

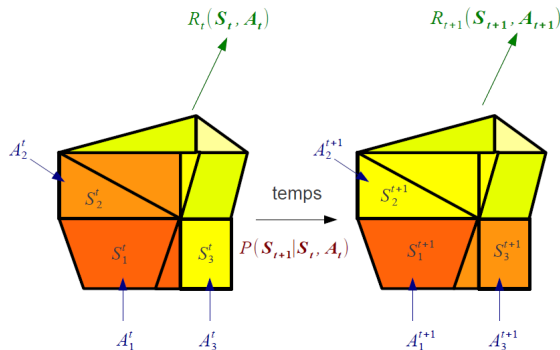
Résolution approchée de processus décisionnels de Markov factorisés

Julia Radoszycki
Nathalie Peyrard, Régis Sabbadin, Sabrina Gaba

séminaire MIAT

12 Septembre 2014

Gestion de phénomènes spatio-temporels stochastiques à l'échelle du paysage

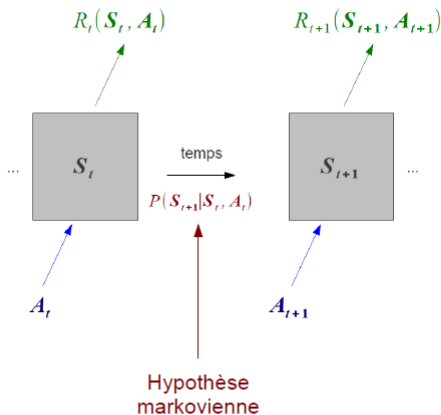


Exemple (problème épidémio) :

S_i^t = présence/absence maladie sur parcelle i

A_i^t = traitement ou non de la maladie sur parcelle i

Processus décisionnel de Markov (PDM)

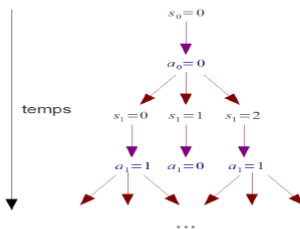


Données :

Fonction de transition

Fonction de récompense

Politique : $\delta = \{\delta_t : S_t \rightarrow A_t\}$

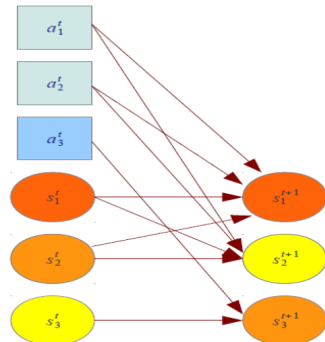
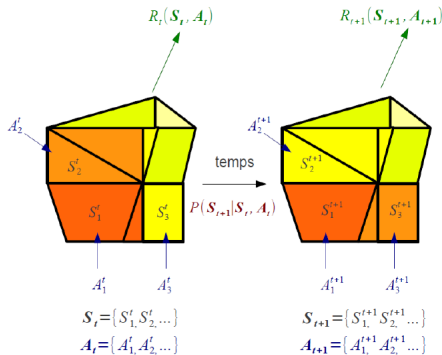


Valeur de la politique :

$$V^\delta(S_0) = E \left[\sum_{t=0}^{T-1} R_t(S_t, A_t) + R_T(S_T) \mid S_0, \delta \right]$$

Résoudre le PDM : trouver δ^* telle que $\forall \delta, \forall S_0, V^{\delta^*}(S_0) \geq V^\delta(S_0)$

Processus décisionnel de Markov à espace d'état et d'action factorisés (PDMF-AF)

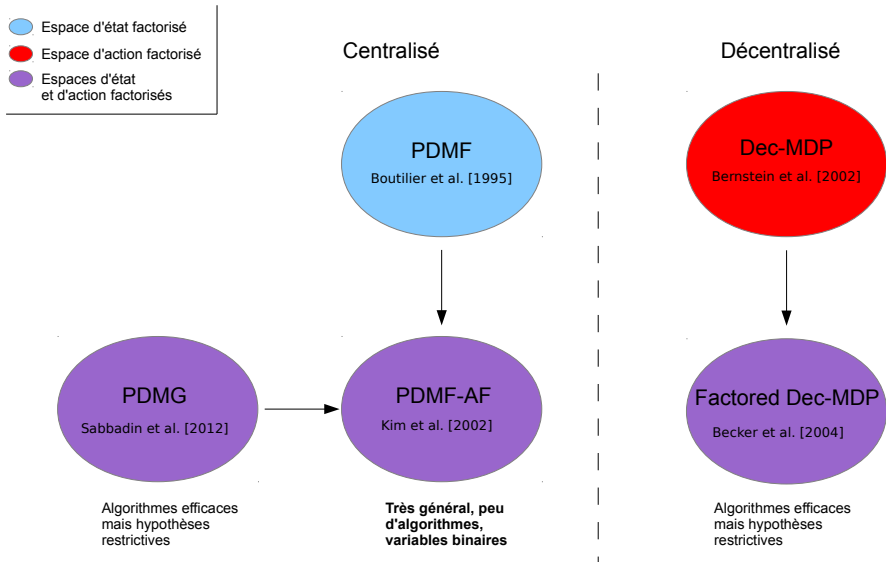


Représentation factorisée sous forme de réseau bayésien dynamique (DBN)

$$P(S^{t+1}|S^t, A^t) = \prod_{i=1}^n P(S_i^{t+1}|pa(S_i^{t+1}))$$

Objectif : Résolution approchée de PDMF-AF

Etat de l'art pour les PDMs factorisés : cadres



Quel espace des politiques considérer ?

- Un PDMF-AF est un PDM particulier
→ en théorie, on peut trouver une politique déterministe optimale pour tout état initial
- Quand $\mathcal{S}$ et $\mathcal{A}$ sont grands, une politique $\delta : \mathcal{S} \rightarrow \mathcal{A}$ ne peut pas être représentée explicitement
→ nous choisissons de la représenter de manière factorisée
- Mais les politiques optimales d'un PDMF-AF ne sont pas forcément factorisées. De ce fait, les politiques factorisées optimales
 - peuvent dépendre de l'état initial
 - peuvent être stochastiques et non déterministes
 - peuvent être non stationnaires
 - peuvent être non markoviennes

Notre approche

Choix

- considérer une distribution factorisée donnée sur l'état initial
- rechercher dans l'espace des politiques factorisées stochastiques (structure donnée)
- hypothèse de récompense additive

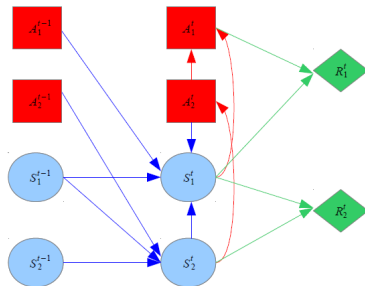
Méthode

- utiliser un algorithme d'optimisation continue (locale)
- utiliser une méthode variationnelle pour l'évaluation approchée de la fonction objectif

→ un algorithme de résolution approchée pour les PDMF-AF

PDMF-AFs et politiques factorisées stochastiques

- Distribution initiale factorisée donnée :
 $P^0(s^0) = \prod_{i=1}^n P_i^0(s_i^0)$
- Recherche de politiques factorisées stochastiques dont la structure est donnée : $\delta(a^t | s^t) = \prod_{j=1}^m \delta_j(a_j^t | pa_\delta(a_j^t))$
 où $pa_\delta(a_j^t) \subset \{s_i^t, a_k^t, k \neq j\}$
- Récompense additive :
 $R^t(s^t, a^t) = \sum_{\alpha=1}^r R_\alpha^t(pa_R(R_\alpha^t))$, où
 $pa_R(R_\alpha^t) \subseteq \{s_i^t, a_j^t\}$



A politique fixée, la distribution de probabilité sur les trajectoires d'horizon fini $(s, a)^{0:t} = \langle s^0, a^0, \dots, s^t, a^t \rangle$ est un DBN :

$$P_\delta^t((s, a)^{0:t}) = P^0(s^0) \delta(a^0 | s^0) \prod_{t'=1}^t \left(\prod_{i=1}^n P_i(s_i^{t'} | pa_P(s_i^{t'})) \prod_{j=1}^m \delta_j(a_j^{t'} | pa_\delta(a_j^{t'})) \right)$$

Evaluation de politiques factorisées stochastiques

La valeur d'une politique factorisée stochastique δ pour P^0 donnée s'exprime en fonction de **marginales** de la distribution $P_\delta^T((s, a)^{0:T})$:

$$\begin{aligned}
 V_\delta^{R,T}(P^0) &= \mathbb{E}_{P_\delta^T} \left[\sum_{t=0}^T R^t(S^t, A^t) \middle| P^0, \delta \right] \\
 &= \sum_{t=0}^T \sum_{\alpha=1}^r \mathbb{E}_{P_\delta^T} \left[R_\alpha^t(pa_R(R_\alpha^t)) \middle| P^0, \delta \right] \\
 &= \sum_{t=0}^T \sum_{\alpha=1}^r \sum_{pa_R(R_\alpha^t)} b_\alpha^t(pa_R(R_\alpha^t)) R_\alpha^t(pa_R(R_\alpha^t))
 \end{aligned}$$

où $b_\alpha^t(pa_R(R_\alpha^t)) = \sum_{(s,a)^{0:T} \setminus (pa_R(R_\alpha^t))} P_\delta^T((s, a)^{0:T})$

→ estimation de ces marginales par des **méthodes variationnelles** (*mean field approximation, loopy belief propagation* ...)

→ pas de garanties théoriques mais bons résultats en pratique

→ plus rapide que Monte-Carlo

Le problème d'optimisation

Problème 1

$$\begin{aligned} & \underset{\delta \in (\mathbb{R}^+)^N}{\text{maximize}} && V_{\delta}^{R,T}(P^0) \\ & \text{subject to} && \sum_{a_j \in \mathcal{A}_j} \delta_j(a_j | pa_{\delta}(a_j)) = 1 \quad \forall j, \forall pa_{\delta}(a_j) \end{aligned}$$

où $N = \sum_{j=1}^m |\mathcal{A}_j| \prod_{k/A_k \in pa_{\delta}^A(A_j)} |\mathcal{A}_k| \prod_{k/S_k \in pa_{\delta}^S(A_j)} |S_k|$.

→ problème de maximisation convexe, problème NP^{PP} -difficile

Reformulation : Problème 2

Problème 2 (diminution du nombre de variables)

$$\text{maximize}_{\delta \in (\mathbb{R}^+)^{N'}} V_{\delta}^{R,T}(P^0)$$

$$\text{subject to} \quad \sum_{a_j \in \{1, \dots, |\mathcal{A}_j| - 1\}} \delta_j(a_j | pa_{\delta}(a_j)) \leq 1 \quad \forall j, \forall pa_{\delta}(a_j)$$

$$\delta_j(|\mathcal{A}_j| | pa_{\delta}(a_j)) = 1 - \sum_{a_j \in \{1, \dots, |\mathcal{A}_j| - 1\}} \delta_j(a_j | pa_{\delta}(a_j)) \quad \forall j, \forall pa_{\delta}(a_j)$$

$$\text{où } \delta^- = \{\delta_j(a_j | pa_{\delta}(a_j)), \forall j, \forall a_j \neq |\mathcal{A}_j|, \forall pa_{\delta}(a_j)\},$$

$$N' = \sum_{j=1}^m (|\mathcal{A}_j| - 1) \prod_{k/A_k \in pa_{\delta}^A(A_j)} |\mathcal{A}_k| \prod_{k/S_k \in pa_{\delta}^S(A_j)} |\mathcal{S}_k|.$$

→ dans le cas de variables d'action binaires, $N' = N/2$

Implémentation d'un premier algo : CA-LBP

- variables d'action binaires
- évaluation par **loopy belief propagation (LBP)**
- algorithme **coordinate ascent (CA)** sur problème 2 avec $N' = N/2$ variables → garantie de convergence vers un maximum local si l'évaluation était exacte

→ comparaison avec MF-API (Sabbadin et al. [2012]), algorithme de résolution approchée des GMDPs

Résultats : petits problèmes

TABLE: Results on 100 random GMDPs and FA-FMDPs. MRE=mean relative error. Scores between brackets are scores for the deterministic policies derived from the SFPs.

		CA-LBP	MF-API	optimum global
GMDP	MRE	0.0065 (0.0066)	0.0015	-
	mean time	4.45s	1.79s	115.16s
FA-FMDP	MRE	0.1228 (0.1230)	-	-
	mean time	91.8s	-	117s

Résultats : gros problèmes

TABLE: Results on the disease management problem (GMDP).

		CA-LBP	MF-API	greedy
5×5	MC value	22434 (+61%)	22434	13932
	exec. time	15min	0.82s	-
10×10	MC value	89102 (+62%)	89102	54866
	exec. time	6h34	3.58s	-

TABLE: Results on the disease management problem (FA-FMDP).

		CA-LBP	greedy	unif.	rand.
5×5	MC value	19142 (+46%)	13092	10436	10801
	exec. time	1h16	-	-	-
10×10	MC value	69175 (+32%)	52191	40495	40014
	exec. time	43h42	-	-	-

Bilan : comparaison entre MF-API et CA-LBP

algorithme	MF-API	CA-LBP
cadre	GMDP	FA-FMDP
espace de recherche	politiques factorisées déterministes	politiques factorisées stochastiques
évaluation	champ moyen	loopy belief propaga- tion
optimisation	itération de la poli- tique approchée	coordinate ascent
vitesse	rapide	lent
efficacité	GMDP de très grande taille	FA-FMDP de très grande taille
qualité sur GMDP	politiques de valeur équivalente	

Perspectives à court terme

Recherche d'un algorithme

- ayant la garantie de convergence vers un maximum local pour des variables d'action non binaires (si l'évaluation était exacte)
- plus rapide

Piste envisagée : montée de gradient sur une reparamétrisation du problème 1 permettant de ne pas avoir de contraintes :

$$\delta_i(a_i|pa_\delta(a_i)) = \frac{e^{\theta_i(a_i|pa_\delta(a_i))}}{\sum_{a_i} e^{\theta_i(a_i|pa_\delta(a_i))}}$$

→ à cause de la reparamétrisation, il peut y avoir plus de maxima locaux

Perspectives à court terme

Estimation du gradient

- différences finies
- utilisation de méthodes variationnelles comme pour l'évaluation

→ même complexité ($\approx N$ évaluations)

→ dans les deux cas, le calcul peut être parallélisé

Perspectives à long terme

Algorithme

- varier la méthode d'évaluation au cours de l'optimisation ('multifidélité')
- mêler évaluation et optimisation (apprentissage par renforcement)

Plus généralement

- cas partiellement observable
- optimisation de la structure de la politique

Application en agroécologie (PDMF-AF)



Oilseed Rape



Wheat weeds



Semi-natural habitats

Pollination
Food

Honey bees



Wild pollinators



Comment arranger les cultures dans l'espace et dans le temps de manière à faire un compromis entre rendement et conservation de la biodiversité ?

Merci de votre attention !



- O. Buffet and D. Aberdeen. The factored policy-gradient planner. *Artificial Intelligence*, 173 :722–747, 2009.
- Q. Cheng, Q. Liu, F. Chen, and A. Ihler. Variational planning for graph-based MDPs. In *Advances in Neural Information Processing Systems 26*, pages 2976–2984, 2013.
- J. S. Dibangoye, C. Amato, and A. Doniec. Scaling Up Decentralized MDPs Through Heuristic Search. In *Proceedings of the 28th Conference on Uncertainty and Artificial Intelligence*, pages 217–226, 2012.
- A. Kumar, S. Zilberstein, and M. Toussaint. Scalable multiagent planning using probabilistic inference. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence*, 2011.
- A. Raghavan, S. Joshi, A. Fern, P. Tadepalli, and R. Khardon. Planning in factored action spaces with symbolic dynamic programming. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

- R. Sabbadin, N. Peyrard, and N. Forsell. A framework and a mean-field algorithm for the local control of spatial processes. *International Journal of Approximate Reasoning*, 53(1) :66–86, 2012.
- B. Sallans and G. E. Hinton. Reinforcement learning with factored states and actions. *Journal of Machine Learning Research*, 5 :1063–1088, 2004.

général mais peu efficace en grande dimension :

- Raghavan et al. [2012] (PDMF-AF)
- Sallans and Hinton [2004], Buffet and Aberdeen [2009] (AR)

efficace mais hypothèses restrictives :

- Sabbadin et al. [2012], Cheng et al. [2013] (PDMG)
- Dibangoye et al. [2012] (Dec-MDP)
- Kumar et al. [2011] (Dec-POMDP)

→ notre objectif : bon compromis généralité-efficacité-qualité !

Comparaison des cadres GMDP et FA-FMDP

	GMDP	FA-FMDP
$S^t =$	$(S_1^t, S_2^t, \dots, S_n^t)$	$(S_1^t, S_2^t, \dots, S_n^t)$
$A^t =$	$(A_1^t, A_2^t, \dots, A_n^t)$	$(A_1^t, A_2^t, \dots, A_m^t), m \neq n$
$P(S^{t+1} S^t, A^t) =$	$\prod_{i=1}^n P_i(S_i^{t+1} S_{N(i)}^t, A_i^t)$	$\prod_{i=1}^n P_i(S_i^{t+1} S_{N_1(i)}^t, A_{N_2(i)}^t)$
$R(S^t, A^t) =$	$\sum_{i=1}^n R_i(S_{N(i)}^t, A_i^t)$	$\sum_{k=1}^K R_k(S_{N_3(k)}^t, A_{N_4(k)}^t)$
politique δ	$\delta_i : S_{N(i)}^t \rightarrow A_i^t, i = 1 \dots n$ (deterministe)	$\prod_{j=1}^m \delta_j(S_{N_5(j)}^t, A_{N_6(j)}^t)$ (stochastique)