



**HAL**  
open science

# Concentration of the Frobenius norms of generalized matrix inverses

Ivan Dokmanić, Rémi Gribonval

► **To cite this version:**

Ivan Dokmanić, Rémi Gribonval. Concentration of the Frobenius norms of generalized matrix inverses. 2018. hal-01897046v1

**HAL Id: hal-01897046**

**<https://inria.hal.science/hal-01897046v1>**

Preprint submitted on 17 Oct 2018 (v1), last revised 22 Nov 2018 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# CONCENTRATION OF THE FROBENIUS NORMS OF PSEUDOINVERSES

IVAN DOKMANIĆ AND RÉMI GRIBONVAL

**Abstract.** In many applications it is useful to replace the Moore-Penrose pseudoinverse (MPP) by another generalized inverse with more favorable properties. We may want, for example, to have many zero entries, but without giving up too much of the stability of the MPP. One way to quantify stability is to look at how much the Frobenius norm of a generalized inverse exceeds that of the MPP. In this paper we derive finite-size concentration bounds for the Frobenius norm of  $\ell^p$ -minimal general inverses of iid Gaussian matrices, with  $1 \leq p \leq 2$ . For  $p = 1$  we prove exponential concentration of the Frobenius norm of the sparse pseudoinverse; for  $p = 2$ , we get similar results for the MPP. Our proof is based on the convex Gaussian min-max theorem, but unlike previous applications we give finite-size concentration results which required development of new techniques.

**1. Introduction.** Generalized inverses are matrices that have some properties of the usual inverse of a regular square matrix. We call<sup>1</sup>  $\mathbf{X} \in \mathbb{R}^{n \times m}$  a generalized inverse of a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  if  $\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}$  and write  $\mathcal{G}(\mathbf{A})$  for the set of all such matrices  $\mathbf{X}$ . In inverse problems, it is desirable to use generalized inverses with a small Frobenius norm because the Frobenius norm controls the output mean-squared error. In this paper, we study Frobenius norms of generalized inverses that are obtained by constrained minimization of a class of matrix norms. Concretely, we look at entrywise  $\ell^p$  norms for  $1 \leq p \leq 2$ , including the Moore-Penrose pseudoinverse (MPP) for  $p = 2$  and the *sparse pseudoinverse* for  $p = 1$ .

For  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $m < n$ ,  $1 \leq p \leq \infty$ , we define the<sup>2</sup>  $\ell^p$ -minimal generalized inverse as

$$\mathbf{ginv}_p(\mathbf{A}) \stackrel{\text{def}}{=} \arg \min_{\mathbf{X}} \|\mathbf{X}\|_p \quad \text{subject to } \mathbf{X} \in \mathcal{G}(\mathbf{A}),$$

with  $\|\mathbf{M}\|_p = (\sum_{ij} |m_{ij}|^p)^{1/p}$ . The MPP  $\mathbf{A}^\dagger$  is obtained by minimizing the Frobenius norm for  $p = 2$ , thus  $\|\mathbf{ginv}_p(\mathbf{A})\|_F \geq \|\mathbf{A}^\dagger\|_F$  for any  $p$ . Computing  $\mathbf{ginv}_p(\mathbf{A})$  involves solving a convex program.

Our initial motivation for this work is the sparse pseudoinverse,  $p = 1$ , since applying a sparse pseudoinverse requires less operations than applying a full one [10, 20, 4, 18]. The sparsest generalized inverse may be formulated as

$$(1.1) \quad \mathbf{ginv}_0(\mathbf{A}) \stackrel{\text{def}}{=} \arg \min_{\mathbf{X}} \|\mathbf{X}\|_0 \quad \text{subject to } \mathbf{X} \in \mathcal{G}(\mathbf{A}),$$

where  $\|\cdot\|_0$  counts the total number of non-zero entries. The non-zero count gives the naive complexity of applying  $\mathbf{X}$  or its adjoint to a vector.

Any optimally sparse generalized inverse of  $\mathbf{A}$  in the sense of the  $\ell^0$  norm (1.1) is by definition in the set  $\mathbf{ginv}_0(\mathbf{A})$ . When  $\mathbf{A} \in \mathbb{C}^{m \times n}$ ,  $m \leq n$ , has full row rank the condition  $\mathbf{X} \in \mathcal{G}(\mathbf{A})$  is equivalent to  $\mathbf{A}\mathbf{X} = \mathbf{I}_m$ , and computing it can be expressed columnwise, as a collection of  $\ell^0$  minimization problems  $[\mathbf{ginv}_0(\mathbf{A})]_i = \arg \min_{\mathbf{Ax}=\mathbf{e}_i} \|\mathbf{x}\|_{\ell^0}$ . Even though optimization problems of this kind are NP-hard [7, 21], generically, finding a solution is simple (see Theorem 2.1): just invert any full-rank  $m \times m$  submatrix and zero the rest. Unfortunately, this strategy leads to poorly conditioned matrices.

It is then natural to try and replace  $\ell^0$  by the tractable  $\ell^1$  norm [12]. Indeed, it was shown in [9] that  $\mathbf{ginv}_1(\mathbf{A})$  provides a minimizer of the  $\ell^0$  norm for almost all matrices, and that this minimizer is unique, motivating the notation  $\mathbf{spinv}(\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{ginv}_1(\mathbf{A})$ . Intuitively, an  $m \times n$  matrix  $\mathbf{A}$  with  $m \leq n$  is generically rank  $m$ , hence  $\mathbf{A}\mathbf{X} = \mathbf{I}_m$  is a system of  $m^2$  independent linear

<sup>1</sup>We only study real matrices.

<sup>2</sup>For  $p = 1$ ,  $\mathbf{ginv}_p(\mathbf{A})$  is a singleton except for a set of matrices of measure zero.

equations. The matrix  $\mathbf{X}$  has  $nm$  entries, leaving us with  $nm - m^2$  degrees of freedom, which one hopes will correspond to  $nm - m^2$  zero entries. The main advantage of  $\text{spinv}$  over  $\ell^0$  minimization is that it yields unique, well-behaved matrices.

**1.1. Prior Art.** There is a one-to-one correspondence between generalized inverses of full rank matrices and dual frames. Several earlier works [4, 18, 20] study existence and explicit constructions of sparse frames and sparse dual frames. Krahmer, Kutyniok, and Lemvig [18] establish sharp bounds on the sparsity of dual frames, showing that generically, for  $\mathbf{A} \in \mathbb{C}^{m \times n}$ , the sparsest dual has  $mn - m^2$  zeros, while Li, Liu, and Mi [20] provide better bounds on the sparsity of dual Gabor frames. They introduce the idea of using  $\ell^1$  minimization to find these dual frames, and show that under certain conditions,  $\ell^1$  minimization yields the sparsest possible dual Gabor frame. Further examples of non-canonical dual frames are given by Perraudin et al., who use convex optimization to derive dual frames with good time-frequency localization [24].

Results on finite-size concentration bounds for norms of pseudoinverses are scarce, with one notable exception being an upper bound on the probability of large deviation for the MPP [16] (we obtain concentration bounds for a complementary regime). On the other hand, a number of results exist for square matrices [25, 33]. The sparse pseudoinverse was previously studied in [10], where it was shown empirically that the minimizer is indeed a sparse matrix, and that it can be used to speed up the resolution of certain inverse problems.

Our proof relies on the *convex Gaussian min-max theorem* (CGMT) [22, 30, 31] which was previously used to quantify performance of regularized M-estimators such as the lasso [22, 30, 31]. Many technical ideas in [30, 22, 31] have been developed in earlier works. The CGMT can be seen as a descendant of Gordon’s Gaussian min-max theorem [14]. Rudelson and Vershynin [26] first recognized that Gordon’s result (more precisely, its consequence known as escape through a mesh [15]) is a useful theoretical device to study sparse regression. Ensuing papers by Stojnić [28, 29], Chandrasekaran et al. [5], Amelunxen et al. [1], Foygel and Mackey [13], and others, give sharper analyses and study more general settings. Their techniques percolated into the work by Thrampoulidis et al. [30] which we primarily refer to.

**1.2. Our Contributions.** In Sections 2 and 3 we prove an exponential concentration result for the Frobenius norm of  $\ell^p$ -minimal pseudoinverses for iid Gaussian matrices. Specializing to  $p = 1$  in Corollary 2.4, we show that unlike simpler strategies that yield a sparsest generalized inverse,  $\ell^1$  minimization produces a well-conditioned matrix; specializing to  $p = 2$  in Corollary 2.3, we get new results for the Frobenius norm of the MPP. Unlike previous applications of the CGMT, we give finite-size concentration bounds rather than asymptotic “in probability” results.

**2. Frobenius Norms of Generalized Inverses.** In this section we state our main results and prepare the proof. We first need to clear a technicality: for some  $\mathbf{A}$ ,  $\text{spinv}(\mathbf{A})$  will have multiple minimizers. This is, however, rare, as we prove in [9]:

**THEOREM 2.1.** *Assume that  $\mathbf{A} \in \mathbb{R}^{m \times n}$  has columns in general position, and that  $\mathbf{A}$  is in general position with respect to the canonical basis vectors  $\mathbf{e}_1, \dots, \mathbf{e}_m$ . Then the sparse pseudoinverse  $\text{spinv}(\mathbf{A})$  of  $\mathbf{A}$  contains a single matrix whose columns are all exactly  $m$ -sparse.*

Operationally, this means that for almost all matrices with respect to the Lebesgue measure on  $\mathbb{R}^{m \times n}$  we will have a unique  $\text{spinv}$ , which simplifies the proofs.<sup>3</sup>

The importance of the Frobenius norm of a generalized inverse can be motivated by considering an overdetermined system of linear equations  $\mathbf{y} = \mathbf{A}^\top \mathbf{x} + \mathbf{z}$ , where  $\mathbf{A} \in \mathbb{C}^{m \times n}$  has full row rank and  $\mathbf{z}$  is noise such that  $\mathbb{E}\{\mathbf{z}\mathbf{z}^\top\} \propto \mathbf{I}_n$ . Then for  $\mathbf{W} \in \mathcal{G}(\mathbf{A})$  we have  $\mathbb{E}_{\mathbf{z}}[\|\mathbf{W}\mathbf{y} - \mathbf{x}\|_2^2] = \mathbb{E}_{\mathbf{z}}[\|\mathbf{W}\mathbf{z}\|_2^2] \propto$

<sup>3</sup>A careful reader will notice that the assumptions of Theorem 2.1 forbid many types of sparse matrices. Indeed, it is known that sparse frames can have sparser duals than generic frames [18, 20].

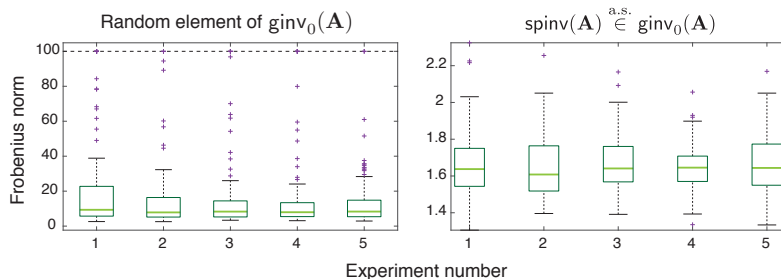


FIG. 2.1. *Frobenius norm of a random element of  $\text{ginv}_0(\mathbf{A})$  for  $\mathbf{A}$  a random Gaussian matrix of size  $20 \times 30$  (left) and a sparse pseudoinverse  $\text{spinv}(\mathbf{A})$  (generically in  $\text{ginv}_0(\mathbf{A})$ ; cf. [9]) of the same matrix (right). The random element of  $\text{ginv}_0(\mathbf{A})$  is computed by selecting  $m$  columns of  $\mathbf{A}$  at random and inverting the obtained submatrix. The plot shows results of 5 identical experiments, each consisting of generating a random  $20 \times 30$  matrix and computing the two inverses 100 times. In the first experiment, the outlier norms extend to 2000 so they were clipped at 100. Green lines denote medians, and boxes denote the second and the third quartile.*

$\|\mathbf{W}\|_F^2$ . Thus the mean-squared error is controlled by the Frobenius norm of  $\mathbf{W}$  and it is desirable to use generalized inverses with small Frobenius norms.

Theorem 2.1 says that for a generic  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with  $m < n$ , a sparsest generalized inverse has  $m^2$  non-zeros. Thus a simple way to compute a sparsest generalized inverse (an element of  $\text{ginv}_0(\mathbf{A})$ ) is to invert any  $m \times m$  submatrix of  $\mathbf{A}$  and set the rest to zero. Unfortunately, this leads to poorly conditioned matrices, unlike computing the  $\text{spinv}$  as shown in Figure 2.1.

The goal of this section is to make the last statement quantitative by developing concentration results for the Frobenius norm of  $\text{spinv}(\mathbf{A})$  as well as  $\text{ginv}_p(\mathbf{A})$  for  $1 \leq p \leq 2$  when  $\mathbf{A}$  is an iid Gaussian random matrix.

Gaussian random matrices may appear as a serious restriction. It is known, however, that Gaussian matrices are a representative of a large class of random matrix models for which many relevant functionals are universal—they concentrate around the same value for a given matrix size [11, 23]. Although it is tempting to justify our model choice by universality, in the case of sparse pseudoinverses we must proceed with care. As Oymak and Tropp point out [23], “...geometric functionals involving non-Euclidean norms need not exhibit universality.” They give an example of  $\ell^1$  restricted minimum singular value. Indeed, as Figure 2.2 shows, while the predictions of Corollary 2.3 for the MPP remain true over a number of ensembles and thus seem to be universal, those of Corollary 2.4 for the  $\text{spinv}$  exhibit various levels of disagreement. For the Rademacher ensemble they collapse completely. In view of our results on pseudoinverses of structured matrices [8], this comes as no surprise since we know that in this case the  $\text{spinv}$  contains the MPP. Still, from Figure 2.2, results for Gaussian matrices are a good qualitative template for many absolutely continuous<sup>4</sup> distributions, and the Gaussian assumption enables us to use sharp tools not available in a more general setting.

Our results rely on the properties of the following geometric functional:

$$(2.2) \quad D(t) = D_p(t; n) \stackrel{\text{def}}{=} \frac{1}{n} \left[ \mathbb{E}_{\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)} \text{dist}(\mathbf{h}, \|\cdot\|_{p^*} \leq t) \right]^2, \quad 1 \leq p \leq \infty, t \geq 0,$$

where  $\text{dist}(\mathbf{h}, \|\cdot\|_{p^*} \leq t) \stackrel{\text{def}}{=} \min_{\mathbf{u}: \|\mathbf{u}\|_{p^*} \leq t} \|\mathbf{h} - \mathbf{u}\|_2$  and  $1/p + 1/p^* = 1$ . In particular, Lemma B.4–Property 7 tells us that  $0 < D(t) < 1$  and for  $0 < \delta < 1$ ,  $\delta = D(t) - \frac{t}{2} D'(t)$  on  $(0, \infty)$  has a unique solution on  $(0, \infty)$ , denoted  $t^* = t_p^*(\delta, n)$ . With this notation we can state our main result.

<sup>4</sup>With respect to the Lebesgue measure.

**THEOREM 2.2.** Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $1 < m \leq n$ , be a standard iid Gaussian matrix and  $\delta \stackrel{\text{def}}{=} (m-1)/n \in (0, 1)$ . For  $1 \leq p \leq 2$ , define  $t^* = t_p^*(\delta, n)$  to be the unique solution of  $\delta = D(t) - \frac{t}{2}D'(t)$  on  $(0, \infty)$  and denote

$$\alpha^* = \alpha_p^*(\delta, n) \stackrel{\text{def}}{=} \sqrt{\frac{D(t^*)}{\delta(\delta - D(t^*))}}.$$

Assume there exist<sup>5</sup>  $\gamma(\delta)$  and  $N(\delta)$  such that  $-t^*D'_p(t^*; n) \geq \gamma(\delta) > 0$  for all  $n \geq N(\delta)$ . Then for any  $n \geq \max(2/(1-\delta), N(\delta))$  we have: for any  $0 < \epsilon \leq 1$

$$(2.3) \quad \mathbb{P} \left[ \left| \frac{n}{m} \|\text{ginv}_p(\mathbf{A})\|_F^2 - (\alpha^*)^2 \right| \geq \epsilon(\alpha^*)^2 \right] \leq \frac{n}{C_1 \epsilon} e^{-C_2 n \epsilon^4},$$

where the constants  $C_1, C_2 > 0$  may depend on  $\delta$  but not on  $n$  or  $\epsilon$ .

From Theorem 2.2 we can derive more explicit results for the two most interesting cases:  $p = 1$  and  $p = 2$ . For  $p = 2$  we get a result about  $\|\mathbf{A}^\dagger\|_F$  complementary to a known large deviation bound [16, Proposition A.5; Theorem A.6] obtained by a different technique.

**COROLLARY 2.3** ( $p = 2$ ). Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,  $1 < m \leq n$ , be a standard iid Gaussian matrix,  $\delta = (m-1)/n \in (0, 1)$ , and  $\alpha^* = \alpha_2^*(\delta, n) \stackrel{\text{def}}{=} \sqrt{\frac{D_2(t^*; n)}{\delta(\delta - D_2(t^*; n))}}$ , with  $t^*$  being the unique solution of  $\delta = D_2(t; n) - \frac{t}{2}D'_2(t; n)$  on  $(0, \infty)$ . Then there exists  $N(\delta)$  such that for  $n \geq N(\delta)$  we have: for any  $0 < \epsilon \leq 1$ ,

$$(2.4) \quad \mathbb{P} \left[ \left| \frac{n}{m} \|\mathbf{A}^\dagger\|_F^2 - (\alpha^*)^2 \right| \geq \epsilon(\alpha^*)^2 \right] \leq \frac{n}{C_1 \epsilon} e^{-C_2 n \epsilon^4},$$

where the constants  $C_1, C_2 > 0$  may depend on  $\delta$  but not on  $n$  or  $\epsilon$ , and  $\alpha^* = \alpha_2^*(\delta; n)$  with

$$(2.5) \quad \lim_{n \rightarrow \infty} \alpha_2^*(\delta; n) = 1/\sqrt{1-\delta}.$$

We remark that Corollary 2.3 covers “small” deviations ( $0 < \epsilon \leq 1$ ). In contrast, the result of [16] establishes<sup>6</sup> that  $\mathbb{E} \frac{n}{m} \|\mathbf{A}^\dagger\|_F^2 = \frac{n}{n-m-1} = \frac{1}{1-\delta}$  and that for any  $\tau \geq 1$ ,

$$\mathbb{P} \left[ \frac{n}{m} \|\mathbf{A}^\dagger\|_F^2 \geq \frac{12n}{n-m} \tau \right] \leq 4\tau^{-(n-m)/4} = 4e^{-n \frac{(1-m/n) \log \tau}{4}}.$$

For large  $n$  we have  $\frac{12n}{n-m} \tau \approx \frac{12\tau}{1-\delta} \approx 12\tau(\alpha_2^*)^2$  and  $\frac{(1-m/n) \log \tau}{4} \approx \frac{(1-\delta) \log \tau}{4}$ , hence this provides a bound for  $1 + \epsilon := 12\tau \geq 12$  with an exponent  $\log \tau$ , of the order of  $\log \epsilon$  (instead of  $\epsilon^4$  we get for  $0 < \epsilon \leq 1$ ). Further, we also show that the probability of  $\frac{n}{m} \|\mathbf{A}^\dagger\|_F^2$  being much smaller than  $(\alpha^*)^2$  is exponentially small.

The most interesting corollary is for the sparse pseudoinverse,  $p = 1$ .

**COROLLARY 2.4** ( $p = 1$ ). With the notation analogous to Corollary 2.3, there exists  $N(\delta)$  such that for all  $n \geq N(\delta)$  we have: for any  $0 < \epsilon \leq 1$ ,

$$(2.6) \quad \mathbb{P} \left[ \left| \frac{n}{m} \|\text{spinv}(\mathbf{A})\|_F^2 - (\alpha^*)^2 \right| \geq \epsilon(\alpha^*)^2 \right] \leq \frac{n}{C_1 \epsilon} e^{-C_2 n \epsilon^4},$$

where the constants  $C_1, C_2 > 0$  may depend on  $\delta$  but not on  $n$  or  $\epsilon$ , and  $\alpha^* = \alpha_1^*(\delta; n)$  with

$$(2.7) \quad \lim_{n \rightarrow \infty} \alpha_1^*(\delta; n) = \sqrt{\left( \sqrt{\frac{2}{\pi}} e^{-\frac{(t_1^*)^2}{2}} t_1^* - \delta(t_1^*)^2 \right)^{-1} - \frac{1}{\delta}}.$$

and  $t_1^* = \sqrt{2} \cdot \text{erfc}^{-1}(\delta)$ .

<sup>5</sup>The existence of  $\gamma(\delta)$  and  $N(\delta)$  will be proved below for  $p \in \{1, 2\}$ .

<sup>6</sup>The results in [16] are designed for the nonasymptotic regime where the matrix is essentially square.

*Several remarks on the corollaries.* Results of Corollaries 2.3 and 2.4 are illustrated<sup>7</sup> in Figures 2.2 and 2.3. Figure 2.2 shows the shape of the limiting  $(\alpha_p^*)^2$  as a function of  $\delta$ , as well as empirical averages for different values of  $n$  and  $\delta$ . As expected, the limiting values get closer to the empirical result as  $n$  grows larger. In Figure 2.3 we also show the individual realizations for different combinations of  $n$  and  $\delta$ . As predicted by the two corollaries, the variance of both  $\frac{n}{m} \|\mathbf{A}^\dagger\|_F^2$  and  $\frac{n}{m} \|\text{spinv}(\mathbf{A})\|_F^2$  reduces with  $n$ . For larger values of  $n$ , all realizations are close to the limiting  $(\alpha_p^*)^2$ .

The  $\text{spinv}$  and the MPP exhibit rather different qualitative behaviors. The Frobenius norm of the MPP monotonically decreases as  $\delta$  gets smaller, while that of the  $\text{spinv}$  turns up below some critical  $\delta$ . Intuitively, for small  $\delta$ , the support of the  $\text{spinv}$  is concentrated on few entries which have to be comparably large to produce the diagonal in  $\mathbf{I} = \mathbf{A}\mathbf{X}$ . Careful analysis of (2.7) using an asymptotic expansion of  $\text{erfc}(t)$  shows that for a sufficiently large  $n$ ,  $\alpha_1^*(\delta; n)$  behaves as  $[\delta \log(1/\delta)]^{-1/2}$ .

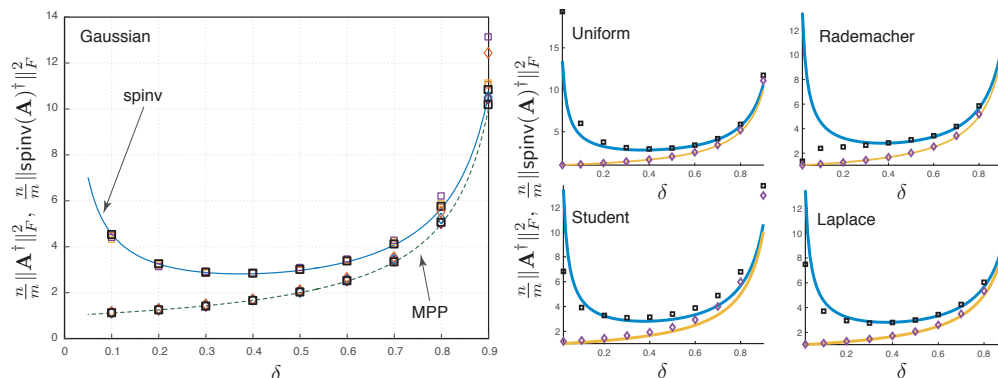


FIG. 2.2. Left: Comparison of the limiting  $(\alpha_p^*)^2$  with the mean of  $\frac{n}{m} \|\mathbf{A}^\dagger\|_F^2$  and  $\frac{n}{m} \|\text{spinv}(\mathbf{A})\|_F^2$  for 10 realizations of  $\mathbf{A}$ . Empirical results are given for  $\delta \in \{0.1, 0.2, \dots, 0.9\}$  and  $n \in \{100, 200, 500, 1000\}$ . Black squares represent the empirical result for  $n = 1000$ ; colored squares represent the empirical mean of  $\frac{n}{m} \|\mathbf{A}^\dagger\|_F^2$  for  $n \in \{100, 200, 500\}$ , with the largest discrepancy (purple squares) for  $n = 100$ ; colored diamonds represent the empirical mean of  $\frac{n}{m} \|\text{spinv}(\mathbf{A})\|_F^2$  with the largest discrepancy (orange diamonds) again for  $n = 100$ . Right: Empirical averages over 100 trials for different matrix ensembles normalized to unit entry variance;  $n = 200$  and  $\delta \in \{0.02, 0.1, 0.2, \dots, 0.9\}$ .

The bound (2.3) and the bounds in Corollaries 2.4 and 2.3 involve  $\epsilon^4$  instead of the more common  $\epsilon^2$  one gets for, e.g., Lipschitz functions: to guarantee a given probability in the right hand side of (2.3),  $\epsilon$  should be of the order  $n^{-1/4}$  instead of the usual  $n^{-1/2}$ , suggesting a comparably higher variance of  $\frac{n}{m} \|\text{ginv}_p(\mathbf{A})\|_F^2$ . This is a consequence of the technique used to bound  $\inf_{|\alpha - \alpha^*| \geq \epsilon} \kappa(\alpha) - \kappa(\alpha^*)$  in Lemma B.8 which relies on strong convexity of  $\kappa$ . Whether a more refined analysis could lead to better error bars remains an open question.

Corollaries 2.3 and 2.4 prove that  $\|\mathbf{A}^\dagger\|_F^2$  and  $\|\text{spinv}(\mathbf{A})\|_F^2$  indeed concentrate and give a closed-form limiting value of the optimal  $\alpha^*$ . It would seem natural that an interpolation to  $p \in (1, 2)$  is possible, although  $\alpha_p^*$  would be specified implicitly and computed numerically. It is less clear whether an extension to  $p > 2$  is possible.

*Proof of Corollaries 2.3–2.4.* For  $p \in \{1, 2\}$ , we lower bound  $-t^* D'_p(t^*) \geq \gamma(\delta)$  for all  $n$  above some  $N(\delta)$  using Lemmata B.6–B.7, and control  $\lim_{n \rightarrow \infty} \alpha_p^*(\delta; n)$ . We conclude using Theorem 2.2.  $\square$

<sup>7</sup>For reproducible research, code is available online at <https://github.com/doksa/altginv>.

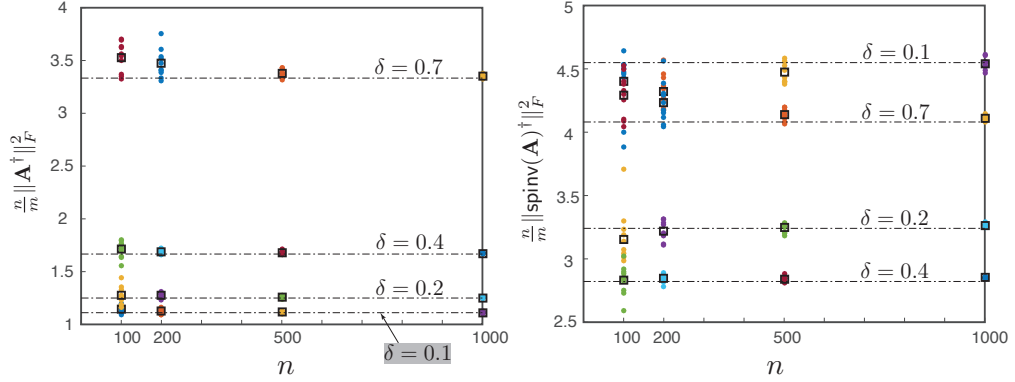


FIG. 2.3. Comparison of the limiting  $(\alpha_p^*)^2$  with the value of  $\frac{n}{m} \|\mathbf{A}^\dagger\|_F^2$  (left) and  $\frac{n}{m} \|\text{spinv}(\mathbf{A})\|_F^2$  (right) for 10 realizations of  $\mathbf{A}$ . Results are shown for  $\delta \in \{0.1, 0.2, 0.4, 0.7\}$  and different values of  $n$ . Values for individual realizations are shown with colored dots with different color for every combination of  $n$  and  $\delta$ . Horizontal dashed lines indicate the limiting value for the considered values of  $\delta$ .

**3. Proof of the Main Concentration Result, Theorem 2.2.** Our proof technique relies on decoupling the optimization for  $\text{ginv}_p(\mathbf{A})$  over columns. A “standard” application of the CGMT would give an asymptotic result for the  $\ell^p$  norm of one column which holds in probability. However, because the squared Frobenius norm is a sum of  $m$  squared  $\ell^2$  column norms, convergence in probability is not enough. To address this shortcoming we developed a number of technical results located mostly in Appendix B that lead to a stronger exponential concentration result which may be of independent interest to the users of the CGMT.

By definition, we have

$$\text{ginv}_p(\mathbf{A}) = \arg \min_{\mathbf{A}\mathbf{X}=\mathbf{I}} \|\mathbf{X}\|_p = \arg \min_{\mathbf{A}\mathbf{X}=\mathbf{I}} \|\mathbf{X}\|_p^p,$$

where we assume that the solution is unique. This is true by strict convexity for  $p > 1$ ; for  $p = 1$  it holds almost surely by Theorem 2.1. This optimization decouples over columns of  $\mathbf{X}$ : denoting  $\mathbf{X}^* = \text{ginv}_p(\mathbf{A})$  we have for the  $i$ th column that  $\mathbf{x}_i^* = \arg \min_{\mathbf{A}\mathbf{x}=\mathbf{e}_i} \|\mathbf{x}\|_p$ . We can thus apply the following lemma proved in Section 3.1:

LEMMA 3.1. *With notations and assumptions as in Theorem 2.2, we have for  $0 < \epsilon' \leq 1$  and  $n \geq \max(2/(1 - \delta), N(\delta))$*

$$\mathbb{P} \left\{ \left| \sqrt{n} \|\mathbf{x}^*\|_2 - \alpha^* \right| \geq \epsilon' \alpha^* \right\} \leq \frac{1}{K_1 \epsilon'} e^{-K_2 n \epsilon'^4},$$

where  $K_1, K_2 > 0$  may depend on  $\delta$  but not on  $n$  or  $\epsilon'$ .

Lemma 3.1 tells us that  $\sqrt{n} \|\mathbf{x}_i^*\|_2$  remains close to  $\alpha^*$  with high probability. To exploit the additivity of the squared Frobenius norms over columns we write  $(n \|\mathbf{x}^*\|_2^2 - (\alpha^*)^2) = (\sqrt{n} \|\mathbf{x}^*\|_2 - \alpha^*)(\sqrt{n} \|\mathbf{x}^*\|_2 + \alpha^*)$ . Then for any  $b > 0$  we have that

$$\mathbb{P} \left\{ \left| n \|\mathbf{x}^*\|_2^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} \leq \mathbb{P} \left\{ \left| \sqrt{n} \|\mathbf{x}^*\|_2 - \alpha^* \right| \geq \frac{\epsilon (\alpha^*)^2}{b} \right\} + \mathbb{P} \left\{ (\sqrt{n} \|\mathbf{x}^*\|_2 + \alpha^*) \geq b \right\}.$$

By taking  $b = 3\alpha^*$ , we bound both terms using Lemma 3.1 to obtain

$$\mathbb{P} \left\{ \left| n \|\mathbf{x}^*\|_2^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} \leq \frac{3}{K_1 \epsilon} e^{-K_2 n (\epsilon/3)^4} + \frac{1}{K_1} e^{-K_2 n} \leq \frac{1}{C_1 \epsilon} e^{-C_2 n \epsilon^4}$$

with an appropriate choice of  $C_1, C_2$ .

This characterizes the squared  $\ell^2$  norm of one column of the MPP. The (scaled) squared Frobenius norm of  $\mathbf{X}^*$  is a sum of  $m$  such terms which are not independent,  $\frac{n}{m} \|\mathbf{X}^*\|_F^2 = \frac{n}{m} \sum_{i=1}^m \|\mathbf{x}_i^*\|_2^2$ , and we want to show that it stays close to  $(\alpha^*)^2$ . We work as follows:

$$\begin{aligned} \mathbb{P} \left\{ \left| \frac{n}{m} \|\mathbf{X}^*\|_F^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} &= \mathbb{P} \left\{ \left| \sum_{i=1}^m n \|\mathbf{x}_i^*\|_2^2 - (\alpha^*)^2 \right| \geq m\epsilon (\alpha^*)^2 \right\} \\ &\leq \mathbb{P} \left\{ \sum_{i=1}^m \left| n \|\mathbf{x}_i^*\|_2^2 - (\alpha^*)^2 \right| \geq m\epsilon (\alpha^*)^2 \right\} = (*). \end{aligned}$$

If the sum of  $m$  terms is to be larger than  $m\epsilon$ , then at least one term must be larger than  $\epsilon$ ,

$$\begin{aligned} (*) &\leq \mathbb{P} \left\{ \exists i, 1 \leq i \leq m : \left| n \|\mathbf{x}_i^*\|_2^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} = \mathbb{P} \left\{ \bigcup_{i=1}^m \left\{ \left| n \|\mathbf{x}_i^*\|_2^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} \right\} \\ &\leq \sum_{i=1}^m \mathbb{P} \left\{ \left| n \|\mathbf{x}_i^*\|_2^2 - (\alpha^*)^2 \right| \geq \epsilon (\alpha^*)^2 \right\} \\ &\leq m \frac{1}{C_1 \epsilon} e^{-C_2 n \epsilon^4} \leq \frac{\delta n}{C_1 \epsilon} e^{-C_2 n \epsilon^4}, \end{aligned}$$

which completes the proof.

### 3.1. Proof of the Main Vector Result, Lemma 3.1.

Define

$$(3.8) \quad \mathbf{x}^* = \arg \min_{\mathbf{A}\mathbf{x}=\mathbf{e}_1} \|\mathbf{x}\|_p$$

$$(3.9) \quad \tilde{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{e}_1\|_2 + \lambda \|\mathbf{x}\|_p$$

By Lemma B.1, since the  $\ell^p$  norm is  $L$ -Lipschitz with respect to the  $\ell^2$  norm (with  $L \stackrel{\text{def}}{=} n^{\max(1/p-1/2, 0)}$ ), if we choose  $\lambda \leq \frac{\sqrt{n}-\sqrt{m}}{L}(1-\epsilon)$ , minimizers  $\mathbf{x}^*$  and  $\tilde{\mathbf{x}}$  of (3.8) and (3.9) coincide<sup>8</sup> with probability at least  $1 - e^{-\epsilon^2(\sqrt{n}-\sqrt{m})^2/2}$ . Using  $\|\mathbf{y}\|_2 = \max_{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{y}$  we get

$$(3.10) \quad \tilde{\mathbf{x}} = \arg \min_{\mathbf{x}} \max_{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{A}\mathbf{x} + \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1.$$

The objective in (3.10) is a sum of a bilinear term involving  $\mathbf{A}$  and a convex-concave function<sup>9</sup>  $\psi(\mathbf{x}, \mathbf{u}) = \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1$  as required by the CGMT (Appendix A.2). The CGMT requires  $\mathbf{x}$  to belong to a compact set so instead of (3.10) we analyze the following bounded modification:

$$(3.11) \quad \tilde{\mathbf{x}}_K = \arg \min_{\mathbf{x}: \|\mathbf{x}\|_2 \leq K} \max_{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{A}\mathbf{x} + \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1$$

We will see that  $\tilde{\mathbf{x}}_K = \tilde{\mathbf{x}}$  with high probability as soon as  $K$  is large enough.

Part (ii) of the CGMT says that the random optimal value of the *principal* optimization (3.11) concentrates if the random value of the following *auxiliary* optimization concentrates:

$$(3.12) \quad \hat{\mathbf{x}}_K = \arg \min_{\mathbf{x}: \|\mathbf{x}\|_2 \leq K} \max_{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1} \|\mathbf{x}\|_2 \mathbf{g}^\top \mathbf{u} - \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{x} + \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1,$$

<sup>8</sup>An analogous result does not hold for the squared lasso (except for  $\lambda = 0^+$ ).

<sup>9</sup>Convex in the first argument, concave in the second one.



where  $\mathbf{g} \sim (\mathbf{0}, \mathbf{I}_m)$  and  $\mathbf{h} \sim (\mathbf{0}, \mathbf{I}_n)$  are independent. This lets us prove that if the norm  $\|\hat{\mathbf{x}}_K\|_2$  of the optimizer of (3.12) concentrates, then the norm  $\|\tilde{\mathbf{x}}_K\|_2$  of the optimizer of (3.11) concentrates around the same value.<sup>10</sup>

We now go through a series of steps to simplify (3.12). For convenience we let  $\mathbf{z}$  be a scaled version of  $\mathbf{x}$ ,  $\mathbf{z} = \mathbf{x}\sqrt{n}$  (accordingly  $A = K\sqrt{n}$ ). Using the variational characterization of the  $\ell^p$  norm we rewrite (3.12) as

$$(3.13) \quad \hat{\mathbf{z}}_A = \arg \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{u}: \|\mathbf{u}\|_2 = \beta \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} 1/\sqrt{n} \cdot \|\mathbf{z}\|_2 \mathbf{g}^\top \mathbf{u} - 1/\sqrt{n} \cdot \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{z} - \mathbf{u}^\top \mathbf{e}_1 + \lambda/\sqrt{n} \cdot \mathbf{w}^\top \mathbf{z}.$$

Let  $\check{\mathbf{z}}_A$  be a (random) value of  $\mathbf{z}$  at the optimum of the following reordered optimization:

$$(3.14) \quad \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} 1/\sqrt{n} \cdot \|\mathbf{z}\|_2 \mathbf{g}^\top \mathbf{u} - 1/\sqrt{n} \cdot \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{z} - \mathbf{u}^\top \mathbf{e}_1 + \lambda/\sqrt{n} \cdot \mathbf{w}^\top \mathbf{z}.$$

By Lemma B.11 and Lemma B.12,  $\|\check{\mathbf{z}}_A\|_2$  and  $\|\hat{\mathbf{z}}_A\|_2$  stay close with high probability (this will be made precise below). We simplify (3.14) further as follows:

$$(3.15) \quad \begin{aligned} (3.14) &= \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \beta \|\|\mathbf{z}\|_2 \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \beta/\sqrt{n} \cdot \mathbf{h}^\top \mathbf{z} + \lambda/\sqrt{n} \cdot \mathbf{w}^\top \mathbf{z} \\ &= \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\alpha: 0 \leq \alpha \leq A} \min_{\mathbf{z}: \|\mathbf{z}\|_2 = \alpha} \beta \|\|\mathbf{z}\|_2 \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \beta/\sqrt{n} \cdot \mathbf{h}^\top \mathbf{z} + \lambda/\sqrt{n} \cdot \mathbf{w}^\top \mathbf{z} \\ &= \max_{\beta: 0 \leq \beta \leq 1} \min_{\alpha: 0 \leq \alpha \leq A} \beta \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \alpha/\sqrt{n} \cdot \|\beta \mathbf{h} - \lambda \mathbf{w}\|_2. \end{aligned}$$

The objective in the last line is convex in  $\alpha$  and jointly concave in  $(\beta, \mathbf{w})$ , and the constraint sets are all convex and bounded, so by [27, Corollary 3.3] we can again exchange min and max:

$$(3.16) \quad \begin{aligned} &= \min_{\alpha: 0 \leq \alpha \leq A} \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \beta \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \alpha/\sqrt{n} \cdot \|\beta \mathbf{h} - \lambda \mathbf{w}\|_2 \\ &= \min_{\alpha: 0 \leq \alpha \leq A} \max_{\beta: 0 \leq \beta \leq 1} \left( \beta \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \alpha/\sqrt{n} \cdot \min_{\mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1} \|\beta \mathbf{h} - \lambda \mathbf{w}\|_2 \right) \\ &= \min_{\alpha: 0 \leq \alpha \leq A} \max_{\beta: 0 \leq \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}) \end{aligned}$$

where

$$(3.17) \quad \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}) \stackrel{\text{def}}{=} \beta \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 - \alpha/\sqrt{n} \cdot \text{dist}(\beta \mathbf{h}, \|\cdot\|_{p^*} \leq \lambda).$$

We thus simplified a high-dimensional vector optimization (3.9) into an optimization over two scalars (3.16), one of these scalars,  $\alpha$ , almost giving us what we seek—the (scaled)  $\ell^2$  norm of  $\mathbf{x}$ . To put the pieces together, there now remains to formally prove that the min-max switches and the concentration results we mentioned actually hold.

<sup>10</sup>The CGMT contains a similar statement, albeit we need a different derivation to get exponential concentration.

**3.2. Combining the ingredients.** By Lemma B.2,  $\phi(\alpha, \beta; \mathbf{g}, \mathbf{h})$  concentrates around some deterministic function  $\kappa(\alpha, \beta) = \kappa_p(\alpha, \beta; n, \delta, \lambda)$  with  $\delta = (m-1)/n$ . By Lemma B.8–3,

$$(3.18) \quad \arg \min_{\alpha: 0 \leq \alpha \leq A} \max_{\beta: 0 \leq \beta \leq 1} \kappa(\alpha, \beta) = \alpha^* = \alpha_p^*(\delta; n) \stackrel{\text{def}}{=} \sqrt{\frac{D_p(t_p^*; n)}{\delta(\delta - D_p(t_p^*; n))}}$$

as soon as  $A = K\sqrt{n} > \alpha^*$  and  $\lambda \leq t^*$ , with  $t^* = t_p^*(\delta, n)$  defined in Lemma B.4–Property 7. By Lemma B.10, for  $0 < \epsilon \leq \max(\alpha^*, A - \alpha^*)$ , the minimizer

$$\alpha_\phi^* \stackrel{\text{def}}{=} \arg \min_{0 \leq \alpha \leq A} \max_{0 \leq \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h})$$

stays  $\epsilon$ -close to  $\alpha^*$  with high probability:

$$\mathbb{P}\{|\alpha_\phi^* - \alpha^*| \geq \epsilon\} \leq \zeta(n, \frac{\omega(\epsilon)}{2}),$$

with  $\zeta(n, \xi) = \zeta(n, \xi; A, \delta) \stackrel{\text{def}}{=} c_1 e^{-c_2 \xi^2 n \cdot c(\delta, A)}$ ,  $c(\delta, A) \stackrel{\text{def}}{=} \frac{\min(\delta, A^{-2})}{1 + \delta A^2}$ ,  $c_1, c_2$  universal constants, and

$$\omega(\epsilon) = \omega_p(\epsilon; n, \delta, \lambda) \stackrel{\text{def}}{=} \frac{\epsilon^2}{2} \frac{\lambda \delta / t^*}{(1 + \delta(\alpha^* + \epsilon)^2)^{3/2}}.$$

As a consequence, by Lemma B.12, the scaled norm  $\sqrt{n} \|\tilde{\mathbf{x}}_K\|_2$  of the minimizer of the (bounded) principal optimization problem (3.11) stays close to  $\alpha^*$  with high probability,

$$\mathbb{P}\{|\sqrt{n} \|\tilde{\mathbf{x}}_K\|_2 - \alpha^*| \geq \epsilon\} \leq 4\zeta(n, \frac{\omega(\epsilon)}{2}).$$

Similarly, for  $0 < \epsilon \leq \min(\alpha^*, A - \alpha^*)$  by invoking Lemma B.13 we have that the norm  $\sqrt{n} \|\tilde{\mathbf{x}}\|$  of the minimizer of the unbounded optimization (3.10) stays close to  $\alpha^*$  with high probability

$$\mathbb{P}\{|\sqrt{n} \|\tilde{\mathbf{x}}\|_2 - \alpha^*| \geq \epsilon\} \leq 4\zeta(n, \frac{\omega(\epsilon)}{2})$$

and in fact  $\mathbb{P}[\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}_K] \leq 4\zeta(n, \frac{\omega(\epsilon)}{2})$ .

Since the  $\ell^p$  norm is  $L$ -Lipschitz with respect to the Euclidean metric in  $\mathbb{R}^n$ , with  $L = n^{\max(1/p-1/2, 0)}$ , Lemma B.1 gives for any  $\lambda < \lambda_{\max}(n, m, t^*) \stackrel{\text{def}}{=} \min\left\{\frac{\sqrt{n}-\sqrt{m}}{2L}, t^*\right\}$  that the minimizer  $\mathbf{x}^*$  of the equality-constrained optimization (3.8) coincides with the minimizer  $\tilde{\mathbf{x}}$  of the lasso formulation (3.9)–(3.10) except with probability at most  $e^{-n(1-\sqrt{\delta+1/n})^2/8}$ .

Overall then, for  $\lambda < \lambda_{\max}$  and  $\epsilon \leq \min(\alpha^*, A - \alpha^*)$ ,

$$(3.19) \quad \mathbb{P}\{|\sqrt{n} \|\mathbf{x}^*\|_2 - \alpha^*| \geq \epsilon\} \leq e^{-n(1-\sqrt{\delta+1/n})^2/8} + 4\zeta(n, \frac{\omega(\epsilon)}{2}).$$

The infimum over admissible values of  $\lambda$  is obtained by taking its value when  $\lambda = \lambda_{\max}$ .

**3.3. Making the bound (3.19) explicit.** From now on we choose  $A \stackrel{\text{def}}{=} 2\alpha^*$  and, for  $0 < \epsilon' \leq 1$ , we consider  $\epsilon \stackrel{\text{def}}{=} \epsilon' \alpha^*$  (which satisfies  $0 < \epsilon \leq \min(\alpha^*, A - \alpha^*) = \alpha^*$ ). We use  $\lesssim_\delta$  and  $\gtrsim_\delta$  to denote inequalities up to a constant that may depend on  $\delta$ , but not on  $n$  or  $\epsilon'$ , provided  $n \geq N(\delta)$ . We specify  $N(\delta)$  where appropriate.

By Lemma B.5–item 1, for any  $1 \leq p \leq 2$  and any  $n \geq 1$ ,

$$\lambda_{\max} = t^* \min\left(\frac{\sqrt{n}-\sqrt{m}}{n^{1/p-1/2}} \frac{1}{2t^*}, 1\right) \gtrsim_\delta t^* \min\left(\frac{\sqrt{n}(1-\sqrt{m/n})}{n^{1/p-1/2}} \frac{1}{n^{1-1/p}}, 1\right) = t^* \min\left(1 - \sqrt{m/n}, 1\right).$$

For  $n \geq \frac{2}{1-\delta}$  we have  $1 - \sqrt{m/n} = 1 - \sqrt{\delta + \frac{1}{n}} \geq 1 - \sqrt{(1+\delta)/2} \gtrsim_{\delta} 1$ , hence  $\lambda_{\max} \gtrsim_{\delta} t^*$ .

With the shorthands  $D(t) = D_p(t; n)$  and  $D = D_p(t^*; n)$ , we have

$$1 + \delta(\alpha^* + \epsilon)^2 \leq 1 + \delta A^2 \leq 4(1 + \delta(\alpha^*)^2) = 4\left(1 + \frac{D}{\delta - D}\right) = \frac{4\delta}{\delta - D} \lesssim_{\delta} (\delta - D)^{-1},$$

hence with  $\lambda = \lambda_{\max}$  we get for  $n \geq 2/(1 - \delta)$ ,

$$\omega(\epsilon) = \frac{\epsilon'^2 \lambda_{\max}}{2} \frac{\delta(\alpha^*)^2}{t^* (1 + \delta(\alpha^* + \epsilon)^2)^{3/2}} \gtrsim_{\delta} \epsilon'^2 \frac{\frac{D}{\delta - D}}{\left(\frac{4\delta}{\delta - D}\right)^{3/2}} \gtrsim_{\delta} \epsilon'^2 D \sqrt{\delta - D}.$$

By Lemma B.5–4 we have  $D = D_p(t^*; n) \geq (\delta/C)^2$  for a universal constant  $C$  independent of  $n$  or  $p$ , hence for  $n \geq 1$ ,  $D \gtrsim_{\delta} 1$  and for  $n \geq 2/(1 - \delta)$ ,

$$\omega(\epsilon) \gtrsim_{\delta} \sqrt{-t^* D'_p(t^*; n)} \epsilon'^2.$$

Moreover since  $\min(\delta, A^{-2}) \geq \frac{1}{4} \min(\delta, (\alpha^*)^{-2}) = \frac{1}{4} \min(\delta, \delta(\delta - D)/D)$  we also get

$$c(\delta, A) = \frac{\min(\delta, A^{-2})}{1 + \delta A^2} \geq \frac{\min(\delta, \delta(\delta - D)/D)}{\delta/(\delta - D)} = \min(\delta - D, \frac{(\delta - D)^2}{D}) \geq (\delta - D)^2 = \frac{1}{4} [-t^* D'_p(t^*; n)]^2.$$

Since  $-t^* D'_p(t^*; n) \geq \gamma(\delta) > 0$  for any  $n \geq N(\delta)$  (recall that  $t^* = t_p^*(\delta; n)$ ), we have

$$(3.20) \quad -t^* D'_p(t^*) \gtrsim_{\delta} 1,$$

and we obtain for  $n \geq \max(2/(1 - \delta), N)$ :  $\omega(\epsilon) \gtrsim_{\delta} \epsilon'^2$  and  $c(\delta, A) \gtrsim_{\delta} 1$ . Combining the above yields, for  $0 < \epsilon' \leq 1$ ,  $n \geq \max(2/(1 - \delta), N(\delta))$ :

$$\mathbb{P} \left\{ \left| \|\mathbf{x}^*\| - \frac{\alpha^*}{\sqrt{n}} \right| \geq \frac{\epsilon' \alpha^*}{\sqrt{n}} \right\} \leq e^{-C_1 n} + 4\zeta(n, C_2 \epsilon'^2) \leq \frac{1}{K_1 (\epsilon')^2} e^{-K_2 (\epsilon')^4 n}$$

with  $K_i, C_i \gtrsim_{\delta} 1$ .

**4. Conclusion.** We studied the concentration of the Frobenius norm of  $\ell^p$ -minimal pseudoinverses for iid Gaussian matrices. In addition to a general result for  $1 \leq p \leq 2$ , we gave explicit bounds for  $p \in \{1, 2\}$ , that is, for the sparse pseudoinverse and the Moore-Penrose pseudoinverse. Our results show that for a large range of  $m/n$  the Frobenius norm of the `spinv` is close to the Frobenius norm of the MPP which is the best possible among all generalized inverses (Figure 2.2). The same does not hold for the various ad hoc strategies that yield generalized inverses with the same non-zero count (Figure 2.1). In applications, this means that the `spinv` will not blow up noise much more than the MPP. Important future directions are extensions of Theorem 2.2 to matrix norms other than  $\ell^p$  with  $1 \leq p \leq 2$ , as well as matrix models other than iid Gaussian.

**5. Acknowledgments.** The authors would like to thank Mihailo Kolundžija, Miki Elad, Jakob Lemvig, and Martin Vetterli for the discussions and input in preparing this manuscript. A special thanks goes to Christos Thrampoulidis for his help in understanding and applying the Gaussian min-max theorem, and to Simon Foucart for discussions leading to the proof of [9, Lemma 2.1].

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

**Appendices. Appendix A. Results about Gaussian processes.**

### A.1. Concentration of measure.

LEMMA A.1. Let  $\mathbf{h}$  be a standard Gaussian random vector of length  $n$ ,  $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ , and  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  a 1-Lipschitz function. Then the following hold:

- (a) For any  $0 < \epsilon < 1$ ,  $\mathbb{P}\{\|\mathbf{h}\|_2^2 \leq n(1 - \epsilon)\} \leq e^{-\epsilon^2 n/4}$ ;  $\mathbb{P}\{\|\mathbf{h}\|_2^2 \geq n/(1 - \epsilon)\} \leq e^{-\epsilon^2 n/4}$ ;
- (b) For any  $0 < \epsilon < 1$ ,  $\mathbb{P}\left\{\|\mathbf{h}\|_2^2 \notin [(1 - \epsilon)n, n/(1 - \epsilon)]\right\} \leq 2e^{-\frac{\epsilon^2 n}{4}}$ ;
- (c) For any  $\epsilon > 0$ ,  $\mathbb{P}\left\{\left|\|\mathbf{h}\|_2^2 - n\right| \geq \sqrt{\epsilon n}\right\} \leq \begin{cases} 2e^{-\frac{\epsilon}{8}} & \text{for } 0 \leq \epsilon \leq n, \\ 2e^{-\frac{\sqrt{\epsilon n}}{8}} & \text{for } \epsilon > n; \end{cases}$
- (d) For any  $u > 0$ ,  $\mathbb{P}\{f(\mathbf{h}) - \mathbb{E}f(\mathbf{h}) \geq u\} \leq e^{-u^2/2}$ ;  $\mathbb{P}\{f(\mathbf{h}) - \mathbb{E}f(\mathbf{h}) \leq -u\} \leq e^{-u^2/2}$ ;
- (e) For any  $u > 0$ ,  $\mathbb{P}\{|f(\mathbf{h}) - \mathbb{E}f(\mathbf{h})| \geq u\} \leq 2e^{-u^2/2}$ ;
- (f)  $\text{Var}\{f(\mathbf{h})\} \leq 1$ .

*Proofs and references.* (a) [2, Corollary 2.3]; (b) (a) with a union bound; (c)  $X_i = h_i^2$  is subexponential with parameters  $\nu = 2, b = 4$  [34, Example 2.4], i.e.,  $\mathbb{E}[e^{\lambda(X_i - \mu)}] \leq e^{\frac{\nu^2 \lambda^2}{2}}$  for all  $|\lambda| \leq \frac{1}{b}$ , and  $\mu = \mathbb{E}[X_i] = 1$ . Applying [34, Proposition 2.2] then yields  $\mathbb{P}\left\{\left|\|\mathbf{h}\|_2^2 - n\right| \geq \sqrt{\epsilon n}\right\} \leq 2e^{-\frac{\epsilon}{8}}$ , for  $0 \leq \epsilon \leq n$ , and  $\mathbb{P}\left\{\left|\|\mathbf{h}\|_2^2 - n\right| \geq \sqrt{\epsilon n}\right\} \leq 2e^{-\frac{\sqrt{\epsilon n}}{8}}$ , for  $\epsilon > n$ . (d) [19, Eq. (1.22)]; (e) union bound applied to (d); (f) a consequence of Poincaré inequality for Gaussian measures [19, Eq. (2.16)]:  $\text{Var}\{f(\mathbf{h})\} \leq \mathbb{E}\{\|\nabla f(\mathbf{h})\|_2^2\}$  for 1-Lipschitz  $f$  for which  $\|\nabla f(\mathbf{h})\|_2 \leq 1$ .  $\square$

We will also use the following facts which can be verified by direct computation:

LEMMA A.2. Let  $\theta(t) \stackrel{\text{def}}{=} \mathbb{E}(|h| - t)_+^2$ , where  $h \sim \mathcal{N}(0, 1)$  and  $t \geq 0$ . Then

1.  $\theta(t) = (t^2 + 1) \text{erfc}\left(\frac{t}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} e^{-\frac{t^2}{2}} t$ ,
2.  $\theta(t) - (t/2)\theta'(t) = \text{erfc}(t/\sqrt{2})$ .

**A.2. Convex Gaussian Min-Max Theorem (CGMT).** Let the principal optimization (PO) and auxiliary optimization (AO) be defined as

$$\begin{aligned} \text{(PO)} \quad \Phi(\mathbf{G}) &\stackrel{\text{def}}{=} \min_{\mathbf{v} \in \mathcal{S}_{\mathbf{v}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} \mathbf{u}^\top \mathbf{G} \mathbf{v} + \psi(\mathbf{v}, \mathbf{u}) \\ \text{(AO)} \quad \phi(\mathbf{g}, \mathbf{h}) &\stackrel{\text{def}}{=} \min_{\mathbf{v} \in \mathcal{S}_{\mathbf{v}}} \max_{\mathbf{u} \in \mathcal{S}_{\mathbf{u}}} \|\mathbf{v}\|_2 \mathbf{g}^\top \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{v} + \psi(\mathbf{v}, \mathbf{u}), \end{aligned}$$

with  $\mathbf{G} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{g} \in \mathbb{R}^m$ ,  $\mathbf{h} \in \mathbb{R}^n$ ,  $\mathcal{S}_{\mathbf{v}} \subset \mathbb{R}^n$ ,  $\mathcal{S}_{\mathbf{u}} \subset \mathbb{R}^m$  and  $\psi : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ . Then we have the following result.

THEOREM A.3. [30, Theorem 6.1]

In (PO) and (AO), let  $\mathcal{S}_{\mathbf{v}}$  and  $\mathcal{S}_{\mathbf{u}}$  be compact and  $\psi$  continuous on  $\mathcal{S}_{\mathbf{v}} \times \mathcal{S}_{\mathbf{u}}$ . Let also  $\mathbf{G}, \mathbf{g}, \mathbf{h}$  have iid standard normal entries. Then the following hold:

- (i) For all  $c \in \mathbb{R}$

$$\mathbb{P}\{\Phi(\mathbf{G}) < c\} \leq 2\mathbb{P}\{\phi(\mathbf{g}, \mathbf{h}) \leq c\}.$$

- (ii) If  $\psi(\mathbf{v}, \mathbf{u})$  is additionally convex-concave on  $\mathcal{S}_{\mathbf{v}} \times \mathcal{S}_{\mathbf{u}}$  where  $\mathcal{S}_{\mathbf{v}}$  and  $\mathcal{S}_{\mathbf{u}}$  are convex, then for all  $c \in \mathbb{R}$

$$\mathbb{P}\{\Phi(\mathbf{G}) > c\} \leq 2\mathbb{P}\{\phi(\mathbf{g}, \mathbf{h}) \geq c\}.$$

In particular, for all  $\mu \in \mathbb{R}$  and  $t > 0$ ,

$$\mathbb{P}\{|\Phi(\mathbf{G}) - \mu| > t\} \leq 2\mathbb{P}\{|\phi(\mathbf{g}, \mathbf{h}) - \mu| \geq t\}.$$

### Appendix B. Lemmata for Section 3.

LEMMA B.1 ([22, Lemma 9.2] with explicit dependence on  $\epsilon$ ). *Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  be a random matrix with iid standard normal entries, and  $m < n$ . Let further  $\mathbf{y} \in \mathbb{R}^m$  and consider the solution of an  $\ell^2$ -lasso with a regularizer  $f$  which is  $L$ -Lipschitz with respect to the  $\ell^2$ -norm:*

$$\mathbf{x}^* \stackrel{\text{def}}{=} \arg \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x}).$$

Then for any  $0 \leq \epsilon < 1$  and  $0 < \lambda < \frac{\sqrt{n}-\sqrt{m}}{L}(1-\epsilon)$  we have  $\mathbf{y} = \mathbf{A}\mathbf{x}^*$  with probability at least  $1 - e^{-\epsilon^2(\sqrt{n}-\sqrt{m})^2/2}$ , that is,  $\ell^2$ -lasso gives the same optimizer as equality-constrained minimization.

*Proof.* Using [32, Corollary 5.35]<sup>11</sup>, we have for any  $\epsilon \geq 0$

$$(B.21) \quad \mathbb{P}[\sigma_{\min}(\mathbf{A}^\top)/(\sqrt{n}-\sqrt{m}) \leq 1-\epsilon] \leq e^{-\epsilon^2(\sqrt{n}-\sqrt{m})^2/2}.$$

Let  $\mathbf{p} \stackrel{\text{def}}{=} \mathbf{y} - \mathbf{A}\mathbf{x}^*$  and  $\mathbf{w} \stackrel{\text{def}}{=} \mathbf{A}^\dagger \mathbf{p}$ , where  $\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}$  denotes the MPP ( $\mathbf{A}\mathbf{A}^\top$  is almost surely invertible). Since  $\|\mathbf{w}\|_2^2 = \mathbf{p}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{p} \leq \|\mathbf{p}\|_2^2 \sigma_{\min}^{-2}(\mathbf{A}^\top)$  we have from (B.21) that for any  $0 < \epsilon < 1$ , with probability at least  $1 - e^{-\epsilon^2(\sqrt{n}-\sqrt{m})^2/2}$ ,  $\|\mathbf{w}\|_2 \leq \frac{\|\mathbf{p}\|_2}{(\sqrt{n}-\sqrt{m})(1-\epsilon)}$ . Let  $\mathbf{x}^\circ \stackrel{\text{def}}{=} \mathbf{x}^* + \mathbf{w}$  so that  $\mathbf{y} - \mathbf{A}\mathbf{x}^\circ = \mathbf{0}$ . Optimality of  $\mathbf{x}^*$  gives

$$(B.22) \quad (*) = [\|\mathbf{y} - \mathbf{A}\mathbf{x}^*\|_2 + \lambda f(\mathbf{x}^*)] - [\|\mathbf{y} - \mathbf{A}\mathbf{x}^\circ\|_2 + \lambda f(\mathbf{x}^\circ)] \leq 0.$$

On the other hand,

$$\begin{aligned} (*) &= \|\mathbf{p}\|_2 + \lambda f(\mathbf{x}^*) - \lambda f(\mathbf{x}^\circ) \geq \|\mathbf{p}\|_2 - \lambda L \|\mathbf{x}^* - \mathbf{x}^\circ\|_2 \\ &= \|\mathbf{p}\|_2 - \lambda L \|\mathbf{w}\|_2 \geq \|\mathbf{p}\|_2 \left(1 - \frac{\lambda L}{(\sqrt{n}-\sqrt{m})(1-\epsilon)}\right), \end{aligned}$$

where, by (B.22), the last expression must be non-positive. But if we choose  $\lambda < \frac{\sqrt{n}-\sqrt{m}}{L}(1-\epsilon)$ , the only way to make it non-positive is that  $\|\mathbf{p}\|_2 = 0$ .  $\square$

LEMMA B.2. *Let  $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ ,  $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ ,  $1 \leq p \leq \infty$ , and define*

$$\Delta_p(\beta; \mathbf{h}, \lambda) \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \text{dist}(\beta \mathbf{h}, \|\cdot\|_{p^*} \leq \lambda) \quad \text{and} \quad \Delta_p(\beta; n, \lambda) \stackrel{\text{def}}{=} \mathbb{E}[\Delta_p(\beta; \mathbf{h}, \lambda)].$$

There exist universal constants  $c_1, c_2 > 0$  such that for any  $0 < \epsilon < 2$ , any integers  $m, n$  and any  $A > 0$  we have, with  $\delta \stackrel{\text{def}}{=} (m-1)/n$  and  $\phi(\alpha, \beta; \mathbf{g}, \mathbf{h})$  defined as in (3.17):

$$\mathbb{P}\{\exists 0 < \alpha \leq A, 0 < \beta < 1, |\phi(\alpha, \beta; \mathbf{g}, \mathbf{h}) - \kappa(\alpha, \beta)| \geq \epsilon\} \leq \frac{c_1}{\epsilon} e^{-c_2 \epsilon^2 n \frac{\min(\delta, 1/A^2)}{1+\delta A^2}} \stackrel{\text{def}}{=} \zeta(n, \epsilon; A, \delta),$$

where

$$(B.23) \quad \kappa(\alpha, \beta) = \kappa_p(\alpha, \beta; n, \delta, \lambda) \stackrel{\text{def}}{=} \beta \sqrt{\delta \alpha^2 + 1} - \alpha \Delta_p(\beta; n, \lambda).$$

*Proof.* We first look at the term  $\|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2$ . Partitioning  $\mathbf{g}$  as  $[g_0, \tilde{\mathbf{g}}^\top]^\top$ ,  $\tilde{\mathbf{g}} \in \mathbb{R}^{m-1}$ ,

$$\|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2^2 = \|\alpha \tilde{\mathbf{g}}/\sqrt{n}\|_2^2 + (\alpha g_0/\sqrt{n} - 1)^2.$$

<sup>11</sup>We actually use a one-sided variant of [32, Corollary 5.35] which can be obtained by combining Lemma A.1(d) with the estimate of the expectation of  $\sigma_{\min}$ , [32, Theorem 5.32].

Using Lemma A.1(b) we get for the first term:

$$\mathbb{P} \left\{ \exists 0 < \alpha \leq A, \quad \|\alpha \tilde{\mathbf{g}}/\sqrt{n}\|_2^2 \notin [(1-\epsilon)\delta\alpha^2, \delta\alpha^2/(1-\epsilon)] \right\} \leq 2e^{-\frac{\epsilon^2(m-1)}{4}} = 2e^{-\frac{\epsilon^2 n \delta}{4}},$$

since the event  $\{\|\alpha \tilde{\mathbf{g}}/\sqrt{n}\| \notin [(1-\epsilon)\delta\alpha^2, \delta\alpha^2/(1-\epsilon)]\}$  does not depend on  $\alpha > 0$ .

Next, we show that the term  $(\alpha g_0/\sqrt{n} - 1)^2$  cannot deviate much from 1: setting  $\epsilon' = \epsilon/2$ , we have  $\sqrt{1-\epsilon} < 1-\epsilon'$  and  $1+\epsilon' < 1/\sqrt{1-\epsilon}$  hence

$$\begin{aligned} \mathbb{P} \left\{ \forall 0 < \alpha \leq A, \quad \left( \frac{\alpha g_0}{\sqrt{n}} - 1 \right)^2 \in [1-\epsilon, 1/(1-\epsilon)] \right\} &\geq \mathbb{P} \left\{ \forall 0 < \alpha \leq A, \quad \left| \frac{\alpha g_0}{\sqrt{n}} - 1 \right| \in [1-\epsilon', 1+\epsilon'] \right\} \\ &\geq \mathbb{P} \left\{ \forall 0 < \alpha \leq A, \quad \alpha g_0/\sqrt{n} \in [-\epsilon', \epsilon'] \right\} \\ &= \mathbb{P} \left\{ A g_0/\sqrt{n} \in [-\epsilon', \epsilon'] \right\}. \end{aligned}$$

Then, with erfc being the complementary error function and using  $\text{erfc}(z) \leq \exp(-z^2)$  [6],

$$\begin{aligned} \mathbb{P} \left\{ \exists 0 < \alpha \leq A, \quad \left( \frac{\alpha g_0}{\sqrt{n}} - 1 \right)^2 \notin [1-\epsilon, 1/(1-\epsilon)] \right\} &\leq \mathbb{P}(|g_0| > \frac{\epsilon' \sqrt{n}}{A}) = \text{erfc} \left[ \frac{\epsilon' \sqrt{n}}{\sqrt{2}A} \right] \\ &\leq e^{-\frac{\epsilon'^2 n}{2A^2}} = e^{-\frac{\epsilon^2 n}{8A^2}}. \end{aligned}$$

Combining the above, we get that: for any  $0 < \epsilon < 1$ , setting  $\epsilon' = 1 - (1-\epsilon)^2 = \epsilon(2-\epsilon) \geq \epsilon$ ,

$$\begin{aligned} &\mathbb{P} \left\{ \exists 0 < \alpha \leq A, \quad \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2 \notin [(1-\epsilon)\sqrt{\delta\alpha^2+1}, \sqrt{\delta\alpha^2+1}/(1-\epsilon)] \right\} \\ &= \mathbb{P} \left\{ \exists 0 < \alpha \leq A, \quad \|\alpha \mathbf{g}/\sqrt{n} - \mathbf{e}_1\|_2^2 \notin [(1-\epsilon')(\delta\alpha^2+1), (\delta\alpha^2+1)/(1-\epsilon')] \right\} \\ \text{(B.24)} \quad &\leq 2e^{-\frac{\epsilon'^2 n \delta}{4}} + e^{-\frac{\epsilon'^2 n}{8A^2}} \leq c'_1 e^{-c'_2 \epsilon'^2 n \cdot \min(\delta, 1/A^2)} \leq c'_1 e^{-c'_2 \epsilon^2 n \cdot \min(\delta, 1/A^2)}, \end{aligned}$$

where the constants  $c'_1, c'_2$  are universal.

For the second term in  $\phi(\alpha, \beta; \mathbf{g}, \mathbf{h})$ , we note that Euclidean distance to a convex set  $\text{dist}(\mathbf{x}, \mathcal{C})$  is 1-Lipschitz in  $\mathbf{x}$  with respect to  $\|\cdot\|_2$  so by Lemma A.1(e) we get for  $0 < \beta \leq 1$  (we omit the dependency in  $\lambda$  for brevity) and any  $\tau \geq 0$  that

$$\text{(B.25)} \quad \mathbb{P} \{ |\Delta_p(\beta; \mathbf{h}) - \Delta_p(\beta)| > \tau \} \leq 2e^{-\frac{\tau^2 n}{2\beta^2}} \leq 2e^{-\frac{\tau^2 n}{2}}.$$

This obviously extends to  $\beta = 0$  since  $\Delta_p(0; \mathbf{h}) = \Delta_p(0) = 0$ .

Next, we want to bound

$$\mathbb{P} \{ \exists \beta \in [0, 1], \quad |\Delta_p(\beta; \mathbf{h}) - \Delta_p(\beta)| > \tau \} = \mathbb{P} \left\{ \sup_{\beta \in [0, 1]} |\Delta_p(\beta; \mathbf{h}) - \Delta_p(\beta)| > \tau \right\}.$$

By Lemma B.14, the function  $\beta \mapsto f_{\mathbf{h}}(\beta) \stackrel{\text{def}}{=} |\Delta_p(\beta; \mathbf{h}) - \Delta_p(\beta)|$  is  $L_{\mathbf{h}}$ -Lipschitz in  $\beta$  with  $L_{\mathbf{h}} \stackrel{\text{def}}{=} \max\{\|\mathbf{h}\|_2/\sqrt{n}, 1\}$ . Hence  $f_{\mathbf{h}}$  is continuous, and its supremum on the closed interval  $[0, 1]$  is indeed a maximum reached at some maximizer  $\beta_{\mathbf{h}}$ .

Let  $b \leq 1/2$  and  $Y_b = \{b\tau/2, 3b\tau/2, \dots, (k-1/2)b\tau\}$  be a uniform sampling of  $[0, 1]$  with spacing  $b\tau$ , with the last segment possibly being shorter. For a given  $\mathbf{h}$ , there exists  $y_{\mathbf{h}} \in Y_b$  such that  $|\beta_{\mathbf{h}} - y_{\mathbf{h}}| \leq b\tau$ . For this  $y_{\mathbf{h}}$  we write

$$f_{\mathbf{h}}(\beta_{\mathbf{h}}) - f_{\mathbf{h}}(y_{\mathbf{h}}) \leq |f_{\mathbf{h}}(\beta_{\mathbf{h}}) - f_{\mathbf{h}}(y_{\mathbf{h}})| \leq L_{\mathbf{h}} |\beta_{\mathbf{h}} - y_{\mathbf{h}}| \leq L_{\mathbf{h}} b\tau$$

so that

$$\begin{aligned} \mathbb{P}\left\{\sup_{\beta \in [0,1]} f_{\mathbf{h}}(\beta) > \tau\right\} &= \mathbb{P}\{f_{\mathbf{h}}(\beta_{\mathbf{h}}) > \tau\} \leq \mathbb{P}\{f_{\mathbf{h}}(y_{\mathbf{h}}) + L_{\mathbf{h}}b\tau > \tau\} \leq \mathbb{P}\left\{f_{\mathbf{h}}(y_{\mathbf{h}}) > \frac{\tau}{2} \text{ or } L_{\mathbf{h}}b\tau > \frac{\tau}{2}\right\} \\ &\leq \mathbb{P}\{f_{\mathbf{h}}(y_{\mathbf{h}}) > \tau/2\} + \mathbb{P}\{\max\{\|\mathbf{h}\|_2/\sqrt{n}, 1\}b > 1/2\} \end{aligned}$$

As we do not know a priori to which  $y \in Y_b$  the maximizer  $\beta_{\mathbf{h}}$  will be close, we continue with a union bound, and we further use that  $b \leq 1/2$  to obtain by (B.25) and Lemma A.1(a)

$$\begin{aligned} \mathbb{P}\left\{\sup_{\beta \in [0,1]} f_{\mathbf{h}}(\beta) > \tau\right\} &\leq \mathbb{P}\left\{\exists y \in Y_b, f_{\mathbf{h}}(y) > \tau/2\right\} + \mathbb{P}\left\{\|\mathbf{h}\|_2^2 > \frac{n}{4b^2}\right\} \leq |Y_b| 2e^{-\frac{\tau^2 n}{8}} + e^{-\frac{(1-4b^2)^2 n}{4}} \\ &\leq (1 + 1/(b\tau))2e^{-\frac{\tau^2 n}{8}} + e^{-\frac{(1-4b^2)^2 n}{4}}. \end{aligned}$$

Setting  $b \stackrel{\text{def}}{=} \frac{1}{2}\sqrt{\frac{2-\sqrt{2}}{2}} \approx 0.27$  we get that for any  $0 < \tau < 1$  we have  $\frac{(1-4b^2)^2}{4} = \frac{1}{8} \geq \frac{\tau^2}{8}$ , hence

$$(B.26) \quad \mathbb{P}\left\{\sup_{\beta \in [0,1]} f_{\mathbf{h}}(\beta) > \tau\right\} \leq (3 + 2/(b\tau))e^{-\frac{\tau^2 n}{8}} < \frac{11}{\tau}e^{-\frac{\tau^2 n}{8}}, \quad 0 < \tau < 1.$$

To conclude we combine concentration bounds for both terms. First, we observe that

$$\sup_{\substack{0 \leq \alpha \leq A \\ 0 \leq \beta \leq 1}} |\phi(\alpha, \beta; \mathbf{g}, \mathbf{h}) - \kappa(\alpha, \beta)| \leq \sup_{0 \leq \alpha \leq A} \left\| \frac{\alpha \mathbf{g}}{\sqrt{n}} - \mathbf{e}_1 \right\|_2 - \sqrt{1 + \delta \alpha^2} + A \sup_{0 \leq \beta \leq 1} |\Delta_p(\beta; \mathbf{h}) - \Delta_p(\beta)|$$

and by a union bound we just need to control the probability that each term exceeds  $\epsilon/2$ . Since we assume that  $0 < \epsilon < 2$ , we can use the multiplicative control (B.24) as follows:

$$\begin{aligned} \mathbb{P}\left\{\exists \alpha \in (0, A], \left| \left\| \frac{\alpha \mathbf{g}}{\sqrt{n}} - \mathbf{e}_1 \right\|_2 - \sqrt{\delta \alpha^2 + 1} \right| > \frac{\epsilon}{2}\right\} &= \mathbb{P}\left\{\exists \alpha \in (0, A], \left| \frac{\left\| \frac{\alpha \mathbf{g}}{\sqrt{n}} - \mathbf{e}_1 \right\|_2}{\sqrt{\delta \alpha^2 + 1}} - 1 \right| > \frac{\epsilon}{2\sqrt{\delta \alpha^2 + 1}}\right\} \\ &\leq \mathbb{P}\left\{\exists \alpha \in (0, A], \left| \frac{\left\| \frac{\alpha \mathbf{g}}{\sqrt{n}} - \mathbf{e}_1 \right\|_2}{\sqrt{\delta \alpha^2 + 1}} - 1 \right| > \frac{\epsilon}{2\sqrt{\delta A^2 + 1}}\right\} \\ &\leq \mathbb{P}\left\{\exists \alpha \in (0, A], \frac{\left\| \frac{\alpha \mathbf{g}}{\sqrt{n}} - \mathbf{e}_1 \right\|_2}{\sqrt{\delta \alpha^2 + 1}} \notin [1 - \epsilon', \frac{1}{1 - \epsilon'}]\right\} \\ &\leq c'_1 e^{-c'_2 \epsilon'^2 n \cdot \min(\delta, 1/A^2)}. \end{aligned}$$

provided that  $[1 - \epsilon', 1/(1 - \epsilon')] \subset [1 - \frac{\epsilon}{2\sqrt{\delta A^2 + 1}}, 1 + \frac{\epsilon}{2\sqrt{\delta A^2 + 1}}]$ . This is achieved with  $\epsilon' = \epsilon/(1 + \sqrt{\delta A^2 + 1})$ . Combining the resulting bound with the bound on  $\mathbb{P}\{\sup_{\beta \in [0,1]} f_{\mathbf{h}}(\beta) > \frac{\epsilon}{2A}\}$  from (B.26) yields the result.  $\square$

COROLLARY B.3. Consider  $A > 0$  and define for any set  $S \subseteq [0, A]$ :

$$\begin{aligned} \phi(\alpha; \mathbf{g}, \mathbf{h}) &\stackrel{\text{def}}{=} \sup_{0 \leq \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}), & \kappa(\alpha) &\stackrel{\text{def}}{=} \sup_{0 \leq \beta \leq 1} \kappa(\alpha, \beta), \\ \phi_S(\mathbf{g}, \mathbf{h}) &\stackrel{\text{def}}{=} \inf_{\alpha \in S} \phi(\alpha; \mathbf{g}, \mathbf{h}), & \kappa_S &\stackrel{\text{def}}{=} \inf_{\alpha \in S} \kappa(\alpha). \end{aligned}$$

With  $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ ,  $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ ,  $\delta = (m - 1)/n$ , we have for  $0 < \epsilon < 2$  that

$$(B.27) \quad \mathbb{P}\left\{\sup_{S \subseteq [0, A]} |\phi_S(\mathbf{g}, \mathbf{h}) - \kappa_S| \geq \epsilon\right\} \leq \zeta(n, \epsilon).$$

with  $\zeta(n, \epsilon) = \zeta(n, \epsilon; A, \delta)$  defined in Lemma B.2. In particular, letting  $\phi(\alpha) \stackrel{\text{def}}{=} \phi(\alpha; \mathbf{g}, \mathbf{h})$ ,

$$(B.28) \quad \mathbb{P} \left\{ \sup_{0 \leq \alpha \leq A} |\phi(\alpha) - \kappa(\alpha)| \geq \epsilon \right\} \leq \zeta(n, \epsilon).$$

*Proof.* To lighten notation we suppress the dependence of the stochastic function  $\phi$  on random vectors  $\mathbf{g}$  and  $\mathbf{h}$ . By Lemma B.2 we have with probability at least  $1 - \zeta(n, \epsilon)$ : for all  $0 \leq \alpha \leq A$  and  $0 \leq \beta \leq 1$ ,  $|\phi(\alpha, \beta) - \kappa(\alpha, \beta)| \leq \epsilon$ . When this holds we have for any  $S \subset [0, A]$ :

$$(B.29) \quad \phi_S = \inf_{\alpha \in S} \phi(\alpha) \leq \inf_{\alpha \in S} [\kappa(\alpha) + \epsilon] = \kappa_S + \epsilon$$

$$(B.30) \quad \phi_S = \inf_{\alpha \in S} \phi(\alpha) \geq \inf_{\alpha \in S} [\kappa(\alpha) - \epsilon] = \kappa_S - \epsilon. \quad \square$$

We will shortly characterize  $\kappa(\alpha, \beta)$  and  $\kappa(\alpha)$  using properties of the following quantity:

$$(B.31) \quad D_p(t; n) \stackrel{\text{def}}{=} \left( \mathbb{E} \left[ \frac{1}{\sqrt{n}} \text{dist}(\mathbf{h}, \|\cdot\|_{p^*} \leq t) \right] \right)^2.$$

LEMMA B.4 (Deterministic properties of  $D_p$ ). *Let  $t \geq 0$ ,  $1 \leq p \leq \infty$ , and define  $\mathcal{C}_t = \mathcal{C}_{t,p} \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_{p^*} \leq t\}$  and  $D_p(t; n)$  as in (B.31). Using  $D_p(t)$  as a shorthand, the following hold:*

1. The sets  $\mathcal{C}_t$  are convex and nested with  $\mathcal{C}_t \subsetneq \mathcal{C}_{t'}$  for  $t < t'$ ;
2. For any vector  $\mathbf{h}$ , the function  $t \mapsto \text{dist}(\mathbf{h}, \mathcal{C}_t)$  is non-increasing and convex;
3.  $D_p(t)$  is a (strictly) decreasing convex function of  $t$ ;
4.  $\lim_{t \rightarrow \infty} D_p(t) = 0$ ;
5.  $\frac{n}{n+1} \leq D_p(0) \leq 1$ ;
6. The function  $t \mapsto D_p(t)$  is infinitely differentiable;
7. Let  $g(t) = g(t; n) \stackrel{\text{def}}{=} D_p(t) - \frac{t}{2} D_p'(t)$ . For any  $0 < \delta < D_p(0)$  there is a unique

$$t^* = t_p^*(\delta; n) \in (0, \infty)$$

such that  $g(t) > \delta$  for  $t < t^*$  and  $g(t) < \delta$  for  $t > t^*$ . It holds that  $D_p(t^*) < g(t^*) = \delta$ .

- Proof.* 1. Obvious. 2. We recall that  $\mathbf{x} \mapsto \text{dist}(\mathbf{x}, \mathcal{C})$  is convex for convex  $\mathcal{C}$  [3, Example 3.16]. Next,  $(\mathbf{x}, t) \mapsto t \text{dist}(\mathbf{x}/t, \mathcal{C})$  is convex in both arguments because it is the perspective of  $\mathbf{x} \mapsto \text{dist}(\mathbf{x}, \mathcal{C})$  [3, Chapter 2]. Applying this to  $\mathcal{C} = \mathcal{C}_1$  and observing that  $\text{dist}(\mathbf{h}, \mathcal{C}_t) = t \text{dist}(\mathbf{h}/t, \mathcal{C}_1)$  we obtain that  $\text{dist}(\mathbf{h}, \mathcal{C}_t)$  is convex in  $t$ . The fact that it is non-increasing follows from Property 1.
3. Since expectation of convex functions is convex, and the square of a non-negative convex function is convex,  $D_p(t)$  is convex as claimed. That it is (strictly) decreasing is obvious.
  4. For any  $\mathbf{y} \in \mathbb{R}^n$  and  $p \geq 1$ ,  $\|\mathbf{y}\|_{p^*} \leq \|\mathbf{y}\|_1 \leq \sqrt{n} \|\mathbf{y}\|_2$ . Hence, for any given  $t > 0$ ,  $\text{dist}(\mathbf{y}, \mathcal{C}_t) = 0$  as soon as  $\|\mathbf{y}\|_2 \leq t/\sqrt{n}$  and we can write for all  $p \geq 1$

$$\begin{aligned} \mathbb{E}[\text{dist}(\mathbf{h}, \mathcal{C}_t)] &= \int_{\mathbf{b} \in \mathbb{S}^{n-1}} \int_{r=t/\sqrt{n}}^{\infty} \text{dist}(r\mathbf{b}, \mathcal{C}_t) p_{\mathbf{h}}(r\mathbf{b}) \mu(d\mathbf{b}) r^{n-1} dr \\ &\leq \int_{\mathbf{b} \in \mathbb{S}^{n-1}} \int_{r=t/\sqrt{n}}^{\infty} r p_{\mathbf{h}}(r\mathbf{b}) \mu(d\mathbf{b}) r^{n-1} dr = Z_n \mu(\mathbb{S}^{n-1}) \int_{r=t/\sqrt{n}}^{\infty} r^n e^{-r^2/2} dr \end{aligned}$$

where  $Z_n = (2\pi)^{-n/2}$ ,  $\mu(\mathbb{S}^{n-1}) = 2\pi^{n/2}/\Gamma(\frac{n}{2})$ , and  $\Gamma(\cdot)$  being the gamma function. The last expression vanishes as  $t \rightarrow \infty$ .



5. The upper bound follows from Property 3 and Jensen's inequality. To get the lower bound we compute  $nD_p(0) = [\mathbb{E}\|\mathbf{h}\|_2]^2$  by integration in polar coordinates,

$$\begin{aligned}\mathbb{E}\{\text{dist}(\mathbf{h}, \mathcal{C}_0)\} &= \mathbb{E}\{\|\mathbf{h}\|_2\} = \int_{\mathbf{b} \in \mathbb{S}^{n-1}} \int_{r=0}^{\infty} r p_{\mathbf{h}}(r\mathbf{b}) \mu(d\mathbf{b}) r^{n-1} dr \\ &= Z_n \mu(\mathbb{S}^{n-1}) \int_{r=0}^{\infty} r^n e^{-r^2/2} dr \stackrel{(a)}{=} (2\pi)^{-n/2} \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})} \int_{s=0}^{\infty} 2^{\frac{n-1}{2}} s^{\frac{n-1}{2}} e^{-s} ds \\ &\stackrel{(b)}{=} \sqrt{2} \Gamma\left(\frac{n+1}{2}\right) / \Gamma\left(\frac{n}{2}\right),\end{aligned}$$

where in (a) we used the substitution  $u = r^2/2$ , and in (b) we invoked the definition of the gamma function,  $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$ . We now use the inequality of Wendel, [35, Eq. (7)],  $\Gamma(x+a)/\Gamma(x) \geq x(x+a)^{a-1}$  to conclude that

$$(B.32) \quad \mathbb{E}\{\|\mathbf{h}\|_2\} \geq n(n+1)^{-1/2}$$

and  $D_p(0) \geq n/(n+1)$ .

6. Using  $\text{dist}(\mathbf{h}, \mathcal{C}_t) = t \text{dist}(\mathbf{h}/t, \mathcal{C}_1)$  and a change of variable  $r = t\rho$  we obtain:

$$(B.33) \quad q(t) \stackrel{\text{def}}{=} \sqrt{nD_p(t)} = Z_n t^{n+1} \int_{\mathbf{b} \in \mathbb{S}^{n-1}} \int_{\rho=0}^{\infty} \text{dist}(\rho\mathbf{b}, \mathcal{C}_1) e^{-\rho^2 t^2/2} \mu(d\mathbf{b}) \rho^{n-1} d\rho$$

so that  $D_p(t)$  is infinitely differentiable (by the dominated convergence theorem).

7. In particular, as  $\left| \frac{\partial}{\partial t} \text{dist}(\rho\mathbf{b}, \mathcal{C}_1) e^{-\rho^2 t^2/2} \rho^{n-1} \right| = \text{dist}(\rho\mathbf{b}, \mathcal{C}_1) t \rho^{n+1} e^{-\rho^2 t^2/2} \leq \rho^{n+2} t e^{-\rho^2 t^2/2}$  where the rightmost expression is integrable for every  $t > 0$  and  $n \in \mathbb{N}$ , we can differentiate under the integral sign in (B.33) to get

$$(B.34) \quad \begin{aligned}-\frac{t}{2} D'_p(t) &= -\frac{t}{n} q(t) q'(t) \\ &= -\frac{1}{n} q(t) \left( (n+1)q(t) - Z_n t^{n+3} \int_{\mathbb{S}^{n-1}} \int_0^{\infty} \text{dist}(\rho\mathbf{b}, \mathcal{C}_1) \rho^{n+1} e^{-\rho^2 t^2/2} \mu(d\mathbf{b}) d\rho \right).\end{aligned}$$

All terms can be seen to vanish as  $t \rightarrow \infty$  by arguments analogous to those in the end of the proof of Property 4, hence  $\lim_{t \rightarrow \infty} [-\frac{t}{2} D'_p(t)] = 0$ .

Since  $D_p(t)$  is strictly decreasing we have  $D'_p(t) < 0$ . Since it is convex,  $D''_p(t) \geq 0$ . Thus

$$g'(t) = D'_p(t) - \frac{1}{2} D'_p(t) - \frac{t}{2} D''_p(t) = \frac{1}{2} D'_p(t) - \frac{t}{2} D''_p(t) < 0$$

for  $t > 0$ , meaning that  $g(t)$  is strictly decreasing. Since  $\lim_{t \rightarrow \infty} [-\frac{t}{2} D'_p(t)] = 0$  and  $\lim_{t \rightarrow \infty} D_p(t) = 0$ , we have  $\lim_{t \rightarrow \infty} g(t) = 0$ . It follows that for  $0 < \delta < \lim_{t \rightarrow 0} g(t)$ , there is a unique  $t^*(\delta)$ , such that  $0 < t^*(\delta) < \infty$  and  $g(t) > \delta$  for  $t < t^*$  and  $g(t) < \delta$  for  $t > t^*$ . Finally, since  $g(t) \geq D_p(t)$  we have  $\lim_{t \rightarrow 0} g(t) \geq D_p(0)$ .  $\square$

LEMMA B.5. *Denote*

$$(B.35) \quad \theta(t) \stackrel{\text{def}}{=} \mathbb{E}(|h| - t)_+^2,$$

with  $h$  being a standard normal variable and  $(\cdot)_+ = \max(\cdot, 0)$  the positive part. This is a strictly decreasing function of  $t$  with  $\theta(0) = 1$  and  $\lim_{t \rightarrow \infty} \theta(t) = 0$ . Recalling that  $t_p^*(\delta; n)$  is the unique solution to  $D_p(t; n) - \frac{t}{2} D'_p(t; n) = \delta$  on  $(0, \infty)$ , the following holds for all  $1 \leq p \leq \infty$ ,  $0 < \delta < 1$ :

1. For any  $n \geq 1$ ,

$$(B.36) \quad t_p^*(\delta; n) \leq 2 \theta^{-1}(\delta) n^{1-1/p}.$$

2. For any  $n \geq \frac{2}{1-\delta}$ ,

$$(B.37) \quad t_p^*(\delta; n) \geq \theta^{-1}\left(\frac{1+\delta}{2}\right) > 0.$$

3. For any  $n \geq \max\left(\frac{2}{1-\delta}, \frac{1}{\delta}\right)$ ,

$$(B.38) \quad t_p^*(\delta; n) \geq \max\left(\theta^{-1}(2\delta), \theta^{-1}\left(\frac{1+\delta}{2}\right)\right) > 0.$$

4. There is a universal constant  $C$  independent of  $\delta$ ,  $p$  and  $n$  such that for all  $n \geq 1$  we have

$$(B.39) \quad D_p(t_p^*; n) \geq (\delta/C)^2,$$

where we use the shorthand  $t_p^* = t_p^*(\delta; n)$ .

LEMMA B.6. With the notations of Theorem 2.2, for  $p = 2$ ,

$$(B.40) \quad \lim_{n \rightarrow \infty} -t_2^* D_2'(t_2^*; n) = (1-\delta)\delta > 0, \quad \lim_{n \rightarrow \infty} \alpha_2^*(\delta; n) = \frac{1}{\sqrt{1-\delta}}.$$

LEMMA B.7. With the notations of Theorem 2.2 for  $p = 1$ , denoting  $t_1^*(\delta) \stackrel{\text{def}}{=} \sqrt{2} \cdot \text{erfc}^{-1}(\delta)$ ,

$$\lim_{n \rightarrow \infty} -t_1^* D_1'(t_1^*; n) > 0, \quad \alpha_1^*(\delta) \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \alpha_1^*(\delta; n) = \sqrt{\left(\sqrt{\frac{2}{\pi}} e^{-\frac{(t_1^*)^2}{2}} t_1^* - \delta(t_1^*)^2\right)^{-1} - \frac{1}{\delta}}.$$

*Proof of Lemmata B.5–B.6–B.7.*

• **Step 1.** We show that for all  $p, n, t$  we have

$$(B.41) \quad D_p(t; n) \geq D_1(t; n) \geq D_p\left(tn^{1-1/p}; n\right).$$

The inequalities  $\|\cdot\|_{p^*} \geq \|\cdot\|_\infty \geq n^{-1/p^*} \|\cdot\|_{p^*} = n^{-(1-1/p)} \|\cdot\|_{p^*}$  imply  $\mathcal{C}_{t,p} \subset \mathcal{C}_{t,1} \subset \mathcal{C}_{tn^{1-1/p},p}$ . It follows that  $d(\cdot, \mathcal{C}_{t,p}) \geq d(\cdot, \mathcal{C}_{t,1}) \geq d(\cdot, \mathcal{C}_{tn^{1-1/p},p})$ , which yields (B.41).

• **Step 2.** We establish that for any  $p, n, \delta$ , with the shorthand  $t_p^* = t_p^*(\delta; n)$ ,

$$(B.42) \quad D_p(t_p^*/2; n) \geq \delta,$$

$$(B.43) \quad D_p(t_p^*; n) \leq \delta.$$

With another shorthand  $D_p(t) \stackrel{\text{def}}{=} D_p(t; n)$  and the convexity of  $D_p(t)$ , we have for all  $t, h$

$$(B.44) \quad D_p(t+h) \geq D_p(t) + hD_p'(t).$$

Applying it to  $t = t_p^*$  and  $h = -t_p^*/2$  and using the definition of  $t_p^*$ , we get

$$D_p(t_p^*/2) \geq D_p(t_p^*) - (t_p^*/2)D_p'(t_p^*) = \delta,$$

i.e., (B.42) holds. Since  $D_p(t)$  is non-increasing, we have  $D_p'(t) \leq 0$  and  $D_p(t) - \frac{t}{2}D_p'(t) \geq D_p(t)$  for any  $t$ . Applying to  $t = t_p^*$  this establishes (B.43) by definition of  $t_p^*$ .

- **Step 3.** Let  $\bar{D}_p(t; n) \stackrel{\text{def}}{=} \frac{1}{n} \mathbb{E} \text{dist}^2(\mathbf{h}, \|\cdot\|_{p^*} \leq t)$ . By Jensen's inequality, for any  $p, n, t$ :

$$(B.45) \quad D_p(t; n) = \frac{1}{n} (\mathbb{E} \text{dist}(\mathbf{h}, \|\cdot\|_{p^*} \leq t))^2 \leq \frac{1}{n} \mathbb{E} \text{dist}^2(\mathbf{h}, \|\cdot\|_{p^*} \leq t) = \bar{D}_p(t; n).$$

For  $p = 1$ ,  $p^* = \infty$  we compute

$$(B.46) \quad \bar{D}_1(t) \stackrel{\text{def}}{=} \frac{1}{n} \mathbb{E} \text{dist}^2(\mathbf{h}, \|\cdot\|_\infty \leq t) = \frac{1}{n} \mathbb{E} \left\| \mathbf{h} - \text{proj}_{\|\cdot\|_\infty \leq t} \mathbf{h} \right\|_2^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} (|h_i| - t)_+^2 = \theta(t).$$

- **Step 4.** We get (B.36) by combining the previous steps. For any  $p, \delta, n$ :

$$\theta(t_p^* n^{1/p-1}/2) \stackrel{(B.46)}{=} \bar{D}_1(t_p^* n^{1/p-1}/2) \stackrel{(B.45)}{\geq} D_1(t_p^* n^{1/p-1}/2) \stackrel{(B.41)}{\geq} D_p(t_p^*/2) \stackrel{(B.42)}{\geq} \delta.$$

- **Step 5.** Since the function  $f : \mathbf{h} \mapsto f(\mathbf{h}) \stackrel{\text{def}}{=} \text{dist}(\mathbf{h}, \|\cdot\|_{p^*} \leq t)$  is 1-Lipschitz, and  $\mathbf{h}$  standard normal, we can apply Lemma A.1(f) to show that for any  $p, n, t$ ,

$$(B.47) \quad D_p(t; n) \geq \bar{D}_p(t; n) - \frac{1}{n}.$$

- **Step 6.** Combining with the previous steps yields for any  $p, \delta$ , and  $n > \frac{1}{1-\delta}$ :

$$\theta(t_p^*) \stackrel{(B.46)}{=} \bar{D}_1(t_p^*) \stackrel{(B.47)}{\leq} D_1(t_p^*) + \frac{1}{n} \stackrel{(B.41)}{\leq} D_p(t_p^*) + \frac{1}{n} \stackrel{(B.43)}{\leq} \delta + \frac{1}{n} < 1.$$

For  $n \geq \frac{2}{1-\delta}$ , we have  $\delta + \frac{1}{n} \leq \frac{1+\delta}{2} < 1$  which yields (B.37). For  $n \geq \max(\frac{2}{1-\delta}, 1/\delta)$ , we have  $\delta + \frac{1}{n} \leq \min(\frac{1+\delta}{2}, 2\delta) < 1$  which yields (B.38).

- **Step 7.** To establish (B.39) we start with the expression (B.34) in Cartesian coordinates,

$$(B.48) \quad \begin{aligned} -\frac{t}{2} D_p'(t) &= -\frac{1}{n} q(t) \left( (n+1)q(t) - \int_{\mathbb{R}^n} t \text{dist}(\mathbf{y}/t, \mathcal{C}_1) \|\mathbf{y}\|_2^2 p_{\mathbf{h}}(\mathbf{y}) d\mathbf{y} \right) \\ &\implies -\frac{t}{2} D_p'(t) + \frac{n+1}{n} q^2(t) = \frac{q(t)}{n} \mathbb{E} \{ \text{dist}(\mathbf{h}, \mathcal{C}_t) \|\mathbf{h}\|_2^2 \}. \end{aligned}$$

Since  $\frac{n+1}{n} q^2(t) = (n+1)D_p(t)$  we can rewrite (B.48) using the definition of  $D_p(t)$  as

$$(B.49) \quad D_p(t) - \frac{t}{2} D_p'(t) = \frac{q(t)}{n} \mathbb{E} [\text{dist}(\mathbf{h}, \mathcal{C}_t) (\|\mathbf{h}\|_2^2 - n)].$$

Observing further that  $\mathbb{E}[\|\mathbf{h}\|_2^2 - n] = 0$  and  $\mathbb{E}[\text{dist}(\mathbf{h}, \mathcal{C}_t)] = q(t)$ , the following holds:

$$\begin{aligned} D_p(t) - \frac{t}{2} D_p'(t) &= \frac{q(t)}{n} \mathbb{E} \left\{ (\text{dist}(\mathbf{h}, \mathcal{C}_t) - q(t)) (\|\mathbf{h}\|_2^2 - n) \right\} \\ &\leq \sqrt{D_p(t)} \mathbb{E} \left\{ \left| \text{dist}(\mathbf{h}, \mathcal{C}_t) - q(t) \right| \left| \frac{\|\mathbf{h}\|_2^2 - n}{\sqrt{n}} \right| \right\} \\ &= \sqrt{D_p(t)} \int_0^\infty \mathbb{P} \left\{ \left| \text{dist}(\mathbf{h}, \mathcal{C}_t) - q(t) \right| \left| \frac{\|\mathbf{h}\|_2^2 - n}{\sqrt{n}} \right| \geq \epsilon \right\} d\epsilon. \end{aligned}$$

The integrand can be controlled by a union bound as

$$\mathbb{P} \left\{ \left| \text{dist}(\mathbf{h}, \mathcal{C}_t) - q(t) \right| \left| \frac{\|\mathbf{h}\|_2^2 - n}{\sqrt{n}} \right| \geq \epsilon \right\} \leq \mathbb{P} \left\{ \left| \text{dist}(\mathbf{h}, \mathcal{C}_t) - q(t) \right| \geq \sqrt{\epsilon} \right\} + \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2^2 - n}{\sqrt{n}} \right| \geq \sqrt{\epsilon} \right\}$$

which together with Lemma A.1(e) for the first term and Lemma A.1(c) for the second term yields  $D_p(t) - \frac{t}{2}D'_p(t) \leq C_n\sqrt{D_p(t)}$ . (B.39) follows by definition of  $t_p^*$  and

$$C_n \stackrel{\text{def}}{=} \int_0^\infty 2 \left( e^{-\epsilon/2} + \begin{cases} e^{-\frac{\epsilon}{8}} & \text{for } 0 \leq \epsilon \leq n, \\ e^{-\frac{\sqrt{\epsilon n}}{8}} & \text{for } \epsilon > n \end{cases} \right) d\epsilon \leq \int_0^\infty 2 \left( e^{-\epsilon/2} + e^{-\frac{\epsilon}{8}} + e^{-\frac{\sqrt{\epsilon}}{8}} \right) d\epsilon \stackrel{\text{def}}{=} C < \infty.$$

- **Step 8. (Proof of Lemma B.6).** Since  $\text{dist}(\mathbf{h}, \|\cdot\|_2 \leq t) = (\|\mathbf{h}\|_2 - t)_+$ , we have  $D_2(t; n) = \frac{1}{n}(\mathbb{E}(\|\mathbf{h}\|_2 - t)_+)^2$ . With  $d(t) \stackrel{\text{def}}{=} \sqrt{D_2(t; n)} = \frac{1}{\sqrt{n}}\mathbb{E}(\|\mathbf{h}\|_2 - t)_+$  we have  $d'(t) = -\frac{1}{\sqrt{n}}\mathbb{E}[\mathbb{I}_{\|\mathbf{h}\|_2 > t}] = -\frac{1}{\sqrt{n}}\mathbb{P}(\|\mathbf{h}\|_2 > t)$  and  $-(t/2)D'_2(t) = -td(t)d'(t)$ . With a change of variables  $\tau = t/\sqrt{n}$ , define  $F(\tau; n) \stackrel{\text{def}}{=} F_1(\tau; n) + F_2(\tau; n)$  with

$$\begin{aligned} F_1(\tau; n) &\stackrel{\text{def}}{=} D_2(\tau\sqrt{n}; n) = \left\{ \mathbb{E} \left( \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right)_+ \right\}^2 \\ F_2(\tau; n) &\stackrel{\text{def}}{=} -\frac{\tau\sqrt{n}}{2}D'_2(\tau\sqrt{n}; n) = \mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}_+ \cdot \tau \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \right\}. \end{aligned}$$

Since  $\|\mathbf{h}\|_2$  concentrates around  $\sqrt{n}$  for large  $n$ , we show below that for any  $0 < \tau < 1$ ,

$$(B.50) \quad \lim_{n \rightarrow \infty} F_1(\tau; n) \stackrel{\text{def}}{=} F_1(\tau) = (1 - \tau)^2 \quad \text{and} \quad \lim_{n \rightarrow \infty} F_2(\tau; n) \stackrel{\text{def}}{=} F_2(\tau) = (1 - \tau)\tau.$$

It follows that for  $\delta \in (0, 1)$  we have  $\lim_{n \rightarrow \infty} F(1 - \delta; n) = \delta$  and

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{t_2^*(\delta; n)}{\sqrt{n}} &= 1 - \delta, & \lim_{n \rightarrow \infty} -(t_2^*/2)D'_2(t_2^*; n) &= F_2(\delta) = (1 - \delta)\delta, \\ \lim_{n \rightarrow \infty} D(t_2^*; n) &= F_1(\delta) = \delta^2, & \lim_{n \rightarrow \infty} \alpha_2^*(\delta; n) &= \sqrt{\frac{\delta^2}{\delta(\delta - \delta^2)}} = 1/\sqrt{1 - \delta}. \end{aligned}$$

which establishes (B.40).

To prove (B.50) we compute for  $0 < \tau < 1$ , and  $\epsilon > 0$

$$\begin{aligned} \mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}_+ &= \mathbb{E} \left\{ \left( \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right)_+ \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| > \epsilon \right\} \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| > \epsilon \right\} \\ &\quad + \mathbb{E} \left\{ \left( \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right)_+ \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\} \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\}, \\ \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \right\} &= \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| > \epsilon \right\} \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| > \epsilon \right\} \\ &\quad + \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\} \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\}. \end{aligned}$$

For any  $0 < \epsilon < \min(1 - \tau, \frac{1}{2})$  we get

$$\begin{aligned} \mathbb{P} \left\{ \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\} &\stackrel{(a)}{\geq} 1 - 2e^{-c n \epsilon^2}, & \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\} &= 1, \\ \mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}_+ &\geq (1 - \tau - \epsilon)(1 - 2e^{-c n \epsilon^2}), & \mathbb{P} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} > \tau \right\} &\geq 1 - 2e^{-c n \epsilon^2}, \\ \mathbb{E} \left\{ \left( \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right)_+ \mid \left| \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - 1 \right| \leq \epsilon \right\} &\geq 1 - \tau - \epsilon > 0, \end{aligned}$$

where (a) follows from Lemma A.1(b) by noting that for  $0 < \epsilon \leq \frac{1}{2}$ ,  $\sqrt{1/(1 - \epsilon)} \leq 1 + \epsilon$ .

Hence, with  $\epsilon = (1 - \tau)/n^{1/4}$ ,

$$(B.51) \quad F_1(\tau; n) \geq (1 - \tau)^2(1 - n^{-1/4})^2(1 - 2e^{-c(1 - \tau)^2 n^{1/2}})^2$$

$$(B.52) \quad F_2(\tau; n) \geq \tau(1 - \tau)(1 - n^{-1/4})(1 - 2e^{-c(1 - \tau)^2 n^{1/2}})^2$$

For an upper bound, let  $c_n \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \mathbb{E} \|\mathbf{h}\|_2$ . Since  $\frac{n}{\sqrt{n+1}} \stackrel{\text{(B.32)}}{\leq} \mathbb{E} \|\mathbf{h}\|_2 \leq \sqrt{n}$  we have  $1 \geq c_n \geq \sqrt{\frac{n}{n+1}} \geq 1 - \frac{1}{2n} \Rightarrow 0 \leq 2(1 - c_n) \leq 1/n$ . By Jensen's inequality, for  $0 < \tau < 1$ ,

$$\begin{aligned} \mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}_+ &\leq \sqrt{\mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}_+^2} \leq \sqrt{\mathbb{E} \left\{ \frac{\|\mathbf{h}\|_2}{\sqrt{n}} - \tau \right\}^2} = \sqrt{1 - 2c_n\tau + \tau^2} \\ &= \sqrt{(1 - \tau)^2 + 2(1 - c_n)\tau} \leq \sqrt{(1 - \tau)^2 + \tau/n} = (1 - \tau) \sqrt{1 + \frac{\tau}{(1 - \tau)^2 n}}, \end{aligned}$$

so that  $F_1(\tau; n) \leq (1 - \tau)^2 + \tau/n$  and  $F_2(\tau; n) \leq \tau(1 - \tau) \sqrt{1 + \frac{\tau}{(1 - \tau)^2 n}} \leq \tau(1 - \tau) \left(1 + \frac{\tau}{(1 - \tau)^2 n}\right)$ . Combining all of the above yields (B.50).

• **Step 9. (Proof of Lemma B.7).** By (B.36) and (B.37) we have for any  $n \geq 2/(1 - \delta)$

$$0 < t_{\min}(\delta) \stackrel{\text{def}}{=} \theta^{-1}\left(\frac{1+\delta}{2}\right) \leq t_1^*(\delta; n) \leq 2\theta^{-1}(\delta) \stackrel{\text{def}}{=} t_{\max}(\delta).$$

By the continuity and strict monotonicity of  $\theta$ ,  $V(\delta) \stackrel{\text{def}}{=} \inf_{t \in [t_{\min}(\delta), t_{\max}(\delta)]} \{\theta(t) - \theta(2t)\} > 0$ . By the convexity of  $D_1(t)$  with  $h = t = t_1^*$  we get  $D_1(2t_1^*) \geq D_1(t_1^*) + t_1^* D_1'(t_1^*)$ , hence

$$\begin{aligned} -t_1^* D_1'(t_1^*; n) &\geq D_1(t_1^*; n) - D_1(2t_1^*; n) \stackrel{\text{(B.45)\&(B.47)}}{\geq} \bar{D}_1(t_1^*; n) - \bar{D}_1(2t_1^*; n) - \frac{1}{n} \stackrel{\text{(B.46)}}{=} \theta(t_1^*) - \theta(2t_1^*) - \frac{1}{n} \\ &\geq V(\delta) - \frac{1}{n}. \end{aligned}$$

For  $n \geq N(\delta) \stackrel{\text{def}}{=} \max(2/V(\delta), 2/(1 - \delta))$  we obtain  $-t_1^* D_1'(t_1^*; n) \geq V(\delta)/2 \stackrel{\text{def}}{=} \gamma(\delta) > 0$  which establishes (B.7).

By (B.45)–(B.46)–(B.47) we have  $\theta(t) - 1/n \leq D_1(t; n) \leq \theta(t)$  for all  $t$  and  $n$ . Hence, the sequence of convex differentiable functions  $\{D_1(\cdot; n)\}_n$  converges uniformly to the convex and smooth function  $\theta(t)$  which implies convergence of the derivatives,  $\lim_{n \rightarrow \infty} D_1'(t; n) = \theta'(t)$ . As for  $p = 2$ , this shows  $\lim_{n \rightarrow \infty} \{D_1(t; n) - (t/2)D_1'(t; n)\} = \theta(t) - (t/2)\theta'(t)$ . By Lemma A.2,  $\theta(t) - (t/2)\theta'(t) = \text{erfc}(t/\sqrt{2})$  so the unique  $t = t_1^*(\delta)$  such that  $\theta(t) - (t/2)\theta'(t) = \delta$  is  $t_1^*(\delta) = \sqrt{2} \cdot \text{erfc}^{-1}(\delta)$ . Reasoning as for the case  $p = 2$  we get that

$$\begin{aligned} \lim_{n \rightarrow \infty} t_1^*(\delta; n) &= t_1^*(\delta) & \lim_{n \rightarrow \infty} -\frac{t_1^*}{2} D_1'(t_1^*; n) &= -\frac{t_1^*(\delta)}{2} \theta'(t_1^*(\delta)) > 0 \\ \lim_{n \rightarrow \infty} D_1(t_1^*; n) &= \theta(t_1^*(\delta)) & \lim_{n \rightarrow \infty} \alpha_1^*(\delta; n) &= \sqrt{\frac{\theta(t_1^*(\delta))}{\delta(\delta - \theta(t_1^*(\delta)))}}. \end{aligned}$$

Since  $\text{erfc}(t_1^*/\sqrt{2}) = \delta$  we have  $\theta(t_1^*) = \delta - \left(\sqrt{\frac{2}{\pi}} e^{-\frac{(t_1^*)^2}{2}} t_1^* - \delta(t_1^*)^2\right)$  and thus

$$(B.53) \quad \lim_{n \rightarrow \infty} \alpha_1^*(\delta; n) = \sqrt{\frac{\delta - (\delta - \theta(t_1^*))}{\delta(\delta - \theta(t_1^*))}} = \sqrt{\left(\sqrt{\frac{2}{\pi}} e^{-\frac{(t_1^*)^2}{2}} t_1^* - \delta(t_1^*)^2\right)^{-1} - \frac{1}{\delta}} \quad \square$$

**LEMMA B.8 (Deterministic properties of  $\kappa$ ).** Let  $m, n, 1 \leq m < n$  be two integers,  $\delta \stackrel{\text{def}}{=} (m - 1)/n$ ,  $1 \leq p \leq \infty$ ,  $\kappa(\alpha, \beta)$  defined in (B.23),  $D_p(t)$  defined in (B.31). The following hold:

1. The function  $\kappa(\alpha, \beta)$  is convex-concave and proper on  $[0, \infty) \times [0, \infty)$ , hence the function  $\kappa(\alpha) \stackrel{\text{def}}{=} \sup_{0 \leq \beta \leq 1} \kappa(\alpha, \beta)$  is convex on  $[0, \infty)$ , and for any  $A > 0$  the function  $\underline{\kappa}_A(\beta) \stackrel{\text{def}}{=} \inf_{0 \leq \alpha \leq A} \kappa(\alpha, \beta)$  is concave on  $[0, \infty)$ .

2. The scalar  $t^* = t_p^*(\delta; n)$  (cf. Lemma B.4–Property 7) is well defined, with  $D_p(t^*) < \delta$ .
3. Define

$$(B.54) \quad \alpha^* = \alpha^*(\delta; n) \stackrel{\text{def}}{=} \sqrt{\frac{D_p(t^*)}{\delta(\delta - D_p(t^*))}}$$

For  $A > \alpha^*$  and  $\lambda \leq t^*$  we have

$$(B.55) \quad \arg \min_{\alpha: 0 \leq \alpha \leq A} \max_{\beta: 0 \leq \beta \leq 1} \kappa(\alpha, \beta) = \alpha^*.$$

The corresponding optimal  $\beta$  is  $\beta^* = \beta^*(\lambda, \delta; n) \stackrel{\text{def}}{=} \lambda/t^*$ .

4. For  $A > \alpha^*$ ,  $\lambda \leq t^*$ ,  $0 < \epsilon \leq \max(\alpha^*, A - \alpha^*)$  we have

$$(B.56) \quad \inf_{|\alpha - \alpha^*| \geq \epsilon} \kappa(\alpha) - \kappa(\alpha^*) = \inf_{|\alpha - \alpha^*| = \epsilon} \kappa(\alpha) - \kappa(\alpha^*) \geq \omega(\epsilon) = \omega_p(\epsilon; n, \delta, \lambda) \stackrel{\text{def}}{=} \frac{\epsilon^2}{2} \frac{\lambda \delta / t^*}{(1 + \delta(\alpha^* + \epsilon)^2)^{3/2}}.$$

For the considered range of  $\lambda$  and  $\epsilon$ , we have  $\omega(\epsilon) \leq 1/2$ .

*Proof of Lemma B.8.*

1. It is obvious that  $\kappa$  is proper. Convexity in  $\alpha$  is easy to check by computing the second derivative. Concavity in  $\beta$  follows from the convexity of  $\beta \mapsto \Delta_p(\beta; \mathbf{h}, \lambda)$  which is a distance to a convex set [3, Example 3.16], and the fact that the expectation  $\Delta_p(\beta, \lambda)$  of a convex function is convex. As a result,  $\kappa(\alpha)$  is convex and  $\underline{\kappa}(\beta)$  is concave.
2. We have  $\delta \stackrel{\text{def}}{=} (m-1)/n < 1$ , and by Lemma B.4, Property 4,  $D_p(0) \geq \frac{n}{n+1}$ . Because we consider the underdetermined case,  $1 \leq m < n$ , we have  $n \geq 2$  and

$$\delta \leq (n-2)/n \leq n/(n+1) \leq D_p(0).$$

3. Since  $\kappa(\alpha, \beta)$  is convex–concave and proper, and the constraint sets in (B.55) convex and compact, we can change the order of maximization and minimization [27, Corollary 3.3]. (**Minimization over  $\alpha$** ) For  $\beta = 0$  and any  $\lambda \geq 0$  we have  $\Delta_p(\beta, \lambda) = 0$  hence  $\underline{\kappa}_A(0) = \inf_{0 \leq \alpha \leq A} \kappa(\alpha, 0) = 0$ . For  $\beta > 0$ , observing that  $\Delta_p(\beta, \lambda) = \beta \sqrt{D_p(\lambda/\beta)}$  we rewrite

$$\kappa(\alpha, \beta) = \beta(\sqrt{\delta\alpha^2 + 1} - \alpha\sqrt{D_p(\lambda/\beta)}).$$

With  $D_p^{-1}(y) \stackrel{\text{def}}{=} \inf\{t : D_p(t) \leq y\}$  and  $D_p$  strictly decreasing, we have  $D_p(\lambda/\beta) \leq \frac{A^2\delta^2}{1+\delta A^2}$  if and only if  $0 < \beta \leq \tilde{\beta} \stackrel{\text{def}}{=} \frac{\lambda}{D_p^{-1}(A^2\delta^2/(1+\delta A^2))}$ . Since  $\frac{A^2\delta^2}{1+\delta A^2} < \delta$ , this implies that for  $0 < \beta \leq \tilde{\beta}$  we can define

$$(B.57) \quad \tilde{\alpha}(\beta) \stackrel{\text{def}}{=} \sqrt{\frac{D_p(\lambda/\beta)}{\delta(\delta - D_p(\lambda/\beta))}}.$$

and check that  $\tilde{\alpha}(\beta) \leq A$ . By studying the sign of  $\frac{\partial \kappa(\alpha, \beta)}{\partial \alpha} = \beta \left( \frac{\delta\alpha}{\sqrt{\delta\alpha^2 + 1}} - \sqrt{D_p(\lambda/\beta)} \right)$ , we get that  $\alpha \mapsto \kappa(\alpha, \beta)$  has a unique minimizer on  $[0, A]$  which is precisely  $\alpha^*(\beta) = \tilde{\alpha}(\beta)$ . It follows that for  $0 < \beta \leq \tilde{\beta}$  we have:

$$(B.58) \quad \underline{\kappa}_A(\beta) = \min_{0 \leq \alpha \leq A} \kappa(\alpha, \beta) = \kappa(\alpha^*(\beta), \beta) = \beta \sqrt{\frac{\delta - D_p(\lambda/\beta)}{\delta}}.$$

**(Maximization over  $\beta$ )** The sign of  $\kappa'_A(\beta)$  is that of  $\delta - g(\lambda/\beta)$  with  $g(t) \stackrel{\text{def}}{=} D_p(t) - \frac{t}{2}D'_p(t)$  as in Lemma B.4–Property 7. Hence, we have:  $\kappa'_A(\beta) > 0$  if  $t = \lambda/\beta > t^*$  (that is to say if  $\beta < \beta^* \stackrel{\text{def}}{=} \lambda/t^*$ );  $\kappa'_A(\beta) < 0$  if  $\beta > \beta^*$ ; and  $D_p(t^*) < \delta$ .

$A > \alpha^*$  implies  $D_p(t^*) < A^2\delta^2/(1 + \delta A^2)$ , i.e.,  $\beta^* < \tilde{\beta}$ . Combined with the fact that  $\kappa(0) = 0$ , this shows that the supremum of  $\kappa(\beta)$  over  $[0, \tilde{\beta}]$  is achieved uniquely at  $\beta^*$ . This also implies that  $\kappa(\beta)$  is strictly decreasing for  $\beta^* < \beta \leq \tilde{\beta}$ . Being concave,  $\kappa_A(\beta)$  must be also strictly decreasing for  $\beta \geq \tilde{\beta}$ , so the supremum over  $[0, \infty)$  is indeed achieved at  $\beta^*$ . Since  $\lambda \leq t^*$  we further have  $\beta^* \leq 1$  hence this is also the supremum over  $\beta \in [0, 1]$ . To summarize, the optimal  $\beta$  is  $\beta^* = \lambda/t^*$ , and the corresponding optimal  $\alpha$  is given as

$$(B.59) \quad \alpha^*(\beta^*) = \tilde{\alpha}(\lambda/t^*) = \sqrt{\frac{D_p(t^*)}{\delta(\delta - D_p(t^*))}}.$$

4. The assumption  $\epsilon \leq \max(\alpha^*, A - \alpha^*)$  ensures that the set  $\{\alpha : |\alpha - \alpha^*| \geq \epsilon, 0 \leq \alpha \leq A\}$  is not empty. Since  $\kappa(\alpha)$  is convex on  $[0, \infty)$  with its minimum at  $\alpha^*$ , we have

$$\omega(\epsilon) \stackrel{\text{def}}{=} \inf_{\substack{\alpha: |\alpha - \alpha^*| \geq \epsilon \\ 0 \leq \alpha \leq A}} \kappa(\alpha) - \kappa(\alpha^*) = \min_{\substack{\alpha: |\alpha - \alpha^*| = \epsilon \\ 0 \leq \alpha \leq A}} \kappa(\alpha) - \kappa(\alpha^*)$$

Since  $A > \alpha^*$  and  $\lambda \leq t^*$ , we have  $0 < \beta^* < 1$ . The second derivative of  $\alpha \mapsto \kappa(\alpha, \beta^*)$  with respect to  $\alpha$  reads  $\frac{\partial^2 \kappa(\alpha, \beta^*)}{\partial \alpha^2} = \frac{\beta^* \delta}{(1 + \delta \alpha^2)^{3/2}} > 0$ , implying that on  $[0, A] \cap [\alpha^* - \epsilon, \alpha^* + \epsilon]$  the function  $\alpha \mapsto \kappa(\alpha, \beta^*)$  is strongly convex with strong convexity modulus  $\frac{\beta^* \delta}{(1 + \delta(\alpha^* + \epsilon)^2)^{3/2}}$ . Since  $\alpha \mapsto \kappa(\alpha, \beta^*)$  is minimum at  $\alpha^*$ , it holds that  $\kappa(\alpha^* \pm \epsilon, \beta^*) \geq \kappa(\alpha^*, \beta^*) + \frac{\epsilon^2}{2} \frac{\beta^* \delta}{(1 + \delta(\alpha^* + \epsilon)^2)^{3/2}}$ . Furthermore, from the definition of  $\kappa(\alpha)$  and  $\beta^*$ , we have that  $\kappa(\alpha) \geq \kappa(\alpha, \beta^*)$  for any  $\alpha$ , with equality for  $\alpha = \alpha^*$ . The claim therefore follows. Since  $\lambda \leq t^*$  and  $\epsilon \leq \alpha^*$  we have  $\omega(\epsilon) \leq \frac{\delta(\alpha^*)^2}{2(1 + \delta(\alpha^*)^2)^{3/2}} = f((1 + \delta(\alpha^*)^2)^{-1/2})$  with  $f(u) \stackrel{\text{def}}{=} \frac{1}{2}(1/u^2 - 1)u^3 = \frac{u - u^3}{2} \leq 1/2$  for  $0 < u \leq 1$ . Hence,  $\omega(\epsilon) \leq 1/2$ .  $\square$

Invoking a lemma from [17] we show that  $\arg \min_{\alpha} \phi(\alpha)$  concentrates around  $\arg \min_{\alpha} \kappa(\alpha)$ .

LEMMA B.9 ([17, Lemma 2]). *Let  $f(t)$  be a random convex function on some open set  $S \subset \mathbb{R}^p$ , and let  $t_f$  be (one of) its minimizer(s). Consider another function  $g(t)$  (which we interpret as approximating  $f$ ), such that it has a unique argmin  $t_g$ . Then for each  $\epsilon > 0$ , we have that:*

$$(B.60) \quad \mathbb{P}\{\|t_f - t_g\|_2 \geq \epsilon\} \leq \mathbb{P}\left\{ \sup_{\|s - t_g\|_2 \leq \epsilon} |f(s) - g(s)| \geq \frac{1}{2} \inf_{\|s - t_g\|_2 = \epsilon} (g(s) - g(t_g)) \right\}.$$

The role of  $f(t)$  and  $t_f$  will be played by  $\phi(\alpha)$  and  $\alpha^*_\phi$ ; the role of  $g(t)$  and  $t_g$  by  $\kappa(\alpha)$  and  $\alpha^*$ .

LEMMA B.10. *Let  $A > \alpha^*$ ,  $\lambda \leq t^*$ , with  $\alpha^*$  defined as in Lemma B.8–Equation (B.55) and  $t^* = t^*_p(\delta; n)$  as in Lemma B.4–Property 7. Consider the random function  $\phi(\alpha)$  defined as in Corollary B.3, and  $\alpha^*_\phi \stackrel{\text{def}}{=} \arg \min_{0 \leq \alpha \leq A} \phi(\alpha)$ . For  $0 < \epsilon \leq \max(\alpha^*, A - \alpha^*)$  we have*

$$(B.61) \quad \mathbb{P}\left[|\alpha^*_\phi - \alpha^*| \geq \epsilon\right] \leq \zeta\left(n, \frac{\omega(\epsilon)}{2}\right),$$

with  $\zeta(n, \epsilon) = \zeta(n, \epsilon; A, \delta)$  defined in Lemma B.2 and  $\omega(\epsilon)$  defined in (B.56).

*Proof.* Since  $0 < \epsilon \leq \max(\alpha^*, A - \alpha^*)$  the set  $\{\alpha : 0 \leq \alpha \leq A, |\alpha - \alpha^*| \leq \epsilon\}$  is a non-empty subset of  $[0, A]$ , and  $\sup_{\alpha: 0 \leq \alpha \leq A, |\alpha - \alpha^*| \leq \epsilon} |\phi(\alpha) - \kappa(\alpha)| \leq \sup_{0 \leq \alpha \leq A} |\phi(\alpha) - \kappa(\alpha)|$ . Since  $\phi$  is a random convex function, we can apply Lemma B.9 to obtain

$$\mathbb{P} \{ |\alpha_\phi^* - \alpha^*| \geq \epsilon \} \leq \mathbb{P} \left\{ \sup_{\alpha: 0 \leq \alpha \leq A, |\alpha - \alpha^*| \leq \epsilon} |\phi(\alpha) - \kappa(\alpha)| \geq \frac{\omega(\epsilon)}{2} \right\} \leq \mathbb{P} \left\{ \sup_{0 \leq \alpha \leq A} |\phi(\alpha) - \kappa(\alpha)| \geq \frac{\omega(\epsilon)}{2} \right\} \leq \zeta(n, \omega(\epsilon)/2),$$

where we used that  $\omega(\epsilon) \leq \inf_{\alpha: 0 \leq \alpha \leq A, |\alpha - \alpha^*| = \epsilon} \kappa(\alpha) - \kappa(\alpha^*)$ , and the last inequality follows from Corollary B.3 which we can use since  $\omega(\epsilon)/2 \leq 1/4 < 2$ .  $\square$

LEMMA B.11. *Let  $A > \alpha^*$ ,  $\lambda \leq t^*$  with  $\alpha^*$  defined as in Lemma B.8–Equation (B.55) and  $t^* = t_p^*(\delta; n)$  as in Lemma B.4–Property 7. For  $0 \leq \epsilon \leq \max(\alpha^*, A - \alpha^*)$ , consider the optimal cost of the auxiliary optimization (3.14) with an altered order of minimization and maximization and  $\|\mathbf{z}\|_2$  further restricted to be at least at distance  $\epsilon$  from  $\alpha^*$*

(B.62)

$$\phi_\epsilon(\mathbf{g}, \mathbf{h}) \stackrel{\text{def}}{=} \max_{\substack{\beta: 0 \leq \beta \leq 1 \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\substack{\mathbf{z}: \|\mathbf{z}\|_2 \leq A \\ \|\mathbf{z}\|_2 \notin (\alpha^* - \epsilon, \alpha^* + \epsilon)}} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} \frac{1}{\sqrt{n}} \|\mathbf{z}\|_2 \mathbf{g}^\top \mathbf{u} - \frac{1}{\sqrt{n}} \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{z} - \mathbf{u}^\top \mathbf{e}_1 + \frac{\lambda}{\sqrt{n}} \mathbf{w}^\top \mathbf{z}.$$

With  $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ ,  $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ ,  $1 \leq m < n$ ,  $\delta = (m - 1)/n$  we have for any  $0 < \eta < 2$ :

$$(B.63) \quad \phi_0(\mathbf{g}, \mathbf{h}) < \kappa(\alpha^*) + \eta \quad \text{and} \quad \phi_\epsilon(\mathbf{g}, \mathbf{h}) > \kappa(\alpha^*) + \omega(\epsilon) - \eta,$$

with probability at least  $1 - \zeta(n, \eta)$ , where  $\zeta(n, \epsilon) = \zeta(n, \epsilon; A, \delta)$  from Lemma B.2,  $\omega(\epsilon)$  from (B.56).

*Proof.* Since  $0 < \epsilon \leq \max(\alpha^*, A - \alpha^*)$  the set  $S_\epsilon \stackrel{\text{def}}{=} \{\alpha : 0 \leq \alpha \leq A, |\alpha - \alpha^*| \leq \epsilon\}$  is a non-empty subset of  $[0, A]$ . Denote  $S_\epsilon^- \stackrel{\text{def}}{=} [0, \alpha^* - \epsilon]$  and  $S_\epsilon^+ = [\alpha^* + \epsilon, A]$  its two convex components (at most one of them may be empty). Let  $\phi_{S_\epsilon^+}(\mathbf{g}, \mathbf{h})$  the value of (B.62), but with  $\|\mathbf{z}\|_2$  constrained to lie in  $S_\epsilon^+$  (by convention, this is  $+\infty$  when  $S_\epsilon^+ = \emptyset$ ). Similarly define  $\phi_{S_\epsilon^-}(\mathbf{g}, \mathbf{h})$ . When  $S_\epsilon^+$  is non-empty, since it is convex, we can effect the same simplifications and min-max swaps as in the proof of Lemma 3.1 (from (3.15) to (3.16)) to arrive at

$$\phi_{S_\epsilon^+}(\mathbf{g}, \mathbf{h}) = \min_{\alpha: \alpha^* + \epsilon \leq \alpha \leq A} \max_{\beta: 0 \leq \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}),$$

and similiary with  $S_\epsilon^-$  we get when it is non-empty that

$$\phi_{S_\epsilon^-}(\mathbf{g}, \mathbf{h}) = \min_{\alpha: 0 \leq \alpha \leq \alpha^* - \epsilon} \max_{\beta: 0 \leq \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}).$$

This shows that  $\phi_\epsilon(\mathbf{g}, \mathbf{h}) = \min(\phi_{S_\epsilon^-}(\mathbf{g}, \mathbf{h}), \phi_{S_\epsilon^+}(\mathbf{g}, \mathbf{h})) = \phi_{S_\epsilon}(\mathbf{g}, \mathbf{h})$  where the notation  $\phi_{S_\epsilon}(\mathbf{g}, \mathbf{h})$  matches that used in Corollary B.3. Moreover by definition (see Lemma B.8) we have

$$\kappa_{S_\epsilon} \stackrel{\text{def}}{=} \min_{\alpha \in S_\epsilon} \max_{0 \leq \beta \leq 1} \kappa(\alpha, \beta) \geq \kappa(\alpha^*) + \omega(\epsilon).$$

By Corollary B.3 we have, for  $0 < \eta < 2$ , with probability at least  $1 - \zeta(n, \eta)$ : for all  $S \subset [0, A]$ ,  $|\phi_S(\mathbf{g}, \mathbf{h}) - \kappa_S| < \eta$ . Specializing to  $S = [0, A]$  and  $S = S_\epsilon$  and combining the above yields

$$\begin{aligned} \phi_0(\mathbf{g}, \mathbf{h}) &= \phi_{[0, A]}(\mathbf{g}, \mathbf{h}) < \kappa_{[0, A]} + \eta = \kappa(\alpha^*) + \eta, \\ \phi_\epsilon(\mathbf{g}, \mathbf{h}) &= \phi_{S_\epsilon}(\mathbf{g}, \mathbf{h}) > \kappa_{S_\epsilon} - \eta \geq \kappa(\alpha^*) + \omega(\epsilon) - \eta. \end{aligned} \quad \square$$



LEMMA B.12. Let  $K > \alpha^*/\sqrt{n}$ ,  $\lambda \leq t^*$  with  $\alpha^*$  defined as in Lemma B.8–Equation (B.55) and  $t^* = t_p^*(\delta; n)$  as in Lemma B.4–Property 7. Denote by  $\tilde{\mathbf{x}}_K$  any optimal solution of (3.11). For  $0 < \epsilon \leq \max(\alpha^*, K\sqrt{n} - \alpha^*)$  we have

$$\mathbb{P}\left\{\left|\|\tilde{\mathbf{x}}_K\| - \frac{\alpha^*}{\sqrt{n}}\right| \geq \frac{\epsilon}{\sqrt{n}}\right\} \leq 4\zeta\left(n, \frac{\omega(\epsilon)}{2}\right).$$

with  $\zeta(n, \epsilon) = \zeta(n, \epsilon; K\sqrt{n}, \delta)$  defined in Lemma B.2 and  $\omega(\epsilon)$  defined in (B.56).

*Proof.* Denote  $\Phi$  the optimal cost of (3.11) and  $\Phi_\epsilon$  the corresponding cost when  $\mathbf{z} = \mathbf{x}\sqrt{n}$  is further restricted to  $\mathcal{S}_\epsilon \stackrel{\text{def}}{=} \{\mathbf{z} : \|\mathbf{z}\|_2 \leq A \text{ and } \|\mathbf{z}\|_2 - \alpha^* \geq \epsilon\}$ , with  $A \stackrel{\text{def}}{=} K\sqrt{n}$ :

$$(B.64) \quad \Phi_\epsilon = \min_{\substack{\mathbf{x}: \|\mathbf{x}\|_2 \leq K, \\ |\sqrt{n}\|\mathbf{x}\|_2 - \alpha^*| \geq \epsilon}} \max_{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{A}\mathbf{x} + \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1.$$

We now want to show that for  $\epsilon > 0$  we have with high probability  $\Phi_\epsilon > \Phi = \Phi_0$ , because this is equivalent to  $\tilde{\mathbf{z}}_K = \tilde{\mathbf{x}}_K\sqrt{n} \in \mathcal{S}'_\epsilon \stackrel{\text{def}}{=} \{\mathbf{z} : \|\mathbf{z}\|_2 \leq A \text{ and } \|\mathbf{z}\|_2 - \alpha^* < \epsilon\}$ .

By Theorem A.3 (i) and (ii), denoting  $\phi_\epsilon^P$  (resp.  $\phi_\epsilon^D$ ) the optimum value of the “primal” (resp. “dual”) auxiliary optimization problem associated to (B.64), we have for any  $c \in \mathbb{R}$

$$(B.65) \quad \mathbb{P}\{\Phi_\epsilon < c\} \leq 2\mathbb{P}\{\phi_\epsilon^P \leq c\} \quad \text{and} \quad \mathbb{P}\{\Phi_0 > c\} \leq 2\mathbb{P}\{\phi_0^D \geq c\},$$

where we additionally used that  $\phi_0^D = \phi_0^P$  since we optimize over convex sets ( $\mathcal{S}_0$  is a convex ball) and the penalty  $\psi(\mathbf{x}, \mathbf{u}) = \lambda \|\mathbf{x}\|_p - \mathbf{u}^\top \mathbf{e}_1$  is convex–concave (see, e.g., [27, Corollary 3.3]).

Let  $C(\mathbf{z}, \mathbf{u}, \mathbf{w}) = C(\mathbf{z}, \mathbf{u}, \mathbf{z}; \mathbf{g}, \mathbf{h})$  be the objective function in (3.14) and (B.62), so that with  $A \stackrel{\text{def}}{=} K\sqrt{n}$  (B.62) becomes  $\phi_\epsilon = \max_{\substack{\beta: 0 \leq \beta \leq 1, \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} C(\mathbf{z}, \mathbf{u}, \mathbf{w})$ , and the optimal cost of

(3.14) reads  $\phi = \phi_0$ . With these notations, we have (since  $\max \min \leq \min \max$  is always true):

$$(B.66) \quad \begin{aligned} \phi_\epsilon^P &\stackrel{\text{def}}{=} \min_{\substack{\mathbf{z}: \|\mathbf{z}\|_2 \leq A \\ \mathbf{z} \in \mathcal{S}_\epsilon}} \max_{\substack{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1, \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) = \min_{\substack{\mathbf{z}: \|\mathbf{z}\|_2 \leq A \\ \mathbf{z} \in \mathcal{S}_\epsilon}} \max_{\beta: 0 \leq \beta \leq 1, \mathbf{u}: \|\mathbf{u}\|_2 = \beta} \max_{\mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) \\ &\geq \max_{\substack{\beta: 0 \leq \beta \leq 1, \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) = \phi_\epsilon \end{aligned}$$

and

$$(B.67) \quad \begin{aligned} \phi_0^D &\stackrel{\text{def}}{=} \max_{\substack{\mathbf{u}: \|\mathbf{u}\|_2 \leq 1, \\ \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1}} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) = \max_{\beta: 0 \leq \beta \leq 1, \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) \\ &\leq \max_{\beta: 0 \leq \beta \leq 1, \mathbf{w}: \|\mathbf{w}\|_{p^*} \leq 1} \min_{\mathbf{z}: \|\mathbf{z}\|_2 \leq A} \max_{\mathbf{u}: \|\mathbf{u}\|_2 = \beta} C(\mathbf{z}, \mathbf{u}, \mathbf{w}) = \phi_0. \end{aligned}$$

Denote  $\underline{\phi}_\epsilon \stackrel{\text{def}}{=} \kappa(\alpha^*) + \omega(\epsilon)$  and  $\bar{\phi} \stackrel{\text{def}}{=} \kappa(\alpha^*)$  and use the above with  $c_1 = \underline{\phi}_\epsilon - \eta$  and  $c_2 = \bar{\phi} + \eta$  where  $\eta > 0$  is arbitrary to get

$$(B.68) \quad \begin{aligned} \mathbb{P}\{\Phi_\epsilon < \underline{\phi}_\epsilon - \eta\} &\leq 2\mathbb{P}\{\phi_\epsilon^P \leq \underline{\phi}_\epsilon - \eta\} \leq 2\mathbb{P}\{\phi_\epsilon \leq \underline{\phi}_\epsilon - \eta\}, \\ \mathbb{P}\{\Phi > \bar{\phi} + \eta\} &\leq 2\mathbb{P}\{\phi_0^D \geq \bar{\phi} + \eta\} \leq 2\mathbb{P}\{\phi_0 \geq \bar{\phi} + \eta\}. \end{aligned}$$

Consider the event  $\mathcal{E} = \{\Phi_\epsilon \geq \underline{\phi}_\epsilon - \eta \text{ and } \Phi \leq \bar{\phi} + \eta\}$ . For  $0 < \eta < (\underline{\phi}_\epsilon - \bar{\phi})/2 = \omega(\epsilon)/2$  we have  $c_1 > c_2$  hence this event implies that  $\tilde{\mathbf{z}}_K \in \mathcal{S}'_\epsilon$ , which is what we wanted to prove. For

such  $\eta$ , since  $\omega(\epsilon)/2 \leq 1/4 < 2$  we can use Lemma B.11 and a union bound to obtain that this event happens with probability at least  $1 - 4\zeta(n, \eta)$ . Hence, for any  $0 < \eta < \omega(\epsilon)/2$  we have  $\mathbb{P}(\tilde{\mathbf{z}}_K \notin \mathcal{S}_\epsilon) \leq 4\zeta(n, \eta)$ . By continuity of  $\eta \mapsto \zeta(n, \eta)$  we take the limit when  $\eta$  tends to  $\omega(\epsilon)/2$ .  $\square$

LEMMA B.13. *Let  $K > \alpha^*/\sqrt{n}$ ,  $\lambda \leq t^*$  with  $\alpha^*$  defined as in Lemma B.8–Equation (B.55) and  $t^* = t_p^*(\delta; n)$  as in Lemma B.4–Property 7. Denote by  $\tilde{\mathbf{x}}_K$  any optimal solution of the random bounded problem (3.11) and  $\tilde{\mathbf{x}}$  any optimal solution of the random unbounded problem (3.9). For  $0 < \epsilon \leq \min(\alpha^*, K\sqrt{n} - \alpha^*)$  we have*

$$\mathbb{P}\{\tilde{\mathbf{x}}_K \neq \tilde{\mathbf{x}}\} \leq 4\zeta\left(n, \frac{\omega(\epsilon)}{2}\right) \quad \text{and} \quad \mathbb{P}\left\{\left\|\tilde{\mathbf{x}}\right\| - \frac{\alpha^*}{\sqrt{n}} \geq \frac{\epsilon}{\sqrt{n}}\right\} \leq 4\zeta\left(n, \frac{\omega(\epsilon)}{2}\right).$$

with  $\zeta(n, \epsilon) = \zeta(n, \epsilon; K\sqrt{n}, \delta)$  defined in Lemma B.2 and  $\omega(\epsilon)$  defined in (B.56).

*Proof.* To handle non-uniqueness,  $\tilde{\mathbf{x}}$  (resp.  $\tilde{\mathbf{x}}_K$ ) may denote the convex set of solutions of the respective convex optimization problems. The property  $\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}_K$  then means that the sets do not intersect, and inequalities such as  $f(\tilde{\mathbf{x}}) > c$  are meant to hold for all elements of the set  $\tilde{\mathbf{x}}$ .

We first prove, by contradiction, that if  $\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}_K$ , then necessarily  $\|\tilde{\mathbf{x}}_K\|_2 = K$ . Suppose that the opposite holds:  $\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}_K$ , but  $\|\tilde{\mathbf{x}}_K\|_2 < K$ . Since  $\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}_K$  we have  $\|\tilde{\mathbf{x}}\|_2 > K$ . Denoting the objective in (3.9) by  $\phi(\mathbf{x})$ , this means that  $\phi(\tilde{\mathbf{x}}) < \phi(\tilde{\mathbf{x}}_K)$ . By convexity of  $\phi$  it follows that all points on the line segment  $\mathbf{x}_\nu = \nu\tilde{\mathbf{x}}_K + (1 - \nu)\tilde{\mathbf{x}}$ ,  $0 \leq \nu \leq 1$  satisfy

$$(B.69) \quad \phi(\mathbf{x}_\nu) \leq \phi(\tilde{\mathbf{x}}_K).$$

Since  $\|\mathbf{x}_0\|_2 > K$  and  $\|\mathbf{x}_1\|_2 < K$ , by continuity there exists  $\nu \in (0, 1)$  such that  $\|\mathbf{x}_\nu\|_2 = K$ . Further, by (B.69),  $\mathbf{x}_\nu$  is optimizing the bounded problem (3.11), contradicting our assumption.

By contraposition, if  $\|\tilde{\mathbf{x}}_K\|_2 < K$  then  $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_K$ . In particular, since we assume  $K\sqrt{n} > \alpha^* + \epsilon$ , we have: if  $|\sqrt{n}\|\tilde{\mathbf{x}}_K\|_2 - \alpha^*| < \epsilon$  then  $\|\tilde{\mathbf{x}}_K\|_2 < K$  hence  $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_K$  and  $|\sqrt{n}\|\tilde{\mathbf{x}}\|_2 - \alpha^*| = |\sqrt{n}\|\tilde{\mathbf{x}}_K\|_2 - \alpha^*| < \epsilon$ . It follows that  $\mathbb{P}\{\tilde{\mathbf{x}}_K \neq \tilde{\mathbf{x}}\} \leq \mathbb{P}\{|\sqrt{n}\|\tilde{\mathbf{x}}_K\|_2 - \alpha^*| \geq \epsilon\}$  and  $\mathbb{P}\{|\sqrt{n}\|\tilde{\mathbf{x}}\|_2 - \alpha^*| \geq \epsilon\} \leq \mathbb{P}\{|\sqrt{n}\|\tilde{\mathbf{x}}_K\|_2 - \alpha^*| \geq \epsilon\}$ . We conclude using Lemma B.12.  $\square$

LEMMA B.14. *With  $\Delta_p$  defined as in Lemma B.2, the function  $f_{\mathbf{h}}(\beta) \stackrel{\text{def}}{=} |\Delta_p(\beta; \mathbf{h}, \lambda) - \Delta_p(\beta; n, \lambda)|$  is  $\max\{\|\mathbf{h}\|/\sqrt{n}, 1\}$ -Lipschitz in  $\beta$ .*

*Proof.* We omit the dependency in  $\lambda$  for brevity and write by definition

$$|\Delta_p(\beta_1; \mathbf{h}) - \Delta_p(\beta_2; \mathbf{h})| = \frac{1}{\sqrt{n}} |\text{dist}(\beta_1 \mathbf{h}, \mathcal{C}) - \text{dist}(\beta_2 \mathbf{h}, \mathcal{C})| \leq \frac{1}{\sqrt{n}} \|\beta_1 \mathbf{h} - \beta_2 \mathbf{h}\|_2 \leq \frac{\|\mathbf{h}\|_2}{\sqrt{n}} |\beta_1 - \beta_2|,$$

since the Euclidean distance to a convex set is 1-Lipschitz with respect to the Euclidean metric. Further,  $|\Delta_p(\beta_1) - \Delta_p(\beta_2)| \leq \mathbb{E} |(\Delta_p(\beta_1; \mathbf{h}) - \Delta_p(\beta_2; \mathbf{h}))| \leq \frac{|\beta_1 - \beta_2|}{\sqrt{n}} \mathbb{E} \|\mathbf{h}\|_2 \leq |\beta_1 - \beta_2|$ . Now observe that the Lipschitz constant of the difference of two Lipschitz functions does not exceed the largest of the two constants, and the Lipschitz constant of  $|f|$  equals that of  $f$ .  $\square$

## References.

- [1] D. AMELUNXEN, M. LOTZ, M. B. MCCOY, AND J. A. TROPP, *Living on the edge: phase transitions in convex programs with random data*, Information and Inference: A Journal of the IMA, 3 (2014), pp. 224–294.
- [2] A. BARVINOK, *Math 710: Measure concentration*, Lecture notes, (2005), <http://www.math.lsa.umich.edu/~barvinok/total710.pdf>.
- [3] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, 2004.
- [4] P. G. CASAZZA, A. HEINECKE, F. KRAHMER, AND G. KUTYNIOK, *Optimally Sparse Frames*, IEEE Trans. Inf. Theory, 57, pp. 7279–7287.
- [5] V. CHANDRASEKARAN, P. A. PARRILO, AND A. S. WILLSKY, *Convex Graph Invariants*, SIAM Rev., 54 (2012), pp. 513–541.

- [6] M. CHIANI, D. DARDARI, AND M. K. SIMON, *New exponential bounds and approximations for the computation of error probability in fading channels*, IEEE Trans. Wirel. Commun., 24 (2003), pp. 840–845.
- [7] G. DAVIS, S. MALLAT, AND M. AVELLANEDA, *Adaptive Greedy Approximations*, Constr. Approx., 13 (1997), pp. 57–98.
- [8] I. DOKMANIC AND R. GRIBONVAL, *Beyond moore-penrose part I: generalized inverses that minimize matrix norms*, CoRR, abs/1706.08349 (2017), <http://arxiv.org/abs/1706.08349>, <https://arxiv.org/abs/1706.08349>.
- [9] I. DOKMANIC AND R. GRIBONVAL, *Beyond moore-penrose part II: the sparse pseudoinverse*, CoRR, abs/1706.08701 (2017), <http://arxiv.org/abs/1706.08701>, <https://arxiv.org/abs/1706.08701>.
- [10] I. DOKMANIĆ, M. KOLUNDŽIJA, AND M. VETTERLI, *Beyond Moore-Penrose: Sparse pseudoinverse*, in IEEE ICASSP, IEEE, 2013, pp. 6526–6530.
- [11] D. DONOHO AND J. TANNER, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 367 (2009), pp. 4273–4293.
- [12] D. L. DONOHO, *Compressed Sensing*, IEEE Trans. Inf. Theory, 52 (2006), pp. 1289–1306.
- [13] R. FOYGEL AND L. MACKEY, *Corrupted Sensing: Novel Guarantees for Separating Structured Signals*, IEEE Trans. Inf. Theory, 60 (2014), pp. 1223–1247.
- [14] Y. GORDON, *Some inequalities for gaussian processes and applications*, Israel Journal of Mathematics, 50 (1985), pp. 265–289.
- [15] Y. GORDON, *On milman’s inequality and random subspaces which escape through a mesh in  $\mathbb{R}^n$* , in Geometric Aspects of Functional Analysis, Berlin, Heidelberg, 1988, Springer Berlin Heidelberg, pp. 84–106.
- [16] N. HALKO, P. G. MARTINSSON, AND J. A. TROPP, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
- [17] N. L. HJORT AND D. POLLARD, *Asymptotics for Minimisers of Convex Processes*, arXiv, (2011), <https://arxiv.org/abs/1107.3806v1>.
- [18] F. KRAHMER, G. KUTYNIOK, AND J. LEMVIG, *Sparsity and Spectral Properties of Dual Frames*, Linear Algebra Appl., 439 (2012), pp. 1–17.
- [19] M. LEDOUX, *Concentration of Measure and Logarithmic Sobolev Inequalities*, in Séminaire de Probabilités XXXIII, Springer, Berlin, Heidelberg, 1999, pp. 120–216.
- [20] S. LI, Y. LIU, AND T. MI, *Sparse Dual Frames and Dual Gabor Functions of Minimal Time and Frequency Supports*, J. Fourier Anal. Appl., 19 (2013), pp. 48–76.
- [21] B. NATARAJAN, *Sparse approximate solutions to linear systems*, SIAM J. Computing, 25 (1995), pp. 227–234.
- [22] S. OYMAK, C. THRAMOULIDIS, AND B. HASSIBI, *The Squared-Error of Generalized LASSO: A Precise Analysis*, arXiv, (2013), <https://arxiv.org/abs/1311.0830v2>.
- [23] S. OYMAK AND J. A. TROPP, *Universality laws for randomized dimension reduction, with applications*, Information and Inference: A Journal of the IMA, (2017), pp. 1–110.
- [24] N. PERRAUDIN, N. HOLIGHAUS, P. L. SØNDERGAARD, AND P. BALAZS, *Designing Gabor Windows Using Convex Optimization*, arXiv, (2014), <https://arxiv.org/abs/1401.6033>.
- [25] M. RUDELSON, *Invertibility of random matrices: norm of the inverse*, Ann. Math., 168 (2008), pp. 575–600.
- [26] M. RUDELSON AND R. VERSHYNIN, *On sparse reconstruction from fourier and gaussian measurements*, Comm. Pure Appl. Math., 61 (2008), pp. 1025–1045.
- [27] M. SION, *On General Minimax Theorems*, Pac. J. Math., 8 (1958), pp. 171–176.

- [28] M. STOJNIC, *Various thresholds for  $\ell^1$ -optimization in compressed sensing*, CoRR, abs/0907.3666 (2009), <http://arxiv.org/abs/0907.3666>, <https://arxiv.org/abs/0907.3666>.
- [29] M. STOJNIC, *A framework to characterize performance of LASSO algorithms*, CoRR, abs/1303.7291 (2013), <http://arxiv.org/abs/1303.7291>, <https://arxiv.org/abs/1303.7291>.
- [30] C. THRAMPOULIDIS, E. ABBASI, AND B. HASSIBI, *Precise Error Analysis of Regularized  $M$ -estimators in High-dimensions*, arXiv, (2016), <https://arxiv.org/abs/1601.06233v1>.
- [31] C. THRAMPOULIDIS, S. OYMAK, AND B. HASSIBI, *Regularized linear regression: A precise analysis of the estimation error*, in Proc. COLT, vol. 40, Paris, France, 03–06 Jul 2015, PMLR, pp. 1683–1709.
- [32] R. VERSHYNIN, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, Cambridge, 2009, pp. 210–268.
- [33] R. VERSHYNIN, *Invertibility of symmetric random matrices*, Random Struct. Algorithms, 44 (2014), pp. 135–182.
- [34] M. WAINWRIGHT, *Basic tail and concentration bounds*, Draft, 2015, <https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2-TailBounds-Jan22-2015.pdf>.
- [35] J. G. WENDEL, *Note on the Gamma Function*, Am. Math. Mon., 55 (1948), p. 563.