Matching pursuit and basis pursuit: a comparison

Albert Cohen

Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
Paris

Collaborators: Andrew Barron, Wolfgang Dahmen, Ron DeVore
+ Rémi Gribonval, Christine de Mol and Jean-Marie Mirebeau

College Station, 2007

A general problem

Consider a dictionary of functions $\mathcal{D} = (g)_{g \in \mathcal{D}}$ in a Hilbert space H.

A general problem

Consider a dictionary of functions $\mathcal{D} = (g)_{g \in \mathcal{D}}$ in a Hilbert space H.

The dictionary is countable, normalized ($\|g\| = 1$), complete, but not orthogonal and possibly redundant.

A general problem

Consider a dictionary of functions $\mathcal{D} = (g)_{g \in \mathcal{D}}$ in a Hilbert space H.

The dictionary is countable, normalized ($\|g\| = 1$), complete, but not orthogonal and possibly redundant.

Sparse approximation problem : given $N > 0$ and $f \in H$, construct an N-term combination

$$f_N = \sum_{k=1,\cdots,N} c_k g_k,$$

with $g_k \in \mathcal{D}$, which approximates f “at best”, and study how fast f_N converges towards f.

A general problem

Consider a dictionary of functions $\mathcal{D} = (g)_{g\in\mathcal{D}}$ in a Hilbert space H.

The dictionary is countable, normalized ($\|g\| = 1$), complete, but not orthogonal and possibly redundant.

Sparse approximation problem : given $N > 0$ and $f \in H$, construct an N-term combination

$$f_N = \sum_{k=1,\cdots,N} c_k g_k,$$

with $g_k \in \mathcal{D}$, which approximates f "at best", and study how fast f_N converges towards f.

Sparse recovery problem : if f has an unknown representation $f = \sum_g c_g g$ with $(c_g)_{g\in\mathcal{D}}$ a (possibly) sparse sequence, recover this sequence exactly or approximately from the data of f.

Orthogonal matching pursuit

f_N is constructed by an greedy algorithm. Initialization: $f_0 = 0$.

At step $k-1$, the approximation is defined by

$$f_{k-1} := P_{k-1} f,$$

with P_{k-1} the orthogonal projection onto $\mathrm{Span}\{g_1, \cdots, g_{k-1}\}$.

Choice of next element based on the residual $r_{k-1} = f - P_{-1} k f$:

$$g_k := \mathrm{Argmax}_{g \in \mathcal{D}} |\langle r_{k-1}, g \rangle|.$$

Orthogonal matching pursuit

f_N is constructed by an greedy algorithm. Initialization: $f_0 = 0$.

At step $k-1$, the approximation is defined by

$$f_{k-1} := P_{k-1} f,$$

with P_{k-1} the orthogonal projection onto $\mathrm{Span}\{g_1, \cdots, g_{k-1}\}$.

Choice of next element based on the residual $r_{k-1} = f - P_{k-1} f$:

$$g_k := \mathrm{Argmax}_{g \in \mathcal{D}} |\langle r_{k-1}, g \rangle|.$$

Many variants : PGA, RGA, WGA, CGA, WCGA, WGAFR... (Acta Numerica survey by Vladimir Temlyakov to appear).

First versions studied in the 1970's by statisticians (Friedman, Huber, Stuetzle, Tukey...).

Relaxed greedy algorithm

Define

$$f_k := \alpha_k f_{k-1} + \beta_k g_k,$$

with

$$(\alpha_k, \beta_k, g_k) := \text{Argmin}_{(\alpha,\beta,g)\in\mathbb{R}^2\times\mathcal{D}} \|f - \alpha f_{k-1} + \beta g\|.$$

Relaxed greedy algorithm

Define

$$f_k := \alpha_k f_{k-1} + \beta_k g_k,$$

with

$$(\alpha_k, \beta_k, g_k) := \text{Argmin}_{(\alpha,\beta,g)\in \mathbb{R}^2\times\mathcal{D}} \|f - \alpha f_{k-1} + \beta g\|.$$

Simpler version: set α_k in advance and optimize only β and g.

If $\alpha_k = 1$, this gives the Pure Greedy Algorithm :

$$g_k := \text{Argmax}_{g\in\mathcal{D}} |\langle r_{k-1}, g\rangle| \ \text{ and } \ f_k := f_{k-1} + \langle r_{k-1}, g_k\rangle g_k.$$

with $r_{k-1} = f - f_{k-1}$.

Relaxed greedy algorithm

Define

$$f_k := \alpha_k f_{k-1} + \beta_k g_k,$$

with

$$(\alpha_k, \beta_k, g_k) := \text{Argmin}_{(\alpha,\beta,g)\in \mathbb{R}^2\times\mathcal{D}} \|f - \alpha f_{k-1} + \beta g\|.$$

Simpler version: set α_k in advance and optimize only β and g.

If $\alpha_k = 1$, this gives the Pure Greedy Algorithm :

$$g_k := \text{Argmax}_{g\in\mathcal{D}} |\langle r_{k-1}, g\rangle| \ \text{ and } \ f_k := f_{k-1} + \langle r_{k-1}, g_k\rangle g_k.$$

with $r_{k-1} = f - f_{k-1}$. It might be wiser to take $\alpha_k = (1 - \frac{c}{k})_+$:

$$g_k := \text{Argmax}_{g\in\mathcal{D}} |\langle \tilde{r}_{k-1}, g\rangle| \ \text{ and } \ f_k := f_{k-1} + \langle \tilde{r}_{k-1}, g_k\rangle g_k.$$

with the modified residual $\tilde{r}_{k-1} := f - (1 - \frac{c}{k})_+ f_{k-1}$.

Basis pursuit

Finding the best approximation of f by N elements of the dictionary is equivalent to the support minimization problem

$$\min\{\|(c_g)\|_{\ell^0} \; ; \; \|f - \sum c_g g\| \leq \varepsilon\}$$

which is usually untractable.

Basis pursuit

Finding the best approximation of f by N elements of the dictionary is equivalent to the support minimization problem

$$\min\{\|(c_g)\|_{\ell^0} \; ; \; \|f - \sum c_g g\| \leq \varepsilon\}$$

which is usually untractable. A convex relaxation to this problem is

$$\min\{\|(c_g)\|_{\ell^1} \; ; \; \|f - \sum c_g g\| \leq \varepsilon\}$$

which can also be formulated as

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

Intuition : the ℓ^1 norm promotes sparse solutions and λ governs the amount of sparsity of the minimizing sequence.

Basis pursuit

Finding the best approximation of f by N elements of the dictionary is equivalent to the support minimization problem

$$\min\{\|(c_g)\|_{\ell^0} \; ; \; \|f - \sum c_g g\| \leq \varepsilon\}$$

which is usually untractable. A convex relaxation to this problem is

$$\min\{\|(c_g)\|_{\ell^1} \; ; \; \|f - \sum c_g g\| \leq \varepsilon\}$$

which can also be formulated as

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

Intuition : the ℓ^1 norm promotes sparse solutions and λ governs the amount of sparsity of the minimizing sequence. Many algorithms have been developped for solving this problem, also introduced in the statistical context as LASSO (Tibshirani).

The case when $\mathcal{D}$ is an orthonormal basis

The target function has a unique expansion $f = \sum c_g(f)g$, with $c_g(f) = \langle f, g \rangle$.

OMP $\Leftrightarrow$ hard thresholding: $f_N = \sum_{g \in E_N} c_g(f)g$, with $E_N = E_N(f)$ corresponding to the N largest $|c_g(f)|$.

The case when $\mathcal{D}$ is an orthonormal basis

The target function has a unique expansion $f = \sum c_g(f)g$, with $c_g(f) = \langle f, g \rangle$.

OMP $\Leftrightarrow$ hard thresholding: $f_N = \sum_{g \in E_N} c_g(f)g$, with $E_N = E_N(f)$ corresponding to the N largest $|c_g(f)|$.

BP $\Leftrightarrow$ soft thresholding: $\min \sum_{g \in \mathcal{D}} \{|c_g(f) - c_g|^2 + \lambda |c_g|\}$ attained by $c_g^* = S_{\lambda/2}(c_g(f))$ with $S_t(x) := \mathrm{sgn}(x)(|x| - t)_+$.
The solution $f_\lambda = \sum c_g^* g$ is N-sparse with $N = N(\lambda)$.

The case when $\mathcal{D}$ is an orthonormal basis

The target function has a unique expansion $f = \sum c_g(f)g$, with $c_g(f) = \langle f, g\rangle$.

OMP $\Leftrightarrow$ hard thresholding: $f_N = \sum_{g\in E_N} c_g(f)g$, with $E_N = E_N(f)$ corresponding to the N largest $|c_g(f)|$.

BP $\Leftrightarrow$ soft thresholding: $\min \sum_{g\in\mathcal{D}} \{|c_g(f) - c_g|^2 + \lambda|c_g|\}$ attained by $c_g^* = S_{\lambda/2}(c_g(f))$ with $S_t(x) := \text{sgn}(x)(|x| - t)_+$.
The solution $f_\lambda = \sum c_g^* g$ is N-sparse with $N = N(\lambda)$.

Rate of convergence: related to the concentration properties of the sequence $(c_g(f))$. For $p < 2$ and $s = \frac{1}{p} - \frac{1}{2}$, equivalence of

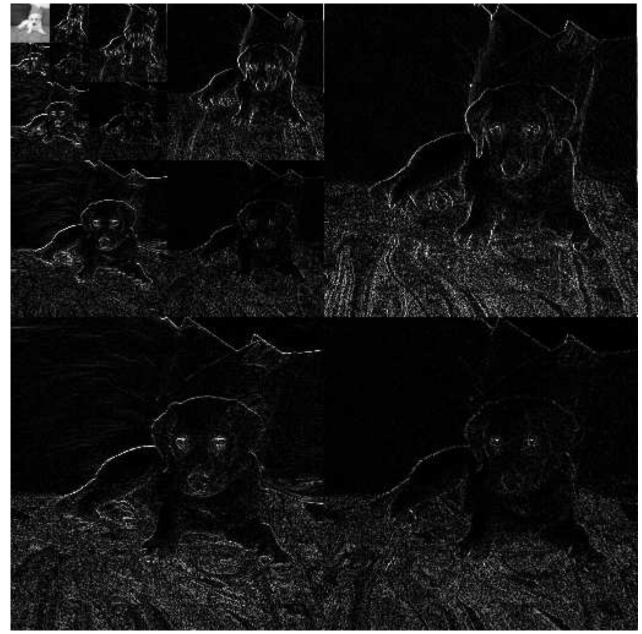
(i) $(c_g(f)) \in w\ell^p$ i.e. $\#(\{g \text{ s.t. } |c_g(f)| > \eta\}) \leq C\eta^{-p}$.

(ii) $c_n \leq Cn^{-\frac{1}{p}}$ with $(c_n)_{n\geq 0}$ decreasing permutation of $(|c_g(f)|)$.

(iii) $\|f - f_N\| = [\sum_{n\geq N} c_n^2]^{\frac{1}{2}} \leq C[\sum_{n\geq N} n^{-\frac{2}{p}}]^{\frac{1}{2}} \leq CN^{-s}$.

Example of concentrated representation

Digital image 512x512



Multiscale decomposition

Multiscale decompositions of natural images into wavelet bases are quasi-sparse: a few numerically significant coefficients concentrate most of the energy and information. This property plays a crucial role in applications such as compression and denoising.

Moving away from orthonormal bases

Many possible intermediate assumptions : measure of uncoherence (Temlyakov), restricted isometry properties (Candes-Tao), frames...

Moving away from orthonormal bases

Many possible intermediate assumptions : measure of uncoherence (Temlyakov), restricted isometry properties (Candes-Tao), frames...

Here we rather want to allow for the most general $\mathcal{D}$.

Signal and image processing (Mallat): composite features might be better captured by dictionary $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_k$ where the $\mathcal{D}_i$ are orthonormal bases. Example: wavelets + Fourier + Diracs...

Statistical learning: (x, y) random variables, observe independent samples $(x_i, y_i)_{i=1,\cdots,n}$ and look for f such that $|f(x) - y|$ is small in some probabilistic sense. Least square method are based on minimizing $\sum_i |f(x_i) - y_i|^2$ and therefore working with the norm $\|u\|^2 := \sum_i |u(x_i)|^2$ for which $\mathcal{D}$ is in general non-orthogonal.

Convergence analysis

Natural question: how do OMP and BP behave when f has an expansion $f = \sum c_g g$ with (c_g) a concentrated sequence ?

Convergence analysis

Natural question: how do OMP and BP behave when f has an expansion $f = \sum c_g g$ with (c_g) a concentrated sequence ?

Convergence analysis

Natural question: how do OMP and BP behave when f has an expansion $f = \sum c_g g$ with (c_g) a concentrated sequence ?

Here concentration cannot be measured by $(c_g) \in w\ell^p$ for some $p < 2$ since $(c_g) \in \ell^p$ does not ensure convergence of $\sum c_g g$ in H

Convergence analysis

Natural question: how do OMP and BP behave when f has an expansion $f = \sum c_g g$ with (c_g) a concentrated sequence ?

Here concentration cannot be measured by $(c_g) \in w\ell^p$ for some $p < 2$ since $(c_g) \in \ell^p$ does not ensure convergence of $\sum c_g g$ in H

...except when $p = 1$: convergence is then ensured by triangle inequality since $\|g\| = 1$ for all $g \in \mathcal{D}$.

Convergence analysis

Natural question: how do OMP and BP behave when f has an expansion $f = \sum c_g g$ with (c_g) a concentrated sequence ?

Here concentration cannot be measured by $(c_g) \in w\ell^p$ for some $p < 2$ since $(c_g) \in \ell^p$ does not ensure convergence of $\sum c_g g$ in H

...except when $p = 1$: convergence is then ensured by triangle inequality since $\|g\| = 1$ for all $g \in \mathcal{D}$.

We define the space $\mathcal{L}^1 \subset H$ of all f having a summable expansion, equiped with the norm

$$\|f\|_{\mathcal{L}^1} := \inf\{\sum |c_g| \, ; \, \sum c_g g = f\}.$$

Note that $\|f\| \leq \|f\|_{\mathcal{L}^1}$.

The case of summable expansions

Theorem (DeVore and Temlyakov, Jones, Maurey): For OMP and RGA with $\alpha_k = (1 - c/k)_+$, there exists $C > 0$ such that for all $f \in \mathcal{L}^1$, we have

$$\|f - f_N\| \le C \|f\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

The case of summable expansions

Theorem (DeVore and Temlyakov, Jones, Maurey): For OMP and RGA with $\alpha_k = (1 - c/k)_+$, there exists $C > 0$ such that for all $f \in \mathcal{L}^1$, we have

$$\|f - f_N\| \leq C\|f\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

Remark 1: the rate $s = 1/2$ is consistent with $s = 1/p - 1/2$ in the case of an orthonormal basis

The case of summable expansions

Theorem (DeVore and Temlyakov, Jones, Maurey): For OMP and RGA with $\alpha_k = (1 - c/k)_+$, there exists $C > 0$ such that for all $f \in \mathcal{L}^1$, we have

$$\|f - f_N\| \le C\|f\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

Remark 1: the rate $s = 1/2$ is consistent with $s = 1/p - 1/2$ in the case of an orthonormal basis

Remark 2: the PGA does not give the optimal rate. It is known that

(i) If $f \in \mathcal{L}^1$, then $\|f - f_N\| \le CN^{-\frac{11}{62}}$ (Konyagin and Temlyakov).

(ii) $\|f - f_N\| \ge cN^{-0.27}$ for some $\mathcal{D}$ and $f \in \mathcal{L}^1$ (Lifchitz and Temlyakov).

Proof for OMP

Since $r_k = f - f_k = f - P_k f$ is the error of orthogonal projection onto Span$\{g_1, \cdots, g_k\}$, we have

$$\|r_k\|^2 \leq \|r_{k-1}\|^2 - |\langle r_{k-1}, g_k \rangle|^2,$$

and

$$\|r_{k-1}\|^2 = \langle r_{k-1}, f \rangle = \sum c_g \langle r_{k-1}, g \rangle \leq \|f\|_{\mathcal{L}^1} |\langle r_{k-1}, g_k \rangle|.$$

Therefore, with the notation $M = \|f\|_{\mathcal{L}^1}^2$ and $a_k = \|r_k\|^2$, we obtain that

$$a_k \leq a_{k-1}(1 - \frac{a_{k-1}}{M}),$$

and $a_0 \leq M$, since $a_0 = \|r_0\|^2 = \|f\|^2 \leq \|f\|_{\mathcal{L}^1}^2$.

This easily implies by induction that $a_N \leq \frac{M}{N+1}$.

The case of a general $f \in \mathcal{H}$

Theorem (Barron, Cohen, Dahmen, DeVore, 2007): for the OMP, there exists $C > 0$ such that for any $f \in H$ and for any $h \in \mathcal{L}^1$, we have

$$\|f - f_N\| \leq \|f - h\| + C\|h\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

The case of a general $f \in \mathcal{H}$

Theorem (Barron, Cohen, Dahmen, DeVore, 2007): for the OMP, there exists $C > 0$ such that for any $f \in H$ and for any $h \in \mathcal{L}^1$, we have

$$\|f - f_N\| \leq \|f - h\| + C\|h\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

Remark: this shows that the convergence behaviour of OMP has some stability, although $f \mapsto f_N$ is intrinsically unstable with respect to a perturbation of f.

The case of a general $f \in \mathcal{H}$

Theorem (Barron, Cohen, Dahmen, DeVore, 2007): for the OMP, there exists $C > 0$ such that for any $f \in H$ and for any $h \in \mathcal{L}^1$, we have

$$\|f - f_N\| \leq \|f - h\| + C\|h\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

Remark: this shows that the convergence behaviour of OMP has some stability, although $f \mapsto f_N$ is intrinsically unstable with respect to a perturbation of f.

A first consequence : for any $f \in \mathcal{H}$ one has $\lim_{N \to +\infty} f_N = f$.

The case of a general $f \in \mathcal{H}$

Theorem (Barron, Cohen, Dahmen, DeVore, 2007): for the OMP, there exists $C > 0$ such that for any $f \in H$ and for any $h \in \mathcal{L}^1$, we have

$$\|f - f_N\| \leq \|f - h\| + C\|h\|_{\mathcal{L}^1} N^{-\frac{1}{2}}.$$

Remark: this shows that the convergence behaviour of OMP has some stability, although $f \mapsto f_N$ is intrinsically unstable with respect to a perturbation of f.

A first consequence : for any $f \in \mathcal{H}$ one has $\lim_{N \to +\infty} f_N = f$.

Also leads to characterization of intermediate rates of convergence N^{-s} with $0 < s < 1/2$, by interpolation spaces between $\mathcal{L}^1$ and H.

What about Basis Pursuit ?

For a general $\mathcal{D}$, consider $f_\lambda = \sum c_g^* \, g$ solution of

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

What about Basis Pursuit ?

For a general $\mathcal{D}$, consider $f_\lambda = \sum c_g^* \, g$ solution of

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

Questions: Does the minimizer always exist ? If it exists is it unique ? If it exists is (c_g^*) finitely supported ?

What about Basis Pursuit ?

For a general $\mathcal{D}$, consider $f_\lambda = \sum c_g^* \, g$ solution of

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

Questions: Does the minimizer always exist ? If it exists is it unique ? If it exists is (c_g^*) finitely supported ?

Answers: NO, NO and NO.

What about Basis Pursuit ?

For a general $\mathcal{D}$, consider $f_\lambda = \sum c_g^* \, g$ solution of

$$\min\{\|f - \sum c_g g\|^2 + \lambda \|(c_g)\|_{\ell^1}\}$$

Questions: Does the minimizer always exist ? If it exists is it unique ? If it exists is (c_g^*) finitely supported ?

Answers: NO, NO and NO.

Finite dimensional case: solution exists if and only if $-\mathcal{D} \cup \mathcal{D}$ is a closed subset of the unit sphere.

Infinite dimensional case: a sufficient condition for existence of a minimizer with (c_g^*) finitely supported is : for all $\varepsilon > 0$ and $f \in H$

$$\#\{g \; ; \; |\langle f, g\rangle| > \varepsilon\} \leq C(\varepsilon) < \infty.$$

Bessel sequences

$\mathcal{D}$ is a Bessel sequence. if for all $(c_g) \in \ell^2$,

$$\|\sum c_g g\|^2 \leq C \sum |c_g|^2,$$

or equivalently for all $f \in H$

$$\sum |\langle f, g\rangle|^2 \leq C\|f\|^2,$$

i.e. the frame operator is bounded.

Bessel sequences

$\mathcal{D}$ is a Bessel sequence. if for all $(c_g) \in \ell^2$,

$$\|\sum c_g g\|^2 \leq C \sum |c_g|^2,$$

or equivalently for all $f \in H$

$$\sum |\langle f, g \rangle|^2 \leq C\|f\|^2,$$

i.e. the frame operator is bounded.

This assumption was considered by Gribonval and Nielsen (2004). It is also the framework considered by Daubechies, Defrise and Demol (2005), when minimizing $\|y - Kx\|_{\ell^2}^2 + \lambda\|x\|_{\ell^1}$ with $\|K\| \leq C$.

Under such an assumption, the minimizer $f_\lambda = \sum c_g^* g$ exists with (c_g^*) finitely supported. Uniqueness might fail.

Approximation properties of BP

Theorem (Cohen, DeMol, Gribonval, 2007): If $\mathcal{D}$ is a Bessel sequence, then

(i) for all $f \in \mathcal{L}^1$,

$$\|f - f_\lambda\| \le C\|f\|_{\mathcal{L}^1} N^{-1/2},$$

with $N = N(\lambda)$ the support size of (c_g^*).

(ii) for all $f \in H$ and $h \in \mathcal{L}^1$,

$$\|f - f_\lambda\| \le C(\|f - h\| + \|f\|_{\mathcal{L}^1} N^{-1/2}).$$

The constant C depends on the constant in the Bessel assumption.

Open question: does this result hold for more general dictionaries ?

Proof ($f \in \mathcal{L}^1$)

Optimality condition gives:

$$c_g^* \neq 0 \Rightarrow |\langle f - f_\lambda, g\rangle| = \frac{\lambda}{2}.$$

Therefore

$$N\frac{\lambda^2}{4} = \sum |\langle f - f_\lambda, g\rangle|^2 \leq C\|f - f_\lambda\|^2,$$

so that $\lambda \leq 2C\|f - f_\lambda\|N^{-1/2}$.

Now if $f = \sum c_g g$ with $\sum |c_g| \leq M$, we have

$$\|f - f_\lambda\|^2 + \lambda\|(c_g^*)\|_{\ell^1} \leq \lambda M,$$

and therefore $\|f - f_\lambda\|^2 \leq \lambda M$.

Combining both gives

$$\|f - f_\lambda\| \leq 2C\|f\|_{\mathcal{L}^1} N^{-1/2}.$$

A result in the regression context for OMP

Let $\hat{f}_k$ be the estimator obtained by running OMP on the data (y_i) with the least square norm $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

A result in the regression context for OMP

Let $\hat{f}_k$ be the estimator obtained by running OMP on the data (y_i) with the least square norm $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Selecting k: one sets

$$k^* := \mathrm{Argmax}_{k>0}\{\|y - \hat{f}_k\|_n^2 + \mathrm{pen}(k,n)\},$$

with $\mathrm{pen}(k,n) \sim \frac{k \log n}{n}$, and defines $\hat{f} = \hat{f}_{k^*}$

A result in the regression context for OMP

Let $\hat{f}_k$ be the estimator obtained by running OMP on the data (y_i) with the least square norm $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Selecting k: one sets

$$k^* := \text{Argmax}_{k>0}\{\|y - \hat{f}_k\|_n^2 + \text{pen}(k,n)\},$$

with $\text{pen}(k,n) \sim \frac{k \log n}{n}$, and defines $\hat{f} = \hat{f}_{k^*}$

We denote by $f(x) := E(y|x)$ the regression function and $\|u\|^2 := E(|u(x)|^2)$.

Theorem (Barron, Cohen, Dahmen, DeVore, 2007):

$$E(\|\hat{f} - f\|^2) \leq C \inf_{k \geq 0,\ h \in \mathcal{L}^1} \{\|h - f\|^2 + k^{-1}\|h\|_{\mathcal{L}^1}^2 + \text{pen}(k,n)\}.$$

A result in the regression context for OMP

Let $\hat{f}_k$ be the estimator obtained by running OMP on the data (y_i) with the least square norm $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Selecting k: one sets

$$k^* := \text{Argmax}_{k>0}\{\|y - \hat{f}_k\|_n^2 + \text{pen}(k,n)\},$$

with $\text{pen}(k,n) \sim \frac{k \log n}{n}$, and defines $\hat{f} = \hat{f}_{k^*}$

We denote by $f(x) := E(y|x)$ the regression function and $\|u\|^2 := E(|u(x)|^2)$.

Theorem (Barron, Cohen, Dahmen, DeVore, 2007):

$$E(\|\hat{f} - f\|^2) \leq C \inf_{k\geq 0,\ h\in\mathcal{L}^1}\{\|h - f\|^2 + k^{-1}\|h\|_{\mathcal{L}^1}^2 + \text{pen}(k,n)\}.$$

Similar oracle estimate for LASSO ? Which assumptions should be needed on $\mathcal{D}$?