Teză de abilitare:

Planificarea optimistă pentru controlul optimal şi pentru sistemele în reţea neliniare

— Rezumat —

Lucian Buşoniu

Activitatea mea de cercetare se concentrează pe obiectivul major de a dezvolta metode de control pentru sisteme complexe, posibil necunoscute. Acest obiectiv este motivat de complexitatea înaltă a sistemelor de control moderne, manifestată prin proprietăţi cum ar fi neliniaritatea, stohasticitatea, scara largă, şi natura lor distribuită. Mai mult decât atât, aceste sisteme nu pot fi oricând modelate precis, de exemplu fiindcă sunt insuficient înţelese. Pentru acest caz în care modelul este necunoscut, am lucrat în timpul doctoratului la clasa promiţătoare a metodelor de învăţare prin recompensă (*reinforcement learning*, RL) (Sutton and Barto, 1998; Szepesvári, 2010; Lewis and Liu, 2012), care învaţă cum să controleze un sistem stohastic neliniar fără a folosi un model. Chiar dacă modelul este cunoscut, problema de control stohastic neliniar rămâne foarte dificilă: metodele de planificare şi programare dinamică, analoage învăţării prin recompensă dar bazate pe model, pot fi aplicate în acest caz (La Valle, 2006; Bertsekas, 2012; Powell, 2012). Eforturile mele recente, ulterioare doctoratului, s-au concentrat pe acest al doilea caz, şi mai exact pe algoritmii de planificare optimistă şi pe aplicaţiile lor.

Planificarea optimistă (*optimistic planning*, OP) (Munos, 2014) rezolvă probleme de control optimal modelate prin procese de decizie Markov (*Markov decision process*, MDP) (Puterman, 1994). În aceste procese, un controller măsoară la fiecare pas de timp discret starea unui sistem, şi aplică o acţiune conform unei legi de control. În urma acestei acţiuni, sistemul trece într-o nouă stare, şi un semnal scalar de recompensă este trimis controllerului evaluând calitatea acestei tranziţii. Controllerul măsoară noua stare, şi întregul ciclu se repetă. Scopul este găsirea unei legi de control optimale, care maximizează recompensa cumulată de-a lungul interacţiunii (returnul). Această paradigmă poate fi aplicată problemelor de control de nivel jos, cum ar fi reglarea stării la o valoare dată, unde recompensele sunt definite de obicei pe baza distanţei de la această valoare. Paradigma funcţionează însă şi pentru probleme de nivel mai înalt, cum ar fi un robot care rezolvă o sarcină într-un mediu necunoscut. În acest caz, recompensele reprezintă succesul sau eşecul în rezolvarea sarcinii dorite.

Tehnicile OP funcţionează într-un mod local, prin căutarea acţiunilor la cerere pentru fiecare stare a sistemului întâlnită. Natura locală a acestor tehnici le reduce dependenţa de dimensiunea vectorului de stare, comparativ cu programarea dinamică sau RL, şi le permite să trateze în mod natural variabilele de stare cu valori continue, care sunt esenţiale în controlul sistemelor fizice. La fiecare pas, OP rulează o căutare exploratorie în spaţiul secvenţelor de acţiuni posibile din starea curentă, reprezentând acest spaţiu sub forma unui arbore. Apoi, cea mai bună primă acţiune găsită este aplicată, şi întregul proces se repetă la pasul următor. Planificarea este aşadar un tip foarte general de control predictiv bazat pe model. Cum resursele

de calcul sunt limitate în acest context online, căutarea trebuie să fie eficientă, şi o modalitate bună pentru a o eficientiza este aplicarea principiului de *optimism în faţa incertitudinii*: date fiind mai multe secvenţe de acţiuni cu valori incerte, secvenţa mai promiţătoare este explorată mai întâi. Matematic, secvenţa cea mai promiţătoare este cea cu marginea superioară pe return maximală. Metodele optimiste combină într-un fel inovativ idei din optimizare, planificare, căutarea în grafuri (La Valle, 2006), şi RL (Sutton and Barto, 1998), aşa-numita teoria banditilor jucând un rol important (Auer et al., 2002) (*bandit theory*, de la numele unui joc de noroc).

Mai mulţi algoritmi de OP erau deja dezvoltaţi când am abordat acest domeniu, mai ales în cazul acţiunilor discrete (Kocsis and Szepesvári, 2006; Hren and Munos, 2008; Bubeck and Munos, 2010; Walsh et al., 2010), dintre care aşa-numitul *Upper Confidence Trees* este probabil cel mai cunoscut (Kocsis and Szepesvári, 2006). Anumiţi algoritmi funcţionau de asemenea pentru acţiuni continue (Mansley et al., 2011; Weinstein and Littman, 2012). Câteva dintre aceste metode au caracteristici foarte utile: sunt aplicabile la sisteme cu dinamici generale neliniare şi la funcţii de cost generale, noncuadratice; şi furnizează garanţii de optimalitate care plasează performanţa soluţiei returnate într-o relaţie foarte strânsă cu bugetul de calcul investit de algoritm (Hren and Munos, 2008; Bubeck and Munos, 2010). Aceste caracteristici apar însă doar în condiţii restrictive: acţiuni discrete; şi fie pentru sisteme deterministe fără incertitudini sau perturbaţii, sau pentru sisteme stohastice dar căutând doar în clasa suboptimală a secvenţelor de acţiuni în buclă deschisă. Prima parte majoră a acestei teze va prezenta aşadar eforturile mele pentru a elimina aceste limitări.

Mai exact, doi algoritmi sunt discutaţi în cazul determinist: unul care funcţionează pentru acţiuni continue, şi altul care adresează perturbaţii printr-o abordare robustă de tip minimax, tratându-le ca pe acţiunile unui oponent. Apoi, cazul sistemelor stohastice este considerat, care poate de exemplu modela zgomotele şi perturbaţiile aleatoare cu distribuţii de probabilitate cunoscute. Un prim algoritm este furnizat pentru distribuţii cu suport discret şi finit. Acest algoritm este apoi extins la o clasă de densităţi continue de probabilitate, prin intermediul discretizării cu puncte sigma (*sigma-point discretization*). Performanţa tuturor algoritmilor este caracterizată, cu excepţia celui pentru acţiuni continue pentru care analiza este în curs. Noi măsuri de complexitate a problemei sunt propuse, potrivite fiecărei clase de probleme adresate. În plus faţă de aceste studii analitice, toţi algoritmii sunt evaluaţi în experimente detaliate de simulare.

Pe lângă semnificaţia lor fundamentală în controlul optimal, generalitatea metodelor de OP le face utile de asemenea pentru a adresa alte provocări în controlul neliniar. În special sistemele în reţea devin extrem de importante în societatea de astăzi: reţelele de comunicare, de energie, de transport, reţelele de calcul distribuite, şi reţelele sociale sunt doar câteva exemple de astfel de sisteme care influenţează viaţa de zi cu zi. Aşadar, a doua parte majoră a tezei investighează aplicaţii ale metodelor optimiste la controlul sistemelor în reţea, tratându-le din două puncte de vedere complementare. Prima perspectivă adresează comportamentul coordonat al unor sisteme multiple interconectate numite agenţi, sub constrângerile impuse de topologia de interconexiune. În acest context, un prim algoritm este propus folosind optimizarea optimistă a secvenţelor de acţiuni de lungime fixă, cu scopul de a atinge un consensus asupra variabilelor de stare ale agenţilor, presupunând o topologie fixă de comunicaţie. Apoi, OP în spaţiul secvenţelor de lungime variabilă este aplicată pentru a obţine *flocking* (numit

astfel după comportamentul stolurilor de păsări), unde topologia este dictată de o relaţie de proximitate între agenţi. Pentru acest al doilea algoritm, principalul scop analitic este de a garanta păstrarea topologiei de comunicaţie dată fiind această constrângere.

A doua perspectivă tratează constrângerile de comunicaţie induse de reţeaua interpusă între un singur sistem şi controlerul său. Două strategii de control folosind OP sunt propuse pentru a reduce numărul de transmisii în reţea. În prima strategie, secvenţe de acţiuni sunt transmise la sistem cu o perioadă fixă. În a doua strategie, algoritmul lucrează cu un buget de calcul fix şi decide momentul următoarei transmisii pe baza ultimei stări măsurate, ducând la o lege auto-declanşată (*self-triggered*). Garanţii analitice de performanţă sunt furnizate pentru ambii algoritmi.

Pe lângă aceste analize teoretice, toţi algoritmii pentru sisteme în reţea sunt evaluaţi în exemple numerice detaliate.

Cele două direcţii discutate mai sus, implicând dezvoltări fundamentale în algoritmii OP pe de o parte, şi aplicaţiile lor la controlul neliniar pe de alta, formează principala mea direcţie de cercetare după doctorat. Direcţii adiţionale în RL, aplicaţii, şi subiecte secundare de planificare şi control sunt trecute în revistă în capitole separate. Cercetările rezultate direct din activitatea doctorală nu sunt discutate, chiar dacă au fost efectuate sau publicate după data doctoratului. În mod similar, nu sunt menţionate activităţile la care am participat dar fără a juca un rol important.

În rezumat, contribuţiile importante prezentate în această teză sunt, în ordinea în care sunt discutate:

- Algoritmul *simultaneous optimistic optimization for planning*, pentru sisteme cu acţiuni continue (Buşoniu et al., 2013a).
- Algoritmul *optimistic minimax search*, pentru abordarea robustă a problemelor cu perturbaţii (Buşoniu et al., 2014).
- Algoritmul *optimistic planning for Markov decision processes*, pentru sisteme stohastice cu zgomote discrete (Buşoniu and Munos, 2012).
- Extensia acestuia numită *optimistic planning with sigma-point discretization*, pentru sisteme stohastice cu zgomote continue (Buşoniu and Tamas, 2014).
- O metodă bazată pe optimizarea optimistă pentru consensus în sisteme multiagent decentralizate (Buşoniu and Morarescu, 2014).
- O metodă bazată pe planificarea optimistă pentru flocking în sisteme multiagent (Buşoniu and Morarescu, 2015, 2013).
- Două strategii bazate pe planificarea optimistă pentru controlul în reţea (Buşoniu et al., 2013b).
- Garanţii analitice de performanţă pentru toate metodele cu excepţia celei pentru acţiuni continue.
- Validări numerice detaliate pentru toate metodele.

În viitor, voi porni de la cercetările de planificare şi control prezentate în această teză şi voi integra noi direcţii de control, învăţare automată şi RL, pentru a mă apropia de obiectivul meu general al unui *cadru algoritmic pentru controlul prin învăţare şi planificare al sistemelor complexe*. O componentă importantă va fi combinarea OP cu idei de RL din experienţa mea anterioară, pentru a obţine algoritmi hibrizi cu avantajele ambelor clase de metode. Garanţiile de stabilitate pentru soluţia obţinută vor fi de asemenea esenţiale. Direcţiile aplicative vor fi continuate şi noi aplicaţii vor fi investigate, atât în contextul general al problemelor de control neliniar, cât şi în domeniul specific al roboticii asistive. Toate aceste rezultate vor crea o platformă solidă pornind de la care se vor explora noi direcţii în control şi luarea deciziilor pe de o parte, şi pe de alta în învăţarea automată şi inteligenţa artificială.

Bibliografie

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 4th edition.

Bubeck, S. and Munos, R. (2010). Open loop optimistic planning. In *Proceedings 23rd Annual Conference on Learning Theory (COLT-10)*, pages 477–489, Haifa, Israel.

Buşoniu, L., Daniels, A., Munos, R., and Babuška, R. (2013a). Optimistic planning for continuous–action deterministic systems. In *2013 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-13)*, Singapore.

Buşoniu, L. and Morarescu, C. (2013). Optimistic planning for consensus. In *Proceedings American Control Conference 2013 (ACC-13)*, pages 6735–6740, Washington, DC.

Buşoniu, L. and Morarescu, C. (2014). Consensus for black-box nonlinear agents using optimistic optimization. *Automatica*, 50(4):1201–1208. .

Buşoniu, L. and Morarescu, C. (2015). Topology-preserving flocking of nonlinear agents using optimistic lanning. *Control Theory and Technology*, 13(1):333–344.

Buşoniu, L. and Munos, R. (2012). Optimistic planning for Markov decision processes. In *Proceedings 15th International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22 of *JMLR Workshop and Conference Proceedings*, pages 182–189, La Palma, Canary Islands, Spain.

Buşoniu, L., Páll, E., and Munos, R. (2014). An analysis of optimstic, best-first search for minimax sequential decision making. In *2014 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-14)*, Orlando.

Buşoniu, L., Postoyan, R., and Daafouz, J. (2013b). Near-optimal strategies for nonlinear networked control systems using optimistic planning. In *Proceedings American Control Conference 2013 (ACC-13)*, pages 3020–3025, Washington, DC.

Buşoniu, L. and Tamas, L. (2014). Optimistic planning for the near-optimal control of general nonlinear systems with continuous transition distributions. In *19th IFAC World Congress (IFAC-14)*, Cape Town, South Africa.

Hren, J.-F. and Munos, R. (2008). Optimistic planning of deterministic systems. In *Proceedings 8th European Workshop on Reinforcement Learning (EWRL-08)*, pages 151–164, Villeneuve d'Ascq, France.

Kocsis, L. and Szepesvári, C. (2006). Bandit based Monte-Carlo planning. In *Proceedings 17th European Conference on Machine Learning (ECML-06)*, pages 282–293, Berlin, Germany.

La Valle, S. M. (2006). *Planning Algorithms*. Cambridge University Press.

Lewis, F. and Liu, D., editors (2012). *Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control*. Wiley.

Mansley, C., Weinstein, A., and Littman, M. L. (2011). Sample-based planning for continuous action Markov decision processes. In *Proceedings 21st International Conference on Automated Planning and Scheduling*, pages 335–338, Freiburg, Germany.

Munos, R. (2014). The optimistic principle applied to games, optimization and planning: Towards foundations of Monte-Carlo tree search. *Foundations and Trends in Machine Learning*, 7(1):1–130.

Powell, W. B. (2012). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2 edition.

Puterman, M. L. (1994). *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. Wiley.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Szepesvári, Cs. (2010). *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers.

Walsh, T. J., Goschin, S., and Littman, M. L. (2010). Integrating sample-based planning and model-based reinforcement learning. In *Proceedings 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, Atlanta, US.

Weinstein, A. and Littman, M. L. (2012). Bandit-based planning and learning in continuous-action Markov decision processes. In *Proceedings 22nd International Conference on Automated Planning and Scheduling (ICAPS-12)*, São Paulo, Brazil.

Despre autor

Lucian Buşoniu deţine din 2014 poziţia de conferenţiar universitar la Universitatea Tehnică din Cluj-Napoca, România, unde a fost şi şef de lucrări în perioada 2011-2014. El a primit doctoratul (*cum laude*) în 2009 de la Universitatea Tehnică din Delft (*Technische Universiteit Delft*), Olanda, şi diploma de inginer (ca şef de promoţie) în 2003 la Universitatea Tehnică din Cluj-Napoca. A activat pe poziţii de cercetător la TUDelft, INRIA Lille, şi Centrul de Cercetare în Automatică din Nancy, Franţa. Interesele lui de cercetare includ planificarea optimistă, controlul neliniar optimal, controlul prin învăţare, şi sistemele în reţea.