Universitatea Babeș-Bolyai
Facultatea de Matematică și
Informatică

Intelligent Models for Robotic Behavior, Decision Making and Environment Interaction

Abstractul tezei de doctorat

Coordonator științific:
Prof. Dr. Horia F. Pop
Dr. Istenes Zoltán

doctorand:
Hunor Sándor Jakab

Cluj Napoca
2012

Cuprinsul rezumatului

Cuprinsul tezei de doctorat

Keywords: robotics, reinforcement learning, machine learning, non-parametric methods.

1 Introducere

Cercetarea în domeniul roboticii a avut parte de o creştere semnificativă în ultimii ani, datorită gradului mare de automatizare prin intermediul roboţilor în aplicaţii industriale şi comerciale. Unul dintre cele mai importante obiective ale roboticii este crearea unor agenţi inteligenţi capabili de a învăţa noi abilităţi şi mecanisme pentru luarea deciziilor în mod autonom. Tema principală a tezei este dezvoltarea unor metode de învăţare pentru realizarea controlului de mişcări în robotică, care este o parte esenţială a oricărui sistem autonom robotic capabil de a opera într-un mediu nestructurat şi stohastic. O caracteristică importantă a scenariului de învăţare, tratată în această lucrare, este natura nesupervizată a instruirii, nevoia de a achizitiona cunoştinţe în mod autonom prin intermediul datelor obţinute în cursul interacţiunii cu mediul de execuţie. În consecinţă metodele propuse în această teză sunt bazate pe metode de instruire automată, din categoria metodelor de instruire cu întărire. Aplicarea metodelor de instruire cu întărire pentru controlul sistemelor robotice realistice, este dificilă chiar şi în cazul sistemelor simple cu un număr mic de stări şi acţiuni. Dificultăţile principale pot fi atribuite creşterii exponenţiale ale spaţiului de căutare cu creşterea complexităţii sistemului şi gradului mare de incertitudine provenită din imperfecţiile fizice ale componentelor robotice şi stohasticitatea mediului. Numărul experimentelor care pot fi executate pe un robot real este deasemenea limitată, datorită atât restricţiilor fizice cât şi celor temporale. Prin urmare, metodele exhaustive de căutare nu pot fi aplicate în cadrul problemelor noastre.

Principalul obiectiv al tezei este de a aborda unele dintre problemele fundamentale care apar în crearea unor modele inteligente de învăţare în cadrul controlului robotic, prin introducerea unor metode noi de genul instruirii cu întărire, concepute special pentru domeniul menţionat. Principalele probleme care sunt investigate în teză sunt urmatoarele: tratarea spaţiilor de stări şi acţiuni continue, tratarea incertitudinii, operare on-line, eficienţă de date şi deducţia proprietăţilor structurale a sistemului din datele obţinute prin interacţiune cu mediul. Pentru validarea metodelor propuse s-au folosit o serie de probleme de control robotic, simulate în mod realistic, cu spaţii de stări şi acţiuni continue, model de tranzacţie zgomotoasă şi multiple grade de libertate.

2 Control robotic prin instruire cu întărire

Controlul mişcărilor în cazul roboţilor este realizat prin transmiterea unor semnale de actuaţie la motoare şi actuatoare pe baza unei strategii de selecţie. Strategiile de control definesc un mecanism de selecţie de acţiuni cu ajutorul căruia robotul generează secvenţe de mişcări potrivite pentru a executa o sarcină, a atinge un scop predefinit. Învăţarea unei strategii de control optimale este realizată prin interacţiune cu mediul, ajustând mechanismul de luare de decizii, pe baza informaţiilor senzoriale obţinute după executarea anumitor acţiuni. Învăţarea autonomă limitează gama metodelor pe care le putem folosi. Pe de o parte nu există un set predefinit de date de antrenare în forma stare-acţiune optimală, pe de altă parte strategia de luare de decizii trebuie îmbunătăţită în paralel cu interacţiunea cu mediul. Datorită acestor restricţii trebuie să lucrăm cu metode de învăţare nesupervizate, cu abilitatea de a opera on-line. Robotul cunoaşte numai o funcţie de utilitate pe baza căreia evaluează observaţiile în urma executării unor acţiuni, şi trebuie să-şi îmbunătăţească strategia maximalizând această funcţie. Însă în majoritatea problemelor de control, maximalizarea imediată a funcţiei de utilite este insuficientă pentru optimalitate, de aceea trebuie luate în considerare şi consecinţele de lungă durată ale acţiunilor. Cu alte cuvinte, agentul robotic trebuie să înveţe care dintre acţiunile lui sunt cele mai optimale, pe baza semnalelor de utilităţi, care pot fi acumulate oricât de departe în viitor. Problemele secvenţiale de luare a deciziilor cu proprietatea Markov sunt modelate matematic prin Procese de Decizie Markov($\mathcal{MDP}$). Fundaţiile matematice pentru primele soluţii ale problemelor de luare de decizii sunt bazate pe teoria ($\mathcal{MDP}$) şi au avut un rol central în dezvoltarea învăţării cu întărire. Ca şi o abstracţie a mecanismului de luare de decizii a robotului, definim *strategia parametrică* $\pi : S \times A \to [0, 1]$ în forma unei distribuţii probabilistice peste spaţiul stare-acţiune, o proiecţie din fiecare stare $s \in S$ şi acţiune $a \in A$ prin probabilitatea condiţională $\pi(a|s)$ de a executa a când sistemul se află în starea s. Soluţia problemei de luare a deciziilor este strategia optimală π^* care maximalizează valoarea aşteptată a recompensei acumulate:

$$\pi^* = \pi_{\theta^*} = \underset{\pi}{\operatorname{argmax}}\,(J^\pi), \quad \text{cu} \qquad J^\pi = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1}\right],$$

unde J^π este funcţia obiectivă a problemei de optimizare, o estimare a recompensei discontate calculată pentru strategia π_θ, cu $r_1, r_2, \ldots$ fiind recompense imediate. În

timpul execuţiei acţiunilor, datele obţinute prin măsurătorile senzoriale sunt folosite pentru a construi modele ale optimalităţii, care pot fi folosite pentru dezvoltarea unui mecanism de luare de decizii. Aceste modele sunt numite funcţii de valori, şi pot fi asociate fie stărilor, fie perechilor stare-acţiune:

$$Q^{\pi}(s,a) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} | s_{0} = s, a_{0} = a\right].$$

Majoritatea metodelor de instruire cu întărire clasice sunt bazate pe estimarea funcţiilor de valori care îndeplinesc rolul de model intermediar între acumularea experienţei şi generarea strategiilor de control.

Extinderea acestor algoritmi pe spaţii continue se face în mod obişnuit prin utilizarea aproximatoarelor de funcţii pentru a reprezenta funcţiile de valori. Însă această procedură duce la apariţia unor probleme de convergenţă şi poate introduce bias în valorile estimate. În această lucrare folosim extensiv metode neparametrice de aproximare de funcţii, mai ales procese Gaussiene de regresie, pentru a reprezenta cunoştinţe şi incertitudine prin procesul de instruire. Metodele noastre noi de control robotic sunt bazate pe construirea unor modele interne de optimalitate prin intermediul proceselor Gaussiene şi folosirea acestora în mod eficient pentru a ghida procesul de interacţie cu mediul şi învăţarea unor strategii de control optimale. În ceea ce urmează, descriem contribuţiile originale ale tezei care pot fi structurate în patru părţi, tratate în capitolele 2, 3, 4, 5 şi 6 din teză.

3 Aproximarea neparametrică a funcţiilor de valori pentru controlul robotic

Construirea modelelor de optimalitate în forma funcţiilor de valori este un element important în problemele de instruire cu întărire. Pentru a trata problema spaţiilor continue de stare-acţiune şi gradul mare de incertitudine în controlul robotic, am analizat utilizarea aproximatoarelor de funcţii neparametrice, în mod special beneficiile folosirii proceselor Gaussiene pentru aproximarea funcţiilor de valori pe spaţii de stare sau stare-acţiune [Jakab and Csató, 2010]. Modelarea funcţiei de valori se face prin punerea unei distribuţii Gaussiene a priori, direct în spaţiul funcţiilor. În timpul interacţiunii cu mediul obţinem o secvenţă de n perechi stare-acţiune şi recompensele imediate în forma unor traiectorii: $\tau = [(s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)]$.

După executarea a m traiectorii obţinem un set de antrenament $\mathcal{D} = [(x_1, r_1) \dots (x_n, r_n)]$, unde $n = mH$. Folosind perechile de stare-acţiune ca şi puncte de antrenament $x_t \stackrel{\text{def}}{=} (s_t, a_t) \in \mathcal{D}$ şi recompensele cumulative $\mathrm{Ret}(x_t) = \sum_{i=t}^{H} \gamma^{i-t} R(s_t, a_t)$ corespunzătoare ca şi etichete obţinem un model probabilistic în forma posteriorului Gaussian pentru funcţia de valori:

$$\begin{aligned} Q_{\mathcal{GP}}|\mathcal{D}, x_{n+1} &\sim \mathcal{N}\left(\mu_{n+1}, \sigma^2_{n+1}\right) \\ \mu_{n+1} = \mathbf{k}_{n+1}\boldsymbol{\alpha}_n \qquad & \sigma^2_{n+1} = k_q\left(x_{n+1}, x_{n+1}\right) - \mathbf{k}_{n+1}\mathbf{C}_n\mathbf{k}^T_{n+1}, \end{aligned}$$

unde $\boldsymbol{\alpha}_n$ şi $\mathbf{C}_n$ sunt parametrii procesului Gaussian $\mathcal{GP}$ – pentru detalii a se consulta [Rasmussen and Williams, 2006] – având următoarea formă:

$$\boldsymbol{\alpha}_n = [\mathbf{K}^n_q + \boldsymbol{\Sigma}_n]^{-1}\hat{\mathbf{Q}}, \quad \mathbf{C}_n = [\mathbf{K}^n_q + \boldsymbol{\Sigma}_n]^{-1}.$$

Setul de date de antrenament pe baza cărora parametrii procesului Gaussian $\boldsymbol{\alpha}$ şi $\mathbf{C}$ sunt calculate, se numesc setul de vectori de bază. Deoarece funcţia medie μ poate fi setată la zero fără a pierde din generalitate, singura parte nespecificată a procesului Gaussian este funcţia de covarianţă. Alegerea unei funcţii de covarianţă şi ajustarea parametrilor acestei funcţii pentru a captura atribuţiile datelor rezultate dintr-o anumită problemă sunt cruciale pentru a obţine performanţă bună şi precizie mare în aproximare. Pentru a adapta metodele noastre la natura secvenţială a obţinerii datelor de antrenament în experimentele de control robotic, aplicăm o versiune on-line a regresiei cu procesele Gaussiene, unde parametrii $\boldsymbol{\alpha}$ şi $\mathbf{C}$ sunt actualizate de fiecare dată când obţinem date noi. Actualizările on-line permit aproximarea funcţiilor de valori în paralel cu interacţiunea robotului cu mediul de execuţie. Folosind funcţii de valori aproximate cu procese Gaussiene, varianţa gradientului estimat poate fi redusă semnificativ în metodele de tip policy gradient, ceea ce duce la convergenţă mai rapidă şi performanţă de vârf mai înaltă. Performanţa metodelor de învăţare bazate pe diferenţe temporale poate fi îmbunătăţită în acelaşi fel. [Jakab et al., 2011] [Jakab and Csató, 2010].

Datorită gradului mare de incertitudine din sistemele de învăţare din lumea reală, performanţa algoritmilor de învăţare este afectată grav de măsurătorile zgomotoase şi perturbaţii exterioare. Teza arată că combinaţia modelelor probabilistice pentru funcţii de valori cu metode existente de tip policy gradient şi metode de învăţare cu întărire clasice rezultă într-o creştere semnificativă a performanţei în cazul problemelor de control în domeniul roboticii.

4 Eficienţă de date şi stabilitate sporită în cazul estimării neparametrice a valorilor

Datorită naturii neparametrice a metodei de aproximare de valori prezentată mai sus, complexitatea computaţională a metodelor noastre creşte cubic cu numărul datelor de antrenament folosite. Pentru a remedia această problemă, am folosit o metodă de rărire bazată pe distanţa Kullback-Leibler între două procese Gaussiene. Am dezvoltat o metodă prin care limitarea numărului maxim de elemente în setul de vectori de bază poate fi combinată cu eliminarea graduală a măsurătorilor vechi şi înlocuirea acestora cu date noi de antrenament. Ideea principală a metodei este de a introduce o variabilă bazată pe timp pentru fiecare vector de bază, semnificând stagiul în care acesta a fost adăugat.

$$D = \{(x_1, y_1), \dots x_n, y_n\} \rightarrow \{(x_1, y_1, t_1), \dots (x_n, y_n, t_n)\} \tag{.1}$$

Modificând criteriul de eliminare bazată pe distanţa Kullback-Leibler prin adăugarea unui termen care penalizează vectorii de bază vechi am eliminat nevoia de a reîncepe estimarea după fiecare pas de îmbunătăţire a strategiei:

$$\varepsilon'(i) = (1 - \lambda)\frac{\alpha^2(i)}{q(i) + c(i)} + \lambda g\left(t(i)\right), \tag{.2}$$

unde termenul $\lambda \in [0, 1]$ serveşte ca şi un factor de echilibru între precizitate şi pierderea de informaţie. Accentuarea conţinutului de informaţie scade procentajul de puncte de date noi în setul de antrenament, iar accentuarea termenului dependent de timp duce la eliminarea mai rapidă a măsurătorilor vechi.

5 Strategii de explorare inteligente

A treia contribuţie majoră a tezei constă în dezvoltarea unor strategii noi de explorare, bazate pe modelul probabilistic estimat al funcţiei de valoare cu ajutorul proceselor Gaussiene. Căutarea unor strategii optimale de control este realizată prin executarea unor acţiuni neîncercate, cu rezultate necunoscute. Selecţionarea acţiunilor exploratorii însă trebuie făcută cu grijă în cazul problemelor de control în robotică, deoarece executarea acţiunilor este costisitoare.

Ideea principală a primului model de explorare adaptivă este bazată pe certi-

tudinea estimării funcţiei de valoare cu procesul Gaussian. Varianţa fixată a zgomotului de explorare din strategiile stohastice obişnuite este înlocuită cu un termen dependent de varianţa predicţiilor valorilor. Strategia stohastică rezultată poate fi exprimată în forma următoare:

$$\begin{aligned} \pi_\theta &= f(s,\theta) + \mathcal{N}(0, \sigma^2_{\mathcal{GP}} I) \\ \sigma^2_{\mathcal{GP}} &= \lambda \left(k_q\left(x^*, x^*\right) - \mathbf{k}^* \mathbf{C}_n \mathbf{k}^{*T} \right) \end{aligned} \qquad \text{cu} \quad x^* = \{s, f(s,\theta)\},$$

unde $x^* \stackrel{\text{def}}{=} (s, f(s,\theta))$, k_q este funcţia de covarianţă, iar $\mathbf{C}_n$ este parametrul procesului Gaussian după procesarea a n puncte de antrenament. Modelul de zgomot adaptiv rezultă în explorare sporită în regiunile spaţiului cu incertitudine mare, ceea ce duce la o explorare mai uniformă a spaţiului întreg.

Al doilea model de explorare este dezvoltat prin influenţarea gradului şi a direcţiei acţiunilor de explorare cu ajutorul valorilor estimate ale acţiunilor din împrejurimea acţiunii selectate de controlorul deterministic $f(s,\theta)$. Ideea este realizată prin propunerea unei stregii bazate $\pi(a|s)$ în forma unei distribuţii Boltzmann peste acţiunile posibile din împrejurimea $f_\theta(s)$:

$$\pi(a|s) = \frac{e^{\beta E(s,a)}}{Z(\beta)}, \quad \text{unde } Z(\beta) = \int da\, e^{\beta E(s,a)}$$

Termenul $Z(\theta)$ este un factor de normalizare, iar β este temperatura inversă. Pentru a include efectul controlorului deterministic în selecţia acţiunilor, funcţia de energie $E(s,a)$ este construită în forma următoare:

$$E(s,a) \;=\; Q_{\mathcal{GP}}(s,a) \cdot \exp\left[-\frac{\| a - f_\theta(s) \|^2}{2\sigma_e^2} \right].$$

Prin combinarea expresiei de sus cu strategia de selecţie Gibbs ajungem la expresia următoare pentru strategia de explorare adaptivă:

$$\pi(a|s) = \frac{\exp\left(\beta Q_{\mathcal{GP}}(s,a) \cdot \exp\left[-\frac{\|a - f_\theta(s)\|^2}{2\sigma_e^2} \right] \right)}{Z(\beta)}$$

$$Z(\beta) = \int da \exp\left(\beta Q_{\mathcal{GP}}(s,a) \cdot \exp\left[-\frac{\| a - f_\theta(s) \|^2}{2\sigma_e^2} \right] \right)$$

Această strategie face posibilă accentuarea explorării în regiunile spaţiului de stare-acţiune care au o utilitate estimată mai mare şi reprezintă o tranziţie între învăţarea de gen off-policy şi selecţionarea lacomă. [Jakab and Csató, 2011],[Jakab and Csató, 2012b].

6 Îmbunătăţirea aproximării prin folosirea informaţiilor structurale

A patra contribuţie majoră a tezei se referă la construcţia unei structurii de graf reprezentative prin datele obţinute prin experimente şi definirea aproximării funcţiilor de valori cu ajutorul proceselor Gaussiene folosind un nou kernel introdus.

În cazul mai multor probleme de control, funcţiile de valori corespunzătoare unei strategii pot prezenta discontinuităţi. Procesele Gaussiene nu pot reprezenta discontinuităţile prezente folosind funcţii kernel obişnuite. Pentru a remedia această problemă am dezvoltat o metodă de construire a unei noi funcţii kernel, bazată pe distanţe calculate dealungul unei suprafeţe aproximate cu ajutorul unei structuri de graf. Graful este construit din perechile stare-acţiune vizitate în urma experimentelor, condiţionat de adăugarea lor la setul de date de antrenament a procesului Gaussian.

Dacă punctul $x_t \stackrel{\text{def}}{=} (s_t, a_t)$ este adăugat la setul de date de antrenament, el constituie un vârf nou în graf, iar muchiile vârfului sunt stabilite după regula următoare:

$$E_{x_t,x_i} = \begin{cases} \|x_i - x_t\|^2 & \text{dacă} \quad \exp\left(-\|x_i - x_t\|^2\right) > \gamma \quad \gamma \in [0,1] \\ 0 & \text{în caz contrar} \end{cases} \qquad i = 1 \dots n$$

Valoarea parametrului γ limitează numărul vecinilor vârfului x_t. Un nou tip de funcţie kernel poate fi construit pe baza structurii obţinute după un număr de experimente, care foloseşte distanţele cele mai scurte dealungul grafului între două puncte:

$$k_{sp}(x, x') = A \exp\left(-\frac{\mathcal{SP}(x, x')^2}{2\sigma_{sp}}\right)$$

unde amplitudinea A şi σ_{sp} sunt hiperparametrii sistemului. Pentru tratarea spaţiilor continue introducem două metode de interpolare, care stabilesc noţiunea de distanţă dintre un punct din spaţiul continuu şi un punct din setul de date de antrenament:

$$\begin{aligned} \mathcal{SP}(x^*, x_j) &\stackrel{(1)}{=} \|x^* - x_i\|^2 + \mathbf{P}_{i,j} \quad \text{unde } x_i = \underset{x_\ell \in BV}{\operatorname{argmin}} \|x^* - x_\ell\|^2 \\ &\stackrel{(2)}{=} \mathbf{k}_{x*}^T \mathbf{P} \boldsymbol{e}_j = \sum_{i=1}^{n} k(x^*, x_i) \mathbf{P}_{i,j} \end{aligned}$$

Aici $\mathbf{P}_{i,j}$ conţine lungimea distanţelor dintre vectorii de bază x_i şi x_j. Folosind tehnicile prezentate mai sus, devine posibilă reprezentarea discontinuităţilor în funcţiile

de valori cu ajutorul proceselor Gaussiene, ceea ce îmbunătăţeşte precizia aproximării.

7 Concluzii

Obiectivul principal al tezei este de a dezvolta metode pentru învăţarea automată a strategiilor de control de către agenţi robotici inteligenţi. În urma cercetării efectuate am dezvoltat un număr de metode pentru a extinde instruirea prin întărire la probleme de control robotic cu spaţii continue şi am arătat că diferitele variante ale aproximării neparametrice cu tratarea adecvată a incertitudinii pot fi folosite cu succes pentru acest scop. Algoritmii prezentaţi tratează unele dintre problemele fundamentale ale realizării controlului robotic prin învăţare. În viitor planificăm efectuarea unor teste pe agenţi robotici reali, îmbunătăţirea eficienţei computaţionale a metodelor prezentate şi investigarea continuă a posibilităţilor de a folosi modelul probabilistic al proceselor Gaussiene în contextul învăţării strategiilor de control în robotică.

Bibliografie

J. S. Albus. A new approach to manipulator control: The cerebellar model articulation controller. *Journal of Dynamic Systems, Measurement, and Control*, pages 220–227, 1975.

A. Antos, R. Munos, and C. Szepesvári. Fitted Q-iteration in continuous action-space mdps. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. MIT Press, 2007.

K. E. Atkinson. *An Introduction to Numerical Analysis*. Wiley, New York, 1978.

J. A. D. Bagnell and J. Schneider. Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings of the International Conference on Robotics and Automation 2001*. IEEE, May 2001.

L. Baird. Residual algorithms: Reinforcement learning with function approximation. In *In Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37. Morgan Kaufmann, 1995.

L. Baird and A. Moore. Gradient descent for general reinforcement learning. In *In Advances in Neural Information Processing Systems 11*, pages 968–974. MIT Press, 1998.

A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to act using real-time dynamic programming. *Artif. Intell.*, 72(1-2):81–138, 1995. ISSN 0004-3702.

D. J. Benbrahim H. and F. J. Real-time learning: A ball on a beam. In *Proceedings of the international joint conference on neural networks*, volume Proceedings of the international joint conference on neural networks, pages 92–103, 1992.

D. A. Berry and B. Fristedt. Bandit problems: Sequential allocation of experiments. In *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1985.

D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 1995.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific,

1996. ISBN 1886529108.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. Boutilier, R. Dearden, and M. Goldszmidt. Stochastic dynamic programming with factored representations. *Artif. Intell.*, 121:49–107, August 2000. ISSN 0004-3702.

J. Boyan and A. Moore. Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. Touretzky, and T. Lee, editors, *Neural Information Processing Systems 7*, pages 369–376, Cambridge, MA, 1995. The MIT Press.

J. A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2-3):233–246, 2002.

S. J. Bradtke, A. G. Barto, and P. Kaelbling. Linear least-squares algorithms for temporal difference learning. In *Machine Learning*, pages 22–33, 1996.

P. Corke. A robotics toolbox for MATLAB. *IEEE Robotics and Automation Magazine*, 3(1):24–32, Mar. 1996.

L. Csató. *Gaussian Processes – Iterative Sparse Approximation*. PhD thesis, Neural Computing Research Group, March 2002.

L. Csató and M. Opper. Sparse representation for Gaussian process models. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, volume 13, pages 444–450. The MIT Press, 2001.

L. Csató and M. Opper. Sparse on-line Gaussian Processes. *Neural Computation*, 14 (3):641–669, 2002.

L. Csató, E. Fokoué, M. Opper, B. Schottky, and O. Winther. Efficient approaches to Gaussian process classification. In *NIPS*, volume 12, pages 251–257. The MIT Press, 2000.

E. E. Dar and Y. Mansour. Learning rates for q-learning. *Journal of Machine Learning Research*, 5:1–25, 2003.

R. Dearden, N. Friedman, and S. Russell. Bayesian q-learning. In *In AAAI/IAAI*, pages 761–768. AAAI Press, 1998.

M. P. Deisenroth and C. E. Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, June 2011.

M. P. Deisenroth, C. E. Rasmussen, and J. Peters. Gaussian process dynamic programming. *Neurocomputing*, 72(7-9):1508–1524, 2009. ISSN 0925-2312. doi:

http://dx.doi.org/10.1016/j.neucom.2008.12.019.

G. Endo, J. Morimoto, T. Matsubara, J. Nakanishi, and G. Cheng. Learning cpg sensory feedback with policy gradient for biped locomotion for a full-body humanoid. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, pages 1267–1273. AAAI Press, 2005. ISBN 1-57735-236-x.

Y. Engel, S. Mannor, and R. Meir. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In *Proc. of the 20th International Conference on Machine Learning*, pages 154–161, 2003.

Y. Engel, S. Mannor, and R. Meir. Reinforcement learning with Gaussian processes. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 201–208, New York, NY, USA, 2005. ACM. doi: http://doi.acm.org/10.1145/1102351.1102377.

Y. Engel, P. Szabo, and D. Volkinshtein. Learning to control an octopus arm with gaussian process temporal difference methods. *Advances in Neural Information Processing Systems*, 14:347–354, 2006.

D. Ernst, P. Geurts, L. Wehenkel, and L. Littman. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

J. Frey. Introduction to stochastic search and optimization: Estimation, simulation, and control. james c. spall. *Journal of the American Statistical Association*, 99:1204–1205, 2004.

C. Gearhart. Genetic programming as policy search in markov decision processes. *Genetic Algorithms and Genetic Programming at Stanford*, page 61?67, 2003.

M. Ghavamzadeh and Y. Engel. Bayesian policy gradient algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *NIPS '07: Advances in Neural Information Processing Systems 19*, pages 457–464, Cambridge, MA, 2007. MIT Press.

G. Gordon. Stable function approximation in dynamic programming. In *Proceedings of IMCL '95*, 1995.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.*, 5:1471–1530, December 2004. ISSN 1532-4435.

H. Hachiya, T. Akiyama, M. Sugiayma, and J. Peters. Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Netw.*, 22:1399–1410, December 2009. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.01.002.

K. Harbick, J. F. Montgomery, and G. S. Sukhatme. Planar spline trajectory following for an autonomous helicopter. In *in IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 237–242, 2001.

M. E. Harmon and S. S. Harmon. Reinforcement learning: A tutorial., 1997.

P. Hennig. Optimal reinforcement learning for gaussian systems. Technical Report arXiv:1106.0800, Max Planck Institute for Intelligent Systems Department of Empirical Inference, Spemannstrasse 38, 72070 Tübingen, Germany, Jun 2011.

T. Herbert. *Modeling and Control of Robot Manipulators, Lorenzo Sciavicco and Bruno-Siciliano*, volume 21. Kluwer Academic Publishers, Hingham, MA, USA, January 1998. doi: 10.1023/A:1007979428654.

G. Hornby, S. Takamura, J. Yokono, O. Hanagata, T. Yamamoto, and M. Fujita. Evolving robust gaits with aibo, 2000.

A. J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. In *Advances in Neural Information Processing Systems 15*, pages 1547–1554. MIT Press, 2002a.

J. A. Ijspeert, J. Nakanishi, and S. Schaal. movement imitation with nonlinear dynamical systems in humanoid robots. In *international conference on robotics and automation (icra2002)*, 2002b.

H. Jakab. A frame based motion system for the aibo four legged robotic agent. Master's thesis, Computer Science Department, Babes-Bolyai University, 2008.

H. Jakab. Guided exploration in policy gradient algorithms using Gaussian process function approximation. In *volume of extended abstracts CSCS2010, Conference of PhD Students in Computer Science*, 2010.

H. Jakab. Controlling the swinging atwood's machine using reinforcement learning. *Müszaki tudományos füzetek: XVI. FMTÜ international scientific conference*, pages 141–145, 2011a. ISSN 2067 - 6808.

H. Jakab. Geodesic distance based kernel construction for Gaussian process value function approximation. *Studia Universitatis Babes-Bolyai Series Informatica*, 61(3): 51–57, 2011b. ISSN 1224-869.

H. Jakab. Geodesic distance based kernel construction for Gaussian process value function approximation. *KEPT-2011:Knowledge Engineering Principles and Techniques International Conference, Selected Papers.*, 2011c. ISSN 2067-1180.

H. Jakab and L. Csató. Q-learning and policy gradient methods. *Studia Universitatis Babes-Bolyai Series Informatica*, 59:175–179, 2009. ISSN 1224-869.

H. Jakab and L. Csató. Using Gaussian processes for variance reduction in policy gradient algorithms. In A. Egri-Nagy, E. Kovács, G. Kovásznai, G. Kusper, and T. Tómács, editors, *ICAI2010: Proceedings of the 8th International Conference on Applied Informatics*, volume 1, pages 55–63, Eger, Hungary, 2010. BVB. ISBN 978-963-989-72-3.

H. Jakab and L. Csató. Guided exploration in direct policy search with Gaussian processes. *Acta Cybernetica*, Under Review, 2011.

H. Jakab and L. Csató. Improving Gaussian process value function approximation in policy gradient algorithms. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, volume 6792 of *Lecture Notes in Computer Science*, pages 221–228. Springer, 2011. ISBN 978-3-642-21737-1.

H. Jakab and L. Csató. Manifold-based non-parametric learning of action-value functions. In *European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium., 2012a.

H. Jakab and L. Csató. Reinforcement learning with guided policy search using Gaussian processes. In *International Joint Conference on Neural Networks (IJCNN)*, 2012b.

H. Jakab, B. Bócsi, and L. Csató. Non-parametric value function approximation in robotics. In H. F. Pop, editor, *MACS2010: The 8th Joint Conference on Mathematics and Computer Science*, volume Selected Papers, pages 235–248. Györ:NOVADAT, 2011. ISBN 978-963-9056-38-1.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

A. Kanarachos, M. Sfantsikopoulos, and P. Vionis. A splines?based control method for robot manipulators. *Robotica*, 7(03):213–221, 1989. doi: 10.1017/S026357470000607X.

M. Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6):718–727, 1999.

M. S. Kim and W. Uther. Automatic gait optimisation for quadruped robots. In *In Australasian Conference on Robotics and Automation*, 2003.

H. Kimura and S. Kobayashi. Reinforcement learning for continuous action using stochastic gradient ascent. *Intelligent Autonomous Systems (IAS-5)*, pages 288–295, 1998.

N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *in Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2619–2624, 2004.

K. J. Kyriakopoulos and G. N. Saridis. Minimum jerk for trajectory planning and control. *Robotica*, 12:109–113, 1994.

M. G. Lagoudakis and R. Parr. Model-free least squares policy iteration. Technical report, Advances in Neural Information Processing Systems, 2001.

M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4: 1107–1149, December 2003. ISSN 1532-4435.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules*1. *Advances in Applied Mathematics*, 6:4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981.

H. Maei, C. Szepesvari, S. Bhatnagar, D. Precup, D. Silver, and R. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in Neural Information Processing Systems NIPS*, 22:1204–1212, 2009.

H. Miyamoto, F. Gandolfo, H. Gomi, S. Schaal, Y. Koike, O. Rieka, E. Nakano, Y. Wada, and M. Kawato. a kendama learning robot based on a dynamic optimization principle. In *preceedings of the international conference on neural information processing*, pages 938–942, 1996a.

H. Miyamoto, S. Schaal, F. Gandolfo, Y. Koike, R. Osu, E. Nakano, Y. Wada, and M. Kawato. a kendama learning robot based on bi-directional theory. *Neural Networks*, 8(8):1281–1302, 1996b.

A. Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. In L. Birnbaum and G. Collins, editors, *Machine Learning: Proceedings of the Eighth International Conference*, 340 Pine Street, 6th Fl., San Francisco, CA 94104, June 1991. Morgan Kaufmann.

A. W. Moore and C. G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Mach. Learn.*, 13(1):103–130, 1993. ISSN 0885-6125. doi: http://dx.doi.org/10.1023/A:1022635613229.

A. W. Moore and C. G. Atkeson. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces. *Mach. Learn.*, 21:199–233, December 1995. ISSN 0885-6125. doi: 10.1023/A:1022656217772.

D. E. Moriarty, A. C. Schultz, and J. J. Grefenstette. Evolutionary algorithms for

reinforcement learning. *Journal of Artificial Intelligence Research*, 11:241–276, 1999.

J. Morimoto and K. Doya. Robust reinforcement learning. *Neural Comput.*, 17:335–359, February 2005. ISSN 0899-7667. doi: 10.1162/0899766053011528.

Y. Nakamura, T. Mori, and S. Ishii. Natural policy gradient reinforcement learning for a cpg control of a biped robot. In *PPSN*, pages 972–981, 2004.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press, 2010.

P. Olivier, P. J.P., S. C., S. S., and W. J.A. Swinging atwood's machine: Experimental and numerical results, and a theoretical study. *Physica D*, 239:1067–1081, 2010.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition, 1991.

J. Peng and R. J. Williams. Efficient learning and planning within the dyna framework. *Adapt. Behav.*, 1(4):437–454, 1993. ISSN 1059-7123. doi: http://dx.doi.org/10.1177/105971239300100403.

J. Peters and S. Schaal. Policy gradient methods for robotics. In *IROS*, pages 2219–2225. IEEE, 2006.

J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 745–750, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3.

J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008a.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71:1180–1190, March 2008b. ISSN 0925-2312.

D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*, pages 417–424, 2001.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

C. E. Rasmussen and M. Kuss. Gaussian processes in reinforcement learning. In L. K. S. Thrun, S. and B. Schölkopf, editors, *NIPS 2003*, pages 751–759. MIT Press,

2004.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

M. Riedmiller. Neural fitted q iteration ? first experiences with a data efficient neural reinforcement learning method. In *In 16th European Conference on Machine Learning*, pages 317–328. Springer, 2005.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387212396.

T. Rückstieß, M. Felder, and J. Schmidhuber. State-dependent exploration for policy gradient methods. In *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 234–249, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87480-5.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, 2003.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, 2002.

A. Schwartz. A reinforcement learning method for maximizing undiscounted rewards. In *ICML*, pages 298–305, 1993.

D. F. Sebastian Thrun, Wolfram Burgard. *Probabilistic Robotics*. The MIT Press, Cambridge, MA, 2005.

M. R. Shaker, S. Yue, and T. Duckett. Vision-based reinforcement learning using approximate policy iteration. In *14th International Conference on Advanced Robotics (ICAR)*, 2009.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

R. Smith. Open dynamics engine v0.5 user guide, 2006.

M. Sugiyama, H. Hachiya, C. Towell, and S. Vijayakumar. Geodesic gaussian kernels for value function approximation. *Auton. Robots*, 25:287–304, October 2008. ISSN 0929-5593.

M. Sugiyama, H. Hachiya, H. Kashima, and T. Morimura. Least absolute policy iteration for robust value function approximation. In *Proceedings of the 2009 IEEE international conference on Robotics and Automation*, ICRA'09, pages 699–704, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-2788-8.

R. S. Sutton. First results with dyna, an integrated architecture for learning, planning

and reacting. *Neural networks for control*, pages 179–189, 1990.

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

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *NIPS*, pages 1057–1063, 1999.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2011.

I. Szita and A. Lörincz. Optimistic initialization and greediness lead to polynomial time learning in factored mdps. In *ICML*, page 126, 2009.

J. N. Tsitsiklis and B. V. Roy. An analysis of temporal-difference learning with function approximation. Technical report, IEEE Transactions on Automatic Control, 1997.

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, London, 1989.

C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

D. Wierstra, A. Foerster, J. Peters, and J. Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 2007.

C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8:514–520, 1996.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

X. Xu, T. Xie, D. Hu, and X. Lu. Kernel least-squares temporal difference learning. *International Journal of Information Technology*, pages 55–63, 2005.

P. Zhang, X. Xu, C. Liu, and Q. Yuan. Reinforcement learning control of a real mobile robot using approximate policy iteration. In *Proceedings of the 6th International Symposium on Neural Networks: Advances in Neural Networks - Part III*, ISNN 2009, pages 278–288, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-01512-0.