

# Sürekli Eylem Alanlarında Politika-Dışı Öğrenme için Birleştirilmiş İçsel Motivasyonlu Keşif

## Unified Intrinsically Motivated Exploration for Off-Policy Learning in Continuous Action Spaces

Baturay Saglam, Furkan B. Mutlu, Onat Dalmaz ve Suleyman S. Kozat  
Elektrik ve Elektronik Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara, Türkiye  
{baturay, burak.mutlu, onat, kozat}@ee.bilkent.edu.tr

**Özetçe** —Keşif, rastgele gürültünün ağ parametrelerini veya seçilen eylemleri bozduğu, yönlendirilmemiş yöntemler kullanılarak sürekli kontrolde sürdürülmektedir. İçsel olarak yönlendirilen keşif, yönlendirilmemiş tekniklere iyi bir alternatiftir ancak yalnızca ayrık eylem alanları için incelenmiştir. Mevcut pekiştirmeli öğrenme literatüründeki içsel teşvikler, bu çalışmada politika-dışı öğrenme için deterministik bir yapay hedef oluşturma kuralıyla birleştirilmiştir. Ajan, kendisini yararlı durum uzaylarına götüren eylemleri seçerse, bu uygulama aracılığıyla ek bir ödül kazanmaktadır. Kapsamlı bir deney seti, tanıtılan yapay ödül kuralının, politika-dışı temel algoritmaların performansını önemli ölçüde geliştirdiğini göstermektedir.

**Anahtar Kelimeler**—*derin pekiştirmeli öğrenme, keşif, içsel motivasyon, sürekli kontrol, politika-dışı öğrenme*

**Abstract**—Exploration is maintained in continuous control using undirected methods, in which random noise perturbs the network parameters or selected actions. Exploration that is intrinsically driven is a good alternative to undirected techniques. However, it is only studied for discrete action domains. The intrinsic incentives in the existing reinforcement learning literature are unified together in this study by a deterministic artificial goal generation rule for off-policy learning. The agent gains additional reward through this practice if it chooses actions that lead it to useful state spaces. An extensive set of experiments indicates that the introduced artificial reward rule significantly improves the performance of the off-policy baseline algorithms.

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

### I. GİRİŞ

Son yıllarda, pekiştirmeli öğrenmede derin yaklaşımların kullanılması, geniş bir uygulama yelpazesinde sayısız başarı elde etmiştir [1]. Ancak, derin sinirsel ağların fonksiyon tahminlerinde kullanılması birçok sorunu da beraberinde getirmektedir [2], [3]. Keşif ve sömürü arasındaki denge, modern derin pekiştirmeli öğrenmede süregelen bu zorluklardan biridir [4]. Keşfin ana amacı, ajanların, erken bir davranışı adapte etmemek için bir dizi farklı deneyim toplamasını sağlamaktır [5]. Keşif yetersizse, yüksek ödüller veren eylem kararları gözden kaçabilir ve bir ajanın politikası yerel bir optimuma

yakınsayabilir [5]. Buna karşılık, aşırı keşifle, ajanlar çok fazla zaman harcayabilir ve toplanan deneyimleri verimli bir şekilde kullanmadan birçok optimal olmayan eylemi denemek için kaynakları boşa harcayabilir [5]. Bu nedenle, *keşif-sömürü ikilemi* olarak da bilinen keşif ve sömürü ödünleşimi için optimal bir denge, etkili ve verimli stratejiler kullanılarak bulunmalıdır. Bununla birlikte, etkili ve verimli bir keşif yapısı tasarlamak önemsiz değildir, çünkü keşif stratejileri temeldeki Markov Karar Sürecinin (MKS) ödül fonksiyonundan çıkarılmamaktadır ve yüksek boyutlu durum ve eylem alanları, keşif için gerekli zamanı ve kaynakları arttırmaktadır [6].

Pekiştirmeli öğrenmede keşif, genellikle yönlendirilmiş ve yönlendirilmemiş yöntemler altında incelenmektedir [6]. Yüksek boyutlu sürekli eylem uzaylarındaki yönlendirilmemiş keşif teknikleri, ek Gauss gürültüsü [7] veya derin parametrelili uzay gürültüsü gibi seçilen eylemleri veya parametrelili politikaların ağırlıklarını bozmayı amaçlamaktadır. Zor keşif görevlerinde yönlendirilmiş keşif stratejilerine bir alternatif olarak, içsel motivasyon üç açıdan etkin bir şekilde kullanılmıştır: tahmin hatası, durum yeniliği ve bilgi kazanımı [8]. İlk motivasyon, ajanları tahminin zor olduğu durum uzaylarına yönlendirmektedir. İkinci sezgi, ajan genellikle gitmediği bir duruma girdiğinde içsel bir bonus eklemektir ve üçüncüsü, ortam dinamikleri üzerindeki belirsizliğin azalmasına bağlı olarak kazanılan içsel bir ödüldür [8].

Bu bildiri, sürekli eylem uzaylarına sahip ortamların kontrolünde yönlendirilmiş bir keşif stratejisi elde etmek için içsel motivasyonun politika-dışı yöntemlerle nasıl etkili bir şekilde birleştirilebileceğini araştırmaktadır. Tahmin hatası, durum yeniliği ve bilgi kazancı motivasyonlarını birleştiren deterministik bir keşif kuralı bu çalışma tarafından tanıtılmaktadır. Önerilen keşif kuralı, ajanların zamansal fark hatası (temporal difference error) ve deneyimlerinden yararlanarak, ajan için bir bonus ödülünü yapay olarak hesaplamaktadır. OpenAI Gym [9] sürekli kontrol görevleri üzerinde yapılan ampirik çalışmalar, tanıtılan içsel olarak motive edilmiş keşif kuralının, politika-dışı temel algoritmaların performansını ajanları fonksiyonel durum uzaylarına yönlendirmeye teşvik ederek kararlılık, öğrenme hızı ve en yüksek değerlendirme getirileri açısından geliştirdiğini göstermektedir. Bu bildirinin katkıları şu şekildedir:

- Önerilen yöntem, literatürdeki üç içsel motivasyonu

birleştiren sürekli kontrol için ilk keşif yöntemidir.

- $O(n)$ 'de çalışmasından dolayı, tanıtılan deterministik yöntem ağır hesaplama karmaşıklığı getirmemektedir.
- Önerilen yöntem, sürekli kontrolde yaygın olarak bilinen ve yönlendirilmemiş stratejilerden olan Ornstein-Uhlenbeck (OU) gürültü sürecini [10] ve ek Gauss gürültüsünü [7] zorlu OpenAI Gym [9] kontrol görevlerinde önemli bir farkla geride bırakmaktadır.

## II. TEKNİK ARKA PLAN

Pekiştirmeli öğrenme, sıralı bir karar verme görevini çözmek için çevresiyle etkileşime giren bir ajanı dikkate almaktadır. Her  $t$  zaman adımında, ajan bir  $s \in \mathcal{S}$  durumunu gözlemlemektedir ve  $a \in \mathcal{A}$  eylemini seçmektedir; burada  $\mathcal{S}$  ve  $\mathcal{A}$  sırasıyla durum ve eylem uzaylarıdır. Eylem kararına bağlı olarak, ajan bir ödül fonksiyonundan  $r$  ödülü almaktadır ve bir sonraki  $s' \in \mathcal{S}$  durumunu gözlemlemektedir. Tamamen gözlemlenebilir ortamlarda, pekiştirmeli öğrenme problemi  $(\mathcal{S}, \mathcal{A}, p_M, \gamma)$  çokuzlusu tarafından temsil edilen bir sonlu MKS olarak varsayılmaktadır; burada  $p_M$  geçiş olasılığıdır,  $s', r \sim p_M(s, a)$ , ve  $\gamma \in [0, 1)$  sabit indirim faktörüdür.

Derin pekiştirmeli öğrenmedeki amaç, beklenen getiri toplamı  $J(\phi) = \mathbb{E}_{s_i \sim p_{\pi}, a_i \sim \pi} [R_0]$ 'yi maksimize eden ve  $\phi$  ile parametrelendirilen  $\pi_{\phi} : \mathcal{S} \rightarrow \mathcal{A}$  optimal politikasını bulmaktır. Sürekli eylem uzaylarında, parametreleştirilmiş politikalar, beklenen getiri  $\nabla_{\phi} J(\phi)$ 'nin gradyanı ile optimize edilmektedir. Eleştirilen veya politika  $\pi$  altında  $s$  durumunda  $a$  eylemini gerçekleştirirken beklenen getiriyi tahmin etmektedir:

$$Q^{\pi}(s, a) = \mathbb{E}_{s_i \sim \pi, a_i \sim \pi} [R_t | s, a]. \quad (1)$$

Derin pekiştirmeli öğrenmede, Q-ağları olarak da bilinen eylem-değer fonksiyonları,  $\theta$  ile parametrelendirilen  $Q_{\theta}(s, a)$  türevlenebilir fonksiyon tahmin edicileri tarafından modellenmektedir. Derin eylem-değer fonksiyonu, eylem-değer işlevini öğrenmek için kullanılan ve temel bir ilişkiyi temsil eden Bellman denklemi'ne dayalı bir güncelleme kuralı olan geçici fark öğrenme [11] ile elde edilmektedir. Mevcut durum-eylem çifti  $(s, a)$  tahmininden sonraki durum-eylem çifti  $(s', a')$ 'ya yapılan önyükleme şu şekilde ifade edilmektedir:

$$Q_{\theta}(s, a) = r + \gamma Q_{\theta'}(s', a'); \quad a' \sim \pi_{\phi'}(s'). \quad (2)$$

Burada,  $\theta'$  ve  $\phi'$  yüksek frekanslı güncellemeler üzerinde kararlılığı ve sabit hedefi korumak için kullanılan ikincil donmuş ağların parametreleridir:  $Q_{\theta'}(s, a)$  ve  $\pi_{\phi'}(s)$ . Bu güncelleme kuralları, ajanların deneyimlerinin deneyim yeniden yürütme arabelleğinde [12] depolandığı, politika-dışı öğrenme için geçerlidir. Politika-dışı öğrenmede ajanlar, verileri ve örnekleme verimliliğini artırmak ve derin ağlar tarafından temsil edilen politikaları üzerinde gradyan adımları gerçekleştirmek için örneklenen deneyimlerini birden çok kez yeniden kullanmaktadırlar [13].

## III. YÖNTEM

*Motivasyon*, uyarılma, güç ve eylem yönünü etkileyen süreçleri tanımlayan bir terimdir [8]. Motive olmak, bir eylemi yapmaya mecbur hissetmektir [14]. Psikologlar motivasyonu iki sınıfa ayırmaktadırlar, *dışsal motivasyon*, dışarıdan sağlanan ödüller nedeniyle harekete geçmek ve *içsel motivasyon*,

bir şeyi doğası gereği zevkli veya ilginç olduğu için kendi iyiliği için yapmaktır [14]. İçsel motivasyon, harici olarak sağlanan ödüllerin yokluğunda veya yetersizliğinde bilincin oyun, yetkinlik, keşif ve diğer merak odaklı davranışlarda bulunmasına neden olmaktadır [15]. Bu çalışma, ajanların eylemleri tekrar etmesini önlemek ve optimal eylemleri keşfetmek için araştırmayı teşvik eden birleşik bir içsel motivasyon yaklaşımı sunmaktadır. Pekiştirmeli öğrenme literatüründeki keşif için kullanılan içsel motivasyonlar şu şekilde tanımlanmaktadır:

- **Bilgi kazanımı:** Çevre dinamiklerinin belirsizliğini azaltmayı amaçlayan keşif problemine ödüle dayalı bir yaklaşımdır [15].
- **Durum yeniliği:** Durum yenilik tabanlı keşif tekniklerindeki sezgi, ajan daha az ziyaret edilen bir durum-uzay bölgesine girdiğinde içsel bir bonus eklemektir [4].
- **Tahmin hatası:** Keşifte tahmin hatası kullanımının arkasındaki fikir, ajan tahminin zor olduğu durum bölgelerine yönlendirmektir [8].

Önerilen yöntem, politika-dışı pekiştirmeli öğrenme eğitim prosedürünün mevcut varlıklarını kullanarak, bahsedilen içsel motivasyonları, ajanları bilgilendirici durum uzaylarına yönlendirmek için tek bir yapay keşif ödül bonusu altında birleştirmektedir. Kritik fonksiyonunun, ajanın çevre dinamikleri üzerindeki kesinliğini tanımlayan temsili bir varlık olduğu bilinmektedir [5]. Eleştirilen durum-eylem çiftlerinin değerini öğrenmesi durumunda, o durum-eylem çifti üzerindeki kesinlik artmaktadır ve ajan bu durum-eylem çifti hakkında bir parça bilgi kazanmaktadır. Derin sinir ağları, kritik fonksiyonlarını tahmin etmek için kullanıldığından dolayı, Q ağlarının gradyanı kazanılan bilgi olarak varsayılabilmektedir [5]. Bu nedenle, önerilen yapay ödül bonusunun ilk varlığı olan bilgi kazanımı, karşılaşılan geçiş üzerindeki Q-ağının gradyanıdır:

$$r_{\text{BK}} = \nabla_{\theta} J(\theta) = \frac{\partial}{\partial \theta} (r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_{\theta}(s, a))^2. \quad (3)$$

Keşif ve ağ güncellemeleri politika-dışı öğrenmede ayrılrsa da, bilgi kazancı ödülü  $r_{\text{BK}}$ 'yi hesaplamak için keşif adımlarında toplanan geçiş, güncelleme simüle edilerek kullanılabilir.

Durum yenilik varlığı, bilgi kazancı kısmına kıyasla daha basittir. Bir durumun yeniliği, ajanın o durumu kaç kez ziyaret ettiği ile ölçülmektedir. Durum uzayı ayrık ise sözde sayım yaklaşımı kullanılabilir. Ancak, durum uzayı kontrol problemlerinde sürekli ve büyük durum uzayları performansı düşürebilir. Bunun yerine, gözlemlenen durumun deneyim tekrar arabelleğinde [12] depolanan durumlara olan mesafesine göre içsel ödül eklenmektedir:

$$r_{\text{DY}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \|s_{\text{yeni}} - s_i\|_1. \quad (4)$$

Burada  $|\mathcal{R}|$ , deneyim arabelleğinde [12] depolanan ve toplanan geçişlerin sayısıdır,  $\|\cdot\|_1$  birinci vektör normudur,  $s_i$ , toplanan geçiş çokuzluların içindeki durumdur ve  $s_{\text{yeni}}$ , yeni gözlemlenen durumu temsil etmektedir. Mesafelerin toplamı ve ikinci vektör normunun kullanımı, kararsızlıklara neden olabileceği ve sınırlandırılmayacağı için kullanılmamaktadır. Ayrıca, ortalama operatörü, yeni gözlemlenen durumun tüm gözlemlenen durumlardan uzak olmasını sağlamaktadır.

---

**Algoritma 1** Politika-Dışı Birleştirilmiş İçsel Motivasyonlu Keşif (PD-BİMK)

---

Aktör ve kritik ağlarını başlat.

Deneyim yeniden oynatma arabelleği  $\mathcal{R}$ 'yi başlat.

**for** her keşif zaman adımı **do**

Seçilen eylemi yürüt ve geçiş  $(s, a, r, s')$ 'i topla.

Kritik güncellemesini simüle ederek bilgi kazanımı ödülünü hesapla:

$$r_{BK} = \frac{\partial}{\partial \theta} (r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_{\theta}(s, a))^2.$$

Durum yenilik ödülünü hesapla:

$$r_{DY} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \|s_{\text{yeni}} - s_i\|_2^2.$$

Tahmin hatası ödülünü hesapla:

$$r_{TH} = |r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_{\theta}(s, a)|.$$

İçsel motivasyon ödülünü oluştur ve gerçek ödülün maksimum ortam ödülüne olan oranıyla ölçeklendir:

$$r_{IM} = \lambda \cdot (r_{BK} + r_{DY} + r_{TH}).$$

Değiştirilmiş ödülü eylem ve durumla birlikte deneyim ara belleğine kaydet:  $(s, a, r_{IM}, s') \leftarrow (s, a, r + r_{IM}, s')$ .

**end for**

**for** her güncelleme adımı **do**

$\mathcal{R}$ 'den bir mini-toplu geçiş örneği al:  $(s, a, r_{IM}, s') \sim \mathcal{R}$ .

Değiştirilmiş ödülü kullanarak derin aktör ve kritik ağlarını bir politika-dışı sürekli kontrol yöntemiyle güncelle.

**end for**

---

Aktör-kritik yöntemlerinde, eleştirmen veya eylem-değer fonksiyonu, politikanın eylem kararlarını, gözlemlenen durumlarda değerlendirmektedir [5]. Ayrıca, zamansal fark hatası, kritik fonksiyonunun tahmin hatası olduğundan, ajanlar, durum-eylem gruplarının tahmininin zor olduğu durumlarda yüksek zamansal fark hatası veren eylemleri seçmeye teşvik edilmektedir. Bu, önerilen yapay içsel ödülün son varlığını oluşturmaktadır:

$$r_{TH} = |r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_{\theta}(s, a)|. \quad (5)$$

Sonuç olarak, yapay olarak motive edilmiş bonus ödülü, bahsedilen ödül varlıklarının kombinasyonu ile hesaplanmaktadır:

$$r_{IM} = r_{BK} + r_{DY} + r_{TH}. \quad (6)$$

Bununla birlikte, oluşturulan ödül bonusu ile ilgili sorun olabilecek bir husus, her bir ödül varlığının toplanan gerçek ödülün önemli ölçüde daha büyük olabileceğidir. Böyle olası bir sorunun üstesinden gelmek için, genel içsel motivasyon ödülü  $r_{IM}$  ayrıca, gerçek ödülün ortamda elde edilebilecek maksimum ödüle oranıyla ölçeklendirilmektedir:

$$r_{IM} = \lambda \cdot (r_{BK} + r_{DY} + r_{TH}); \quad \lambda = \frac{r}{r_{\max}}, \quad (7)$$

burada  $r$  elde edilen gerçek ödüdür ve  $r_{\max}$  erişilebilecek maksimum ortam ödülüdür. Maksimum ortam ödülüne erişim, mevcut pekiştirmeli öğrenme simülatörlerinde ve pekiştirmeli öğrenmenin pratik uygulamalarındaki el yapımı ödül fonksiyonlarından elde edilebilmektedir. Genel olarak önerilen yöntemde,  $r_{IM}$  hesaplanmaktadır ve keşif adımlarında gerçek ödüle eklenmektedir. Ardından, değiştirilen ödül, eylem ve durumlarla birlikte deneyim tekrar arabelleğinde saklanmaktadır. Önerilen algoritma Politika-Dışı Birleştirilmiş İçsel Motivasyonlu Keşif (PD-BİMK) olarak adlandırılmaktadır ve Algoritma 1'de özetlenmektedir.

Derin bir sinir ağı üzerindeki geri yayılımın,  $\mathcal{O}(nmd)$  zaman karmaşıklığına sahip olduğu bilinmektedir; burada  $d$  mini-parti boyutudur,  $n$  ve  $m$  sırasıyla giriş ve çıkış vektörlerinin boyutlarıdır. Bilgi kazancı hesaplamaları için, gradyan tek bir geçiş üzerinden hesaplandığından mini-parti boyutu 1'dir. Ek olarak, Q-ağlarının çıktısı skaler Q değeridir, dolayısıyla  $m$  de 1'dir. Bu nedenle, bilgi kazancı ödül hesaplamasının çalışma zamanı durum uzay boyutluluğu  $n$ 'e bağlı olarak  $\mathcal{O}(n)$ 'dir. Benzer şekilde, ileri geçiş ve mesafe hesaplamaları da zaman içinde doğrusaldır. Sonuç olarak, önerilen yapay ödül hesaplaması basitleştirilmiş  $\mathcal{O}(n)$  değeri ile çalışmaktadır. Bu değer, genellikle aşağıdan  $\mathcal{O}(2^n)$  ile sınırlanan [6], iyi bilinen üstel keşif yöntemlerinden oldukça kısadır. Bu nedenle, önerilen algoritmanın hesaplama karmaşıklığı açısından verimli olduğu ve bir sonraki bölümde gösterildiği gibi etkili olduğu sonucuna varılmaktadır.

#### IV. DENEYLER

##### A. Deneysel Kurulumlar ve Uygulama Ayrıntıları

Önerilen yöntemin etkinliğini değerlendirmek için OpenAI Gym [9] kütüphanesindeki zorlu sürekli eylem uzayına sahip ortamlar dikkate alınmıştır. Önerilen algoritma, literatürdeki en iyi sonuçları veren iki sürekli kontrol algoritmasına uygulanmaktadır: Twin Delayed Deep Deterministic Policy Gradient (TD3) [16] ve Soft Actor-Critic (SAC) [17]. Aynı zamanda sunulan yöntem sıkça kullanılan iki keşif yöntemi olan OU-gürültü süreci [10] ve Gauss gürültüsü [7] ile karşılaştırılmıştır.

Tüm simülasyonlarda değerlendirmeler 1 milyon zaman adımı boyunca her 1000 zaman adımında gerçekleştirilmiştir. Her bir değerlendirme, güncellemeler ve keşif olmadan farklı bir değerlendirme ortamında 10 bölüm üzerinden ortalama olarak hesaplanmıştır. Her ortamdaki deneyler, pekiştirmeli öğrenme yüksek ölçüde rastgeleliğe bağlı olduğu için rastgele 10 tohum üzerinden yapılmıştır. Yeniden üretilebilirlik ve adil bir değerlendirme için ortamların simülasyon dinamiklerinde ve ödül fonksiyonlarında herhangi bir değişiklik yapılmamıştır.

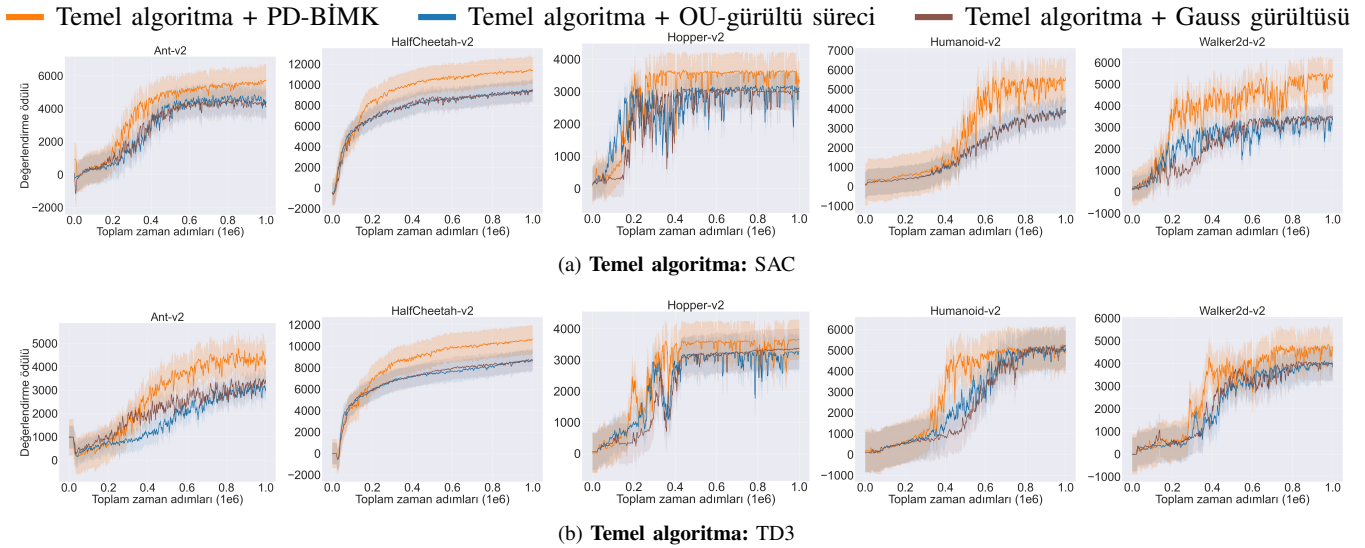
TD3 algoritması [16], yazarların GitHub depolarından alınan kodla uygulanmıştır<sup>1</sup>. SAC algoritmasının [17] uygulanması, orijinal bildiriye belirtildiği gibi tam üst değişkenleri ve mimari ayarını takip etmektedir. Önerilen yöntem, Algoritma 1'de belirtildiği gibi baz alınan algoritmaların kodları üzerine uygulanmıştır.

##### B. Değerlendirmeler

Karşılaştırmalı değerlendirme sonuçları Şekil 1'de rapor edilmiştir. İlk olarak, OU ve Gauss gürültüsünün ortaya çıkan performansının benzer olduğu gözlemlenmektedir. Bu nedenle, keşif gürültüsündeki zamansal korelasyonun sürekli kontrol görevlerinde farklılık göstermediği sonucuna varılmaktadır. Ayrıca, önerilen yöntemin test edilen tüm görevlerde Gauss ve OU gürültüsünden daha iyi performans gösterdiği gözlenmektedir. Bu nedenle, keşifte derin ağların varlığı olmasa bile, keşfe ilişkin kural tabanlı birleşik içsel teşviklerin, sürekli eylem alanlarında politika-dışı öğrenme performansını önemli ölçüde iyileştirebileceği sonucuna varılmaktadır.

---

<sup>1</sup><https://github.com/sfujim/TD3>



Şekil 1: OpenAI Gym sürekli kontrol görevleri için tanıtılan algoritmanın değerlendirme sonuçları. 10 rastgele tohum üzerinden ortalama değerlendirme getirisinin standart sapması gölgeli bölge tarafından temsil edilmektedir. Görsel netlik için boyutu 5 olan sürgülü pencere kullanılmıştır.

## V. SONUÇ

Bu bildiri, hayvan psikolojisindeki içsel motivasyonların pekiştirmeli öğrenme paradigmasında nasıl etkili bir şekilde birleştirilebileceğini araştırmaktadır. Tanıtılan yöntem, mevcut literatürdeki üç motivasyonu kural tabanlı bir şekilde tek bir keşif hedefi altında birleştirmektedir. Ampirik çalışmalar, önerilen algoritmanın baz alınan algoritmaların performansını önemli ölçüde geliştirdiğini, OU-gürültü sürecinin ve ek Gauss gürültüsünün gibi iyi bilinen yönsüz keşif yöntemlerinden daha iyi performans sergilediğini göstermektedir.

## KAYNAKLAR

- [1] 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.
- [2] 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.
- [3] 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>
- [4] 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>
- [5] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [6] S. B. Thrun, "Efficient exploration in reinforcement learning," Carnegie Mellon University, USA, Tech. Rep. CMU-CS-92-102, January 1992.
- [7] 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>
- [8] 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](https://doi.org/10.1007/978-3-642-32375-1_2)
- [9] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.
- [10] 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>
- [11] 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>
- [12] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Mach. Learn.*, vol. 8, no. 3–4, p. 293–321, May 1992. [Online]. Available: <https://doi.org/10.1007/BF00992699>
- [13] 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.
- [14] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54–67, 2000. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0361476X99910202>
- [15] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in neurobotics*, vol. 1, pp. 6–6, Nov 2007, 18958277[pmid]. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/18958277>
- [16] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1587–1596. [Online]. Available: <https://proceedings.mlr.press/v80/fujimoto18a.html>
- [17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: <https://proceedings.mlr.press/v80/haarnoja18b.html>