The Annals of Statistics
2021, Vol. 49, No. 4, 2313–2335
https://doi.org/10.1214/20-AOS2038

THE DISTRIBUTION OF THE LASSO: UNIFORM CONTROL OVER SPARSE BALLS AND ADAPTIVE PARAMETER TUNING

BY LÉO MIOLANE[1] AND ANDREA MONTANARI[2]

[1]*Courant Institute of Mathematical Sciences and Center for Data Science, New York University, leo.miolane@gmail.com*
[2]*Department of Statistics and Department of Electrical Engineering, Stanford University, montanar@stanford.edu*

The Lasso is a popular regression method for high-dimensional problems in which the number of parameters $\theta_1, \dots, \theta_N$, is larger than the number n of samples: $N > n$. A useful heuristics relates the statistical properties of the Lasso estimator to that of a simple soft-thresholding denoiser, in a denoising problem in which the parameters $(\theta_i)_{i \le N}$ are observed in Gaussian noise, with a carefully tuned variance. Earlier work confirmed this picture in the limit $n, N \to \infty$, pointwise in the parameters θ and in the value of the regularization parameter.

Here, we consider a standard random design model and prove exponential concentration of its empirical distribution around the prediction provided by the Gaussian denoising model. Crucially, our results are uniform with respect to θ belonging to ℓ_q balls, $q \in [0, 1]$, and with respect to the regularization parameter. This allows us to derive sharp results for the performances of various data-driven procedures to tune the regularization.

Our proofs make use of Gaussian comparison inequalities, and in particular of a version of Gordon's minimax theorem developed by Thrampoulidis, Oymak and Hassibi, which controls the optimum value of the Lasso optimization problem. Crucially, we prove a stability property of the minimizer in Wasserstein distance that allows one to characterize properties of the minimizer itself.

1. Introduction. Given data (x_i, y_i), $1 \le i \le n$, with $x_i \in \mathbb{R}^N$, $y_i \in \mathbb{R}$, the Lasso [15, 48] fits a linear model by minimizing the cost function

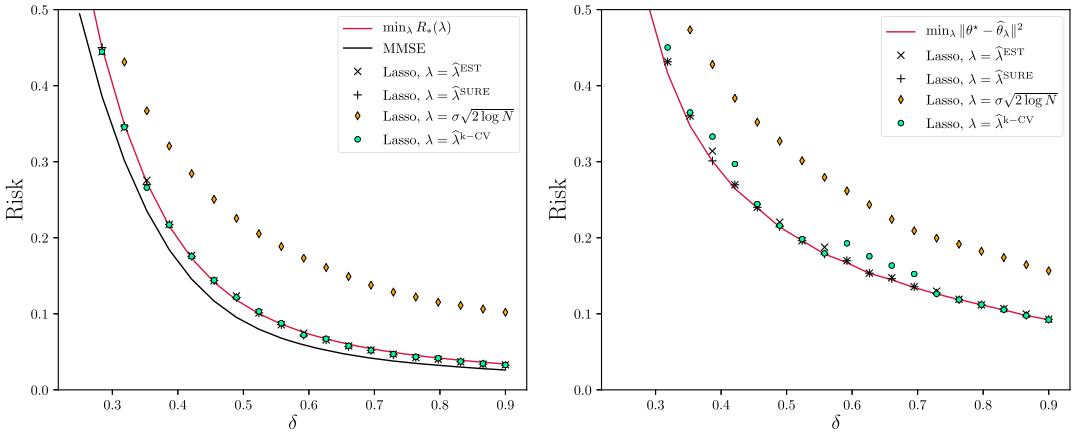
$$\begin{aligned} \mathcal{L}_\lambda(\theta) &= \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2 + \frac{\lambda}{\sqrt{n}} |\theta| \\ &= \frac{1}{2n} \|y - X\theta\|^2 + \frac{\lambda}{\sqrt{n}} |\theta|. \end{aligned} \tag{1.1}$$

Here, $X \in \mathbb{R}^{n \times N}$ is the matrix with rows $x_1, \dots, x_n$, $y = (y_1, \dots, y_n)$, $\|v\|$ denotes the ℓ_2 norm of vector v and $|v|$ its ℓ_1 norm.

A large body of theoretical work supports the use of ℓ_1 regularization in the high-dimensional regime $n \lesssim N$, when only a small subset of the coefficients θ are expected to be large. Broadly speaking, we can distinguish two types of theoretical approaches. A first line of work makes deterministic assumptions about the design matrix X, such as the restricted isometry property and its generalizations [10, 13]. Under such conditions, minimax optimal estimation rates as well as oracle inequalities have been proved in a remarkable sequence of papers [9, 12, 35, 38, 52]. As an example, assume that that the linear model is correct. Namely,

$$y = X\theta^\star + \sigma z, \tag{1.2}$$

Received July 2020; revised November 2020.
MSC2020 subject classifications. 62J05, 62J07.
Key words and phrases. Linear regression, sparsity, lasso, cross-validation.



(a) Independent columns (b) Correlated columns

FIG. 1. *Estimation risk of the Lasso for different choices of* λ, *as a function of* δ. $N = 5000$. *In both plots*, $\sigma = 0.2$. *The true coefficients vector* $\theta^\star$ *is chosen to be* sN*-sparse with* $s = 0.2$. *The entries on the support of* $\theta^\star$ *are drawn i.i.d.* $\mathcal{N}(0, 1/n)$. *Cross-validation is carried out using* 4 *folds. SURE is computed using the estimator* $\widehat{\sigma}$ *for the plot on the left, and the true value of* σ *on the right. Left*: *A standard random design with* $(X_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. *Right*: *The rows of the design matrix* X *are i.i.d. Gaussian, with correlation structure given by an autoregressive process*; *see equation* (4.6). *Here, we used* $\phi = 2$.

for $\sigma > 0$, $z \sim \mathcal{N}(0, \mathrm{I}_n)$, and $\theta^\star$ a vector with k_0 nonzero entries. Then a theorem of Bickel, Ritov and Tsybakov [9] implies that, with high probability,

$$\lambda = c_0 \sigma \sqrt{\log N} \quad \Rightarrow \quad \|\widehat{\theta}_\lambda - \theta^\star\|^2 \le \frac{C k_0 \sigma^2}{n} \log N, \tag{1.3}$$

for some constants c_0, C that depend on the specific assumptions on the design.

Unfortunately, this analysis provides limited insight into the choice of the regularization parameter λ which—in practice—can impact significantly the estimation accuracy. As an example, Figure 1 reports the result of a small simulation in which we compare four different methods of selecting λ. The bound of equation (1.3) suggests to set $\lambda = c_0 \sigma \sqrt{\log N}$. For the standard random design used in the left frame, the optimal constant is expected to be $c_0 = \sqrt{2}$ [19, 21]. We compare this method to three procedures that adapt the choice of λ to the data: (i) cross-validation (CV), which splits the data in $k = 4$ folds and fits a model over 3 of the folds choosing λ as to minimize the prediction error over the 4-th fold; (ii) Stein's Unbiased Risk Estimate (SURE): we will prove in Section 4.1 that SURE provides a consistent estimator $\widehat{P}^{\text{SURE}}(\lambda)$ of the prediction error $\frac{1}{n}\|X(\widehat{\theta}_\lambda - \theta^\star\|_2^2 + \sigma^2$. We set λ by as to minimize $\widehat{P}^{\text{SURE}}(\lambda)$; (iii) a new procedure (EST) that is based on minimizing over λ a consistent estimate of the ℓ_2 error $\|\widehat{\theta}_\lambda - \theta^\star\|^2$, which we denote by $\widehat{\tau}(\lambda)$. We refer to Section 4.1 for a description of this method and consistency results. We compare the estimation error of these methods with the predicted asymptotics for the oracle risk $\min_\lambda \|\widehat{\theta}_\lambda - \theta^*\|^2$ developed in Section 3: the agreement is excellent already at moderate sizes.

Note that all of these adaptive procedures significantly outperform the "theory driven" λ: over a broad range of sample sizes n, the resulting estimation error is 2 to 3 times smaller. While the choice $\lambda \asymp \sqrt{\log N}$ retains a useful role for asymptotic guidance, it is also important to develop a theory for adaptive choices. (We refer to Sections 4.2 and 4.3 for further discussion.)

In the simulation of Figure 1, we generate the true parameter θ^* with i.i.d. coefficients $\theta_i^* \sim P_0 := (1 - s)\delta_0 + s\mathcal{N}(0, 1/n)$. The scale of the nonzeros is chosen of the same order as the noise level on each of them, which is $\sigma/\sqrt{n}$. The ultimate lower bound on

the mean square error of any statistical procedure is given by the error of the posterior mean $\widehat{\theta}^{\mathrm{Bayes}}(y, X) := \mathbb{E}\{\theta^*|y, X\}$ (with the prior given by $P_0^{\otimes N}$). While—in general—we cannot compute $\widehat{\theta}^{\mathrm{Bayes}}(y, X)$ efficiently, the Bayes mean square error $\mathrm{MMSE}_N := \mathbb{E}\{\|\widehat{\theta}^{\mathrm{Bayes}}(y, X) - \theta^*\|^2\}$ is known to converge in the proportional asymptotic $n, N \to \infty$, $n/N \to \delta$, namely $\mathrm{MMSE}_N \to \mathrm{MMSE}$ in this limit. An explicit formula for the limiting Bayes error MMSE was proved in [2, 39] and is also plotted in Figure 1. We refer to Section 4.3 for further details on this prediction. Remarkably, the error achieved by the three adaptive methods for selecting λ is very close to the Bayes error.

These empirical observations are not captured by the bound (1.3), or by similar results. An alternative style of analysis postulates an idealized model for the data and derives asymptotically exact results. This type of analysis was first carried out in the context of the Lasso in [6] and then extended to a number of other problems; see, for example, [17, 24, 25, 44, 46, 47].

We develop our theory in the case of uncorrelated covariates $x_i \sim \mathcal{N}(0, \mathrm{I}_N)$, which is also the setting of Figure 1, left frame. Figure 1 reports the predictions of our theory for the risk of the three adaptive procedure for selecting λ. The agreement with the numerical simulations is excellent. It is natural to wonder whether the insights developed in this case might apply to general correlation structures $x_i \sim \mathcal{N}(0, \Sigma_N)$. In the right frame, we consider the case of a nonsingular covariance $\Sigma_N \neq \mathrm{I}_N$ corresponding to the correlation structure of an autoregressive process. The qualitative picture in this case is very similar to the one obtained for uncorrelated designs: data adaptive methods outperform the standard choice $\lambda = c_0\sigma\sqrt{\log N}$. We refer to Section 4.3 for further simulations with correlated designs supporting this point: methods for selecting λ developed with uncorrelated designs seem to perform well more generally.

Finally, while assumption $x_i \sim \mathcal{N}(0, \mathrm{I}_N)$, is likely to be violated in practice, we believe that the general mathematical approach developed here can be used to attack the general case as well.

Unfortunately, the results in [6] (and in follow-up work) do not allow one to derive in a mathematically rigorous way curves such as the ones in Figure 1. In fact, earlier results hold "pointwise" over λ, and hence do not apply to adaptive procedures to select λ. Further, they provide asymptotic estimates "pointwise" over θ, and hence do not allow one to compute, for instance, minimax risk.

In order to clarify these points, it is useful to overview informally the picture emerging from [6, 20]. Fix $\theta \in \mathbb{R}^N$, $\lambda \in \mathbb{R}_{>0}$, and let $\eta(x; b) = (|x| - b)_+ \operatorname{sign}(x)$ be the soft thresholding function. By the KKT conditions, the Lasso estimator $\widehat{\theta}_\lambda$ satisfies

$$\widehat{\theta}_\lambda = \eta\big(\widehat{\theta}_\lambda^d; \alpha\tau/\sqrt{n}\big), \quad \widehat{\theta}_\lambda^d = \widehat{\theta}_\lambda + \frac{\alpha\tau}{\lambda n} X^{\mathsf{T}}(y - X\widehat{\theta}_\lambda), \tag{1.4}$$

where the vector $\widehat{\theta}_\lambda^d$ is also referred to as the "debiased Lasso" [28, 51, 54]. The above identity holds for arbitrary $\alpha, \tau > 0$. However, [6] predicts that the distribution of the debiased estimator $\widehat{\theta}_\lambda^d$ simplifies dramatically for specific choices of these parameters.

Namely, let Θ be a random variable with distribution given by the empirical distribution of $(\theta_i^*)_{i\le N}$ (i.e., $\Theta = \theta_i^*$ with probability $1/N$, for $i \in \{1, \dots, N\}$) and let $Z \sim \mathcal{N}(0, 1)$ be independent of Θ. Define α_*, τ_* to be the solution of the following system of equations (we refer to Section 3.1 for a discussion of existence and uniqueness):

$$\begin{cases} \tau^2 = \sigma^2 + \dfrac{1}{\delta}\mathbb{E}\big[\big(\eta(\sqrt{n}\Theta + \tau Z, \alpha\tau) - \sqrt{n}\Theta\big)^2\big], \\ \lambda = \alpha\tau\Big(1 - \dfrac{1}{\delta}\mathbb{P}\big(|\sqrt{n}\Theta + \tau Z| > \alpha\tau\big)\Big). \end{cases} \tag{1.5}$$

When α, τ are selected in this way, $\widehat{\theta}_\lambda^d$ is approximately normal with mean $\theta^\star$ (the true parameters vector) and variance τ_*^2/n: $\widehat{\theta}^d \approx \mathcal{N}(\theta^\star, \tau_*^2\mathrm{I}/n)$. More precisely, for any function

$\mathsf{f}:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$, with $|\mathsf{f}(x)-\mathsf{f}(y)|\le L(1+\|x\|+\|y\|)\|x-y\|$ (note that for $x=(x_1,x_2)\in\mathbb{R}\times\mathbb{R}$, $\|x\|\equiv\sqrt{x_1^2+x_2^2}$), we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathsf{f}\big(\sqrt{n}\theta_i^\star,\sqrt{n}\widehat{\theta}^d_{\lambda,i}\big)=\mathbb{E}\big\{\mathsf{f}\big(\sqrt{n}\Theta,\sqrt{n}\Theta+\tau_*Z\big)\big\}+o_{\mathbb{P}}(1), \tag{1.6}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathsf{f}\big(\sqrt{n}\theta_i^\star,\sqrt{n}\widehat{\theta}_{\lambda,i}\big)=\mathbb{E}\big\{\mathsf{f}\big(\sqrt{n}\Theta,\eta\big(\sqrt{n}\Theta+\tau_*Z;\alpha_*\tau_*\big)\big)\big\}+o_{\mathbb{P}}(1), \tag{1.7}$$

where $o_{\mathbb{P}}(1)$ denotes a quantity going to zero in probability as $N,n\to\infty$, while $n/N\to\delta$. This is an asymptotic result, which holds along sequences of problems with: (i) converging aspect ratio $n/N\to\delta\in(0,\infty)$; (ii) fixed regularization $\lambda\in(0,\infty)$; (iii) parameter vectors $\sqrt{n}\theta^\star=\sqrt{n}\theta^\star(n)$ whose empirical distribution converges (weakly) to a limit law $p_{\bar\Theta}$ (equivalently $\sqrt{n}\Theta$ converges in distribution to $\bar\Theta$). As emphasized above, this does not allow one to deduce the behavior of the Lasso with adaptive choices of λ (there could be deviations from the above limits for exceptional values of λ), or to compute the minimax risk (there could be deviations for exceptional vectors $\theta^\star$).

REMARK 1.1. Notice that the arguments of f in equations (1.6), (1.7) are scaled in such a way to probe the distribution of $\widehat{\theta}^d_{\lambda,i}$, and $\widehat{\theta}_{\lambda_i}$ on a scale of the same order as the noise level, that is, $1/\sqrt{n}$. As shown from the right-hand side, the resulting distribution is nontrivial when Θ is of order $1/\sqrt{n}$. In an asymptotic setting, this corresponds to taking $\sqrt{n}\Theta$ that converges in distribution. In this paper, we will obtain nonasymptotic results and provide explicit conditions at finite n,N.

The importance of establishing uniform convergence with respect to the regularization parameter λ was recently emphasized by Mousavi, Maleki and Baraniuk [34]. Among other results, these authors derive a uniform convergence statement for the related approximate message passing (AMP) algorithm. However, in order to establish uniform convergence, they have to construct an ad hoc smoothing of the quantity of interest, which is roughly equivalent to discretizing the corresponding tuning parameter.

In this paper, we obtain uniform (in λ) convergence results for the Lasso, hence providing a sound mathematical basis to the comparison of various adaptive procedures, as well as to the study of minimax risk. Further, we establish explicit nonasymptotic bounds that hold at finite n,N, without requiring assumptions about the asymptotic behavior of the aspect ratio n/N, or on the empirical distribution of the entries of $\sqrt{n}\theta^*$.

The rest of the paper is organized as follows. Section 2 reviews related work. We state our main theoretical results in Section 3. In Section 4, we apply these results to two types of statistical questions: estimating the risk and noise level, and selecting λ through adaptive procedures. Further, we illustrate our results in numerical simulations. Finally, Section 5 outlines the main proof ideas, with most technical legwork deferred to the Appendices [32].

2. Related work. There is by now a substantial literature on determining exact asymptotics in high-dimensional statistical models, and a number of mathematical techniques have been developed for this task. We will only provide a few pointers focusing on high-dimensional regression problems.

The original proof of [6] was based on an asymptotically exact analysis of an approximate message passing (AMP) algorithm [5] that was first proposed in [20] to minimize the Lasso cost function. Variants of AMP have been developed in a number of contexts, opening the way to the analysis of various statistical estimation problems. A short list includes generalized

linear models [37], phase retrieval [31, 40], robust regression [17], logistic regression [44], generalized compressed sensing [8]. This approach is technically less direct than others, but has the advantage of providing an efficient algorithm, and is and not necessarily limited to convex problems (see [33] for a nonconvex example).

As mentioned above, our work was partially motivated by the recent results of Mousavi, Maleki and Baraniuk [34] that establish a form of uniformity for the AMP estimates but not for the Lasso solution. It would be interesting to understand whether the approach of [34] could also be used to obtain uniform results for the Lasso or other statistical estimators.

Here, we follow a different route that exploits powerful Gaussian comparison inequalities first proved by Gordon [26, 27]. Gordon inequality allows one to bound the distribution of a minimax value, that is, the value of a random variable $G_* = \min_{i\le N} \max_{j\le M} G_{ij}$, where $(G_{ij})_{i\le N, j\le M}$ is a Gaussian process, in terms of a similar quantity for a "simpler" Gaussian process. The use of Gordon's inequality in this context was pioneered by Stojnic [43] and then developed by a number of authors in the context of regularized regression [47], M-estimation [46], generalized compressed sensing [1], binary compressed sensing [42] and so on. The key idea is to write the optimization problem of interest as a minimax problem, and then apply a suitable version of Gordon's inequality. A matching bound is obtained by convex duality and then a second application of Gordon's inequality. In particular, convexity of the cost function of interest is a crucial ingredient.

While the Gaussian comparison inequality provides direct access to the value of the optimization problem, understanding the properties of the estimator can be more challenging. In this paper, we identify a property (that we call *local stability*) that allows one to transfer information on the minimum (the Lasso cost) into information about the minimizer (the Lasso estimator). We believe this strategy can be applied to other examples beyond the Lasso.

Independently, a different approach based on leave-one-out techniques was developed by El Karoui in the context of ridge-regularized robust regression [24, 25].

Finally, a parallel line of research determines exact asymptotics for Bayes optimal estimation, under a model in which the coordinates of $\sqrt{n}\theta$ are i.i.d. with common distribution p_Θ. In particular, the asymptotic Bayes optimal error for linear regression with random designs was recently determined in [2, 39]. Of course, in general, Bayes optimal estimation requires knowledge of the distribution p_Θ, and is not computationally efficient. We will use this Bayes-optimal error as a benchmark of our adaptive procedures, as we have already done in Figure 1. Generalizations of these results were also obtained in [3] for other regression problems. A successful approach to these models uses smart interpolation techniques that generalize ideas in spin-glass theory.

3. Main results.

3.1. *Definitions.* As stated above, we consider the standard linear model (1.2) where $y = X\theta^\star + \sigma z$, with noise $z \sim \mathcal{N}(0, \mathrm{I}_n)$, and X a Gaussian design: $(X_{i,j})_{i\le n, j\le N} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The Lasso estimator is defined by

$$\widehat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^N} \mathcal{L}_\lambda(\theta). \tag{3.1}$$

(The minimizer is almost surely unique since the columns of X are in generic positions.) We set $\delta = n/N$ to be the number of samples per dimension. We are interested in uniform estimation over sparse vectors $\theta^\star$. Following [19, 30], we formalize this notion using ℓ_p-balls (which are convex sets only for $p \ge 1$).

DEFINITION 3.1. Define for $p, \xi > 0$ the ℓ_p-ball

$$\mathcal{F}_p(\xi) = \{x \in \mathbb{R}^N \,|\, \|x_i\|_p^p \le N^{1-p/2}\xi^p\},$$

and for $s \in [0, 1]$

$$\mathcal{F}_0(s) = \{x \in \mathbb{R}^N \,|\, \|x\|_0 \le sN\}.$$

By Jensen's inequality, we have for $p \ge p' > 0$, $\mathcal{F}_p(\xi) \subset \mathcal{F}_{p'}(\xi)$.

Let $\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ be the standard Gaussian density and $\Phi(x) = \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t$ be the associated cumulative function. In the case of ℓ_0 balls (sparse vectors), a crucial role is played by the following sparsity level.

DEFINITION 3.2. Define the critical sparsity as follows for $\delta \in [0, 1]$:

$$s_{\max}(\delta) = \delta \max_{\alpha \ge 0} \left\{ \frac{1 - \frac{2}{\delta}((1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha))}{1 + \alpha^2 - 2((1+\alpha^2)\Phi(-\alpha) - \alpha\phi(\alpha))} \right\}.$$

For $\delta > 1$, choose $s_{\max}(\delta) > 1$ arbitrarily (in particular, for $\delta > 1$, $s < s_{\max}(\delta)$ for any $s \in [0, 1]$).

The critical sparsity curve first appears in the seminal work by Donoho and Tanner on compressed sensing [18, 22]. These authors consider the noiseless case ($z = 0$) of model (1.2) and reconstruction via ℓ_1 minimization (which corresponds to the $\lambda \to 0$ limit of the Lasso). They prove that ℓ_1 minimization reconstructs exactly $\theta^\star$ with high probability, if $\|\theta^\star\|_0 \le N(s_{\max}(\delta) - \varepsilon)$, and fails with high probability if $\|\theta^\star\|_0 \ge N(s_{\max}(\delta) + \varepsilon)$ (for any $\varepsilon > 0$). A second interpretation of the critical sparsity $s_{\max}(\delta)$ was given in [21, 47, 50]. For $\|\theta^\star\|_0 \le N(s_{\max}(\delta) - \varepsilon)$, the Lasso achieves stable reconstruction. Namely, there exists $M = M(s, \delta) < \infty$ for $s < s_{\max}(\delta)$, such that, if $\|\theta^\star\|_0 \le Ns$, then $\|\widehat{\theta}_\lambda - \theta^\star\|_2 \le M(s, \delta)\sigma^2$. Our results provide a third interpretation: For $\delta \in [0, 1]$, uniform limit laws for the Lasso will be obtained on ℓ_0 balls only for $s < s_{\max}(\delta)$.

The following max-min problem plays an important role in our results:

$$\max_{\beta \ge 0} \min_{\tau \ge \sigma} \psi_\lambda(\beta, \tau),$$

$$\psi_\lambda(\beta, \tau) \equiv \left(\frac{\sigma^2}{\tau} + \tau\right)\frac{\beta}{2} - \frac{1}{2}\beta^2 + \frac{1}{\delta}\mathbb{E}\min_{w \in \mathbb{R}}\left\{\frac{w^2}{2\tau}\beta - \beta Z w + \lambda|w + \sqrt{n}\Theta| - \lambda|\sqrt{n}\Theta|\right\}. \tag{3.2}$$

The expectation above is with respect to $(\Theta, Z) \sim \widehat{\mu}_{\theta^\star} \otimes \mathcal{N}(0, 1)$, where $\widehat{\mu}_{\theta^\star}$ denotes the empirical distribution of the entries of the vector $\theta^\star$:

$$\widehat{\mu}_{\theta^\star} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i^\star}.$$

PROPOSITION 3.1. *The max-min* (3.2) *is achieved at a unique couple* $(\beta_*(\lambda), \tau_*(\lambda))$. *Moreover,* $(\tau_*(\lambda), \beta_*(\lambda))$ *is also the unique couple* $(\beta, \tau) \in (0, +\infty)^2$ *that verify*

$$\begin{cases} \tau^2 = \sigma^2 + \dfrac{1}{\delta}\mathbb{E}\left[\left(\eta\left(\sqrt{n}\Theta + \tau Z, \tau\dfrac{\lambda}{\beta}\right) - \sqrt{n}\Theta\right)^2\right], \\ \beta = \tau\left(1 - \dfrac{1}{\delta}\mathbb{E}\left[\eta'\left(\sqrt{n}\Theta + \tau Z, \dfrac{\tau\lambda}{\beta}\right)\right]\right). \end{cases} \tag{3.3}$$

We will also use the notation $\alpha_*(\lambda) = \lambda/\beta_*(\lambda)$ *and*

$$s_*(\lambda) = \mathbb{E}\left[\eta'\left(\sqrt{n}\Theta + \tau_*(\lambda)Z, \tau_*(\lambda)\alpha_*(\lambda)\right)\right] = \mathbb{P}\left(\left|\sqrt{n}\Theta + \tau_*(\lambda)Z\right| \ge \alpha_*(\lambda)\tau_*(\lambda)\right). \tag{3.4}$$

We will sometimes omit the dependency on λ and write simply $\alpha_*, \beta_*, \tau_*, s_*$. The distribution μ_λ^* defined below will correspond (see Theorem 3.1 in the next section) to the limit of the empirical distribution of the entries of $(\widehat{\theta}_\lambda, \theta^\star)$.

DEFINITION 3.3. Let $(\Theta, Z) \sim \widehat{\mu}_{\theta^\star} \otimes \mathcal{N}(0,1)$. We denote by μ_λ^* the law of the couple $(\eta(\Theta + \tau_*(\lambda)Z/\sqrt{n}, \alpha_*(\lambda)\tau_*(\lambda)/\sqrt{n}), \Theta)$.

3.2. *Results.* We fix from now on $0 < \lambda_{\min} \le \lambda_{\max}$ and $\mathcal{D} \subset \mathbb{R}^N$ that can be either $\mathcal{F}_p(\xi)$ for some $\xi, p > 0$, or $\mathcal{F}_0(s)$ for some $s < s_{\max}(\delta)$. Our uniformity domain is defined by $\Omega = (\delta, \sigma, \mathcal{D}, \lambda_{\min}, \lambda_{\max})$. Namely, we will control $\widehat{\theta}_\lambda$ uniformly with respect to $\theta^\star \in \mathcal{D}$ and $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, with $n/N = \delta$. We will call *constant* any quantity that only depends on Ω. In absence of further specifications, C, c will be constants (that depend only on Ω) that are allowed to change from one line to another.

REMARK 3.1. In what follows, we control the behavior of the Lasso for λ in the bounded interval $[\lambda_{\min}, \lambda_{\max}]$ with $\lambda_{\min}, \lambda_{\max}$ of order 1 and bounded away from 0 as $N, n \to \infty$. This rules out the more classical prescription $\lambda = c_0\sigma\sqrt{\log N}$. We will show in Lemma 4.1 that any choice of the regularization such that $\lambda \to \infty$ as $N, n \to \infty$ is suboptimal in the present setting, and therefore there is no loss of generality in assuming $\lambda_{\min}, \lambda_{\max}$ bounded. Namely, for $\lambda = O(1)$ the ℓ_2-estimation error is uniformly bounded by a quantity of order one (In fact, we characterize precisely its limit in Theorem 3.2). In contrast, for $\lambda \to \infty$ there exists sequences of sparse vectors $\theta^\star \in \mathbb{R}^N$ such the risk $\|\widehat{\theta}_\lambda - \theta^\star\|$ diverges. We also notice that [16] points out that—empirically—the value of λ selected by cross-validation is often smaller than the one from classical prescriptions.

The assumption of $\lambda_{\min}$ bounded away from 0 is motivated by the need to control the solution of equation (3.3) uniformly over the law of Θ, and the given range of λ. The case $\lambda = 0$ is singular (in that case we are performing unregularized least squares), and therefore can lead to nonuniformity. While it might be possible to extend our results to $(0, \lambda_{\max}]$ under additional assumptions, we also expect that the optimal λ will be bounded away from 0 as long as $\sigma > 0$, so we regard $\lambda_{\min} > 0$ as a minor limitation.

Our first result shows that the empirical distribution of the entries $\{(\widehat{\theta}_{\lambda,i}, \theta_i^\star)\}_{i \le N}$ is uniformly close to the model μ_λ^*. We quantify deviations using the Wasserstein distance. Recall that, given two probability measures μ, ν on $\mathbb{R}^d$ with finite second moment, their Wasserstein distance of order 2 is

$$W_2(\mu,\nu) = \Big(\inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int \|x-y\|_2^2 \gamma(\mathrm{d}x, \mathrm{d}y)\Big)^{1/2}, \tag{3.5}$$

where the infimum is taken over all couplings of μ and ν. Note that W_2 metrizes the convergence in equation (1.7). Namely $\lim_{n\to\infty} W_2(\mu_n, \mu_*) = 0$ if and only if, for any function $\mathsf{f}: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, with $|\mathsf{f}(x) - \mathsf{f}(y)| \le L(1 + \|x\| + \|y\|)\|x - y\|$, we have $\lim_{n\to\infty} \int \mathsf{f}(x)\mu_n(\mathrm{d}x) = \int \mathsf{f}(x)\mu_*(\mathrm{d}x)$ [53]. It provides therefore a natural way to extend earlier results to a nonasymptotic regime.

THEOREM 3.1. *Assume that $\mathcal{D} = \mathcal{F}_p(\xi)$ for some $\xi > 0$ and $p > 0$. Then there exists a constant $c > 0$ that only depends on Ω, such that for all $\epsilon \in (0, \frac{1}{2}]$*

$$\sup_{\theta^\star \in \mathcal{D}} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} W_2\big(\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}, \mu_\lambda^*\big)^2 \ge \epsilon/n \Big) \le N \exp\big(-cN\epsilon^a \log(\epsilon)^{-2}\big),$$

where $a = \frac{1}{2} + \frac{1}{p}$.

Theorem 3.1 is proved in Section C.2 of the Supplementary Material [32].

REMARK 3.2. It is worth emphasizing in what sense Theorem 3.1 is uniform with respect to $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and to $\theta^\star \in \mathcal{D}$:

- *Uniformity with respect to* λ. We bound (in probability) the maximum (over λ) deviation between the empirical distribution $\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}$ and the predicted distribution μ_λ^*. (The supremum over λ is "inside" the probability.)
- *Uniformity with respect to* $\theta^\star$. We bound the maximum probability (over $\theta^\star$) of a deviation between $\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}$ and μ_λ^*. (The supremum over $\theta^\star$ is "outside" the probability.)

The reader might wonder whether it is possible to strengthen this result and bound the maximum deviation over $\theta^\star$ ("move the supremum over $\theta^\star$ inside"). The answer is negative. In particular, we can choose the support of $\theta^\star$ to coincide with a submatrix of X with atypically small minimum singular value. This will result in larger estimation error $\|\widehat{\theta}_\lambda - \theta^\star\|_2$, and hence in a large Wasserstein distance $W_2(\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}, \mu_\lambda^*)$.

REMARK 3.3. Theorem 3.1 compares the Wasserstein distance $W_2(\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}, \mu_\lambda^*)$ with the scale $1/\sqrt{n}$, that is the scale of the noise level. In particular, it implies that, form most coordinates i,

$$\widehat{\theta}_{\lambda,i} = \eta\left(\theta_i^\star + \frac{\tau_* Z_i}{\sqrt{n}}, \frac{\alpha_* \tau_*}{\sqrt{n}}\right) + O\left(\sqrt{\frac{\epsilon}{n}}\right), \tag{3.6}$$

for $Z_i \sim \mathcal{N}(0, 1)$. Hence the factor $1/n$ the theorem's statement is crucial for the error term to be negligible compared to the noise $\tau_* Z_i/\sqrt{n}$.

REMARK 3.4. Note that Theorem 3.1 does not hold for ℓ_0 balls. This is probably a fundamental problem, since controlling W_2 distance uniformly over ℓ_0 balls is impossible even in the simple sequence model (or, equivalently, for orthogonal designs X). Namely, consider the case in which we observe $y_i = \theta_i^\star + z_i$, $i \le N$, where $(z_i)_{i \le N} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau_*^2/n)$, and we try to estimate $\theta^\star$ by computing $\widehat{\theta}_{\lambda,i} = \eta(y_i; \lambda/\sqrt{n})$. Then there are vectors $\theta^\star \in \mathcal{F}_0(s)$ such that the empirical law $\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}$ does not concentrate in Wasserstein distance around its expectation μ_λ^*, that is, the law of $(\Theta, \eta(\Theta + Z; \lambda/\sqrt{n}))$ for $Z \sim \mathcal{N}(0, \tau_*/n)$.

In order to see this, it is sufficient to consider the vector

$$\theta^\star = (N, 2N, \dots, kN, 0, \dots, 0)/\sqrt{n} \in \mathcal{F}_0(s),$$

for $k = sN$. In Section F.1 of the Supplementary Material [32], we prove that (for this choice of $\theta^\star$) there exists a constant c_0 such that $W_2(\widehat{\mu}_{(\widehat{\theta}_\lambda, \theta^\star)}, \mu_\lambda^*)^2 \ge k/(Nn) = s/n$ with probability at least $1 - e^{-c_0 k}$ for all N large enough. This means that we cannot hope for Theorem 3.1 to hold for $\epsilon < s$, leading to a nonnegligible error term in (3.6).

We can think of several possibilities to overcome this intrinsic nonuniformity over ℓ_0 balls. One option would be to consider a weaker notion of distance between probability measures. Here, we follow a different route, and prove uniform estimates over ℓ_0 balls for several specific quantities of interest. In order to state these results, we introduce the following quantities, which correspond to the risk and the prediction error (and are expressed in terms of the solution (τ_*, β_*) of (3.3))

$$R_*(\lambda) = \tau_*(\lambda)^2 - \sigma^2, \tag{3.7}$$

$$P_*(\lambda) = \beta_*(\lambda)^2 + \frac{2\sigma^2}{\delta} s_*(\lambda) - \frac{\sigma^2}{\delta}. \tag{3.8}$$

THEOREM 3.2. *Assume here that $\mathcal{D}$ is either $\mathcal{F}_0(s)$ or $\mathcal{F}_p(\xi)$ for some $0 \le s < s_{\max}(\delta)$ and $\xi > 0$, $p > 0$. There exist a constant $c > 0$ that only depends on Ω, such that for all $\epsilon \in (0, 1]$*

$$\sup_{\theta^\star \in \mathcal{D}} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \big(\|\widehat{\theta}_\lambda - \theta^\star\|^2 - R_*(\lambda) \big)^2 \ge \epsilon \Big) \le N e^{-cN\epsilon^2}, \tag{3.9}$$

$$\sup_{\theta^\star \in \mathcal{D}} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \Big(\frac{1}{n} \|y - X\widehat{\theta}_\lambda\|^2 - \beta_*(\lambda)^2 \Big)^2 \ge \epsilon \Big) \le N e^{-cN\epsilon^2}, \tag{3.10}$$

$$\sup_{\theta^\star \in \mathcal{D}} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \Big(\frac{1}{n} \|X(\theta^\star - \widehat{\theta}_\lambda)\|^2 - P_*(\lambda) \Big)^2 \ge \epsilon \Big) \le N e^{-cN\epsilon^2}. \tag{3.11}$$

The statement (3.9) is proved in Section C.2, while (3.10)–(3.11) are proved in Section D of the Supplementary Material [32].

So far we focused on the Lasso estimator $\widehat{\theta}_\lambda$. The *debiased Lasso* estimator is defined as

$$\widehat{\theta}^d_\lambda = \widehat{\theta}_\lambda + \frac{X^{\mathsf{T}}(y - X\widehat{\theta}_\lambda)}{n - \|\widehat{\theta}_\lambda\|_0}.$$

This estimator plays a crucial role in the construction of confidence intervals and p-values [28, 45, 51, 54], and provide an explicit construction of the "direct observations" model in the sense that $\widehat{\theta}^d_\lambda$ is approximately distributed as $\mathcal{N}(\theta^\star, \tau_* \mathrm{I}/\sqrt{n})$. We let $\mu^{(d)}_\lambda$ be the law of the couple $(\Theta + \tau_*(\lambda) Z/\sqrt{n}, \Theta)$, where $(\Theta, Z) \sim \widehat{\mu}_{\theta^\star} \otimes \mathcal{N}(0, 1)$.

THEOREM 3.3. *Let $\widehat{\mu}_{(\widehat{\theta}^d_\lambda, \theta^\star)}$ denote the empirical distribution (on $\mathbb{R}^2$) of the entries of $(\widehat{\theta}^d_\lambda, \theta^\star)$. There exists a constant $c > 0$ such that for all $\epsilon \in (0, 1]$,*

$$\sup_{\theta^\star \in \mathcal{F}_4(\xi)} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} W_2\big(\widehat{\mu}_{(\widehat{\theta}^d_\lambda, \theta^\star)}, \mu^{(d)}_\lambda\big)^2 \ge \epsilon/n \Big) \le N e^{-cN\epsilon^{17/2}}.$$

Theorem 3.3 is proved in Section F.6 of the Supplementary Material [32].

REMARK 3.5. Theorems 3.1, 3.2, 3.3 appear to capture the correct probability decay as $N \to \infty$ for ϵ fixed, which is exponentially vanishing in N. On the other hand, our bounds are not always tight when $N \to \infty$ and $\varepsilon \to 0$ at the same time. In particular, the exponent $17/2$ in Theorem 3.3 is most likely a weakness of the proof. Notice in particular that the probability bound in Theorem 3.3 is less precise than the one in Theorem 3.1, probably because the more intricate structure of $\widehat{\theta}^d_\lambda$ makes our proof less direct.

We developed our analysis for the case of uncorrelated covariates $x_i \sim \mathcal{N}(0, \mathrm{I}_N)$. However, we believe that the general approach developed in this paper can be extended to general correlation structures. Indeed a characterization similar to the present one is expected to hold for correlated features [29]. The correlated case poses new technical challenges as well, in particular to prove a generalization of Proposition 3.1.

Nevertheless, we believe the theory for i.i.d. designs to be reasonably accurate in a broader domain, as illustrated by numerical simulations in the next section.

4. Applications.

4.1. *Estimation of the risk and the noise level.* In order to select the regularization parameter and to evaluate the quality of the Lasso solution $\widehat{\theta}_\lambda$, it is useful to estimate the risk and noise level. The paper [4] developed a suite of estimators of these quantities based on the asymptotic theory of [6]. The same paper also proposed generalizations of these estimators to correlated designs. Here, we revisit these estimators and prove stronger guarantees. First, we obtain quantitative bound on the consistency rate of our estimators. Second, our results are uniform over λ, which justifies using these estimators to select λ.

Let us start with the estimation of $\tau_*(\lambda)$ which plays a crucial role in the asymptotic theory. We define

$$\widehat{\tau}(\lambda) = \sqrt{n}\frac{\|y - X\widehat{\theta}_\lambda\|}{n - \|\widehat{\theta}_\lambda\|_0}.$$

We will see with Theorem F.1 presented in Section F.5 of the Supplementary Material [32] that

$$\lim_{N,n\to\infty} \frac{1}{N}\|\widehat{\theta}_\lambda\|_0 = \mathbb{P}\big(|\sqrt{n}\Theta + \tau_* Z| \geq \tau_*\lambda/\beta_*\big) \equiv s_*(\lambda).$$

Further, by Theorem 3.2, we have $\frac{1}{\sqrt{n}}\|y - X\widehat{\theta}_\lambda\| = \beta_*(\lambda) + o_n(1)$. Recall that by (3.3) we have $\beta_*(\lambda) = \tau_*(\lambda)(1 - \frac{1}{\delta}s_*(\lambda))$. We deduce $\widehat{\tau}(\lambda) = \tau_*(\lambda) + o_n(1)$. More precisely we have the following consistency result.

COROLLARY 4.1. *Assume here that $\mathcal{D}$ is either $\mathcal{F}_0(s)$ or $\mathcal{F}_p(\xi)$ for some $0 \leq s < s_{\max}(\delta)$ and $\xi > 0$, $p > 0$. There exists a constant $c > 0$ that only depend on Ω such that for all $\epsilon \in (0, 1]$,*

$$\sup_{\theta^\star\in\mathcal{D}} \mathbb{P}\Big(\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \big|\widehat{\tau}(\lambda) - \tau_*(\lambda)\big| \geq \epsilon\Big) \leq N\exp(-cN\epsilon^6).$$

We next consider estimating the ℓ_2 error of the Lasso. Following [6], we define

$$\widehat{R}(\lambda) = \widehat{\tau}(\lambda)^2\bigg(\frac{2}{n}\|\widehat{\theta}_\lambda\|_0 - \frac{N}{n}\bigg) + \frac{\|X^{\mathsf{T}}(y - X\widehat{\theta}_\lambda)\|^2}{(n - \|\widehat{\theta}_\lambda\|_0)^2}.$$

COROLLARY 4.2. *Assume here that $\mathcal{D}$ is either $\mathcal{F}_0(s)$ or $\mathcal{F}_p(\xi)$ for some $0 \leq s < s_{\max}(\delta)$ and $\xi > 0$, $p > 0$. There exists a constant $c > 0$ such that for all $\epsilon \in (0, 1]$,*

$$\sup_{\theta^\star\in\mathcal{D}} \mathbb{P}\Big(\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \big|\widehat{R}(\lambda) - \|\widehat{\theta}_\lambda - \theta^\star\|^2\big| \geq \epsilon\Big) \leq Ne^{-cN\epsilon^6},$$

Corollary 4.2 is proved in Section F.7 of [32]. Since by Corollary 4.2, Corollary 4.1, Theorem 3.2 we have with high probability $\widehat{R}(\lambda) \simeq \|\widehat{\theta}_\lambda - \theta^\star\|^2 \simeq \tau_*(\lambda)^2 - \sigma^2 \simeq \widehat{\tau}(\lambda)^2 - \sigma^2$, the estimator

$$\widehat{\sigma}^2(\lambda) = \widehat{\tau}(\lambda)^2 - \widehat{R}(\lambda) = \widehat{\tau}(\lambda)^2\bigg(1 + \frac{N}{n} - \frac{2}{n}\|\widehat{\theta}_\lambda\|_0\bigg) - \frac{\|X^{\mathsf{T}}(y - X\widehat{\theta}_\lambda)\|^2}{(n - \|\widehat{\theta}_\lambda\|_0)^2} \tag{4.1}$$

is a consistent estimator of the noise level σ^2.

COROLLARY 4.3. *There exists a constant $c > 0$ that only depends on Ω, such that for all $\epsilon \in (0, 1]$,*

$$\sup_{\theta^\star\in\mathcal{D}} \mathbb{P}\Big(\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \big|\widehat{\sigma}^2(\lambda) - \sigma^2\big| > \epsilon\Big) \leq Ne^{-cN\epsilon^6}.$$

Finally, we consider the prediction error $\|X\theta^\star - X\widehat{\theta}_\lambda\|$. Stein Unbiased Risk Estimator (SURE) provides a general method to estimate the prediction error, see, for example, [23, 41, 49]. In the present case, it takes the form

$$\widehat{P}^{\mathrm{SURE}}(\lambda) = \frac{1}{n}\|y - X\widehat{\theta}_\lambda\|^2 + \frac{2\sigma^2}{n}\|\widehat{\theta}_\lambda\|_0. \tag{4.2}$$

Tibshirani and Taylor [49] proved that $\widehat{P}^{\mathrm{SURE}}(\lambda)$ is an unbiased estimator of the prediction error, namely

$$\mathbb{E}\{\widehat{P}^{\mathrm{SURE}}(\lambda)\} = \frac{1}{n}\|X\theta^\star - X\widehat{\theta}_\lambda\|^2 + \sigma^2. \tag{4.3}$$

The next result establishes consistency, uniformly over λ and $\theta^\star$, with quantitative concentration estimates.

COROLLARY 4.4. *Assume here that $\mathcal{D}$ is either $\mathcal{F}_0(s)$ or $\mathcal{F}_p(\xi)$ for some $0 \le s < s_{\max}(\delta)$ and $\xi > 0$, $p > 0$. There exists a constant $c > 0$ that only depends on Ω such that for all $\epsilon \in (0, 1]$*

$$\sup_{\theta^\star\in\mathcal{D}} \mathbb{P}\left(\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \left| \frac{1}{n}\|X\theta^\star - X\widehat{\theta}_\lambda\|^2 + \sigma^2 - \widehat{P}^{\mathrm{SURE}}(\lambda) \right| \ge \epsilon \right) \le Ne^{-cN\epsilon^6}.$$

The same result holds if σ in (4.2) is replaced by an estimator of the noise level satisfying the same consistency condition as $\widehat{\sigma}$ defined by (4.1) (cf. Corollary 4.3).

This corollary follows simply from Theorem F.1 from [32] and Theorem 3.2.

REMARK 4.1. Notice that exact unbiasedness of $\widehat{P}^{\mathrm{SURE}}(\lambda)$ only holds if the noise z in the linear model (1.2) is Gaussian [49]. In contrast, it is not hard to generalize the proofs in the present paper to include other noise distributions.

REMARK 4.2. In Corollaries 4.1, 4.2, 4.3, 4.4, we need to take $\epsilon \ge C((\log N)/N)^{1/6}$ in order for the probability bounds on the right-hand side to vanish asymptotically. Therefore, considering for instance Corollary 4.2, we get

$$\|\widehat{\theta}_\lambda - \theta^\star\|^2 = \widehat{R}(\lambda) + O_{\mathbb{P}}\left(\left(\frac{\log N}{N}\right)^{1/6}\right). \tag{4.4}$$

We do not expect the exponent $1/6$ in this result to be tight. Nevertheless, in the proportional regime which is our focus here ($n \asymp N \asymp k_0$, with k_0 the number of nonzeros in θ^*), the risk $\|\widehat{\theta}_\lambda - \theta^\star\|^2$ is typically of order one, and therefore the $N^{-1/6}$ term is negligible.

If $k_0 \ll N$, the risk is of order $(k_0 \log N)/n$. Keeping to the proportional regime[1] $n \asymp N$ the error term in equation (4.4) is negligible provided $k_0 \gg (N/\log N)^{5/6}$.

We can also compare equation (4.4) with the results of [11], Theorem 1, establishing that the minimax rates for estimating $\|\widehat{\theta}_\lambda - \theta^\star\|^2$ is $\min(k \log N)/n; 1/\sqrt{n})$. In the present proportional regime $n \asymp N$, this is much smaller than the error bound in equation (4.4). On the other hand, [11], Theorem 1, requires $k_0 \ll \sqrt{N}$, while our guarantee holds up to linear sparsity.

[1] In the proof of Corollary 4.2, some of the inequalities are less precise outside the regime $n \asymp N$.

4.2. *Adaptive selection of* λ. As anticipated, we can use our uniform bounds to select λ through an adaptive procedure. We discuss here three such procedures, that have already been illustrated in Figure 1 (we refer to Sections 4.1 to 4.3 for further details on these quantities): (i) selecting λ by minimizing the estimate $\widehat{\tau}(\lambda)$, we denote this by $\widehat{\lambda}^{\mathrm{EST}}$; (ii) select λ as to minimize Stein's Unbiased Risk Estimate $\widehat{P}^{\mathrm{SURE}}(\lambda)$, $\widehat{\lambda}^{\mathrm{SURE}}$; (iii) select λ by k-fold cross-validation, $\widehat{\lambda}^{k\text{-CV}}$. We will next describe these procedures in greater detail, and state the corresponding guarantees. Before getting into the analysis of these adaptive procedures, it is useful to discuss why the minimax choice $\lambda \asymp \sqrt{\log(N)}$ is not satisfactory.

LEMMA 4.1. *Let $\gamma \in \mathbb{R}_{>0}$ be such that the largest singular value of X is at most γ with probability at least $1-q$. For all $s \in (0,1]$ and all $\lambda \geq 4\sigma\gamma/\sqrt{sN}$, there exists $\theta^\star \in \mathcal{F}_0(s)$ such that*

$$\|\widehat{\theta}_\lambda - \theta^\star\|^2 \geq \frac{nN}{4\gamma^4} s\lambda^2, \tag{4.5}$$

with probability at least $1 - e^{-2n} - q$.

Lemma 4.1 is proved in Section F.2 of the Supplementary Material [32]. For independent Gaussian design $X_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ we have $\gamma \simeq \sqrt{N} + \sqrt{n}$ with high probability. Hence if $\lambda \xrightarrow[N,n\to\infty]{} +\infty$ we have $\lambda \geq 4\sigma\gamma/\sqrt{sN}$ with high probability for N, n large enough. Lemma 4.1 gives that with high probability

$$\|\widehat{\theta}_\lambda - \theta^\star\|^2 \geq \frac{nN}{4\gamma^4} s\lambda^2 \simeq \frac{\delta}{4(1+\sqrt{\delta})^4} s\lambda^2,$$

which diverges as $N, n \to \infty$. We conclude that a diverging λ leads to a risk that goes to infinity. The underlying mechanism is well known: If the nonzero entries of θ^* are large enough, the Lasso incurs a bias of order λ on those entries, and hence a mean square error of order $s\lambda^2$.

As this point a few remarks are in order:

(i) The minimax choice $\lambda \asymp \sqrt{\log N}$ is motivated by cases in which the sparsity $\|\theta^*\|_0 = k_0$ is such that $k_0 \leq N^{1-\varepsilon}$, and $(k_0 \log N)/n \to 0$. An alternative would be to use the minimax regularization for the proportional asymptotics $k_0, N, n \to \infty$, with $k_0/N = s$, $n/N = \delta$ as determined in [21]. In the next section, we carry out such a comparison. For $s \lesssim 0.1$, the minimax choice is, again, substantially suboptimal.

(ii) Both in the proportional regime $k_0 \asymp n \asymp N$, and in the sparse regime $k_0 \leq N^{1-\varepsilon}$, the minimax and adaptive choices of λ can differ by an arbitrarily large constant factor. The resulting estimation errors also differ by an arbitrarily large constant factor. To see this, consider the case of an s_0 sparse vector in which $s \leq s_0$ of the nonzero entries are very large, and the other are extremely small. The minimax choice $\lambda \asymp \sqrt{\log N/s_0}$ leads to $\|\widehat{\theta}_\lambda - \theta^\star\|_2^2 \asymp s\log(N/s_0)/n$, while the adaptive selection $\lambda \asymp \sqrt{\log(N/s)}$ leads to[2] $\|\widehat{\theta}_\lambda - \theta^\star\|_2^2 \asymp s\log(N/s)/n$.

In practice, the regularization selected by cross-validation is often smaller than this classical value [16].

[2]An arbitrarily constant factor is obtained, for instance, by taking $s_0 = N^{\alpha_0}$ and $s = N^\alpha$ for some $0 < \alpha < \alpha_0 < 1$. Notice that in fact a diverging factor can also arise by taking $s_0 = N/\log N$, $s = N^\alpha$, $0 < \alpha < 1$.

Minimization of $\widehat{\tau}(\lambda)$. Since the ℓ_2 risk of the Lasso is by Theorem 3.2 approximately equal to $R_*(\lambda) = \tau_*(\lambda)^2 - \sigma^2$ and since by Corollary 4.1, $\widehat{\tau}$ is a consistent estimator (uniformly in λ) of τ_*, a natural procedure for selecting λ is to minimize $\widehat{\tau}$. We then define

$$\widehat{\lambda}^{\mathrm{EST}} = \operatorname*{arg\,min}_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \widehat{\tau}(\lambda).$$

The next result is an immediate consequence of Theorem 3.2 and Corollary 4.1:

PROPOSITION 4.1. *Assume here that* $\mathcal{D}$ *is either* $\mathcal{F}_0(s)$ *or* $\mathcal{F}_p(\xi)$ *for some* $0 \le s < s_{\max}(\delta)$ *and* $\xi > 0$, $p > 0$. *There exists a constants* $c > 0$ *that only depends on* Ω *such that for all* $\epsilon \in (0, 1]$,

$$\inf_{\theta^\star\in\mathcal{D}} \mathbb{P}\Big(\big\|\widehat{\theta}_{\widehat{\lambda}^{\mathrm{EST}}} - \theta^\star\big\|^2 \le \inf_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \big\{\big\|\widehat{\theta}_\lambda - \theta^\star\big\|^2\big\} + \epsilon\Big) \ge 1 - Ne^{-cN\epsilon^6}.$$

Minimization of SURE. We define

$$\widehat{\lambda}^{\mathrm{SURE}} = \operatorname*{arg\,min}_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \widehat{P}^{\mathrm{SURE}}(\lambda).$$

Here, it is understood that we can use either σ or $\widehat{\sigma}(\lambda)$ (cf. equation (4.1), in the definition of $\widehat{P}^{\mathrm{SURE}}$. We deduce from Corollary 4.4 the following.

PROPOSITION 4.2. *Assume here that* $\mathcal{D}$ *is either* $\mathcal{F}_0(s)$ *or* $\mathcal{F}_p(\xi)$ *for some* $0 \le s < s_{\max}(\delta)$ *and* $\xi > 0$, $p > 0$. *There exists a constant* $c > 0$ *that only depends on* Ω *such that for all* $\epsilon \in (0, 1]$,

$$\inf_{\theta^\star\in\mathcal{D}} \mathbb{P}\left(\frac{1}{n}\big\|X\widehat{\theta}_{\widehat{\lambda}^{\mathrm{SURE}}} - X\theta^\star\big\|^2 \le \inf_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \left\{\frac{1}{n}\big\|X\widehat{\theta}_\lambda - X\theta^\star\big\|^2\right\} + \epsilon\right)$$
$$\ge 1 - N\exp\big(-cN\epsilon^6\big).$$

Cross-validation. We analyze now k-fold cross-validation. Let $k \ge 2$ and define $n_k = n(k-1)/k$. We partition the rows of X in k groups: we obtain k-submatrices of size $(n/k) \times N$ that we denote $X^{(1)}, \dots, X^{(k)}$. Let us also write for $i \in \{1, \dots, k\}$, $X^{(-i)}$ for the submatrix of X obtained by removing the rows $X^{(i)}$. We denote by $y^{(i)}$, $z^{(i)}$ and $y^{(-i)}$, $z^{(-i)}$ the corresponding subvectors of y and z.

The estimator $\widehat{R}^{k\text{-CV}}$ of the risk using k-fold cross-validation if defined as follows. For $i = 1, \dots, k$ solve the Lasso problem

$$\widehat{\theta}^i_\lambda = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^N} \left\{\frac{1}{2n_k}\big\|y^{(-i)} - X^{(-i)}\theta\big\|^2 + \frac{\lambda}{\sqrt{n}}|\theta|\right\},$$

and then compute

$$\widehat{R}^{k\text{-CV}}(\lambda) = \frac{1}{n}\sum_{i=1}^{k}\big\|y^{(i)} - X^{(i)}\widehat{\theta}^i_\lambda\big\|^2.$$

Finally, we set λ as follows:

$$\widehat{\lambda}^{k\text{-CV}} = \operatorname*{arg\,min}_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \widehat{R}^{k\text{-CV}}(\lambda).$$

The next proposition shows that $\widehat{R}^{k\text{-CV}}(\lambda)$ is equal to the true risk (shifted by σ^2) up to $O(k^{-1/2})$.

PROPOSITION 4.3. *There exists a constant $c > 0$ that depends only on Ω, such that for all $k \geq 2$ such that $s_{\max}((k-1)\delta/k) > s$ in the case where $\mathcal{D} = \mathcal{F}_0(s)$, we have*

$$\sup_{\theta^\star \in \mathcal{D}} \mathbb{P}\Big(\sup_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |\widehat{R}^{k\text{-CV}}(\lambda) - \|\widehat{\theta}_\lambda - \theta^\star\|^2 - \sigma^2| \geq \frac{C}{\sqrt{k}} \Big) \leq N e^{-cN/k^6}.$$

Proposition 4.3 is proved in Section F.8 from [32]. It follows from Proposition 4.3 that with high probability,

$$\|\widehat{\theta}_{\widehat{\lambda}^{k\text{-CV}}} - \theta^\star\|^2 \leq \inf_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \|\widehat{\theta}_\lambda - \theta^\star\|^2 + O(k^{-1/2}).$$

Discussion. The three methods discussed in this section (EST, SURE, CV) present important differences: for this reason it is useful for the statistician to have multiple approaches available.

First of all, these three approaches minimize estimates of different quantities. SURE estimates the average prediction error at the points in the training sample, namely $n^{-1}\|X(\widehat{\theta}_\lambda - \theta^\star)\|_2^2 + \sigma^2$.

CV estimates the prediction error on a test sample $(x^{\text{test}}, y^{\text{test}})$, namely $\mathbb{E}_{\text{test}}\{(y^{\text{test}} - \langle \widehat{\theta}_\lambda, x^{\text{test}} \rangle)^2\}$. Under the linear model $y = \langle \theta^\star, x \rangle + \sigma z$, this coincides with $\|\Sigma^{1/2}(\widehat{\theta}_\lambda - \theta^\star)\|_2^2 + \sigma^2$, where $\Sigma = \mathbb{E}(xx^{\mathsf{T}})$ is the population covariance. While in this paper, we are focusing on $\Sigma = I_p$, this quantity is in general different from the estimation error $\|\widehat{\theta}_\lambda - \theta^\star\|_2^2$.

Finally, $\widehat{\tau}(\lambda)^2$ was conjectured in [4] to be an alternative estimate of the same quantity $\|\Sigma^{1/2}(\widehat{\theta}_\lambda - \theta^\star)\|_2^2 + \sigma^2$ for general Gaussian designs.[3]

CV is the most robust: we expect it to be consistent under significantly weaker assumptions than the ones in Proposition 4.3. On the other hand, it presents an inconvenient computation-accuracy tradeoff. For small k, it is biased since it use the prediction error from a sample of size $n(k-1)/k$ to estimate the prediction error corresponding to the full sample. We expect this bias to be at least of order k^{-1} and Proposition 4.3 shows that it is at most of order $k^{-1/2}$. On the other hand, for large k it is computationally expensive (it requires solving k Lassos). SURE and EST are likely to be more sensitive to the model assumptions, but do not have a large bias (in fact SURE is unbiased) and are very inexpensive. The bias of CV is clearly visible in Figure 2 below.

4.3. *Numerical experiments*. In this section, we compare numerically various different choices for the regularization parameter λ, namely $\widehat{\lambda}^{\text{EST}}$, $\widehat{\lambda}^{\text{SURE}}$ and $\widehat{\lambda}^{k\text{-CV}}$, presented in the previous section. For these experiments, we take the components $\theta_1^\star, \dots, \theta_N^\star$ to be i.i.d. from

$$P_0 = s\mathcal{N}(0, 1/n) + (1-s)\delta_0.$$

Within this probabilistic model, we can compare achieved by our various choice of λ to the Bayes optimal error (Minimal Mean Squared Error):

$$\text{MMSE}_N = \min_{\widehat{\theta}} \mathbb{E}\big[\|\theta^\star - \widehat{\theta}(y, X)\|^2\big] = \mathbb{E}\big[\|\theta^\star - \mathbb{E}[\theta^\star | y, X]\|^2\big],$$

where the minimum is taken over all estimators $\widehat{\theta}$ (i.e., measurable functions of X, y). The limit of the MMSE has been recently computed by [2] and [39]. Recall, that given two random variables U, V, their mutual information is the Kullback–Leibler divergence between their joint distribution and the product of the marginals: $I(U; V) \equiv D_{\text{KL}}(p_{U,V} \| p_U \times p_V)$.

[3] This conjecture was proved long after a first submission of this manuscript in two independent papers [7, 14].

THEOREM 4.1 (Information-theoretic limit, from [2, 39]). *Consider the linear regression model* (1.2) *with standard Gaussian designs and parameter vector* θ^* *such that* $(\sqrt{n}\theta_i^*)_{i\le n} \sim_{iid} P$, *with* P *a probability distribution with finite second moment. Define the function*

$$\Psi_{\delta,\sigma}(m) = I_P\left(\frac{\sigma^{-2}}{1+m}\right) + \frac{\delta}{2}\left(\log(1+m) - \frac{m}{1+m}\right),$$

where $I_P(r) = I(\Theta; \sqrt{r}\bar{\Theta} + Z)$ *for* $(\bar{\Theta}, Z) \sim P \otimes \mathcal{N}(0,1)$. *Then, for almost every* $\delta, \sigma > 0$ *the function* $\Psi_{\delta,\sigma}$ *admits a unique maximizer* $m^*(\delta,\sigma)$ *on* $\mathbb{R}_{\ge 0}$. *Further, in the limit* $N, n \to \infty$ *with* $n/N \to \delta$, *we have*

$$\mathrm{MMSE}_N \xrightarrow[N\to\infty]{} \delta\sigma^2 m^*(\delta,\sigma).$$

We also refer to the MMSE predicted in this theorem as to the Bayes optimum. Figure 1 reports the risk achieved by the various choices of λ as a function of the number of samples per dimension δ. We also compare the data-driven procedures of the previous section to the theory-driven choice $\lambda = \sigma\sqrt{2\log N}$. In the left frame, we consider uncorrelated random designs: $X_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. On the right, we consider i.i.d. Gaussian rows with covariance structure determined by an autoregressive model. Explicitly, the columns $(X_j)_{1\le j\le N}$ of X are generated according to

$$X_1 = u_0,\ X_{j+1} = \frac{1}{\sqrt{1+\phi^2}}(\phi X_j + u_j), \tag{4.6}$$

where $u_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\phi = 2$. For both types of designs, $\widehat{\lambda}^{\mathrm{EST}}$, $\widehat{\lambda}^{\mathrm{SURE}}$ and $\widehat{\lambda}^{k\text{-CV}}$ perform similarly, and substantially outperform the theoretical choice $\lambda = \sigma\sqrt{2\log N}$.

For uncorrelated designs, the resulting risk is closely tracked by the asymptotic theory, and is surprisingly close to the asymptotic prediction for the Bayes risk MMSE_N. While our theory does not cover the case of correlated designs, the qualitative behavior is remarkably similar.

In Figure 2, we investigate in greater detail the effect of correlations among the covariates. We consider again the autoregressive correlation structure of equation (4.6) and plot the estimation error $\|\widehat{\theta}_\lambda - \theta^*\|_2^2$ and test error $\langle \widehat{\theta}_\lambda - \theta^*, \Sigma(\widehat{\theta}_\lambda - \theta^*)\rangle$, as functions of λ. We also plot the cross-validation estimator $\widehat{R}^{k\text{-CV}}(\lambda)$ as well as the estimator $\widehat{R}(\lambda)$ introduced in Section 4.1. Notice that the cross-validation estimator is expected to be a consistent estimator of the test error $\langle \widehat{\theta}_\lambda - \theta^*, \Sigma(\widehat{\theta}_\lambda - \theta^*)\rangle$, for large k, but we to not expect $\widehat{R}(\lambda)$ to be necessarily consistent for correlated designs.

It is worth emphasizing two observations that seem generalize to other examples. The risk changes significantly as the correlation strength increases, but gracefully so. For instance, for $\phi = 1$ (which means that consecutive covariates have correlation $1/\sqrt{2} \approx 0.71$), the error estimator $\widehat{R}(\lambda)$ is nearly identical to the actual prediction error. Further, it is only a factor 2 larger than the estimation error. Second, and practically more important, the value of λ selected by minimizing $\widehat{R}(\lambda)$ is very close to optimal. This is to be compared with the standard theory prescription $\lambda = \sigma\sqrt{2\log N} \approx 0.82$ (beyond the axis limit in these figures).

We also observed that in this case, the risk estimator $\widehat{R}(\lambda)$ is not consistent but its minimum is roughly located at the same value of λ as for uncorrelated designs.

Next, we study adaptivity to sparsity. On Figure 3, we plot the risk as a function of the sparsity of the signal $\theta^\star$. We compare the three adaptive procedures (namely, $\widehat{\lambda}^{\mathrm{EST}}$, $\widehat{\lambda}^{\mathrm{SURE}}$

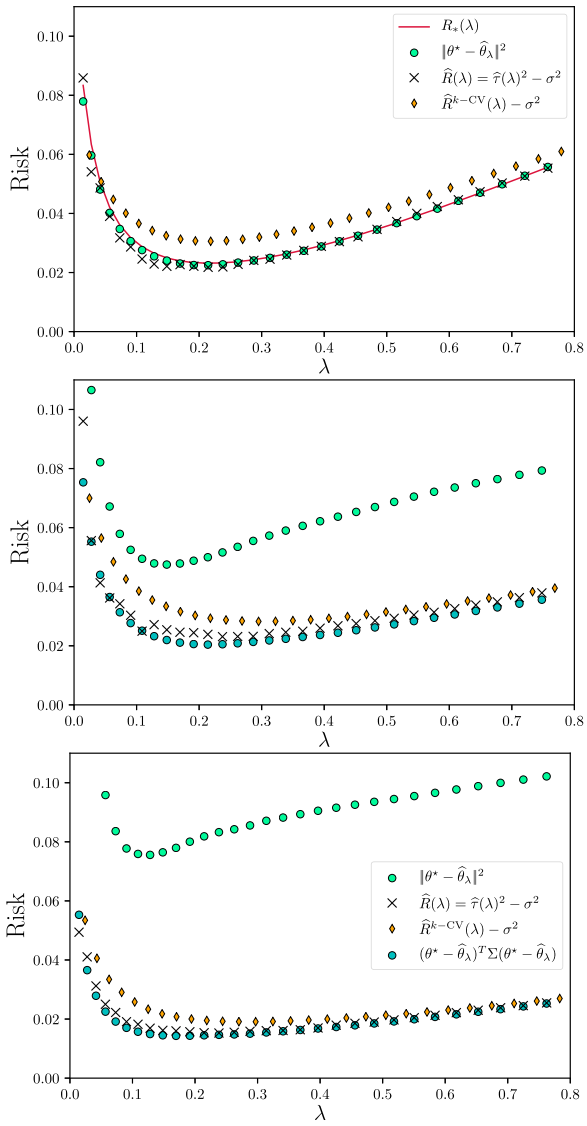


FIG. 2. *Estimation risk of the Lasso as a function of* λ. *Here,* $N = 5000$, $\delta = 0.7$, $\sigma = 0.2$. *The true coefficients vector* $\theta^\star$ *is chosen to be* sN*-sparse with* $s = 0.1$. *The entries on the support of* $\theta^\star$ *are i.i.d.* $\mathcal{N}(0, 1/n)$. *Cross-validation is carried out using* 4 *folds. SURE is computed using the estimator* $\widehat{\sigma}$ *for the plot on the left, and the true value of* σ *on the right. Left*: *A standard random design with* $(X_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. *Center*: *The rows of the design matrix* X *are i.i.d. Gaussian, with correlation structure given by an autoregressive process* (*see equation* (4.6)) *with* $\phi = 1$. *Bottom*: *Same as for the previous plot, but with* $\phi = 2$.

and $\widehat{\lambda}^{k\text{-CV}}$), to the following choice:

$$\lambda^{\mathrm{MM}}(s_0) = \alpha_0 \sigma \sqrt{1 - \frac{1}{\delta} M_{s_0}(\alpha_0)},$$

$$M_s(\alpha) = s\big(1 + \alpha^2\big) + 2(1 - s)\big(\big(1 + \alpha^2\big)\Phi(-\alpha) - \alpha\phi(\alpha)\big),$$

$$\alpha_0 = \arg\min_{\alpha \geq 0} M_{s_0}(\alpha),$$

where $s_0 < s_{\max}(\delta)$ is a nominal value for the sparsity (in Figure 3, we use $s_0 = 0.3$). The value $\lambda^{\mathrm{MM}}(s_0)$ is expected to be asymptotically minimax optimal over $\mathcal{F}_0(s_0)$ [21].

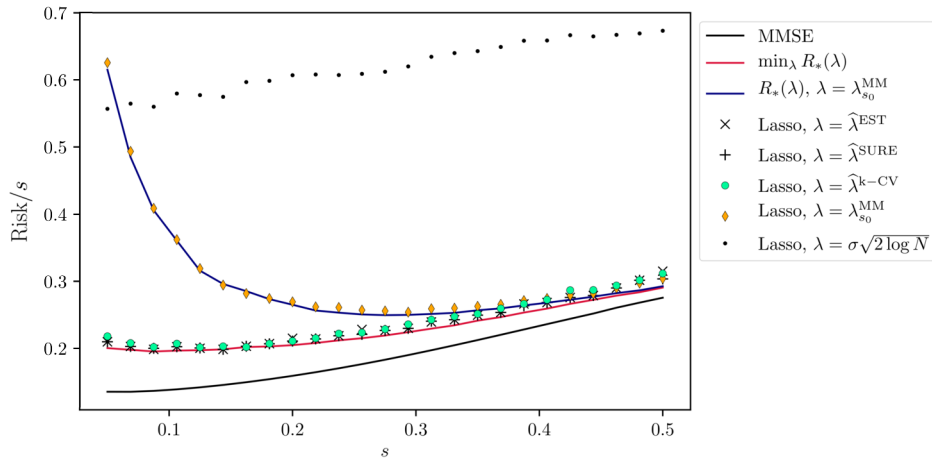


FIG. 3. *Risk of the Lasso for different choices of* λ. $N = 10{,}000$, $\sigma = 0.2$, $\delta = 0.8$. *Here,* $\theta^\star$ *is chosen to be* sN*-sparse, and we vary the sparsity level* s. *The entries on the support of* $\theta^\star$ *are i.i.d.* $\mathcal{N}(0,1)$. *Cross-validation is carried out using* 4 *folds. SURE is computed using the estimator* $\widehat{\sigma}$. *The minimax regularization* $\lambda^{\mathrm{MM}}(s_0)$ *is used at the nominal level* $s_0 = 0.3$.

Also in this example, adaptive procedures dramatically outperform the fixed choice $\lambda = \sigma\sqrt{2\log N}$, and also the minimax optimal λ at the nominal sparsity level.

5. Proof strategy. As mentioned above, our proofs are based on Gaussian comparison inequalities, and in particular on Gordon's min-max theorem [26, 27]. In this section, we review the application of this inequality to the Lasso as developed in [47]. We then discuss the limitations of earlier work, which does not characterize the empirical distribution of the Lasso estimator $\widehat{\theta}_\lambda$ (or need extra sparsity assumptions [36]) nor uniform bounds as in Theorem 3.1. A key challenge is related to the fact that the Lasso cost function (1.1) is convex but not strongly convex. Hence, a small change in λ could cause *a priori* a large change in the minimizer $\widehat{\theta}_\lambda$.

In order to overcome these problems, we establish a property that we call "local stability." Namely, if the empirical distribution of $(\widehat{\theta}_\lambda, \theta^\star)$ deviates from our prediction, then the value of the optimization problem increases significantly. This implies that the empirical distribution is stable with respect to perturbations of the cost (e.g., changes in λ). Gordon's comparison is again crucial to prove this stability property.

Finally, we describe how local stability is used to prove the theorems in the previous sections. A full description of the proofs is provided in the Appendices [32].

5.1. *Tight Gaussian min-max theorem.* For convenience, we will use a different (but equivalent) scaling for the proofs. Instead of taking $X_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ as we did above, we should consider from now (and in the Supplementary Material) that $X_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1/n)$. This amounts of replacing θ by $\theta/\sqrt{n}$, so that $X\theta$ remains unchanged. With this new normalization, a statement like $\|\widehat{\theta}_\lambda - \theta^\star\|^2 \simeq R_*(\lambda)$ becomes $\frac{1}{N}\|\widehat{\theta}_\lambda - \theta^\star\|^2 \simeq \delta R_*(\lambda)$

It is more convenient (but equivalent) to study $\widehat{w}_\lambda = \widehat{\theta}_\lambda - \theta^\star$ instead of $\widehat{\theta}_\lambda$. The vector $\widehat{w}_\lambda$ is the minimizer of the cost function

$$\mathcal{C}_\lambda(w) = \frac{1}{2n}\|Xw - \sigma z\|^2 + \frac{\lambda}{n}\big(|w+\theta^\star| - |\theta^\star|\big). \tag{5.1}$$

Following [47], we rewrite the minimization of $\mathcal{C}_\lambda$ as a saddle point problem:

$$\min_{w\in\mathbb{R}^N} \mathcal{C}_\lambda(w) = \min_{w\in\mathbb{R}^N}\max_{u\in\mathbb{R}^n}\left\{\frac{1}{n}u^{\mathsf{T}}(Xw - \sigma z) - \frac{1}{2n}\|u\|^2 + \frac{\lambda}{n}\big(|w+\theta^\star| - |\theta^\star|\big)\right\}. \tag{5.2}$$

We apply the following Theorem from [47] which improves over Gordon's Theorem [27] by exploiting convex duality.

THEOREM 5.1 (Theorem 3 from [47]). *Let $S_w \subset \mathbb{R}^N$ and $S_u \subset \mathbb{R}^n$ be two compact sets and let $Q : S_w \times S_u \to \mathbb{R}$ be a continuous function. Let $G = (G_{i,j}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $g \sim \mathcal{N}(0, \mathbf{I}_N)$ and $h \sim \mathcal{N}(0, \mathbf{I}_n)$ be independent standard Gaussian vectors. Define*

$$\begin{cases} \mathcal{C}^*(G) = \min\limits_{w \in S_w} \max\limits_{u \in S_u} u^\mathsf{T} G w + Q(w,u), \\ L^*(g,h) = \min\limits_{w \in S_w} \max\limits_{u \in S_u} \|u\|_2 g^\mathsf{T} w + \|w\|_2 h^\mathsf{T} u + Q(w,u). \end{cases}$$

Then we have:

- *For all $t \in \mathbb{R}$,*

$$\mathbb{P}\big(\mathcal{C}^*(G) \le t\big) \le 2\mathbb{P}\big(L^*(g,h) \le t\big).$$

- *If S_w and S_u are convex and if Q is convex concave, then for all $t \in \mathbb{R}$,*

$$\mathbb{P}\big(\mathcal{C}^*(G) \ge t\big) \le 2\mathbb{P}\big(L^*(g,h) \ge t\big).$$

For the reader's convenience, we provide in Section G.3 of the Supplementary Material [32] a proof of this theorem.

Because of Gordon's theorem, it suffices now to study (see Corollary 5.1 below) for $(g, g', h) \sim \mathcal{N}(0, \mathbf{I}_N) \otimes \mathcal{N}(0,1) \otimes \mathcal{N}(0, \mathbf{I}_n)$.

$$L_\lambda(w) = \frac{1}{2}\left(\sqrt{\frac{\|w\|^2}{n} + \sigma^2}\frac{\|h\|}{\sqrt{n}} - \frac{1}{n} g^\mathsf{T} w + \frac{g'\sigma}{\sqrt{n}}\right)_+^2 + \frac{\lambda}{n}|w + \theta^\star| - \frac{\lambda}{n}|\theta^\star|. \tag{5.3}$$

COROLLARY 5.1.

(a) *Let $D \subset \mathbb{R}^N$ be a closed set. We have for all $t \in \mathbb{R}$*

$$\mathbb{P}\Big(\min_{w \in D} \mathcal{C}_\lambda(w) \le t\Big) \le 2\mathbb{P}\Big(\min_{w \in D} L_\lambda(w) \le t\Big).$$

(b) *Let $D \subset \mathbb{R}^N$ be a convex closed set. We have for all $t \in \mathbb{R}$*

$$\mathbb{P}\Big(\min_{w \in D} \mathcal{C}_\lambda(w) \ge t\Big) \le 2\mathbb{P}\Big(\min_{w \in D} L_\lambda(w) \ge t\Big).$$

PROOF. We will only prove the first point, since the second follows from the same arguments. Define for $(w,u) \in \mathbb{R}^N \times \mathbb{R}^n$,

$$\begin{aligned} c_\lambda(w,u) &= \frac{1}{n} u^\mathsf{T} X w - \frac{\sigma}{n} u^\mathsf{T} z - \frac{1}{2n}\|u\|^2 + \frac{\lambda}{n}\big(|w + \theta^\star| - |\theta^\star|\big), \\ l_\lambda(w,u) &= -\frac{1}{n^{3/2}}\|u\| g^\mathsf{T} w + \frac{1}{n}\|u\| g'\sigma + \sqrt{\frac{\|w\|^2}{n} + \sigma^2}\frac{h^\mathsf{T} u}{n} \\ &\quad - \frac{1}{2n}\|u\|^2 + \frac{\lambda}{n}\big(|w + \theta^\star| - |\theta^\star|\big). \end{aligned}$$

Notice that for all $w \in \mathbb{R}^N$, $L_\lambda(w) = \max_{u \in \mathbb{R}^n} l_\lambda(w,u)$ and $\mathcal{C}_\lambda(w) = \max_{u \in \mathbb{R}^n} c_\lambda(w,u)$.

Let us suppose that X, z, g, h, g' live on the same probability space and are independent. Let $\epsilon \in (0,1]$. Let $\sigma_{\max}(X)$ denote the largest singular value of the matrix X. By tightness, we can find $K > 0$ such that the event

$$\big\{\sigma_{\max}(X) \le K, \|z\| \le K, \|g\| \le K, \|h\| \le K, |g'| \le K\big\} \tag{5.4}$$

has probability at least $1-\epsilon$. Let $D \subset \mathbb{R}^N$ be a (nonempty, otherwise the result is trivial) closed set. Let us fix $w_0 \in D$. On the event (5.4) $\mathcal{C}_\lambda(w_0)$ and $L_\lambda(w_0)$ are both upper bounded by some nonrandom quantity R. Let now $w \in D$ such that $\mathcal{C}_\lambda(w) \le R$. We have then $\frac{\lambda}{n}|w+\theta^\star| \le R + \frac{\lambda}{n}|\theta^\star|$, which implies that $\|w\|$ is upper bounded by some nonrandom quantity R_1. This implies that, on the event (5.4), the minimum of $\mathcal{C}_\lambda$ over D is achieved on $D \cap B(0,R_1)$. Similarly on (5.4), the minimum of L_λ over D is achieved on $D \cap B(0,R_2)$, for some nonrandom quantity R_2. Without loss of generalities, one can assume $R_1 = R_2$. On the event (5.4), we have

$$\min_{w\in D} \mathcal{C}_\lambda(w) = \min_{w\in D\cap B(0,R_1)} \mathcal{C}_\lambda(w) = \min_{w\in D\cap B(0,R_1)} \max_{u\in B(0,R_3)} c_\lambda(w,u),$$

for some nonrandom $R_3 > 0$. This gives that for all $t \in \mathbb{R}$, we have

$$\mathbb{P}\Big(\min_{w\in D} \mathcal{C}_\lambda(w) \le t\Big) \le \mathbb{P}\Big(\min_{w\in D\cap B(0,R_1)} \max_{u\in B(0,R_3)} c_\lambda(w,u) \le t\Big) + \epsilon,$$

and similarly

$$\mathbb{P}\Big(\min_{w\in D\cap B(0,R_1)} \max_{u\in B(0,R_3)} l_\lambda(w,u) \le t\Big) \le \mathbb{P}\Big(\min_{w\in D} L_\lambda(w) \le t\Big) + \epsilon.$$

Since the sets $D \cap B(0,R_1)$ and $B(0,R_3)$ are compact, one can apply Theorem 5.1 to c_λ and l_λ and obtain

$$\mathbb{P}\Big(\min_{w\in D} \mathcal{C}_\lambda(w) \le t\Big) \le 2\mathbb{P}\Big(\min_{w\in D} L_\lambda(w) \le t\Big) + 2\epsilon.$$

The corollary follows then from the fact one can take ϵ arbitrarily small. □

5.2. *Local stability.* In order to prove that (for instance) $\widehat{w}_\lambda$ verifies with high probability some property, let us say for instance that the empirical distribution of $(\widehat{\theta}_\lambda = \theta^\star + \widehat{w}_\lambda, \theta^\star)$ is close to $\mu_\lambda^\star$, we define a set $D_\epsilon \subset \mathbb{R}^N$ that contains all the vectors that do not verify this property, for example, $D_\epsilon = \{w \in \mathbb{R}^N | W_2(\widehat{\mu}_{(\theta^\star+w,\theta^\star)}, \mu_\lambda^*)^2 \ge \epsilon\}$, for some $\epsilon \in (0,1)$. The goal now is to prove that with high probability

$$\min_{w\in D} \mathcal{C}_\lambda(w) \ge \min_{w\in\mathbb{R}^N} \mathcal{C}_\lambda(w) + \epsilon,$$

for some $\epsilon > 0$. Using Gordon's min-max theorem (Corollary 5.1), we will be able to show

$$\mathbb{P}\Big(\min_{w\in D_\epsilon} \mathcal{C}_\lambda(w) \le \min_{w\in\mathbb{R}^N} \mathcal{C}_\lambda(w) + \epsilon\Big) \le 2\mathbb{P}\Big(\min_{w\in D_\epsilon} L_\lambda(w) \le \min_{w\in\mathbb{R}^N} L_\lambda(w) + \epsilon\Big) + o_N(1). \tag{5.5}$$

Informally, this is a consequence of the following two remarks. First, by applying parts (a) and (b) of Corollary 5.1 to the convex domain $\mathbb{R}^N$, we deduce that $\min_{w\in\mathbb{R}^N} \mathcal{C}_\lambda(w) \approx \min_{w\in\mathbb{R}^N} L_\lambda(w)$. Second, by applying part (a) to the closed domain D, we obtain $\min_{w\in D_\epsilon} \mathcal{C}_\lambda(w) \gtrsim \min_{w\in D_\epsilon} L_\lambda(w)$

It remains now to study the cost function L_λ, which is much simpler. This is done in Section B of [32]. The key step will be to establish the following "local stability" result (the next statement is an immediate consequence of Proposition B.1 and Theorem B.1 in the Supplementary Material [32]. We prove in fact that the cost function L_λ is strongly convex on a neighborhood of its minimizer.).

THEOREM 5.2. *The minimizer $w_\lambda^* = \arg\min_w L_\lambda(w)$ exists and is almost surely unique. Further, there exists constants $\gamma, c > 0$ that only depend on Ω such that for all $\theta^\star \in \mathcal{D}$, all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and all $\epsilon \in (0,1]$,*

$$\mathbb{P}\Big(\exists w \in \mathbb{R}^N, \frac{1}{N}\|w - w_\lambda^*\|^2 > \epsilon \text{ and } L_\lambda(w) \le \min_{v\in\mathbb{R}^N} L_\lambda(v) + \gamma\epsilon\Big) \le Ne^{-cN\epsilon^2}.$$

We do not obtain an equally strong result for the cost function $\mathcal{C}_\lambda(w)$, but we prove the following statement, which is sufficient for obtaining uniform control (for the sake of argument, we focus here on the domain $\mathcal{F}_p(\xi)$ and control of the empirical distribution).

THEOREM 5.3. *Assume that $\mathcal{D} = \mathcal{F}_p(\xi)$ for some ξ, $p > 0$. There exists constants $c, \gamma > 0$ that only depend on Ω such that for all $\epsilon \in (0, \frac{1}{2}]$,*

$$\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \sup_{\theta^\star\in\mathcal{D}} \mathbb{P}\big(\exists\theta\in\mathbb{R}^N, W_2\big(\widehat{\mu}_{(\theta,\theta^\star)}, \mu^*_\lambda\big)^2 \geq \epsilon \text{ and } \mathcal{L}_\lambda(\theta) \leq \min \mathcal{L}_\lambda + \gamma\epsilon\big) \leq N\exp\big(-cN\epsilon^a \log(\epsilon)^{-2}\big),$$

where $a = \frac{5}{2} + \frac{1}{p}$.

Theorem 5.3 is proved in Section C.1 of the Supplementary Material [32].

5.3. *Sketch of proof of main results.* For the sake of simplicity, we will illustrate the prove strategy by considering the empirical distribution of $\widehat{w}_\lambda = \widehat{\theta}_\lambda - \theta^\star$, as the argument is similar for other quantities. According to Theorem 3.1, this should be well approximated by $\overline{\mu}_\lambda$ that is the law of $\widehat{\Theta} - \Theta$, when $(\widehat{\Theta}, \Theta) \sim \mu^*_\lambda$; cf. Definition 3.3.

As anticipated, equation (5.5) and Theorem 5.2, allow one to control $W_2(\widehat{\mu}_{\widehat{w}_\lambda}, \overline{\mu}_\lambda)$ for a fixed λ ($\widehat{\mu}_{\widehat{w}_\lambda}$ denotes the empirical distribution of the entries of $\widehat{w}_\lambda$). Namely, we can define D_ε to be the set of vectors w such that $W_2(\widehat{\mu}_w, \overline{\mu}_\lambda) \geq \varepsilon > 0$. We then prove that the minimizer w^*_λ of L_λ has empirical distribution close to $\overline{\mu}_\lambda$ and, therefore, by Theorem 5.2, $L_\lambda(w) > L_\lambda(w^*_\lambda) + \gamma\epsilon$ for all $w \in D_\epsilon$, with high probability. This implies that the right-hand side of (5.5) is very small and we deduce that, with high probability, all minimizers or near minimizers of $\mathcal{C}_\lambda(w)$ have empirical distribution close to $\overline{\mu}_\lambda$,

We now would like to prove Theorem 3.1 and show that with high probability $\widehat{\mu}_{\widehat{w}_\lambda} \approx \overline{\mu}_\lambda$, uniformly in $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. To do so, we apply the above argument for $\lambda = \lambda_1, \dots, \lambda_k$, where $\lambda_1, \dots, \lambda_k$ is an ϵ-net of $[\lambda_{\min}, \lambda_{\max}]$. This implies that, with high probability for $\lambda \in \{\lambda_1, \dots, \lambda_k\}$, $W_2(\widehat{\mu}_{\widehat{w}_{\lambda_i}}, \overline{\mu}_{\lambda_i}) \leq \varepsilon$. Next, for $\lambda \in [\lambda_i, \lambda_{i+1}]$, we show that

$$\mathcal{C}_{\lambda_i}(\widehat{w}_\lambda) = \min_{w\in\mathbb{R}^N} \mathcal{C}_{\lambda_i}(w) + O\big(|\lambda_{i+1} - \lambda_i|\big).$$

Consequently, if $|\lambda_{i+1} - \lambda_i| = O(\epsilon)$ (using again equation (5.5) and Theorem 5.2), we obtain that $W_2(\widehat{\mu}_{\widehat{w}_\lambda}, \overline{\mu}_{\lambda_i}) = O(\epsilon)$ and, therefore, $W_2(\widehat{\mu}_{\widehat{w}_\lambda}, \overline{\mu}_\lambda) = O(\epsilon)$. We conclude that $W_2(\widehat{\mu}_{\widehat{w}_\lambda}, \overline{\mu}_\lambda) = O(\epsilon)$ for all $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, with high probability, which is the desired claim.

If the strategy exposed above allows one to obtain the risk of the Lasso and the empirical distribution of its coordinates, it is not enough to get its sparsity $\|\widehat{\theta}_\lambda\|_0$ or to obtain the empirical distribution of the debiased lasso

$$\widehat{\theta}^d_\lambda = \widehat{\theta}_\lambda + \frac{X^\mathsf{T}(y - X\widehat{\theta}_\lambda)}{1 - \frac{1}{n}\|\widehat{\theta}_\lambda\|_0}.$$

Therefore, we will need to analyze the vector

$$\widehat{v}_\lambda = \frac{1}{\lambda} X^\mathsf{T}(y - X\widehat{\theta}_\lambda),$$

which is a subgradient of the ℓ_1-norm at $\widehat{\theta}_\lambda$. We are able to study $\widehat{v}_\lambda$ using Gordon's min-max theorem because $\widehat{v}_\lambda$ is the unique maximizer of

$$v \mapsto \min_{w\in\mathbb{R}^N} \left\{ \frac{1}{2n}\|Xw - \sigma z\|^2 + \frac{\lambda}{n} v^\mathsf{T}\big(w + \theta^\star\big) \right\}.$$

The detailed analysis is done in Section E from [32].

Funding. This work was partially supported by grants NSF DMS-1613091, NSF CCF-1714305 and NSF IIS-1741162 and ONR N00014-18-1-2729.

SUPPLEMENTARY MATERIAL

Supplement to "The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning" (DOI: 10.1214/20-AOS2038SUPP; .pdf). Supplementary information containing omitted proofs.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[2] BARBIER, J., DIA, M., MACRIS, N. and KRZAKALA, F. (2016). The mutual information in random linear estimation. In 54*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 625–632. IEEE.

[3] BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2018). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. USA* **116** 5451–5460.

[4] BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[5] 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

[6] 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

[7] BELLEC, P. C. (2020). Out-of-sample error estimate for robust m-estimators with convex penalty. Available at arXiv:2008.11840.

[8] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[9] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[10] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[11] CAI, T. T. and GUO, Z. (2018). Accuracy assessment for high-dimensional linear regression. *Ann. Statist.* **46** 1807–1836. MR3819118 https://doi.org/10.1214/17-AOS1604

[12] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[13] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[14] CELENTANO, M., MONTANARI, A. and WEI, Y. (2020). The lasso with general gaussian designs with applications to hypothesis testing. Available at arXiv:2007.13716.

[15] CHEN, S. and DONOHO, D. L. (1995). Examples of basis pursuit. In *Wavelet Applications in Signal and Image Processing III* **2569** 564–575. International Society for Optics and Photonics.

[16] CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2016). On cross-validated lasso. Available at arXiv:1605.02214.

[17] 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

[18] DONOHO, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35** 617–652. MR2225676 https://doi.org/10.1007/s00454-005-1220-0

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. MR1278886 https://doi.org/10.1007/BF01199026

[20] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[21] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823

[22] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716 https://doi.org/10.1073/pnas.0502258102

[23] EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR2090899 https://doi.org/10.1198/016214504000000692

[24] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. Available at arXiv:1311.2445.

[25] 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

[26] GORDON, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50** 265–289. MR0800188 https://doi.org/10.1007/BF02759761

[27] 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

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

[29] 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

[30] JOHNSTONE, I. M. (2002). Function estimation and gaussian sequence models. Unpublished manuscript, 2(5.3):2.

[31] MA, J., XU, J. and MALEKI, A. (2019). Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Trans. Inf. Theory* **65** 3600–3629. MR3959008 https://doi.org/10.1109/TIT.2019.2893254

[32] MIOLANE, L. and MONTANARI, A. (2021). Supplement to "The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning." https://doi.org/10.1214/20-AOS2038SUPP

[33] MONTANARI, A. and RICHARD, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inf. Theory* **62** 1458–1484. MR3472260 https://doi.org/10.1109/TIT.2015.2457942

[34] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2017). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **45** 2427–2454. MR3737897 https://doi.org/10.1214/16-AOS1529

[35] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

[36] PANAHI, A. and HASSIBI, B. (2017). A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems* 3381–3390.

[37] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings* (*ISIT*), 2011 *IEEE International Symposium on* 2168–2172. IEEE, New York.

[38] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

[39] REEVES, G. and PFISTER, H. D. (2016). The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *Information Theory* (*ISIT*), 2016 *IEEE International Symposium on* 665–669. IEEE, New York.

[40] SCHNITER, P. and RANGAN, S. (2015). Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.* **63** 1043–1055. MR3311635 https://doi.org/10.1109/TSP.2014.2386294

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

[42] STOJNIC, M. (2010). Recovery thresholds for ℓ_1 optimization in binary compressed sensing. In *Information Theory Proceedings* (*ISIT*), 2010 *IEEE International Symposium on* 1593–1597. IEEE, New York.

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

[44] 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

[45] TAKAHASHI, T. and KABASHIMA, Y. (2018). A statistical mechanics approach to de-biasing and uncertainty estimation in LASSO for random measurements. *J. Stat. Mech. Theory Exp.* **7** 073405, 25. MR3845486 https://doi.org/10.1088/1742-5468/aace2e

[46] 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

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

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

[49] TIBSHIRANI, R. J. and TAYLOR, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948 https://doi.org/10.1214/12-AOS1003

[50] TROPP, J. A. (2015). Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance. Appl. Numer. Harmon. Anal.* 67–101. Birkhäuser/Springer, Cham. MR3467419

[51] 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

[52] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[53] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[54] 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