Notes on sparsity

Alexandre Tsybakov

Laboratoire de Statistique, CREST ,
Laboratoire de Probabilités et Modèles Aléatoires,
Université Paris 6
and
CMAP, Ecole Polytechnique

Goettingen, November 18-20 , 2009

Nonparametric regression model (fixed design)

Assume that we observe the pairs $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ where

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n.$$

- Regression function $f : \mathbb{R}^d \to \mathbb{R}$ is unknown
- Errors ξ_i are independent Gaussian $\mathcal{N}(0, \sigma^2)$ random variables.
- $X_i \in \mathbb{R}^d$ are arbitrary fixed (non-random) points.

We want to estimate f based on the data $(X_1, Y_1), \dots, (X_n, Y_n)$.

Approximating function, dictionary

We assume that there exists a function $\mathrm{f}_\theta(x)$ (known as a function of θ and x) such that

$$f \approx \mathrm{f}_\theta$$

for some $\theta = (\theta_1, \dots, \theta_M)$.

Possibly $M \gg n$

Example: linear approximation, dictionary

Let $f_1, \dots, f_M$ be a finite **dictionary of functions**, $f_j : \mathbb{R}^d \to \mathbb{R}$. We approximate the regression function f by linear combination

$$\mathrm{f}_\theta(x) = \sum_{j=1}^{M} \theta_j f_j(x) \quad \text{with weights} \quad \theta = (\theta_1, \dots, \theta_M).$$

We believe that

$$f(x) \approx \sum_{j=1}^{M} \theta_j f_j(x)$$

for some $\theta = (\theta_1, \dots, \theta_M)$.

Scenarios for linear approximation

(LinReg) Exact equality: there exists $\theta^* \in \mathbb{R}^M$ such that
$$f = \mathrm{f}_{\theta^*} = \sum_{j=1}^M \theta_j^* f_j$$
(**linear regression**, with possibly $M \gg n$ parameters);

(NPReg) $f_1, \dots, f_M$ are the first M functions of a basis (usually orthonormal) and $M \le n$, there exists θ^* such that $f - \mathrm{f}_{\theta^*}$ is small: **nonparametric estimation of regression**;

(Agg) **aggregation of arbitrary estimators**: in this case $f_1, \dots, f_M$ are preliminary estimators of f based on a training sample independent of the observations $(X_1, Y_1), \dots, (X_n, Y_n)$;

Weak learning, additive models etc.

Example: nonlinear approximation

We consider the generalized linear (or single-index) model

$$f_\theta(x) = G(\theta^T x)$$

where $G : \mathbb{R} \to \mathbb{R}$ is a known (or unknown) function.

Thus, we believe that

$$f(x) \approx G(\theta^T x)$$

for some $\theta = (\theta_1, \dots, \theta_M)$, possibly with $M \gg n$.

Sparsity of a vector

The number of non-zero coordinates of θ:

$$M(\theta) = \sum_{j=1}^{M} \mathbb{I}_{\{\theta_j \neq 0\}}$$

The value $M(\theta)$ characterizes the **sparsity** of vector $\theta \in \mathbb{R}^M$: the smaller $M(\theta)$, the "sparser" θ.

Sparsity of the model

Intuitive formulation of sparsity assumption:

$f(x) \approx \mathrm{f}_\theta \quad$ ("f is well approximated by f_θ")

where the vector $\theta = (\theta_1, \dots, \theta_M)$ is sparse:

$$M(\theta) \ll M.$$

Sparsity and dimension reduction

Let $\widehat{\theta}_{\mathrm{OLS}}$ be the ordinary least squares (OLS) estimator. Let f_θ be **linear** approximation. Elementary result:

$$\mathbb{E}\|\mathrm{f}_{\widehat{\theta}_{\mathrm{OLS}}} - f\|_n^2 \le \|f - \mathrm{f}_\theta\|_n^2 + \frac{\sigma^2 M}{n}$$

for any $\theta \in \mathbb{R}^M$ where $\|\cdot\|_n$ is the empirical norm:

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)}.$$

Sparsity and dimension reduction

For any $\theta \in \mathbb{R}^M$ the "oracular" OLS that acts only on the relevant $M(\theta)$ coordinates satisfies

$$\mathbb{E}\|\mathrm{f}^{\text{oracle}}_{\widehat{\theta}_{\text{OLS}}} - f\|_n^2 \le \|f - \mathrm{f}_\theta\|_n^2 + \frac{\sigma^2 M(\theta)}{n}.$$

This is only an OLS oracle, not an estimator. The set of relevant coordinates should be known.

Sparsity oracle inequalities

Do there exist true estimators with similar behavior? Basic idea: Choose some suitable data-driven weights $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_M)$ and estimate f by

$$\widehat{f}(x) = \mathrm{f}_{\widehat{\theta}}(x) = \sum_{j=1}^{M} \widehat{\theta}_j f_j(x).$$

- What to do when the approximation is non-linear (ex. $G(\theta^T x)$)? Should we also plug in an estimator $\widehat{\theta}$?
- Can we find $\widehat{\theta}$ such that $\tilde{f} = \mathrm{f}_{\widehat{\theta}}$ or $\tilde{f}$ defined in differently satisfies

$$\mathbb{E}\|\tilde{f} - f\|_n^2 \lesssim \|f - \mathrm{f}_\theta\|_n^2 + \frac{\sigma^2 M(\theta)}{n}, \quad \forall \theta?$$

Sparsity oracle inequalities (SOI)

Realizable task: Construct an estimator $\tilde{f}$ satisfying a **sparsity oracle inequality** (SOI)

$$\mathbb{E}\|\tilde{f} - f\|_n^2 \le \inf_{\theta \in \mathbb{R}^M} \left\{ C\|f - \mathrm{f}_\theta\|_n^2 + C' \frac{M(\theta)(''\log M'')}{n} \right\}$$

with some constants $C \ge 1$, $C' > 0$ and an inevitable extra $''\log M''$ in the variance term.

$$C = 1 \Rightarrow \textbf{sharp SOI}.$$

Implications of SOI: Scenario (LinReg)

Assume that we have found an estimator $\mathrm{f}_{\widehat{\theta}}$ satisfying SOI. Some consequences for different scenarios:

(LinReg) **linear regression**: $f = \mathrm{f}_{\theta^*}$ for some θ^*. Using SOI:

$$\begin{aligned} \mathbb{E}\|\mathrm{f}_{\widehat{\theta}} - f\|_n^2 &\leq C\left\{\|f - \mathrm{f}_{\theta^*}\|_n^2 + \frac{M(\theta^*)\log M}{n}\right\} \\ &= \frac{CM(\theta^*)\log M}{n} \end{aligned}$$

(the desired result for Scenario (LinReg)).

Implications of SOI: Scenario (NPReg)

(NPReg) **nonparametric regression**. If f belongs to standard smoothness classes of functions, $\min_{\theta\in\Theta_m}\|f-\mathrm{f}_\theta\|_n \le Cm^{-\beta}$ for some $\beta>0$ (Θ_m = the set of vectors with only first m non-zero coefficients, $m\le M$). Using SOI:

$$\begin{aligned}
\mathbb{E}\|\mathrm{f}_{\widehat{\theta}}-f\|_n^2 &\le C\inf_{m\ge 1}\left\{\min_{\theta\in\theta_m}\|f-\mathrm{f}_\theta\|_n^2+\frac{m\log M}{n}\right\}\\
&\le C\inf_{m\ge 1}\left\{\frac{1}{m^{2\beta}}+\frac{m\log M}{n}\right\}\\
&= O\left(\left(\frac{\log n}{n}\right)^{2\beta/(2\beta+1)}\right)\quad \text{for } M\le n
\end{aligned}$$

(optimal rate of convergence, up to logs, in Scenario (NPReg)).

Implications of SOI: Scenario (Agg)

(Agg) **aggregation of arbitrary estimators**: in this case $f_1, \dots, f_M$ are preliminary estimators of f based on a pilot (training) sample independent of the observations $(X_1, Y_1), ..., (X_n, Y_n)$. The training sample is considered as frozen. Assume that SOI holds with leading constant 1. Then:

$$\begin{aligned} \mathbb{E}\|\mathrm{f}_{\widehat{\theta}} - f\|_n^2 &\leq \inf_{\theta \in \mathbb{R}^M} \left\{ \|f - \mathrm{f}_\theta\|_n^2 + \frac{CM(\theta)\log M}{n} \right\} \\ &\leq \min_{1 \leq j \leq M} \|f - f_j\|_n^2 + \frac{C \log M}{n} \end{aligned}$$

$\Longrightarrow$ $\mathrm{f}_{\widehat{\theta}}$ attains optimal rate of Model Selection type aggregation $\frac{\log M}{n}$ (T., 2003).

Implications of SOI: Scenario (Agg)

Similar conclusion holds for Convex aggregation. We restrict θ to the simplex

$$\Theta^M = \{\theta \in \mathbb{R}^M : \theta_j \geq 0, \sum_{j=1}^M \theta_j = 1\}.$$

From SOI with leading constant 1 + "Maurey argument":

$$\begin{aligned} \mathbb{E}\|\mathrm{f}_{\widehat{\theta}} - f\|_n^2 &\leq \inf_{\theta \in \mathbb{R}^M} \left\{ \|f - \mathrm{f}_\theta\|_n^2 + \frac{CM(\theta)\log M}{n} \right\} \\ &\leq \inf_{\theta \in \Theta^M} \|f - \mathrm{f}_\theta\|_n^2 + C'\sqrt{\frac{\log M}{n}}. \end{aligned}$$

$\Longrightarrow$ $\mathrm{f}_{\widehat{\theta}}$ attains optimal rate of Convex aggregation $\sqrt{\frac{\log M}{n}}$ [Nemirovski (2000)].

"Ideal" requirements for SOI

We would like to construct an estimator $\tilde{f}$ such that it satisfies:

- SOI with leading constant 1 (sharp SOI);
- this holds under no assumptions on the approximation function; in the linear case, under no assumptions on the dictionary $f_1, \dots, f_M$;
- the estimator is computationally feasible

Penalized techniques (BIC, Lasso)

Penalize the residual sum of squares directly by $M(\theta)$ (BIC criterion, Schwarz (1978), Foster and George (1994)):

$$\widehat{\theta}^{BIC} = \arg\min_{\theta\in\mathbb{R}^M} \left\{ \|\mathbf{y} - \mathrm{f}_\theta\|_n^2 + \gamma \frac{M(\theta)\log M}{n} \right\},$$

where $\gamma > 0$ and

$$\|\mathbf{y} - \mathrm{f}_\theta\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^n \Big(Y_i - \mathrm{f}_\theta(X_i)\Big)^2, \quad \mathbf{y} = (Y_1, \dots, Y_n).$$

Remarks:

- If the matrix $X = (f_j(X_i))_{i,j}$ has orthnormal columns, BIC is equivalent to hard thresholding of the components of $X^T\mathbf{y}/n$ at the level $\sqrt{\gamma(\log M)/n}$.
- Non-convex, discontinuous minimization problem.

Sparsity oracle inequality for BIC (linear approximation)

Theorem. [Bunea/ T/ Wegkamp (2004)]: if $\gamma > K_0\sigma^2$ for an absolute constant K_0, and **with no assumption on the dictionary** $f_1, \dots, f_M$, the BIC estimator satisfies, with probability close to 1,

$$\|\mathrm{f}_{\widehat{\theta}^{BIC}} - f\|_n^2 \le (1+\varepsilon) \inf_{\theta \in \mathbb{R}^M} \left\{ \|f - \mathrm{f}_\theta\|_n^2 + C(\varepsilon) \frac{M(\theta) \log M}{n} \right\}, \quad \forall \varepsilon > 0.$$

Remarks:

- the BIC is realizable only for small M (say, $M \le 20$),
- the leading constant is **not** 1,
- $C(\varepsilon) \sim 1/\varepsilon$.
- no result for non-linear approximation

LASSO

Second penalization technique: LASSO [Frank and Friedman (1993, Bridge regression), Tibshirani (1996), Chen and Donoho (1998, basis pursuit)]. Penalize the residual sum of squares not by $M(\theta)$, as in the BIC, but by the ℓ_1-norm of θ:

$$\widehat{\theta}^L = \arg\min_{\theta \in \mathbb{R}^M} \left\{ \|\mathbf{y} - \mathrm{f}_\theta\|_n^2 + 2r|\theta|_1 \right\},$$

where $|\theta|_1 = \sum_{j=1}^M |\theta_j|$, $r > 0$ a tuning constant. A sensible choice:

$$r \sim \sqrt{\frac{\log M}{n}}.$$

- If the matrix $X = (f_j(X_i))_{i,j}$ has orthonormal columns, LASSO is equivalent to soft thresholding of the components of $X^T\mathbf{y}/n$ at the level r.

Restricted eigenvalue assumption

For a vector $\boldsymbol{\Delta} = (a_j)_{j=1,\dots,M}$ and a subset of indices $J \subseteq \{1, \dots, M\}$ write

$$\boldsymbol{\Delta}_J = (a_j \mathbf{1}\{j \in J\})_{j=1,\dots,M}.$$

The Gram matrix: $\Psi_M = \left(\langle f_j, f_{j'} \rangle_n\right)_{1 \le j,j' \le M} (= X^T X/n)$.

Assumption RE(s, c_0). (Bickel, Ritov and T., 2007)

For an integer $1 \le s \le M$ and $c_0 > 0$ there exists $\kappa = \kappa(s, c_0)$:

$$\boldsymbol{\Delta}^T \Psi_M \boldsymbol{\Delta} \ge \kappa |\boldsymbol{\Delta}_J|_2^2$$

for all $J \subseteq \{1, \dots, M\}$ such that $|J| \le s$ and $|\boldsymbol{\Delta}_{J^c}|_1 \le c_0 |\boldsymbol{\Delta}_J|_1$.

More specific assumptions

Assumption RE is more general than several other assumptions on the Gram matrix:

- Coherence assumption (Donoho/Elad/Temlyakov),
- "Uniform uncertainty principle" (Candès/Tao),
- Incoherent design assumption (Meinshausen/Yu, Zhang/Huang).

These papers focus on the linear regression scenario (LinReg).

Sparsity oracle inequality for the LASSO

Theorem [Bickel, Ritov and T., 2009]

Let $\|f_j\|_n = 1, j = 1, \dots, M$. Fix some $\varepsilon > 0$. Let Assumption RE(s, c_0) be satisfied with $c_0 = 3 + 4/\varepsilon$. Consider the LASSO estimator $\mathrm{f}_{\widehat{\theta}^L}$ with the tuning constant

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

for some $A > 2\sqrt{2}$. Then, for all $M \geq 3$, $n \geq 1$ with probability at least $1 - M^{1-A^2/8}$ we have: $\forall\, \theta \in \mathbb{R}^M : M(\theta) = s$,

$$\|\mathrm{f}_{\widehat{\theta}^L} - f\|_n^2 \leq (1+\varepsilon)\|\mathrm{f}_\theta - f\|_n^2 + C(\varepsilon)\left(\frac{M(\theta)\log M}{\boxed{\kappa} n}\right).$$

Advantages of the LASSO: computationally simple, selects the sparsity pattern [Bühlmann and Meinshausen (2004), Zhao and Yu (2006), Lounici (2008)], ...

Disadvantages of the LASSO:

- SOI for the LASSO holds under strong assumptions on the dictionary involving minimal "restricted eigenvalues". Moreover, the assumptions depend on the (unknown) number s of non-zero components of the oracle vector, or eventually on the upper bound on this number. Such assumptions are unavoidable: Candès and Plan (2009).
- The leading constant in SOI is **not** 1.
- How to deal with non-linear approximations?

Same problems with other ℓ_1 penalized techniques (Dantzig selector, modifications of the Lasso).

Dantzig selector and LASSO for linear regression

Scenario (LinReg): $f = \mathrm{f}_{\theta^*}$ for some θ^*, so that we can rewrite our model as the standard linear regression:

$$\mathbf{y} = X\theta^* + \xi$$

where the matrix $X = (f_j(X_i))_{i,j}, i = 1, \dots, n,\ j = 1, \dots, M$ and ξ is the Gaussian random vector of noise.
Dantzig selector (Candès and Tao, 2007):

$$\widehat{\theta}_D \triangleq \arg\min \left\{ |\theta|_1 : \ \left| \frac{1}{n} X^T (\mathbf{y} - X\theta) \right|_\infty \le r \right\}.$$

where $|\cdot|_p$ denotes the ℓ_p norm in $\mathbb{R}^M$.

Theorem [Bickel, Ritov and T., 2009]

Let $\|f_j\|_n = 1, j = 1, \dots, M$. *Let Assumption RE*$(s, 3)$ *hold and let* $\widehat{\theta}$ *be either LASSO or Dantzig selector with tuning parameter* $r = A\sigma\sqrt{\frac{\log M}{n}}$ *and* $A > 2\sqrt{2}$. *Then, for all* $M \geq 3$, $n \geq 1$, *with probability at least* $1 - M^{1-A^2/8}$ *we have*

$$|X(\widehat{\theta} - \theta^*)|_2^2/n \leq \frac{C'}{\kappa}\frac{M(\theta^*)\log M}{n} \quad \text{(SOI for LASSO /Dantzig)}$$

$$|\widehat{\theta} - \theta^*|_p^p \leq \frac{C}{\kappa} M(\theta^*)\left(\sqrt{\frac{\log M}{n}}\right)^p, \quad \forall\ 1 \leq p \leq 2.$$

Selection of the sparsity pattern

Selection of the sparsity pattern [Lounici (2008)]: under the coherence assumption, with probability close to 1,

$$|\widehat{\theta} - \theta^*|_\infty \le \frac{C}{\kappa}\sqrt{\frac{\log M}{n}}$$

where $\widehat{\theta}$ is LASSO or Dantzig estimator; their thresholded versions $\tilde{\theta}$ satisfy:

$$P(J_{\tilde{\theta}} = J_{\theta^*}) \to 1 \quad \text{if } \min_{j \in J_{\theta^*}} |\theta_j^*| > \frac{C'}{\kappa}\sqrt{\frac{\log M}{n}}\,.$$

Remarks on the Dantzig selector

- Advantages: extreme comutational simplicity. The computation reduces to linear programming and can be realized in higher dimensional models than for the Lasso.
- Disadvantages: the same as for the Lasso.
- Slightly less convenient than the Lasso when the model is not exactly the linear one (e.g., in nonparametric regression scenario). Needs extra conditions guaranteeing that the target satisfies the Dantzig constraint.

Exponential weighting

On the difference from Lasso and Dantzig selector, the method of sparse exponential weighting requires no assumption on the dictionary. Estimate $f(x)$ by

$$\tilde{f}^{EW}(x) = \int_{\mathbb{R}^M} \mathrm{f}_\theta(x) S_n(d\theta)$$

where the probability measure S_n is given by

$$S_n(d\theta) = \frac{\exp\left\{ -n\|\mathbf{y} - \mathrm{f}_\theta\|_n^2/\beta \right\}\pi(d\theta)}{\int_{\mathbb{R}^M} \exp\left\{ -n\|\mathbf{y} - \mathrm{f}_w\|_n^2/\beta \right\}\pi(dw)}$$

with some $\beta > 0$ and some prior measure π.

Exponential weighting

- For the linear approximation: $\tilde{f}^{EW} = \mathrm{f}_{\widehat{\theta}^{EW}}$ where

$$\widehat{\theta}_j^{EW} = \int_{\mathbb{R}^M} \theta_j S_n(d\theta), \quad j = 1, \dots, M,$$

- Bayesian estimator if $\beta = 2\sigma^2$, but we need a larger β.
- Non-discrete π: Computational issues?

A PAC-Bayesian bound

Lemma [Dalalyan and T., 2007]

The estimator with exponential weights $\tilde{f}^{EW}$ *defined with* $\beta \geq 4\sigma^2$ *and any prior* π *satisfies:*

$$\mathbb{E}\|\tilde{f}^{EW} - f\|_n^2 \leq \inf_P \left\{ \int \|f_\theta - f\|_n^2 \, P(d\theta) + \frac{\beta \, \mathcal{K}(P, \pi)}{n} \right\}$$

where the infimum is taken over all probability measures P *on* $\mathbb{R}^M$ *and* $\mathcal{K}(P, \pi)$ *denotes the Kullback-Leibler divergence between* P *and* π*.*

Sparsity prior

Choose a specific prior measure π with Lebesgue density q:

$$q(\theta) = \prod_{j=1}^{M} \tau^{-1} \, q_0\big(\theta_j/\tau\big), \ \forall \theta \in \mathbb{R}^M,$$

where q_0 is the Student t_3 density,

$$q_0(t) \sim |t|^{-4}, \ \text{for large } |t|$$

and $\tau \sim (Mn)^{-1/2}$. We will call this prior the **sparsity prior**. The resulting estimator $\tilde{f}^{EW}$ is called the **Sparse Exponential Weighting (SEW)** estimator.

SOI for the SEW estimator: Linear approximation case

Theorem [Dalalyan and T., 2007]

Let $\max_{1\le j\le M}\|f_j\|_n \le c_0 < \infty$. *Let* f_θ *be linear in* θ. *Then for* $\beta \ge 4\sigma^2$ *the estimator* $\mathrm{f}_{\widehat{\theta}^{EW}}$ *with the* **sparsity prior** *satisfies:*

$$\mathbb{E}\|\mathrm{f}_{\widehat{\theta}^{EW}} - f\|_n^2 \le \inf_{\theta\in\mathbb{R}^M}\left\{\|\mathrm{f}_\theta - f\|_n^2 + \frac{CM(\theta)}{n}\log\left(1+\frac{|\theta|_1\sqrt{Mn}}{M(\theta)}\right)\right\}$$

where $|\theta|_1$ *is the* ℓ_1*-norm of* θ.

- **No assumption on the dictionary**.
- **Leading constant 1**.
- ℓ_1-norm of θ, but under the log.
- Fast computation for at least $M \sim 10^3$.

SOI for the SEW estimator: generalized linear models

Assume now:

$$f_\theta(x) = G(x^T\theta).$$

Then we have the same result as for the linear approximation case provided that

$$\sup_\theta \mathrm{Spec}\left\{\frac{1}{n}\sum_{i=1}^n G''(X_i^T\theta)X_iX_i^T\right\} \le c_0 < \infty.$$

SEW estimator: discussion

- *SEW is* **not** *a penalized estimator.*

$$\widehat{\theta}_j^{EW} = \int_{\mathbb{R}^M} \theta_j S_n(d\theta) = \int_{\mathbb{R}^M} \theta_j g_n(\theta) d\theta, \quad j = 1, \dots, M,$$

with posterior density $g_n(\theta) = S_n(d\theta)/d\theta$:

$$g_n(\theta) \quad \propto \quad \exp\Big\{ - n\|\mathbf{y} - \mathrm{f}_\theta\|_n^2/\beta - C\sum_{j=1}^M \log(1+\theta_j^2/\tau)\Big\}$$

Maximizer of this density (the MAP estimator):

$$\widehat{\theta}^{MAP} = \arg\min_{\theta\in\mathbb{R}^M} \Big\{ \|\mathbf{y} - \mathrm{f}_\theta\|_n^2 + \frac{\gamma}{n}\sum_{j=1}^M \log(1+\theta_j^2/\tau)\Big\} \neq \widehat{\theta}^{EW}.$$

Exponential weights: models with i.i.d. data

- An i.i.d. sample $Z_1, \dots, Z_n$ from the distribution of an abstract random variable $Z \in \mathcal{Z}$.
- $Q(Z, \mathrm{f}_\theta)$ a given real-valued loss (prediction loss).

Define the probability measure S_n on $\mathbb{R}^M$ by

$$S_n(d\theta) = \frac{\exp\Big\{-\sum_{i=1}^n Q(Z_i, \mathrm{f}_\theta)/\beta\Big\}\pi(d\theta)}{\int_{\mathbb{R}^M} \exp\Big\{-\sum_{i=1}^n Q(Z_i, \mathrm{f}_w)/\beta\Big\}\pi(dw)}$$

with some $\beta > 0$ and some prior measure π. Generalization of the previous definition: we replace

$$n\|\mathbf{y} - \mathrm{f}_\theta\|_n^2 \quad \rightsquigarrow \quad \sum_{i=1}^n Q(Z_i, \mathrm{f}_\theta).$$

Mirror averaging

- Cumulative exponential weights (**mirror averaging**):

$$\widehat{\theta}_j^{MA} = \int_{\mathbb{R}^M} \theta_j S(d\theta), \quad j = 1, \dots, M, \quad \text{with} \quad S = \frac{1}{n} \sum_{i=1}^{n} S_i$$

  cf. Juditsky/Rigollet/T (2008) [even more general method: Juditsky/Nazin/T/Vayatis (2005)]. In a particular case we get the "progressive mixture method" of Catoni and Yang.

- Choose a prior measure π supported on a convex compact $\theta \subset \mathbb{R}^M$ (e.g., on an ℓ_1 ball).

Assumption JRT (2008).

The mapping $\theta \mapsto Q(Z, \mathrm{f}_\theta)$ is convex for all Z and there exists $\beta > 0$ such that the function

$$\theta \mapsto \mathbb{E} \exp\left(\frac{Q(Z, \mathrm{f}_{\theta'}) - Q(Z, \mathrm{f}_\theta)}{\beta} \right)$$

is concave on a convex compact set $\theta \subset \mathbb{R}^M$ for all $\theta' \in \theta$.

Roughly: "strong convexity on the average".

PAC-Bayesian bound for mirror averaging

Define the average risk: $A(\theta) = \mathbb{E}Q(Z, \mathrm{f}_\theta)$.

Lemma (PAC-Bayesian bound).

Let $\mathrm{f}_{\widehat{\theta}^{MA}}$ be a mirror averaging estimator defined with β satisfying Assumption JRT and any prior π supported on a convex compact set θ. Then

$$\mathbb{E}\, A(\widehat{\theta}^{MA}) \le \inf_P \left\{ \int A(\theta)\, P(d\theta) + \frac{\beta\, \mathcal{K}(P, \pi)}{n+1} \right\}$$

where the infimum is taken over all probability measures P on θ and $\mathcal{K}(P, \pi)$ is the Kullback-Leibler divergence between P and π.

Proof follows the scheme of Juditsky, Rigollet and T. (2008), cf. also Lounici (2007), Audibert (2009).

SOI for Mirror Averaging

Theorem

Assume that $\sup_{|\theta|_1 \le 2R} \mathrm{Spec}\{\nabla^2 A(\theta)\} < \infty$ *for some* $R > 0$. *Let* $\mathsf{f}_{\widehat{\theta}^{MA}}$ *be a mirror averaging estimator satisfying assumptions of the PAC lemma, with the* **sparsity prior** π *truncated to* $\{\theta : |\theta|_1 \le 2R\}$ *and* $\tau \sim 1/\sqrt{M(n \vee M)}$. *Then*

$$\mathbb{E}\, A(\widehat{\theta}^{MA}) \le \inf_{|\theta|_1 \le R} \left\{ A(\theta) + \frac{CR^2 M(\theta)}{n} \log\left(\frac{C' R \sqrt{M(n \vee M)}}{M(\theta)} \right) \right\}.$$

- **No restrictive assumption on the dictionary.**
- **Leading constant 1.**

Comparison with SOI for the LASSO

The LASSO type estimators

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, f_\theta) + r \sum_{j=1}^{M} |\theta_j| \right\} .$$

van de Geer (2008), Koltchinskii (2008,2009):

$$\mathbb{E}\, A(\widehat{\theta}) \le \inf_{|\theta|_1 \le R} \left(\boxed{3} A(\theta) + \frac{C R^2 M(\theta) \log M}{\boxed{\kappa}\, n} \right)$$

where κ is a "Restricted Eigenvalue", can be very small.

Example: Gaussian regression, squared loss

- Gaussian regression with random design :
  $Z = (X, Y), \quad X \in \mathbb{R}^d, \quad Y \in \mathbb{R}$ such that

$$Y = f(X) + \xi,$$

  $\xi | X \sim \mathcal{N}(0, \sigma^2), \ X \sim P_X, \ \|f\|_\infty \le L.$

- Assumption on the dictionary: $\|f_j\|_\infty \le L, \ j = 1, \dots, M.$
- The loss function
  $Q(Z, \mathrm{f}_\theta) = \big(Y - \mathrm{f}_\theta(X)\big)^2$ where $\mathrm{f}_\theta = \sum_{j=1}^M \theta_j f_j.$
- Then $A(\theta) = \mathbb{E}\, Q(Z, \mathrm{f}_\theta) = \|\mathrm{f}_\theta - f\|_X^2 + \sigma^2$, $\|f\|_X^2 \triangleq \int f^2 dP_X.$

SOI for regression with squared loss

Corollary

Under the conditions of this example, for all $\beta \geq 2\sigma^2 + 8L^2$,

$$\mathbb{E}\,\|f_{\widehat{\theta}^{MA}} - f\|_X^2 \leq \inf_{\theta \in \Theta^M} \left\{ \|f_\theta - f\|_X^2 + \frac{CM(\theta)}{n} \log\left(\frac{C'\sqrt{M(n \vee M)}}{M(\theta)} \right) \right\}.$$

Here Θ^M is the simplex:

$$\Theta^M = \{\theta \in \mathbb{R}^M : \theta_j \geq 0, \sum_{j=1}^M \theta_j = 1\}.$$

Example: density estimation with L_2 loss

- $Z = X \in \mathbb{R}^d$ with density f, such that $\|f\|_\infty \le L$.
- Assumption on the dictionary: $f_1, \dots, f_M$ are probability densities such that $\|f_j\|_\infty \le L$.
- The loss function:

$$Q(X, \mathrm{f}_\theta) = \|\mathrm{f}_\theta\|^2 - 2\mathrm{f}_\theta(X) \quad \text{where} \quad \|f\|^2 = \int f^2(x)dx\,.$$

- The associated risk:

$$A(\theta) = \mathbb{E}\, Q(X, \mathrm{f}_\theta) = \|f - \mathrm{f}_\theta\|^2 - \|f\|^2.$$

SOI for density estimation with L_2 loss

Corollary

Under the conditions of this example, for all $\beta > 12L$,

$$\mathbb{E}\,\|\mathrm{f}_{\widehat{\theta}^{MA}} - f\|^2 \le \inf_{\theta\in\Theta^M}\left\{\|\mathrm{f}_\theta - f\|^2 + \frac{CM(\theta)}{n}\log\left(\frac{C'\sqrt{M(n\vee M)}}{M(\theta)}\right)\right\}.$$

Here Θ^M is the simplex:

$$\Theta^M = \{\theta\in\mathbb{R}^M : \theta_j \ge 0, \sum_{j=1}^M \theta_j = 1\}.$$

Modified SEW estimators

Take the modified sparsity prior

$$q(\theta) \propto \left(\prod_{j=1}^{M} \frac{e^{-\omega(\alpha\theta_j)}}{\left(1+|\theta_j|/\tau\right)^2} \right) \mathbf{1}\{|\theta|_1 \le R\}$$

where $\omega(\cdot)$ is Huber's function

$$\omega(t) = \begin{cases} t^2, & \text{if } |t| \le 1, \\ 2|t|, & \text{if } |t| > 1, \end{cases}$$

α and τ are small (ex.: $\alpha \sim M^{-1}$, $\tau \sim n^{-1/2}$), R is large ($R \sim M$).

Computation of SEW estimators

Consider the linear regression scenario:

$$\mathbf{y} = X\theta + \xi.$$

X is a $n \times M$ deterministic design matrix, $\theta \in \mathbb{R}^M$ is an unknown vector and $\xi \in \mathbb{R}^M$ is a Gaussian vector with i.i.d. components, with variances σ^2. The SEW estimator

$$\hat{\theta}^{EW} \triangleq \int_{\mathbb{R}^M} \mathbf{u}\, g(\mathbf{u})\, d\mathbf{u}$$

where the posterior density

$$g(\mathbf{u}) \propto \exp(-V(\mathbf{u}))$$

$$V(\mathbf{u}) = \beta^{-1}\|\mathbf{y} - X\mathbf{u}\|^2 + 2\sum_{j=1}^{M} \log(\tau^2 + u_j^2).$$

Langevin Monte Carlo

Remark: the posterior density $g(\cdot)$ is the invariant density of the Langevin diffusion

$$\mathbf{L}_t = -\nabla V(\mathbf{L}_t)\, dt + \sqrt{2}\, d\mathbf{W}_t, \quad \mathbf{L}_0 = 0, \quad t > 0.$$

Here $\mathbf{W}_t$ is the M-dimensional Brownian motion.

Let now $\eta_1, \eta_2, \ldots$ be i.i.d. standard normal random vectors. Set

$$\overline{\mathbf{L}}_0 = 0, \quad \overline{\mathbf{L}}_{k+1} = \overline{\mathbf{L}}_k - h\nabla V(\overline{\mathbf{L}}_k) + \sqrt{2h}\, \eta_k, \qquad k = 0, 1, \ldots.$$

Then

$$\frac{1}{[Th^{-1}]} \sum_{k=1}^{[Th^{-1}]} \overline{\mathbf{L}}_k \approx \frac{1}{T} \int_0^T \mathbf{L}_t\, dt \xrightarrow[T\to\infty]{a.s.} \int_{\mathbb{R}^M} \mathbf{u} g(\mathbf{u})\, d\mathbf{u} = \hat{\theta}^{EW}.$$

Simulations

Example 1: Compressed sensing

The entries of the matrix X are i.i.d. Rademacher random variables independent of the noise ξ.

$$\theta_j = \mathbf{1}\{j \le S\} \qquad \text{and} \qquad \sigma^2 = \frac{S}{9n}.$$

We apply the SEW estimator using Langevin Monte-Carlo with

$$\tau = 4\sigma/\sqrt{M}, \qquad \beta = 4\sigma^2, \qquad h = 0.0001.$$

Simulations

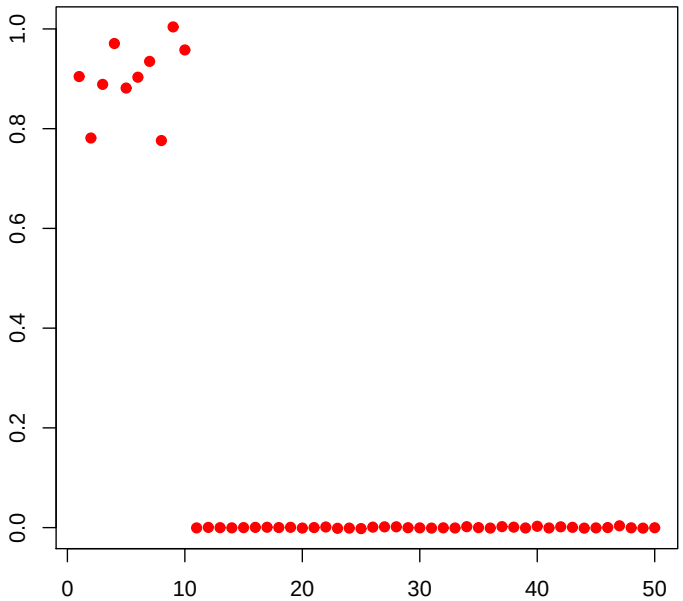
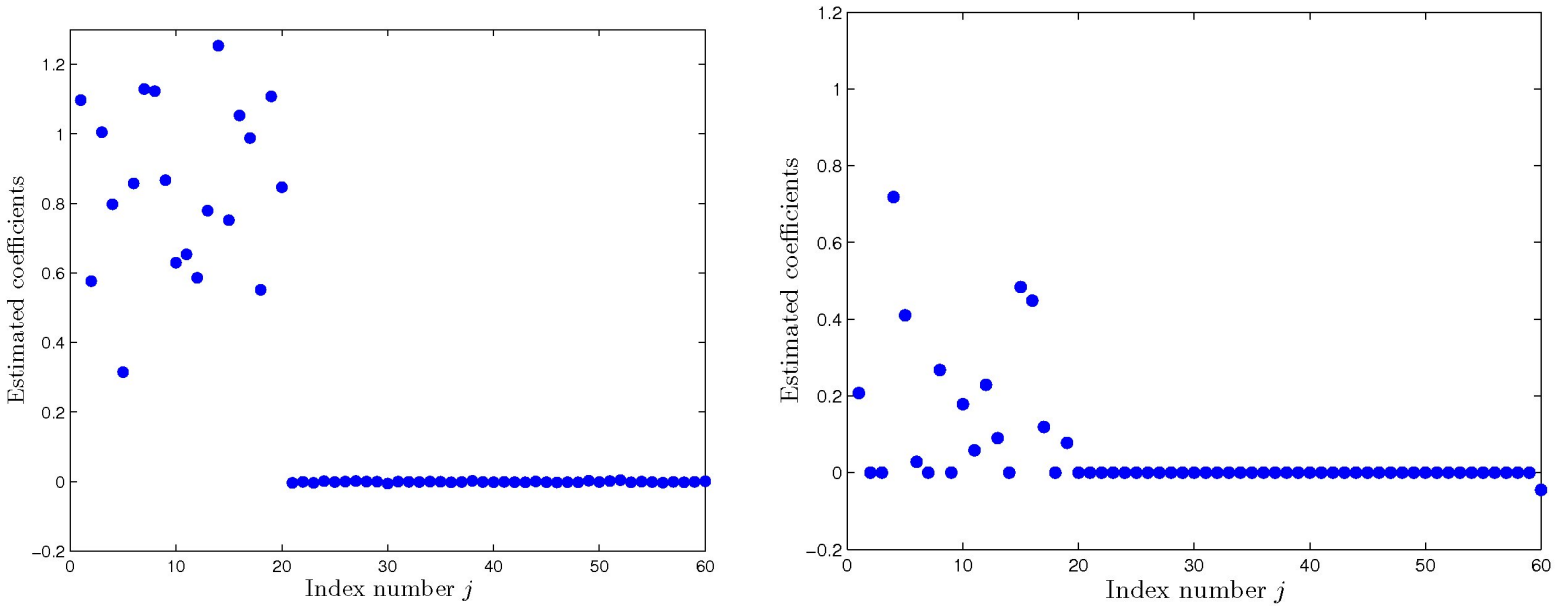


Figure: Typical result for Example 1 with $n = 200$, $M = 500$, $S = 10$, $h = 10^{-4}$, $T = 5$. SEW estimates of the first 50 coefficients are plotted. The prediction error $\frac{1}{n}|X(\hat{\theta} - \theta)|_2^2 = 0.0021$. The time of computation $\sim$ 30 seconds.

Example 1: Compressed sensing

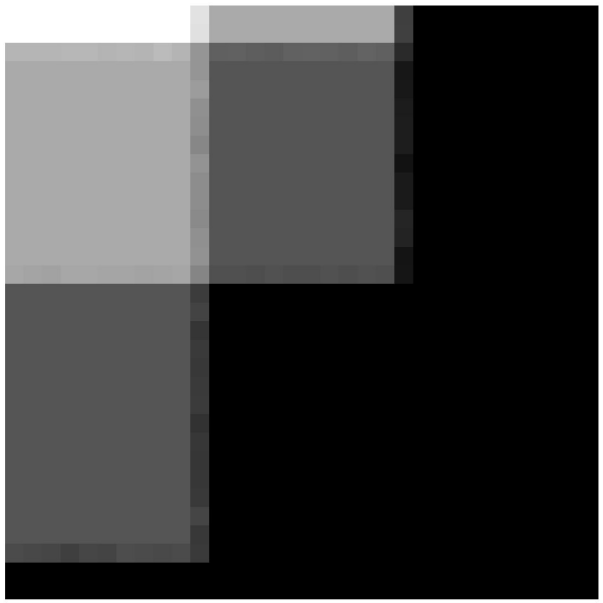


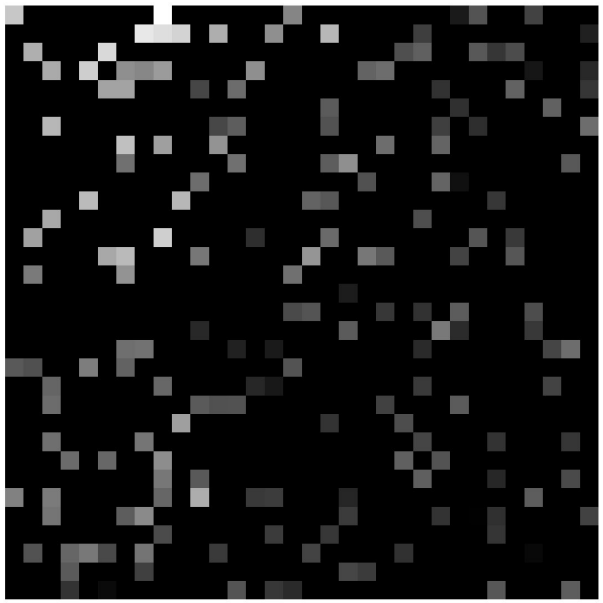
Typical outcome for $n = 200$, $M = 500$ and $S = 20$. Left panel: SEW, right panel: LASSO

Example 2: Image denoising

- Input: n, k positive integers and $\sigma > 0$.
- We generate n vectors U_i of $\mathbb{R}^2$ uniformly distributed in $[0,1]^2$.
- Covariates $\phi_j(u) = \mathbf{1}_{\{[0,j_1/k]\times[0,j_2/k]}(u)$.
- Errors: we generate a centered Gaussian vector ξ with covariance matrix $\sigma^2 I$.
- Response: $Y_i = (\phi_1(U_i), \dots, \phi_{k^2}(U_i))^T \theta + \xi_i$ where $\theta = [\mathbf{1}\{(j \in \{10, 100, 200\})]'$.
- Tuning parameters: the same rule as previously.

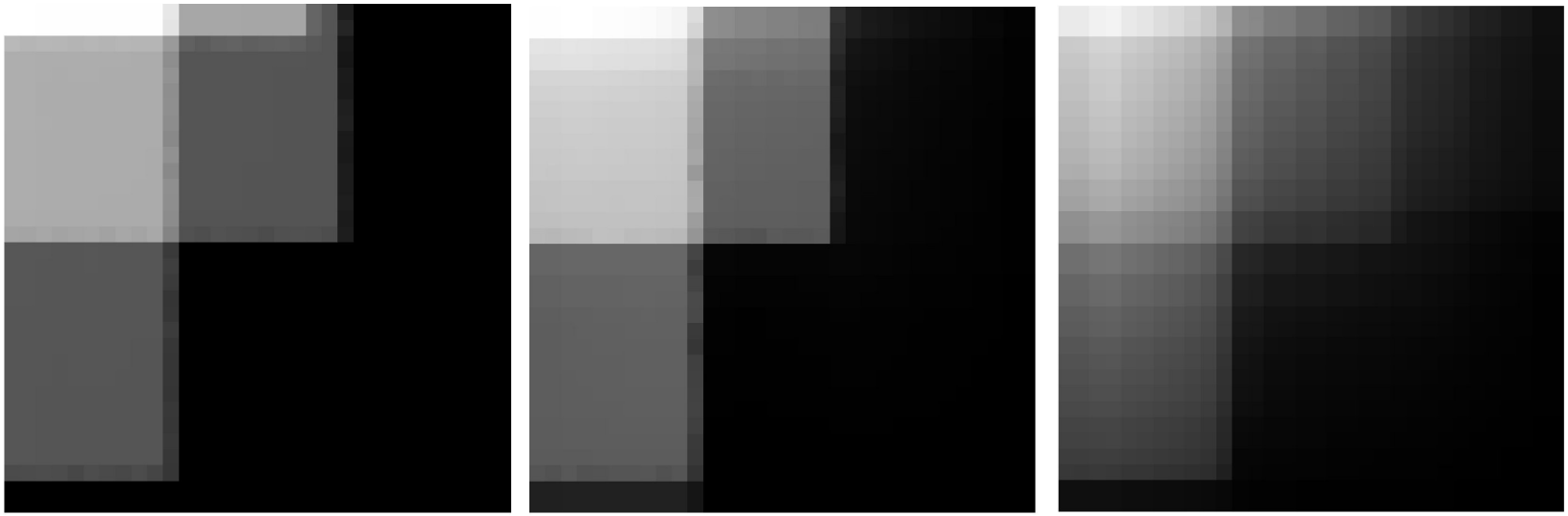
Image denoising





The original image and its sampled noisy version.

Image denoising



Estimated images from observations with noise magnitudes 0.1, 0.5 and 1.

Image denoising

Prediction errors and their standard deviations

σ	$n = 100$			$n = 200$		
	SEW	Lasso	Ideal LG	SEW	Lasso	Ideal LG
2	0.210 (0.072)	0.759 (0.562)	0.330 (0.145)	0.187 (0.048)	0.661 (0.503)	0.203 (0.086)
4	0.420 (0.222)	2.323 (1.257)	0.938 (0.631)	0.278 (0.132)	2.230 (1.137)	0.571 (0.324)

Bickel,P.J., Ritov, Y. and Tsybakov, A.B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, v.37, 1705–1732.

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007) Aggregation for Gaussian regression. *Annals of Statistics*, v.35, 1674-1697.

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007) Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, v.1, 169-194.

Dalalyan, A. and Tsybakov, A.B. (2007) Aggregation by exponential weighting and sharp oracle inequalities. *Proceedings of COLT-2007*, 97-111.

Dalalyan, A. and Tsybakov, A.B. (2008) Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, v.72, 39-61.

Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008) Learning by mirror averaging. *Annals of Statistics*, v. 36, 2183-2206.