Sparse Exponential Weighting and Langevin Monte-Carlo

Alexandre Tsybakov,
joint work with Arnak Dalalyan

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

Cumberland Lodge, April 1, 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 $\mathrm{f}_\lambda(x)$ (known as a function of λ and x) such that

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

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

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

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

We believe that

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

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

Scenarios for linear approximation

(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 for linear approximation

(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 $M \le n$, there exists λ^* such that $f - f_{\lambda^*}$ is small: **nonparametric estimation of regression**;

Scenarios for linear approximation

(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 $M \le n$, 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, additive models etc.

Example: nonlinear approximation

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

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

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

Thus, we believe that

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

for some $\lambda = (\lambda_1, \dots, \lambda_M)$, possibly with $M \gg n$.

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:

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

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

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

Sparsity and dimension reduction

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

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

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

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

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

Sparsity oracle inequalities

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

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

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

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

Sparsity oracle inequalities (SOI)

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

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

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

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

"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 $M(\lambda)$ (BIC criterion, Schwarz (1978), Foster and George (1994)):

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

where $\gamma > 0$ and

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

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,

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

Remarks:

- the BIC is realizable only for small M (say, $M \le 20$),
- 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 $M(\lambda)$, as in the BIC, but by the ℓ_1-norm of λ:

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

where $|\lambda|_1 = \sum_{j=1}^M |\lambda_j|$, $r > 0$ 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.

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

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 (2008).
- 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).

Exponential weighting

Estimate $f(x)$ by

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

where the probability measure S_n is given by

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

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

Exponential weighting

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

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

- 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:*

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

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:

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

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_λ *be linear in* λ. *Then for* $\beta \ge 4\sigma^2$ *the estimator* $\mathrm{f}_{\widehat{\lambda}^{EW}}$ *with the* **sparsity prior** *satisfies:*

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

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 $M \sim 10^3$.

SOI for the SEW estimator: generalized linear models

Assume now:

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

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

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

SEW estimator: discussion

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

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

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

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

Maximizer of this density (the MAP estimator):

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

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

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

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

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

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

Mirror averaging

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

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

  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".

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

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

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

- Assumption on the dictionary $\|f_j\|_\infty \leq L, \ j = 1, \dots, M.$
- The loss function
  $Q(Z, \mathrm{f}_\lambda) = \big(Y - \mathrm{f}_\lambda(X)\big)^2$ where $\mathrm{f}_\lambda = \sum_{j=1}^M \lambda_j f_j.$
- Then $A(\lambda) = \mathbb{E}\, Q(Z, \mathrm{f}_\lambda) = \|\mathrm{f}_\lambda - f\|_X^2 + \sigma^2$, $\|f\|_X^2 \triangleq \int f^2 dP_X.$
- $\beta \geq 2\sigma^2 + 8L^2.$

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:

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

- The associated risk:

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

- $\beta \ge 12L$.

PAC-Bayesian bound for mirror averaging

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

Lemma (PAC-Bayesian bound).

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

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

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

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

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

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

Comparison with SOI for the LASSO

The LASSO type estimators

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

van de Geer (2007,2008), Koltchinskii (2007,2008):

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

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

Modified SEW estimators

Take the modified sparsity prior

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

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:

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

X is a $n \times M$ deterministic design matrix, $\lambda \in \mathbb{R}^M$ is an unknown vector and $\xi \in \mathbb{R}^M$ 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 $\eta_1, \eta_2, \ldots$ 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

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

Simulations

Example 1: Compressive sensing

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

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

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

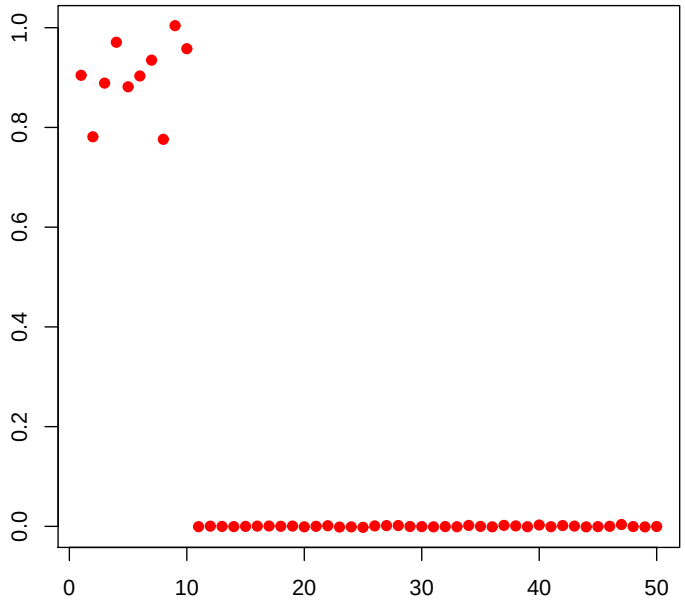
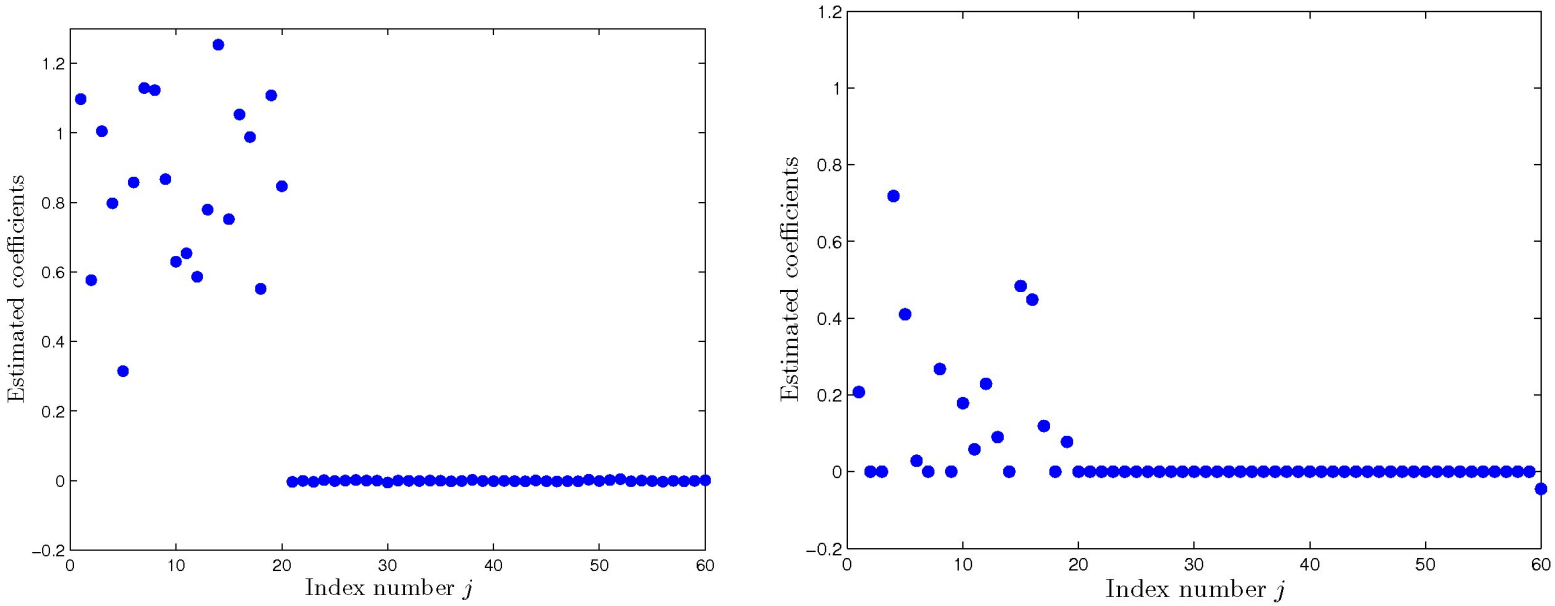


Figure: Typical result for Example 1 with $n = 200$, $M = 500$, $S = 10$, $h = 10^{-4}$, $T = 5$. The estimates of first 50 coefficients are plotted. In this example, we have $\frac{1}{n}\|X(\hat{\lambda} - \lambda)\|^2 = 0.0021$. The time of computation of the estimator was about 30 seconds.

Example 1: Compressive sensing

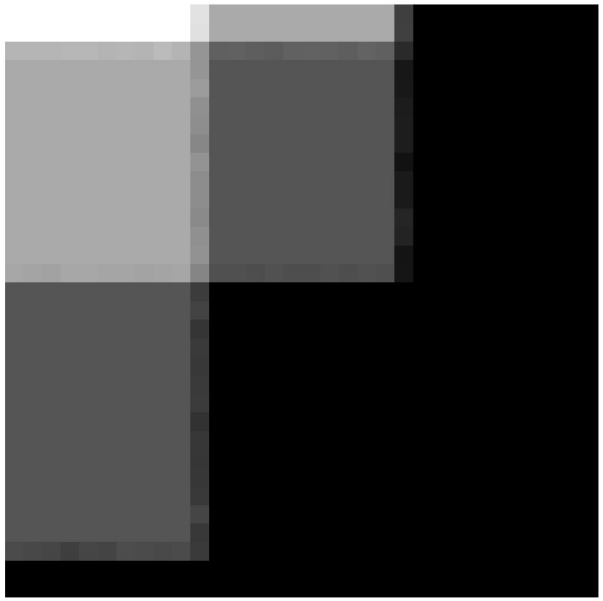


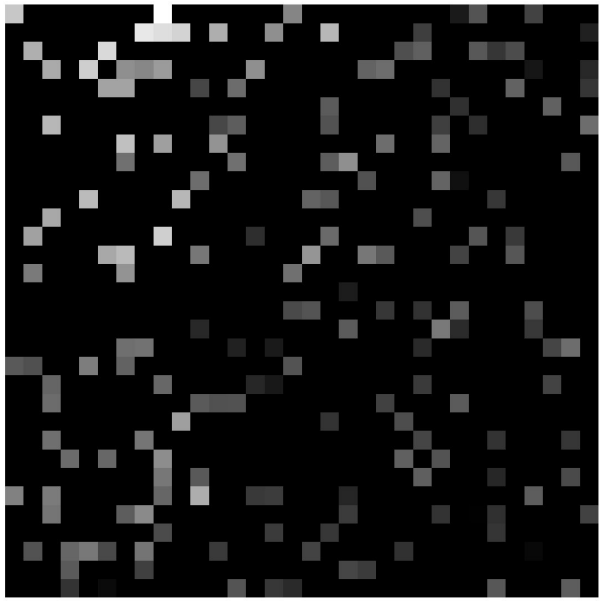
Typical outcome for $n = 200$, $M = 500$ and $S = 20$.

Image denoising: A simple example

- 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: $Y_i = (\phi_1(U_i), \dots, \phi_{k^2}(U_i))^T \lambda + \xi_i$ where $\lambda = [\mathbf{1}\{(j \in \{10, 100, 200\})]'$.
- Tuning parameters: the same rule as previously.

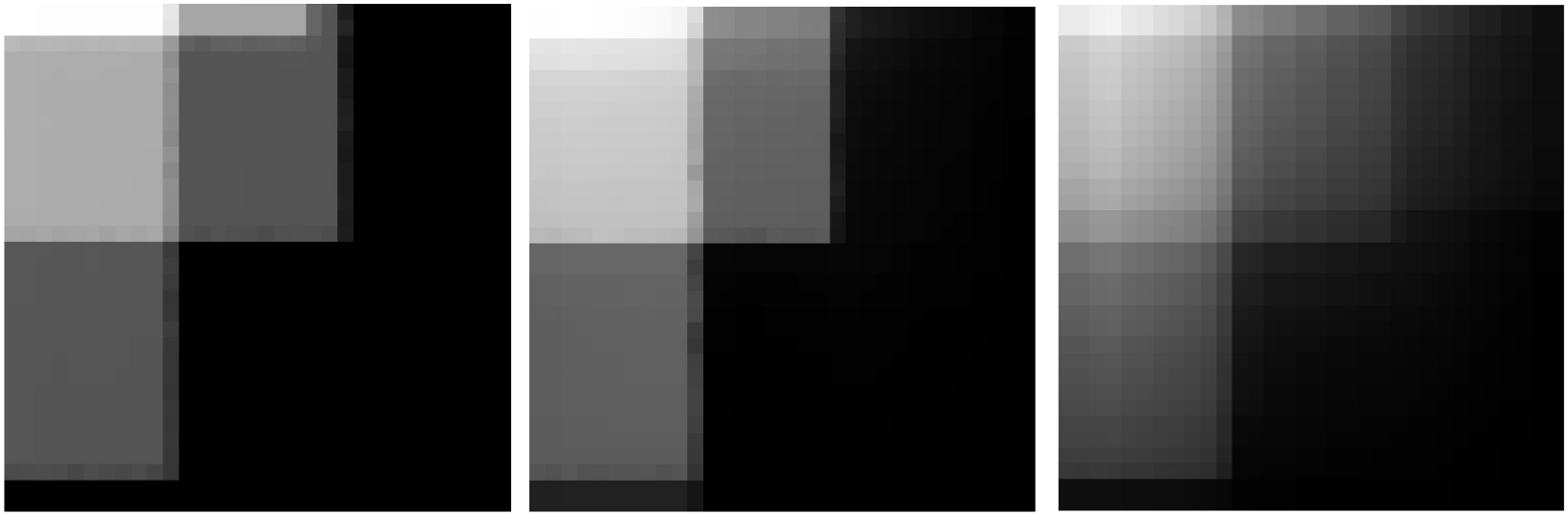
Image denoising





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

σ	$n = 100$			$n = 200$		
	EWA	Lasso	Ideal LG	EWA	Lasso	Ideal LG
2	0.210 (0.072) $T = 1$	0.759 (0.562)	0.330 (0.145)	0.187 (0.048) $T = 1$	0.661 (0.503)	0.203 (0.086)
4	0.420 (0.222) $T = 1$	2.323 (1.257)	0.938 (0.631)	0.278 (0.132) $T = 1$	2.230 (1.137)	0.571 (0.324)

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.