False Discoveries occur Early on the Lasso Path
Paper Analysis

ABOU EL QASSIME Mehdi
mehdi.abou-el-qassime@polytechnique.edu

MAJDOUBI Ahmed Amine
ahmed.majdoubi@polytechnique.edu

February 15,2020

Abstract

The following report is an analysis of the paper ”False Discoveries occur Early on the Lasso Path” as part of the course 'Theoretical guidelines for high for high-dimensional data analysis'. The goal of this analysis is to present the context and the main results of the paper followed by some comments that discuss the key results of the paper as well as their limitations.

First, we will begin by briefly introducing the paper by describing the problem encountered by statisticians with some real-world datasets which leads to the use of the Lasso regression. Then will describe the limitations of the Lasso when dealing with these problems, as well as discuss the results mentioned in the paper. This will be followed by enplaning the concepts and hypotheses taken to get those results, while also discussing the limitations of these hypotheses.

Our work also includes a numerical implementation that aims to reproduce the author's results in the paper and can serve for the reader to do more experimentation and validation.

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

Contents

1 Introduction

Nowadays, datasets collected for modeling contain a huge amount of features that are noisy most of the time, it become then difficult to select among all of those features the ones that are really useful in a regression setting problem. One appealing feature of the Lasso over earlier techniques such as ridge regression is that it automatically performs variable reduction, since it produces models where a lot of —if not most— regression coefficients are estimated to be exactly zero. For a problem of size $n \times p$ with a design matrix **X** and a vector of labels **y** of dimension n, the standard linear model is defined as:

$$y = X\beta + z \tag{1}$$

where z is the noise term. The Lasso regression is then defined as:

$$\hat{\beta}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \frac{1}{2} \|y - Xb\|^2 + \lambda \|b\|_1 \tag{2}$$

The author states that the Lasso is mostly used in situations where the regression coefficient sequence is sparse, and in the case of sufficiently strong coefficients (compared to the noise level) of features that are not strongly correlated with each other, the Lasso is believed to select most of the useful variables and only few of those that represent only noise.

While this assertion is supported by theoretical asymptotic results, and is valid for extreme asymptotic regimes, it is then obvious that it would also perform well in settings of practical interest. In contrast, the paper comes with examples of studies that shows the incapability of the Lasso to selecting a proper model in practical use-cases, and that it also may include some false discoveries during the selection process.

Other previous works showed that the problem of false discoveries in selecting the predictors mainly occurs because of the correlation between the variables and the small effect sizes. The paper that we are analysing today rather proves that the problem still appears in the case of totally independent features and with no noise, meaning that early false discoveries of the Lasso will be present even in perfect conditions. Moreover the article aims to explain the main source of the problem and to present an approximate quantification of the phenomenon by making numerical experiments and theoretical conclusions.

To do so, the authors proposed to observe simultaneously the values of the false discovery proportion **FDP** (defined as the ratio between the number of false discoveries and the total number of discoveries) and the true positive proportion **TPP** (defined as the ratio between the number of true discoveries and that of potential true discoveries to be made, within several cases. The problem of identifying what true and false discoveries really are raises in the case of high dimensional correlated features, the study is then performed for the special case of independent variables, that is in the same time clear and easy to distinguish between false and true discoveries and to observe the phenomenon in the case of independent regressors.

2 Presentation of main experimental results

The article defines and clarifies the settings for the theoretical and experimental findings, the false discovery as a selected feature X_j , while being stochastically independent from the output vector, which means that its regression coefficient is actually equal to 0 i.e. $\beta_J = 0$.

The experiment has been run under a design matrix X of size 1010×1000 as a random Gaussian design where all column entries are independent to guarantee that the hypothesis cited above are true. All of the 200 first coefficients are then set to a constant value different from 0, and all other coefficients are then set to 0 for sparsity. $\beta_1 = ... = \beta_{200} = 4, \beta_{201} = ... = \beta_{1000} = 0$, with a high signal to noise ratio. Taking in consideration all the settings cited before, the article states that with an ordinary least squares estimate, all the first 200 discoveries will correspond to true discoveries, which can be explained by the fact that the features are non-correlated .

The experiment run by the authors of the article confirm that this is not the case for a Lasso regression, which would select null variables rather early. To give a quantitative estimation of this statement, the article present the numerical percentages observed when working with a Lasso. To be specific, 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. **Figure 1** illustrates better the phenomenon and the numerical findings, and shows a comparison between the Lasso and the Ordinary Least Squares regression (**OLS**).

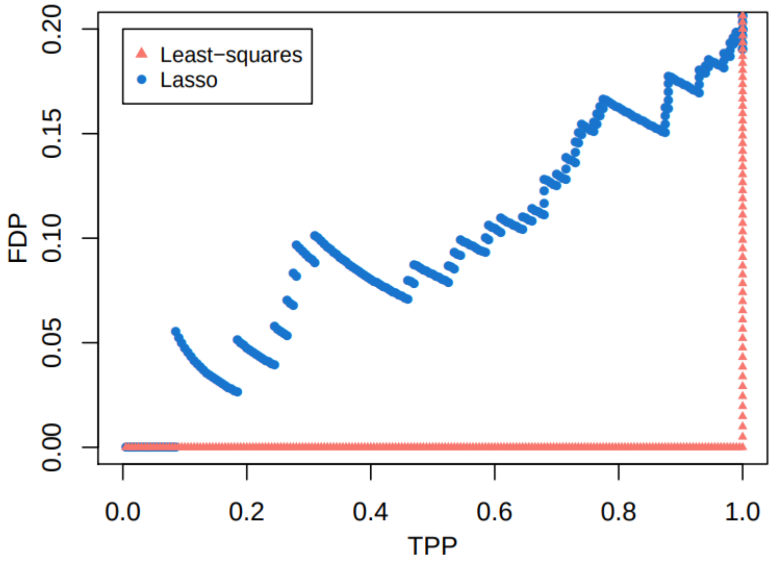

Figure 1: True positive and false positive rates for Lasso and OLS.

So as to have more significant insights about the performance of the Lasso on this type of problems, the writers decided to run more independent similar experiments (About

100), to figure out that 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%. This can be better observed by **figure 2**.

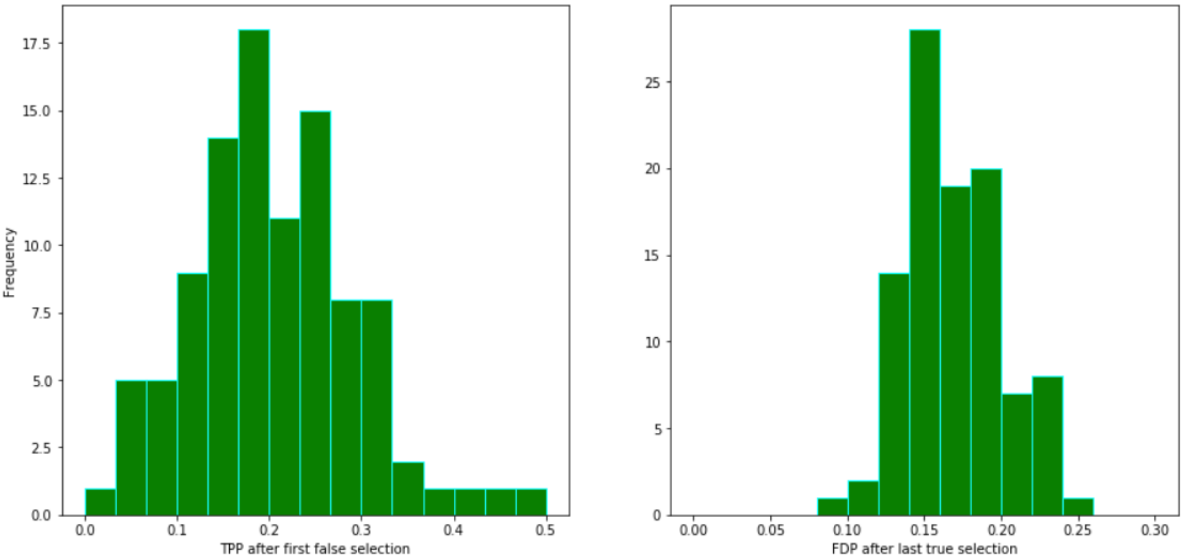

Figure 2: power when the first false variable enters the Lasso model and false discovery proportion when the power reaches 1.

This was a general overview of the problem encountered by the Lasso when performing variable selection within high-dimensional datasets, and a quantitative insight of theses results. The aim of this paper is discussing and providing an exact quantitative representation of the phenomenon, and derive an essential trade-off between power and error of type I defined to be the false positive rate, and to define theoretically an exact boundary separating the region of pairs (**FDP,TPP**) said to be impossible to achieve, and feasible set of pairs.

In the next section, we will see how the authors defined the unfeasible region and how to set the boundary between the two sets of pairs.

3 The Lasso Trade-Off Diagram

We will now give a quick reminder of assumptions and working conditions of the paper including the design of the model that is defined by the matrix of features $X \in \mathbb{R}^{n \times p}$ set as an i.i.d $\mathcal{N}(0, \frac{1}{n})$ column entries. The regression coefficients $\beta_1 ... \beta_p$ are set to be independent copies of the random variable Π such that $E[\Pi^2] < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon \in [0, 1]$. The interesting case for the authors is when $p, n \to \infty$ and $\frac{n}{p} \to \delta$.

3.1 linear sparsity

The authors of the paper set the number of non-zero coefficients to be linear in p and equal to $\epsilon.p$ for some $\epsilon \geq 0$, where it excludes the asymptotic discussion cited in previous works. This hypothesis of linear sparsity is made by the authors to mimic the actual setting of practical data set, so as they can derive conclusions from their experiments that are valid for practically large scale cases as well.

3.2 Gaussian design

The article sets the features with a Gaussian design and independent columns, as they affirm that is easier for the problem of variable selection to perform well with a Gaussian independent distribution. Again they set the most favorable conditions so as their conclusion could be valid for generalization in most of the other cases.

3.3 Regression coefficients

The assumptions for the regression coefficients are rather weakened, as the sequence of $\beta_1, ..., \beta_p$ only needs to have a convergent empirical distribution with a bounded second moment.

3.4 Main result of the paper

The paper claims, that under the given assumptions, and in a regime of linear sparsity, the Lasso algorithm not perform as good as expected for variable selection. The paper shows in particular that there is a region that contains pairs of (FDP,TPP) that cannot be reached, where we tend to increase the TPP and decrease the FPD, which means in the perfect case scenario, selecting only the true variables without including any of the false discoveries. Theoretically, they show that there exist an asymptotic trade-off between the FPD and the TPP. We will first give a formula definition for both the FPD and TPP :

$$FPD = \frac{\left|\left\{j : \hat{\beta}_j(\lambda) \neq 0, \beta_j = 0\right\}\right|}{max(\left|\left\{j : \hat{\beta}_j(\lambda) \neq 0\right\}\right|, 1)} \tag{3}$$

$$TPP = \frac{\left|\left\{j : \hat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0\right\}\right|}{max(\left|\left\{j : \hat{\beta}_j(\lambda) \neq 0\right\}\right|, 1)} \tag{4}$$

Then for a fixed $\delta \in [0, \infty[$ and $\epsilon \in [0, 1]$ and considering a function $q^*(.) > 0$ parameterized by δ and ϵ, under the hypothesis stated before, we have that :

$$\bigcap_{\lambda > \lambda_0} \{FPD(\lambda) \geq TPP(\lambda) - \eta\} \tag{5}$$

holds with a probability asymptotically tending to one for both the noisy observations case and noiseless observations case. The authors define the two types of errors, the type error I that is measured by the FPD, and the type error II which is measured by 1-TPP ie False negative proportion, to give anothor intuitive explanation to their theorem, which claims that nowhere in the Lasso path can both of the error types be low at the same time.

The figure below, illustrates better the trade-off diagram, where $q^*(.)$ is the boundary curve that separates the tow regions.

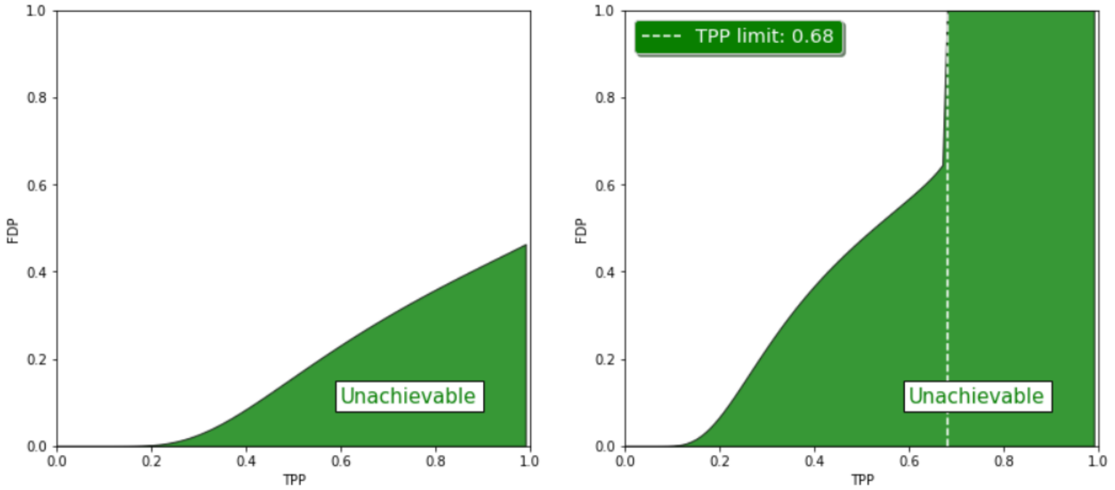

Figure 3: Lasso trade-off diagram.

It is now the time to give an exact definition of that boundary curve or function that separates the possible and impossible regions:

For a fixed λ, let $t^*(u)$ be the largest positive root of the equation in t,

$$\frac{2(1-\epsilon)[(1+t^2)\phi(-t)-t\phi(t)]+\epsilon(1+t^2)-\delta}{\epsilon[(1+t^2)(1-2\phi(-t))+t\phi(t)]}=\frac{1-u}{1-2\phi(-t)} \tag{6}$$

then $q^*(.)$ is defined by :

$$q^*(u;\delta,\epsilon)=\frac{2(1-\epsilon)\phi(-t^*(u)))}{2(1-\epsilon)\phi(-t^*(u))+\epsilon u} \tag{7}$$

Figure 4 displays examples of the boundaries 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 ϵ = k/p increases and the dimensionality $1/\delta$ = p/n increases.

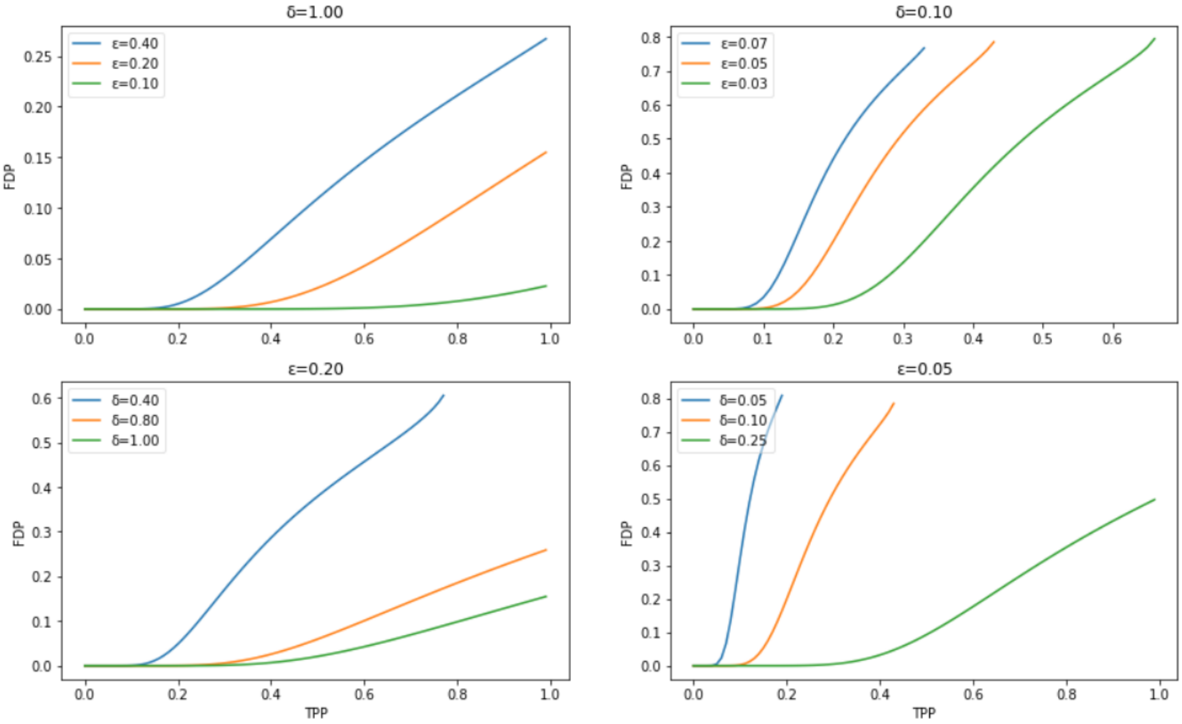

Figure 4: Different simulations for boundary function.

Figure 5 shows the results of different simulations with finite number of n and p under the assumption of noiseless observations. We plot all pairs (TPP, FDP) of 10 simulations for each of the cases n=p=1000 and n=p=5000 to establish a comparison for the high dimensional case.

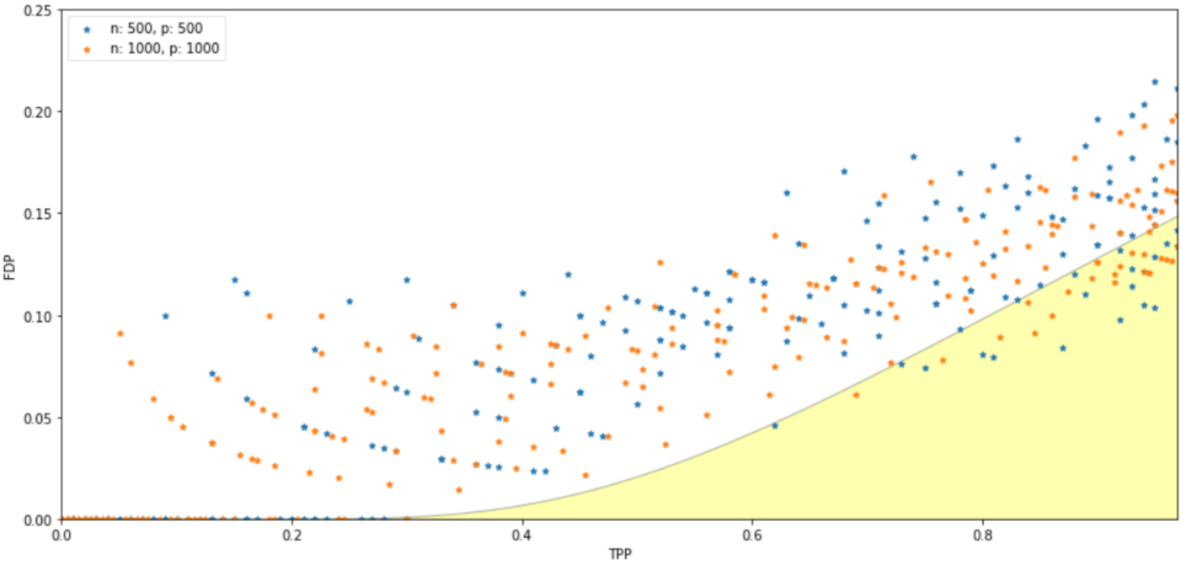

Figure 5: Different numerical Simulations .

One important observation is that when the average of FDP gets closer to the boundary as TPP gets closer to one, we see that a fraction of the paths fall below the boundary, which contradicts what it was claimed by the theorem. In fact the theorem states that

$\bigcap_{\lambda>\lambda_0} \{FPD(\lambda) \geq TPP(\lambda) - \eta\}$ holds with a probability converging to one, but for every small value of η, the authors didn't give an exact quantitative estimation of the value of η and whether it is depending on ϵ or δ. The main result asserts that all the region below the boundary is impossible to achieve, while in the graph some of the paths are within that region, and no explanation or quantification of those special cases is given.

4 Shrinkage noise

In this section, we will describe the explanation of the reasons behind of the false discoveries of the Lasso regression, as well as identifying some other methods that can substitute the Lasso on some cases, but with a different computational cost.

There are in fact many other methods that can help remedy the problem of sparsity and variables selection in the case of sufficiently strong signals. The example that is further presented by the article is the l_0 regularization even if it comes with an exponential computational cost but can achieve good model selection it is formulated as below :

$$\hat{\beta}_0(\lambda) = \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|^2 + \lambda\|b\|_0 \tag{8}$$

Based on this formulation, the authors of the article give the second theorem serving as a result for the l_0 penalization.

Under the same hypothesis, and for $\epsilon \leq \delta$, by considering the following prior distribution:

$$\Pi = \begin{cases} M, w.p.\epsilon \\ 0, w.p.1-\epsilon \end{cases} \tag{9}$$

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

$$\lim_{M\to+\infty}\lim_{n,p\to+\infty} TPP(\lambda) = 1 \lim_{M\to+\infty}\lim_{n,p\to+\infty} FDP(\lambda) = 0 \tag{10}$$

This affirms, contrary to what was showed before, that there is no asymptotic behaviour regarding the trade-off between the FPD and the TPP. But this comes with computational costs for non-convexity of the function to be optimized.

Going back to the main cause of the Lasso regression false discoveries problem, it is presented to be a pseudo-noise generated by the shrinkage. In other words, when the factor of regularization is high, the estimates of the Lasso tend to be small, so when strong variables are selected the noise is amplified and its projection along the null variables may hide the signal of the strong variables, it is at that time when the false discoveries appear.

For further analysis in the article, the authors consider a reduced version of the Lasso assuming that the real support τ is a deterministic subset of size ϵp, the values of non-null coefficient regressors is equal to M. The reduced Lasso problem is then defined by :

$$\hat{\beta}_\tau(\lambda) = \arg\min_{b\tau\in\mathbb{R}^{\epsilon p}} \frac{1}{2}\|y - X_\tau b_\tau\|^2 + \lambda\|b_\tau\|_1 \tag{11}$$

The reduced solution $\hat{\beta}_{\mathcal{T}}(\lambda)$ is independent from the other columns, and the authors suggest to select a value of λ to of the same magnitude of M so that half of the signal variables are selected. And must verify the KKT conditions given by :

$$\lambda 1 < X_{\tau}^{T}(y - X_{\tau}\hat{\beta}_{\tau}) \leq \lambda 1 \tag{12}$$

So as if we have $\left|X_{\tau}^{T}(y - X_{\tau}\hat{\beta}_{\tau})\right| \leq \lambda$ for every $j \in \tau$, completing the solution with null values would give the exact solution of the Lasso regression. We should also keep in mind that each X_j is selected by the incremental Lasso. This can be reformulated for better understanding as that if the Lasso selected only a few variables from τ, this means that the proportion of variables selected from $\bar{\tau}$ will be high and will result in a high number of false discoveries.

5 Conclusion and discussion

In this report, we have discussed the main results of the paper that shows that for a dataset in perfect conditions (non correlated features and linear sparsity of the regression sequence of coefficients), the Lasso doesn't perform very well the task that it is famous for, which is that it can't select all the strong variables without integrating a quantified rate of noise from the false variables, which is called false discoveries. The authors of the paper went further to give an approximation of this proportion of noise by introducing the two key terms, FDP and TPP , and displaying the trade-off between them. The main objective from this was to show that there exist a region for perfect performance of the Lasso which is impossible to attain, and is truncated by means of a defined boundary. The figure illustrating the performance of the Lasso paths for a simulation of n=p=1000 and n=p=5000 showed that there exist some of the paths that are within the impossible region, this makes us wonder if there is an existing gap between the theory and the practical cases. In fact the hypotheses are taken to be the perfect case for better generalization, but what if the fact that all the columns of the design matrix lie within the same distribution avoid the findings to be better generalized ? and how can we valid in practical cases the theorems proved within precise assumptions?

On the other hand, if our main goal is not to achieve better solutions on the optimization of coefficients, but rather select only pertinent variables, we could avoid the Lasso problem of false discoveries by using other available methods that could achieve better performance, keeping in mind that the computational will be higher compared to that of the Lasso.