Habilitation à diriger des recherches

de l'Université Paris-Est

Agrégation PAC-Bayésienne et bandits à plusieurs bras

Jean-Yves Audibert

Imagine (ENPC/CSTB) - LIGM - Université Paris Est,
&
Willow - CNRS/ENS/INRIA

Statistical learning

- Training data = n input-output pairs :

$$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$$

- A new input X comes
- General goal: predict the corresponding output Y
- Probabilistic assumption :

$$Z = (X, Y), Z_1, \ldots, Z_n \qquad \text{i.i.d.}$$

from some unknown distribution P

Some typical examples

- Computer Vision
  - object recognition
    $X =$ an image
    $Y = +1$ if the image contains the object, $Y = 0$ otherwise
- Textual document
  - $X =$ a mail $\quad Y =$ spam vs non spam
- Insurance
  - $X =$ data of a future policy holder $\quad Y =$ premium
- Finance
  - $X =$ data of a loanee $\quad Y =$ loan rate
  - $X =$ data of a company $\quad Y =$ buy or sell
- Many others...

Some powerful machine learning algorithms

- k-nearest neighbor algorithms
- Artificial neural networks
- Support vector machines
- Aggregation methods (“boosting”)

Aggregation

- Real-valued outputs
- $R(g) = \mathbb{E}[Y - g(X)]^2$
- Given $g_1, \ldots, g_d$, predict as well as

$$g^*_{\mathbf{MS}} \in \operatorname*{argmin}_{g \in \{g_1,\ldots,g_d\}} R(g),$$

$$g^*_{\mathbf{C}} \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \geq 0,\ldots,\theta_d \geq 0, \sum_{j=1}^d \theta_j = 1\}} R(g),$$

$$g^*_{\mathbf{L}} \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \in \mathbb{R},\ldots,\theta_d \in \mathbb{R}\}} R(g).$$

- Combining estimators (Nemirovski, 1998; Juditsky & Nemirovski, 2000; Yang, 2001)

Optimal rates of aggregation

Under suitable assumptions, several works have shown that there exist $\hat{g}_{\mathbf{MS}}$, $\hat{g}_{\mathbf{C}}$ and $\hat{g}_{\mathbf{L}}$ such that

$$\mathbb{E}R(\hat{g}_{\mathbf{MS}}) - R(g^*_{\mathbf{MS}}) \leq C \min\left(\frac{\log d}{n} , 1 \right),$$

$$\mathbb{E}R(\hat{g}_{\mathbf{C}}) - R(g^*_{\mathbf{C}}) \leq C \min\left(\sqrt{\frac{\log(1 + d/\sqrt{n})}{n}} , \frac{d}{n} , 1 \right),$$

$$\mathbb{E}R(\hat{g}_{\mathbf{L}}) - R(g^*_{\mathbf{L}}) \leq C \min\left(\frac{d}{n} , 1 \right),$$

where $\hat{g}_L$ requires the knowledge of the input distribution.

Optimal rates of aggregation (Tsybakov, 2003)

- $\sigma > 0$
- $\mathcal{P}_\sigma$ = set of proba. on $\mathcal{X} \times \mathbb{R}$ such that $Y = g(X) + \xi$, with $\|g\|_\infty \leq 1$, and $\xi \sim \mathcal{N}(0, \sigma^2)$
- For appropriate choices of $g_1, \dots, g_d$:

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \left\{ \mathbb{E} R(\hat{g}) - R(g^*_{\mathbf{MS}}) \right\} \geq C \min\left(\frac{\log d}{n}, 1 \right),$$

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \left\{ \mathbb{E} R(\hat{g}) - R(g^*_{\mathbf{C}}) \right\} \geq C \min\left(\sqrt{\frac{\log(1 + d/\sqrt{n})}{n}}, \frac{d}{n}, 1 \right),$$

$$\inf_{\hat{g}} \sup_{P \in \mathcal{P}_\sigma} \left\{ \mathbb{E} R(\hat{g}) - R(g^*_{\mathbf{L}}) \right\} \geq C \min\left(\frac{d}{n}, 1 \right).$$

Model selection type aggregation: unusual properties

$$g^*_{\mathbf{MS}} \in \underset{g \in \{g_1,\dots,g_d\}}{\operatorname{argmin}} \quad R(g)$$

- To be "optimal", we need to choose $\hat{g}$ outside the model. In particular, $\underset{g \in \{g_1,\dots,g_d\}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]^2$ is suboptimal.
  (Lee, Bartlett, Williamson, 1998; Catoni, 1999; A., 2007; Juditsky, Rigollet, Tsybakov, 2008; Lecué, 2007; Mendelson, 2008)
- Up to recently, the only known optimal algorithm was the progressive mixture rule
- The proof is neither based on bounds on the supremum of empirical processes nor on the PAC-Bayesian analysis
  (Barron, 1987; Catoni, 1997 & 1999; Barron & Yang, 1999; Yang 2000; Juditsky, Rigollet, Tsybakov, 2008; A., 2009)

Progressive mixture rule (Barron, 1987; Catoni, 1999; Yang, 2000)

- π uniform distribution on the finite set $\{g_1, \dots, g_d\}$
- Let $h : \{g_1, \dots, g_d\} \to \mathbb{R}$. Define

$$\pi_h(g) = \frac{\exp[h(g)]}{\sum_{j=1}^d \exp[h(g_j)]} \propto e^{h(g)} \cdot \pi(g)$$

- $\lambda > 0$
- $\Sigma_i(g) = \sum_{k=1}^i [Y_k - g(X_k)]^2$: cumulative loss on the first i data points
- The progressive mixture rule: $\hat{g}_{\mathrm{PM}} = \frac{1}{n+1} \sum_{i=0}^n \mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}} g$,

i.e.,
$$\hat{g}_{\mathrm{PM}}(x) = \frac{1}{n+1} \sum_{j=1}^d \sum_{i=0}^n \frac{e^{-\lambda \Sigma_i(g_j)}}{\sum_{j=1}^d e^{-\lambda \Sigma_i(g_j)}} g_j(x)$$

- Theoretical guarantee for $\mathcal{Y} = [-1, 1]$ and $\lambda = \frac{1}{8}$:

$$\mathbb{E} R(\hat{g}_{\mathrm{PM}}) - R(g^*_{\mathbf{MS}}) \leq \frac{8 \log d}{n+1}$$

Progressive indirect mixture rules (A., 2009)

- $\lambda > 0$
- For any $i \in \{0, \dots, n\}$, let $\hat{h}_i$ be a prediction function s.t.

$$(1) \qquad \forall (x, y) \qquad \left[y - \hat{h}_i(x)\right]^2 \leq -\frac{1}{\lambda} \log \left(\mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}} \exp\left\{-\lambda [y - g(x)]^2\right\} \right)$$

- Progressive indirect mixture rule: $\hat{g}_\lambda = \frac{1}{n+1} \sum_{i=0}^n \hat{h}_i$.
- $\hat{h}_i = \mathbb{E}_{g \sim \pi_{-\lambda \Sigma_i}} g$ satisfies (1) for $\lambda \leq 1/8$.
- $\hat{h}_i$ exists even for $\lambda = 1/2$, and then

$$\mathbb{E} R(\hat{g}_{1/2}) - R(g^*_{\mathbf{MS}}) \leq \frac{2 \log d}{n+1}$$

Excess risk deviations abnormally high (A., 2007)

- $\mathbb{E}R(\hat{g}_\lambda) - R(g^*_{\mathbf{MS}}) = \mathrm{O}\big(\frac{1}{n}\big) \not\Rightarrow R(\hat{g}) - R(g^*_{\mathbf{MS}}) = \mathrm{O}\big(\frac{1}{n}\big)$ w.h.p.
- $g_1 = 1$ and $g_2 = -1$
- For any $\lambda > 0$ and any training set size n large enough, there exist $\epsilon > 0$ and a distribution generating the data for which with probability larger than ϵ, we have

$$R(\hat{g}_\lambda) - R(g^*_{\mathbf{MS}}) \;\geq\; c\sqrt{\frac{\log(e\epsilon^{-1})}{n}}$$

Getting round the previous limitation (A., 2007)

- $r(g) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2$.
- $\hat{g}_{\text{ERM}} \in \underset{g \in \{g_1,\dots,g_d\}}{\operatorname{argmin}} \; r(g)$.
- $[g, g'] = \left\{ \alpha g + (1-\alpha) g' : \alpha \in [0,1] \right\}$.
- The empirical star estimator is

$$\hat{g} \in \underset{g \in [\hat{g}_{\text{ERM}}, g_1] \cup \dots \cup [\hat{g}_{\text{ERM}}, g_d]}{\operatorname{argmin}} r(g).$$

- Theoretical guarantee: with probability at least $1 - \epsilon$,

$$R(\hat{g}) - R(g^*_{\mathbf{MS}}) \le \frac{600 \log(d\epsilon^{-1})}{n}.$$

See also Lecué & Mendelson (2009)

Convex aggregation in high dimension

$$g^*_{\mathbf{C}} \in \underset{g \in \{\sum_{j=1}^d \theta_j g_j ; \theta_1 \geq 0, \dots, \theta_d \geq 0, \sum_{j=1}^d \theta_j = 1\}}{\operatorname{argmin}} R(g)$$

$$\sqrt{n} \ll d \ll e^n$$

- Apply the previous progressive mixture rule on an appropriate grid (Tsybakov, 2003)
- Use the exponentiated gradient algorithm (Kivinen & Warmuth, 1997; Cesa-Bianchi, 1999)
- Use a stochastic version of the mirror descent algorithm (Juditsky, Nazin, Tsybakov, Vayatis, 2005)

Results in expectation, based on a sequential procedure

A PAC-Bayesian approach to convex aggregation (A., 2004)

- $\hat{\rho}_{\mathbf{C}}$ = distribution minimizing a PAC-Bayesian upper bound on $R(\mathbb{E}_{g\sim\hat{\rho}}g) - R(g^*_{\mathbf{C}})$ for any $\hat{\rho}$
- $g^*_{\mathbf{C}} = \mathbb{E}_{g\sim\rho^*_{\mathbf{C}}} g$.
- Theoretical guarantee: with probability at least $1-\epsilon$,

$$R(\mathbb{E}_{g\sim\hat{\rho}_{\mathbf{C}}}g) - R(g^*_{\mathbf{C}}) \le C\sqrt{\frac{\log(d\epsilon^{-1})}{n}\mathbb{E}\mathbf{Var}_{g\sim\rho^*_{\mathbf{C}}} g(X)} + C\frac{\log(d\epsilon^{-1})}{n},$$

- Excess risk at most of order $\sqrt{\frac{\log(d)}{n}}$
- If $\rho^*_{\mathbf{C}}$ is a Dirac, excess risk at most of order $\frac{\log(d)}{n}$
- Exponentially small deviations of the excess risk

Linear aggregation

$$g^*_{\mathbf{L}} \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^{d} \theta_j g_j ; \theta_1 \in \mathbb{R}, \dots, \theta_d \in \mathbb{R}\}} R(g).$$

- Linear aggregation = linear least squares regression
- $f^{(\text{reg})} : x \mapsto \mathbb{E}(Y|X = x)$ not necessarily in the span of $\{g_1, \dots, g_d\}$

Projection estimator (Tsybakov, 2003)

Let $\phi_1, \dots, \phi_d$ be an o.n.b. of span $\{g_1, \dots, g_d\}$ for $\langle f_1, f_2 \rangle = \mathbb{E} f_1(X) f_2(X)$. The projection estimator on this basis is $\hat{f}^{(\text{proj})} = \sum_{j=1}^{d} \hat{\theta}_j^{(\text{proj})} \phi_j$, with

$$\hat{\theta}^{(\text{proj})} = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).$$

If

$$\sup_{x \in \mathcal{X}} \mathbf{V}\text{ar}\,(Y | X = x) = \sigma^2 < +\infty$$

and

$$\|f^{(\text{reg})}\|_\infty = \sup_{x \in \mathcal{X}} |f^{(\text{reg})}(x)| \le H < +\infty,$$

then we have

$$\mathbb{E} R(\hat{f}^{(\text{proj})}) - R(g^*_{\mathbf{L}}) \le (\sigma^2 + H^2) \frac{d}{n}.$$

Empirical risk minimization (Birgé & Massart, 1998)

Assume $\|f^{(\text{reg})}\|_\infty \le H$ and

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[|Y|/A\Big] \Big| X = x\Big\} \le M,$$

for some positive constants A and M. Then, for any $\epsilon > 0$, with probability at least $1 - \epsilon$:

$$R(\hat{f}^{(\text{erm})}) - R(f^*) \le \kappa(A^2 + H^2)\frac{d\log(n) + \log(\epsilon^{-1})}{n},$$

where κ is a positive constant depending only on M.

Empirical risk minimization (Birgé & Massart, 1998)

Assume $\|f^{(\mathrm{reg})}\|_\infty \leq H$ and

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[|Y|/A\Big] \,\Big|\, X = x\Big\} \leq M,$$

for some positive constants A and M. Then, for any $\epsilon > 0$, with probability at least $1 - \epsilon$:

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \leq \kappa(A^2 + H^2)\frac{d \log(n) + \log(\epsilon^{-1})}{n},$$

where κ is a positive constant depending only on M.

Let

$$B = \inf_{\phi_1,\dots,\phi_d} \sup_{\theta\in\mathbb{R}^d-\{0\}} \frac{\| \sum_{j=1}^{d} \theta_j \phi_j \|_\infty^2}{\|\theta\|_\infty^2}$$

where the infimum is taken w.r.t. all possible o.n.b. of span $\{g_1, \dots, g_d\}$ for $\langle f_1, f_2 \rangle = \mathbb{E} f_1(X) f_2(X)$. Then, with probability at least $1 - \epsilon$:

$$R(\hat{f}^{(\mathrm{erm})}) - R(f^*) \le \kappa(A^2 + H^2) \frac{d \log \left[2 + \min\left(\frac{B}{n}, \frac{n}{d}\right)\right] + \log(\epsilon^{-1})}{n},$$

where κ is a positive constant depending only on M.

A PAC-Bayesian approach (A. and Catoni, 2009)

- Assume $\|f^{(\text{reg})}\|_\infty \leq H$ and

$$\text{for any } x \in \mathcal{X}, \quad \mathbb{E}\Big\{\exp\Big[|Y|/A\Big] \Big| X = x\Big\} \leq M,$$

  for some positive constants A and M.
- Let π = uniform distrib. on $B_\infty(0, H) \cap \text{span}\{g_1, \dots, g_d\}$.
- For an appropriate $\lambda > 0$, with probability at least $1 - \epsilon$,

$$R(\mathbb{E}_{g\sim\pi_{-\lambda r}} g) - R(g^*) \leq C\,\frac{d + \log(2\epsilon^{-1})}{n}.$$

- Shrinking effect of $\pi_{-\lambda r}$ when compared to $\hat{g}_{\text{ERM}}$.

Robust estimation to heavy noise (A. and Catoni, 2009)

$$\mathcal{G} \subset \mathrm{span}\left\{g_1, \ldots, g_d\right\} \text{ bounded}, \qquad g^* \in \mathrm{argmin}_{g \in \mathcal{G}} R(g)$$

- Truncation function:

$$\psi(x) = \min\big(1, \max(x, -1)\big).$$

- For the truncation parameter $\alpha > 0$, define

$$\mathcal{D}(f, f') = \sum_{i=1}^{n} \psi\Big(\alpha\big[Y_i - f(X_i)\big]^2 - \alpha\big[Y_i - f'(X_i)\big]^2\Big).$$

and

$$\hat{f} = \underset{f \in \mathcal{G}}{\mathrm{argmin}} \ \max_{f' \in \mathcal{G}} \mathcal{D}(f, f')$$

Robustness of the truncated min-max estimator

$$\mathcal{G} \subset \text{span}\left\{g_1, \ldots, g_d\right\} \qquad g^* \in \text{argmin}_{g\in\mathcal{G}} R(g)$$

$$\sigma = \sqrt{\mathbb{E}\left\{[Y - g^*(X)]^2\right\}} = \sqrt{R(f^*)}, \qquad \kappa = \frac{\sqrt{\mathbb{E}\left\{[\vec{g}(X)^T Q^{-1} \vec{g}(X)]^2\right\}}}{\mathbb{E}\left[\vec{g}(X)^T Q^{-1} \vec{g}(X)\right]},$$

$$\mathcal{S} = \{f \in \text{span}\{g_1^d\} : \mathbb{E}[f(X)^2] = 1\}, \qquad \chi = \max_{f\in\mathcal{S}} \sqrt{\mathbb{E}[f(X)^4]},$$

$$\kappa' = \frac{\sqrt{\mathbb{E}\left\{[Y - g^*(X)]^4\right\}}}{\mathbb{E}\left\{[Y - g^*(X)]^2\right\}}, \qquad T = \max_{f\in\mathcal{G}, f'\in\mathcal{G}} \sqrt{\mathbb{E}[f(X) - f'(X)]^2}.$$

Theorem

For some numerical constants c and c', for $n > c\kappa\chi d$, and an appropriate choice of α, for any $\epsilon > 0$, with proba. at least $1 - \epsilon$,

$$R(\hat{f}) - R(g^*) \leq c\frac{\kappa\kappa' d\sigma^2}{n} + c'\chi\left(\frac{\log(\epsilon^{-1})}{n} + \frac{\kappa^2 d^2}{n^2}\right)\left[\sqrt{\kappa'}\sigma + \sqrt{\chi}T\right]^2.$$

Implementation of the estimator

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{G}} \max_{f' \in \mathcal{G}} \mathcal{D}(f, f')$$

- Solving the min-max problem is nontrivial
- Iterative scheme starting at the ordinary least squares estimator, and considering the functions

  $$\hat{f}_I = \operatorname{argmin}_{f \in \mathcal{F}_{\text{lin}}} \sum_{i \in I} \big(f(X_i) - Y_i\big)^2,$$

  for $I \subset \{1, \dots, n\}$.
- Computational complexity: 50 times the one of the OLS

Input-output functional relationships

- Independent Normalized Covariates (INC(n, d)) and Highly Correlated Covariates (HCC(n, d)):

  X is a d-dimensional centered normal random vector

$$g_j(X) = X^{(j)} \qquad \text{and} \qquad Y = \langle \theta^*, X \rangle + \sigma W,$$

- Trigonometric series (TS(n, d)):
  - $X \sim \mathcal{U}([0,1])$,

$$\big(g_1(X), \dots, g_d(X)\big) = \big(\cos(2\pi X), \dots, \cos(d\pi X), \sin(2\pi X), \dots, \sin(d\pi X)\big),$$

  - $Y = 20X^2 - 10X - \frac{5}{3} + \sigma W.$

Noise = 95% Gaussian + 5% Dirac

	nb of iterations	iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})})\vert\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f})\vert\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	419	405	0.567($\pm$0.083)	0.178($\pm$0.025)	1.191($\pm$0.178)	0.262($\pm$0.052)
INC(n=200,d=2)	1000	506	498	1.055($\pm$0.112)	0.271($\pm$0.030)	1.884($\pm$0.193)	0.334($\pm$0.050)
HCC(n=200,d=2)	1000	502	494	1.045($\pm$0.103)	0.267($\pm$0.024)	1.866($\pm$0.174)	0.316($\pm$0.032)
TS(n=200,d=2)	1000	561	554	1.069($\pm$0.089)	0.310($\pm$0.027)	1.720($\pm$0.132)	0.367($\pm$0.036)
INC(n=1000,d=2)	1000	402	392	0.204($\pm$0.015)	0.109($\pm$0.008)	0.316($\pm$0.029)	0.081($\pm$0.011)
INC(n=1000,d=10)	1000	950	946	1.030($\pm$0.041)	0.228($\pm$0.016)	1.051($\pm$0.042)	0.207($\pm$0.014)
HCC(n=1000,d=10)	1000	942	942	0.980($\pm$0.038)	0.222($\pm$0.015)	1.008($\pm$0.039)	0.203($\pm$0.015)
TS(n=1000,d=10)	1000	976	973	1.009($\pm$0.037)	0.228($\pm$0.017)	1.018($\pm$0.038)	0.217($\pm$0.016)
INC(n=2000,d=2)	1000	209	207	0.104($\pm$0.007)	0.078($\pm$0.005)	0.206($\pm$0.021)	0.082($\pm$0.012)
HCC(n=2000,d=2)	1000	184	183	0.099($\pm$0.007)	0.076($\pm$0.005)	0.196($\pm$0.023)	0.070($\pm$0.010)
TS(n=2000,d=2)	1000	172	171	0.101($\pm$0.007)	0.080($\pm$0.005)	0.206($\pm$0.020)	0.083($\pm$0.012)
INC(n=2000,d=10)	1000	669	669	0.510($\pm$0.018)	0.206($\pm$0.012)	0.572($\pm$0.023)	0.117($\pm$0.009)
HCC(n=2000,d=10)	1000	669	669	0.499($\pm$0.018)	0.207($\pm$0.013)	0.561($\pm$0.023)	0.125($\pm$0.011)
TS(n=2000,d=10)	1000	754	753	0.516($\pm$0.018)	0.195($\pm$0.013)	0.558($\pm$0.022)	0.131($\pm$0.011)

Heavy tailed noise: $\mathbb{E}|Y|^{2.01} = +\infty$

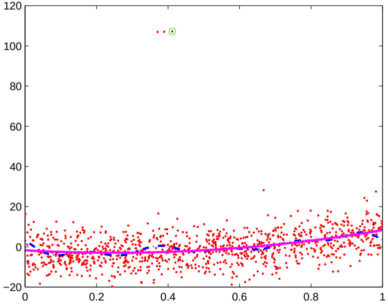
	nb of iterations	iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})})\vert\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f})\vert\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	163	145	7.72(±3.46)	3.92(±0.409)	30.52(±20.8)	7.20(±1.61)
INC(n=200,d=2)	1000	104	98	22.69(±23.14)	19.18(±23.09)	45.36(±14.1)	11.63(±2.19)
HCC(n=200,d=2)	1000	120	117	18.16(±12.68)	8.07(±0.718)	99.39(±105)	15.34(±4.41)
TS(n=200,d=2)	1000	110	105	43.89(±63.79)	39.71(±63.76)	48.55(±18.4)	10.59(±2.01)
INC(n=1000,d=2)	1000	104	100	3.98(±2.25)	1.78(±0.128)	23.18(±21.3)	2.03(±0.56)
INC(n=1000,d=10)	1000	253	242	16.36(±5.10)	7.90(±0.278)	41.25(±19.8)	7.81(±0.69)
HCC(n=1000,d=10)	1000	220	211	13.57(±1.93)	7.88(±0.255)	33.13(±8.2)	7.28(±0.59)
TS(n=1000,d=10)	1000	214	211	18.67(±11.62)	13.79(±11.52)	30.34(±7.2)	7.53(±0.58)
INC(n=2000,d=2)	1000	113	103	1.56(±0.41)	0.89(±0.059)	6.74(±3.4)	0.86(±0.18)
HCC(n=2000,d=2)	1000	105	97	1.66(±0.43)	0.95(±0.062)	7.87(±3.8)	1.13(±0.23)
TS(n=2000,d=2)	1000	101	95	1.59(±0.64)	0.88(±0.058)	8.03(±6.2)	1.04(±0.22)
INC(n=2000,d=10)	1000	259	255	8.77(±4.02)	4.23(±0.154)	21.54(±15.4)	4.03(±0.39)
HCC(n=2000,d=10)	1000	250	242	6.98(±1.17)	4.13(±0.127)	15.35(±4.5)	3.94(±0.25)
TS(n=2000,d=10)	1000	238	233	8.49(±3.61)	5.95(±3.486)	14.82(±3.8)	4.17(±0.30)

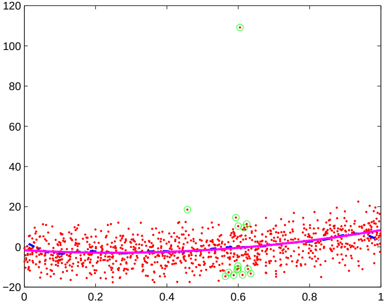
Standard Gaussian noise

	nb of iterations	iter. with $R(\hat{f}) \neq R(\hat{f}^{(\text{ols})})$	iter. with $R(\hat{f}) < R(\hat{f}^{(\text{ols})})$	$\mathbb{E}R(\hat{f}^{(\text{ols})}) - R(f^*)$	$\mathbb{E}R(\hat{f}) - R(f^*)$	$\mathbb{E}R[(\hat{f}^{(\text{ols})})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$	$\mathbb{E}[R(\hat{f})\|\hat{f} \neq \hat{f}^{(\text{ols})}] - R(f^*)$
INC(n=200,d=1)	1000	20	8	0.541($\pm$0.048)	0.541($\pm$0.048)	0.401($\pm$0.168)	0.397($\pm$0.167)
INC(n=200,d=2)	1000	1	0	1.051($\pm$0.067)	1.051($\pm$0.067)	2.566	2.757
HCC(n=200,d=2)	1000	1	0	1.051($\pm$0.067)	1.051($\pm$0.067)	2.566	2.757
TS(n=200,d=2)	1000	0	0	1.068($\pm$0.067)	1.068($\pm$0.067)	–	–
INC(n=1000,d=2)	1000	0	0	0.203($\pm$0.013)	0.203($\pm$0.013)	–	–
INC(n=1000,d=10)	1000	0	0	1.023($\pm$0.029)	1.023($\pm$0.029)	–	–
HCC(n=1000,d=10)	1000	0	0	1.023($\pm$0.029)	1.023($\pm$0.029)	–	–
TS(n=1000,d=10)	1000	0	0	0.997($\pm$0.028)	0.997($\pm$0.028)	–	–
INC(n=2000,d=2)	1000	0	0	0.112($\pm$0.007)	0.112($\pm$0.007)	–	–
HCC(n=2000,d=2)	1000	0	0	0.112($\pm$0.007)	0.112($\pm$0.007)	–	–
TS(n=2000,d=2)	1000	0	0	0.098($\pm$0.006)	0.098($\pm$0.006)	–	–
INC(n=2000,d=10)	1000	0	0	0.517($\pm$0.015)	0.517($\pm$0.015)	–	–
HCC(n=2000,d=10)	1000	0	0	0.517($\pm$0.015)	0.517($\pm$0.015)	–	–
TS(n=2000,d=10)	1000	0	0	0.501($\pm$0.015)	0.501($\pm$0.015)	–	–

Group of points disregarded by the min-max estimator

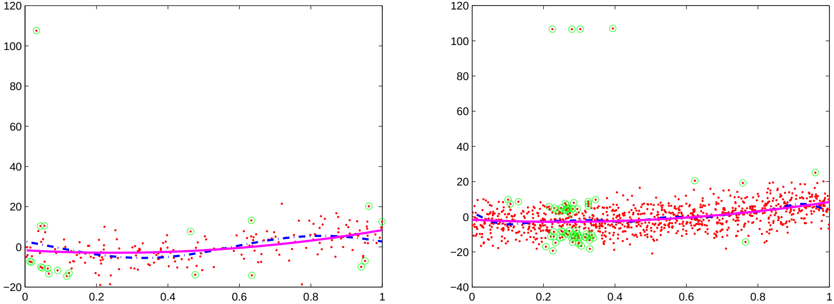
- $TS(n = 1000, d = 10)$
- Mixture noise = 95% Gaussian + 5% Dirac





Group of points disregarded by the min-max estimator

- $TS(n = 1000, d = 10)$
- Mixture noise = 95% Gaussian + 5% Dirac



High-dimensional input and sparsity

$$n \ll d \ll e^n$$

- predicting as $g^*_{\mathbf{C}}$ = achievable : $\sqrt{\frac{\log d}{n}}$
- predicting as $g^*_{\mathbf{L}}$ = not achievable : $\frac{d}{n}$

$$g^* \in \operatorname*{argmin}_{g \in \{\sum_{j=1}^{d} \theta_j g_j ; \theta_1 \in \mathbb{R}, \dots, \theta_d \in \mathbb{R}, \sum_{j=1}^{d} \mathbf{1}_{\theta_j \neq 0} \leq s\}} R(g).$$

- g^* achievable by Lasso under strong assumptions on the correlations of $g_1(X), \dots, g_d(X)$: rate = $\frac{s \log(d)}{n}$
- g^* should be achievable by penalization proportional to the number of nonzero coefficients but with rate $\sqrt{\frac{s \log(d)}{n}}$ (Bunea, Tsybakov, Wegkamp, 2007; Birgé and Massart, 2007; Raskutti, Wainwright, Yu, 2009)

A model selection approach

- $\mathcal{L}_1 = \{Z_1, \dots, Z_{n/2}\}$, and $\mathcal{L}_2 = \{Z_{n/2+1}, \dots, Z_n\}$
- For any $I \subset \{1, \dots, d\}$ of size s, let $\hat{g}_I$ be the Gibbs estimator for linear aggregation of $(g_j)_{j \in I}$ trained on $\mathcal{L}_1$
- Let $\hat{g}$ be the empirical star estimator trained on $\mathcal{L}_2$ and associated with the $\binom{d}{s}$ functions $\hat{g}_I$

$$R(\hat{g}) - R(g^*) \leq C \frac{s \log(d/s) + \log(2\epsilon^{-1})}{n}$$

Conclusion

- New estimators solving the three aggregation problems
- **L** and **MS** are central problems: building blocks for getting nontrivial results
- Open problems: provide robust efficient estimators with small and concentrated excess risk
  - problem **L**
  - Learning sparse models