La prise de décisions séquentielles multi-objectif

Thèse de doctorat

École Doctorale d'Informatique, ED 427

Université Paris-Sud 11

Par **Weijia WANG**

Présentée et soutenue publiquement

Le 11 juillet 2014

À Orsay, France

Devant le jury ci-dessous :

Yann Chevaleyre,	Professeur, Université Paris 13, LIPN	(Examinateur)
Cécile Germain-Renaud,	Professeur, Université Paris-Sud, LRI/TAO	(Examinateur)
Dominique Gouyou-Beauchamps,	Professeur, Université Paris-Sud, LRI	(Examinateur)
Jin-Kao Hao,	Professeur, Université d'Angers, LERIA	(Rapporteur)
Philippe Preux,	Professeur, Université de Lille 3, INRIA Lille	(Rapporteur)
Michèle Sebag,	DR CNRS, Université Paris-Sud, LRI/TAO	(Directrice de Thèse)

Rapporteurs :

Jin-Kao Hao,	Professeur, Université de Angers, LERIA, France
Philippe Preux,	Professeur, Université de Lille 3, INRIA Lille, France

Résumé en Français (extended abstract in French)

Cette thèse porte sur le problème de la prise de décision séquentielle multi-objectif. Les algorithmes de la prise de décision séquentielle, plus spéciquement fondés sur la recherche Monte-Carlo arborescente (MCTS) [Kocsis and Szepesvári, 2006], ont été étendus au cas multi-objectif en s'inspirant des indicateurs à base de population proposés dans la littérature de l'optimisation multi-objectif, l'indicateur d'hyper-volume et la relation de dominance.

0.1 Contexte / Motivation

0.1.1 Décision séquentielle

La prise de décision compose une partie importante de nos activités quotidiennes. La prise de décision repose généralement sur une mesure d'ordre total (comme la fonction de récompense), indiquant la qualité des décisions à optimiser. Le problème de la prise de décision séquentielle (SDM) est plus complexe en ce sens que les séquences de décision optimales, aussi appelées politiques de décision ou stratégies, ne sont généralement pas formées en sélectionnant la meilleure décision individuelle à chaque étape : les décisions qui composent la séquence optimale ne sont pas indépendantes. Les applications typiques de SDM incluent les jeux [Aliprantis and Chakrabarti, 2000], la programmation [Zhang and Dietterich, 1995], et la robotique [Mahadevan and Connell, 1992].

Une des principales difficultés du problème SDM est la taille de l'espace de recherche, exponentielle en fonction de la longueur des séquences considérées. Le but de l'apprentissage par renforcement (RL), produire des séquences de décisions optimales à l'échelle globale, procède dans le cas général en identifiant la fonction de valeur attachée à un état ou une paire (état,action), i.e. la somme des récompenses que l'on peut espérer recevoir après avoir visité cet état, ou après avoir effectué cette action dans cet état.

0.1.2 Optimisation multi-objectif

Indépendamment, de nombreux problèmes de décision dans le monde réel impliquent de multiples objectifs ; par exemple, un processus de fabrication cherchera souvent à minimiser simultanément le coût et le risque de la production. Ces problèmes sont appelés optimisation multi-objectif (MOO). Pour un problème de MOO non trival, il n'existe pas de solution unique qui optimise simultanément chaque objectif. Les fonctions objectifs sont antagonistes. Deux solutions ne sont pas nécessairement comparables ; par exemple, un plan de production pourrait être d'un coût élevé et à faible risque, et un autre de faible coût et à haut risque. Les solutions qui ne peuvent pas être améliorées relativement à un objectif sans dégrader les autres objectifs sont appelées les solutions Pareto optimales ; leur ensemble forme le front de Pareto. L'optimisation multi-objectif est largement appliquée dans de nombreux domaines de la science, y compris en économie, en finances et

en ingénierie.

0.1.3 La prise de décision séquentielle multi-objectif

Cette thèse est au carrefour de l'apprentissage par renforcement (RL) et l'optimisation multi-objectif (MOO). L'apprentissage par renforcement (RL) [Sutton and Barto, 1998; Szepesvári, 2010] est un domaine mature où de nombreux algorithmes avec des garanties d'optimalité ont été proposés au prix d'un passage à l'échelle quelque peu limité. Il traite des problèmes SDM dans le cadre de processus de décision de Markov (MDP). La recherche Monte-Carlo arborescente (MCTS), ancrée sur le cadre de bandit manchot, ou bandit à bras multiples (MAB) [Robbins, 1985], résout le problème du passage à l'échelle des algorithmes RL standard, avec d'excellents résultats dans nombreux problèmes de SDM de taille moyenne, comme des jeux [Ciancarini and Favini, 2009] et la planification [Nakhost and Müller, 2009]. Il procède par la construction itérative de l'arbre formalisant la séquence des décisions. Son efficacité algorithmique est notamment reconnue par son application au jeu de Go ; le programme MoGo a été salué comme une avancée fondamentale dans le domaine du jeu de Go par ordinateur [Gelly and Silver, 2007].

Motivée par le fait que de nombreuses applications dans le monde réel sont naturellement formulées dans le cadre de l'optimisation multi-objectif (MOO), cette thèse étudie le problème de la prise de décision séquentielle multi-objectif (MOSDM) où la récompense associée à un état donné dans le MDP est d-dimensionelle au lieu de scalaire. L'apprentissage par renforcement multi-objectif (MORL) a été appliqué aux tâches MOSDM telles que le contrôle du niveau d'eau du lac [Castelletti et al., 2002], l'équilibre entre la consommation d'énergie dans les serveurs web [Tesauro et al., 2007], planification de grille [Yu et al., 2008] et job-shop planification [Adibi et al., 2010].

0.2 Contributions Principales

Le présent travail concerne la prise de décision multi-objectif dans le cadre de MCTS. Il releve le défi de définir un règle de sélection de nœud lorsque les récompenses cumulées sont d-dimensionnelles, en s'appuyant sur des indicateurs bien étudiés de la littérature MOO. Les principales contributions sont les suivantes.

0.2.1 Algorithme MOMCTS

L'algorithme de la Recherche Monte-Carlo Arborescente Multi-Objectif (MOMCTS) a été proposé dans ce travail, dans lequel l'exploration de l'arbre MCTS a été modifié pour tenir compte de l'ordre partiel entre les nœuds dans l'espace d'objectif multidimensionnel, et le fait que le résultat souhaité est un ensemble de solutions Pareto-optimales (par opposition à une solution optimale unique).

Dans chaque nœud l'arbre de recherche de MOMCTS, une récompense vectorielle $\hat{\mathbf{r}}_{s,a} = (r_{s,a;1}, r_{s,a;2}, \cdots, r_{s,a;d})$ représentant la récompense moyenne dans chaque objectif est maintenue, ainsi que le nombre $n_{s,a}$ de visites sur le nœud. Chaque arbre dans MOMCTS est construit en suivant les trois mêmes phases que MCTS – la phase de sélection, la phase de construction de l'arbre et la phase aléatoire. Afin de s'adapter à la configuration MOO, les modifications apportées dans les trois phases sont présentées dans les sections suivantes.

La phase de sélection

La sélection de nœud MOMCTS dépend d'un score scalaire, qui définit un ordre total entre les nœuds avec des récompenses multi-dimensionnelles. Dans ce travail, nous proposons deux scores dans la phase de sélection de MOMCTS – l'indicateur de hypervolume et la récompense de dominance Pareto. Les deux scores appartiennent à la catégorie des fonctions de scalarisation fondées sur la population (section 4.2.3). Ils s'appuient sur l'archive P, qui maintient les récompenses vectorielles recueillies pendant le processus de recherche de MOMCTS.

La phase de construction de l'arbre

Dans la phase de construction de l'arbre, les heuristics d'élargissement progressif (Progressif Widening, PW) et d'estimation rapide de la valeur d'action (RAVE) qui sont optionellement utilisés dans MCTS (section 2.6.2) sont régulièrement intégrées dans MOMCTS. PW limite le nombre d'actions admissibles d'un nœud à une valeur entière $\lfloor n_{s,a}^{1/b} \rfloor$, avec b généralement fixé à 2 ou 4. La sélection de l'action dans la phase de construction de l'arbre repose sur l'heuristique RAVE.

La phase aléatoire

La phase aléatoire est réalisée de la même manière que dans MCTS, sauf que à la fin, une récompense vectorielle $\mathbf{R}$ est retournée. L'autre modification est que la fonction de

scalarisation basée sur la population qui maintient l'archive P des récompenses vectorielles reçues durant la recherche de MOMCTS[1] Sans perte de généralité, les points dominés sont supprimés de l'archive P.

MOMCTS

Par rapport à MCTS, le modification principlale apportée dans MOMCTS concerne l'étape de sélection de nœud. Le défi est d'étendre le critère mono-objectif de sélection de nœud au contexte multi-objectif. Comme indiqué, le noyau de la MOO est de récupérer l'ordre total entre les points de l'espace d'objectif multi-dimensionnel. La façon la plus simple de traiter avec l'optimisation multi-objectif est de revenir à l'optimisation mono-objectif, grâce à l'utilisation de la fonction de scalarisation. MOMCTS est caractérisé par la scalarisation des récompenses vectorielles basée sur la population des solutions précédentes, l'archive P. Contrairement à MCTS, qui estime la valeur de nœuds selon la distribution de récompenses fixe sur un seul objectif, MOMCTS estime la valeur de nœuds avec des récompenses à plusieurs dimensions en fonction de leur contribution à l'archive P. Notons que cette archive évolue au cours du processus, définissant un objectif non-stationnaire au long du processus de recherche.

Grâce à l'utilisation de la fonction de scalarisation basée sur la population, MOMCTS traite un problème d'optimisation mono-objectif dans chaque parcours d'arbre, dans lequel la qualité de l'ensemble des solution sauvegardées dans l'archive P est améliorée par la recherche répétitive de solutions simples. Plusieurs parcours d'arbres fournissent un ensemble de solutions optimales au sens de dominance Pareto dans MOMCTS.

L'algorithme MOMCTS est résumé par l'algorithme 0.1. Les hyper-paramètres communs à tous les algorithmes MOMCTS comprennent le budget de calcul N, le paramètre B utilisé dans l'heuristique d'élargissement progressive PW, et le modèle génératif $\mathcal{M}_D$ du problème MOSDM considéré. La valeur du nœud (s, a) noté par $g_x(s, a)$ est une fonction de scalarization basée sur la population, où x identifie le choix de la méthode de scalarisation.

Dans MOMCTS, l'estimation rapide de la valeur d'action (RAVE) prend une forme vectorielle ($\mathbf{RAVE}(a) \in \mathbb{R}^d, a \in \mathcal{A}$). Une fonction de scalarisation est donc nécessaire pour définir un ordre total entre les actions en se fondant sur l'estimation RAVE. Dans MOMCTS, la valeur scalarisée des vecteurs RAVE $g_{x;rave}(a), a \in \mathcal{A}$ se fonde sur la même fonction de scalarisation $g_x(s, a)$. La description des fonctions $g_x(s, a)$ et $g_{x;rave}(a)$ est donnée dans les sections 5.2 et 5.3.

Une propriété importante de MCTS est la propriété de consistance définie comme la capacité de l'algorithme de converger vers la politique optimale lorsque le nombre de parcours d'arbres N tend vers l'infini [Berthier et al., 2010]. La propiété de consistance est vérifiée dans le cas stationnaire, i.e. lorsque la distribution de la fonction récompense est

[1]Lorsque le nombre d'objectifs est faible ($d \leq 3$), les ressources de calcul et de mémoire nécessaires pour maintenir l'archive P sont limitées. Certaines heuristiques supplémentaires doivent être conçues pour préserver le passage à l'échelle de l'approche basée sur la population scalarisation dans le cadre de problèmes MOO faisant intervenir de nombreux objectifs (many objective optimization, MaOO). L'extension de MOMCTS au cas MaOO est une perspective de recherche future.

Algorithm 0.1: Algorithme MOMCTS

MOMCTS
Entrée: Nombre N de simulations
Sortie: Arbre de recherche $\mathcal{T}$
Initializer $\mathcal{T} \leftarrow$ racine (état initial), $P \leftarrow \{\}$
for $t = 1$ **to** N **do**
  Simulation($\mathcal{T}, P,$ root node)
end for
retourner $\mathcal{T}$

Simulation
Entrée: Arbre de recherche $\mathcal{T}$, archive P, nœud s
Sortie: récompense vectorielle $\mathbf{r}_u$
if s n'est pas une feuille, et $\neg(\lfloor (n_s+1)^{1/b} \rfloor > \lfloor (n_s)^{1/b} \rfloor)$ // (test PW non déclenché) **then**
  Selectionner $a^* = \arg\max\{g_x(s,a), (s,a) \in \mathcal{T}\}$
  $\mathbf{r}_u \leftarrow$ Simulation($\mathcal{T}, P, (s, a^*)$)
else
  $\mathcal{A}_s = \{$ actions disponibles non-visitées sous état $s\}$
  Selectionner $a^* = \arg\max\{g_{x;rave}(a), a \in \mathcal{A}_s\}$
  Ajouter (s, a^*) comme fils de s
  $\mathbf{r}_u \leftarrow$ SimulationAleatoire($P, (s, a^*)$)
end if
Mettre à jour n_s, n_{s,a^*}, $\hat{\mathbf{r}}_{s,a}$ et $\mathbf{RAVE}(a^*)$
retourner $\mathbf{r}_u$

SimulationAleatoire
Entrée: archive P, état u
Sortie: récompense vectorielle $\mathbf{r}_u$
$\mathcal{A}_{rnd} \leftarrow \{\}$ //sauvegarder l'ensemble des actions visitées durant la phase aléatoire
while u n'est pas l'état final **do**
  Selectionner uniformement une action disponible a pour u
  $\mathcal{A}_{rnd} \leftarrow \mathcal{A}_{rnd} \cup \{a\}$
  $u \leftarrow (u, a)$
end while
$\mathbf{r}_u \leftarrow \mathcal{M}_d(u)$ //obtenir la récompense vectorielle de la simulation
if $\mathbf{r}_u$ n'est pas dominé pas les points dans P **then**
  Eliminer tous les points dominés par r_u dans P
  $P \leftarrow P \cup \{\mathbf{r}_u\}$
end if
Mettre à jour $\mathbf{RAVE}(a)$ pour $a \in \mathcal{A}_{rnd}$
retourner $\mathbf{r}_u$

fixe au cours du temps. Dans le cas de MOMCTS, cependant, la fonction de scalarisation basée sur la population dépend de l'archive de P, et donc elle est non-stationnaire. L'étude de la consistance de l'approche proposée est une perspective de recherche future.

0.2.2 Indicateurs de qualité de solution multi-objectif

Les approches existantes en MORL [Gábor et al., 1998; Castelletti et al., 2002; Mannor and Shimkin, 2004; Natarajan and Tadepalli, 2005; Tesauro et al., 2007] sont pour la plupart basées sur la scalarisation linéaire de récompenses multidimensionnelles, avec la limitation qu'elle ne permet pas de découvrir des solutions sur les parties non-convexes du front de Pareto. Ces approches n'utilisent pas les indicateurs de qualité qui ont été définis et utilisés dans le domaine des Algorithmes Evolutionaires Multi-Objectif (MOEA) [Zitzler et al., 2003]. Ce travail établit un pont entre les deux domaines de MORL et MOEA, en introduisant deux de ces indicateurs d'évaluation des performances des politiques dans l'algorithme MOMCTS.

Spécifiquement, l'indicateur de hypervolume [Zitzler and Thiele, 1998] a été utilisé pour définir la performance scalaire d'un nœud. Comme montré par [Fleischer, 2003], l'indicateur de hypervolume est maximisée si et seulement si les points dans P^* appartiennent au front Pareto du problème MOO considéré. Auger et al. [2009] montrent que, pour $d = 2$, pour un certain nombre K de points, l'indicateur hypervolume projecte un problème d'optimisation multi-objectif défini dans $\mathbb{R}^d$, sur un problème d'optimisation mono-objectif dans $\mathbb{R}^{d\times K}$, dans le sens où il existe au moins un ensemble de K points dans $\mathbb{R}^d$ qui maximise l'indicateur hypervolume. Le mérite de cette approche est d'aller au-delà de la scalarisation linéaire standard. L'indicateur d'hyper-volume souffre toutefois de deux limitations. D'une part, les coûts de calcul d'indicateur de hypervolume augmentent de façon exponentielle avec le nombre d'objectifs. Deuxièmement, l'indicateur de hypervolume n'est pas invariant par la transformation monotone des objectifs. La propriété d'invariance (satisfaite par exemple par les algorithmes d'optimisation à base de comparaison) donne des garanties de robustesse extrêmement importantes pour les problèmes d'optimisation mal conditionnés [Hansen, 2006].

Par conséquent, un autre indicateur a été considéré : la récompense de dominance Pareto. Cette récompense peut être considé ré comme un compteur du nombre de découvertes de solutions non dominées, qui est cumulé de manière actualisée. Par rapport à la première approche − appelée MOMCTS-hv dans le reste de cette thèse, la deuxième approche − appelée MOMCTS-dom − a une complexité linéaire de calcul par rapport au nombre d'objectifs, et est invariante par rapport à la transformation monotone des objectifs. Le prix à payer pour l'amélioration de l'évolutivité de MOMCTS-dom est que la récompense de la dominance peut moins favoriser la diversité de l'archive Pareto, qui est une mesure essentielle de la qualité de l'ensemble de solutions non-dominés : un point non-dominé a la même récompense de dominance Pareto alors que l'indicateur de hypervolume favorise les points non-dominés situés dans les régions peu peuplées de l'archive Pareto.

0.2.3 Validation expérimentale

Les deux algorithms MOMCTS-hv et MOMCTS-dom ont été validés expéprimentalement sur quatre problèmes : Deep Sea Treasure (DST) [Vamplew et al., 2010], Resources Gathering (RG) [Barrett and Narayanan, 2008], Grid Scheduling [Yu et al., 2008] et Physical Travelling Salesman Problem (PTSP) [Powley et al., 2012]. Les deux premiérs problèmes artificiels sont conçus pour comparer les approches MOMCTS à l'état de l'art en MORL (les méthodes basées sur la scalarisation liné aire). Les deux dernièrs problèmes plus applicatifs sont utilisé pour tester le passage à l'échelle de MOMCTS. Les proprietés des problèmes considérés sont résumées par la Table 1.

Table 1: Problèmes de la prise de décision séquentielle multi-objectif

Problème	Forme du front Pareto	Fonction de transition déterministe or non-déterministe	Nombre d'objectifs	Decision en temps réel
Deep Sea Treasure	Non-convexe	Deterministe et Non-déterministe	2	Non
Resource Gathering	Convexe	Non-déterministe	3	Non
Grid Scheduling	Inconnu	Deterministe	2	Non
Physical Travelling Salesman	Inconnu	Deterministe	7	Oui

Les résultats expérimentaux sur le problème de Deep Sea Treasure confirment un mérite principal des approches proposées, leur capcité de découvrir des politiques se trouvant dans les régions non-convexes de front dePareto. À notre connaissance, cette fonctionnalité est unique dans la littérature MORL. Les expériences sur le probléme de Resource Gathering montrent que MOMCTS-dom bénéficie d'un meilleur passage à l'échelle que MOMCTS-hv en raison du coût de calcul de test de dominance Pareto qui est linéaire par rapport au nombre d'objectifs. Cette robustesse de MOMCTS-dom est en outre confirmée par les Physical Travelling Salesman Problem expériences, dont 7 objectifs sont optimisés de manière on-line.

En contrepartie, les approches MOMCTS souffrent de deux faiblesses principales. Tout d'abord, comme indiqué sur le Grid Scheduling et Physical Travelling Salesman Problem, une certaine connaissance préalable est nécessaire pour appliquer une exploration de MOMCTS avec efficacité. Deuxièmement, comme témoigné par le problème de Resource Gathering, les approches présentées découvrent peu de politiques "à risque" qui se trouvent dans une région peu prometteuse (la découverte d'optima de type "chapeau mexicain").

En conclusion, ce travail peut être considéré comme une preuve de concept de l'application du cadre MOMCTS pour les problème MOO. Les résultats obtenus peuvent être considérés comme prometteurs : les performances sont décentes comparativement à l'état de l'art, en dépit du fait qu'il s'agit d'approches beaucoup moins matures que les approches RL standard.

0.3 Les perspectives futures

Ce travail ouvre plusieurs perspectives de recherche, de type à la fois théorique et applicatif.

La perspective théorique principale concerne l'analyse des propriétés du mécanisme de mise à jour de la récompense cumulable dans le contexte général de l'optimisation (mono-objectif) dynamique. En outre, l'analyse de la consistance des critères de sélection de nœud actuels (y compris l'indicateur d'hypervolume et la récompense de dominance Pareto) permettra de définir des lignes directrices pour la conception de nouvelles récompenses scalarisées dans le cadre MOMCTS.

Du côté applicatif, d'une part, la fonction de préférence de scalarization linéaire utilisé dans les expérience de Physical Travelling Salesman Problem peut être étendue à un contexte plus général (par exemple non-linéaire), ce qui peut permettre à l'utilisateur d'exprimer ses préférences d'une façon plus naturelle et interactive.

Un perspective algorithmique concerne l'ajustement du mécanisme de mise à jour cumulatif actualisé de la récompense de dominance Pareto (Equation (5.7)). Vu que la découverte de solutions non-dominées est de plus en plus rare au cours du temps, l'ajustement du paramètre d'actualisation δ devrait être dynamique pour compenser cet effet de rareté. Une approche serait de considérer la découverte de nouvelles solutions non-dominées dans le cadre de la théorie de la valeur extrême [De Haan and Ferreira, 2007], et d'ajuster δ en conséquence.