Optimal rates of sparse estimation and universal aggregation

Philippe Rigollet

Princeton University

with A. Tsybakov (Paris VI and CREST)

Prologue: sparsity in linear model

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

$$|\mathbf{X}(\hat{\theta}_L - \theta)|_2^2/n \le C|\theta|_0 \frac{\log M}{n},$$ restrictive assumptions on $\mathbf{X}$.

$$|\mathbf{X}(\hat{\theta}_L - \theta)|_2^2/n \le C|\theta|_1 \sqrt{\frac{\log M}{n}},$$ NO assumption on $\mathbf{X}$.

Here $|\cdot|_p, p \ge 1$ is the ℓ_p norm, $|\theta|_0 =$ number of non-zero components of θ.

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

Setup

- Regression with fixed design.
- We observe

$$Y_i = \eta(x_i) + \xi_i, \quad i = 1, \dots, n$$

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

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

Aggregation

- Given a dictionary $\mathcal{H} = \{f_1, \dots, f_M\}$, $f_j : \mathcal{X} \to \mathbb{R}$,
- we are interested in finding the best linear combination of the f_j's:

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

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

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

is as small as possible.

Oracle inequalities

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

$$\mathbb{E}\|\hat{\eta} - \eta\|^2 \leq (1+\varepsilon) \min_{\theta \in \mathbb{R}^M} \|\mathsf{f}_\theta - \eta\|^2 + \Delta_{n,M} ,$$

- 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):

$$\mathbb{E}\|\hat{\eta} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|\mathsf{f}_\theta - \eta\|^2 + \Delta_{n,M}(\theta) \right\} ,$$

  where $\Delta_{n,M}(\theta)$ is smaller for "sparser" θ.
- Notice that the oracle $\theta^* = \mathrm{argmin}_\theta \|\mathsf{f}_\theta - \eta\|^2$ 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 $\mathsf{p} \in \{0,1\}^M$.
- Define the set $\mathbb{R}^{\mathsf{p}}$ of vectors with sparsity pattern p as

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

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

$$\hat{\theta}_{\mathsf{p}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{\mathsf{p}}} |\mathbf{Y} - \mathbf{X}\theta|_2^2\,,$$

  where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} f_1(x_1) & \dots & f_M(x_1) \\ \vdots & & \vdots \\ f_1(x_n) & \dots & f_M(x_n) \end{pmatrix}$$

Sparsity pattern aggregation

- A first simple oracle inequality gives

$$\mathbb{E}\|\mathsf{f}_{\hat{\theta}_{\mathsf{p}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^{\mathsf{p}}} \|\mathsf{f}_\theta - \eta\|^2 + \sigma^2 \frac{|\mathsf{p}|_1 \wedge R}{n}$$

where $R = \mathrm{rank}(\mathbf{X})$.

- $M \gg n$: $\dfrac{M}{n}$ is useless but $\dfrac{|\mathsf{p}|_1 \wedge R}{n}$ can be good $\rightsquigarrow$ which p to choose?
- Define the sparsity pattern aggregate $\tilde{\theta}^{\mathrm{SPA}}$ by

$$\tilde{\theta}^{\mathrm{SPA}} := \sum_{\mathsf{p} \in \{0,1\}^M} \hat{\theta}_{\mathsf{p}} \nu_{\mathsf{p}} \,,$$

where $\nu = (\nu_{\mathsf{p}})_{\mathsf{p}}$ is a probability measure on $\{0, 1\}^M$.

Exponential screening

- To choose ν, we should downweight sparsity patterns with large SSE and large $|\mathsf{p}|_1$.
- Define the probability measure

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

- The SPA with this ν: Exponential screening $\tilde{\theta}^{\mathrm{ES}}$.
- 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 weigthing with continuous priors.

Sparsity in terms of ℓ_1 norm

- Several methods based on ℓ_1 penalization (Lasso, Dantzig) are very efficient.
- SOI for those measure sparsity in terms of ℓ_1 norm (as opposed to ℓ_0-norm).
- Becomes an advantage if $|\theta|_1 \ll |\theta|_0$ (many small coefficients, power decay, ...).
- Exponential screening adapts to both measures of sparsity.

Sparsity oracle inequality for ES

Theorem 1

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

$$\mathbb{E}\|\mathsf{f}_{\tilde{\theta}^{\mathrm{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{\|\mathsf{f}_\theta - \eta\|^2 + \varphi_{n,M}(\theta)\right\} + \frac{\sigma^2}{n}(9 \log(1 + eM) + 4 \log 2)$$

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

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

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

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

Sparsity oracle inequality for ES

Theorem 1

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

$$\mathbb{E}\|\mathsf{f}_{\tilde{\theta}^{\mathrm{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \{\|\mathsf{f}_\theta - \eta\|^2 + \varphi_{n,M}(\theta)\}$$
$$+ \frac{\sigma^2}{n}(9\log(1+eM) + 4\log 2)$$

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

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

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

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

Sparsity oracle inequality for ES

Theorem 1

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

$$\mathbb{E}\|\mathsf{f}_{\tilde{\theta}^{\mathrm{ES}}} - \eta\|^2 \leq \min_{\theta \in \mathbb{R}^M} \left\{ \|\mathsf{f}_\theta - \eta\|^2 + \varphi_{n,M}(\theta) \right\} + \frac{\sigma^2}{n}(9\log(1+eM) + 4\log 2)$$

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

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

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

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

Discussion

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

- small number of non-zero entries of θ (ℓ_0 norm)
- small global weight (ℓ_1 norm)
- small rank of the matrix $\mathbf{X}$

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.

Minimax lower bounds

- We want to prove that $\psi_{n,M}(\theta) = \varphi_{n,M}(\theta) \wedge |\theta|_1^2$ is optimal in a minimax sense.
- Define the rate function

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

$\leadsto$ $\zeta_{n,M}(S,\delta) = \psi_{n,M}(\theta)$ with $\widetilde{M}(\theta) = S$ and $|\theta|_1 = \delta$.

Minimax lower bound on the intersection of ℓ_0 and ℓ_1 balls

Theorem 3

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

$$\sup_{\eta} \sup_{\substack{\theta \in \mathbb{R}_+^M \setminus \{0\} \\ M(\theta) \leq S \\ |\theta|_1 \leq \delta}} \left\{ E_\eta \|T_n - \eta\|^2 - \|\mathsf{f}_\theta - \eta\|^2 \right\} \geq c^* \kappa \zeta_{n,M}(S \wedge R, \delta) \,,$$

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 ℓ_p ball $B_p(a)$ with radius a.
- Cases: $p = 0$ and $p = 1$. Asymptotic minimax rate

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

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

- 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

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

- Different choices of Θ 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

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

- We call this universal aggregation (one estimator for all problems).

Different types of aggregation

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

Problem	Θ	Description
(MS)	$\Theta_{(\mathrm{MS})} = \{e_1, \dots, e_M\}$	Best in dictionary
(C)	$\Theta_{(\mathrm{C})} = B_1(1)$	Best convex comb.
(L)	$\Theta_{(\mathrm{L})} = \mathbb{R}^M$	Best linear comb.
(L_D)	$\Theta_{(\mathrm{L}_D)} = B_0(D)$	Best D-sparse linear comb.
(C_D)	$\Theta_{(\mathrm{C}_D)} = B_0(D) \cap B_1(1)$	Best D-sparse convex comb.

[Bunea *et al.* (07)]

ES solves all aggregation problems

Theorem 3

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

$$\mathbb{E}\|\mathsf{f}_{\tilde\theta^{\mathrm{ES}}} - \eta\|^2 \le \min_{\theta\in\Theta} \|\mathsf{f}_\theta - \eta\|^2 + C\Delta^*_{n,M}(\Theta)\,,$$

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

Optimal rates of aggregation $\Delta^*_{n,M}(\Theta)$

A refinement of the rates with R and σ gives

Problem	$\Delta^*_{n,M}(\Theta)$
(MS)	$\frac{\sigma^2 \log M}{n}$
(C)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 (M \wedge R)}{n} \log\left(1 + \frac{eM}{M \wedge R}\right)$
(L)	$\frac{\sigma^2 (M \wedge R)}{n} \log\left(1 + \frac{eM}{M \wedge R}\right)$
(L_D)	$\frac{\sigma^2 (D \wedge R)}{n} \log\left(1 + \frac{eM}{D \wedge R}\right)$
(C_D)	$\sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2 (D \wedge R)}{n} \log\left(1 + \frac{eM}{D \wedge R}\right)$

Metropolis-Hastings algorithm

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

$$\tilde{\theta}^{\text{ES}} = \sum_{\mathsf{p}\in\{0,1\}^M} \hat{\theta}_{\mathsf{p}}\nu_{\mathsf{p}}$$

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

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

- We use the uniform proposal but can be improved for faster convergence.

Convergence of the Metroplis-Hastings algorithm

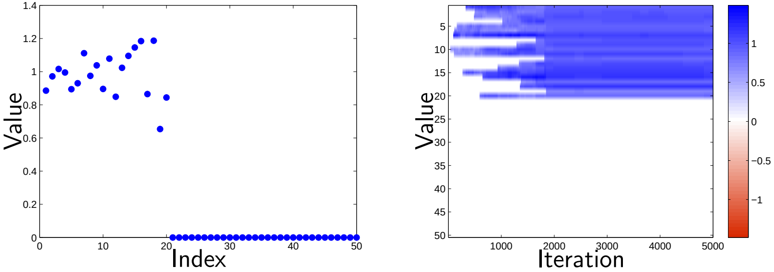


Figure: Typical realization for $(M, n, S) = (500, 200, 20)$. *Left:* Value of the $\tilde{\tilde{\theta}}_T^{\text{ES}}$, $T = 7,000$, $T_0 = 3,000$. *Right:* Value of iterate for $t = 1, \ldots, 5000$. Only the first 50 coordinates are shown for each vector.

Prediction under restricted isometry

- Compare our results in a sparse recovery setting, i.e., when RI property is satisfied.
- Consider the model $\mathbf{Y} = \mathbf{X}\theta^* + \sigma\xi$ where
  1. $\mathbf{X}$ 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 $\mathbf{X}$
  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

$$|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n = \|\mathsf{f}_{\hat{\theta}} - \mathsf{f}_{\theta^*}\|^2.$$

(Setup of Candes & Tao (07))

Results

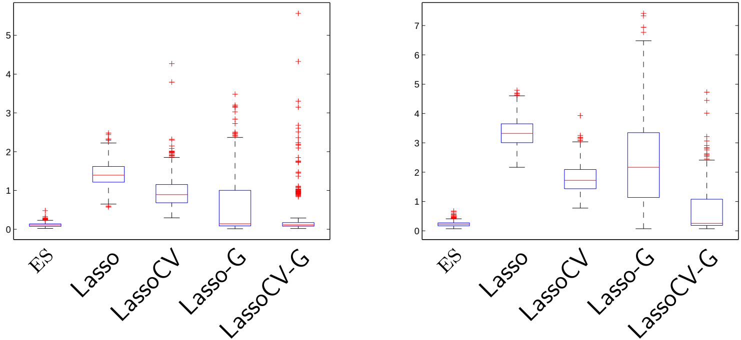
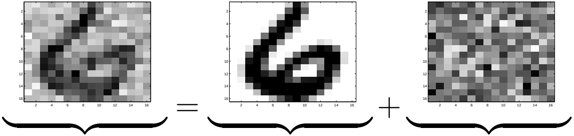


Figure: Boxplots of $|\mathbf{X}(\hat{\theta} - \theta^*)|_2^2/n$ over 500 realizations for the ES, Lasso, cross-validated Lasso (LassoCV), Lasso-Gauss (Lasso-G) and cross-validated Lasso-Gauss (LassoCV-G) estimators. *Left:* $(n, M, S) = (100, 200, 10)$, *right:* $(n, M, S) = (200, 500, 20)$.

Reconstruction of the digit "6"

- Difficult to actually find $\mathbf{X}$ 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 $\mathbf{Y}$ in $\mathbb{R}^{256}$ and the dictionary to be the remaining 7,290 images.
- Formally



$$\underbrace{\mathbf{Y}} = \underbrace{\mu} + \underbrace{\sigma\xi}$$

- We try to approximate μ with linear combinations of the other images in the dataset.

Correlated dictionary

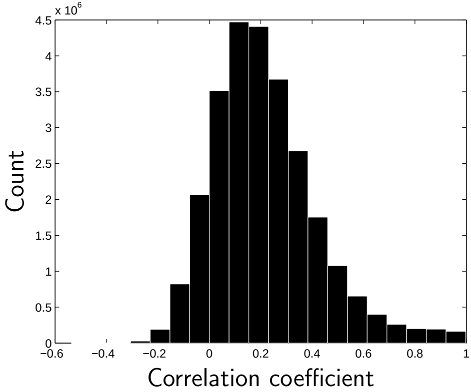


Figure: Histogram of the $M(M-1)/2$ correlation coefficients between different images in the database.

Prediction performance

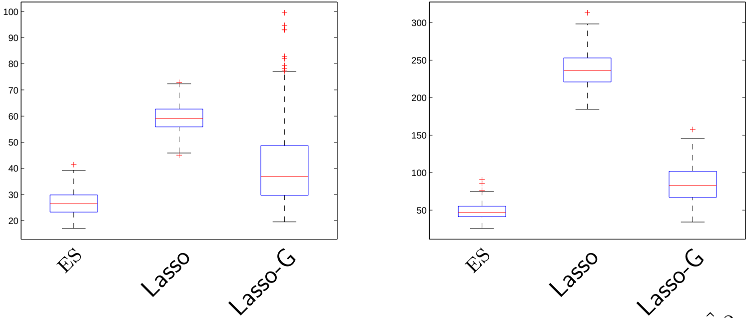
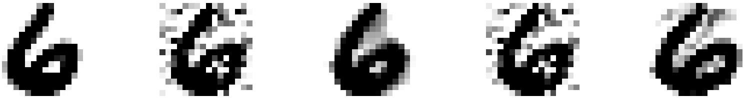


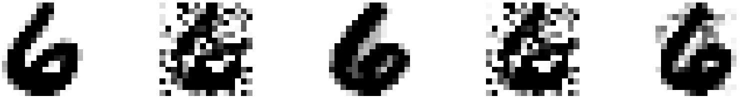
Figure: *Left:* Boxplots of the predictive performance $|\mu - \mathbf{X}\hat{\theta}|_2^2$ of the ES, Lasso and Lasso-Gauss (Lasso-G) estimators computed from 250 replications. *Left:* $\sigma = 0.5$. *Right:* $\sigma = 1$.

Examples of reconstructions



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

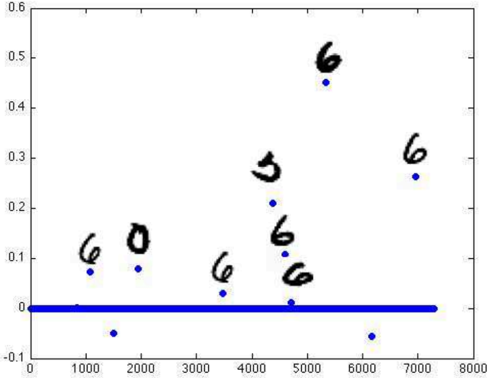
Figure: Reconstruction of the digit "6" with $\sigma = 0.5$



(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 $\tilde{\theta}^{\mathrm{ES}}$



Metropolis-Hastings on the cube

Set

$$\nu_{\mathsf{p}} \propto \exp\Big(- \frac{1}{4\sigma^2} \sum_{i=1}^{n} (Y_i - \mathsf{f}_{\hat{\theta}_{\mathsf{p}}}(x_i))^2 \Big) \pi_{\mathsf{p}} \,, \quad \mathsf{p} \in \mathcal{P} \,.$$

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 $\mathsf{p} \in \mathcal{P}$, define the instrumental distribution $q(\cdot|\mathsf{p})$ as the uniform distribution on the neighbors of p in $\mathcal{G}$.

Metropolis-Hastings on the cube

Fix $\mathsf{p}_0 = 0 \in \mathbb{R}^M$. For any $t \geq 0$, given $\mathsf{p}_t \in \mathcal{P}$,

1. Generate a random variable Q_t with distribution $q(\cdot|\mathsf{p}_t)$.
2. Generate a random variable

$$P_{t+1} = \begin{cases} Q_t & \text{with probability} \quad r(\mathsf{p}_t, Q_t) \\ \mathsf{p}_t & \text{with probability} \quad 1 - r(\mathsf{p}_t, Q_t) \end{cases}$$

   where

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

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