Algorithmes greedy pour l'approximation et l'apprentissage

Albert Cohen

Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
Paris

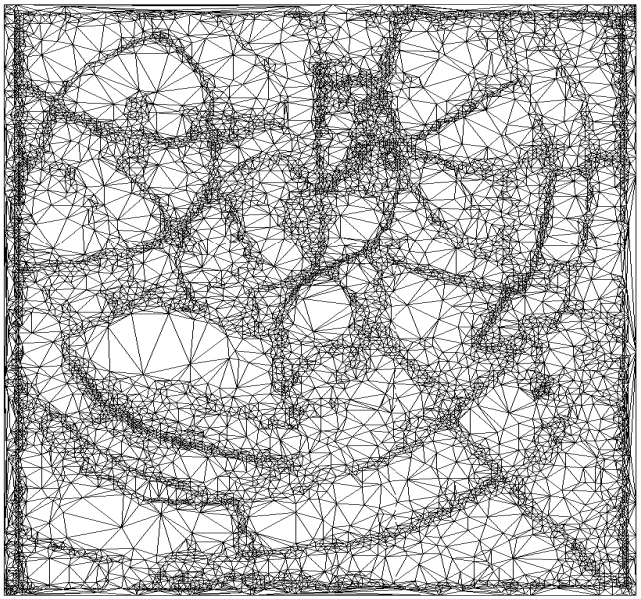
Collaborateurs: Andrew Barron, Wolfgang Dahmen,
Ron DeVore, Nira Dyn, Frédéric Hecht et Jean-Marie Mirebeau

College de France, 30-11-07

Un problème d'approximation

But: étant donné une fonction f et $N > 0$, construire une triangulation $\mathcal{T}$ avec N points qui minimise la distance (L^2) entre f et l'espace d'éléments finis (affines) sur $\mathcal{T}$.

Image numérique

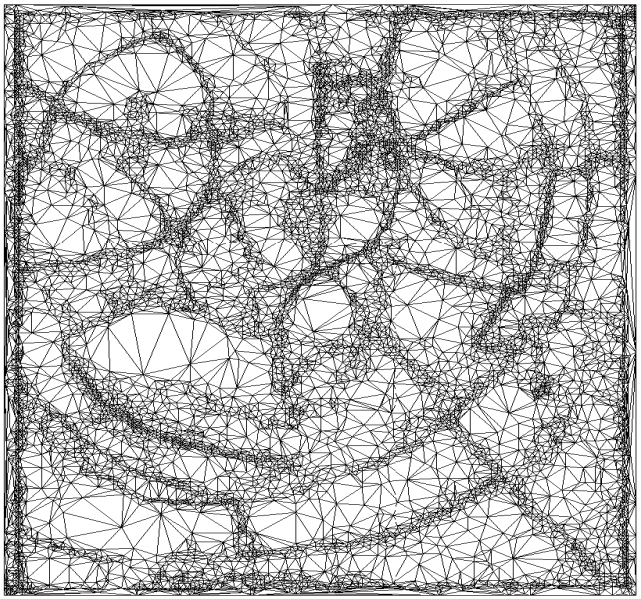


Triangulation (Pascal Frey)

Un problème d'approximation

But: étant donné une fonction f et $N > 0$, construire une triangulation $\mathcal{T}$ avec N points qui minimise la distance (L^2) entre f et l'espace d'éléments finis (affines) sur $\mathcal{T}$. problème NP complet.

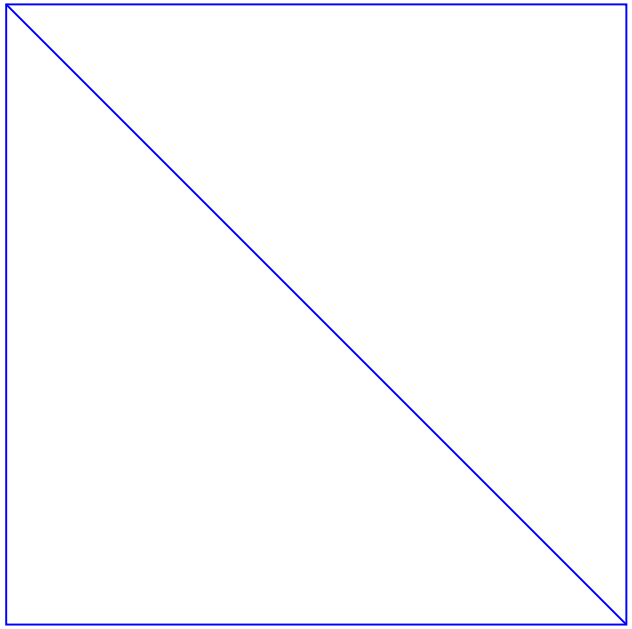
Image numérique



Triangulation (Pascal Frey)

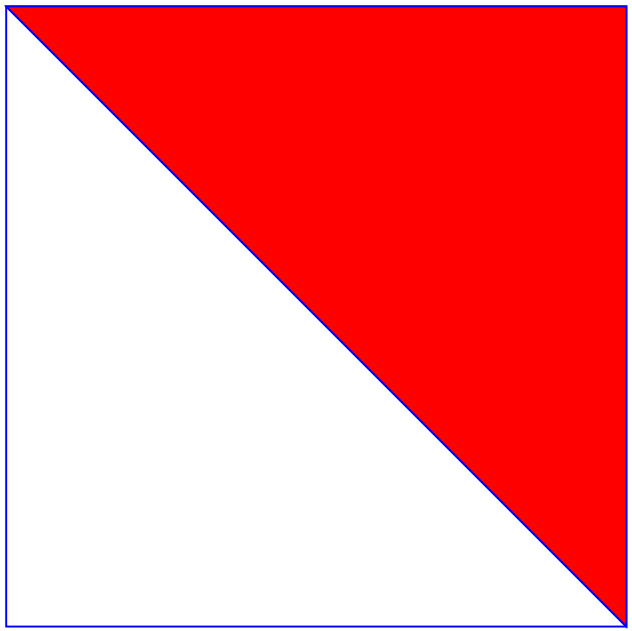
Triangulation optimale: raffinement anisotrope près des contours.

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



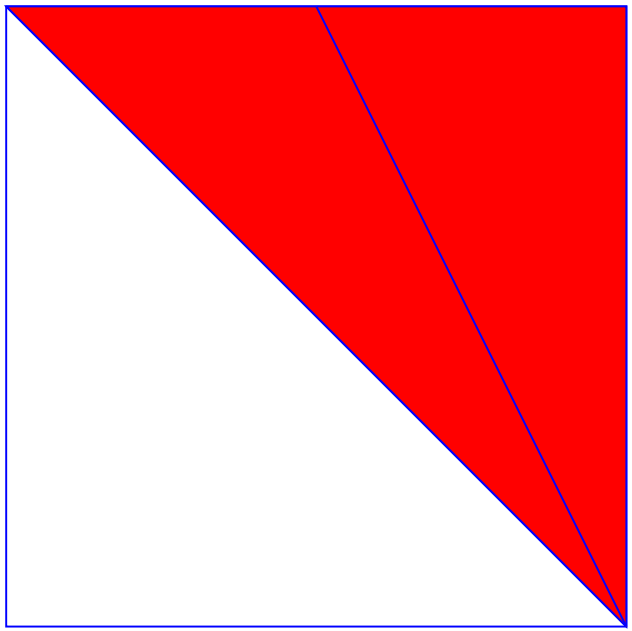
Triangulation grossière

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



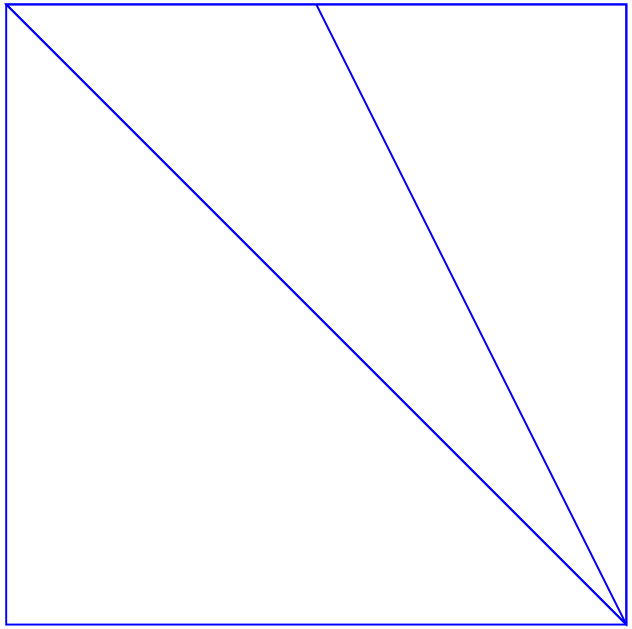
Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



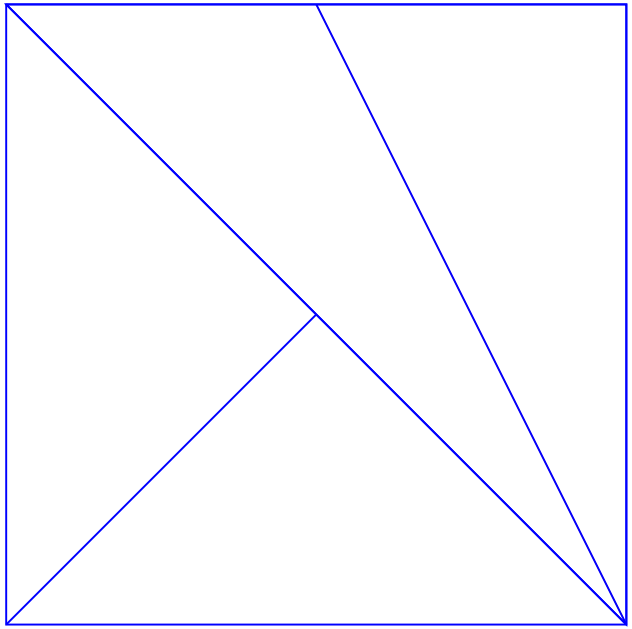
Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2 $\Rightarrow$ choix de la bisection qui réduit au mieux l'erreur

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



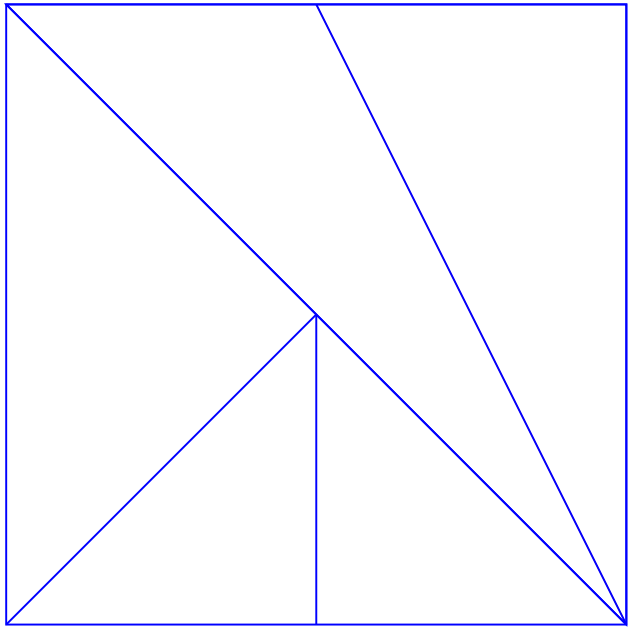
Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2 $\Rightarrow$ choix de la bisection qui réduit au mieux l'erreur $\Rightarrow$ découpage

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



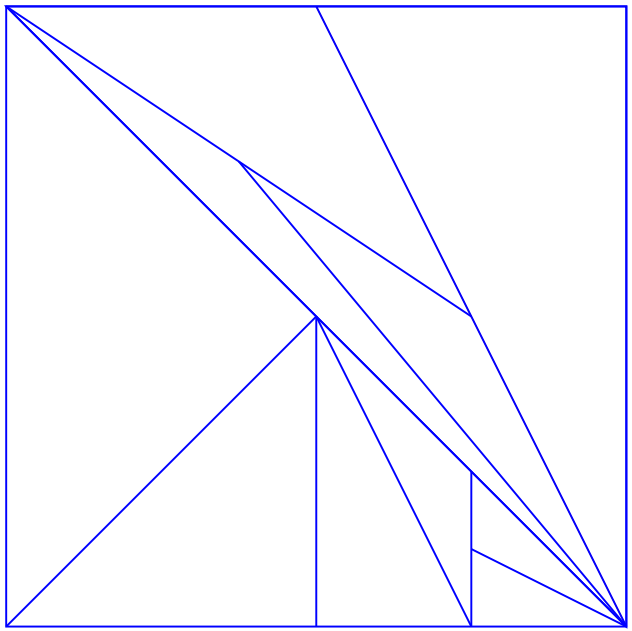
Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2 $\Rightarrow$ choix de la bisection qui réduit au mieux l'erreur $\Rightarrow$ découpage $\Rightarrow$ itération ...

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)



Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2 $\Rightarrow$ choix de la bisection qui réduit au mieux l'erreur $\Rightarrow$ découpage $\Rightarrow$ itération ...

Algorithme greedy: bisection adaptative (Dyn, Hecht, AC)

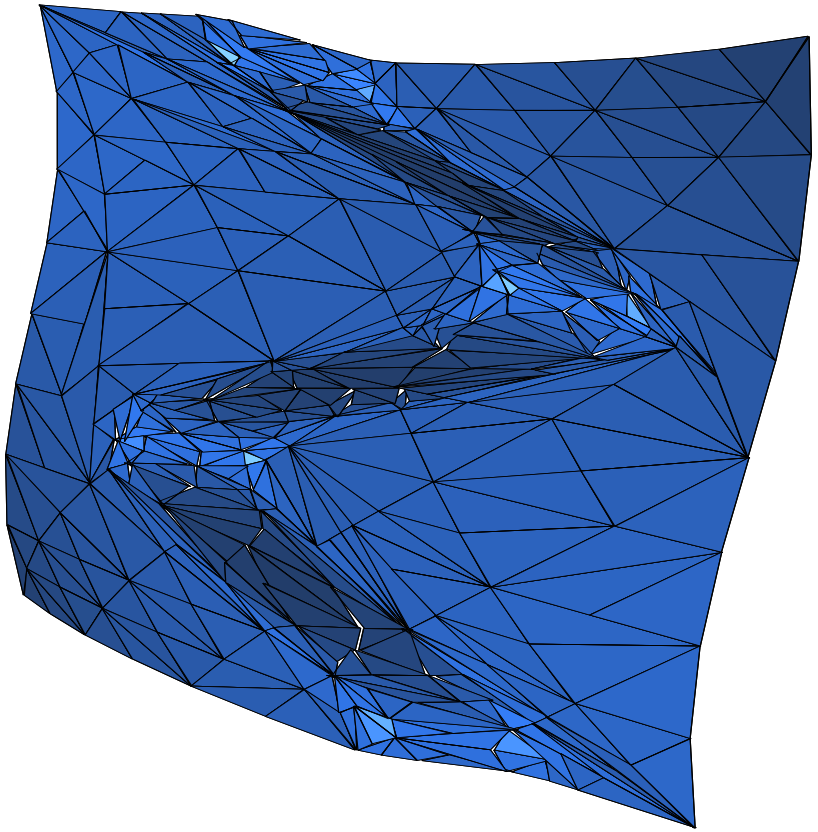


Triangulation grossière $\Rightarrow$ triangle maximisant l'erreur locale L^2 $\Rightarrow$ choix de la bisection qui réduit au mieux l'erreur $\Rightarrow$ découpage $\Rightarrow$ itération ....

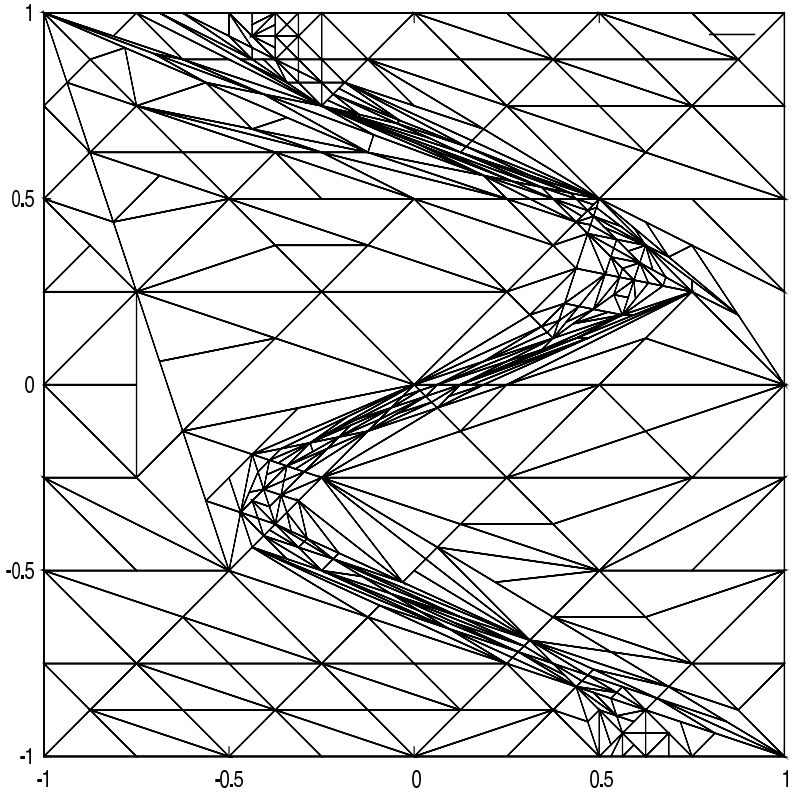
... jusqu'à atteindre une précision ou nombre de triangle prescrit.

L'agorithme génere des triangles anisotropes

Exemple: zone de transition le long d'une courbe.



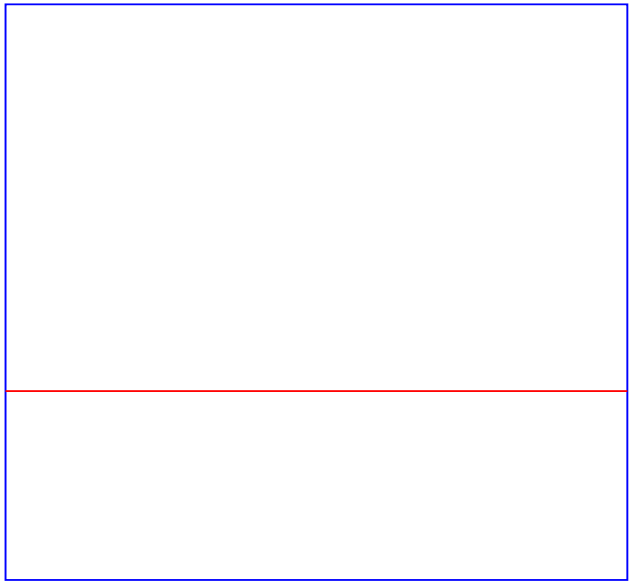
Approximation



Triangulation

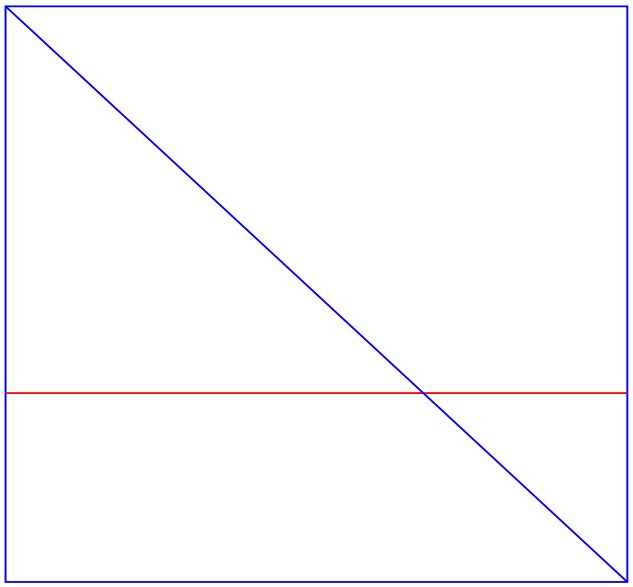
Le cas d'une fonction "step"

On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Le cas d'une fonction "step"

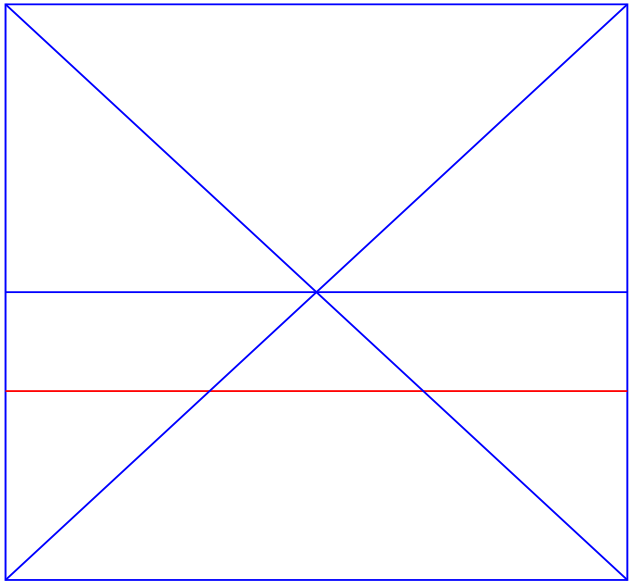
On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1$

Le cas d'une fonction "step"

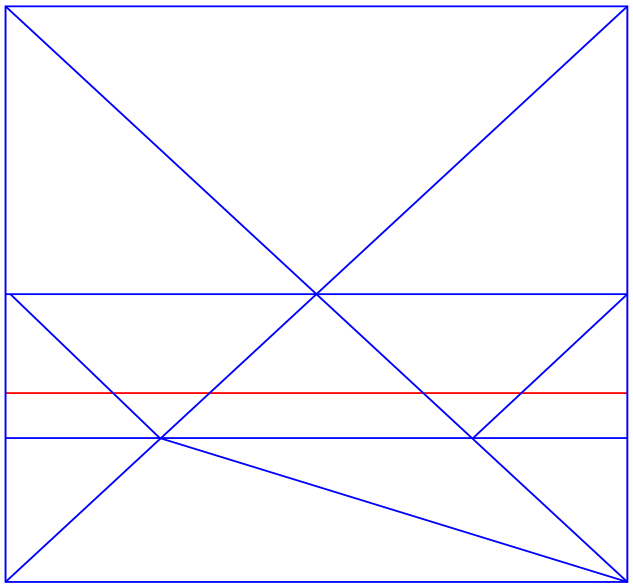
On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1, \frac{1}{2}$

Le cas d'une fonction "step"

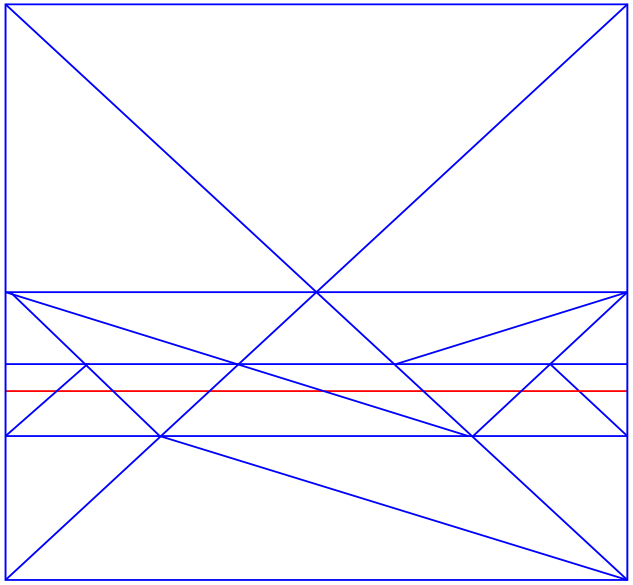
On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1, \frac{1}{2}, \frac{1}{4}$

Le cas d'une fonction "step"

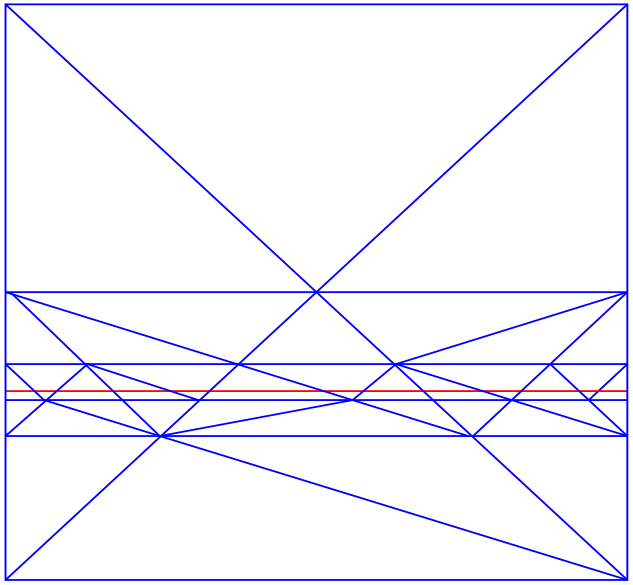
On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$

Le cas d'une fonction "step"

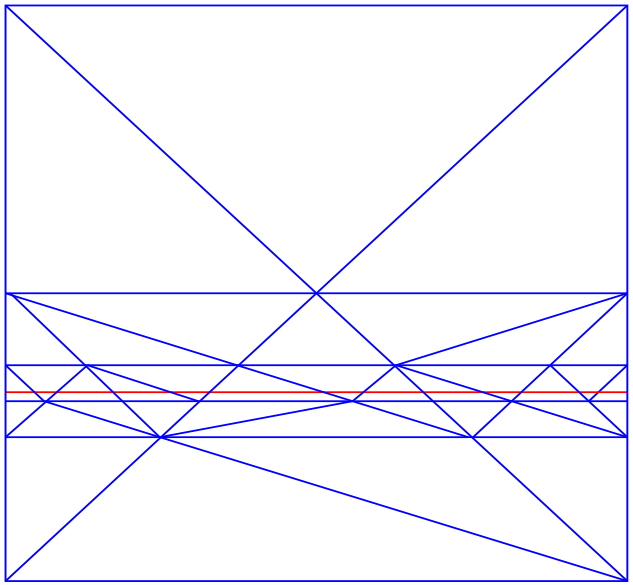
On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16} \cdots$

Le cas d'une fonction "step"

On prend ici $f(x,y) = \chi_{\{y>1/3\}}$ sur $[0,1]^2$.



Découpage supervisé : approcher $y = 1/3$ à la résolution $\Delta y = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16} \cdots$

Nombre de triangles $N = N(j)$ générés à la résolution 2^{-j} :

$$N(j) = N(j-1) + N(j-2) \sim \mathrm{G}^j$$

Vitesse de convergence

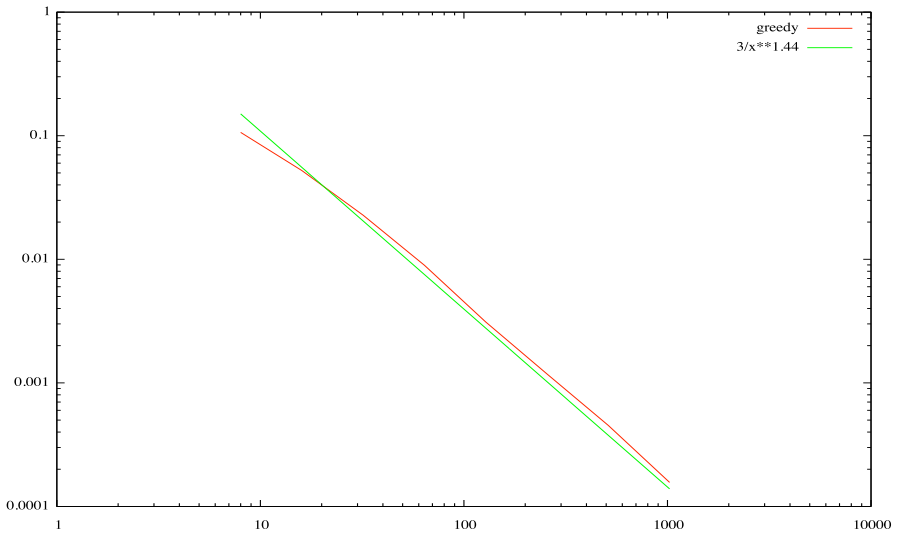
Découpage supervisé : à la résolution 2^{-j}, erreur $\|f - f_N\|_{L^2}^2$ au mieux controllée par

$$E \leq 2^{-j} \leq CN^{-r}, \quad r = \frac{\log 2}{\log G} \approx 1.44$$

Vitesse de convergence

Découpage supervisé : à la résolution 2^{-j}, erreur $\|f - f_N\|_{L^2}^2$ au mieux controllée par

$$E \leq 2^{-j} \leq CN^{-r}, \quad r = \frac{\log 2}{\log G} \approx 1.44$$



Même vitesse pour l'algorithme greedy. résultats plus généraux ?

Comportement local de l'algorithme

Deux résultats de Jean-Marie Mirebeau (2007) :

Théorème 1 (comportement au voisinage d'un contour): si $f = \chi_{\{ax+by \geq c\}}$ avec (a, b, c) arbitraire, l'erreur de l'algorithme greedy vérifie toujours

$$E \leq CN^{-r}, \quad r = \frac{\log 2}{\log G} \approx 1.44$$

Théorème 2 (comportement dans une zone régulière): si f est quadratique de hessienne $D^2 f$ positive, la proportion $\rho(N) := N_{\text{opt}}/N$ de triangles d'aspect optimal, i.e. tel que

$$\{D^2 f(z - x_T, z - x_T) \leq c h_T^2\} \subset T \subset \{D^2 f(z - x_T, z - x_T) \leq C h_T^2\}$$

avec $x_T = \text{bar}(T)$ et $h_T = \text{diam}_H(T)$ vérifie

$$\rho(N) \geq 1 - N^{-\delta}$$

Un problème d'approximation

$\mathcal{D} = (g)_{g \in \mathcal{D}}$ dictionnaire de fonctions dans un espace de Hilbert H.

Un problème d'approximation

$\mathcal{D} = (g)_{g \in \mathcal{D}}$ dictionnaire de fonctions dans un espace de Hilbert H.

Dictionnaire normalisé ($\|g\| = 1$), complet, mais pas nécessairement une base orthonormale.

Un problème d'approximation

$\mathcal{D} = (g)_{g \in \mathcal{D}}$ dictionnaire de fonctions dans un espace de Hilbert H.

Dictionnaire normalisé ($\|g\| = 1$), complet, mais pas nécessairement une base orthonormale.

Problème : étant donné $N > 0$ et $f \in H$, trouver une combinaison de N termes

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

avec $g_k \in \mathcal{D}$, qui approche au mieux f.

Un problème d'approximation

$\mathcal{D} = (g)_{g \in \mathcal{D}}$ dictionnaire de fonctions dans un espace de Hilbert H.

Dictionnaire normalisé ($\|g\| = 1$), complet, mais pas nécessairement une base orthonormale.

Problème : étant donné $N > 0$ et $f \in H$, trouver une combinaison de N termes

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

avec $g_k \in \mathcal{D}$, qui approche au mieux f.

Adaptatif : les fonctions choisies $\{g_1, \cdots, g_N\}$ dépendent de f.

Un problème d'approximation

$\mathcal{D} = (g)_{g \in \mathcal{D}}$ dictionnaire de fonctions dans un espace de Hilbert H.

Dictionnaire normalisé ($\|g\| = 1$), complet, mais pas nécessairement une base orthonormale.

Problème : étant donné $N > 0$ et $f \in H$, trouver une combinaison de N termes

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

avec $g_k \in \mathcal{D}$, qui approche au mieux f.

Adaptatif : les fonctions choisies $\{g_1, \cdots, g_N\}$ dépendent de f.

On s'intéresse aux vitesse de décroissance de $\|f - f_N\|$ quand $N \to +\infty$ et à des algorithmes simples et rapides pour construire f_N.

Le cas d'une base orthonormale

Algorithme : (i) Calculer la décomposition $f = \sum c_g g$, avec $c_g = \langle f, g \rangle$.

(ii) On prend $f_N = \sum_{g \in E_N} c_g g$, avec $E_N = E_N(f)$ les indices des N plus grands $|c_g|$.

Le cas d'une base orthonormale

Algorithme : (i) Calculer la décomposition $f = \sum c_g g$, avec $c_g = \langle f, g \rangle$.

(ii) On prend $f_N = \sum_{g \in E_N} c_g g$, avec $E_N = E_N(f)$ les indices des N plus grands $|c_g|$.

Vitesse de convergence : liées aux propriétés de concentrations de la suite (c_g). Pour $p < 2$ et $s = \frac{1}{p} - \frac{1}{2}$, propriétés équivalentes :

(i) $(c_g) \in w\ell^p$, i.e. $c_n \leq C n^{-\frac{1}{p}}$ avec $(c_n)_{n \geq 0}$ le réarrangement décroissant de $(|c_g|)$.

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

Le cas d'une base orthonormale

Algorithme : (i) Calculer la décomposition $f = \sum c_g g$, avec $c_g = \langle f, g \rangle$.

(ii) On prend $f_N = \sum_{g \in E_N} c_g g$, avec $E_N = E_N(f)$ les indices des N plus grands $|c_g|$.

Vitesse de convergence : liées aux propriétés de concentrations de la suite (c_g). Pour $p < 2$ et $s = \frac{1}{p} - \frac{1}{2}$, propriétés équivalentes :

(i) $(c_g) \in w\ell^p$, i.e. $c_n \leq C n^{-\frac{1}{p}}$ avec $(c_n)_{n \geq 0}$ le réarrangement décroissant de $(|c_g|)$.

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

Remarque : Les conditions sur f qui assurent $(c_g) \in w\ell^p$ dépendent du dictionnaire. Exemple: si $\mathcal{D}$ est une base d'ondelette, f appartient à un espace de Besov.

Représentations des images dans des bases d'ondelettes

Image digitale 512x512

Décomposition multiéchelle

Les représentations multiéchelles des images naturelles sont parcimonieuses : un petit nombre de coefficients numériquement significatifs concentrent l'essentiel de l'énergie et de l'information.

Et si $\mathcal{D}$ est un dictionnaire non-orthogonal et redondant ?

1. Traitement du signal et de l'image : aspects composites mieux capturés par $\mathcal{D} = \cup \mathcal{D}_i$ où $\mathcal{D}_i$ sont des bases différentes. Exemple: Ondelettes + Fourier + Diracs...

Et si $\mathcal{D}$ est un dictionnaire non-orthogonal et redondant ?

1. Traitement du signal et de l'image : aspects composites mieux capturés par $\mathcal{D} = \cup \mathcal{D}_i$ où $\mathcal{D}_i$ sont des bases différentes. Exemple: Ondelettes + Fourier + Diracs...

2. Régression : on observe $(x_i, y_i)_{i=1,\cdots,n}$ tirages indépendants de variables (x, y) et on cherche f tel que $|f(x) - y|$ est petit en un sens probabiliste. Moindres carrés : on minimise $\sum_i |f(x_i) - y_i|^2$ et on travaille donc avec la norme $\|u\|_n^2 := \frac{1}{n} \sum_{i=1}^n |u(x_i)|^2$ pour laquelle $\mathcal{D}$ n'est pas en général orthogonal.

Et si $\mathcal{D}$ est un dictionnaire non-orthogonal et redondant ?

1. Traitement du signal et de l'image : aspects composites mieux capturés par $\mathcal{D} = \cup \mathcal{D}_i$ où $\mathcal{D}_i$ sont des bases différentes. Exemple: Ondelettes + Fourier + Diracs...

2. Régression : on observe $(x_i, y_i)_{i=1,\cdots,n}$ tirages indépendants de variables (x, y) et on cherche f tel que $|f(x) - y|$ est petit en un sens probabiliste. Moindres carrés : on minimise $\sum_i |f(x_i) - y_i|^2$ et on travaille donc avec la norme $\|u\|_n^2 := \frac{1}{n} \sum_{i=1}^n |u(x_i)|^2$ pour laquelle $\mathcal{D}$ n'est pas en général orthogonal.

Algorithme incorrect : prendre $c_g := \langle f, g \rangle$ et $f_N = \sum_{g \in E_N} c_g g$, avec $E_N = E_N(f)$ l'ensemble des N plus grands $|c_g|$. En général f_N n'approche pas f.

Algorithme irréaliste : inspecter tous les ensembles $\{g_1, \cdots, g_N\}$ et minimiser l'erreur entre f et sa projection sur Vect$\{g_1, \cdots, g_N\}$. En général trop coûteux en calcul.

Orthogonal matching pursuit (OMP)

f_N construit itérativement. Initialisation: $f_0 = 0$.

Au pas $k-1$, l'approximation est

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

avec P_{k-1} la projection orthogonale sur $\mathrm{Vect}\{g_1, \cdots, g_{k-1}\}$.

Choix de l'élément suivant basé sur l'erreur $r_{k-1} = f - P_{k-1} f$:

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

Orthogonal matching pursuit (OMP)

f_N construit itérativement. Initialisation: $f_0 = 0$.

Au pas $k-1$, l'approximation est

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

avec P_{k-1} la projection orthogonale sur $\mathrm{Vect}\{g_1, \cdots, g_{k-1}\}$.

Choix de l'élément suivant basé sur l'erreur $r_{k-1} = f - P_{k-1} f$:

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

Nombreuses variantes : PGA, RGA, WGA, CGA, WCGA, WGAFR... (Survey par Vladimir Temlyakov dans Acta Numerica 2008).

Introduits dans les années 1970 en statistiques (Friedman, Huber, Stuetzle, Tukey...).

Algorithme greedy relaxé (RGA)

On pose

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

avec

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

Algorithme greedy relaxé (RGA)

On pose

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

avec

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

Version simple : α_k fixè à l'avance, on optimise β et g.

Choix $\alpha_k = 1$: algorithme greedy pur (PGA)

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

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

Algorithme greedy relaxé (RGA)

On pose

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

avec

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

Version simple : α_k fixè à l'avance, on optimise β et g.

Choix $\alpha_k = 1$: algorithme greedy pur (PGA)

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

avec $r_{k-1} = f - f_{k-1}$. On choisit plutôt $\alpha_k = (1 - \frac{c}{k})_+$:

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

avec résidu modifié $\tilde{r}_{k-1} := f - (1 - \frac{c}{k})_+ f_{k-1}$.

Analyse de convergence

Question : ces algorithmes convergent-ils rapidement vers f, si celle-ci admet une décomposition $f = \sum c_g g$ où (c_g) est bien concentrée ?

Analyse de convergence

Question : ces algorithmes convergent-ils rapidement vers f, si celle-ci admet une décomposition $f = \sum c_g g$ où (c_g) est bien concentrée ?

Par analogie avec une base orthonormée, on pourrait supposer $(c_g) \in w\ell^p$ pour un $p < 2$ et chercher à comprendre si $\|f - f_N\| \leq CN^{-s}$ avec $s = \frac{1}{p} - \frac{1}{2}$.

Problème : $\mathcal{D}$ n'est pas une base orthonormée, et la condition $(c_g) \in \ell^p$ n'assure pas en général la convergence de $\sum c_g g$ dans H.

Analyse de convergence

Question : ces algorithmes convergent-ils rapidement vers f, si celle-ci admet une décomposition $f = \sum c_g g$ où (c_g) est bien concentrée ?

Par analogie avec une base orthonormée, on pourrait supposer $(c_g) \in w\ell^p$ pour un $p < 2$ et chercher à comprendre si $\|f - f_N\| \leq CN^{-s}$ avec $s = \frac{1}{p} - \frac{1}{2}$.

Problème : $\mathcal{D}$ n'est pas une base orthonormée, et la condition $(c_g) \in \ell^p$ n'assure pas en général la convergence de $\sum c_g g$ dans H.

Sauf cas $p = 1$: convergence triviale par l'inégalité triangulaire puisque $\|g\| = 1$ pour tout $g \in \mathcal{D}$.

Cas des décompositions sommables

On définit l'espace $\mathcal{L}_1$ par

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

Injection continue et dense de $\mathcal{L}^1$ dans H avec $\|f\| \leq \|f\|_{\mathcal{L}_1}$.

Cas des décompositions sommables

On définit l'espace $\mathcal{L}_1$ par

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

Injection continue et dense de $\mathcal{L}^1$ dans H avec $\|f\| \leq \|f\|_{\mathcal{L}_1}$.

Théorème (Maurey 1982, Jones 1988, DeVore et Temlyakov 1998) : pour OMP et RGA avec $\alpha_k = (1 + \frac{c}{k})_+$, il existe $C > 0$ telle que pour tout $f \in \mathcal{L}^1$,

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

Exposant $s = \frac{1}{2}$ consistant avec le case d'une base orthonormale pour lequel $s = \frac{1}{p} - \frac{1}{2} = \frac{1}{2}$.

Preuve pour OMP

Le résidu $r_k = f - f_k = f - P_k f$ est l'erreur de projection orthogonale sur $\text{Vect}\{g_1, \cdots, g_k\}$. On a

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

et

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

Preuve pour OMP

Le résidu $r_k = f - f_k = f - P_k f$ est l'erreur de projection orthogonale sur $\mathrm{Vect}\{g_1, \cdots, g_k\}$. On a

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

et

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

En posant $M = \|f\|^2_{\mathcal{L}^1}$ et $a_k = \|r_k\|^2$, on obtient donc

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

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

Ceci entraine $a_N \leq \frac{M}{N+1}$.

Cas d'une fonction arbitraire $f \in H$

Théorème (Barron, Cohen, Dahmen, DeVore 2006) : pour OMP et RGA avec $\alpha_k = (1 + \frac{c}{k})_+$, il existe $C > 0$ telle que pour tout $f \in H$ et pour tout $h \in \mathcal{L}^1$, on a

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

Cas d'une fonction arbitraire $f \in H$

Théorème (Barron, Cohen, Dahmen, DeVore 2006) : pour OMP et RGA avec $\alpha_k = (1 + \frac{c}{k})_+$, il existe $C > 0$ telle que pour tout $f \in H$ et pour tout $h \in \mathcal{L}^1$, on a

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

Interpretation : "stabilité" de la convergence de l'algorithme greedy, bien que $f \mapsto f_N$ est instable par perturbation de f.

Cas d'une fonction arbitraire $f \in H$

Théorème (Barron, Cohen, Dahmen, DeVore 2006) : pour OMP et RGA avec $\alpha_k = (1 + \frac{c}{k})_+$, il existe $C > 0$ telle que pour tout $f \in H$ et pour tout $h \in \mathcal{L}^1$, on a

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

Interpretation : "stabilité" de la convergence de l'algorithme greedy, bien que $f \mapsto f_N$ est instable par perturbation de f.

Première conséquence : pour tout $f \in \mathcal{H}$, on a $\lim_{N \to +\infty} f_N = f$.

Permet aussi d'obtenir des vitesses de convergence N^{-s} avec $0 < s < 1/2$, pour des espaces intermédiaires entre $\mathcal{L}^1$ et H.

Vitesse de convergence

Puisque h est arbitraire, on a

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

Vitesse de convergence

Puisque h est arbitraire, on a

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

K-fonctionelle (Lions-Peetre): pour (X, Y) espaces de Banach avec $Y \subset X$, on pose pour $f \in X$ et $t \geq 0$

$$K(f, t) = K(f, t, X, Y) = \inf_{h \in Y} \{\|f - h\|_X + t\|h\|_Y\}.$$

Espace d'interpolation : pour $0 < \theta < 1$ la fonction f appartient à $[X, Y]_{\theta,\infty}$ si et seulement si $K(f, t) \leq Ct^\theta$.

Vitesse de convergence

Puisque h est arbitraire, on a

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

K-fonctionelle (Lions-Peetre): pour (X, Y) espaces de Banach avec $Y \subset X$, on pose pour $f \in X$ et $t \geq 0$

$$K(f, t) = K(f, t, X, Y) = \inf_{h \in Y} \{\|f - h\|_X + t\|h\|_Y\}.$$

Espace d'interpolation : pour $0 < \theta < 1$ la fonction f appartient à $[X, Y]_{\theta,\infty}$ si et seulement si $K(f, t) \leq Ct^\theta$.

Conséquence: si $f \in [H, \mathcal{L}^1]_{\theta,\infty}$, on a pour OMP et RGA avec $\alpha_k = (1 - c/k)_+$,

$$\|f - f_N\| \leq K(f, CN^{-\frac{1}{2}}) \leq CN^{-\theta/2}$$

Si $\mathcal{D}$ est une base orthonormale $H \sim \ell^2(\mathcal{D})$ et $\mathcal{L}^1 \sim \ell^1(\mathcal{D})$, et

$$[H, \mathcal{L}^1]_{\theta,\infty} \sim [\ell^2, \ell^1]_{\theta,\infty} = w\ell^p, \quad \frac{1}{p} = \frac{1-\theta}{2} + \theta = \frac{1+\theta}{2}.$$

On retrouve ainsi la vitesse optimale

$$\|f - f_N\| \leq CN^{-s}, \quad s = \frac{1}{p} - \frac{1}{2}.$$

Si $\mathcal{D}$ est une base orthonormale $H \sim \ell^2(\mathcal{D})$ et $\mathcal{L}^1 \sim \ell^1(\mathcal{D})$, et

$$[H, \mathcal{L}^1]_{\theta,\infty} \sim [\ell^2, \ell^1]_{\theta,\infty} = w\ell^p, \quad \frac{1}{p} = \frac{1-\theta}{2} + \theta = \frac{1+\theta}{2}.$$

On retrouve ainsi la vitesse optimale

$$\|f - f_N\| \leq CN^{-s}, \quad s = \frac{1}{p} - \frac{1}{2}.$$

Pour d'autres dictionnaires, on peut parfois caractériser l'espace $\mathcal{L}^1$ et les espaces intermédiaires $[H, \mathcal{L}^1]_\theta$. Exemple : fonctions ridges

$$\mathcal{D} := \{\varphi_{a,b}(x) = \varphi(a \cdot x + b), \quad a \in \mathbb{R}^{\mathrm{d}}, \ \|\mathrm{a}\| = 1, \ \mathrm{b} \in \mathbb{R}\},$$

avec φ fonction d'une variable du type $\varphi := \chi_{\{x>0\}}$. On a pour ce dictionnaire (Barron 1992) :

$$\int |\omega \hat{f}(\omega)| d\omega < +\infty \Rightarrow f \in \mathcal{L}^1.$$

Un résultat dans le cadre de la régression

On note $\hat{f}_k$ l'estimateur obtenu en appliquant OGA ou RGA pour approcher les données (y_i) pour la norme hilbertienne $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Un résultat dans le cadre de la régression

On note $\hat{f}_k$ l'estimateur obtenu en appliquant OGA ou RGA pour approcher les données (y_i) pour la norme hilbertienne $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Sélection de k : on pose

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

avec $\text{pen}(k,n) \sim \frac{k \log n}{n}$, et on définit $\hat{f} = \hat{f}_{k^*}$

Un résultat dans le cadre de la régression

On note $\hat{f}_k$ l'estimateur obtenu en appliquant OGA ou RGA pour approcher les données (y_i) pour la norme hilbertienne $\|u\|_n^2 := \frac{1}{n}\sum_{i=1}^n |u(x_i)|^2$.

Sélection de k : on pose

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

avec $\text{pen}(k,n) \sim \frac{k \log n}{n}$, et on définit $\hat{f} = \hat{f}_{k^*}$

On note $f = E(y|x)$ la fonction de régression et $\|u\|^2 = E(|u(x)|^2)$.

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

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

Fonctions très concentrées

Et si $f = \sum c_g g$ avec $(c_g) \in w\ell^p$ et $p < 1$?

Fonctions très concentrées

Et si $f = \sum c_g g$ avec $(c_g) \in w\ell^p$ et $p < 1$?

Théorème (DeVore et Temlyakov 1998) : il existe une suite $f_N = \sum_{k=1}^{N} c_k g_k$ telle que

$$\|f - f_N\| \leq C N^{-s}, \quad \text{avec} \quad s = \frac{1}{p} - \frac{1}{2}.$$

Fonctions très concentrées

Et si $f = \sum c_g g$ avec $(c_g) \in w\ell^p$ et $p < 1$?

Théorème (DeVore et Temlyakov 1998) : il existe une suite $f_N = \sum_{k=1}^{N} c_k g_k$ telle que

$$\|f - f_N\| \leq CN^{-s}, \quad \text{avec} \quad s = \frac{1}{p} - \frac{1}{2}.$$

Cependant, les algorithmes de poursuites ne convergent pas mieux en général qu'en $N^{-\frac{1}{2}}$ pour de telles fonctions.

Problème ouvert : identifier des conditions sur le dictionnaire $\mathcal{D}$ qui permettent d'avoir la vitesse optimale N^{-s} pour de telles fonctions.

L'algorithme PGA

- DeVore et Temlyakov (1998) : si $f \in \mathcal{L}^1$, l'algorithme PGA converge avec

$$\|f - f_N\| \leq C N^{-\frac{1}{6}}.$$

L'algorithme PGA

- DeVore et Temlyakov (1998) : si $f \in \mathcal{L}^1$, l'algorithme PGA converge avec

$$\|f - f_N\| \leq CN^{-\frac{1}{6}}.$$

- Konyagin et Temlyakov (2002) : $\|f - f_N\| \leq CN^{-\frac{11}{62}}$.

L'algorithme PGA

- DeVore et Temlyakov (1998) : si $f \in \mathcal{L}^1$, l'algorithme PGA converge avec

$$\|f - f_N\| \leq CN^{-\frac{1}{6}}.$$

- Konyagin et Temlyakov (2002) : $\|f - f_N\| \leq CN^{-\frac{11}{62}}$.

- Lifchitz et Temlyakov (2005) : il existe un dictionnaire $\mathcal{D}$ et $f \in \mathcal{L}^1$ tels que

$$\|f - f_N\| \geq cN^{-0.27}.$$

L'algorithme PGA

- DeVore et Temlyakov (1998) : si $f \in \mathcal{L}^1$, l'algorithme PGA converge avec

$$\|f - f_N\| \leq CN^{-\frac{1}{6}}.$$

- Konyagin et Temlyakov (2002) : $\|f - f_N\| \leq CN^{-\frac{11}{62}}$.

- Lifchitz et Temlyakov (2005) : il existe un dictionnaire $\mathcal{D}$ et $f \in \mathcal{L}^1$ tels que

$$\|f - f_N\| \geq cN^{-0.27}.$$

Vitesse optimale inconnue !

Conclusion : rôle crucial du facteur $\alpha_k = (1 - \frac{c}{k})_+$ dans l'algorithme RGA à la place de $\alpha_k = 1$ pour l'algorithme PGA.

Matching pursuit vs basis pursuit

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

$$\min_{\|f-\sum c_g g\|_{L^2}\leq\varepsilon} \|(c_g)\|_{\ell^0}$$

Matching pursuit vs basis pursuit

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

$$\min_{\|f-\sum c_g g\|_{L^2}\leq\varepsilon} \|(c_g)\|_{\ell^0}$$

A convex relaxation to this problem is

$$\min_{\|f-\sum c_g g\|_{L^2}\leq\varepsilon} \|(c_g)\|_{\ell^1}$$

known as basis pursuit (closely related to LASSO).

Matching pursuit vs basis pursuit

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

$$\min_{\|f-\sum c_g g\|_{L^2}\leq\varepsilon} \|(c_g)\|_{\ell^0}$$

A convex relaxation to this problem is

$$\min_{\|f-\sum c_g g\|_{L^2}\leq\varepsilon} \|(c_g)\|_{\ell^1}$$

known as basis pursuit (closely related to LASSO).

In the case of an orthonormal basis, basis pursuit and matching pursuit are essentially equivalent to coefficient thresholding.

Basis pursuit has the same approximation properties as matching pursuit under restricted assumptions on the dictionary (Cohen, Demol, Gribonval).

Autres directions

- Classification : algorithmes de poursuites mis en oeuvre pour des mesures de risque $L(f, f_N)$ non hilbertiennes (boosting, SVM).

Autres directions

- Classification : algorithmes de poursuites mis en oeuvre pour des mesures de risque $L(f, f_N)$ non hilbertiennes (boosting, SVM).

-Analyse de convergence dans les espaces de Banach : théorie disponible (Temlyakov).

Autres directions

- Classification : algorithmes de poursuites mis en oeuvre pour des mesures de risque $L(f, f_N)$ non hilbertiennes (boosting, SVM).

-Analyse de convergence dans les espaces de Banach : théorie disponible (Temlyakov).

- Problèmes inverses (compressed sensing) : pour $x \in \mathbb{R}^n$ avec $n >> 1$, on observe $y = \Phi x \in \mathbb{R}^m$ avec Φ une matrice $m \times n$ et $m << n$. Problème : approcher x à la précision de ses m plus grands coefficients (Candes-Tao-Romberg, Donoho-Elad, Gilbert-Strauss, Mutukrishnan-Cormode..). On peut écrire $y = \sum_{i=1}^{n} x_i \phi_i$ avec ϕ_i les colonnes de Φ, et chercher à capturer les plus grands x_i en approchant y par une combinaison linéaire du dictionnaire $\mathcal{D} = (\phi_i)$.

Article (Ann. of Stats) : www.ann.jussieu.fr/~cohen