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]:

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

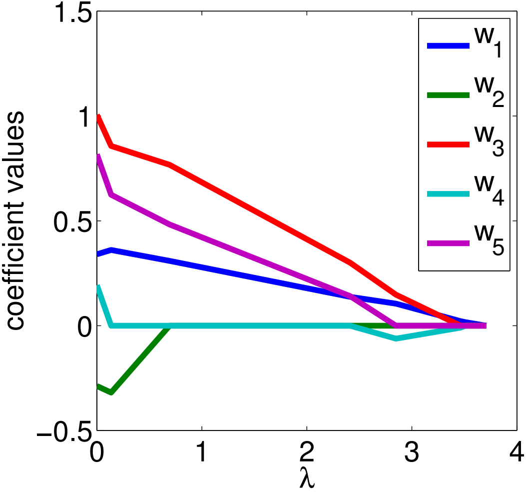
- the first complexity analysis of the homotopy method [Ritter, 1962, Osborne et al., 2000, Efron et al., 2004] for solving (1);

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 ε*-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 $\lambda \mapsto \mathbf{w}^\star(\lambda)$ is continuous and piecewise linear.

Brief Introduction to the Homotopy Algorithm

Piecewise linearity

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

Recipe of the homotopy method - main ideas

1. finds a trivial solution $\mathbf{w}^\star(\lambda_\infty) = 0$ with $\lambda_\infty = \|\mathbf{X}^\top \mathbf{y}\|_\infty$;
2. compute the direction of the current linear segment of the path;
3. follow the direction of the path by decreasing λ;
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 $\lambda \mapsto \mathbf{w}^\star(\lambda)$ is continuous and piecewise linear.

Recipe of the homotopy method - main ideas

1. finds a trivial solution $\mathbf{w}^\star(\lambda_\infty) = 0$ with $\lambda_\infty = \|\mathbf{X}^\top \mathbf{y}\|_\infty$;
2. compute the direction of the current linear segment of the path;
3. follow the direction of the path by decreasing λ;
4. stop at the next "kink" and go back to 2.

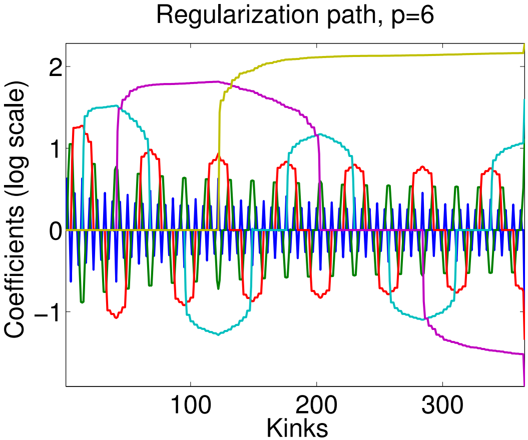
Caveats

- 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.*



Worst case analysis

Consider a Lasso problem ($\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$).
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 $0 < \alpha < \lambda_1/(2\mathbf{y}^\top\mathbf{y} + y_{n+1}^2)$.

Adverserial strategy

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

Worst case analysis

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

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

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

$$\Bigg\{ \overbrace{\begin{bmatrix} \boldsymbol{\eta}^1 = 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\eta}^2 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} \boldsymbol{\eta}^k \\ 0 \end{bmatrix}}^{\text{first } k \text{ patterns}}, \overbrace{\begin{bmatrix} \boldsymbol{\eta}^k \\ 1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\eta}^{k-1} \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} \boldsymbol{\eta}^1 = 0 \\ 1 \end{bmatrix}}^{\text{middle } k \text{ patterns}},$$

$$\underbrace{\begin{bmatrix} -\boldsymbol{\eta}^2 \\ 1 \end{bmatrix}, \begin{bmatrix} -\boldsymbol{\eta}^3 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} -\boldsymbol{\eta}^k \\ 1 \end{bmatrix}}_{\text{last } k-1 \text{ patterns}} \Bigg\}.$$

Worst case analysis

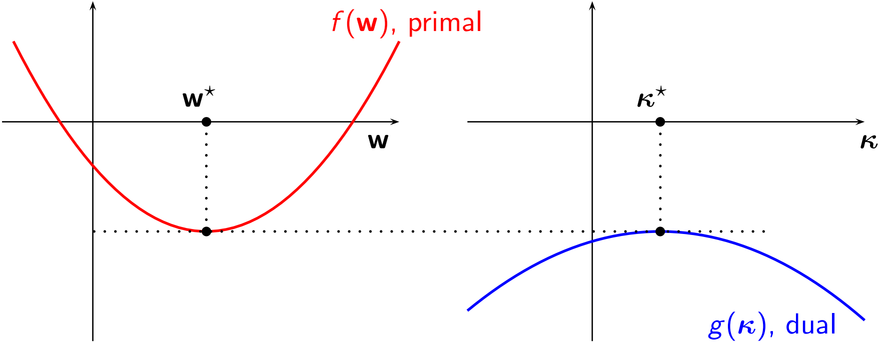
We are now in shape to build a pathological path with $(3^p+1)/2$ linear segments. Note that this lower-bound complexity is tight.

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

Approximate Complexity

Refinement of Giesen, Jaggi, and Laue [2010] for the Lasso

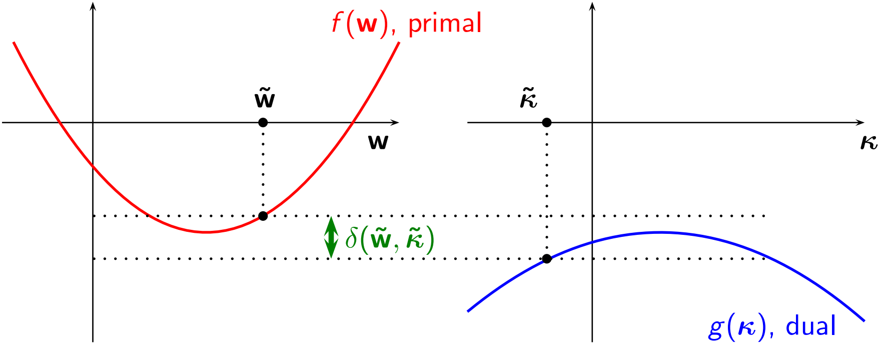
Strong Duality



Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\mathbf{w}} f(\mathbf{w})$

Approximate Complexity

Duality Gaps



Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\mathbf{w}} f(\mathbf{w})$

The duality gap guarantees us that $0 \leq f(\tilde{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \delta(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\kappa}})$.

Approximate Complexity

$$\min_{\mathbf{w}} \left\{ f_\lambda(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \right\}, \qquad \text{(primal)}$$

$$\max_{\boldsymbol{\kappa}} \left\{ g_\lambda(\boldsymbol{\kappa}) \triangleq -\frac{1}{2}\boldsymbol{\kappa}^\top\boldsymbol{\kappa} - \boldsymbol{\kappa}^\top\mathbf{y} \;\; \text{s.t.} \;\; \|\mathbf{X}^\top\boldsymbol{\kappa}\|_\infty \leq \lambda \right\}. \qquad \text{(dual)}$$

ε-approximate solution

$\mathbf{w}$ satisfies $APPROX_\lambda(\varepsilon)$ when there exists a dual variable $\boldsymbol{\kappa}$ s.t.

$$\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) = f_\lambda(\mathbf{w}) - g_\lambda(\boldsymbol{\kappa}) \leq \varepsilon f_\lambda(\mathbf{w}).$$

ε-approximate path

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

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

Approximate Complexity

Main relation

$$APPROX_{\lambda}(0) \Longrightarrow APPROX_{\lambda(1-\sqrt{\varepsilon})}(\varepsilon)$$

Key: find an appropriate dual variable $\boldsymbol{\kappa}(\mathbf{w})$ + simple calculation;

Proposition - approximate analysis

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

Approximate Homotopy

Recipe - main ideas/features

- Maintain approximate optimality conditions along the path;
- Make steps in λ 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 λ_∞ and λ_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 ℓ_0, ℓ_1, **elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-ℓ_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.

Try it! `http://www.di.ens.fr/willow/SPAMS/`

References I

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

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

B. Gärtner, M. Jaggi, and C. Maria. An exponential lower bound on the complexity of regularization paths. *preprint arXiv:0903.4817v2*, 2010.

J. Giesen, M. Jaggi, and S. Laue. Approximating parameterized convex optimization problems. In *Algorithms - ESA*, Lectures Notes Comp. Sci. 2010.

V. Klee and G. J. Minty. How good is the simplex algorithm? In O. Shisha, editor, *Inequalities*, volume III, pages 159–175. Academic Press, New York, 1972.

References II

M. R. Osborne, B. Presnell, and B. A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–37, 2000.

K. Ritter. Ein verfahren zur lösung parameterabhängiger, nichtlinearer maximum-probleme. *Mathematical Methods of Operations Research*, 6(4):149–166, 1962.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Worst case analysis - Backup Slide

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

Some intuition about the adverserial strategy:

1. the patterns of the new path must be $[\boldsymbol{\eta}^{i\top}, 0]^\top$ or $[\pm\boldsymbol{\eta}^{i\top}, 1]^\top$;
2. the factor α ensures the $(p+1)$-th variable to enter late the path;
3. after the k first kinks, we have $\mathbf{y} \approx \mathbf{X}\mathbf{w}^\star(\lambda)$ and thus

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

Worst case analysis - Backup Slide 2

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

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

is equivalent to

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

and then

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