Le problème des signes : son lien avec le problème des phases et algorithme de reconstruction

The sign problem: its connection with the phase problem and retrieval algorithm

Françoise BROUAYE

École Supérieure d'Électricité, plateau du Moulon, 91192 GIF-SUR-YVETTE.

F. Brouaye est actuellement professeur de Mathématiques à l'école Supérieure d'Électricité et travaille au service électromagnétisme de cette école, qui est dirigé par J. C. Bolomey. Par ailleurs elle fait partie de l'équipe CNRS UA 743 (Bâtiment de Mathématiques, Université Paris-XI, 91405 Orsay Cedex) où elle a effectué des travaux de recherche dans le domaine de la Statistique bayésienne empirique.

RÉSUMÉ

Dans certaines expériences de physique, on mesure le carré du champ électrique par échantillonnage le long d'un axe et on veut reconstruire le champ électrique le long de cet axe. Le problème se formalise en appelant $f(x)$ la fonction mesurée le long de l'axe et en supposant que f est une fonction de $\mathbb{R}$ dans $\mathbb{C}$ de carré intégrable, à spectre borné dans un intervalle $[a, b]$ connu, et dont on connaît le carré aux points d'échantillonnage.

Ce problème est à rapprocher du problème des phases [1] où sous les mêmes hypothèses on mesure $|f(x)|$ sur $\mathbb{R}$ ou sur un intervalle, et on cherche à reconstruire f partout. Le problème des phases a été abondamment étudié et on sait [2] qu'il est mal posé si on ne rajoute pas l'hypothèse sur la fonction f; par contre, le problème des signes admet comme uniques solutions $\pm f$ lorsque Δt est suffisamment petit.

L'article est consacré à l'étude des relations entre les deux problèmes, à l'étude asymptotique des solutions données par des algorithmes de minimisation et à la construction d'un algorithme pour le problème des signes; en ce qui concerne le problème des signes, je donne une démonstration élémentaire de l'unicité basée sur la théorie de l'échantillonnage. Les démonstrations sont mises en annexe.

MOTS CLÉS

Algorithmes, échantillonnage, problème des phases.

SUMMARY

In some physical experiments, the square of the electrical field is mesured along an axis trough sampling techniques. The problem can be formalised, by considering the function $f(x)$, which represents the electrical field, and assuming this function to be square-integrable, band-limited to a known interval $[a, b]$, and with square known at the sampling points.

The paper is devoted to relations between this problem, the sign problem, and the well-known phase problem, asymptotic behavior of the solutions from minimisation algorithms, and to the construction of an algorithm for the sign problem. An elementary proof of unicity in case of sign problem is given.

KEY WORDS

Algorithms, sampling, phase problem.

1. Introduction

Soit f une fonction de $\mathbb{R}$ dans $\mathbb{C}$. On suppose que f est de carré intégrable et que sa transformée de Fourier, $\hat{f}$, est nulle en dehors d'un intervalle borné $[a, b]$. On sait que cette hypothèse implique que f est la restriction à $\mathbb{R}$ d'une fonction analytique, de type exponentiel [3].

Lorsqu'on choisit un pas d'échantillonnage Δt, on pose $T=(\Delta t)^{-1}$. D'après le théorème d'échantillonnage [4] (*voir* aussi dans les démonstrations du paragraphe 2), si T est supérieur à la largeur du spectre, $b-a$, on peut reconstituer la fonction f à partir de la suite de valeurs $f(n\Delta t)$ par la formule d'extrapolation :

$$(1) \qquad f(t)=\sum_{n=-\infty}^{n=+\infty} f(n\Delta t)\,\varphi_n(t)$$

où

$$(2) \qquad \varphi_n(t)=(-1)^n e^{-(2ian/T)}\frac{\sin(\pi(Tt-n))}{\pi(Tt-n)}e^{i\pi t(T+2a)}$$

φ_n est la fonction qui a pour transformée de Fourier,

$$\frac{1}{T}e^{2i\pi(\lambda/T)n}1_{[a,\,a+T]}(\lambda)$$

Dans de nombreux problèmes pratiques [5], on connaît les quantités $f(n\Delta t)$ uniquement par leur module. La situation où $f(n\Delta t)$ est connu au signe près, se présente notamment en Électromagnétisme lorsqu'on fait des mesures d'antennes en configuration monostatique, c'est-à-dire lorsque l'antenne testée est aussi l'antenne qui sert à faire les mesures; dans ce cas, au lieu de mesurer une composante du champ électrique, on mesure son carré; la phase est donc connue à π près, ce qui revient à résoudre un problème de signe.

Dans les deux cas, le problème est alors de reconstituer la fonction f à partir d'informations partielles. Les deux problèmes, problème des phases et problème des signes, admettent des formulations voisines et il y a diverses relations entre eux. Dans toute la suite, on supposera T strictement supérieur à la largeur du spectre $b-a$.

2. Unicité de la solution

Dans les deux cas, la fonction f est à déterminer dans un espace de fonctions que nous noterons $\mathscr{E}$, qui est l'espace des fonctions accessibles à l'observateur :

$$\mathscr{E}=\left\{g:\mathbb{R}\to\mathbb{C}/g(t)=\sum_{n=-\infty}^{n=+\infty}\varepsilon_n c_n\,\varphi_n(t);\ \varepsilon_n\in K\right\}$$

en posant

(i) dans le premier cas :

$$c_n=|f(n\Delta t)| \qquad \text{et} \qquad K=\mathbb{T}=\{z\in\mathbb{C}/|z|=1\};$$

(ii) dans le deuxième : $c_n=$ une racine carrée arbitraire de $f^2(n\Delta t)$ et $K=\{-1, 1\}$.

En effet, f est dans $\mathscr{E}$ grâce à l'égalité,

$$\mathscr{E}=\left\{g:\mathbb{R}\to\mathbb{C}/g(t)=\sum_{n=-\infty}^{n=+\infty}\varepsilon_n f(n\Delta t)\,\varphi_n(t);\ \varepsilon_n\in K\right\}$$

Lorsque les ε_n sont choisis au hasard dans $\{-1, 1\}$, et indépendamment les uns des autres, les éléments de $\mathscr{E}$ sont appelés des séries de Rademacher [6].

Pour éviter les cas triviaux, on dira que deux configurations ε_n et ε'_n sont équivalentes (resp. les fonctions g construites à partir de ces configurations) si elles sont proportionnelles, c'est-à-dire si il existe un élément a de K tel que $\varepsilon_n=a\,\varepsilon'_n$, $\forall n$.

Dans ce cas, les fonctions construites sont aussi proportionnelles et ont même module aussi bien dans le domaine temporel que dans le domaine spectral puisque la constante a est toujours de module 1 :

$$\left|\sum_{n=-\infty}^{n=+\infty}\varepsilon_n c_n\,\hat{\varphi}_n(\lambda)\right|=\left|\sum_{n=-\infty}^{n=+\infty}\varepsilon'_n c_n\,\hat{\varphi}_n(\lambda)\right|$$

On a le résultat suivant :

Proposition : *Si Δt est strictement inférieur à $(b-a)^{-1}$ alors pour toute fonction g de $\mathscr{E}$ il y a un nombre infini d'indices n pour lesquels $g(n\Delta t)$ est non nul. D'autre part, toutes les fonctions de $\mathscr{E}$ sont de carré intégrable, à spectre dans l'intervalle $[a, a+T]$ et de même énergie* E *que la fonction f :*

$$(3) \qquad E=\Delta t\left\{\sum_{n=-\infty}^{n=+\infty}|f(n\Delta t)|^2\right\}$$

Lorsqu'on modifie ε_n, on change la répartition dans le domaine spectral de cette énergie. Lorsque tous les ε_n sont égaux, g est équivalente à f, $|g(t)|$ est égal à $|f(t)|$ et l'énergie s'écrit comme une intégrale pour λ variant de a à b uniquement, alors que pour d'autres configurations de la suite ε_n, g a une plus grande largeur de bande que la fonction f initiale (*voir* § 3).

On remarque que f réalise le minimum de l'intégrale, que l'on notera $I(g)$:

$$(4) \qquad I(g)=\int_b^{a+T}|\hat{g}(\lambda)|^2\,d\lambda$$

parmi toutes les fonctions g de $\mathscr{E}$:

$$(5) \qquad I(f)=\inf_{g\in\mathscr{E}} I(g)$$

Si le minimum est atteint pour un seul élément de $\mathscr{E}$ (*unicité à l'équivalence près*) alors on peut reconstruire la fonction f par cette méthode. Dans la pratique on minimise $I(g)$ sur un espace $\mathscr{E}_N$ (§ 5).

Lemme : *Pour le problème de reconstruction des signes, quel que soit Δt strictement inférieur à $(b-a)^{-1}$, si deux configurations non équivalentes ε_n et ε'_n réalisent le minimum, alors les deux ensembles disjoints*

$$A=\{n/\varepsilon_n=\varepsilon'_n \text{ et } c_n\neq 0\}$$

et

$$B=\{n/\varepsilon_n=-\varepsilon'_n \text{ et } c_n\neq 0\}$$

sont infinis.

La construction de deux tels ensembles d'indices est impossible lorsque Δt est inférieur à $[2(b-a)]^{-1}$ comme le montre le théorème ci-dessous.

Si $\Delta t > [2(b-a)]^{-1}$, on peut construire une fonction f pour laquelle le minimum n'est pas unique. Ce résultat n'est pas surprenant puisque dans le cas de la détermination des signes, $f^2(t)$ a un spectre de largeur $2(b-a)$; il faut donc imposer $\Delta t \leqq [2(b-a)]^{-1}$ pour être sur de reconstruire $f^2(t)$ à partir de l'échantillonnage.

Théorème 1 : (*a*) *Pour le problème de détermination de signes, le minimum de* $I(g)$ *est atteint par un seul élément de* $\mathscr{E}$ *dès que* $\Delta t \leqq [2(b-a)]^{-1}$.

(*b*) *Pour le problème de détermination des phases, le minimum n'est pas atteint de façon unique dès que f admet un zéro non réel.*

Le résultat (*a*) peut se généraliser comme suit : supposons une puissance de f, $f^p(t)$, connue aux points d'échantillonnage, c'est-à-dire qu'en ces points le module est connu tandis que la phase est connue à $2\pi/p$ près. En prenant K égal à l'ensemble des racines p-ièmes de l'unité dans la définition de l'espace $\mathscr{E}$, le minimum de $I(g)$ sur $\mathscr{E}$ est unique dès que Δt est inférieur à $[p(b-a)]^{-1}$.

Dans le cas du problème des signes, on sait qu'il faut choisir une seule et même détermination de la racine carrée entre deux zéros réels. Cela n'apporte pas d'information quand aux « signes » à choisir aux points d'échantillonnage; en effet, même si on localise les zéros réels de f^2, donc de f, l'argument de f^2 peut varier d'un multiple de 2π entre deux zéros réels; dans ce cas, une détermination de la racine carrée est obtenue en choisissant l'argument continûment le long de la portion $t \to f^2(t)$ et en le divisant par deux.

Entre deux points d'échantillonnage, le « signe » peut avoir changé, c'est-à-dire que l'argument peut avoir varié de $\pm\pi$, sans que la fonction se soit annulée.

On peut néanmoins avoir des informations supplémentaires aux points où le module est suffisamment grand (§ 4). Dans la pratique, une méthode « empirique » pour faire le suivi de la phase entre deux points d'échantillonnage, consiste, par exemple, à choisir la détermination de la phase au point d'indice $n+1$, qui minimise l'écart des phases.

Remarque. On peut minimiser l'intégrale

$$(6) \qquad \int_b^{a+T} h(\lambda) \left| \hat{g}(\lambda) \right|^2 d\lambda$$

où $h(\lambda)$ est une fonction mesurable strictement positive et bornée, ce qui revient à introduire une pondération sur les différentes valeurs de λ entre b et $a+T$.

3. Effet des erreurs sur le spectre : localisation et largeur

On peut se demander quel est l'effet des erreurs de signes, dans le domaine spectral; les exemples ci-dessous montrent que l'on peut translater complètement le spectre ou l'élargir.

3.1. Effet d'une erreur ou de deux erreurs sur le spectre

Si tous les ε_n sont égaux à 1 sauf pour l'indice n_0, l'intégrale $I(g)$ est égale à

$$(7) \qquad \Delta t[1-\Delta t(b-a)] \left| \varepsilon_{n_0} - 1 \right|^2 \left| f(n_0 \Delta t) \right|^2$$

$I(g)$ est faible lorsque Δt est proche de 0 ou de $(b-a)^{-1}$ ou lorsque le module $\left|f(n_0 \Delta t)\right|$ est petit. Comparons $I(g)$ avec l'énergie totale de la fonction. Le rapport $\left|f(n_0 \Delta t)\right|^2/E$ est maximum pour la fonction $\Psi(t-n_0 \Delta t)$ où

$$(8) \qquad \Psi(t) = e^{i\pi(a+b)t} \frac{\sin \pi(b-a)t}{\pi t}$$

On trouve alors

$$(9) \qquad I(g)/E = [1-\Delta t(b-a)] \Delta t(b-a) \left| \varepsilon_{n_0} - 1 \right|^2$$

Le rapport $I(g)/E$ atteint son maximum, 1, lorsqu'on prend $T = 2(b-a)$ et $\varepsilon_{n_0} = -1$; dans ce cas le spectre de $g(t)$ est concentré sur l'intervalle $[b, a+T]$. Le spectre de f a donc été translaté par une seule erreur de signe, erreur qui a lieu sur un point important; c'est par ailleurs le cas le plus favorable pour détecter une erreur puisqu'on connaît la localisation du spectre. De même, $I(g)$ peut être égal à E avec deux erreurs exactement; c'est le cas par exemple avec une combinaison de deux sinus cardinaux :

$$a = -b = -0{,}5; \qquad T = 2;$$
$$f(t) = \Psi(t-\Delta t) + \Psi(t+\Delta t);$$
$$\varepsilon_n = 1 \qquad \text{si} \quad n \neq \pm 1;$$
$$\varepsilon_1 = \varepsilon_{-1} = -1.$$

3.2. Exemple avec une infinité d'erreurs

Étant donnée une configuration de signes, en général on ne sait pas donner l'expression de la fonction $g(t)$ ainsi obtenue. Considérons la fonction sinus cardinal

$$f(x) = \frac{\sin(\pi x)}{\pi x}$$

qui a pour transformée de Fourier $1_{[-0{,}5,\, 0{,}5]}(\lambda)$.

Alors si on prend $\Delta t = 1/4$, la configuration de signes

$$\varepsilon_n = -1 \qquad \text{si} \quad n = 2 \bmod 4$$

permet de calculer la fonction $g(t)$ en comparant avec la formule d'échantillonnage (1) écrite pour $\Delta t = 3/4$:

$$g(t) = \sum_{n=-\infty}^{+\infty} \varepsilon_n f(n\Delta t)\, \varphi_n(t)$$
$$= -\frac{1}{2} \frac{\sin(4\pi t)}{\pi t} + \frac{\sin(3\pi t)}{\pi t}$$

Sur cet exemple la largeur du spectre a été multipliée par 4 et les 3/4 de l'énergie se sont déplacés en dehors de la zone de fréquences initiale $[-1/2, 1/2]$.

4. Étude des contraintes entre points voisins (problème des signes)

Dans les deux problèmes étudiés, on connaît la suite des modules $|f(n\Delta t)|$; on en déduit l'énergie E par l'expression (3). La connaissance de E permet de majorer les variations de f sur un intervalle $[t, t+\Delta t]$; en effet, si on pose

$$M = \Delta t\, 2\pi \sqrt{E} \sqrt{\frac{b^3 - a^3}{3}}$$

on montre que le module de la variation de f entre deux points d'échantillonnage, est inférieur ou égal à M. On connaît les nombres complexes $z_n = (f(n\Delta t))^2$. Si $|z_n| > M^2$ le disque centré sur $f(n\Delta t)$ et de rayon M ne contient pas l'origine; $f((n-1)\Delta t)$ est donc l'unique racine carrée de z_{n-1} qui vérifie

$$|f(n\Delta t) - \sqrt{z_{n-1}}| \leqq M$$

De même, $f((n+1)\Delta t)$ est l'unique racine carrée de z_{n+1} qui vérifie

$$|f(n\Delta t) - \sqrt{z_{n+1}}| \leqq M$$

Cela veut dire que l'on doit choisir les racines carrées de z_{n-1}, z_n et z_{n+1} de façon à les rapprocher au maximum dans le plan complexe (écart de phase minimal); dans le cas réel cela veut dire prendre le même signe pour $f(n\Delta t)$, $f((n-1)\Delta t)$ et $f((n+1)\Delta t)$; c'est la méthode pratique que l'on utilise en l'absence d'algorithme pour faire le suivi de la phase.

On montre que cette inégalité n'est possible que si Δt est suffisamment petit, plus précisément

$$\Delta t < \frac{1}{b-a}\left(\frac{2}{\pi^2}\right)^{3/2}$$

Dans le cas de la fonction $\sin(\pi x)/(\pi x)$, on trouve $M = \pi\Delta t/\sqrt{3}$ l'inégalité $M < |f(n\Delta t)|$ est vérifiée pour $n=0$, ce qui veut dire que l'on doit prendre le même signe pour des points d'indices -1,0 et 1, dès que Δt est inférieur à $\sqrt{3}/\pi$.

Proposition : *Soit* $A > 0$; *lorsque* Δt *tend vers* 0, *la proportion de points dans l'intervalle* $[-A, A]$ *qui satisfont l'inégalité* $M < |f(n\Delta t)|$, *tend vers* 1.

Dans la pratique, on estime E puis M à l'aide des valeurs mesurées; on impose alors les contraintes de signe entre les points qui vérifient l'inégalité et leurs plus proches voisins; comme en sous-estime E, on surestime le nombre de points soumis à la contrainte.

5. Comportement asymptotique des solutions données par des algorithmes de minimisation : reconstruction des angles ou des signes

Si on se restreint à l'ensemble $\mathscr{E}_N$ des fonctions g qui s'écrivent comme la somme finie,

$$\sum_{n=-N}^{n=N} \varepsilon_n c_n \varphi_n(t)$$

le problème de recherche du minimum de l'intégrale (4) sur $\mathscr{E}_N$, qui est un problème d'optimisation non linéaire, devient alors équivalent à la minimisation de

$$\mathscr{R}e\Big\{ \sum_{n \neq m} \varepsilon_m \bar{\varepsilon}_n A_{m,n} \Big\}$$

avec,

$$A_{n,m} = c_n \bar{c}_m \langle \hat{\varphi}_n, \hat{\varphi}_m \rangle$$

et où les produits scalaires sont pris dans $L^2([b, a+T])$.

On a alors le résultat suivant :

Théorème 2 : *Soit* g_N *réalisant le minimum de* $I(g)$ *sur* $\mathscr{E}_N$; *alors :*

(*a*) $I(g_N)$ *tend vers* 0 *lorsque* N *tend vers l'infini.*

(*b*) *Il existe une sous-suite* N_k *tendant vers* $+\infty$ *et une suite* ε_n *d'éléments de* K *telle que*

$$\hat{g}(\lambda) = \sum_{n=-\infty}^{+\infty} \varepsilon_n c_n \hat{\varphi}_n(\lambda)$$

est identiquement nulle hors de $[a, b]$ *et* g_{N_k} *converge vers g en norme* L^2.

(*c*) *Dans le cas du problème de détermination de signes, et sous l'hypothèse* $\Delta t \leqq [2(b-a)]^{-1}$, *supposons* $c_0 \neq 0$ *et* g_N *choisi de façon à avoir toujours* $\varepsilon_0 = +1$, *alors à partir d'un certain rang* $\varepsilon_{1,N}$ *et* $\varepsilon_{-1,N}$ *sont constants et égaux à* $+1$ *si f est non nulle en* $\pm\Delta t$, *à partir d'un certain rang* $\varepsilon_{1,N}$, $\varepsilon_{-1,N}$, $\varepsilon_{2,N}$, $\varepsilon_{-2,N}$ *sont égaux à* $+1$, ... *et* g_N *converge vers f.*

Ces résultats restent vrais s'il s'agit de minimum sous contrainte du paragraphe 4.

La méthode théorique de reconstruction des signes, qui consiste à trouver le minimum sur $\mathscr{E}_N$ et faire tendre N vers $+\infty$, converge vers la fonction f quelque soit l'algorithme utilisé pour trouver le minimum que $\mathscr{E}_N$. La difficulté est de trouver un algorithme qui donne le minimum global.

6. Simulations pour le problème de reconstruction des signes

D'après le résultat (*c*) du théorème 2, si on sait trouver le minimum avec N points de mesure, en augmentant N, on doit voir les signes se stabiliser. En l'absence d'algorithme donnant le minimum global pour ce problème, la méthode utilisée dans ce paragraphe est une adaptation de l'algorithme de Hebb [7] où l'on fait intervenir les contraintes entre points voisins si il y a lieu. Il s'agit d'une méthode itérative où la quantité à minimiser diminue à chaque pas, et qui converge en quelques itérations (*voir* les quatre exemples sur la figure 1). Dans certains cas, il y a plusieurs limites possibles, c'est-à-dire que la limite dépend de la configuration initiale de signes qui a été choisie. L'étude a été faite en simulant à chaque fois, de façon aléatoire, 1 000 configurations initiales de signes afin de tester la convergence vers la bonne solution. Les résultats montrent que dans le cas où il y a plusieurs limites possibles, celles-ci restent en nombre très limité. Cette méthode a été choisie de

préférence à celle du recuit simulé [9] en raison des durées d'exécution; une comparaison des algorithmes est néanmoins à l'étude.

On peut toujours se ramener au cas où le spectre est symétrique, de largeur 1 en considérant

$$\frac{1}{b-a} e^{-i\pi x (a+b)} f\left(\frac{x}{b-a}\right)$$

Les simulations ont été faites sur les trois fonctions f_1, f_2, f_3 suivantes,

$$f_1(x) = \frac{\sin(\pi x)}{\pi x}$$

$$f_2(x) = \left(\frac{\sin^{(\pi x/2)}}{\pi x}\right)$$

$$f_3(x) = \frac{1}{2}\left[\frac{\sin \pi (x-0{,}5)}{\pi(x-0{,}5)} + \frac{\sin \pi (x+0{,}5)}{\pi(x+0{,}5)}\right]$$

en prenant N points de chaque côté de 0, soit $2N+1$ points de mesure au total. Le pas d'échantillonnage Δt a été choisi de façon à ne pas annuler les sinus cardinaux de manière systématique; N a été choisi de façon à avoir des mesures à 40 dB (1% du maximum).

Les mesures sont utilisées pour estimer l'énergie E et la borne M du paragraphe 4. Lorsque l'inégalité $M < |f(x_i)|$ est vérifiée, cela impose de prendre le même signe pour $f(x_{i-1})$, $f(x_i)$, $f(x_{i+1})$. Grâce à ces contraintes, les signes d'un certain nombre d'éléments sont fixés à $+1$. Le tableau I donne le nombre de points où l'inégalité est vérifiée.

TABLEAU I

Fonction	N	Δt^{-1}	$N\Delta t$	Nombre de points qui vérifient l'inégalité
f_1	40	2,564	15,6	3
	63	4,0384	»	5
f_2	15	2,8037	5,35	3
	17	3,1775	»	5
	22	4,1111	»	7
f_3	8	2,5316	3,16	3
	10	3,1646	»	5
	13	4,1111	»	7

Description de l'algorithme: Dans le cas de la reconstruction des signes, la quantité à minimiser s'écrit comme la somme d'un terme qui ne dépend pas de ε_i et de $-\varepsilon_i V_i$ avec

$$V_i = \sum_{n \neq i} \varepsilon_n C_{n,i}$$

$$C_{n,i} = -\mathcal{R}e(c_n \bar{c}_i) \langle \hat{\varphi}_n, \hat{\varphi}_i \rangle.$$

Dans cette expression, le produit scalaire est écrit dans L^2 ($[0{,}5,\ 0{,}5+T]$). On regarde tout d'abord si la configuration exacte correspond à un minimum local, c'est-à-dire si elle ne peut être améliorée en changeant le signe d'un élément quelconque, puis si c'est le cas, on regarde son pouvoir d'attraction en choisissant les signes des éléments non fixes, de façon aléatoire et en faisant évoluer le système suivant la règle proche de celle de Hebb [7]; pour cela, on parcourt les indices i jusqu'à ce que l'on puisse plus améliorer par la règle :

Si ε_i n'est pas fixé à $+1$,

$$\varepsilon_i \text{ (temps } n+1) = -\varepsilon_i \text{ (temps } n)$$

si et seulement si $\varepsilon_i V_i < 0$.

Dans le cas où l'on change le signe, on en tient compte dans le calcul du potentiel de l'élément suivant.

On passe ensuite au signe de l'élément suivant... lorsqu'on arrive au dernier élément, on revient au premier élément de la liste. On appelle itération, un parcours de tous les sites.

Cette règle d'évolution donne accès à des minimums locaux uniquement. La seule différence avec la règle d'évolution de Hebb, c'est que l'on fixe certains signes à cause des contraintes entre points voisins. De même on tient compte de ces contraintes dans la configuration initiale. Les résultats sont résumés dans le tableau II.

Dans le cas où il n'y a pas convergence vers la bonne solution, il est intéressant de comparer la distance de Hamming (nombre de signes faux) initiale et la distance de Hamming finale (tableau III). La figure 1 montre l'évolution du nombre de signes faux en fonction du nombre d'itérations (nombre de fois où l'on parcourt tous les sites); la dernière itération correspond à un parcours des sites où aucune amélioration n'est possible.

TABLEAU II

Fonction	N	Δt^{-1}	Minimum local	Pouvoir attractif (1000 simulations)
f_1	40	2,564	Non	0%
	12	»	Oui	100%
	63	4,0384	Non	0%
	20	»	Oui	100%
f_2	15	2,8037	Oui	31,9%
	5	»	Oui	100%
	17	3,1775	Oui	29,4%
	12	»	Oui	65,3%
	22	4,1111	Oui	21,7%
	15	»	Oui	56,5%
f_3	8	2,5316	Oui	100%
	10	3,1646	»	»
	13	4,1111	»	»

TABLEAU III

Fonction	N	Δt^{-1}	Distance initiale moyenne	Distance finale moyenne
f_1	40	2,564	37,8	8,4
	63	4,0384	59,8	6
f_2	15	2,8037	10,9	4,2
	17	3,1775	13,9	5,1
	22	4,1111	17,9	8,1

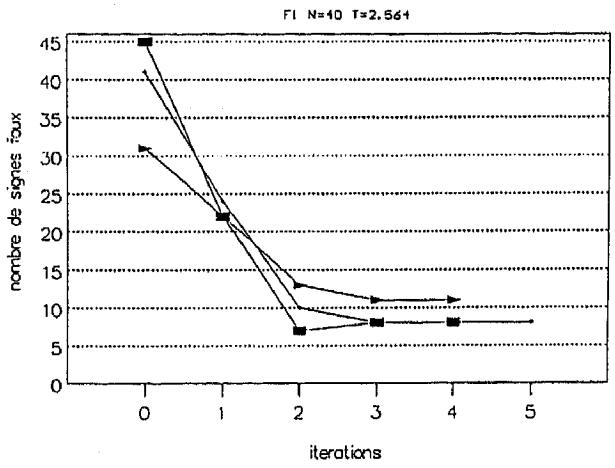

Fig. 1.

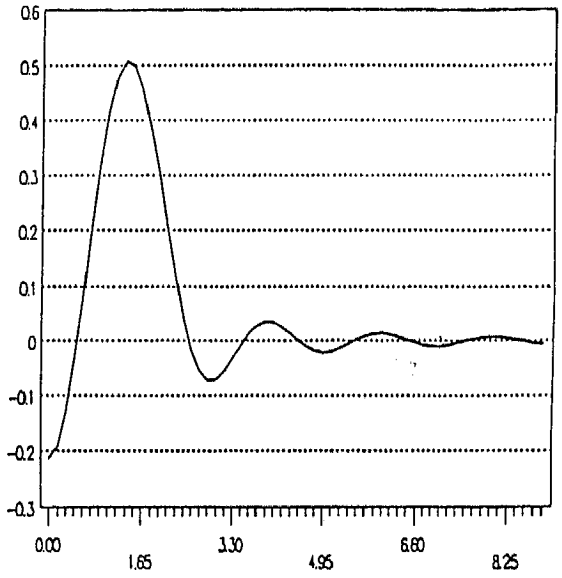

Fig. 2.

Pour la fonction f_1, avec N = 63 et T = 4,0384, on reconstitue correctement la fonction entre $-12,8$ et 12,8, ce qui correspond à 11 lobes secondaires de chaque côté. D'autre part, il faut tenir compte de l'importance des points où lieu une erreur; dans le cas de la fonction f_2 avec N = 17 et T = 3,1775, la limite la plus défavorable qui ait été trouvée, présentait 14 erreurs de signes sur les 35 points au total, mais ces 14 erreurs ne représentaient que 0,25% de l'énergie, proportion calculée par la formule (3).

Les trois fonctions ci-dessus présentent un lobe principal et des lobes secondaires. Il est intéressant de considérer la fonction f_4 suivante :

$$f_4(t) = \frac{1}{2}\left\{\frac{\sin(\pi(t-(3/2)))}{\pi(t-(3/2))} + \frac{\sin(\pi(t+(3/2)))}{\pi(t+(3/2))}\right\}$$

Cette fonction est paire, à spectre de largeur 1 et présente deux lobes principaux d'égale importance (figure 2 pour $t \geqq 0$).

Les simulations ont été faites pour N = 80 et $\Delta t = 0,15$, en forçant les cinq points de chaque lobe principal à être de même signe et en affectant du signe $+1$ le lobe de droite. Suivant les signes initiaux, la limite obtenue est proche de la solution et on passe en moyenne de 74,9 à 12,5 signes faux, ou bien la limite est proche au contraire d'une fonction impaire et on passe en moyenne de 80,4 à 73,5 signes faux. Dans ce cas l'examen des différentes configurations de signes limites (attracteurs), qui sont en nombre limité, et une connaissance *a priori* du phénomène physique, permet de choisir la solution la plus vraisemblable.

Remarque : La règle d'évolution ci-dessus peut être adaptée au problème de la reconstruction de phases; en effet, la quantité à minimiser s'écrit dans ce cas, comme la somme d'un terme qui ne dépend pas ε_i et de $\mathcal{R}e(-\varepsilon_i V_i)$ avec

$$V_i = \sum_{m \neq i} A_{i,m}\,\varepsilon_m$$

On modifie ε_i de façon à obtenir

$$\mathrm{Arg}(\varepsilon_i) + \mathrm{Arg}(V_i) = 0 \bmod 2\pi$$

Au cours de cette évolution, la fonction diminue strictement à chaque modification et si elle converge en temps fini, la solution trouvée réalise un minimum par rapport à chaque variable.

7. Conclusion

L'algorithme construit pour le problème des signes est très simple à mettre en œuvre; d'autre part, on ne connait pas d'adaptation au problème des signes, de l'algorithme de Gerchberg-Saxton [8] utilisé pour le problème des phases. Les simulations effectuées sont encourageantes en vue d'applications puisque dans le cas d'une fonction avec un lobe principal, il y a convergence soit vers la bonne solution, soit vers une fonction proche; les simulations montrent que l'algorithme permet de corriger les signes faux dans une proportion de 54% à 100% lorsque ceux-ci sont choisis initialement au hasard, et de reconstituer jusqu'à 11 lobes secondaires. Les simulations effectuées

sur une fonction à deux lobes principaux met en évidence deux solutions possibles; le choix final se fera alors en utilisant des connaissances *a priori* sur le phénomène. Les essais sur des données physique sont en cours.

Manuscrit reçu le 21 *novembre* 1988.

BIBLIOGRAPHIE

[1] G. Ross, M. A. Fiddy, M. Nieto-Vesperinas et M. W. L. Wheeler, The phase problem in scattering phenomena: the zeros of entire functions and their significance, *Proc. Roy. Soc. London.*, A 360, 1978, p. 25-45.

[2] R. E. Burge, M. A. Fiddy, A. H. Grenaway et G. Rossi, The phase problem, *Proc. Roy. Soc. London*, A 350, 1976, p. 191-212.

[3] R. P. Boas, *Entire functions*, Academic Press, New York. Levin, *Distribution of zeros of entire functions*, Am. Math. Soc., Providence, RI, 1964. A. G. Requicha, The zeros of entire functions: theory and engineering applications, *Proc. IEEE*, 68, n° 3, March 1980, p. 308-328.

[4] B. Picinbono, *Éléments de théorie du signal*, Dunod. H. Reinhard, *Mathématiques du signal*, Dunod.

[5] M. H. Hayes, J. S. Lim et A. Oppenheim, Signal reconstruction from phase and magnitude, *IEEE Trans. ASSP*, 28, n° 6, December 1980. P. L. Van Hove, M. H. Hayes, J. S. Lim et A. V. Oppenheim, Signal reconstruction from signed Fourier Transform Magnitude, *IEEE Trans ASSP*, 31, n° 5, October 1983, p. 1286. A. V. Oppenheim, J. S. Lim et S. R. Curtis, Signal synthesis from partial Fourier-domain transformation, *J. Opt. Soc. Am.*, 73, n° 11, 1983, p. 1413-1419. H. A. Fawerda, *The phase reconstruction problem for wave amplitudes and coherence functions in inverse source problems in Optics*, H. P. Baltes éd., Berlin, Springer Verlag, 1983.

[6] J. P. Kahane, *Random series*, Cambridge University Press.

[7] Hebb, *The organization of behavior*, Wiley, 1949. Hopfield, *J. Proc. Nat. Aca. of Sc. USA*, 1979, p. 2554.

[8] R. W. Gerschberg et W. O. Saxton, A pratical algorithm for the determination of phase from image and diffraction plane pictures, *Optik*, 35, n° 2, 1972, p. 237-246.

[9] S. Geman et D. Geman, Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images, *IEEE TAMI*, 6, n° 6, November 1984, p. 721-741. R. Azencott, Optimisation Stochastique et recuit simulé, Recueil des communications de la journée d'étude du 10 juin 1987 : *Le recuit simulé (annealing) : quelques applications-clés*, École Normale Supérieure.

ANNEXE

Paragraphe 2. Démonstration de la proposition

La transformée de Fourier est une isométrie sur $L^2(\mathbb{R})$; avec la condition de support, f est dans $L^1(\mathbb{R})$ et f est la transformée inverse de $\hat{f}$.

$$f(n\Delta t)=\int_a^b \hat{f}(\lambda)\, e^{2i\pi\lambda\,(n/T)}\, d\lambda$$

Donc la fonction périodique de période T et égale à f sur $[a, a+T]$ a pour coefficient de Fourier d'indice n le rapport $f(n\Delta t)$ sur T.

La relation (3) provient de l'application de la formule de Parseval-Plancherel.

Les fonctions de $\mathscr{E}$ sont écrites sous forme d'une série convergente de fonctions orthogonales.

Si $f(n\Delta t)$ était nul sauf pour un ensemble fini d'indices, le polynôme trigonométrique

$$\sum_n f(n\Delta t)\, e^{-2i\pi\lambda\,(n/T)}$$

serait identiquement nul sur l'intervalle $]b, a+T[$.

Démonstration du lemme

Le minimum de I est atteint pour f et vaut 0. Si deux configurations réalisent le minimum, les transformées de Fourier sont nulles sur l'intervalle $]b, a+T[$; en faisant la somme et la différence de ces deux fonctions, on obtient

$$\sum_{n\in A} f(n\Delta t)\, e^{-2i\pi\lambda\,(n/T)}=0 \qquad \text{sur }]b, a+T[$$

$$\sum_{n\in B} f(n\Delta t)\, e^{-2i\pi\lambda\,(n/T)}=0 \qquad \text{sur }]b, a+T[$$

Ces deux fonctions ne peuvent être des polynômes trigonométriques sans être identiquement nulles.

Construction d'un contre-exemple

Soit $g(\lambda)$ une fonction continue, non identiquement nulle, à support dans $[0, d]$ avec $d \leqq T/4$. On appelle g_1, la fonction périodique, de période T égale à g sur $[0, T]$ et on pose

$$h(\lambda)=g_1(2\lambda)$$

$$\hat{f}(\lambda)=h(\lambda)(1+e^{2i\pi\,(\lambda/T)}) \qquad \text{sur } \left[-\frac{T}{2}, \frac{T}{2}\right]$$

$$=0 \quad \text{en dehors}$$

On a alors

$$f(2n\Delta t)=f((2n+1)\Delta t)$$

La configuration de signes $\ldots +1\ -1\ +1\ -1\ldots$ réalise aussi le minimum.

Démonstration du théorème 1

(*a*) Par translation et homothétie dans le domaine spectral, on se ramène au cas où $a=-1$ et $b=1$. Si g, non équivalente à f, réalise le minimum, en faisant la somme et la différence on a encore une transformée

de Fourier nulle sur l'intervalle]1, $-1+T$[, soit

$$h_A(\lambda)=\sum_{n\in A} f(n\Delta t)\,e^{-2i\pi(n\lambda/T)}=0$$

sur]1, $-1+T$[

$$h_B(\lambda)=\sum_{n\in B} f(n\Delta t)\,e^{-2i\pi(n\lambda/T)}=0 \qquad \text{sur }]1,\ -1+T[$$

Le produit de convolution des fonctions périodiques h_A et h_B est nul puisque le produit de leurs coefficients de Fourier est toujours nul. Considérons les fonctions f_A et f_B

$$f_A(t)=\sum_{n\in A} f(n\Delta t)\,\varphi_n(t)$$

$$f_B(t)=\sum_{n\in B} f(n\Delta t)\,\varphi_n(t)$$

Le produit de convolution de leurs transformées de Fourier est nul en dehors de l'intervalle $[-2,2]$ et est égal à $h_A * h_B$ à l'intérieur de cet intervalle si on suppose Δt inférieur ou égal à 1/4. On en déduit que le produit ordinaire des fonctions analytiques f_A et f_B est identiquement nul. L'une au moins des deux fonctions est identiquement nulle, ce qui contredit l'hypothèse que g n'est pas équivalente à f.

(*b*) Si z est un zéro non réel, on sait (*voir* [1]) que la fonction

$$g(x)=f(x)\,[(x-\bar{z})/(x-z)]$$

vérifie toutes les hypothèses et réalise le minimum sans être équivalente à f.

Paragraphe 4. Majoration des variations de f

$$f'(t)=\int_a^b \hat{f}(\lambda)\,2i\pi\lambda\,e^{2i\pi\lambda t}\,d\lambda$$

$$|f'(t)|^2\leqq E(2\pi)^2\frac{b^3-a^3}{3}$$

Démonstration de la proposition

La fonction f étant analytique et non identiquement nulle, a un nombre fini de zéros, n_A, dans l'intervalle $[-A, A]$. Soit ε vérifiant $0<\varepsilon<A$. On entoure les zéros de f par des intervalles ouverts non vides dont la longueur totale est inférieure ou égale à ε. Le nombre de points du type $k\ \Delta t$ appartenant à ces intervalles est inférieur ou égal à $n_A+(\varepsilon/\Delta t)$.

Posons

$$m=\inf_{x\in D}|f(x)|$$

D étant l'ensemble des points de $[-A, A]$ qui ne sont pas dans l'un des intervalles qui entourent les zéros de f. $M<m$ pour Δt suffisamment petit. Dans ce cas la proportion de points qui ne satisfont pas l'inégalité $M<|f(x_i)|$ est inférieure ou égale à

$$\frac{\varepsilon}{2A-2\Delta t}+\frac{n_A}{2N+1}$$

On choisit donc ε puis Δt pour rendre cette somme aussi petite que l'on veut.

Paragraphe 5. Démonstration du théorème 2

Si on pose

$$f_N(t)=\sum_{|n|\leqq N} f(n\Delta t)\,\Delta t)\,\varphi_n(t)$$

(*a*) $I(g_N)\leqq I(f_N)\xrightarrow[N\to+\infty]{} 0$.

(*b*) La configuration qui réalise le minimum sur $\mathscr{E}_N$ dépend de N. Pour chaque N, on complète la configuration par $+1$ en dehors de $\{-N, \ldots, 0, 1, \ldots, N\}$; appelons $(\varepsilon_{n,\,N})_{-\infty<n<+\infty}$, la configuration ainsi obtenue.

Dans les deux cas étudiés, K est compact. On peut donc extraire une sous-suite d'indices N_k tendant vers $+\infty$ telle que pour chaque n, $\varepsilon_{n,\,N_k}$ converge. On appelle ε_n la limite.

Soit $\varepsilon>0$. Il existe $p>0$ tel que

$$T^{-1}\sum_{|n|>p}|c_n|^2\leqq\varepsilon^2$$

Il existe alors k_1 tel que pour $k\geqq k_1$ on ait

$$|\varepsilon_{n,\,N_k}-\varepsilon_n|\leqq\varepsilon \qquad \text{pour} \quad n=-p, \ldots, 0, \ldots, p$$

Dans $L^2([a, a+T])$, pour $k\geqq k_1$, on a :

$$\|\hat{g}_{N_k}-\sum_n \varepsilon_n c_n\hat{\varphi}_n\|^2=T^{-1}\sum_n|\varepsilon_{n,\,N_k}-\varepsilon_n|^2|c_n|^2$$

$$\leqq\varepsilon^2 T^{-1}\sum_{|n|\leqq p}|c_n|^2+4T^{-1}\sum_{|n|>p}|c_n|^2\leqq(E+4)\,\varepsilon^2$$

D'autre part,

$$\|\sum_n \varepsilon_n c_n\hat{\varphi}_n\|_{L^2([b,\ a+T])}$$

$$\leqq\|\hat{g}_{N_k}-\sum_n\varepsilon_n c_n\hat{\varphi}_n\|_{L^2([b,\ a+T])}+\|\hat{g}_{N_k}\|_{L^2([b,\ a+T])}$$

$$\leqq|\hat{g}_{N_k}-\sum_n\varepsilon_n c_n\hat{\varphi}_n\|_{L^2([a,\ a+T])}+I(g_{N_k})$$

Ces deux quantités tendent vers 0 lorsque k tend vers l'infini, ce qui montre que $\sum\varepsilon_n c_n\hat{\varphi}_n$ est nulle sur $]b, a+T[$.

(*c*) On se place maintenant dans le cas du problème des signes; quitte à changer la numérotation des indices on peut toujours supposer c_0 non nul et choisir g_N de façon à avoir $+1$ sur l'indice 0. Si c_1 est aussi non nul, et si on peut extraire une sous-suite de façon à avoir -1 pour le deuxième indice, en utilisant la méthode de (*b*) on construirait une fonction g solution; en raison de l'unicité de la solution à l'équivalence près, on aboutit à une contradiction puisque f ne s'annule pas sur les deux indices considérés. Ce raisonnement peut être utilisé pour n'importe quel indice n tel que $f(n\Delta t)$ est non nul.