

A NEW PERSPECTIVE ON LEAST SQUARES UNDER CONVEX CONSTRAINT¹

BY SOURAV CHATTERJEE

Stanford University

Consider the problem of estimating the mean of a Gaussian random vector when the mean vector is assumed to be in a given convex set. The most natural solution is to take the Euclidean projection of the data vector on to this convex set; in other words, performing “least squares under a convex constraint.” Many problems in modern statistics and statistical signal processing theory are special cases of this general situation. Examples include the lasso and other high-dimensional regression techniques, function estimation problems, matrix estimation and completion, shape-restricted regression, constrained denoising, linear inverse problems, etc. This paper presents three general results about this problem, namely, (a) an exact computation of the main term in the estimation error by relating it to expected maxima of Gaussian processes (existing results only give upper bounds), (b) a theorem showing that the least squares estimator is always admissible up to a universal constant in any problem of the above kind and (c) a counterexample showing that least squares estimator may not always be minimax rate-optimal. The result from part (a) is then used to compute the error of the least squares estimator in two examples of contemporary interest.

1. Theory.

1.1. *The setup.* Throughout this manuscript, $Z = (Z_1, \dots, Z_n)$ denotes an n -dimensional standard Gaussian random vector. Let $\mu = (\mu_1, \dots, \mu_n) \in \mathbb{R}^n$ be a point in \mathbb{R}^n , and let $Y = Z + \mu$. We are interested in estimating μ from the data vector Y . If nothing more is known, the vector Y itself is the maximum likelihood estimate of μ .

Suppose now that μ is known to belong to a closed convex set $K \subseteq \mathbb{R}^n$. Let P_K denote the Euclidean projection on to K . That is, for a vector $x \in \mathbb{R}^n$, $P_K(x)$ is the point in K that is closest to x in the Euclidean distance. It is a standard fact about closed convex sets (see Lemma 4.2 in Section 4) that P_K is a well-defined map. Under the assumption that $\mu \in K$, the maximum likelihood estimate of μ in the Gaussian model is $\hat{\mu} := P_K(Y)$. We will refer to $\hat{\mu}$ as the least squares estimator (LSE) of μ under the convex constraint K . As mentioned in the abstract, many

Received March 2014; revised June 2014.

¹Supported in part by NSF Grant DMS-10-05312.

MSC2010 subject classifications. 62F10, 62F12, 62F30, 62G08.

Key words and phrases. Least squares, maximum likelihood, convex constraint, empirical process, lasso, isotonic regression, denoising.

problems in modern statistics are special cases of this general setup, including the lasso and other high-dimensional regression techniques, function estimation problems, matrix estimation and completion, shape-restricted regression, etc.

Let $\|x\|$ denote the Euclidean norm of a vector $x \in \mathbb{R}^n$. Our main goal is to understand the magnitude of the estimation error $\|\hat{\mu} - \mu\|$. The standard approach to computing upper bounds on the expected squared value of this error (the “risk”) is via empirical process theory and related entropy computations. As a consequence of path-breaking contributions from a number of authors over a period of more than thirty years, including Birgé [5], Tsirelson [62–64], Pollard [50], van de Geer [65–67], Birgé and Massart [6], van der Vaart and Wellner [72] and many others, we now have a fairly good idea about how to convert results for expected maxima of empirical processes to upper bounds on estimation errors in problems of the above type, especially in the context of regression. To know more about this important branch of theoretical statistics and machine learning, see the monographs of Bühlmann and van de Geer [9], Massart [41], van de Geer [68] and van der Vaart and Wellner [72].

In a different direction, this general problem has recently gained prominence in the statistical signal processing literature. The least squares problem outlined above is essentially equivalent to the problem of constrained denoising in signal processing. It is also intimately connected with the so-called linear inverse problems. The history of this line of investigation, although of relatively recent origin, is already quite formidable. Important papers include those of Rudelson and Vershynin [54], Stojnic [56], Oymak and Hassibi [47, 48], Chandrasekaran et al. [16], Chandrasekaran and Jordan [15] and Amelunxen et al. [1]. For some interesting recent developments, see McCoy and Tropp [42, 43], Foygel and Mackey [31] and Tropp [61].

In the signal processing context, the expected squared error $\mathbb{E}\|\hat{\mu} - \mu\|^2$ is closely related to the concept of “statistical dimension” recently introduced in Amelunxen et al. [1]. It is also related to the older existing notions of “Gaussian width” in probability, functional analysis and signal processing, “mean width” in convex geometry and “Gaussian complexity” in learning theory and machine learning.

1.2. Main result. One limitation of the theory based on empirical processes in its current form is that it only gives upper bounds on the error. There are some lower bounds “in spirit,” in the form of necessary and sufficient conditions for consistency (e.g., in Tsirelson [62] and van de Geer and Wegkamp [70]) but the lower bounds are not explicit. The first main result of this manuscript, presented below, shows that if one looks at expected maxima of certain Gaussian processes (instead of upper bounds on these maxima) then one can get an approximation for the actual error instead of just an upper bound. Not only that, the theorem also shows that the error $\|\hat{\mu} - \mu\|$ is typically concentrated around its expected value.

Let $x \cdot y$ denote the usual inner product on \mathbb{R}^n and let K be any nonempty closed convex set. For any $\mu \in \mathbb{R}^n$ and any $t \geq 0$, let

$$f_\mu(t) := \mathbb{E} \left(\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - \mu) \right) - \frac{t^2}{2},$$

where Z is an n -dimensional standard Gaussian random vector. If $\mu \notin K$, then there is no $v \in K$ satisfying $\|\mu - v\| \leq t$ if t is strictly less than the distance of μ from K . In that case, define $f_\mu(t)$ to be $-\infty$, following the standard convention that the supremum of an empty set is $-\infty$.

Let t_μ be the point in $[0, \infty)$ where f_μ attains its maximum. We will show below that t_μ exists and is unique. Recall that P_K denotes the projection on to K , and that

$$\hat{\mu} := P_K(Z + \mu)$$

is the least squares estimate of μ based on the data vector $Z + \mu$. The following theorem shows that irrespective of the dimension n and the convex set K , it is always true that

$$\|\hat{\mu} - \mu\| = t_\mu + O(\max\{\sqrt{t_\mu}, 1\}).$$

In particular, if t_μ is large, then the random quantity $\|\hat{\mu} - \mu\|$ is concentrated around the nonrandom value t_μ .

THEOREM 1.1. *Let $K, \mu, \hat{\mu}, f_\mu$ and t_μ be as above. Let $t_c := \inf_{v \in K} \|v - \mu\|$. Then $f_\mu(t)$ is equal to $-\infty$ when $t < t_c$, is a finite and strictly concave function of t when $t \in [t_c, \infty)$, and decays to $-\infty$ as $t \rightarrow \infty$. Consequently, t_μ exists and is unique. Moreover, for any $x \geq 0$,*

$$\mathbb{P}(\|\hat{\mu} - \mu\| - t_\mu \geq x\sqrt{t_\mu}) \leq 3 \exp\left(-\frac{x^4}{32(1 + x/\sqrt{t_\mu})^2}\right).$$

Note that μ is not required to be in K in this theorem. The tail bound is valid even if μ is a point lying outside K .

The above theorem can potentially give rise to many corollaries. One basic corollary, presented below, gives estimates for the expected squared error of $\hat{\mu}$. Although Theorem 1.1 contains a lot more information than this corollary, expected squared errors are culturally important.

COROLLARY 1.2. *Let all notation be as in Theorem 1.1. Then there is a universal constant C such that if $t_\mu \geq 1$, then*

$$t_\mu^2 - Ct_\mu^{3/2} \leq \mathbb{E}\|\hat{\mu} - \mu\|^2 \leq t_\mu^2 + Ct_\mu^{3/2},$$

and if $t_\mu < 1$, then

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 \leq C.$$

It may be illuminating to see an example at this point. Consider the simplest possible example, namely, that K is a p -dimensional subspace of \mathbb{R}^n , where $p \leq n$. This is nothing but the linear regression setup, assuming that $\mu = X\beta$, where X is an $n \times p$ matrix of full rank and $\beta \in \mathbb{R}^p$ is arbitrary.

Since K is a subspace, $Z \cdot x = P_K(Z) \cdot x$ for any $x \in K$. Moreover, $P_K(Z)$ is a standard Gaussian random vector in K . A simple application of the rotational invariance of Z shows that we may assume, without loss of generality, that K is simply a copy of \mathbb{R}^p contained in \mathbb{R}^n . Combining these observations, we see that for any $\mu \in K$ and $t \geq 0$,

$$f_\mu(t) = \mathbb{E}\left(\sup_{x \in \mathbb{R}^p, \|x\| \leq t} W \cdot x\right) - \frac{t^2}{2},$$

where W is a p -dimensional standard Gaussian random vector. The above expression can be exactly evaluated, to give

$$f_\mu(t) = \mathbb{E}(t\|W\|) - \frac{t^2}{2}.$$

Clearly, f_μ is maximized at

$$t_\mu = \mathbb{E}\|W\| = \sqrt{p} + O(1),$$

where $O(1)$ denotes a quantity that may be bounded by a constant that does not depend on p or n . By Theorem 1.1, this shows that when K is a p -dimensional subspace of \mathbb{R}^n , then with high probability,

$$\|\hat{\mu} - \mu\| = \sqrt{p} + O(p^{1/4}).$$

Of course, this result may be derived by other means. It is included here only to serve as a simple illustration.

The above example is, in some sense, exceptionally simple. In general, it will be very difficult to compute t_μ exactly, since we have only limited tools at our disposal to compute expected maxima of high-dimensional Gaussian processes. However, the strict concavity of the function f_μ gives an easy way to calculate upper and lower bounds on t_μ (and hence, upper and lower bounds on the estimation error $\|\hat{\mu} - \mu\|$) by calculating bounds on f_μ at a small number of points.

PROPOSITION 1.3. *If $0 \leq r_1 < r_2$ are such that $f_\mu(r_1) \leq f_\mu(r_2)$, then $t_\mu \geq r_1$. On the other hand, if $f_\mu(r_1) \geq f_\mu(r_2)$, then $t_\mu \leq r_2$. In particular, if $\mu \in K$ and $r > 0$ is such that $f_\mu(r) \leq 0$, then $t_\mu \leq r$.*

In Section 2, we will see applications of this proposition in computing matching upper and lower bounds for estimation errors in two nontrivial problems.

1.3. *The LSE is admissible up to a universal constant.* Does the constrained least squares estimator $\hat{\mu}$ enjoy any kind of general optimality property that holds for any K ? A priori, this may sound like a hopeless question due to the completely arbitrary nature of the convex set K . One may hope that the LSE is minimax optimal in some asymptotic sense, but as we will show later, this is not the case. Fortunately, it turns out that $\hat{\mu}$ indeed enjoys a certain other kind of universal optimality property, as shown in Theorem 1.4 below. From a purely mathematical point of view, this is the deepest result of this paper.

The famous Stein paradox [55] shows that the least squares estimate $\hat{\mu}$ is inadmissible under square loss when $K = \mathbb{R}^n$. Stein's example gave birth to the flourishing field of shrinkage estimates. The following theorem shows that although the LSE $\hat{\mu}$ may be inadmissible, it is always “admissible up to a universal constant,” whatever be the set K . In particular, shrinkage—or any other clever idea—cannot improve the risk beyond a universal constant factor everywhere on the parameter space.

THEOREM 1.4. *There is a universal constant $C > 0$ such that the following is true. Take any n and any nonempty closed convex set $K \subseteq \mathbb{R}^n$. Let $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be any Borel measurable map, and for each $\mu \in \mathbb{R}^n$ define the estimate $\tilde{\mu} := g(Z + \mu)$, where Z is a standard Gaussian random vector. Let $\hat{\mu}$ be the least squares estimate $P_K(Z + \mu)$, as in Theorem 1.1. Then there exists $\mu \in K$ such that $\mathbb{E}\|\tilde{\mu} - \mu\|^2 \geq C\mathbb{E}\|\hat{\mu} - \mu\|^2$.*

Again, it may be a good idea to understand the impact of Theorem 1.4 through an example. Consider the problem of ℓ^1 -penalized regression with p covariates, where p may be bigger than n . Here, K is the set of all μ of the form $X\beta$, where X is a given $n \times p$ matrix and β is a point in \mathbb{R}^p with ℓ^1 norm bounded by some pre-specified constant L . The convex-constrained least squares estimate in this problem is the same as the lasso estimate of Tibshirani [58] in its primal form. One may consider various other procedures for computing estimates of β in this problem. Theorem 1.4 says that no matter what procedure one considers, there is always some β with ℓ^1 norm $\leq L$ where the prediction error of the new procedure is at least as big as the prediction error of the lasso, multiplied by a universal constant.

It is interesting to figure out the optimal value of the universal constant in Theorem 1.4. Note that by the Stein paradox, the largest possible value is strictly less than 1.

1.4. *The LSE may not be minimax rate-optimal.* Theorem 1.4 shows that there is always some region of the parameter space where the least squares estimate $\hat{\mu}$ does not perform too badly in comparison to any given competitor. This immediately raises the question as to whether the same is true about the maximum risk: is the maximum risk of the least squares estimate always within a universal constant multiple of the minimax risk? (Here the “risk” of an estimator $\tilde{\mu}$ under square loss

is defined, as usual, to be $\mathbb{E}\|\tilde{\mu} - \mu\|^2$.) Surprisingly, the answer turns out to be negative, as shown by the following counterexample.

Take any n . Define a closed convex set $K \subseteq \mathbb{R}^n$ as follows: take any $\alpha \in [0, 1]$, $\theta_1, \dots, \theta_n \in [-1, 1]$, and let

$$\mu_i := \alpha n^{-1/4} + \alpha \theta_i n^{-1/2}, \quad i = 1, \dots, n.$$

Let K be the set of all $\mu = (\mu_1, \dots, \mu_n)$ obtained as above.

PROPOSITION 1.5. *The set K defined above is closed and convex. As before, let $\hat{\mu} = P_K(Z + \mu)$ be the least squares estimate of $\mu \in K$ obtained by projecting the data vector $Y = Z + \mu$ on to K . Let $\tilde{\mu}$ be the estimate whose coordinates are all equal to the average of the coordinates of Y . Then, under square loss, the maximum risk of $\hat{\mu}$ is bounded below by $C_1 n^{1/2}$ whereas the maximum risk of $\tilde{\mu}$ is bounded above by C_2 , where C_1 and C_2 are positive constants that do not depend on n .*

It is interesting to understand whether this example is a pathological exception, or if there is a general rule that dictates whether the LSE is minimax rate-optimal or not in a given problem. Theorem 1.4 gives a sufficient condition for minimax rate-optimality, namely, that the risk is of the same order everywhere on K . This is expressed quantitatively in the following proposition. But this condition may be difficult to verify in examples.

PROPOSITION 1.6. *Let $\tilde{\mu}$ and $\hat{\mu}$ be as in Theorem 1.4. For each $\mu \in K$, let $R_1(\mu)$ be the risk of $\hat{\mu}$ at μ and $R_2(\mu)$ be the risk of $\tilde{\mu}$ at μ . Then*

$$\sup_{\mu \in K} R_2(\mu) \geq CL \sup_{\mu \in K} R_1(\mu),$$

where C is the universal constant from Theorem 1.4 and

$$L := \frac{\inf_{\mu \in K} R_1(\mu)}{\sup_{\mu \in K} R_1(\mu)}.$$

The counterexample given in Proposition 1.5 also raises the question as to whether there is a general estimator that is guaranteed to be minimax up to a universal constant.

2. Examples. This section contains two nontrivial applications of Theorem 1.1, to supplement the easy example worked out in Section 1.2. We only present the results here. The details are worked out in Section 4.

2.1. *Lasso with nonsingular design.* Let $p \geq 1$ and $n \geq 2$ be two integers, and let X be a given $n \times p$ matrix with real entries. Let L be a positive real number, and let

$$(1) \quad K_0 := \{\beta \in \mathbb{R}^p : |\beta|_1 \leq L\},$$

where $|\beta|_1$ stands for the ℓ^1 norm of β , that is, the sum of the absolute values of the components of β . Let

$$(2) \quad K := \{X\beta : \beta \in K_0\}.$$

The least squares estimator for the convex constraint K is nothing but the lasso estimator in its primal form as defined by Tibshirani [58]. The number L is called the “penalty parameter.”

The theoretical properties of the lasso and related procedures have been extensively studied over the last ten years, notably by Donoho and coauthors [20–26], Knight and Fu [38], Zou [78], Wainwright [73], Candès and Tao [11, 12], Meinshausen and Bühlmann [44], Meinshausen and Yu [45], Koltchinskii [39], Wang and Leng [74], Zhao and Yu [77], Bunea et al. [10], van de Geer [71], Greenshtein and Ritov [32], Bickel et al. [4], Bartlett et al. [3], Rigollet and Tsybakov [52], Oymak et al. [49] and many others. For a more complete set of references and a clear exposition of the results and techniques, see the wonderful recent monograph of Bühlmann and van de Geer [9]. The investigators have tried to understand a number of different kinds of consistency for the lasso estimator. The expected squared error $\mathbb{E}\|\hat{\mu} - \mu\|^2$ translates into what is known as the “squared prediction error” in the lasso literature. Among the papers cited above, the ones dealing mainly with the behavior of the prediction error are [10, 32, 71]. If the prediction error vanishes on an appropriate scale, the lasso procedure is called “risk consistent.”

Risk consistency does not require too many assumptions [9, 17, 32], but the available bounds on the expected squared prediction error are solely upper bounds. Matching lower bounds are not known in any case. In particular, it is not known how the error depends on the choice of the penalty parameter L . Practitioners believe from experience that choosing the penalty parameter correctly is of crucial importance, and this is usually done using cross-validation of some sort, for example, in Tibshirani [58, 59], Greenshtein and Ritov [32], Hastie et al. [35], Efron et al. [30], and van de Geer and Lederer [69], although some other techniques have also been proposed, for example, in Tibshirani and Taylor [60] and Zou et al. [79]. For some nascent theoretical progress on cross-validation for the lasso and further references, see Homrighausen and McDonald [36].

The following theorem demonstrates, for the first time, the critical importance of choosing the correct penalty parameter value. If the penalty parameter L is chosen to be equal to $|\beta|_1$, then the prediction error is vastly smaller than if the two quantities are unequal. Although the theorem is restricted to the case of nonsingular design matrices, we may expect the phenomenon to hold in greater generality.

THEOREM 2.1. *Take any $L > 0$ and let K be defined as in (2). Let $\Sigma := X^T X/n$, and let a and b be the smallest and largest eigenvalues of Σ . Assume that $a > 0$, and that all the diagonal entries of Σ are equal to 1. Take any $\beta \in \mathbb{R}^p$ and let $\mu := X\beta$. Let s be the number of nonzero entries of β . Let*

$$\delta := L - |\beta|_1$$

and $r := p/n$. Let t_μ be as in Theorem 1.1, for the set K defined in (2). If $\delta > 0$, then given any $\varepsilon > 0$ there is a constant C_1 depending only on $\delta, \varepsilon, a, b, s, r$ and L such that whenever $n > C_1$, we have

$$n^{1/4-\varepsilon} \leq t_\mu \leq n^{1/4+\varepsilon}.$$

If $\delta = 0$, then there is a constant C_2 depending only on a, b, s, r and L such that

$$t_\mu \leq C_2 \sqrt{\log n}.$$

Finally, if $\delta < 0$, there are positive constants C_3 and C_4 depending only on δ, a, b, s, r and L such that

$$C_3 \sqrt{n} \leq t_\mu \leq C_4 \sqrt{n}.$$

The reader may easily check the implications of the above bounds on the prediction error by looking back at Theorem 1.1. In particular, they show that the squared prediction error $\mathbb{E}\|X\hat{\beta} - X\beta\|^2$ equals $n^{1/2+o(1)}$ if the penalty parameter L is greater than $|\beta|_1$, is of order n if L is less than $|\beta|_1$, and is bounded above by some constant multiple of $\log n$ if the penalty parameter is chosen correctly, to be equal to $|\beta|_1$. Therefore, it is very important that L is chosen correctly when implementing the lasso procedure. However, there is a caveat: Theorem 2.1 does not prove anything in the case where L is chosen using the data. It only shows the importance of choosing the correct value of the penalty parameter, besides being the first result that establishes a lower bound on the lasso error. To prove an analogous result for the case where L is chosen using the data requires further work.

2.2. Isotonic regression. Define the convex set

$$(3) \quad K := \{(\mu_1, \dots, \mu_n) \in \mathbb{R}^n : \mu_1 \leq \mu_2 \leq \dots \leq \mu_n\}.$$

The least squares problem for this convex constraint, popularly known as “isotonic regression” or “monotone regression,” has a long history in the statistics literature, possibly beginning in Ayer et al. [2] and Grenander [33]. The LSE is easily computed using the so-called “pool adjusted violators algorithm” (see Robertson et al. [53], Chapter 1).

There is substantial literature on the properties of individual $\hat{\mu}_i$, as i/n is fixed and n goes to infinity, with some appropriate limiting behavior assumed for the mean vector μ . Some notable papers on such local errors are those of Prakasa Rao [51], Brunk [8], Groeneboom and Pyke [34], Durot [28], Carolan and Dykstra [13],

Cator [14] and Jankowski [37]. The global error $\|\hat{\mu} - \mu\|$ has also received considerable attention, notably in van de Geer [66, 67], Donoho [20], Birgé and Massart [6], Wang [75], Meyer and Woodroffe [46], Zhang [76] and Chatterjee, Guntoboyina and Sen [18].

It is now generally understood that if the μ_i 's are "strictly increasing" in some limiting sense, then $\hat{\mu}_i - \mu_i$ is typically of order $n^{-1/3}$, whereas the error is smaller if the μ_i 's have "flat stretches" [18]. Therefore, it is natural to expect that in the strictly increasing case, $\|\hat{\mu} - \mu\|$ should be of order $n^{1/6}$. Using Theorem 1.1, it turns out that we may not only get finite sample upper and lower bounds for the global risk $\mathbb{E}\|\hat{\mu} - \mu\|^2$, but also show that $\|\hat{\mu} - \mu\|$ is concentrated around its mean value; that is, there is some constant $C(\mu)$ depending on μ such that with high probability,

$$(4) \quad \|\hat{\mu} - \mu\| = C(\mu)n^{1/6} + O(n^{1/12}).$$

The following theorem makes this precise.

THEOREM 2.2. *Let K be the convex set defined in (3). Take any $\mu \in K$ and let $\hat{\mu} = P_K(Z + \mu)$ be the LSE of μ obtained from the data vector $Z + \mu$. Let*

$$D := \max\{\mu_n - \mu_1, 1\},$$

$$A := \min_{1 \leq i \leq n-1} n(\mu_{i+1} - \mu_i),$$

$$B := \max_{1 \leq i \leq n-1} n(\mu_{i+1} - \mu_i).$$

Let t_μ be as in Theorem 1.1, for the set K defined in (3). Then

$$\frac{C_1 A^{8/3} n^{1/6}}{B^{4/3} D} \leq t_\mu \leq C_2 D^{1/3} n^{1/6},$$

where C_1 and C_2 are positive universal constants.

The reader may easily check the consequences of the above bounds on t_μ by looking back at Theorem 1.1 and Corollary 1.2, and in particular, that it proves (4) when D , A and B are all of constant order.

Just to be clear, the upper bound on the expected mean-squared error that we get from Theorem 2.2 can be derived from existing results such as those in Zhang [76]. The new contribution of Theorem 2.2 is the lower bound, and also the conclusion (in combination with Theorem 1.1) that the squared error concentrates around its expected value.

3. Proof sketches. Since the proofs of the main results (Theorems 1.1 and 1.4) are somewhat technical, I will try to give a readable sketch of the main ideas in this section. The details are given in Section 4.

The proof of Theorem 1.1 goes roughly as follows. Define a random function

$$F_\mu(t) = \sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - \mu) - \frac{t^2}{2},$$

so that $f_\mu(t) = \mathbb{E}(F_\mu(t))$. Using the convexity of K , prove that F_μ and f_μ are both strictly concave functions. Let t^* be the unique point at which F_μ is maximized. Again, use convexity of K and some algebraic manipulation to prove the key identity

$$\|\hat{\mu} - \mu\| = t^*.$$

Note that this is a purely deterministic identity, having nothing to do with the modeling assumptions.

Next, using the concentration of Gaussian maxima, show that $F_\mu(t)$, although random, is close to $f_\mu(t)$ with high probability. Since F_μ and f_μ are two strictly concave functions that are close to each other with high probability, their points of maxima must also be close. That is, $t_\mu \approx t^* = \|\hat{\mu} - \mu\|$ with high probability.

The proof of Theorem 1.4 is more complex and it is quite hard to present the ideas in a nutshell. Still, a high level overview of the main steps may be given as follows.

Throughout this proof sketch, “constant” will mean “positive universal constant.” Take any $\mu^* \in K$ and let B_0 be a ball of radius $C_1 t_{\mu^*}$ around μ^* , where C_1 is a small constant that will be chosen later. Let ρ be a probability measure on B_0 , also to be chosen later. Let $g(Z + \mu)$ be any estimator of μ . Suppose that we are able to prove

$$(5) \quad \int_{B_0} \mathbb{E} \|\mu - g(Z + \mu)\|^2 d\rho(\mu) \geq C_2 t_{\mu^*}^2$$

for some constant C_2 . Upon solving some technical hurdles, it can be shown that the dependence of t_μ on μ is smooth enough to guarantee that if the constant C_1 is chosen small enough, then there is a constant C_3 such that $t_{\mu^*} \geq C_3 t_\mu$ for all $\mu \in B_0$. Combined with (5) and Corollary 1.2, this implies the existence of $\mu_0 \in B_0$ such that

$$\mathbb{E} \|\mu_0 - g(Z + \mu_0)\|^2 \geq C_2 t_{\mu^*}^2 \geq C_2 C_3^2 t_{\mu_0}^2 \geq C_4 \mathbb{E} \|\mu_0 - P_K(Z + \mu_0)\|^2$$

for some constant C_4 , completing the proof.

The main challenge, therefore, is to show (5). We now make a specific choice of ρ . Let ρ be the probability measure of the point v^* that maximizes $Z \cdot (\mu - \mu^*)$ among all $\mu \in B_0$. Let Z' be an independent copy of Z , and let $Y' = Z' + v^*$. Then observe that the expression on the left-hand side of (5) is nothing but $\mathbb{E} \|v^* - g(Y')\|^2$.

The main trick now is the following. Let v' be another K -valued random variable, such that v^* and v' are i.i.d. given Y' . Then it is not difficult to argue that for any measurable function h ,

$$\mathbb{E}\|v^* - h(Y')\|^2 \geq \frac{1}{2}\mathbb{E}\|v^* - v'\|^2.$$

In particular, this holds for $h = g$. Thus, it suffices to show that

$$(6) \quad \mathbb{E}\|v^* - v'\|^2 \geq C_5 t_{\mu^*}^2$$

for some constant C_5 .

Let B_1 be the ball of radius $C_6 t_{\mu^*}$ around v^* , where C_6 is a constant that will be chosen later. Note that unlike B_0 , B_1 is a random set. If we can show that $\mathbb{P}(v' \notin B_1)$ is larger than a universal threshold, it will complete the proof of (6).

To prove this, the first step is to explicitly write down

$$\mathbb{P}(v' \in B_1 | Y', v^*) = \frac{\int_{B_1} e^{-Z' \cdot (v^* - v) - 1/2 \|v^* - v\|^2} d\rho(v)}{\int_{B_0} e^{-Z' \cdot (v^* - v) - 1/2 \|v^* - v\|^2} d\rho(v)}.$$

After a sequence of relatively complicated technical steps involving concentration inequalities and second moment lower bounds, one can produce an upper bound on the expectation of the right-hand side. The complications arise from the fact that the right-hand side is a ratio of random variables. When the dust settles, we get the inequality

$$\mathbb{P}(v' \in B_1) \leq C_7 \sqrt{\mathbb{E}(\rho(B_1))}.$$

The proof, therefore, will be complete if we can show that $\mathbb{E}(\rho(B_1))$ can be made as small as we like by choosing C_6 small enough. By the definition of ρ , it is clear that

$$\rho(B_1) \leq \mathbb{P}(M_2 \geq M_1 | v^*),$$

where

$$M_1 = \sup_{\mu \in B_0} Z' \cdot (\mu - \mu^*), \quad M_2 = \sup_{\mu \in B_1} Z' \cdot (\mu - \mu^*).$$

Consequently, $\mathbb{E}(\rho(B_1)) \leq \mathbb{P}(M_2 \geq M_1)$. To make the right-hand side as small as we need it to be, it makes sense to choose μ^* such that $\mathbb{E}(M_1)$ is as large as possible, and then choose C_6 so small that $\mathbb{E}(M_2)$ is small enough. Working out the details of this step involves delicate technical problems. Carefully solving these problems leads to the completion of the proof of Theorem 1.4.

4. Proofs. This section contains the proofs of all the results stated in Sections 1 and 2. We will follow a certain notational convention about universal constants throughout this section. Within the proof of each lemma or theorem or proposition, C_1, C_2, \dots will denote positive universal constants. The values of the C_i 's may change from one lemma to the next. On the other hand, c_1, c_2, \dots will denote universal constants whose values are important; once defined, they will not change.

The first goal is to prove Theorem 1.1. We need the following ingredient from measure concentration theory.

LEMMA 4.1 (Cirelson, Ibragimov and Sudakov [19]). *Let V_1, \dots, V_n be jointly Gaussian random variables, each with mean zero and second moment bounded above by 1 (but not necessarily independent). Let $M := \max_{1 \leq i \leq n} V_i$. Then for any $t \geq 0$,*

$$\max\{\mathbb{P}(M - \mathbb{E}(M) \geq t), \mathbb{P}(M - \mathbb{E}(M) \leq -t)\} \leq e^{-t^2/2}.$$

The above inequalities were proved in [19], although they follow (with slightly worse constants) from the earlier papers [7] and [57].

We also need a standard fact from convex geometry. A proof is included for the sake of completeness.

LEMMA 4.2 (Projection on to convex sets). *Let K be a nonempty closed convex subset of \mathbb{R}^n . For any $x \in \mathbb{R}^n$, there is a unique point in K , that we call $P_K(x)$, which is closest to x .*

PROOF. Let $s := \inf_{y \in K} \|x - y\|$. Since K is nonempty, s is finite. Let K' be the set of all points in K that are within distance $s + 1$ from x . This is clearly nonempty, convex and bounded. Furthermore, since K is closed, so is K' . The compactness of K' ensures the existence of at least one point in K' that is at distance exactly s from x . This proves the existence of a projection. Suppose now that there are two points y and z in K that are both at distance exactly s from x . Then the three points x, y and z form an isosceles triangle with the line segment joining y and z as the base. But this line segment is contained in K , because K is convex. Since $y \neq z$, this proves that there is a point in K that is at distance strictly less than s from x , which is impossible. \square

We are now ready to prove Theorem 1.1, Corollary 1.2 and Proposition 1.3.

PROOF OF THEOREM 1.1. Fix $\mu \in \mathbb{R}^n$ and let $Y = Z + \mu$. Define two random functions M and F from $[0, \infty)$ into $[-\infty, \infty)$ as

$$M(t) := \sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - \mu)$$

and

$$F(t) := M(t) - \frac{t^2}{2},$$

with the usual convention that the supremum of an empty set is $-\infty$. Let $m(t) := \mathbb{E}(M(t))$. Note that $\mathbb{E}(F(t)) = f_\mu(t) = m(t) - t^2/2$.

Note that $M(t)$, $F(t)$, $m(t)$ and $f_\mu(t)$ are all finite if $t \geq t_c$, and $-\infty$ if $t < t_c$. Take any $t_c \leq s \leq t$. Let v_1 and v_2 be points in K such that

$$(7) \quad \|v_1 - \mu\| \leq s \quad \text{and} \quad \|v_2 - \mu\| \leq t.$$

Take any $u \in [0, 1]$ and let $v := uv_1 + (1 - u)v_2$. Then $\|v - \mu\| \leq r := us + (1 - u)t$. On the other hand,

$$Z \cdot (v - \mu) = uZ \cdot (v_1 - \mu) + (1 - u)Z \cdot (v_2 - \mu).$$

Maximizing over all v_1 and v_2 satisfying (7), this gives

$$(8) \quad M(r) \geq uM(s) + (1 - u)M(t).$$

Thus, M is a concave function of t . Consequently, F is strictly concave. Note that $\lim_{t \rightarrow \infty} F(t) = -\infty$, since

$$(9) \quad M(t) \leq \sup_{v \in \mathbb{R}^n : \|v - \mu\| \leq t} Z \cdot (v - \mu) = t\|Z\|.$$

The strict concavity and the decay to $-\infty$ prove the existence and uniqueness of a (random) point $t^* \in [t_c, \infty)$ where F is maximized.

Taking expectation on both sides in (8) implies that m is also concave, and therefore f_μ is strictly concave. Similarly by (9), $m(t) \leq t\mathbb{E}\|Z\|$, which proves that $\lim_{t \rightarrow \infty} f_\mu(t) = -\infty$. Therefore, t_μ exists and is unique.

Let v^* be a point in K that maximizes $Z \cdot (v - \mu)$ among all $v \in K$ satisfying $\|v - \mu\| \leq t^*$. Let $t_0 := \|v^* - \mu\|$. If $t_0 < t^*$, then

$$F(t_0) \geq Z \cdot (v^* - \mu) - \frac{t_0^2}{2} = M(t^*) - \frac{t_0^2}{2} > F(t^*),$$

which is false. Therefore, $t_0 = t^*$. This shows that for any $v \in K$,

$$\begin{aligned} Z \cdot (v - \mu) - \frac{\|v - \mu\|^2}{2} &\leq F(\|v - \mu\|) \\ &\leq F(t^*) = Z \cdot (v^* - \mu) - \frac{\|v^* - \mu\|^2}{2}. \end{aligned}$$

Since

$$\|Y - v\|^2 = \|Y - \mu\|^2 - 2\left(Z \cdot (v - \mu) - \frac{\|v - \mu\|^2}{2}\right),$$

this proves that $\|Y - v\| \geq \|Y - v^*\|$ for all $v \in K$. Therefore, by the uniqueness of projection on to closed convex sets, $\hat{\mu} = v^*$. In particular,

$$\|\mu - \hat{\mu}\| = t^*.$$

Now note that for any $t \geq t_c$, the inequality $f_\mu(t) \leq f_\mu(t_\mu)$ may be rewritten as

$$(10) \quad m(t) \leq m(t_\mu) + \frac{t^2 - t_\mu^2}{2}.$$

By concavity of m , for any $\varepsilon \in (0, 1)$,

$$(11) \quad m((1 - \varepsilon)t_\mu + \varepsilon t) \geq (1 - \varepsilon)m(t_\mu) + \varepsilon m(t).$$

Applying (10) to $(1 - \varepsilon)t_\mu + \varepsilon t$ instead of t gives

$$m((1 - \varepsilon)t_\mu + \varepsilon t) \leq m(t_\mu) + \frac{(-2\varepsilon + \varepsilon^2)t_\mu^2 + 2(1 - \varepsilon)\varepsilon t_\mu t + \varepsilon^2 t^2}{2}.$$

Combining this inequality with (11) gives

$$\varepsilon m(t) \leq \varepsilon m(t_\mu) + \frac{(-2\varepsilon + \varepsilon^2)t_\mu^2 + 2(1 - \varepsilon)\varepsilon t_\mu t + \varepsilon^2 t^2}{2}.$$

Dividing both sides by ε and taking $\varepsilon \rightarrow 0$, we get

$$(12) \quad m(t) \leq m(t_\mu) - t_\mu^2 + t_\mu t,$$

which may be rewritten as

$$f_\mu(t) \leq f_\mu(t_\mu) - \frac{(t - t_\mu)^2}{2}.$$

Note that the above two inequalities hold even if $t < t_c$. Take any $x > 0$ and let $r_1 := t_\mu - x\sqrt{t_\mu}$ and $r_2 := t_\mu + x\sqrt{t_\mu}$. First assume that $r_1 \geq t_c$. Then by the above inequality,

$$\max\{f_\mu(r_1), f_\mu(r_2)\} \leq f_\mu(t_\mu) - \frac{x^2 t_\mu}{2}.$$

By the concentration inequality for maxima of Gaussian random variables (Lemma 4.1), for any $t \geq 0$ and $y \geq 0$,

$$\max\{\mathbb{P}(F(t) \geq f_\mu(t) + y), \mathbb{P}(F(t) \leq f_\mu(t) - y)\} \leq e^{-y^2/2t^2}.$$

Taking $y = x^2 t_\mu / 4$ and $z = f_\mu(t_\mu) - y$, a combination of the last two displays gives the inequalities

$$\mathbb{P}(F(r_1) \geq z) \leq \mathbb{P}(F(r_1) \geq f_\mu(r_1) + y) \leq e^{-y^2/2r_1^2},$$

$$\mathbb{P}(F(r_2) \geq z) \leq \mathbb{P}(F(r_2) \geq f_\mu(r_2) + y) \leq e^{-y^2/2r_2^2},$$

$$\mathbb{P}(F(t_\mu) \leq z) = \mathbb{P}(F(t_\mu) \leq f_\mu(t_\mu) - y) \leq e^{-y^2/2t_\mu^2}.$$

Let E be the event that $F(r_1) < z$, $F(r_2) < z$ and $F(t_\mu) > z$. By the above three inequalities,

$$\mathbb{P}(E^c) \leq e^{-y^2/2r_1^2} + e^{-y^2/2r_2^2} + e^{-y^2/2t_\mu^2} \leq 3e^{-y^2/2r_2^2}.$$

On the other hand, by the concavity of F , if E happens then t^* must lie in the interval (r_1, r_2) . Together with our previous observation that $t^* = \|\mu - \hat{\mu}\|$, this completes the proof of the theorem when $r_1 \geq t_c$.

If $r_1 < t_c$, the inequality $f_\mu(r_2) \leq f_\mu(t_\mu) - x^2 t_\mu / 2$ is still true. Redefine E to be the event that $F(r_2) < z$ and $F(t_\mu) > z$. Then the upper bound on $\mathbb{P}(E^c)$ is still valid, and the occurrence of E implies that $t^* \in [t_c, r_2) \subseteq (r_1, r_2)$. This finishes the argument in the case $r_1 < t_c$. \square

PROOF OF COROLLARY 1.2. Throughout this proof, C denotes an arbitrary universal constant whose value may change from line to line. First, suppose that $t_\mu \geq 1$. Then by Theorem 1.1,

$$\mathbb{P}(\|\hat{\mu} - \mu\| - t_\mu \geq x\sqrt{t_\mu}) \leq 3e^{-x^4/32(1+x)^2}.$$

This shows that

$$\mathbb{E}(\|\hat{\mu} - \mu\| - t_\mu)^2 \leq C t_\mu,$$

which gives the first set of inequalities. On the other hand, if $t_\mu < 1$, then putting $z = x\sqrt{t_\mu}$, Theorem 1.1 gives

$$\mathbb{P}(\|\hat{\mu} - \mu\| - t_\mu \geq z) \leq 3e^{-z^4/32(t_\mu+z)^2} \leq 3e^{-z^4/32(1+z)^2},$$

which gives the second inequality. \square

PROOF OF PROPOSITION 1.3. The first two assertions are obvious by the strict concavity of f_μ . For the third one, observe that if $\mu \in K$, then $f_\mu(0) = 0$, and apply the second assertion. \square

The next goal is to prove Theorem 1.4. In addition to Lemmas 4.1 and 4.2, we need a few more standard results. The first result, stated below, is called the ‘‘Gaussian concentration inequality.’’

LEMMA 4.3 (Gaussian concentration inequality). *Let Z be an n -dimensional standard Gaussian random vector, and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that satisfies $|f(x) - f(y)| \leq L\|x - y\|$ for all x and y , where L is a positive constant. Then for any $\theta \in \mathbb{R}$,*

$$\mathbb{E}(e^{\theta(f(Z) - \mathbb{E}(f(Z)))}) \leq e^{L^2\theta^2/2}.$$

Consequently, for any $t \geq 0$,

$$\max\{\mathbb{P}(f(Z) - \mathbb{E}(f(Z)) \geq t), \mathbb{P}(f(Z) - \mathbb{E}(f(Z)) \leq -t)\} \leq e^{-t^2/2L^2}.$$

This famous result possibly appeared for the first time as an implied consequence of the theorems in [7, 19, 57]. For a simple proof, originally appearing in [19], see the argument following equation (2.35) in [40].

We also need the fact that the projection P_K on to a closed convex set is a contraction with respect to the Euclidean norm. This, again, is quite standard but we provide a short proof for the sake of completeness.

LEMMA 4.4. *For any closed convex set $K \subseteq \mathbb{R}^n$ and any x, y , $\|P_K(x) - P_K(y)\| \leq \|x - y\|$.*

PROOF. Let $z = P_K(x)$ and $w = P_K(y)$. If $z = w$, there is nothing to prove. So assume that $z \neq w$. Let S be the line segment joining z and w . Then S is entirely contained in K . Let H_1 be the hyperplane passing through z that is orthogonal to S , and H_2 be the hyperplane passing through w that is orthogonal to S .

The hyperplane H_1 divides $\mathbb{R}^n \setminus H_1$ into two open half-spaces, one of which contains w . If x belongs to the half-space that contains w , then there is a point on S that is closer to x than z . This is impossible. Similarly, H_2 divides $\mathbb{R}^n \setminus H_2$ into two open half-spaces, and y cannot belong to the one that contains z . Therefore, both of the parallel hyperplanes H_1 and H_2 must lie between x and y . This proves that $\|x - y\| \geq$ the distance between H_1 and H_2 , which is equal to $\|z - w\|$. \square

Finally, we need the so-called “second moment inequality,” also known as the “Paley–Zygmund inequality.”

LEMMA 4.5 (Second moment inequality). *If X is a nonnegative random variable with $\mathbb{E}(X) > 0$ and finite second moment, then for any $a \in [0, \mathbb{E}(X)]$,*

$$\mathbb{P}(X > a) \geq \frac{(\mathbb{E}(X) - a)^2}{\mathbb{E}(X^2)}.$$

The proof of this standard inequality may be found in graduate probability text books such as [29].

We now embark on the proof of Theorem 1.4. Several preparatory lemmas are required.

LEMMA 4.6. *Let $K \subseteq \mathbb{R}^n$ be a line segment of length l . Let Z be an n -dimensional standard Gaussian random vector. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be any Borel measurable map. Then there exists $\mu \in K$ such that*

$$\mathbb{E}\|f(Z + \mu) - \mu\|^2 \geq c_1 \min\{l^2, 4\},$$

where c_1 is a positive universal constant.

PROOF. Let λ denote the uniform distribution on K . Let v be point chosen uniformly at random from K . Let $Y := Z + v$. Given Y and v , let v' be drawn from the posterior distribution of v given Y . Explicitly, if θ denotes the joint law of (v, Y, v') , then

$$(13) \quad d\theta(\mu, y, \mu') = \frac{e^{-1/2\|y-\mu'\|^2} e^{-1/2\|y-\mu\|^2}}{(2\pi)^{n/2} \int_K e^{-1/2\|y-x\|^2} d\lambda(x)} d\lambda(\mu) dy d\lambda(\mu').$$

The above expression clearly shows that v and v' are i.i.d. given Y and, therefore,

$$(14) \quad \begin{aligned} \mathbb{E}(\|f(Y) - v\|^2 | Y) &\geq \mathbb{E}(\|\mathbb{E}(v|Y) - v\|^2 | Y) \\ &= \frac{1}{2} \mathbb{E}(\|v' - v\|^2 | Y). \end{aligned}$$

[In the above display, $\mathbb{E}(v|Y)$ denotes the random vector whose i th coordinate is $\mathbb{E}(v_i|Y)$. The inequality in the first line is simply a consequence of the fact that for any random variable X , $\mathbb{E}(X - a)^2$ is minimized when $a = \mathbb{E}(X)$.] Next, let

$$M := \sup_{x \in K} |Z \cdot (x - v)|.$$

Note that since the function being maximized is convex and the set K is a line segment, therefore the maximum is necessarily attained at one of the endpoints of the line segment K . From this, it easy to see that

$$(15) \quad \mathbb{E}(e^{2M}) \leq C_1 e^{C_1 l^2}.$$

Take any $\varepsilon > 0$. Given Y and v , let I denote the set of all points in K that are within distance ε from v . Then by (13),

$$\mathbb{P}(\|v' - v\| \leq \varepsilon | Y, v) = \frac{\int_I e^{Z \cdot (x-v) - 1/2\|x-v\|^2} d\lambda(x)}{\int_K e^{Z \cdot (x-v) - 1/2\|x-v\|^2} d\lambda(x)} \leq \frac{2e^{2M+l^2} \varepsilon}{l}.$$

Taking expectation and applying (15), we get

$$\mathbb{P}(\|v' - v\| \leq \varepsilon) \leq \frac{C_2 e^{C_2 l^2} \varepsilon}{l},$$

and, therefore,

$$\mathbb{E}\|v' - v\|^2 \geq \varepsilon^2 \mathbb{P}(\|v' - v\| > \varepsilon) \geq \varepsilon^2 \left(1 - \frac{C_2 e^{C_2 l^2} \varepsilon}{l}\right).$$

If $l \leq 2$, then combined with (14) and taking $\varepsilon = C_3 l$ for some small enough C_3 , this proves that

$$\mathbb{E}\|f(Z + v) - v\|^2 \geq C_4 l^2.$$

In particular, there exists $\mu \in K$ such that

$$\mathbb{E}\|f(Z + \mu) - \mu\|^2 \geq C_4 l^2.$$

If $l > 2$, then choose a subinterval $K' \subseteq K$ of length ≤ 2 and work with K' instead of K . \square

LEMMA 4.7. *There is a positive universal constant c_2 such that following is true. Let K be a closed convex subset of \mathbb{R}^n with diameter ≥ 2 . For each $\mu \in K$, let t_μ be defined as in Theorem 1.1. Then $t_\mu \geq c_2 n^{-1/2}$ for all $\mu \in K$.*

PROOF. Take any $\mu \in K$. Since the diameter of K is ≥ 2 , there exists $v \in K$ such that $\|v - \mu\| \geq 1$. By the convexity of K , this implies that there exists $v \in K$ such that $\|v - \mu\| = 1$. For each $t \in [0, 1]$ let $v_t := (1 - t)\mu + tv$. Then $v_t \in K$ and $\|v_t - \mu\| = t$. Therefore, there exists positive C_1 and C_2 such that if $t \leq C_1$ then

$$f_\mu(t) \geq \mathbb{E}(\max\{0, Z \cdot (v_t - \mu)\}) - \frac{t^2}{2} \geq C_2 t.$$

On the other hand, by (9),

$$f_\mu(t) \leq C_3 t \sqrt{n}.$$

Thus, with $C_4 := C_1 C_2 / C_3$,

$$f_\mu(C_4 n^{-1/2}) \leq C_2 C_1 \leq f_\mu(C_1).$$

Taking C_3 large enough, we have $C_4 n^{-1/2} < C_1$. By Proposition 1.3, this shows that $t_\mu \geq C_4 n^{-1/2}$. \square

LEMMA 4.8. *Let K be a closed convex subset of \mathbb{R}^n and let t_μ be defined as in Theorem 1.1. Then for any $\mu, v \in K$ such that $\|\mu - v\| \leq t_\mu / 24$,*

$$\frac{11t_\mu}{24} \leq t_v \leq \frac{50t_\mu}{24}.$$

PROOF. If $t_\mu = 0$ there is nothing to prove. So assume that $t_\mu > 0$. For any $\gamma \in K$ and $t \geq 0$, let

$$(16) \quad B(\gamma, t) := \{\gamma' \in K : \|\gamma' - \gamma\| \leq t\}$$

and

$$(17) \quad m_\gamma(t) := \mathbb{E}\left(\sup_{\gamma' \in B(\gamma, t)} Z \cdot (\gamma' - \gamma)\right) = \mathbb{E}\left(\sup_{\gamma' \in B(\gamma, t)} Z \cdot \gamma'\right).$$

Let $B_0 := B(\mu, r)$, where $r := t_\mu / 24$. Take any $v \in B_0$. Note that for any positive integer k ,

$$B(\mu, (k - 1)r) \subseteq B(v, kr) \subseteq B(\mu, (k + 1)r),$$

and, therefore,

$$(18) \quad m_\mu((k - 1)r) \leq m_v(kr) \leq m_\mu((k + 1)r).$$

Applying (18) with $k = 11$ gives

$$m_v(11r) \leq m_\mu(12r) = m_\mu(t_\mu / 2),$$

and with $k = 25$, we get

$$m_v(25r) \geq m_\mu(24r) = m_\mu(t_\mu).$$

Therefore, by the inequality (12) from the proof of Theorem 1.1,

$$\begin{aligned} f_v(25r) - f_v(11r) &= m_v(25r) - m_v(11r) - \frac{(25^2 - 11^2)r^2}{2} \\ &\geq m_\mu(t_\mu) - m_\mu(t_\mu/2) - \frac{252t_\mu^2}{576} \geq \frac{t_\mu^2}{2} - \frac{7t_\mu^2}{16} \geq 0. \end{aligned}$$

Therefore, by Proposition 1.3,

$$t_v \geq 11r = \frac{11t_\mu}{24}.$$

Next, note that by (18),

$$\begin{aligned} f_v(50r) - f_v(25r) &= m_v(50r) - m_v(25r) - \frac{1875r^2}{2} \\ &\leq m_\mu(51r) - m_\mu(24r) - \frac{1875r^2}{2}. \end{aligned}$$

By the inequality (12),

$$m_\mu(51r) - m_\mu(24r) = m_\mu(27r + t_\mu) - m_\mu(t_\mu) \leq 27rt_\mu = 648r^2.$$

Combining the last two displays gives

$$f_v(50r) - f_v(25r) \leq 648r^2 - \frac{1875r^2}{2} \leq 0.$$

By Proposition 1.3, this proves that $t_v \leq 50r$. \square

We are now ready to prove Theorem 1.4.

PROOF OF THEOREM 1.4. First, suppose that $l := \text{diam}(K) \leq 2$. Choose a line segment $I \subseteq K$ of length l . By Lemma 4.6, there exists $\mu_0 \in I$ such that $\mathbb{E}\|g(Z + \mu_0) - \mu_0\|^2 \geq c_1l^2$. But $\mathbb{E}\|P_K(Z + \mu_0) - \mu_0\|^2 \leq l^2$, since any two elements of K are within distance l of each other. This completes the proof of the theorem when $\text{diam}(K) \leq 2$. For the rest of the proof, assume that $\text{diam}(K) > 2$.

For $\mu \in K$ and $t \geq 0$, let $m_\mu(t)$ be defined as in (17). Since $\text{diam}(K) > 2$, Lemma 4.7 implies that for all $\mu \in K$,

$$(19) \quad t_\mu \geq c_2n^{-1/2}.$$

Let

$$s := \sup_{\mu \in K} m_\mu(10^{-3}t_\mu).$$

Then there exists at least one point $\mu^* \in K$ such that

$$m_{\mu^*}(10^{-3}t_{\mu^*}) \geq s - \frac{c_2^2}{10^6 n}.$$

For $v \in K$ and $t \geq 0$ let $B(v, t)$ be defined as in (16). Let $B_0 := B(\mu^*, r)$, where $r := 10^{-3}t_{\mu^*}$. Lemma 4.8 implies that for all $v \in B_0$,

$$(20) \quad \frac{11t_{\mu^*}}{24} \leq t_v \leq \frac{50t_{\mu^*}}{24}.$$

Define a probability measure ρ on B_0 as follows. Let v^* be the point that maximizes $Z \cdot (v - \mu^*)$ among all $v \in B_0$. If there are more than one such points, take the one that is the least in the lexicographic ordering (it is easy to prove that there is a least element since the set of maximizers is closed). Let ρ be the law of v^* . Let Z' be a standard Gaussian random vector, independent of Z . Let $Y' := Z' + v^*$. Let ρ' be the conditional distribution of v^* given Y' . It is easy to see that

$$d\rho'(v) = L^{-1} e^{-1/2\|Y'-v\|^2} d\rho(v), \quad v \in B_0,$$

where

$$L := \int_{B_0} e^{-1/2\|Y'-v\|^2} d\rho(v).$$

Given Y' and v^* , let v' be a random point generated from the distribution ρ' . Then v' and v^* are conditionally i.i.d. given Y' , as is evident from the joint law θ of the triple (v^*, Y', v') :

$$d\theta(v_1, y, v_2) = \frac{e^{-1/2\|y-v_2\|^2} e^{-1/2\|y-v_1\|^2}}{(2\pi)^{n/2} \int_{B_0} e^{-1/2\|y-v\|^2} d\rho(v)} d\rho(v_1) dy d\rho(v_2),$$

where $v_1, v_2 \in B_0$ and $y \in \mathbb{R}^n$.

Let $\mathbb{E}(v^*|Y')$ be the random vector whose i th coordinate is $\mathbb{E}(v_i^*|Y')$ and g be an arbitrary Borel measurable map from \mathbb{R}^n into itself, as in the statement of Theorem 1.4. Then

$$\begin{aligned} \mathbb{E}(\|v^* - v'\|^2|Y') &= 2\mathbb{E}(\|v^* - \mathbb{E}(v^*|Y')\|^2|Y') \\ &\leq 2\mathbb{E}(\|v^* - g(Y')\|^2|Y'). \end{aligned}$$

Thus,

$$(21) \quad \mathbb{E}\|v^* - v'\|^2 \leq 2\mathbb{E}\|v^* - g(Y')\|^2.$$

Let B_1 denote the (random) set $B(v^*, 10^{-3}r) \cap B_0$. Then

$$(22) \quad \begin{aligned} \mathbb{P}(v' \in B_1|Y', v^*) &= L^{-1} \int_{B_1} e^{-1/2\|Y'-v\|^2} d\rho(v) \\ &= \frac{\int_{B_1} e^{-Z' \cdot (v^*-v) - 1/2\|v^*-v\|^2} d\rho(v)}{\int_{B_0} e^{-Z' \cdot (v^*-v) - 1/2\|v^*-v\|^2} d\rho(v)}. \end{aligned}$$

Let L_1 and L_2 denote the numerator and the denominator in the last expression. First, note that

$$\begin{aligned}
 \mathbb{E}(L_1^2|v^*) &\leq \int_{B_1} \mathbb{E}(e^{-2Z' \cdot (v^* - v) - \|v^* - v\|^2} | v^*) d\rho(v) \\
 (23) \qquad \qquad &= \int_{B_1} e^{\|v^* - v\|^2} d\rho(v) \leq e^{10^{-6}r^2} \rho(B_1).
 \end{aligned}$$

Next, note that $\mathbb{E}(L_2|v^*) = 1$, and

$$\begin{aligned}
 &\mathbb{E}(L_2^2|v^*) \\
 &= \int_{B_0} \int_{B_0} \mathbb{E}(e^{-Z' \cdot ((v^* - v_1) + (v^* - v_2)) - 1/2(\|v^* - v_1\|^2 + \|v^* - v_2\|^2)}) d\rho(v_1) d\rho(v_2) \\
 &= \int_{B_0} \int_{B_0} e^{(v^* - v_1) \cdot (v^* - v_2)} d\rho(v_1) d\rho(v_2) \leq e^{4r^2}.
 \end{aligned}$$

Therefore by the second moment inequality (Lemma 4.5),

$$(24) \qquad \mathbb{P}(L_2 > 1/2|v^*) \geq \frac{(\mathbb{E}(L_2|v^*))^2}{4\mathbb{E}(L_2^2|v^*)} \geq \frac{1}{4}e^{-4r^2}.$$

Now note that, by a slight abuse of notation,

$$\frac{\partial}{\partial Z'_i} \log L_2 = -\frac{1}{L_2} \int_{B_0} (v_i^* - v_i) e^{-Z' \cdot (v^* - v) - 1/2\|v^* - v\|^2} d\rho(v).$$

Consequently,

$$\begin{aligned}
 (25) \qquad \sum_{i=1}^n \left(\frac{\partial}{\partial Z'_i} \log L_2 \right)^2 &\leq \frac{\int_{B_0} \|v^* - v\|^2 e^{-Z' \cdot (v^* - v) - 1/2\|v^* - v\|^2} d\rho(v)}{\int_{B_0} e^{-Z' \cdot (v^* - v) - 1/2\|v^* - v\|^2} d\rho(v)} \\
 &\leq 4r^2.
 \end{aligned}$$

Therefore, by the Gaussian concentration inequality (Lemma 4.3), for any $x \geq 0$,

$$(26) \qquad \mathbb{P}(\log L_2 \geq \mathbb{E}(\log L_2|v^*) + x|v^*) \leq e^{-x^2/8r^2}.$$

Now suppose that $4r^2 > \log 4$, or in other words,

$$(27) \qquad t_{\mu^*} > 500\sqrt{2 \log 2}.$$

Under the above condition, taking $x = 8r^2$ in (26) gives

$$\mathbb{P}(\log L_2 \geq \mathbb{E}(\log L_2|v^*) + 8r^2|v^*) \leq e^{-8r^2} < \frac{1}{4}e^{-4r^2}.$$

Comparing this with (24), we realize that under (27), it must be true that

$$\mathbb{E}(\log L_2|v^*) \geq -8r^2 - \log 2.$$

Therefore, if (27) holds, then

$$\begin{aligned} \mathbb{E}(L_2^{-2}|v^*) &= e^{-2\mathbb{E}(\log L_2|v^*)} \mathbb{E}(e^{-2(\log L_2 - \mathbb{E}(\log L_2|v^*))}|v^*) \\ &\leq 4e^{16r^2} \mathbb{E}(e^{-2(\log L_2 - \mathbb{E}(\log L_2|v^*))}|v^*). \end{aligned}$$

But by the Gaussian concentration inequality (Lemma 4.3) and the estimate (25),

$$\mathbb{E}(e^{-2(\log L_2 - \mathbb{E}(\log L_2|v^*))}|v^*) \leq e^{8r^2}.$$

Combining the last two displays gives

$$(28) \quad \mathbb{E}(L_2^{-2}|v^*) \leq 4e^{24r^2}.$$

By (22), (23) and (28), we see that under condition (27),

$$(29) \quad \begin{aligned} \mathbb{P}(v' \in B_1|v^*) &= \mathbb{E}(L_1 L_2^{-1}|v^*) \\ &\leq (\mathbb{E}(L_1^2|v^*) \mathbb{E}(L_2^{-2}|v^*))^{1/2} \\ &\leq 2e^{13r^2} \sqrt{\rho(B_1)}. \end{aligned}$$

Define

$$\begin{aligned} M_1 &:= \sup_{v \in B_0} Z' \cdot (v - \mu^*), & M_2 &:= \sup_{v \in B_1} Z' \cdot (v - v^*), \\ M_3 &:= Z' \cdot (v^* - \mu^*). \end{aligned}$$

The basic fact, easy to see, is that

$$(30) \quad \begin{aligned} \rho(B_1) &\leq \mathbb{P}\left(\sup_{v \in B_1} Z' \cdot (v - \mu^*) \geq \sup_{v \in B_0} Z' \cdot (v - \mu^*)|v^*\right) \\ &\leq \mathbb{P}(M_2 + M_3 \geq M_1|v^*). \end{aligned}$$

Having understood this, note that by the definitions of μ^* and s and the lower bounds (19) and (20),

$$(31) \quad \begin{aligned} \mathbb{E}(M_1|v^*) &= m_{\mu^*}(10^{-3}t_{\mu^*}) \geq s - \frac{c_2^2}{10^6 n} \\ &\geq m_{v^*}(10^{-3}t_{v^*}) - \frac{t_{\mu^*}^2}{10^6} \geq m_{v^*}(11r/24) - r^2. \end{aligned}$$

On the other hand,

$$(32) \quad \mathbb{E}(M_2|v^*) \leq m_{v^*}(10^{-3}r).$$

Let $\delta := 11r/24 - 10^{-3}r$. By the concavity of m_{v^*} , and the inequalities (12) and (20),

$$\begin{aligned}
 & m_{v^*}(11r/24) - m_{v^*}(10^{-3}r) \\
 &= m_{v^*}(11r/24) - m_{v^*}(11r/24 - \delta) \\
 (33) \quad & \geq m_{v^*}(t_{v^*}) - m_{v^*}(t_{v^*} - \delta) \\
 & \geq t_{v^*}\delta \geq \frac{11t_{\mu^*}\delta}{24} \geq \frac{110t_{\mu^*}r}{24^2} \geq 100r^2.
 \end{aligned}$$

By (31), (32) and (33), we see that

$$\mathbb{E}(M_1|v^*) - \mathbb{E}(M_2|v^*) \geq 99r^2.$$

Let $x = 33r^2$. Then by the above inequality,

$$\begin{aligned}
 & \mathbb{P}(M_2 + M_3 \geq M_1|v^*) \\
 & \leq \mathbb{P}(M_1 \leq \mathbb{E}(M_1|v^*) - x|v^*) \\
 & \quad + \mathbb{P}(M_2 \geq \mathbb{E}(M_1|v^*) - 2x|v^*) + \mathbb{P}(M_3 \geq x|v^*) \\
 & \leq \mathbb{P}(M_1 \leq \mathbb{E}(M_1|v^*) - x|v^*) \\
 & \quad + \mathbb{P}(M_2 \geq \mathbb{E}(M_2|v^*) + x|v^*) + \mathbb{P}(M_3 \geq x|v^*).
 \end{aligned}$$

By the concentration inequality for Gaussian maxima (Lemma 4.1) and the fact that $\mathbb{E}(M_3|v^*) = 0$, this shows that

$$\begin{aligned}
 \mathbb{P}(M_2 + M_3 \geq M_1|v^*) & \leq e^{-x^2/2r^2} + e^{-x^2/2(10^{-3}r)^2} + e^{-x^2/2r^2} \\
 & \leq 3 \exp(-500r^2).
 \end{aligned}$$

Combined with (29) and (30), this shows that if (27) holds, then

$$\mathbb{P}(v' \in B_1|v^*) \leq C_1 \exp(-C_2 t_{\mu^*}^2).$$

Therefore, there is a universal constant $C_3 \geq 500\sqrt{2\log 2}$ such that if $t_{\mu^*} \geq C_3$, then

$$\mathbb{E}\|v' - v^*\|^2 \geq (10^{-3}r)^2 \mathbb{P}(v' \notin B_1) \geq C_4 t_{\mu^*}^2,$$

and so by (21),

$$\mathbb{E}\|v^* - g(Z' + v^*)\|^2 \geq C_5 t_{\mu^*}^2.$$

Since

$$\mathbb{E}\|v^* - g(Z' + v^*)\|^2 = \int_{B_0} \mathbb{E}\|\mu - g(Z' + \mu)\|^2 d\rho(\mu),$$

this shows that there exists $\mu_0 \in B_0$ such that

$$\mathbb{E}\|\mu_0 - g(Z + \mu_0)\|^2 \geq C_5 t_{\mu^*}^2.$$

By (20), $t_{\mu^*} \geq 24t_{\mu_0}/50$. On the other hand if $t_{\mu^*} \geq C_3$, then by (20), $t_{\mu_0} \geq 11t_{\mu^*}/24 \geq 200\sqrt{2\log 2}$. Therefore, by Corollary 1.2,

$$\mathbb{E}\|\mu_0 - g(Z + \mu_0)\|^2 \geq C_6 t_{\mu_0}^2 \geq C_7 \mathbb{E}\|\mu_0 - P_K(Z + \mu_0)\|^2.$$

This completes the proof of the theorem when $t_{\mu^*} \geq C_3$ and $\text{diam}(K) > 2$.

Suppose now that $t_{\mu^*} < C_3$ and $\text{diam}(K) > 2$. For each μ , let

$$l_\mu^2 := \mathbb{E}\|P_K(Z + \mu) - \mu\|^2.$$

Then by Corollary 1.2, $l_{\mu^*} \leq C_8$. Let I be a line segment in K of length 1, with one endpoint at μ^* . By Lemma 4.6, there exists $\mu_0 \in I$ such that

$$(34) \quad \mathbb{E}\|g(Z + \mu_0) - \mu_0\|^2 \geq c_1.$$

On the other hand, by Lemma 4.4,

$$\begin{aligned} & \|P_K(Z + \mu_0) - \mu_0\| \\ & \leq \|P_K(Z + \mu_0) - P_K(Z + \mu^*)\| + \|P_K(Z + \mu^*) - \mu^*\| + \|\mu^* - \mu_0\| \\ & \leq \|P_K(Z + \mu^*) - \mu^*\| + 2\|\mu^* - \mu_0\| \\ & \leq \|P_K(Z + \mu^*) - \mu^*\| + 2. \end{aligned}$$

Consequently,

$$\mathbb{E}\|P_K(Z + \mu_0) - \mu_0\|^2 \leq 2l_{\mu^*}^2 + 8 \leq C_9.$$

Together with (34), this completes the proof of the theorem when $t_{\mu^*} < C_3$ and $\text{diam}(K) > 2$. \square

The next goal is to prove Proposition 1.5. The proof is a simple consequence of Proposition 1.3. We just have to carry out some computations to verify the conditions of Proposition 1.3.

PROOF OF PROPOSITION 1.5. We have to first prove that the set K is closed and convex. It is obviously closed, and it is convex because for any $\alpha, \alpha' \in [0, 1]$ and $\theta_i, \theta'_i \in [-1, 1]$,

$$\begin{aligned} & t(\alpha n^{-1/4} + \alpha \theta_i n^{-1/2}) + (1-t)(\alpha' n^{-1/4} + \alpha' \theta'_i n^{-1/2}) \\ & = \alpha_t n^{-1/4} + \alpha_t \theta_{i,t} n^{-1/2}, \end{aligned}$$

where

$$\alpha_t = t\alpha + (1-t)\alpha' \in [0, 1]$$

and

$$\theta_{i,t} = \frac{t\alpha\theta_i + (1-t)\alpha'\theta'_i}{t\alpha + (1-t)\alpha'} \in [-1, 1].$$

Let $\bar{Y} := \sum_{i=1}^n Y_i/n$, so that the components of $\tilde{\mu}$ are all equal to \bar{Y} . Defining $\bar{\mu} = \sum_{i=1}^n \mu_i/n$ and $\bar{\theta} = \sum_{i=1}^n \theta_i/n$, we have

$$\mathbb{E}(\tilde{\mu}_i - \mu_i)^2 = \text{Var}(\tilde{\mu}_i) + (\bar{\mu} - \mu_i)^2 = \frac{1 + \alpha^2(\theta_i - \bar{\theta})^2}{n} \leq \frac{5}{n}.$$

Therefore,

$$\mathbb{E}\|\tilde{\mu} - \mu\|^2 \leq 5,$$

which proves one part of the proposition.

Next, let $\mu = (0, 0, \dots, 0)$. Take any $t \geq 0$ and any $v \in K$ such that $\|v - \mu\| \leq t$. Suppose that

$$v_i := \alpha n^{-1/4} + \alpha \theta_i n^{-1/2},$$

where $\alpha \in [0, 1]$ and $\theta_i \in [-1, 1]$. Note that

$$\|v - \mu\|^2 \geq \alpha^2 \sqrt{n} (1 - n^{-1/4})^2.$$

Therefore, $\alpha \leq tn^{-1/4}/(1 - n^{-1/4})$. Since

$$Z \cdot (v - \mu) = \alpha n^{-1/4} \sum_{i=1}^n Z_i + \alpha n^{-1/2} \sum_{i=1}^n Z_i \theta_i,$$

this proves that

$$\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - \mu) \leq \frac{tn^{-1/2}}{1 - n^{-1/4}} \left| \sum_{i=1}^n Z_i \right| + \frac{tn^{-3/4}}{1 - n^{-1/4}} \sum_{i=1}^n |Z_i|.$$

Consequently,

$$(35) \quad f_\mu(t) \leq C_1 t n^{1/4} - \frac{t^2}{2}.$$

On the other hand, if $t \leq n^{1/4}$, then taking $\theta_i = \text{sign}(Z_i)$ and $\alpha = tn^{-1/4}/2$, we get $\|v - \mu\| \leq t$ and

$$Z \cdot (v - \mu) = \frac{tn^{-1/2}}{2} \sum_{i=1}^n Z_i + \frac{tn^{-3/4}}{2} \sum_{i=1}^n |Z_i|,$$

proving that

$$(36) \quad f_\mu(t) \geq C_2 t n^{1/4} - \frac{t^2}{2}.$$

Without loss of generality, assume that $C_2 < 1 < C_1$. Let

$$r_1 := \frac{C_2^2 n^{1/4}}{4C_1}$$

and $r_2 := C_2 n^{1/4}$. Then by (35),

$$f_\mu(r_1) \leq \frac{C_2^2 n^{1/2}}{4}.$$

On the other hand, since $r_2 \leq n^{1/4}$, therefore, by (36),

$$f_\mu(r_2) \geq \frac{C_2^2 n^{1/2}}{2}.$$

Since $r_1 < r_2$, Proposition 1.3 shows that $t_\mu \geq r_1$. \square

Finally, we give the proof of Proposition 1.6, which is an easy corollary of Theorem 1.4.

PROOF OF PROPOSITION 1.6. By Theorem 1.4, there exists $\mu_0 \in K$ such that $R_2(\mu_0) \geq C R_1(\mu_0)$, where C is a universal constant. Therefore,

$$\sup_{\mu \in K} R_2(\mu) \geq R_2(\mu_0) \geq C R_1(\mu_0) \geq C \inf_{\mu \in K} R_1(\mu) = CL \sup_{\mu \in K} R_1(\mu).$$

This completes the proof of the proposition. \square

We now turn to the proofs of the theorems from Section 2. The first goal is to prove Theorem 2.1. Let us begin with some basic facts about Gaussian random variables.

LEMMA 4.9 (Gaussian tails). *Let V be a standard Gaussian random variable. Then for any $x > 0$,*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{2e^{-x^2/2}}{\sqrt{2\pi}} \leq \mathbb{P}(|V| > x) \leq \frac{2e^{-x^2/2}}{x\sqrt{2\pi}},$$

$$\mathbb{E}(|V|; |V| > x) = \frac{2e^{-x^2/2}}{\sqrt{2\pi}} \quad \text{and}$$

$$\mathbb{E}(V^2; |V| > x) \leq \frac{2(x^2 + 1)e^{-x^2/2}}{x\sqrt{2\pi}}.$$

PROOF. The upper bound in the first inequality is well known as the Mills ratio upper bound for the Gaussian tail. To prove this, just note that

$$\mathbb{P}(|V| > x) = 2 \int_x^\infty \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \leq 2 \int_x^\infty \frac{ye^{-y^2/2}}{x\sqrt{2\pi}} dy = \frac{2e^{-x^2/2}}{x\sqrt{2\pi}}.$$

For the lower bound, we apply integration by parts two times to get

$$\int_x^\infty e^{-y^2/2} dy = \left(\frac{1}{x} - \frac{1}{x^3}\right) e^{-x^2/2} + \int_x^\infty \frac{3e^{-y^2/2}}{y^4} dy.$$

For the second assertion, note that

$$\mathbb{E}(|V|; |V| > x) = 2 \int_x^\infty \frac{ye^{-y^2/2}}{\sqrt{2\pi}} dy = \frac{2e^{-x^2/2}}{\sqrt{2\pi}}.$$

Finally, for the third claim, note that

$$\mathbb{E}(V^2; |V| > x) = 2 \int_x^\infty \frac{y^2 e^{-y^2/2}}{\sqrt{2\pi}} = \frac{2xe^{-x^2/2}}{\sqrt{2\pi}} + 2 \int_x^\infty \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$$

and apply the first inequality to bound the second term on the right-hand side. \square

LEMMA 4.10 (Size of Gaussian maxima). *Let V_1, \dots, V_n be standard Gaussian random variables, not necessarily independent. Then*

$$\mathbb{E}\left(\max_{1 \leq i \leq n} |V_i|\right) \leq \sqrt{2 \log(2n)}.$$

PROOF. Take any $\beta > 0$. Then by Jensen’s inequality,

$$\begin{aligned} \mathbb{E}\left(\max_{1 \leq i \leq n} |V_i|\right) &= \frac{1}{\beta} \mathbb{E}(\log e^{\beta \max_{1 \leq i \leq n} |V_i|}) \\ &\leq \frac{1}{\beta} \mathbb{E}\left(\log \sum_{i=1}^n e^{\beta |V_i|}\right) \leq \frac{1}{\beta} \log \sum_{i=1}^n \mathbb{E}(e^{\beta |V_i|}) \\ &\leq \frac{1}{\beta} \log \sum_{i=1}^n (\mathbb{E}(e^{\beta V_i}) + \mathbb{E}(e^{-\beta V_i})) = \frac{\log(2n)}{\beta} + \frac{\beta}{2}. \end{aligned}$$

The proof is completed by taking $\beta = \sqrt{2 \log(2n)}$. \square

For any n and r , let $C^r(\mathbb{R}^n)$ be the set of r -times continuously differentiable functions from \mathbb{R}^n into \mathbb{R} , and let $C_b^r(\mathbb{R}^n)$ be the set of all $g \in C^r(\mathbb{R}^n)$ such that g and all its derivatives up to order r are bounded. For any $g \in C_b^1(\mathbb{R})$, let Ug be the solution to the differential equation

$$f'(x) - xf(x) = g(x) - \mathbb{E}(g(V)),$$

where $V \sim N(0, 1)$. Explicitly, we have

$$Ug(x) = e^{x^2/2} \int_{-\infty}^x e^{-u^2/2} (g(u) - \mathbb{E}(g(V))) du.$$

It is not difficult to prove that Ug maps $C_b^1(\mathbb{R})$ into $C_b^2(\mathbb{R})$. The following lemma is well known, and follows directly from integration by parts:

LEMMA 4.11. *Let $V = (V_1, \dots, V_n)$ be a Gaussian random vector with zero mean and arbitrary covariance matrix. Then for any $g \in C_b^1(\mathbb{R}^n)$ and any i , we have*

$$\mathbb{E}(V_i g(V)) = \sum_{j=1}^n \mathbb{E}(V_i V_j) \mathbb{E}\left(\frac{\partial g}{\partial x_j}(V)\right).$$

Using this, we easily get the following lemma.

LEMMA 4.12. *Take any $g_1, g_2 \in C_b^2(\mathbb{R})$, and let $f_1 = Ug_1, f_2 = Ug_2$. Suppose V_1 and V_2 are jointly Gaussian random variables with $\mathbb{E}(V_1) = \mathbb{E}(V_2) = 0, \mathbb{E}(V_1^2) = \mathbb{E}(V_2^2) = 1$ and $\mathbb{E}(V_1V_2) = \rho$. Then*

$$\text{Cov}(g_1(V_1), g_2(V_2)) = \rho\mathbb{E}(f_1(V_1)f_2(V_2)) + \rho^2\mathbb{E}(f_1'(V_1)f_2'(V_2)).$$

PROOF. Using Lemma 4.11 in two steps, we have

$$\begin{aligned} \text{Cov}(g_1(V_1), g_2(V_2)) &= \mathbb{E}((f_1'(V_1) - V_1f_1(V_1))(f_2'(V_2) - V_2f_2(V_2))) \\ &= -\rho\mathbb{E}(f_1(V_1)(f_2''(V_2) - f_2(V_2) - V_2f_2'(V_2))) \\ &= -\rho\mathbb{E}(f_1(V_1)(f_2''(V_2) - f_2(V_2))) \\ &\quad + \rho\mathbb{E}(f_1(V_1)f_2''(V_2)) + \rho^2\mathbb{E}(f_1'(V_1)f_2'(V_2)) \\ &= \rho\mathbb{E}(f_1(V_1)f_2(V_2)) + \rho^2\mathbb{E}(f_1'(V_1)f_2'(V_2)). \end{aligned}$$

This completes the proof of the lemma. \square

Using Lemma 4.12, we now prove the following set of inequalities for additive functions of Gaussian random variables. This is probably a new result.

LEMMA 4.13. *Let $V = (V_1, \dots, V_n)$ be a Gaussian random vector with mean zero and covariance matrix Σ . Let λ_{\max} and λ_{\min} be the largest and smallest eigenvalues of Σ . Assume that $\mathbb{E}(V_i^2) = 1$ for each i . Let g_1, \dots, g_n be functions such that $\mathbb{E}(g_i(V_i)^2) < \infty$ for each i . Then*

$$\lambda_{\min} \sum_{i=1}^n \text{Var}(g_i(V_i)) \leq \text{Var}\left(\sum_{i=1}^n g_i(V_i)\right) \leq \lambda_{\max} \sum_{i=1}^n \text{Var}(g_i(V_i)).$$

PROOF. First, let us make some reductions. Recall that we have assumed that $\mathbb{E}(V_i^2) = 1$ for each i . Next, note that if g is a function such that $\mathbb{E}(g(Z)^2) < \infty$, where $Z \sim N(0, 1)$, then there is a sequence of step functions $\{g_n\}$ such that $g_n(Z)$ converges to $g(Z)$ in L^2 . Again, if g is a step function, then there is a sequence $\{g_n\}$ of C_b^1 functions such that $g_n(Z)$ converges to $g(Z)$ in L^2 . Hence, assume without loss of generality that g_i 's are elements of $C_b^1(\mathbb{R})$.

Now let $f_i := Ug_i$ and $\sigma_{ij} := \mathbb{E}(V_iV_j)$. Let (Y_1, \dots, Y_n) be an independent copy of (V_1, \dots, V_n) . Then by Lemma 4.12, we have

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n g_i(V_i)\right) &= \sum_{i,j} (\sigma_{ij}\mathbb{E}(f_i(V_i)f_j(V_j)) + \sigma_{ij}^2\mathbb{E}(f_i'(V_i)f_j'(V_j))) \\ &= \mathbb{E}\left(\sum_{i,j} \sigma_{ij}(f_i(V_i)f_j(V_j) + Y_i f_i'(V_i)Y_j f_j'(V_j))\right) \end{aligned}$$

$$\begin{aligned} &\leq \lambda_{\max} \mathbb{E} \left(\sum_{i=1}^n (f_i(V_i)^2 + Y_i^2 f_i'(V_i)^2) \right) \\ &= \lambda_{\max} \sum_{i=1}^n \mathbb{E}(f_i(V_i)^2 + f_i'(V_i)^2). \end{aligned}$$

But by Lemma 4.12, $\text{Var}(g_i(V_i)) = \mathbb{E}(f(V_i)^2) + \mathbb{E}(f_i'(V_i)^2)$. This gives the upper bound. The lower bound follows similarly. \square

We need a few more lemmas before proving Theorem 2.1. Let all notation be as in the statement of the theorem. Additionally, let $S := \{i : \beta_i \neq 0\}$, and let $V := n^{-1/2} X^T Z$. Then V is a Gaussian random vector with mean zero and covariance matrix Σ .

LEMMA 4.14. *Suppose that $\delta > 0$. Take any $\alpha > 0$. Then there is a constant c_3 depending only on $\alpha, \delta, a, b, s, r$ and L such that*

$$f_\mu(n^\alpha) \leq c_3 \sqrt{n} (\log n)^{1/4} + c_3 n^\alpha \sqrt{\log n} + 2\delta \sqrt{\alpha n \log n} - \frac{n^{2\alpha}}{2}.$$

PROOF. Throughout this proof, we will use C_1, C_2, \dots to denote constants that may depend only on $\alpha, \delta, a, b, s, r$ and L . Let K'_0 be the set of all $\gamma \in K_0$ such that $\|X\gamma - X\beta\| \leq n^\alpha$. Let

$$M := \sup_{\gamma \in K'_0} Z \cdot (X\gamma - X\beta) = \sqrt{n} \sup_{\gamma \in K'_0} V \cdot (\gamma - \beta).$$

Note that for any $\gamma \in K'_0$,

$$(37) \quad na \|\gamma - \beta\|^2 \leq \|X\gamma - X\beta\|^2 \leq n^{2\alpha}.$$

Next, note that

$$(38) \quad V \cdot (\gamma - \beta) \leq \sum_{i \in S} |V_i| |\gamma_i - \beta_i| + \sum_{i \notin S} |V_i| |\gamma_i|.$$

Now, by (37),

$$\begin{aligned} \sum_{i \in S} |V_i| |\gamma_i - \beta_i| &\leq \left(\sum_{i \in S} V_i^2 \sum_{i \in S} (\gamma_i - \beta_i)^2 \right)^{1/2} \\ &\leq \left(\sum_{i \in S} V_i^2 \right)^{1/2} \|\gamma - \beta\| \\ &\leq \frac{n^\alpha}{\sqrt{na}} \left(\sum_{i \in S} V_i^2 \right)^{1/2}. \end{aligned}$$

Since $\mathbb{E}(V_i^2) = 1$ for each i , this shows that

$$(39) \quad \mathbb{E}\left(\sup_{\gamma \in K'_0} \sum_{i \in S} |V_i| |\gamma_i - \beta_i|\right) \leq n^\alpha \sqrt{\frac{s}{na}}.$$

Define the random set

$$T := \{i \notin S : |V_i| \geq 2\sqrt{\alpha \log n}\}.$$

Then by (37) and the fact that $|\gamma|_1 \leq L$,

$$\begin{aligned} \sum_{i \notin S} |V_i| |\gamma_i| &\leq \sum_{i \in T} |V_i| |\gamma_i| + 2\sqrt{\alpha \log n} \sum_{i \notin S \cup T} |\gamma_i| \\ &\leq \left(\sum_{i \in T} V_i^2\right)^{1/2} \|\gamma - \beta\| + 2\sqrt{\alpha \log n} \left(L - \sum_{i \in S} |\gamma_i|\right) \\ &\leq \left(\sum_{i \in T} V_i^2\right)^{1/2} \frac{n^\alpha}{\sqrt{na}} + 2\sqrt{\alpha \log n} \left(\delta + \sum_{i \in S} |\gamma_i - \beta_i|\right). \end{aligned}$$

Again, by the Cauchy–Schwarz inequality and (37),

$$\sum_{i \in S} |\gamma_i - \beta_i| \leq \sqrt{s} \|\gamma - \beta\| \leq n^\alpha \sqrt{\frac{s}{na}}.$$

From the last two displays, we get

$$\sum_{i \notin S} |V_i| |\gamma_i| \leq \left[\left(\sum_{i \in T} V_i^2\right)^{1/2} + 2\sqrt{s\alpha \log n}\right] \frac{n^\alpha}{\sqrt{na}} + 2\delta\sqrt{\alpha \log n}.$$

Therefore, by Lemma 4.9,

$$\begin{aligned} &\mathbb{E}\left(\sup_{\gamma \in K'_0} \sum_{i \notin S} |V_i| |\gamma_i|\right) \\ &\leq \left[\left(\sum_{i=1}^p \mathbb{E}(V_i^2; |V_i| \geq 2\sqrt{\alpha \log n})\right)^{1/2} + 2\sqrt{s\alpha \log n}\right] \frac{n^\alpha}{\sqrt{na}} + 2\delta\sqrt{\alpha \log n} \\ &\leq [C_1 n^{(1-2\alpha)/2} (\log n)^{1/4} + 2\sqrt{s\alpha \log n}] \frac{n^\alpha}{\sqrt{na}} + 2\delta\sqrt{\alpha \log n} \\ &\leq C_2 (\log n)^{1/4} + C_3 n^\alpha \sqrt{\frac{\log n}{n}} + 2\delta\sqrt{\alpha \log n}. \end{aligned}$$

From the above display, and the inequalities (38) and (39), we get

$$\begin{aligned} f_\mu(n^\alpha) &= \mathbb{E}(M) - \frac{n^{2\alpha}}{2} = \sqrt{n} \mathbb{E}\left(\sup_{\gamma \in K'_0} V \cdot (\gamma - \beta)\right) - \frac{n^{2\alpha}}{2} \\ &\leq C_2 \sqrt{n} (\log n)^{1/4} + C_3 n^\alpha \sqrt{\log n} + 2\delta \sqrt{\alpha n \log n} - \frac{n^{2\alpha}}{2}. \end{aligned}$$

This completes the proof of the lemma. \square

LEMMA 4.15. *Suppose that $\delta > 0$. Take any $0 < \alpha_1 < \alpha_2 < 1/2$. Then there is a constant c_4 depending only on $\alpha_1, \alpha_2, \delta, a, b, s, r$ and L such that if $n > c_4$, then*

$$f_\mu(n^{\alpha_2}) \geq 2\delta\sqrt{\alpha_1 n \log n} - 2n^{\alpha_2} - \frac{n^{2\alpha_2}}{2}.$$

PROOF. Choose some $\alpha \in (\alpha_1, \alpha_2)$. Throughout this proof, we will use C_1, C_2, \dots to denote constants that may depend only on $\alpha, \alpha_1, \alpha_2, \delta, a, b, s, r$ and L .

Let V and T be as in the proof of Lemma 4.14. Let K'_0 and M be as in the proof of Lemma 4.14, with α replaced by α_2 . Let us make the following specific choice of γ :

$$\gamma_i := \begin{cases} \text{sign}(V_i)\delta/|T|, & \text{if } i \in T, \\ \beta_i, & \text{if } i \in S, \\ 0, & \text{in all other cases.} \end{cases}$$

Then note that

$$(40) \quad |\gamma|_1 \leq |\beta|_1 + \delta = L.$$

(The above inequality is an equality if T is nonempty, but we are allowing for the possibility that T may be empty.) Next, note that if T is nonempty, then

$$(41) \quad \|X\beta - X\gamma\| \leq \sqrt{bn}\|\beta - \gamma\| \leq \delta\sqrt{\frac{bn}{|T|}}$$

and

$$(42) \quad V \cdot (\gamma - \beta) = \frac{\delta}{|T|} \sum_{i \in T} |V_i|.$$

By Lemma 4.9,

$$(43) \quad \left(1 - \frac{1}{4\alpha \log n}\right) \frac{(p-s)n^{-2\alpha}}{\sqrt{2\pi\alpha \log n}} \leq \mathbb{E}|T| \leq \frac{pn^{-2\alpha}}{\sqrt{2\pi\alpha \log n}}$$

and

$$(44) \quad \mathbb{E}\left(\sum_{i \in T} |V_i|\right) = \sum_{i \notin S} \mathbb{E}(|V_i|; |V_i| \geq 2\sqrt{\alpha \log n}) = \frac{2(p-s)n^{-2\alpha}}{\sqrt{2\pi}}.$$

On the other hand, by Lemma 4.13,

$$(45) \quad \text{Var}(|T|) \leq b \sum_{i \notin S} \mathbb{P}(|V_i| \geq 2\sqrt{\alpha \log n}) \leq \frac{C_1 pn^{-2\alpha}}{\sqrt{\log n}}$$

and

$$(46) \quad \text{Var}\left(\sum_{i \in T} |V_i|\right) = b \sum_{i \notin S} \mathbb{E}(V_i^2; |V_i| \geq 2\sqrt{\alpha \log n}) \leq C_2 p n^{-2\alpha} \sqrt{\log n}.$$

Let ε' be a positive constant depending only on $\alpha, \alpha_1, \alpha_2, \delta, a, b, s, r$ and L . The value of ε' will be determined later. As a consequence of (41), (43), (44), (45), (46), the fact that $\alpha_2 < 1/2$, and Chebychev's inequality, it follows that there exists C_3 depending only on $\alpha, \alpha_1, \alpha_2, \delta, a, b, s, r$ and L and our choice of ε' , such that if $n > C_3$, then

$$\begin{aligned} \mathbb{P}\left((1 - \varepsilon')\left(1 - \frac{1}{4\alpha \log n}\right) \frac{(p - s)n^{-2\alpha}}{\sqrt{2\pi\alpha \log n}} \leq |T| \leq (1 + \varepsilon') \frac{pn^{-2\alpha}}{\sqrt{2\pi\alpha \log n}} \right. \\ \left. \text{and } \sum_{i \in T} |V_i| \geq (1 - \varepsilon'^2) \frac{2(p - s)n^{-2\alpha}}{\sqrt{2\pi}}\right) \geq \frac{1}{2}. \end{aligned}$$

Note that if C_3 is chosen large enough, and $|T|$ indeed turns out to be bigger than the lower bound on $|T|$ in the above expression, then $|T| \geq \delta^2 b n^{1-2\alpha_2}$ since $\alpha_2 > \alpha$. Thus, under this circumstance (40) and (41) imply that $\gamma \in K'_0$. Combined with (42) and the lower bound on the probability displayed above, this gives

$$\mathbb{P}\left(M \geq \frac{(1 - \varepsilon')(p - s)}{p} 2\delta\sqrt{\alpha n \log n}\right) \geq \frac{1}{2}.$$

By the concentration of Gaussian maxima (Lemma 4.1) and the above inequality, it follows that

$$\mathbb{E}(M) \geq \frac{(1 - \varepsilon')(p - s)}{p} 2\delta\sqrt{\alpha n \log n} - 2n^{\alpha_2}.$$

The proof is now completed by taking ε' small enough and C_3 large enough to satisfy the required inequality. \square

LEMMA 4.16. *Suppose that $\delta = 0$. Then there is a constant c_5 depending only on a, b, s, r and L such that for any $u > 0$,*

$$f_\mu(u\sqrt{\log n}) \leq c_5 u \log n - \frac{u^2 \log n}{2}.$$

PROOF. Throughout this proof, we will use C_1, C_2, \dots to denote constants that may depend only on δ, a, b, s, r and L . Fix $u > 0$ and let K'_0 be the set of all $\gamma \in K_0$ such that $\|X\gamma - X\beta\| \leq u\sqrt{\log n}$. Let M and V be as in the proof of Lemma 4.14. Additionally, let $G := \max_{1 \leq i \leq p} |V_i|$.

Take any $\gamma \in K'_0$. Note that the inequality (38) from the proof of Lemma 4.14 is still valid, and that (37) and (39) are also valid, after replacing n^α with $u\sqrt{\log n}$.

In addition to that, note that by Lemma 4.10,

$$\begin{aligned} \mathbb{E}\left(\sum_{i \notin S} |V_i| |\gamma_i|\right) &\leq \mathbb{E}(G) \sum_{i \notin S} |\gamma_i| \\ &\leq \sqrt{2 \log(2p)} \left(L - \sum_{i \in S} |\gamma_i|\right) \\ &= \sqrt{2 \log(2p)} \sum_{i \in S} (|\beta_i| - |\gamma_i|) \leq \sqrt{2 \log(2p)} \sum_{i \in S} |\beta_i - \gamma_i| \\ &\leq \sqrt{2s \log(2p)} \|\beta - \gamma\| \leq \frac{C_1 u \log n}{\sqrt{n}}. \end{aligned}$$

Combining the above observations, we get

$$\mathbb{E}(M) \leq C_2 u \sqrt{\log n} + C_1 u \log n - \frac{u^2 \log n}{2}.$$

This completes the proof of the lemma. \square

LEMMA 4.17. *Suppose that $\delta < 0$. Then there are positive constants c_6 and c_7 depending only on δ, a, b, s, r and L such that $c_6 \sqrt{n} \leq t_\mu \leq c_7 \sqrt{n}$.*

PROOF. Note that for any $\gamma \in K_0$,

$$\begin{aligned} \|X\gamma - X\beta\|^2 &\geq na \|\gamma - \beta\|^2 \geq na \sum_{i \in S} (\gamma_i - \beta_i)^2 \\ &\geq \frac{na}{s} \left(\sum_{i \in S} |\gamma_i - \beta_i|\right)^2 \geq \frac{na\delta^2}{s}. \end{aligned}$$

This shows that there is a small enough C_1 depending only on δ, a and s such that $f_\mu(t) = -\infty$ if $t \leq C_1 \sqrt{n}$. By Proposition 1.3 and the fact that $f_\mu(t)$ is finite for at least one t (from Theorem 1.1), this implies the lower bound on t_μ .

Next, note that since $0 \in K_0$,

$$\begin{aligned} \|\mu - \hat{\mu}\| &\leq \|\mu - Y\| + \|Y - \hat{\mu}\| \\ &\leq \|\mu - Y\| + \|Y\| \\ &\leq 2\|\mu - Y\| + \|\mu\|. \end{aligned}$$

But $\mathbb{E}\|\mu - Y\|^2 = n$ and

$$\|\mu\| = \|X\beta\| \leq \sqrt{nb} \|\beta\| \leq \sqrt{nb} |\beta|_1 \leq \sqrt{nb} L.$$

Thus, $\mathbb{E}\|\mu - \hat{\mu}\|^2 \leq (8 + 2bL^2)n$. By Corollary 1.2, this shows that $t_\mu \leq C_2 \sqrt{n}$ for some constant C_2 depending only on b and L . This completes the proof of the lemma. \square

We are now ready to prove Theorem 2.1.

PROOF OF THEOREM 2.1. First, suppose that $\delta > 0$. Take any $0 < \alpha < \alpha_1 < \alpha_2 < 1/4$. By Lemmas 4.14 and 4.15, it follows that if n is large enough (depending only on $\alpha, \alpha_1, \alpha_2, \delta, a, b, s, r$ and L), then $f_\mu(n^\alpha) \leq f_\mu(n^{\alpha_2})$ and, therefore, by Proposition 1.3, $t_\mu \geq n^\alpha$. Next, take any $\alpha > 1/4$. Lemma 4.14 implies that if n is large enough, then $f_\mu(n^\alpha) \leq 0$ and, therefore, by Proposition 1.3, $t \leq n^\alpha$.

If $\delta = 0$, the conclusion follows directly from a combination of Lemma 4.16 and Proposition 1.3. If $\delta < 0$, simply invoke Lemma 4.17. \square

Our final task is to prove Theorem 2.2. As before, we need some standard results and notations from the literature.

If \mathcal{F} is a subset of a normed space with norm $\| \cdot \|$ and ε is a positive real number, the covering number $N(\varepsilon, \mathcal{F}, \| \cdot \|)$ is defined as the minimum number of open balls of radius ε (with respect to the norm $\| \cdot \|$) with centers in \mathcal{F} that are needed to cover \mathcal{F} .

The following result, known as ‘‘Dudley’s entropy bound,’’ connects the covering numbers of \mathcal{F} with the expected maximum of a certain Gaussian process.

LEMMA 4.18 (Dudley’s entropy bound [27]). *Let \mathcal{F} be as above. Suppose that $(X_f)_{f \in \mathcal{F}}$ is a Gaussian process on \mathcal{F} such that $\mathbb{E}(X_f) = 0$ for each $f \in \mathcal{F}$, and $\mathbb{E}(X_f - X_g)^2 = \|f - g\|^2$ for each $f, g \in \mathcal{F}$. Then*

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} X_f\right) \leq C \int_0^{\text{diam}(\mathcal{F})} \sqrt{\log N(\varepsilon, \mathcal{F}, \| \cdot \|)} d\varepsilon,$$

where C is a universal constant.

Suppose now that \mathcal{F} is a set of functions from some set S into \mathbb{R} , and $\| \cdot \|$ is a norm on a vector space of functions containing \mathcal{F} . Suppose that l and u are two elements of \mathcal{F} such that $l \leq u$ everywhere on S . If $\|l - u\| \leq \varepsilon$, then the set of all $f \in \mathcal{F}$ such that $l \leq f \leq u$ everywhere on S is called an ε -bracket, and is denoted by $[l, u]$. The bracketing number $N_{[\cdot]}(\varepsilon, \mathcal{F}, \| \cdot \|)$ is the minimum number of ε -brackets needed to cover \mathcal{F} . It is quite easy to see that

$$(47) \quad N(\varepsilon, \mathcal{F}, \| \cdot \|) \leq N_{[\cdot]}(2\varepsilon, \mathcal{F}, \| \cdot \|).$$

The following result is quoted from van der Vaart and Wellner [72], Theorem 2.7.5, page 159.

LEMMA 4.19 (van der Vaart and Wellner [72]). *Let P be any probability measure on \mathbb{R} and let $\| \cdot \|_r$ denote the $L^r(P)$ norm. Let \mathcal{F} be the set of all monotone functions from \mathbb{R} into $[0, 1]$. Then for any $\varepsilon > 0$,*

$$\log N_{[\cdot]}(\varepsilon, \mathcal{F}, \| \cdot \|_r) \leq C\varepsilon^{-1},$$

where C is a constant that depends on r only.

The statement of Lemma 4.19 has to be modified in a certain way to suit our purpose in the proof of Theorem 2.2. The following lemma gives the modified statement.

LEMMA 4.20. *Take any two real numbers $a < b$, and a positive integer n . Let Q denote the set of all vectors $\mu \in \mathbb{R}^n$ such that*

$$a \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_n \leq b.$$

Let $\|\cdot\|$ denote the Euclidean norm on Q . Then for any $t > 0$,

$$\log N(t, Q, \|\cdot\|) \leq \frac{C\sqrt{n}(b-a)}{t},$$

where C is a universal constant.

PROOF. First, assume that $a = 0$ and $b = 1$. Let

$$\varepsilon := \frac{t}{2\sqrt{n}}.$$

Let P be the uniform probability distribution on $[0, 1]$, and let $\|\cdot\|_{L^2(P)}$ denote the L^2 norm induced by P . Let \mathcal{F} be the set of all monotone functions from \mathbb{R} into $[0, 1]$. Let \mathcal{G} be a finite subset of \mathcal{F} such that for any $f \in \mathcal{F}$ there exists $g \in \mathcal{G}$ such that $\|f - g\|_{L^2(P)} \leq \varepsilon$. By Lemma 4.19 and the inequality (47), \mathcal{G} can be chosen such that $\log |\mathcal{G}| \leq C\varepsilon^{-1}$, where C is a universal constant.

Now take any $\mu \in Q$. Define a function $f^\mu : \mathbb{R} \rightarrow [0, 1]$ as

$$f^\mu(x) = \begin{cases} 0, & \text{if } x < 0, \\ \mu_i, & \text{if } (i-1)/n \leq x < i/n, \\ 1, & \text{if } x \geq 1. \end{cases}$$

Then clearly $f^\mu \in \mathcal{F}$. For each $g \in \mathcal{G}$, inspect whether there exists some $\mu \in Q$ such that $\|f^\mu - g\|_{L^2(P)} < \varepsilon$. If there exists such a μ , choose one according to some pre-specified rule and call it $\mu(g)$. Let Q' be the subset of Q consisting of all such $\mu(g)$. Then clearly $|Q'| \leq |\mathcal{G}|$. On the other hand, for any $\mu \in Q$, there exists $g \in \mathcal{G}$ such that $\|f^\mu - g\|_{L^2(P)} < \varepsilon$. Consequently,

$$\|f^\mu - f^{\mu(g)}\|_{L^2(P)} < 2\varepsilon.$$

But

$$\|f^\mu - f^{\mu(g)}\|_{L^2(P)}^2 = \int_0^1 (f^\mu(x) - f^{\mu(g)}(x))^2 dx = \frac{1}{n} \sum_{i=1}^n (\mu_i - \mu_i(g))^2.$$

Thus, $\|\mu - \mu(g)\| = \sqrt{n}\|f^\mu - f^{\mu(g)}\|_{L^2(P)} < 2\sqrt{n}\varepsilon = t$. This completes the proof of the lemma when $a = 0$ and $b = 1$.

For general a and b , let l be the unique linear map that takes a to 0 and b to 1. Let $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the map that applies l to each coordinate. Given $t > 0$, we now

know that there exists a set $Q_0 \subseteq L(Q)$ of size $\leq C\sqrt{n}(b-a)/t$ such that for any $\mu \in Q$, there exists $\nu \in Q_0$ satisfying

$$\|L(\mu) - \nu\| \leq \frac{t}{b-a}.$$

To complete the proof, note that $L^{-1}(Q_0) \subseteq Q$, and $\|\mu - L^{-1}(\nu)\| \leq t$. \square

We are now ready to prove Theorem 2.2.

PROOF OF THEOREM 2.2. Fix $\mu \in K$. Let l be a positive integer, to be chosen later. Let K' be the subset of K consisting of all ν such that

$$\nu_1 \geq \mu_1 - 2^l, \nu_n \leq \mu_n + 2^l.$$

Fix $t > 0$. Let

$$K'' := \{\nu \in K' : \|\nu - \mu\| \leq t\}$$

and

$$m := \mathbb{E}\left(\sup_{\nu \in K''} Z \cdot (\nu - \mu)\right).$$

Given any $s > 0$, Lemma 4.20 implies that there exists a set $A \subseteq K'$ of size $\leq \exp(C_0 2^l D \sqrt{n}/s)$ such that for any $\nu \in K'$ there exists $\gamma \in A$ satisfying $\|\nu - \gamma\| < s$. Combined with Dudley's entropy bound (Lemma 4.18), this gives

$$(48) \quad m \leq C_1 \sqrt{2^l D n}^{1/4} \int_0^t \frac{ds}{\sqrt{s}} = 2C_1 \sqrt{2^l D t n}^{1/4}.$$

Now take any $\nu \in K$ such that $\|\nu - \mu\| \leq t$. For any $L > 0$,

$$|\{i : |v_i - \mu_i| > L\}| \leq \frac{1}{L^2} \sum_{i=1}^n (v_i - \mu_i)^2 \leq \frac{t^2}{L^2}.$$

Consequently, if $r(L)$ is the largest i such that $|v_i - \mu_i| \leq L$, then

$$r(L) \geq n - \frac{t^2}{L^2}.$$

Similarly, if $s(L)$ is the smallest i such that $|v_i - \mu_i| \leq L$, then

$$s(L) \leq 1 + \frac{t^2}{L^2}.$$

Define ν' as

$$\nu'_i := \begin{cases} \mu_i + 2^l, & \text{if } i > r(2^l), \\ \mu_i - 2^l, & \text{if } i < s(2^l), \\ \nu_i, & \text{if } s(2^l) \leq i \leq r(2^l). \end{cases}$$

Since $v_{r(2^l)} \leq \mu_{r(2^l)} + 2^l$ and $v_{s(2^l)} \geq \mu_{s(2^l)} - 2^l$, we see that $v' \in K$. Again by definition it is clear that $v'_n \leq \mu_n + 2^l$ and $v'_1 \geq \mu_1 - 2^l$. Therefore, $v' \in K'$. Finally, note that for any i , $|\mu_i - v'_i| \leq |\mu_i - v_i|$, implying that $v' \in K''$. Thus,

$$(49) \quad Z \cdot (v' - \mu) \leq \sup_{\gamma \in K''} Z \cdot (\gamma - \mu).$$

Next, note that

$$\begin{aligned} Z \cdot (v - v') &\leq \sum_{i > r(2^l)} |Z_i| |v_i - v'_i| + \sum_{i < s(2^l)} |Z_i| |v_i - v'_i| \\ &\leq \sum_{k=l}^{\infty} \sum_{r(2^k) < i \leq r(2^{k+1})} |Z_i| |v_i - v'_i| + \sum_{k=l}^{\infty} \sum_{s(2^{k+1}) \leq i < s(2^k)} |Z_i| |v_i - v'_i| \\ &\leq \sum_{k=l}^{\infty} \sum_{r(2^k) < i \leq r(2^{k+1})} |Z_i| 2^{k+2} + \sum_{k=l}^{\infty} \sum_{s(2^{k+1}) \leq i < s(2^k)} |Z_i| 2^{k+2} \\ &\leq \sum_{k=l}^{\infty} \sum_{i > n-t^2/2^{2k}} |Z_i| 2^{k+2} + \sum_{k=l}^{\infty} \sum_{i < 1+t^2/2^{2k}} |Z_i| 2^{k+2}. \end{aligned}$$

This shows that

$$(50) \quad \mathbb{E} \left(\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - v') \right) \leq \sum_{k=l}^{\infty} \frac{C_2 t^2}{2^k} = \frac{C_2 t^2}{2^{l-1}}.$$

Combining (48), (49) and (50) gives

$$\begin{aligned} \mathbb{E} \left(\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - \mu) \right) &\leq \mathbb{E} \left(\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v' - \mu) \right) \\ &\quad + \mathbb{E} \left(\sup_{v \in K : \|v - \mu\| \leq t} Z \cdot (v - v') \right) \\ &\leq \mathbb{E} \left(\sup_{\gamma \in K''} Z \cdot (\gamma - \mu) \right) + \frac{C_2 t^2}{2^{l-1}} \\ &\leq 2C_1 \sqrt{2^l D t n}^{1/4} + \frac{C_2 t^2}{2^{l-1}}. \end{aligned}$$

Now choose l so large that $C_2 2^{-(l-1)} \leq 1/4$. With this choice of l , the above inequality implies that

$$(51) \quad f_{\mu}(t) \leq C_3 \sqrt{D t n}^{1/4} - \frac{t^2}{4}.$$

In particular, $f_{\mu}(r) \leq 0$, where $r = (4C_3 \sqrt{D n}^{1/4})^{2/3}$. By Proposition 1.3, this implies that $t_{\mu} \leq r$. This completes the proof of the upper bound for t_{μ} in the statement of the theorem.

Next, fix $t \in [Bn^{-1/2}, \sqrt{n}]$. Let $k := \lceil t\sqrt{n}/B \rceil$ and $m := \lfloor n/k \rfloor$. For $j = 1, 2, \dots, m$, let

$$S_j := \sum_{(j-1)k < i \leq jk} Z_i, \quad a_j := \mu_{(j-1)k+1}, \quad b_j := \mu_{jk}$$

and if $mk < n$, let

$$S_{m+1} := \sum_{mk < i \leq n} Z_i, \quad a_{m+1} := \mu_{mk+1}, \quad b_{m+1} := \mu_n.$$

For each i , let

$$v_i := \frac{a_j + b_j}{2} \quad \text{if } (j-1)k < i \leq jk.$$

Additionally, define

$$\gamma_i := \begin{cases} a_j, & \text{if } (j-1)k < i \leq jk \text{ and } S_j < 0, \\ b_j, & \text{if } (j-1)k < i \leq jk \text{ and } S_j > 0. \end{cases}$$

Notice that for each i ,

$$|\gamma_i - \mu_i| \leq \frac{Bk}{n} \leq \frac{t}{\sqrt{n}}.$$

Consequently,

$$\|\gamma - \mu\| \leq t.$$

Moreover, $\gamma \in K$. Next, note that

$$Z \cdot (\gamma - v) = \frac{1}{2} \sum_{j=1}^{m+1} |S_j| (b_j - a_j) \geq \frac{Ak}{2n} \sum_{j=1}^m |S_j|.$$

Therefore,

$$\begin{aligned} \mathbb{E} \left(\sup_{\theta \in K : \|\theta - \mu\| \leq t} Z \cdot (\theta - \mu) \right) &= \mathbb{E} \left(\sup_{\theta \in K : \|\theta - \mu\| \leq t} Z \cdot (\theta - v) \right) \\ &\geq \mathbb{E}(Z \cdot (\gamma - v)) \\ &\geq \frac{Ak}{2n} \sum_{j=1}^m \mathbb{E}|S_j| \geq \frac{C_4 Ak m \sqrt{k}}{n} \\ &\geq C_5 A \sqrt{k} \geq C_5 AB^{-1/2} t^{1/2} n^{1/4}. \end{aligned}$$

Thus,

$$(52) \quad f_\mu(t) \geq C_5 AB^{-1/2} t^{1/2} n^{1/4} - \frac{t^2}{2}.$$

Let α and β be two positive constants, to be chosen later. Let

$$r_1 := \alpha A^{8/3} B^{-4/3} D^{-1} n^{1/6}, \quad r_2 := \beta A^{2/3} B^{-1/3} n^{1/6}.$$

Then by (52),

$$f_\mu(r_2) \geq (C_5 \sqrt{\beta} - \beta^2/2) A^{4/3} B^{-2/3} n^{1/3},$$

and by (51),

$$f_\mu(r_1) \leq C_3 \alpha^{1/2} A^{4/3} B^{-2/3} n^{1/3}.$$

Suppose that $A > 0$. Choosing β sufficiently small, and then choosing α even smaller (depending on β), it is now easy to arrange that $r_1 < r_2$ and $f_\mu(r_1) \leq f_\mu(r_2)$. By Proposition 1.3, this implies that $t_\mu \geq r_1$. If $A = 0$, the lower bound in the statement of the theorem is automatically true. \square

Acknowledgments. The author thanks Bodhisattva Sen for introducing him to this area and many useful discussions, Xi Chen for pointing out a mistake in the proof of Lemma 4.15 in an earlier draft, Joel Tropp for pointing out the relevant signal processing literature, and Adityanand Guntuboyina and Sara van de Geer for helpful comments. The author also thanks the anonymous referees and the Associate Editor for several useful suggestions.

REFERENCES

- [1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2013). Living on the edge: A geometric theory of phase transitions in convex optimization. Preprint. Available at [arXiv:1303.6672](https://arxiv.org/abs/1303.6672).
- [2] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Inst. Statist. Math.* **26** 641–647. [MR0073895](https://doi.org/10.1080/00137885508839552)
- [3] BARTLETT, P. L., MENDELSON, S. and NEEMAN, J. (2012). ℓ_1 -regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields* **154** 193–224. [MR2981422](https://doi.org/10.1007/s00440-011-9422-2)
- [4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469](https://doi.org/10.1214/09-AOS729)
- [5] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. [MR0722129](https://doi.org/10.1007/BF0122129)
- [6] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. [MR1240719](https://doi.org/10.1007/BF01240719)
- [7] BORELL, C. (1975). The Brunn–Minkowski inequality in Gauss space. *Invent. Math.* **30** 207–216. [MR0399402](https://doi.org/10.1007/BF01239940)
- [8] BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)* 177–197. Cambridge Univ. Press, London. [MR0277070](https://doi.org/10.1017/CBO9780511524427.010)
- [9] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. [MR2807761](https://doi.org/10.1007/978-3-642-12539-0)
- [10] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. [MR2312149](https://doi.org/10.1214/07-EJS1149)
- [11] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. [MR2382644](https://doi.org/10.1214/07-AN129)

- [12] CANDÈS, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. [MR2243152](#)
- [13] CAROLAN, C. and DYKSTRA, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.* **27** 557–566. [MR1745821](#)
- [14] CATOR, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli* **17** 714–735. [MR2787612](#)
- [15] CHANDRASEKARAN, V. and JORDAN, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci. USA* **110** E1181–E1190. [MR3047651](#)
- [16] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. [MR2989474](#)
- [17] CHATTERJEE, S. (2013). Assumptionless consistency of the lasso. Preprint. Available at [arXiv:1303.5817](#).
- [18] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2013). Improved risk bounds in isotonic regression. Preprint. Available at [arXiv:1311.3765](#).
- [19] CIREL'SON, B. S., IBRAGIMOV, I. A. and SUDAKOV, V. N. (1976). Norms of Gaussian sample functions. In *Proceedings of the Third Japan–USSR Symposium on Probability Theory (Tashkent, 1975)*. *Lecture Notes in Math.* **550** 20–41. Springer, Berlin. [MR0458556](#)
- [20] DONOHO, D. (1991). Gelfand n -widths and the method of least squares. Technical report, Dept. Statistics, Univ. California, Berkeley.
- [21] DONOHO, D. L. (2006). For most large underdetermined systems of equations, the minimal l_1 -norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** 907–934. [MR2222440](#)
- [22] DONOHO, D. L. and ELAD, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA* **100** 2197–2202 (electronic). [MR1963681](#)
- [23] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. [MR2237332](#)
- [24] DONOHO, D. L. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** 2845–2862. [MR1872845](#)
- [25] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. [MR1311089](#)
- [26] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369. [MR1323344](#)
- [27] DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330. [MR0220340](#)
- [28] DUROT, C. (2002). Sharp asymptotics for isotonic regression. *Probab. Theory Related Fields* **122** 222–240. [MR1894068](#)
- [29] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th ed. Cambridge Univ. Press, Cambridge. [MR2722836](#)
- [30] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. [MR2060166](#)
- [31] FOYGEL, R. and MACKEY, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Trans. Inform. Theory* **60** 1223–1247. [MR3164972](#)
- [32] GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. [MR2108039](#)
- [33] GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. [MR0093415](#)
- [34] GROENEBOOM, P. and PYKE, R. (1983). Asymptotic normality of statistics based on the convex minors of empirical distribution functions. *Ann. Probab.* **11** 328–345. [MR0690131](#)
- [35] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)

- [36] HOMRIGHAUSEN, D. and MCDONALD, D. (2013). The lasso, persistence, and cross-validation. *JMLR W&CP* **28** 1031–1039.
- [37] JANKOWSKI, H. K. (2012). Convergence of linear functionals of the Grenander estimator under misspecification. Preprint. Available at [arXiv:1207.6614](https://arxiv.org/abs/1207.6614).
- [38] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. [MR1805787](https://arxiv.org/abs/1805787)
- [39] KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. [MR2555200](https://arxiv.org/abs/2555200)
- [40] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. Amer. Math. Soc., Providence, RI. [MR1849347](https://arxiv.org/abs/1849347)
- [41] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. Springer, Berlin. [MR2319879](https://arxiv.org/abs/2319879)
- [42] MCCOY, M. B. and TROPP, J. A. (2013). The achievable performance of convex demixing. Preprint. Available at [arXiv:1309.7478](https://arxiv.org/abs/1309.7478).
- [43] MCCOY, M. B. and TROPP, J. A. (2014). From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete Comput. Geom.* **51** 926–963. [MR3216671](https://arxiv.org/abs/3216671)
- [44] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](https://arxiv.org/abs/2278363)
- [45] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. [MR2488351](https://arxiv.org/abs/2488351)
- [46] MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. [MR1810920](https://arxiv.org/abs/1810920)
- [47] OYMAK, S. and HASSIBI, B. (2010). New null space results and recovery thresholds for matrix rank minimization. Preprint. Available at [arXiv:1011.6326](https://arxiv.org/abs/1011.6326).
- [48] OYMAK, S. and HASSIBI, B. (2013). Sharp MSE bounds for proximal denoising. Preprint. Available at [arXiv:1305.2714](https://arxiv.org/abs/1305.2714).
- [49] OYMAK, S., THRAMOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized LASSO: A precise analysis. Preprint. Available at [arXiv:1311.0830](https://arxiv.org/abs/1311.0830).
- [50] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York. [MR0762984](https://arxiv.org/abs/0762984)
- [51] PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. [MR0267677](https://arxiv.org/abs/0267677)
- [52] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. [MR2816337](https://arxiv.org/abs/2816337)
- [53] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. [MR0961262](https://arxiv.org/abs/0961262)
- [54] RUDELSON, M. and VERSHYNIN, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.* **61** 1025–1045. [MR2417886](https://arxiv.org/abs/2417886)
- [55] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1** 197–206. Univ. California Press, Berkeley and Los Angeles. [MR0084922](https://arxiv.org/abs/0084922)
- [56] STOJNIC, M. (2009). Various thresholds for ℓ_1 -optimization in compressed sensing. Preprint. Available at [arXiv:0907.3666](https://arxiv.org/abs/0907.3666).
- [57] SUDAKOV, V. N. and CIREL'SON, B. S. (1974). Extremal properties of half-spaces for spherically invariant measures. Problems in the theory of probability distributions, II. *Zap. Nauch. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **41** 14–24, 165. [MR0365680](https://arxiv.org/abs/0365680)
- [58] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](https://arxiv.org/abs/1379242)
- [59] TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 273–282. [MR2815776](https://arxiv.org/abs/2815776)

- [60] TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. [MR2985948](#)
- [61] TROPP, J. A. (2014). Forthcoming article. Private communication.
- [62] TSIREL'SON, B. S. (1982). A geometric approach to maximum likelihood estimation for an infinite-dimensional Gaussian location. I. *Teor. Veroyatn. Primen.* **27** 388–395. [MR0657940](#)
- [63] TSIRELSON, B. S. (1985). A geometric approach to maximum likelihood estimation for an infinite-dimensional Gaussian location. II. *Teor. Veroyatn. Primen.* **30** 772–779. [MR0816291](#)
- [64] TSIRELSON, B. S. (1986). A geometric approach to maximum likelihood estimation for an infinite-dimensional Gaussian location. III. *Teor. Veroyatn. Primen.* **31** 537–549. [MR0866873](#)
- [65] VAN DE GEER, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602. [MR0888427](#)
- [66] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. [MR1056343](#)
- [67] VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. [MR1212164](#)
- [68] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.
- [69] VAN DE GEER, S. and LEDERER, J. (2013). The Lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. Inst. Math. Stat. (IMS) Collect.* **9** 303–316. IMS, Beachwood, OH. [MR3202642](#)
- [70] VAN DE GEER, S. and WEGKAMP, M. (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24** 2513–2523. [MR1425964](#)
- [71] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. [MR2396809](#)
- [72] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, New York. [MR1385671](#)
- [73] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. [MR2729873](#)
- [74] WANG, H. and LENG, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102** 1039–1048. [MR2411663](#)
- [75] WANG, Y. (1996). The L_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods* **25** 281–294. [MR1379445](#)
- [76] ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. [MR1902898](#)
- [77] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](#)
- [78] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#)
- [79] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the “degrees of freedom” of the lasso. *Ann. Statist.* **35** 2173–2192. [MR2363967](#)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL, 390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: souravc@stanford.edu