how close sparse vectors are to it.

Definition 1 (Well-spreadness) *A subspace $V \subseteq \mathbb{R}^n$ is m-spread if for any $v \in V$ and any $S \subseteq [n]$ of size $|S| \geqslant n - m$, we have*

$$\|v_S\|_2 \geqslant \Omega(1) \cdot \|v\|_2 ,$$

where v_s denotes the projection of v onto the coordinates in S. We say that a matrix is m-spread if its column span is.

Due to its connection to distortion Guruswami et al. (2010), Euclidean section properties Baraniuk et al. (2008) and restricted isometry properties (RIP) Allen-Zhu et al. (2016); Guruswami et al. (2021), well-spread subspaces have been studied in the context of error-correction over the reals Candes and Tao (2005); Guruswami et al. (2008), compressed sensing matrices for low compression factors Kashin and Temlyakov (2007); Donoho (2006) convex geometry Gluskin (1984);

Kashin and Temlyakov (2007) and metric embeddings Indyk (2007). Recently, an unforeseen connection between well-spreadness and oblivious adversarial regression models has emerged d'Orsi et al. (2021a,b).

While relations between properties of the design matrix and algorithmic guarantees are not new —restricted eigenvalue condition, restricted isometry property (RIP) and distortion are all known to be sufficient to design efficient algorithms for recovering the encoded sparse vector (see Kashin and Temlyakov (2007); Zhang et al. (2014))— the connection between well-spreadness and oblivious regression appears intriguing as: (i) there is currently no significant evidence of the necessity of this property for recovery, and (ii) there is no indication of a gap between exponential time and polynomial time algorithms depending on the well-spreadness of the design. Investigating this relation is the main focus of this paper.

Oblivious regression Oblivious adversarial models offer a convenient framework to find the weakest assumptions under which one can efficiently recover structured signal from noisy data with *vanishing error*.[1] Once the observations are sampled, an adversary is allowed to add arbitrary noise *without* accessing the data and with the additional constraint that for an α fraction of the observations (possibly vanishing small) the noise must have small magnitude. In the context of regression this idea can be formalized into the following problem.

Problem 2 (Oblivious linear regression) *Given observations*[2] $(X_1, \boldsymbol{y}_1), \dots, (X_n, \boldsymbol{y}_n)$ *following the linear model* $\boldsymbol{y}_i = \langle X_i, \beta^* \rangle + \boldsymbol{\eta}_i$*, where* $X_i \in \mathbb{R}^d$*,* $\beta^* \in \mathbb{R}^d$*, and* $\boldsymbol{\eta}_i$ *is a symmetrically distributed random variable with* $\min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant 1\} = \alpha$*, the goal is to to find an estimator* $\hat{\beta}$ *for* β^* *achieving small squared parameter error* $\|\hat{\beta} - \beta^*\|_2^2$.[3]

We may conveniently think of the (possibly vanishingly small) α fraction of entries of $\boldsymbol{y}$ with small noise as the uncorrupted observations. Moreover, as moments are not required to exist, this noise model captures heavy-tailed distributions.

A flurry of works Tsakonas et al. (2014); Bhatia et al. (2017); Suggala et al. (2019); Pesme and Flammarion (2020); d'Orsi et al. (2021a,b) has led to the design of efficient and consistent[4] algorithms that achieve provably optimal error guarantees and sample complexity for oblivious regression. The guarantees of these algorithms are rather surprising. For classical regression with Gaussian noise $\mathcal{N}\left(0, \sigma^2\right)$, it is known that the optimal error convergence is $O(\sigma^2 \cdot d/n)$ Wainwright (2019). For oblivious regression, efficient algorithms obtain squared parameter error bounded by $O(d/(\alpha^2 \cdot n))$ and thus are consistent for $n \geqslant \omega\left(d/\alpha^2\right)$. [5] As Gaussian distributions $\mathcal{N}\left(0, \sigma^2\right)$ can be modeled as noise in Problem 2 with $\alpha = O(1/\sigma)$, these error convergence rates are the same up to constant factors. In other words, even though Problem 2 allows for a large variety of complicated

1. *Adaptive* adversarial models, where the adversary has access to the data, are not suitable to study these questions as part of the signal may be removed and hence impossible to reconstruct.
2. We use bold face to denote random variables.
3. We remark that we may analogously ask for small squared prediction error $\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2$, at the coarseness of this discussion the two may be considered equivalent. We also remark that for small values of α, symmetry of the noise is necessary d'Orsi et al. (2021b).
4. An estimator is said to be consistent if its error tends to zero as the number of observations grows.
5. More generally, we may assume in Problem 2 that $\min_{i\in[n]} \mathbb{P}\{|\boldsymbol{\eta}_i| \leqslant \tau\} = \alpha$ by introducing another parameter $\tau > 0$. In this case, the error bound becomes $O(\tau^2 d/(\alpha^2 \cdot n))$ while the analysis of the error bound is essentially the same. In this paper, we set $\tau = 1$ for simplicity.

noise distributions, it is possible to achieve error guarantees similar to those one would be able to achieve under the special case of Gaussian noise.

It turns out that the catch is in the design matrix $X \in \mathbb{R}^{n\times d}$. Algorithms for oblivious regression require the column span $\mathrm{cspan}(X)$ of X to be well-spread. If $\mathrm{cspan}(X)$ is $\Omega\left(d/\alpha^2\right)$-spread, then the above guarantees can be achieved efficiently. On the other hand, *no algorithm* is known to obtain non-trivial error guarantees as soon as the design matrix is only $o(d/\alpha^2)$-spread, even in exponential time. This picture raises an important question concerning the relation between oblivious regression and well-spreadness:

> *Is the well-spreadness requirement a fundamental limitation of current algorithms or is there a sharp phase transition in the landscape of the problem? Is this phase transition a computational or statistical phenomenon?*

In this paper we provide, to a large extent, answers to these and related questions.

1.1. Results

Information-theoretic lower bounds regarding well-spreadness Our first result is the non-existence of algorithms with non-trivial error guarantees for oblivious regression with design matrices lacking well-spreadness.

Theorem 3 *Let $\alpha = \alpha(n) \in (0,1)$. For arbitrary $\gamma = \gamma(n) > 0$, there exist:*

1. *a matrix $X \in \mathbb{R}^{n\times d}$ with $\max\left\{\Omega\left(\frac{\log d}{\alpha^2}\right), \Omega\left(\frac{d}{\alpha}\right)\right\}$-spreadness and $X^\top X = n \cdot \mathrm{Id}$,*
2. *a distribution $\mathcal{D}_\beta$ over d-dimensional vectors, and*
3. *a distribution $\mathcal{D}_\eta$ —independent of $\mathcal{D}_\beta$— over n-dimensional vectors with independent, symmetrically distributed entries satisfying $\min_{i\in[n]} \mathbb{P}_{\boldsymbol{\eta}\sim\mathcal{D}_\eta}\left(|\boldsymbol{\eta}_i| \leqslant 1\right) = \alpha$,*

such that for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, given as input X and $\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\eta}$ with $\boldsymbol{\beta}^* \sim \mathcal{D}_\beta$ and $\boldsymbol{\eta} \sim \mathcal{D}_\eta$ sampled independently, one has

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma\,.$$

A more precise version of Theorem 3 is given by Theorem 7. It states that there exists a natural distribution $\mathcal{D}_X$ over $\mathbb{R}^{n\times d}$ such that with high probability (i) $\boldsymbol{X} \sim \mathcal{D}_X$ is $\max\{\Omega(\frac{\log d}{\alpha^2}), \Omega(\frac{d}{\alpha})\}$-spread, and (ii) given a matrix X sampled from $\mathcal{D}_X$ with $\max\{\Omega(\frac{\log d}{\alpha^2}), \Omega(\frac{d}{\alpha})\}$-spreadness as input for Problem 2, no estimator can obtain bounded error guarantees, for *any* number of observations. We remark that, in our construction we utilize the condition $X^\mathsf{T}X = n \cdot \mathrm{Id}$ only to make the squared parameter error $\|\hat{\beta}-\beta^*\|_2^2$ and squared prediction error $\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2$ equivalent. Thus, Theorem 3 immediately yields a lower bound for the prediction error as well. This shows there is a fundamental difference between the statistical hardness of oblivious regression and its classical counterpart (with sub-gaussian noise): a statistical price to pay for robustness against oblivious adversaries.

It is fascinating to notice that, on one hand, current algorithms d'Orsi et al. (2021a) obtain non-trivial error guarantees only for $\Omega(d/\alpha^2)$-spread design matrices; on the other hand, although those hard oblivious regression instances constructed in Theorem 3 defy any non-trivial error guarantees,

they do not rule out the existence of consistent estimators for other families of $o(d/\alpha^2)$-spread design matrices. Thus, there remains a family of design matrices for which it is not known whether consistent oblivious regression can be achieved, and if so whether it can be done efficiently. This remains a pressing open question.

Certifying well-spreadness As the well-spreadness of design matrices can guarantee efficient recovery in oblivious regression, it is natural to ask whether one can efficiently *certify* the well-spreadness of a matrix. Similar questions have indeed been investigated for RIP, due to its application in compressed sensing Bandeira et al. (2013); Tillmann and Pfetsch (2014); Natarajan and Wu (2014); Koiran and Zouzias (2014); Wang et al. (2016); Weed (2018); Ding et al. (2021). Unsurprisingly, certifying well-spreadness turns out to be NP-hard in the worst case (see Theorem 29 for a proof). On the other hand, in the context of average-case analysis, there exists a regime where efficient algorithms can certify well-spreadness.

Theorem 4 (Algorithms for certifying well-spreadness) *Fix arbitrary constants $\delta \in (0,1)$ and $C > 0$. Let $\boldsymbol{X} \sim \mathcal{N}(0,1)^{n\times d}$ with $n \geqslant Cd^2$. There exist a polynomial-time algorithm and a constant $C' = C'(C,\delta) \in (0,1)$ such that*

1. *$\boldsymbol{X}$ is $(C'n,\delta)$-spread with probability $1-o(1)$;*
2. *if $\boldsymbol{X}$ is not $(C'n,\delta)$-spread, the algorithm outputs NO;*
3. *if $\boldsymbol{X}$ is $(C'n,\delta)$-spread, the algorithm outputs YES with probability $1-o(1)$.*

We want to emphasize that, under the assumptions of Theorem 4, (i) an n-by-d Gaussian random matrix is $\Omega(n)$-spread with high probability; (ii) if the sampled matrix is indeed $\Omega(n)$-spread, the algorithm can efficiently certify this fact with high probability; (iii) the algorithm *never* outputs false positives and thus guarantees the serviceability of the sampled matrix as a design matrix for oblivious regression. Theorem 4 also directly implies that, in the regime $n \geqslant \omega\left(d/\alpha^2\right)$ where there exist efficient algorithms for consistent oblivious regression, we may certify the well-spreadness of random design matrices, given $\Omega\left(d^2\right)$ samples.

It is tempting to ask whether a similar verification algorithm can be designed with fewer observations $n \leqslant o(d^2)$. However, we provide evidence of the computational hardness of this problem in the form of a lower bound against low degree polynomials. This computational model captures state-of-the-art algorithms for many average-case problems such as sparse PCA, tensor PCA or community detection (e.g. see Hopkins and Steurer (2017); Hopkins (2018); Kunisky et al. (2019); Ding et al. (2019); d'Orsi et al. (2020); Choo and d'Orsi (2021)).

Theorem 5 (Lower bounds against low-degree polynomials) *Let $t \leqslant O(\log n)^C$ for some arbitrary constant $C > 1$. Let $\alpha = \alpha(n) \in (0,1)$. For[6] any $\alpha \gg d^{-1/2}$ and $d/\alpha^2 \ll n \ll (d/\alpha)^{4/3}$, there exist two distributions over n-by-d matrices, $\mathcal{D}_0$ and $\mathcal{D}_1$, such that*

1. *$\mathcal{D}_0$ is the standard Gaussian distribution;*
2. *$\boldsymbol{X} \sim \mathcal{D}_0$ is $\Omega\left(d/\alpha^2\right)$-spread with probability $1-o(1)$;*

6. We use the notation $a \ll b$ for inequalities of the form $a \leqslant O(b/\operatorname{polylog}(b))$. The assumption $\alpha \gg d^{-1/2}$ derives from $d/\alpha^2 \ll (d/\alpha)^{4/3}$.

3. $\boldsymbol{X} \sim \mathcal{D}_1$ *is not* $\Omega\left(d/\alpha^2\right)$*-spread with probability* $1 - o(1)$*;*

4. *the two distributions are indistinguishable with respect to all polynomials* $p : \mathbb{R}^{n\times d} \to \mathbb{R}$ *of degree at most* t *in the sense that:*

$$\frac{\mathbb{E}_{\mathcal{D}_0} p(\boldsymbol{X}) - \mathbb{E}_{\mathcal{D}_1} p(\boldsymbol{X})}{\sqrt{\mathbb{V}_{\mathcal{D}_0} p(\boldsymbol{X})}} \leqslant O(1)\,.$$

In other words, Theorem 5 shows that low-degree polynomials cannot be used to distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$ as typical values of such polynomials look the same (up to a small difference) under both distributions.[7] An immediate consequence of this result is that there exists a family of $\Omega\left(d/\alpha^2\right)$-spread matrices $X \in \mathbb{R}^{n\times d}$ with $d/\alpha^2 \ll n \ll (d/\alpha)^{4/3}$, for which consistent oblivious regression is possible, but verifying whether the design matrix satisfy the required well-spread condition is hard.

We remark that there is no gap between the algorithmic result in Theorem 4 and the lower bound in Theorem 5, since Theorem 5 is a corollary of Theorem 39 wihch provides evidence of computational hardness for the entire regime $n \ll d^2$.

1.2. Organization

The rest of the paper is organized as follows. We present the high level ideas behind our results in Section 2. We prove Theorem 3 in Section 3. We obtain Theorem 4 and Theorem 5 in Appendix D. We show NP-hardness of well-spreadness certification in Appendix C. Finally, necessary background notions can be found in Appendix A.

2. Techniques

We present here the main ideas behind our results.

2.1. Statistical lower bounds for regression

Recall the linear model in Problem 2,

$$\boldsymbol{y} = X\beta^* + \boldsymbol{\eta},$$

where we observe (a realization of) the random vector $\boldsymbol{y}$, the matrix $X \in \mathbb{R}^{n\times d}$ is a known fixed design, the vector $\beta^* \in \mathbb{R}^d$ is the unknown parameter of interest, and the noise vector $\boldsymbol{\eta}$ has independent, symmetrically distributed coordinates with $\min_{i\in[n]} \mathbb{P}\left\{|\boldsymbol{\eta}_i| \leqslant 1\right\} = \alpha$. We will restrict our discussion to matrices satisfying $X^\mathsf{T} X = n \cdot \mathrm{Id}$, so that —up to scaling— there is no difference between prediction and parameter error.

To obtain an information-theoretic lower bound, we cast the problem as a distinguishing problem among ℓ hypotheses of the form:

$$H_i : \quad \boldsymbol{y} = X\beta_i + \boldsymbol{\eta}\,, \tag{2.1}$$

where we ought to make the vectors $\beta_1, \ldots, \beta_\ell \in \mathbb{R}^d$ as far as possible from each other. It is remarkably easy to see that a small degree of spreadness is necessary to obtain any error guarantee

7. See Appendix A.4 for a more in-depth discussion concerning the low-degree likelihood ratio.

in oblivious regression d'Orsi et al. (2021b). Let $X \in \mathbb{R}^{n\times d}$ be $o(1/\alpha)$-spread, then there exists a vector $\beta \in \mathbb{R}^d$ and a set S of cardinality $n - O(1/\alpha)$ such that $\|X_S\beta\|_2 \leqslant o(1) \cdot \|X\beta\|_2$.[8] The problem with such a design matrix is that with probability $\Omega(1)$ all nonzero entries in $(X - X_S)\beta$ will be corrupted by (possibly unbounded) noise. As a result no estimator can provide guarantees of the form $\mathbb{E}\|\hat{\beta}(\boldsymbol{y}) - \beta^*\|_2^2 \leqslant \gamma$ for any $\gamma > 0$. In other words, approximate recovery of the parameter vector is impossible.

Going beyond this $o(1/\alpha)$ barrier, however, turns out to be non-trivial. One issue with the above reasoning is that *even if we knew* the uncorrupted entries, we would not be able to recover the hidden vector (with any bounded error), as no such entry contains information over β^*. In contrast, for any X that is m-spread, with $m \geqslant \Omega(1/\alpha)$, if we knew the uncorrupted entries, after filtering out the corrupted ones, the classical least squares estimator would yield error guarantees

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \beta^*\right\|_2^2 \leqslant O\left(\frac{d}{m\alpha}\right).$$

That is, knowing the uncorrupted entries one could achieve constant error for $\Omega\left(d/\alpha\right)$-spread design matrices. Notice Theorem 3 implies that, there exist oblivious regression instances where no estimator can achieve these guarantees.

We overcome this barrier with a construction consisting of two main ingredients:

1. An m-spread matrix X and a set of vectors β in $\mathbb{R}^d$ such that

$$\|X_S\beta\|_2 \leqslant o(1) \cdot \|X\beta\|_2 \tag{2.2}$$

   for some $S \subseteq [n]$ with $|S| \geqslant n - m$, and the subspace spanned by these vectors is high dimensional. That is, a matrix whose column span contains many nearly orthogonal sparse vectors.

2. A distribution D_η over $\mathbb{R}$, for the entries of $\boldsymbol{\eta}$, satisfying the constraints in Problem 2, and with the additional properties:

   - *Low shift-sensitivity:* the distribution looks approximately the same after an additive shift in the following sense. If $D_\eta(k)$ is the distribution shifted by k, then the Kullback-Leibler divergence $D_{\mathrm{KL}}\left(D_\eta \| D_\eta(k)\right)$ is small.
   - *Insensitivity to scaling:* the Kullback-Leibler divergence does not change significantly upon scaling in the sense that $D_{\mathrm{KL}}\left(D_\eta \| D_\eta(k)\right) \approx D_{\mathrm{KL}}\left(\rho \cdot D_\eta \| \rho \cdot D_\eta(k)\right)$, for any $\rho > 0$.

Sparsity of the noiseless observation vectors $X\beta_i$ as in Eq. (2.2), combined with low shift-sensitivity of the noise distribution will allow us to make the different hypotheses indistinguishable. Then insensitivity to scaling will make the prediction error arbitrarily large.

Noise distributions with low sensitivity It turns out that constructing a distribution with low sensitivity is straightforward. We consider the symmetric geometric distribution $\mathcal{G}\left(c, \lambda\right)$ with probability mass function

$$p(k) = \begin{cases} \alpha, & k = c \\ \frac{1-\alpha}{2} \cdot \lambda(1-\lambda)^{|k|-1}, & k = c \pm 1, c \pm 2, c \pm 3, \ldots \end{cases}$$

8. $X_S \in \mathbb{R}^{n\times d}$ is the matrix obtained from X by zeroing rows with index in $[n] \setminus S$.

Clearly, $\rho \cdot \mathcal{G}(0,\lambda)$ is symmetric and satisfies $\mathbb{P}_{\boldsymbol{z}\sim\rho\cdot\mathcal{G}(0,\lambda)}(|\boldsymbol{z}| \leqslant 1) = \alpha$ for any $\rho > 1$ and any $\lambda \in (0,1)$. Moreover, as $D_{\mathrm{KL}}(\mathcal{G}(0,\lambda) \| \mathcal{G}(c,\lambda)) = D_{\mathrm{KL}}(\rho \cdot \mathcal{G}(0,\lambda) \| \rho \cdot \mathcal{G}(c,\lambda))$, we have the desired insensitivity to scaling. Finally, for small enough values of λ low shift-sensitivity holds for integer-valued shifts, as the distribution is discrete in nature. (See Lemma 11 for a formal statement.)

Matrices spanning sparse subspaces To construct the aforementioned design matrix, we wish to find two rectangular matrices $A \in \mathbb{R}^{m\times d}$ and $B \in \mathbb{R}^{(n-m)\times d}$ such that A is $\Omega(m)$-spread and the row spans of A and B are orthogonal to each other. Let $X \in \mathbb{R}^{n\times d}$ be the matrix obtained by stacking A onto B. Then for any vector β in the row span of A, i.e. $\beta \in \mathrm{rspan}(A)$, we have

$$\|X\beta\|_2^2 = \left\| \begin{bmatrix} A \\ B \end{bmatrix} \beta \right\|_2^2 = \|A\beta\|_2^2 + \|B\beta\|_2^2 = \|A\beta\|_2^2 \,,$$

and thus X also has the required m-spreadness. To find two such matrices A and B, the following observation turns out to be crucial:[9]

An m-by-d Rademacher matrix is $\Omega(m)$-spread with high probability.

With this ingredient we are now ready to construct the design matrix X. Let $\mathcal{R}$ denote Rademacher distribution. Let $\boldsymbol{A}^* \sim \mathcal{R}^{m\times d}$ and $\boldsymbol{B}^* \sim \mathcal{R}^{(n-m)\times d}$ be independently sampled, and let $V \subseteq \mathbb{R}^d$ be an $\Omega(d)$-dimensional subspace. We construct

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^*\Pi_V \\ \boldsymbol{B}^*\Pi_{V^\perp} \end{bmatrix} \,,$$

where Π_V denotes the projector onto the subspace V. Then the row spans of $\boldsymbol{A}, \boldsymbol{B}$ are orthogonal.

Putting things together The ideas presented above allow us to construct hypotheses as in Eq. (2.1) that are indistinguishable from each other even though the corresponding parameter vectors are far from each other. Let $\beta, \beta' \in \mathbb{R}^d$ be distinct vectors in the row span of $\boldsymbol{A}$ with integer coordinates satisfying $\|\beta - \beta'\|_2 \geqslant \Omega\left(\sqrt{d}\right)$ and $\|\beta\|_2, \|\beta'\|_2 \leqslant O\left(\sqrt{d}\right)$. By construction, $\boldsymbol{X}\beta$ and $\boldsymbol{X}\beta'$ are both n-dimensional vectors with $\Omega(m)$ nonzero entries, each integer-valued. So, by low shift-sensitivity of the noise distribution,

$$\begin{aligned} H: &\quad \boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\eta}\,, \\ H': &\quad \boldsymbol{y} = \boldsymbol{X}\beta' + \boldsymbol{\eta} \end{aligned}$$

are stastistically indistinguishable. Finally, expanding on these ideas Theorem 3 will follow. Furthermore, by insensitivity to scaling of D_η we can now blow up the error by scaling up $\sigma \cdot \boldsymbol{y} = \boldsymbol{X}(\sigma \cdot \beta) + \sigma \cdot \boldsymbol{\eta}$ and $\sigma \cdot \boldsymbol{y}' = \boldsymbol{X}(\sigma \cdot \beta') + \sigma \cdot \boldsymbol{\eta}$, for any $\sigma > 0$, without making the distinguishing problem easier.

9. We remark that a similar observation holds for other distributions (e.g. Gaussian). See Theorem 26 for a formal proof. The value of integer values will become evident in the interplay between the design matrix and the noise distribution.

2.2. Differences between well-spreadness and RIP, RE properties

In the context of compressed sensing we are given $n \leqslant d$ observations of the form $\boldsymbol{y}_i = \langle M_i, \beta \rangle + \boldsymbol{\eta}_i$ with $M_i, \beta \in \mathbb{R}^d$ and $\boldsymbol{\eta}$ being additive noise. In order to guarantee recovery of the compressed vector β, RIP Candes and Tao (2005); Donoho (2006); Candes et al. (2006); Kashin and Temlyakov (2007) is arguably the most popular condition to enforce on the sensing matrix:

Definition 6 (Restricted isometry property) *We say a matrix $M \in \mathbb{R}^{n\times d}$ satisfies the (k,δ)-restricted isometry property (RIP) if*

$$(1-\delta)\|v\|_2^2 \leqslant \|Mv\|_2^2 \leqslant (1+\delta)\|v\|_2^2$$

for every vector v with at most k nonzero entries.

We argue here that the relation between well-spreadness and oblivious regression fundamentally differs from that of RIP and compressed sensing in two ways.

First, while state-of-the-art algorithms for compressed sensing rely on RIP in order to filter out the noise in the observations and recover the hidden vector, it is known that small prediction error can be achieved in exponential time *without* any constraint on the sensing matrix Bunea et al. (2007). In contrast, Theorem 3 shows that in the context of oblivious regression, no algorithm can achieve even small prediction error for a family of design matrices that are not sufficiently well-spread.

Second, RIP is not purely a condition of the column span of the sensing matrix. In particular, if M satisfies (k,δ)-RIP, then the kernel of M must be $\Omega(k)$-spread Guruswami et al. (2021).[10] Conversely, it is easy to construct matrices with well-spread column span and kernel containing sparse vectors.

As the following examples show, it is easy to construct matrices that satisfy RIP but are not well-spread and vice versa.

Example 1 (RIP but not even 1-spread) *Let $\boldsymbol{W} \sim \mathcal{N}(0,1)^{(n-1)\times(d-1)}$ and consider the following n-by-d matrix (we do not fix the relation between n and d),*

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{n-1}}\boldsymbol{W} \end{bmatrix}.$$

If $n \gtrsim \delta^{-2} k \log d$, then with high probability, $\boldsymbol{M}$ satisfies (k,δ)-RIP. However, $\boldsymbol{M}$ is not even 1-spread, since its column span contains the canonical basis vector $e_1 \in \mathbb{R}^n$.

Example 2 (Well-spread but not satisfying RIP) *Let $\boldsymbol{W} \sim \mathcal{N}(0,1)^{n\times(d-1)}$ and consider the following n-by-d matrix,*

$$\boldsymbol{M} = \begin{bmatrix} v & \frac{1}{\sqrt{n}}\boldsymbol{W} \end{bmatrix},$$

where $v \in \mathbb{R}^n$ is a unit vector parallel or highly correlated to the first column of $\boldsymbol{W}$. Then $\boldsymbol{M}$ cannot satisfy RIP or RE. However, if $n \geqslant Cd$ for some sufficiently large absolute constant C, then it is easy to verify that, $\boldsymbol{M}$ is $\Omega(n)$-spread with high probability.

10. The careful reader may have noticed that δ seems to play no role in this implication. In fact, the relation is more general than what we consider here. See Guruswami et al. (2021).

3. Information-theoretic bounds for oblivious regression

We state and prove the more precise and technical version of Theorem 3 that shows, there exists a family of $\max\left\{\Omega\left(\frac{\log d}{\alpha^2}\right), \Omega\left(\frac{d}{\alpha}\right)\right\}$-spread design matrices such that, consistent estimation is information-theoretically impossible in oblivious linear regression.

Theorem 7 *Let $\alpha = \alpha(n) \in (0,1)$. For arbitrary $\gamma = \gamma(n) > 0$, there exist:*

1. *a distribution $\mathcal{D}_X$ over $n \times d$ matrices X with $X^\top X = n \cdot \mathrm{Id}$,*
2. *a distribution $\mathcal{D}_\beta$ over d-dimensional vectors, and*
3. *a distribution $\mathcal{D}_\eta$ —independent of $\mathcal{D}_X$ and $\mathcal{D}_\beta$— over n-dimensional vectors with independent, symmetrically distributed entries satisfying $\min_{i\in[n]} \mathbb{P}_{\boldsymbol{\eta}\sim\mathcal{D}_\eta}(|\boldsymbol{\eta}_i| \leqslant 1) = \alpha$,*

such that,

1. *$\boldsymbol{X} \sim \mathcal{D}_X$ is $\max\left\{\Omega\left(\frac{\log d}{\alpha^2}\right), \Omega\left(\frac{d}{\alpha}\right)\right\}$-spread with high probability; and*
2. *for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, given as input $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\eta}$ with $\boldsymbol{X} \sim \mathcal{D}_X$, $\boldsymbol{\eta} \sim \mathcal{D}_\eta$, and $\boldsymbol{\beta}^* \sim \mathcal{D}_\beta$ sampled independently, one has*

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma\,,$$

conditioning on $\boldsymbol{X}$ being $\max\left\{\Omega\left(\frac{\log d}{\alpha^2}\right), \Omega\left(\frac{d}{\alpha}\right)\right\}$-spread.

To prove Theorem 7, we provide the following two lemmas which we will prove in Section 3.1 and Section 3.2 respectively.

Lemma 8 shows that, there exists a family of $\Omega\left(\frac{d}{\alpha}\right)$-spread design matrices such that, consistent estimation is information-theoretically impossible in oblivious linear regression.

Lemma 8 *Let $\alpha = \alpha(n) \leqslant O(1)$. For arbitrary $\gamma = \gamma(n) > 0$, there exist:*

1. *a distribution $\mathcal{D}_X$ over $n \times d$ matrices X with $X^\top X = n \cdot \mathrm{Id}$,*
2. *a distribution $\mathcal{D}_\beta$ over d-dimensional vectors, and*
3. *a distribution $\mathcal{D}_\eta$ —independent of $\mathcal{D}_X$ and $\mathcal{D}_\beta$— over n-dimensional vectors with independent, symmetrically distributed entries satisfying $\min_{i\in[n]} \mathbb{P}_{\boldsymbol{\eta}\sim\mathcal{D}_\eta}(|\boldsymbol{\eta}_i| \leqslant 1) = \alpha$,*

such that,

1. *$\boldsymbol{X} \sim \mathcal{D}_X$ is $\Omega\left(\frac{d}{\alpha}\right)$-spread with high probability; and*
2. *for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, given as input $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\eta}$ with $\boldsymbol{X} \sim \mathcal{D}_X$, $\boldsymbol{\eta} \sim \mathcal{D}_\eta$, and $\boldsymbol{\beta}^* \sim \mathcal{D}_\beta$ sampled independently, one has*

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma\,,$$

conditioning on $\boldsymbol{X}$ being $\Omega\left(\frac{d}{\alpha}\right)$-spread.

Lemma 9 shows that, there exists a family of $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread design matrices such that, consistent estimation is information-theoretically impossible in oblivious linear regression.

Lemma 9 *Let $\alpha = \alpha(n) \leqslant O(1)$. For arbitrary $\gamma = \gamma(n) > 0$, there exist:*

1. *a distribution $\mathcal{D}_X$ over $n \times d$ matrices X with $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spreadness and $X^\top X = n \cdot \mathrm{Id}$,*
2. *a distribution $\mathcal{D}_\beta$ over d-dimensional vectors, and*
3. *a distribution $\mathcal{D}_\eta$ —independent of $\mathcal{D}_X$ and $\mathcal{D}_\beta$— over n-dimensional vectors with independent, symmetrically distributed entries satisfying $\min_{i\in[n]} \mathbb{P}_{\boldsymbol{\eta}\sim\mathcal{D}_\eta}\left(|\boldsymbol{\eta}_i| \leqslant 1\right) = \alpha$,*

such that for every estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, given as input $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\eta}$ with $\boldsymbol{X} \sim \mathcal{D}_X$, $\boldsymbol{\eta} \sim \mathcal{D}_\eta$, and $\boldsymbol{\beta}^* \sim \mathcal{D}_\beta$ sampled independently, one has

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma\,.$$

Theorem 7 follows directly from the above two lemmas.

Proof By Lemma 8 and Lemma 9. ■

We introduce here the noise distribution (i.e. D_η) which will play a crucial role in our proof of Lemma 8 and Lemma 9.

Definition 10 (Symmetric geometric distribution) *The symmetric geometric distribution with location parameter $c \in \mathbb{Z}$ and scale parameter $\lambda \in (0,1)$, denoted by $\mathcal{G}(c,\lambda)$, is a discrete distribution supported on $\mathbb{Z}$. Its probability mass function is defined as*

$$p(k) = \begin{cases} \alpha, & k = c, \\ \frac{1-\alpha}{2}\cdot\lambda(1-\lambda)^{|k|-1}, & k = c \pm 1, c \pm 2, c \pm 3, \cdots, \end{cases} \tag{3.1}$$

where α is the same α in Problem 2. Let $\mathcal{G}(\lambda) = \mathcal{G}(0,\lambda)$ by default.

We collect several useful facts about symmetric geometric distributions in the following lemma.

Lemma 11 *Let $\mathcal{G}(c,\lambda)$ be the symmetric geometric distribution with parameters c and λ, as defined in Definition 10.*

1. *For any $\sigma > 0$, $c \in \mathbb{Z}$, and $\lambda \in (0,1)$, we have*

$$\mathrm{D}_{\mathrm{KL}}\left(\sigma\cdot\mathcal{G}(\lambda)\,\|\,\sigma\cdot\mathcal{G}(c,\lambda)\right) = \mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda)\,\|\,\mathcal{G}(c,\lambda)\right).$$

2. *Suppose $\alpha \leqslant 1/4$. Let $\lambda = 2\alpha$. Then,*

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda)\,\|\,\mathcal{G}(1,\lambda)\right) \leqslant 4\alpha^2.$$

3. *Suppose $d \geqslant 4$ and $\alpha \leqslant 1/2$. Let $\lambda = 2\alpha d^{-5}$. Then for any $\Delta \in [d^4]$, we have*

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda)\,\|\,\mathcal{G}(\Delta,\lambda)\right) \leqslant 8\alpha\cdot\log d.$$

Proof Given $\lambda \in (0,1)$ and $\Delta \in \mathbb{Z}$, let p and q be the probability mass functions of $\mathcal{G}(\lambda)$ and $\mathcal{G}(\Delta,\lambda)$ respectively. By definition,

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda) \,\|\, \mathcal{G}(\Delta,\lambda)\right) = \sum_{k=-\infty}^{\infty} p(k)\log\frac{p(k)}{q(k)} = \underbrace{\sum_{k\neq 0,\Delta} p(k)\log\frac{p(k)}{q(k)}}_{=:D(\lambda,\Delta)} + \underbrace{\sum_{k=0,\Delta} p(k)\log\frac{p(k)}{q(k)}}_{=:D'(\lambda,\Delta)}.$$

After some direct computations, we have

$$\begin{aligned} D(\lambda,\Delta) &= \frac{1-\alpha}{2}\cdot\frac{1}{\lambda}\cdot\log\frac{1}{1-\lambda}\cdot\left[2\lambda\Delta + 2(1-\lambda)^{\Delta} - 2 + \lambda^2\Delta(1-\lambda)^{\Delta-1}\right], \text{ and} \\ D'(\lambda,\Delta) &= \alpha\cdot\left(1 - \frac{(1-\alpha)\lambda(1-\lambda)^{\Delta-1}}{2\alpha}\right)\cdot\log\frac{2\alpha}{(1-\alpha)\lambda(1-\lambda)^{\Delta-1}}. \end{aligned} \tag{3.2}$$

We remark that both $D(\lambda,\Delta)$ and $D'(\lambda,\Delta)$ can be viewed as the Kullback-Leibler divergence between two probabilistic distributions up to a positive scaling factor. Thus, $D(\lambda,\Delta)$ and $D'(\lambda,\Delta)$ are always non-negative regardless of λ and Δ.

1. By definition.

2. Substituting λ by 2α and Δ by 1 in Eq. (3.2), we have

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(2\alpha) \,\|\, \mathcal{G}(1,2\alpha)\right) = \alpha^2\cdot\log\frac{1}{1-2\alpha}.$$

   Using the assumption $\alpha \leqslant 1/4$ and the fact $\log\frac{1}{1-x} \leqslant 2x$ for $0 \leqslant x \leqslant 1/2$, we have

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda) \,\|\, \mathcal{G}(1,\lambda)\right) \leqslant 4\alpha^2.$$

3. Fix an arbitrary $\Delta \in [d^4]$. By the assumption $d \geqslant 4$ and $\alpha \leqslant 1/2$, one has $\lambda\Delta \leqslant 2\alpha d^{-1} \leqslant 1/4$ and hence

$$(1-\lambda)^{\Delta} = \sum_{i=0}^{\Delta}\binom{\Delta}{i}(-\lambda)^i \leqslant 1 - \lambda\Delta + (\lambda\Delta)^2.$$

   Then, it is not difficult to show

$$\begin{aligned} D(\lambda,\Delta) &\leqslant \lambda^2\Delta\left(2\Delta + \frac{1}{1-\lambda}\right) \leqslant 4\lambda^2\Delta^2 \leqslant 4\alpha^2 d^{-2}, \text{ and} \\ D'(\lambda,\Delta) &\leqslant \alpha\left(2\alpha + 5\log d + 2\lambda\Delta\right) \leqslant 7\alpha\cdot\log d. \end{aligned}$$

   Therefore, we have

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda) \,\|\, \mathcal{G}(\Delta,\lambda)\right) = D(\lambda,\Delta) + D'(\lambda,\Delta) \leqslant 8\alpha\cdot\log d.$$

■

3.1. Proof of Lemma 8

To prove Lemma 8, we apply Fano's method as introduced in Appendix A.2. We first construct an $\Omega\left(\frac{d}{\alpha}\right)$-spread design matrix $X \in \mathbb{R}^{n\times d}$ and a set $\mathcal{B} \subset \mathbb{R}^d$ of $\Omega\left(d^d\right)$ parameter vectors. We set $m = d/(50\alpha)$ throughout Section 3.1.

Design matrix Let $\boldsymbol{R}$ be an $m \times d$ Rademacher matrix. By Theorem 26, there exists an absolute constant $c \in (0,1)$ such that $\boldsymbol{R}$ is $\Omega\left(\frac{d}{\alpha}\right)$-spread with high probability for $\alpha \leqslant c$. Suppose $\alpha \leqslant c$. Thus, "most" $m \times d$ $\{\pm 1\}$-matrices are $\Omega\left(\frac{d}{\alpha}\right)$-spread. Let Y be such a matrix, i.e. $Y \in \{\pm 1\}^{m\times d}$ and Y is $\Omega\left(\frac{d}{\alpha}\right)$-spread.

Let X_1 be an arbitrary orthonormal basis matrix of subspace $\operatorname{cspan}(Y)$. Then scale $X_1 \in \mathbb{R}^{m\times d}$ properly such that $X_1^\top X_1 = n \cdot \mathrm{Id}$. Let $X^\top = \begin{bmatrix} X_1^\top & X_2^\top \end{bmatrix}$ where X_2 is a zero matrix. Then the design matrix X is $\Omega\left(\frac{d}{\alpha}\right)$-spread and satisfies $X^\top X = n \cdot \mathrm{Id}$.

Hard-to-distinguish parameter vectors The set of parameter vectors is constructed by reverse engineering. Let $\ell_{X_1} : \mathbb{R}^d \to \mathbb{R}^m$ be a linear mapping defined by $\ell_{X_1}(v) := X_1 v$. We first construct a set $\mathcal{U} \subset \operatorname{cspan}(Y)$ with several desired properties and then let $\mathcal{B}$ be a scaled preimage of $\mathcal{U}$ under the injective linear mapping ℓ_{X_1}. Let $\mathcal{U} = \left\{Yv \mid v \in [d]^d\right\}$. Note that for any $u \in \mathcal{U}$, we have $u \in \mathbb{Z}^m$ and $\|u\|_\infty \leqslant d^2$. Choose the set of parameter vectors to be

$$\mathcal{B} = \sigma \cdot \ell_{X_1}^{-1}(\mathcal{U}) = \sigma \cdot \left\{X_1^{-1} u \mid u \in \mathcal{U}\right\}, \tag{3.3}$$

where $\sigma > 0$ is a scaling factor. Clearly, σ controls the separateness of set $\mathcal{B}$. Then for any two distinct vectors $\beta, \beta' \in \mathcal{B}$, we have

$$X(\beta - \beta') \in \sigma \cdot \left\{-2d^2, -2d^2+1, ..., 2d^2\right\}^m \times \{0\}^{n-m}. \tag{3.4}$$

We remark that, although the design matrix X we constructed above is rather sparse, it is not necessarily this case and we can easily make X non-sparse via the following trick. Let $R \in \mathbb{R}^{d\times d}$ be a dense orthogonal matrix, e.g. a uniformly random one. Now Let $X' = XR$ be the design matrix and $\mathcal{B}' = \left\{R^\top \beta : \beta \in \mathcal{B}\right\}$ be the set of parameter vectors. Clearly, the spreadness of X' is identical to the spreadness of X, since $\operatorname{cspan}(X') = \operatorname{cspan}(X)$. Also, $(X')^\top (X') = n \cdot \mathrm{Id}$ and Eq. (3.4) is preserved as well.

Putting things together Now we are ready to prove Lemma 8.

Proof Consider the following hypothesis testing problem. Let D_β be the uniform distribution over set $\mathcal{B}$ in Eq. (3.3) and $\boldsymbol{\beta}^* \sim D_\beta$. Let X be the $\Omega\left(\frac{d}{\alpha}\right)$-spread design matrix as constructed above. Set $\lambda = 2\alpha d^{-5}$ and use the same σ in Eq. (3.3). Let the noise vector be $\boldsymbol{\eta} = (\boldsymbol{\eta}_i)_{i=1}^n$ where $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_n \sim \sigma \cdot \mathcal{G}(\lambda)$ are independent symmetric geometric random variables as defined in Definition 10. Observing $\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\eta}$, the goal is to distinguish d^d hypotheses $\{\boldsymbol{y} = X\beta + \boldsymbol{\eta} : \beta \in \mathcal{B}\}$. Now we apply Fano's method by reducing this hypothesis testing problem to oblivious linear regression.

Given two distinct vectors $\beta, \beta' \in \mathcal{B}$, let $\Delta_i := \sigma^{-1}|(X\beta)_i - (X\beta')_i|$ for $i \in [n]$. By Eq. (3.4), we have $\Delta_i \in \left\{-2d^2, -2d^2+1, ..., 2d^2\right\}$. By independence of random variables $\{\boldsymbol{\eta}_i\}_{i=1}^n$ and the chain rule of Kullback-Leibler divergence, we have

$$\mathrm{D}_{\mathrm{KL}}\left(X\beta + \boldsymbol{\eta} \,\|\, X\beta' + \boldsymbol{\eta}\right) = \sum_{i=1}^n \mathrm{D}_{\mathrm{KL}}\left((X\beta)_i + \boldsymbol{\eta}_i \,\|\, (X\beta')_i + \boldsymbol{\eta}_i\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \mathrm{D}_{\mathrm{KL}}\left(\sigma \cdot \mathcal{G}(\lambda) \,\|\, \sigma \cdot \mathcal{G}(\Delta_i, \lambda)\right) \\
&= \sum_{i=1}^{m} \mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(\lambda) \,\|\, \mathcal{G}(\Delta_i, \lambda)\right) \\
&\leqslant m \cdot 8\alpha \log d = 0.16 d \log d, \qquad (3.5)
\end{aligned}
$$

where the second equality uses Eq. (3.4), the third equality and the inequality is due to Lemma 11.

Let $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$ be an arbitrary estimator for oblivious linear regression and $\gamma > 0$ be an arbitrary given error bound. Note $\mathcal{B}$ is $\sigma\sqrt{m/n}$-separated and $|\mathcal{B}| = d^d$. Combining Eq. (A.1) with Eq. (3.5), and setting $\sigma^2 = 8\gamma n/m = 400\gamma n\alpha/d$, we have

$$
\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma, \qquad (3.6)
$$

for any $d \geqslant 3$. ■

Some remarks To show any estimator is inconsistent, it is enough to set $\sigma^2 = n\alpha/d$ in the above proof. In this case[11], the set $\mathcal{B}$ is $\Omega(1)$-separated and the error lower bound is $\Omega(1)$, which does not vanish as n goes to infinity. Moreover, since the lower bound Eq. (3.6) holds for any $\gamma > 0$, we have actually showed that no estimator can obtain bounded estimation error.

Eq. (3.4) is crucial in the above proof. In fact, to prove Lemma 8, it is enough to construct an $\Omega\left(\frac{d}{\alpha}\right)$-spread design matrix $X \in \mathbb{R}^{n\times d}$ and a set $\mathcal{B} \subset \mathbb{R}^d$ of parameter vectors such that, for any $\beta, \beta' \in \mathcal{B}$, one has (i) $X\beta \in \mathbb{Z}^n$, (ii) $\|X(\beta - \beta')\|_\infty \leqslant \mathrm{poly}(d)$, and (iii) $\|X(\beta - \beta')\|_0 \lesssim \log|\mathcal{B}|$.

3.2. Proof of Lemma 9

To prove Lemma 9, we apply Fano's method as introduced in Appendix A.2. We first construct an $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread design matrix $X \in \mathbb{R}^{n\times d}$ and a set $\mathcal{B} \subset \mathbb{R}^d$ of $\Omega(d)$ parameter vectors. We set $k = \log(d)/(200\alpha^2)$ throughout Section 3.2.

Design matrix Pick a random orthogonal matrix $Q \in \mathbb{R}^{d\times d}$. Let $Y^\top = \begin{bmatrix} Q^\top & Q^\top \end{bmatrix}$. It is straightforward to see Y is 1-spread. Let $X_1^\top = \begin{bmatrix} Y_1^\top & \cdots & Y_k^\top \end{bmatrix}$ where $Y_i = Y$ for $i \in [k]$. Then X_1 is $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread. Then scale X_1 properly such that $X_1^\top X_1 = n \cdot \mathrm{Id}$. Obviously, scaling a matrix by a nonzero factor does not change its spreadness. Let $X^\top = \begin{bmatrix} X_1^\top & X_2^\top \end{bmatrix}$ where X_2 is a zero matrix. Note this requires $n \geqslant k \cdot 2d = d\log(d)/(100\alpha^2)$. Then the design matrix X is $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread and satisfies $X^\top X = n \cdot \mathrm{Id}$.

Hard-to-distinguish parameter vectors Let $\{q_1, \ldots q_d\} \subset \mathbb{R}^d$ be the columns of Q. Let

$$
\mathcal{B} = \sigma\sqrt{\frac{k}{n}} \cdot \{q_1, \ldots, q_d\} \qquad (3.7)
$$

11. Note that σ^2 is proportional to the variance of the noise distribution. Setting $\sigma^2 = n\alpha/d$, then the signal-to-noise ratio does not grow with n, which provides one evidence why consistent estimation is impossible.

be the set of parameter vectors to be distinguish where $\sigma > 0$ is a scaling factor. The $\sqrt{k/n}$ term in Eq. (3.7) is just to make the subsequent notations cleaner. It is worth noting that for each $\beta \in \mathcal{B}$, $X\beta$ is the "least-spread" vector in $\mathrm{cspan}\,(X)$. For any two distinct vectors $\beta, \beta' \in \mathcal{B}$, we have

$$X(\beta - \beta') \in \{0, \sigma\}^n\,, \quad \left\|X(\beta - \beta')\right\|_0 = 2k. \tag{3.8}$$

In other words, $X\beta$ and $X\beta'$ differ on exactly $2k$ coordinates and all the differences are equal to σ.

Putting things together Now we are ready to prove Lemma 9.

Proof Consider the following hypothesis testing problem. Let D_β be the uniform distribution over set $\mathcal{B}$ in Eq. (3.7) and $\boldsymbol{\beta}^* \sim D_\beta$. Let X be the $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread design matrix as constructed above. Let the noise vector be $\boldsymbol{\eta} = (\boldsymbol{\eta}_i)_{i=1}^n$ where $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_n \sim \sigma \cdot \mathcal{G}(2\alpha)$ are independent symmetric geometric random variables as defined in Definition 10. Here the scaling factor $\sigma > 0$ is the same σ in Eq. (3.7). Observing $\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\eta}$, the goal is to distinguish d hypotheses $\{\boldsymbol{y} = X\beta + \boldsymbol{\eta} : \beta \in \mathcal{B}\}$. Now we apply Fano's method by reducing this hypothesis testing problem to oblivious linear regression.

For any two distinct vectors $\beta, \beta' \in \mathcal{B}$, by independence of random variables $\{\boldsymbol{\eta}_i\}_{i=1}^n$ and the chain rule of Kullback-Leibler divergence, we have

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(X\beta + \boldsymbol{\eta} \,\|\, X\beta' + \boldsymbol{\eta}\right) &= \sum_{i=1}^n \mathrm{D}_{\mathrm{KL}}\left((X\beta)_i + \boldsymbol{\eta}_i \,\|\, (X\beta')_i + \boldsymbol{\eta}_i\right) \\
&= 2k \cdot \mathrm{D}_{\mathrm{KL}}\left(\sigma \cdot \mathcal{G}(2\alpha) \,\|\, \sigma \cdot \mathcal{G}(1, 2\alpha)\right) \\
&= 2k \cdot \mathrm{D}_{\mathrm{KL}}\left(\mathcal{G}(2\alpha) \,\|\, \mathcal{G}(1, 2\alpha)\right) \\
&\leqslant 2k \cdot 4\alpha^2 = 0.04 \log d,
\end{aligned} \tag{3.9}$$

where the second equality uses Eq. (3.8), the third equality and the inequality is due to Lemma 11.

Let $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$ be an arbitrary estimator for oblivious linear regression and $\gamma > 0$ be an arbitrary given error bound. Note $\mathcal{B}$ is $\sigma\sqrt{2k/n}$-separated and $|\mathcal{B}| = d$. Combining Eq. (A.1) with Eq. (3.9), and setting $\sigma^2 = 4\gamma n/k = 800\gamma n\alpha^2/\log d$, we have

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \gamma, \tag{3.10}$$

for any $d \geqslant 5$. ∎

Some remarks To show inconsistency, it is enough to set $\sigma^2 = n\alpha^2/\log d$ in the above proof. The error lower bound Eq. (3.10) can get arbitrarily large. To prove Lemma 9, it suffices to construct an $\Omega\left(\frac{\log d}{\alpha^2}\right)$-spread design matrix $X \in \mathbb{R}^{n\times d}$ and a set $\mathcal{B} \subset \mathbb{R}^d$ of parameter vectors such that, for any $\beta, \beta' \in \mathcal{B}$, one has (i) $X\beta \in \mathbb{Z}^n$, (ii) $\|X(\beta - \beta')\|_\infty \leqslant O(1)$, and (iii) $\|X(\beta - \beta')\|_0 \lesssim \log|\mathcal{B}|$.

Acknowledgments

We thank David Steurer and Zhihan Jin for helpful discussions.

References

Zeyuan Allen-Zhu, Rati Gelashvili, and Ilya Razenshteyn. Restricted isometry property for general p-norms. *IEEE Transactions on Information Theory*, 62(10):5839–5854, 2016.

Afonso S Bandeira, Edgar Dobriban, Dustin G Mixon, and William F Sawin. Certifying the restricted isometry property is hard. *IEEE transactions on information theory*, 59(6):3448–3450, 2013.

Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326, 2012.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *NIPS*, pages 2110–2119, 2017.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic journal of statistics*, 1:169–194, 2007.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Davin Choo and Tommaso d'Orsi. The complexity of sparse tensor PCA. *Advances in Neural Information Processing Systems*, 34, 2021.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-time algorithms for sparse PCA. *arXiv preprint arXiv:1907.11635*, 2019.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. The average-case time complexity of certifying the restricted isometry property. *IEEE Transactions on Information Theory*, 2021.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Tommaso d'Orsi, Pravesh K Kothari, Gleb Novikov, and David Steurer. Sparse PCA: Algorithms, adversarial perturbations and certificates. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 553–564. IEEE, 2020.

Tommaso d'Orsi, Chih-Hung Liu, Rajai Nasser, Gleb Novikov, David Steurer, and Stefan Tiegel. Consistent estimation for PCA and sparse regression with oblivious outliers. *Advances in Neural Information Processing Systems*, 34, 2021a.

Tommaso d'Orsi, Gleb Novikov, and David Steurer. Consistent regression when oblivious outliers overwhelm. In *International Conference on Machine Learning*, pages 2297–2306. PMLR, 2021b.

Efim Davydovich Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Mathematics of the USSR-Sbornik*, 48(1):173, 1984.

Venkatesan Guruswami, James R Lee, and Avi Wigderson. Euclidean sections of ℓ_1^N with sublinear randomness and error-correction over the reals. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 444–454. Springer, 2008.

Venkatesan Guruswami, James R Lee, and Alexander Razborov. Almost Euclidean subspaces of ℓ_1^N via expander codes. *Combinatorica*, 30(1):47–68, 2010.

Venkatesan Guruswami, Peter Manohar, and Jonathan Mosheiff. ℓ_p-spread properties of sparse matrices. *CoRR*, abs/2108.13578, 2021. URL https://arxiv.org/abs/2108.13578.

Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and David Steurer. Bayesian estimation from few samples: community detection and related problems. *arXiv preprint arXiv:1710.00264*, 2017.

Piotr Indyk. Uncertainty principles, extractors, and explicit embeddings of L2 into L1. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 615–620, 2007.

Boris S Kashin and Vladimir N Temlyakov. A remark on compressed sensing. *Mathematical notes*, 82(5):748–755, 2007.

Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE transactions on information theory*, 60(8):4999–5006, 2014.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

Abhiram Natarajan and Yi Wu. Computational complexity of certifying restricted isometry property. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2014, September 4-6, 2014, Barcelona, Spain*, pages 371–380, 2014. doi: 10.4230/LIPIcs.APPROX-RANDOM.2014.371. URL https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2014.371.

Scott Pesme and Nicolas Flammarion. Online robust regression via SGD on the l1 loss. *Advances in Neural Information Processing Systems*, 33:2540–2552, 2020.

Jonathan Scarlett and Volkan Cevher. An introductory guide to Fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014. doi: 10.1109/TIT.2013.2290112.

Efthymios Tsakonas, Joakim Jaldén, Nicholas D Sidiropoulos, and Björn Ottersten. Convergence of the huber regression M-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, 2014.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of RIP certification. *Advances in Neural Information Processing Systems*, 29, 2016.

Jonathan Weed. Approximately certifying the restricted isometry property is hard. *IEEE Trans. Inf. Theory*, 64(8):5488–5497, 2018. doi: 10.1109/TIT.2017.2776131. URL https://doi.org/10.1109/TIT.2017.2776131.

Ilias Zadik, Min Jae Song, Alexander S Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. *arXiv preprint arXiv:2112.03898*, 2021.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948. PMLR, 2014.

Appendix A. Background

A.1. Basic notation

We use the convention $\mathbb{N} = \{0, 1, 2, 3, ...\}$. For a positive integer n, let $[n] := \{1, 2, ..., n\}$. For $\alpha \in \mathbb{N}^n$, define $|\alpha| := \sum_{i=1}^n \alpha_i$. For a vector $v \in \mathbb{R}^n$, let $\mathrm{supp}(v) := \{i \in [n] : v_i \neq 0\}$ be its support, $\|v\|_p := (\sum_{i=1}^n |v_i|^p)^{1/p}$ be its ℓ_p-norm ($p \geqslant 1$), and $\|v\|_0 := |\mathrm{supp}(v)|$. Given a vector $v \in \mathbb{R}^n$ and a subset $S \subseteq [n]$, let $v_S \in \mathbb{R}^{|S|}$ denote the projection of v onto the coordinates in S. For a matrix X, let $\mathrm{cspan}\,(X)$ denote its column span, $\mathrm{rspan}(A)$ denote its row span, and $\ker(X)$ denote its kernel or null space. Let $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ denote its minimum and maximum singular values respectively. We use standard asymptotic notations $\Omega(\cdot), O(\cdot), \lesssim, \gtrsim$ to hide absolute multiplicative constants. Throughout this paper, we write random variables in boldface. We say an event happens *with high probability* if it happens with probability $1 - o(1)$. Given two distributions ν and μ, let $\mathrm{D_{KL}}(\nu \,\|\, \mu)$ denote their Kullback-Leibler divergence. For two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, we write $\mathrm{D_{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y})$ to denote the Kullback-Leibler divergence between their distributions. By saying Gaussian (or Rademacher) matrix, we mean a matrix whose entries are independent standard

Gaussian (or Rademacher) random variables. Unless explicitly stated, the base of logarithm is the natural number e.

A.2. Fano's method

Fano's method is a classical approach to proving lower bounds for statistical estimation problems, which we apply to prove the information-theoretic lower bounds in Section 3.

Suppose we are given an δ-separated set $\mathcal{B} \subset \mathbb{R}^d$. That is, $\|\beta - \beta'\|_2 \geqslant \delta$ for any distinct $\beta, \beta' \in \mathcal{B}$. Let D_β be the uniform distribution over $\mathcal{B}$ and $\boldsymbol{\beta}^* \sim D_\beta$. Let $X \in \mathbb{R}^{n\times d}$ be a known design matrix and $\boldsymbol{\eta}$ be the noise vector. Observing $\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\eta}$, the hypothesis testing problem is to distinguish $|\mathcal{B}|$ distributions $\{X\beta + \boldsymbol{\eta} : \beta \in \mathcal{B}\}$. Let $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$ be an arbitrary estimator for the linear regression problem. By a reduction from the hypothesis testing problem and applying Fano's inequality (Lemma 12) combined with the convexity of the Kullback-Leibler divergence, one has[12]

$$\mathbb{E}\left\|\hat{\beta}(\boldsymbol{y}) - \boldsymbol{\beta}^*\right\|_2^2 \geqslant \frac{\delta^2}{4}\left(1 - \frac{\max\limits_{\beta,\beta'\in\mathcal{B}} \mathrm{D}_{\mathrm{KL}}\left(X\beta + \boldsymbol{\eta} \,\|\, X\beta' + \boldsymbol{\eta}\right) + \log 2}{\log|\mathcal{B}|}\right). \tag{A.1}$$

Lemma 12 (Fano's inequality) *Let Σ be a finite set and $\boldsymbol{J}$ be a random variable uniformly distributed over Σ. Suppose $\boldsymbol{J} \to \boldsymbol{Z} \to \hat{\boldsymbol{J}}$ is a Markov chain. Then,*

$$\mathbb{P}\left(\boldsymbol{J} \neq \hat{\boldsymbol{J}}\right) \geqslant 1 - \frac{I(\boldsymbol{J};\boldsymbol{Z}) + \log 2}{\log|\Sigma|},$$

where $I(\boldsymbol{J};\boldsymbol{Z})$ denotes the mutual information between $\boldsymbol{J}$ and $\boldsymbol{Z}$.

A.3. Spreadness and distortion

Definition 13 (ℓ_p-spreadness) *Let $p \geqslant 1$, $\delta \in [0,1]$, $n \in \mathbb{N}$, and $m \leqslant n$. A vector $v \in \mathbb{R}^n$ is said to be (m,δ)-ℓ_p-spread if for every subset $S \subseteq [n]$ with $|S| \leqslant m$, we have*

$$\|v_S\|_p \leqslant \delta \cdot \|v\|_p.$$

A subspace $V \subseteq \mathbb{R}^n$ is said to be (m,δ)-ℓ_p-spread if every vector $v \in V$ is (m,δ)-ℓ_p-spread. A matrix is said to be (m,δ)-ℓ_p-spread if its column span $\operatorname{cspan}(X)$ is (m,δ)-ℓ_p-spread.

When the ambient dimension n is clear from the context, there are three parameters, i.e. p, m, δ, to be specified in Definition 13. If the value of p is not specified, set $p = 2$ by default. That is, (m,δ)-spreadness means (m,δ)-ℓ_2-spreadness. In certain cases (e.g. oblivious linear regression), we are more interested in capturing the dependence of m than the dependence of δ on other paramters (e.g. the ambient dimension n). Then it is more convenient to hide δ as long as it is $\Omega(1)$. Concretly, we say a vector, or a subspace, or a matrix is m-ℓ_p-spread if there exists a absolute constant $c \in (0,1)$ such that it is (m,c)-ℓ_p-spread.

We introduce the following definition that is closely related to spreadness and has important algorithmic implications in Appendix D.1.

12. We refer interested readers to Scarlett and Cevher (2019) for a proof of Eq. (A.1) as well as more applications of Fano's method.

Definition 14 (ℓ_p-vs-ℓ_q distortion) *Given $1 \leqslant p < q$, the ℓ_p-vs-ℓ_q distortion of a nonzero vector $v \in \mathbb{R}^n$ is defined by*

$$\Delta_{p,q}(v) := \frac{\|v\|_q}{\|v\|_p} \cdot n^{\frac{1}{p}-\frac{1}{q}}.$$

The ℓ_p-vs-ℓ_q distortion of a subspace $V \subseteq \mathbb{R}^n$ is defined by

$$\Delta_{p,q}(V) := \max_{v \in V, v \neq 0} \Delta_{p,q}(v).$$

The ℓ_p-vs-ℓ_q distortion of a matrix X is defined by

$$\Delta_{p,q}(X) := \Delta_{p,q}(\operatorname{cspan}(X)).$$

By Hölder's inequality and monotonicity of ℓ_p norm, it is easy to check $1 \leqslant \Delta_{p,q}(v) \leqslant n^{\frac{1}{p}-\frac{1}{q}}$ for any nonzero vector $v \in \mathbb{R}^n$. Note that for a nonzero vector $v \in \mathbb{R}^n$, $\Delta_{p,q}(v) = 1$ if and only if $|v_1| = \cdots = |v_n|$; $\Delta_{p,q}(v) = n^{\frac{1}{p}-\frac{1}{q}}$ if and only if $\|v\|_0 = 1$. Intuitively, low distortion implies well-spreadness, which is formalized in the following proposition.

Proposition 15 *Let $1 \leqslant p < q$ and V be a subspace of $\mathbb{R}^n$.*

1. *If $\Delta_{p,q}(V) \leqslant \Delta$, then V is (m, δ_p)-ℓ_p-spread with*

$$\delta_p = (m/n)^{\frac{1}{p}-\frac{1}{q}}\Delta.$$

2. *If V is not (m, δ)-ℓ_p-spread, then*

$$\Delta_{p,q}(V) > \delta(n/m)^{\frac{1}{p}-\frac{1}{q}}.$$

3. *If $\Delta_{p,q}(V) \leqslant \Delta$, then V is (m, δ_q)-ℓ_q-spread with*

$$\delta_q^q = 1 - \left(\Delta^{-p} - (m/n)^p\right)^{\frac{q}{p}}.$$

Proof

1. By Hölder's inequality, for any nonzero vector x,

$$|\operatorname{supp} x|^{\frac{1}{q}-\frac{1}{p}} \leqslant \frac{\|x\|_q}{\|x\|_p} \leqslant 1.$$

   For any nonzero vector $x \in V$ and any subset $S \subseteq [n]$ with $|S| \leqslant m$, we have

$$\|x_S\|_p \leqslant |S|^{\frac{1}{p}-\frac{1}{q}} \cdot \|x_S\|_q \leqslant |S|^{\frac{1}{p}-\frac{1}{q}} \cdot \|x\|_q \leqslant \Delta(m/n)^{\frac{1}{p}-\frac{1}{q}} \|x\|_p.$$

2. Since V is not (m, δ)-ℓ_p-spread, then by definition there exist $x \in V$ with $\|x\|_p = 1$ and $S \subset [n]$ with $|S| \leqslant m$ such that

$$\|x_S\|_p > \delta \|x\|_p = \delta.$$

Applying Hölder's inequality,

$$\|x\|_q \geqslant \|x_S\|_q \geqslant |S|^{\frac{1}{q}-\frac{1}{p}} \|x_S\|_p > \delta |S|^{\frac{1}{q}-\frac{1}{p}}.$$

Then,

$$\Delta_{p,q}(V) \geqslant \Delta_{p,q}(x) = n^{\frac{1}{p}-\frac{1}{q}} \frac{\|x\|_q}{\|x\|_p} > \delta (n/m)^{\frac{1}{p}-\frac{1}{q}}.$$

3. Fix an arbitrary vector $x \in V$ with $\|x\|_q = 1$. Then for any subset $S \subseteq [n]$ with $|S| \leqslant m$, we have

$$\|x_S\|_p \leqslant |S|^{\frac{1}{p}-\frac{1}{q}} \cdot \|x_S\|_q \leqslant m^{\frac{1}{p}-\frac{1}{q}}.$$

As $\|x\|_p \geqslant \Delta^{-1} n^{\frac{1}{p}-\frac{1}{q}}$, then

$$\|x_{\bar{S}}\|_p = \left(\|x\|_p^p - \|x_S\|_p^p\right)^{\frac{1}{p}} \geqslant \left(\Delta^{-p} - (m/n)^p\right)^{\frac{1}{p}} n^{\frac{1}{p}-\frac{1}{q}}.$$

Applying Hölder's inequality again,

$$\|x_{\bar{S}}\|_q \geqslant |\bar{S}|^{\frac{1}{q}-\frac{1}{p}} \|x_{\bar{S}}\|_p \geqslant \left(\Delta^{-p} - (m/n)^p\right)^{\frac{1}{p}}.$$

Thus,

$$\|x_S\|_q^q = \|x\|_q^q - \|x_{\bar{S}}\|_q^q \leqslant 1 - \left(\Delta^{-p} - (m/n)^p\right)^{\frac{q}{p}}.$$

■

In particular, given $1 \leqslant p < q$ and a subspace $V \subseteq \mathbb{R}^n$, if $\Delta_{p,q}(V) \leqslant O(1)$, then V is both $\Omega(n)$-ℓ_p-spread and $\Omega(n)$-ℓ_q-spread. On the other hand, if V is not $\Omega(n)$-ℓ_p-spread, then $\Delta_{p,q}(V) \geqslant \omega(1)$.

A.4. Low-degree likelihood ratio

To better understand the hardness result in Appendix D.2, we briefly introduce the *low-degree polynomial method* Hopkins (2018) that is developed for studying computational complexity of high-dimensional statistical inference problems. For further details about the low-degree polynomial method, we refer interested readers to Kunisky et al. (2019).

Consider in an asymptotic regime ($N \to \infty$) the hypothesis testing problem of distinguishing two sequences of hypotheses $\mu = \{\mu_N\}_{N\in\mathbb{N}}$ and $\nu = \{\nu_N\}_{N\in\mathbb{N}}$, where μ_N and ν_N are probability distributions over $\mathbb{R}^N$. We are interested in the case where ν, the *null distribution*, contains pure noise (e.g. $\nu_N = \mathcal{N}(0,1)^N$), and μ, the *planted distribution*, contains planted signal. A sequence of test functions $f = \{f_N\}_{N\in\mathbb{N}}$ with $f_N : \mathbb{R}^N \to \{0,1\}$ is said to *strongly distinguish* μ and ν if

$$\lim_{N\to\infty} \mathop{\mathbb{P}}_{\mu}(f_N(\boldsymbol{X}) = 1) = 1 \text{ and } \lim_{N\to\infty} \mathop{\mathbb{P}}_{\nu}(f_N(\boldsymbol{X}) = 0) = 1. \tag{A.2}$$

In other words, strong distinguishability means both type I and type II errors go to 0 as $N \to \infty$. We only consider the case where μ is absolutely continuous with respect to ν. The *likelihood ratio* defined by

$$L(X) := \frac{\mathrm{d}\,\mu}{\mathrm{d}\,\nu}(X)$$

is an optimal test function in the following sense.

Proposition 16 *Suppose μ is absolutely continuous with respect to ν. The unique solution of the optimization problem*

$$\max \mathop{\mathbb{E}}_{\mu}[f(\boldsymbol{X})] \quad \textit{subject to } \mathop{\mathbb{E}}_{\nu}\left[f(\boldsymbol{X})^2\right] = 1$$

is $L(X)/\sqrt{\mathbb{E}_\nu\left[L(\boldsymbol{X})^2\right]}$ and the value of the optimization problem is $\sqrt{\mathbb{E}_\nu\left[L(\boldsymbol{X})^2\right]}$.

Furthermore, classical decision theory tells us $\mathbb{E}_\nu\left[L(\boldsymbol{X})^2\right]$ characterizes strong distinguishability in the following way.

Proposition 17 *If $\mathbb{E}_\nu\left[L(\boldsymbol{X})^2\right]$ remains bounded as $N \to \infty$, then μ and ν is not strongly distinguishable in the sense of Eq. (A.2).*

One limitation of the above classical decision theory is that no computational-complexity considerations are involved. With the goal of studying whether a hypothesis testing problem is strongly distinguishable computation-efficiently, the low-degree polynomial method uses low-degree multivariate polynomials in the entries of $\boldsymbol{X}$ sampled from either μ or ν as a proxy for efficiently-computable functions.

Definition 18 (Low-degree likelihood ratio) *The degree-D likelihood ratio, denoted by $L^{\leqslant D}$, is the orthogonal projection[13] of the likelihood ratio $L = \mathrm{d}\,\mu/\,\mathrm{d}\,\nu$ onto the subspace of polynomials of degree at most D.*

We have the following low-degree analogue of Proposition 16.

Proposition 19 *Suppose μ is absolutely continuous with respect to ν. The unique solution of the optimization problem*

$$\max_{f \in \mathbb{R}[X]_{\leqslant D}} \mathop{\mathbb{E}}_{\mu}[f(\boldsymbol{X})] \quad \textit{subject to } \mathop{\mathbb{E}}_{\nu}\left[f(\boldsymbol{X})^2\right] = 1$$

is $L^{\leqslant D}(X)/\sqrt{\mathbb{E}_\nu\left[L^{\leqslant D}(\boldsymbol{X})^2\right]}$ and the value of the optimization problem is $\sqrt{\mathbb{E}_\nu\left[L^{\leqslant D}(\boldsymbol{X})^2\right]}$.

The following informal conjecture (Kunisky et al., 2019, Conjecture 1.16), which itself is based on (Hopkins, 2018, Conjecture 2.2.4), can be thought of as an computational analogue of Proposition 17.

Conjecture 20 (Informal) *For "sufficiently nice" sequences of probability distributions μ and ν, if there exists $\varepsilon > 0$ and $D = D(N) \geqslant (\log N)^{1+\varepsilon}$ for which $\mathbb{E}_\nu\left[L^{\leqslant D}(\boldsymbol{X})^2\right]$ remains bounded as $N \to \infty$, then there is no polynomial-time algorithm that strongly distinguishes μ and ν.*

Appendix B. Concentration bounds

Theorem 21 (Chernoff bound) *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be independent Bernoulli ranodm variables with parameter p. Then for any $t \in [0, np]$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \boldsymbol{X}_i - np\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{3np}\right).$$

13. We consider the Hilbert space endowed with inner product $\langle f, g\rangle := \mathbb{E}_\nu\left[f(\boldsymbol{X})g(\boldsymbol{X})\right]$.

Theorem 22 *Let $\boldsymbol{v} \sim \mathcal{N}(0, \mathrm{Id}_n)$. Then for any $t \geqslant 0$, one has*

$$\mathbb{P}\left(\|\boldsymbol{v}\|_2 \geqslant \sqrt{n} + t\right) \leqslant \exp(-t^2/2).$$

Theorem 23 *Let $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n \times d}$. Then for any $t \geqslant 0$, with probability at least $1 - 2\exp(-t^2/2)$,*

$$\sqrt{n} - \sqrt{d} - t \leqslant \sigma_{\min}(\boldsymbol{A}) \leqslant \sigma_{\max}(\boldsymbol{A}) \leqslant \sqrt{n} + \sqrt{d} + t.$$

Definition 24 (Sub-Gaussian norm) *The sub-Gaussian norm of a d-dimensional random vector $\boldsymbol{x}$ is defined by*

$$\|\boldsymbol{x}\|_{\psi_2} := \sup_{\substack{v \in \mathbb{R}^d \\ \|v\|_2 = 1}} \inf \left\{ t > 0 : \mathbb{E} \exp\left(\frac{\langle \boldsymbol{x}, v \rangle^2}{t^2} \right) \leqslant 2 \right\}.$$

Theorem 25 (Vershynin (2018), Theorem 4.6.1) *Let $\boldsymbol{A}$ be an $n \times d$ random matrix with independent rows $\boldsymbol{A}_1, ..., \boldsymbol{A}_n$. Suppose $\boldsymbol{A}_i$'s have zero mean, identity covariance matrix, and $K := \max_{i \in [n]} \|\boldsymbol{A}_i\|_{\psi_2} < \infty$ (see Definition 24). Then for any $t \geqslant 0$, with probability at least $1 - 2\exp(-t^2)$,*

$$\sqrt{n} - CK^2\left(\sqrt{d} + t\right) \leqslant \sigma_{\min}(\boldsymbol{A}) \leqslant \sigma_{\max}(\boldsymbol{A}) \leqslant \sqrt{n} + CK^2\left(\sqrt{d} + t\right),$$

where $C > 0$ is an absolute constant.

Theorem 26 (Well-spreadness of sub-Gaussian matrices) *Let $\boldsymbol{A}$ be an $n \times d$ random matrix with independent rows $\boldsymbol{A}_1, ..., \boldsymbol{A}_n$. Suppose $\boldsymbol{A}_i$'s have zero mean, identity covariance, and $K := \max_{i \in [n]} \|\boldsymbol{A}_i\|_{\psi_2} \leqslant O(1)$ (see Definition 24). Then there exist absolute constants $c_1, c_1, c_3 \in (0,1)$ such that $\boldsymbol{A}$ is $(c_1 n, c_2)$-spread with probability at least $1 - \exp(-\Omega(n))$ for $d \leqslant c_3 n$.*

Proof In the following, $c_1, c_2, c_3, c_4, c_5 \in (0,1)$ are sufficiently small constants that only depend on K and the absolute constant C in Theorem 25. Suppose $d \leqslant c_3 n$ and let $k = c_1 n$. We will show that, with high probability, for any nonzero $v \in \mathbb{R}^d$ and any $S \subset [n]$ with $|S| = k$, one has $\|\boldsymbol{A}_S v\|_2 \leqslant c_2 \|\boldsymbol{A} v\|_2$, where $\boldsymbol{A}_S$ is a $|S| \times d$ submatrix of $\boldsymbol{A}$ with rows indexed by S.

Fix a set $S \subset [n]$ with $|S| = k$. By Theorem 25, with probability at least $1 - 2\exp(-c_4 n)$,

$$\sigma_{\max}(\boldsymbol{A}_S) \leqslant \sqrt{k} + C'\left(\sqrt{d} + \sqrt{c_4 n}\right) \leqslant \left(\sqrt{c_1} + C'\left(\sqrt{c_3} + \sqrt{c_4}\right)\right)\sqrt{n},$$

where $C' = CK^2$ is a constant. Using $\binom{n}{k} \leqslant \left(\frac{en}{k}\right)^k$ and applying union bound, we have with probability at least $1 - 2\exp\left\{\left(-c_4 + c_1(1 - \log c_1)\right) n\right\}$ that,

$$\sigma_{\max}(\boldsymbol{A}_S) \leqslant \left(\sqrt{c_1} + C'\left(\sqrt{c_3} + \sqrt{c_4}\right)\right)\sqrt{n}$$

for any $S \subset [n]$ with $|S| = k$. Using Theorem 25 again, we have

$$\sigma_{\min}(\boldsymbol{A}) \geqslant \sqrt{n} - C'\left(\sqrt{d} + \sqrt{c_5 n}\right) \geqslant \left(1 - C'\left(\sqrt{c_3} + \sqrt{c_5}\right)\right)\sqrt{n}$$

with probability at least $1 - 2\exp(-c_5 n)$.

Given any constant $C' > 0$, we can always choose sufficiently small constants $c_1, c_2, c_3, c_4, c_5 \in (0,1)$ such that (i) $\frac{\sqrt{c_1}+C'(\sqrt{c_3}+\sqrt{c_4})}{1-C'(\sqrt{c_3}+\sqrt{c_5})} \leqslant c_2$ and (ii) $-c_4 + c_1(1-\log c_1) < 0$. Then with probability at least $1 - \exp(-\Omega(n))$, one has for any nonzero $v \in \mathbb{R}^d$ and any $S \subset [n]$ with $|S| = k$ that,

$$\frac{\|\boldsymbol{A}_S v\|_2}{\|\boldsymbol{A} v\|_2} \leqslant \frac{\sigma_{\max}(\boldsymbol{A}_S)}{\sigma_{\min}(\boldsymbol{A})} \leqslant c_2.$$

■

Remark 27 *For a random matrix with i.i.d. standard Gaussian or Rademacher random variables, it is easy to check $K \leqslant O(1)$.*

Appendix C. NP-hardness of deciding well-spreadness

We prove Theorem 29 that shows deciding whether a matrix satisfies a given well-spreadness condition is NP-hard. To cope with computational complexity issues with numbers, we will assume all input numbers to be rational. For a rational number $r \in \mathbb{Q}$, let $\langle r \rangle$ denote its encoding length, i.e. the length of its representation. For a rational matrix $A \in \mathbb{Q}^{n\times d}$, let $\langle A \rangle := \sum_{i=1}^{n}\sum_{j=1}^{d} \langle A_{ij} \rangle$ denote its encoding length.

Problem 28 *Given as input $A \in \mathbb{Q}^{n\times d}$, $m \in [n]$, and $\delta \in \mathbb{Q}$, decide whether A is (m,δ)-spread.*

Theorem 29 *Problem 28 is NP-hard.*

To prove Theorem 29, we will show the following problem is NP-hard and there exists a polynomial-time reduction from this problem to Problem 28.

Problem 30 *Given as input $A \in \mathbb{Q}^{p\times n}$, $m \in [n]$, and $\delta \in \mathbb{Q}$, decide whether $\ker(A)$ is (m,δ)-spread.*

Following Bandeira et al. (2013); Tillmann and Pfetsch (2014), our proof of the NP-hardness of Problem 30 is based on a reduction from the problem of deciding *matrix spark* (Problem 32).

Definition 31 (Matrix spark) *The spark of a matrix A is the smallest number k such that there exists a set of k columns of A that are linearly dependent. Equivalently,*

$$\operatorname{spark}(A) := \min\{\|x\|_0 : Ax = 0, x \neq 0\}.$$

Problem 32 *Given as input $A \in \mathbb{Q}^{p\times n}$ and $m \in \mathbb{N}$, decide whether $\operatorname{spark}(A) > m$.*

By a reduction from the NP-complete k-clique problem, i.e. deciding whether a given simple graph has a clique of size k, Problem 32 is proven to be NP-hard in Tillmann and Pfetsch (2014). Moreover, the matrices in the hard instances of Problem 32 are integer matrices whose entry-wise encoding length is bounded by a polynomial in p and n.

Theorem 33 *Problem 30 is NP-hard.*

Proof Let (A, m) be a hard instance of Problem 32 given by Tillmann and Pfetsch (2014). Let $P = \|A\|_\infty$. It is known that $\langle P \rangle$ is bounded by some polynomial in p and n. Our strategy is to choose an appropriate rational number $\delta \in (0, 1)$ with $\langle \delta \rangle$ bounded by some polynomial in p and n such that the following is true. When we give the instance (A, m, δ) to an oracle of Problem 30, if the answer is YES, then $\operatorname{spark}(A) > m$; if the answer is NO, then $\operatorname{spark}(A) \leqslant m$. If such a δ exists, then we have a polynomial-time reduction from Problem 32 to Problem 30, and as a result, Problem 30 is NP-hard.

In the following, we show how to construct such a $\delta \in (0, 1)$. For the case when the oracle answers YES, it is straightforward to see $\operatorname{spark}(A) > m$ for any $\delta \in (0, 1)$. For the case when the oracle answers NO, we consider the contrapositive. Assume $\operatorname{spark}(A) > m$. We want a $\delta \in (0, 1)$ such that $\|v_S\|_2 \leqslant \delta \|v\|_2$ for any nonzero $v \in \ker(A)$ and any $S \subseteq [n]$ with $|S| \leqslant m$.

Take an arbitrary nonzero vector $x \in \ker(A)$. Without loss of generality, assume $|x_1| \geqslant |x_2| \geqslant \cdots \geqslant |x_n|$. Let $S = [m]$ and $\bar{S} = [n] \setminus S$. Then it suffices to upper bound $\|x_S\|_2 / \|x\|_2$ by δ. Let A_S be the $m \times k$ submatrix of A with columns indexed by S and define $A_{\bar{S}}$ likewise. Since $x \in \ker(A)$, we have

$$\begin{aligned} Ax = 0 &\iff A_S x_S + A_{\bar{S}} x_{\bar{S}} = 0 \\ &\implies \|A_S x_S\|_2 = \|A_{\bar{S}} x_{\bar{S}}\|_2 \\ &\implies \frac{\|x_S\|_2}{\|x_{\bar{S}}\|_2} \leqslant \frac{\sigma_{\max}(A_{\bar{S}})}{\sigma_{\min}(A_S)}. \end{aligned}$$

It is easy to see

$$\sigma_{\max}(A_{\bar{S}}) \leqslant \|A_{\bar{S}}\|_F \leqslant \|A\|_F \leqslant \sqrt{pn} \cdot P.$$

From the proof of (Bandeira et al., 2013, Theorem 4), we know

$$\sigma_{\min}(A_S)^2 \geqslant \left(pmP^2\right)^{1-m} \geqslant \left(pnP^2\right)^{1-p}.$$

Therefore,

$$\frac{\|x_S\|_2}{\|x\|_2} \leqslant \sqrt{\frac{1}{1 + \sigma_{\min}(A_S)^2 / \sigma_{\max}(A_{\bar{S}})^2}} \leqslant \sqrt{\frac{1}{1 + (pnP^2)^{-p}}} \leqslant 1 - \frac{1}{2\left((pnP^2)^p + 1\right)}.$$

Set $\delta = 1 - \frac{1}{2((pnP^2)^p + 1)}$. Then $\ker(A)$ is (m, δ)-spread. Moreover, $\langle \delta \rangle \leqslant f(p, n)$ for some polynomial f. ■

Now we are ready to prove Theorem 29.

Proof Given Theorem 33, it only remains to show there exists a polynomial-time reduction from Problem 30 to Problem 28. It is well-known that, given as input a matrix $X \in \mathbb{Q}^{p \times n}$, Gaussian elimination is able to produce in polynomial time a matrix $Y \in \mathbb{Q}^{n \times (n-p)}$ such that $\ker(X) = \operatorname{cspan}(Y)$ and $\langle Y \rangle$ is polynomial in $\langle X \rangle$. ■

Theorem 29 establishes the NP-hardness of deciding whether a given matrix is (m, δ)-spreadness when m and δ are also inputs. Nevertheless, this result has a major limitation in the conext of oblivious regression. That is, the parameter $\delta \in (0, 1)$ used in the above proof is $1 - o(1)$ that this result reveals almost nothing about the hardness of the more interesting case when δ is a constant.

Appendix D. Computational aspects of certifying well-spreadness

In this section we prove Theorem 4 and Theorem 5. Concretly, in Appendix D.1, we provide an efficient algorithm, based on known sum-of-squares algorithms, that can certify an $n \times d$ Gaussian matrices is $\Omega(n)$-spread when $n \gtrsim d^2$. On the other hand, in Appendix D.2, we provide strong evidence, based on the low-degree polynomial method, which suggests no polynomial-time algorithm is able to certify an $n \times d$ Gaussian matrix is $\Omega(n)$-spread when $n \ll d^2$.

D.1. Algorithms for certifiying well-spreadness

We prove Theorem 34 that shows we can efficiently certify an $n \times d$ Gaussian matrix is $\Omega\left(n\right)$-spread with high probability whenever $n \gtrsim d^2$. We consider the regime where d is growing.

Theorem 34 *Let $\delta \in (0,1)$ and $C > 0$ be arbitrary constants. Let $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n \times d}$ with $n \geqslant Cd^2$. There exists a polynomial-time algorithm based on sum-of-squares relaxation and a constant $C' = C'(C)$ such that*

1. *if $\boldsymbol{A}$ is not $\left((\delta/C')^4 n, \delta\right)$-spread, the algotithm outputs NO;*
2. *with high probability, $\boldsymbol{A}$ is $\left((\delta/C')^4 n, \delta\right)$-spread and the algotithm outputs YES.*

Theorem 4 is a direct application of Theorem 34 to oblivious linear regression with Gaussian design.

To prove Theorem 34, we make use of the following result which shows that, with high probability, the 2-to-4 norm[14] of an $n \times d$ Gaussian matrix can be efficiently upper bounded by $O\left(n^{1/4}\right)$, given $n \gtrsim d^2$.

Theorem 35 (Barak et al. (2012), Theorem 7.1) *Let $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n \times d}$. There exists a polynomial-time algorithm based on sum-of-squares relaxation that outputs an upper bound $\mathfrak{U}$ of the 2-to-4 norm of $\boldsymbol{A}$, i.e. $\max_{\|u\|_2 = 1} \|\boldsymbol{A}u\|_4$, which satisfies*

$$\mathfrak{U} \leqslant n^{1/4} \left(3 + c \cdot \max\left(\frac{d}{\sqrt{n}}, \frac{d^2}{n}\right)\right)^{1/4}$$

with high probability. Here, $c > 0$ is an absolute constant.

Then it is straightforward to show that, with high probability the ℓ_2-vs-ℓ_4 distortion of an $n \times d$ Gaussian matrix can be efficiently upper bounded by $O(1)$ given $n \gtrsim d^2$, which we formalize in the following corollary.

Corollary 36 *Let $C > 0$ be an arbitrary constant. Let $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n \times d}$ with $n \geqslant Cd^2$. There exists a polynomial-time algorithm based on sum-of-squares relaxation that outputs an upper bound $\mathfrak{U}'$ of the ℓ_2-vs-ℓ_4 distortion of $\boldsymbol{A}$, i.e. $\max\left\{n^{1/4} \cdot \|v\|_4 / \|v\|_2 : v \in \operatorname{cspan}(A), v \neq 0\right\}$, which satisfies $\mathfrak{U}' \leqslant C'$ with high probability. Here $C' > 1$ is a constant only depending on C.*

14. The p-to-q norm of a matrix X is defined by $\|X\|_{p \to q} := \max_{u \neq 0} \|Xu\|_q / \|u\|_p$.

Proof For any non-singular matrix $X \in \mathbb{R}^{n\times d}$, one has

$$\begin{aligned}\Delta_{2,4}(X) &= n^{\frac{1}{4}} \max_{u\neq 0} \frac{\|Xu\|_4}{\|Xu\|_2} = n^{\frac{1}{4}} \max_{u\neq 0} \frac{\|Xu\|_4 / \|u\|_2}{\|Xu\|_2 / \|u\|_2} \\ &\leqslant n^{\frac{1}{4}} \frac{\max_{u\neq 0} \|Xu\|_4 / \|u\|_2}{\min_{u\neq 0} \|Xu\|_2 / \|u\|_2} \\ &= \frac{n^{\frac{1}{4}}}{\sigma_{\min}(X)} \cdot \max_{\|u\|_2=1} \|Xu\|_4 .\end{aligned}$$

Now consider $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n\times d}$ which is non-singular almost surely as long as $n \geqslant d$. By Theorem 23, for any $n \gg d$, one has $\sigma_{\min}(\boldsymbol{A}) = (1-o(1))\sqrt{n}$ with high probability. And singular values can be efficiently computed. By Theorem 35, there is an efficiently-computable upper bound $\mathfrak{U}$ of $\max_{\|u\|_2=1} \|\boldsymbol{A}u\|_4$ that satisfies $\mathfrak{U} \leqslant C'' n^{1/4}$ with high probability. Here C'' only depends on C.

Therefore, there exist a constant C' only depending on C and an efficiently-computable upper bound $\mathfrak{U}'$ of $\Delta_{2,4}(\boldsymbol{A})$ such that $\mathfrak{U}' \leqslant C'$ with high probability. ■

Now, we combine Corollary 36 and Proposition 15 to prove Theorem 34.

Proof We first describe the algorithm $\mathcal{A}$. Given an input $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n\times d}$, we use the efficient algorithm given by Corollary 36 to compute an upper bound $\mathfrak{U}'$ of the ℓ_2-vs-ℓ_4 distortion $\Delta_{2,4}(\boldsymbol{A})$. Let $C' = C'(C)$ be the constant given by Corollary 36. If $\mathfrak{U}' \leqslant C'$, algorithm $\mathcal{A}$ outputs YES. Otherwise, algorithm $\mathcal{A}$ outputs NO.

Then we show algorithm $\mathcal{A}$ satisfies the two requirements. Instantiate Proposition 15 with $p = 2, q = 4$ and let $\boldsymbol{\Delta} = \Delta_{2,4}(\boldsymbol{A})$. Then $\boldsymbol{A}$ is $\big(\delta^4 \boldsymbol{\Delta}^{-4} n, \delta\big)$-spread for any $\delta \in (0,1)$. By contrapositivity, if $\boldsymbol{A}$ is not $\big((\delta/C')^4 n, \delta\big)$-spread, then $\mathfrak{U}' > C'$ and algorithm $\mathcal{A}$ will output NO. By Corollary 36, $\mathfrak{U}' \leqslant C'$ with high probability. Thus, with high probability, $\boldsymbol{A}$ is $\big((\delta/C')^4 n, \delta\big)$-spread and algorithm $\mathcal{A}$ outputs YES. ■

D.2. Hardness of certifiying well-spreadness

We provide here formal evidence suggesting the computational hardness of certifiying well-spreadness in average case. We consider the regime where d is growing and $n \gtrsim d$.

To state our hardness result, we first introduce the noisy Bernoulli-Rademacher distribution (over $\mathbb{R}$) and a distinguishing problem.

Definition 37 (Noisy Bernoulli-Rademacher distribution) *A random variable $\boldsymbol{x}$ following noisy Bernoulli-Rademacher distribution with parameter $\rho \in (0,1)$ and $\sigma \in [0, 1/\sqrt{1-\rho})$, denoted by $\boldsymbol{x} \sim \mathrm{nBR}(\rho,\sigma)$, is defined by*

$$\boldsymbol{x} = \begin{cases} \mathcal{N}\left(0,\sigma^2\right), & \text{with probability } 1-\rho, \\ +\frac{1}{\sqrt{\rho'}}, & \text{with probability } \frac{\rho}{2}, \\ -\frac{1}{\sqrt{\rho'}}, & \text{with probability } \frac{\rho}{2}, \end{cases}$$

where $\rho' = \frac{\rho}{1-(1-\rho)\sigma^2}$.

We remark that the particular choice of ρ' in the above definition is to make $\mathbb{E}\,\boldsymbol{x}^2 = 1$ for $\boldsymbol{x} \sim \mathrm{nBR}\,(\rho, \sigma)$.

Problem 38 (Distinguishing) *Let $n, d \in \mathbb{N}$, $\rho \in (0,1)$, and $\sigma \in [0, 1/\sqrt{1-\rho})$.*

- *Under the null distribution ν, observe $\boldsymbol{A} \sim \mathcal{N}\,(0,1)^{n\times d}$.*
- *Under the planted distribution μ, first sample a hidden vector $\boldsymbol{v}$ whose entries are i.i.d. noisy Bernoulli-Rademacher random variables with parameter (ρ, σ). Let $\boldsymbol{Y}$ be an $n \times d$ matrix of which the first column is $\boldsymbol{v}$ and the rest entries are independent $\mathcal{N}\,(0,1)$. Then sample a random orthogonal matrix $\boldsymbol{Q}$ and observe $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Q}$.*

Given a sample $\boldsymbol{A}$ from either ν or μ, decide from which distribution $\boldsymbol{A}$ is sampled.

Now we state our computational hardness result.

Theorem 39 *Let ν and μ be the null an planted distributions defined in Problem 38 respectively. Let $C > 1$ be an arbitrary constant. There exist absolute constants $c_1, c_2, c_3 \in (0,1)$ and $C_4 > 1$ such that the following holds. For any $\rho \gg \frac{1}{n}$, $\sigma^2 \leqslant \frac{1}{2}\,(\log n)^{-C}$, $d \in \left(C_4\rho^{-1}\sqrt{n}\,(\log n)^{2C}\,, c_3 n\right)$, $m \in (1.5\rho n, c_1 n)$, constant $\delta \in (c_2, 1)$, and $D \leqslant (\log n)^C$, one has*

1. *$\boldsymbol{A} \sim \nu$ is (m,δ)-spread with high probability;*
2. *$\boldsymbol{A} \sim \mu$ is not (m,δ)-spread with high probability;*
3. *$\mathbb{E}_\nu\left[L^{\leqslant D}(\boldsymbol{A})^2\right] \leqslant O(1)$ where $L^{\leqslant D}$ is the degree-D likelihood ratio defined in Definition 18.*

Proof By Lemma 40 and Lemma 41. ■

Implications of Theorem 39 Before proving Theorem 39, we discuss some of its implications. First we set $C = 2$, $\rho = 1/\log n$, and $m = 2n/\log n$ in Theorem 39. It follows that, in the regime where $\sqrt{n} \ll d \lesssim n$, we have (i) $\boldsymbol{A} \sim \nu$, i.e. $\boldsymbol{A} \sim \mathcal{N}\,(0,1)^{n\times d}$, is $\Omega\,(n)$-spread with high probability; (ii) $\boldsymbol{A} \sim \mu$ is $o(n)$-spread with high probability; and (iii) it is very likely that no polynomial-time algorithm can distinguish ν and μ, based on the discussion of low-degree polynomial method in Appendix A.4.

Then we apply Theorem 39 to oblivious linear regression with Gaussian design matrix and thus prove Theorem 5. By (d'Orsi et al., 2021b, Theorem 1.2), the sufficent conditions for consistent oblivious regression are (i) $n \gg \frac{d}{\alpha^2}$ and (ii) the design matrix is $\Omega\left(\frac{d}{\alpha^2}\right)$-spread. In the following, we will characterize a regime over (n, d, α) where (i) $n \gg \frac{d}{\alpha^2}$; (ii) $\boldsymbol{A} \sim \mathcal{N}\,(0,1)^{n\times d}$ is $\Omega\left(\frac{d}{\alpha^2}\right)$-spread with high probability; and (iii) there exists strong evidence suggesting certifiying $\Omega\left(\frac{d}{\alpha^2}\right)$-spreadness of $\boldsymbol{A}$ is computationally difficult. To this end, let $n \gg \frac{d}{\alpha^2}$ and fix two arbitrary constants $C > 0$ and $\delta \in (0,1)$. Let ν and μ be the null and planted distributions considered in Theorem 39. It is not difficult to see from the proof of Theorem 26 that $\boldsymbol{A} \sim \nu$ is $\left(C\frac{d}{\alpha^2}, \delta\right)$-spread with high probability given $n \gg \frac{d}{\alpha^2}$. From the proof of Lemma 40 we know, if $\rho \gg \frac{1}{n}$, $\sigma = o(1)$, and $\rho n \lesssim \frac{d}{\alpha^2}$, then $\boldsymbol{A} \sim \mu$ is not $\left(C\frac{d}{\alpha^2}, \delta\right)$-spread with high probability. Set $C = 2$ in Lemma 41 and

we have the following: if $\sigma^2 \leqslant \frac{1}{2}(\log n)^{-2}$, $\rho^{-1}n^{1/2}(\log n)^4 \lesssim d$, then $\mathbb{E}_\nu\left[L^{\leqslant D}(\boldsymbol{A})^2\right] \leqslant O(1)$ for any $D \leqslant (\log n)^2$. Therefore, such a regime over (n, d, α) can be characterized by

$$\left\{(n,d,\alpha) : \exists \rho \text{ such that } n \gg \frac{d}{\alpha^2}, \rho \gg \frac{1}{n}, \rho n \lesssim \frac{d}{\alpha^2}, \rho^{-1}n^{1/2}\operatorname{polylog}(n) \lesssim d\right\},$$

or equivalently,

$$\left\{(n,d,\alpha) : n^{3/4}\operatorname{polylog}(n)\alpha \lesssim d \ll n\alpha^2\right\}.$$

Finally, we remark that the "noiseless" Bernoulli-Rademacher distribution (i.e. $\sigma = 0$) already appeared in the literaure (e.g. d'Orsi et al. (2020); Mao and Wein (2021)). In the "noiseless" setting, Problem 38 can be efficiently solved even when n is only linear in d Zadik et al. (2021). Although the algorithm proposed in Zadik et al. (2021) surpasses the lower bound for low-degree polynomial method, their algorithm relies heavily on the extact and brittle structure of the hidden vector. If we add a little noise to the hidden vector, like what we did here, then their algorithm is likely to fail.

Proof of Theorem 39 The following two lemmas together directly imply Theorem 39.

Lemma 40 *Let ν and μ be the null and planted distributions defined in Problem 38 respectively. There exist absolute constants $c_1, c_2, c_3 \in (0,1)$ such that the following holds. For any $\rho \gg \frac{1}{n}$, $\sigma = o(1)$, $d \leqslant c_3 n$, $m \in (1.5\rho n, c_1 n)$, and constant $\delta \in (c_2, 1)$, one has*

1. *$\boldsymbol{A} \sim \nu$ is (m,δ)-spread with high probability;*
2. *$\boldsymbol{A} \sim \mu$ is not (m,δ)-spread with high probability.*

Proof The existence of absolute constants c_1, c_2, c_3 is guaranteed by Theorem 26. That is, if $\boldsymbol{A} \sim \mathcal{N}(0,1)^{n\times d}$ and $d \leqslant c_3 n$, then $\boldsymbol{A}$ is $(c_1 n, c_2)$-spread with high probability. Observe that (m_1,δ_1)-spreadness implies (m_2,δ_2)-spreadness for any $m_2 \leqslant m_1$ and $\delta_2 \geqslant \delta_1$. Thus for any $m \leqslant c_1 n$ and $\delta \geqslant c_2$, $\boldsymbol{A} \sim \nu$ is (m,δ)-spread with high probability.

Now consider $\boldsymbol{A} \sim \mu$ and let $\boldsymbol{v}$ be the hidden vector of μ. Clearly, $\boldsymbol{v} \in \operatorname{cspan}(\boldsymbol{A})$. We decompose $\boldsymbol{v}$ into two parts with disjoint supports, $\boldsymbol{v} = \boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{b}$ is the Bernoulli-Rademacher part and $\boldsymbol{\varepsilon}$ is the Gaussian part. Let $S = \operatorname{supp}\boldsymbol{b}$. Then,

$$\frac{\|\boldsymbol{v}_S\|_2}{\|\boldsymbol{v}\|_2} = \frac{\|\boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2 + \|\boldsymbol{\varepsilon}\|_2}.$$

By Theorem 21,

$$\mathbb{P}\left(0.5\rho n \leqslant \|\boldsymbol{b}\|_0 \leqslant 1.5\rho n\right) \geqslant 1 - 2\exp\left(-\frac{\rho n}{12}\right).$$

By Theorem 22,

$$\mathbb{P}\left(\|\boldsymbol{\varepsilon}\|_2 \geqslant 2\sigma\sqrt{n}\right) \leqslant \exp\left(-\frac{n}{2}\right).$$

If $\rho \gg 1/n$ and $\sigma = o(1)$, then with high probability, we have $|S| \leqslant 1.5\rho n$ and

$$\frac{\|\boldsymbol{v}_S\|_2}{\|\boldsymbol{v}\|_2} \geqslant \frac{1}{1+4\sigma} = 1 - o(1).$$

Thus for any $m \geqslant 1.5\rho n$ and any constant $\delta < 1$, $\boldsymbol{A}$ is not (m,δ)-spread with high probability. ■

Lemma 41 *Let ν and μ be the null and planted distributions defined in Problem 38 respectively. Let $C > 1$ be an arbitrary constant. For any $D \leqslant (\log n)^C$, $\sigma^2 \leqslant \frac{1}{2}(\log n)^{-C}$, and $d \geqslant C_4 \rho^{-1}\sqrt{n}(\log n)^{2C}$, one has*

$$\mathop{\mathbb{E}}_{\nu}\left[L^{\leqslant D}(\boldsymbol{A})^2\right] \leqslant O(1),$$

where $C_4 > 1$ is an absolute constant and $L^{\leqslant D}$ is the degree-D likelihood ratio defined in Definition 18.

The proof of Lemma 41 is an adaptation of the proof of (Mao and Wein, 2021, Theorem 3.4)[15] which we include here for completeness. The proof relies on the following three lemmas.

Lemma 42 (Mao and Wein (2021), Lemma 4.23) *Let ν and μ be the null and planted distributions defined in Problem 38 respectively. Let $\boldsymbol{u}, \boldsymbol{u}'$ be independent uniformly random vectors on the unit sphere in $\mathbb{R}^d$ and $\boldsymbol{x} \sim \mathrm{nBR}(\rho, \sigma)$. Then,*

$$\mathop{\mathbb{E}}_{\nu}\left[L^{\leqslant D}(\boldsymbol{A})^2\right] = \sum_{k=0}^{D} \mathbb{E}\left\langle \boldsymbol{u}, \boldsymbol{u}'\right\rangle^k \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}\, h_{\alpha_i}(\boldsymbol{x})\right)^2,$$

where $h_k : \mathbb{R} \to \mathbb{R}$ is the k-th normalized Hermite polynomial and where $L^{\leqslant D}$ is the degree-D likelihood ratio defined in Definition 18.

Lemma 43 (Mao and Wein (2021), Lemma 4.25) *Let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be independent uniformly random vectors on the unit sphere in $\mathbb{R}^d$. For odd $k \in \mathbb{N}$, $\mathbb{E}\left\langle \boldsymbol{u}, \boldsymbol{u}'\right\rangle^k = 0$. For even $k \in \mathbb{N}$,*

$$\mathbb{E}\left\langle \boldsymbol{u}, \boldsymbol{u}'\right\rangle^k \leqslant (k/d)^{k/2}.$$

Lemma 44 (Adapted from Mao and Wein (2021), Lemma 4.26) *For a noisy Bernoulli-Rademacher random variable $\boldsymbol{x} \sim \mathrm{nBR}(\rho, \sigma)$, we have*

1. *$\mathbb{E}\, h_k(\boldsymbol{x}) = 0$ for odd $k \in \mathbb{N}$;*
2. *$\mathbb{E}\, h_0(\boldsymbol{x}) = 1$;*
3. *$\mathbb{E}\, h_2(\boldsymbol{x}) = 0$;*
4. *$\left(\mathbb{E}\, h_k(\boldsymbol{x})\right)^2 \leqslant 8^k \rho^{2-k}$ for $k \geqslant 4$ and $\sigma^2 \leqslant \frac{1}{k-1}$.*

Proof Since the noisy Bernoulli-Rademacher distribution is symmetric and odd-degree Hermite polynomials are odd functions, one has $\mathbb{E}\, h_k(\boldsymbol{x}) = 0$ for odd $k \in \mathbb{N}$. It is straightforward to check by definition that

$$\mathbb{E}\, h_0(\boldsymbol{x}) = 1, \quad \mathbb{E}\, h_2(\boldsymbol{x}) = \frac{1}{\sqrt{2}}\mathbb{E}\left[\boldsymbol{x}^2 - 1\right] = 0.$$

Fix an even integer $k \geqslant 4$ and let $\sigma^2 \leqslant \frac{1}{k-1}$. Then for any even integer $r \in [k]$,

$$\mathbb{E}\, \boldsymbol{x}^r = (1-\rho)\sigma^r (r-1)!! + \rho(\rho')^{-r/2} \leqslant \sigma^r (r-1)!! + \rho^{1-r/2} \leqslant \sigma^2 + \rho^{1-k/2} \leqslant 2\rho^{1-k/2}.$$

15. A related result was previously shown in (d'Orsi et al., 2020, Theorem 6.7).

Also, $\mathbb{E}\,\boldsymbol{x}^0 = 1 \leqslant 2\rho^{1-k/2}$. Let c_r be the coefficient of z^r in the polynomial $\sqrt{k!}\cdot h_k(z)$. Then,

$$|\mathbb{E}\,h_k(\boldsymbol{x})| = \frac{1}{\sqrt{k!}}\left|\sum_{r=0}^{k} c_r\,\mathbb{E}\,\boldsymbol{x}^r\right| \leqslant \frac{2\rho^{1-k/2}}{\sqrt{k!}}\sum_{r=0}^{k}|c_r|\,.$$

Define $T(k) := \sum_{r=0}^{k}|c_r|$. Note that $T(k)$ is the k-th telephone number which satisfies the following recurrence,

$$T(n) = T(n-1) + (n-1)\cdot T(n-2) \quad \forall n \geqslant 2,$$

and $T(0) = T(1) = 1$. It is easy to show by induction that,

$$T(n) \leqslant C^n n^{n/2}, \quad \forall n \geqslant 1, \quad \forall C \geqslant \frac{1+\sqrt{5}}{2}.$$

Now fix some $C \geqslant \frac{1+\sqrt{5}}{2}$. Using Stirling's approximation, $n! \geqslant \sqrt{2\pi n}(n/e)^n$ for any $n \geqslant 1$, we have

$$(\mathbb{E}\,h_k(\boldsymbol{x}))^2 \leqslant 4\rho^{2-k}\cdot\frac{T(k)^2}{k!} \leqslant \frac{4}{\sqrt{2\pi k}}\left(C^2 e\right)^k\rho^{2-k}.$$

Therefore, for $k \geqslant 4$ and $\sigma^2 \leqslant \frac{1}{k-1}$, we have

$$(\mathbb{E}\,h_k(\boldsymbol{x}))^2 \leqslant 8^k\rho^{2-k}.$$

■

Now we are ready to prove Lemma 41.

Proof Let $\boldsymbol{x} \sim \mathrm{nBR}\,(\rho,\sigma)$ be a noisy Bernoulli-Rademacher random variable. Given $\alpha \in \mathbb{N}^n$, if there exists $i \in [n]$ such that α_i is odd or $\alpha_i = 2$, then $\mathbb{E}\,h_{\alpha_i}(\boldsymbol{x}) = 0$ by Lemma 44. Thus, we define the following set

$$S(k,m) := \{\alpha \in \mathbb{N}^n : |\alpha| = k, \|\alpha\|_0 = m, \alpha_i \in \{0\}\cup\{4,6,8,...\} \text{ for all } i \in [n]\}\,.$$

As $\sigma^2 \leqslant \frac{1}{2}(\log n)^{-C}$ and $D \leqslant (\log n)^C$, we have $\sigma^2 \leqslant 1/(k-1)$ for any $k \leqslant D$. Using Lemma 44, for $\alpha \in S(k,m)$, we have

$$\prod_{i=1}^{n}(\mathbb{E}\,h_{\alpha_i}(\boldsymbol{x}))^2 \leqslant \prod_{\alpha_i\neq 0} 8^{\alpha_i}\rho^{2-\alpha_i} \leqslant 8^k\rho^{2m-k}.$$

Note that $S(k,m)$ is empty if $m > \lfloor k/4\rfloor$. And it is easy to see

$$|S(k,m)| \leqslant \binom{n}{m}m^{k/2} \leqslant n^m\,(k/4)^{k/2}\,.$$

Then, for $k \geqslant 4$, we have

$$\sum_{\substack{\alpha\in\mathbb{N}^n\\|\alpha|=k}}\prod_{i=1}^{n}(\mathbb{E}\,h_{\alpha_i}(\boldsymbol{x}))^2 = \sum_{m=1}^{\lfloor k/4\rfloor}\sum_{\alpha\in S(k,m)}\prod_{i=1}^{n}(\mathbb{E}\,h_{\alpha_i}(\boldsymbol{x}))^2 \leqslant \sum_{m=1}^{\lfloor k/4\rfloor} n^m\,(k/4)^{k/2}\,8^k\rho^{2m-k}$$

$$= (k/4)^{k/2}\, 8^k \rho^{-k} \frac{\left(n\rho^2\right)^{\lfloor k/4 \rfloor + 1} - n\rho^2}{n\rho^2 - 1} \leqslant (k/4)^{k/2}\, 8^k \rho^{-k} \frac{\left(n\rho^2\right)^{k/4+1}}{n\rho^2/2}$$
$$\leqslant 2 \cdot k^{k/2} n^{k/4} \rho^{-k/2} 4^k.$$

Let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be independent uniformly random vectors on the unit sphere in $\mathbb{R}^d$. Using Lemma 43, for $k \geqslant 4$, we have

$$\mathbb{E} \left\langle \boldsymbol{u}, \boldsymbol{u}' \right\rangle^k \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}\, h_{\alpha_i}(\boldsymbol{x})\right)^2 \leqslant (k/d)^{k/2} \cdot 2 \cdot k^{k/2} n^{k/4} \rho^{-k/2} 4^k = \left(\frac{512 k^4 n}{d^2 \rho^2}\right)^{k/4}.$$

Finally, by Lemma 42, we have

$$\mathop{\mathbb{E}}_{\nu} \left[L^{\leqslant D}(\boldsymbol{A})^2\right] = \sum_{k=0}^{D} \mathbb{E} \left\langle \boldsymbol{u}, \boldsymbol{u}' \right\rangle^k \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}\, h_{\alpha_i}(\boldsymbol{x})\right)^2 = 1 + \sum_{k \geqslant 4}^{D} \mathbb{E} \left\langle \boldsymbol{u}, \boldsymbol{u}' \right\rangle^k \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = k}} \prod_{i=1}^{n} \left(\mathbb{E}\, h_{\alpha_i}(\boldsymbol{x})\right)^2$$
$$\leqslant 1 + \sum_{k \geqslant 4}^{D} \left(\frac{512 k^4 n}{d^2 \rho^2}\right)^{k/4} \leqslant 1 + \sum_{k \geqslant 4}^{\infty} \left(\frac{512 n \left(\log n\right)^{4C}}{d^2 \rho^2}\right)^{k/4}.$$

If there exists a constant $c \in (0, 1)$ such that $\frac{512 n (\log n)^{4C}}{d^2 \rho^2} \leqslant c$, i.e. $d \geqslant \sqrt{512/c} \sqrt{n} \left(\log n\right)^{2C}$, then we have

$$\mathop{\mathbb{E}}_{\nu} \left[L^{\leqslant D}(\boldsymbol{A})^2\right] \leqslant O(1).$$

■