
Complexity Analysis of the Lasso Regularization Path

Julien Mairal

Bin Yu

Department of Statistics, University of California, Berkeley.

JULIEN@STAT.BERKELEY.EDU

BINYU@STAT.BERKELEY.EDU

Abstract

The regularization path of the Lasso can be shown to be piecewise linear, making it possible to “follow” and explicitly compute the entire path. We analyze in this paper this popular strategy, and prove that its worst case complexity is exponential in the number of variables. We then oppose this pessimistic result to an (optimistic) approximate analysis: We show that an approximate path with at most $O(1/\sqrt{\varepsilon})$ linear segments can always be obtained, where every point on the path is guaranteed to be optimal up to a relative ε -duality gap. We complete our theoretical analysis with a practical algorithm to compute these approximate paths.

1. Introduction

Without a priori knowledge about data, it is often difficult to estimate a model or make predictions, either because the number of observations is too small, or the problem dimension too high. When a problem solution is known to be sparse, sparsity-inducing penalties have proven to be useful to improve both the quality of the prediction and its interpretability. In particular, the ℓ_1 -norm has been used for that purpose in the Lasso formulation (Tibshirani, 1996).

Controlling the regularization often requires to tune a parameter. In a few cases, the regularization path—that is, the set of solutions for all values of the regularization parameter, can be shown to be piecewise linear (Rosset & Zhu, 2007). This property is exploited in homotopy methods, which consist of following the piecewise linear path by computing the direction of the current linear segment and the points where the direction changes (also known as kinks). Piecewise linearity of regularization paths was discovered by Markowitz (1952) for portfolio selection; it was similarly exploited

by Osborne et al. (2000) and Efron et al. (2004) for the Lasso, and by Hastie et al. (2004) for the support vector machine (SVM). As observed by Gärtner et al. (2010), all of these examples are in fact particular instances of *parametric quadratic programming* formulations, for which path-following algorithms appear early in the optimization literature (Ritter, 1962).

In this paper, we study the number of linear segments of the Lasso regularization path. Even though experience with data suggests that this number is linear in the problem size (Rosset & Zhu, 2007), it is known that discrepancies can be observed between worst-case and empirical complexities. This is notably the case for the simplex algorithm (Dantzig, 1951), which performs empirically well for solving linear programs even though it suffers from exponential worst-case complexity (Klee & Minty, 1972). Similarly, by using geometrical tools originally developed to analyze the simplex algorithm, Gärtner et al. (2010) have shown that the complexity of the SVM regularization path can be exponential. However, to the best of our knowledge, none of these results do apply to the Lasso regularization path, whose theoretical complexity remains unknown. The goal of our paper is to fill in this gap.

Our first contribution is to show that in the worst-case the number of linear segments of the Lasso regularization path is exactly $(3^p + 1)/2$, where p is the number of variables (predictors). We remark that our proof is constructive and significantly different than the ones proposed by Klee & Minty (1972) for the simplex algorithm and by Gärtner et al. (2010) for SVMs. Our approach does not rely on geometry but on an adversarial scheme. Given a Lasso problem with p variables, we show how to build a new problem with $p + 1$ variables increasing the complexity of the path by a multiplicative factor. It results in explicit pathological examples that are surprisingly simple, unlike pathological examples for the simplex algorithm or SVMs.

Worst-case complexity analyses are by nature pessimistic. Our second contribution on approximate regularization paths is more optimistic. In fact, we show

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s).

that an approximate path for the Lasso with at most $O(1/\sqrt{\varepsilon})$ segments can always be obtained, where every point on the path is guaranteed to be optimal up to a relative ε -duality gap. We follow in part the methodology of Giesen et al. (2010) and Jaggi (2011), who have presented weaker results but in a more general setting for parameterized convex optimization problems. Our analysis builds upon approximate optimality conditions, which we maintain along the path, leading to a practical approximate homotopy algorithm.

The paper is organized as follows: Section 2 presents some brief overview of the Lasso. Section 3 is devoted to our worst-case complexity analysis, and Section 4 to our results on approximate regularization paths.

2. Background on the Lasso

In this section, we present the Lasso formulation of Tibshirani (1996) and well known facts, which we exploit later in our analysis. For self-containedness and clarity reasons we include simple proofs of these results. Let \mathbf{y} be a vector in \mathbb{R}^n and $\mathbf{X} = [\mathbf{x}^1, \dots, \mathbf{x}^p]$ be a matrix in $\mathbb{R}^{n \times p}$. The Lasso is formulated as:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1, \quad (1)$$

where the ℓ_1 -norm induces sparsity in the solution \mathbf{w} and $\lambda > 0$ controls the amount of regularization. Under a few assumptions, which are detailed in the sequel, the solution of this problem is unique. We denote it by $\mathbf{w}^*(\lambda)$ and define the *regularization path* \mathcal{P} as the set of all solutions for all positive values of λ :¹

$$\mathcal{P} \triangleq \{\mathbf{w}^*(\lambda) : \lambda > 0\}.$$

The following lemma presents classical optimality and uniqueness conditions for the Lasso solution (see Fuchs, 2005), which are useful to characterize \mathcal{P} :

Lemma 1 (Optimality Conditions of the Lasso). *A vector \mathbf{w}^* in \mathbb{R}^p is a solution of Eq. (1) if and only if for all j in $\{1, \dots, p\}$,*

$$\begin{aligned} \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^*) &= \lambda \text{sign}(\mathbf{w}_j^*) \quad \text{if } \mathbf{w}_j^* \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^*)| &\leq \lambda \quad \text{otherwise.} \end{aligned} \quad (2)$$

Define $J \triangleq \{j \in \{1, \dots, p\} : |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^*)| = \lambda\}$. Assuming the matrix $\mathbf{X}_J = [\mathbf{x}^j]_{j \in J}$ to be full rank, the solution is unique and we have

$$\mathbf{w}_J^* = (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} (\mathbf{X}_J^\top \mathbf{y} - \lambda \boldsymbol{\eta}_J), \quad (3)$$

where $\boldsymbol{\eta} \triangleq \text{sign}(\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}^*))$ is in $\{-1; 0; +1\}^p$, and the notation \mathbf{u}_J for a vector \mathbf{u} denotes the vector of size $|J|$ recording the entries of \mathbf{u} indexed by J .

¹For technicality reasons, we enforce $\lambda > 0$ even though the limit $\mathbf{w}^*(0^+) \triangleq \lim_{\lambda \rightarrow 0^+} \mathbf{w}^*(\lambda)$ may exist.

Proof. Eq. (2) can be obtained by considering subgradient optimality conditions. These can be written as $0 \in \{-\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}^*) + \lambda \mathbf{p} : \mathbf{p} \in \partial \|\mathbf{w}^*\|_1\}$, where $\partial \|\mathbf{w}^*\|_1$ denotes the subdifferential of the ℓ_1 -norm at \mathbf{w}^* . A classical result (Borwein & Lewis, 2006) says that the subgradients \mathbf{p} are the vectors in \mathbb{R}^p such that for all j in $\{1, \dots, p\}$, $\mathbf{p}_j = \text{sign}(\mathbf{w}_j^*)$ if $\mathbf{w}_j^* \neq 0$, and $|\mathbf{p}_j| \leq 1$ otherwise. This gives Eq. (2). The equalities in Eq. (2) define a linear system that has a unique solution given by (3) when \mathbf{X}_J is full rank.

Let us now show the uniqueness of the Lasso solution. Consider another solution $\mathbf{w}^{*\prime}$ and choose a scalar θ in $(0, 1)$. By convexity, $\mathbf{w}^{\theta*} \triangleq \theta \mathbf{w}^* + (1 - \theta) \mathbf{w}^{*\prime}$ is also a solution. For all $j \notin J$, we have $|\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^{\theta*})| \leq \theta |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^*)| + (1 - \theta) |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^{*\prime})| < \lambda$. Combining this inequality with the conditions (2), we necessarily have $\mathbf{w}_{J^c}^{\theta*} = \mathbf{w}_{J^c}^* = 0$,² and the vector $\mathbf{w}_J^{\theta*}$ is also a solution of the following reduced problem:

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^{|J|}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_J \tilde{\mathbf{w}}\|_2^2 + \lambda \|\tilde{\mathbf{w}}\|_1.$$

When \mathbf{X}_J is full rank, the Hessian $\mathbf{X}_J^\top \mathbf{X}_J$ is positive definite and this reduced problem is strictly convex. Thus, it admits a unique solution $\mathbf{w}_J^{\theta*} = \mathbf{w}_J^*$. It is then easy to conclude that $\mathbf{w}^* = \mathbf{w}^{\theta*} = \mathbf{w}^{*\prime}$. \square

With the assumption that the matrix \mathbf{X}_J is always full-rank, we can formally recall a well-known property of the Lasso (see Markowitz, 1952; Osborne et al., 2000; Efron et al., 2004) in the following lemma:

Lemma 2 (Piecewise Linearity of the Path).

Assume that for any $\lambda > 0$ and solution of Eq. (1) the matrix \mathbf{X}_J defined in Lemma 1 is full-rank. Then, the regularization path $\{\mathbf{w}^(\lambda) : \lambda > 0\}$ is well defined, unique and continuous piecewise linear.*

Proof. The existence/uniqueness of the regularization path was shown in Lemma 1.

Let us define $\{\boldsymbol{\eta}^*(\lambda) \triangleq \text{sign}(\mathbf{w}^*(\lambda)) : \lambda > 0\}$ the set of sparsity patterns. Let us now consider $\lambda_1 < \lambda_2$ such that $\boldsymbol{\eta}^*(\lambda_1) = \boldsymbol{\eta}^*(\lambda_2)$. For all $\theta \in [0, 1]$, it is easy to see that the solution $\mathbf{w}^{\theta*} \triangleq \theta \mathbf{w}^*(\lambda_1) + (1 - \theta) \mathbf{w}^*(\lambda_2)$ satisfies the optimality conditions of Lemma 1 for $\lambda = \theta \lambda_1 + (1 - \theta) \lambda_2$, and that $\mathbf{w}^*(\theta \lambda_1 + (1 - \theta) \lambda_2) = \mathbf{w}^{\theta*}$.

This shows that whenever two solutions $\mathbf{w}^*(\lambda_1)$ and $\mathbf{w}^*(\lambda_2)$ have the same signs for $\lambda_1 \neq \lambda_2$, the regularization path between λ_1 and λ_2 is a linear segment. As an important consequence, the number of linear segments of the path is smaller than 3^p , the number of possible sparsity patterns in $\{-1, 0, 1\}^p$. The path \mathcal{P} is therefore piecewise linear with a finite number of kinks.

² J^c denotes the complement of the set J in $\{1, \dots, p\}$.

Moreover, since the function $\lambda \rightarrow \mathbf{w}^*(\lambda)$ is piecewise linear, it is piecewise continuous and has right and left limits for every $\lambda > 0$. It is easy to show that these limits satisfy the optimality conditions of Eq. (2). By uniqueness of the Lasso solution, they are equal to $\mathbf{w}^*(\lambda)$ and the function is in fact continuous. \square

Assuming again that \mathbf{X}_J is always full rank, we can now present in Algorithm 1 the homotopy method (Osborne et al., 2000; Efron et al., 2004).

Algorithm 1 Homotopy Algorithm for the Lasso.

- 1: **Inputs:** a vector \mathbf{y} in \mathbb{R}^n ; a matrix \mathbf{X} in $\mathbb{R}^{n \times p}$;
 - 2: **initialization:** set λ to $\|\mathbf{X}^\top \mathbf{y}\|_\infty$; we have $\mathbf{w}^*(\lambda) = 0$ (trivial solution);
 - 3: set $J \triangleq \{j_0\}$ such that $|\mathbf{x}^{j_0 \top} \mathbf{y}| = \lambda$;
 - 4: **while** $\lambda > 0$ **do**
 - 5: Set $\boldsymbol{\eta} \triangleq \text{sign}(\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}^*(\lambda)))$;
 - 6: compute the direction of the path:

$$\begin{cases} \mathbf{w}_j^*(\lambda) &= (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} (\mathbf{X}_J^\top \mathbf{y} - \lambda \boldsymbol{\eta}_J) \\ \mathbf{w}_{j^c}^*(\lambda) &= 0. \end{cases}$$
 - 7: Find the smallest step $\tau > 0$ such that:
 - there exists $j \in J^c$ such that $|\mathbf{x}^{j \top} (\mathbf{y} - \mathbf{X}\mathbf{w}^*(\lambda - \tau))| = \lambda - \tau$; add j to J ;
 - there exists j in J such that $\mathbf{w}_j^*(\lambda) \neq 0$ and $\mathbf{w}_j^*(\lambda - \tau) = 0$; remove j from J ;
 - 8: replace λ by $\lambda - \tau$; record the pair $(\lambda, \mathbf{w}^*(\lambda))$;
 - 9: **end while**
 - 10: **Return:** sequence of recorded values $(\lambda, \mathbf{w}^*(\lambda))$.
-

It can be shown that this algorithm maintains the optimality conditions of Lemma 1 when λ decreases. Two assumptions have nevertheless to be made for the algorithm to be correct. First, $(\mathbf{X}_J^\top \mathbf{X}_J)$ has to be invertible, which is a reasonable assumption commonly made when working with real data and when one is interested in sparse solutions. When $(\mathbf{X}_J^\top \mathbf{X}_J)$ becomes ill-conditioned, which may typically occur for small values of λ , the algorithm has to stop and the path is truncated. Second, one assumes in Step 7 of the algorithm that the value τ corresponds to a single event $|\mathbf{x}^{j \top} (\mathbf{y} - \mathbf{X}\mathbf{w}^*(\lambda - \tau))| = \lambda - \tau$ for j in J^c or $\mathbf{w}_j^*(\lambda - \tau)$ hits zero for j in J . In other words, variables enter or exit the path one at a time. Even though this assumption is reasonable most of the time, it can be problematic from a numerical point of view in rare cases. When the length of a linear segment of \mathcal{P} is smaller than the numerical precision, the algorithm can fail. In contrast, our approximate homotopy algorithm presented in Section 4 is robust to this issue. In the next section, we present our worst-case complexity analysis of the regularization path, showing that Algorithm 1 can have exponential complexity.

3. Worst-Case Complexity

We denote by $\{\boldsymbol{\eta}^*(\lambda) \triangleq \text{sign}(\mathbf{w}^*(\lambda)) : \lambda > 0\}$ the set of sparsity patterns in $\{-1, 0, 1\}^p$ encountered along the path \mathcal{P} . We have seen in the proof of Lemma 2 that whenever $\boldsymbol{\eta}^*(\lambda_1) = \boldsymbol{\eta}^*(\lambda_2)$, for $\lambda_1, \lambda_2 > 0$, then $\boldsymbol{\eta}^*(\lambda) = \boldsymbol{\eta}^*(\lambda_1)$ for all $\lambda \in [\lambda_1, \lambda_2]$, and thus the number of linear segments of \mathcal{P} is upper-bounded by 3^p . With an additional argument, we can further reduce this number, as stated in the following proposition:

Proposition 1 (Upper-bound Complexity).

Let assume the same conditions as in Lemma 2. The number of linear segments in the regularization path of the Lasso is less than $(3^p + 1)/2$.

Proof. We have already noticed that the number of linear segments of the path is at most 3^p . Let us consider $\boldsymbol{\eta}^*(\lambda_1) \neq 0$ for $\lambda_1 > 0$. We now show that for all $\lambda_2 > 0$, we have $\boldsymbol{\eta}^*(\lambda_2) \neq -\boldsymbol{\eta}^*(\lambda_1)$, and therefore the number of different sparsity patterns on the path \mathcal{P} is in fact less than or equal to $(3^p + 1)/2$.

Let us assume that there exists $\lambda_2 > 0$ with $\boldsymbol{\eta}^*(\lambda_2) = -\boldsymbol{\eta}^*(\lambda_1)$, and look for a contradiction. We define the set $J' \triangleq \{j \in \{1, \dots, p\} : \boldsymbol{\eta}_j^*(\lambda_1) \neq 0\}$, and consider the solution of the reduced problem for all $\lambda \geq 0$:

$$\tilde{\mathbf{w}}^*(\lambda) \triangleq \arg \min_{\tilde{\mathbf{w}} \in \mathbb{R}^{|J'|}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{J'} \tilde{\mathbf{w}}\|_2^2 + \lambda \|\tilde{\mathbf{w}}\|_1,$$

which is well defined since the optimization problem is strictly convex (the conditions of Lemma 2 imply that $\mathbf{X}_{J'}$ is full rank). We remark that $\tilde{\mathbf{w}}^*(\lambda_1) = \mathbf{w}_{J'}^*(\lambda_1)$, and $\tilde{\mathbf{w}}^*(\lambda_2) = \mathbf{w}_{J'}^*(\lambda_2)$. Given the optimality conditions of Lemma 1, it is then easy to show that $\tilde{\mathbf{w}}^*(0) = (\mathbf{X}_{J'}^\top \mathbf{X}_{J'})^{-1} \mathbf{X}_{J'}^\top \mathbf{y} = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tilde{\mathbf{w}}^*(\lambda_1) + \frac{\lambda_1}{\lambda_1 + \lambda_2} \tilde{\mathbf{w}}^*(\lambda_2)$. Since the signs of $\tilde{\mathbf{w}}^*(\lambda_1)$ and $\tilde{\mathbf{w}}^*(\lambda_2)$ are opposite to each other and non-zero, we have $\|\tilde{\mathbf{w}}^*(0)\|_1 < \|\tilde{\mathbf{w}}^*(\lambda_1)\|_1$. Independently, it is also easy to show that the function $\lambda \rightarrow \|\tilde{\mathbf{w}}^*(\lambda)\|_1$ should be non-increasing, and we obtain a contradiction. \square

In the next proposition, we present our adversarial strategy to build a pathological regularization path. Given a Lasso problem with p variables and a path \mathcal{P} , we design an additional variable along with an extra dimension, such that the number of kinks of the new path $\tilde{\mathcal{P}}$ increases by a multiplicative factor compared to \mathcal{P} . We call our strategy ‘‘adversarial’’ since it consists of iteratively designing ‘‘pathological’’ variables.

Proposition 2 (Adversarial Strategy).

Let us consider \mathbf{y} in \mathbb{R}^n and \mathbf{X} in $\mathbb{R}^{n \times p}$ such that the conditions of Lemma 2 are satisfied and \mathbf{y} is in the span of \mathbf{X} . We denote by \mathcal{P} the regularization path of the Lasso problem corresponding to (\mathbf{y}, \mathbf{X}) , by k the

number of linear segments of \mathcal{P} , and by $\lambda_1 > 0$ the smallest value of the parameter λ corresponding to a kink of \mathcal{P} . We define the vector $\tilde{\mathbf{y}}$ in \mathbb{R}^{n+1} and the matrix $\tilde{\mathbf{X}}$ in $\mathbb{R}^{(n+1) \times (p+1)}$ as follows:

$$\tilde{\mathbf{y}} \triangleq \begin{bmatrix} \mathbf{y} \\ y_{n+1} \end{bmatrix}, \quad \tilde{\mathbf{X}} \triangleq \begin{bmatrix} \mathbf{X} & 2\alpha\mathbf{y} \\ 0 & \alpha y_{n+1} \end{bmatrix},$$

where $y_{n+1} \neq 0$ and $0 < \alpha < \lambda_1 / (2\mathbf{y}^\top \mathbf{y} + y_{n+1}^2)$.

Then, the regularization path $\tilde{\mathcal{P}}$ of the Lasso problem associated to $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ exists and has $3k-1$ linear segments. Moreover, let us consider $\{\boldsymbol{\eta}^1=0, \boldsymbol{\eta}^2, \dots, \boldsymbol{\eta}^k\}$ the sequence of sparsity patterns in $\{-1, 0, 1\}^p$ of \mathcal{P} (the signs of the solutions $\mathbf{w}^*(\lambda)$), ordered from large to small values of λ . The sequence of sparsity patterns in $\{-1, 0, 1\}^{p+1}$ of the new path $\tilde{\mathcal{P}}$ is the following:

$$\left\{ \underbrace{\left[\begin{array}{c} \boldsymbol{\eta}^1 \\ 0 \end{array} \right], \left[\begin{array}{c} \boldsymbol{\eta}^2 \\ 0 \end{array} \right], \dots, \left[\begin{array}{c} \boldsymbol{\eta}^k \\ 0 \end{array} \right]}_{\text{first } k \text{ patterns}}, \underbrace{\left[\begin{array}{c} \boldsymbol{\eta}^k \\ 1 \end{array} \right], \left[\begin{array}{c} \boldsymbol{\eta}^{k-1} \\ 1 \end{array} \right], \dots, \left[\begin{array}{c} \boldsymbol{\eta}^1=0 \\ 1 \end{array} \right]}_{\text{middle } k \text{ patterns}}, \underbrace{\left[\begin{array}{c} -\boldsymbol{\eta}^2 \\ 1 \end{array} \right], \left[\begin{array}{c} -\boldsymbol{\eta}^3 \\ 1 \end{array} \right], \dots, \left[\begin{array}{c} -\boldsymbol{\eta}^k \\ 1 \end{array} \right]}_{\text{last } k-1 \text{ patterns}} \right\}. \quad (4)$$

Let us first make some remarks about this proposition:

- According to Eq. (4) the sparsity patterns of the new path $\tilde{\mathcal{P}}$ are related to those of \mathcal{P} . More precisely, they have either the form $[\boldsymbol{\eta}^i, 0]^\top$ or $[\pm\boldsymbol{\eta}^i, 1]^\top$, where $\boldsymbol{\eta}^i$ is a sparsity pattern in $\{-1, 0, 1\}^p$ of \mathcal{P} .

- The last column of $\tilde{\mathbf{X}}$ involves a factor α that controls its norm. With α small enough, the $(p+1)$ -th variable enters late the path $\tilde{\mathcal{P}}$. As shown in Eq. (4), the first k sparsity patterns of $\tilde{\mathcal{P}}$ do not involve this variable and are exactly the same as those of \mathcal{P} .

- Let us give some intuition about the pathological behavior of the path $\tilde{\mathcal{P}}$. The first k kinks of $\tilde{\mathcal{P}}$ are the same as those of \mathcal{P} , and after these first k kinks we have $\mathbf{y} \approx \mathbf{X}\mathbf{w}^*(\lambda)$. Then, the $(p+1)$ -th variable enters the path and we heuristically have

$$\tilde{\mathbf{X}} \begin{bmatrix} \mathbf{w}^*(\lambda) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ y_{n+1} \end{bmatrix} \approx \tilde{\mathbf{y}} \approx \tilde{\mathbf{X}} \begin{bmatrix} -\mathbf{w}^*(\lambda) \\ 1/\alpha \end{bmatrix}. \quad (5)$$

The left side of Eq. (5) tells us that when the $(p+1)$ -th variable is inactive, the coefficients associated to the first p variables should be close to $\mathbf{w}^*(\lambda)$. At the same time, the right side of Eq. (5) tells us that when the $(p+1)$ -th variable is active, these same p coefficients should be instead close to $-\mathbf{w}^*(\lambda)$. According to Eq. (4), the signs of these p coefficients along the path switch from $\boldsymbol{\eta}^k = \text{sign}(\mathbf{w}^*(\lambda))$ to $-\boldsymbol{\eta}^k$ by following the sequence $\boldsymbol{\eta}^k, \boldsymbol{\eta}^{k-1}, \dots, (\boldsymbol{\eta}^1=0=-\boldsymbol{\eta}^1), -\boldsymbol{\eta}^2, \dots, -\boldsymbol{\eta}^k$, resulting in a path with $3k-1$ linear segments. The proof below more rigorously describes this strategy:

Proof. Existence of the new regularization path:

Let us rewrite the Lasso problem for $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$.

$$\begin{aligned} & \min_{\tilde{\mathbf{w}} \in \mathbb{R}^p, \tilde{w} \in \mathbb{R}} \frac{1}{2} \left\| \tilde{\mathbf{y}} - \tilde{\mathbf{X}} \begin{bmatrix} \tilde{\mathbf{w}} \\ \tilde{w} \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \tilde{\mathbf{w}} \\ \tilde{w} \end{bmatrix} \right\|_1, \\ & = \min_{\tilde{\mathbf{w}} \in \mathbb{R}^p, \tilde{w} \in \mathbb{R}} \frac{1}{2} \|(1-2\alpha\tilde{w})\mathbf{y} - \mathbf{X}\tilde{\mathbf{w}}\|_2^2 + \frac{1}{2} (y_{n+1} - \alpha y_{n+1} \tilde{w})^2 \\ & \quad + \lambda \|\tilde{\mathbf{w}}\|_1 + \lambda |\tilde{w}|. \quad (6) \end{aligned}$$

Let $(\tilde{\mathbf{w}}^*, \tilde{w}^*)$ be a solution for a given $\lambda > 0$. By fixing $\tilde{w} = \tilde{w}^*$ in Eq. (6) and optimizing with respect to $\tilde{\mathbf{w}}$, we obtain an equivalent problem to (6):

$$\min_{\tilde{\mathbf{w}}' \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\tilde{\mathbf{w}}'\|_2^2 + \frac{\lambda}{|1-2\alpha\tilde{w}^*|} \|\tilde{\mathbf{w}}'\|_1,$$

with the change of variable $\tilde{\mathbf{w}} = (1-2\alpha\tilde{w}^*)\tilde{\mathbf{w}}'$ and assuming $1-2\alpha\tilde{w}^* \neq 0$. The solution of this problem is unique since it is a point of \mathcal{P} and we therefore have

$$\tilde{\mathbf{w}}^* = \begin{cases} (1-2\alpha\tilde{w}^*)\mathbf{w}^* \left(\frac{\lambda}{|1-2\alpha\tilde{w}^*|} \right) & \text{if } \tilde{w}^* \neq \frac{1}{2\alpha} \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

Since the last column of $\tilde{\mathbf{X}}$ is not in the span of the first p columns by construction of $\tilde{\mathbf{X}}$, it is then easy to see that the conditions of Lemma 2 are necessarily satisfied and therefore $(\tilde{\mathbf{w}}^*, \tilde{w}^*)$ is in fact the unique solution of Eq. (6). Since this is true for all $\lambda > 0$, the regularization path is well defined, and we denote from now on the above solutions by $\tilde{\mathbf{w}}^*(\lambda)$ and $\tilde{w}^*(\lambda)$.

Maximum number of linear segments:

We now show that the number of linear segments of the path is upper-bounded by $3k-1$. Eq. (7) shows that $\text{sign}(\tilde{\mathbf{w}}^*(\lambda))$ has the form $\pm\boldsymbol{\eta}^i$, where $\boldsymbol{\eta}^i$ in $\{-1, 0, 1\}^p$ is one of the k sparsity patterns from \mathcal{P} , whereas we have three possibilities for $\text{sign}(\tilde{w}^*(\lambda))$, namely $\{-1, 0, +1\}$. Since one can not have two non-zero sparsity patterns that are opposite to each other on the same path, as shown in the proof of Proposition 1, the number of possible sparsity patterns reduces to $3k-1$.

Characterization of the first k linear segments:

Let us consider $\lambda \geq \lambda_1$ and show that $\tilde{\mathbf{w}}^*(\lambda) = \mathbf{w}^*(\lambda)$ and $\tilde{w}^*(\lambda) = 0$ by checking the optimality conditions of Lemma 1. The first p equalities/inequalities in Eq. (2) are easy to verify, the last one being also satisfied:

$$|2\alpha\mathbf{y}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}^*(\lambda)) + \alpha y_{n+1}^2| \leq 2\alpha \|\mathbf{y}\|_2^2 + \alpha y_{n+1}^2 < \lambda_1,$$

where the last inequality is obtained from the definition of α . Since this inequality is strict, this also ensures that there exists $0 < \lambda'_1 < \lambda_1$ such that $\tilde{\mathbf{w}}^*(\lambda) = \mathbf{w}^*(\lambda)$ and $\tilde{w}^*(\lambda) = 0$ for all $\lambda \geq \lambda'_1$. We have therefore shown that the first k sparsity patterns of the regularization path are given in Eq. (4).

Characterization of the last $2k-1$ segments:

We mainly use here the form of Eq. (7) and a few continuity arguments to characterize the rest of the path. First, we remark that for all β in $[0, \frac{1}{\alpha}]$, there exists a value for $\lambda > 0$ such that $\tilde{w}^*(\lambda) = \beta$. This is true because: (i) $\lambda \rightarrow \tilde{w}^*(\lambda)$ is continuous; (ii) $\tilde{w}^*(\lambda_1) = 0$; (iii) $\tilde{w}^*(0^+) = \frac{1}{\alpha}$. Point (i) was shown in Lemma 2, point (ii) in the previous paragraph, and point (iii) is necessary to have the term $(y_{n+1} - \alpha y_{n+1} \tilde{w})^2$ in Eq. (6) go to 0 when λ goes to 0^+ .

We now consider two values $\lambda'_1, \lambda'_2 > 0$ such that $\tilde{w}^*(\lambda'_1) = 0$, $\tilde{w}^*(\lambda'_2) = \frac{1}{2\alpha}$ and $\tilde{w}^*(\lambda) \in (0, \frac{1}{2\alpha})$ for all $\lambda \in (\lambda'_1, \lambda'_2)$. On this open interval, we have that $(1 - 2\alpha\tilde{w}^*(\lambda)) > 0$, and the continuous function $\lambda \rightarrow \lambda / |1 - 2\alpha\tilde{w}^*(\lambda)|$ ranges from λ'_1 to $+\infty$. Combining this observation with Eq. (7), we obtain that all sparsity patterns of the form $[\eta^{i\top}, 1]^\top$ for i in $\{1, \dots, k\}$ appear on the regularization path. With similar continuity arguments, it is easy to show that all sparsity patterns of the form $[-\eta^{i\top}, 1]^\top$ for i in $\{1, \dots, k\}$ appear on the path as well.

We had previously identified k of the sparsity patterns, and now have identified $2k-1$ different ones. Since we have at most $3k-1$ linear segments, the set of sparsity patterns on the path \tilde{P} is entirely characterized. The fact that the sequence of sparsity patterns is the one given in Eq. (4) can easily be shown by reusing similar continuity arguments. \square

With this proposition in hand, we can now state the main result of this section:

Theorem 1 (Worst-case Complexity).

In the worst case, the regularization path of the Lasso has exactly $(3^p + 1)/2$ linear segments.

Proof. We start with $n = p = 1$, and define $\mathbf{y} = [1]$, and $\mathbf{X} = [1]$, leading to a path with $k = 2$ segments. We then recursively apply Proposition 2, keeping $n = p$, choosing at iteration $p + 1$, $y_{p+1} = 1$, and a factor $\alpha = \alpha_{p+1}$ satisfying the conditions of Proposition 2. Denoting by k_p the number of linear segments at iteration p , we have that $k_{p+1} = 3k_p - 1$, and it is easy to show that $k_p = (3^p + 1)/2$. According to Proposition 1, this is the longest possible regularization path. Note that this example has a particularly simple shape:

$$\mathbf{y} \triangleq \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{X} \triangleq \begin{bmatrix} \alpha_1 & 2\alpha_2 & 2\alpha_3 & \dots & 2\alpha_p \\ 0 & \alpha_2 & 2\alpha_3 & \dots & 2\alpha_p \\ 0 & 0 & \alpha_3 & \dots & 2\alpha_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \alpha_p \end{bmatrix}.$$

\square

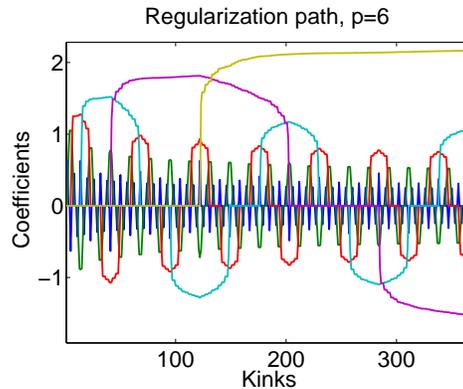


Figure 1. Pathological regularization path with $p=6$ variables and $(3^6 + 1)/2 = 365$ kinks. The curves represent the values of the coefficients at every kink of the path. For visibility purposes, we use a non-linear scale and report the values $\text{sign}(w)|w|^{0.1}$ for a coefficient w . Best seen in color.

3.1. Numerical Simulations

We have implemented Algorithm 1 in Matlab, optimizing numerical precision regardless of computational efficiency, which has allowed us to check our theoretical results for small values of p . For instance, we obtain a path with $(3^p + 1)/2 = 88574$ linear segments for $p = 11$, and present such a pathological path in Figure 1. Note that when p gets larger, these examples quickly lead to precision issues where some kinks are very close to each other. Our implementation and our pathological examples will be made publicly available. In the next section, we present more optimistic results on approximate regularization paths.

4. Approximate Homotopy

We now present another complexity analysis when exact solutions of Eq. (1) are not required. We follow in part the methodology of Giesen et al. (2010), later refined by Jaggi (2011), on approximate regularization paths of parameterized convex functions. Their results are quite general but, as we show later, we obtain stronger results with an analysis tailored to the Lasso.

A natural tool to guarantee the quality of approximate solutions is the duality gap. Writing the Lagrangian of problem (1) and minimizing with respect to the primal variable \mathbf{w} yields the following dual formulation of (1):

$$\max_{\boldsymbol{\kappa} \in \mathbb{R}^n} -\frac{1}{2} \boldsymbol{\kappa}^\top \boldsymbol{\kappa} - \boldsymbol{\kappa}^\top \mathbf{y} \quad \text{s.t.} \quad \|\mathbf{X}^\top \boldsymbol{\kappa}\|_\infty \leq \lambda, \quad (8)$$

where $\boldsymbol{\kappa}$ in \mathbb{R}^n is a dual variable. Let us denote by $f_\lambda(\mathbf{w})$ the objective function of the primal problem (1) and by $g_\lambda(\boldsymbol{\kappa})$ the objective function of the dual (8). Given a pair of feasible primal and dual variables $(\mathbf{w}, \boldsymbol{\kappa})$, the difference $\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) \triangleq f_\lambda(\mathbf{w}) - g_\lambda(\boldsymbol{\kappa})$ is called a duality gap and provides an optimality guar-

antee (see Borwein & Lewis, 2006):

$$0 \leq f_\lambda(\mathbf{w}) - f_\lambda(\mathbf{w}^*(\lambda)) \leq \delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}).$$

In plain words, it upper bounds the difference between the current value of the objective function $f_\lambda(\mathbf{w})$ and the optimal value of the objective function $f_\lambda(\mathbf{w}^*(\lambda))$. In this paper, we use a relative duality gap criterion to guarantee the quality of an approximate solution:³

Definition 1 (ε -approximate Solution).

Let ε be in $[0, 1]$. A vector \mathbf{w} in \mathbb{R}^p is said to be an ε -approximate solution of problem (1) if there exists $\boldsymbol{\kappa}$ in \mathbb{R}^n such that $\|\mathbf{X}^\top \boldsymbol{\kappa}\|_\infty \leq \lambda$ and $\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) \leq \varepsilon f_\lambda(\mathbf{w})$.

Given a set $\tilde{\mathcal{P}} \triangleq \{\tilde{\mathbf{w}}(\lambda) \in \mathbb{R}^p : \lambda > 0\}$, we say that $\tilde{\mathcal{P}}$ is an ε -approximate regularization path if any point $\tilde{\mathbf{w}}(\lambda)$ of $\tilde{\mathcal{P}}$ is an ε -approximate solution for problem (1).

Our goal is now to build ε -approximate regularization paths and study their complexity. To that effect, we introduce approximate optimality conditions based on small perturbations of those given in Lemma 1:

Definition 2 ($OPT_\lambda(\varepsilon_1, \varepsilon_2)$ Condition).

Let $\varepsilon_1 \geq 0$ and $\varepsilon_2 \geq -\varepsilon_1$. A vector \mathbf{w} in \mathbb{R}^p satisfies the $OPT_\lambda(\varepsilon_1, \varepsilon_2)$ condition if and only if for all $1 \leq j \leq p$,

$$\begin{aligned} \lambda(1 - \varepsilon_2) \leq \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}) \operatorname{sign}(\mathbf{w}_j) \leq \lambda(1 + \varepsilon_1) \text{ if } \mathbf{w}_j \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w})| \leq \lambda(1 + \varepsilon_1) \text{ otherwise.} \end{aligned} \quad (9)$$

Note that when $\varepsilon_1 = \varepsilon_2 = 0$, this condition reduces to the exact optimality conditions of Lemma 1. Of interest for us is the relation between Definitions 1 and 2. Let us consider a vector \mathbf{w} such that $OPT_\lambda(\varepsilon_1, \varepsilon_2)$ is satisfied. Then, the vector $\boldsymbol{\kappa} \triangleq \frac{1}{1 + \varepsilon_1}(\mathbf{X}\mathbf{w} - \mathbf{y})$ is feasible for the dual (8) and we can compute a duality gap:

$$\begin{aligned} \delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) &= f_\lambda(\mathbf{w}) - g_\lambda(\boldsymbol{\kappa}) \\ &= \frac{1}{2}(1 + \varepsilon_1)^2 \boldsymbol{\kappa}^\top \boldsymbol{\kappa} + \lambda \|\mathbf{w}\|_1 + \frac{1}{2} \boldsymbol{\kappa}^\top \boldsymbol{\kappa} + \boldsymbol{\kappa}^\top \mathbf{y} \\ &= \frac{\varepsilon_1^2}{2} \boldsymbol{\kappa}^\top \boldsymbol{\kappa} + \lambda \|\mathbf{w}\|_1 + \boldsymbol{\kappa}^\top (\mathbf{y} + (1 + \varepsilon_1) \boldsymbol{\kappa}) \\ &= \frac{\varepsilon_1^2}{(1 + \varepsilon_1)^2} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 + \boldsymbol{\kappa}^\top \mathbf{X}\mathbf{w}. \end{aligned}$$

From Eq. (9), it is easy to show that $\lambda \|\mathbf{w}\|_1 + \boldsymbol{\kappa}^\top \mathbf{X}\mathbf{w} \leq \frac{\varepsilon_1 + \varepsilon_2}{1 + \varepsilon_1} \lambda \|\mathbf{w}\|_1$, and we can obtain the following bound:

$$\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) \leq \max \left(\frac{\varepsilon_1^2}{(1 + \varepsilon_1)^2}, \frac{\varepsilon_1 + \varepsilon_2}{1 + \varepsilon_1} \right) f_\lambda(\mathbf{w}). \quad (10)$$

From this upper bound, we derive our first result:

³Note that our criterion is not exactly the same as in Jaggi (2011). Whereas Jaggi (2011) consider a formulation where the ℓ_1 -norm appears in a constraint, Eq. (1) involves an ℓ_1 -penalty. Even though these formulations have the same regularization path, they involve slightly different objective functions, dual formulations, and duality gaps.

Proposition 3 (Approximate Analysis).

Let \mathbf{y} be in \mathbb{R}^n and \mathbf{X} in $\mathbb{R}^{n \times p}$ such that the conditions of Lemma 2 are satisfied. Let $\lambda_\infty \triangleq \|\mathbf{X}^\top \mathbf{y}\|_\infty$ be the value of λ corresponding to the start of the path, and $\lambda_1 > 0$ be the one corresponding to the last kink. For all $\varepsilon \in (0, 1)$, there exists an ε -approximate regularization path with at most $\left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\sqrt{\varepsilon}} \right\rceil$ linear segments.

Proof. From Eq. (9), one can show by a simple calculation that an exact solution $\mathbf{w}^*(\lambda)$ for a given λ satisfies $OPT_{\lambda(1-\varepsilon_3)}(\varepsilon_3/(1-\varepsilon_3), -\varepsilon_3/(1-\varepsilon_3))$. According to Eq. (10), there exists a dual variable $\boldsymbol{\kappa}$ such that $\delta_{\lambda(1-\varepsilon_3)}(\mathbf{w}^*(\lambda), \boldsymbol{\kappa}) \leq \varepsilon_3^2$. Thus, for any λ' chosen in $[\lambda, \lambda(1-\sqrt{\varepsilon})]$, the solution $\mathbf{w}^*(\lambda)$ is an ε -approximate solution for the parameter λ' . Between λ_∞ and λ_1 , we can obtain an ε -approximate piecewise linear (in fact piecewise constant) regularization path by sampling solutions $\mathbf{w}^*(\lambda)$ for λ in $\{\lambda_\infty, \lambda_\infty(1-\sqrt{\varepsilon}), \dots, \lambda_\infty(1-\sqrt{\varepsilon})^k, \lambda_1\}$ with $\lambda_\infty(1-\sqrt{\varepsilon})^{k+1} \leq \lambda_1$. The number of segments of the corresponding approximate path is at most $\left\lceil \frac{-\log(\lambda_\infty/\lambda_1)}{\log(1-\sqrt{\varepsilon})} \right\rceil + 1 \leq \left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\sqrt{\varepsilon}} \right\rceil$. \square

Note that the term λ_∞/λ_1 is possibly large, but it is controlled by a logarithmic function and can be considered as constant for finite precision machines. In other words, the complexity of the approximate path is upper-bounded by $O(1/\sqrt{\varepsilon})$. In contrast, the analysis of Giesen et al. (2010) and Jaggi (2011) give us:

- an approximate path with $O(1/\varepsilon)$ linear segments can be obtained with a weaker approximation guarantee than ours. Namely, a bound $\delta \leq \varepsilon$ along the path, where δ is a duality gap, whereas we use *relative* duality gaps of the form $\delta \leq \varepsilon f_\lambda(\mathbf{w})$;⁴ Interestingly, this bound is proven to be optimal in the context of parameterized convex functions on the ℓ_1 -ball. Our result show that such bound can be improved for the Lasso.
- a methodology to obtain relative duality gaps along the path, which can easily provide complexity bounds for the full path of different problems, notably support vector machines, but not for the Lasso.

Proposition 3 is optimistic, but not practical since it requires sampling *exact* solutions of the path \mathcal{P} . We introduce an approximate homotopy method in Algorithm 2 which does not require computing exact solutions and still enjoys a similar complexity. It exploits the piecewise linearity of the path, but uses a first-order method (Beck & Teboulle, 2009; Fu, 1998) when the linear segments of the path are too short.

⁴When there exists $m, M > 0$ such that $m < f_\lambda < M$, the relative duality gap guarantee is similar (up to a constant) to the simple bound $\delta \leq \varepsilon$. However, we have for the Lasso that $f_\lambda(\mathbf{w}^*(\lambda)) \rightarrow 0$ when λ goes to 0^+ , as long as \mathbf{y} is in the span of \mathbf{X} . Note that as noticed in footnote 3, Jaggi (2011) uses a slightly different duality gap than ours.

Algorithm 2 Approximate Homotopy for the Lasso.

- 1: **Inputs:** a vector \mathbf{y} in \mathbb{R}^n , a matrix \mathbf{X} in $\mathbb{R}^{n \times p}$, the required precision $\varepsilon \in [0, 1]$; $\lambda_1 > 0$;
 - 2: **initialization:** set λ to $\|\mathbf{X}^\top \mathbf{y}\|_\infty$; set $\tilde{\mathbf{w}}(\lambda) = 0$;
 - 3: set $\theta = 1 + \varepsilon/2 - \sqrt{\varepsilon}/2$;
 - 4: set $J \triangleq \{j_0\}$ such that $|\mathbf{x}^{j_0 \top} \mathbf{y}| = \lambda$;
 - 5: **while** $\lambda \geq \lambda_1$ **do**
 - 6: **if** $(\mathbf{X}_J^\top \mathbf{X}_J)$ is not invertible **then** go to 12;
 - 7: set $\tilde{\boldsymbol{\eta}} \triangleq (1/\lambda)\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\tilde{\mathbf{w}}(\lambda))$;
 - 8: compute the approximate direction of the path:

$$\begin{cases} \tilde{\mathbf{w}}_J(\lambda) &= (\mathbf{X}_J^\top \mathbf{X}_J)^{-1}(\mathbf{X}_J^\top \mathbf{y} - \lambda \tilde{\boldsymbol{\eta}}_J) \\ \tilde{\mathbf{w}}_{J^c}(\lambda) &= 0. \end{cases}$$
 - Find the smallest step $\tau > 0$ such that:
 - there exists j in J^c such that $|\mathbf{x}^{j \top}(\mathbf{y} - \mathbf{X}\tilde{\mathbf{w}}(\lambda - \tau))| = (\lambda - \tau)(1 + \frac{\varepsilon}{2})$; add j to J ;
 - there exists j in J such that $\tilde{\mathbf{w}}_j(\lambda) \neq 0$ and $\tilde{\mathbf{w}}_j(\lambda - \tau) = 0$; remove j from J ;
 - 9: **if** $\tau \geq \lambda\theta\sqrt{\varepsilon}$ **then**
 - 10: replace λ by $\lambda - \tau$;
 - 11: **else**
 - 12: replace λ by $\lambda(1 - \theta\sqrt{\varepsilon})$;
 - 13: use a first-order optimization method to find a solution $\tilde{\mathbf{w}}(\lambda)$ satisfying $OPT_\lambda(\varepsilon/2, \varepsilon/2)$;
 - 14: set $J = \{j \in \{1, \dots, p\} : \tilde{\mathbf{w}}_j(\lambda) \neq 0\}$.
 - 15: **end if**
 - 16: record the pair $(\lambda, \tilde{\mathbf{w}}(\lambda))$;
 - 17: **end while**
 - 18: **Return:** sequence of recorded values $(\lambda, \tilde{\mathbf{w}}(\lambda))$.
-

Note that when $\varepsilon = 0$, Algorithm 2 reduces to Algorithm 1. Our approach exploits the following ideas, which we formally prove in the sequel. Assume that $\tilde{\mathbf{w}}(\lambda)$ satisfies $OPT_\lambda(\varepsilon/2, \varepsilon/2)$. Then,

- $\tilde{\mathbf{w}}(\lambda)$ is an ε -approximation for all λ' in $[\lambda, \lambda(1 - \theta\sqrt{\varepsilon})]$. This guarantees us that one can always make step sizes for λ greater than or equal to $\lambda\theta\sqrt{\varepsilon}$;

- the direction followed in Step 8 maintains $OPT_\lambda(\varepsilon/2, \varepsilon/2)$, but when two kinks are too close to each other—that is, $\tau < \lambda\theta\sqrt{\varepsilon}$, we directly look for a solution for the parameter $\lambda' = \lambda(1 - \theta\sqrt{\varepsilon})$ that satisfies $OPT_{\lambda'}(\varepsilon/2, \varepsilon/2)$. Any first-order method can be used for that purpose, e.g., a proximal gradient method (Beck & Teboulle, 2009), using the current value $\tilde{\mathbf{w}}(\lambda)$ as a warm start.

Note also that when $(\mathbf{X}_J^\top \mathbf{X}_J)$ is not invertible, the method uses first-order steps. The next proposition precisely describes the guarantees of our algorithm.

Proposition 4 (Analysis of Algorithm 2).

Let \mathbf{y} be in \mathbb{R}^n and \mathbf{X} in $\mathbb{R}^{n \times p}$. For all $\lambda_1 > 0$ and $\varepsilon \in (0, 1)$, Algorithm 2 returns an ε -approximate regularization path on $[\lambda_\infty, \lambda_1]$. Moreover, it terminates in at most $\left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\theta\sqrt{\varepsilon}} \right\rceil$ iterations, where $\lambda_\infty \triangleq \|\mathbf{X}^\top \mathbf{y}\|_\infty$.

Proof. We first show that any solution on the path is an ε -approximate solution. First, it is easy to check that $OPT_\lambda(\varepsilon/2, \varepsilon/2)$ is always satisfied at Step 6. This is either a consequence of Step 13, or because the direction $\tilde{\mathbf{w}}_J(\lambda') = (\mathbf{X}_J^\top \mathbf{X}_J)^{-1}(\mathbf{X}_J^\top \mathbf{y} - \lambda' \tilde{\boldsymbol{\eta}}_J)$ maintains $OPT_{\lambda'}(\varepsilon/2, \varepsilon/2)$ when λ' varies between λ and $\lambda - \tau$. From Eq. (10), we obtain that $\tilde{\mathbf{w}}(\lambda)$ is an ε -approximate solution whenever $OPT_\lambda(\varepsilon/2, \varepsilon/2)$ is satisfied. Thus, we only need to check that $\tilde{\mathbf{w}}(\lambda)$ is also an ε -approximate solution for λ' in $[\lambda, \lambda(1 - \theta\sqrt{\varepsilon})]$: for $\varepsilon_3 \geq 0$, it is easy to check that $OPT_\lambda(\varepsilon/2, \varepsilon/2)$ implies $OPT_{\lambda(1-\varepsilon_3)}((\varepsilon/2+\varepsilon_3)/(1-\varepsilon_3), (\varepsilon/2-\varepsilon_3)/(1-\varepsilon_3))$. Setting $\varepsilon_3 = \theta\sqrt{\varepsilon}$ and using Eq. (10), it is possible to show that the desired condition is satisfied.

Since the step size for λ is always greater than $\lambda\theta\sqrt{\varepsilon}$, the maximum number of iterations is upper-bounded by $\left\lceil \frac{-\log(\lambda_\infty/\lambda_1)}{\log(1-\theta\sqrt{\varepsilon})} \right\rceil + 1 \leq \left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\theta\sqrt{\varepsilon}} \right\rceil$ \square

We remark that the scalar θ is very close to 1 and therefore the complexity is similar to the one of Proposition 3, with a logarithmic function controlling the possibly large term λ_∞/λ_1 . This algorithm is practical in different aspects: (i) it is almost as simple to implement as the homotopy method; (ii) it is robust to cases where two kinks are too close for the classical homotopy method to work; (iii) it provides optimality guarantees along the path; (iv) whenever possible, it explicitly exploits the piecewise linearity of the path. We next present experiments to verify our analysis.

4.1. Numerical Simulations

We have implemented Algorithm 2 with a few modifications to the code used in Section 3.1. The inner solver is a coordinate descent algorithm (see Fu, 1998), with a stopping criterion based on Definition 2.

We consider 4 datasets. The first one dubbed SYNTH consists of a pure noise fitting scenario with no statistical meaning. The entries of the corresponding vector \mathbf{y} and matrix \mathbf{X} are i.i.d. draws from a standard normal distribution. The next dataset is called PATHOL and is a pathological example obtained from the analysis of Section 3. Finally, we consider two datasets based on real data, respectively dubbed MADELON⁵ and PC-MAC⁶. For each dataset, we center and normalize the columns of \mathbf{X} and the vector \mathbf{y} , and choose the parameter λ_1 corresponding to the last kink of the true path.

For all datasets, we compute the full regularization path using Algorithm 1 and several ε -approximate regularization paths using Algorithm 2. Note that

⁵<http://www.nipsfsc.ecs.soton.ac.uk/datasets/>.

⁶<http://featureselection.asu.edu/datasets.php>.

Table 1. Complexity results of ε -approximated regularization paths for four datasets with n observations and p variables. The number of linear segments is denoted by k .

	SYNTH	PATHOL	MADELON	PCMAC
n	1 100	11	2 000	1 943
p	1 000	11	500	3 289
k , full path	1 615	88 574	517	2 561
k , $\varepsilon = 10^{-5}$	1 297	2 744	468	1 254
k , $\varepsilon = 10^{-4}$	686	1 071	327	444
k , $\varepsilon = 10^{-3}$	268	405	152	155
k , $\varepsilon = 10^{-2}$	96	146	61	53
k , $\varepsilon = 0.1$	34	51	22	18
k , $\varepsilon = 0.25$	21	32	15	11
k , $\varepsilon = 0.5$	14	20	10	7

the path of PCMAC was stopped around $\lambda \approx 10^{-4}$ where the matrix $\mathbf{X}_J^T \mathbf{X}_J$ became ill-conditioned and the Lasso solution dense. As a simple sanity check, we first experimentally verify the correctness of Propositions 3 and 4, by sampling solutions on the approximate path we obtain, computing duality gaps, and checking that the solutions are indeed ε -approximate. We conclude that our experimental results match our theoretical analysis. We present the different path complexities in Table 1.

Interestingly, the complexity of the pathological example significantly reduces when one is looking for an approximate solution. For example, for $\varepsilon = 10^{-3}$, the complexity of the approximate path is less than 0.5% the one of the full path. This significantly contrasts with the pessimistic result obtained in Section 3. As expected, the two examples based on real data exhibit a path complexity of the same order of the problem size, which also significantly reduces when ε increases.

5. Conclusion

We have presented new results on the regularization path and thus on homotopy methods for the Lasso. First, we have shown that the path has an exponential worst-case complexity, which, as far as we know, had never been formally proved before. Our second result is more optimistic, and shows that when an exact path is not required, only a relatively small number of points on the path need to be computed. Finally, we propose a practical approximate homotopy algorithm, which can provide such approximate paths at a desired precision.

Acknowledgments

This paper was supported in part by NSF grants SES-0835531, CCF-0939370, DMS-1107000, DMS-0907632, and by ARO-W911NF-11-1-0114.

References

- Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
- Borwein, J. M. and Lewis, A. S. *Convex analysis and nonlinear optimization: theory and examples*. Springer, 2006.
- Dantzig, G. B. Maximization of a linear function of variables subject to linear inequalities. In Koopmans, T. C. (ed.), *Activity Analysis of Production and Allocation*, pp. 339–347. Wiley, New York, 1951.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004.
- Fu, W. J. Penalized regressions: The bridge versus the Lasso. *J. Comput. Graph. Stat.*, 7(3):397–416, 1998.
- Fuchs, J. J. Recovery of exact sparse representations in the presence of bounded noise. *IEEE T. Inform. Theory.*, 51(10):3601–3608, 2005.
- Gärtner, B., Jaggi, M., and Maria, C. An exponential lower bound on the complexity of regularization paths. *preprint arXiv:0903.4817v2*, 2010.
- Giesen, J., Jaggi, M., and Laue, S. Approximating parameterized convex optimization problems. In *Algorithms - ESA, Lectures Notes Comp. Sci.* 2010.
- Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. The entire regularization path for the support vector machine. *J. Mach. Learn. Res.*, 5:1391–1415, 2004.
- Jaggi, M. *Sparse Convex Optimization Methods for Machine Learning*. PhD thesis, ETH Zürich, 2011.
- Klee, V. and Minty, G. J. How good is the simplex algorithm? In Shisha, O. (ed.), *Inequalities*, volume III, pp. 159–175. Academic Press, New York, 1972.
- Markowitz, H. Portfolio selection. *J. Financ.*, 7(1):77–91, 1952.
- Osborne, M., Presnell, B., and Turlach, B. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.
- Ritter, K. Ein verfahren zur lösung parameterabhängiger, nichtlinearer maximum-probleme. *Math. Method Oper. Res.*, 6(4):149–166, 1962.
- Rosset, S. and Zhu, J. Piecewise linear regularized solution paths. *Ann. Stat.*, 35(3):1012–1030, 2007.
- Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, 58(1):267–288, 1996.