Techniques d'apprentissage non-linéaires en ligne avec contraintes de positivité

Jie Chen[1,2], Cédric Richard[2], Paul Honeine[1], Hichem Snoussi[1], Henri Lantéri[2], Céline Theys[2]

[1]Institut Charles Delaunay, LM2S
FRE 2848 CNRS
Université de Technologie de Troyes, 12 rue Marie Curie. 10010 Troyes, France.
jie.chen@utt.fr, paul.honeine@utt.fr, hichem.snoussi@utt.fr

[2]Laboratoire Fizeau
UMR 6525 CNRS
Université de Nice Sophia-Antipolis, Observatoire de la Côte d'Azur
Parc Valrose. 06108 Nice, France.
cedric.richard@unice.fr, henri.lanteri@unice.fr, celine.theys@unice.fr

Résumé— **Cet article décrit une nouvelle classe d'algorithmes d'apprentissage non-linéaires en ligne avec contrainte de positivité sur la solution. Ceux-ci sont appliqués au problème d'identification distribuée d'un champ scalaire positif, par exemple de rayonnement thermique ou de concentration d'une espèce chimique, par un réseau de capteurs. La question du suivi de l'évolution de la grandeur physique surveillée au cours du temps est également considérée. Les algorithmes proposés sont testés sur des données synthétiques régies par des équations de diffusion. Ils démontrent une excellente capacité de suivi des évolutions du système, tout en affichant un coût calculatoire réduit.**

Mots-clés— **apprentissage, régression, non-linéaire, positivité, adaptatif, réseau de capteurs.**

I. Introduction

Le domaine des réseaux de capteurs sans fil fait actuellement l'objet d'un intérêt considérable de la part des communautés académique et industrielle. La dispersion d'une multitude de capteurs bon marché dans une région donnée, l'élaboration d'un protocole de routage adéquat, et une implémentation algorithmique efficace, ouvrent en effet de nombreuses perspectives d'applications civiles et militaires. Dans cet article, on s'intéresse au problème d'identification distribuée d'un *champ scalaire positif*, par exemple un rayonnement thermique ou la concentration d'une espèce chimique, et le suivi de son évolution au cours du temps.

Le mode de calcul distribué est inhérent au caractère réparti des nœuds du réseau sur l'aire surveillée, dont la tâche est d'acquérir et traiter localement les mesures. L'efficacité de la procédure d'identification est conditionnée par les interactions des nœuds, dictées par la topologie du réseau. Deux principes de coopération ont été principalement considérés dans la littérature. En mode incrémental, l'information transite de façon séquentielle et cyclique d'un nœud voisin à l'autre. Le coût énergétique, largement dicté par le volume des communications, tend à être minimum [9], [13]. En mode de diffusion, chaque nœud coopère avec l'ensemble de ses voisins pour une meilleure qualité d'estimation [8]. Il en résulte un nombre accru d'échanges d'informations, qu'il est possible de restreindre en limitant artificiellement le nombre de voisins concernés. On parle dans ce cas de diffusion probabiliste.

La complexité des applications considérées ici, telles que la diffusion de chaleur mesurée par rayonnement thermique qui illustre la suite de cet article, nécessite de recourir à des méthodes d'identification adéquates. Fondées sur les travaux précurseurs d'Aronszajn [1], les récentes avancées de l'estimation fonctionnelle basée sur les espaces de Hilbert à noyau reproduisant, et de la théorie de la régularisation, ont apporté des réponses convaincantes aux problèmes implicant des systèmes non-linéaires. La stratégie d'apprentissage en ligne considérée, reposant sur un mode coopératif de type incrémental, démontre d'excellentes capacités de suivi des évolutions du système tout en affichant un coût calculatoire réduit. Si plusieurs approches similaires existent dans la littérature, l'article présenté tire en partie son originalité de l'usage du formalisme des espaces de Hilbert à noyau reproduisant.

Ce travail fait suite à [6], [7] dont il se distingue par la contrainte de positivité pesant sur le champ scalaire estimé, ce qui constitue une originalité par rapport à l'ensemble de la littérature du domaine. On s'inspire pour cela de méthodes de déconvolution, où la prise en compte d'informations *a priori* sur l'espace des solutions admissibles est primordiale pour garantir la stabilité du résultat et traduire le phénomène physique associé. Les algorithmes multiplicatifs jouent un rôle central dans la résolution de problèmes inverses sous contrainte de positivité, en raison de leur formulation simple assurant très naturellement la non-négativité de la solution. Les plus populaires sont l'algorithme de Richardson-Lucy [12], [15], obtenu par maximisation de la vraisemblance sous hypothèse poissonienne, particulièrement utilisé en imagerie astrophysique et médicale, et l'algorithme ISRA de Daude-Witterspoon et Muehlenner [4], adapté à des données gaussiennes. La méthode SGM – Split Gradient Method – de Lantéri et coll. [10] généralise ces techniques en les groupant sous une formulation unique, dédiée à la minimisation de divergences entre données et modèles sous contrainte de positivité.

L'article est organisé ainsi. On adapte d'abord l'algorithme SGM au problème d'identification considéré. Son implémentation est distribuée de sorte à satisfaire aux contraintes d'un réseau de capteurs coopérant en mode

incrémental, et ses capacités de poursuite développées par un calcul de gradient stochastique. On rappelle ensuite les fondements de l'estimation fonctionnelle dans les espaces de Hilbert à noyau reproduisant, que l'on traduit algorithmiquement afin d'être à même de traiter des phénomènes non-linéaires complexes. L'approche obtenue est enfin mise en œuvre sur des données synthétiques. Une conclusion et des perspectives viennent clore la discussion.

II. Formulation générale de la méthode

Etant donné un ensemble de couples $(\boldsymbol{x}_i, y_i) \in \mathbb{R}_+^p \times \mathbb{R}_+$, on s'intéresse au problème d'identification d'un modèle de la forme générale $y = \psi(\boldsymbol{x})$, où ψ appartient à un espace fonctionnel $\mathcal{H}$ défini ultérieurement. Afin de fixer les idées, on considère dans un premier temps un modèle linéaire $y_i = \boldsymbol{\alpha}^\top \boldsymbol{x}_i$. Pour de nombreuses applications, une contrainte de non-négativité sur les composantes de $\boldsymbol{\alpha}$ s'impose naturellement. Voir [5], [10]. Celle-ci est au cœur de la méthode développée ici, appliquée en fin d'article à l'estimation d'un champ scalaire positif par un réseau de capteurs. Le problème à résoudre se formule ainsi

$$\boldsymbol{\alpha}^o = \arg\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) \tag{1}$$

$$\text{sous contrainte de } \boldsymbol{\alpha} \geq 0 \tag{2}$$

où $J(\boldsymbol{\alpha})$ désigne une fonction coût donnée. Le lagrangien correspondant s'écrit

$$Q(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = J(\boldsymbol{\alpha}) - \boldsymbol{\lambda}^\top \boldsymbol{\alpha} \tag{3}$$

avec $\boldsymbol{\lambda}$ le vecteur constitué des facteurs de Lagrange, tous non-négatifs. Les conditions de Kuhn-Tucker à l'optimum se traduisent par

$$\nabla_\alpha Q(\boldsymbol{\alpha}^o, \boldsymbol{\lambda}^o) = 0 \tag{4}$$

$$\lambda_i^o \, \alpha_i^o = 0 \qquad \forall i, \tag{5}$$

avec α_i^o (resp. λ_i^o) la i-ème composante de $\boldsymbol{\alpha}^o$ (resp. $\boldsymbol{\lambda}^o$). Ceci se résume par le problème suivant à résoudre

$$\alpha_i^o \, [\nabla J(\boldsymbol{\alpha})]_i = 0, \tag{6}$$

où $[\nabla J(\boldsymbol{\alpha})]_i$ désigne la i-ème composante de $\nabla J(\boldsymbol{\alpha})$. Pour aboutir à un algorithme d'optimisation de type multiplicatif décrit dans la suite, les auteurs de [10] suggèrent d'appliquer la méthode du point fixe à (6). Celle-ci mène à [1]

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} + \eta_i^{(k)} f_i(\boldsymbol{\alpha}^{(k)}) \, \alpha_i^{(k)} [-\nabla J(\boldsymbol{\alpha}^{(k)})]_i, \tag{7}$$

avec $\eta_i^{(k)} > 0$ un facteur de relaxation et $f_i(\boldsymbol{\alpha})$ une fonction positive sur le domaine admissible de $\boldsymbol{\alpha}$. Ces derniers et le signe négatif introduit dans l'expression ci-dessus sont respectivement destinés, d'une part à pouvoir agir localement sur le caractère contractant de la fonction ainsi traitée par la méthode du point fixe, d'autre part à adopter pour direction de descente celle qui est opposée au gradient.

Garantir la positivité de $\alpha_i^{(k+1)}$ à partir de celle de $\alpha_i^{(k)}$ impose la condition

$$\eta_i^{(k)} \leq \frac{1}{f_i(\boldsymbol{\alpha}^{(k)}) \, [\nabla J(\boldsymbol{\alpha})]_i} \tag{8}$$

1. Afin de résoudre l'équation (6), de la forme $\varphi(\alpha) = 0$, on considère le problème $\varphi(\alpha) + \alpha = \alpha$ auquel on applique une méthode de point fixe.

si $[\nabla J(\boldsymbol{\alpha})]_i > 0$. Dans le cas contraire, où $[\nabla J(\boldsymbol{\alpha})]_i \leq 0$, aucune restriction relative à la contrainte de positivité ne s'impose dans le choix de ce pas. On en déduit finalement l'expression générale de l'algorithme

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} + \eta^{(k)} \, \boldsymbol{d}^{(k)}, \tag{9}$$

avec $\boldsymbol{d}^{(k)}$ la direction de descente définie par

$$\boldsymbol{d}^{(k)} = -\text{diag}[f_i(\boldsymbol{\alpha}^{(k)}) \, \alpha_i^{(k)}] \, \nabla J(\boldsymbol{\alpha}^{(k)}), \tag{10}$$

et $\eta^{(k)}$ le pas résultant éventuellement d'un algorithme de recherche linéaire [2] dans l'intervalle $]0, \eta_{\max}^{(k)}]$ avec

$$\eta_{\max}^{(k)} = \min_i \frac{1}{f_i(\boldsymbol{\alpha}^{(k)}) \, [\nabla J(\boldsymbol{\alpha}^{(k)})]_i}. \tag{11}$$

On s'intéresse à présent à l'algorithme multiplicatif associé [10]. Cette forme est récemment devenue très populaire en raison notamment de sa simplicité de mise en œuvre. Elle sera exploitée par la suite dans le cadre des réseaux de capteurs. Pour ce faire, on décompose l'opposé du gradient $-\nabla J(\boldsymbol{\alpha}^{(k)})$ ainsi

$$-\nabla J(\boldsymbol{\alpha}^{(k)}) = U(\boldsymbol{\alpha}^{(k)}) - V(\boldsymbol{\alpha}^{(k)}), \tag{12}$$

où U et V sont deux fonctions à valeurs strictement positives. Voir [11] pour une discussion sur la non-unicité de cette décomposition et ses interprétations. En reportant celle-ci dans l'expression (7), on obtient

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} + \eta_i^{(k)} f_i(\boldsymbol{\alpha}^{(k)}) \, \alpha_i^{(k)} [U(\boldsymbol{\alpha}^{(k)}) - V(\boldsymbol{\alpha}^{(k)})]_i. \tag{13}$$

En posant $f_i(\boldsymbol{\alpha}^{(k)}) = 1/[V(\boldsymbol{\alpha}^{(k)})]_i$, on aboutit à

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} + \eta_i^{(k)} \, \alpha_i^{(k)} [U(\boldsymbol{\alpha}^{(k)})/V(\boldsymbol{\alpha}^{(k)}) - 1]_i. \tag{14}$$

Comme dans (8), garantir la positivité de $\alpha_i^{(k+1)}$ impose la condition

$$\eta_i^{(k)} \leq \frac{1}{1 - [U(\boldsymbol{\alpha}^{(k)}]_i / [V(\boldsymbol{\alpha}^{(k)})]_i} \tag{15}$$

si $[\nabla J(\boldsymbol{\alpha})]_i > 0$, c'est-à-dire $[V(\boldsymbol{\alpha}^{(k)})]_i > [U(\boldsymbol{\alpha}^{(k)})]_i$. Dans le cas contraire, aucune condition ne s'impose. On note que le second membre de l'inégalité (15) est strictement supérieur à 1. En prenant un pas unitaire dans (14), qui garantit en conséquence la positivité du résultat, on aboutit à la formulation multiplicative suivante

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} [U(\boldsymbol{\alpha}^{(k)})]_i / [V(\boldsymbol{\alpha}^{(k)})]_i. \tag{16}$$

En notant $\otimes$ (resp. $\oslash$) le produit (resp. la division) au sens d'Hadamard, on reformule enfin ce résultat ainsi

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} \otimes U(\boldsymbol{\alpha}^{(k)}) \oslash V(\boldsymbol{\alpha}^{(k)}). \tag{17}$$

Il convient de noter le peu d'opérations mises en jeu par cet algorithme, qui assure naturellement la positivité de la solution tout au long du processus d'optimisation si la condition initiale l'est. On relève également que les composantes nulles de $\boldsymbol{\alpha}^{(k)}$ se propagent au cours des itérations, ce qui peut constituer un avantage dans la recherche de solutions parcimonieuses. Enfin, on constate que la formulation multiplicative (17) ne repose sur aucune expression particulière du critère $J(\boldsymbol{\alpha})$ utilisé, ni du modèle sous-jacent.

III. Description de l'algorithme distribué

Dans le contexte de l'apprentissage distribué dans un réseau de N capteurs, on s'intéresse au problème d'identification d'un *champ scalaire positif*, par exemple un rayonnement thermique ou la concentration d'une espèce chimique, et le suivi de son évolution au cours du temps. Chaque nœud n acquière pour cela des mesures $(\boldsymbol{x}_{n,\ell}, y_{n,\ell})$ au cours du temps. Pour simplifier l'exposé, on suppose que le critère $J(\boldsymbol{\alpha})$ peut être décomposé en une somme de critères individuels $J_n(\boldsymbol{\alpha})$, un en chaque nœud n, soit

$$J(\boldsymbol{\alpha}) = \sum_{n=1}^{N} J_n(\boldsymbol{\alpha}). \tag{18}$$

Pour déterminer la solution $\boldsymbol{\alpha}^o$, une approche naturelle consiste à mettre en œuvre l'algorithme de descente (9) sous la forme

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} + \sum_{n=1}^{N} \eta^{(k)} \boldsymbol{d}_n^{(k)}, \tag{19}$$

avec $\boldsymbol{d}_n^{(k)}$ la direction de descente donnée ici par

$$\boldsymbol{d}_n^{(k)} = -\mathrm{diag}\left[f_i(\boldsymbol{\alpha}^{(k)})\, \alpha_i^{(k)}\right] \nabla J_n(\boldsymbol{\alpha}^{(k)}), \tag{20}$$

et $\eta^{(k)}$ un pas choisi dans l'intervalle $]0, \eta_{\max}^{(k)}]$ avec

$$\eta_{\max}^{(k)} = \min_i \frac{1}{f_i(\boldsymbol{\alpha}^{(k)}) \left[\sum_{n=1}^{N} \nabla J_n(\boldsymbol{\alpha}^{(k)})\right]_i}. \tag{21}$$

Il en résulte l'algorithme incrémental provisoire suivant où les paramètres du modèle sont mis à jour, successivement, d'un nœud voisin à l'autre.

1. $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\alpha}^{(k)}$
2. $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} + \eta^{(k)} \boldsymbol{d}_n^{(k)}, \quad n = 1, \ldots, N$
3. $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\beta}_N^{(k)}$

Cette approche nécessite cependant, par l'instruction 2., que chaque nœud ait accès à l'ensemble des données et à $\boldsymbol{\alpha}^{(k)}$ pour le calcul de $\eta^{(k)}$ et $\boldsymbol{d}_n^{(k)}$. Cet inconvénient se retrouve dans la formulation multiplicative qui, après qu'on ait posé $-\nabla J_n(\boldsymbol{\alpha}^{(k)}) = U_n(\boldsymbol{\alpha}^{(k)}) - V_n(\boldsymbol{\alpha}^{(k)})$, s'écrit donc

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} \otimes \sum_{n=1}^{N} U_n(\boldsymbol{\alpha}^{(k)}) \oslash \sum_{n=1}^{N} V_n(\boldsymbol{\alpha}^{(k)}). \tag{22}$$

Afin de résoudre ce problème, qui entraînerait un coût de communication prohibitif, on préconise une évaluation locale du pas et de la direction de descente au point reçu du précédent capteur. L'équation de mise à jour est alors donnée par

$$\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} + \eta_n^{(k)} \boldsymbol{d}_n^{(k)}, \tag{23}$$

avec $\boldsymbol{d}_n^{(k)}$ la direction de descente définie par

$$\boldsymbol{d}_n^{(k)} = -\mathrm{diag}\left[f_i(\boldsymbol{\beta}_{n-1}^{(k)})\, \beta_{n-1,i}^{(k)}\right] \nabla J_n(\boldsymbol{\beta}_{n-1}^{(k)}), \tag{24}$$

et $\eta_n^{(k)}$ un pas choisi dans l'intervalle $]0, \eta_{\max}^{(k)}]$ avec

$$\eta_{\max}^{(k)} = \min_i \frac{1}{f_i(\boldsymbol{\beta}_{n-1}^{(k)}) \left[\nabla J_n(\boldsymbol{\beta}_{n-1}^{(k)})\right]_i}. \tag{25}$$

L'algorithme itératif proposé, élaboré sur le mode de coopération incrémental, est finalement donné par

---

Algorithme itératif

A chaque instant k, répéter

1. $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\alpha}^{(k)}$
2. $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} + \eta_n^{(k)} \boldsymbol{d}_n^{(k)}, \quad n = 1, \ldots, N$
3. $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\beta}_N^{(k)}$

Le réseau est parcouru séquentiellement, et les paramètres du modèle transmis d'un nœud voisin à l'autre, afin de suivre l'évolution du phénomène physique au cours temps.

En posant $f_i(\boldsymbol{\beta}_{n-1}^{(k)}) = 1/[V_n(\boldsymbol{\beta}_{n-1}^{(k)})]_i$ et $\eta_n^{(k)} = 1$, on aboutit finalement à l'algorithme multiplicatif suivant.

Algorithme multiplicatif

A chaque instant k, répéter

1. $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\alpha}^{(k)}$
2. $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} \otimes U_n(\boldsymbol{\beta}_{n-1}^{(k)}) \oslash V_n(\boldsymbol{\beta}_{n-1}^{(k)}), \quad n = 1, \ldots, N$
3. $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\beta}_N^{(k)}$

La section suivante vise à préciser ces algorithmes sous des hypothèses statistiques particulières, par une sélection appropriée de la fonction coût utilisée.

IV. Choix de fonctions coût

Il convient de préciser maintenant quelques choix possibles pour la fonction coût $J(\boldsymbol{\alpha})$. On motive ceux-ci par des hypothèses statistiques, en supposant dans un premier temps que les données sont poissoniennes, en considérant dans un second temps qu'elles sont entachées d'un bruit gaussien additif. Pour autant, ces choix ne se limitent pas à ceux développés ci-dessous.

A. Cas de données poissoniennes

Dans le cadre d'un procédé parfait de détection de photons par un réseau de capteurs, destiné à la mesure d'un rayonnement thermique par exemple, l'intensité $y_{n,\ell}$ en chaque capteur n est supposée suivre une loi de Poisson. On fait l'hypothèse que la moyenne de celle-ci est régie par un modèle linéaire $y_{n,\ell} = \boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}$, où $(\boldsymbol{x}_{n,\ell}, y_{n,\ell})$ est un couple de données acquises par le capteur n, supposées non-négatives. A titre de généralisation, on suppose que L_n de ces couples de données sont disponibles au moment de la mise à jour des paramètre du modèle par le capteur n. La fonction de vraisemblance s'écrit aisément

$$F(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \prod_{\ell=1}^{L_n} \frac{\left(\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}\right)^{y_{n,\ell}}}{y_{n,\ell}!} \exp(-\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}) \tag{26}$$

Par la formule de Stirling, la fonction de log-vraisemblance s'exprime selon

$$-\log F(\boldsymbol{\alpha}) \approx \sum_{n=1}^{N} \sum_{\ell=1}^{L_n} \Big((\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell} - y_{n,\ell}) + y_{n,\ell} \log \frac{y_{n,\ell}}{\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}} \Big). \tag{27}$$

Il s'agit là d'une I-divergence de Csiszär entre $\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}$ et $y_{n,\ell}$. Cette mesure généralise la divergence de Kullback à des fonctions non-négatives qui ne sont pas des distributions. Le problème à résoudre se formule ainsi

$$\boldsymbol{\alpha}^o = \arg\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} J_n(\boldsymbol{\alpha}) \tag{28}$$

$$\text{sous contrainte que } \boldsymbol{\alpha} \geq 0 \tag{29}$$

où $J_n(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L_n} \left[(\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}) - y_{n,\ell} \log(\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell})\right]$ désigne la fonction coût associée au capteur n. Les résultats qui suivent découlent directement des développements de la Section III. En l'occurrence, il convient de calculer le gradient de $J_n(\boldsymbol{\alpha})$ au préalable

$$\nabla J_n(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L_n} \left(\boldsymbol{x}_{n,\ell} - \frac{y_{n,\ell}}{\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}} \boldsymbol{x}_{n,\ell} \right). \tag{30}$$

Si la décomposition de $-\nabla J_n(\boldsymbol{\alpha})$ n'est pas unique, il vient immédiatement que l'on peut toutefois poser

$$U_n(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L_n} \frac{y_{n,\ell}}{\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}} \boldsymbol{x}_{n,\ell} \qquad V_n(\boldsymbol{\alpha}) = \sum_{\ell=1}^{L_n} \boldsymbol{x}_{n,\ell}. \tag{31}$$

L'algorithme correspondant, sous sa forme multiplicative, s'écrit finalement ainsi

---

Algorithme multiplicatif (I-divergence de Csiszär)

A chaque instant k, répéter

1. $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\alpha}^{(k)}$
2. $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} \otimes \sum_{\ell=1}^{L_n} \frac{y_{n,\ell}}{\boldsymbol{x}_{n,\ell}^\top \boldsymbol{\beta}_{n-1}^{(k)}} \boldsymbol{x}_{n,\ell} \oslash \sum_{\ell=1}^{L_n} \boldsymbol{x}_{n,\ell}$
3. $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\beta}_N^{(k)}$

Il convient de noter la simplicité de cet algorithme, qui vise pourtant à minimiser une I-divergence de Csiszär en assurant la positivité de la solution tout au long du processus d'optimisation.

B. Cas de données gaussiennes

On suppose que l'intensité $y_{n,k}$ en chaque capteur n est ici encore régie par un modèle linéaire $y_{n,\ell=}\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell}$, à présent corrompue par un bruit additif gaussien i.i.d. centré de variance σ_n^2. La fonction de vraisemblance est

$$F(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \prod_{\ell=1}^{L_n} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(- \frac{(\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell} - y_{n,\ell})^2}{2\sigma_n^2} \right). \tag{32}$$

La fonction de log-vraisemblance correspondante s'écrit dans ce cas

$$-\log F(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{\ell=1}^{L_n} \frac{(\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,\ell} - y_{n,\ell})^2}{\sigma_n^2}, \tag{33}$$

d'où l'on a écarter les termes constants additifs. Comme précédemment, le problème à résoudre se formule ainsi

$$\boldsymbol{\alpha}^o = \arg\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} J_n(\boldsymbol{\alpha}) \tag{34}$$

$$\text{sous contrainte que } \boldsymbol{\alpha} \geq 0 \tag{35}$$

où $J_n(\boldsymbol{\alpha}) = \frac{1}{2\sigma_n^2} \sum_k (\boldsymbol{\alpha}^\top \boldsymbol{x}_{n,k} - y_{n,k})^2$ désigne la fonction coût associée au capteur n. Le calcul du gradient de cette dernière conduit à

$$\nabla J_n(\boldsymbol{\alpha}) = \frac{1}{\sigma_n^2} \sum_{\ell=1}^{L_n} (\boldsymbol{x}_{n,\ell}\, \boldsymbol{x}_{n,\ell}^\top\, \boldsymbol{\alpha} - y_{n,\ell}\, \boldsymbol{x}_{n,\ell}). \tag{36}$$

Il apparaît immédiatement que la décomposition suivante de $-\nabla J_n(\boldsymbol{\alpha})$ en une différence de fonctions strictement positives convient

$$U_n(\boldsymbol{\alpha}) = \frac{1}{\sigma_n^2} \sum_{\ell=1}^{L_n} y_{n,\ell}\, \boldsymbol{x}_{n,\ell} \tag{37}$$

$$V_n(\boldsymbol{\alpha}) = \frac{1}{\sigma_n^2} \sum_{\ell=1}^{L_n} \boldsymbol{x}_{n,\ell}\, \boldsymbol{x}_{n,\ell}^\top\, \boldsymbol{\alpha}. \tag{38}$$

On en déduit l'algorithme multiplicatif suivant

Algorithme multiplicatif (erreur quadratique)

A chaque instant k, répéter

1. $\boldsymbol{\beta}_0^{(k)} = \boldsymbol{\alpha}^{(k)}$
2. $\boldsymbol{\beta}_n^{(k)} = \boldsymbol{\beta}_{n-1}^{(k)} \otimes \sum_{\ell=1}^{L_n} y_{n,\ell}\, \boldsymbol{x}_{n,\ell} \oslash \sum_{\ell=1}^{L_n} \boldsymbol{x}_{n,\ell}\, \boldsymbol{x}_{n,\ell}^\top\, \boldsymbol{\beta}_{n-1}^{(k)}$
3. $\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\beta}_N^{(k)}$

Il est intéressant et rassurant de noter que l'algorithme ci-dessus fait intervenir la matrice de corrélation empirique des données d'entrée, ainsi que le vecteur d'intercorrélation empirique entre les entrées et sorties. On constate enfin que les deux algorithmes développés au cours de cette section se confondent lorsqu'une unique donnée est disponible par capteur, au moment de la mise à jour. L'instruction 2. s'écrit alors simplement

$$\boldsymbol{\beta}_n^{(k)} = \frac{y_{n,\ell}}{\boldsymbol{x}_{n,\ell}^\top\, \boldsymbol{\beta}_{n-1}^{(k)}}\, \boldsymbol{\beta}_{n-1}^{(k)}. \tag{39}$$

Ces éléments seront approfondis dans un prochain article.

V. Méthodes d'identification non-linéaires

La complexité des applications envisagées, telles que la diffusion de chaleur mesurée par rayonnement thermique qui illustre la fin de cet article, nécessite de recourir à des méthodes d'identification adéquates. La présente section vise à dépasser le cadre restrictif des modèles linéaires pour considérer des formes $y_\ell = \psi(\boldsymbol{x}_\ell)$ plus générales, où ψ appartient à un espace fonctionnel $\mathcal{H}$ qu'il convient de définir attentivement afin de rendre les calculs aisés. Fondées sur les travaux précurseurs d'Aronszajn [1], les avancées de l'estimation fonctionnelle basée sur les espaces de Hilbert à noyau reproduisant, et de la théorie de la régularisation, ont apporté des réponses convaincantes aux problèmes impliquant des systèmes non-linéaires. La stratégie d'apprentissage considérée ci-après reposent sur cette théorie.

A. Principe général

On considère un espace de Hilbert $\mathcal{H}$ à noyau reproduisant constitué de fonctions réelles sur un compact $\mathcal{X} \subset \mathbb{R}^p$.

On note $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ son produit scalaire. Soit $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ le noyau de cet espace, qui vérifie par conséquent la propriété reproduisante suivante : $\psi(\boldsymbol{x}_\ell) = \langle \psi, \kappa(\boldsymbol{x}_\ell, \cdot) \rangle_{\mathcal{H}}$ pour toute fonction ψ de $\mathcal{H}$ et tout $\boldsymbol{x}_\ell$ de $\mathcal{X}$. Cette dernière étant vérifiée par le noyau lui-même, on a évidemment

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot) \rangle_{\mathcal{H}}, \tag{40}$$

ce qui justifie l'appellation de noyau reproduisant. Le noyau gaussien, donné par $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / 2\sigma_0^2)$ avec σ_0 la largeur de bande, en est un exemple.

Etant donné un ensemble $\{(\boldsymbol{x}_\ell, y_\ell)\}_{\ell=1,\dots,L}$ de données d'apprentissage i.i.d., on considère le problème de minimisation d'un critère J entre les sorties du modèle $\psi(\boldsymbol{x}_\ell)$ et les réponses désirées y_ℓ

$$\psi^o = \arg\min_{\psi \in \mathcal{H}} J\big((\psi(\boldsymbol{x}_\ell), y_\ell)_{\ell=1,\dots,L}\big). \tag{41}$$

Le Théorème de Représentation Généralisé [16] établit que la solution ψ^o peut être recherchée dans l'espace engendré par les L fonctions noyau $\kappa(\boldsymbol{x}_\ell, \cdot)$, c'est-à-dire

$$\psi^o(\boldsymbol{x}) = \sum_{\ell=1}^{L} \alpha_\ell^o \, \kappa(\boldsymbol{x}_\ell, \boldsymbol{x}), \tag{42}$$

dont il faut déterminer les coefficients $\boldsymbol{\alpha}^o = [\alpha_1^o \dots \alpha_L^o]^\top$ optimaux au sens de J. Ainsi note-t-on que le problème devient linéaire par rapport à ces derniers, permettant de reprendre le formalisme développé jusqu'ici. En particulier, pour les données poissoniennes et gaussiennes vues précédemment, les fonctions coût associées s'écrivent respectivement

$$J(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \sum_{\ell=1}^{L_n} \Big((\boldsymbol{\alpha}^\top \boldsymbol{\kappa}_{n,\ell} - y_{n,\ell}) + y_{n,\ell} \, \log \frac{y_{n,\ell}}{\boldsymbol{\alpha}^\top \boldsymbol{\kappa}_{n,\ell}} \Big)$$

$$J(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{\ell=1}^{L_n} \frac{(\boldsymbol{\alpha}^\top \boldsymbol{\kappa}_{n,\ell} - y_{n,\ell})^2}{\sigma_n^2}$$

avec $\boldsymbol{\kappa}_{n,\ell} = [\kappa(\boldsymbol{x}_{n,\ell}, \boldsymbol{x}_{1,1}) \, \kappa(\boldsymbol{x}_{n,\ell}, \boldsymbol{x}_{1,2}) \dots \kappa(\boldsymbol{x}_{n,\ell}, \boldsymbol{x}_{N,L_N})]^\top$ conformément au Théorème de Représentation Généralisé. La contrainte de positivité de $\psi^o(\boldsymbol{x})$, pour tout $\boldsymbol{x}$ de $\mathcal{X}$, se traduit alors par l'usage de noyaux positifs et la recherche d'une solution $\boldsymbol{\alpha}^o$ positive.

B. Critères de parcimonie

On note que le modèle linéaire généralisé (42) compte autant de termes qu'il y a de données disponibles, limitant son intérêt pratique. Il s'agit là d'une difficulté rencontrée de manière récurrente avec les méthodes à noyau reproduisant. La vaste littérature sur le sujet préconise généralement l'usage de modèles approchés à taille réduite

$$\psi^{(k)}(\boldsymbol{x}) = \sum_{i=1}^{|\mathcal{D}|} \alpha_i^{(k)} \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}) \tag{43}$$

où les éléments $\boldsymbol{x}_{\omega_i}$ composent un dictionnaire $\mathcal{D}$. Ce dernier est adapté au cours des itérations, à partir d'un critère complémentaire de parcimonie. Voir [14] et bibliographie incluse. On s'en tient ici à une règle simple, tout en sachant que de nombreux raffinements sont possibles. Lors de l'acquisition de tout nouveau couple de données $(\boldsymbol{x}_\ell, y_\ell)$, on s'attache dans un premier temps à la mise à jour des coefficients constituant le modèle étendu

$$\psi^{(k)}(\boldsymbol{x}) + \epsilon \, \kappa(\boldsymbol{x}_\ell, \boldsymbol{x}) \tag{44}$$

avec $\epsilon = \max\{(y_\ell - \psi^{(k)}(\boldsymbol{x}_\ell))/\kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_\ell); 0\}$. On utilise l'un des algorithmes présentés au cours des sections précédentes pour cela, itératif ou multiplicatif, distribué ou non selon le contexte. Dans un second temps, afin de limiter la taille du dictionnaire $\mathcal{D}$, on ne retient finalement que les composantes excédant un seuil ν_0 donné.

VI. Expérimentations

Pour illustrer la pertinence des algorithmes proposés, on s'intéresse à l'estimation d'un champ de température régi par l'équation différentielle suivante

$$\frac{\partial \Theta(\boldsymbol{x}, t)}{\partial t} - c \nabla_{\boldsymbol{x}}^2 \Theta(\boldsymbol{x}, t) = Q(\boldsymbol{x}, t)$$

à partir du rayonnement thermique associé. Dans cette expression, $\Theta(\boldsymbol{x}, t)$ est la température dépendant de la position $\boldsymbol{x}$ et du temps t, $\nabla_{\boldsymbol{x}}^2 \Theta(\boldsymbol{x}, t)$ l'opérateur spatial de Laplace, et $Q(\boldsymbol{x}, t)$ la quantité de chaleur apportée. On considère une surface rectangulaire, de conductivité thermique $c = 0.1$, sur laquelle N capteurs sont tirés uniformément sur une grille de taille 21×21. Deux sources de chaleur de 200 W sont activées successivement, la première des instants $t = 1$ à $t = 100$ située dans le quart-plan inférieur-droit, la seconde de $t = 100$ à $t = 200$ située dans le quart-plan supérieur-gauche.

L'objectif de la simulation proposée est, étant donné des mesures $d_{n,\ell}$ telles que

$$d_{n,\ell} \sim \mathcal{P}(\Theta(\boldsymbol{x}_n, t_\ell)) \tag{45}$$

avec $\mathcal{P}(\lambda)$ un bruit de Poisson de paramètre λ traduisant un processus de détection de photons par un capteur, d'estimer $\Theta(\boldsymbol{x}, t)$ via $\psi^{(t)}(\boldsymbol{x})$. Les simulations ont été réalisées avec l'algorithme multiplicatif distribué, développé pour les données poissoniennes, associé à un noyau gaussien de largeur de bande σ égal à 0.18. Le seuil ν_0 destiné à sélectionner les éléments les plus significatifs pour le dictionnaire a été fixé à $\max_i(\alpha_i)/10$.

Comme l'indique la figure 1, l'algorithme incrémental suit parfaitement les évolutions de la distribution de chaleur. Le choix des capteurs représentés dans le modèle via le dictionnaire $\mathcal{D}$, cerclés de rouge sur chaque représentation, est approprié. On rappelle que leur nombre, ici restreint, correspond à la taille du modèle. La convergence de l'erreur quadratique moyenne normalisée, définie par

$$\frac{\sum_{n=1}^{N} (d_{n,\ell} - \psi^{(\ell)}(\boldsymbol{x}_n))^2}{\sum_{n=1}^{N} d_{n,\ell}^2}.$$

est illustrée par la figure 2. Elle est tracée pour différentes valeurs du nombre N de capteurs présents sur la zone surveillée. On constate que, plus ils sont nombreux dans une certaine limite, plus l'erreur converge rapidement. Il est à noter enfin la croissance de cette erreur à l'instant $t = 100$, provoquée par l'extinction de la première source et l'allumage de la seconde.

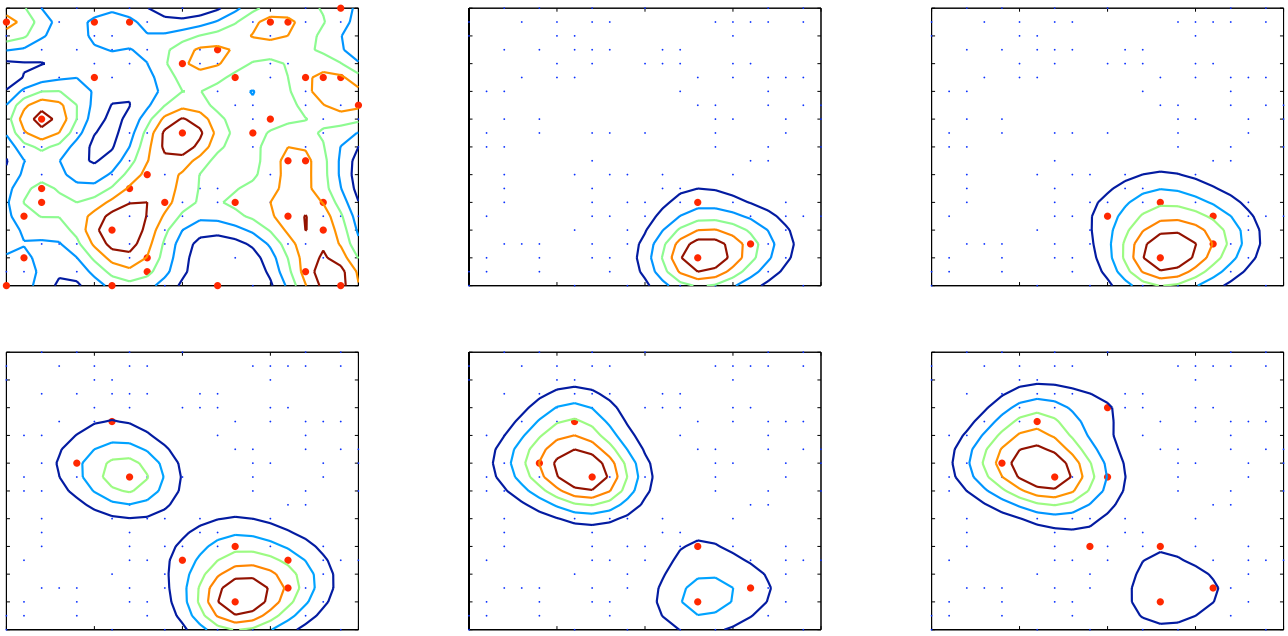

Fig. 1. Champs de température estimés à différents instants. De gauche à droite, et de haut en bas : $t = 1$ (init.), $t = 50$, $t = 100$ (changement de source), $t = 110$, $t = 150$, $t = 200$. Les capteurs appartenant au dictionnaire sont marqués d'un point rouge, les autres d'un point bleu.

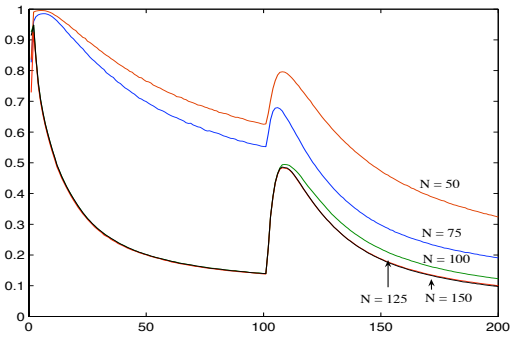

Fig. 2. Evolution de l'erreur quadratique moyenne normalisée au cours du temps, pour différentes valeurs du nombre N de capteurs disponibles sur la zone surveillée.

VII. Conclusion

Nous avons décrit une nouvelle classe d'algorithmes d'apprentissage non-linéaires en ligne avec contrainte de positivité sur la solution. Ceux-ci ont été appliqués au problème d'identification distribuée d'un champ scalaire positif par un réseau de capteurs. Les perspectives de ce travail concernent l'étude de la convergence de ces méthodes et, compte tenu de leur caractère générique vis-à-vis du critère optimisé, leur mise en œuvre avec d'autres divergences.

Références

[1] Aronszajn, N. *Theory of reproducing kernels.* Trans. Amer. Math. Soc., vol. 68, pp. 337-404, 1950.

[2] Bertsekas, D. *Nonlinear Programming.* Athena Scientific, New Jersey, 1999.

[3] F. Cattivelli, F. et Sayed, A. Diffusion LMS algorithms with information exchange. Workshop Asilomar'2008, Pacific Grove, USA, 2008.

[4] Daude-Witherspoon M. E. et Muehllehner, G. An iterative image space reconstruction algorithm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, vol. 5, n° 2, pp. 61-66, 1986.

[5] Fevotte, C., Bertin, N. et Durrieu, J.-L. Nonnegative matrix factorization with the Itakura-Saito divergence. Application to Music Analysis. *Neural Computation*, vol. 21, n° 3, pp. 793-830, 2009.

[6] Honeine, P., Essoloh, M., Richard, C. et Snoussi, H. Distributed regression in sensor networks with a reduced-order kernel model. Congrès IEEE Globecom'2008, New Orleans, USA, 2008.

[7] Honeine, P., Richard, C., Bermudez, J.-C., Essoloh, M., Snoussi, H. et Vincent, F. Functional estimation in Hilbert space for distributed learning in wireless sensor networks. Congrès IEEE ICASSP'2009, Taipei, Taiwan, 2009.

[8] Lopes, C. et Sayed, A. Diffusion least-mean squares over adaptive networks. Congrès IEEE ICASSP'2007, Hawaï, USA, 2007.

[9] Lopes, C. et Sayed, A. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, vol. 55, n° 8, pp. 4064-4077, 2007.

[10] Lantéri, H., Roche, M., Cuevas, O. et Aime, C. A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, vol. 81, n° 5, pp. 945-974, 2001.

[11] Lantéri, H., Theys, C., Benvenuto, F. et Mary, D. Méthode algorithmique de minimisation de fonctions d'écart entre champs de données. Application à la reconstruction d'images astrophysiques. Colloque GRETSI'2009, Dijon, France, 8-11 septembre 2009.

[12] Lucy, L. B. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, vol. 79, pp. 745-754, 1974.

[13] Rabbat, M. et Nowak, R. Distributed optimization in sensor networks. Congrès IPSN'2004, Berkeley, USA, 2004.

[14] Richard, C., Bermudez, J.-C. et Honeine, P. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, vol. 57, n° 3, pp. 1058-1067, 2009.

[15] Richardson, W.H. Bayesian based iterative method of image restauration. *Journal of the Optical Society of America*, vol. 62, n° 1, pp. 55-59, 1972.

[16] Schölkopf, B., Herbrich, R. et Williamson, R. A generalized Representer Theorem. NeuroCOLT, Royal Holloway College, Univ. London, UK, Tech. Rep. NC2-TR-2000-81, 2000.