ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ ΡΟΜΠΟΤΙΚΟΥ ΕΛΕΓΧΟΥ ΜΕΣΩ ΠΙΘΑΝΟΤΙΚΟΥ ΣΥΜΠΕΡΑΣΜΟΥ

Νίκος Βλάσσης **Γεώργιος Κόντες** **Σάββας Πιπερίδης**

Εργαστήριο Ευφυών Συστημάτων και Ρομποτικής
Τμήμα Μηχανικών Παραγωγής και Διοίκησης
Πολυτεχνείο Κρήτης, Χανιά
Email: *vlassis@dpem.tuc.gr*

ΠΕΡΙΛΗΨΗ

Παρουσιάζουμε μία νέα προσέγγιση στο πρόβλημα της αυτόματης μάθησης ρομποτικού ελέγχου με Ενισχυτική Μάθηση (Reinforcement Learning, RL). Πρόσφατες εργασίες στη βιβλιογραφία έχουν δείξει ότι ένα πρόβλημα Βέλτιστου Ελέγχου Διακριτού Χρόνου (Discrete Time Optimal Control) μπορεί να αναχθεί σε ένα πρόβλημα Πιθανοτικού Συμπερασμού (Probabilistic Inference) και να λυθεί με αντίστοιχες τεχνικές. Στην παρούσα εργασία δείχνουμε ότι μια τέτοια αναγωγή είναι επίσης δυνατή στην περίπτωση που το δυναμικό μοντέλου του συστήματος είναι άγνωστο, οπότε η μάθηση του ρομποτικού ελέγχου θα πρέπει να γίνει με μεθοδολογίες δοκιμής-και-σφάλματος (trial-and-error). Η ανάλυση που προτείνουμε οδηγεί σε ένα Monte-Carlo αλγόριθμο Προσδοκίας-Μεγιστοποίησης (Expectation-Maximization, EM) σε ένα μοντέλο μικτής κατανομής πιθανότητας (probabilistic mixture model). Παραθέτουμε αποτελέσματα από την εφαρμογή του προτεινόμενου αλγορίθμου σε ένα πρόβλημα ισορροπίας κινούμενου ρομπότ.

Λέξεις κλειδιά: Μάθηση ρομποτικού ελέγχου, Ενισχυτική μάθηση, Πιθανοτικός συμπερασμός, Ισορροπούμενο ρομπότ.

1 ΕΙΣΑΓΩΓΗ

Η Ενισχυτική Μάθηση (Reinforcement Learning, RL) αποτελεί αξιοσημείωτο παράδειγμα στην έρευνα για αυτόματη μάθηση ρομποτικού ελέγχου, με πολλά πρόσφατα αποτελέσματα (Tedrake et al., 2005; Abbeel et al., 2007; Kober and Peters, 2009). Οι περισσότερες εφαρμογές RL στη ρομποτική βασίζονται σε αλγορίθμους πολιτικής κλίσης (policy gradients) (Ng and Jordan, 2000; Peters and Schaal, 2006). Ένας αλγόριθμος πολιτικής κλίσης υπολογίζει και ακολουθεί σε κάθε βήμα το στοχαστικό διάνυσμα κλίσης (gradient) της απόδοσης της πολιτικής, μέχρι να συγκλίνει σε κάποιο (τοπικό) μέγιστο. Αλγόριθμοι τέτοιου είδους έχουν αναλυθεί εκτενώς και έχουν αποδειχτεί ιδιαίτερα επιτυχημένοι στην πράξη, ωστόσο πάσχουν επίσης και από αρκετά μειονεκτήματα, όπως είναι η εξάρτησή τους από το ρυθμό μάθησης (learning rate), η αργή τους σύγκλιση, και η ευαισθησία τους σε τοπικά βέλτιστα.

Οι Dayan and Hinton (1997) έδειξαν πως ένα RL πρόβλημα μπορεί να αντιμετωπιστεί με τον αλγόριθμο Expectation-Maximization (EM) (Dempster et al., 1977). Η βασική ιδέα είναι η

μοντελοποίηση των άμεσων ανταμοιβών (rewards) του RL ως πιθανότητες κάποιων εικονικών γεγονότων, οπότε μπορούμε να χρησιμοποιήσουμε τεχνικές πιθανοτικού συμπερασμού για βελτιστοποίηση, όπως ο αλγόριθμος EM. Πρόσφατα, οι Kober and Peters (2009) ανέπτυξαν έναν αλγόριθμο βασισμένο στον EM, ο οποίος ονομάζεται PoWER, για τη μάθηση παραμετροποιημένων πολιτικών ρομποτικού ελέγχου σε ένα επεισοδιακό RL σενάριο. Ο PoWER κληρονομεί πολλά από τα πλεονεκτήματα του αλγορίθμου EM, όπως η απλότητα υλοποίησης, η μη ανάγκη ύπαρξης και υπολογισμού ρυθμού μάθησης, η ταχύτερη σύγκλιση από αλγορίθμους πολιτικής κλίσης, και η καλύτερη συμπεριφορά σε τοπικά βέλτιστα. Σε πολλά ρομποτικά προβλήματα ο PoWER επέδειξε καλύτερη απόδοση από αρκετούς state-of-the-art αλγορίθμους πολιτικής κλίσης.

Στην παρούσα εργασία προτείνουμε μια νέα προσέγγιση στη μοντελοποίηση και επίλυση RL προβλημάτων στη ρομποτική. Η προσέγγισή μας βασίζεται στην εργασία των Toussaint and Storkey (2006) οι οποίοι έδειξαν έναν τρόπο αναγωγής ενός προβλήματος Βέλτιστου Ελέγχου Διακριτού Χρόνου (Discrete Time Optimal Control) σε ένα πρόβλημα συμπερασμού σε ένα πιθανοτικό μοντέλο (Probabilistic Inference). Οι Toussaint and Storkey (2006) αντιμετώπισαν την περίπτωση στην οποία γνωρίζουμε το δυναμικό μοντέλο του φυσικού συστήματος. Στην παρούσα εργασία προτείνουμε μια προσέγγιση που επιτρέπει να αντιμετωπίσουμε και την περίπτωση που δεν διαθέτουμε (ή είναι δύσκολο να εκτιμήσουμε) το δυναμικό μοντέλο του συστήματος, μια ρεαλιστική υπόθεση για πολύπλοκα ρομποτικά συστήματα. Το παραγόμενο πιθανοτικό μοντέλο είναι μία μικτή κατανομή πιθανότητας (probabilistic mixture model). Η μάθηση του μοντέλου αυτού αντιστοιχεί στην επίλυση του αρχικού RL προβλήματος και μπορεί να γίνει με τη βοήθεια ενός Monte-Carlo EM αλγορίθμου (Wei and Tanner, 1990). Μάλιστα μια εκδοχή του Monte-Carlo EM ανάγεται ακριβώς στον αλγόριθμο PoWER των Kober and Peters (2009).

2 ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ ΩΣ ΠΙΘΑΝΟΤΙΚΟΣ ΣΥΜΠΕΡΑΣΜΟΣ

Μοντελοποιούμε το ρομποτικό πρόβλημα ως μια διακριτού χρόνου και πεπερασμένου ορίζοντα Μαρκοβιανή Διαδικασία Απόφασης (Markov Decision Process, MDP) με συνεχείς καταστάσεις $x \in \mathbb{R}^n$ και ενέργειες $u \in \mathbb{R}$. Το ρομπότ εκκινεί από μια κατάσταση x_0 (ή μια κατανομή γύρω από αυτήν την κατάσταση) και ακολουθεί στοχαστική πολιτική $\pi_\theta(u_t|x_t)$ παραμετροποιημένη με παραμέτρους θ. Θεωρούμε ότι σε κάθε χρονικό βήμα t το ρομπότ συλλέγει άμεση ανταμοιβή r_t, η οποία είναι μια συνάρτηση της κατάστασης x_t και της ενέργειας u_t. Υποθέτουμε πως δεν έχουμε πρόσβαση στο μοντέλο μετάβασης του MDP, αλλά μπορούμε να λάβουμε δείγματα από τροχιές ξεκινώντας από την κατάσταση x_0 και ακολουθώντας κάποια πολιτική. Χρησιμοποιώντας μόνο δειγματική εμπειρία από το MDP, θέλουμε να εκτιμήσουμε εκείνο το θ που μεγιστοποιεί την αναμενόμενη ανταμοιβή (expected cumulative reward, value of policy)

$$J(\theta) = E\Big[\sum_{t=0}^{H} r_t; \theta\Big], \tag{1}$$

όπου H είναι ο ορίζοντας, και ο τελεστής προσδοκίας $E[\cdot]$ αφορά όλες τις πιθανές τροχιές που μπορούν να προκύψουν ξεκινώντας από την κατάσταση x_0 και ακολουθώντας πολιτική π_θ. Στην παρούσα εργασία ενδιαφερόμαστε για αλγορίθμους RL που δεν χρησιμοποιούν ούτε προσπαθούν να εκτιμήσουν το δυναμικό μοντέλο του συστήματος, ούτε βασίζονται σε κάποια συνάρτηση αξίας (στη βιβλιογραφία τέτοιου τύπου αλγόριθμοι αναφέρονται ως model-free RL).

Όταν ένα μοντέλο του MDP είναι διαθέσιμο, και οι ανταμοιβές r_t είναι μη αρνητικές ποσότητες (π.χ. μετά από κανονικοποίηση ισχύει $r_t \in [0, 1]$), οι Toussaint and Storkey (2006) έδειξαν ότι είναι εφικτό να αναγάγουμε τη βελτιστοποίηση της $J(\theta)$ σε ένα πρόβλημα πιθανοτικού συμπερασμού πάνω σε ένα μείγμα από MDPs πεπερασμένου ή άπειρου ορίζοντα. Σε αυτή την προσέγγιση ο ορίζοντας του MDP λαμβάνεται ως διακριτή τυχαία μεταβλητή T, η οποία στην περίπτωση πεπερασμένου ορίζοντα θεωρούμε ότι έχει ομοιόμορφη εκ των προτέρων κατανομή $p(T) = 1/(H + 1)$, για $T = 0, 1, \ldots, H$. Η κεντρική ιδέα, η οποία ανάγεται στους Dayan and Hinton (1997), είναι η αντιμετώπιση των άμεσων ανταμοιβών ως πιθανότητες κάποιων εικονικών γεγονότων.

Στην παρούσα εργασία θεωρούμε ότι δεν γνωρίζουμε το μοντέλο του MDP αλλά το ρομπότ μπορεί να αλληλεπιδρά με το περιβάλλον του και να συλλέγει δεδομένα. Υιοθετούμε την προσέγγιση των Toussaint and Storkey (2006) όπου η ανταμοιβή r_T που συλλέγεται σε κάποιο βήμα T εκλαμβάνεται ως η πιθανότητα το εικονικό γεγονός R να συμβεί στο τελικό βήμα μιας τροχιάς μήκους T. Έστω ξ_T μια τέτοια τροχιά και $p(\xi_T|T; \theta)$ η πιθανοφάνεια (παραμετροποιημένη ως προς θ) να παρατηρήσουμε την ξ_T υπό πολιτική π_θ. Έστω επίσης $p(R|\xi_T)$ η πιθανότητα να συμβεί το γεγονός R στο τελικό βήμα της τροχιάς ξ_T, και $r_{\xi T} \equiv p(R|\xi_T)$. Τότε είναι εύκολο να δειχθεί ότι η αναμενόμενη ανταμοιβή $J(\theta)$ είναι ανάλογη της συνάρτησης πιθανοφάνειας κάποιας μικτής κατανομής:

$$J(\theta) \propto \sum_{T=0}^{H} p(T) \int_{\xi_T} p(\xi_T|T; \theta)\, p(R|\xi_T)\, d\xi_T = E_{p(\xi_T, T; \theta)}\, p(R|\xi_T). \tag{2}$$

Το μοντέλο αυτό επιτρέπει την αναγωγή ενός RL προβλήματος σε ένα πρόβλημα πιθανοτικού συμπερασμού: υποθέτουμε ότι το εικονικό γεγονός R παρατηρήθηκε, και θέλουμε να συμπεράνουμε τις παραμέτρους θ του μοντέλου ώστε να μεγιστοποιήσουμε την πιθανοφάνεια αυτής της παρατήρησης. Η συνάρτηση $J(\theta)$ στην (2) είναι ένα μείγμα πιθανοφανειών, οπότε μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο EM για βελτιστοποίηση, όπως εξηγούμε παρακάτω.

3 Ο ΑΛΓΟΡΙΘΜΟΣ ΠΡΟΣΔΟΚΙΑΣ-ΜΕΓΙΣΤΟΠΟΙΗΣΗΣ

Θέλουμε να μεγιστοποιήσουμε ως προς θ τη συνάρτηση $J(\theta)$ από την (2). Ισοδύναμα μπορούμε να μεγιστοποιήσουμε το λογάριθμο της συνάρτησης πιθανοφάνειας $L(\theta) = \log J(\theta)$. Ο αλγόριθμος προσδοκίας-μεγιστοποίησης (EM) μεγιστοποιεί επαναληπτικά μια συνάρτηση ενέργειας $F(\theta, q) = L(\theta) - D_{KL}\big[q(\xi_T, T) \,||\, p(\xi_T, T|R; \theta)\big]$ η οποία αποτελεί κάτω φράγμα της $L(\theta)$ (Neal and Hinton, 1998). Η ενέργεια F είναι μία συνάρτηση των αγνώστων παραμέτρων θ και μιας αυθαίρετης κατανομής $q \equiv q(\xi_T, T)$ πάνω στις 'λανθάνουσες' μεταβλητές ξ_T, T.

Ο αλγόριθμος EM εναλλάσσεται μεταξύ δύο βημάτων. Στο βήμα Ε κρατάμε σταθερές τις παραμέτρους θ και μεγιστοποιούμε την F ως προς την κατανομή q. Στο βήμα Μ κρατάμε σταθερή την κατανομή q και μεγιστοποιούμε την F ως προς τις παραμέτρους θ. Αυτή η επαναληπτική διαδικασία συγκλίνει σε ένα τοπικό μέγιστο της F (το οποίο συχνά είναι και τοπικό μέγιστο της L). Στην περίπτωσή μας, στο βήμα Ε η βέλτιστη κατανομή q^* που μεγιστοποιεί την F είναι η εκ των υστέρων κατανομή Bayes υπολογισμένη για τις παραμέτρους θ_{old} από το προηγούμενο βήμα Μ:

$$q^*(\xi_T, T) = p(\xi_T, T|R; \theta_{old}) \propto p(T)\; p(\xi_T|T; \theta_{old})\; p(R|\xi_T) = p(T)\; p(\xi_T|T; \theta_{old})\; r_{\xi T}. \tag{3}$$

Για $q = q^*$ η ενέργεια δίνεται από τη σχέση

$$F(\theta, q^*) = E_{T\sim p(T)} E_{\xi_T \sim p(\xi_T|T;\theta_{old})} \big[r_{\xi T} \log p(\xi_T|T;\theta) \big]. \tag{4}$$

Καθώς δεν διαθέτουμε δυναμικό μοντέλο του MDP δεν μπορούμε να μεγιστοποιήσουμε την F ακριβώς. Μπορούμε όμως να την προσεγγίσουμε με δειγματοληψία τροχιών από το MDP. Συγκεκριμένα, εκτελούμε την πολιτική $\pi_{\theta_{old}}$ για H βήματα, ξεκινώντας από το x_0 και για τις παραμέτρους θ_{old} που είχαμε υπολογίσει στο προηγούμενο βήμα Μ, και συλλέγουμε δειγματικές τροχιές ξ μήκους H. Στη συνέχεια χρησιμοποιούμε όλες τις υπο-τροχιές ξ_T, $T = 0, \ldots, H$, κάθε δειγματικής τροχιάς ξ, για τον υπολογισμό μιας εκτιμήτριας της F:

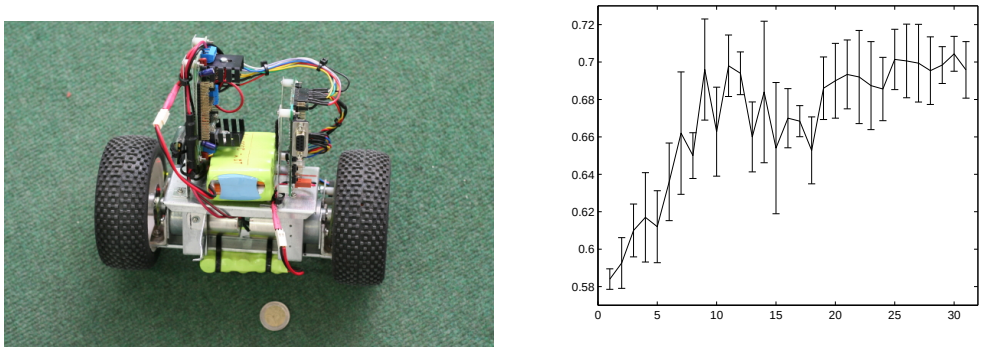
$$F(\theta, q^*) \approx \left\langle \sum_{T=0}^{H} r_{\xi T} \log p(\xi_T|T;\theta) \right\rangle_{\xi}, \tag{5}$$

όπου $\langle\cdot\rangle_{\xi}$ δηλώνει δειγματικό μέσο ως προς το πλήθος των τροχιών. Η χρήση των υπο-τροχιών ως δείγματα από την κατανομή $q^*(\xi_T, T) \propto p(T)p(\xi_T|T;\theta_{old})$ δικαιολογείται από το γεγονός ότι η κατανομή μηκών τροχιών $p(T)$ είναι ομοιόμορφη, οπότε οι υπο-τροχιές μιας τροχιάς αποτελούν μη ανεξάρτητα αλλά ομοίως κατανεμημένα δείγματα της q^*. Αλγόριθμοι τέτοιας μορφής, όπου στο βήμα Μ υπολογίζεται ένας δειγματικός μέσος της ενέργειας, συναντώνται στη βιβλιογραφία με το όνομα Monte-Carlo EM (Wei and Tanner, 1990).

Στην παρούσα μελέτη παραμετροποιούμε την πολιτική όπως στην εργασία των Kober and Peters (2009), με τη μορφή $u_t = (\theta + \varepsilon_t)\phi(x_t)$, όπου $\phi(\cdot)$ είναι σταθερές συναρτήσεις βάσης και ε_t είναι τυχαίος θόρυβος με κανονική κατανομή $\varepsilon_t \sim \mathcal{N}(\varepsilon_t; 0, \sigma^2)$. Στην περίπτωση αυτή η μεγιστοποίηση της F επιδέχεται αναλυτική λύση:

$$\theta = \theta_{old} + \frac{\left\langle \sum_{t=0}^{H} Q_{\xi t} \varepsilon_{\xi t} \right\rangle_{\xi}}{\left\langle \sum_{t=0}^{H} Q_{\xi t} \right\rangle_{\xi}}, \qquad \text{όπου} \qquad Q_{\xi t} = \sum_{T=t}^{H} r_{\xi T}. \tag{6}$$

Η συγκεκριμένη εκδοχή του αλγορίθμου Monte-Carlo EM που παρουσιάσαμε ανάγεται ακριβώς στον αλγόριθμο PoWER των Kober and Peters (2009) ο οποίος έχει προκύψει με διαφορετική μαθηματική προσέγγιση.



Σχήμα 1: Αριστερά: Το ισορροπούμενο ρομπότ Robba. Δεξιά: Τα αποτελέσματα της μάθησης.

4 ΠΕΙΡΑΜΑΤΑ

Παραθέτουμε αποτελέσματα του αλγορίθμου σε ένα πρόβλημα ισορροπίας δίτροχου ρομπότ. Έχουμε φτιάξει το δικό μας δίτροχο ισορροπούμενο ρομπότ, το οποίο λέγεται Robba και φαίνεται στο Σχήμα 1. Το Robba χρησιμοποιεί διαφορική οδήγηση και έχει σχεδιαστεί με σκοπό να αποτελέσει ένα μικρού μεγέθους, χαμηλού κόστους ισορροπούμενο ρομπότ. Το όχημα περιλαμβάνει ένα πλαίσιο αλουμινίου με τα ακόλουθα εξαρτήματα: Δύο 12 Vdc, 152 RpM κινητήρες, έναν ooPIC μικροελεγκτή, έναν dual PWM οδηγό κινητήρα, δύο οδόμετρα 64 παλμών ανά περιστροφή, ένα γυροσκόπιο ενός άξονα (CRS-10 από την Silicon Sensing), και δύο μπαταρίες (μια 12V 2700mAh επαναφορτιζόμενη για τους κινητήρες και μια 6V 2700mAh επαναφορτιζόμενη για όλα τα ηλεκτρονικά συστήματα του ρομπότ). Ένα από τα κύρια θέματα που καθόρισαν τη σχεδίαση του ρομπότ ήταν οι μικρές διαστάσεις και η ανθεκτική κατασκευή, ώστε να ανταπεξέλθει τις καταπονήσεις κατά την περίοδο μάθησης. Το Robba έχει δύο τροχούς διαμέτρου 12εκ., έχει μήκος 12εκ., πλάτος 24εκ., ύψος 21εκ., και ζυγίζει 2 κιλά.

Στο πείραμά μας ξεκινάμε το ρομπότ από γωνία 0 (κάθετη θέση) και στιγμιαία δίνουμε μεγάλη ροπή και στους δύο κινητήρες. Αυτό έχει σαν αποτέλεσμα την κλίση του ρομπότ προς τα πίσω. Στόχος μας είναι να επιστρέψει το ρομπότ στην αρχική του θέση όσο πιο γρήγορα γίνεται και να ισορροπήσει. Σε κάθε βήμα 'τιμωρούμε' κάθε γωνία διαφορετική από τις 0 μοίρες και κάθε περιστροφή των τροχών χρησιμοποιώντας εκθετικές ανταμοιβές (και κανονικοποιώντας τις ανταμοιβές στο διάστημα $[0, 1]$). Ο χώρος καταστάσεων είναι διδιάστατος και περιλαμβάνει τη γωνία του ρομπότ x_1 και τη γωνιακή του ταχύτητα x_2. Ο έλεγχος του ρομπότ επιτυγχάνεται με απευθείας ανάθεση της (κοινής) ροπής u των κινητήρων με πολιτική ελέγχου $u = \theta_1 x_1 + \theta_2 x_2$. Σε κάθε επεισόδιο χρησιμοποιούμε μόνο δύο δειγματικές τροχιές, που είναι αποδεκτό μέγεθος για το δείγμα μας αν οι παράμετροι θορύβου ε_t είναι συμμετρικές μεταξύ τους, έτσι ώστε η μέση τιμή τους να είναι μηδέν (αυτό αποδείχτηκε ένα αποδοτικό 'τέχνασμα' που μας βοήθησε να μειώσουμε την πολυπλοκότητα της δειγματοληψίας στο πραγματικό πείραμα). Η καμπύλη μάθησης του προτεινόμενου αλγορίθμου φαίνεται στο Σχήμα 1. Κάθε πολιτική αξιολογήθηκε δοκιμάζοντάς την 10 φορές στο πραγματικό ρομπότ. Μετά τη μάθηση το ρομπότ ήταν ικανό να επανέρχεται γρήγορα από την αρχική διαταραχή και να σταθεροποιεί τη θέση του σε γωνία 0 μοιρών.

5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Περιγράψαμε μια νέα προσέγγιση στο πρόβλημα της μάθησης ρομποτικού ελέγχου με ενισχυτική μάθηση (RL). Η προσέγγισή μας βασίζεται στην τεχνική πιθανοτικού συμπερασμού των Toussaint and Storkey (2006) για μάθηση βέλτιστου ελέγχου με γνώση του δυναμικού μοντέλου του συστήματος, την οποία επεκτείναμε για τις περιπτώσεις που το μοντέλο του συστήματος είναι άγνωστο. Δείξαμε ότι ο αλγόριθμος PoWER των Kober and Peters (2009) μπορεί να προκύψει ως μία εκδοχή ενός Monte-Carlo EM αλγορίθμου για συμπερασμό σε ένα πιθανοτικό μοντέλο μικτής κατανομής, και δοκιμάσαμε τον αλγόριθμο σε ένα πραγματικό πρόβλημα ισορροπίας ρομπότ με ενθαρρυντικά αποτελέσματα. Η τρέχουσα έρευνα επικεντρώνεται στην επέκταση του αλγορίθμου και σε άλλα RL προβλήματα.

ΑΝΑΦΟΡΕΣ

Abbeel, P., Coates, A., Quigley, M., and Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. In *Proc. Neural Information Processing Systems*.

Dayan, P. and Hinton, G. E. (1997). Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38.

Kober, J. and Peters, J. (2009). Policy search for motor primitives in robotics. In *Proc. Neural Information Processing Systems*.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in graphical models*, pages 355–368. Kluwer Academic Publishers.

Ng, A. Y. and Jordan, M. I. (2000). PEGASUS: A policy search method for large MDPs and POMDPs. In *Proc. Uncertainty in Artificial Intelligence*.

Peters, J. and Schaal, S. (2006). Policy gradient methods for robotics. In *Proc. Int. Conf. on Intelligent Robots and Systems*.

Tedrake, R., Zhang, T. W., and Seung, H. S. (2005). Learning to walk in 20 minutes. In *Proc. 14th Yale Workshop on Adaptive and Learning Systems*.

Toussaint, M. and Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proc. Int. Conf. on Machine Learning*.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.*, 85:699–704.