SPARSE RECOVERY BY AGGREGATION AND LANGEVIN MONTE-CARLO

Joint work with A. Tsybakov

Arnak DALALYAN
CERTIS, Ecole des Ponts ParisTech

Outline

Part I: Introduction

- Motivation and basic concepts
- Penalized LSE and aggregation
- Bayesian interpretation

Part II: Risk bounds

- General oracle inequality
- Sparsity oracle inequality
- Discussion

Part III: Langevin Monte-Carlo

- Diffusions and stationarity
- Implementation
- Numerical experiments

Part I: Introduction

High-dimensional data

Technological innovations allow us to collect massive amount of data with low cost:

- microarray data or magnetic resonance images in biomedical studies,
- high resolution satellite imagery used in natural resource discovery and agriculture,
- financial data (option prices, bond yields, etc) used in financial engineering and risk management.

Statistical methods are vital for analyzing these data :

- finding sparse representations,
- performing variable selection,
- prediction and model estimation.

Curse of dimensionality and sparsity

High-dimensionality has significantly challenged traditional statistical theory.

- In linear regression, the accuracy of the LSE is of order M/n, where M is the number of covariates.
- Applied statisticians are often interested in the case where ***M* is much larger than *n***.

Sparsity assumption provides compelling theoretical framework for dealing with high dimension.

- Even if the number of parameters describing the model in general setup is large, only few of them contribute to the process of data generation.
- No a priori information on the set of relevant parameters is available. It is only known that the cardinality of the set of relevant parameters is small.

Basic concepts

- Data: $\mathcal{D}_n = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\} \subset \mathcal{Z} \times \mathbf{R}$.
- Model: $\{Z_i\}$ are deterministic and for some function $f \in \mathcal{F}_0$,
$$\xi_i = Y_i - f(Z_i), \quad i = 1, \ldots, n$$
are iid with zero mean and finite variance σ^2.
- Loss function: for a set $\mathcal{F}$ and for every $g \in \mathcal{F}$,
$$\ell(f, g) = \frac{1}{n} \sum_{i=1}^{n} [g(Z_i) - f(Z_i)]^2 := \|g - f\|_n^2,$$
is the loss we suffer when we use a procedure g.
- Unbiased estimate: for every fixed g,
$$\mathrm{R}[\mathcal{D}_n, g] = \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(Z_i)]^2 - \sigma^2$$
is an unbiased estimator of the loss $\ell(f, g)$.

Penalized RSS minimization

- **Extended parametric setup:** we are given a vast but parametric family $\mathcal{F}_\Lambda = \{f_\lambda : \lambda \in \Lambda\}$, with $\Lambda \subset \mathbf{R}^M$ such that $d_{\|\cdot\|_n}(f, \mathcal{F}_\Lambda)$ is small.
- **Penalized RSS estimator:** $\hat{f}^{\text{PLSE}} = f_{\hat{\lambda}^{\text{PLSE}}}$ with

$$\hat{\lambda}^{\text{PLSE}} = \arg\min_{\lambda\in\Lambda} \Big(\underbrace{\mathrm{R}[\mathcal{D}_n, f_\lambda]}_{\text{data fidelity term}} + \underbrace{\mathrm{Pen}(\lambda)}_{\text{a priori penalization}} \Big).$$

- **Common penalties:**
  - $\mathrm{Pen}(\lambda) = \kappa\|\lambda\|_0$ – BIC penalty,
  - $\mathrm{Pen}(\lambda) = \kappa\|\lambda\|_2^2$ – ridge penalty,
  - $\mathrm{Pen}(\lambda) = \kappa\|\lambda\|_1$ – Lasso penalty.
  - SCAD, Elastic Net, etc.

Exponentially weighted aggregate (EWA)

- Main idea: extend the search space and change the penalty.
- Search space: $\mathcal{P} = \{p: \text{ prob. s.t. } \int_\Lambda \|f_\lambda\|_n^2\, p(d\lambda) < \infty\}$.
- Rewriting PLSE: $\hat f^{\text{PLSE}} = \int_\Lambda f_\lambda\, \hat\pi^{\text{PLSE}}(d\lambda)$ with

$$\hat\pi^{\text{PLSE}} = \arg\min_{p\in\mathcal{P}} \Big\{ \int_\Lambda \text{R}[\mathcal{D}_n, f_\lambda]\, p(d\lambda) + \int_\Lambda \text{Pen}(\lambda)\, p(d\lambda)\Big\}.$$

- KL-penalization: Let $\boldsymbol{\pi} \in \mathcal{P}$ be a prior on Λ. Define the EWA as $\hat f_n^{\text{EWA}} = \int_\Lambda f_\lambda \hat\pi_n(d\lambda)$ where

$$\hat\pi_n = \arg\min_{p\in\mathcal{P}} \Big\{ \int_\Lambda \text{R}[\mathcal{D}_n, f_\lambda]\, p(d\lambda) + \kappa\mathcal{K}(p, \boldsymbol{\pi})\Big\}.$$

- Explicit form: $\boxed{\hat\pi_n(d\lambda) \propto \exp\{-\kappa^{-1}\text{R}[\mathcal{D}_n, f_\lambda]\}\pi(d\lambda)}$.

Bayesian interpretation

- Terminology: $\hat{\pi}_n$ – posterior, κ – temperature.
- Bayesian posterior mean: If we consider the parametric model

$$Y_i = f_\lambda(Z_i) + \tilde{\xi}_i, \quad i = 1, \ldots, n,$$

with $\tilde{\xi}_i \overset{iid}{\sim} \mathcal{N}(0, n\kappa/2)$ and prior π on the parameter set Λ, then $\hat{\pi}_n$ is the posterior probability and $\hat{f}_n^{EWA}$ is the posterior mean:

$$\hat{f}_n^{EWA}(Z_i) = \mathbf{E}_\pi[f_\lambda(Z_i)|\mathcal{D}_n], \quad i = 1, \ldots, n.$$

- Notation: In what follows, we take

$$\kappa = \beta/n.$$

Part II: Risk bounds

General oracle inequality: assumptions

Assumption N

For any $\gamma > 0$ small enough, $\exists$ probability space and 2 r. v. ξ and ζ defined on it such that

i) ξ has the same distribution as the errors ξ_i,

ii) $\xi + \zeta \overset{\mathscr{D}}{=} (1+\gamma)\xi$ and $\mathbf{E}[\zeta|\xi] = 0$,

iii) $\exists$ bounded Borel function $v : \mathbf{R} \to \mathbf{R}_+$ such that,

$$\mathbf{E}[e^{t\zeta}|\xi = a] \approx e^{t^2\gamma v(a)}, \quad \gamma \to 0$$

for every a and $\forall t \in [-t_0, t_0]$.

Assumption L

The set Λ satisfies

$$(\lambda, \lambda') \in \Lambda^2 \quad \Longrightarrow \quad \max_i |f_\lambda(Z_i) - f_{\lambda'}(Z_i)| \le L$$

for some $L \in [0, \infty]$.

General oracle inequality

Theorem (PAC-Bayesian bound)

Let Assumptions N and L be satisfied. Then for any prior π and for any $\beta \geq max(4\|v\|_\infty, 2L/t_0)$ we have

$$\mathbf{E}_f[\ell(\hat{f}_n^{EWA}, f)] \leq \inf_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \ell(f_\lambda, f)\, p(d\lambda) + \frac{\beta \mathcal{K}(p,\pi)}{n} \right), \quad (1)$$

where $\mathcal{K}(p,\pi)$ stands for the Kullback-Leibler divergence

$$\mathcal{K}(p,\pi) = \begin{cases} \int_\Lambda \log\left(\frac{dp}{d\pi}(\lambda)\right) p(d\lambda), & \text{if } p \ll \pi, \\ +\infty, & \text{otherwise} \end{cases}.$$

Comments

- If the cardinality of Λ is finite, say $\Lambda = \{1, \ldots, N\}$, and π is uniform, then inequality (1) implies that

$$\mathbf{E}_f[\ell(\hat{f}_n^{\text{EWA}}, f)] \leq \min_{j=1,\ldots,N} \ell(f_j, f) + \frac{\beta \log N}{n}.$$

  This type of inequalities are usually called oracle inequalities.

- If the noise is Gaussian, Rademacher, Uniform or a countable convolution of these distributions, then one can take $L = +\infty$ and (1) holds for every $\beta \geq 4\mathbf{E}[\xi_1^2]$.

- For regression with Gaussian noise and finite set Λ, bounds similar to (1) have been established in an earlier work by Leung and Barron (2006).

Linear family and sparsity prior

- Aim: by a proper choice of the prior, to adapt the EWA to the setting of sparse estimation.
- Linear family: Assume that $\|\phi_j\|_n = 1$ and

$$\mathcal{F}_\Lambda = \Big\{ \sum_{j=1}^{M} \lambda_j \phi_j : \ \boldsymbol{\lambda} \in \mathbf{R}^M \Big\}.$$

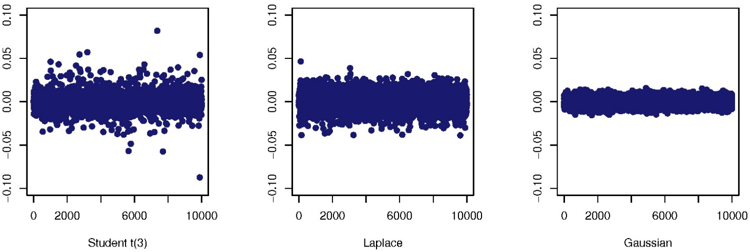
- Huber function: Define

$$\omega(t) = \begin{cases} t^2, & \text{if } |t| \leq 1 \\ 2|t| - 1, & \text{otherwise} \end{cases}.$$

- Sparsity prior: Let τ, α and R be > 0, we define the prior

$$\pi(d\boldsymbol{\lambda}) \propto \Big\{ \prod_{j=1}^{M} \frac{e^{-\omega(\alpha\lambda_j)}}{(\tau^2 + \lambda_j^2)^2} \Big\} d\boldsymbol{\lambda}.$$

Does π favor the sparsity ?



The scatter plots of a sample of size 10000 drawn from scaled $t(3)$-distribution (left panel), Laplace distribution (central panel) and Gaussian distribution (right panel). In all three cases the location parameter is equal to zero and the scale parameter is set to 10^{-2}.

Sparsity Oracle inequality

Theorem

If Assumption N holds with $t_0 = +\infty$, then for every $\beta \geq 4\|v\|_\infty$ the EWA based on the sparsity prior satisfies

$$\mathbf{E}_f[\ell(\hat{f}_n^{EWA}, f)] \leq \ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{4\beta}{n}\Big\{\alpha\|\boldsymbol{\lambda}^*\|_1 + \sum_{j=1}^{M} \log\Big(1 + \Big|\frac{\lambda_j^*}{\tau}\Big|\Big)\Big\}$$

$$+\mathrm{R}(M,\tau,\alpha), \qquad \forall \boldsymbol{\lambda}^* \in \mathbf{R}^M,$$

where $\mathrm{R}(M,\tau,\alpha) = 12\tau^2 M + \frac{2\beta}{n}$.

Remarks

- **Corollary**: If there is a sparse $\boldsymbol{\lambda}^* \in \mathbf{R}^M$ such that $f_{\boldsymbol{\lambda}^*}$ is close to f, then by choosing $\alpha \sim 1$ and $\tau^2 \sim (Mn)^{-1}$, we get

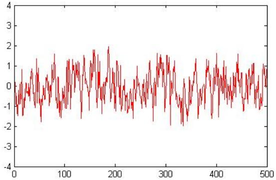
$$\mathbf{E}_f[\ell(\hat{f}_n^{\mathrm{EWA}}, f)] \leq \ell(f_{\boldsymbol{\lambda}^*}, f) + \frac{\mathrm{C} \cdot \|\boldsymbol{\lambda}^*\|_0 \log(Mn)}{n}.$$

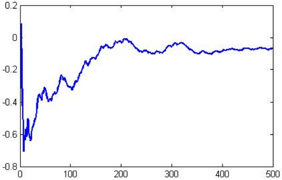
- **Optimality**: the last inequality is an oracle inequality with leading constant equal to one and a remainder term which is “optimal”.
- **Important**: this result is obtained under no assumption on the dictionary $\{\phi_j\}$!

Part III: Langevin Monte-Carlo

Motivation

- Although the EWA can be written in an explicit form, its computation is not trivial because of the M-fold integral.
- Naive Monte-Carlo methods fail in moderately large dimensions ($M = 50$).
- A specific type of Markov Chain Monte-Carlo technique, called Langevin Monte-Carlo, turns out to be very efficient.
- A path of a 1D diffusion process and its averaged version:





The set-up

- We have $Y_i = \boldsymbol{X}_i^T \boldsymbol{\lambda}^* + \xi_i$, $i = 1, \ldots, n$, where ξ_i are i.i.d. and $\boldsymbol{\lambda}^* \in \mathbf{R}^M$ is the parameter of interest.
- We wish to compute the EWA, which can be written as

$$\hat{\boldsymbol{\lambda}}_n = \hat{\boldsymbol{\lambda}}_n^{\text{EWA}} = C \int \boldsymbol{\lambda} e^{-\beta^{-1}\|Y - X\boldsymbol{\lambda}\|_2^2} \pi(d\boldsymbol{\lambda}),$$

  where C is the constant of normalization.
- We can rewrite $\hat{\boldsymbol{\lambda}}_n = \int_{\mathbf{R}^M} \boldsymbol{\lambda} p_V(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$, where $p_V(\boldsymbol{\lambda}) \propto e^{V(\boldsymbol{\lambda})}$ is a density function and

$$V(\boldsymbol{\lambda}) = -\frac{\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\lambda}\|_2^2}{\beta} - \sum_{j=1}^{M} \Big\{ 2\log(\tau^2 + \lambda_j^2) + \omega(\alpha\lambda_j) \Big\},$$

  with $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

Multidimensional Langevin Diffusions

- Let $\boldsymbol{L}_0 \in \mathbf{R}^M$ and $\boldsymbol{W}$ be an M-dimensional BM. For any $V \in C^2(\mathbf{R}^M; \mathbf{R})$ we call the solution to the SDE
$$d\boldsymbol{L}_t = \nabla V(\boldsymbol{L}_t)\,dt + \sqrt{2}\,d\boldsymbol{W}_t,$$
the Langevin diffusion with potential V.
- **Drift condition**: There is a $D \in C^2(\mathbf{R}^M; [1,\infty))$ and $a, b, r > 0$ such that, for every $\boldsymbol{\lambda} \in \mathbf{R}^M$,
$$\nabla V(\boldsymbol{\lambda})^T \nabla D(\boldsymbol{\lambda}) + \Delta D(\boldsymbol{\lambda}) \leq -aD(\boldsymbol{\lambda}) + b\mathbb{1}(\|\boldsymbol{\lambda}\|_2 \leq r).$$
- If $\sup_{\boldsymbol{\lambda}} V(\boldsymbol{\lambda}) < \infty$ and the drift condition is fulfilled, then $\boldsymbol{L}$ is D-geometrically ergodic: $\exists R, \rho > 0$ s.t.
$$\sup_{\|h/D\|_\infty \leq 1} \left| \mathbf{E}[h(\boldsymbol{L}_t)] - \int_{\mathbf{R}^M} h(\boldsymbol{\lambda})\, p_V(d\boldsymbol{\lambda}) \right| \leq R\, D(\boldsymbol{L}_0) e^{-\rho t}.$$
with $\boxed{p_V(\boldsymbol{\lambda}) \propto e^{-V(\boldsymbol{\lambda})}}$.

Langevin Diffusion in our setup

- In our case

$$V(\boldsymbol{\lambda}) = -\frac{\|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\lambda}\|_2^2}{\beta} - \sum_{j=1}^{M}\Big\{2\log(\tau^2+\lambda_j^2)+\omega(\alpha\lambda_j)\Big\},$$

satisfies the drift condition with $D(\boldsymbol{\lambda}) = e^{\|\boldsymbol{\lambda}\|_2}$ <u>if $\alpha > 0$</u>. Thus the Langevin diffusion with potential V is geometrically ergodic and mixing.

- Therefore,

$$\bar{\boldsymbol{L}}_T = \frac{1}{T}\int_0^T \boldsymbol{L}_t\,dt \xrightarrow[T\to\infty]{L^2} \int_{\mathbf{R}^M} \boldsymbol{\lambda} p_V(\boldsymbol{\lambda})\,d\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_n.$$

This convergence takes place with the rate $1/\sqrt{T}$.

- Since the mean value $\bar{\boldsymbol{L}}_T$ is impossible to compute exactly, we replace it by Riemann sums and approximate $\boldsymbol{L}$ by its Euler discretization.

Implementation

- Fix a step of discretization $h > 0$ and set

$$L^E_{k+1} = L^E_k + h\nabla V(L^E_k) + \sqrt{2h}\,W_k, \qquad L^E_0 = 0,$$

where $k = 0, 1, \ldots, [T/h] - 1$, W_1, $W_2, \ldots$ are i.i.d. standard Gaussian random vectors in $\mathbf{R}^M$ and $[x]$ stands for the integer part of $x \in \mathbf{R}$.

- Approximate $\bar{L}_T$ by

$$\bar{L}^E_{T,h} = \frac{1}{[T/h]} \sum_{k=0}^{[T/h]-1} L^E_k.$$

- When $h \to 0$, $\bar{L}^E_{T,h}$ tends to $\bar{L}_T$.

Variants of Langevin Monte-Carlo

- Use non-constant step Euler scheme with a step depending on $\nabla^2 V$.

Variants of Langevin Monte-Carlo

- Use non-constant step Euler scheme with a step depending on $\nabla^2 V$.
- Use Ozaki discretization: more accurate but time consuming.

Variants of Langevin Monte-Carlo

- Use non-constant step Euler scheme with a step depending on $\nabla^2 V$.
- Use Ozaki discretization: more accurate but time consuming.
- Use tempered Langevin diffusions (having non-constant diffusion coefficient).

Variants of Langevin Monte-Carlo

- Use non-constant step Euler scheme with a step depending on $\nabla^2 V$.
- Use Ozaki discretization: more accurate but time consuming.
- Use tempered Langevin diffusions (having non-constant diffusion coefficient).
- Apply a Metropolis-Hastings correction.

Variants of Langevin Monte-Carlo

- Use non-constant step Euler scheme with a step depending on $\nabla^2 V$.
- Use Ozaki discretization: more accurate but time consuming.
- Use tempered Langevin diffusions (having non-constant diffusion coefficient).
- Apply a Metropolis-Hastings correction.

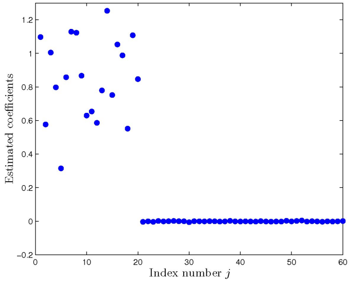
It seems that the simplest LMC is the best !

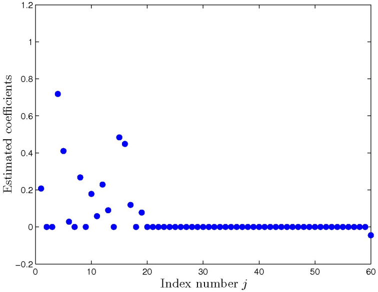
Numerical experiments
Example 1: Compressive sensing

- Input: n, M and S, all positive integers.
- Covariates: we generate an $n \times M$ matrix $\mathbb{X}$ with iid Rademacher entries.
- Errors: we generate a standard Gaussian vector $\boldsymbol{\xi}$.
- Noise magnitude: $\sigma = \sqrt{S/9}$.
- Response: $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\lambda}^* + \boldsymbol{\xi}$ where $\boldsymbol{\lambda}^* = [\mathbb{1}(j \le S); j \le M]$.
- Tuning parameters:

$$\beta = 4\sigma^2, \quad \tau = 4\sigma/\|\mathbb{X}\|_2, \quad h = 4\sigma^2/\|\mathbb{X}\|_2^2.$$

Numerical experiments
Example 1: Compressive sensing





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

Numerical experiments

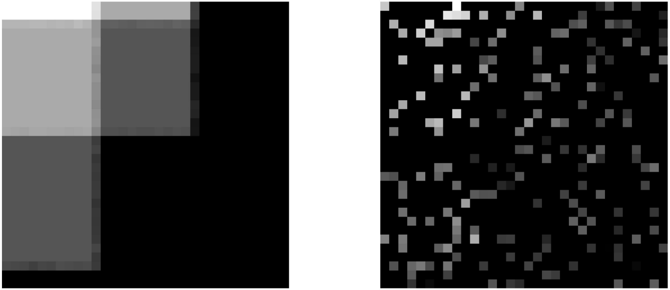
Example 1: Compressive sensing

	$M = 200$		$M = 500$	
	EWA	Lasso	EWA	Lasso
$n = 100$ $S = 5$	0.064 (0.043) $T = 1$	1.442 (0.461)	0.087 (0.054) $T = 1$	1.616 (0.491)
$n = 100$ $S = 10$	1.153 (1.091) $T = 2$	5.712 (1.157)	1.891 (1.522) $T = 5$	6.508 (1.196)
$n = 100$ $S = 15$	6.839 (1.896) $T = 5$	11.149 (1.303)	8.917 (2.186) $T = 10$	11.82 (1.256)

Image denoising
A simple example

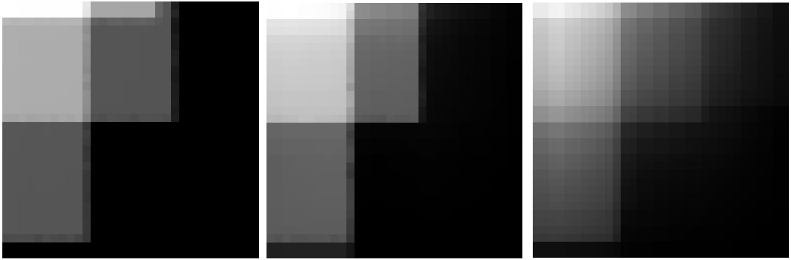
- Input: n, k positive integers and $\sigma > 0$.
- We generate n vectors U_i of $\mathbf{R}^2$ uniformly distributed in $[0, 1]^2$.
- Covariates $\phi_j(u) = \mathbb{1}_{[0,j_1/k]\times[0,j_2/k]}(u)$.
- Errors: we generate a centered Gaussian vector $\boldsymbol{\xi}$ with covariance matrix $\sigma^2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

Dalalyan, A.S.

Image denoising

σ	$n = 100$ EWA	$n = 100$ Lasso	$n = 100$ Ideal LG	$n = 200$ EWA	$n = 200$ Lasso	$n = 200$ 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)

Open questions for future work

- Is it possible to perform model selection with the EWA ?
- What is the rate of ergodicity when $\alpha = 0$?
- How to rigorously justify the choice of T and h?
- ...