Complexity Analysis of the Lasso Regularization Path

Julien Mairal and Bin Yu

UC Berkeley

Edinburgh, ICML, June 2012
What this work is about

- another paper about the Lasso/Basis Pursuit [Tibshirani, 1996, Chen et al., 1999]:

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

- the first complexity analysis of the homotopy method [Ritter, 1962, Osborne et al., 2000, Efron et al., 2004] for solving (1);

Some conclusions reminiscent of

- the simplex algorithm [Klee and Minty, 1972];
- the SVM regularization path [Gärtner, Jaggi, and Maria, 2010].
The Lasso Regularization Path and the Homotopy

Under uniqueness assumption of the Lasso solution, the regularization path is piecewise linear:
Our Main Results

Theorem - worst case analysis

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

Proposition - approximate analysis

*There exists an $\varepsilon$-approximate path with $O(1/\sqrt{\varepsilon})$ linear segments.*
Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto w^*(\lambda)$ is continuous and piecewise linear.
Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto w^*(\lambda)$ is continuous and piecewise linear.

Recipe of the homotopy method - main ideas

1. finds a trivial solution $w^*(\lambda_\infty) = 0$ with $\lambda_\infty = \|X^\top y\|_\infty$;
2. compute the direction of the current linear segment of the path;
3. follow the direction of the path by decreasing $\lambda$;
4. stop at the next “kink” and go back to 2.
Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto w^*(\lambda)$ is continuous and piecewise linear.

Recipe of the homotopy method - main ideas

1. finds a trivial solution $w^*(\lambda_\infty) = 0$ with $\lambda_\infty = \|X^T y\|_\infty$;
2. compute the direction of the current linear segment of the path;
3. follow the direction of the path by decreasing $\lambda$;
4. stop at the next “kink” and go back to 2.

Caveats

- kinks can be very close to each other;
- the direction of the path can involve ill-conditioned matrices;
- worst-case exponential complexity (main result of this work).
Worst case analysis

Theorem - worst case analysis

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

![Regularization path, p=6](image_url)
Worst case analysis

Consider a Lasso problem \((y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p})\).
Define the vector \(\tilde{y}\) in \(\mathbb{R}^{n+1}\) and the matrix \(\tilde{X}\) in \(\mathbb{R}^{(n+1) \times (p+1)}\) as follows:

\[
\tilde{y} \triangleq \begin{bmatrix} y & \gamma_{n+1} \end{bmatrix}, \quad \tilde{X} \triangleq \begin{bmatrix} X & 2\alpha y \\ 0 & \alpha \gamma_{n+1} \end{bmatrix},
\]

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

Adverserial strategy

If the regularization path of the Lasso \((y, X)\) has \(k\) linear segments, the path of \((\tilde{y}, \tilde{X})\) has \(3k - 1\) linear segments.
Worst case analysis

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

Let us denote by \( \{\eta^1, \ldots, \eta^k\} \) the sequence of \( k \) sparsity patterns in \( \{-1, 0, 1\}^p \) encountered along the path of the Lasso \((y, X)\).

The new sequence of sparsity patterns for \((\tilde{y}, \tilde{X})\) is

- **first** \( k \) patterns
  \[ \begin{bmatrix} \eta^1 = 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta^2 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} \eta^k \\ 0 \end{bmatrix}, \]

- **middle** \( k \) patterns
  \[ \begin{bmatrix} \eta^k \\ 1 \end{bmatrix}, \begin{bmatrix} \eta^{k-1} \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} \eta^1 = 0 \\ 1 \end{bmatrix}, \]

- **last** \( k-1 \) patterns
  \[ \begin{bmatrix} -\eta^2 \\ 1 \end{bmatrix}, \begin{bmatrix} -\eta^3 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} -\eta^k \\ 1 \end{bmatrix} \]
Worst case analysis

We are now in shape to build a pathological path with \((3^p + 1)/2\) linear segments. Note that this lower-bound complexity is tight.

\[
y \triangleq \begin{bmatrix}
1 \\
1 \\
1 \\
\vdots \\
1
\end{bmatrix}, \quad
x \triangleq \begin{bmatrix}
\alpha_1 & 2\alpha_2 & 2\alpha_3 & \ldots & 2\alpha_p \\
0 & \alpha_2 & 2\alpha_3 & \ldots & 2\alpha_p \\
0 & 0 & \alpha_3 & \ldots & 2\alpha_p \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \alpha_p
\end{bmatrix},
\]
Strong Duality

Strong duality means that $\max_\kappa g(\kappa) = \min_w f(w)$.
Strong duality means that $\max_\kappa g(\kappa) = \min_w f(w)$.

The duality gap guarantees us that $0 \leq f(\tilde{w}) - f(w^*) \leq \delta(\tilde{w}, \tilde{\kappa})$. 
Approximate Complexity

\[ \min_w \left\{ f_\lambda(w) \triangleq \frac{1}{2} \| y - Xw \|^2_2 + \lambda \| w \|_1 \right\}, \quad \text{(primal)} \]

\[ \max_\kappa \left\{ g_\lambda(\kappa) \triangleq -\frac{1}{2} \kappa^\top \kappa - \kappa^\top y \text{ s.t. } \| X^\top \kappa \|_\infty \leq \lambda \right\}. \quad \text{(dual)} \]

\( \varepsilon \)-approximate solution

\( w \) satisfies \( APPROX_\lambda(\varepsilon) \) when there exists a dual variable \( \kappa \) s.t.

\[ \delta_\lambda(w, \kappa) = f_\lambda(w) - g_\lambda(\kappa) \leq \varepsilon f_\lambda(w). \]

\( \varepsilon \)-approximate path

A path \( \mathcal{P} : \lambda \mapsto w(\lambda) \) is an approximate path if it always contains \( \varepsilon \)-approximate solutions.

(see Giesen et al. [2010] for generic results on approximate paths)
Approximate Complexity

Main relation

\[ \text{APPROX}_\lambda(0) \iff \text{APPROX}_\lambda(1 - \sqrt{\varepsilon})(\varepsilon) \]

Key: find an appropriate dual variable \( \kappa(w) \) + simple calculation;

Proposition - approximate analysis

there exists an \( \varepsilon \)-approximate path with at most \( \left\lceil \frac{\log(\frac{\lambda_\infty}{\lambda_1})}{\sqrt{\varepsilon}} \right\rceil \) segments.
Approximate Homotopy

Recipe - main ideas/features

- Maintain approximate optimality conditions along the path;
- Make steps in $\lambda$ greater than or equal to $\lambda(1 - \theta \sqrt{\varepsilon})$;
- When the kinks are too close to each other, make a large step and use a first-order method instead;
- Between $\lambda_\infty$ and $\lambda_1$, the number of iterations is upper-bounded by $\left\lceil \frac{\log(\lambda_\infty / \lambda_1)}{\theta \sqrt{\varepsilon}} \right\rceil$. 
A Few Messages to Conclude

- Despite its exponential complexity, the homotopy algorithm remains extremely powerful in practice;
- the main issue of the homotopy algorithm might be its numerical stability;
- when one does not care about precision, the worst-case complexity of the path can significantly be reduced.
Advertisement SPAMS toolbox (open-source)

- C++ interfaced with Matlab, R, Python.
- proximal gradient methods for $\ell_0$, $\ell_1$, elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-$\ell_0$, sparse group Lasso, overlapping group Lasso...
- ...for square, logistic, multi-class logistic loss functions.
- handles sparse matrices, provides duality gaps.
- fast implementations of OMP and LARS - homotopy.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- coordinate descent, block coordinate descent algorithms.
- fast projections onto some convex sets.

References I


References II


Worst case analysis - Backup Slide

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

Some intuition about the adversarial strategy:

1. the patterns of the new path must be \([\eta^i, 0]^\top\) or \([\pm \eta^i, 1]^\top\);
2. the factor \(\alpha\) ensures the \((p + 1)\)-th variable to enter late the path;
3. after the \(k\) first kinks, we have \(y \approx Xw^*(\lambda)\) and thus

\[ \tilde{X} \begin{bmatrix} w^*(\lambda) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ y_{n+1} \end{bmatrix} \approx \tilde{y} \approx \tilde{X} \begin{bmatrix} -w^*(\lambda) \\ 1/\alpha \end{bmatrix}. \]
Worst case analysis - Backup Slide 2

\[ \min_{\tilde{w} \in \mathbb{R}^p, \tilde{w} \in \mathbb{R}} \frac{1}{2} \left\| \tilde{y} - \tilde{X} \begin{bmatrix} \tilde{w} \\ \tilde{w} \end{bmatrix} \right\|^2_2 + \lambda \left\| \begin{bmatrix} \tilde{w} \\ \tilde{w} \end{bmatrix} \right\|_1 =, \]

\[ \min_{\tilde{w} \in \mathbb{R}^p, \tilde{w} \in \mathbb{R}} \frac{1}{2} \left\| (1 - 2\alpha \tilde{w}) y - X\tilde{w} \right\|^2_2 + \frac{1}{2} \left( y_{n+1} - \alpha y_{n+1} \tilde{w} \right)^2 + \lambda \|\tilde{w}\|_1 + \lambda |\tilde{w}|. \]

is equivalent to

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

and then

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