Heuristique de pente pour des M-estimateurs à contraste régulier

Journées MAS 2010, Bordeaux

Adrien Saumard

Université Rennes 1, IRMAR

Sept 3, 2010

1. The Slope Phenomenon, introduction and first heuristics

2. Optimal control of the excess risks when the contrast is "regular", fixed model case

1 - The Slope Phenomenon, introduction and first heuristics

Some general notations

- Unknown law P on a measurable space $(\mathcal{Z}, \mathcal{T})$, generic random variable Z of law P.
- We are given $(Z_1, ..., Z_n)$ i.i.d. sample of law $P^{\otimes n}$ (also independent of Z).
- Empirical measure associated to the sample

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$$

- Expectations :

$$\begin{aligned} P(s) &= \mathbb{E}[s(Z)] \ , \\ P_n(s) &= \frac{1}{n} \sum_{i=1}^{n} s(Z_i) \ . \end{aligned}$$

- Norms :

$$\|s\|_{2,\mu} = \sqrt{\mu\left(s^2\right)} \qquad ; \qquad \|s\|_2 := \|s\|_{L_2(P)}$$

$$\|s\|_\infty = \mathop{ess\,sup}_{z \in \mathcal{Z}} |s(z)| \ .$$

- Positive and negative parts :

$$(x)_+ := \max\{x; 0\} \ ; \ (x)_- := \max\{-x; 0\} \geq 0 \qquad \forall x \in \mathbb{R}$$

$$(f)_\pm : x \in \mathcal{D}_f \longmapsto (f(x))_\pm \ .$$

- A functional space : (not a vector space !)

$$L_1^-(P) := \left\{ f : (\mathcal{Z}, \mathcal{T}) \to \overline{\mathbb{R}}, \ P(f)_- < +\infty \right\} (\supset L_1(P)) \ .$$

Expectation is well-defined on $L_1^-(P)$,

$$Pf := P(f)_+ - P(f)_- \in (-\infty; +\infty] \ .$$

M–estimation

Definitions (Contrast, Target, Risk)

Given $(\mathcal{Z}, \mathcal{T}, P)$, a **Contrast** is a functional K defined from a set $\mathcal{S}$ of functions to $L_1^-(P)$,

$$K : \begin{cases} \mathcal{S} \longrightarrow L_1^-(P) := \{f : (\mathcal{Z}, \mathcal{T}) \to \mathbb{R},\ P(f)_- < +\infty\} \\ s \longmapsto (Ks : z \longmapsto (Ks)(z)) \end{cases} ,$$

such that the **risk** function (for any $s \in \mathcal{S}$, $P(Ks)$ is called the **risk** of s)

$$PK : \begin{cases} \mathcal{S} \longrightarrow (-\infty; +\infty] \\ s \longmapsto P(Ks) := \mathbb{E}[(Ks)(Z)] \end{cases}$$

is proper (i.e. not identically equal to $+\infty$) and admits a unique minimum. The argument of this minimum is called the **target**, denoted by s_*.

Definition (M-estimator)

Let $K : \mathcal{S} \to L_1^- (P)$ be a contrast and let $M \subset \mathcal{S}$ such the restriction of the risk function PK to M is proper. M is called a **model**. We call **M-estimator** associated to the contrast K and to the model M, a random variable $s_n (M)$ satisfying

$$s_n (M) \in \arg\min_{s \in M} P_n (Ks) \quad , \quad |P_n (Ks_n (M))| < +\infty \;\; a.s.$$

- Quality of a M-estimator : measured by its **excess risk**,

$$\ell (s_*, s_n (M)) := P (Ks_n (M)) - P (Ks_*) = P (Ks_n (M) - Ks_*) \geq 0 \ .$$

Examples

- Maximum likelihood estimation of density (**MLE**) :

$$s_* = \frac{dP}{d\mu} \; ; \; K(s) = -\ln s$$

Excess risk : **Kullback-Leibler divergence** of s w.r.t. s_*.

$$\ell(s_*, s) = \mathcal{K}(s_*, s) = \int_{\mathcal{Z}} s_* \ln\left(\frac{s_*}{s}\right) d\mu$$

Examples

- Maximum likelihood estimation of density (**MLE**) :

$$s_* = \frac{dP}{d\mu} \ ; \ K(s) = -\ln s$$

Excess risk : **Kullback-Leibler divergence** of s w.r.t. s_*.

$$\ell(s_*, s) = \mathcal{K}(s_*, s) = \int_{\mathcal{Z}} s_* \ln\left(\frac{s_*}{s}\right) d\mu$$

- Least-square estimation of density (**LSE**):

$$s_* = \frac{dP}{d\mu}.$$

Goal : $\ell(s_*, s) = P(Ks - Ks_*) = \|s - s_*\|_{2,\mu}^2$ (excess risk given by the **quadratic norm** of $L_2(\mu)$).
Contrast : $K(s) = \|s\|_{2,\mu}^2 - 2Ps$.

- Least-squares **heteroscedastic Regression** :

$$Y = s_*(X) + \sigma(X)\varepsilon, \quad \mathbb{E}[\varepsilon \,|X] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon^2 \,|X\right] = 1 \; .$$

If $\mathcal{S} = L_2\left(P^X\right)$ and $K : \mathcal{S} \longmapsto L_1(P)\left(\subset L_1^-(P)\right)$, with

$$Ks : z = (x,y) \mapsto (Ks)(z) = (Ks)(x,y) = (y - s(x))^2 \; .$$

$$\ell(s_*, s) = P(Ks - Ks_*) = P^X(s - s_*)^2 = \|s - s_*\|_2^2 \; .$$

Excess risk : **quadratic norm** of $L_2\left(P^X\right)$.

- Binary Classification : $Z = (X, Y) \in \mathcal{X} \times \{-1, 1\}$.
Target : **Bayes classifier**.

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) = \arg\min_{s \in \mathcal{S}} P(Y \neq s(X)) \; .$$

Contrast : $K(s)(x,y) = \mathbf{1}_{y \cdot s(x) \geq 0}$.

- Least-squares **heteroscedastic Regression** :

$$Y = s_*(X) + \sigma(X)\varepsilon, \quad \mathbb{E}[\varepsilon \,|X] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon^2 \,|X\right] = 1 \ .$$

  If $\mathcal{S} = L_2\left(P^X\right)$ and $K : \mathcal{S} \longmapsto L_1(P)\left(\subset L_1^-(P)\right)$, with

$$Ks : z = (x,y) \mapsto (Ks)(z) = (Ks)(x,y) = (y - s(x))^2 \ .$$

$$\ell(s_*, s) = P(Ks - Ks_*) = P^X(s - s_*)^2 = \|s - s_*\|_2^2 \ .$$

  Excess risk : **quadratic norm** of $L_2\left(P^X\right)$.

- Binary Classification : $Z = (X, Y) \in \mathcal{X} \times \{-1, 1\}$.
  Target : **Bayes classifier**.

$$s_* = \arg\min_{s \in \mathcal{S}} P(Ks) = \arg\min_{s \in \mathcal{S}} P(Y \neq s(X)) \ .$$

  Contrast : $K(s)(x,y) = \mathbf{1}_{y \cdot s(x) \geq 0}$.

- Convex binary classification : SVM, Boosting, Logistic regression, etc...
  Contrasts : **convex surrogate** of the Bayes contrast.

$$K_\phi : s \mapsto [K_\phi(s) : z = (x,y) \mapsto K_\phi(s)(z) = \phi(y \cdot s(x))] \ .$$

Model Selection in M-estimation, via penalization.

- Contrast : $K : \mathcal{S} \rightarrow L_1^-(P)$, target $s_* = \arg\min_{s \in \mathcal{S}} P(Ks)$.

Model Selection in M-estimation, via penalization.

- Contrast : $K : \mathcal{S} \rightarrow L_1^- (P)$, target $s_* = \arg\min_{s \in \mathcal{S}} P(Ks)$.
- Collection of models : $\mathcal{M}_n$. Associated collection of M-estimators : $\{s_n(M) \ ; \ M \in \mathcal{M}_n\}$,

$$s_n(M) \in \arg\min_{s \in M} P_n(Ks) \ , \qquad \forall M \in \mathcal{M}_n \ .$$

Model Selection in M-estimation, via penalization.

- Contrast : $K : \mathcal{S} \rightarrow L_1^- (P)$, target $s_* = \arg\min_{s \in \mathcal{S}} P(Ks)$.
- Collection of models : $\mathcal{M}_n$. Associated collection of M-estimators : $\{s_n(M) \;;\; M \in \mathcal{M}_n\}$,

$$s_n(M) \in \arg\min_{s \in M} P_n(Ks) \ , \qquad \forall M \in \mathcal{M}_n \ .$$

- Oracle model (target of the model selection procedure) :

$$\begin{aligned} M_* &\in \arg\min_{M \in \mathcal{M}_n} P(Ks_n(M)) \\ &= \arg\min_{M \in \mathcal{M}_n} P(Ks_n(M) - Ks_*) \\ &= \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M) - Ks_*) + (P - P_n)(Ks_n(M) - Ks_*)\} \ . \end{aligned}$$

Model Selection in M-estimation, via penalization.

- Contrast : $K : \mathcal{S} \to L_1^-(P)$, target $s_* = \arg\min_{s \in \mathcal{S}} P(Ks)$.
- Collection of models : $\mathcal{M}_n$. Associated collection of M-estimators : $\{s_n(M) \ ; \ M \in \mathcal{M}_n\}$,

$$s_n(M) \in \arg\min_{s \in M} P_n(Ks) \ , \qquad \forall M \in \mathcal{M}_n \ .$$

- Oracle model (target of the model selection procedure) :

$$\begin{aligned} M_* &\in \arg\min_{M \in \mathcal{M}_n} P(Ks_n(M)) \\ &= \arg\min_{M \in \mathcal{M}_n} P(Ks_n(M) - Ks_*) \\ &= \arg\min_{M \in \mathcal{M}_n} \{P_n(Ks_n(M) - Ks_*) + (P - P_n)(Ks_n(M) - Ks_*)\} \ . \end{aligned}$$

- If $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$, then we set the **ideal penalty (**S. Arlot, PhD Thesis, 2007),

$$\mathrm{pen}_{\mathrm{id}} : M \in \mathcal{M}_n \mapsto \mathrm{pen}_{\mathrm{id}}(M) = (P - P_n)(Ks_n(M) - Ks_*) \geq 0 \ .$$

Hence,

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(K s_n (M) - K s_* \right) + \mathrm{pen}_{\mathrm{id}} (M) \right\} \ .$$

- Selected model : Choose pen : $M \in \mathcal{M}_n \mapsto \mathrm{pen}(M) \geq 0$ and select

$$\begin{aligned} \widehat{M} &\in \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(K s_n (M) \right) + \mathrm{pen}(M) \right\} \\ &= \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(K s_n (M) - K s_* \right) + \mathrm{pen}(M) \right\} \ . \end{aligned}$$

Hence,

$$M_* \in \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(Ks_n(M) - Ks_* \right) + \mathrm{pen}_{\mathrm{id}}(M) \right\} \ .$$

- Selected model : Choose pen : $M \in \mathcal{M}_n \mapsto \mathrm{pen}(M) \geq 0$ and select

$$\begin{aligned} \widehat{M} &\in \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(Ks_n(M) \right) + \mathrm{pen}(M) \right\} \\ &= \arg\min_{M \in \mathcal{M}_n} \left\{ P_n \left(Ks_n(M) - Ks_* \right) + \mathrm{pen}(M) \right\} \ . \end{aligned}$$

- Quality of the procedure : measured by an **oracle inequality**. With large probability (of order $1 - Ln^{-2}$),

$$\ell \left(s_*, s_n \left(\widehat{M} \right) \right) \leq C \times \ell \left(s_*, s_n(M_*) \right) \ .$$

The smaller is $C \geq 1$ (under a fixed probability), the better is the model selection procedure in terms of **prediction** (measured by the excess risk).

Optimal Model Selection, Slope Heurisitics

- A model selection procedure is **optimal** - or nearly optimal - if, with probability at least $1 - Ln^{-2}$, we have

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \leq (1 + \varepsilon_n) \times \ell\left(s_*, s_n\left(M_*\right)\right) \ , \ \ \varepsilon_n \to 0 \ .$$

<u>**Slope Heuristics** :</u> (Birgé & Massart, 2007, extended by Arlot & Massart, 2009) There exists a penalty, called **minimal penalty** and denoted $\mathrm{pen}_{\min}$, such that:

(I) If a penalty $\mathrm{pen} : \mathcal{M}_n \longrightarrow \mathbb{R}_+$ is such that, for all model $M \in \mathcal{M}_n$,

$$\mathrm{pen}\left(M\right) \leq (1 - \delta)\, \mathrm{pen}_{\min}$$

with $\delta > 0$, then the dimension of the selected model $\widehat{M}$ is "very large" and the excess risk of the selected estimator $s_n\left(\widehat{M}\right)$ is "much larger" than the excess risk of the oracle.

(II) If pen $\approx (1+\delta)\,\mathrm{pen}_{\min}$ with $\delta > 0$, then the corresponding model selection procedure satisfies an oracle inequality with a leading constant $C(\delta) < +\infty$ and the dimension of the selected model is "not too large".
(III) Moreover,

$$\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\min}$$

is a (quasi)-optimal penalty.

If the **projection** s_M of the target s_* exists and is unique, i.e.

$$s_M = \arg\min_{s \in M} P(Ks)\,,\ Ks_M \in L_1(P)\ ,$$

then

$$P_n(Ks_M - Ks_n(M)) \geq 0$$

is called the **empirical excess risk on** M. In this case, Arlot & Massart (09) conjecture that the following equality holds with great generality,

$$\text{for all } M \in \mathcal{M}_n, \qquad \mathrm{pen}_{\min}(M) = \mathbb{E}\left[P_n(Ks_M - Ks_n(M))\right]\ .$$

- **Question :** In what extend the conjecture of Arlot & Massart is true ? Find a general positive answer, find (nontrivial) counter-examples...

Practice ? Listen to the talk of J-P. Baudry

- Baudry, Maugis & Michel (10) : Survey. Overview of the theoritical and practical results about the Slope heuristics. Logiciel CAPUSHE. **Already many conclusive empirical study (simulations and real data) !**

- When it is possible, use the Slope Heuristics to calibrate your penalty in practice, it seems to work quite well !...

One Heuristic on the slope phenomenon

If we take $\mathrm{pen} \approx 2 \times \mathrm{pen}_{\min} = 2\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right]$. $\widehat{M}$ minimizes

$$
\begin{aligned}
& P_n\left(Ks_n\left(M\right) - Ks_*\right) + \mathrm{pen}\left(M\right) \\
& \approx \ell\left(s_*, s_M\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) + \left(P_n - P\right)\left(Ks_M - Ks_*\right) \\
& \quad + 2\underbrace{\left(\mathbb{E}\left[P_n\left(Ks_M - Ks_n\left(M\right)\right)\right] - P_n\left(Ks_M - Ks_n\left(M\right)\right)\right)}_{\text{Boucheron \& Massart, 2010.}} \\
& \approx \ell\left(s_*, s_M\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) \; .
\end{aligned}
$$

If

$$\boxed{P_n\left(Ks_M - Ks_n\left(M\right)\right) \sim P\left(Ks_n\left(M\right) - Ks_M\right)} \qquad (*)$$

then

$$\begin{aligned} P\left(Ks_M - Ks_*\right) + P_n\left(Ks_M - Ks_n\left(M\right)\right) &\approx \ell\left(s_*, s_M\right) + P\left(Ks_n\left(M\right) - Ks_M\right) \\ &\approx \ell\left(s_*, s_n\left(M\right)\right) \ . \end{aligned}$$

Hence,

$$\ell\left(s_*, s_n\left(\widehat{M}\right)\right) \approx \ell\left(s_*, s_n\left(M_*\right)\right)$$

and the procedure is nearly optimal.

The **keystone of the slope heuristics** is the equivalence (*) with high probability between the true and empirical excess risk, for the model of interest.

2 - Optimal control of the excess risks when the contrast is "regular", fixed model case

The notion of regular contrast

Definition (Regular Contrast w.r.t. a model)

Let $K : \mathcal{S} \rightarrow L_1^-(P)$ be a contrast with $\mathcal{S}$. Take $M \subset \mathcal{S}$ a convex model. Then K is said to be regular w.r.t. M if there exists a projection s_M of the target s_* onto M,

$$s_M \in \arg\min_{s \in M} P(Ks) \quad ,$$

if the restriction $PK_{|M} : M \rightarrow (-\infty, +\infty]$ is strictly convex and if there exists $c > 0$ such that, by denoting

$$B_c := \{s \in \mathrm{Aff}(M) \; ; \; \|s - s_M\|_\infty < c\}$$

we have

$$B_c \subset M$$

and the restriction $K_{|B_c} : B_c \rightarrow L_\infty(P)$ is $\mathcal{C}^3$ in the sense of the Fréchet-differentiability.

Optimal bounds for the excess risks

Theorem

Let $\alpha, A_-, A_+, A_H, A_{cons} > 0$ *and let* $K : \mathcal{S} \to L_1^-(P)$, *be a regular contrast w.r.t. a model* M. *Denote by* M_0 *the underlying vector space of* $\mathrm{Aff}(M)$. *Assume that*

$$0 < A_- (\ln n)^2 \leq \dim(M_0) = D \leq A_+ \frac{n}{(\ln n)^2} < +\infty$$

and that there exists a positive constant $A_H > 0$ *such that*

$$\text{for all } s \in B_c,\ \|s - s_M\|_2 \leq A_H P\left(K''(s_M)(s - s_M, s - s_M)\right) .$$

Hence the norm defined by

$$\|h\|_{H,M} = \sqrt{P\left(K''(s_M)(h,h)\right)},\ h \in M_0,$$

is an Hilbertian norm on M_0.

Theorem

Moreover, assume that there exists $R_{n,D,\alpha} \leq A_{cons} (\ln n)^{-1/2}$ *such that for all* $n \geq n_1$,

$$\mathbb{P}\left[\|s_n(M) - s_M\|_\infty > R_{n,D,\alpha}\right] \leq n^{-\alpha} \ .$$

Finally, assume that $\left(M_0, \|\cdot\|_{H,M}\right)$ *has a localized basis structure : there exists an orthonormal basis* $\varphi = (\varphi_k)_{k=1}^D$ *in* $\left(M_0, \|\cdot\|_{H,M}\right)$ *that satisfies, for a positive constant* $r_M(\varphi)$ *and all* $\beta = (\beta_k)_{k=1}^D \in \mathbb{R}^D$,

$$\left\|\sum_{k=1}^{D} \beta_k \varphi_k\right\|_\infty \leq r_M(\varphi)\sqrt{D}\,|\beta|_\infty \ ,$$

where $|\beta|_\infty = \max\{|\beta_k|\,; k \in \{1, ..., D\}\}$ *is the sup-norm of the D-dimensional vector* β.

Theorem

Then there exists $A_0 > 0$ and a positive number n_0 depending on the constants of the problem such that by setting

$$\varepsilon_n = A_0 \max\left\{ \left(\frac{\ln n}{D}\right)^{1/4}, \left(\frac{D \ln n}{n}\right)^{1/4}, \sqrt{R_{n,D,\alpha}} \right\} ,$$

we have for all $n \geq n_0$, with probability at least $1 - 15n^{-\alpha}$,

$$\left(1 - \varepsilon_n\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \leq P\left(Ks_n\left(M\right) - Ks_M\right) \leq \left(1 + \varepsilon_n\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 ,$$

$$\left(1 - \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 \leq P_n\left(Ks_M\left(M\right) - Ks_n\right) \leq \left(1 + \varepsilon_n^2\right) \frac{1}{4} \frac{D}{n} \mathcal{K}_{1,M}^2 ,$$

where $\mathcal{K}_{1,M}^2 = D^{-1} \sum_{k=1}^{D} \operatorname{Var}\left(K'\left(s_M\right)\left(\varphi_k\right)\right)$.

- **Conclusion :** In this case, we have proved

$$P\left(Ks_n\left(M\right) - Ks_M\right) \sim P_n\left(Ks_M - Ks_n\left(M\right)\right) .$$

References

Preprints:

- A.S., Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression, 2010, hal-00512304, v1.
- A.S., The Slope Heuristics in Heteroscedastic Regression, 2010, hal-00512306, v1.
- A.S., Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models, 2010, hal-00512310, v1.

In preparation:

- Convergence in sup-norm of the least-squares estimator of a regression function with heteroscedastic noise.
- Regular Contrast Estimation on a fixed convex model.
- Regular Contrast Estimation and the Slope Heuristics.