

On the performance of the Lasso in terms of prediction loss

Joint work with M. Hebiri and J. Lederer

Van Dantzig seminar, Amsterdam
October 9, 2014

Arnak S. Dalalyan

ENSAE / CREST / GENES

I. Overcomplete dictionaries and Lasso

Classical problem of regression

- ▶ **Observations** : feature-label pairs $\{(\mathbf{z}_i, y_i); i = 1, \dots, n\}$
 - $\mathbf{z}_i \in \mathbb{R}^d$ multidimensional feature vector ;
 - $y_i \in \mathbb{R}$ real valued label.
- ▶ **Regression function** : for some $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$ it holds that

$$y_i = f^*(\mathbf{z}_i) + \xi_i; \quad \text{with i.i.d. noise } \{\xi_i\}.$$

We will always assume that $\mathbf{E}[\xi_1] = 0$, $\mathbf{Var}[\xi_1] = \sigma^2$.

The feature vectors \mathbf{z}_i are assumed deterministic.

- ▶ **Dictionary approach** : for a given family (called dictionary) of functions $\{\varphi_j\}_{j \in [p]}$, it is assumed that for some $\bar{\beta} \in \mathbb{R}^p$,

$$f^* \approx f_{\bar{\beta}} := \sum_{j=1}^p \bar{\beta}_j \varphi_j.$$

- ▶ **Sparsity** : the dimensionality of $\bar{\beta}$ is large, possibly much larger than n , but it has only a few nonzero entries ($s = \|\bar{\beta}\|_0 \ll p$).

Classical problem of regression

- ▶ **Observations** : feature-label pairs $\{(\mathbf{z}_i, y_i); i = 1, \dots, n\}$
- ▶ **Regression function** : for some $f^* : \mathbb{R}^d \rightarrow \mathbb{R}$ it holds that

$$y_i = f^*(\mathbf{z}_i) + \xi_i; \quad \text{with i.i.d. noise } \{\xi_i\}.$$

- ▶ **Dictionary approach** : for a dictionary $\{\varphi_j\}_{j \in [p]}$,

$$f^* \approx f_{\bar{\beta}} := \sum_{j=1}^p \bar{\beta}_j \varphi_j.$$

- ▶ **Sparsity** : the dimensionality of $\bar{\beta}$ is large, possibly much larger than n , but it has only a few nonzero entries ($s = \|\bar{\beta}\|_0 \ll p$).
- ▶ **Prediction loss** : the quality of recovery is measured by the normalized Euclidean norm :

$$\ell_n(\hat{f}, f^*) = \frac{1}{n} \sum_{i=1}^n \{\hat{f}(\mathbf{z}_i) - f^*(\mathbf{z}_i)\}^2.$$

The goal is to propose an estimator $\hat{\beta}$ such that $\ell_n(f_{\hat{\beta}}, f^*)$ is small.

Equivalence with multiple linear regression

- ✓ Set $\mathbf{y} = (y_1, \dots, y_n)^\top$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$.
- ✓ Define the design matrix $\mathbf{X} = [\varphi_j(\mathbf{z}_i)]_{i \in [n], j \in [p]}$.
- ✓ Assume, for notational convenience, that $f^* = f_{\beta^*}$. We get then the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}.$$

- ✓ The prediction loss of an estimator $\hat{\boldsymbol{\beta}}$ is then

$$\ell_n(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) := \frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2.$$

- ✓ The columns of \mathbf{X} (dictionary elements) satisfy $\frac{1}{n} \|\mathbf{X}^j\|_2^2 \leq 1$.

\mathbf{Y} \mathbf{X} $\boldsymbol{\beta}^*$ $\boldsymbol{\xi}$

$n \times 1$ $n \times p$ $p \times 1$ $n \times 1$

Lasso and its prediction error

- ✓ **Definition** : Given $\lambda > 0$, the Lasso estimator is

$$\hat{\beta}_\lambda^{\text{Lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

- ✓ **Risk bound with “slow” rate** : if $\lambda \geq \sigma \left(\frac{2}{n} \log(p/\delta) \right)^{1/2}$, then

$$\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \leq \min_{\bar{\beta}} \left\{ \ell_n(\bar{\beta}, \beta^*) + 4\lambda \|\bar{\beta}\|_1 \right\}, \quad (1)$$

with probability at least $1 - \delta$ (see, for instance, [Rigollet and Tsybakov, 2011]).

- ✓ For fixed sparsity s , the remainder term is of order $n^{-1/2}$, up to a log factor. This is called “slow” rate.
- ✓ Slow-rate bound holds even if the columns of \mathbf{X} are strongly correlated.

Fast rates for the Lasso

- ✓ Recall the Restricted Eigenvalue condition $\mathbf{RE}(T, 5) : \forall \delta \in \mathbb{R}^p$

$$\|\delta_{T^c}\|_1 \leq 5\|\delta_T\|_1 \quad \Rightarrow \quad \frac{1}{n}\|\mathbf{X}\delta\|_2^2 \geq \kappa_{T,5}^2\|\delta_T\|_2^2.$$

- ✓ **Risk bound with “fast” rate** : according to [Koltchinskii, Lounici and Tsybakov, AoS, 2011], if for some $T \subset [p]$ the matrix \mathbf{X} satisfies $\mathbf{RE}(T, 5)$ and the noise distribution is Gaussian, then $\lambda = 3\sigma\left(\frac{2\log(p/\delta)}{n}\right)^{1/2}$ leads to

$$\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \leq \inf_{\bar{\beta} \in \mathbb{R}^p} \left\{ \ell_n(\bar{\beta}, \beta^*) + 4\lambda\|\bar{\beta}_{T^c}\|_1 + \frac{\sigma^2\|\bar{\beta}\|_0}{n} \frac{18\log(p/\delta)}{\kappa_{T,5}^2} \right\},$$

with probability at least $1 - \delta$ (see also [Sun and Zhang, 2012]).

- ✓ The remainder term above is of order s/n , called fast rate, if $\kappa_{T,5}$ is bounded away from zero. This constrains the correlations between the columns of \mathbf{X} .

II. Some questions

Question 1

- ✓ For really sparse vectors (for example, s is fixed and $n \rightarrow \infty$), there are methods that satisfy fast rate bounds for prediction irrespective of the correlations between the covariates [BTW07a, DT07, RT11].
- ✓ Fast rate bounds for Lasso prediction, in contrast, usually rely on assumptions on the correlations of the covariates such as low coherence [CP09], restricted eigenvalues [BRT09, RWY10], restricted isometry [CT07], compatibility [vdG07], *etc.*
- ✓ **Question** : is it possible to establish fast rate bounds for the Lasso that are valid irrespective of the correlations between the covariates. This question is open even if we allow for oracle choices of the tuning parameter λ , that is, if we allow for λ that depends on the true regression vector β^* , the noise vector ξ , and the noise level σ .

Question 2

- ✓ Known results imply fast rates for prediction with the Lasso in the following two extreme cases : First, when the covariates are mutually orthogonal, and second, when the covariates are all collinear.
- ✓ **Question** : how far from these two extreme cases can a design be such that it still permits fast rates for prediction with the Lasso ?
- ✓ For the first case, the case of mutually orthogonal covariates, this question has been thoroughly studied [BRT09, BTW07b, Zha09, vdGB09, Wai09, CWX10, JN11].
- ✓ For the second case, the case of collinear covariates, this question has received much less attention and is therefore one of our main topics.

Question 3

A particular case of the Lasso is the least squares estimator with the total variation penalty :

$$\hat{\mathbf{f}}^{\text{TV}} \in \arg \min_{\mathbf{f} \in \mathbb{R}^n} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda \|\mathbf{f}\|_{\text{TV}} \right\}, \quad (2)$$

which corresponds to the Lasso estimator for the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad \mathbf{f} = \mathbf{X}\boldsymbol{\beta}, \quad \|\mathbf{f}\|_{\text{TV}} = \|\boldsymbol{\beta}\|_1.$$

- ✓ It is known that if \mathbf{f}^* is piecewise constant, then the minimax rate of estimation is parametric $O(n^{-1})$.
- ✓ According to [MvdG97], the risk of the TV-estimator is $O(n^{-2/3})$.
- ✓ **Question** : Is the TV-estimator indeed suboptimal for estimating piece-wise constant functions or this gap is just an artifact of the proof ?

III. A counter-example

Fast rates : a negative result

- Let $n \geq 2$ and $m = \lfloor \sqrt{2n} \rfloor$. Define the design matrix $\mathbf{X} \in \mathbb{R}^{n \times 2m}$ by

$$\mathbf{X} = \sqrt{\frac{n}{2}} \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}.$$

- We assume in this example that ξ is composed of i.i.d. Rademacher random variables.
- Let $\beta^* \in \mathbb{R}^{2m}$ such that $\beta_1^* = \beta_2^* = 1$ and $\beta_j^* = 0$ for every $j > 2$.

Proposition

For any $\lambda > 0$, the prediction loss of $\hat{\beta}_\lambda^{\text{Lasso}}$ satisfies

$$\mathbf{P}\left(\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \geq (8n)^{-1/2}\right) \geq \frac{1}{2}.$$

Fast rates : a negative result

Other negative results can be found in [CP09], but the specificities of the last proposition are that :

- ◀ the sparsity is fixed and small : $s = 2$, while $p \approx \sqrt{8n}$.
- ◀ the correlations are fixed and bounded away from zero and one : $\langle \mathbf{X}^j, \mathbf{X}^{j'} \rangle = 1/2$ for most j, j' .
- ◀ the result is true for all values of λ .

Conclusion

The statistical complexity of the Lasso is definitely worse than that of the Exponential Screening [RT11] and Exponentially Weighted Aggregate with sparsity prior [DT07].

IV. Taking advantage of correlations : intermediate rates

A measure of (high) correlations and a sharp OI

Recall “slow” rate : if $\lambda \geq \sigma \left(\frac{2}{n} \log(p/\delta) \right)^{1/2}$, then w.p. $\geq 1 - \delta$,

$$\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \leq \min_{\bar{\beta}} \left\{ \ell_n(\bar{\beta}, \beta^*) + 4\lambda \|\bar{\beta}\|_1 \right\}. \quad (3)$$

This bound can be substantially improved when some columns of \mathbf{X} are nearly collinear (very strongly correlated).

For every set $T \subset [p]$, we introduce the quantity

$$\rho_T = n^{-1/2} \max_{j \in [p]} \|(\mathbf{I}_n - \Pi_T) \mathbf{X}^j\|_2,$$

where Π_T is the projector onto $\text{span}(\mathbf{X}_T)$.

Theorem 1

If $\lambda \geq \rho_T \sigma \left(\frac{2}{n} \log(p/\delta) \right)^{1/2}$, with prob. $\geq 1 - 2\delta$ the Lasso fulfills

$$\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \leq \inf_{\bar{\beta} \in \mathbb{R}^p} \left\{ \ell_n(\bar{\beta}, \beta^*) + 4\lambda \|\bar{\beta}\|_1 \right\} + \frac{2\sigma^2(|T| + 2 \log(1/\delta))}{n}.$$

Discussion

- ▶ “Slow” rates meet “fast” rates when the quantity ρ_T is $O(n^{-1/2})$.
- ▶ For designs containing highly correlated covariates (as in the case of the TV-estimator), choosing the tuning parameter substantially smaller than the universal value $\sigma(\frac{2}{n} \log(p/\delta))^{1/2}$ may considerably improve the rate.
- ▶ Applying Theorem 1 in the case of the TV-estimator, we get sharp OI's with a minimax-rate-optimal remainder term in the case of Hölder continuous and monotone functions f .

V. Fast rates and weighted compatibility

Weighted compatibility factors

For any $T \subset [p]$, let us introduce the weights

$$\omega_j(T, \mathbf{X}) = \frac{1}{\sqrt{n}} \|(\mathbf{I}_n - \Pi_T)\mathbf{X}^j\|_2.$$

- ▶ the weights $\omega_j(T, \mathbf{X})$ are all between zero and one,
- ▶ they vanish whenever \mathbf{X}^j belongs to $\text{Span}\{\mathbf{X}^\ell, \ell \in T\}$.

For any $\gamma > 0$, we define the sets

$$\mathcal{C}_0(T, \gamma, \boldsymbol{\omega}) = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|(\mathbf{1}_p - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1 < \|\boldsymbol{\delta}_T\|_1 \right\}.$$

For every vector $\boldsymbol{\omega} \in \mathbb{R}^p$ with nonnegative entries, we call the weighted compatibility factor the quantity

$$\bar{\kappa}_{T, \gamma, \boldsymbol{\omega}} = \inf_{\boldsymbol{\delta} \in \mathcal{C}_0(T, \gamma, \boldsymbol{\omega})} \frac{|T| \cdot \|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n \left\{ \|\boldsymbol{\delta}_T\|_1 - \|(\mathbf{1}_p - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1 \right\}^2}.$$

When $\boldsymbol{\omega} = \mathbf{1}_p$, this coincides with the compatibility factors [vdG07].

Theorem 2

If for some $\gamma > 1$, $\lambda = \gamma\sigma\left(\frac{2}{n}\log(p/\delta)\right)^{1/2}$, then with prob. $\geq 1 - 2\delta$:

$$\ell_n(\hat{\beta}_\lambda^{\text{Lasso}}, \beta^*) \leq \inf_{\bar{\beta}, T} \left\{ \ell_n(\bar{\beta}, \beta^*) + 4\lambda \|\bar{\beta}_{T^c}\|_1 + \frac{4\sigma^2 |T| \log(p/\delta)}{n} \cdot r_{n,p,T} \right\},$$

where $r_{n,p,T} = \log^{-1}(p/\delta) + 2|T|^{-1} + \gamma^2 \bar{\kappa}_{T,\gamma,\omega}^{-1}$.

- ▶ The remainder term converges to zero at the (fast) rate s/n if the weighted compatibility factor is bounded away from zero.
- ▶ The weighted compatibility factor is significantly larger than the unweighted one and can be bounded away from zero even if the columns of \mathbf{X} are strongly correlated.

TV-estimator and piece-wise constant functions

TV-estimator :

$$\hat{\mathbf{f}}^{\text{TV}} \in \arg \min_{\mathbf{f} \in \mathbb{R}^n} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{f}\|_2^2 + \lambda \|\mathbf{f}\|_{\text{TV}} \right\}, \quad (4)$$

which corresponds to the Lasso estimator for the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad \mathbf{f} = \mathbf{X}\boldsymbol{\beta}, \quad \|\mathbf{f}\|_{\text{TV}} = \|\boldsymbol{\beta}\|_1.$$

- ✓ Assume that $f_i^* = f^*(i/n)$ for a piece-wise constant function f^* .
- ✓ Let T be the set of “jumps” of f^* .
- ✓ We managed to prove that the weighted comp. factor satisfies : $\bar{\kappa}_{T,\gamma,\omega}^2 \geq (\log(n) \vee \Delta^{-1})^{-1}$, where Δ is the smallest distance between the jumps of the function f^* .

Proposition 2











Let \mathbf{f}^* be a piecewise constant vector and $J^* = \{j \in [n] : f_j^* \neq f_{j+1}^*\}$. If $\lambda = 2\sigma \left\{ \frac{2}{n} \log(n/\delta) \right\}^{1/2}$, then w. p. $\geq 1 - 2\delta$,

$$\frac{1}{n} \|\hat{\mathbf{f}}^{\text{TV}} - \mathbf{f}^*\|_2^2 \leq \frac{4\sigma^{*2} |J^*| \log(n/\delta)}{n} \cdot (3 + 256(\log(n) + \Delta^{-1})).$$

Some take away messages

- ◀ Generally, the statistical complexity of the Lasso is strictly worse than that of Exponential Screening.
- ◀ Presence of highly correlated covariates may be very helpful when predicting (denoising) with the Lasso.
- ◀ If all the irrelevant covariates are within a distance $O(n^{-1/2})$ of the linear span of the relevant covariates, then the Lasso achieves the fast rate of prediction.
- ◀ (Known) Prediction risk bounds for the Lasso are strictly better than those for the Dantzig selector.
- ◀ TV-estimator does achieve the optimal rate on the class of piece-wise constant functions.

References I

-  Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov, [Simultaneous analysis of lasso and Dantzig selector](#), *Ann. Statist.* **37** (2009), no. 4, 1705–1732. MR 2533469 (2010j :62118)
-  Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp, [Aggregation for Gaussian regression](#), *Ann. Statist.* **35** (2007), no. 4, 1674–1697.
-  _____, [Sparsity oracle inequalities for the Lasso](#), *Electron. J. Stat.* **1** (2007), 169–194. MR 2312149 (2008h :62101)
-  Emmanuel J. Candès and Yaniv Plan, [Near-ideal model selection by \$\ell_1\$ minimization](#), *Ann. Statist.* **37** (2009), no. 5A, 2145–2177. MR 2543688 (2010j :62017)
-  Emmanuel Candès and Terence Tao, [The Dantzig selector : statistical estimation when \$p\$ is much larger than \$n\$](#) , *Ann. Statist.* **35** (2007), no. 6, 2313–2351. MR 2382644 (2009b :62016)
-  Tony Cai, Lie Wang, and Guangwu Xu, [Shifting inequality and recovery of sparse signals](#), *IEEE Trans. Signal Process.* **58** (2010), no. 3, part 1, 1300–1308. MR 2730209 (2011f :94035)
-  Arnak S. Dalalyan and Alexandre B. Tsybakov, [Aggregation by exponential weighting and sharp oracle inequalities](#), *Learning theory (COLT2007)*, *Lecture Notes in Comput. Sci.*, Vol. 4539, 2007, pp. 97–111.
-  Anatoli Juditsky and Arkadi Nemirovski, [Accuracy guarantees for \$\ell_1\$ -recovery](#), *IEEE Trans. Inform. Theory* **57** (2011), no. 12, 7818–7839. MR 2895363
-  Enno Mammen and Sara van de Geer, [Locally adaptive regression splines](#), *The Annals of Statistics* **25** (1997), no. 1, 387–413.
-  Philippe Rigollet and Alexandre Tsybakov, [Exponential Screening and optimal rates of sparse estimation](#), *Ann. Statist.* **39** (2011), no. 2, 731–771.

References II



Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, Restricted eigenvalue properties for correlated Gaussian designs, J. Mach. Learn. Res. **11** (2010), 2241–2259. MR 2719855 (2011h :62272)



Sara van de Geer, The deterministic lasso, Proc. of Joint Statistical Meeting, 2007.



Sara van de Geer and Peter Bühlmann, On the conditions used to prove oracle results for the Lasso, Electron. J. Stat. **3** (2009), 1360–1392.



Martin J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso), IEEE Trans. Inform. Theory **55** (2009), no. 5, 2183–2202. MR 2729873 (2011f :62084)



Tong Zhang, Some sharp performance bounds for least squares regression with L_1 regularization, Ann. Statist. **37** (2009), no. 5A, 2109–2144. MR 2543687 (2010k :62136)