HAVA ROBOTU İÇİN GELİŞTİRİLEN PEKİŞTİRMELİ ÖĞRENME POLİTİKASININ GÖZLEM UZAYI, ÖDÜL FONKSİYONU VE MÜFREDAT ÖĞRENME AÇISINDAN İNCELENMESİ

Dr. Öğr. Üyesi Ali Tahir KARAŞAHİN

ISBN: 978-625-5923-92-9

Ankara -2025

HAVA ROBOTU İÇİN GELİŞTİRİLEN PEKİŞTİRMELİ ÖĞRENME POLİTİKASININ GÖZLEM UZAYI, ÖDÜL FONKSİYONU VE MÜFREDAT ÖĞRENME AÇISINDAN İNCELENMESİ

YAZAR
Dr. Öğr. Üyesi Ali Tahir KARAŞAHİN

Necmettin Erbakan Üniversitesi, Mühendislik Fakültesi, Mekatronik Mühendisliği, Konya, Türkiye
alitahir.karasahin@erbakan.edu.tr
ORCID ID: 0000-0002-7440-1312

DOI: https://doi.org/10.5281/zenodo.17083954

Copyright © 2025 by UBAK publishing house
All rights reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by
any means, including photocopying, recording or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law. UBAK International Academy of Sciences Association Publishing House®
(The Licence Number of Publicator: 2018/42945)

E mail: ubakyayinevi@gmail.com
www.ubakyayinevi.org
It is responsibility of the author to abide by the publishing ethics rules.
UBAK Publishing House – 2025©

ISBN: 978-625-5923-92-9

September / 2025
Ankara / Turkey

ÖNSÖZ

Hava robotları hem maliyet etkin olmaları sebebiyle hem de birçok alanda başarılı uygulamalar gerçekleştirmelerinden dolayı sıklıkla tercih edilmektedirler. Uygulamalar içerisindeki başarılarının artırılmasına yönelik çeşitli geliştirmeler devam etmektedir. Klasik kontrol tekniklerinin yanı sıra öğrenme tabanlı kontrol yaklaşımları yaygın bir şekilde kullanılmaya başlanmıştır. Özellikle pekiştirmeli öğrenme teknikleri hem donanımların ulaşılabilir olması nedeniyle hem de simülasyon ortamlarının yeteneklerinin gelişmesiyle güncel çalışmalar bu konuda yoğunlaşmıştır. Bu kapsamda kitabın, araştırmacılara ve uygulama içerisinde olanlara pekiştirmeli öğrenme tekniği içerisinde yer alan mekanizmaları anlama, mevcut yaklaşımları analiz etme ve gereksinimlere göre bir politika geliştirilmesi hususunda yardımcı olması amaçlanmaktadır.

09.09.2025

Dr. Öğr. Üyesi Ali Tahir KARAŞAHİN

İÇİNDEKİLER

1. GİRİŞ

Hava robotları bakım onarım çalışmalarından, arama ve kurtarma faaliyetlerine, otonom hava aracı yarışlarından lojistik ve çevresel gözlem amaçlarına kadar birçok alanda sıklıkla tercih edilmektedir (Ho vd., 2022; Kaufmann vd., 2023a; Li vd., 2021; Song vd., 2023; Suarez vd., 2022). Hava robotları bu belirtilen görevleri gerçekleştirmesi esnasında hem pozisyon hem de yönelim referans bilgilerini hassas bir şekilde takip etmesi gerekmektedir. Bu girdilerin takip edilmesinde kontrol teknikleriyle geliştirilen modelden bağımsız veya model tabanlı olacak şekilde çalışabilen kontrolörler devreye girmektedir (Amin vd., 2016).

Hava robotlarıyla gerçekleştirilen uygulamalarda çoğunlukla geleneksel PID kontrolör kullanılmaktadır (ArduPilot n.d.; PX4 Autopilot Software, n.d.; Salih vd., 2010; Szafranski & Czyba, 2011). Bunun arkasında yatan önemli sebeplerden birisi olarak açık-kaynak uçuş kontrol kartlarının (PX4 ve ArduPilot) bu kontrol yaklaşımını kullanması olarak ifade edilebilmektedir. Bu kontrolörün geliştirilmesi için sistem modelinin elde edilmesi gerekmektedir. Sistem modelinin gerçeğe yakın bir şekilde elde edilmesi kontrolör başarıyla doğrudan ilişki olduğu söylenebilmektedir.

Geleneksel PID kontrolör tekniği haricinde adaptif kontrolör yaklaşımı da tercih edilmektedir (Dydek vd., 2012; Mo & Farid, 2019; Wang vd., 2023). (Dydek et al., 2012) tarafından gerçekleştirilen

çalışmada parametre belirsizliği altında ve itki kaybından kaynaklanan hata durumlarında dayanıklılığı artırmaya yönelik birtakım yaklaşımlar ele alınmıştır. Bu çalışma ile adaptif kontrolörün arıza toleransını iyileştirmede ve referans yörüngenin takip edilmesinde başarılı sonuçlar elde edildiği belirtilmiştir. (Mo & Farid, 2019) tarafından gerçekleştirilen çalışmada hava robotunun doğrusal olmayan karakteristiğinden gelen dinamiklerle başa çıkabilecek kontrol stratejileri incelenmiştir. Bu kapsamda doğrusal olmayan kontrol, gürbüz, adaptif ve zeki kontrol tekniklerinden bahsedilmiştir. Her sistemin birtakım avantajlarından ve dezavantajları ele alınmıştır. Bu kontrol teknikleri hesaplama karmaşıklığı, gürültüye dayanıklılık, uygulama kolaylığı ve gerçek dünya performansı gibi değerlendirme kriterlerine göre incelenmiştir. Hibrit ve öğrenme tabanlı kontrol tekniklerinin kontrol performansını artırma noktasından daha başarılı olabileceği ifade edilmiştir. (J. Wang vd., 2023) tarafından gerçekleştirilen çalışmada ölçülemeyen aerodinamik parametreler ve ölçüm gürültüsü altında kontrol performansını devam ettirebilen bir adaptif kontrolör üzerine çalışmıştır. Elde edilen sonuçlara göre sağlam-adaptif kontrolör yaklaşımı ile ölçüm gürültüsü ve belirsiz aerodinamik parametrele altında dahi kararlı bir şekilde referans yörünge takibi gerçekleştirildiği paylaşılmıştır.

Model öngörülü kontrol gibi daha gelişmiş kontrol teknikleri de hava robotu alanında sıklıkla tercih edilmektedir (Kang & Hedrick, 2009; Lindqvist vd., 2020; Prach & Kayacan, 2018; Romero vd., 2022; Torrente vd., 2021). (Prach & Kayacan, 2018) tarafından

gerçekleştirilen çalışmada bir hava aracı için iki katmanlı bir hibrit yaklaşım ele alınmıştır. Konum kontrolü için geleneksel PID kontrolörü kullanılırken iç döngüde tutum kontrolü model öngörülü kontrol ile gerçekleştirilmiştir. Bu sayede hesaplama yükü azaltılmış ve gerçek zamanda uygulanabilirliği artırılmıştır. (Romero vd., 2022) tarafından gerçekleştirilen çalışmada karmaşık referans yörüngelerin takip edilebilmesi için model öngülü kontrol tabanlı yeni bir yöntem önerilmiştir. Klasik yöntemin haricinde referans yörüngenin takip edilmesi süresince ilerleme hatası şeklinde tanımlanan bir bileşen sürece dahil edilmektedir. Bu sayede sadece hatayı en aza indirmekle beraber en kısa sürede ilerlemeyi de dikkate alınacak şekilde bir yapı kurgulanmıştır. Elde edilen sonuçlara göre klasik yörünge takibi yöntemlerine göre daha kısa sürede ve daha az hata ile görevlerin gerçekleştirildiği gösterilmiştir. (Torrente vd., 2021) tarafından gerçekleştirilen çalışmada sistem modelindeki eksik veya belirsizlik noktalarında veri odaklı bir model öngörülü kontrol yaklaşımı üzerine çalışılmıştır. Gerçekleştirilen çalışmadaki hava robotu dinamikleri uçuş verileri üzerinden elde edilmiştir. Karmaşık dinamiklerin matematiksel olarak ifade edilmeden veri odaklı bir yaklaşım ile temsil edilebileceği ve bir kontrol yaklaşımı için zemin oluşturabileceği gösterilmiştir.

Belirtilen modern kontrol tekniklerinin kullanılmasıyla hava robotlarının yörünge takip performansı ve bozucu etkilere karşı daha gürbüz performans göstermeleri gerçekleştirilmiştir (Brunner vd., 2020; Z. Chen vd., 2024; Kaufmann vd., 2022; O'Connell vd., 2022).

Bu modern kontrol tekniklerine ek olarak öğrenme tabanlı kontrol olarak ifade edebileceğimiz pekiştirmeli öğrenme teknikleri de robotik sistemlerde sıklıkla tercih edilmektedir (Hoeller vd., 2024; Hwangbo vd., 2019; Ibarz vd., 2021; Kober vd., 2013; Miki vd., 2022; Rudin vd., 2022; Yu vd., 2024). (Rudin vd., 2022) tarafından gerçekleştirilen çalışmada dört ayaklı robotların dakikalar içerisinde yürümeyi öğrenecek şekilde derin pekiştirmeli öğrenme tekniği ve binlerce paralel ortam sayesinde hızlandırmayı hedeflemiştir. Ek olarak eğitim süresince oyun teorisinden esinlenen müfredat öğrenme stratejisi kullanılarak eğitimi ilerleyişi desteklenmiştir. Eğitim sonucunda öğrenilen politikanın gerçek dünya koşullarında robotları üzerinde de başarılı bir şekilde çalıştığı gösterilmiştir. Çalışmada kullanılan kodlar açık kaynak olarak paylaşılarak gelecekteki araştırmalar için bir zemin oluşturulmuştur. (Miki vd., 2022) tarafından gerçekleştirilen çalışmada dört ayaklı robotların arazi koşullarında görev yapabilmesine odaklanan yeni bir yöntem önerilmiştir. Eklem ve gövde üzerinde bulunan sensörlerden alınan veriler ile kamera ve lazer sensörleri gibi farklı kaynaklardan gelen veriler bir enkoder üzerinden birleştirilmektedir. Elde edilen veriler öğretmen-öğrenci yaklaşımı tabanlı pekiştirmeli öğrenme tekniği ile eğitilmiştir. Elde edilen politika hem gerçek dünya koşullarında doğrudan kullanılabilmiş hem de arazi koşullarında görsel verilerde bulunan bozucu etkilere karşı hızlı ve güvenli bir yürüme davranışı sergileyebilecek kontrol yöntemi olarak kullanılmıştır. (Hoeller vd., 2024) tarafından gerçekleştirilen çalışmada çevik ve güvenilir bir biçimde karmaşık ortamlarda hareket edebilecek bir dört ayaklı robot için öğrenme tabanlı bir kontrol

yönteminin geliştirilmesine odaklanılmıştır. Robotların yürüyüş gibi temel hareketlerden ziyade engelden aşma, tırmanma ve sıçrama gibi zorlayıcı hareket yeteneklerine ulaşması amaçlanmıştır. Eğitim süreci bu zorlayıcı hareket profillerini kapsayacak şekilde düzenlenmesinden sonra öğrenilen politika doğrudan gerçek dünya koşullarındaki dört ayaklı robot üzerinde test edilmiştir. Elde edilen sonuçlara göre robotun engel üzerinden atlama, yüksek basamaklara erişme ve dar alanlardan geçiş gibi zorlayıcı hareket profillerini başarıyla gerçekleştirdiği gösterilmiştir. Pekiştirme öğrenme tabanlı bir teknik üzerinden çevik parkur hareketlerinin öğrenilmesi ve gerçek dünya koşullarından doğrulanması yönüyle literatüre önemli bir katkı sunulmuştur.

Pekiştirmeli öğrenme hem verilen gözlem ve ödül fonksiyonuna bağlı olarak istenilen davranışı gerçekleştirebilecek bir potansiyele sahip olması hem de doğrudan deneyim üzerinden öğrenme kabiliyetiyle güçlü bir yaklaşım olarak ortaya çıkmaktadır. Robotik alanında pekiştirmeli öğrenme tabanlı birtakım araştırmalar gerçekleştirilmiştir ve bu tekniğinin uygulanması durumunda ümit verici sonuçlara ulaşılmıştır (Hwangbo vd., 2019; Lee vd., 2020). Bu yaklaşımı kullanarak hava robotlarıyla otonom yarış alanında insandan daha başarılı performans sergileyen politika geliştirilmiştir (Kaufmann vd., 2023b).

Hava robotları alanında belirtilen performansın sergilenmesi pekiştirmeli öğrenme tekniğinin potansiyeli ortaya koymuştur (Azar vd., 2021; Hwangbo vd., 2017; Lambert vd., 2019; Liu vd., 2019,

2024; Loquercio vd., 2021; Ma, Liu, Dang, vd., 2023; Ma, Liu, Zhao, vd., 2023; Romero vd., 2024). (Hwangbo vd., 2017) tarafından gerçekleştirilen çalışmada bir hava robotunun kontrolü pekiştirmeli öğrenme tekniği ile gerçekleştirilmiştir. Farklı zorlayıcı başlangıç noktalarından (ters çevrilmiş veya yüksek hızlarda fırlatılması gibi) başlatılmasının ardından kararlı bir şekilde görevi yerine getirebildiği gösterilmiştir. Elde edilen sonuçlara göre tamamen pekiştirmeli öğrenme tabanli bir kontrol yaklaşımı ile bir hava robotunun başarılı bir şekilde kontrol edilebileceği ifade edilmiştir. Aşırı zor olan başlangıç koşulları altında gösterdiği toparlanma davranışı ile literatüre önemli bir katkı sunulmuştur. (Lambert vd., 2019) tarafından gerçekleştirilen çalışmada eylem uzayı olarak düşük seviyeli bir kontrol yaklaşımı kullanılarak model-tabanlı pekiştirmeli öğrenme tekniği ile bir hava robotunun kontrolü gerçekleştirilmiştir. Klasik kontrol teknikleri yerine öğrenilmiş bir dinamik model üzerinden bir süreç işletilmiştir. Öncelikle hava robotunun dinamik modeli bir sinir ağı üzerinden elde edilmiştir. Ardından öğrenilmiş dinamik model üzerinden model öngörülü kontrol benzeri bir yaklaşım kullanılarak planlama işlemi gerçekleştirilmiştir. Hava robotunun farklı başlangıç koşullarından kararlı uçuşa geçebildiği gösterilmiştir. Öğrenilen politikanın düşük seviyeli kontrol komutları ile gerçek dünya koşullarında çalıştığı ifade edilmiştir. (Loquercio vd., 2021) tarafından gerçekleştirilen çalışmada yüksek hızlarda ve karmaşık gerçek dünya koşulları altında uçtan-uca olacak şekilde çalışabilen pekiştirmeli öğrenme tabanlı bir kontrol yaklaşımı önerilmiştir. Öğrenilen politika kameralar üzerinden elde edilen görsel girdiler üzerinden ve hava

robotuna ait sensörlerden gelen bilgilere göre çalışmaktadır. Eğitim ortamı gerçek dünya koşullarını yansıtacak şekilde dar geçitler, birtakım engeller ve ortamları kapsayacak şekilde düzenlenmiştir. Elde edilen sonuçlara göre yüksek hızlarda gerçek dünya koşullarında bir uçuş sergileyebilen literatürdeki ilk çalışmalardan birisi olacak şekilde bir başarıya ulaşılmıştır. (Romero vd., 2024) tarafından gerçekleştirilen çalışmada model öngörülü kontrol ile pekiştirmeli öğrenme yaklaşımlarını birleştirmeye odaklanılmıştır. Model öngörülü kontrol tekniğinden gelen öngörülebilirlik ve kısıtlara göre kontrol çıktısının hesaplanması ve pekiştirmeli öğrenme tekniğinden gelen öğrenme ve adaptasyon kabiliyetlerinin bir araya getirilmesi hususunda bir yaklaşım önerilmiştir. Organize edilen yapı içerisine aktör politikasının son katmanı olarak bir model öngörülü kontrol mekanizması yerleştirilmiştir. Geliştirilen mekanizma ile daha güvenli ve daha kararlı bir öğrenme süreci meydana geldiği ifade edilmiştir. Elde edilen sonuçlara göre önerilen yöntemin klasik kontrol tekniklerinden daha başarılı sonuçlar ürettiği gösterilmiştir. (Geles vd., 2024) tarafından gerçekleştirilen çalışmada hava robotuna sunulan görsel girdilerden herhangi bir durum kestirimi yapılmadan doğrudan kontrol çıktısının üretilmesine yönelik bir yaklaşım önerilmiştir. Uçtan-uca öğrenme yaklaşımını ile görsel veri üzerinden hareket edilen çalışmada derin pekiştirmeli öğrenme tekniği kullanılmıştır. Öğrenilen politika çevik uçuş gerektiren dar engellerden geçiş, hızlı manevralar altında ve karmaşık gerçek dünya ortamlarındaki uçuşlar ile test edilmiştir. Hem simülasyon hem de gerçek dünya koşullarında elde edilen sonuçlara göre durum kestirimi kullanılmadan bir uçuşun

gerçekleştirilmesi sebebiyle bilindiği kadarıyla literatürdeki ilk uçtan-uca pekiştirmeli öğrenme tabanlı çalışması olmuştur. Elde edilen sonuçlara göre durum kestirimi tabanlı yaklaşıma yakın sonuçlara ulaşıldığı gösterilmiştir.

Pekiştirmeli öğrenme teknikleriyle gerçekleştirilen başarılı birçok çalışma bulunmasına rağmen bu kontrol yaklaşımının dikkatli bir şekilde geliştirilmesi ve incelenmesi gereken birtakım konular bulunmaktadır. Pekiştirmeli öğrenme süresince ajan tarafından öğrenilen politikanın daha kararlı olması veya daha hızlı öğrenmesi için eğitim işleminin nasıl tasarlanacağı dikkat edilmesi gereken bir konu olarak ifade edilebilmektedir. Bu konuyla ilgili olarak müfredat öğrenme yaklaşımı pekiştirmeli öğrenmede tercih edilen bir yaklaşım olarak karşımıza çıkmaktadır.

Bu kapsamda hava robotu ile asılı yükün taşınması için geliştirilen bir pekiştirmeli öğrenme politikasında kullanılan müfredat öğrenme yaklaşımı, giriş uzayı ve ödül fonksiyonu tasarımı incelenmiştir. İnsanların ve hayvanların öğrenme biçimlerinden esinlenilerek geliştirilen müfredat öğrenme yaklaşımı hava robotunun yörünge takip performansında kullanılan pekiştirmeli öğrenme politikasındaki eğitiminde kullanılmıştır. Hava robotundan farklı yörünge profillerinin takip edilmesi istenebileceğinden dolayı pekiştirmeli öğrenme ajan eğitiminde müfredat öğrenme yaklaşımı uygulanmıştır. Elde edilen sonuçlara göre pürüzsüz yörüngelerin geliştirilen politika ile başarı ile takip edilebildiği gösterilmiştir.

Çalışmanın kalan kısmı şu şekilde organize edilmiştir: hava robotunun dinamik hareket denklemleri elde edilmiştir, 6 serbestlik dereceli sisten tanımı ve fiziksel parametrelerinden bahsedilmiştir. Pekiştirmeli öğrenme yaklaşımı başlığında belirtilen kontrol yaklaşımına ait mekanizmalara değinilmiştir. Müfredat öğrenme stratejileri başlığında robotik sistemlerde kullanılan yaklaşımlardan bahsedilmiş ve bu çalışmada kullanılan stratejiyle ilgili gerekli bilgiler paylaşılmıştır. Deneysel bulgular ve değerlendirme kısmında eğitim sonucunda elde edilen politikanın test edilmesiyle ortaya çıkan deney verilerinin incelenmesi gerçekleştirilmiştir. Sonuç ve gelecek çalışmalar bölümünde ise hava robotuyla asılı yükün taşınması için geliştirilen müfredat öğrenme yaklaşımına göre eğitim süreci gerçekleştirilen pekiştirmeli öğrenme politikasıyla elde edilen veriler değerlendirilmiştir ve mevcut yaklaşımlar altında gelecekte yapılabilecek çalışmalardan bahsedilmiştir.

2. HAVA ROBOTUNUN DİNAMİĞİ VE MODELLEME

Bu başlık içerisinde hava robotunun ifade edilmesinde kullanılan serbestlik derecesinden, hava robotunun modellenmesinde ele alınan dinamik hareket denklemlerinden ve hava robotuna ait birtakım fiziksel parametrelerinden bahsedilmiştir.

2.1. Dinamik Hareket Denklemleri

Bu araştırmada hava robotu 6 serbestlik dereceli model ile ifade edilmiştir. Bu serbestlik dereceleri sırasıyla: ileri/geri, sağ/sol, yukarı aşağı olmak üzere pozisyon değişiklilerini kapsarken, sağa/sola yuvarlanma, öne/arkaya eğilme ve sağa/sola dönme olmak üzere yönelim değişikleriyle beraber 6 serbestlik derecesi tanımlanmaktadır. Hava robotunun pozisyon ve yönelim hareketlerinin ifade edilmesi aşağıdaki denklem takımları kullanılmıştır.

$$\begin{aligned} \dot{r} &= v \\ \dot{v} &= -gk_z + \frac{1}{m}R(f+\delta) \\ \dot{R} &= R \cdot S\Omega \\ \dot{\Omega} &= I^{-1}(\tau - \Omega \times I\Omega) \end{aligned} \tag{1}$$

Denklem 1'de r hava robotunun pozisyonunu, v doğrusal hızı (m/s), R dönüşüm matrisini (3x3), k_z [0,0,1] yönünde birim vektörü ifade etmektedir. I ifadesi eylemsizlik matrisini, m hava robotunun kütlesini, δ bozucu etki vektörünü ve f toplam itki kuvvetini temsil etmektedir. S ifadesi skew-simetrik matrisi ifade etmektedir. Hava robotundan elde edilen tork ve açısal hız değerler ise τ ve Ω ile temsil edilmiştir.

2.2.Fiziksel Parametreler

Hava robotuna verilen herhangi bir referansın kontrol yaklaşımı ile takip edilmesi istenmesi durumunda sistem modelinin elde edilmesi oldukça önemli bir yere sahiptir. Öncelikle uygulama içerisinde kullanılacak hedef platforma ait sistem tanımlama işlemlerinin gerçekleştirilmesi gerekmektedir. Bu çalışmada hava robotuna ait fiziksel parametreler SimpleFlight çerçevesi önerilen çalışma esas alınmıştır (J. Chen vd., 2025). Pekiştirmeli öğrenme tekniği kullanılarak eğitimi gerçekleştirilen politikanın ortaya çıkması süresince kullanılan hava robotuna ait fiziksel parametreler Tablo 1'de gösterilmiştir.

Tablo 1. Eğitim süresince kullanılan hava robotuna ait fiziksel parametreler

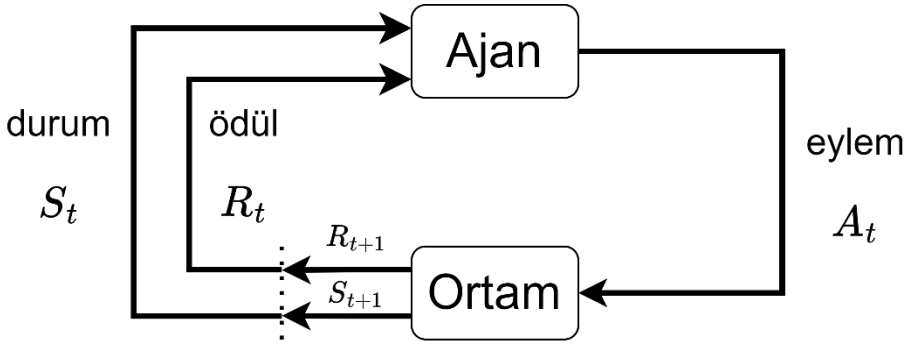
Parametreler (Birim)	**Crazyflie 2.1+**
Kütle, M[Kg]	0.0321
Eylemsizlik, I [Kg.M^2]	[1.43e-5, 1.43e-5, 2.89e-5]
Kol Uzunluğu [M]	0.043
İtki/Ağırlık Oranı	1.95
Maksimum İtki [N]	0.575
Motor Zaman Sabiti [S]	0.05

Pekiştirmeli öğrenme tekniği üzerinden elde edilen politikanın simülasyon ortamında değerlendirilmesi ve geliştirilmesi Tablo 1'de paylaşılan fiziksel parametreler referans alınarak gerçekleştirilmiştir.

3. PEKİŞTİRMELİ ÖĞRENME YAKLAŞIMI

3.1. Markov Karar Süreci Tanımı

Pekiştirmeli öğrenme algoritmasının eğitiminin gerçekleştirilebilmesi için ortam ve ajan araçlarına ihtiyaç duyulmaktadır. Ortam ve ajan arasındaki ilişki Şekil 1'de gösterilmiştir.



Şekil 1. Pekiştirmeli öğrenmede ortam ve ajan arasındaki etkileşime ait şematik gösterim (Sutton & Barto, 2018)

Ortam içerisinde kendisine verilen görevleri yerine getirmek için eylemde bulunan ajan, eylem sonucunda elde ettiği ödülü maksimize edecek şekilde bir politikanın öğrenilmesi şeklinde devam eden bir süreç gerçekleştirilmektedir. Bu kapsamda pekiştirmeli öğrenmenin ihtiyaç duyduğu ortam Denklem 1'de gösterilen dinamik hareket denklemleri üzerinden oluşturulmuştur. Bu çalışma hava robotunun istenilen yörüngenin takip edilmesi için oluşturulan görev bir Markov Karar Süreci (MDP) olarak tanımlanmıştır. MDP ile gerçekleştirilen modelleme $M = (S, A, P, R, \gamma)$ içerisindeki karar alma süreci aşağıdaki bileşenler üzerinden gerçekleştirilmektedir:

- Durum uzayı (S): hava robotunun mevcut fiziksel durumu,
- Eylem uzayı (A): ajan tarafından uygulanan kontrol çıktıları,
- Geçiş modeli (P): durumlar arası geçişi temsil eden dinamikler,
- Ödül fonksiyonu (R): kontrol çıktısına karşılık gelen geri bildirim,
- İndirim oranı (γ): gelecekteki ödüllerin mevcut karar üzerindeki etki oranı ($0 < \gamma \leq 1$).

Eğitim süresince geliştirilen politikanın temel hedefi ödül fonksiyonu sonucunda elde edilen geri bildirimin maksimize edilecek şekilde uygun ajan kontrol çıktısının üretilmesidir. Bunun gerçekleştirilebilmesi için Denklem 2'de gösterilen kazanç fonksiyonu maksimize edilmek istenmektedir.

$$J(\pi) = Beklenen\ Toplam\ \text{Ödül} = \mathrm{E}_{\pi}\left(\sum_{t=0}^{T} \gamma^{t}\, R(s_t, a_t)\right) \qquad (2)$$

3.2.Durum, Eylem ve Geçiş Uzayları

Pekiştirmeli öğrenme yaklaşımı içerisinde politikanın geliştirilmesi süresince tasarlanan ortam içerisinde ajanın yani hava robotunun bir eylemi tercih etmesi durumunda dinamik hareket denklemlerine göre hem pozisyon hem de yönelim eksenlerinde birtakım değişimleri meydana gelmektedir. Bu değişimler ajanın çalıştığı ortam içerisinde durum uzayı olarak ifade edilmektedir. Söz konusu çalışmada hava robotunun bir asılı yükle beraber bir yörüngenin takip edilmesi görevinde tanımlanan durumlar aşağıdaki gibidir:

- Hava robotunun pozisyon bilgisi: x, y, z
- Asılı yükün pozisyon bilgisi: x_p, y_p, z_p
- Hava robotunun dünya koordinat sistemindeki doğrusal hızı: v_x, v_y, v_z
- Asılı yükün dünya koordinat sistemindeli doğrusal hızı: v_{px}, v_{py}, v_{pz}
- Hava robotunun gövde koordinat sistemindeki doğrusal hızı: v_{bx}, v_{by}, v_{bz}
- Hava robotunun dönüşüm matrisi: R
- Hava robotunun ivme bilgisi: a_x, a_y, a_z
- Asılı yükün ivme bilgisi: a_{px}, a_{py}, a_{pz}

Yukarıdaki belirtilen tanımlamalara göre toplam durum vektörü 30 boyutlu bir uzayda olacak şekilde belirlenmiştir ve aşağıdaki tanımdaki gibi ifade edilebilmektedir:

$$s_t = [x, y, z, x_p, x_p, x_p, v_x, v_y, v_z, v_{px}, v_{py}, v_{pz}\ R, a_x, a_y, a_z, a_{px}, a_{py}, a_{pz}] \quad (3)$$

Pekiştirmeli öğrenme algoritması ortamdan elde edilen duruma veya gözlem uzayına bakarak ödülü maksimize edecek şekilde bir eylem uzayını belirlemektedir. Bundan dolayı eğitimden önce eylem uzayının da belirlenmesi gerekmektedir. Bu çalışmada politika çıktısı ve eylem uzayı olarak kolektif itme ve gövde hızlarının (CTBR) kullanılması tercih edilmiştir. (Kaufmann vd., 2022) tarafından gerçekleştirilen çalışmada özellikle simülasyon ve gerçek dünya arasındaki boşluğu azaltmada CTBR eylem uzayının daha başarılı

olduğu gösterilmiştir. Bu çalışmada eylem uzayı olarak CTBR aşağıdaki gibi tanımlanmıştır:

- Yuvarlanma hızı (roll rate) komutu
- Eğilme hızı (pitch rate) komutu
- Dönme hızı (yaw rate) komutu
- Toplam itki (thrust) komutu

Yönelim hız komutları $[-1\ 1]$ arasında sınırlandırılmıştır. Toplam itki komutu ise $[0\ 1]$ arasında olacak şekilde politika çıktısı olarak ortam içerisindeki ajana uygulanmaktadır.

Pekiştirmeli öğrenme içerisinde eğitim sürecinin sağlıklı bir şekilde devam ettirilebilmesi için geçiş uzayının da tanımlanması gerekmektedir. Geçiş uzayı, politika tarafından hesaplanan eylem uzayının ajan tarafından uygulanmasından sonra bir sonraki duruma nasıl geçileceğinin modellenmesi şeklinde çalışmaktadır. Bu araştırmada Isaac Sim tabanlı çalışan SimpleFlight çerçevesi tarafından geçiş fonksiyonu çalıştırılmaktadır. Hava robotunun bir sonraki duruma geçişi için aşağıdaki yaklaşım kullanılmaktadır:

$$s_{t+1} = f(s_t, a_t) \tag{4}$$

Bu geçişler esnasında gerçek dünyadaki etkilerin simülasyon ortamında karşılıklarının oluşturulması politikanın doğru bir şekilde çalışması üzerinden doğrudan bir etkisi bulunmaktadır. Bundan dolayı hava robotunun fiziksel parametrelerinin sistem tanımlama gibi yaklaşımlar ile elde edilmesi son derece önem arz etmektedir.

3.3.Gözlem Uzayı

Pekiştirmeli öğrenme tekniğiyi ile bir politikanın öğrenilmesi süresince oluşturulan ortamda bulunan tüm durumlar genellikle politikaya doğrudan verilmemektedir. Bu durumda politikaya ortam bilgileri gözlem uzayı aracılığıyla sunulmaktadır. Gözlem uzayı politikanın başarısını etkileyen önemli bir tasarım parametresi olarak karşımıza çıkmaktadır.

(Dionigi vd., 2024) tarafından gerçekleştirilen çalışmada hava robotu kontrolü için tasarlanan pekiştirmeli öğrenme kontrol politikası için farklı gözlem uzaylarının değerlendirildiği bir karşılaştırma gerçekleştirilmiştir. Öncelikle pekiştirmeli öğrenme politikasına verilecek girdiler belirlenmiştir. Politikaya verilen girdiler olarak pozisyon hatası, dönüşüm matrisi, açısal hızlar ve bir önceki kontrol çıktısı olarak tasarlanmıştır. Ardından bu girdiler farklı konfigürasyonlar üzerinden çeşitli gözlem uzayları oluşturulmuştur. Bu gözlem uzayları havada asılı durma, iki boyutlu ve üç boyutlu referans yörüngeler olmak üzere farklı deneysel senaryolar altından değerlendirilmiştir. Gerçekleştirilen değerlendirmeler neticesinde en başarılı gözlem uzayı konfigürasyonu olarak dünya koordinat sistemindeki pozisyon hatası, dönüşüm matrisi ve bir önceki kontrol çıktısı olarak ifade edilmiştir. Diğer önemli bir tasarım kriteri olarak gözlem uzayının uzunluğu olarak belirtilmiştir. Gözlem uzayı uzunluğu olarak {1,2,5,10,15} şeklinde farklı konfigürasyonlar değerlendirilmiştir. Politikanın eğitim süresince elde edilen verilere

göre en yüksek ödül ile gözlem uzayı uzunluğunun 10 olarak kullanılmasında elde edilmiştir.

(J. Chen vd., 2025) tarafından gerçekleştirilen çalışmada hava robotu kontrolü için tasarlanan pekiştirmeli öğrenme politikasının herhangi bir ek ayar yapmadan gerçek dünyada kullanılmasına odaklanılmıştır. Bu kapsamda aktör ağına verilen girdilere, kritik ağına girdiyle beraber zaman vektörünün sunulmasına, kontrol çıktısının anlamlı bir şekilde sürekliliğini sağlamak için ödül fonksiyonunda ilgili bileşen tasarımına, sistem tanımlamasının uygulanmasına ve eğitim süresince farklı yığın sayılarının kullanılmasına olacak şekilde 5 farklı kritik tasarım kriteri üzerinde çalışılmıştır. Bunlardan birisi olan gözlem uzayıyla ilgili olarak dördey (quaternion) yerine dönüşüm matrisinin kullanılması ve doğrusal hız bilgisinin aktör ağı girişinde beraber kullanılması ve kritik ağı girişinde zaman vektörünün dahil edilmesiyle beraber öğrenme eğrisinde iyileşme meydana geldiği belirtilmiştir.

Literatürde gözlem uzayıyla ilgili gerçekleştirilen çalışmalar incelendikten ve dikkate alındıktan sonra politikanın eğitimi süresince kullanılan gözlem uzayının belirlenmesi gerçekleştirilmiştir. Bu kapsamda hava robotunun bir asılı yükle beraber bir yörüngenin takip edilmesi görevinde tanımlanan gözlem uzayı aşağıdaki gibidir:

- Hava robotunun asılı yüke göre pozisyon bilgisi: x, y, z
- Gelecekteki 10 referans noktasının asılı yükün mevcut pozisyonu ile arasındaki farkı,

- Hava robotunun doğrusal hızı: v_x, v_y, v_z
- Asılı yükün doğrusal hızı: v_{px}, v_{py}, v_{pz}
- Hava robotunun dönüşüm matrisi: R

Yukarıdaki belirtilen tanımlamalara göre toplam gözlem uzayı vektörü 48 boyutlu bir uzayda olacak şekilde belirlenmiştir ve aşağıdaki tanımdaki gibi ifade edilebilmektedir:

$$o_t = [x, y, z, \Delta_x, \Delta_y, \Delta_z, v_x, v_y, v_z, R] \quad (5)$$

3.4. Ödül Fonksiyonu

Pekiştirmeli öğrenme yaklaşımı üzerinden elde edilen ve beklenen çıktıları üretmek üzere eğitilen politikanın belirtilen şekilde davranabilmesi ödül fonksiyonu ile gerçekleşmektedir. Ödül fonksiyonu pekiştirmeli öğrenmede hedeflenen davranışın modellenmesi şeklinde ifade edilebilmektedir.

Pekiştirmeli öğrenme yaklaşımı ile bir dört ayaklı robotun düz bir şekilde yürüyebilmesi, hava robotunun belirli bir yükseklik referansında asılı kalabilmesi, insansız kara robotunun engellerden kaçınarak yörüngesini takip edilmesi gibi uygulama örnekleri doğrudan ödül fonksiyonu ile ilişkilidir. Bundan dolayı eğitimi gerçekleştirilecek politikanın uygulama alanı ve gereksinimlerinin iyi bir şekilde analiz edilmesi gerekmektedir.

Uygulamaya yönelik gereksinimler analiz edildikten sonra politikanın hedeflenen davranışı gerçekleştirmesini sağlayacak şekilde bir ödül fonksiyonu tasarımı gerçekleştirilmesi gerekmektedir.

Bu araştırmada kullanılan ödül fonksiyonu aşağıdaki gibidir:

$$r = r_{pos} + r_{pos} * \left(r_{up} + r_{spin}\right) + r_{act_{norm}} + r_{act_{smooth}} \quad (6)$$

Denklem 6'da belirtilen ödül fonksiyonu içerisinde yer alan r_{pos} bileşeni aşağıdaki gibi tasarlanmıştır:

$$r_{pos} = k_d . e^{-\|p^* - p\|} \quad (7)$$

Denklem 7'de belirtilen k_d katsayısı hedef pozisyonu takip etme kabiliyetine ne kadar önem verileceğini ayarlamak için kullanılmaktadır. p^* bileşeni asılı yükün takip etmesi istenilen referans yörüngeyi tanımlamaktadır. p ifadesi asılı yükün mevcut pozisyonunu belirtmektedir. Hava robotunun taşıdığı asılı yük için verilen referans yörünge gelecekteki 10 yörünge bilgisini kapsayacak şekilde kullanılmıştır. r_{pos} bileşeni üzerinden politikanın referans yörüngeyi takip etme kabiliyetini kazanması hedeflenmiştir. Bu bileşen içerisine gelecekteki 10 referans yörünge noktasının dahil edilmesiyle daha yüksek ödül fonksiyonuna ulaşmak mümkün olmasından dolayı tercih edilmiştir.

Ödül fonksiyonu içerisindeki r_{up} bileşeni, hava robotunun asılı yüke verilen referans yörüngenin takip edilmesi esnasında dikey duruşunu kararlı bir şekilde sürdürmesini ödüllendirmektedir. r_{up} bileşeni aşağıdaki gibi tanımlanmıştır:

$$r_{up} = k_u . \frac{0.5}{1 + |1 - u_z|^2} \quad (8)$$

Denklem 8'de bulunan k_u katsayısı hava robotunun asılı yüke verilen referans yörüngenin takip edilmesi esnasında z-eksenindeki dik

durmasına verilen önemi ifade edecek şekilde kullanılmaktadır. u_z bileşeni hava robotunun dünya koordinat sistemindeki z-eksenindeki bileşenini ifade etmektedir. r_{up} bileşeni pozisyon hatasını azaltmaktan ziyade hava robotuna verilen görevin gerçekleştirilmesi esnasında yan yatma veya devrilmesini engellemek üzere ödül fonksiyonu içerisinde yer verilmiştir.

Ödül fonksiyonu içerisinde yer alan r_{spin} bileşeni, hava robotunun z-ekseni etrafında anlamsız bir şekilde dönmesini engellemek şeklinde değerlendirmeye dahil edilmiştir. r_{spin} bileşeni aşağıdaki gibi tanımlanmıştır:

$$spin = (w_z^b)^2$$
$$r_{spin} = k_s \cdot \frac{0.5}{1 + spin^2} \tag{9}$$

Denklem 9'da bulunan w_z^b ifadesi hava robotunun gövde koordinat sistemindeki z-ekseni etrafındaki açısal hızını belirtmektedir. k_s terimi ile hava robotunun z-ekseni etrafındaki kontrolsüz dönmesini engellemenin hangi oranda gerçekleştirileceği ayarlanmaktadır. Pekiştirmeli öğrenmenin eğitim süresince istenilen davranışların gerçekleştirilmesi teşvik edilirken istenmeyen davranışlardan kaçınılması şeklinde ödül fonksiyonu ile gerçekleştirilmektedir. Bundan dolayı r_{spin} gibi bileşenler dahil edilerek politikanın davranışı şekillendirilmektedir.

Ödül fonksiyonu içerisinde yer alan $r_{act_{norm}}$ bileşeni, politikanın belirlediği eylem uzayındaki bileşenlerin büyüklüğünü cezalandırmak için kullanılmaktadır. $r_{act_{norm}}$ bileşeni aşağıda gösterilmiştir:

$$r_{act_{norm}} = k_{norm}.e^{-\|a\|} \quad (10)$$

Denklem 10'da belirtilen k_{norm} katsayısı eylem uzayının büyüklüğünün ödül fonksiyonuna ne oranda etki edeceğinin belirlenmesinde kullanılmaktadır. a bileşeni ise politika tarafından üretilen kontrol komutunu ifade etmektedir. $r_{act_{norm}}$ bileşeni dolayısıyla daha küçük kontrol komutlarıyla bir eylem uzayının kullanılmasını teşvik edecek şekilde tasarlanmış olmaktadır. Dolayısıyla politika çıktısı için enerji verimliliği sağlanmaktadır. Ayrıca aşırı güç kullanımından da kaçınılmaktadır.

Ödül fonksiyonu içerisinde yer alan $r_{act_{smooth}}$ bileşeni, politikanın belirlediği eylem uzayındaki ardışık kontrol komutlarının sürekliliğini teşvik edecek şekilde kullanılmaktadır. $r_{act_{smooth}}$ ifadesi aşağıdaki gibi tanımlanmıştır:

$$r_{act_{smooth}} = k_{smooth}.e^{-\|\Delta a\|} \quad (11)$$

Denklem 11'de bulunan k_{smooth} ifadesi ödül fonksiyonuna kontrol komutunun ani değişikliklerinin ne oranda engelleneceğini ayarlamada kullanılmaktadır. Δa terimi ise ardışık kontrol komutları arasındaki farkı temsil etmektedir. $r_{act_{smooth}}$ bileşeni üzerinden kontrol komutlarının kararlı ve pürüzsüz olması gerçekleştirilmektedir. Dolayısıyla hem titreşim azaltılmakta hem de ani kontrol komutu değişiklikleri engellenmektedir.

4. MÜFREDAT ÖĞRENME STRATEJİLERİ

4.1.Müfredat Öğrenme Kavramı

Pekiştirmeli öğrenme tekniği içerisinde yer alan eğitim süreci birçok farklı şekilde düzenlenmektedir. Eğitim sürecine durum, eylem, gözlem ve geçiş uzaylarının tasarlanmasıyla müdahil olunabilmektedir. Ek olarak ödül fonksiyonu tasarımıyla da eğitim sürecinin akışına etki edilebilmektedir. Belirtilen bileşenlere ek olarak müfredat öğrenme kavramı karşımıza çıkmaktadır.

Müfredat öğrenmesi ajanın ortam içerisinde karşılaştığı görevlerin zorlukların eğitimin ilerlemesine göre giderek zorlaşması veya ortam içerisinde çevresel koşulların eğitimin belirli görevleri gerçekleştirebildikten sonra uygulanması şeklinde farklı yaklaşımlar ile gerçekleştirilmektedir. Literatürdeki müfredat öğrenme kavramı incelenmesi durumunda pekiştirmeli öğrenme eğitimi süresince görevlerin kolaydan zora doğru gösterilmesi şeklinde bir yaklaşımın kullanıldığı görülmektedir. Bu yaklaşıma ilave olarak ajanın ortam içerisindeki davranışları sonucunda elde ettiği ödüllere bağlı olarak zorluk seviyesinin dinamik olarak belirlenmesi gibi yaklaşımlar da tercih edilmektedir. Ödül fonksiyonunun da aşamalı öğrenmeyi teşvik edecek şekilde müfredat öğrenme stratejisine uygun olarak tasarlandığı yaklaşımlarda kullanılmaktadır. Müfredat öğrenme stratejisi kapsamında ek bir algoritma ile hangi görevin ne zaman ajana uygulanması gerektiğinin belirlenmesi şeklinde çalışan uygulamalarda söz konusudur.

Bu çalışmada geliştirilen politikanın farklı referans yörüngelere cevap verebilecek şekilde kontrol çıktılarının üretilebilmesi için karmaşık referans yörünge doğrudan verilmesiyle beraber ödül fonksiyonu üzerinden aşamalı öğrenmeyi gerçekleştirecek şekilde müfredat öğrenme stratejisi uygulanmıştır.

4.2.Literatürdeki Çalışmaların İncelenmesi

(Hwangbo vd., 2019) tarafından gerçekleştirilen çalışmada dört ayaklı robotların politikalarındaki iyileştirmeler için iki aşamalı bir strateji uygulanmıştır. Öncellikle robotlar dengeli bir yürüyüş sergilenmesi gibi basit görevler altında eğitilmiş ardından ise karmaşık ortam şartlarına öğrenilen politikanın korunması ve geliştirilmesini sağlayacak şekilde eğitim sürdürülmüştür. (Lee vd., 2020) tarafından gerçekleştirilen bir diğer dört ayaklı robot uygulamasında zorlu arazilerde robotların pekiştirmeli öğrenme teknikleriyle bir politika geliştirilmesi üzerine çalışmıştır. Manuel olarak uygulanan müfredat öğrenme stratejisi eğimli, çakıllı ve düzensiz yüzeylerde dört ayaklı robotların verimli bir şekilde politika geliştirilmesi sağlanmıştır. (Song vd., 2021) tarafından geliştirilen pekiştirmeli öğrenme politikası hava robotlarıyla gerçekleştirilen otonom hava robotu uygulamasına odaklanmıştır. Eğitim süresince hava robotu ilk aşamalarda hedef kapılara yakın bir pozisyonda başlatılmış, eğitim ilerledikçe başlangıç pozisyonu güncellenerek politikanın kararlılığı artırılmıştır.

(Messikommer vd., 2024) tarafından gerçekleştirilen müfredat öğrenme stratejisi ise zorlayıcı başlangıç durumlarını

önceliklendirerek politikanın öğrenme sürecini hızlandıracak bir şekilde uygulanmıştır.

(Xiao vd., 2023) tarafından gerçekleştirilen çalışmada eğimli ve dar bir boşluktan hava robotunun geçmesine yönelik bir konuya odaklanmıştır. Bu çalışmada müfredat öğrenme stratejisi olarak eğitimin ilk aşamalarında göreceli olarak daha kolay görevler gerçekleştirilirken eğitimin ilerlemesiyle görev tanımı zorlaşmaktadır. İlk başta hava robotu boşluğa yaklaşmayı öğrenirken eğitim ilerledikçe hava robotunun boşluğun arkasındaki hedefe ulaşmaya çalışmaktadır. Belirtilen müfredat öğrenme stratejisini kullanarak gerçek uçuş verisine gerek kalmadan boşluk içerisinden geçecek bir politikanın geliştirilmesi sağlanmıştır.

(M. Wang vd., 2024) tarafından gerçekleştirilen çalışmada hareket eden ve eğimli bir boşluktan geçişte müfredat öğrenme stratejisi ele alınmıştır. Belirtilen çalışmadaki müfredat öğrenme stratejisi içerisinde tanımlanan görev zorluğu, hava robotunun eğitim süresince elde ettiği başarısına göre güncellenmektedir. Eğitimdeki ilerleyişe göre boşluğun açısı ve hareket hızı adaptif bir şekilde güncellenmektedir. Bu çalışma ile müfredat öğrenme stratejisi sayesinde dinamik ve zorlu ortamlara yönelik bir politikanın geliştirilmesine ve sıfır-ayarlama (zero-shot) ile öğrenilen politikanın hava robotunda kullanılmasına odaklanılmıştır.

(Eschmann vd., 2024) tarafından gerçekleştirilen çalışmada hava robotu için havada asılı kalma görevinin saniyeler içerisinde öğrenilmesi ele alınmıştır. Çalışma içerisinde müfredat öğrenme

stratejisi kademeli ödül değerlendirilmesinde kullanılmıştır. Eğitim süresince kullanılan ödül fonksiyonu içerisindeki katsayılar ilerleyişe göre güncellenmektedir. Başlangıçta hava robotu pozisyon hatasına karşı toleranslı iken eğitim ilerledikçe cezalandırma eğiliminde bulunacak şekilde katsayıları güncellenmektedir. Bu şekilde ajanın ilk aşamada havada asılı kalmayı öğrenmesi teşvik edilmektedir. Düşük seviyeli eylem uzayına sahip olması çalışmanın gerçek dünyada uygulanması açısından bir dezavantaj iken saniyeler içerisinde bir politikanın öğrenilmesi değerlendirilmesi gereken önemli bir çıktı olarak ifade edilmiştir.

4.3. Çalışmada Kullanılan Müfredat Öğrenme Stratejisi

Bu çalışmada kullanılan müfredat öğrenme stratejisi olarak eğitim süresince takibi zor olan bir karmaşık yörüngenin ajana referans olarak verilmesi şeklinde bir yaklaşım tercih edilmiştir. Hava robotunun asılı yükün taşıması esnasında kendisine birçok farklı karakteristiğe sahip yörünge referans olarak verilebilir. Bundan dolayı pekiştirmeli öğrenme ile geliştirilen bir politikanın farklı referans yörüngelere cevap verebilecek şekilde bir eğitim sürecinin planlanması düşünülmüştür.

Tercih edilen yaklaşım ile hem politikanın zorlayıcı yörünge takip performansı hem de politikanın farklı yörüngeler altında genelleme yeteneği geliştirilmiştir. Eğitimde kullanılan referans yörünge hem takibi mümkün olan düz hareket profilini hem de referans takibini zorlaştıran köşeli hareket profillerini barındırmaktadır.

5. SİMÜLASYON ORTAMI VE EĞİTİM SÜRECİ

Bu çalışmada pekiştirmeli öğrenme yaklaşımı için gerekli olan ortam ve ajan temsili için SimpleFlight çerçevesi kullanılmıştır (J. Chen vd., 2025). SimpleFlight ortamı Isaac Sim yazılımı üzerine inşa edilen bir çerçeve olarak hava robotuna ait ortamı sunmaktadır. Isaac Sim ortamı robotik, otonom sistemler ve pekiştirmeli öğrenme uygulamaları için geliştirilen bir alternatif olarak karşımıza çıkmaktadır. PhysX tabanlı fizik motoru, GPU (Graphics Processing Unit) sayesinde yüksek düzeyde ölçeklenebilirlik desteği ve çeşitli sensör entegrasyonları sayesinde araştırmacılar için güçlü bir platform olarak kullanılmaktadır. Belirtilen simülasyon ortamındaki tasarım dosyaları üzerinde çeşitli özelleştirmeler gerçekleştirilebilmektedir. Belirtilen ortamında kullanılması planan hava robotuna ait tasarım dosyası ve sistem parametrelerinin elde edilmesi yeterli olmaktadır.

Pekiştirmeli öğrenme tekniğine dayalı olarak tasarlanan ajan ise Proximal Policy Optimization (PPO) algoritmasıyla politikayı öğrenmektedir (Schulman vd., 2017). PPO algoritması sürekli eylem uzayına sahip görevlerde sıklıkla tercih edilmektedir. PPO algoritması eğitim süresince büyük güncellemelerden kaçınması sebebiyle daha kararlı bir öğrenme gerçekleştirmektedir. PPO algoritmasında kullanılan entropi düzenlemesiyle ajanın keşif yeteneği sürdürülürken aynı zaman lokal minimuma yönelme ihtimali de azalma eğiliminde olduğunu söylenebilmektedir.

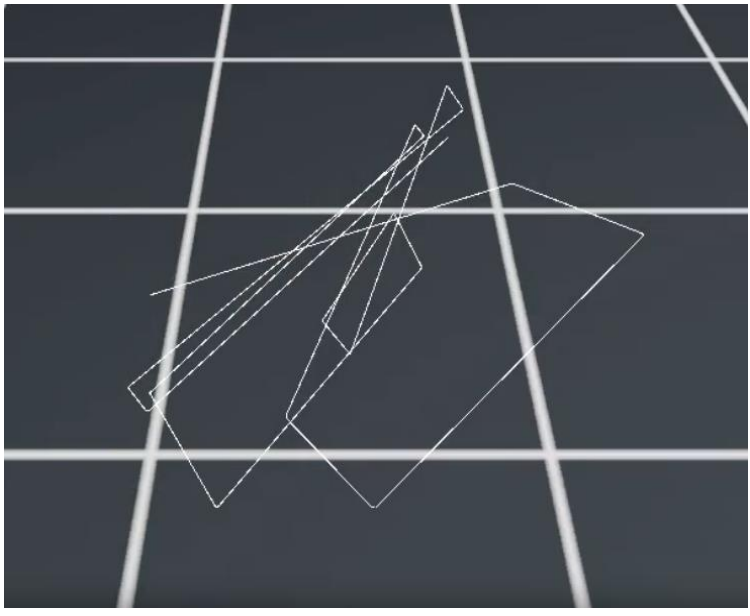
Asimetrik aktör-kritik ağı mimarisine göre tasarlanan SimpleFlight çerçevesi hem yüksek paralel ortam yeteneği hem de ayarlama

gerekmeksizin politikanın dünya ortamında kullanılmasına izin verecek şekilde tasarlanmasından dolayı tercih edilmiştir.

Politikanın eğitimi süresince 512 paralel ortam kullanılmıştır ve 7500 bölümden oluşan eğitim NVIDIA 4060 GPU cihazı ile gerçekleştirilmiştir. Paralel ortamların kullanılması pekiştirmeli öğrenme eğitim süresi için bir hızlandırıcı yaklaşım olarak kullanılmaktadır. Çoklu ortam desteğiyle hava robotu farklı paralel ortamlar üzerinden farklı başlangıç koşulları üzerinden ilerleyişine devam edeceğinden dolayı karşılaşacağı durumlar çeşitlenmektedir. Dolayısıyla öğrenilen politikanın lokal minimuma sıkışmaktan kurtarmaya ve genelleme kabiliyetini artıracak şekilde bir katkı sunulmaktadır. Belirtilen donanım özellikleriyle ve paralel ortam sayısıyla politikanın eğitilmesi 6 saat sürmüştür. Paralel ortam sayısı doğrudan GPU'nun sahip olduğu donanım sınırlamasıyla ilişkili olmasından dolayı 512 ortam olarak tercih edilmiştir. GPU donanımının yeterli olması durumunda Isaac Sim ortamında farklı çalışmalar 8192 ortam ile eğitimlerini gerçekleştirmiştir.

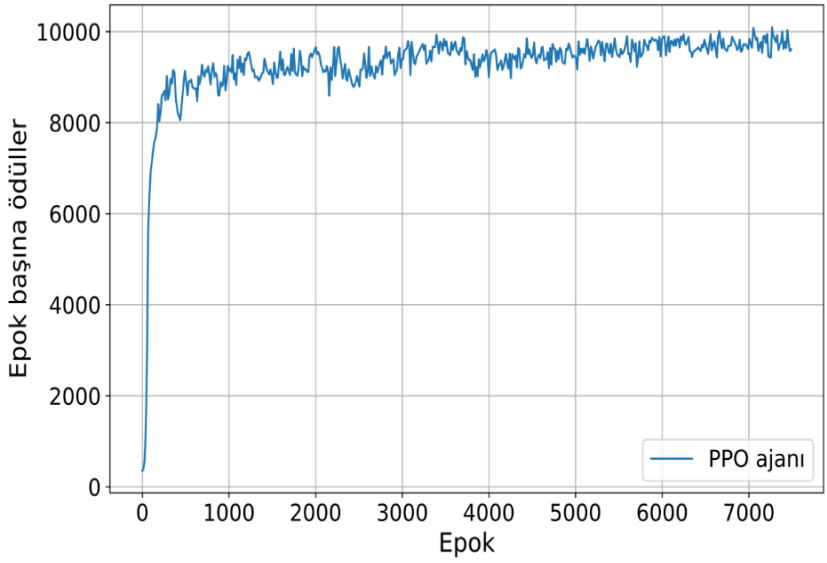
6. DENEYSEL BULGULAR VE DEĞERLENDİRME

Bu başlık altında müfredat öğrenme stratejisine göre eğitilen politika üzerinden elde edilen sonuçlar paylaşılmıştır. Eğitilen politikanın gerçek dünya koşulları altında veya simülasyon ortamında değerlendirilmeden önce doğrulanması gerekmektedir. Eğitimin başarılı bir şekilde gerçekleştiği eğitim süresince kayıt altına alınan performans kriterleri üzerinden takip edilirken aynı zamanda ilave bir doğrulamaya da ihtiyaç duyulmaktadır. Bu aşamada politikanın eğitiminde kullanılan simülasyon ortamında bu işin gerçekleştirilmesi ilk başvurulması gereken yöntem olarak karşımıza çıkmaktadır. Eğitim süresince asılı yükü taşıyan bir hava robotuna takip etmesi için verilen referans yörünge Şekil 2'de gösterilmiştir.



Şekil 2. Eğitim süresince kullanılan zorlayıcı yörüngenin referans olarak verilmesi

Müfredat öğrenme stratejisine bağlı olarak doğrudan hedef referans yörüngenin eğitimde kullanılmasından ziyade zorlayıcı bir yörünge ile eğitim başlatılmıştır. Eğer ajan zorlayıcı bir yörünge takibinde kararlı bir politika geliştirmesi durumunda pürüzsüz olarak kendisine verilen bir yörüngenin takip edilmesini de gerçekleştirilmesi beklenmektedir. Ajan tarafından referans yörüngenin öğrenilmesi süresince elde edilen ödül değerleri Şekil 3'te gösterilmiştir.



Şekil 3. Eğitim süresince ajan tarafından elde edilen ödüllerin eğitim sürecindeki değişimi

Eğitim süresince elde edilen ödül grafiği incelendiğinde ajan tarafından politikanın kararlı bir şekilde geliştirildiği ve ödülü maksimize edecek şekilde bir davranışın öğrenildiği gözlemlenmiştir. Burada eğitimin ne kadar süre boyuna devam ettirileceği pekiştirmeli öğrenme tasarımını gerçekleştiren uzman kişinin vereceği bir karar

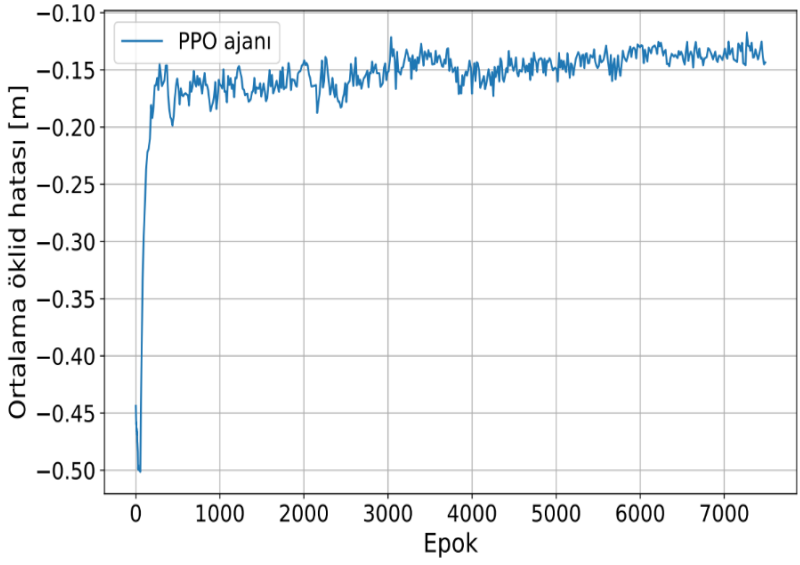
olarak karşımıza çıkmaktadır. Eğitimin süresi sabit bir epok (epoch) sayısı olarak kullanılabileceği gibi ödülün belirli bir değere ulaşması durumunda eğitimin sonlandırılması şeklinde de kullanılabilmektedir. Politikanın elde ettiği ödül fonksiyonun değişimi hem artma eğiliminde olması beklenirken hem de elde edilen kazanımın korunması da beklenmektedir. Eğer politika belirli bir ödüle ulaştıktan sonra bunu koruyamazsa politika çöküşü olarak ifade edilen bir problem gerçekleşmiş olmaktadır.

Pekiştirmeli öğrenme eğitimi devam ettirilirken politika çöküşü ve yıkıcı unutma gibi problemlerin gerçekleşip gerçekleşmediğinin de takip edilmesi oldukça önemli bir husustur. Politika çöküşü genellikle uzun süreli eğitimlerde sinir ağı içerisindeki âtıl nöronların sayısının artması sonucundan, keşif ve sömürü arasındaki dengenin bozulmasından, hatalı ödül fonksiyonu katsayılarından, yanlış belirlenen öğrenme oranından ve ajanın ortam içerisindeki yetersiz deneyim çeşitliliğiden ortaya çıktığı ifade edilebilmektedir. Yıkıcı unutma ise politikanın daha önce öğrendiği bilgileri yeni görev ve mevcut görev içerisindeki değişimlerle karşılaşılması durumunda ortaya çıkmaktadır. Belirtilen durum genellikle sürekli öğrenme ve çoklu görevli pekiştirmeli öğrenme tekniği uygulamalarında karşılaşılan bir problem olarak belirtilmektedir. Mevcut öğrenilen bilgilere göre eğitilen sinir ağının yeni görev veya ortam değişiklikleri durumunda ağırlıkların güncellenmesi sonucunda performansın ciddi bir şekilde düşmesiyle sonuçlanmaktadır. Hem politika çöküşü hem de

yıkıcı unutma için önerilen yöntemler ile belirtilen problemlerle başa çıkmak mümkündür.

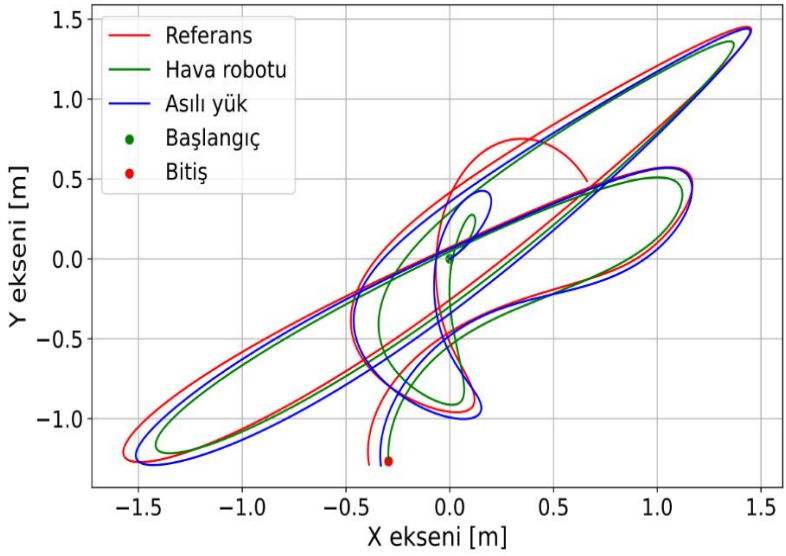
Eğitimi tamamlanan politikanın değerlendirilmesi hususunda toplam ödül fonksiyonunun değişimi inceleneceği gibi ödül fonksiyonu içerisindeki her bileşenin eğitim sürecindeki elde ettiği değerlerin incelenmesi daha sağlıklı bir yaklaşım olarak ifade edilebilmektedir.

Politikanın eğitim süresince müfredat öğrenme stratejisine göre gerçekleştirilmesi esnasında ajanın elde ettiği takip hatası da incelenmiştir. Takip hatası performans metriği, ajana verilen referans ile mevcut pozisyonu arasındaki fark ortalama öklid hatasına göre hesaplanmaktadır. Ortalama öklid hatasına göre eğitim süresince öğrenilen politikanın performansı Şekil 4'te gösterilmiştir.



Şekil 4. Eğitim süresince ajan tarafından elde edilen ortalama öklid hatasının eğitim sürecindeki değişimi

Eğitim süresince elde edilen verilere göre ortalama öklid hatasının zamanla minimize edilerek kararlı bir şekilde eğitimin tamamlandığı gözlemlenmiştir. Eğitimin tamamlanmasından sonra politikanın simülasyon ortamında değerlendirilmesi pürüzsüz yörünge olarak ifade edilebilecek polinom yörüngesi altındaki performansı ile ele alınmıştır. Hava robotuna verilen polinom referans yörüngenin takip edilmesi esnasında elde edilen sonuçlar Şekil 5'te gösterilmiştir.



Şekil 5. Politikanın değerlendirilmesi aşamasında elde edilen yörünge takibi performansı

Elde edilen sonuçlara göre politikaya referans olarak verilen polinom yörüngesini başarılı bir şekilde takip ettiği gözlemlenmiştir. Eğitimde farklı bir referans yörüngenin kullanılması ve değerlendirme aşamasında farklı bir referans yörünge üzerinde politikanın performansının incelenmesiyle ajanın genelleme yeteneği

gözlemlenmiştir. Simülasyon ortamında elde edilen başarı oranı ve ortalama öklid hatası (Mean Euclidean Distance Error (MED)) Tablo 2'de gösterilmiştir.

Tablo 2. Politikanın değerlendirme sürecinde performans metrikleri üzerinden elde edilen sonuçlar

Metot	**Başarı Oranı (%)**	**MED Hatası [M]**
Simülasyon	100	0.1024

Elde edilen sonuçlara göre politikanın polinom yörüngesini her denemede başarıyla tamamladığını göstermektedir. Ek olarak MED performansına göre asılı yükün verilen referansın takip edilmesi 0.1024 m gibi bir hata ile gerçekleştirilmiştir. Bu görev esnasında hava robotunun takip performansı ise 0.1408 m olarak gözlemlenmiştir. Referansın takip edilmesi sırasında hava robotunun ortalama hızı ise 1.4212 m/s olarak tespit edilmiştir. Elde edilen sonuçlara göre politikanın genelleme yeteneği gösterilmiştir.

Tablo 2'de belirtilen MED hatasının daha da azaltılması için eğitim süresince kullanılan paralel ortam sayısının artırılması yöntemine başvurulabilir. (J. Chen vd., 2025) tarafından gerçekleştirien çalışmada politikanın eğitimi süresince farklı paralel ortamların (256, 2048, 4096 ve 8192) kullanılması durumunda MED hatasının daha da azaltılabileceği gösterilmiştir. Bu iyileşme hava robotuyla uygulanan yavaş, normal ve hızlı olarak ifade edilebilecek uçuşların gerçekleştirilmesi senaryolarında incelenmiştir. Paralel ortamların

sayısının artırılmasının GPU donanım belleğiyle doğrudan ilişkili bir husus olduğu da belirtilmelidir. Bundan dolayı eğitimin gerçekleştirildiği fiziki donanım da bir kısıtlayıcı unsur olarak uygulayıcıların karşısına çıkmaktadır. Aynı çalışmada dikkat çekilen başka bir hususta kritik sinir ağı girişine zaman bilgisinin verilmesi olduğu gösterilmiştir. Bu sinir ağına giriş olarak zaman bilgisinin verilmesi durumunda daha düşük MED hatasının elde edilebileceği ifade edilmiştir. Bu ilave bilgi sayesinde kritik ağının durum değerlerini daha iyi bir şekilde tahmin etmesine imkân tanıyabildiği şeklinde değerlendirilmiştir.

Tablo 2'de belirtilen MED hatasının daha da düşürülmesi için asılı yükü temsil eden ip modelindeki iyileştirmelere de gidilebilir. Pekiştirmeli öğrenme politikasının eğitimi süresince ortam içerisinde tasarlanan hava robotu ve yük arasındaki ip sabit bir çubuk ile temsil edilmiştir. Dolayısıyla, asılı yükün herhangi bir çevresel etkene temas etmemesi, etkileşimde bulunmaması adına büyük bir sorun olarak düşünülmemektedir. Fakat esnek ip modelinin hem etkileşim gerektiren uygulamalarda kullanılması gerekliliği hem de gerçeğe daha yakın bir temsil açısından kullanılması takip performansında bir iyileşme sağlayabileceği düşünülmektedir.

7. SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada pekiştirmeli öğrenme tekniği içerisinde yer alan durum, gözlem, eylem ve geçiş uzayları ele alınmış ve eğitim aşamasına doğrudan etkisi olan ödül fonksiyonu incelenmiştir. Hem durum, gözlem, eylem ve geçiş uzaylarının hem de ödül fonksiyonunun pekiştirmeli öğrenme içerisindeki öneminden bahsedilmiştir. Robotik alanında gerçekleştirilen çalışmalar bu bakış açısıyla ele alınmıştır. Mevcut literatür içerisinde yer alan önemli çıkarımlar bu çalışma içerisine dahil edilerek pekiştirmeli öğrenme tekniği üzerindeki güncel araştırmalar ve tartışmalar değerlendirilmiştir.

Bu kapsamda hava robotuyla asılı bir yükün taşınması problemi ele alınmıştır. Hava robotları farklı görevlerde başarılı bir şekilde performans sergilemelerinden dolayı güncel literatür bu alana ilgi duymaktadır. Asılı bir yük ile hava robotunun ilişkilendirilmesi sistem dinamiklerini zorlaştırırken pekiştirmeli öğrenme tekniğinin bu problem üzerindeki performansını sergilemek için özellikle tercih edilmiştir.

Asılı yükün bir hava robotuyla taşınması uygulaması için pekiştirmeli öğrenme tekniği kullanılarak referans yörüngenin takip edilmesini gerçekleştirilecek bir politika eğitimi ve değerlendirilmesi gerçekleştirilmiştir. Eğitim süresince müfredat öğrenme stratejisine göre hava robotuna bir referans girdisi sağlanmıştır. SimpleFlight çerçevesiyle gerçekleştirilen eğitim neticesinde politika polinom yörünge altında değerlendirilmiştir. Elde edilen sonuçlara göre eğitim

ve değerlendirme esnasında farklı yörünge profilinin kullanılması durumunda politikanın bir genelleme sağlayabileceği gösterilmiştir.

Pekiştirmeli öğrenme tekniği üzerinden bir hava robotunun kontrol edilmesiyle ilgili çeşitli çalışmalar devam ederken asılı yükün bu sisteme uygulanması yeni birçok güncel araştırma problemini gündeme getirmektedir. Hava robotunun asılı bir yük ile dar geçitlerden geçirilmesi, yüksek hızlarda kararlı yörünge takibinin gerçekleştirilmesi hem hava robotu hem de asılı yük için verilen durak noktalarının takip edilmesiyle ilgili yeni birtakım pekiştirmeli öğrenme yaklaşımlarına ihtiyaç duyulmaktadır. Belirtilen senaryolara uygun olacak şekilde gözlem uzayı ve ödül fonksiyonu tasarımı incelenmesi ve yeni geliştirmelerin yapılabileceği potansiyel konular olarak ifade edilebilir.

Pekiştirmeli öğrenme tekniği üzerinden elde edilen politikanın gerçek dünya koşullarında herhangi bir ek ayar gerektirmeksizin çalışabilmesi gibi konular güncel literatürün ilgi gösterdiği çalışmalar olarak karşımıza çıkmaktadır ve mevcut geliştirmelerin daha iyi performans sergileyebilmesi için gelişmeye ihtiyaç duyulan konular olarak gelecekteki çalışmalar için zemin hazırlamaktadır.

KAYNAKÇA

Amin, R., Aijun, L., & Shamshirband, S. (2016). A review of quadrotor UAV: control methodologies and performance evaluation. *International Journal of Automation and Control*, *10*(2), 87–103.

ArduPilot/ardupilot: ArduPlane, ArduCopter, ArduRover, ArduSub source. (n.d.). Retrieved August 18, 2025, from https://github.com/ArduPilot/ardupilot

Azar, A. T., Koubaa, A., Ali Mohamed, N., Ibrahim, H. A., Ibrahim, Z. F., Kazim, M., Ammar, A., Benjdira, B., Khamis, A. M., Hameed, I. A., & others. (2021). Drone deep reinforcement learning: A review. *Electronics*, *10*(9), 999.

Brunner, M., Bodie, K., Kamel, M., Pantic, M., Zhang, W., Nieto, J., & Siegwart, R. (2020). Trajectory Tracking Nonlinear Model Predictive Control for an Overactuated MAV. *Proceedings - IEEE International Conference on Robotics and Automation*, 5342–5348. https://doi.org/10.1109/ICRA40945.2020.9197005

Chen, J., Yu, C., Xie, Y., Gao, F., Chen, Y., Yu, S., Tang, W., Ji, S., Mu, M., Wu, Y., & others. (2025). What Matters in Learning A Zero-Shot Sim-to-Real RL Policy for Quadrotor Control? A Comprehensive Study. *IEEE Robotics and Automation Letters*.

Chen, Z., Lin, X., & Jiang, W. (2024). Coaxial Helicopter Attitude Control System Design by Advanced Model Predictive Control

under Disturbance. *Aerospace 2024, Vol. 11, Page 486*, *11*(6), 486. https://doi.org/10.3390/AEROSPACE11060486

Dionigi, A., Costante, G., & Loianno, G. (2024). The Power of Input: Benchmarking Zero-Shot Sim-to-Real Transfer of Reinforcement Learning Control Policies for Quadrotor Control. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11812–11818. https://doi.org/10.1109/IROS58592.2024.10802831

Dydek, Z. T., Annaswamy, A. M., & Lavretsky, E. (2012). Adaptive control of quadrotor UAVs: A design trade study with flight evaluations. *IEEE Transactions on Control Systems Technology*, *21*(4), 1400–1406.

Eschmann, J., Albani, D., & Loianno, G. (2024). Learning to Fly in Seconds. *IEEE Robotics and Automation Letters*, *9*(7), 6336–6343. https://doi.org/10.1109/LRA.2024.3396025

Geles, I., Bauersfeld, L., Romero, A., Xing, J., & Scaramuzza, D. (2024). *Demonstrating Agile Flight from Pixels without State Estimation*. https://doi.org/10.48550/ARXIV.2406.12505

Ho, B., Kocer, B. B., & Kovac, M. (2022). Vision based crown loss estimation for individual trees with remote aerial robots. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 75–88.

Hoeller, D., Rudin, N., Sako, D., & Hutter, M. (2024). Anymal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics*, *9*(88), eadi7566.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., & Hutter, M. (2019). Learning agile and dynamic motor skills for legged robots. *Science Robotics*, *4*(26), eaau5872. https://doi.org/10.1126/scirobotics.aau5872

Hwangbo, J., Sa, I., Siegwart, R., & Hutter, M. (2017a). Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters*, *2*(4), 2096–2103. https://doi.org/10.1109/LRA.2017.2720851

Hwangbo, J., Sa, I., Siegwart, R., & Hutter, M. (2017b). Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters*, *2*(4), 2096–2103. https://doi.org/10.1109/LRA.2017.2720851

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, *40*(4–5), 698–721.

Kang, Y., & Hedrick, J. K. (2009). Linear tracking for a fixed-wing UAV using nonlinear model predictive control. *IEEE Transactions on Control Systems Technology*, *17*(5), 1202–1210.

Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023a). Champion-level drone racing using deep reinforcement learning. *Nature 2023 620:7976*, *620*(7976), 982–987. https://doi.org/10.1038/s41586-023-06419-4

Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023b). Champion-level drone racing using deep reinforcement learning. *Nature*, *620*(7976), 982–987. https://doi.org/10.1038/s41586-023-06419-4

Kaufmann, E., Bauersfeld, L., & Scaramuzza, D. (2022). A benchmark comparison of learned control policies for agile quadrotor flight. *2022 International Conference on Robotics and Automation (ICRA)*, 10504–10510.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

Lambert, N. O., Drew, D. S., Yaconelli, J., Levine, S., Calandra, R., & Pister, K. S. J. (2019). Low-level control of a quadrotor with deep model-based reinforcement learning. *IEEE Robotics and Automation Letters*, *4*(4), 4224–4230.

Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2020). Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, *5*(47), eabc5986. https://doi.org/10.1126/scirobotics.abc5986

Li, G., Ge, R., & Loianno, G. (2021). Cooperative transportation of cable suspended payloads with mavs using monocular vision and inertial sensing. *IEEE Robotics and Automation Letters*, *6*(3), 5316–5323.

Lindqvist, B., Mansouri, S. S., Agha-mohammadi, A., & Nikolakopoulos, G. (2020). Nonlinear MPC for collision avoidance and control of UAVs with dynamic obstacles. *IEEE Robotics and Automation Letters*, *5*(4), 6001–6008.

Liu, X., Liu, Y., & Chen, Y. (2019). Reinforcement learning in multiple-UAV networks: Deployment and movement design. *IEEE Transactions on Vehicular Technology*, *68*(8), 8036–8049.

Liu, X., Yuan, Z., Gao, Z., & Zhang, W. (2024). Reinforcement learning-based fault-tolerant control for quadrotor UAVs under actuator fault. *IEEE Transactions on Industrial Informatics*.

Loquercio, A., Kaufmann, E., Ranftl, R., Müller, M., Koltun, V., & Scaramuzza, D. (2021). Learning high-speed flight in the wild. *Science Robotics*, *6*(59), eabg5810.

Ma, B., Liu, Z., Dang, Q., Zhao, W., Wang, J., Cheng, Y., & Yuan, Z. (2023). Deep reinforcement learning of UAV tracking control under wind disturbances environments. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–13.

Ma, B., Liu, Z., Zhao, W., Yuan, J., Long, H., Wang, X., & Yuan, Z. (2023). Target tracking control of UAV through deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, *24*(6), 5983–6000.

Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2022). Learning robust perceptive locomotion for

quadrupedal robots in the wild. *Science Robotics*, *7*(62), eabk2822.

Mo, H., & Farid, G. (2019). Nonlinear and adaptive intelligent control techniques for quadrotor uav–a survey. *Asian Journal of Control*, *21*(2), 989–1008.

O'Connell, M., Shi, G., Shi, X., Azizzadenesheli, K., Anandkumar, A., Yue, Y., & Chung, S. J. (2022). Neural-Fly enables rapid learning for agile flight in strong winds. *Science Robotics*, *7*(66), 6597. https://doi.org/10.1126/SCIROBOTICS.ABM6597/SUPPL_FILE/SCIROBOTICS.ABM6597_SM.PDF

Prach, A., & Kayacan, E. (2018). An MPC-based position controller for a tilt-rotor tricopter VTOL UAV. *Optimal Control Applications and Methods*, *39*(1), 343–356.

PX4/PX4-Autopilot: PX4 Autopilot Software. (n.d.). Retrieved August 18, 2025, from https://github.com/PX4/PX4-Autopilot

Romero, A., Song, Y., & Scaramuzza, D. (2024). Actor-critic model predictive control. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14777–14784.

Romero, A., Sun, S., Foehn, P., & Scaramuzza, D. (2022). Model predictive contouring control for time-optimal quadrotor flight. *IEEE Transactions on Robotics*, *38*(6), 3340–3356.

Rudin, N., Hoeller, D., Reist, P., & Hutter, M. (2022). Learning to walk in minutes using massively parallel deep reinforcement learning. *Conference on Robot Learning*, 91–100.

Salih, A., Moghavvemi, M., & Mohamed, H. A. F. (2010). Flight PID controller design for a UAV quadrotor. *Scientific Research and Essays (SRE)*, *5*(23), 3660–3667.

Song, Y., Romero, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Reaching the limit in autonomous racing: Optimal control versus reinforcement learning. *Science Robotics*, *8*(82), eadg1462.

Suarez, A., Salmoral, R., Garofano-Soldado, A., Heredia, G., & Ollero, A. (2022). Aerial device delivery for power line inspection and maintenance. *2022 International Conference on Unmanned Aircraft Systems, ICUAS 2022*, 30–38. https://doi.org/10.1109/ICUAS54217.2022.9836039

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book.

Szafranski, G., & Czyba, R. (2011). Different approaches of PID control UAV type quadrotor. *International Micro Air Vehicle Conference and Competitions 2011 (IMAV 2011),'t Harde, The Netherlands, September 12-15, 2011*.

Torrente, G., Kaufmann, E., Föhn, P., & Scaramuzza, D. (2021). Data-driven MPC for quadrotors. *IEEE Robotics and Automation Letters*, *6*(2), 3769–3776.

Wang, J., Zhu, B., & Zheng, Z. (2023). Robust adaptive control for a quadrotor UAV with uncertain aerodynamic parameters. *IEEE Transactions on Aerospace and Electronic Systems*, *59*(6), 8313–8326.

Wang, M., Jia, S., Niu, Y., Liu, Y., Yan, C., & Wang, C. (2024). Agile Flights Through a Moving Narrow Gap for Quadrotors Using Adaptive Curriculum Learning. *IEEE Transactions on Intelligent Vehicles*, *9*(11), 6936–6949. https://doi.org/10.1109/TIV.2024.3391384

Xiao, C., Lu, P., & He, Q. (2023). Flying Through a Narrow Gap Using End-to-End Deep Reinforcement Learning Augmented With Curriculum Learning and Sim2Real. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(5), 2701–2708. https://doi.org/10.1109/TNNLS.2021.3107742

Yu, H., De Wagter, C., & de Croon, G. C. H. E. (2024). MAVRL: Learn to fly in cluttered environments with varying speed. *IEEE Robotics and Automation Letters*.

HAVA ROBOTU İÇİN GELİŞTİRİLEN PEKİŞTİRMELİ ÖĞRENME POLİTİKASININ GÖZLEM UZAYI, ÖDÜL FONKSİYONU VE MÜFREDAT ÖĞRENME AÇISINDAN İNCELENMESİ