Review of Statistics Paper: ”False Discoveries Occur Early on the Lasso Path”

Tuan-Binh Nguyen [*]

Abstract

In the paper ”False Discoveries Occur Early on the Lasso Path” by Su, Bogdan and Candès [1], the authors proved that under some specific assumptions, there is a sharp asymptotic trade-off between false and true positive rates (which is equivalently the measure of type I and type II errors) along the Lasso paths. The authors then gave a heuristic explanation to this phenomenon, which relates to the 'shrinkage noise' that the Lasso introduce when doing variable selection. Furthermore, they proved that ℓ_0-penalized regression does not suffer from the same problem.

1 Introduction

Introduced by Tibshirani in 1994 [2], the Least Absolute Shrinkage and Selection Operator, or Lasso, has since then become a widely-used tool for fitting regression models, especially in the setting of high-dimension statistics where the number of variables p is much greater than number of observations n. This setting can often be found in the field of genetics and neuroimaging, where one often encounters data samples of few subjects but with large quantity of variables (and that the true regression coefficients are sparse), which are DNA sequences in the former case or human brain-images in the latter case. In such cases, Lasso is expected regression to perform well by select true important signals and shrink the unimportant ones to zero. However, the recent study of Su, Bogdan and Candès show that as soon as the Lasso regression correctly selects all the true signal, it will also include a number of false discoveries which cannot be reduce to zero. The study provide a theoretical description of this phenomenon given particular assumptions.

The content of our review is as follows: section §2 gives an overview about the results and contributions of the study, section §3 is our attempt to reproduce the numerical experiments from the study, and the last section §4 is some of our remarks and conclusions for the review.

To make it easier for readers to follow, important points in the review will be marked with **[Important]** symbol.

[*] Paris-Saclay University & UEVE (tuan-binh.nguyen@ens.univ-evry.fr)

2 Results of the Study

2.1 Notations

Before going into results of the study, it is beneficial to mention its settings and assumptions. Formally, the definition of Lasso is as follows. Suppose we have a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and its corresponding dependent variables vector $\boldsymbol{y} \in \mathbb{R}^n$ such that:

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

with $\boldsymbol{\beta}$ is vector of true signal and $\boldsymbol{z}$ is random noise. The Lasso estimator of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}(\lambda)$, is then the solution of:

$$\underset{\boldsymbol{b} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_1$$

where $\|\cdot\|_1$ the ℓ_1 norm. Concisely, this is the definition for least square Lasso regression with standard linear model assumption. Some other notations are:

- $k = |\{j : \beta_j \neq 0\}|$ is the number of true signals and its estimator $\hat{k} = |\{j : \hat{\beta}_j \neq 0\}|$.
- $\epsilon = \dfrac{k}{p}$ is the sparsity ratio and $\delta = \dfrac{n}{p}$ is the dimensionality of design matrix.
- $V(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j = 0\}|$: the numbers of *false discoveries.*
- $T(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j \neq 0\}|$: the numbers of *true discoveries.*
- False Discoveries Proportion: $\text{FDP}(\lambda) = \dfrac{V(\lambda)}{\max\{\hat{k}, 1\}}$
- True Positive Proportion: $\text{TPP}(\lambda) = \dfrac{T(\lambda)}{\max\{k, 1\}}$

2.2 Key assumptions

[Important] The following assumptions were made by the paper to ensure its theoretical results:

- **Assumption 1** (linear sparsity): Expected number of non-zero regression coefficients is linear in p. In other words, we will always have $k = \epsilon p$ for some $\epsilon \in (0, 1)$.
- **Assumption 2** (Gaussian design): The design matrix $\boldsymbol{X}$ has i.i.d $\mathcal{N}(0, 1/n)$ entries, and the errors z_i are i.i.d $\mathcal{N}(0, \sigma^2)$.
- **Assumption 3** (distribution of regression coefficient): The coefficient sequence $\beta_1, \beta_2, \ldots, \beta_p$ are independent copies of a random variable Π with $\mathbb{E}(\Pi^2) < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$.

2.3 Main results

We will state the main results of the paper without going into details of technical proofs. Readers can refer to the Appendix section of 'False discoveries occur Early on the Lasso Path' for the full proofs of the theorem listed below.

[Important] Under the previous assumptions, the main points of the paper are:

1. **(Theorem 1)** There exists a tight boundary curve $q^*(u)$ such that $\mathrm{FDP}(\hat{\lambda}) > q^*(\mathrm{TPP}(\hat{\lambda}))$ with probability tending to one. In other words, we will never have both small False Discoveries Proportion and small True Positive Proportion at the same time no matter how carefully we choose the estimator $\hat{\lambda}$ of λ.
2. When the sparsity ratio $\epsilon = k/p$ and the dimensionality $1/\delta = p/n$ increases the problem becomes more severe.
3. A *heuristic explanation* was given: under the linear sparsity regime (Assumption 1), the Lasso introduce pseudo-noise which we can call *shrinkage noise*. The situation gets worse as many strong variables get picked up, which leads to wrong selection of null variables along the Lasso path.
4. Variants of the Lasso and other regression methods that utilize the ℓ_1 regularization also suffer the same shrinkage to noise problem.
5. However, not all of methods of model fitting using regularization suffer from the same trade-off as the Lasso regression.
6. **(Theorem 2)** In particular, we can find the an estimator λ for regression with ℓ_0 regularization such that asymptotically, both type I and type II error rate is reduced to 0.

Following is the elaboration on some of these points.

2.3.1 There exists a tight boundary curve $q^*(u)$ such that $\mathrm{FDP}(\hat{\lambda}) > q^*(\mathrm{TPP}(\hat{\lambda}))$ with probability tending to one

The boundary curve q^* was defined in the paper as a function of u given fixed δ, ϵ :

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

where $t^*(u)$ is the largest positive root of the following equation in t:

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

and $\Phi(\cdot)$ is the cumulative distribution function (cdf), $\phi(\cdot)$ the probability density function (pdf) of the standard normal distribution.

Notice that $q^*(\cdot; \delta, \epsilon) = q^*(\cdot)$ is positive, strictly increasing and infinitely many times differentiable. Moreover, $q^*(0) = 0$, which fits into the theory that when $\mathrm{TPP}(\lambda) = 0$ one must also have

$\mathrm{FDP}(\lambda) = 0$. The statement of Theorem 1 in the paper is then as follow:

Theorem 1: under the assumptions, for fixed $\delta \in (0, \infty), \epsilon \in (0, 1)$ and consider the above function $q^*(\cdot) > 0$. For any arbitrary small constants $\lambda_0 > 0$ and $\eta > 0$ we have:

(a) In both case of noiseless ($\sigma = 0$) or noisy data ($\sigma \neq 0$), we have:

$$\mathbb{P}\left(\bigcap_{\lambda > \lambda_0} \{\mathrm{FDP}(\lambda) \leq q^*(\mathrm{TPP}(\lambda)) - \eta\}\right) \underset{n\to\infty}{=} 1$$

(b) No matter how we choose $\hat{\lambda}(\boldsymbol{y}, \boldsymbol{X})$ adaptively, we will never have $\mathrm{FDP}(\hat{\lambda}) < q^*(\mathrm{TPP}(\hat{\lambda}) - C)$ with C constant.

(c) The boundary curve q^* is tight: any continuous $q(u) \geq q^*(u)$ will fail (a).

The Proof of Theorem 1 can be found in Appendix section of the paper. It was developed using the theory of approximate message passing (AMP) [3].

Moreover:

With a constant ϵ', if we set the prior distribution of regression coefficients to

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

then for any $u \in (0, 1)$, there is some fixed $\epsilon' = \epsilon'(u) > 0$ such that

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

2.3.2 A heuristic explanation for the FDR trade-off problem in Lasso regression

In the heuristic explanation part, the authors use simplified case of Lasso regression to prove that by the time half of the true variables are selected by this method, the False Discoveries Proportion cannot be reduced to zero.

For simplicity, in this part the authors assume:

- The true support $\mathcal{T}$ is a subset of size $\epsilon \cdot p$ and $\bar{\mathcal{T}}$ its compliment.
- Each non-zero regression coefficient in $\mathcal{T}$ take a constant value $M > 0$
- $\delta > \epsilon$
- The data is noiseless ($\sigma_{z_i} = 0$)

The Lasso regression thus can be reduced to the form

$$\hat{\boldsymbol{\beta}}_{\mathcal{T}}(\lambda) = \underset{\boldsymbol{b}_{\mathcal{T}} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\boldsymbol{\mathcal{T}}}\|^2 + \lambda\|\boldsymbol{b}_{\mathcal{T}}\|_1$$

The reduced solution $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mathcal{T}}}(\lambda)$ is independent from the other columns $\boldsymbol{X}_{\bar{\mathcal{T}}}$. At this point we take λ to be the same magnitude with M so that around half of the signals are selected by the Lasso.

The Karush-Kuhn-Tucker condition states that

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

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

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

Because we have chose λ so that half of the discoveries were made, the sub-gradient will takes on value 1 in magnitude about $\epsilon \cdot p/2$ times. So with high probability

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

with c_0, c_1 constant depends on ϵ and δ. Then for any $j \notin \mathcal{T}$ we have $\boldsymbol{X}_j^{\top}(\boldsymbol{y} - \boldsymbol{X}_{\mathcal{T}}\hat{\boldsymbol{\beta}}_{\mathcal{T}}) = \boldsymbol{X}_j^{\top}\boldsymbol{X}_{\mathcal{T}}(\boldsymbol{\beta}_{\mathcal{T}} - \hat{\boldsymbol{\beta}}_{\mathcal{T}})$ follows a normal distribution with zero mean and variance

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

In other words we will always have

$$\mathbb{P}\left(|\boldsymbol{X}_j^{\top}(\boldsymbol{\beta}_{\mathcal{T}} - \hat{\boldsymbol{\beta}}_{\mathcal{T}})| > \lambda\right) > 0$$

We conclude with a remark that if $|\boldsymbol{X}_j^{\top}(\boldsymbol{\beta}_{\mathcal{T}} - \hat{\boldsymbol{\beta}}_{\mathcal{T}})| > \lambda$ for any $j \notin \mathcal{T}$ then $\boldsymbol{X}_j$ must be selected by the incremental Lasso with design variables indexed by $\mathcal{T} \cup \{j\}$. This means when half of true discoveries were found, there must be at least one false discovery included in the variables selected by the Lasso.

2.3.3 Regression with ℓ_0 regularization will not suffer from the same trade-off as the Lasso regression when the signals are sufficiently strong

The regression with ℓ_0-penalized term is:

$$\underset{\boldsymbol{b} \in \mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_0$$

Under such regularization and consider when $\epsilon < \delta$ (or equivalently $k < n$) with the regression coefficients follow a two-point prior

$$\Pi = \begin{cases} M, & \text{with prob. } \epsilon \\ 0, & \text{with prob. } 1 - \epsilon \end{cases}$$

the authors prove that:

Theorem 2: We can find $\lambda_{\ell_0}(M)$ such that in probability

$$\lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{FDP}(\lambda_{\ell_0}) = 0 \quad \text{and} \quad \lim_{M\to\infty} \lim_{n,p\to\infty} \mathrm{TPP}(\lambda_{\ell_0}) = 1$$

In other words, theorem 2 states that asymptotically, the regularization parameter in ℓ_0-regression can be tuned so that we can effectively reduce both type I error rate and type II error rate to 0 without the trade-off observed in Lasso regression. The proof of theorem 2 can be found in Appendix E of the paper.

3 Reproduction of Numerical Experiments

Given that reproducibility of scientific studies is one of the main themes of our course, it is of importance for us to reproduce the numerical experiments shown in the paper. All the reproduction is implemented in Python using SciPy [4] and Scikit-learn [5] library. The trade-off curve q^* is rewritten in Python using Matlab code provided by the authors [1].

In the same folder with this review is a folder called '`numerical_code/`' that contains all the code we use to do the experiments and plot all the figures in this section, which are `experiment_1.py`, `experiment_2.py`, `experiment_3.py` and `trade_off_diagram.py`.

Overview: All 3 reimplemented experiments gave us almost identical results with the original ones introduced in the paper.

3.1 First experiments (Figure 1 and 2 in the paper) - `experiment_1.py`

The first experiment settings are:

- Gaussian design matrix $\boldsymbol{X} \in \mathbb{R}^{1010\times1000}$, each entries are drawn independently from $\mathcal{N}(0, 1)$.
- $\beta_1 = \cdots = \beta_{200} = 4$, $\beta_{201} = \cdots = \beta_{1000} = 0$.
- This means $k = 200$, $\epsilon = k/p = 0.2$, $\delta = n/p = 1.01$.

Figure 1 and 2 describes the reimplementation. We achieve a result that follows very closely the results from the original paper. In figure 1, when the true positive proportion (TPP) reaches 50% mark, the false discovery proportion has almost surpassed 7.5%. At the first time that the Lasso achieve 100% TPP, its FDP has reached 15%. We also observed that the TPP and FDP along the Lasso path follows very closely its asymptotic trade-off curve $q^*(\mathrm{TPP}(\lambda); \delta = 1.01, \epsilon = 0.2)$.

In figure 2, we redo the same experiment 100 times independently and plot the histogram of TPP at the first time a false discovery was detected, and FDP when TPP first reach 1.0. On the left, we can see that in all simulation, the first false discovery was wrongly selected before 40% of true signal was detected. The figure 2 (right) shows that after 100 independent experiments, the average

[1] `https://github.com/wjsu/fdrlasso`

FDP at time of last true detection is around 0.15, which follows closely the asymptotical results that states:

$$q^* \left(\mathrm{TPP}(\hat{\lambda}) = 1.0; \delta = 1.01, \epsilon = 0.2 \right) \approx 0.1537$$

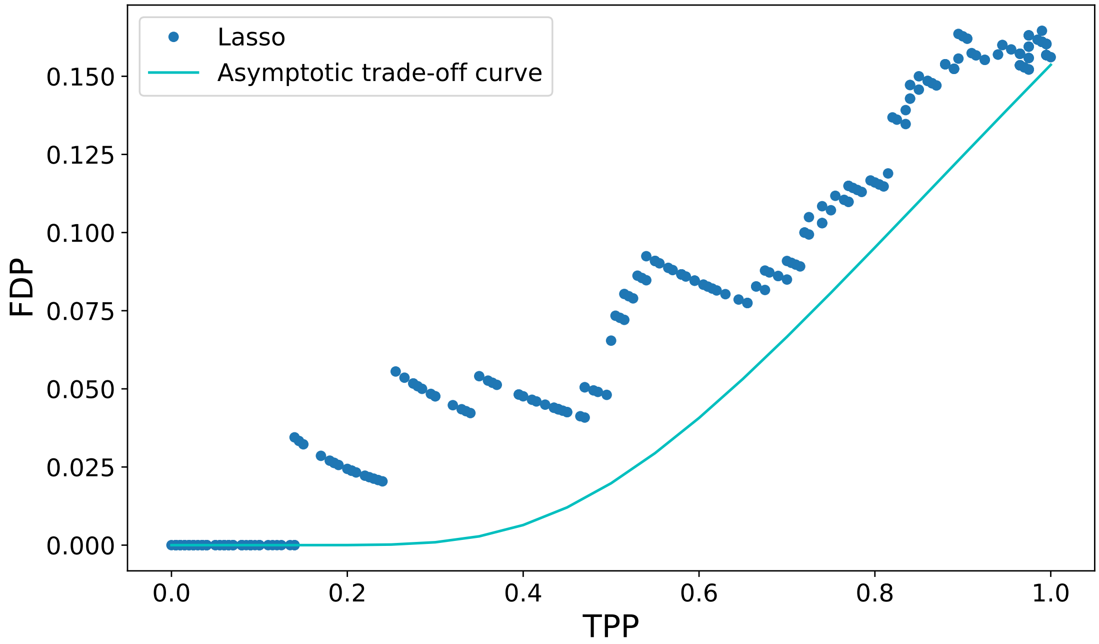

Figure 1: Reimplementation of Experiment 1: True positive and false positive rates along the Lasso path with their asymptotic boundary curve.

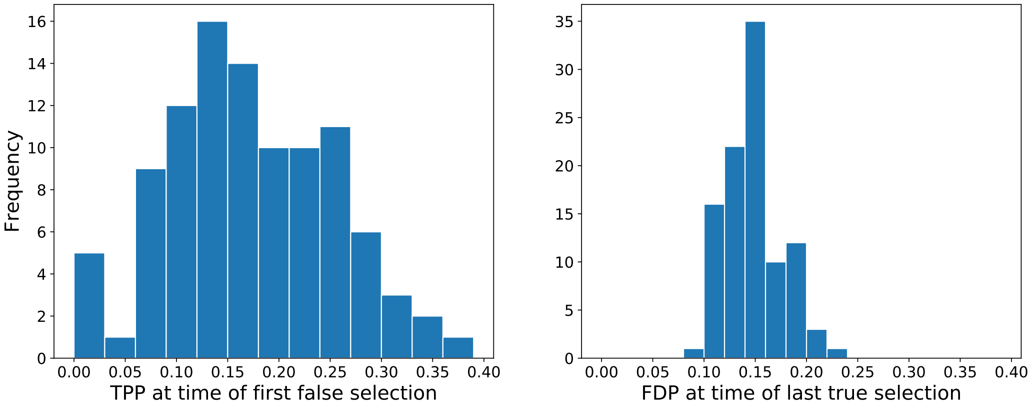

Figure 2: Histogram from reimplementation of experiment 1 in 100 times. Left: power when the first false variable enters the Lasso model. Right: FDP at the first time power reaches one.

3.2 Examples of boundary curve q^* - `trade_off_diagram.py`

For comparison purpose only, we were also able replicate exactly figure 3 and 4 in the paper thanks to the idea from the code that the authors provided.

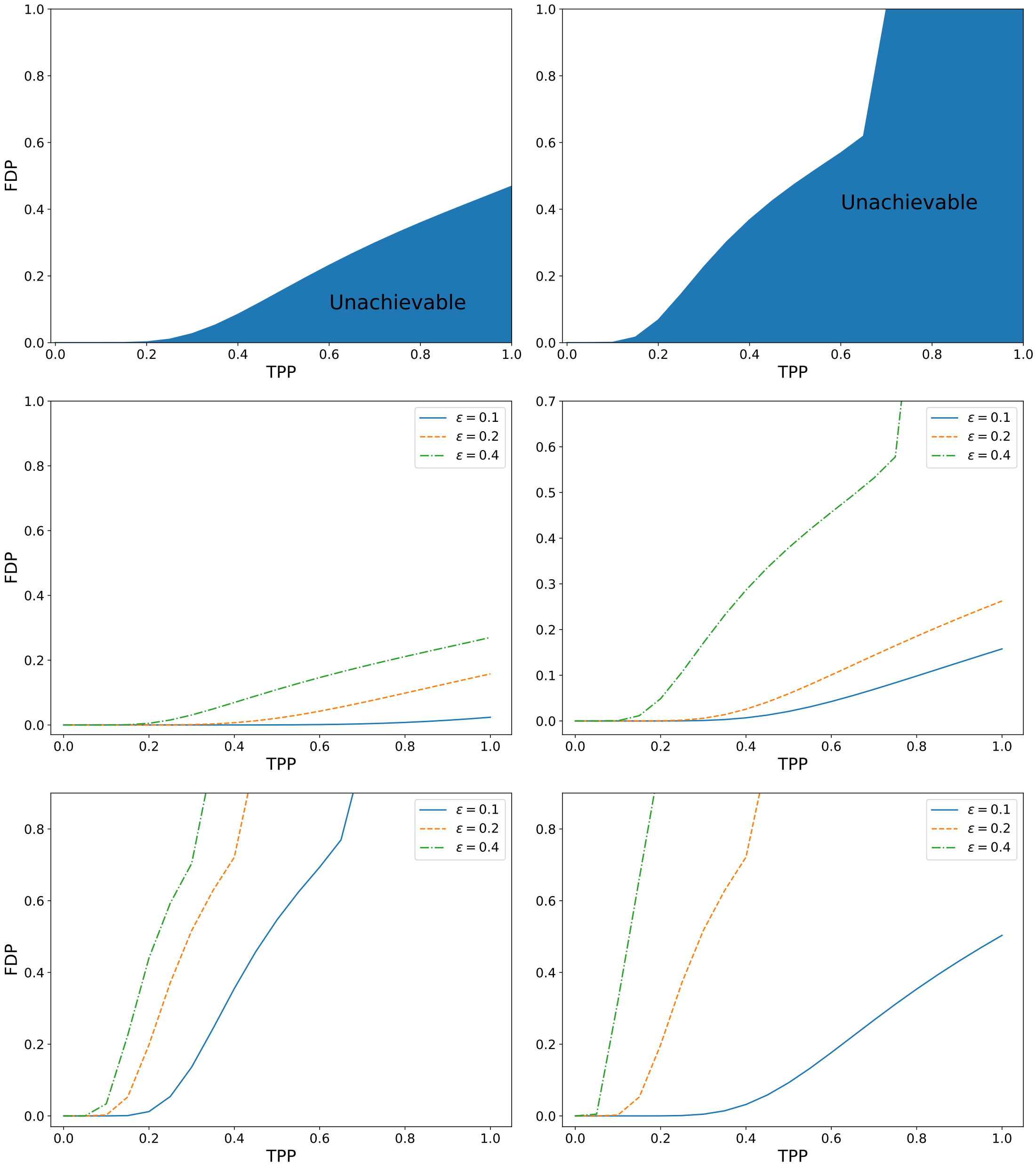

Figure 3: Replotting figure 3 and figure 4 in the paper: Unachievable region of (TPP, FDP pairs) and examples of the boundary curve q^* for different value of ϵ and δ.

3.3 Second experiment (figure 5 in the paper) - `experiment_2.py`

In the second experiment, the settings are:

- $n/p = \delta = 1.0$ and $\epsilon = 0.2$.
- Noiseless data: $\sigma = 0$.
- For figure 5a: 10 independent Lasso paths with $n = p = 1000$ and $n = p = 5000$, coefficients follow prior of $\mathbb{P}(\Pi = 50) = 1 - \mathbb{P}(\Pi = 0) = \epsilon$.
- For figure 5b: average of 100 replicates with $n = p = 1000$, the prior distribution follows eq.1 with $\epsilon' \in \{0.3, 0.5, 0.7, 0.9\}$, except instead of $\mathbb{P}(\Pi = M^{-1} = 1/50) = \epsilon \cdot \; (1 - \epsilon')$ we have $\mathbb{P}(\Pi = 0.1) = \epsilon \cdot (1 - \epsilon')$.

Remark: Without lack of generality and more importantly because of limit in computational resource, we only re-implemented the experiment with the case of $n = p = 1000$. In the end this experiment still, took around 2 hours to finish.

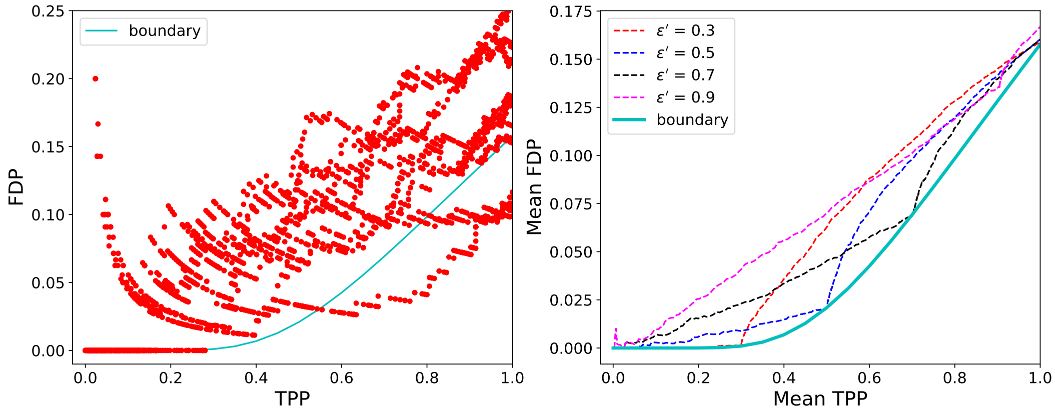

Figure 4: Reimplentation of figure 5 in the paper. Left: 10 independent Lasso and its boundary curve. Right: Mean FDP vs. mean TPP averaged at different values over 100 replicates for $n = p = 1000$, which shows the sharpness of the boundary with its boundary curve.

Figure 4 summarizes the result of the experiment. It is clear that for 10 independent Lasso paths (left figure), the large majority of the (TPP, FDP) pairs stay above the boundary. Taking average of the (TPP, FDP) pairs of 100 (right figure) shows that even with different value of ϵ' (*i.e.* different prior distribution of regression coefficients), when TPP reaches 1.0 different FDP values all stays close to its asymptotic value. Besides that, the average curves only touch the boundary curve and always run above it, which confirms the theoretical result.

3.4 Third experiment (Figure 6 in the paper) - `experiment_3.py`

The settings for the last experiment are:

- $n = 250, p = 1000$.
- $\beta_1 = \cdots = \beta_{18} = 2.5\sqrt{2\log p} \approx 9.3$, other coefficients equal 0.
- Which means $\delta = n/p = 0.25$, $\epsilon = 18/1000 = 0.018$.
- With noisy data: $\sigma_z = 1$, and noiseless data: $\sigma_z = 0$.

Remark: this experiment in the paper is somewhat closest to the reality, when we have the case with $n \ll p$, which is the situation when the Lasso would be utilize the most. The true coefficient vector generated is also strongly sparse. Again, due to lack of computational resource, we only do 200 trials to calculate mean FDP and TPP instead of 500 trials as in the original study.

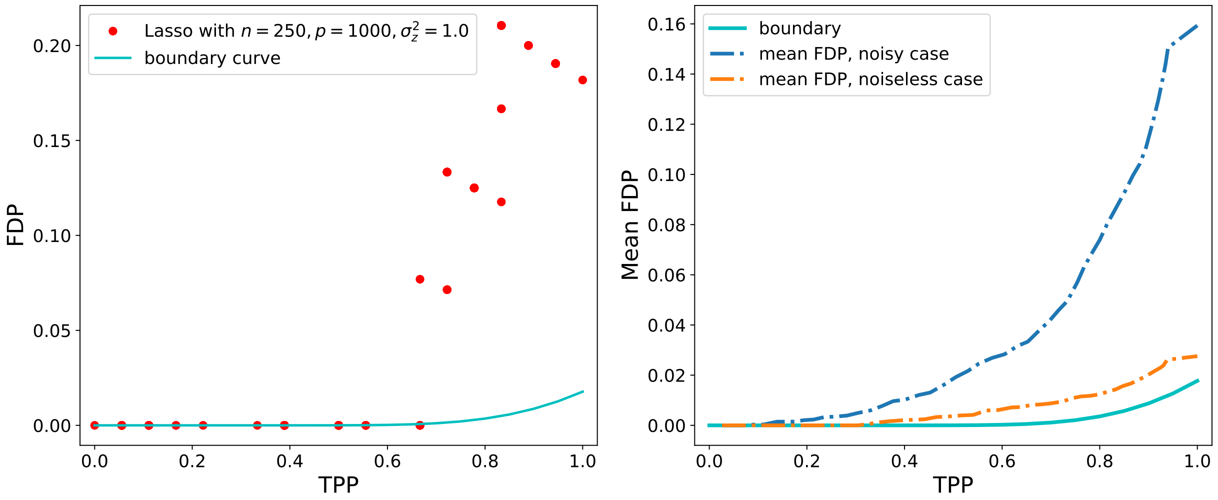

Figure 5: Reimplementation of figure 6 in the paper. High-dimension settings with strongly sparse signals: $n = 250, p = 1000, \epsilon = 0.018$. Left: single realization of the Lasso path in the noisy case ($\sigma_z^2 = 1$). Right: average FDP as a function of TPP with 200 iterations of the same experiment.

Figure 5 is the result of reimplementation of toy experiment 3. This result confirms the remark in the study: even in the case of high-dimension settings with very sparse signals, where one might expect the Lasso to perform very well, there are still false discovery included in the detected signals. With a single realization of Lasso path (left), when TPP reaches 100% the FDP is over 17%. Checking the average of FDP with 200 iterations gave us a clearer look: even in the noiseless case, the average FDP curve always lies above the boundary curve. The situation with noisy case is worse, where the mean FDP when TPP reaches 100% is around 16%.

4 Discussion on the study

In our opinion (the reviewer), the study of Su, Bogdan & Candès have clearly shown a rigorous formula for the trade-off curve between true positive rates and false discoveries rates by the Lasso regression. All of the numerical experiments the authors implemented confirmed this fact, and also can be reproduced. However, following are a few points we think that are the shortfall of the study.

Remark about limit of the study

1. Gaussian design matrix assumption (Assumption 1): in our opinion this assumption is very strong and will not be the case of most of all real datasets. Unfortunately, without this important assumption the result of Theorem 1 and 2 in the paper does not hold. At the point of the original study the authors only stated that it is likely that one can extend the Theorem for i.i.d sub-Gaussian entries, but no further elaboration was made.
2. We notice that the heuristic explanation only works in the case of linear sparsity. The study noted that under extreme sparsity and high Signal-to-Noise ratio, both type I and type II errors can be controlled at low levels. An example of research under this setting can be found in [6].
3. The paper also does not pursue generalization in the assumption of distribution of regression coefficients (Assumption 3). The authors stated that this assumption can be weakened to only empirical distribution with bounded second moment, but does not elaborated more on this point.
4. About the settings of *numerical experiment 1*: the design matrix $\boldsymbol{X}$ has entries follow $\mathcal{N}(0,1)$. However, in the next section of the paper, the authors clearly stated that the Gaussian design matrix setting are **all entries sampled from i.i.d** $\mathcal{N}(0,1/n)$. Indeed, the latter setting was used by the reviewer for reimplementation of experiment 2 and 3, since it would not return the same results if we used Gaussian design of $\mathcal{N}(0,1)$ for i.i.d entries.

Acknowledgments

We would like to express our gratitude to Prof. Christophe Giraud, who provides us the necessary background on specific topics of high-dimensional statistics, especially the exposure to the work of Su, Bogdan and Candès. This experience has inspired the reviewer to pursue further understanding about topic on selective inference in high-dimensional settings by working with the internship project related to the statistical control of sparse linear model for neuroimaging.

The reviewer also greatly benefit from paper review methodology presented from [7]. The Latex format in this report is taken from Prof. Stephen Boyd's webpage [8].

References

[1] W. Su, M. Bogdan, and E. Candès, "False discoveries occur early on the lasso path," *Ann. Statist.*, vol. 45, pp. 2133–2150, 10 2017.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[3] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[4] E. Jones, T. Oliphant, P. Peterson, *et al.*, "SciPy: Open source scientific tools for Python." `http://www.scipy.org/`, 2001–.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] P. Ji and Z. Zhao, "Rate optimal multiple testing procedure in high-dimensional regression," 2014.

[7] S. Keshav, "How to read a paper." `http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf`, 2016.

[8] S. Boyd, "Latex template for ee364b." `http://stanford.edu/class/ee364b/latex_templates/`, 2014.