A propos de l'utilisation des méthodes de Monte Carlo séquentielles pour l'estimation de paramètres dans les modèles de Markov cachés

Olivier Cappé & Eric Moulines
Centre National de la Recherche Scientifique & Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75634 Paris cedex 13
cappe, moulines à enst.fr

Résumé − A part dans quelques cas particuliers (modèles à espace d'état fini, modèles d'état linéaires gaussiens), l'estimation de paramètres au sens du maximum de vraisemblance dans les modèles de Markov cachés est une tâche difficile. La possibilité d'utiliser les techniques de Monte Carlo séquentielles (dites également de filtrage particulaire) à cette fin semble naturelle mais n'a, à ce jour, pas trouvé de concrétisation réellement convaincante. Dans cette communication, nous détaillons les liens entre l'estimation au sens du maximum de vraisemblance et la tâche de lissage de fonctionnelles additives de l'état caché. Nous proposons une solution simple permettant d'accroître la fiabilité des estimateurs en utilisant les propriétés d'oubli du filtre exact.

Abstract − Apart from a few specific cases (finite state space models, linear Gaussian state-space models), maximum likelihood parameter estimation in hidden Markov models is a difficult task. Using sequential Monte Carlo (or particle filtering) techniques for this task sounds appealing but is faced with some difficulties. In this contribution, we show that several recently proposed methods share the common feature of requiring the approximation of the expectation of a sum functional of the hidden states, conditionally on all the available observations. We propose a robustification of the basic particle estimator which is based on forgetting ideas.

1 Introduction

Les modèles de Markov cachés forment une classe suffisamment flexible pour avoir donné lieu à des applications très significatives dans des domaines aussi variés que le traitement de la parole, la bioinformatique, la poursuite et la localisation, la vision, les finances quantitatives, etc. [1] Les méthodes de Monte Carlo séquentielles (ou de filtrage particulaire) constituent un outil important dans ce cadre car elles forment la seule approche véritablement générique permettant d'approximer numériquement les relations de filtrage dès que l'on sort de cas spécifiques comme les modèles à état fini ou les modèles d'états linéaires gaussiens [2].

Paradoxalement, la question de l'estimation de paramètres fixes du modèle à l'aide des simulations obtenues par filtrage particulaire reste un problème difficile. Une des explications en est sans aucun doute le fait que bien que la log-vraisemblance des observations soit facilement estimable à l'aide du système de particules [3, 4], cette possibilité ne résout pas réellement la question de l'optimisation de la vraisemblance. En particulier, l'approximation de la log-vraisemblance n'est jamais régulière (vis-à-vis des variations des paramètres) dès lors qu'on procède à des rééchantillonnages [3], or on sait que le rééchantillonnage est un élément crucial pour garantir la stabilité des techniques de Monte Carlo séquentielles [1, 5].

Le but de cette contribution est de discuter la façon dont on peut, à l'aide de techniques de filtrage particulaire, approximer les quantités nécessaires à l'inférence statistique, qu'il s'agisse de la quantité intermédiaire de l'algorithme EM (Expectation-Maximization) ou du gradient de la log-vraisemblance (fonction de score). Pour dissiper tout malentendu, soulignons que nous ne nous intéressons pas ici à l'estimation récursive au sens, par exemple, de [6] même s'il est clair qu'être capable d'approximer récursivement les quantités d'intérêt pour l'inférence constitue effectivement un premier pas dans cette direction.

2 Notations et définitions

Nous commençons par définir formellement le modèle de Markov caché :

1. $\{X_k\}_{k\geq 0}$ est markovien de loi initiale ν et de noyau de transition (on ne travaille ici qu'avec des densités de probabilité) q tels que $\mathrm{E}[f(X_0)] = \int f(x)\,\nu(x)dx$ et $\mathrm{E}[f(X_n)|X_{0:n-1}] = \int f(x)\,q(X_{n-1},x)dx$, pour toute fonction f ; $X_{0:n-1}$ est notre notation simplifiée pour désigner l'ensemble des variables consécutives $X_0,\ldots,X_{n-1}$.
2. $\{Y_k\}_{k\geq 0}$ est conditionnellement indépendant sachant $\{X_k\}_{k\geq 0}$ de telle façon que
$$\mathrm{E}\left[\prod_{k=0}^{n} f_k(Y_k)\,\middle|\, X_{0:n}\right] = \prod_{k=0}^{n}\int f_k(y)\,g(X_k,y)dy$$
pour toutes fonctions $f_0,\ldots,f_n$; g est dite fonction de vraisemblance conditionnelle et dans la suite nous considérons toujours sa dépendance en son premier argument en écrivant $g_k(x)$ plutôt que $g(x,Y_k)$.

En utilisant la règle de Bayes, on vérifie directement que la densité jointe de lissage (loi de $X_{0:n}$ sachant $Y_{0:n}$) est donnée par

$$\phi_{0:n|n}(x_{0:n}) = \mathrm{L}_n^{-1}\nu(x_0)g_0(x_0)\prod_{k=1}^{n} q(x_{k-1},x_k)g_k(x_k) \quad (1)$$

où le facteur de normalisation L_n est la vraisemblance des observations $Y_{0:n}$. A partir de (1) on obtient aisément une formulation récursive (en n)

$$\phi_{0:n|n}(x_{0:n}) = c_n^{-1}\phi_{0:n-1|n-1}(x_{0:n-1})q(x_{n-1},x_n)g_n(x_n) \quad (2)$$

où $c_n = \mathrm{L}_n/\mathrm{L}_{n-1}$. Au vu de (2), le facteur de normalisation c_n peut également s'écrire

$$c_n = \iint \phi_{n-1|n-1}(x_{n-1})q(x_{n-1},x_n)g_n(x_n)dx_{n-1}dx_n \quad (3)$$

où $\phi_{n|n}$ désigne la loi marginale de filtrage (densité de probabilité de X_n sachant $Y_{0:n}$). Cette relation associée à l'identité

$$\ell_n \stackrel{\text{def}}{=} \log \mathrm{L}_n = \sum_{k=0}^{n} \log c_k \quad (4)$$

montre que l'évaluation de la log-vraisemblance est, pour les modèles de Markov cachés, un sous-produit des relations de filtrage.

Le filtrage particulaire approxime (2) à l'aide du mécanisme de simulation suivant :

- Initialement, on tire N particules $\{\xi_0^i\}_{1\leq i\leq N}$ sous une loi commune ρ_0 et l'on calcule les poids d'importance $\omega_0^i = \nu(\xi_0^i)g_0(\xi_0^i)/\rho_0(\xi_0^i)$.
- Par la suite, ξ_{k+1}^i est simulé à l'aide d'un noyau de transition $r(\xi_k^i,\cdot)$ et les poids sont mis à jour selon $\omega_{k+1}^i = \omega_k^i q(\xi_k^i,\xi_{k+1}^i)g_{k+1}(\xi_{k+1}^i)/r(\xi_k^i,\xi_{k+1}^i)$. La trajectoire de la $i$ème particule est simplement prolongée par $\xi_{0:k+1}^i = (\xi_{0:k}^i,\xi_{k+1}^i)$.
- Le rééchantillonnage consiste à normaliser les poids par leur somme de telle façon que $\sum_{i=1}^N \omega_k^i = 1$ puis à tirer des indices $I_k^1,\ldots,I_k^N$ dans l'ensemble $\{1,\ldots,N\}$ tels que $\mathrm{E}(\#\{1\leq j\leq N : I_k^j = i_0\}) = N\omega_k^{i_0}$. Les poids redeviennent alors tous égaux et la $j$ème trajectoire est mise à jour par recopie : $\xi_{0:k}^j = \xi_{0:k}^{I_k^j}$.

Du fait de ce rééchantillonnage, la notation ξ_k^i devient ambiguë et on utilise la notation $\xi_{0:k}^i(l)$ pour désigner le point d'indice l dans la $i$ème trajectoire finissant en k ; pour $\xi_{0:k}^i(k)$ on conserve cependant la notation ξ_k^i. L'estimateur

$$\sum_{i=1}^{N} \frac{\omega_k^i}{\sum_{j=1}^N \omega_k^j} f(\xi_{0:k}^i) \quad (5)$$

constitue une approximation de $\mathrm{E}\left[f(X_{0:k})\,|\,Y_{0:k}\right]$ pour toute fonction f.

Par rapport à d'autres techniques plus classiques d'approximation des relations de filtrage non-linéaires, comme l'EKF (Extended Kalman Filter) ou l'UKF (Unscented Kalman Filter) [7], l'intérêt du filtrage particulaire est qu'il est possible de montrer que (5) est une estimation consistante lorsque le nombre N de particules augmente. Le rééchantillonnage joue un rôle primordial pour garantir que la méthode ne dégénère pas lorsque le nombre n d'observations augmente [1, 2, 5]. Il complique néanmoins significativement l'analyse théorique de la méthode en rendant les trajectoires de particules dépendantes.

3 Estimation au sens du maximum de vraisemblance

Du point de vue des principes, la remarque importante est ici que la détermination de l'estimateur du maximum de vraisemblance prend nécessairement la forme d'un algorithme itératif dans lequel l'étape clé revient à calculer (ou, à défaut, approximer) une quantité de la forme

$$\tau_n = \mathrm{E}\left[\sum_{k=0}^{n-1} s_k(X_k,X_{k+1}) \,\middle|\, Y_{0:n}\right] \quad (6)$$

où $s_0,\ldots,s_n$ sont des fonctions bien choisies[1].

Dans le cas de l'algorithme EM l'affirmation précédente provient simplement de la forme Markovienne de la loi jointe donnée par (1), au facteur de normalisation L_n près. Ainsi la quantité intermédiaire de l'algorithme EM s'écrit

$$\mathrm{E}\left[\sum_{k=0}^{n-1} s_k(X_k,X_{k+1}\,;\theta) \,\middle|\, Y_{0:n}\,;\theta'\right] \quad (7)$$

avec

$$s_k(x_k,x_{k+1}\,;\theta) = \begin{cases} \log\left[\nu(x_0\,;\theta)g_0(x_0\,;\theta)q(x_0,x_1\,;\theta)g_1(x_1\,;\theta)\right] & \text{si } k=0 \\ \log\left[q(x_k,x_{k+1}\,;\theta)g_{k+1}(x_{k+1}\,;\theta)\right] & \text{sinon} \end{cases} \quad (8)$$

Notons que dans les expression ci-dessus, on fait figurer explicitement la dépendance des différentes quantités vis à vis du paramètre θ et de la valeur courante θ' de ce dernier pour éviter l'ambiguïté.

On peut également songer à utiliser un algorithme d'optimisation itérative directe de la log-vraisemblance sans passer par l'intermédiaire défini en (7) ci-dessus. Sauf dans les cas très simples, il est dans ce cas nécessaire d'évaluer le gradient de la log-vraisemblance pour garantir une convergence raisonnablement rapide de l'optimisation (l'alternative consistant à utiliser une méthode reposant uniquement sur l'évaluation de la log-vraisemblance–méthode qui devient de plus difficile à envisager lorsque l'évaluation de la log-vraisemblance est entachée d'erreurs Monte Carlo). La relation dite de Fisher indique que le gradient de la log-vraisemblance s'écrit

$$\mathrm{E}\left[\sum_{k=0}^{n-1} \nabla_\theta s_k(X_k,X_{k+1}\,;\theta) \,\middle|\, Y_{0:n}\,;\theta\right] \quad (9)$$

où ∇_θ désigne la dérivation par rapport à θ et les s_k sont les fonctions définies en (8). Il existe manifestement une autre approche, plus directe, qui permet de déterminer le gradient de la log-vraisemblance consistant à dériver la représentation (3)–(4) par rapport au paramètre θ. On obtient alors une forme récursive du gradient de la log-vraisemblance dite des "équations de sensibilité" [10]. On peut toutefois montrer que ces équations de sensibilité sont équivalentes à une réécriture récursive de (9) [1, 11].

[1] Une exception importante est le cas de la méthode du rapport de vraisemblance simulé de [8] qui toutefois ne résout pas explicitement le problème de l'optimisation. De surcroît, les résultats obtenus par cette approche sont peu fiables en pratique pour les raisons exposées dans [9].

Ainsi si l'on excepte la question de la reformulation récursive, dont on verra ci-dessous qu'elle est traitée de façon très intuitive dans le cadre de l'approximation particulaire, (9) correspond bien à la forme prise par le gradient de la log-vraisemblance dans le modèle de Markov caché.

4 Approximation particulaire

La tâche qui nous est fixée consiste donc à approximer des expressions de la forme générique (6) dans le contexte des méthodes de Monte Carlo séquentielles. L'approximation naturelle de (6) à partir du système de particules consiste, cf. (5), à utiliser l'estimateur

$$\hat{\tau}_n^N = \sum_{i=1}^N \frac{\omega_n^i}{\sum_{j=1}^N \omega_n^j} \sum_{k=0}^{n-1} s_k \left[\xi_{0:n}^i(k), \xi_{0:n}^i(k+1)\right] \qquad (10)$$

dont on constate facilement qu'il peut se calculer récursivement (et avec une capacité de stockage fixe, indépendante de l'horizon n) en stockant pour chaque trajectoire i ($1 \leq i \leq N$) non seulement la position finale ξ_n^i et le poids trajectoriel ω_n^i mais également l'évaluation cumulée de la fonctionnelle d'intérêt le long de la trajectoire

$$\gamma_n^i = \sum_{k=0}^{n-1} s_k \left[\xi_{0:n}^i(k), \xi_{0:n}^i(k+1)\right]$$

Pour des choix particuliers de $\{s_k\}_{k\geq 0}$ (cf. discussion dans la section précédente), on retrouve en procédant de cette façon les algorithmes d'approximation particulaire proposés dans [12, 13, 14].

De façon, pour l'instant, empirique, nous avons constaté en expérimentant avec cette approche sur plusieurs type de modèles qu'une modification assez simple de (10) permettait d'obtenir un estimateur de (6) nettement plus fiable. Cette modification consiste à viser non pas (6) mais

$$\tau_n^\delta = \sum_{k=0}^{n-1} \mathrm{E}\left[s_k\left(X_k, X_{k+1}\right) | Y_{0:(k+\delta)\wedge n}\right] \qquad (11)$$

qui sera approximé par

$$\hat{\tau}_n^{\delta,N} = \sum_{i=1}^N \frac{\omega_n^i}{\sum_{j=1}^N \omega_n^j} \sum_{k=0}^{n} s_k \left[\xi_{0:(k+\delta)\wedge n}^i(k), \xi_{0:(k+\delta)\wedge n}^i(k+1)\right] \qquad (12)$$

où $a \wedge b$ désigne le minimum de a et de b et δ est un délai. Comme dans le cas de (10), on vérifie aisément que (12) peut être mis à jour récursivement en stockant un "cache" de l'histoire récente des trajectoires de particules $\{\xi_{0:n}^i(n-\delta+1:n)\}_{1\leq i\leq n}$ ainsi que la contribution cumulée des termes destinés *à ne plus être mis à jour* :

$$\sum_{k=0}^{n-\delta} s_k \left[\xi_{0:k+\delta}^i(k), \xi_{0:k+\delta}^i(k+1)\right]$$

Cette idée de remplacer les lois de lissage par des lois de lissage à délai fixe (en anglais, "fixed-lag smoothing") n'est pas nouvelle [6]. Il est important toutefois de souligner qu'elle est très naturelle dans le cas de l'approximation particulaire puisque l'examen des preuves existantes de la stabilité (en le nombre d'échantillons n) des techniques de filtrage particulaire montre que c'est précisément ce point (convergence rapide des lois de lissage à délai fixe vers les lois de lissage, lorsque le délai δ augmente) qui joue un rôle clé [1, 5]. En pratique, le choix de δ relève véritablement d'un compromis de type biais/variance : si δ est faible, il peut exister un biais important entre $\hat{\tau}_n^{\delta,N}$ et τ_n ; mais plus δ augmente plus la variabilité de $\hat{\tau}_n^{\delta,N}$ croît. Nous illustrons ce point sur l'exemple considéré dans le chapitre 11 de [1].

5 Un exemple

L'exemple en question est celui du modèle de volatilité stochastique utilisé en économétrie qui correspond à une représentation d'état dont l'équation d'observation est fortement non-linéaire :

$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma U_k \ , & U_k &\sim \mathrm{N}(0,1) \\ Y_k &= \beta \exp(X_k/2) V_k \ , & V_k &\sim \mathrm{N}(0,1) \end{aligned}$$

Nous considérons une série d'observations de longueur 945 utilisée classiquement dans la littérature pour laquelle la valeur de l'estimateur du maximum de vraisemblance des paramètres a été évalué (numériquement) à $\hat{\beta} = 0.64$, $\hat{\phi} = 0.975$, $\hat{\sigma} = 0.17$.

Nous présentons ci-après les trajectoires d'itérations successives de l'algorithme dit SEM (Stochastic EM) dans lequel on alterne entre

- L'approximation de l'espérance conditionnelle de fonctionnelles additives qui dans ce modèle sont au nombre de quatre :

$$t_0(x_{0:n}) = x_0^2,\ t_1(x_{0:n}) = \sum_{k=0}^{n-1} x_k^2,\ t_2(x_{0:n}) = \sum_{k=1}^{n} x_k^2$$

$$t_3(x_{0:n}) = \sum_{k=1}^{n} x_k x_{k-1},\ s_4(x_{0:n}) = \sum_{k=0}^{n} Y_k^2 \exp(-x_k)$$

- La réestimation des paramètres β, ϕ et σ à partir des estimations obtenues dans l'étape précédente par une formule qui correspond à l'étape M de l'algorithme EM.

On sait que cet algorithme (le SEM) n'est pas un algorithme convergent, ceci-dit une modification assez simple (le SAEM, pour Stochastic Approximation EM) permet de garantir la convergence vers un point stationnaire de la log-vraisemblance [1]. L'idée est ici simplement d'illustrer la variabilité due aux simulations qui se manifeste de façon plus visible avec l'algorithme SEM dans la mesure où celui-ci ne comporte aucun effet de moyennage destiné à garantir la convergence ponctuelle.

On constate facilement en comparant les figures 1 et 2, sur lesquelles on a superposé trois trajectoires indépendantes pour donner une idée de la variabilité, que l'utilisation de l'approximation à délai fixe permet de réduire grandement la variance sans augmenter le biais de façon visible. Pour donner une idée du gain obtenu par cette approche on a représenté sur la figure 3 l'analogue de la

figure 1 obtenu avec un nombre 10 fois supérieur de particules (c'est à dire concrètement un coût de calcul 10 fois plus élevé). On constate que les figures 3 et 2 sont relativement comparables ce qui montre que le gain obtenu dans le cas de la figure 2 par utilisation de l'approximation à délai fixe est réellement très significatif.

Pour les autres modèles sur lesquels cette approche a été mise en œuvre les gains obtenus sont du même ordre de grandeur. Il est intéressant de souligner que la forme d'approximation avec oubli décrite dans cette contribution est également potentiellement plus simple à étudier d'un point de vue théorique dans la mesure où elle n'implique qu'une mémoire finie de l'historique du système de particules.

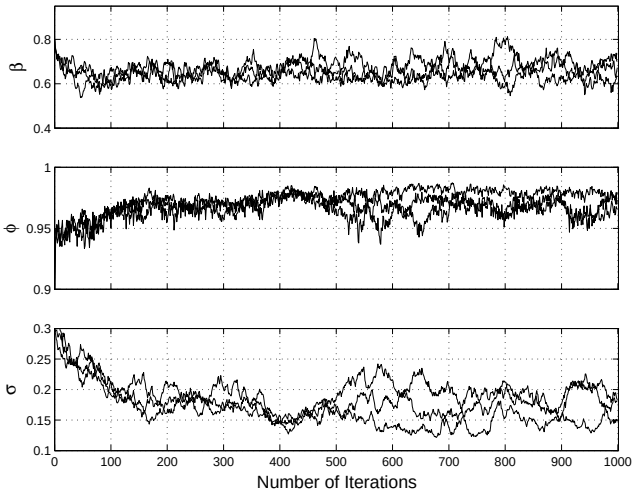

FIG. 1 – Trois trajectoires superposées de l'algorithme SEM avec $N = 50$ particules (paramètres β, ϕ et σ de haut en bas).

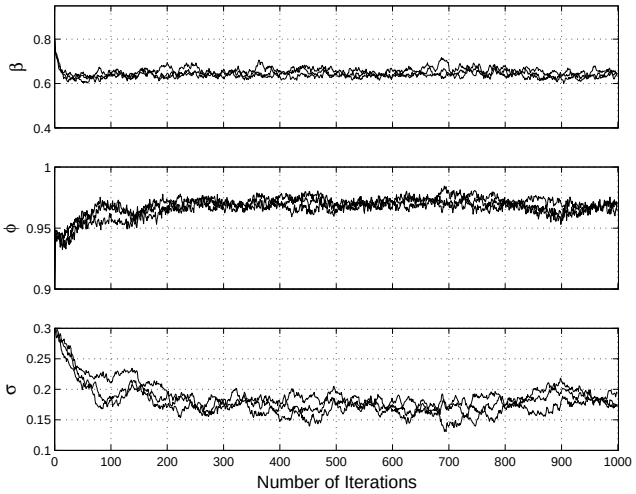

FIG. 2 – Même représentation avec $N = 50$ particules et *en utilisant l'approximation à délai fixe avec* $\delta = 20$.

Références

[1] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

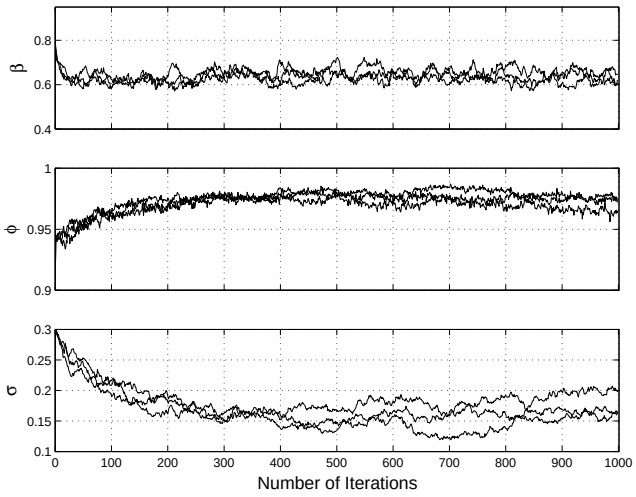

FIG. 3 – Méthode de la Fig. 1 avec $N = 500$ particules.

[2] A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[3] M. K. Pitt and N. Shephard. Filtering via simulation : Auxiliary particle filters. 94(446) :590–599, 1999.

[4] M. Hürzeler and H. R. Künsch. Monte Carlo approximations for general state-space models. *J. Comput. Graph. Statist.*, 7 :175–193, 1998.

[5] P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[6] V. Krishnamurthy and J. B. Moore. On-line estimation of hidden markov model parameters based on the kullback-leibler information measure. *IEEE Trans. Signal Process.*, 41(8) :2557–2573, 1993.

[7] B. Ristic, M. Arulampalam, and A. Gordon. *Beyond Kalman Filters : Particle Filters for Target Tracking*. Artech House, 2004.

[8] C. J. Geyer and E. A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B*, 54(3) :657–699, 1992.

[9] R. Douc, O. Cappé, E. Moulines, and C. P. Robert. On the convergence of the monte carlo maximum likelihood method for latent variable models. *Scand. J. Statist.*, 29(4), 2002.

[10] F. Campillo and F. Le Gland. MLE for patially observed diffusions : Direct maximization vs. the EM algorithm. *Stoch. Proc. App.*, 33 :245–274, 1989.

[11] O. Cappé and E. Moulines. Recursive computation of the score and observed information matrix in hidden markov models. In *IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, jul 2005.

[12] O. Cappé. Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods Appl.*, 7(1-2) :81–92, 2001.

[13] A. Doucet and V. B. Tadić. Parameter estimation in general state-space models using particle methods. *Ann. Inst. Statist. Math.*, 55(2) :409–422, 2003.

[14] F. Cérou, F. Le Gland, and N. Newton. Stochastic particle methods for linear tangent filtering equations. In J.-L. Menaldi, E. Rofman, and A. Sulem, editors, *Optimal Control and PDE's*. IOS Press, Amsterdam, 2001.