

نویسندگان: لورا گراسر | والون کنگ

# اصول و مبانی یادگیری تقویتی ژرف

---

تئوری و عمل

در Python

مترجمان:

دکتر مهدی اسماعیلی

دکتر روجیار پیرمحمدیانی

# اصول و مبانی یادگیری تقویتی ژرف

## تئوری و عمل در Python

مترجمان: دکتر مهدی اسماعیلی، دکتر روحیار پیرمحمدیانی

ویراستار علمی: دکتر رامین مولاناپور

ناشر: انتشارات آتی نگر

مدیر هنری و طراح جلد: همتا بیداریان

چاپ اول، ۱۴۰۳

شمارگان: ۵۰۰ نسخه

قیمت: ۵,۷۵۰,۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۸۲۴۵-۳۳-۱

ISBN: 978-622-8245-33-1

حق چاپ برای انتشارات آتی نگر محفوظ است.

نشانی دفتر فروش: خیابان جمالزاده جنوبی، روبه روی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نمابر: ۶۶۵۶۵۳۳۷

تلفن: ۸-۶۶۵۶۵۳۳۶



www.ati-negar.com \* info@ati-negar.com

سرشناسه: گراسر، لورا، Graesser, Laura

اصول و مبانی یادگیری تقویتی ژرف «تئوری و عمل در Python» / نویسندگان: لورا گراسر، وا لون کنگ / مترجمان: مهدی اسماعیلی،

روحیار پیرمحمدیانی / ویراستار علمی: رامین مولاناپور

تهران: آتی نگر ۱۴۰۳

۴۶۰ ص: مصور، جدول، نمودار.

ISBN: 978-622-8245-33-1

فیبا.

یادداشت: عنوان اصلی کتاب: [2020] Foundations of deep reinforcement learning : theory and practice in Python,

موضوع: پایتون (زبان برنامه نویسی کامپیوتر) - Python (Computer program language)

موضوع: یادگیری تقویتی - Reinforcement learning

موضوع: فراگیری ماشینی - Machine learning

موضوع: شبکه های عصبی (کامپیوتر) - هوش مصنوعی - Artificial intelligence - Neural networks (Computer science)

شناسه افزوده: کنگ، وا لون، Keng, Wah Loon

شناسه افزوده: اسماعیلی، مهدی، ۱۳۵۰ - مترجم

شناسه افزوده: پیرمحمدیانی، روحیار، ۱۳۶۵ - مترجم

شناسه افزوده: بیداریان، همتا، ۱۳۶۱ - مدیر هنری

رده بندی کنگره

رده بندی دیویی

شماره کتابشناسی ملی

Q۳۲۵/۶

۰۰۶/۳۱

۹۷۴۳۱۳۷

# فهرست مطالب

مقدمه.....	۱۳
پیشگفتار.....	۱۵
فصل ۱: مقدمه‌ای بر یادگیری تقویتی.....	۱۹
۱-۱ یادگیری تقویتی.....	۱۹
۱-۲ یادگیری تقویتی به‌عنوان MDP.....	۲۵
۱-۳ توابع یادگیرنده در یادگیری تقویتی.....	۳۰
۱-۴ الگوریتم یادگیری تقویتی ژرف.....	۳۳
۱-۴-۱ الگوریتم‌های مبتنی بر سیاست.....	۳۳
۱-۴-۲ الگوریتم‌های مبتنی بر ارزش.....	۳۵
۱-۴-۳ الگوریتم‌های مبتنی بر مدل.....	۳۵
۱-۴-۴ روش‌های ترکیبی.....	۳۷
۱-۴-۵ الگوریتم‌های پوشش داده‌شده در این کتاب.....	۳۸
۱-۴-۶ الگوریتم‌های برسیاست و برون‌سیاست.....	۳۹
۱-۴-۷ خلاصه.....	۴۰
۱-۵ یادگیری ژرف برای یادگیری تقویتی.....	۴۰
۱-۶ یادگیری تقویتی و یادگیری باناظر.....	۴۳
۱-۶-۱ نبود مقادیر صحیح هدف.....	۴۳
۱-۶-۲ تنگ بودن بازخورد.....	۴۴
۱-۶-۳ داده‌های تولیدشده در طول آموزش.....	۴۵
۱-۷ خلاصه.....	۴۶

## بخش ۱: الگوریتم‌های مبتنی برسیاست و مبتنی بر ارزش

فصل ۲: الگوریتم REINFORCE.....	۴۹
۲-۱ سیاست.....	۵۰

۵۱	۲-۲ تابع هدف
۵۲	۲-۳ گرادیان سیاست
۵۳	۲-۳-۱ به‌دست آوردن گرادیان سیاست
۵۶	۲-۴ نمونه‌گیری مونت کارلو
۵۷	۲-۵ الگوریتم REINFORCE
۵۸	۲-۵-۱ بهبود الگوریتم REINFORCE
۶۰	۲-۶ پیاده‌سازی REINFORCE
۶۰	۲-۶-۱ پیاده‌سازی کمینه REINFORCE
۶۳	۲-۶-۲ ساختن سیاست‌ها با PyTorch
۶۶	۲-۶-۳ نمونه‌گیری عمل‌ها
۶۷	۲-۶-۴ محاسبه تابع زیان سیاست
۶۸	۲-۶-۵ حلقه آموزش REINFORCE
۶۹	۲-۶-۶ حافظه برسیاست
۷۳	۲-۷ آموزش یک عامل REINFORCE
۷۶	۲-۸ نتایج تجربی
۷۶	۲-۸-۱ بررسی اثر عامل تنزیل $\gamma$
۷۹	۲-۸-۲ بررسی تأثیر مینا
۸۱	۲-۹ خلاصه
۸۱	۲-۱۰ برای مطالعه بیشتر
۸۱	۲-۱۱ تاریخچه

### فصل ۳: SARSA ۸۳

۸۴	۳-۱ تابع Q و تابع V
۸۷	۳-۲ یادگیری اختلاف زمانی
۹۱	۳-۲-۱ شهود برای یادگیری اختلاف زمانی
۹۸	۳-۳ انتخاب عمل در SARSA
۱۰۰	۳-۳-۱ کاوش و بهره‌برداری
۱۰۱	۳-۴ الگوریتم SARSA
۱۰۲	۳-۴-۱ الگوریتم برسیاست
۱۰۳	۳-۵ پیاده‌سازی SARSA
۱۰۴	۳-۵-۱ تابع عمل: شبه‌حریصانه

۱۰۵	..... ۳-۵-۲ محاسبه تابع زیان Q
۱۰۶	..... ۳-۵-۳ حلقه آموزش SARSA
۱۰۸	..... ۳-۵-۴ حافظه تکرار دسته‌ای برسیاست
۱۱۰	..... ۳-۶ آموزش یک عامل SARSA
۱۱۳	..... ۳-۷ نتایج تجربی
۱۱۴	..... ۳-۷-۱ بررسی اثر نرخ یادگیری
۱۱۶	..... ۳-۸ خلاصه
۱۱۷	..... ۳-۹ منابع بیشتر
۱۱۷	..... ۳-۱۰ تاریخچه

#### فصل ۴: شبکه‌های Q ژرف ..... ۱۱۹

۱۲۰	..... ۴-۱ یادگیری تابع Q در DQN
۱۲۲	..... ۴-۲ انتخاب عمل در DQN
۱۲۵	..... ۴-۲-۱ سیاست بولتزمان
۱۲۸	..... ۴-۳ تکرار تجربه
۱۳۰	..... ۴-۳ الگوریتم DQN
۱۳۲	..... ۴-۵ پیاده‌سازی DQN
۱۳۲	..... ۴-۵-۱ محاسبه تابع زیان Q
۱۳۳	..... ۴-۵-۲ حلقه آموزش DQN
۱۳۴	..... ۴-۵-۳ حافظه تکرار
۱۳۸	..... ۴-۶ آموزش یک عامل DQN
۱۴۲	..... ۴-۷ نتایج تجربی
۱۴۲	..... ۴-۷-۱ تأثیر معماری شبکه
۱۴۴	..... ۴-۸ خلاصه
۱۴۴	..... ۴-۹ بیشتر بخوانیم
۱۴۵	..... ۴-۱۰ تاریخچه

#### فصل ۵: بهبود DQN ..... ۱۴۷

۱۴۸	..... ۵-۱ شبکه‌های هدف
۱۵۱	..... ۵-۲ DQN دوگانه
۱۵۵	..... ۵-۳ تکرار تجربه اولویت‌دار

۱۵۷	..... ۵-۳-۱ نمونه‌گیری مهم‌ترها
۱۵۹	..... ۵-۴ پیاده‌سازی DQN اصلاح‌شده
۱۵۹	..... ۵-۴-۱ مقداردهی اولیه شبکه
۱۶۰	..... ۵-۴-۲ محاسبه زیان Q
۱۶۲	..... ۵-۴-۳ به‌روزرسانی شبکه هدف
۱۶۳	..... ۵-۴-۴ DQN با شبکه‌های هدف
۱۶۳	..... ۵-۴-۵ DQN دوگانه
۱۶۴	..... ۵-۴-۶ تکرار تجربه اولویت‌دار
۱۷۱	..... ۵-۵ آموزش یک عامل DQN برای بازی‌های آتاری
۱۷۷	..... ۵-۶ نتایج تجربی
۱۷۸	..... ۵-۶-۱ اثر DQN دوگانه و PER
۱۸۲	..... ۵-۷ خلاصه
۱۸۳	..... ۵-۸ بیشتر بخوانیم

## بخش ۲: روش‌های ترکیبی

۱۸۷	..... فصل ۴: الگوریتم‌های عملگر-نقاد منفعت‌گرا (A2C)
۱۸۸	..... ۶-۱ عملگر
۱۸۹	..... ۶-۲ نقاد
۱۸۹	..... ۶-۲-۱ تابع منفعت
۱۹۵	..... ۶-۲-۲ یادگیری تابع منفعت
۱۹۶	..... ۶-۳ الگوریتم A2C
۱۹۸	..... ۶-۴ پیاده‌سازی A2C
۱۹۹	..... ۶-۴-۱ تخمین منفعت
۲۰۲	..... ۶-۴-۲ محاسبه تابع زیان ارزش و تابع زیان سیاست
۲۰۳	..... ۶-۴-۳ حلقه آموزش عملگر-نقاد
۲۰۴	..... ۶-۵ معماری شبکه
۲۰۶	..... ۶-۶ آموزش یک عامل A2C
۲۰۶	..... ۶-۶-۱ اجرای A2C با برگشت‌های n مرحله‌ای روی Pong
۲۱۱	..... ۶-۶-۲ اجرای A2C با GAE روی Pong
۲۱۲	..... ۶-۶-۳ اجرای A2C با برگشت‌های n مرحله‌ای روی BipedalWalker

۲۱۵	۶-۷ نتایج تجربی
۲۱۵	۶-۷-۱ اثر برگشت‌های n مرحله‌ای
۲۱۸	۶-۷-۲ اثر $\lambda$ در روش GAE
۲۲۰	۶-۸ خلاصه
۲۲۰	۶-۹ مطالعه بیشتر
۲۲۱	۶-۱۰ تاریخچه

## فصل ۷: بهینه‌سازی سیاست مجاور (PPO)..... ۲۲۳

۲۲۴	۷-۱ هدف جایگزین
۲۲۴	۷-۱-۱ فروپاشی عملکرد
۲۲۷	۷-۱-۲ اصلاح تابع هدف
۲۳۴	۷-۲ بهینه‌سازی سیاست مجاور (PPO)
۲۳۹	۷-۳ الگوریتم PPO
۲۴۱	۷-۴ پیاده‌سازی PPO
۲۴۱	۷-۴-۱ محاسبه زیان سیاست PPO
۲۴۳	۷-۴-۲ حلقه آموزش PPO
۲۴۴	۷-۵ آموزش یک عامل PPO
۲۴۴	۷-۵-۱ PPO روی Pong
۲۴۹	۷-۵-۲ PPO روی BipedalWalker
۲۵۲	۷-۶ نتایج تجربی
۲۵۲	۷-۶-۱ اثر $\lambda$ از GAE
۲۵۴	۷-۶-۲ اثر متغیر $\epsilon$ برش
۲۵۶	۷-۷ خلاصه
۲۵۶	۷-۸ مطالعه بیشتر

## فصل ۸: موازی‌سازی روش‌ها..... ۲۵۹

۲۶۰	۸-۱ موازی‌سازی همگام
۲۶۱	۸-۲ موازی‌سازی ناهمگام
۲۶۳	۸-۲-۱ Hogwild!
۲۶۶	۸-۳ آموزش یک عامل A3C
۲۷۰	۸-۴ خلاصه

۲۷۰..... ۸-۵ مطالعه بیشتر .....

فصل ۹: مروری بر الگوریتم‌ها ..... ۲۷۱

### بخش ۳: جزئیات عملی

فصل ۱۰: به‌کارگیری یادگیری تقویتی ژرف ..... ۲۷۷

۱۰-۱ روش‌های مهندسی نرم‌افزار ..... ۲۷۷

۱۰-۱-۱ آزمون‌های واحد ..... ۲۷۸

۱۰-۱-۲ کیفیت کُد ..... ۲۸۵

۱۰-۱-۳ گردش کار Git ..... ۲۸۶

۱۰-۲ نکاتی درباره اشکال‌زدایی ..... ۲۸۹

۱۰-۲-۱ علائم حیات ..... ۲۸۹

۱۰-۲-۲ تشخیص گرادیان سیاست ..... ۲۹۰

۱۰-۲-۳ تشخیص داده‌ها ..... ۲۹۱

۱۰-۲-۴ پیش‌پردازنده ..... ۲۹۳

۱۰-۲-۵ حافظه ..... ۲۹۳

۱۰-۲-۶ توابع الگوریتمیک ..... ۲۹۳

۱۰-۲-۷ شبکه‌های عصبی ..... ۲۹۴

۱۰-۲-۸ ساده‌سازی الگوریتم ..... ۲۹۷

۱۰-۲-۹ ساده‌سازی مسئله ..... ۲۹۸

۱۰-۲-۱۰ آپرپارامترها ..... ۲۹۸

۱۰-۲-۱۱ گردش کار ..... ۲۹۸

۱۰-۳ ترندهایی برای محیط‌های Atari ..... ۳۰۰

۱۰-۴ یادگیری تقویتی ژرف و آپرپارامترها ..... ۳۰۴

۱۰-۴-۱ جداول آپرپارامترها ..... ۳۰۴

۱۰-۴-۲ مقایسه عملکرد الگوریتم‌ها ..... ۳۰۸

۱۰-۵ خلاصه ..... ۳۱۱

## فصل ۱۱: SLM Lab ..... ۳۱۳

- ۳۱۳ ..... ۱۱-۱ الگوریتم‌های پیاده‌سازی‌شده در SLM Lab
- ۳۱۶ ..... ۱۱-۲ فایل spec
- ۳۱۸ ..... ۱۱-۲-۱ جست‌وجو در فایل spec
- ۳۲۲ ..... ۱۱-۳ اجرای SLM Lab
- ۳۲۲ ..... ۱۱-۳-۱ دستورات SLM Lab
- ۳۲۳ ..... ۱۱-۴ تحلیل نتایج تجربی
- ۳۲۴ ..... ۱۱-۴-۱ مروری بر داده‌های Experiment
- ۳۲۶ ..... ۱۱-۵ خلاصه

## فصل ۱۲: معماری‌های شبکه ..... ۳۲۷

- ۳۲۷ ..... ۱۲-۱ انواع شبکه‌های عصبی
- ۳۲۸ ..... ۱۲-۱-۱ پرسپترون چند لایه
- ۳۳۰ ..... ۱۲-۱-۲ شبکه‌های عصبی کانولوشن (CNNs)
- ۳۳۲ ..... ۱۲-۱-۳ شبکه‌های عصبی برگشتی (RNNs)
- ۳۳۴ ..... ۱۲-۲ رهنمون‌هایی برای انتخاب یک نوع شبکه
- ۳۳۴ ..... ۱۲-۲-۱ MDP در مقابل POMDP
- ۳۳۸ ..... ۱۲-۲-۲ انتخاب شبکه‌ها برای محیط‌ها
- ۳۴۲ ..... ۱۲-۳ Net API
- ۳۴۴ ..... ۱۲-۳-۱ استنباط شکل لایه ورودی و خروجی
- ۳۴۶ ..... ۱۲-۳-۲ ساخت خودکار شبکه
- ۳۵۰ ..... ۱۲-۳-۳ گام آموزش
- ۳۵۱ ..... ۱۲-۳-۴ ارائه متدهای رایج مرتبط
- ۳۵۲ ..... ۱۲-۴ خلاصه
- ۳۵۳ ..... ۱۲-۵ برای مطالعه بیشتر

## فصل ۱۳: سخت‌افزار ..... ۳۵۵

- ۳۵۵ ..... ۱۳-۱ کامپیوتر
- ۳۶۲ ..... ۱۳-۲ انواع داده‌ها
- ۳۶۵ ..... ۱۳-۳ بهینه‌سازی انواع داده‌ها در یادگیری تقویتی
- ۳۷۰ ..... ۱۳-۴ انتخاب سخت‌افزار

۱۳-۵ خلاصه ..... ۳۷۱

## بخش ۴: طراحی محیط

فصل ۱۴: حالت‌ها ..... ۳۷۵

۱۴-۱ مثال‌هایی از حالت‌ها ..... ۳۷۶

۱۴-۲ کامل بودن حالت ..... ۳۸۴

۱۴-۳ پیچیدگی حالت ..... ۳۸۵

۱۴-۴ از دست دادن اطلاعات حالت ..... ۳۹۰

۱۴-۴-۱ مقیاس‌بندی خاکستری تصویر ..... ۳۹۱

۱۴-۴-۲ گسسته‌سازی ..... ۳۹۱

۱۴-۴-۳ تصادم درهم‌سازی ..... ۳۹۲

۱۴-۴-۴ از دست دادن فراطلاعات ..... ۳۹۳

۱۴-۵ پیش‌پردازش ..... ۳۹۷

۱۴-۵-۱ استانداردسازی ..... ۳۹۸

۱۴-۵-۲ پیش‌پردازش تصویر ..... ۴۰۰

۱۴-۵-۳ پیش‌پردازش زمانی ..... ۴۰۲

۱۴-۶ خلاصه ..... ۴۰۶

فصل ۱۵: عمل‌ها ..... ۴۰۹

۱۵-۱ چند مثال از عمل ..... ۴۰۹

۱۵-۲ کامل بودن عمل ..... ۴۱۳

۱۵-۳ پیچیدگی عمل ..... ۴۱۵

۱۵-۴ خلاصه ..... ۴۲۱

۱۵-۵ مطالعه بیشتر: طراحی عمل در موضوعات روزانه ..... ۴۲۱

فصل ۱۶: یاداشتها ..... ۴۲۷

۱۶-۱ نقش یاداشتها ..... ۴۲۷

۱۶-۲ رهنمون‌هایی برای طراحی یاداش ..... ۴۲۹

۱۶-۳ خلاصه ..... ۴۳۵

۴۳۷	..... فصل ۱۷: تابع انتقال
۴۳۸	..... ۱۷-۱ امکان‌سنجی
۴۴۰	..... ۱۷-۲ بررسی واقعیت
۴۴۳	..... ۱۷-۳ خلاصه
۴۴۵	..... سخن پایانی
۴۴۹	..... پیوست الف: جدول زمانی یادگیری تقویتی ژرف
۴۵۱	..... پیوست ب: چند محیط نمونه
۴۵۲	..... ب-۱- محیط‌های گسسته
۴۵۳	..... ب-۱-۱ محیط CartPole-v0
۴۵۴	..... ب-۱-۲ محیط MountainCar-v0
۴۵۴	..... ب-۱-۳ محیط LunarLander-v2
۴۵۶	..... ب-۱-۴ محیط PongNoFrameskip-v4
۴۵۷	..... ب-۱-۵ محیط BreakoutNoFrameskip-v4
۴۵۷	..... ب-۲- محیط‌های پیوسته
۴۵۸	..... ب-۲-۱ محیط Pendulum-v0
۴۵۸	..... ب-۲-۲ محیط BipedalWalker-v2