Optimal rates of sparse estimation and universal aggregation

Philippe Rigollet

Princeton University

with A. Tsybakov (Paris VI and CREST)
Prologue: sparsity in linear model

- \( Y = X\theta + \xi \), standard normal \( \xi \).
- \( \dim \theta = M \gg n = \text{sample size} \).
- The Lasso estimator \( \hat{\theta}_L \) w.p. close to 1 satisfies:

\[
|X(\hat{\theta}_L - \theta)|_2^2/n \leq C|\theta|_0 \frac{\log M}{n}, \quad \text{restrictive assumptions on } X.
\]
\[
|X(\hat{\theta}_L - \theta)|_2^2/n \leq C|\theta|_1 \sqrt{\frac{\log M}{n}}, \quad \text{NO assumption on } X.
\]

Here \( |\cdot|_p, p \geq 1 \) is the \( \ell_p \) norm, \( |\theta|_0 = \text{number of non-zero components of } \theta \).

- **Question:** How optimal are these bounds?
Setup

- Regression with **fixed** design.
- We observe

\[ Y_i = \eta(x_i) + \xi_i, \quad i = 1, \ldots, n \]

- where:
  - \( \eta : \mathcal{X} \to \mathbb{R} \) is the unknown regression function,
  - \( x_i, i = 1, \ldots, n \) are known deterministic points in \( \mathcal{X} \),
  - \( \xi_i, i = 1, \ldots, n \) are i.i.d \( \mathcal{N}(0, \sigma^2) \), \( \sigma^2 \) known.
- Performance of an estimator \( \hat{\eta} \)

\[
\| \hat{\eta} - \eta \|^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{\eta}(x_i) - \eta(x_i) \right]^2 \quad \text{(MSE)}
\]
Aggregation

• Given a dictionary $\mathcal{H} = \{f_1, \ldots, f_M\}$, $f_j : \mathcal{X} \rightarrow \mathbb{R}$,

• we are interested in finding the best linear combination of the $f_j$’s:

$$f_\theta = \sum_{j=1}^{M} \theta_j f_j, \quad \theta \in \mathbb{R}^M$$

• More precisely we want to find $\hat{\eta}$ such that

$$\mathbb{E} \|\hat{\eta} - \eta\|^2 - \min_{\theta \in \mathbb{R}^M} \|f_\theta - \eta\|^2$$

is as small as possible.
Oracle inequalities

- Upper bounds for the risk of (linear) aggregation are presented as oracle inequalities of the form

\[ \mathbb{E} \| \hat{\eta} - \eta \|^2 \leq (1 + \varepsilon) \min_{\theta \in \mathbb{R}^M} \| f_\theta - \eta \|^2 + \Delta_{n,M}, \]

- We are interested specifically in the case \( \varepsilon = 0 \) (exact oracle inequalities).

- The smallest possible remainder term \( \Delta_{n,M} \) (optimal rate of linear aggregation)

\[ \Delta_{M,n} = \mathcal{O} \left( \frac{M}{n} \right) \]

and is attained by least squares.
Sparse oracle inequalities

- For good approximation properties: \( M \gg n \) so the rate \( \frac{M}{n} \) is useless.
- Solution: assume sparsity.
- Sparse Oracle Inequality (SOI):

\[
\mathbb{E}\|\hat{\eta} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|f_\theta - \eta\|^2 + \Delta_{n,M}(\theta) \right\},
\]

where \( \Delta_{n,M}(\theta) \) is smaller for “sparser” \( \theta \).
- Notice that the oracle \( \theta^* = \text{argmin}_\theta \|f_\theta - \eta\|^2 \) need not be sparse. Only the best balance between the two terms (approximation and remainder) matters.
Outline

Sparse oracle inequalities when $M \gg n$

- Sparsity pattern aggregation
- Exponential screening

Optimality

Universal aggregation

Implementation and numerical illustration
Sparsity patterns

- A sparsity pattern is a vector $p \in \{0, 1\}^M$.
- Define the set $\mathbb{R}^p$ of vectors with sparsity pattern $p$ as

$$
\mathbb{R}^p = \{ \theta \cdot p : \theta \in \mathbb{R}^M \} \subset \mathbb{R}^M,
$$

where $\theta \cdot p \in \mathbb{R}^M$ denotes the Hadamard product.
- For any $p \in \{0, 1\}^M$ define the least squares estimator

$$
\hat{\theta}_p \in \arg\min_{\theta \in \mathbb{R}^p} |Y - X\theta|^2,
$$

where

$$
Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} f_1(x_1) & \cdots & f_M(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \cdots & f_M(x_n) \end{pmatrix}
$$
Sparsity pattern aggregation

- A first simple oracle inequality gives

\[
\mathbb{E} \| \hat{f}_{\hat{\theta}}^p - \eta \|^2 \leq \min_{\theta \in \mathbb{R}^p} \| f_{\theta} - \eta \|^2 + \sigma^2 \frac{|p|_1 \land R}{n}
\]

where \( R = \text{rank}(X) \).

- \( M \gg n \): \( \frac{M}{n} \) is useless but \( \frac{|p|_1 \land R}{n} \) can be good \( \rightsquigarrow \) which \( p \) to choose?

- Define the sparsity pattern aggregate \( \tilde{\theta}^{SPA} \) by

\[
\tilde{\theta}^{SPA} := \sum_{p \in \{0,1\}^M} \hat{\theta}_p \nu_p,
\]

where \( \nu = (\nu_p)_p \) is a probability measure on \( \{0,1\}^M \).
Exponential screening

• To choose $\nu$, we should downweight sparsity patterns with large SSE and large $|p|_1$.

• Define the probability measure

$$
\nu_p \propto \exp \left( -\frac{1}{4\sigma^2} \sum_{i=1}^{n} (Y_i - f_{\hat{\theta}_p}(x_i))^2 - \frac{|p|}{2} \right) \left( \frac{|p|_1}{2eM} \right)^{|p|_1} I(|p|_1 \leq R)
$$

• The SPA with this $\nu$: Exponential screening $\tilde{\theta}^\text{ES}$.

• George (86), Leung & Barron (06), Giraud (08), Alquier & Lounici (10): exponential weighting with other initial estimators or other discrete priors. Dalalyan & Tsybakov. (07,08,09): exponential weighing with continuous priors.
Sparsity in terms of $\ell_1$ norm

- Several methods based on $\ell_1$ penalization (Lasso, Dantzig) are very efficient.
- SOI for those measure sparsity in terms of $\ell_1$ norm (as opposed to $\ell_0$-norm).
- Becomes an advantage if $|\theta|_1 \ll |\theta|_0$ (many small coefficients, power decay, ...).
- Exponential screening adapts to both measures of sparsity.
Theorem 1

For any $M \geq 1$, $n \geq 1$, if $\max_j \|f_j\| \leq 1$, the sparsity oracle inequality for ES is

$$\mathbb{E}\|f_{\tilde{\theta}_{\text{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|f_\theta - \eta\|^2 + \varphi_{n,M}(\theta) \right\}$$

$$+ \frac{\sigma^2}{n} \left( 9 \log(1 + eM) + 4 \log 2 \right)$$

where the remainder term $\varphi_{n,M}(\theta)$ is equal to

$$\frac{9\sigma^2 \tilde{M}(\theta)}{n} \log \left( \frac{eM}{\tilde{M}(\theta) \lor 1} \right) \land \frac{11\sigma |\theta|_1}{\sqrt{n}} \sqrt{\log \left( 1 + \frac{3eM\sigma}{|\theta|_1 \sqrt{n}} \right)}.$$ 

where $\tilde{M}(\theta) := \min(|\theta|_0, R)$.

Moreover, if $\eta = f_{\theta^*}$, we can take $\varphi_{n,M}(\theta^*) \land |\theta^*|_1^2$ in the remainder term.
Sparsity oracle inequality for ES

**Theorem 1**

For any $M \geq 1, n \geq 1$, if $\max_j \|f_j\| \leq 1$,

$$
\mathbb{E}\|f_{\tilde{\theta}_{\text{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|f_\theta - \eta\|^2 + \varphi_{n,M}(\theta) \right\}
$$

$$
+ \frac{\sigma^2}{n} (9 \log(1 + eM) + 4 \log 2)
$$

where the remainder term $\varphi_{n,M}(\theta)$ is equal to

$$
\frac{9\sigma^2 \tilde{M}(\theta)}{n} \log \left( \frac{eM}{\tilde{M}(\theta) \lor 1} \right) \land \frac{11\sigma |\theta|_1}{\sqrt{n}} \sqrt{\log \left( 1 + \frac{3eM \sigma}{|\theta|_1 \sqrt{n}} \right)}.
$$

where $\tilde{M}(\theta) := \min(|\theta|_0, R)$.

Moreover, if $\eta = f_{\theta^*}$, we can take $\varphi_{n,M}(\theta^*) \land |\theta^*|_1^2$ in the remainder term.
Sparsity oracle inequality for ES

Theorem 1

For any $M \geq 1, n \geq 1$, if $\max_j \|f_j\| \leq 1$,

$$
\mathbb{E}\|f_{\tilde{\theta}_{\text{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|f_\theta - \eta\|^2 + \varphi_{n,M}(\theta) \right\}
$$

$$
+ \frac{\sigma^2}{n} (9 \log(1 + eM) + 4 \log 2)
$$

where the remainder term $\varphi_{n,M}(\theta)$ is equal to

$$
\frac{9\sigma^2 \tilde{M}(\theta)}{n} \log \left( \frac{eM}{\tilde{M}(\theta) \lor 1} \right) \land \frac{11\sigma |\theta|_1}{\sqrt{n}} \sqrt{\log \left( 1 + \frac{3eM\sigma}{|\theta|_1 \sqrt{n}} \right)}.
$$

where $\tilde{M}(\theta) := \min(|\theta|_0, R)$.

Moreover, if $\eta = f_{\theta^*}$, we can take $\varphi_{n,M}(\theta^*) \land |\theta^*|_1^2$ in the remainder term.
Discussion

One and the same estimator takes advantage of three types of sparsity:

- small number of non-zero entries of $\theta$ ($\ell_0$ norm)
- small global weight ($\ell_1$ norm)
- small rank of the matrix $X$
Related results

- SOI have been obtained by Bickel et al. (09), Bunea et al. (07, 07), Candes & Tao (07), Koltchinskii (08, 09, 09), van de Geer (08), Zhang & Huang (08), Zhang (09), ... (other references in those papers).
- Most of those results have the term \((1 + \varepsilon), \varepsilon > 0\) in front of RHS.
- They deal with only one measure of sparsity (either \(|\theta|_0\) or \(|\theta|_1\)) at a time.
- The rates there are slower than in Theorem 1.
- SOI of Theorem 1 holds with no assumption on the dictionary.
We want to prove that \( \psi_{n,M}(\theta) = \varphi_{n,M}(\theta) \wedge |\theta|^2_1 \) is optimal in a minimax sense.

Define the rate function

\[
\zeta_{n,M}(S, \delta) = \frac{\sigma^2 S}{n} \log \left( 1 + \frac{eM}{S} \right) \wedge \frac{\sigma \delta}{\sqrt{n}} \sqrt{\log \left( 1 + \frac{eM \sigma}{\delta \sqrt{n}} \right)} \wedge \delta^2
\]

\( \Rightarrow \zeta_{n,M}(S, \delta) = \psi_{n,M}(\theta) \) with \( \widetilde{M}(\theta) = S \) and \( |\theta|_1 = \delta \).
Minimax lower bound on the intersection of $\ell_0$ and $\ell_1$ balls

Theorem 3

There exists a large class of dictionaries such that for any estimator $T_n$, possibly depending on $\delta, S, n, M, R$ and $H$, there exists a numerical constant $c^* > 0$, such that

$$
\sup_{\eta} \sup_{\theta \in \mathbb{R}_+^M \setminus \{0\}} \left\{ E_{\eta} \|T_n - \eta\|^2 - \|f_\theta - \eta\|^2 \right\} \geq c^* \kappa \zeta_{n,M}(S \land R, \delta),
$$

where $\mathbb{R}_+^M$ is the positive cone of $\mathbb{R}_+^M$.

Least favorable dictionaries satisfy a weak version of restricted isometry (RI) property.
Comparison with asymptotic bounds

- Donoho and Johnstone (92, 94), Abramovich et al. (06)
  - diagonal model: $M = n$, $\mathbf{X}^\top \mathbf{X}/n = I$,
  - asymptotics as $n \to \infty$ of the minimax risk on $\ell_p$ ball $B_p(a)$ with radius $a$.

- Cases: $p = 0$ and $p = 1$. Asymptotic minimax rate

\[
\inf_{\hat{\theta}} \sup_{\theta \in B_0(S)} \mathbb{E}|\mathbf{X}(\hat{\theta} - \theta)|_2^2/n \sim 2\sigma^2 S \log \left( \frac{n}{S} \right)
\]

\[
\inf_{\hat{\theta}} \sup_{\theta \in B_1(\delta)} \mathbb{E}|\mathbf{X}(\hat{\theta} - \theta)|_2^2/n \sim \frac{\delta \sigma}{\sqrt{n}} \sqrt{2 \log \left( \frac{\sigma \sqrt{n}}{\delta} \right)} \wedge \delta^2
\]

- Raskutti et al. (09): $M \neq n$, asymptotic rates $\frac{S}{n} \log \left( \frac{M}{S} \right)$ and $\delta \sqrt{\frac{\log M}{n}}$. Non-asymptotic effects wiped out.
Universal aggregation

• Given $\Theta \subset \mathbb{R}^M$, the goal of aggregation is to construct $\hat{\eta}$ such that

$$
\mathbb{E}\|\hat{\eta} - \eta\|^2 \leq \min_{\theta \in \Theta} \|f_\theta - \eta\|^2 + C\Delta_{n,M}(\Theta), \quad C > 0,
$$

• Different choices of $\Theta$ have been proposed and studied by Nemirovskii (00), Tsybakov (03), Bunea et al. (07) and Lounici (07).

• Optimal rates of aggregations were obtained by Bunea et al. (07) where they showed that the BIC estimator satisfies

$$
\mathbb{E}\|f_{\hat{\theta}_{BIC}} - \eta\|^2 \leq (1 + a) \min_{\theta \in \Theta} \|f_\theta - \eta\|^2 + C \frac{1 + a}{a^2} \Delta_{n,M}
$$

• We call this universal aggregation (one estimator for all problems).
Different types of aggregation

$$\mathbb{E}\|\hat{\eta} - \eta\|^2 \leq \min_{\theta \in \Theta} \|f_{\theta} - \eta\|^2 + C\Delta_{n,M}(\Theta), \quad C > 0,$$

<table>
<thead>
<tr>
<th>Problem</th>
<th>$\Theta$</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(MS)</td>
<td>$\Theta_{(MS)} = {e_1, \ldots, e_M}$</td>
<td>Best in dictionary</td>
</tr>
<tr>
<td>(C)</td>
<td>$\Theta_{(C)} = B_1(1)$</td>
<td>Best convex comb.</td>
</tr>
<tr>
<td>(L)</td>
<td>$\Theta_{(L)} = \mathbb{R}^M$</td>
<td>Best linear comb.</td>
</tr>
<tr>
<td>(LD)</td>
<td>$\Theta_{(LD)} = B_0(D)$</td>
<td>Best $D$-sparse linear comb.</td>
</tr>
<tr>
<td>(CD)</td>
<td>$\Theta_{(CD)} = B_0(D) \cap B_1(1)$</td>
<td>Best $D$-sparse convex comb.</td>
</tr>
</tbody>
</table>

[Bunea et al. (07)]
Theorem 3

Assume that \( \max_{1 \leq j \leq M} \| f_j \| \leq 1 \). Then for any \( M \geq 2, n \geq 1, D \leq M \), and \( \Theta \in \{ \Theta_{(MS)}, \Theta_{(C)}, \Theta_{(L)}, \Theta_{(L_D)}, \Theta_{(C_D)} \} \) the Exponential Screening estimator satisfies the following oracle inequality

\[
\mathbb{E} \| \hat{\theta}_{ES} - \eta \|^2 \leq \min_{\theta \in \Theta} \| \theta - \eta \|^2 + C \Delta^*_n,M(\Theta),
\]

where \( C > 0 \) is a numerical constant and \( \Delta^*_n,M(\Theta) \) is the optimal rate of aggregation on \( \Theta \) given on the next slide.
Optimal rates of aggregation $\Delta^*_{n,M}(\Theta)$

A refinement of the rates with $R$ and $\sigma$ gives

<table>
<thead>
<tr>
<th>Problem</th>
<th>$\Delta^*_{n,M}(\Theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(MS)</td>
<td>$\frac{\sigma^2 \log M}{n}$</td>
</tr>
<tr>
<td>(C)</td>
<td>$\sqrt{\frac{\sigma^2}{n} \log \left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 (M \wedge R)}{n} \log \left(1 + \frac{eM}{M \wedge R}\right)$</td>
</tr>
<tr>
<td>(L)</td>
<td>$\frac{\sigma^2 (M \wedge R)}{n} \log \left(1 + \frac{eM}{M \wedge R}\right)$</td>
</tr>
<tr>
<td>(L_D)</td>
<td>$\frac{\sigma^2 (D \wedge R)}{n} \log \left(1 + \frac{eM}{D \wedge R}\right)$</td>
</tr>
<tr>
<td>(C_D)</td>
<td>$\sqrt{\frac{\sigma^2}{n} \log \left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 (D \wedge R)}{n} \log \left(1 + \frac{eM}{D \wedge R}\right)$</td>
</tr>
</tbody>
</table>
Metropolis-Hastings algorithm

- Recall that the ES estimator $\tilde{\theta}^{ES}$ is:

$$\tilde{\theta}^{ES} = \sum_{p \in \{0,1\}^M} \hat{\theta}_p \nu_p$$

- Virtually $2^M$ least squares estimators to compute.
- Overcome by finding a Markov chain on the vertices $\{0, 1\}^M$ and with stationary distribution

$$\nu_p \propto \exp \left( -\frac{1}{4\sigma^2} \sum_{i=1}^{n} (Y_i - f_{\hat{\theta}_p}(x_i))^2 \right) \left( \frac{|p|_1}{2eM} \right)^{|p|_1} I(|p|_1 \leq R)$$

- We use the uniform proposal but can be improved for faster convergence.
Convergence of the Metropolis-Hastings algorithm

Figure: Typical realization for \((M, n, S') = (500, 200, 20)\). Left: Value of the \(\widetilde{\theta}_T^{ES}\), \(T = 7,000\), \(T_0 = 3,000\). Right: Value of iterate for \(t = 1, \ldots, 5000\). Only the first 50 coordinates are shown for each vector.
Prediction under restricted isometry

• Compare our results in a sparse recovery setting, i.e., when RI property is satisfied.

• Consider the model $Y = X\theta^* + \sigma \xi$ where
  1. $X$ is an $n \times M$ matrix with independent Rademacher entries
  2. $\xi \in \mathbb{R}^n$ is a vector of independent standard Gaussian random variables and is independent of $X$
  3. $\theta_j^* = \mathbb{1}(j \leq S)$ for some fixed $S$ so that $M(\theta^*) = S$
  4. $\sigma^2 = S/9$

• We consider the prediction error

$$|X(\hat{\theta} - \theta^*)|_2^2/n = \|f_{\hat{\theta}} - f_{\theta^*}\|^2.$$ 

(Setup of Candes & Tao (07))
Results

Figure: Boxplots of $|X(\hat{\theta} - \theta^*)|^2/n$ over 500 realizations for the ES, Lasso, cross-validated Lasso (LassoCV), Lasso-Gauss (Lasso-G) and cross-validated Lasso-Gauss (LassoCV-G) estimators. **Left:** $(n, M, S) = (100, 200, 10)$, **right:** $(n, M, S) = (200, 500, 20)$. 
Reconstruction of the digit “6”

- Difficult to actually find $X$ which does not satisfy RI condition and with $M \gg n$.
- Solution: handwritten digit dataset of LeCun et al. (90). Consists of 256 pixels grayscale images.
- Idea: take one image + noise to be $Y$ in $\mathbb{R}^{256}$ and the dictionary to be the remaining 7,290 images.
- Formally

$$Y = \mu + \sigma \xi$$

- We try to approximate $\mu$ with linear combinations of the other images in the dataset.
Correlated dictionary

Figure: Histogram of the $M(M - 1)/2$ correlation coefficients between different images in the database.
Prediction performance

Figure: *Left:* Boxplots of the predictive performance $|\mu - X\hat{\theta}|^2_2$ of the ES, Lasso and Lasso-Gauss (Lasso-G) estimators computed from 250 replications. *Left:* $\sigma = 0.5$. *Right:* $\sigma = 1$. 
Examples of reconstructions

(a) True  (b) Noisy  (c) ES  (d) Lasso  (e) Lasso-G

Figure: Reconstruction of the digit “6” with $\sigma = 0.5$

(a) True  (b) Noisy  (c) ES  (d) Lasso  (e) Lasso-G

Figure: Reconstruction of the digit “6” with $\sigma = 1.0$
Interpretations of the coefficients in $\tilde{\theta}^{ES}$
Set

\[ \nu_p \propto \exp \left( - \frac{1}{4\sigma^2} \sum_{i=1}^{n} (Y_i - f_{\hat{\theta}_p}(x_i))^2 \right) \pi_p, \quad p \in \mathcal{P}. \]

This Gibbs-type distribution can be expressed as the stationary distribution of the Markov chain generated by a Metropolis-Hastings algorithm. Consider the $M$-hypercube graph $\mathcal{G}$ with vertices given by $\mathcal{P}$. For any $p \in \mathcal{P}$, define the instrumental distribution $q(\cdot | p)$ as the uniform distribution on the neighbors of $p$ in $\mathcal{G}$.
Fix $p_0 = 0 \in \mathbb{R}^M$. For any $t \geq 0$, given $p_t \in \mathcal{P}$,

1. Generate a random variable $Q_t$ with distribution $q(\cdot | p_t)$.
2. Generate a random variable

$$P_{t+1} = \begin{cases} Q_t \text{ with probability } r(p_t, Q_t) \\ p_t \text{ with probability } 1 - r(p_t, Q_t) \end{cases}$$

where

$$r(p, q) = \min \left( \frac{\nu_q}{\nu_p}, 1 \right).$$

3. Compute the least squares estimator $\hat{\theta}_{P_{t+1}}$. 

Metropolis-Hastings on the cube