The Annals of Statistics
2023, Vol. 51, No. 4, 1799–1823
https://doi.org/10.1214/23-AOS2309

UNIVERSALITY OF REGULARIZED REGRESSION ESTIMATORS IN HIGH DIMENSIONS

BY QIYANG HAN[1,a] AND YANDI SHEN[2,b]

[1]*Department of Statistics, Rutgers University,* [a]*qh85@stat.rutgers.edu*
[2]*Department of Statistics, University of Chicago,* [b]*ydshen@uchicago.edu*

The Convex Gaussian Min–Max Theorem (CGMT) has emerged as a prominent theoretical tool for analyzing the precise stochastic behavior of various statistical estimators in the so-called high-dimensional proportional regime, where the sample size and the signal dimension are of the same order. However, a well-recognized limitation of the existing CGMT machinery rests in its stringent requirement on the exact Gaussianity of the design matrix, therefore rendering the obtained precise high-dimensional asymptotics, largely a specific Gaussian theory in various important statistical models.

This paper provides a structural universality framework for a broad class of regularized regression estimators that is particularly compatible with the CGMT machinery. Here, universality means that if a "structure" is satisfied by the regression estimator $\widehat{\mu}_G$ for a standard Gaussian design G, then it will also be satisfied by $\widehat{\mu}_A$ for a general non-Gaussian design A with independent entries. In particular, we show that with a good enough ℓ_∞ bound for the regression estimator $\widehat{\mu}_A$, any "structural property" that can be detected via the CGMT for $\widehat{\mu}_G$ also holds for $\widehat{\mu}_A$ under a general design A with independent entries.

As a proof of concept, we demonstrate our new universality framework in three key examples of regularized regression estimators: the Ridge, Lasso and regularized robust regression estimators, where new universality properties of risk asymptotics and/or distributions of regression estimators and other related quantities are proved. As a major statistical implication of the Lasso universality results, we validate inference procedures using the degrees-of-freedom adjusted debiased Lasso under general design and error distributions. We also provide a counterexample, showing that universality properties for regularized regression estimators do not extend to general isotropic designs.

The proof of our universality results relies on new comparison inequalities for the optimum of a broad class of cost functions and Gordon's max–min (or min–max) costs, over arbitrary structure sets subject to ℓ_∞ constraints. These results may be of independent interest and broader applicability.

1. Introduction.

1.1. *Overview.*

Consider the standard linear model

$$Y = A\mu_0 + \xi, \tag{1.1}$$

where $\mu_0 \in \mathbb{R}^n$ is the signal of interest, $A \in \mathbb{R}^{m\times n}$ is the design matrix, $\xi \in \mathbb{R}^m$ is the error vector and $Y \in \mathbb{R}^m$ stands for the response vector. Here and below, we reserve the notation n for signal dimension and m for sample size. We will be interested in understanding the precise stochastic behavior of a broad class of regularized estimators (of μ_0) taking the following

Received June 2022; revised March 2023.
MSC2020 subject classifications. Primary 60F17; secondary 62E17.
Key words and phrases. Gaussian comparison inequalities, high-dimensional asymptotics, Lasso, Lindeberg's principle, random matrix theory, robust regression, ridge regression, universality.

generic form:

$$\widehat{\mu}_A \in \arg\min_{\mu\in\mathbb{R}^n}\left\{\frac{1}{m}\sum_{i=1}^{m}\psi_0\big((A\mu)_i - Y_i\big) + \mathsf{f}(\mu)\right\}. \tag{1.2}$$

Here, $\psi_0 : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a loss function, and $\mathsf{f} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a structure-promoting regularizer for μ_0.

As a canonical example of the regularized regression estimators in (1.2), the Lasso estimator $\widehat{\mu}_A^{\mathsf{L}}$ (cf. [70]) can be realized by taking $\psi_0(x) = x^2/2$ and $\mathsf{f}(\mu) = \lambda\|\mu\|_1/m$ with a tuning parameter $\lambda > 0$. A notable recent line of Lasso theory attempts to characterize its exact behavior under certain specific settings, in an "average" sense to be specified below. This line (i) postulates an exact distributional assumption on the design matrix, where

$$A \text{ is a standard Gaussian design } G \text{ with i.i.d. } \mathcal{N}(0, 1/m) \text{ entries} \tag{1.3}$$

and (ii) works in the so-called "proportional regime," where

$$\text{the sample size } m \text{ is proportional to the signal dimension } n. \tag{1.4}$$

In particular, [54] showed that under (1.3)–(1.4), among with other conditions, with the (Gaussian) error ξ possessing a noise level $\sigma > 0$ and a tuning parameter λ, there exist some $\sigma_*, \lambda_* > 0$ such that the distribution of the Lasso estimator $\widehat{\mu}_G^{\mathsf{L}}$ can be identified as $\eta_1(\mu_0 + \sigma_* Z_n; \lambda_*)$ in the following sense:[1] for any 1-Lipschitz function $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$, it holds with high probability that

$$\mathsf{g}\big(\widehat{\mu}_G^{\mathsf{L}}/\sqrt{n}\big) \approx \mathbb{E}\,\mathsf{g}\big(\eta_1(\mu_0 + \sigma_* Z_n; \lambda_*)/\sqrt{n}\big). \tag{1.5}$$

Here, η_1 is the soft-thresholding function (formally defined in (1.9)), and $Z_n \overset{d}{=} \mathcal{N}(0, I_n)$ is a standard Gaussian vector in $\mathbb{R}^n$. Typically, (1.5) is most informative when the test function g takes certain average over the n coordinates (e.g., ℓ_2 norm).

The method of proof for (1.5) in [54] is based on a two-sided version of Gordon's Gaussian min–max theorem (cf. [32]) now known as the Convex Gaussian Min–Max Theorem (CGMT) (cf. [64, 68]); see Theorem A.3 for a formal statement. The CGMT approach is a flexible theoretical framework that reduces a given, complicated "primal min–max optimization problem" involving a standard Gaussian design matrix, to a much simpler "Gordon's min–max optimization problem" involving Gaussian vectors only. For the Lasso estimator, the CGMT machinery executed by [12, 54] substantially improves a weaker version of (1.5) obtained in [5],[2] by providing precise nonasymptotic descriptions of (1.5) and the distributions of other quantities associated with the Lasso. These results are not only theoretically interesting in their own rights as they also provide an important foundation for statistical inference using the Lasso estimator in the proportional regime (1.4).

The flexible and principled nature of the CGMT method has led to systematic progress in understanding the precise risk/distributional behavior of canonical statistical estimators across a wide array of important statistical models; see, for example, [12, 17, 33, 35, 38, 48, 51, 57, 61, 63, 68, 69, 73, 76, 77] for some samples. The power of the CGMT method is further demonstrated in some of the above cited works that deal with either general correlated Gaussian designs (cf. [12, 48, 51, 57]) or the "maximal" problem aspect ratio beyond the proportional regime (1.4); cf. [33].

[1] Precisely, formulation (1.5) is taken from [12]; however, the proofs in [54] also lead to (1.5) with appropriate modifications.

[2] The proof of a weaker version of (1.5) in [5] is based on the so-called state evolution analysis (cf. [4]) of an approximate message passing (AMP) algorithm for Lasso that also relies crucially on the exact Gaussianity of the design matrix as in (1.3).

Unfortunately, while being a powerful theoretical tool, the CGMT machinery relies on the Gaussianity of the design in an essential way via the use of Gaussian comparison inequalities and, therefore, precise high-dimensional asymptotics results derived from the CGMT remain largely a specific Gaussian theory.

The main goal of this paper is to provide a general universality framework for "structural properties" of regularized regression estimators (1.2) that is compatible with the CGMT machinery. Here, universality means that if a "structure" is satisfied by $\widehat{\mu}_G$ for a standard Gaussian design G, then it will also be satisfied by $\widehat{\mu}_A$ for a general non-Gaussian design A with independent entries. A more concrete example for the prescribed structural universality is to establish the validity of the distribution (1.5) of the Lasso estimator for general non-Gaussian designs.

As already hinted above, a major theoretical advantage of our universality framework lies in its compatibility with the CGMT method. Roughly speaking, we show that, with a good enough ℓ_∞ bound for $\widehat{\mu}_A$ in (1.2), any structural property that can be detected via the CGMT for $\widehat{\mu}_G$ also holds for $\widehat{\mu}_A$ under a general design A with independent entries. Due to the widespread use of the CGMT approach as mentioned above, we expect our universality framework to be of much broader applicability beyond the examples worked out in the current paper.

1.2. *Structural universality framework.* In the sequel, we will work with $\widehat{w}_A \equiv \widehat{\mu}_A - \mu_0$ instead of $\widehat{\mu}_A$ for consistent presentation with the main results in Sections 2 and 3. Clearly,

$$\widehat{w}_A \in \operatorname*{arg\,min}_{w\in\mathbb{R}^n} H_{\psi_0,\mathsf{f}}(w, A, \xi) \equiv \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ \frac{1}{m}\sum_{i=1}^{m} \psi_0\big((Aw)_i - \xi_i\big) + \mathsf{f}(\mu_0 + w) \right\}. \tag{1.6}$$

Now we may formulate the universality problem precisely.

QUESTION 1. *Take any "structural property" $\mathcal{T}_n \subset \mathbb{R}^n$ such that $\mathbb{P}(\widehat{w}_G \in \mathcal{T}_n) \approx 1$. Then is it true that $\mathbb{P}(\widehat{w}_A \in \mathcal{T}_n) \approx 1$ when the design matrix A has independent entries with matching first two moments as those of G?*

Our main abstract universality framework for the regularized regression estimator $\widehat{w}_A$, Theorem 3.1, answers the above question in the affirmative in the proportional regime (1.4), provided the following hold:

(U0) The entries of A and ξ have "enough" moments, the loss function ψ_0 is "self-similar" and the regularizer f possesses "enough" continuity.

(U1) With high probability $\|\widehat{w}_A\|_\infty \vee \|\widehat{w}_G\|_\infty \le L_n$ for some $L_n > 0$ that grows mildly, say, $L_n = n^\varepsilon$ for small enough $\varepsilon > 0$.

(U2) $\mathbb{P}(\widehat{w}_G \in \mathcal{T}_n) \approx 1$ holds at the level of the cost function $H_{\psi_0,\mathsf{f}}$: for some nonrandom $z > 0$ and small $\rho_0 > 0$, with high probability

$$\min_{w\in\mathcal{T}_n^c} H_{\psi_0,\mathsf{f}}(w, G, \xi) \ge z + 2\rho_0 > z + \rho_0 \ge \min_{w\in\mathbb{R}^n} H_{\psi_0,\mathsf{f}}(w, G, \xi).$$

Here, (U0) should be viewed as regularity conditions. In particular, Theorem 3.1 is established for the square loss case and the (possibly nondifferentiable) robust loss case, but as will be clear below, other loss functions ψ_0 whose derivatives are "similar" to itself would also work. Furthermore, the precise number of moments needed for A and ξ depends on the choice of the loss function ψ_0, and the moduli of continuity needed for f is almost minimal. Consequently, the essential conditions to apply the machinery of Theorem 3.1 are (U1) and (U2):

- (U1) requires ℓ_∞ bounds for the regression estimator under both a standard Gaussian design G and the targeted design A. Heuristically, a near constant order ℓ_∞ bound is not too surprising when the design A is nearly isotropic. For instance, if the risk stabilizes in the sense $\|\widehat{w}_A\| = \mathcal{O}_{\mathbf{P}}(\sqrt{n})$ (which must hold true if any risk characterization is desired), due to comparable magnitudes of A's columns, it is natural to expect $\widehat{w}_A$ to be "delocalized" in that each coordinate fluctuates roughly on the same order, that is, $|(\widehat{w}_A)_j| = \mathcal{O}_{\mathbf{P}}(1)$.

  To formally verify (U1), a particular useful general method we adopt in this paper is to study perturbations of $\widehat{w}_A$ by its column and row leave-one-out versions (cf. [23, 24]). In essence, these leave-one-out perturbations are both close enough to $\widehat{w}_A$ while creating sufficient independence that reduces the difficult coordinatewise controls for $\widehat{w}_A$ on a near constant order, to the much easier problem of ℓ_2 norm controls on the (almost trivial) order $\mathcal{O}(\sqrt{n})$. This leave-one-out method appears most easily implemented in the presence of strong convexity, but otherwise requires additional case specific techniques.
- (U2) requires high probability detection of the structural property $\mathcal{T}_n$ for $\widehat{w}_G$ via the cost function $H_{\psi_0,\mathfrak{f}}$. A particularly appealing feature of (U2) lies in its compatibility with the CGMT approach, as one then only needs to verify for the simpler Gordon's problem a constant order gap ($= \rho_0$) between its cost optimum over $\mathcal{T}_n^c$ and the global cost optimum ($= z$).

In summary, for a given structural universality problem of $\widehat{w}_A$, once a good enough ℓ_∞ bound is verified, the problem is almost completely reduced to the standard Gaussian design in which the powerful CGMT can be directly applied.

1.3. *Examples.* As a proof of concept, we apply the aforementioned universality framework to three canonical examples of regularized regression estimators (1.2) in the linear model (1.1), namely:

(E1) the Ridge estimator,
(E2) the Lasso estimator and
(E3) regularized robust regression estimators.

In particular, we prove that the ℓ_∞ bounds required in (U1) hold for all the above three examples (under appropriate moment conditions on A and ξ) and, therefore, universality holds for any structural properties $\mathcal{T}_n$ of these estimators that can be verified under a standard Gaussian design in the sense of (U2).

For the Lasso estimator, our distributional universality of $\widehat{\mu}_A^{\mathsf{L}}$ shows that (1.5) is valid with $\mathsf{g}(\widehat{\mu}_G^{\mathsf{L}}/\sqrt{n})$ replaced by $\mathsf{g}(\widehat{\mu}_A^{\mathsf{L}}/\sqrt{n})$ for any 1-Lipschitz function $\mathsf{g}:\mathbb{R}^n \to \mathbb{R}$. Using the same formulation as (1.5), universality is also confirmed for the distributions of the Lasso residual $\widehat{r}_A^{\mathsf{L}} \equiv Y - A\widehat{\mu}_A^{\mathsf{L}}$ as a scaled convolution of ξ and an extra Gaussian noise of the subgradient $\widehat{v}_A^{\mathsf{L}} \equiv \lambda^{-1}A^\top(Y - A\widehat{\mu}_A^{\mathsf{L}})$ as a random variable taking value in the hypercube $[-1,1]^n$, and of the sparsity $\widehat{s}_A^{\mathsf{L}} \equiv \|\widehat{\mu}_A^{\mathsf{L}}\|_0/n$ as a discrete random variable; see Theorem 3.8 for precise statements. Similar distributional universality properties are proved for the Ridge estimator and its residual; see Theorem 3.4 for details. Using these Lasso universality results, we further verify asymptotic normality of the so-called degrees-of-freedom (dof) adjusted debiased Lasso (cf. [9, 10, 12, 43, 54]) under general design and error distributions; see Theorem 3.9 for details. This universality result validates statistical inference procedures based on dof adjusted debiased Lasso methodologies in the proportional regime (1.4) beyond the exclusive focus on Gaussian designs in previous works (cited above).

It is worth mentioning that using the CGMT machinery (or AMP techniques), the emergence of the Gaussian component in (1.5) for $\widehat{\mu}_A^{\mathsf{L}}$ (or other quantities above) is crucially tied to the Gaussianity of the design matrix. As such, an interesting conceptual consequence of

our universality results is to retrieve—in the challenging proportional regime (1.4)—the "traditional wisdom" that the Gaussianity in $\widehat{\mu}_A^{\mathsf{L}}$ origins from aggregation effects of the errors (or the design entries for some of the other quantities) rather than the specificity of design distributions.

For robust regression estimators, our universality results in Theorems 3.10 and 3.12, although proved using the general purpose universality framework, compare favorably to previous attempts by [23, 24] using problem-specific techniques. In particular, [23, 24] require strong regularity conditions on the loss function that exclude the canonical Huber/absolute losses, along with a strong exponential moment condition on the design. In contrast, our results hold under a wide range of nonsmooth robust loss functions (including the canonical Huber/absolute losses), a much weaker $6+\varepsilon$ moment assumption on the design matrix A, and no moment assumption on the error ξ.

1.4. *Universality of general cost optimum.* The proof of our universality framework relies on comparison inequalities for the optimum of the cost function $w \mapsto H_{\psi_0,\mathsf{f}}(w, A, \xi)$ over a generic structure set $\mathcal{S}_n \subset \mathbb{R}^n$. In particular, we show that for any structure set $\mathcal{S}_n \subset [-L_n, L_n]^n$ with $L_n \geq 1$ growing mildly,

$$\mathbb{E}\,\mathsf{g}\Big(\min_{w\in\mathcal{S}_n} H_{\psi_0,\mathsf{f}}(w, A, \xi)\Big) \approx \mathbb{E}\,\mathsf{g}\Big(\min_{w\in\mathcal{S}_n} H_{\psi_0,\mathsf{f}}(w, B, \xi)\Big) \quad \text{for all } \mathsf{g}\in C^3(\mathbb{R}), \tag{1.7}$$

whenever (i) the design matrices A, B possess independent entries with matching first two moments to the standard Gaussian design G in (1.3), (ii) the loss function ψ_0 grows mildly at ∞ and its derivatives satisfy certain "self-similarity" properties and (iii) f enjoys a certain degree of continuity; see Theorem 2.3 for a formal statement that holds for a more general class of cost functions.

The proof of the comparison inequality (1.7) is based on the quantitative Lindeberg's method (cf. [13]), coupled with an almost dimension-free third derivative bound for every $\mathcal{S}_n$ with the prescribed ℓ_∞ constraint. The ℓ_∞ constraint plays a crucial role in circumventing the undesirable yet unavoidable logarithmic dependence on the "effective size" in the minimum that scales exponentially in n, previously obtained in the high-dimensional central limit theorem literature (see, e.g., [15] for a recent review). These techniques are further generalized to a class of Gordon's max–min (or min–max) cost optimum. Let $X_n(u, w; A) \equiv m^{-1}u^\top Aw + Q_n(u, w)$. We show that for any pair of structure sets $\mathcal{S}_u \subset [-L_u, L_u]^m$ and $\mathcal{S}_w \subset [-L_w, L_w]^n$ with $L_u, L_w \geq 1$ growing mildly,

$$\mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; A)\Big) \approx \mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; B)\Big) \quad \text{for all } \mathsf{g}\in C^3(\mathbb{R}), \tag{1.8}$$

again whenever (i) the design matrices A, B possess independent entries with matching first two moments to the standard Gaussian design G in (1.3) and (ii) Q_n enjoys a certain degree of continuity; see Theorem 2.5 and Corollary 2.6 for formal statements. In the regression examples we study here, we use the comparison inequality (1.8) to derive universality properties beyond the regression estimator itself, but we also expect it to be of broader applicability in view of its intimate resemblance to the "primal optimization problem" in the CGMT machinery (cf. Theorem A.3).

1.5. *Related literature and nonuniversality for general isotropic designs.* A number of universality results are obtained for design matrices consisting of independent entries in the proportional regime (1.4). The work [46] obtained, among other results, asymptotic universality of box-constrained Lasso cost optimum; [56] obtained asymptotic universality for the elastic net; [60] obtained asymptotic universality for certain special test functions applied to

the least squares regression coefficients with strongly convex penalties, along with some results on Lasso, see Section 3.3 for a more detailed comparison. The work [23, 24] obtained universality of precise risk asymptotics and residual distributions in the context of robust regression. On related topics, universality results for various quantities of interest are obtained in [1, 3, 59] for noiseless random linear inverse problems, and in [3, 14, 21, 22, 25, 74] for a class of AMP algorithms. To the best of our knowledge, none of these methods are generally compatible with the CGMT, and nor are applicable for studying universality properties of the broad class of regularized regression estimators (1.2).

Going beyond independent components, universality results are also obtained in several interesting models, under designs (features) whose rows have matching first two moments. For instance, [39] obtained asymptotic universality results concerning training/generalization errors in the random feature model, between non-Gaussian features (nonlinear transforms of underlying Gaussian feature matrix and input vector) and "linearized" Gaussian features, thus verifying a so-called Gaussian equivalence conjecture (cf. [28, 31, 51]). The work [58] obtained further asymptotic universality results for these errors, under an "asymptotic Gaussian" assumption on the feature vectors (see Assumption 6 therein), that apply to other significant models including the two-layer neural tangent model. The work [27] obtained universality for the training loss of ridge regularized generalized linear classification with random labels, under a similar asymptotic Gaussian assumption (see Assumption 4 therein). These results motivate the natural question.

QUESTION 2. *Do the foregoing structural universality properties for regularized regression estimators* (1.2) *proved for entrywise independent designs, extend to isotropic designs or more general row independent designs with matching first two moments?*

In Section 3.5, we answer the above question in the negative, by showing that risk universality for the simple ordinary least squares estimator in the basic linear model (1.1) already fails to hold under an explicitly constructed row independent isotropic design. As will be clear therein, the failure of risk universality is intrinsically due to a well-known fact in random matrix theory—the spectrum of the sample covariance matrix does not exhibit universality for general i.i.d. samples of isotropic random vectors. Simulation results in Section 3.6 further confirm this risk nonuniversality phenomenon for the Ridge and Lasso estimators under the same isotropic design used in the construction of the counterexample.

Universality results of a different nature, for instance under rotational invariance assumptions on the design matrix, are obtained in [29] for a class of regularized least squares problems with convex penalties (depending on the universality target, strong convexity may be required), and in [30] for a broader class of generalized linear estimation problems.

1.6. *Organization.* The rest of the paper is organized as follows. Section 2 presents comparison inequalities for general cost optimum in (1.7) and Gordon's min–max cost optimum in (1.8). As an application of these comparison inequalities, we establish the structural universality framework in Section 3.1. Examples on the Ridge, Lasso and regularized robust regression estimators are detailed in Sections 3.2–3.4. The nonuniversality counterexample is given in Section 3.5. Simulation results that confirm both universality and nonuniversality results are provided in Section 3.6. Most proofs are collected in Sections 4–7 and the appendices in the Supplementary Material [34].

1.7. *Notation.* For any positive integer n, let $[n] = [1 : n]$ denote the set $\{1, \dots, n\}$. For $a, b \in \mathbb{R}$, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. For $a \in \mathbb{R}$, let $a_{\pm} \equiv (\pm a) \vee 0$. For $a > 0$, let $\log_+(a) \equiv 1 \vee \log(a)$. For $x \in \mathbb{R}^n$, let $\|x\|_p$ denote its p-norm $(0 \le p \le \infty)$,

and $B_{n;p}(R) \equiv \{x \in \mathbb{R}^n : \|x\|_p \le R\}$. We simply write $\|x\| \equiv \|x\|_2$ and $B_n(R) \equiv B_{n;2}(R)$. For a matrix $M \in \mathbb{R}^{m\times n}$, let $\|M\|_{\mathrm{op}}$ denote the spectral norm of M. For a measurable map $f : \mathbb{R}^n \to \mathbb{R}$, let $\|f\|_{\mathrm{Lip}} \equiv \sup_{x\neq y}|f(x)-f(y)|/\|x-y\|$. f is called *L-Lipschitz* iff $\|f\|_{\mathrm{Lip}} \le L$.

We use C_x to denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$, respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ ($a \lesssim b$ means $a \le Cb$ for some absolute constant C). For two nonnegative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ (resp., $a_n \gg b_n$) if $\lim_{n\to\infty}(a_n/b_n) = 0$ (resp., $\lim_{n\to\infty}(a_n/b_n) = \infty$). We follow the convention that $0/0 = 0$. $\mathcal{O}$ and $\mathfrak{o}$ (resp., $\mathcal{O}_{\mathbf{P}}$ and $\mathfrak{o}_{\mathbf{P}}$) denote the usual big and small O notation (resp., in probability).

For a proper, closed convex function f defined on $\mathbb{R}$, its *Moreau envelope* $\mathsf{e}_f(\cdot;\tau)$ and *proximal operator* $\mathsf{prox}_f(\cdot;\tau)$ for any $\tau > 0$ are defined by

$$\mathsf{e}_f(x;\tau) \equiv \min_{z\in\mathbb{R}}\left\{\frac{1}{2\tau}(x-z)^2 + f(z)\right\}, \qquad \mathsf{prox}_f(x;\tau) \equiv \arg\min_{z\in\mathbb{R}}\left\{\frac{1}{2\tau}(x-z)^2 + f(z)\right\}.$$

Finally, let for $p > 0$, $z \in \mathbb{R}$, $\lambda \ge 0$,

$$\eta_p(z;\lambda) \equiv \arg\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|z-x\|^2 + \lambda\cdot\frac{\|x\|_p^p}{p}\right\} = \mathsf{prox}_{\|\cdot\|_p^p/p}(z;\lambda). \tag{1.9}$$

We will only use $p = 1, 2$ in this paper.

2. Universality of general cost optimum.

2.1. *Basic setup and assumptions.* Let $A \in \mathbb{R}^{m\times n}$ be a $m \times n$ matrix, $\psi_i : \mathbb{R} \to \mathbb{R}_{\ge 0}$, $\mathsf{f} : \mathbb{R}^n \to \mathbb{R}$ be measurable functions, and

$$H_\psi(w, A) \equiv \frac{1}{m}\sum_{i=1}^m \psi_i\big((Aw)_i\big) + \mathsf{f}(w). \tag{2.1}$$

Let $\bar{H}_\psi \equiv m\cdot H_\psi$ be the unnormalized version of H_ψ. We will be interested in the universality properties related to the optimum and optimizers of H_ψ with respect to the law of the random matrix A (with independent entries).

First, we formalize the precise meaning of the "proportional regime" in (1.4).

ASSUMPTION I (Proportional regime). $\tau \le m/n \le 1/\tau$ holds for some $\tau \in (0, 1)$.

Next, we state the assumptions on the loss functions $\{\psi_i\}$.

ASSUMPTION II (Loss function). There exist reals $q_\ell \ge 0$ ($\ell = 0, 1, 2, 3$), constants $\bar{\rho} \in (0, 1]$, $\{L_{\psi_i} \ge 1 : i \in [m]\}$ and two measurable functions $\mathscr{D}_\psi, \mathscr{M}_\psi : (0, \bar{\rho}) \to \mathbb{R}_{\ge 0}$, $\mathscr{M}_\psi(\rho) \le 1 \le \mathscr{D}_\psi(\rho)$, with the following properties:

1. ψ_i's grow at mostly polynomially in the sense that for $i \in [m]$,
$$\sup_{x\in\mathbb{R}}\frac{|\psi_i(x)|}{1+|x|^{q_0}} \le L_{\psi_i}.$$
2. Smooth approximations $\{\psi_{i;\rho} : \mathbb{R} \to \mathbb{R}_{\ge 0}\}_{\rho\in(0,\bar{\rho})}$ of ψ_i exist so that (i) $\psi_{i;\rho} \in C^3(\mathbb{R})$, (ii) $\max_{i\in[m]}\|\psi_{i;\rho} - \psi_i\|_\infty \le \mathscr{M}_\psi(\rho)$ and (iii) derivatives of $\{\psi_{i;\rho}\}$ satisfy the following self-bounding property:
$$\max_{\ell=1,2,3}\max_{i\in[m]}\sup_{x\in\mathbb{R}}\frac{|\partial^\ell\psi_{i;\rho}(x)|}{1+|\psi_{i;\rho}(x)|^{q_\ell}} \le \mathscr{D}_\psi(\rho).$$

The first requirement (1) says that ψ_i cannot grow too fast at ∞. The constants L_{ψ_i} will be important as well; in applications to regression problems in Section 3, these constants are typically related to the moment of the "errors." The thrust of the second requirement (2) is that the form of the derivatives of (smoothed versions of) ψ_i should be "similar" to itself. Its statement appears however slightly involved; the purpose of this is to include nonsmooth loss functions that occur frequently in robust regression problems.

For later purposes, we define

$$\mathsf{q} \equiv \max\{q_3, q_1 + q_2, 3q_1\}, \qquad \bar{\mathsf{q}} \equiv q_0\mathsf{q} + 3. \tag{2.2}$$

We now give two examples of loss functions $\{\psi_i\}$ that satisfy Assumption II above.

EXAMPLE 2.1 (Square-type loss). Let $\psi_i(x) \equiv (x-\xi_i)^{2s}$ for some real $\xi_i \in \mathbb{R}$ and $s \in \mathbb{N}$. It is easy to verify Assumption II with $\psi_{i;\rho} \equiv \psi_i$, $q_0 = 2s$, $L_{\psi_i} = C(1+\xi_i^{2s})$, $q_\ell = (2s-\ell)_+/(2s)$ for $\ell = 1, 2, 3$ and $\mathscr{D}_\psi(\cdot) \equiv C$, $\mathscr{M}_\psi(\cdot) \equiv 0$ for some constant $C > 1$ depending on s only. As no smoothing is required, the choice of $\bar{\rho}$ is arbitrary. In the most common square loss case ($s = 1$), $\mathsf{q} = 3/2$ and $\bar{\mathsf{q}} = 6$.

EXAMPLE 2.2 (Robust loss). Let $\psi_i(x) \equiv \psi_0(x-\xi_i)$ for some absolute continuous function $\psi_0 : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $|\psi_0(0)| \vee \operatorname{ess\,sup}|\psi_0'| \leq L_0$, and real $\xi_i \in \mathbb{R}$. Under this condition, Assumption II-(1) is satisfied with $q_0 = 1$ and $L_{\psi_i} = CL_0(1+|\xi_i|)$ for some absolute constant $C > 1$. Two concrete examples:

- (Least absolute loss) $\psi_0(x) = |x|$, so $q_0 = 1$, $L_0 = 1$, $L_{\psi_i} = C(1+|\xi_i|)$.
- (Huber loss) For any $\eta > 0$, let $\psi_0(x;\eta) \equiv (x^2/2)\mathbf{1}_{|x|\leq\eta} + (\eta|x| - (\eta^2/2))\mathbf{1}_{|x|>\eta}$, so $q_0 = 1$, $L_0 = \eta$, $L_{\psi_i} = C\eta(1+|\xi_i|)$.

Consider the smooth approximation $\psi_{i;\rho}(x) \equiv \mathbb{E}\,\psi_0(x - \xi_i + \rho Z)$, where $Z \sim \mathcal{N}(0,1)$ and $\rho \in (0,1)$. Lemma B.1 entails that Assumption II is satisfied with $q_1 = q_2 = q_3 = 0$, $\mathscr{D}_\psi(\rho) = C \cdot L_0/\rho^2$, $\mathscr{M}_\psi(\rho) = C \cdot L_0\rho$ and $\bar{\rho} = 1/(CL_0)$ for some absolute constant $C > 1$. Consequently, $\mathsf{q} = 0$ and $\bar{\mathsf{q}} = 3$.

Finally, we state the assumption on the random design matrix A.

ASSUMPTION III (Design matrix). Let $A \equiv A_0/\sqrt{m}$, where $A_0 \in \mathbb{R}^{m\times n}$ is a random matrix with independent entries $\{A_{0;ij}\}$ such that $\mathbb{E}\,A_{0;ij} = 0$, $\mathbb{E}\,A_{0;ij}^2 = 1$ for all $i \in [m]$, $j \in [n]$ and $M \equiv \max_{i\in[m],\,j\in[n]} \mathbb{E}|A_{0;ij}|^{\bar{\mathsf{q}}} < \infty$ ($\bar{\mathsf{q}}$ defined in (2.2)).

Here, A_0 is the standardized version of A with entrywise variance 1. The variance scaling $1/m$ (or equivalently $1/n$ under Assumption I) is quite common in the high-dimensional asymptotics literature; see, for example, [2, 5, 10, 11, 17, 20, 24, 33, 35, 47, 48, 51, 52, 54, 55, 57, 61, 65, 68, 69, 73, 76] for an incomplete list of recent statistical papers on this topic.

2.2. *Universality of the optimum.* First, we establish universality of the cost optimum $\min_{w\in\mathcal{S}_n} H_\psi(w, A)$ with respect to A, over an arbitrary structure set $\mathcal{S}_n$ that is "compact" in an ℓ_∞ sense. Its proof can be found in Section 4.2.

THEOREM 2.3. *Suppose Assumptions I and II hold. Let $A_0, B_0 \in \mathbb{R}^{m\times n}$ be two random matrices with independent components, such that $\mathbb{E}\,A_{0;ij} = \mathbb{E}\,B_{0;ij} = 0$ and $\mathbb{E}\,A_{0;ij}^2 = \mathbb{E}\,B_{0;ij}^2$ for all $i \in [m]$, $j \in [n]$. Further assume that*

$$M \equiv \max_{i\in[m],\,j\in[n]} \big(\mathbb{E}|A_{0;ij}|^{\bar{\mathsf{q}}} + \mathbb{E}|B_{0;ij}|^{\bar{\mathsf{q}}}\big) < \infty.$$

Let $A \equiv A_0/\sqrt{m}$ and $B \equiv B_0/\sqrt{m}$. Then there exists some $C_0 = C_0(\tau, q, M) > 0$ such that the following hold:[3] *For any $\mathcal{S}_n \subset [-L_n, L_n]^n$ with $L_n \geq 1$, and any $\mathsf{g} \in C^3(\mathbb{R})$, we have*

$$\Big|\mathbb{E}\,\mathsf{g}\Big(\min_{w\in\mathcal{S}_n} H_\psi(w, A)\Big) - \mathbb{E}\,\mathsf{g}\Big(\min_{w\in\mathcal{S}_n} H_\psi(w, B)\Big)\Big| \leq C_0 \cdot K_{\mathsf{g}} \cdot \mathsf{r}_{\mathsf{f}}(L_n).$$

Here, $K_{\mathsf{g}} \equiv 1 + \max_{\ell\in[0:3]}\|\mathsf{g}^{(\ell)}\|_\infty$, and $\mathsf{r}_{\mathsf{f}}(L_n)$ is defined by

$$\mathsf{r}_{\mathsf{f}}(L_n) \equiv \inf_{\rho\in(0,\bar\rho)}\Big\{\mathscr{M}_\psi(\rho) + \mathscr{D}^3_\psi(\rho)\inf_{\delta\in(0,\omega_n)}\Big[\mathscr{N}_{\mathsf{f}}(L_n,\delta) + \mathsf{avg}^{1/3}\big(\{L^{\mathsf{q}}_{\psi_i}\}\big)\cdot\frac{L_n^{\bar{\mathsf{q}}/3}\log_+^{2/3}(L_n/\delta)}{n^{1/6}}\Big]\Big\},$$

where $\omega_n \equiv n^{-(q_0q_1+4)/2}$, $\mathsf{avg}(\{L^{\mathsf{q}}_{\psi_i}\}) \equiv m^{-1}\sum_{i=1}^m L^{\mathsf{q}}_{\psi_i}$ and

$$\mathscr{N}_{\mathsf{f}}(L_n,\delta) \equiv \sup|\mathsf{f}(w) - \mathsf{f}(w')|$$

with the supremum taken over all $w, w' \in [-L_n, L_n]^n$ such that $\|w - w'\|_\infty \leq \delta$. Consequently, for any $z \in \mathbb{R}$, $\varepsilon > 0$,

$$\mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H_\psi(w, A) > z + 3\varepsilon\Big) \leq \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H_\psi(w, B) > z + \varepsilon\Big) + C_1(1\vee\varepsilon^{-3})\mathsf{r}_{\mathsf{f}}(L_n).$$

Here, $C_1 > 0$ is an absolute multiple of C_0.

The strength of Theorem 2.3 rests in allowing arbitrary structure sets $\mathcal{S}_n \subset [-L_n, L_n]^n$, where L_n grows slowly enough in the high-dimensional limit $n \to \infty$.

To put this result in a broader context, Theorem 2.3 is closely related to recent developments on the high-dimensional central limit theorems (see, e.g., the review article [15] for many references). This line of works considers universality of the maxima $S_m(X) \equiv \max_{1\leq j\leq p} m^{-1/2}\sum_{i=1}^m X_{ij}$ (or its variants), where $X \equiv \{X_i \in \mathbb{R}^p\}$ contains m independent p-dimensional random vectors X_i's. A crucial step therein is to exploit the ℓ_∞-like structure, that is, the maximum over $j \in [p]$, so that the resulting Berry–Esseen bounds scale as a multiple of $(\log^a p/m)^b$, with the optimal choice $a = 3$, $b = 1/2$ (e.g., [16, 26, 50]). The situation in Theorem 2.3 is more subtle: in the proportional regime $m \asymp n$, the typical "effective dimension p" of $\mathcal{S}_n$ in the minimum scales as $\log p \asymp n \asymp m$, so existing techniques in the above references do not (cannot) yield meaningful bounds. Inspired by the derivative calculations in [46, 59] embedded in the quantitative Lindeberg's method (cf. [13, 49]), here we get around this issue by providing an almost dimension-free third derivative bound along the entire Lindeberg path, using essentially the ℓ_∞ constraint on $\mathcal{S}_n$; see Section 4.2 for details.

Of course, there is no a priori reason to believe that the exponent $1/6$ in the rate $\mathsf{r}_{\mathsf{f}}(L_n)$ is optimal. In view of the recent progress in high-dimensional central limit theorems mentioned above, we conjecture that the optimal exponent is $1/2$. While theoretically interesting, we are however not aware of practical merits of this potential improvement in the applications in Section 3.

2.3. *Universality of the optimizer.* Next, we will establish universality properties for the optimizer of the cost function $w \mapsto H_\psi(w, A)$, defined as any minimizer

$$\widehat{w}_A \in \operatorname*{arg\,min}_{w\in\mathbb{R}^n} H_\psi(w, A). \tag{2.3}$$

Recall the standard Gaussian design G in (1.3). The following result is proved in Section 4.3.

[3]Here, we abbreviate "C_0 depends on $\{q_\ell : \ell \in [0:3]\}$ in Assumption II" as "C_0 depends on q," and simply write "$C_0 = C_0(\{q_\ell : \ell \in [0:3]\})$" as "$C_0 = C_0(q)$." The same convention will be adopted in the statements of other results.

THEOREM 2.4. *Suppose Assumptions* I, II *and* III *hold. Fix a measurable subset* $\mathcal{S}_n \subset \mathbb{R}^n$. *Suppose there exist* $z \in \mathbb{R}$, $\rho_0 > 0$, $L_n \geq 1$ *and* $\varepsilon_n \in [0, 1/4)$ *such that the following hold*:

(O1) *Both* $\|\widehat{w}_G\|_\infty$ *and* $\|\widehat{w}_A\|_\infty$ *grow mildly in the sense that*

$$\mathbb{P}\big(\|\widehat{w}_G\|_\infty > L_n\big) \vee \mathbb{P}\big(\|\widehat{w}_A\|_\infty > L_n\big) \leq \varepsilon_n.$$

(O2) $\widehat{w}_G$ *violates the structural property* $\mathcal{S}_n$ *in the sense that*

$$\mathbb{P}\Big(\min_{w\in\mathbb{R}^n} H_\psi(w, G) \geq z + \rho_0\Big) \vee \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H_\psi(w, G) \leq z + 2\rho_0\Big) \leq \varepsilon_n.$$

Then $\widehat{w}_A$ *also violates* $\mathcal{S}_n$ *with high probability*:

$$\mathbb{P}(\widehat{w}_A \in \mathcal{S}_n) \leq 4\varepsilon_n + C_0\big(1 \vee \rho_0^{-3}\big)\mathsf{r}_\mathsf{f}(L_n).$$

Here, $\mathsf{r}_\mathsf{f}(L_n)$ *is defined in Theorem* 2.3, *and* $C_0 > 0$ *depends on* τ, q, M *only.*

Here, $\mathcal{S}_n$ is regarded as the "exceptional set" into which the optimizer $\widehat{w}_A$ is unlikely to fall. As mentioned in the Introduction, while condition (O1) typically requires application-specific techniques, this will be verified in a unified manner via "leave-one-out" techniques in the regression examples in Section 3. To verify condition (O2), z is usually regarded as the "high-dimensional limit" of the global minimum $\min_{w\in\mathbb{R}^n} H_\psi(w, G)$, while $\rho_0 > 0$ is a small enough constant (typically of constant order) that guarantees the cost function admits a strict gap between the global optimum and the optimum over the exceptional set $\mathcal{S}_n$; see the discussion after Theorem 3.1 for more details in applications to regression problems.

2.4. *Universality of the Gordon's max–min* (*min–max*) *cost optimum.* The techniques in proving Theorem 2.3 can be generalized to establish universality for Gordon's max–min (min–max) cost optimum, defined as below. Let for $u \in \mathbb{R}^m$, $w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and a measurable function $Q : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$,

$$X(u, w; A) \equiv u^\top A w + Q(u, w). \tag{2.4}$$

THEOREM 2.5. *Let* $A, B \in \mathbb{R}^{m\times n}$ *be two random matrices with independent entries and matching first two moments, that is,* $\mathbb{E}\,A_{ij}^\ell = \mathbb{E}\,B_{ij}^\ell$ *for all* $i \in [m]$, $j \in [n]$, $\ell = 1, 2$. *There exists a universal constant* $C_0 > 0$ *such that the following hold. For any measurable subsets* $\mathcal{S}_u \subset [-L_u, L_u]^m$, $\mathcal{S}_w \subset [-L_w, L_w]^n$ *with* $L_u, L_w \geq 1$, *and any* $\mathsf{g} \in C^3(\mathbb{R})$, *we have*

$$\begin{aligned}&\Big|\mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u, w; A)\Big) - \mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u, w; B)\Big)\Big| \\ &\quad\leq C_0 \cdot K_\mathsf{g} \cdot \inf_{\delta\in(0,1)}\big\{M_1 L\delta + \mathscr{N}_Q(L, \delta) + \log_+^{2/3}(L/\delta)\cdot(m+n)^{2/3}M_3^{1/3}L^2\big\}.\end{aligned}$$

Here, $K_\mathsf{g} \equiv 1 + \max_{\ell\in[0:3]}\|\mathsf{g}^{(\ell)}\|_\infty$, $L \equiv L_u + L_w$, $M_\ell \equiv \sum_{i\in[m],\,j\in[n]}(\mathbb{E}|A_{ij}|^\ell + \mathbb{E}|B_{ij}|^\ell)$ *and*

$$\mathscr{N}_Q(L, \delta) \equiv \sup\big|Q(u, w) - Q\big(u', w'\big)\big|$$

with the supremum taken over all $u, u' \in [-L, L]^m$, $w, w' \in [-L, L]^n$ *such that* $\|u - u'\|_\infty \vee \|w - w'\|_\infty \leq \delta$. *The conclusion continues to hold when max–min is flipped to min–max.*

The proof of the above theorem can be found in Section 4.4. Due to widespread applications of the CGMT in the theoretical analysis of high-dimensional/overparametrized statistical models (as mentioned in the Introduction), we expect Theorem 2.5 to be useful in establishing universality properties for statistical estimators in other high-dimensional problems beyond the ones considered in this paper. Pertinent to this paper, this result will also be useful in some of the applications in Section 3 below.

To facilitate easy applications of Theorem 2.5, below we work out a particularly useful version where the design matrices have centered entries with variance $1/m$. Its proof is contained in Section 4.5.

COROLLARY 2.6. *Suppose Assumption* I *holds. Let* $A_0, B_0 \in \mathbb{R}^{m\times n}$ *be two random matrices with independent entries,* $\mathbb{E}\,A_{0,ij}^{\ell} = \mathbb{E}\,B_{0,ij}^{\ell} = \mathbf{1}_{\ell=2}$ *for all* $i \in [m]$, $j \in [n]$, $\ell = 1, 2$ *and* $M_0 \equiv \max_{i\in[m],\,j\in[n]}(\mathbb{E}|A_{0,ij}|^3 + \mathbb{E}|B_{0,ij}|^3) < \infty$. *Let* $A \equiv A_0/\sqrt{m}$, $B \equiv B_0/\sqrt{m}$ *and recall*

$$X_n(u, w; A) = \frac{1}{m} u^\top A w + Q_n(u, w).$$

Then there exists $C_0 = C_0(\tau, M_0) > 0$ *such that for any measurable subsets* $\mathcal{S}_u \subset [-L_u, L_u]^m$, $\mathcal{S}_w \subset [-L_w, L_w]^n$ *with* $L_u, L_w \geq 1$ *and any* $\mathsf{g} \in C^3(\mathbb{R})$, *we have*

$$\Big|\mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; A)\Big) - \mathbb{E}\,\mathsf{g}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; B)\Big)\Big|$$
$$\leq C_0 \cdot K_{\mathsf{g}} \cdot \mathsf{r}_n, \quad \textit{with } \mathsf{r}_n \equiv \inf_{0<\delta\leq n^{-1}}\Big\{\mathscr{N}_{Q_n}(L, \delta) + \frac{L^2 \log^{2/3}(L/\delta)}{n^{1/6}}\Big\}.$$

Here, $K_{\mathsf{g}} \equiv 1 + \max_{\ell\in[0:3]}\|\mathsf{g}^{(\ell)}\|_\infty$ *and* $L \equiv L_u + L_w$. *Consequently,*

$$\mathbb{P}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; A) > z + 3\varepsilon\Big)$$
$$\leq \mathbb{P}\Big(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X_n(u, w; B) > z + \varepsilon\Big) + C_1(1 \vee \varepsilon^{-3})\mathsf{r}_n$$

holds for any $z \in \mathbb{R}$ *and* $\varepsilon > 0$. *Here,* $C_1 > 0$ *is an absolute multiple of* C_0. *The conclusion continues to hold both when* (*i*) *max–min is flipped to min–max and* (*ii*) *there exists some set* $S \subset [m] \times [n]$ *such that* $A_{ij} = B_{ij} = 0$ *for* $(i, j) \in S$.

The extension to scenario (ii) will be useful in situations where the drift function Q_n contains a certain extra variable, say, v over which the maximum is also taken. This will be used in the proof of some applications (in particular, the distribution of Lasso subgradient) in Section 3.

3. Applications to high-dimensional regression.

3.1. *General regression setting.* In the linear regression model (1.1), recall $A = A_0/\sqrt{m}$ and we also write $\xi = \sigma\xi_0$, where the variance of the entries of A_0 and ξ_0 are standardized to be 1. Further recall that $\widehat{\mu}_A \equiv \mu_0 + \widehat{w}_A$, where the estimator of interest $\widehat{\mu}_A$ is defined in (1.2) and $\widehat{w}_A$ is defined in (1.6).

Below we will work out Theorem 2.4 in the above regression setting for the square loss $\psi_0(x) = x^2/2$ (Example 2.1) and the robust loss ψ_0 (Example 2.2). While we do not pursue the most general possible form here, adaptation to other loss functions is straightforward.

THEOREM 3.1. *Consider the above regression setting with either* (*i*) *square loss* $\psi_0(x) = x^2/2$ *or* (*ii*) *robust loss* ψ_0 *satisfying* $|\psi_0(0)| \vee \operatorname{ess\,sup}|\psi_0'| \leq L_0$ *for some* $L_0 > 0$. *Suppose Assumption* I *holds, Assumption* III *holds with*

$$M \equiv \begin{cases} \max_{i,j}\mathbb{E}|A_{0;ij}|^6 < \infty & \textit{square loss case;} \\ \max_{i,j}\mathbb{E}|A_{0;ij}|^3 < \infty & \textit{robust loss case,} \end{cases}$$

and ξ_0 *has independent components that are also independent of* A_0. *Further assume that there exists some* $K_{\mathsf{f}} > 1$ *such that*

$$\log \mathscr{N}_{\mathsf{f}}(L, \delta) \leq K_{\mathsf{f}}(\log L + \log n) - \log(1/\delta)/K_{\mathsf{f}} \quad \forall L \geq 1, \delta \in (0, 1). \tag{3.1}$$

Fix a measurable subset $\mathcal{S}_n \subset \mathbb{R}^n$. Suppose there exist $z \in \mathbb{R}$, $\rho_0 > 0$, $1 \le L_n \le n$ and $\varepsilon_n \in [0, 1/4)$ such that (O1) in Theorem 2.4 is fulfilled under the joint probability of (A, ξ) and (G, ξ), and (O2) fulfilled for $H_{\psi_0,\mathsf{f}}(\cdot, G, \xi)$ under the joint probability of (G, ξ). Then under the joint probability of (A, ξ),

$$\mathbb{P}(\widehat{w}_A \in \mathcal{S}_n) \le 4\varepsilon_n + C_0\big(1 \vee \rho_0^{-3}\big) \begin{cases} M_\xi^{1/3} \dfrac{L_n^2 \log^{2/3} n}{n^{1/6}} & \text{square loss case;} \\ \Big(\dfrac{L_n \log^{2/3} n}{n^{1/6}}\Big)^{1/7} & \text{robust loss case.} \end{cases}$$

Here, $M_\xi \equiv 1 + m^{-1}\sum_{i=1}^m \mathbb{E}|\sigma\xi_{0,i}|^3$. The constant $C_0 > 0$ depends on τ, M, K_{f} only in the square loss case and depends further on L_0 in the robust loss case.

The condition (3.1) on the penalty function f is imposed here to simplify the final bound, and is easily verified as long as f has some degree of global moduli of continuity. The major nontrivial condition is the ℓ_∞ bounds for $\widehat{w}_A$ and $\widehat{w}_G$ required in (O1). In the examples to be detailed below, we will use the so-called "leave-one-out" method to establish the desired ℓ_∞ bounds. Formally, we study perturbation of $\widehat{w}_A$ by (i) its column leave-one-out version

$$\widehat{w}_A^{(s)} \equiv \mathop{\arg\min}_{w \in \mathbb{R}^n : w_s = 0} \bigg\{ \frac{1}{m}\sum_{i=1}^m \psi_0\big((Aw)_i - \xi_i\big) + \mathsf{f}(\mu_0 + w) \bigg\}, \quad s \in [n]$$

and (ii) its row leave-one-out version

$$\widehat{w}_A^{[t]} \equiv \mathop{\arg\min}_{w \in \mathbb{R}^n} \bigg\{ \frac{1}{m} \sum_{i \in [m], i \ne t} \psi_0\big((Aw)_i - \xi_i\big) + \mathsf{f}(\mu_0 + w) \bigg\}, \quad t \in [m].$$

Intuitively, both $\widehat{w}_A^{(s)}$ and $\widehat{w}_A^{[t]}$ should be very close to $\widehat{w}_A$ for designs A with independent entries. We will show that indeed in many examples the orders of $\|\widehat{w}_A^{(s)} - \widehat{w}_A\|$, $\|\widehat{w}_A^{[t]} - \widehat{w}_A\|$ are almost $\mathcal{O}_{\mathbf{P}}(1)$, while the typical order of $\|\widehat{w}_A\|$ is $\mathcal{O}_{\mathbf{P}}(n^{1/2})$. The independence of $\widehat{w}_A^{(s)}$ (resp., $\widehat{w}_A^{[t]}$) with respect to the sth column (resp., tth row) of A will then play a crucial role in establishing elementwise bounds for $\widehat{w}_A$.

The method described above is closely related to the one used in [24] under the name "leave-one-observation/predictor out approximations" (also known as *cavity method* in statistical physics [53, 66, 67]); see also [44, 56] for related techniques.

Once the ℓ_∞ bound condition (O1) is verified, we then only need to study the behavior of $\widehat{w}_G$ for the standard Gaussian design G (1.3) with i.i.d. $\mathcal{N}(0, 1/m)$ entries, by creating an $\mathcal{O}(1)$ gap between $\min_{w \in \mathcal{S}_n} H_{\psi_0,\mathsf{f}}(w, G, \xi)$ over the "exceptional set" $\mathcal{S}_n$ and the global optimum $\min_{w \in \mathbb{R}^n} H_{\psi_0,\mathsf{f}}(w, G, \xi)$, where the choice of $\mathcal{S}_n$ reflects certain property of the estimator that we try to understand. Such a goal is particularly amenable to analysis via the CGMT, as it reduces the analysis of $H_{\psi_0,\mathsf{f}}(w, G, \xi)$ that involves a standard Gaussian design matrix G to a Gordon problem that involves two Gaussian vectors only. The desired gap can then be created by exploiting the (local) strong convexity of the Gordon problem.

PROOF OF THEOREM 3.1. By the assumed moduli of continuity in f, we may take $\delta = (L_n n)^{-K_{\mathsf{f}}-1}$ in the definition of $\mathsf{r}_{\mathsf{f}}(L_n)$. Now we apply Theorem 2.4 first conditionally on ξ and then take expectation. There in the square loss case, $L_{\psi_i} = C(1 + \xi_i^2)$, $\mathsf{q} = 3/2$, $\bar{\mathsf{q}} = 6$, $\mathscr{D}_\psi(\rho) \equiv C$ and $\mathscr{M}_\psi(\rho) = 0$, so

$$\mathsf{r}_{\mathsf{f}}(L_n) \lesssim M_\xi^{1/3} \cdot \frac{L_n^2 \log^{2/3}(nL_n)}{n^{1/6}}.$$

In the robust loss case, $L_{\psi_i} = CL_0(1 + |\xi_i|)$, $\mathsf{q} = 0$, $\bar{\mathsf{q}} = 3$, $\mathscr{D}_\psi(\rho) \equiv CL_0/\rho^2$, $\mathscr{M}_\psi(\rho) = CL_0\rho$ and $\bar{\rho} = 1/(CL_0)$, so

$$\mathsf{r}_{\mathsf{f}}(L_n) \lesssim_{L_0} \inf_{\rho\in(0,c')} \left\{\rho + \rho^{-6}\frac{L_n \log^{2/3}(nL_n)}{n^{1/6}}\right\} \asymp_{L_0} \left(\frac{L_n \log^{2/3}(nL_n)}{n^{1/6}}\right)^{1/7}.$$

In both displays, the term $\log^{2/3}(nL_n) \asymp \log^{2/3} n$ thanks to $1 \le L_n \le n$. $\square$

3.2. *Example I*: *Ridge regression*. In this section, we consider universality properties for the Ridge estimator [37]. Formally, let the Ridge cost function be

$$\bar{H}^{\mathsf{R}}(w, A, \xi) \equiv \frac{1}{2}\|Aw - \xi\|^2 + \frac{\lambda}{2}(\|w + \mu_0\|^2 - \|\mu_0\|^2), \tag{3.2}$$

and its normalized version $H^{\mathsf{R}} \equiv \bar{H}^{\mathsf{R}}/m$. Here, the constant term $-\lambda\|\mu_0\|^2/2$ is added to simplify the expression of the Gordon cost function. The Ridge solution is given by $\widehat{\mu}_A^{\mathsf{R}} = \widehat{w}_A^{\mathsf{R}} + \mu_0$ with

$$\widehat{w}_A^{\mathsf{R}} \equiv \operatorname*{arg\,min}_{w\in\mathbb{R}^n} H^{\mathsf{R}}(w, A, \xi) = (A^\top A + \lambda I)^{-1}(A^\top \xi - \lambda\mu_0). \tag{3.3}$$

We will work with the following conditions instead of referring back to the assumptions listed in Section 2:

(R1) $\tau \le m/n \le 1/\tau$ holds for some $\tau \in (0, 1)$, and $\lambda > 0$.
(R2) $\|\mu_0\|^2/n \le M_2$ for some $M_2 > 0$.
(R3) $A_0 = \sqrt{m}A$ and $\xi_0 = \xi/\sigma$ are independent, and their entries are all independent, mean 0, variance 1 and uniformly sub-Gaussian variables.

The precise mathematical meaning of uniform sub-Gaussianity in (R3) is that

$$\sup_n \max_{i\in[m], j\in[n]} (\|A_{0,ij}\|_{\psi_2} + \|\xi_{0,i}\|_{\psi_2}) < \infty,$$

where $\|\cdot\|_{\psi_2}$ is the Orlicz-2 norm or the sub-Gaussian norm (definition see [72], Section 2.1).

The following theorem establishes the generic universality of $\widehat{w}_A^{\mathsf{R}}$ with respect to the design matrix A. All proofs in this section can be found in Section 5.

THEOREM 3.2. *Suppose* (R1)–(R3) *hold. Fix* $\mathcal{S}_n \subset \mathbb{R}^n$. *Suppose there exist* $z \in \mathbb{R}$, $\rho_0 > 0$ *and* $\varepsilon_n \in [0, 1/4)$ *such that*

$$\mathbb{P}\Big(\min_{w\in\mathbb{R}^n} H^{\mathsf{R}}(w, G, \xi) \ge z + \rho_0\Big) \vee \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H^{\mathsf{R}}(w, G, \xi) \le z + 2\rho_0\Big) \le \varepsilon_n. \tag{3.4}$$

Then there exists some $K = K(\sigma, \lambda, \tau, M_2) > 0$ *such that*

$$\mathbb{P}(\widehat{w}_A^{\mathsf{R}} \in \mathcal{S}_n) \le 4\varepsilon_n + K \cdot (1 \vee \rho_0^{-3})(1 \vee \|\mu_0\|_\infty^2) \cdot n^{-1/6}\log^2 n.$$

The sub-Gaussian moments in (R3) are assumed for simplicity; easy modifications of the proofs allow for weaker conditions, for example, the existence of high enough moments, at the cost of a possible worsened probability bound. We do not pursue these nonessential refinements here for clarity of presentation and proofs.

The key to the proof of Theorem 3.2 is the following ℓ_∞ bound (and other risk bounds) for $\widehat{w}_A^{\mathsf{R}}$, which may be of independent interest.

PROPOSITION 3.3. *Assume the same conditions as in Theorem* 3.2. *Then the following hold with probability at least* $1 - Cn^{-100}$ *with respect to the randomness of* (A, ξ):

1. (*Prediction risk*) $\|A\widehat{w}_A^{\mathsf{R}}\|^2 \le K \cdot n$.
2. (ℓ_∞ *risk*) $\|\widehat{w}_A^{\mathsf{R}}\|_\infty \le K\sqrt{\log n} + 2\|\mu_0\|_\infty$.
3. (*Prediction* ℓ_∞ *risk*) $\|A\widehat{w}_A^{\mathsf{R}}\|_\infty \le K\sqrt{\log n}$.

Here, $C, K > 0$ *depend on* $\sigma, \lambda, \tau, M_2$ *only.*

Thanks to the closed-form formula for the Ridge estimator in (3.3), some of its properties that can be related to the spectrum of $A^\top A$ including estimation error/prediction error (in the Euclidean norm) can be established directly via existing random matrix theory (RMT) (cf. [18, 19, 36, 52]). Here, we will illustrate the power of Theorem 3.2 and its compatibility with the CGMT, by establishing nonasymptotic distributional approximations of $\widehat{w}_A^{\mathsf{R}}$ and its residual given by $\widehat{r}_A^{\mathsf{R}} \equiv Y - A\widehat{\mu}_A^{\mathsf{R}}$. These results appear to be less amenable to direct applications of existing RMT techniques. More important, the formulation of these results suggests natural generalizations to the Lasso case in which no closed-form formulas are available (cf. Section 3.3).

Recall $\eta_2(\cdot;\cdot)$ defined in (1.9). By Proposition 5.2(2), the system of equations

$$\begin{aligned}(\gamma_*^{\mathsf{R}})^2 &= \sigma^2 + \frac{1}{m/n}\cdot\mathbb{E}\bigg[\eta_2\Big(\Pi_{\mu_0} + \gamma_*^{\mathsf{R}} Z; \frac{\gamma_*^{\mathsf{R}}\lambda}{\beta_*^{\mathsf{R}}}\Big) - \Pi_{\mu_0}\bigg]^2,\\ \beta_*^{\mathsf{R}} &= \gamma_*^{\mathsf{R}}\bigg[1 - \frac{1}{m/n}\cdot\mathbb{E}\,\eta_2'\Big(\Pi_{\mu_0} + \gamma_*^{\mathsf{R}} Z; \frac{\gamma_*^{\mathsf{R}}\lambda}{\beta_*^{\mathsf{R}}}\Big)\bigg],\end{aligned} \tag{3.5}$$

where $\Pi_{\mu_0} \otimes Z \equiv (n^{-1}\sum_{j=1}^n \delta_{\mu_{0,j}}) \otimes \mathcal{N}(0,1)$, admits a unique solution $(\beta_*^{\mathsf{R}}, \gamma_*^{\mathsf{R}})$ within compacta of $[0,\infty)^2$ provided (R1)–(R2) are satisfied. Again, one may give explicit formulae for $(\beta_*^{\mathsf{R}}, \gamma_*^{\mathsf{R}})$ defined using (3.5), but we stick to the above fixed-point equation formulation for transparent comparison to the Lasso case in (3.9) below.

Now we define the 'population version' of $\widehat{w}_A^{\mathsf{R}}$ and $\widehat{r}_A^{\mathsf{R}}$ via $(\beta_*^{\mathsf{R}}, \gamma_*^{\mathsf{R}})$ by

$$w_*^{\mathsf{R}} \equiv \eta_2\Big(\mu_0 + \gamma_*^{\mathsf{R}} Z_n; \frac{\gamma_*^{\mathsf{R}}\lambda}{\beta_*^{\mathsf{R}}}\Big) - \mu_0, \qquad r_*^{\mathsf{R}} \equiv \frac{\beta_*^{\mathsf{R}}}{\gamma_*^{\mathsf{R}}}\big(\sigma\cdot\xi_0 + \sqrt{(\gamma_*^{\mathsf{R}})^2 - \sigma^2}\cdot Z_m'\big), \tag{3.6}$$

where $Z_n \sim \mathcal{N}(0, I_n)$ and $Z_m' \sim \mathcal{N}(0, I_m)$ are independent standard Gaussian vectors that are also independent of the noise vector $\xi = \sigma\xi_0$.

THEOREM 3.4. *Assume the same conditions as in Theorem 3.2. Then there exists some* $K = K(\sigma, \lambda, \tau, M_2) > 0$ *such that for all* 1*-Lipschitz functions* $\mathsf{g}: \mathbb{R}^n \to \mathbb{R}$, $\mathsf{h}: \mathbb{R}^m \to \mathbb{R}$ *and* $\varepsilon > 0$,

$$\begin{aligned}&\mathbb{P}\big(|\mathsf{g}(\widehat{w}_A^{\mathsf{R}}/\sqrt{n}) - \mathbb{E}\,\mathsf{g}(w_*^{\mathsf{R}}/\sqrt{n})| \ge \varepsilon\big) \vee \mathbb{P}\big(|\mathsf{h}(\widehat{r}_A^{\mathsf{R}}/\sqrt{m}) - \mathbb{E}'\,\mathsf{h}(r_*^{\mathsf{R}}/\sqrt{m})| \ge \varepsilon\big)\\ &\quad\le K\cdot\big(1\vee\varepsilon^{-6}\big)\big(1\vee\|\mu_0\|_\infty^2\big)\cdot n^{-1/6}\log^2 n.\end{aligned}$$

Here, $\mathbb{E}'$ *indicates that the expectation is taken with respect to* Z_m' *only.*

The first claim on $\widehat{w}_A^{\mathsf{R}}$ in the above theorem is proved via an application of Theorem 3.2, by analyzing the corresponding Gaussian design problem via the CGMT. The proof for the second claim on $\widehat{r}_A^{\mathsf{R}}$ in the above theorem is more involved, and requires an application of the universality result for Gordon's max–min (min–max) cost in Theorem 2.5.

As a quick application of Theorem 3.4 above, we may obtain universality of the distribution of $\widehat{w}_A^{\mathsf{R}}$ and $\widehat{r}_A^{\mathsf{R}}$ in an average sense as follows:

- Let $\mathsf{g}(v) \equiv n^{-1}\sum_{j=1}^{n}\phi(\sqrt{n}v_j + \mu_{0,j}, \mu_{0,j})$ for some 1-Lipschitz function $\phi : \mathbb{R}^2 \to \mathbb{R}$ in the first probability that we have with high probability

$$\frac{1}{n}\sum_{j=1}^{n}\phi\big(\widehat{\mu}^{\mathsf{R}}_{A,j}, \mu_{0,j}\big) \approx \mathbb{E}\,\phi\big(\eta_2(\Pi_{\mu_0} + \gamma^{\mathsf{R}}_* Z; \lambda\gamma^{\mathsf{R}}_*/\beta^{\mathsf{R}}_*), \Pi_{\mu_0}\big).$$

  Here, the expectation is taken over $\Pi_{\mu_0} \otimes Z = (n^{-1}\sum_{j=1}^{n}\delta_{\mu_{0,j}}) \otimes \mathcal{N}(0,1)$.
- Let $\mathsf{h}(v) = m^{-1}\sum_{i=1}^{m}\phi(\sqrt{m}v_i)$ for some 1-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$ in the second probability that we have with high (unconditional) probability

$$\frac{1}{m}\sum_{i=1}^{m}\phi\big(\widehat{r}^{\mathsf{R}}_{A,i}\big) \approx \mathbb{E}\,\phi\bigg[\frac{\beta^{\mathsf{R}}_*}{\gamma^{\mathsf{R}}_*}\big(\sigma\cdot\Pi_{\xi_0} + \sqrt{(\gamma^{\mathsf{R}}_*)^2 - \sigma^2}\cdot Z\big)\bigg].$$

  Here, the expectation is taken over $\Pi_{\xi_0} \otimes Z = (m^{-1}\sum_{i=1}^{m}\delta_{\xi_{0,i}}) \otimes \mathcal{N}(0,1)$.

REMARK 3.5. We compare Theorem 3.4 to several results in the literature:

- For the distribution of $\widehat{w}^{\mathsf{R}}_A$, [60] obtained the following special version of universality for design matrices consisting of i.i.d. entries with vanishing third/fifth moments (almost symmetry): for convex $\mathsf{g}_0 : \mathbb{R} \to \mathbb{R}$ with bounded second and third derivatives, or $\mathsf{g}_0 = \mathbf{1}_{\cdot \ge x}$ for any $x \in \mathbb{R}$, $n^{-1}\sum_{j=1}^{n}\mathsf{g}_0(\widehat{w}^{\mathsf{R}}_{A,j})$ and $n^{-1}\sum_{j=1}^{n}\mathsf{g}_0(\widehat{w}^{\mathsf{R}}_{G,j})$ converge to the same limit. Our results are nonasymptotic allowing for arbitrary nonseparable Lipschitz test functions, and do not require prior distributions on μ_0 and vanishing third/fifth moments (almost symmetry) of the design entries.
- For the distribution of $\widehat{r}^{\mathsf{R}}_A$, [7], Theorem 3.1, obtained stochastic representation of $\widehat{r}^{\mathsf{R}}_G$ under (correlated) Gaussian designs in a broader class of problems. The results in [7] depend on the Gaussian design assumption crucially via repeated applications of Gaussian integration by parts (e.g., the Stein's identity and the second-order Stein formula in [8, 62]).

3.3. *Example II*: *Lasso*. In this section, we consider universality properties for the Lasso estimator [70]. Formally, let the Lasso cost function be

$$\bar{H}^{\mathsf{L}}(w, A, \xi) \equiv \frac{1}{2}\|Aw - \xi\|^2 + \lambda\big(\|w + \mu_0\|_1 - \|\mu_0\|_1\big), \tag{3.7}$$

and its normalized version $H^{\mathsf{L}} \equiv \bar{H}^{\mathsf{L}}/m$. Again, here the constant term $-\lambda\|\mu_0\|_1$ is added to simplify the expression of the Gordon cost function, which matches exactly to that used in [54]. The Lasso solution is $\widehat{\mu}^{\mathsf{L}}_A \equiv \widehat{w}^{\mathsf{L}}_A + \mu_0$ with

$$\widehat{w}^{\mathsf{L}}_A \equiv \operatorname*{arg\,min}_{w\in\mathbb{R}^n} H^{\mathsf{L}}(w, A, \xi).$$

We continue working with the conditions (R1)–(R3) in Section 3.2. The following theorem establishes the generic universality of $\widehat{w}^{\mathsf{L}}_A$ with respect to the design matrix A. All proofs in this section can be found in Section 6.

THEOREM 3.6. *Suppose* (*R*1)–(*R*3) *hold. Suppose further that* $\lambda \ge K_0(1 \vee \sigma)$ *for some* $K_0 = K_0(M_2, \tau) > 0$. *Fix* $\mathcal{S}_n \subset \mathbb{R}^n$. *Suppose there exist* $z \in \mathbb{R}$, $\rho_0 > 0$ *and* $\varepsilon_n \in [0, 1/4)$ *such that*

$$\mathbb{P}\Big(\min_{w\in\mathbb{R}^n} H^{\mathsf{L}}(w, G, \xi) \ge z + \rho_0\Big) \vee \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H^{\mathsf{L}}(w, G, \xi) \le z + 2\rho_0\Big) \le \varepsilon_n. \tag{3.8}$$

Then there exists some $K = K(\sigma, \lambda, \tau, M_2) > 0$ *such that*

$$\mathbb{P}\big(\widehat{w}^{\mathsf{R}}_A \in \mathcal{S}_n\big) \le 4\varepsilon_n + K\cdot\big(1 \vee \rho_0^{-3}\big)\cdot n^{-1/6}\log^2 n.$$

The lower bound on λ *can be eliminated when* $m/n \ge 1 + \varepsilon$ *for some* $\varepsilon > 0$ *at the cost of possibly enlarged constants* K *depending further on* ε.

Note that a lower bound on the tuning parameter λ is imposed only in the regime $m/n < 1$; a precise value for this lower bound can be found in the statement of Lemma 6.3. Such a condition renders sufficient linear-order sparsity of the regression estimator in the proportional regime $m \asymp n$, and is quite common in the literature; see, for example, [9, 10] for related results in the Gaussian design case.

The key to the proof of Theorem 3.6 is the following ℓ_∞ bound (and other risk bounds) for $\widehat{w}_A^{\mathsf{L}}$, which may be of independent interest.

PROPOSITION 3.7 (Lasso risk bounds). *Assume the same conditions as in Theorem 3.6. Suppose $\lambda \geq K_0(1 \vee \sigma)$ for some $K_0 = K_0(M_2, \tau) > 0$. Then the following holds with probability at least $1 - Cn^{-100}$ with respect to the randomness of (A, ξ):*

1. *(Prediction risk)* $\|A\widehat{w}_A^{\mathsf{L}}\|^2 \leq K \cdot n$.
2. *(ℓ_∞ risk)* $\|\widehat{w}_A^{\mathsf{L}}\|_\infty \leq K\sqrt{\log n}$.
3. *(Prediction ℓ_∞ risk)* $\|A\widehat{w}_A^{\mathsf{L}}\|_\infty \leq K\sqrt{\log n}$.

Here, $C, K > 0$ depend on $\sigma, \lambda, \tau, M_2$ only. The lower bound on λ can be eliminated when $m/n \geq 1 + \varepsilon$ for some $\varepsilon > 0$ at the cost of possibly enlarged constants C, K depending further on ε.

To the best of our knowledge, ℓ_∞ bounds for Lasso in the proportional regime $m \asymp n$ without exact sparsity conditions on μ_0 (or in the linear order sparsity regime) are available only in the Gaussian design case, under a similar lower bound requirement on λ when $m/n < 1$; see [9], Theorem 5.1, for a precise statement. The "interpolation" proof techniques used therein are specific to the Gaussianity of the design matrix, and cannot be extended easily to non-Gaussian design matrices. Here, we use leave-one-out methods (as mentioned after Theorem 3.1) to establish ℓ_∞ bounds for Lasso for general design matrices in the proportional regime.

Now we give an application of Theorem 3.6, coupled with the CGMT method and the comparison inequalities in Theorem 2.5 or Corollary 2.6, that establishes the universality of the distributions of:

- the error $\widehat{w}_A^{\mathsf{L}} = \widehat{\mu}_A^{\mathsf{L}} - \mu_0$,
- the residual $\widehat{r}_A^{\mathsf{L}} \equiv Y - A\widehat{\mu}_A^{\mathsf{L}}$,
- the subgradient $\widehat{v}_A^{\mathsf{L}} \equiv \lambda^{-1}A^\top(Y - A\widehat{\mu}_A^{\mathsf{L}})$, and
- the sparsity $\widehat{s}_A^{\mathsf{L}} \equiv \|\widehat{\mu}_A^{\mathsf{L}}\|_0/n$.

Recall $\eta_1(\cdot;\cdot)$ defined in (1.9) and $\Pi_{\mu_0} \otimes Z \equiv (n^{-1}\sum_{j=1}^n \delta_{\mu_{0,j}}) \otimes \mathcal{N}(0,1)$. By [54], the system of equations

$$
\begin{aligned}
(\gamma_*^{\mathsf{L}})^2 &= \sigma^2 + \frac{1}{m/n} \cdot \mathbb{E}\Big[\eta_1\Big(\Pi_{\mu_0} + \gamma_*^{\mathsf{L}} Z; \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big) - \Pi_{\mu_0}\Big]^2, \\
\beta_*^{\mathsf{L}} &= \gamma_*^{\mathsf{L}}\Big[1 - \frac{1}{m/n} \cdot \mathbb{E}\,\eta_1'\Big(\Pi_{\mu_0} + \gamma_*^{\mathsf{L}} Z; \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big)\Big]
\end{aligned} \tag{3.9}
$$

admits a unique solution $(\beta_*^{\mathsf{L}}, \gamma_*^{\mathsf{L}})$ within compacta of $[0,\infty)^2$ provided (R1)–(R2) are satisfied. Let the "population version" of $\widehat{w}_A^{\mathsf{L}}, \widehat{r}_A^{\mathsf{L}}, \widehat{v}_A^{\mathsf{L}}, \widehat{s}_A^{\mathsf{L}}$ be

$$
w_*^{\mathsf{L}} \equiv \eta_1\Big(\mu_0 + \gamma_*^{\mathsf{L}} Z_n; \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big) - \mu_0, \qquad r_*^{\mathsf{L}} \equiv \frac{\beta_*^{\mathsf{L}}}{\gamma_*^{\mathsf{L}}}(\sigma \cdot \xi_0 + \sqrt{(\gamma_*^{\mathsf{L}})^2 - \sigma^2} \cdot Z_m'), \tag{3.10}
$$

$$
v_*^{\mathsf{L}} \equiv -\frac{\beta_*^{\mathsf{L}}}{\gamma_*^{\mathsf{L}}\lambda}\Big[\eta_1\Big(\mu_0 + \gamma_*^{\mathsf{L}} Z_n; \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big) - (\mu_0 + \gamma_*^{\mathsf{L}} Z_n)\Big] = -\frac{\beta_*^{\mathsf{L}}}{\gamma_*^{\mathsf{L}}\lambda}(w_*^{\mathsf{L}} - \gamma_*^{\mathsf{L}} Z_n), \tag{3.11}
$$

$$s_*^{\mathsf{L}} = \mathbb{E}\,\eta_1'\Big(\Pi_{\mu_0} + \gamma_*^{\mathsf{L}} Z; \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big) = \mathbb{P}\Big(|\Pi_{\mu_0} + \gamma_*^{\mathsf{L}} Z| \geq \frac{\gamma_*^{\mathsf{L}}\lambda}{\beta_*^{\mathsf{L}}}\Big), \tag{3.12}$$

where $Z_n \sim \mathcal{N}(0, I_n)$ and $Z_m' \sim \mathcal{N}(0, I_m)$ are independent standard Gaussian vectors that are also independent of the noise vector ξ. Clearly, the fixed-point equation (3.9) and the "population" quantities w_*^{L}, r_*^{L} in (3.10) for the Lasso estimator are in complete analogue to those for the Ridge estimator defined in (3.5) and (3.6). The "population" quantities v_*^{L}, s_*^{L} are of special interest to the Lasso.

THEOREM 3.8. *Assume the same conditions as in Theorem 3.6. Then there exists some* $K = K(\sigma, \lambda, \tau, M_2) > 0$ *such that for all* 1*-Lipschitz functions* $\mathsf{g} : \mathbb{R}^n \to \mathbb{R}$, $\mathsf{h} : \mathbb{R}^m \to \mathbb{R}$ *and* $\varepsilon > 0$, *all the following probabilities*:

- $\mathbb{P}(|\mathsf{g}(\widehat{w}_A^{\mathsf{L}}/\sqrt{n}) - \mathbb{E}\,\mathsf{g}(w_*^{\mathsf{L}}/\sqrt{n})| \geq \varepsilon)$,
- $\mathbb{P}(|\mathsf{h}(\widehat{r}_A^{\mathsf{L}}/\sqrt{m}) - \mathbb{E}'\,\mathsf{h}(r_*^{\mathsf{L}}/\sqrt{m})| \geq \varepsilon)$,
- $\mathbb{P}(|\mathsf{g}(\widehat{v}_A^{\mathsf{L}}/\sqrt{n}) - \mathbb{E}\,\mathsf{g}(v_*^{\mathsf{L}}/\sqrt{n})| \geq \varepsilon)$,
- $\mathbb{P}(|\widehat{s}_A^{\mathsf{L}} - s_*^{\mathsf{L}}| \geq \varepsilon^{1/2})$

are bounded by $K \cdot (1 \vee \varepsilon^{-6}) \cdot n^{-1/6} \log^3 n$.

The proofs of the results in Theorem 3.8 are fairly involved, even given Theorem 3.6, Proposition 3.7 and the results in [54]—one needs to pay special attention to (suitable versions of) ℓ_∞ constrained "Gordon problems" over exception sets. We also note that the result for $\widehat{s}_A^{\mathsf{L}}$ does not follow directly from $\widehat{w}_A^{\mathsf{L}}$—in fact, similar to [54], the distributional characterization for $\widehat{w}_A^{\mathsf{L}}$ only provides a lower bound for $\widehat{s}_A^{\mathsf{L}}$, while a matching upper bound is provided by the control of the subgradient $\widehat{v}_A^{\mathsf{L}}$.

To put Theorem 3.8 in the literature, [60] obtained universality for $\widehat{w}_A^{\mathsf{L}}$ in a quite restrictive sense under several strong conditions on the design distributions (for details see Remark 3.5). The work [54] obtained distributional characterizations in the isotropic Gaussian design and Gaussian error case; our results here extend those of [54] to general designs and errors.

As an immediate application of Theorem 3.8, we may use the observable quantities $\|\widehat{r}_A^{\mathsf{L}}\|$, $\|\widehat{v}_A^{\mathsf{L}}\|$, $\widehat{s}_A^{\mathsf{L}}$ to form consistent estimators for the estimation error $\|\widehat{w}_A^{\mathsf{L}}\|^2/n$, the prediction error $\|A\widehat{w}_A^{\mathsf{L}}\|^2/m$, the original noise level σ and the effective noise level γ_*^{L} under general designs A and errors ξ. For instance, we may use

$$\widehat{\gamma}_A^{\mathsf{L}} \equiv \frac{\|\widehat{r}_A^{\mathsf{L}}\|/\sqrt{m}}{1 - \frac{1}{m/n}\widehat{s}_A^{\mathsf{L}}} = \frac{\sqrt{m}\|Y - A\widehat{\mu}_A^{\mathsf{L}}\|}{m - \|\widehat{\mu}_A^{\mathsf{L}}\|_0} \tag{3.13}$$

as a consistent estimator for γ_*^{L}; see [54], Section 4.1, for precise formulae of estimators for other quantities mentioned above.

As another important outlet of the proofs of Theorem 3.8, we consider the distribution of the degrees-of-freedom (dof) adjusted debiased Lasso $\widehat{\mu}_A^{\mathsf{dL}}$ (cf. [9, 10, 42–44, 71, 75]) defined by

$$\widehat{\mu}_A^{\mathsf{dL}} \equiv \widehat{\mu}_A^{\mathsf{L}} + \frac{A^\top(Y - A\widehat{\mu}_A^{\mathsf{L}})}{1 - \|\widehat{\mu}_A^{\mathsf{L}}\|_0/m}, \tag{3.14}$$

and the validity of the following $(1-\alpha)$ confidence intervals for $\{\mu_{0,j}\}$:

$$\mathsf{CI}_j^{\mathsf{dL}} \equiv [\widehat{\mu}_{A,j}^{\mathsf{dL}} - z_{\alpha/2} \cdot \widehat{\gamma}_A^{\mathsf{L}}, \widehat{\mu}_{A,j}^{\mathsf{dL}} + z_{\alpha/2} \cdot \widehat{\gamma}_A^{\mathsf{L}}], \quad j \in [n]. \tag{3.15}$$

Here, z_α is the normal upper α-quantile defined via $\mathbb{P}(\mathcal{N}(0,1) > z_\alpha) = \alpha$.

THEOREM 3.9. *Assume the same conditions as in Theorem 3.6. Then there exists some* $K = K(\sigma, \lambda, \tau, M_2) > 0$ *such that for any* $\mathsf{g} : \mathbb{R}^2 \to \mathbb{R}$ *and* $\varepsilon \in (0,1)$,

$$\mathbb{P}\big(\big|\mathbb{E}^\circ \mathsf{g}(\Pi_{\widehat{\mu}_A^{\mathsf{dL}}}, \Pi_{\mu_0}) - \mathbb{E}\mathsf{g}(\Pi_{\mu_0} + \gamma_*^{\mathsf{L}} Z, \Pi_{\mu_0})\big| \ge (\|\mathsf{g}\|_{\mathrm{Lip}} \vee \|\mathsf{g}\|_\infty) \cdot \varepsilon\big) \le K \cdot \varepsilon^{-12} \cdot n^{-1/6} \log^3 n.$$

Here, we write $\mathbb{E}^\circ[\cdot] = \mathbb{E}[\cdot|A, \xi]$ *and* $(\Pi_{\widehat{\mu}_A^{\mathsf{dL}}}, \Pi_{\mu_0}) = n^{-1}\sum_{j=1}^n \delta_{(\widehat{\mu}_{A,j}^{\mathsf{dL}}, \mu_{0,j})}$, $\Pi_{\mu_0} \otimes Z = (n^{-1}\sum_{j=1}^n \delta_{\mu_{0,j}}) \otimes \mathcal{N}(0,1)$. *Consequently, with the averaged empirical coverage for* $\{\mathsf{CI}_j^{\mathsf{dL}}\}$ *defined as* $\widehat{\mathscr{C}}_A^{\mathsf{dL}} \equiv n^{-1}\sum_{j=1}^n \mathbf{1}(\mu_{0,j} \in \mathsf{CI}_j^{\mathsf{dL}})$, *for any* $\varepsilon \in (0,1)$,

$$\mathbb{P}\big(\big|\widehat{\mathscr{C}}_A^{\mathsf{dL}} - (1-\alpha)\big| > \varepsilon\big) \le K \cdot \varepsilon^{-24} \cdot n^{-1/6} \log^3 n.$$

Note that the above theorem does not directly follow from Theorem 3.8 due to the lack of the joint distributional characterizations for $(\widehat{w}_A^{\mathsf{L}}, \widehat{r}_A^{\mathsf{L}})$. Inspired by [54], this technical issue is overcome by establishing distributional characterizations of $(\widehat{w}_A^{\mathsf{L}}, \widehat{r}_A^{\mathsf{L}})$ in Wasserstein-2 distance that provide couplings to relate the joint distribution of $(\widehat{w}_A^{\mathsf{L}}, \widehat{r}_A^{\mathsf{L}})$; see Proposition 6.7 for details.

To put Theorem 3.9 in the literature, for the dof adjusted debiased Lasso (3.14), [43] obtained an asymptotic version and [54] obtained an improved nonasymptotic version, of the above theorem in the isotropic Gaussian design and Gaussian error case. Distributional characterizations for dof adjusted debiased Lasso under general correlated Gaussian designs and Gaussian errors are obtained in [9, 10, 12]. These works rely crucially on the Gaussianity of the design via either the CGMT (cf. [12, 54]) or Gaussian integration by parts techniques (cf. [9, 10]). To the best of our knowledge, Theorem 3.9 provides the first theoretical justification for the dof adjusted debiased Lasso beyond Gaussian designs.

A limitation of the coverage guarantee for $\{\mathsf{CI}_j^{\mathsf{dL}}\}$ in Theorem 3.9 above is its average nature. In the (general correlated) Gaussian design and Gaussian error case, [10], Theorem 3.10, obtained stronger coverage guarantees for $\{\mathsf{CI}_j^{\mathsf{dL}}\}$ that hold for individual coordinates; see also the discussion after [6], Theorem 4.1. Whether such stronger guarantees also hold for general designs and errors remains an interesting open question.

3.4. *Example III*: *Regularized robust regression*. In this section, we consider universality properties for robust regression estimators [40, 41]. Let the robust cost function be

$$\bar{H}^{\mathsf{M}}(w, A, \xi) \equiv \sum_{i=1}^m \psi_0\big((Aw)_i - \xi_i\big) + \frac{\lambda}{2}\big(\|w + \mu_0\|^2 - \|\mu_0\|^2\big),$$

and its normalized version $H^{\mathsf{M}} \equiv \bar{H}^{\mathsf{M}}/m$. The robust regression solution is given by $\widehat{\mu}_A^{\mathsf{M}} = \widehat{w}_A^{\mathsf{M}} + \mu_0$ with

$$\widehat{w}_A^{\mathsf{M}} \equiv \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \bar{H}^{\mathsf{M}}(w, A, \xi).$$

Instead of the conditions (R1)–(R3), we work with the following alternative set of conditions:

(M1) $\tau \le m/n \le 1/\tau$ holds for some $\tau \in (0,1)$ and $\lambda > 0$.
(M2) $\psi_0 : \mathbb{R} \to \mathbb{R}$ is convex with weak derivative ψ_0' satisfying $|\psi_0(0)| \vee \operatorname{ess\,sup}|\psi_0'| \le L_0$ for some $L_0 > 0$.
(M3) $A_0 = \sqrt{m}A$ and $\xi_0 = \xi$ are independent. The entries of A_0 are independent, mean 0, variance 1 with $M_{6+\delta;A} \equiv \max_{i\in[m],\, j\in[n]} \mathbb{E}|A_{0;ij}|^{6+\delta} < \infty$ for some $\delta \in (0,1)$. The entries of ξ_0 are independent.

Note that under the above assumption on $\psi_0(\cdot)$, the Ridge penalty guarantees the existence and uniqueness of $\widehat{w}_A^{\mathsf{M}}$.

The following theorem establishes the generic universality of $\widehat{w}_A^{\mathsf{M}}$ with respect to the design matrix A. All proofs in this section can be found in Section 7.

THEOREM 3.10. *Suppose* (*M*1)–(*M*3) *hold. Fix* $\mathcal{S}_n \subset \mathbb{R}^n$. *Suppose there exist* $z \in \mathbb{R}$, $\rho_0 > 0$ *and* $\varepsilon_n \in [0, 1/4)$ *such that*

$$\mathbb{P}\Big(\min_{w\in\mathbb{R}^n} H^{\mathsf{M}}(w, G, \xi) \geq z + \rho_0\Big) \vee \mathbb{P}\Big(\min_{w\in\mathcal{S}_n} H^{\mathsf{M}}(w, G, \xi) \leq z + 2\rho_0\Big) \leq \varepsilon_n. \tag{3.16}$$

Then there exists some $K = K(\lambda, \tau, M_{6+\delta;A}, \delta, L_0) > 0$ *such that*

$$\mathbb{P}\big(\widehat{w}_A^{\mathsf{M}} \in \mathcal{S}_n\big) \leq 4\varepsilon_n + K\big(1 + \|\mu_0\|_\infty^{6+\delta} + \rho_0^{-3}\big) \cdot n^{-(1\wedge\delta)/500}.$$

A significant feature of Theorem 3.10 above is that no a priori moment conditions on the error vector ξ are required. This is particularly appealing from the perspective of robust regression [40, 41].

The next proposition establishes an elementwise bound for $\widehat{w}_A^{\mathsf{M}}$ that serves as the key to the proof of Theorem 3.10.

PROPOSITION 3.11. *Suppose* (*M*1)–(*M*3) *hold. Then for any* $p \geq 2$, *there exists some* $K = K(p) > 0$ *such that*

$$\max_{j\in[n]} \mathbb{E}\big|\widehat{w}_{A,j}^{\mathsf{M}}\big|^p \leq K \cdot \big\{(L_0/\lambda)^p M_{p;A} + \|\mu_0\|_\infty^p\big\}.$$

Here, $M_{p;A} \equiv \max_{i\in[m],\, j\in[n]} \mathbb{E}|A_{0;ij}|^p$.

As a quick demonstration of the power of Theorem 3.10 above, below we establish the asymptotic risk universality for $\widehat{\mu}_A^{\mathsf{M}}$ with the help of essentially existing Gaussian design results in [68] proved via the CGMT method.

THEOREM 3.12. *Suppose the following hold*:

1. $m/n \to \tau_0 \in (0, \infty)$ *and* $\lambda > 0$ *is fixed.*
2. ψ_0 *satisfies* (*M*2) *and either* (*i*) ψ_0 *is not differentiable at certain point, or* (*ii*) ψ_0 *contains an interval on which* ψ_0 *is differentiable with strictly increasing derivative.*
3. *The entries of* A_0 *are independent, mean* 0, *variance* 1 *with*
$$\sup_n \max_{i\in[m],\, j\in[n]} \mathbb{E}|A_{0;ij}|^{6+\delta} < \infty$$
for some $\delta \in (0, 1)$.
4. $\xi = (\xi_i)$ *contains i.i.d. components with a continuous Lebesgue density.*
5. μ_0 *contains i.i.d. components whose law* Π_0 *possesses moment of any order.*

Then with $Z \sim \mathcal{N}(0, 1)$, *the system of equations*

$$\begin{aligned} \big(\gamma_*^{\mathsf{M}}\big)^2/\tau_0 &= \mathbb{E}\big[\gamma_*^{\mathsf{M}} Z + \xi_1 - \mathsf{prox}_{\psi_0}\big(\gamma_*^{\mathsf{M}} Z + \xi_1;\, \beta_*^{\mathsf{M}}\big)\big]^2 + \lambda^2\big(\beta_*^{\mathsf{M}}\big)^2 \cdot \mathbb{E}\,\Pi_0^2, \\ 1 - \tau_0^{-1} + \lambda\beta_*^{\mathsf{M}} &= \mathbb{E}\,\mathsf{prox}'_{\psi_0}\big(\gamma_*^{\mathsf{M}} Z + \xi_1;\, \beta_*^{\mathsf{M}}\big) \end{aligned} \tag{3.17}$$

admits a unique nontrivial solution $(\beta_*^{\mathsf{M}}, \gamma_*^{\mathsf{M}}) \in (0, \infty)^2$ *such that*

$$\frac{\|\widehat{\mu}_A^{\mathsf{M}} - \mu_0\|^2}{n} \to \tau_0\big(\gamma_*^{\mathsf{M}}\big)^2 \quad \textit{in probability.}$$

In the second equation of (3.17), $\mathsf{prox}'_{\psi_0}(x;\tau) = (\mathrm{d}/\mathrm{d}x)\,\mathsf{prox}_{\psi_0}(x;\tau)$ is interpreted as the weak derivative thanks to the 1-Lipschitz property of the proximal map $x \mapsto \mathsf{prox}_{\psi_0}(x;\tau)$ for any $\tau > 0$ (cf. Lemma B.3).

We now compare Theorem 3.12 to the risk results in [24]. The most significant advantage of Theorem 3.12 rests in its much weaker condition on the loss function ψ_0. In particular, [24] requires strong regularity assumptions on ψ_0 (e.g., ψ_0'' is required to be Lipschitz), which exclude the two canonical examples in Example 2.2 in robust regression that are covered by our theory. In addition, the $6+\delta$ moment assumption on the design matrix is also much weaker in Theorem 3.12 compared to the exponential moments required in [24].

An interesting question is whether the universality results in Theorem 3.12 hold for the unregularized case $\lambda = 0$ when $\tau_0 > 1$. The only result in this direction appears to be [23], Section 6 (some heuristics are presented in [45]), where strong convexity of ψ_0 and local Lipschitzness of ψ_0'' are required; see also related results in [20] under Gaussian designs. Under these strong assumptions on ψ_0, it is possible to establish ℓ_∞ bounds for $\widehat{w}_A^{\mathsf{M}}$ using similar techniques as in Proposition 3.11 and, therefore, the risk universality in Theorem 3.12. It however remains an open question to establish such ℓ_∞ bounds under the weak conditions on ψ_0 as in Theorem 3.12. To the best knowledge of the authors, this problem remains open even in the Gaussian design case.

3.5. *Nonuniversality for general isotropic designs.* Consider the regression model (1.1) with $m/n > 1$ and the ordinary least squares estimator (LSE):

$$\widehat{\mu}_A^{\mathsf{LSE}} \equiv \mathop{\arg\min}_{\mu\in\mathbb{R}^n} \|Y - A\mu\|^2 = \big(A^\top A\big)^{-1} A^\top Y.$$

Suppose that the error vector satisfies $\xi \sim \mathcal{N}(0, I_m)$ for simplicity. Define the sample covariance matrix $\widehat{\Sigma}_a \equiv m^{-1} A^\top A = m^{-1}\sum_{i=1}^m a_i a_i^\top \in \mathbb{R}^{n\times n}$, where $\{a_i\}_{i=1}^m$ are the rows of the design matrix A. Since the squared ℓ_2 risk of $\widehat{\mu}_A^{\mathsf{LSE}}$ can be written as a linear spectral statistic of $\widehat{\Sigma}_a$, it is easy to prove risk universality of $\widehat{\mu}_A^{\mathsf{LSE}}$ for design matrices A with independent entries satisfying Assumption III, by either using results in this paper or directly resorting to random matrix theory. On the other hand, it is a well-known fact in random matrix theory that the spectrum of $\widehat{\Sigma}_a$ does not exhibit universality for A with i.i.d. isotropic rows (see discussion below), thereby also negating the risk universality of $\widehat{\mu}_A^{\mathsf{LSE}}$ for this class of designs. The following proposition provides a simple explicit counterexample.

PROPOSITION 3.13. *Fix $m, n \in \mathbb{N}$ with $m \geq 2$, $m > n$ and $L_n > 1$. There exists some centered random vector $b_0 \in \mathbb{R}^n$ with $\mathbb{E}\, b_0^{\otimes 2} = \mathbb{E}(\mathcal{N}(0, I_n))^{\otimes 2}$ such that the following hold: With $B_0 \in \mathbb{R}^{m\times n}$ denoting a random matrix whose rows are i.i.d. as b_0, and $B \equiv B_0/\sqrt{m}$, we have $n^{-1}\,\mathbb{E}\|\widehat{\mu}_B^{\mathsf{LSE}} - \mu_0\|^2 \geq L_n \cdot n^{-1}\,\mathbb{E}\|\widehat{\mu}_G^{\mathsf{LSE}} - \mu_0\|^2$.*

PROOF. Take $L_n > 1$. Let U be a discrete distribution supported on $\{\pm L_n^{-1}, \pm S_n\}$ with $2\,\mathbb{P}(U = \pm L_n^{-1}) = 1 - 1/m$ and $2\,\mathbb{P}(U = \pm S_n) = 1/m$. Here, $S_n > 0$ is determined by the condition $\mathbb{E}\,U^2 = 1$; in fact, some simple calculation shows that $S_n = \{m(1 - L_n^{-2}(1 - 1/m))\}^{1/2}$. Now let $b_0 \equiv U \cdot Z$, where $Z \sim \mathcal{N}(0, I_n)$ is independent of U. Then by construction $\mathbb{E}\, b_0^{\otimes 2} = \mathbb{E}(\mathcal{N}(0, I_n))^{\otimes 2}$. With U_i's and Z_i's being independent copies of U and Z, We have

$$\begin{aligned} n^{-1}\,\mathbb{E}\big\|\widehat{\mu}_B^{\mathsf{LSE}} - \mu_0\big\|^2 &= n^{-1}\,\mathbb{E}\,\mathrm{tr}\big(\big(B^\top B\big)^{-1}\big) \\ &\geq n^{-1}\,\mathbb{E}\,\mathrm{tr}\left[\left(\frac{1}{m}\sum_{i=1}^m U_i^2 Z_i Z_i^\top\right)^{-1}\right]\mathbf{1}_{U_i^2 = L_n^{-2}, \forall i\in[m]} \end{aligned}$$

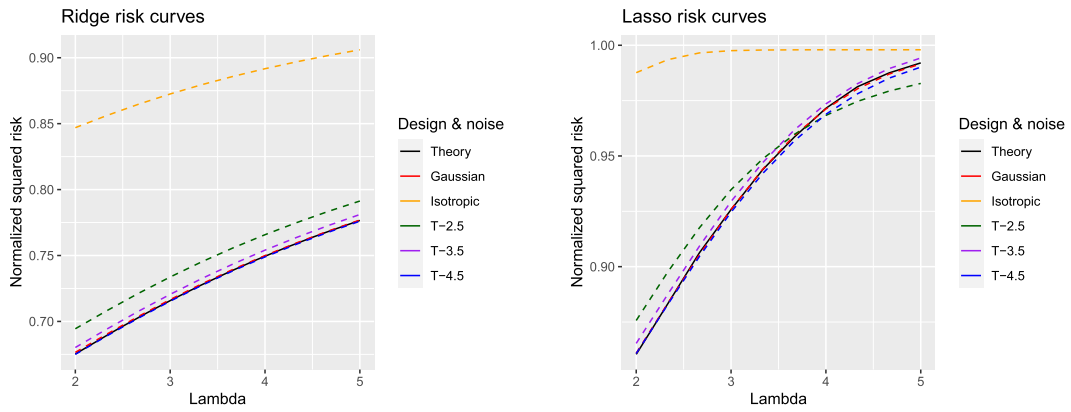
$$\geq \left(1-\frac{1}{m}\right)^m \cdot L_n^2 \cdot \mathbb{E}\,\mathrm{tr}\left[\left(\frac{1}{m}\sum_{i=1}^m Z_i Z_i^\top\right)^{-1}\right]$$
$$\geq e^{-2} L_n^2 \cdot n^{-1}\,\mathbb{E}\|\widehat{\mu}_G^{\mathsf{LSE}} - \mu_0\|^2,$$

where the last inequality used the simple fact that $(1-1/m)^m \geq e^{-2}$ for $m \geq 2$. Now the claim follows by adjusting constants. □

While the ℓ_2 risk in the above counterexample only serves as a special case of "structure properties" studied in previous sections, we conjecture that universality fails for a large class of structure properties associated with regularized regression estimators, when the spectrum of the design matrix differs significantly from that of the standard Gaussian design. Some positive results along this conceptual line are obtained in [22], proved using a completely different AMP method developed in [21]. However, it remains a wildly open question to provide an exact relation between the spectrum of the design matrix (with "generically positioned" singular vectors), and the behavior of a general structure property of a regularized regression estimator. A good and fairly nontrivial test case in this regard would be the dof adjusted debiased Lasso studied in Section 3.3.

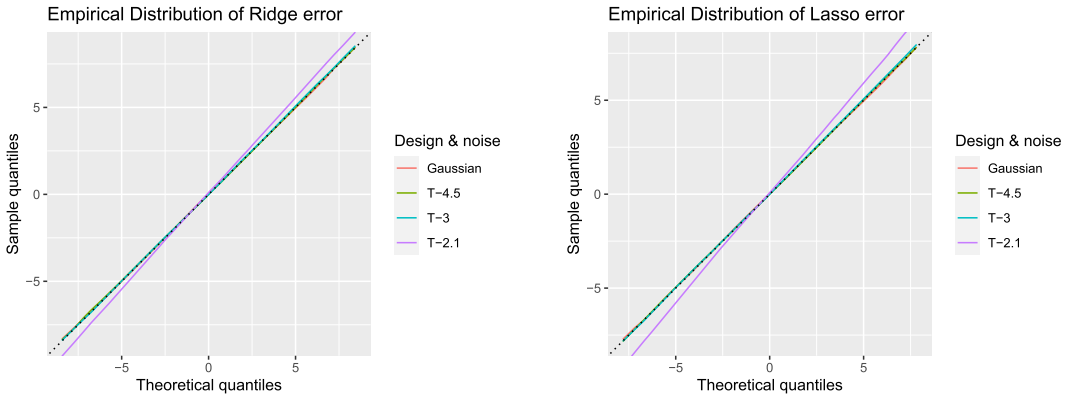
3.6. *Some illustrative simulation.* We perform a small scale simulation here to illustrate the (non)universality results proved in the previous subsections.

First, we examine (non)universality of risk asymptotics under different distributions of (A, ξ). As can be seen from Figure 1, for both Ridge and Lasso estimators, universality of risk asymptotics holds for $(\sqrt{m}A, \xi)$ with i.i.d. entries from a t distribution with only 4.5 dof, and then gradually breaks down when the dof approaches 3.5. It seems reasonable to conjecture that a phase transition near $t(4)$ occurs for the risk universality for both Ridge and Lasso estimators. On the other hand, under the setup of (nonuniversality Section 3.5) with a simple three-point delta prior, we see a matching second moment of the design does not guarantee universality.



(a) Ridge risk curves (b) Lasso risk curves

FIG. 1. *The black solid line marks the theoretical risk given by* w_*^{L} *and* w_*^{R}. *Simulation parameters*: $m = 1200$, $n = 1500$, $\mu_0 \in \mathbb{R}^n$ *are i.i.d.* $\mathcal{N}(0,1)$, $(\sqrt{m}A, \xi)$ *have i.i.d. entries following* $\mathcal{N}(0,1)$ (*red*), t *distribution with df* 4.5 (*blue*), *df* 3.5 (*purple*), *df* 2.5 (*green*), *with proper normalization in the latter two cases so that the variance is* 1. *In the isotropic case* (*orange*), ξ *has i.i.d.* $\mathcal{N}(0,1)$ *entries and* $(\sqrt{m}A)_{i,\cdot} = Z_i U_i$ *with* $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_n)$ *and* $U_i \overset{\text{i.i.d.}}{\sim} 0.25\delta_{\sqrt{2}} + 0.25\delta_{-\sqrt{2}} + 0.5\delta_0$. *The empirical risk curves are averaged over* 50 *replications*.



(a) QQ plot of Ridge errors

(b) QQ plot of Lasso errors

FIG. 2. *Comparison of the empirical quantiles of the error with theoretical quantiles given by* w_*^{L} *and* w_*^{R}. *Simulation parameters*: $m = 1200$, $n = 1500$, $\mu_0 \in \mathbb{R}^n$ *are i.i.d.* $\mathcal{N}(0, 5^2)$, $(\sqrt{m}A, \xi)$ *have i.i.d. entries following* $\mathcal{N}(0, 1)$ (*red*), t *distribution with df* 4.5 (*green*), *df* 3 (*blue*), *df* 2.1 (*purple*), *with proper normalization in the latter two cases so that the variance is* 1. *The empirical quantiles are averaged over* 50 *replications*.

Next, we examine the distributional universality proved for Ridge and Lasso estimators in Theorems 3.4 and 3.8. By the QQ plots in Figure 2, we see that such closeness holds all the way down to the very heavy-tailed situation where $(\sqrt{m}A, \xi)$ have i.i.d. entries following a t-distribution with only about 3 dof. Here, the simulation setup is similar to that used in Figure 1 with the exception that the variance level of μ_0 is enlarged to ensure a visible difference in the QQ plots.

Acknowledgments. The authors are indebted to Cun-Hui Zhang for a number of stimulating discussions during various stages of this research. The authors also thank three referees, an Associate Editor and the Editor for helpful comments and suggestions that significantly improved the quality of the paper.

Funding. The research of Q. Han is partially supported by NSF Grants DMS-1916221 and DMS-2143468.

SUPPLEMENTARY MATERIAL

Supplement: Proofs (DOI: 10.1214/23-AOS2309SUPP; .pdf). In the supplement, we provide proofs for the results in this paper.

REFERENCES

[1] ABBASI, E., SALEHI, F. and HASSIBI, B. (2019). Universality in learning from linear measurements. *Adv. Neural Inf. Process. Syst.* **32**.

[2] BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. USA* **116** 5451–5460. MR3939767 https://doi.org/10.1073/pnas.1802705116

[3] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[4] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[5] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[6] BELLEC, P. C. (2022). Observable adjustments in single-index models for regularized M-estimators. arXiv preprint. Available at arXiv:2204.06990.
[7] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory* **178** 1912–1947. PMLR.
[8] BELLEC, P. C. and ZHANG, C.-H. (2021). Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *Ann. Statist.* **49** 1864–1903. MR4319234 https://doi.org/10.1214/20-aos2005
[9] BELLEC, P. C. and ZHANG, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli* **28** 713–743. MR4389062 https://doi.org/10.3150/21-BEJ1348
[10] BELLEC, P. C. and ZHANG, C.-H. (2023). Debiasing convex regularized estimators and interval estimation in linear models. *Ann. Statist.* **51** 391–436. MR4600987 https://doi.org/10.1214/22-aos2243
[11] CELENTANO, M. and MONTANARI, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Statist.* **50** 170–196. MR4382013 https://doi.org/10.1214/21-aos2100
[12] CELENTANO, M., MONTANARI, A. and WEI, Y. (2022). The lasso with general gaussian designs with applications to hypothesis testing. arXiv preprint. Available at arXiv:2007.13716v2.
[13] CHATTERJEE, S. (2006). A generalization of the Lindeberg principle. *Ann. Probab.* **34** 2061–2076. MR2294976 https://doi.org/10.1214/009117906000000575
[14] CHEN, W.-K. and LAM, W.-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.* **26** Paper No. 36, 44. MR4235487 https://doi.org/10.1214/21-EJP604
[15] CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. and KOIKE, Y. (2023). High-dimensional data bootstrap. *Annu. Rev. Stat. Appl.* **10** 427–449. MR4567800 https://doi.org/10.1146/annurev-statistics-040120-022239
[16] CHERNOZHUKOV, V., CHETVERIKOV, D. and KOIKE, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Ann. Appl. Probab.* **33** 2374–2425. MR4583674 https://doi.org/10.1214/22-aap1870
[17] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Inf. Inference* **11** 435–495. MR4474343 https://doi.org/10.1093/imaiai/iaab002
[18] DICKER, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37. MR3449775 https://doi.org/10.3150/14-BEJ609
[19] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279. MR3766952 https://doi.org/10.1214/17-AOS1549
[20] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[21] DUDEJA, R., LU, Y. M. and SEN, S. (2022). Universality of approximate message passing with semi-random matrices. arXiv preprint. Available at arXiv:2204.04281.
[22] DUDEJA, R., SEN, S. and LU, Y. M. (2022). Spectral universality of regularized linear regression with nearly deterministic sensing matrices. arXiv preprint. Available at arXiv:2208.02753.
[23] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. arXiv preprint. Available at arXiv:1311.2445.
[24] EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9
[25] FAN, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Ann. Statist.* **50** 197–224. MR4382014 https://doi.org/10.1214/21-aos2101
[26] FANG, X. and KOIKE, Y. (2021). High-dimensional central limit theorems by Stein's method. *Ann. Appl. Probab.* **31** 1660–1686. MR4312842 https://doi.org/10.1214/20-aap1629
[27] GERACE, F., KRZAKALA, F., LOUREIRO, B., STEPHAN, L. and ZDEBOROVÁ, L. (2022). Gaussian universality of linear classifiers with random labels in high-dimension. arXiv preprint. Available at arXiv:2205.13303.
[28] GERACE, F., LOUREIRO, B., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning* **119** 3452–3462. PMLR.
[29] GERBELOT, C., ABBARA, A. and KRZAKALA, F. (2020). Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices. In *Conference on Learning Theory* **125** 1682–1713 PMLR.
[30] GERBELOT, C., ABBARA, A. and KRZAKALA, F. (2023). Asymptotic errors for teacher–student convex generalized linear models (or: How to prove Kabashima's replica formula). *IEEE Trans. Inf. Theory* **69** 1824–1852. MR4564683 https://doi.org/10.1109/tit.2022.3222913

[31] GOLDT, S., LOUREIRO, B., REEVES, G., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2022). The Gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning* **145** 426–471. PMLR.

[32] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric Aspects of Functional Analysis* (1986/87). *Lecture Notes in Math*. **1317** 84–106. Springer, Berlin. MR0950977 https://doi.org/10.1007/BFb0081737

[33] HAN, Q. (2022). Noisy linear inverse problems under convex constraints: Exact risk asymptotics in high dimensions. arXiv preprint. Available at arXiv:2201.08435.

[34] HAN, Q. and SHEN, Y. (2023). Supplement to "Universality of regularized regression estimators in high dimensions." https://doi.org/10.1214/23-AOS2309SUPP

[35] HAN, Q. and XU, X. (2023). The distribution of ridgeless least squares interpolators. arXiv preprint. Available at arXiv:2307.02044.

[36] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133

[37] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[38] HU, H. and LU, Y. M. (2019). Asymptotics and optimal designs of slope for sparse linear regression. In 2019 *IEEE International Symposium on Information Theory* (*ISIT*) 375–379. IEEE, Los Alamitos.

[39] HU, H. and LU, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theory* **69** 1932–1964. MR4564688

[40] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

[41] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

[42] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[43] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629

[44] JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630

[45] KAROUI, N. E., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562. https://doi.org/10.1073/pnas.1307842110

[46] KORADA, S. B. and MONTANARI, A. (2011). Applications of the Lindeberg principle in communications and statistical learning. *IEEE Trans. Inf. Theory* **57** 2440–2450. MR2809100 https://doi.org/10.1109/TIT.2011.2112231

[47] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[48] LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-$\ell 1$-norm interpolated classifiers. *Ann. Statist.* **50** 1669–1695. MR4441136 https://doi.org/10.1214/22-aos2170

[49] LINDEBERG, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211–225. MR1544569 https://doi.org/10.1007/BF01494395

[50] LOPES, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions: Near $1/\sqrt{n}$ rates via implicit smoothing. *Ann. Statist.* **50** 2492–2513. MR4505371 https://doi.org/10.1214/22-aos2184

[51] LOUREIRO, B., GERBELOT, C., CUI, H., GOLDT, S., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems* **34** 18137–18151.

[52] MEI, S. and MONTANARI, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.* **75** 667–766. MR4400901 https://doi.org/10.1002/cpa.22008

[53] MÉZARD, M., PARISI, G. and VIRASORO, M. A. (1987). *Spin Glass Theory and Beyond*. *World Scientific Lecture Notes in Physics* **9**. World Scientific Co., Inc., Teaneck, NJ. MR1026102

[54] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[55] MONTANARI, A. (2018). Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro* 2018. *Invited Lectures* **4** 2973–2994. World Sci. Publ., Hackensack, NJ. MR3966519

[56] MONTANARI, A. and NGUYEN, P.-M. (2017). Universality of the elastic net error. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 2338–2342. IEEE, Los Alamitos.

[57] MONTANARI, A., RUAN, F., SOHN, Y. and YAN, J. (2023). The generalization error of max-margin linear classifiers: Benign overfitting and high-dimensional asymptotics in the overparametrized regime. arXiv preprint. Available at arXiv:1911.01544v3.

[58] MONTANARI, A. and SAEED, B. (2022). Universality of empirical risk minimization. In *Conference on Learning Theory* **178** 4310–4312. PMLR.

[59] OYMAK, S. and TROPP, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Inf. Inference* **7** 337–446. MR3858331 https://doi.org/10.1093/imaiai/iax011

[60] PANAHI, A. and HASSIBI, B. (2017). A universal analysis of large-scale regularized least squares solutions. *Adv. Neural Inf. Process. Syst.* **30**.

[61] SALEHI, F., ABBASI, E. and HASSIBI, B. (2019). The impact of regularization on high-dimensional logistic regression. *Adv. Neural Inf. Process. Syst.* **32**.

[62] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[63] STOJANOVIC, S., DONHAUSER, K. and YANG, F. (2022). Tight bounds for maximum ℓ_1-margin classifiers. arXiv preprint. Available at arXiv:2212.03783.

[64] STOJNIC, M. (2013). A framework to characterize performance of lasso algorithms. arXiv preprint. Available at arXiv:1303.7291.

[65] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[66] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses, Vol. I: Basic Examples. Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge. A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series. A Series of Modern Surveys in Mathematics*] **54**. Springer, Berlin. MR2731561 https://doi.org/10.1007/978-3-642-15202-3

[67] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses, Vol. II: Advanced Replica-Symmetry and Low Temperature. Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge. A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas*. 3*rd Series. A Series of Modern Surveys in Mathematics*] **55**. Springer, Heidelberg. MR3024566

[68] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720

[69] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* **40** 1683–1709. PMLR.

[70] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[71] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[72] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[73] WANG, S., WENG, H. and MALEKI, A. (2022). Does SLOPE outperform bridge regression? *Inf. Inference* **11** 1–54. MR4409197 https://doi.org/10.1093/imaiai/iaab025

[74] WANG, T., ZHONG, X. and FAN, Z. (2022). Universality of approximate message passing algorithms and tensor networks. arXiv preprint. Available at arXiv:2206.13037.

[75] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[76] ZHANG, X., ZHOU, H. and YE, H. (2022). A modern theory for high-dimensional Cox regression models. arXiv preprint. Available at arXiv:2204.01161.

[77] ZHOU, L., KOEHLER, F., SUR, P., SUTHERLAND, D. J. and SREBRO, N. (2022). A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. arXiv preprint. Available at arXiv:2210.12082.