MODELE DAYALI PEKİŞTİRME İLE ÖĞRENME İÇİN ÖNEM ÖRNEKLEMESİ
IMPORTANCE SAMPLING FOR MODEL-BASED REINFORCEMENT LEARNING

Orhan Sönmez, A. Taylan Cemgil

Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi
{orhan.sonmez,taylan.cemgil}@boun.edu.tr

ÖZETÇE

ıEn gelişmiş pekiştirme ile öğrenme algoritmalarının bir çoğu Bellman denklemlerini temel alır ve sabit noktalı yineleme metodlarını kullanarak kısmi en iyi sonuçlara yakınsarlar. Fakat, son dönemdeki bazı yöntemler uygun grafik modelleri kullanarak pekiştirme ile öğrenme problemini eşdeğer bir olabilirlik enbüyütme problemine çevirmekte ve böylelikle olasılıksal çıkarım metodlarının kullanımına olanak sağlamaktadır. Biz de burada beklenti adımı bir önem örnekleyicisi olan bir beklenti-enbüyütme metodu öneriyoruz ve bu metodu olabilirliği tahmin etmede ve sonrasında da en iyi ilkeyi belirlemede kullanıyoruz.

ABSTRACT

Most of the state-of-the-art reinforcement learning algorithms are based on Bellman equations and make use of fixed-point iteration methods to converge to suboptimal solutions. However, some of the recent approaches transform the reinforcement learning problem into an equivalent likelihood maximization problem with using appropriate graphical models. Hence, it allows the adoption of probabilistic inference methods. Here, we propose an expectation-maximization method that employs importance sampling in its E-step in order to estimate the likelihood and then to determine the optimal policy.

1. GİRİŞ

Bir çok kontrol ve planlama problemlerinin üzerinde tanımlandığı Markov karar süreçlerinin, karesel maliyetli doğrusal dinamik sistemler gibi özel durumlar dışında kapalı biçimde bir çözümü bulunmamaktadır [1]. Bu yüzden de, bu süreçler üzerinde tanımlanan pekiştirme ile öğrenme (PÖ) probleminin çözümünde genel olarak yakınsama metodlarına başvurulur. Öyle ki, en gelişmiş PÖ algoritmalarının bir çoğu Bellman denklemlerini temel alır ve sabit nokta yineleme metodlarını kullanarak kısmi en iyi sonuçlara yakınsarlar [2],[3].

Fakat, son dönemdeki bazı yöntemler uygun grafik modelleri kullanarak PÖ problemini eşdeğer bir olabilirlik enbüyütme problemine çevirmekte ve böylelikle olasılıksal çıkarım metodlarının kullanımına olanak sağlamaktadır. Örneğin, ilk olarak Dayan ve Hinton [4] olasılıksal bayır algoritmalarına alternatif olarak bir beklenti-enbüyütme algoritması önermiştir. Ancak, bizim makalemizin temelini oluşturan esas çalışma Toussaint ve Storkey [5] tarafından önerilen PÖ problemine eşdeğer grafik modeli ve bu grafik modelinin çözümü için sundukları beklenti-enbüyütme algoritmasıdır. Daha sonra, Furmston ve Barber [6],[7] ise önerilen grafik modelinin Markov özelliklerinden faydalanıp beklenti adımındaki tam tamına çıkarım metodunu iyileştirmişlerdir.

Fakat, çözmek istediğimiz problemdeki durum uzayı büyüdükçe önerilen bu beklenti-enbüyütme metodunun beklenti adımınında tam tamına çıkarım yapmak pratik olarak mümkün olmamaktadır. Hoffman *v.d.*[8],[9] beklenti adımında tersinir atlama Markov zinciri Monte Carlo kullanarak yaklaşık çıkarımların da bu problemin çözümünde kullanılabileceğini göstermiştir. Biz ise, beklenti adımı için bir önem örneklemesi metodu öneriyoruz ve olabilirliği tahmin etmede ve sonrasında da en iyi ilkeyi belirlemede kullanıyoruz.

Öncelikle, 2. bölümde Markov karar süreçleri ile ilgili önbilgi verip pekiştirme ile öğrenme problemini tanımlıyoruz. Daha sonra ise, bu problem için beklenti-enbüyütme algoritmasını ve beklenti adımı için önerdiğimiz önem örneklemesini 3. bölümde anlatıyoruz.

2. PROBLEM

2.1. Markov Karar Süreçleri ve Pekiştirme ile Öğrenme

Markov karar süreçleri (MKS), bir sistem içinde fayda tabanlı karar veren ajanların ardışık karar verme süreçlerini modellemek için kullanılan olasılıksal araçlardır. Bu süreç boyunca, bir x_0 durumdan başlamak kaydıyla, ajan her t anında bir $x_t \in \mathcal{X}$ durumunda bulunur. Daha sonra, π ilkesini kullanarak içinde bulunduğu x_t durumuna göre bir $a_t \in \mathcal{A}$ eylemini gerçekleştirir. Bunun sonucu olarak ise ajan bir $r_t \geq 0$ ödülü alır ve $t+1$ anı için bir x_{t+1} durumuna geçer.

Daha biçimsel olmak gerekirse, bir MKS $t = 0, 1, 2, ..., T$

Bu çalışma TÜBİTAK BİDEB Yurt İçi Doktora Burs Programı (2211) tarafından desteklenmektedir.

978-1-4673-0056-8/12/$26.00 ©2012 IEEE

için aşağıda tanımlanan olasılık modeline göre işler.

$$\begin{aligned} x_0 &\sim P(x_0) \\ a_t &\sim P(a_t|x_t;\pi) \\ r_t &\sim P(r_t|x_t,a_t) \\ x_{t+1} &\sim P(x_{t+1}|x_t,a_t) \end{aligned} \tag{1}$$

Burada $P(a_t|x_t;\pi)$, π ilkesi ile,

$$\pi_{i,a} = P(a_t = a|x_t = i;\pi) \tag{2}$$

şeklinde parametrelendirilmiş bir çokterimli olasılık dağılımını ifade etmektedir.

Bunun sonucu olarak da, belirli bir π ilkesi için verilen herhangi bir durum eylem gezinge ikilisinin $x_{0:T}, a_{0:T}$ bir MKS üzerindeki önsel dağılımı q_π aşağıdaki şekilde hesaplanmaktadır.

$$\begin{aligned} q_\pi(x_{0:T}, a_{0:T}) =& P(x_0)P(a_T|x_T;\pi) \\ & \cdot \prod_{t=0}^{T-1} P(a_t|x_t;\pi)P(x_{t+1}|x_t,a_t) \end{aligned} \tag{3}$$

PÖ problemi ise, MKSler üzerinde bir ajanın toplam ödülünü enbüyüten ilkeyi bulmak olarak tanımlanır. Fakat, MKSler olasılıksal süreçler olduğundan dolayı, toplam ödülün tüm durum-eylem gezingeleri üzerinden beklenen değerinin hesaplanması gerekmektedir.

Ayrıca, zamanın sonsuza gittiği durumlarda, toplam ödül değerinin ıraksamaması için bir $0 < \gamma \leq 1$ indirim faktörü tanımlamak gerekir.

Böylece, genel bir PÖ problemi, $\langle\cdot\rangle$ beklenen değer operatörü olmak üzere,

$$\pi^* = \arg\max_\pi \left\langle \sum_{t=0}^{\infty} \gamma^t r_t \right\rangle_{q_\pi} \tag{4}$$

biçimde tanımlanmaktadır ve çözüm π^* en iyi ilke olarak adlandırılır.

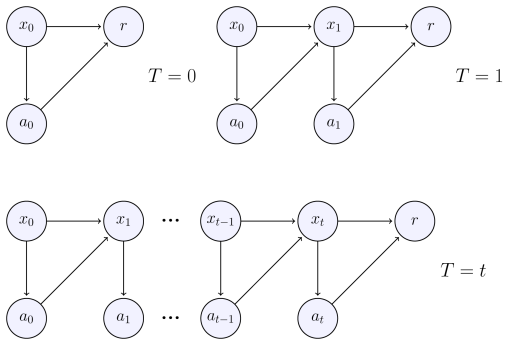
En genel PÖ algoritmaları, dinamik programlama tekniklerini kullanabilmek için durumlar veya durum-eylem ikilileri üzerinde değer fonksiyonları tanımlarlar. Daha sonra da, bu fonksiyonlar üzerinden Bellman denklemlerini yazıp değer veya ilke yinelemesi ve Q, SARSA, TD öğrenmesi gibi yinelemeli yöntemlerle [3] o denklemleri çözümüne yakınsarlar.

2.2. Grafik Modeli ve Olabilirlik

PÖ problemine son dönemlerde çıkan yeni bir bakış açısı ise bu problemi bir dinamik programlama problemi yerine bir olasılıksal çıkarım problemi olarak ele almaktır. Toussaint ve Storkey [5] Şekil 1'de grafik modeli görülen olasılık karışımınında tanımladıkları olabilirliği enbüyütmenin PÖ problemi ile eşdeğer olduğunu göstermişlerdir.

Bu olasılık karışımı, her biri sonlu sayıda olan sonsuz sayıda MKS'den oluşmaktadır. Buna göre, her bir MKS'nin bu karışımdaki önsel dağılımı sadece ve sadece kendisinin uzunluğu T'ye bağlı olup, uzunluğu $T = t$ olan MKS'nin karışımdaki önsel dağılımı aşağıdaki şekilde tanımlanmıştır.

$$P(T = t) = \gamma^t(1-\gamma) \tag{5}$$



Şekil 1: $T = 0, 1, ...$ için Sonlu Zamanlı Markov Karar Süreçleri Olasılık Karışımının Grafik Modeli

Ayrıca, bu karışımdaki her bir MKS, bölüm 2.1'de bahsedilmiş olan genel MKS tanımından farklı olarak sadece sonlandıkları zaman adımında bir ödül almaktadır. Buna göre T uzunluğundaki bir MKS için tam birleşik dağılım,

$$\begin{aligned} P(r, x_{0:T}, a_{0:T}|T;\pi) =& P(r|x_T,a_T)P(a_T|x_T;\pi)P(x_0) \\ & \cdot \prod_{t=0}^{T-1} P(x_{t+1}|x_t,a_t)P(a_t|x_t;\pi) \end{aligned} \tag{6}$$

şeklinde elde edilebilir.

Böylelikle, olasılık karışımının tam ifadesi,

$$P(r, x_{0:T}, a_{0:T};\pi) = \sum_{t=0}^{\infty} P(r, x_{0:T}, a_{0:T}|T = t;\pi)P(T = t) \tag{7}$$

şeklinde olacaktır.

Genelliği bozmadan, MKSlerdeki ödüllerin $r \in \{0, 1\}$ şeklinde olduğunu varsayalım. Zira, herhangi bir ödül sistemi uygun düzgeleme ile $[0, 1]$ aralığına eşlemlenip ödüllerin beklenen değerleri de pozitif ödül alma olasılığı olarak atanarak benzetimlenebilir. Buna göre, Toussaint ve Storkey [5] bir π ilkesinin olabilirliğini $\mathcal{L}(\pi)$, ifadesi verilmiş olasılık karışımından π ilkesini izleyerek pozitif ödül alma olasığı şeklinde tanımlamıştır.

$$\mathcal{L}(\pi) = P(r = 1;\pi) \tag{8}$$

Ve bu olabilirliği enbüyüten ilke, aynı zamanda PÖ probleminin çözümü olan en iyi ilke π^*dır.

$$\pi^* = \arg\max_\pi \mathcal{L}(\pi) \tag{9}$$

3. YÖNTEM

Bu bildiride PÖ probleminin bir özel hali olan modele dayalı PÖ probleminin çözümü için beklenti adımı bir önem örneklemesi olan bir beklenti-enbüyütme algoritması sunuyoruz. PÖ probleminin bu ayarında, sistemin bileşenleri olan başlangıç dağılımı $P(x_0)$, geçiş dağılımı $P(x_{t+1}|x_t,a_t)$ ve ödül dağılımı $P(r_t|x_t,a_t)$ ajan tarafından bilinmektedir. Gerektiği durumlarda bunların da çeşitli yöntemlerle kestirilmesi mümkün olmasına rağmen, bu konu bu bildirinin kapsamının dışındadır.

3.1. Beklenti-Enbüyütme Algoritması

Toussaint ve Storkey [5], (9) ile tanımladıkları olasılıksal çıkarım probleminin çözümü için, bir de beklenti-enbüyütme algoritma çıkarmışlardır. Rasgele bir ilke seçilerek başlayan bu algoritmanın her k adımında, bir sonraki adımdaki ilke $\pi^{(k+1)}$ aşağıdaki enbüyütme problemini çözerek bulunur.

$$\pi^{(k+1)} \leftarrow \arg\max_{\pi} \langle \log P\left(r=1, x_{0:T}, a_{0:T}, T; \pi\right) \rangle \quad (10)$$

Yukarıdaki enbüyütme problemindeki beklenen değer şu anki ilke $\pi^{(k)}$ kullanılarak,

$$P(x_{0:T}, a_{0:T}, T | r=1; \pi^{(k)}) \quad (11)$$

sonsal dağılımına göre hesaplanmaktadır.

Yinelemeler sonunda da, bu beklenti-enbüyütme algoritması bir kısmi en iyi ilkeye yakınsayacaktır.

Aslında, (10) ile tanımlanmış beklenti-enbüyütme algoritmasının enbüyütme adımı görece basittir ve kapalı biçimdeki çözümü Lagrange çarpanları ile verilen bir π ilkesinin her bir $\pi_{i,a}$ parametresi için,

$$\pi_{i,a} = \left\langle \frac{\sum_{t=0}^{T}[x_t = i \wedge a_t = a]}{\sum_{t=0}^{T}[x_t = i]} \right\rangle_{P(x_{0:T}, a_{0:T}, T | r=1; \pi)} \quad (12)$$

şeklinde elde edilebilir.

Beklenti adımında ise (11) nolu denklemde ifade edilen sonsal dağılımı hesaplamak gerekmektedir.

[5], [6], [7], farklı yaklaşımlarla bu sonsal dağılımın tamı tamına çıkarımını yapmışlardır. Ancak, bu çıkarım durum uzayı büyüdükçe zorlaşmakta ve pratik olarak imkansızlaşmaktadır.

3.2. Önem Örneklemesi

Farklılık olarak, biz bu çıkarımı tamı tamına yapmak yerine yaklaşık olarak kestiren bir önem örneklemesi metodu öneriyoruz.

Metodumuzda, teklif fonksiyonu olarak grafik modelimizin Markov özelliğinden dolayı rahatlıkla örnekleme yapabileceğimiz,

$$q(x_{0:T}, a_{0:T}, T) = P(x_{0:T}, a_{0:T}, T; \pi) \quad (13)$$

önsel dağılımını seçtik. Böylece, önem örneklemesindeki her bir $s = (x_{0:T}, a_{0:T}, T)$ örneği, (14) nolu denklemdeki olasılık dağılımlarına göre ardışık olarak örneklenebiliyor.

$$\begin{aligned} T^{(s)} &\sim P(T) \\ x_t^{(s)} &\sim \begin{cases} P(x_0) & t = 0 \\ P(x_t | x_{t-1} = x_{t-1}^{(s)}, a_{t-1} = a_{t-1}^{(s)}) & 0 < t \leq T^{(s)} \end{cases} \\ a_t^{(s)} &\sim P(a_t | x_t = x_t^{(s)}; \pi) \qquad 0 \leq t \leq T^{(s)} \end{aligned} \quad (14)$$

Teklif fonksiyonu seçimimizin sonucu olarak önem örneklemesindeki ağırlık fonksiyonu $W(x)$,

$$W(x_{0:T}, a_{0:T}, T) = \frac{P(x_{0:T}, a_{0:T}, T | r=1; \pi)}{P(x_{0:T}, a_{0:T}, T; \pi)} \quad (15)$$

olarak oluşuyor.

Ayrıca, Bayes teoremini kullanarak örnekleme yapmak istediğimiz (11) nolu denklemdeki sonsal dağılımı, bir olabilirlik ve bir önsel dağılımın çarpımıyla orantılı olarak aşağıdaki şekilde ifade etmemiz mümkün.

$$\begin{aligned} P(x_{0:T}, a_{0:T}, T | r=1; \pi) \propto & P(r=1 | x_{0:T}, a_{0:T}, T; \pi) \\ & \cdot P(x_{0:T}, a_{0:T}, T; \pi) \end{aligned} \quad (16)$$

Buradaki olabilirliği, Şekil 1'deki grafik modelinki MKSlerinin Markov özelliği dolayısıyla,

$$P(r=1 | x_{0:T}, a_{0:T}, T; \pi) = P(r=1 | x_T, a_T) \quad (17)$$

şeklinde basitleştirebiliriz. Böylece, gerekli sadeleştirmeler yapıldıktan sonra ağırlık fonksiyonu,

$$W(x_{0:T}, a_{0:T}, T) = P(r=1 | x_T, a_T) \quad (18)$$

biçimde bulunmuş oluyor.

Böylece, tanımladığımız bu önem örneklemesi metodu ile beklenti-enbüyütme algoritmasının her bir yinelemesinde (12) nolu denklemde tanımlanmış olan $\pi_{i,a}$ ilke parametrelerini, önsel dağılımdan (13), (14) nolu denklemdeki şekilde S adet örnek çektikten sonra aşağıdaki biçimde yaklaşık olarak kestirebiliriz.

$$\pi_{i,a} \approx \frac{\sum_{s=1}^{S} W^{(s)} \sum_{t=0}^{T}[x_t^{(s)} = i \wedge a_t^{(s)} = a]}{\sum_{s=1}^{S} W^{(s)} \sum_{t=0}^{T}[x_t^{(s)} = i]} \quad (19)$$

3.3. Farklı Teklif Fonksiyonları

Tanımladığımız basit önem örneklemesi algoritmasının verimliliğini arttırmak için farklı teklif fonksiyonları denenebilir.

Öncelikle, durum-eylem gezingesinin $x_0, a_0, x_1, a_1, ..., x_T$ elemanlarını önsel dağılımdan çektikten sonra, sadece a_T eylemini ödüle bağlı aşağıdaki sonsal dağılımdan çekmek sezgisel olarak verimi arttıracaktır.

$$a_T \sim P(a_T | x_T, r=1, T) \quad (20)$$

Değişen teklif fonsiyonuna göre de yeni ağırlık fonksiyonu W_1 aşağıdaki biçimde hesaplanabilir.

$$W_1(x_{0:T}, a_{0:T}, T) = \sum_{a_T \in \mathcal{A}} P(r=1 | x_T, a_T) P(a_T | x_T; \pi) \quad (21)$$

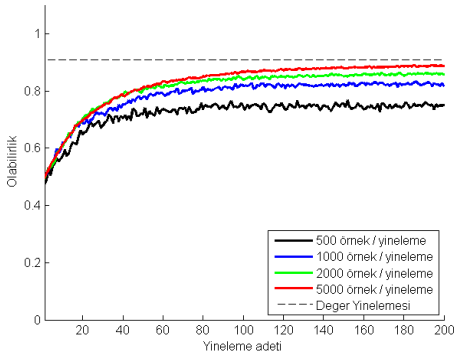
Bundan bir sonraki adım ise, sadece a_T için yaptığımız bu geliştirmeyi sanki her eylem son eylemmiş gibi yorumlayarak diğer eylemlere taşımak olabilir. Böylece, yeni teklif fonksiyonumuzda her a_t eylemi,

$$a_t \sim P(a_t | x_t, r=1, T=t) \quad (22)$$

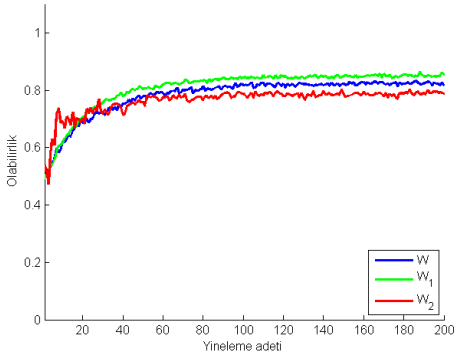
sonsal dağılımından örneklenecektir. Bunun sonucu olarak oluşacak yeni ağırlık fonksiyonumuz W_2 ise,

$$\begin{aligned} W_2(x_{0:T}, a_{0:T}, T) = & \prod_{t=0}^{T-1} \frac{\sum_{a_t \in \mathcal{A}} P(r=1 | x_t, a_T) P(a_t | x_t; \pi)}{P(r=1 | x_t, a_t)} \\ & \cdot \sum_{a_T \in \mathcal{A}} P(r=1 | x_T, a_T) P(a_T | x_T; \pi) \end{aligned} \quad (23)$$

biçiminde hesaplanabilir.



Şekil 2: Farklı Örnek Sayısına göre Önem Örneklemesinin Performansı



Şekil 3: Farklı Teklif Fonksiyonlarının Önem Örneklemesine Katkısı

4. SONUÇLAR

Önerdiğimiz yöntemi 100 durumlu 10 eylemli sentetik bir MKS üzerinde rastgele olarak tanımlanmış bir PÖ problemi için farklı örnek sayıları için uyguladık. Her yinelemedeki ilkelerin olabilirliğini de,

$$P(r=1;\pi) \approx \frac{1}{S}\sum_{s=1}^{N} W^{(s)} \tag{24}$$

biçimde yaklaşık olarak kestirdik. Şekil 2'deki grafikte çıkan sonuçları da, klasik bir modele dayalı PÖ algoritması olan değer yinelemesi yönetiminin sonucu ile karşılaştırdık. Açık olarak, önem örneklemesinin özellikle de örnek sayısı arttıkça bu kısmi en iyi sonuca yakınsadığı görülmektedir.

Ayrıca, önerdiğimiz teklif fonksiyonlarının önem örneklemesinin performansındaki etkisini görmek için, aynı sayıda örnek için farklı teklif fonksiyonları için deneyler yaptık. Şekil 3'teki sonuçlara göre, W_1 ağırlık fonksiyonuna sahip teklif fonksiyonu ile örneklemenin doğrudan önsel dağılımdan örneklemeye göre daha iyi yakınsadığını görülüyor. Bunun yanı sıra, önerdiğimiz diğer farklı teklif fonksiyonunun ise başlarda hızlı bir biçimde yakınsamasına rağmen sonuç olarak örneklemenin performansını düşürdüğünü gördük.

5. VARGILAR

Bu bildiriyi genel olarak PÖ problemini bir olasılıksal çıkarım problemi olarak ele alan ve en iyi ilkenin bulunmasında olasılıksal yaklaşık çıkarım metodlarının kullanımını inceleyecek bir araştırmanın ilk adımı olarak görüyoruz.

Farkındayız ki, önerdiğimiz önem örneklemesi yöntemi durum-eylem gezingelerinin önsel dağılımından örnekleme yaptığından, ödülün seyrek olduğu uzaylarda çok verimsiz olacaktır. Bu yüzden, ileriki çalışmalarda daha verimli bir olasılıksal çıkarım algoritması türetmek için doğrudan sonsal dağılımdan örnekleme yapmayı düşünüyoruz.

Ayrıca, örnekleme yaptığımız olasılık karışımının içiçe yapısından faydalanabilen farklı olasılıksal çıkarım metodları ile örneklemenin verimliliğini arttırmayı hedefliyoruz.

Son olarak, bu hali ile önem örneklemesi görece basit bir örnekleme metodudur. Daha farklı olasılıksal yaklaşık çıkarım metodları ile daha başarılı sonuçlar ve daha hızlı bir yakınsama elde edilebileceğini düşünüyoruz. Bunun için, doğrudan sonsal dağılımdan ardışık Monte Carlo örneklemesi ve tersinir atlama Monte Carlo örneklemesi gibi daha ileri yöntemlerle örnekleme yapmayı planlıyoruz.

6. KAYNAKÇA

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control 3rd Edition, Vol. I*, vol. 2 of *Athena Scientific optimization and computation series*. Athena Scientific, 2007.

[2] C. Szepesvári, *Algorithms for Reinforcement Learning*, vol. 4 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2010.

[3] E. Alpaydin, "Introduction to Machine Learning," *Machine Learning*, vol. 56, no. 2, pp. 387–99, 2004.

[4] P. Dayan and G. E. Hinton, "Using Expectation-Maximization for Reinforcement Learning," *Neural Computation*, vol. 9, pp. 271–278, Feb. 1997.

[5] M. Toussaint and A. Storkey, "Probabilistic inference for solving discrete and continuous state Markov Decision Processes," in *Proceedings of the 23rd international conference on Machine learning*, (New York, New York, USA), pp. 945–952, ACM, 2006.

[6] T. Furmston and D. Barber, "Efficient Inference in Markov Control Problems," in *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pp. 221–229, AUAI Press, 2011.

[7] T. Furmston and D. Barber, "Lagrange dual decomposition for finite horizon Markov decision processes," *Machine Learning and Knowledge Discovery in Databases*, no. 1, pp. 487–502, 2011.

[8] M. Hoffman and A. Jasra, "Trans-dimensional MCMC for Bayesian Policy Learning," *Neural Information Processing Systems*, vol. 20, pp. 1–8, 2008.

[9] M. Hoffman, H. Kueck, N. D. Freitas, and A. Doucet, "New inference strategies for solving Markov Decision Processes using reversible jump MCMC," in *Conference on Uncertainty in Artificial Intelligence*, 2009.