

دانشکده مهندسی برق

یادگیری تقویتی در کنترل

تمرین هشتم: LQR مبتنی بر Actor-Critic

استاد: دکتر سعید شمقدری

دانشجو: سیده ستاره خسروی

زمستان 1403

چکیده

در تمرین سری هشتم یادگیری تقویتی در کنترل با 2 سوال از مبحث کنترل بهینه، یادگیری تقویتی مبتنی بر نقاد و بازیگر مواجه هستیم، که در هر فصل به سوال و یا سوالات مطرح شده پاسخ داده شده است.

واژه‌های کلیدی: یادگیری تقویتی، کنترل بهینه، نقاد، بازیگر

# فهرست مطالب

عنوان صفحه

[فهرست مطالب ‌ب](#_Toc188540075)

[فهرست تصاویر و نمودارها ‌ج](#_Toc188540076)

[فصل 1: کنترل بهینه مبتنی بر نقاد و بازیگر 1](#_Toc188540077)

[1‌.1‌ مقدمه 1](#_Toc188540078)

[1‌.2‌ سوال اول............. 1](#_Toc188540079)

[1‌.3‌ سوال دوم 7](#_Toc188540080)

# فهرست تصاویر و نمودارها

عنوان صفحه

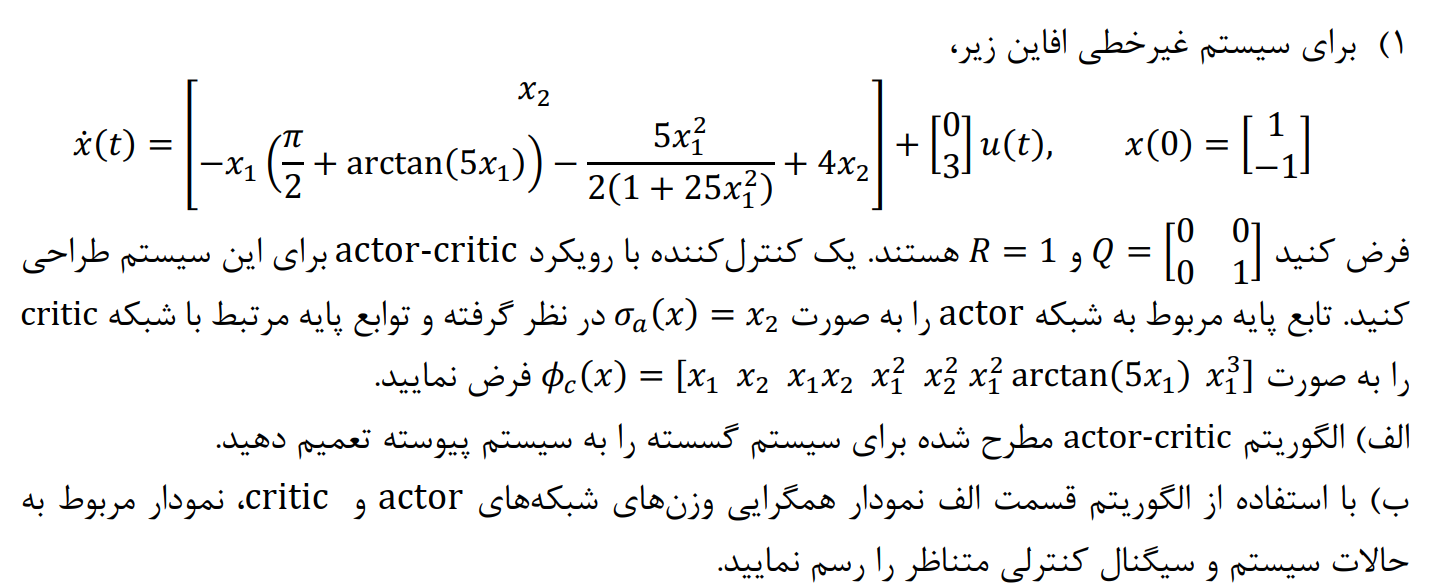
[شکل 1: نمودار حالت‌ها و سیگنال کنترلی 7](#_Toc188540092)

# کنترل بهینه مبتنی بر نقاد و بازیگر

## مقدمه

در این فصل به 2 سوال مربوط به بادگیری تقویتی و حل مسئله کنترل بهینه مبتنی بر Actor Critic پاسخ داده می‌شود.

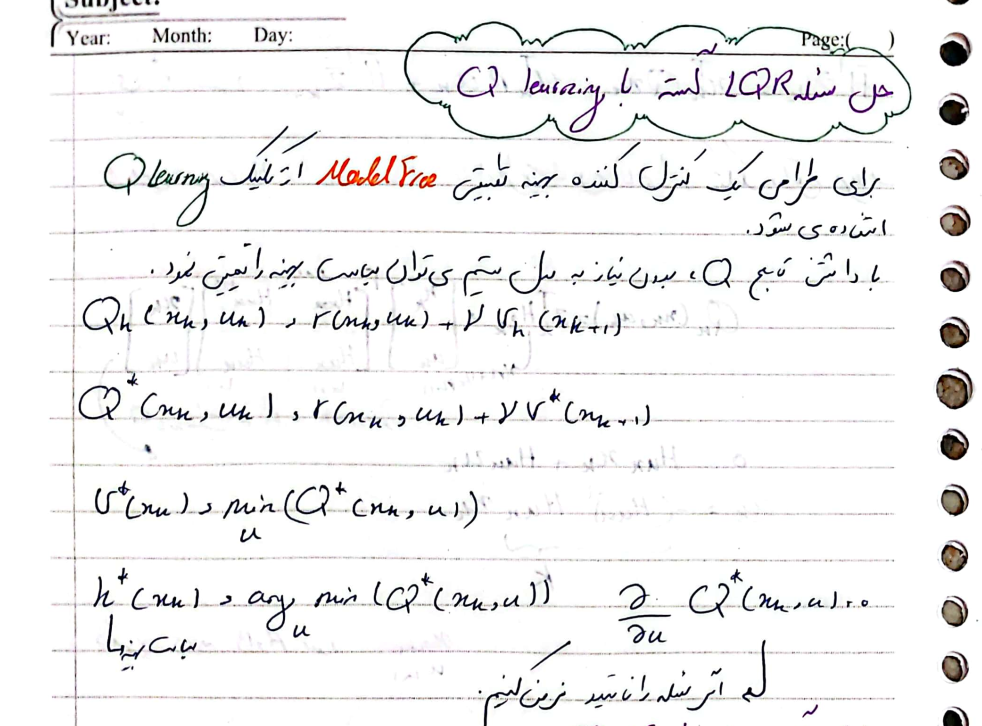
## سوال اول

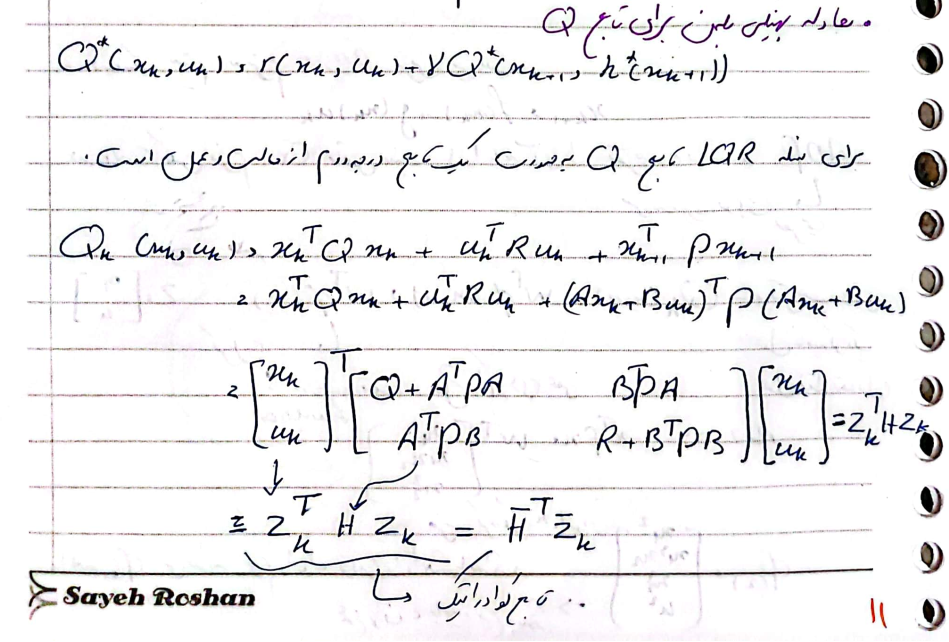
صورت سوال:

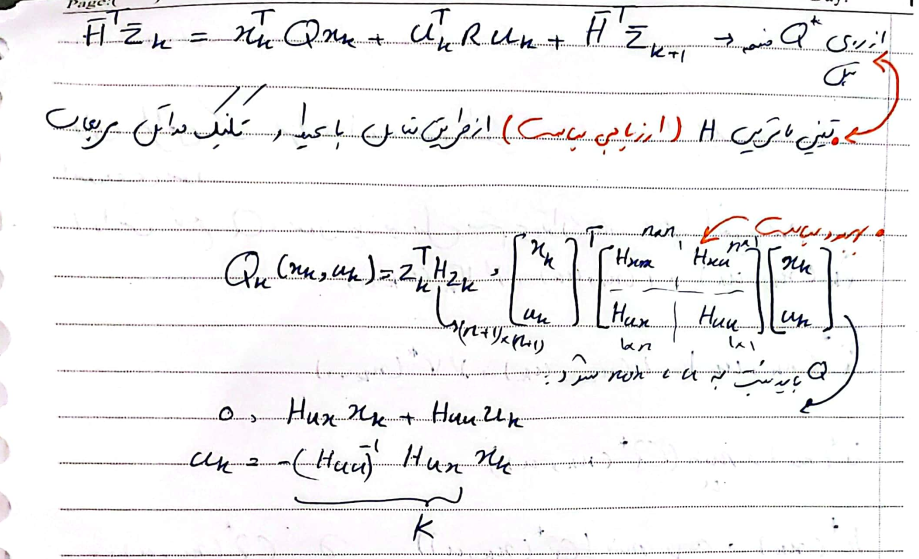
پاسخ بخش‌های الف و ب:

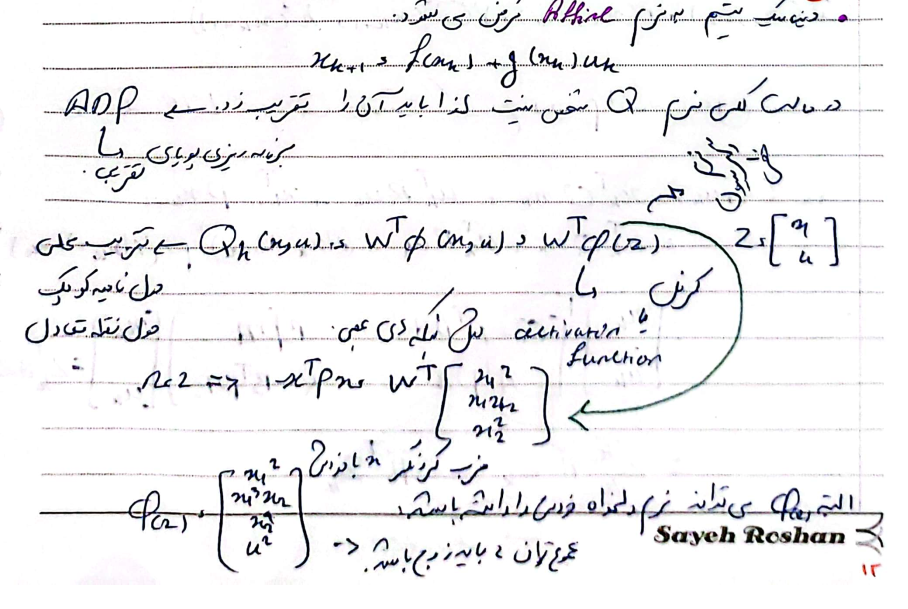
در این بخش برای حل سوال چون نتواستیم از روش مطرح شده در کلاس برای حل سوال پیش برویم، از روش Actor Critic مبتنی بر Q learning استفاده کردیم. ابتدا این الگوریتم را شرح داده، تعمیم آن به پیوسته را مطرح کرده و کد آن را می‌نویسیم.

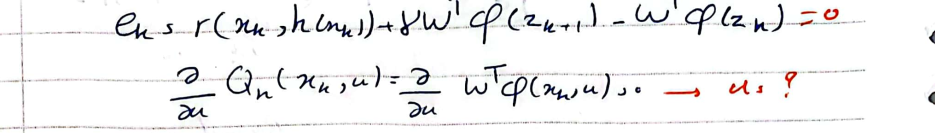
می‌دانیم که برای یک سیستم گسسته Q- learning به صورت زیر بدست می‌آید.

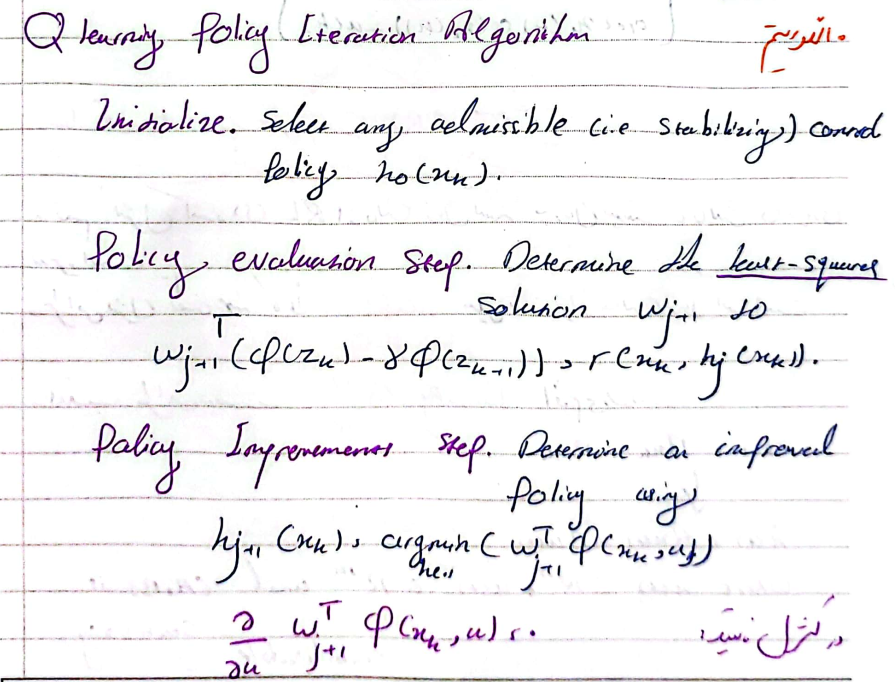


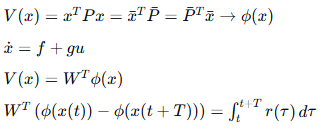
در ادامه داریم:

در نهایت بحث بهبود سیاست نیز به صورت زیر می‌گردد.

حال اگر این سیستم گسسته غیرخطی اما به صورت افاین باشد، داریم:

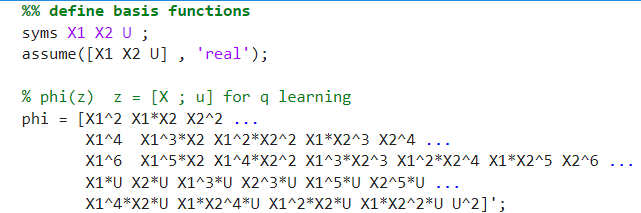


که نهایتا الگوریتم آن به صورت زیر خواهد بود:

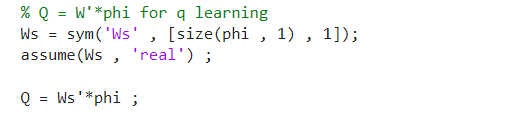
برای سیستم گسسته زمان که دیدیم به چه صورت می‌شود. در پیوسته زمان نیز مشابه قبل است با این تفاوت:

که البته کدنویسی آن مشابه حالت گسسته است.

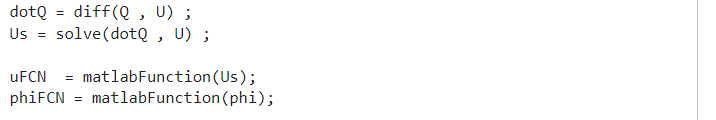
برای کدنویسی باید دقت شود، برخلاف روش ارائه شده در کلاس، در این روش باید وزن‌های مورد استفاده، برحسب u هم باشند، بعلاوه اینکه هر سطر باید توان زوج از حالت‌ها و ورودی‌ها باشد. به همین دلیل از توابع پایه‌ای صورت سوال نمی‌توانیم استفاده کنیم، و توابع دیگری را باید لحاظ کنیم.

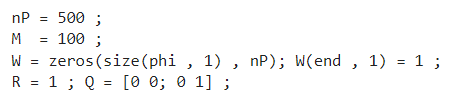
به صورت زیر توابع پایه‌ای را تشکیل می‌دهیم.

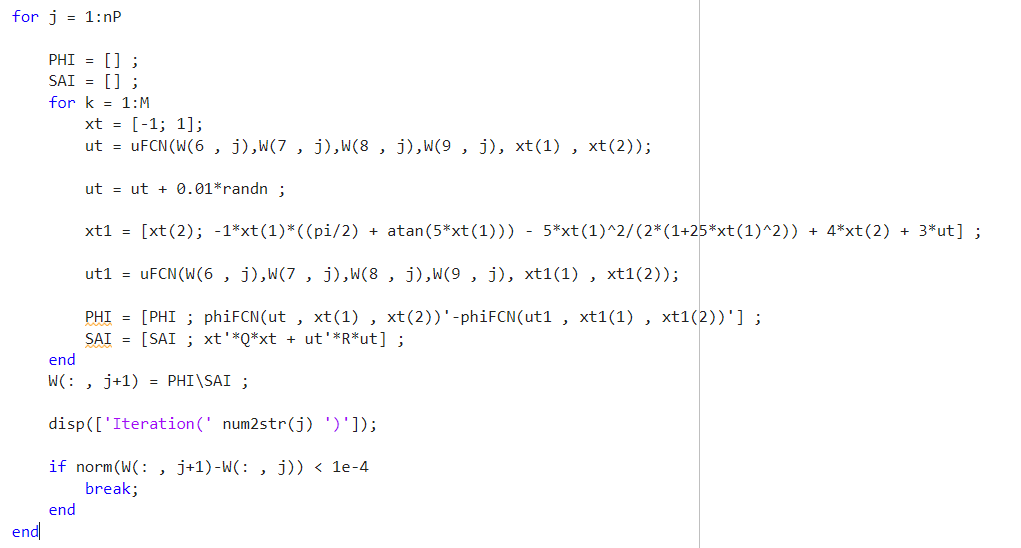
چون قرار است از استراتژی Q Learning برای حل استفاده کنیم، باید بدانیم که از فرمت z باید استفاده شود، و به همین ترتیب در تابع پایه‌ای باید U هم وجود داشته باشد، چون باید از Q نسبت به U مشتق بگیریم تا بتوانیم مسئله را حل کنیم.

در ادامه به تعداد مقادیر توابع پایه‌ای نیاز به وزن‌های حقیقی است، این وزن‌ها را تعریف و تابع Q را تشکیل می‌دهیم.

سپس می‌دانیم که با استفاده از مشتق Q بر حسب U مقدار بهینه بدست می‌آید، که به صورت زیر مشتق گرفته و مقادیر بهینه را نیز بدست آورده و ذخیره می‌کنیم و به گونه‌ای ذخیره سازی را تغییر می‌دهیم که محاسبات سریعتر انجام شود.



تنظیمات آموزش را به صورت زیر تعیین می‌کنیم:

سپس در ادامه مطابق الگوریتمی که در صفحه چهار آوردیم، آموزش را کدنویسی کردیم.

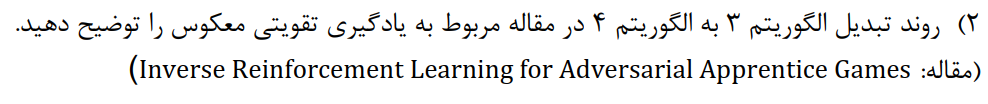
دینامیک سیستم و شرایط اولیه را درون باکس‌ها مشخص کردیم.

پس از انجام آموزش، نتایج به صورت زیر است:

## سوال دوم

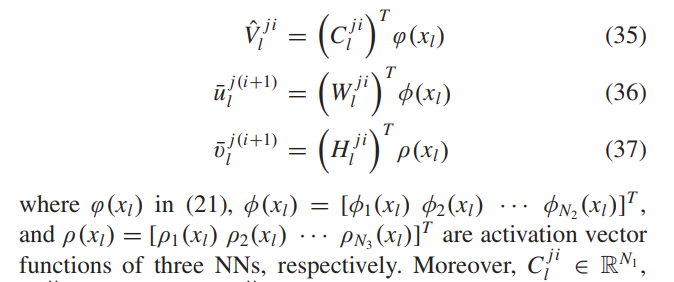
شکل 1: نمودار حالت‌ها و سیگنال کنترلی

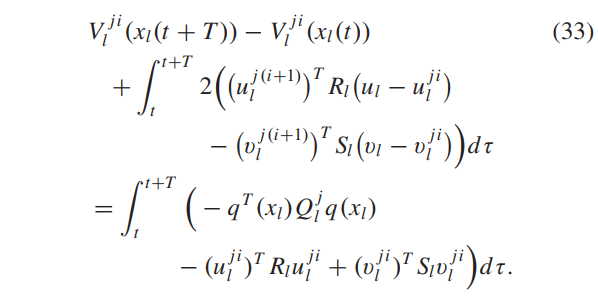
**صورت سوال:**

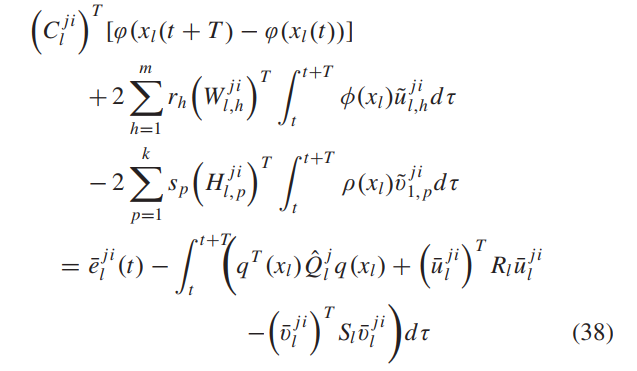


**پاسخ:**

اولین گام، تقریب توابع ارزش با استفاده از شبکه‌های عصبی است، به این صورت که برای تابع ارزش، ورودی کنترلی و ورودی مخرب، به صورت زیر تقریب می‌زنیم.

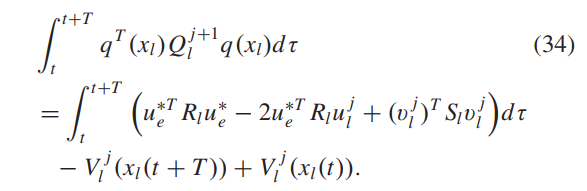


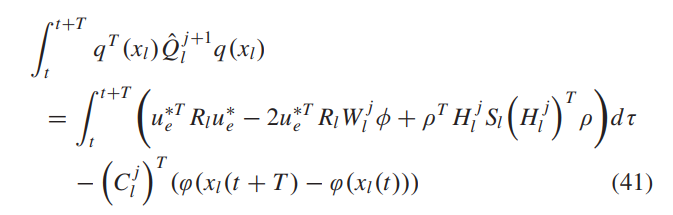
در ادامه، معادله بلمن اصلی در الگوریتم 3 گام چهارم، که معادله 33 می‌باشد:

با جایگزینی تقریب‌های شبکه‌ عصبی به معادله 38 تبدیل می‌گردد:

در اینجا انتگرال‌ها و رابطه زیر:

از طریق داده‌های مشاهده شده، محاسبه می‌شوند. روش BLS نیز، وزن‌ها را با لحاظ یک معادله خطی از روی داده‌های مشاهده شده حل می‌کند.

سپس به روزرسانی وزن جریمه حالت، یا درواقع معادله 34 در الگوریتم 3:

با جایگزینی خروجی‌های شبکه عصبی که در 41 آمده تبدیل می‌شود:

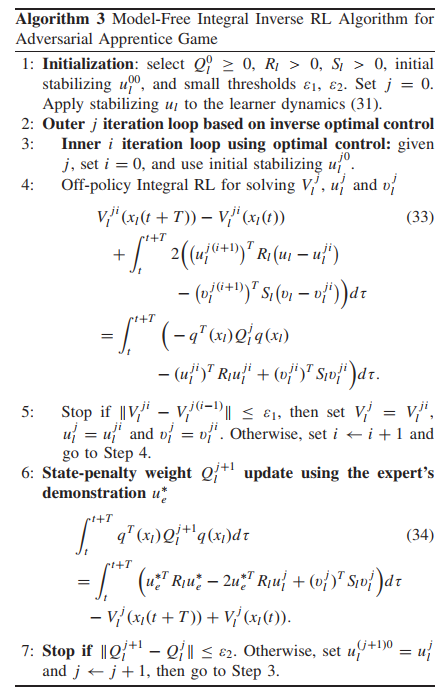
وزن‌های Q نیز، از طریق رابطه 43 بدست می‌آیند:

باید به میزان زیر:

نقطده داده جمع شود تا ماتریس‌های مورد استفاده در بدست آوردن Q، معکوس پذیر شوند.

از اینتروال‌های زمانی کوچک T باید جهت گسسته سازی استفاده شود.

در نهایت الگوریتم 3:



به الگوریتم 4 تبدیل می‌شود:

