2022 30th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-5092-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIU55565.2022.9864957

Politikaya Dayalı Sürekli Kontrolde İçsel Motivasyona Dayalı Yapay Hedef Oluşturma
An Intrinsic Motivation Based Artificial Goal Generation in On-Policy Continuous Control

Baturay Saglam[1], Furkan B. Mutlu[1], Kaan Gonc[2], Onat Dalmaz[1] ve Suleyman S. Kozat[1]
[1]Elektrik ve Elektronik Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara, Türkiye
{baturay, burak.mutlu, onat, kozat}@ee.bilkent.edu.tr
[2]Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara, Türkiye
{kaan.gonc}@bilkent.edu.tr

***Özetçe* —Bu çalışma, politikaya dayalı sürekli kontrolde yönlendirilmiş bir keşif stratejisi oluşturmak için hayvan motivasyon sistemleri hakkındaki mevcut teorileri pekiştirmeli öğrenme (RL) paradigmasına uyarlamaktadır. Ajanları faydalı durum alanlarını ziyaret etmeye teşvik eden yeni ve ölçeklenebilir bir yapay bonus ödül kuralı sunulmaktadır. Pekiştirmeli öğrenme paradigmasındaki içsel teşvikleri, tanıtılan deterministik ödül kuralı altında birleştirerek değer işlevini, görülmeyen veya daha az bilinen durum değerlerini öğrenmeye ve çevreyi yeterince öğrenmeden önce erken davranışı önlemeye zorlamaktadır. Simülasyon sonuçları, önerilen algoritmanın literatürdeki en iyi sonuçları veren politikaya dayalı yöntemleri önemli ölçüde geliştirdiğini ve içsel entropi tabanlı keşfi iyileştirdiğini göstermektedir.**

Anahtar Kelimeler—derin pekiştirmeli öğrenme, keşif, içsel motivasyon, sürekli kontrol, politikaya dayalı öğrenme

***Abstract*—This work adapts the existing theories on animal motivational systems into the reinforcement learning (RL) paradigm to constitute a directed exploration strategy in on-policy continuous control. We introduce a novel and scalable artificial bonus reward rule that encourages agents to visit useful state spaces. By unifying the intrinsic incentives in the reinforcement learning paradigm under the introduced deterministic reward rule, our method forces the value function to learn the values of unseen or less-known states and prevent premature behavior before sufficiently learning the environment. The simulation results show that the proposed algorithm considerably improves the state-of-the-art on-policy methods and improves the inherent entropy-based exploration.**

Keywords—deep reinforcement learning, exploration, intrinsic motivation, continuous control, on-policy learning

I. Giriş

Pekiştirmeli öğrenme alanı, sürekli sistemlerin kontrol edilmesi gibi etkileyici gelişmeler nedeniyle son zamanlarda büyük ilgi görmüştür [1]–[4]. Bununla birlikte, pek çok araştırmacının odaklandığı, pekiştirmeli öğrenmenin derin ortamında birkaç sorun vardır. Keşif-sömürme ikilemi, modern pekiştirmeli öğrenme ve çok kollu haydut problemlerinde zorlu ve uzun süredir devam eden bir süreçtir [5]. Keşfin arkasındaki ana motivasyon, olası eylemler dizisi arasından en uygun eylemi keşfetmek iken sömürü, zaman veya hesaplama gücü gibi kaynakları harcamamak için her zaman en iyi bilinen eylemi seçmeyi sağlamaktadır [6]. Bu nedenle, bu iki aşırı uç arasındaki optimal bir denge, etkili stratejiler aracılığıyla verimli bir şekilde bulunmalıdır. Bununla birlikte, pekiştirmeli öğrenme ortamlarının Markov Karar Süreci (MKS) tarafından temsil edildiği kabul edildiğinden, çevrenin altında yatan ödül işlevinden etkili bir keşif stratejisi çıkarmak imkansız hale gelmektedir [6].

İçsel hayvan motivasyonlarını pekiştirici öğrenmeye uyarlamak, keşif-sömürme ikilemine uygun çözümler sunmaktadır, ancak bunlar genellikle ayrık eylem alanları için çalışılmıştır [7]. Öte yandan, Q-öğrenme ya da zamansal fark öğrenmesi [8] tabanlı yöntemler, sürekli eylem uzaylarında politika dışı öğrenme için rastgele eylem pertürbasyonlarına etkili bir alternatiftir [9]. Bununla birlikte, politikaya dayalı algoritmalar Q-öğrenme bazlı zamansal fark öğrenmesi [8] kullanmadığından ve politikalar zaten doğal olarak eylemlerin entropisini en üst düzeye çıkarmayı hedeflediğinden, derin pertürbasyon ağları bazlı modeller mevcut tek keşif seçenekleridir [7].

Bu çalışma, sürekli eylem alanlarında politikaya dayalı kontrol için tek bir yapay ödül bonusu altında mevcut pekiştirmeli öğrenme literatüründeki içsel motivasyonları birleştirmektedir. Tanıtılan deterministik hedef oluşturma yaklaşımı, ajanları işlevsel durum alanlarını gözlemlemeye yönlendirmek ve bilgilendirici bir yapay ödül oluşturmak için politikaya dayalı yöntemlerde kullanılan sunum arabelleklerini ve derin değer fonksiyonlarını kullanmaktadır. Önerilen yaklaşım, $\mathcal{O}(n)$ gibi kısa ve basit bir çalışma zamanı ve kural tabanlı olmasına karşın, deneysel çalışmalarımız, bir dizi zorlu MuJoCo [10] sürekli kontrol görevinde Proksimal Politika Optimizasyonu (PPO) algoritmasının [11] performansını büyük ölçüde geliştirdiğini göstermektedir.

II. İlgili Çalışmalar

Sürekli eylem uzaylarında keşif, genellikle öğrenilebilir ve rastgele eylem pertürbasyonları altında incelenmektedir [12]. Rastgele pertürbasyonlar için iyi bilinen örneklerden biri sıfır

978-1-6654-5092-8/22/$31.00 ©2022 IEEE

ortalamalı Gauss [13] ya da Ornstein-Uhlenbeck gürültüsüdür [14]. Öğrenilebilir bir keşif için, rastgele bir derin pertürbasyon ağının, ajan tarafından seçilen eylemleri ya da derin değer ve politika ağlarının parametrelerini bozduğu yaklaşımlar önerilmiştir [15], [16]. Bununla birlikte, aşırı bir hesaplama karmaşıklığı getirdikleri ve pertürbasyon ağlarını eğitmek için önemli miktarda zaman gerektirdiği gösterilmiştir [17]. İçsel motivasyon, sayıma dayalı yaklaşımlar üzerine kurulu ayrık eylem alanları için de kapsamlı bir şekilde incelenmiştir [18]. Bu tür yöntemler ziyaret edilen durumları, ödülleri ve seçilen eylemleri sayabilmektedir.

III. Teknik Arka Plan

A. Derin Pekiştirmeli Öğrenme

Bu çalışma, bir ortamla etkileşime giren bir ajandan oluşan standart pekiştirmeli öğrenme çerçevesini dikkate almaktadır. Açıklamayı basitleştirmek için ortamın tamamen gözlemlenebilir olduğu varsayılmaktadır. Bir ortam, bir MKS olarak modellenmiştir ve bir dizi durum $\mathcal{S}$, bir dizi eylem $\mathcal{A}$, başlangıç durumları $p(s_0)$ üzerinde bir dağılım, bir ödül işlevi $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, geçiş olasılıkları $p(s_{t+1}|s_t, a_t)$, bir zaman ufku T ve bir indirim faktörü $\gamma \in [0, 1)$ tarafından tanımlanmaktadır. Sürekli eylem uzaylarında bir deterministik ya da stokastik politika $\pi_\theta : \mathcal{S} \to \mathcal{A}$, θ tarafından parametrelenen bir derin ağ tarafından temsil edilmektedir. Ajanın amacı, beklenen indirimli getiriyi en yüksek seviyeye çıkarmaktır:

$$\eta(\pi_\theta) = \mathbb{E}_\tau[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)] \tag{1}$$

burada $\tau = (s_0, a_0, \ldots, s_T)$, $s_0 \sim p(s_0)$, $a_t \sim \pi_\theta(a_t|s_t)$ ve $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ ile birlikte bir yörüngeyi belirtmektedir. Deneysel değerlendirme, indirgenmemiş $\mathbb{E}_\tau[\sum_{t=0}^{T} r(s_t, a_t)]$ getirisine dayanmaktadır.

B. Politikaya Dayalı Yöntemler

Politika dışı algoritmaların aksine, politikaya dayalı yöntemler, geçerli olarak izlenen politikaya göre derin ağlar tarafından temsil edilen değer ve politika fonksiyonlarının güncellenmesini gerektirmektedir. Bu çalışmada iyi bilinen ve literatürde en iyi performansı gösteren politikaya dayalı PPO [11] algoritması ele alınmaktadır. Tanıtılan yöntem herhangi bir politikaya dayalı yönteme kolayca uygulanabilmektedir.

PPO [11], politika dağılımında küçük bir değişiklik sağlayan bir yükselme yönü hesaplayarak REINFORCE'u [13] geliştirmektedir. Daha spesifik olarak, politikayı optimize etmek için PPO [11] aşağıdaki kısıtlı optimizasyon problemini çözmektedir:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \frac{1}{T} \sum_{t=0}^{T} f(c, \epsilon, A^{\pi_{\theta_k}}(s_t, a_t)); \tag{2}$$

$$f(c, \epsilon, A^{\pi_{\theta_k}}) = \min\left(cA^{\pi_{\theta_k}}, g(\epsilon, A^{\pi_{\theta_k}})\right), \tag{3}$$

$$c = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(s_t, a_t)}; \tag{4}$$

$$g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) = \operatorname{clip}(-\epsilon, \epsilon, A^{\pi_{\theta_k}}(s_t, a_t)), \tag{5}$$

burada $\operatorname{clip}(-\epsilon, \epsilon, \cdot)$ verilen girdiyi $[-\epsilon, \epsilon]$ aralığında olacak şekilde kırpmaktadır ve $A^{\pi_{\theta_k}}(s_t, a_t)$ genellikle Genel Avantaj Tahmini [19] yöntemiyle tahmin edilen avantaj fonksiyonudur. Avantaj fonksiyonu, bir durumdaki bir eylemin beklenen getirisinden çıkarılan bir durumun (değerin) beklenen getirisidir. Daha sezgisel olarak, alınan eylemin toplam beklenen getiriye kıyasla ne kadar iyi veya kötü olduğunu temsil etmektedir. Son olarak, politika tarafından seçilen eylemleri değerlerlendiren ve bir durumun olabilecek bütün eylemler üzerindeki beklenen toplam getirisini tahmin eden değer fonksiyonu V_ϕ, şu şekilde güncellenmektedir:

$$\phi_{t+1} = \underset{\phi}{\operatorname{argmin}} \frac{1}{T} \sum_{t=0}^{T} (V_\phi(s) - \hat{R}_t)^2, \tag{6}$$

burada $\hat{R}_t$, avantaj taminleri kullanılarak hesaplanmış ve modifiye edilmiş ödüllerin toplamıdır.

Genel politikaya dayalı öğrenme yaklaşımında, ajanlar, (s_t, a_t, r_t, s'_t) ile belirtilen geçişleri T ufkuna kadar toplamaktadır ve geçişleri herhangi bir ağ güncellemesi olmadan kullanıma sunma arabelleğinde depolamaktadır. Her güncelleme adımında, ajanlar, mini toplu öğrenme yoluyla kullanıma sunma arabelleği içindeki geçişlerden öğrenmektedir. Her güncelleme adımından sonra, kullanıma sunma arabelleği temizlenmektedir.

C. İçsel Motivasyonlu Keşif

Özünde ödüllendirici olayı kurtarmada yetenekleri hızlı bir şekilde geliştirmek için ajanlar, içsel olarak motive edilmiş keşif konusundaki davranışlarını ayarlamak zorundadırlar [7]. Bunun bir sonucu, aktivite tekrarlandığında, içsel ödülün kademeli olarak azalması, ajanların sonunda "sıkılması" ve yeni bir alternatif inşa etmeye ve öğrenmeye devam etmesidir [20]. Mevcut pekiştirmeli öğrenme literatüründe, içsel motivasyon üç farklı açıdan kullanılmıştır: durum yeniliği, bilgi kazanımı ve tahmin hatası [7]. İlk teşvik, ajanların genellikle gitmediği durumlara yönlendirmektir. Böyle bir yenilik farklı yöntemlerle ölçülebilmektedir [7]. Bilgi kazanımı, ajanların en fazla bilgiyi elde ettikleri durumları, örneğin politikalarındaki güncelleme miktarını en yüksek seviyeye çıkaracak durumlar, ziyaret etmelerini sağlamaktadır [7]. Son olarak, tahmin hatasına dayalı teşvikler, ajanları dürtüsel davranışlara uyum sağlamamak için hata yapmaya zorlamaktadır [7], [7]. Bu çalışma, bu teşvikleri kural tabanlı bir şekilde politikaya dayalı öğrenmede birleştirmektedir.

IV. Yöntem

Bahsedilen içsel motivasyonları birleştiren yapay hedef oluşturmayı inşa etmek için öncelikle durum yenilik motivasyonu dikkate alınmaktadır. Tanım gereği, ajanlar genellikle gitmedikleri durumları ziyaret etmeye teşvik edilmektedir. Her ufkun tamamlanmasından sonra, her geçişe bir bonus ödül eklenmektedir. Her bir geçiş için bu bonus ödül, bir geçiş durumunun geri kalan geçişlerin durumlarına olan ortalama mutlak mesafesi ile hesaplanmaktadır:

$$r'_{1,i} = \frac{1}{T-1} \sum_{\substack{t=0 \\ i \neq t}}^{T} |s_t - s_i|. \tag{7}$$

Bilgi kazanımında, bir durumun ziyareti ile elde edilen bilgiler maksimize edilmektedir. Daha önce açıklandığı gibi,

politika, gözlemlenen durumlar üzerinde optimal eylemleri seçmek için eğitilmektedir ve değer fonksiyonu, politika tarafından seçilen eylemleri eleştirmeyi öğrenmektedir. Bu nedenle elde edilen bilgiler politika yerine değer fonksiyonu üzerinden ölçülebilir [6]. Spesifik olarak, belirli bir durum üzerindeki değer fonksiyonunun gradyanının büyük olması, böyle bir durumun öğrenme için daha faydalı olabileceği ve önemli bir bilgi kazancına işaret ettiği anlamına gelmektedir. Bu nedenle, her bir geçiş için bilgi kazanımı içsel ödülü durum fonksiyonunun gradyanı olarak doğrudan kullanılabilir:

$$\begin{aligned} r'_{2,i} &= \phi_{t+1,i} - \phi_{t,i}, \\ &= \sum \underset{\phi}{\operatorname{argmin}} (V_\phi(s_i) - \hat{R}_t)^2 - \phi_{t,i}, \end{aligned} \tag{8}$$

burada gradyan, derin değer fonksiyonun bütün parametreleri üzerinden toplamı alınarak hesaplanmıştır. Daha önce de belirtildiği gibi sunum arabelleğini doldurmak için gerçekleştirilen keşif adımlarında ağ güncellemesi yapılmasa da, bu içsel ödül ağ güncellemesi simüle edilerek hesaplanabilir.

Son olarak, bir politikaya dayalı yöntemin tahmin hatası, değer fonksiyonunun tahmin hatasıdır. Bu nedenle, tahmin hatasına dayalı teşvik için ödül bonusu, her geçiş için mutlak hata değeri olarak hesaplanmaktadır:

$$r'_{3,i} = |V_\phi(s_i) - \hat{R}_t|. \tag{9}$$

Bununla birlikte, politikaya dayalı yöntemler için genel yapay bonus ödülü şu şekilde oluşturulmaktadır:

$$r'_i = \frac{r_i}{r_{\max}}(r_{1,i} + r_{2,i} + r_{3,i}), \tag{10}$$

burada bonus ödül r'_i, i tarafından indekslenen her bir geçiş için ayrı olarak hesaplanmaktadır. Ayrıca, ödül hesaplamalarındaki kararsızlığı ve sınırsızlığı önlemek için elde edilen yapay ödül, o geçişteki gerçek ödülün ortamdaki en yüksek ödüle olan oranıyla ölçeklendirilmektedir. Bu sayede, eğer ajan optimal ödüle yaklaşırsa çok daha büyük bir ödül elde etmektedir. Sonuç olarak, hesaplanan içsel motivasyon bazlı ödül, ajanın kazandığı gerçek ödüle her keşif adımında eklenip sonrasında derin aktör ve değer ağlarını eğitmek için kullanılmaktadır.

Bilgi kazancı ödül hesaplamasında kullanılan tek değer çıkışı ve tek örnek nedeniyle, eğer durum uzayının boyutu n ise, her geçiş için r'_i'in $\mathcal{O}(n)$ cinsinden hesaplandığı görülmektedir. Benzer şekilde, kalan hesaplamalar da $\mathcal{O}(n)$ cinsindendir. Bu nedenle, ödül hesaplamalarının tek bir geçiş için $\mathcal{O}(3n)$ ve tek bir ufuk için $\mathcal{O}(3nT)$ içerisinde çalıştığı gözlemlenmektedir. T sabit ve eğitim zaman adımlarına kıyasla nispeten küçük olduğundan önerilen algoritma basitleştirilmiş bir versiyonda $\mathcal{O}(3nT) \approx \mathcal{O}(n)$'de çalışmaktadır.

V. Deneyler

A. Deney Protokolü

Tanıtılan yöntemin etkinliği, MuJoCo [10] fizik simülatöründeki zorlu sürekli kontrol ortamlarında test edilmiştir. Daha önce bahsedildiği gibi, önerilen yöntem, iyi bilinen bir politikaya dayalı sürekli kontrol algoritması olan PPO'ya [11] uygulanmıştır. A2C [21] gibi diğer politikaya dayalı algoritmalar, MuJoCo [10] görevlerinde başarısız oldukları bilindiği için dikkate alınmamıştır. Önerilen algoritma, keşfin iki strateji aracılığıyla gerçekleştirildiği temel PPO [11] ile karşılaştırılmıştır:

Tablo I: On rastgele tohum üzerinden son on değerlendirme ödülünün ortalaması. Her ortam için maksimum değer kalın yazılmıştır. ±, denemeler üzerindeki yarım standart sapmayı belirtmektedir.

Ortam	Algoritma	
	PPO	Tanıtılan Algoritma
Ant-v2	2601.91 ± 408.75	**3135.85 ± 527.43**
HalfCheetah-v2	1857.63 ± 265.51	**4774.32 ± 747.85**
Hopper-v2	3448.15 ± 681.41	**3651.41 ± 587.61**
Humanoid-v2	667.13 ± 55.21	**3502.48 ± 607.17**
InvertedDoubPendulum-v2	**7837.56 ± 1285.36**	7721.15 ± 1141.01
InvertedPendulum-v2	999.96 ± 68.21	**999.97 ± 82.96**
Swimmer-v2	103.52 ± 112.44 ± 14.37	**130.05 ± 14.26**
Walker2d-v2	3800.76 ± 762.81	**4879.8 ± 922.28**

eylemlerin örneklendiği bir Gauss politikası ve ek bir entropi maksimizasyon hedefi. Bir olasılık dağılımından örnekleme yoluyla rastgelelik enjekte edilirken, entropi maksimizasyonu hedefi, örneklenen eylemlerin çeşitli olmasını sağlamaktadır.

Her algoritma, tüm ortamlar için 1 milyon zaman adımı için çalıştırılmıştır. Her yöntemin değerlendirilmesi, ajanların performansının her 1000 zaman adımında ortalama 5 rastgele tohum için ayrı bir değerlendirme ortamında kaydedilmesiyle gerçekleştirilmiştir. Değerlendirmelerde güncelleme ve keşif yapılmamıştır. Ödül fonksiyonları, geçiş olasılıkları gibi ortam dinamikleri, adil bir değerlendirme prosedürü için önceden işlenmemiştir veya değiştirilmemiştir.

PPO algoritması [11], birçok araştırmacı tarafından kullanılan ve iyi bilinen bir GitHub deposundan alınan kod aracılığıyla uygulanmıştır[1]. Depo, PPO [11] için ince ayarlanmış üst değişkenleri içermektedir. Yapay ödül bonusu, bu depodaki kodun üzerine inşa edilelerek uygulanmıştır.

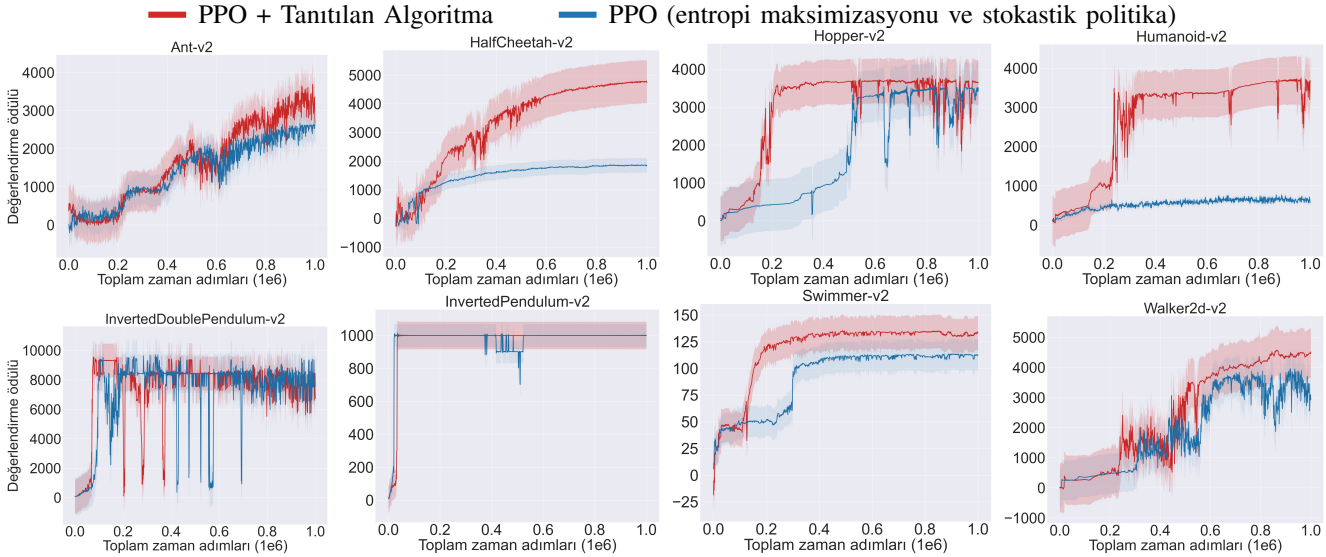
B. Sonuçlar ve Değerlendirmeler

Sonuçlar Tablo I ve Şekil 1'de verilmiştir. Tablo I ajanların elde ettiği ortalama son on değerlendirme ödülünü sunarken, Şekil 1 öğrenme eğrilerini göstermektedir. Önerilen yöntemin ziyaret edilen durum uzaylarını çeşitlendirerek yakınsama oranı ve en yüksek değerlendirme getirileri açısından PPO'nun [11] performansını önemli ölçüde iyileştirdiği görülmektedir. PPO [11] ve tanıtılan yöntemin nihai getirileri, InvertedPendulum ve Swimmer ortamlarında olduğu gibi aynı olsa bile, algoritmamız daha hızlı bir yakınsama oranına sahiptir. Bu bulgudan, yöntemimizin aynı zaman aralığında daha çeşitli durum uzaylarını ziyaret ettiği sonucuna varılmıştır. Bu aynı zamanda ortamların geri kalanında da gösterilmiştir. Tanıtılan algoritma tarafından sağlanan çeşitli durumlara olan ziyaret, bir Gauss politikası ve entropi maksimizasyonunun ajanların bilgilendirici durum uzaylarına yönlendirilmesi için yeterli olmayabileceğinden, temel algoritma üzerinde daha fazla iyileştirmeye izin vermektedir. Bu nedenle, bu ortamlarda, ajan, içsel teşvikler tarafından motive edilmesi gibi kural tabanlı bir keşfe ihtiyaç duymaktadır.

VI. Vargılar

Bu çalışma, sürekli eylem alanlarında politikaya dayalı öğrenmeye alternatif bir kural tabanlı keşif yaklaşımı sunmaktadır. Önerilen bonus ödül kuralı, mevcut pekiştirmeli öğrenme literatüründeki bilgi kazanımı, durum yeniliği ve tahmin hatası içsel motivasyonlarını tek bir katkı ödül işlevi altında

[1] https://github.com/ikostrikov/pytorch-a2c-ppo-acktr-gail



Şekil 1: Tanıtılan algoritmanın PPO üzerindeki performans sonuçları ve temel PPO ile MuJoCo sürekli kontrol ortamlarında karşılaştırılması. Gölgeli bölge 5 rastgele tohum üzerindeki sonuçların standart sapmasını temsil etmektedir.

birleştirmektedir. Kapsamlı bir deney seti ve bir hesaplama karmaşıklığı analizi, yöntemimizin yalnızca $\mathcal{O}(n)$'de çalışarak PPO algoritmasındaki Gauss politikası ve entropi maksimizasyonuna dayalı keşiften önemli ölçüde daha iyi performans sergilediğini göstermektedir.

KAYNAKLAR

[1] B. Saglam, E. Duran, D. C. Cicek, F. B. Mutlu, and S. S. Kozat, "Estimation error correction in deep reinforcement learning for deterministic actor-critic methods," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 137–144.

[2] B. Saglam, E. Duran, D. Cicek, F. Mutlu, and S. Kozat, "Parameter-free deterministic reduction of the estimation bias in continuous control," 2021. [Online]. Available: https://arxiv.org/abs/2109.11788

[3] D. C. Cicek, E. Duran, B. Saglam, K. Kaya, F. Mutlu, and S. S. Kozat, "Awd3: Dynamic reduction of the estimation bias," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 775–779.

[4] D. C. Cicek, E. Duran, B. Saglam, F. B. Mutlu, and S. S. Kozat, "Off-policy correction for deep deterministic policy gradient algorithms via batch prioritized experience replay," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 1255–1262.

[5] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine Learning*, vol. 49, no. 2, pp. 209–232, Nov 2002. [Online]. Available: https://doi.org/10.1023/A:1017984413808

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[7] A. G. Barto, *Intrinsic Motivation and Reinforcement Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 17–47. [Online]. Available: https://doi.org/10.1007/978-3-642-32375-1_2

[8] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, Aug 1988. [Online]. Available: https://doi.org/10.1007/BF00115009

[9] A. Barto and O. Imek, "Intrinsic motivation for reinforcement learning systems," *Intrinsically Motivated Learning in Natural and Artificial Systems*, vol. 4, 01 2005.

[10] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[12] S. Thrun, "The role of exploration in learning control," in *Handbook for Intelligent Control: Neural, Fuzzy and Adaptive Approaches*. Florence, Kentucky: Van Nostrand Reinhold, June 1992.

[13] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229–256, May 1992. [Online]. Available: https://doi.org/10.1007/BF00992696

[14] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Phys. Rev.*, vol. 36, pp. 823–841, Sep 1930. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.36.823

[15] Y. Zhang and H. Van Hoof, "Deep coherent exploration for continuous control," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 12 567–12 577. [Online]. Available: https://proceedings.mlr.press/v139/zhang21t.html

[16] M. Fortunato, M. G. Azar, B. Piot, J. Menick, M. Hessel, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg, "Noisy networks for exploration," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rywHCPkAW

[17] S. Amin, M. Gomrokchi, H. Satija, H. van Hoof, and D. Precup, "A survey of exploration methods in reinforcement learning," 2021.

[18] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, "Count-based exploration with neural density models," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 2721–2730.

[19] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2018.

[20] F. Kaplan and P.-Y. Oudeyer, "Intrinsically motivated machines," in *50 Years of AI*, Lungarella, M. Iida, F. Bongard, J. Pfeifer, and R., Eds. Lungarella, M.; Iida, F.; Bongard, J.; Pfeifer, R., 2008, p. n/a. [Online]. Available: https://hal.inria.fr/inria-00420223

[21] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1928–1937. [Online]. Available: https://proceedings.mlr.press/v48/mniha16.html