

Sparse Exponential Weighting as an alternative to LASSO and Dantzig selector

Alexandre Tsybakov

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

Vienna, July 24, 2008

Nonparametric regression model

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $(X_1, Y_1), \dots, (X_n, Y_n)$.

Dictionary, linear approximation

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

$$f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x) \quad \text{with weights} \quad \lambda = (\lambda_1, \dots, \lambda_M).$$

We believe that

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

for some $\lambda = (\lambda_1, \dots, \lambda_M)$.

Dictionary, linear approximation

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

$$f_\lambda(x) = \sum_{j=1}^M \lambda_j f_j(x) \quad \text{with weights} \quad \lambda = (\lambda_1, \dots, \lambda_M).$$

We believe that

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

for some $\lambda = (\lambda_1, \dots, \lambda_M)$.

Possibly $M \gg n$

Scenarios

(LinReg) Exact equality: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = f_{\lambda^*} = \sum_{j=1}^M \lambda_j^* f_j$$

(**linear regression**, with possibly $M \gg n$ parameters);

Scenarios

(LinReg) Exact equality: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = f_{\lambda^*} = \sum_{j=1}^M \lambda_j^* f_j$$

(**linear regression**, with possibly $M \gg n$ parameters);

(NPRReg) f_1, \dots, f_M are the first M functions of a basis (usually orthonormal) and $M \leq n$, there exists λ^* such that $f - f_{\lambda^*}$ is small: **nonparametric estimation of regression**;

Scenarios

(LinReg) Exact equality: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = f_{\lambda^*} = \sum_{j=1}^M \lambda_j^* f_j$$

(**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 $M \leq n$, there exists λ^* such that $f - f_{\lambda^*}$ 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 $(X_1, Y_1), \dots, (X_n, Y_n)$;

Scenarios

(LinReg) Exact equality: there exists $\lambda^* \in \mathbb{R}^M$ such that

$$f = f_{\lambda^*} = \sum_{j=1}^M \lambda_j^* f_j$$

(**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 $M \leq n$, there exists λ^* such that $f - f_{\lambda^*}$ 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**: f_1, \dots, f_M are “weak learners”, i.e., some rough approximations to f ; M is extremely large.

Sparsity of a vector

The number of non-zero coordinates of λ :

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

The value $M(\lambda)$ characterizes the **sparsity** of vector $\lambda \in \mathbb{R}^M$: the smaller $M(\lambda)$, the “sparser” λ .

Sparsity of the model

Intuitive formulation of sparsity assumption:

$$f(x) \approx \sum_{j=1}^M \lambda_j f_j(x) \quad ("f \text{ is well approximated by } f_\lambda")$$

where the vector $\lambda = (\lambda_1, \dots, \lambda_M)$ is sparse:

$$M(\lambda) \ll M.$$

Strong sparsity

Strong sparsity:

f admits an exact sparse representation

$$f = f_{\lambda^*}$$

for some $\lambda^* \in \mathbb{R}^M$, with

$$M(\lambda^*) \ll M$$

⇒ Scenario (LinReg)

Sparsity and dimension reduction

Let $\hat{\lambda}_{\text{OLS}}$ be the ordinary least squares (OLS) estimator.
Elementary result:

$$\mathbb{E} \|f_{\hat{\lambda}_{\text{OLS}}} - f\|_n^2 \leq \|f - f_{\lambda}\|_n^2 + \frac{\sigma^2 M}{n}$$

for any $\lambda \in \mathbb{R}^M$ 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 $\lambda \in \mathbb{R}^M$ the “oracular” OLS that acts only on the relevant $M(\lambda)$ coordinates satisfies

$$\mathbb{E} \|\hat{f}_{\hat{\lambda}_{\text{OLS}}}^{\text{oracle}} - f\|_n^2 \leq \|f - f_\lambda\|_n^2 + \frac{\sigma^2 M(\lambda)}{n}.$$

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

Sparsity oracle inequalities

Do there exist estimators with similar behavior? Choose some other data-driven weights $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_M)$ and estimate f by

$$\hat{f}(x) = f_{\hat{\lambda}}(x) = \sum_{j=1}^M \hat{\lambda}_j f_j(x).$$

Can we find $\hat{\lambda}$ such that

$$\mathbb{E} \|f_{\hat{\lambda}} - f\|_n^2 \lesssim \|f - f_{\lambda}\|_n^2 + \frac{\sigma^2 M(\lambda)}{n}, \quad \forall \lambda?$$

Sparsity oracle inequalities (SOI)

Realizable task: look for an estimator $f_{\hat{\lambda}}$ satisfying a **sparsity oracle inequality (SOI)**

$$\mathbb{E} \|f_{\hat{\lambda}} - f\|_n^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ C \|f - f_{\lambda}\|_n^2 + C' \frac{M(\lambda) \log M}{n} \right\}$$

with some constants $C \geq 1$, $C' > 0$ and an inevitable extra $\log M$ in the variance term. $C = 1 \Rightarrow$ **sharp SOI**.

Sparsity oracle inequalities (SOI)

Realizable task: look for an estimator $f_{\hat{\lambda}}$ satisfying a **sparsity oracle inequality (SOI)**

$$\mathbb{E} \|f_{\hat{\lambda}} - f\|_n^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ C \|f - f_{\lambda}\|_n^2 + C' \frac{M(\lambda) \log M}{n} \right\}$$

with some constants $C \geq 1$, $C' > 0$ and an inevitable extra $\log M$ in the variance term. $C = 1 \Rightarrow$ **sharp SOI**.

“In probability” form of sparsity oracle inequalities:

with probability close to 1,

$$\|f_{\hat{\lambda}} - f\|_n^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ C \|f - f_{\lambda}\|_n^2 + C' \frac{M(\lambda) \log M}{n} \right\}.$$

Implications of SOI: Scenario (LinReg)

Assume that we have found an estimator $f_{\hat{\lambda}}$ satisfying SOI. Some consequences for different scenarios:

(LinReg) **linear regression**: $f = f_{\lambda^*}$ for some λ^* . Using SOI:

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

(the desired result for Scenario (LinReg)).

Implications of SOI: Scenario (NPReg)

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

$$\begin{aligned} \mathbb{E} \|\hat{f}_\lambda - f\|_n^2 &\leq C \inf_{m \geq 1} \left\{ \min_{\lambda \in \Lambda_m} \|f - f_\lambda\|_n^2 + \frac{m \log M}{n} \right\} \\ &\leq C \inf_{m \geq 1} \left\{ \frac{1}{m^{2\beta}} + \frac{m \log M}{n} \right\} \\ &= O \left(\left(\frac{\log n}{n} \right)^{2\beta/(2\beta+1)} \right) \quad \text{for } M \leq n \end{aligned}$$

(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 $(X_1, Y_1), \dots, (X_n, Y_n)$. The training sample is considered as frozen. Assume that SOI holds with leading constant 1. Then:

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

$\implies f_{\hat{\lambda}}$ 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 λ to the simplex

$$\Lambda^M = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1\}.$$

From SOI with leading constant 1 + “Maurey argument”:

$$\begin{aligned} \mathbb{E} \|f_{\hat{\lambda}} - f\|_n^2 &\leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \|f - f_\lambda\|_n^2 + \frac{CM(\lambda) \log M}{n} \right\} \\ &\leq \inf_{\lambda \in \Lambda^M} \|f - f_\lambda\|_n^2 + C' \sqrt{\frac{\log M}{n}}. \end{aligned}$$

$\implies f_{\hat{\lambda}}$ attains optimal rate of Convex aggregation $\sqrt{\frac{\log M}{n}}$
[Nemirovski (2000), Juditsky and Nemirovski (2000)].

Sparsity oracle inequalities

Conclusion: all these nice properties are simultaneously satisfied for one and the same procedure, whenever it obeys a SOI.

Ultimate target:

- no assumptions on the dictionary f_1, \dots, f_M
- SOI with leading constant 1
- computational feasibility

Definition of the BIC

First idea: penalize least squares directly by $M(\lambda)$ (BIC criterion, Schwarz (1978), Foster and George (1994)).

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

where $\gamma > 0$ and

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

Definition of the BIC

First idea: penalize least squares directly by $M(\lambda)$ (BIC criterion, Schwarz (1978), Foster and George (1994)).

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

where $\gamma > 0$ and

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

Remarks:

- If the matrix $X = (f_j(X_i))_{i,j}$ has orthonormal 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

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,

$$\|f_{\hat{\lambda}^{BIC}} - f\|_n^2 \leq (1+\varepsilon) \inf_{\lambda \in \mathbb{R}^M} \left\{ \|f - f_\lambda\|_n^2 + C(\varepsilon) \frac{M(\lambda) \log M}{n} \right\}, \quad \forall \varepsilon > 0.$$

Sparsity oracle inequality for BIC

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,

$$\|f_{\hat{\lambda}^{BIC}} - f\|_n^2 \leq (1+\varepsilon) \inf_{\lambda \in \mathbb{R}^M} \left\{ \|f - f_\lambda\|_n^2 + C(\varepsilon) \frac{M(\lambda) \log M}{n} \right\}, \quad \forall \varepsilon > 0.$$

Remarks:

- the BIC is realizable only for small M (say, $M \leq 20$),
- the leading constant is **not** 1,
- $C(\varepsilon) \sim 1/\varepsilon$.

LASSO

Second popular idea: LASSO [Frank and Friedman (1993, Bridge regression), Tibshirani (1996), Chen and Donoho (1998, basis pursuit)]: instead of penalizing the residual sum of squares by $M(\lambda)$, as in the BIC, penalize by the ℓ_1 norm of λ :

$$\hat{\lambda}^L = \arg \min_{\lambda \in \mathbb{R}^M} \{ \|\mathbf{y} - f_\lambda\|_n^2 + 2r|\lambda|_1 \},$$

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

$$r = A \sqrt{\frac{\log M}{n}} \quad \text{for } A > 0 \text{ large enough.}$$

- 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 .

LASSO

- LASSO is computationally feasible, even for $M \gg n$. Convex optimization algorithms, such as LARS [Efron, Hastie, Johnstone, and Tibshirani (2004)].
- “Selection of variables” property: $\hat{\lambda}^L$ always has some components $\hat{\lambda}_j^L$ that are **exactly** equal to zero. For linear regression (\equiv Scenario (LinReg)) the selection is asymptotically correct: Bühlmann and Meinshausen (2006), Zhao and Yu (2006).

Restricted eigenvalue assumption

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

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

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

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

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

$$\Delta^T \Psi_M \Delta \geq \kappa |\Delta_J|_2^2$$

for all $J \subseteq \{1, \dots, M\}$ such that $|J| \leq s$ and $|\Delta_{J^c}|_1 \leq c_0 |\Delta_J|_1$.

More specific assumptions

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

- Coherence assumption (Donoho/Elad/Temlyakov),
- “Uniform uncertainty principle” (Candes/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., 2007]

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 $f_{\hat{\lambda}_L}$ 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: $\forall \lambda \in \mathbb{R}^M : M(\lambda) = s$,

$$\|f_{\hat{\lambda}_L} - f\|_n^2 \leq (1 + \varepsilon)\|f_\lambda - f\|_n^2 + C(\varepsilon) \left(\frac{M(\lambda) \log M}{\kappa n} \right).$$

Dantzig selector and LASSO for linear regression

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

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

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 and LASSO for linear regression

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

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

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 (Candes and Tao, 2005):

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

where $|\cdot|_{\infty}$ is the ℓ_{∞} norm in \mathbb{R}^M .

Theorem [Bickel, Ritov and T., 2007]

Let $\|f_j\|_n = 1, j = 1, \dots, M$. Let Assumption $RE(s, 3)$ hold and let $\hat{\lambda}$ 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

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

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

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

$$|\hat{\lambda} - \lambda^*|_\infty \leq \frac{C}{\kappa} \sqrt{\frac{\log M}{n}}$$

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

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

Disadvantages of the LASSO:

- SOI for the LASSO holds under very restrictive assumptions on the dictionary involving κ . 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 (2008).
- **Bad behavior when κ is small.**
- The leading constant in SOI is **not** 1.

Same problems with the Dantzig selector: the properties of Dantzig selector are essentially the same as for the LASSO, cf. Bickel, Ritov and T. (2007).

Sparse exponential weighting

Choose $\widehat{\lambda}^{EW}$ according to:

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

where the probability measure S_n is given by

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

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

- Bayesian estimator if $\beta = 2\sigma^2$, but we need a larger β .
- Non-discrete π : is the fast computation possible?

A PAC-Bayesian bound

Lemma [Dalalyan and T., 2007]

The estimator with exponential weights $f_{\hat{\lambda}^{EW}}$ defined with $\beta \geq 4\sigma^2$ and any prior π satisfies:

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

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 :

$$q(\lambda) = \prod_{j=1}^M \tau^{-1} q_0(\lambda_j/\tau), \quad \forall \lambda \in \mathbb{R}^M,$$

where q_0 is the Student t_3 density,

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

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

SOI for the SEW estimator

Theorem [Dalalyan and T., 2007]

Let $\max_{1 \leq j \leq M} \|f_j\|_n \leq c_0 < \infty$. Then the SEW estimator $f_{\hat{\lambda}^{EW}}$ defined with $\beta \geq 4\sigma^2$ and with the **sparsity prior** π satisfies:

$$\mathbb{E} \|f_{\hat{\lambda}^{EW}} - f\|_n^2 \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \|f_\lambda - f\|_n^2 + \frac{CM(\lambda)}{n} \log \left(1 + \frac{|\lambda|_1 \sqrt{Mn}}{M(\lambda)} \right) \right\}$$

where $|\lambda|_1$ is the ℓ_1 -norm of λ .

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

SEW estimator: discussion

- *SEW is not a penalized estimator.*

$$\hat{\lambda}_j^{EW} = \int_{\mathbb{R}^M} \lambda_j S_n(d\lambda) = \int_{\mathbb{R}^M} \lambda_j g_n(\lambda) d\lambda, \quad j = 1, \dots, M,$$

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

$$g_n(\lambda) \propto \exp \left\{ -n \|\mathbf{y} - f_\lambda\|_n^2 / \beta - C \sum_{j=1}^M \log(1 + \lambda_j^2 / \tau) \right\}$$

Maximizer of this density (the MAP estimator):

$$\hat{\lambda}^{MAP} = \arg \min_{\lambda \in \mathbb{R}^M} \left\{ \|\mathbf{y} - f_\lambda\|_n^2 + \frac{\gamma}{n} \sum_{j=1}^M \log(1 + \lambda_j^2 / \tau) \right\} \neq \hat{\lambda}^{EW}.$$

SEW estimator: discussion

- *Precursors of SEW for the “diagonal” sequence model.*

Rivoirard (2004): minimax Bayes priors with heavy tails,
Johnstone and Silverman (2005): “quasi-Cauchy” prior.

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 $Z \in \mathcal{Z}$.
- $Q(Z, f_\lambda)$ a given real-valued loss (prediction loss).

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

$$S_n(d\lambda) = \frac{\exp \left\{ - \sum_{i=1}^n Q(Z_i, f_\lambda) / \beta \right\} \pi(d\lambda)}{\int_{\mathbb{R}^M} \exp \left\{ - \sum_{i=1}^n Q(Z_i, f_w) / \beta \right\} \pi(dw)}$$

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

$$n \|\mathbf{y} - f_\lambda\|_n^2 \rightsquigarrow \sum_{i=1}^n Q(Z_i, f_\lambda).$$

Mirror averaging

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

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

cf. Juditsky/Rigollet/T (2005) [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 $\Lambda \subset \mathbb{R}^M$ (e.g., on an ℓ_1 ball).

Assumption JRT (2005).

The mapping $\lambda \mapsto Q(Z, f_\lambda)$ is convex for all Z and there exists $\beta > 0$ such that the function

$$\lambda \mapsto \mathbb{E} \exp \left(\frac{Q(Z, f_{\lambda'}) - Q(Z, f_\lambda)}{\beta} \right)$$

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

Roughly: “strong convexity on the average”.

PAC-Bayesian bound for mirror averaging

Define the average risk: $A(\lambda) = \mathbb{E}Q(Z, f_\lambda)$.

Lemma (PAC-Bayesian bound).

Let $f_{\hat{\lambda}^{MA}}$ be a mirror averaging estimator defined with β satisfying Assumption JRT and any prior π supported on a convex compact set Λ . Then

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

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

Proof follows the scheme of Juditsky, Rigollet and T. (2005), cf. Rigollet and Zhao (2006), Audibert (2006), Lounici (2007).

SOI for Mirror Averaging

Theorem [Dalalyan, Rigollet and T., 2007]

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

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

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

Comparison with SOI for the LASSO

The LASSO type estimators

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

van de Geer (2007), Koltchinskii (2007):

$$\mathbb{E} A(\hat{\lambda}) \leq \inf_{|\lambda|_1 \leq R} \left(\boxed{3} A(\lambda) + \frac{CR^2 M(\lambda) \log M}{\boxed{\kappa(\lambda)} n} \right)$$

where $\kappa(\lambda)$ is a quantity analogous to κ in Assumption RE (Restricted Eigenvalue). To get the correct rate, we need to consider only λ such that $\kappa(\lambda) \geq c$, which is equivalent to RE.

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

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

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

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

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

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

- Assumption on the dictionary: $\|f_j\|_\infty \leq L, \quad j = 1, \dots, M.$

Example: Gaussian regression, squared loss

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

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

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

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

SOI for regression with squared loss

Corollary

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

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

Here Λ^M is the simplex:

$$\Lambda^M = \left\{ \lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1 \right\}.$$

Example: density estimation with L_2 loss

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

Example: density estimation with L_2 loss

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

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

- The associated risk:

$$A(\lambda) = \mathbb{E} Q(X, f_\lambda) = \|f - f_\lambda\|^2 - \|f\|^2.$$

SOI for density estimation with L_2 loss

Corollary

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

$$\mathbb{E} \|f_{\hat{\lambda}^{MA}} - f\|^2 \leq \inf_{\lambda \in \Lambda^M} \left\{ \|f_{\lambda} - f\|^2 + \frac{CM(\lambda)}{n} \log \left(\frac{C' \sqrt{M(n \vee M)}}{M(\lambda)} \right) \right\}.$$

Here Λ^M is the simplex:

$$\Lambda^M = \left\{ \lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1 \right\}.$$

Computation of SEW estimators

Consider the linear regression scenario:

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

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

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

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 η_1, η_2, \dots be i.i.d. standard normal random vectors. Set

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

Then

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

Simulations

Example 1: selection properties when the Gram matrix is nice

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

$$\lambda_j = \mathbf{1}\{j \leq S\} \quad \text{and} \quad \sigma^2 = \frac{S}{9n}.$$

We apply the SEW estimator using Langevin Monte-Carlo with

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

Simulations

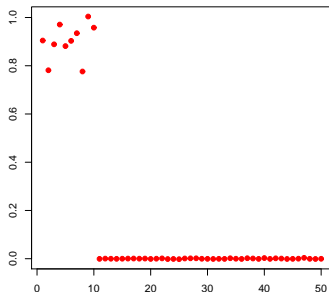


Figure: Typical result for Example 1 with $n = 200$, $M = 500$, $S = 10$, $h = 10^{-4}$, $T = 5$. The estimates of first 50 coefficients are plotted. In this example, we have $\frac{1}{n} \|X(\hat{\lambda} - \lambda)\|^2 = 0.0021$. The time of computation of the estimator was about 30 seconds.

Simulations

Example 2: Comparison with the LASSO/LARS

Choose X_1, \dots, X_n i.i.d. uniformly distributed in $[0, 1]^2$ and set $f_j(t) = \mathbf{1}\{[0, j_1/k] \times [0, j_2/k]\}(t)$, $j = (j_1, j_2) \in \{1, \dots, k\}^2$, $t \in [0, 1]^2$.

We get a matrix $X = (f_j(X_i))_{i,j}$ with k^2 columns some of which are nearly collinear. The number of covariates is $M = k^2$. Set $\sigma = 1$, $k = 15$, $n = 100$, $\lambda_j^* = 0$ for $j \in \{1, \dots, k\} \setminus \{87, 110, 200\}$, $\lambda_j^* = 1$ for $j \in \{87, 110\}$ and $\lambda_{200}^* = -2$. Applying the SEW estimator with Langevin Monte-Carlo and

$$\tau = \frac{4\sigma}{\sqrt{\sum_{j,i} f_j^2(X_i)}}, \quad \beta = 4\sigma^2, \quad h = 0.0005.$$

Simulations

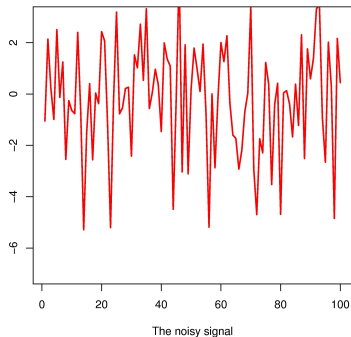
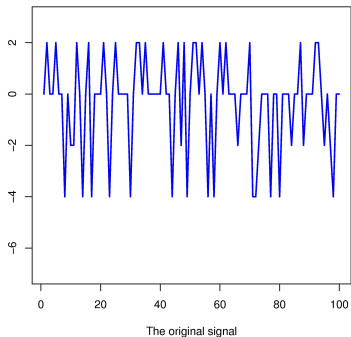
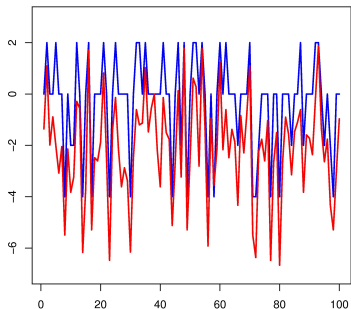
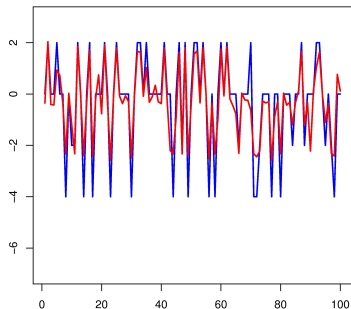


Figure: Example 2 with $n = 100$, $M = 225$, $M(\lambda^*) = 3$, $h = 5 \cdot 10^{-4}$, $T = 2$.

Sim



The original signal and the LASSO-Cp



The original signal and the EW-aggregate

Figure: Typical result for Example 2 with $n = 100$, $M = 225$, $M(\lambda^*) = 3$, $h = 5 \cdot 10^{-4}$, $T = 2$. In this example, we have $\frac{1}{n} \|X(\hat{\lambda} - \lambda^*)\|^2 = 0.28$ for our estimator and $\frac{1}{n} \|X(\hat{\lambda} - \lambda^*)\|^2 = 1.72$ for the LASSO. The time of computation of the SEW estimator was about 5 seconds.

BICKEL, P.J., RITOV, Y. and TSYBAKOV, A.B. (2007) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, to appear.

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. *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. Learning by mirror averaging. *Annals of Statistics*, to appear.