The Annals of Statistics
arXiv: arXiv:1511.01957

FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

By Weijie Su[*] and Małgorzata Bogdan[†] and Emmanuel Candès[‡]

University of Pennsylvania, University of Wroclaw, and Stanford University

In regression settings where explanatory variables have very low correlations and there are relatively few effects, each of large magnitude, we expect the Lasso to find the important variables with few errors, if any. This paper shows that in a regime of linear sparsity—meaning that the fraction of variables with a non-vanishing effect tends to a constant, however small—this cannot really be the case, even when the design variables are stochastically independent. We demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the effect sizes are. We derive a sharp asymptotic trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the Lasso path. This trade-off states that if we ever want to achieve a type II error (false negative rate) under a critical value, then anywhere on the Lasso path the type I error (false positive rate) will need to exceed a given threshold so that we can never have both errors at a low level at the same time. Our analysis uses tools from approximate message passing (AMP) theory as well as novel elements to deal with a possibly adaptive selection of the Lasso regularizing parameter.

1. Introduction.

Almost all data scientists know about and routinely use the Lasso [30, 31] to fit regression models. In the big data era, where the number p of explanatory variables often exceeds the number n of observational units, it may even supersede the method of least-squares. One appealing feature of the Lasso over earlier techniques such as ridge regression is that it automatically performs variable reduction, since it produces

[*] Supported in part by NSF under grant CCF-0963835.

[†] Supported in part by the European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement No. 602552 and co-financed by the Polish Ministry of Science and Higher Education under Grant Agreement 2932/7.PR/2013/2

[‡] Supported in part by NSF under grant CCF-0963835 and by the Math + X Award from the Simons Foundation.

MSC 2010 subject classifications: Primary 62F03; secondary 62J07, 62J05.

Keywords and phrases: Lasso, Lasso path, false discovery rate, false negative rate, power, approximate message passing (AMP), adaptive selection of parameters

models where lots of—if not most—regression coefficients are estimated to be exactly zero. In high-dimensional problems where p is either comparable to n or even much larger, the Lasso is believed to select those important variables out of a sea of potentially many irrelevant features.

Imagine we have an $n \times p$ design matrix $\boldsymbol{X}$ of features, and an n-dimensional response $\boldsymbol{y}$ obeying the standard linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z},$$

where $\boldsymbol{z}$ is a noise term. The Lasso is the solution to

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\operatorname{argmin}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \, \|\boldsymbol{b}\|_1; \tag{1.1}$$

if we think of the noise term as being Gaussian, we interpret it as a penalized maximum likelihood estimate, in which the fitted coefficients are penalized in an ℓ_1 sense, thereby encouraging sparsity. (There are nowadays many variants on this idea including ℓ_1-penalized logistic regression [31], elastic nets [40], graphical Lasso [36], adaptive Lasso [39], and many others.) As is clear from (1.1), the Lasso depends upon a regularizing parameter λ, which must be chosen in some fashion: in a great number of applications this is typically done via adaptive or data-driven methods; for instance, by cross-validation [15, 23, 37, 27]. Below, we will refer to the Lasso path as the family of solutions $\widehat{\boldsymbol{\beta}}(\lambda)$ as λ varies between 0 and ∞. We say that a variable j is selected at λ if $\widehat{\beta}_j(\lambda) \neq 0$.[1]

The Lasso is, of course, mostly used in situations where the true regression coefficient sequence is suspected to be sparse or nearly sparse. In such settings, researchers often believe—or, at least, wish—that as long as the true signals (the nonzero regression coefficients) are sufficiently strong compared to the noise level and the regressor variables weakly correlated, the Lasso with a carefully tuned value of λ will select most of the true signals while picking out very few, if any, noise variables. This belief is supported by theoretical asymptotic results discussed below, which provide conditions for perfect support recovery, i.e. for perfectly identifying which variables have a non-zero effect, see [35, 34, 29] for instance. Since these results guarantee that the Lasso works well in an extreme asymptotic regime, it is tempting to over-interpret what they actually say, and think that the Lasso will behave correctly in regimes of practical interest and offer some guarantees there as well. However, some recent works such as [18] have observed that the Lasso

[1] We also say that a variable j enters the Lasso path at λ_0 if there is there is $\varepsilon > 0$ such that $\widehat{\beta}_j(\lambda) = 0$ for $\lambda \in [\lambda_0 - \varepsilon, \lambda_0]$ and $\widehat{\beta}_j(\lambda) \neq 0$ for $\lambda \in (\lambda_0, \lambda_0 + \varepsilon]$. Similarly a variable is dropped at λ_0 if $\widehat{\beta}_j(\lambda) \neq 0$ for $\lambda \in [\lambda_0 - \varepsilon, \lambda_0)$ and $\widehat{\beta}_j(\lambda) = 0$ for $\lambda \in [\lambda_0, \lambda_0 + \varepsilon]$.

has problems in selecting the proper model in practical applications, and that false discoveries may appear very early on the Lasso path. This is the reason why [7, 6, 28] suggest that the Lasso should merely be considered as a *variable screener* rather than a *model selector*.

While the problems with the Lasso ordering of predictor variables are recognized, they are often attributed to (1) correlations between predictor variables, and (2) small effect sizes. In contrast, the novelty and message of our paper is that the selection problem also occurs when the signal-to-noise ratio is infinitely large (no noise) and the regressors are stochastically independent; we consider a random design $\boldsymbol{X}$ with independent columns, and as a result, all population correlations vanish (so the sample correlations are small). We also explain that this phenomenon is mainly due to the shrinkage of regression coefficients, and does not occur when using other methods, e.g. an ℓ_0 penalty in (1.1) rather than the ℓ_1 norm, compare Theorem 3.1 below.

Formally, we study the value of the false discovery proportion (FDP), the ratio between the number of false discoveries and the total number of discoveries, along the Lasso path.[2] This requires notions of true/false discoveries, and we pause to discuss this important point. In high dimensions, it is not a trivial task to define what are true and false discoveries, see e.g. [4, 20, 33, 19, 22]; these works are concerned with a large number of correlated regressors, where it is not clear which of these should be selected in a model. In response, we have selected to work in the very special case of *independent* regressors precisely to analyze a context where such complications do not arise and it is, instead, quite clear what true and false discoveries are. We classify a selected regressor $\boldsymbol{X}_j$ to be a false discovery if it is stochastically independent from the response, which in our setting is equivalent to $\beta_j = 0$. Indeed, under no circumstance can we say that that such a variable, which has zero explanatory power, is a true discovery.

Having clarified this point and as a setup for our theoretical findings, Figure 1 studies the performance of the Lasso under a 1010×1000 a random Gaussian design, where the entries of $\boldsymbol{X}$ are independent draws from $\mathcal{N}(0, 1)$. Set $\beta_1 = \cdots = \beta_{200} = 4$, $\beta_{201} = \cdots = \beta_{1000} = 0$ and the errors to be independent standard normals. Hence, we have 200 nonzero coefficients out of 1000 (a relatively sparse setting), and a very large signal-to-noise ratio (SNR). For instance, if we order the variables by the magnitude of the least-squares estimate, which we can run since $n = 1010 > 1000 = p$, then with probability practically equal to one, all the top 200 least-squares discoveries

[2] Similarly, the TPP is defined as the ratio between the number of true discoveries and that of potential true discoveries to be made.

correspond to true discoveries, i.e. variables for which $\beta_j = 4$. This is in sharp contrast with the Lasso, which selects null variables rather early. To be sure, when the Lasso includes half of the true predictors so that the false negative proportion falls below 50% or true positive proportion (TPP) passes the 50% mark, the FDP has already passed 8% meaning that we have already made 9 false discoveries. The FDP further increases to 19% the first time the Lasso model includes all true predictors, i.e. achieves full power (false negative proportion vanishes).

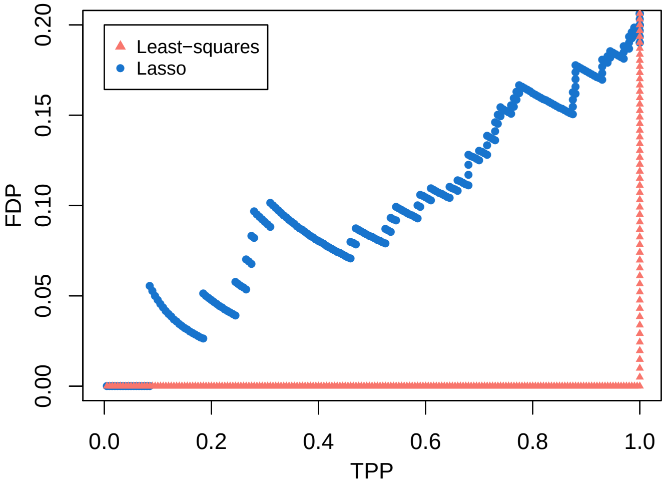


Fig 1: True positive and false positive rates along the Lasso path as compared to the ordering provided by the least-squares estimate.

Figure 2 provides a closer look at this phenomenon, and summarizes the outcomes from 100 independent experiments under the same Gaussian random design setting. In all the simulations, the first noise variable enters the Lasso model before 44% of the true signals are detected, and the last true signal is preceded by at least 22 and, sometimes, even 54 false discoveries. On average, the Lasso detects about 32 signals before the first false variable enters; to put it differently, the TPP is only 16% at the time the first false discovery is made. The average FDP evaluated the first time all signals are detected is 15%. For related empirical results, see e.g. [18].

The main contribution of this paper is to provide a quantitative description of this phenomenon in the asymptotic framework of *linear sparsity* defined below and previously studied e.g. in [3]. Assuming a random design with independent Gaussian predictors as above, we derive a fundamental Lasso trade-off between power (the ability to detect signals) and type I er-

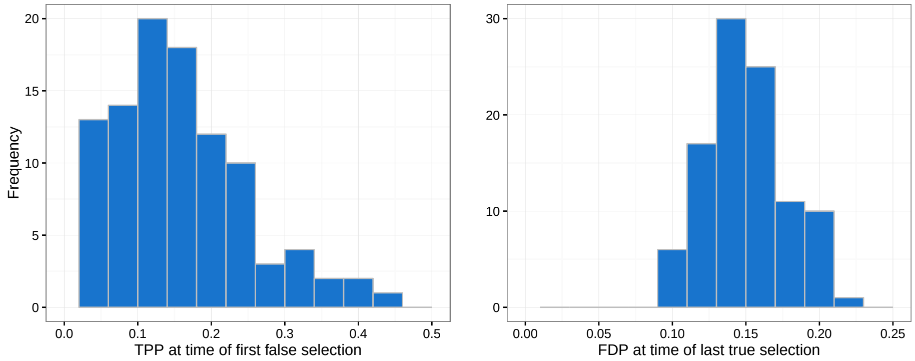


Fig 2: Left: power when the first false variable enters the Lasso model. Right: false discovery proportion the first time power reaches one (false negative proportion vanishes).

rors or, said differently, between the true positive and the false positive rates. This trade-off says that it is impossible to achieve high power and a low false positive rate simultaneously. Formally, we compute the formula for an exact boundary curve separating achievable $(\mathrm{TPP}, \mathrm{FDP})$ pairs from pairs that are impossible to achieve no matter the value of the signal-to-noise ratio (SNR). Hence, we prove that there is a whole favorable region in the $(\mathrm{TPP}, \mathrm{FDP})$ plane that cannot be reached, see Figure 3 for an illustration.

2. The Lasso Trade-off Diagram.

2.1. *Linear sparsity and the working model.* We mostly work in the setting of [3], which specifies the design $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, the parameter sequence $\boldsymbol{\beta} \in \mathbb{R}^p$ and the errors $\boldsymbol{z} \in \mathbb{R}^n$. The design matrix $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries so that the columns are approximately normalized, and the errors z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$, where σ is fixed but otherwise arbitrary. Note that we do not exclude the value $\sigma = 0$ corresponding to noiseless observations. The regression coefficients $\beta_1, \ldots, \beta_p$ are independent copies of a random variable Π obeying $\mathbb{E}\,\Pi^2 < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$ for some constant ϵ. For completeness, $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{z}$ are all independent from each other. As in [3], we are interested in the limiting case where $p, n \to \infty$ with $n/p \to \delta$ for some positive constant δ. A few comments are in order.

Linear sparsity. The first concerns the degree of sparsity. In our model, the expected number of nonzero regression coefficients is linear in p and equal to $\epsilon \cdot p$ for some $\epsilon > 0$. Hence, this model excludes a form of asymptotics discussed in [35, 34, 29], for instance, where the fraction of nonzero coefficients vanishes in the limit of large problem sizes. Specifically, our results do not contradict asymptotic results from [35] predicting perfect support recovery in an asymptotic regime, where the number of k of variables in the model obeys $k/p \le \delta/(2\log p) \cdot (1 + o(1))$ and the effect sizes all grow like $c \cdot \sigma\sqrt{2\log p}$, where c is an unknown numerical constant. The merit of the linear sparsity regime lies in the fact that our theory makes accurate predictions when describing the performance of the Lasso in practical settings with moderately large dimensions and reasonable values of the degree of sparsity, including rather sparse signals. The precision of these predictions is illustrated in Figure 5 and in Section 4. In the latter case, $n = 250$, $p = 1000$ and the number of k of signals is very small, i.e. $k = 18$.

Gaussian designs. Second, Gaussian designs with independent columns are believed to be "easy" or favorable for model selection due to weak correlations between distinct features. (Such designs happen to obey restricted isometry properties [8] or restricted eigenvalue conditions [5] with high probability, which have been shown to be useful in settings sparser than those considered in this paper.) Hence, negative results under the working hypothesis are likely to extend more generally.

Regression coefficients. Third, the assumption concerning the distribution of the regression coefficients can be slightly weakened: all we need is that the sequence $\beta_1, \ldots, \beta_p$ has a convergent empirical distribution with bounded second moment. We shall not pursue this generalization here.

2.2. *Main result.* Throughout the paper, V (resp. T) denotes the number of Lasso false (resp. true) discoveries while $k = |\{j : \beta_j \neq 0\}|$ denotes the number of true signals; formally, $V(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j = 0\}|$ whereas $T(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j \neq 0\}|$. With this, we define the FDP as usual,

$$\text{FDP}(\lambda) = \frac{V(\lambda)}{|\{j : \widehat{\beta}_j(\lambda) \neq 0\}| \vee 1} \tag{2.1}$$

and, similarly, the TPP is defined as

$$\text{TPP}(\lambda) = \frac{T(\lambda)}{k \vee 1} \tag{2.2}$$

(above, $a \vee b = \max\{a, b\}$). The dependency on λ shall often be suppressed when clear from the context. Our main result provides an explicit trade-off between FDP and TPP.

Theorem 2.1. *Fix $\delta \in (0, \infty)$ and $\epsilon \in (0, 1)$, and consider the function $q^\star(\cdot) = q^\star(\cdot; \delta, \epsilon) > 0$ given in (2.4). Then under the working hypothesis and for any arbitrary small constants $\lambda_0 > 0$ and $\eta > 0$, the following conclusions hold:*

(a) In the **noiseless case** *($\sigma = 0$), the event*

$$\bigcap_{\lambda \geq \lambda_0} \left\{ \mathrm{FDP}(\lambda) \geq q^\star\left(\mathrm{TPP}(\lambda)\right) - \eta \right\} \tag{2.3}$$

holds with probability tending to one. (The lower bound on λ in (2.3) does not impede interpretability since we are not interested in variables entering the path last.)

(b) With **noisy data** *($\sigma > 0$) the conclusion is exactly the same as in (a).*

(c) Therefore, in both the noiseless and noisy cases, no matter how we choose $\widehat{\lambda}(\boldsymbol{y}, \boldsymbol{X}) \geq c_1$ **adaptively** *by looking at the response $\boldsymbol{y}$ and design $\boldsymbol{X}$, with probability tending to one we will never have $\mathrm{FDP}(\widehat{\lambda}) < q^\star(\mathrm{TPP}(\widehat{\lambda})) - c_2$.*

(d) The boundary curve $q^\star$ is **tight***: any continuous curve $q(u) \geq q^\star(u)$ with strict inequality for some u will fail (a) and (b) for some prior distribution Π on the regression coefficients.*

A different way to phrase the trade-off is via false discovery and false negative rates. Here, the FDP is a natural measure of type I error while $1-\mathrm{TPP}$ (often called the false negative proportion) is the fraction of missed signals, a natural notion of type II error. In this language, our results simply say that nowhere on the Lasso path can both types of error rates be simultaneously low.

Remark 1. We would like to emphasize that the boundary is derived from a best-case point of view. For a fixed prior Π, we also provide in Theorem D.2 from Appendix D a trade-off curve q^Π between TPP and FDP, which always lies above the boundary $q^\star$. Hence, the trade-off is of course less favorable when we deal with a specific Lasso problem. In fact, $q^\star$ is nothing else but the lower envelope of all the instance-specific curves q^Π with $\mathbb{P}(\Pi \neq 0) = \epsilon$.

Figure 3 presents two instances of the *Lasso trade-off diagram*, where the curve $q^\star(\cdot)$ separates the red region, where both type I and type II errors are small, from the rest (the white region). Looking at this picture, Theorem 2.1 says that nowhere on the Lasso path we will find ourselves in the red region, and that this statement continues to hold true even when there is no noise. Our theorem also says that we cannot move the boundary upward. As we shall see, we can come arbitrarily close to any point on the curve by specifying a prior Π and a value of λ. Note that the right plot is vertically truncated at 0.6791, implying that TPP cannot even approach 1 in the regime of $\delta = 0.3, \epsilon = 0.15$. This upper limit is where the Donoho-Tanner phase transition occurs [14], see the discussion in Section 2.6 and Appendix C.

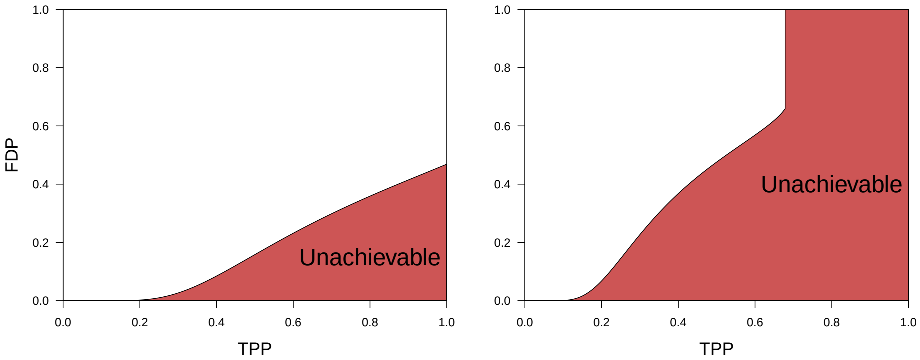


Fig 3: The Lasso trade-off diagram: left is with $\delta = 0.5$ and $\epsilon = 0.15$, and right is with $\delta = 0.3$ and $\epsilon = 0.15$ (the vertical truncation occurs at 0.6791).

Support recovery from noiseless data is presumably the most ideal scenario. Yet, the trade-off remains the same as seen in the first claim of the theorem. As explained in Section 3, this can be understood by considering that the root cause underlying the trade-off in both the noiseless and noisy cases come from the pseudo-noise introduced by shrinkage.

2.3. *The boundary curve $q^\star$.* We now turn to specify $q^\star$. For a fixed u, let $t^\star(u)$ be the largest positive root[3] of the equation in t,

$$\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2t\phi(t)\right]} = \frac{1-u}{1-2\Phi(-t)}.$$

Then

$$q^\star(u;\delta,\epsilon) = \frac{2(1-\epsilon)\Phi(-t^\star(u))}{2(1-\epsilon)\Phi(-t^\star(u)) + \epsilon u}. \tag{2.4}$$

[3] If $u = 0$, treat $+\infty$ as a root of the equation, and in (2.4) conventionally set $0/0 = 0$. In the case where $\delta \geq 1$, or $\delta < 1$ and ϵ is no larger than a threshold determined only by δ, the range of u is the unit interval $[0, 1]$. Otherwise, the range of u is the interval with endpoints 0 and some number strictly smaller than 1, see the discussion in Appendix C.

It can be shown that this function is infinitely many times differentiable over its domain, always strictly increasing, and vanishes at $u = 0$. Matlab code to calculate $q^\star$ is available at `https://github.com/wjsu/fdrlasso`.

Figure 4 displays examples of the function $q^\star$ for different values of ϵ (sparsity), and δ (dimensionality). It can be observed that the issue of FDR control becomes more severe when the sparsity ratio $\epsilon = k/p$ increases and the dimensionality $1/\delta = p/n$ increases.

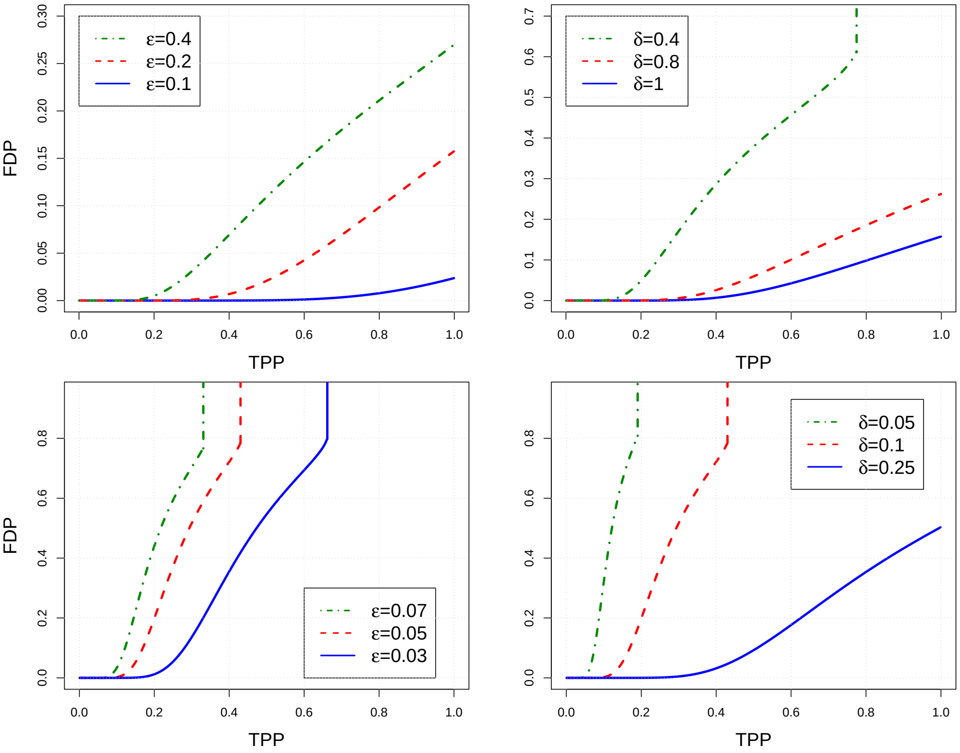
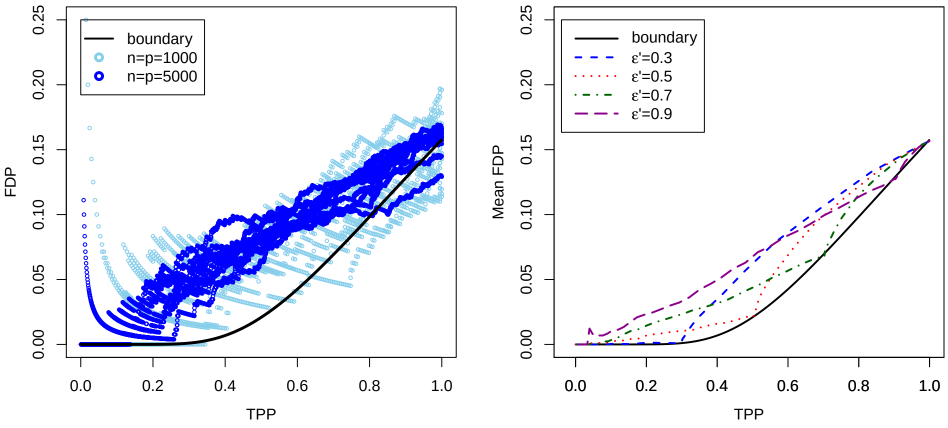


Fig 4: Top-left is with $\delta = 1$; top-right is with $\epsilon = 0.2$; bottom-left is with $\delta = 0.1$; and bottom-right is with $\epsilon = 0.05$.

2.4. *Numerical illustration.* Figure 5 provides the outcomes of numerical simulations for finite values of n and p in the noiseless setup where $\sigma = 0$. For each of $n = p = 1000$ and $n = p = 5000$, we compute 10 independent Lasso paths and plot all pairs $(\mathrm{TPP}, \mathrm{FDP})$ along the way. In Figure 5a we can see that when $\mathrm{TPP} < 0.8$, then the large majority of pairs $(\mathrm{TPP}, \mathrm{FDP})$ along these 10 paths are above the boundary. When TPP approaches one, the average FDP becomes closer to the boundary and a fraction of the paths fall below the line. As expected this proportion is substantially smaller for the larger problem size.



(a) Realized (TPP, FDP) pairs **(b)** Sharpness of the boundary

Fig 5: In both (a) and (b), $n/p = \delta = 1$, $\epsilon = 0.2$, and the noise level is $\sigma = 0$ (noiseless). (a) FDP vs. TPP along 10 independent Lasso paths with $\mathbb{P}(\Pi = 50) = 1 - \mathbb{P}(\Pi = 0) = \epsilon$. (b) Mean FDP vs. mean TPP averaged at different values of λ over 100 replicates for $n = p = 1000$, $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ as before, and $\mathbb{P}(\Pi = 50|\Pi \neq 0) = 1 - \mathbb{P}(\Pi = 0.1|\Pi \neq 0) = \epsilon'$.

2.5. *Sharpness.* The last conclusion from the theorem stems from the following fact: take any point $(u, q^\star(u))$ on the boundary curve; then we can approach this point by fixing $\epsilon' \in (0, 1)$ and setting the prior to be

$$\Pi = \begin{cases} M, & \text{w.p. } \epsilon \cdot \epsilon', \\ M^{-1}, & \text{w.p. } \epsilon \cdot (1 - \epsilon'), \\ 0, & \text{w.p. } 1 - \epsilon. \end{cases}$$

We think of M as being very large so that the (nonzero) signals are either very strong or very weak. In Appendix C, we prove that for any u between 0 and 1 there is some fixed $\epsilon' = \epsilon'(u) > 0$ such that[4]

$$\lim_{M \to \infty} \lim_{n,p \to \infty} (\mathrm{TPP}(\lambda), \mathrm{FDP}(\lambda)) \to (u, q^\star(u)), \tag{2.5}$$

[4] In some cases u should be bounded above by some constant strictly smaller than 1. See the previous footnote for details.

where convergence occurs in probability. This holds provided that $\lambda \to \infty$ in such a way that $M/\lambda \to \infty$; e.g. $\lambda = \sqrt{M}$. Hence, the most favorable configuration is when the signal is a mixture of very strong and very weak effect sizes because weak effects cannot be counted as false positives, thus reducing the FDP.

Figure 5b provides an illustration of (2.5). The setting is as in Figure 5a with $n = p = 1000$ and $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ except that, here, conditionally on being nonzero the prior takes on the values 50 and 0.1 with probability $\epsilon' \in \{0.3, 0.5, 0.7, 0.9\}$ and $1 - \epsilon'$, respectively, so that we have a mixture of strong and weak signals. We observe that the true/false positive rate curve nicely touches *only* one point on the boundary depending on the proportion ϵ' of strong signals .

2.6. *Technical novelties and comparisons with other works.* The proof of Theorem 2.1 is built on top of the approximate message passing (AMP) theory developed in [11, 2, 1], and requires nontrivial extensions. AMP was originally designed as an algorithmic solution to compressive sensing problems under random Gaussian designs. In recent years, AMP has also found applications in robust statistics [12, 13], structured principal component analysis [10, 25], and the analysis of the stochastic block model [9]. Having said this, AMP theory is of crucial importance to us because it turns out to be a very useful technique to rigorously study various statistical properties of the Lasso solution whenever we employ a *fixed* value of the regularizing parameter λ [3, 24, 26].

There are, however, major differences between our work and AMP research. First and foremost, our paper is concerned with situations where λ is selected adaptively, i.e. from the data; this is clearly outside of the envelope of current AMP results. Second, we are also concerned with situations where the noise variance can be zero. Likewise, this is outside of current knowledge. These differences are significant and as far as we know, our main result cannot be seen as a straightforward extension of AMP theory. In particular, we introduce a host of novel elements to deal, for instance, with the *irregularity* of the Lasso path. The irregularity means that a variable can enter and leave the model multiple times along the Lasso path [16, 32] so that natural sequences of Lasso models are not nested. This implies that a naive application of sandwiching inequalities does not give the type of statements holding uniformly over all λ's that we are looking for.

Instead, we develop new tools to understand the "continuity" of the support of $\widehat{\boldsymbol{\beta}}(\lambda)$ as a function of λ. Since the support can be characterized by the Karush-Kuhn-Tucker (KKT) conditions, this requires establishing some sort of continuity of the KKT conditions. Ultimately, we shall see that this comes down to understanding the maximum distance—uniformly over λ and λ'—between Lasso estimates $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\boldsymbol{\beta}}(\lambda')$ at close values of the regularizing parameter. A complete statement of this important intermediate result is provided in Lemma B.2 from Appendix B.

Our results can also be compared to the phase-transition curve from [14], which was obtained under the same asymptotic regime and describes conditions for perfect signal recovery in the noiseless case. The solution algorithm there is the linear program, which minimizes the ℓ_1 norm of the fitted coefficients under equality constraints, and corresponds to the Lasso solution in the limit of $\lambda \to 0$ (the end or bottom of the Lasso path). The conditions for perfect signal recovery by the Lasso turn out to be far more restrictive than those related to this linear program. For example, our FDP-TPP trade-off curves show that perfect recovery of an infinitely large signal by Lasso is often practically impossible even when $n \geq p$ (see Figure 4). Interestingly, the phase-transition curve also plays a role in describing the performance of the Lasso, since it turns out that for signals dense enough not to be recovered by the linear program, not only does the Lasso face the problem of early false discoveries, it also hits a power limit for arbitrary small values of λ (see the discussion in Appendix C).

Finally, we would like also to point out that some existing works have investigated support recovery in regimes including linear sparsity under random designs (see e.g. [34, 29]). These interesting results were, however, obtained by taking an information-theoretic point of view and do not apply to computationally feasible methods such as the Lasso.

3. What's Wrong with Shrinkage?.

3.1. *Performance of ℓ_0 methods.* We wrote earlier that not all methods share the same difficulties in identifying those variables in the model. If the signals are sufficiently strong, some other methods, perhaps with exponential computational cost, can achieve good model selection performance, see e.g. [29]. As an example, consider the simple ℓ_0-penalized maximum likelihood estimate,

$$\widehat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \, \|\boldsymbol{b}\|_0. \tag{3.1}$$

Methods known as AIC, BIC and RIC (short for risk inflation criterion) are all of this type and correspond to distinct values of the regularizing parameter λ. It turns out that such fitting strategies can achieve perfect separation in some cases.

THEOREM 3.1. *Under our working hypothesis, take $\epsilon < \delta$ for identifiability, and consider the two-point prior*

$$\Pi = \begin{cases} M, & w.p.\ \epsilon, \\ 0, & w.p.\ 1-\epsilon. \end{cases}$$

Then we can find $\lambda(M)$ such that in probability, the discoveries of the ℓ_0 estimator (3.1) *obey*

$$\lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{FDP} = 0 \quad \textit{and} \quad \lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{TPP} = 1.$$

The proof of the theorem is in Appendix E. Similar conclusions will certainly hold for many other non-convex methods, including SCAD and MC+ with properly tuned parameters [17, 38].

3.2. *Some heuristic explanation.* In light of Theorem 3.1, we pause to discuss the cause underlying the limitations of the Lasso for variable selection, which comes from the pseudo-noise introduced by shrinkage. As is well-known, the Lasso applies some form of soft-thresholding. This means that when the regularization parameter λ is large, the Lasso estimates are seriously biased downwards. Another way to put this is that the residuals still contain much of the effects associated with the selected variables. This can be thought of as extra noise that we may want to call *shrinkage noise*. Now as many strong variables get picked up, the shrinkage noise gets inflated and its projection along the directions of some of the null variables may actually dwarf the signals coming from the strong regression coefficients; this is why null variables get picked up. Although our exposition below dramatically lacks in rigor, it nevertheless formalizes this point in some *qualitative* fashion. It is important to note, however, that this phenomenon occurs in the linear sparsity regime considered in this paper so that we have sufficiently many variables for the shrinkage noise to build up and have a fold on other variables that becomes competitive with the signal. In contrast, under extreme sparsity and high SNR, both type I and II errors can be controlled at low levels, see e.g. [21].

For simplicity, we fix the true support $\mathcal{T}$ to be a deterministic subset of size $\epsilon \cdot p$, each nonzero coefficient in $\mathcal{T}$ taking on a constant value $M > 0$. Also, assume $\delta > \epsilon$. Finally, since the noiseless case $\boldsymbol{z} = \boldsymbol{0}$ is conceptually perhaps the most difficult, suppose $\sigma = 0$. Consider the reduced Lasso problem first:

$$\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda) = \operatorname*{argmin}_{\boldsymbol{b}_{\mathcal{T}} \in \mathbb{R}^{\epsilon p}} \ \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\mathcal{T}}\|^2 + \lambda\,\|\boldsymbol{b}_{\mathcal{T}}\|_1.$$

This (reduced) solution $\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda)$ is independent from the other columns $\boldsymbol{X}_{\overline{\mathcal{T}}}$ (here and below $\overline{\mathcal{T}}$ is the complement of $\mathcal{T}$). Now take λ to be of the same magnitude as M so that roughly half of the signal variables are selected. The KKT conditions state that

$$-\lambda \mathbf{1} \le \boldsymbol{X}_{\mathcal{T}}^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}}) \le \lambda \mathbf{1},$$

where $\mathbf{1}$ is the vectors of all ones. Note that if $|\boldsymbol{X}_j^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}})| \le \lambda$ for all $j \in \overline{\mathcal{T}}$, then extending $\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda)$ with zeros would be the solution to the full Lasso problem—with all variables included as potential predictors—since it would obey the KKT conditions for the full problem. A first simple fact is this: for $j \in \overline{\mathcal{T}}$, if

$$|\boldsymbol{X}_j^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}})| > \lambda, \tag{3.2}$$

then $\boldsymbol{X}_j$ must be selected by the incremental Lasso with design variables indexed by $\mathcal{T} \cup \{j\}$. Now we make an assumption which is heuristically reasonable: any j obeying (3.2) has a reasonable chance to be selected in the full Lasso problem with the same λ (by this, we mean with some probability bounded away from zero). We argue in favor of this heuristic later.

Following our heuristic, we would need to argue that (3.2) holds for a number of variables in $\overline{\mathcal{T}}$ *linear* in p. Write

$$\boldsymbol{X}_{\mathcal{T}}^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}}) = \boldsymbol{X}_{\mathcal{T}}^\top (\boldsymbol{X}_{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{T}} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}}) = \lambda \boldsymbol{g}_{\mathcal{T}},$$

where $\boldsymbol{g}_{\mathcal{T}}$ is a subgradient of the ℓ_1 norm at $\widehat{\boldsymbol{\beta}}_{\mathcal{T}}$. Hence, $\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}} = \lambda (\boldsymbol{X}_{\mathcal{T}}^\top \boldsymbol{X}_{\mathcal{T}})^{-1} \boldsymbol{g}_{\mathcal{T}}$ and

$$\boldsymbol{X}_{\mathcal{T}} (\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}}) = \lambda \boldsymbol{X}_{\mathcal{T}} (\boldsymbol{X}_{\mathcal{T}}^\top \boldsymbol{X}_{\mathcal{T}})^{-1} \boldsymbol{g}_{\mathcal{T}}.$$

Since $\delta > \epsilon$, $\boldsymbol{X}_{\mathcal{T}} (\boldsymbol{X}_{\mathcal{T}}^\top \boldsymbol{X}_{\mathcal{T}})^{-1}$ has a smallest singular value bounded away from zero (since $\boldsymbol{X}_{\mathcal{T}}$ is a fixed random matrix with more rows than columns). Now because we make about half discoveries, the subgradient takes on the value one (in magnitude) at about $\epsilon \cdot p/2$ times. Hence, with high probability,

$$\|\boldsymbol{X}_{\mathcal{T}} (\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}})\| \ge \lambda \cdot c_0 \cdot \|\boldsymbol{g}_{\mathcal{T}}\| \ge \lambda \cdot c_1 \cdot p$$

for some constants c_0, c_1 depending on ϵ and δ.

Now we use the fact that $\widehat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda)$ is independent of $\boldsymbol{X}_{\overline{\mathcal{T}}}$. For any $j \notin \mathcal{T}$, it follows that

$$\boldsymbol{X}_j^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}} \widehat{\boldsymbol{\beta}}_{\mathcal{T}}) = \boldsymbol{X}_j^\top \boldsymbol{X}_{\mathcal{T}} (\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}})$$

is conditionally normally distributed with mean zero and variance

$$\frac{\|\boldsymbol{X}_{\mathcal{T}} (\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}})\|^2}{n} \ge \frac{c_1 \lambda^2 p}{n} = c_2 \cdot \lambda^2.$$

In conclusion, the probability that $\boldsymbol{X}_j^\top(\boldsymbol{y}-\boldsymbol{X}_{\mathcal{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{T}})$ has absolute value larger than λ is bounded away from 0. Since there are $(1-\epsilon)p$ such j's, their expected number is linear in p. This implies that by the time half of the true variables are selected, we already have a non-vanishing FDP. Note that when $|\mathcal{T}|$ is not linear in p but smaller, e.g. $|\mathcal{T}| \le c_0 n/\log p$ for some sufficiently small constant c_0, the variance is much smaller because the estimation error $\|\boldsymbol{X}_{\mathcal{T}}(\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\beta}}_{\mathcal{T}})\|^2$ is much lower, and this phenomenon does not occur.

Returning to our heuristic, we make things simpler by considering alternatives: (a) if very few extra variables in $\overline{\mathcal{T}}$ were selected by the full Lasso, then the value of the prediction $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ would presumably be close to that obtained from the reduced model. In other words, the residuals $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ from the full problem should not differ much from those in the reduced problem. Hence, for any j obeying (3.2), $\boldsymbol{X}_j$ would have a high correlation with $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Thus this correlation has a good chance to be close to λ, or actually be equal to λ. Equivalently, $\boldsymbol{X}_j$ would likely be selected by the full Lasso problem. (b) If on the other hand, the number of variables selected from $\overline{\mathcal{T}}$ by the full Lasso were a sizeable proportion of $|\mathcal{T}|$, we would have lots of false discoveries, which is our claim.

In a more rigorous way, AMP claims that under our working hypothesis, the Lasso estimates $\widehat{\beta}_j(\lambda)$ are, in a certain sense, asymptotically distributed as $\eta_{\alpha\tau}(\beta_j+\tau W_j)$ for most j and W_j's independently drawn from $\mathcal{N}(0,1)$. The positive constants α and τ are uniquely determined by a pair of nonlinear equations parameterized by $\epsilon, \delta, \Pi, \sigma^2$, and λ. Suppose as before that all the nonzero coefficients of $\boldsymbol{\beta}$ are large in magnitude, say they are all equal to M. When about half of them appear on the path, we have that λ is just about equal to M. A consequence of the AMP equations is that τ is also of this order of magnitude. Hence, under the null we have that $(\beta_j+\tau W_j)/M \sim \mathcal{N}(0,(\tau/M)^2)$ while under the alternative, it is distributed as $\mathcal{N}(1,(\tau/M)^2)$. Because, τ/M is bounded away from zero, we see that false discoveries are bound to happen.

Variants of the Lasso and other ℓ_1-penalized methods, including ℓ_1-penalized logistic regression and the Dantzig selector, also suffer from this "shrinkage to noise" issue.

4. Discussion.

We have evidenced a clear trade-off between false and true positive rates under the assumption that the design matrix has i.i.d. Gaussian entries. It is likely that there would be extensions of this result to designs with general i.i.d. sub-Gaussian entries as strong evidence suggests that the AMP theory may be valid for such larger classes, see [1]. It might also be of interest to study the Lasso trade-off diagram under correlated random designs.

As we previously mentioned in the introduction, a copious body of literature considers the Lasso support recovery under Gaussian random designs, where the sparsity of the signal is often assumed to be *sub-linear* in the ambient dimension p. Recall that if all the nonzero signal components have magnitudes at least $c\sigma\sqrt{2\log p}$ for some unspecified numerical constant c (which would have to exceed one), the results from [35] conclude that, asymptotically, a sample size of $n \ge (2 + o(1))k \log p$ is both necessary and sufficient for the Lasso to obtain perfect support recovery. What does these results say for finite values of n and p? Figure 6 demonstrates the performance of the Lasso under a moderately large 250×1000 random Gaussian design. Here, we consider a very sparse signal, where only $k = 18$ regression coefficients are nonzero, $\beta_1 = \cdots = \beta_{18} = 2.5\sqrt{2\log p} \approx 9.3$, $\beta_{19} = \cdots = \beta_{1000} = 0$, and the noise variance is $\sigma^2 = 1$. Since $k = 18$ is smaller than $n/2\log p$ and β is substantially larger than $\sqrt{2\log p}$ one might expect that Lasso would recover the signal support. However, Figure 1 (left) shows that this might not be the case. We see that the Lasso includes five false discoveries before all true predictors are included, which leads to an FDP of 21.7% by the time the power (TPP) reaches 1. Figure 6 (right) summarizes the outcomes from 500 independent experiments, and shows that the average FDP reaches 13.4% when TPP = 1. With these dimensions, perfect recovery is not guaranteed even in the case of 'infinitely' large signals (no noise). In this case, perfect recovery occurs in only 75% of all replicates and the averaged FDP at the point of full power is equal to 1.7%, which almost perfectly agrees with the boundary FDP provided in Theorem 2.1. Thus, quite surprisingly, our results obtained under a *linear sparsity regime* apply to sparser regimes, and might prove useful across a wide range of sparsity levels.

Of concern in this paper are statistical properties regarding the number of true and false discoveries along the Lasso path but it would also be interesting to study perhaps finer questions such as this: when does the first noise variable get selected? Consider Figure 7: there, $n = p = 1000$, $\sigma^2 = 1$, $\beta_1 = \cdots = \beta_k = 50$ (very large SNR) and k varies from 5 to 150. In the very low sparsity regime, all the signal variables are selected before any noise variable. When the number k of signals increases we observe early false discoveries, which may occur for values of k smaller than $n/(2\log p)$. However, the average rank of the first false discovery is substantially smaller than k only after k exceeds $n/(2\log p)$. Then it keeps on decreasing as k continues to increase, a phenomenon not explained by any result we are aware of. In the linear sparsity regime, it would be interesting to derive a prediction for the average time of the first false entry, at least in the noiseless case.

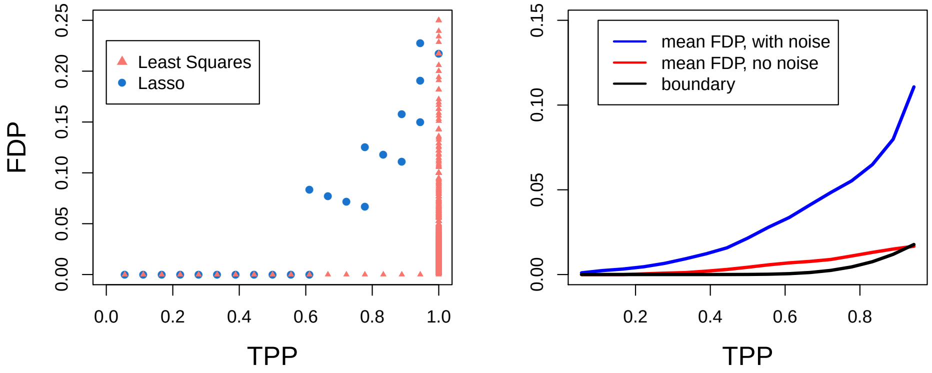


Fig 6: Simulation setup: $n = 250$, $p = 1000$, $\beta_1 = \cdots = \beta_{18} = 2.5\sqrt{2\log p} \approx 9.3$ (the other coefficients all vanish), $\sigma^2 = 1$ (with noise) and $\sigma^2 = 0$ (no noise). Left: noisy case. True positive and false positive rates along a single realization of the Lasso path. The least squares path is obtained by ordering least squares estimates from a model including the first 50 variables selected by the Lasso. Right: mean FDP as a function of TPP. The mean FDP was obtained by averaging over 500 independent trials.

Methods that are computationally efficient and also enjoy good model performance in the linear sparsity regime would be highly desirable. (The Lasso and the ℓ_0 method each enjoys one property but not the other.) While it is beyond our scope to address this problem, we conclude the discussion by considering marginal regression, a technique widely used in practice and not computationally demanding. A simple analysis shows that marginal regression suffers from the same issue as the Lasso under our working hypothesis. To show this, examine the noiseless case ($\sigma = 0$) and assume $\beta_1 = \cdots = \beta_k = M$ for some constant $M > 0$. It is easy to see that the marginal statistic $\boldsymbol{X}_j^\top \boldsymbol{y}$ for the jth variable is asymptotically distributed as $\mathcal{N}(0, \tilde{\sigma}^2)$, where $\tilde{\sigma} = M\sqrt{(k-1)/n}$, if j is a true null and $\mathcal{N}(M, \tilde{\sigma}^2)$ otherwise. In the linear sparsity regime, where k/n tends to a constant, the mean shift M and standard deviation $\tilde{\sigma}$ have comparable magnitudes. As a result, nulls and non-nulls are also interspersed on the marginal regression path, so that we would have either high FDR or low power.

Acknowledgements. W. S. would like to thank Andrea Montanari for helpful discussions. We thank Yuxin Chen for helpful comments about an early version of the manuscript.

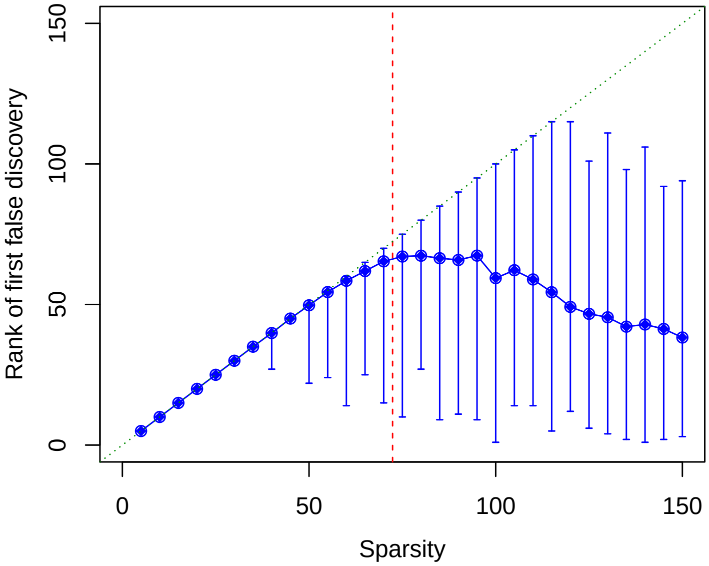


Fig 7: Rank of the first false discovery. Here, $n = p = 1000$ and $\beta_1 = \cdots = \beta_k = 50$ for k ranging from 5 to 150 ($\beta_i = 0$ for $i > k$). We plot averages from 100 independent replicates and display the range between minimal and maximal realized values. The vertical line is placed at $k = n/(2 \log p)$ and the 45° line passing through the origin is shown for convenience.

SUPPLEMENTARY MATERIAL

Supplement to "False Discoveries Occur Early on the Lasso Path"
(; .pdf). The supplementary materials contain proofs of some technical results in this paper.

References.

[1] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[2] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory*, 57(2):764–785, 2011.

[3] M. Bayati and A. Montanari. The Lasso risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

[4] R. Berk, B. Lawrence, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[6] P. Bühlmann. Invited discussion on "regression shrinkage and selection via the lasso: a retrospective (r. tibshirani)". *Journal of the Royal Statistical Society: Series B*, 73:277–279, 2011.

[7] P. Bühlmann and van de Geer. *Statistics for High-dimensional Data*. Springer, New York, 2011.

[8] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, Dec. 2005.

[9] Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

[10] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201, 2014.

[11] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[12] D. L. Donoho and A. Montanari. High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *arXiv preprint arXiv:1310.7320*, 2013.

[13] D. L. Donoho and A. Montanari. Variance breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. *arXiv preprint arXiv:1503.02106*, 2015.

[14] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Trans. R. Soc. A*, 367(1906):4273–4293, 2009.

[15] L. S. Eberlin et al. Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proceedings of the National Academy of Sciences*, 111(7):2436–2441, 2014.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[18] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[19] R. Foygel Barber and E. J. Candes. A knockoff filter for high-dimensional selective inference. *ArXiv e-prints*, Feb. 2016.

[20] M. G'Sell, T. Hastie, and R. Tibshirani. False variable selection rates in regression. *ArXiv e-prints*, 2013.

[21] P. Ji and Z. Zhao. Rate optimal multiple testing procedure in high-dimensional regression. *arXiv preprint arXiv:1404.2961*, 2014.

[22] J. D. Lee, D. L. Sun, S. Y., and T. J. E. Exact post-selection inference, with application to the Lasso. *Annals of Statistics*, 44(2):802–837, 2016.

[23] W. Lee et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics*, 39(10):1235–1244, 2007.

[24] A. Maleki, L. Anitori, Z. Yang, and R. Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Trans. Inform. Theory*, 59(7):4290–4308, 2013.

[25] A. Montanari and E. Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *arXiv preprint arXiv:1406.4775*, 2014.

[26] A. Mousavi, A. Maleki, and R. G. Baraniuk. Asymptotic analysis of LASSO's solution path with implications for approximate message passing. *arXiv preprint arXiv:1309.5979*, 2013.

[27] F. S. Paolo, H. A. Fricker, and L. Padman. Volume loss from Antarctic ice shelves is

accelerating. *Science*, 348(6232):327–331, 2015.

[28] P. Pokarowski and J. Mielniczuk. Combined ℓ_1 and greedy ℓ_0 penalized least squares for linear model selection. *Journal of Machine Learning Research*, 16:991–992, 2015.

[29] G. Reeves and M. C. Gastpar. Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inform. Theory*, 59(6):3451–3465, 2013.

[30] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, Feb. 1994.

[32] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[33] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *ArXiv e-prints*, 2014.

[34] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12), 2009.

[35] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[36] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[37] Y. Yuan et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, 32(7):644–652, 2014.

[38] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[39] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[40] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Departments of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104
USA
E-mail: suw@wharton.upenn.edu

Institute of Mathematics
University of Wroclaw
50-137, Wroclaw
Poland
E-mail: malgorzata.bogdan@pwr.edu.pl

Departments of Statistics and Mathematics
Stanford University
Stanford, CA 94305
USA
E-mail: candes@stanford.edu