

دانشکده مهندسی برق

یادگیری تقویتی در کنترل

تمرین ششم: یادگیری تقویتی در کنترل بهینه

استاد: دکتر سعید شمقدری

دانشجو: سیده ستاره خسروی

زمستان 1403

چکیده

در تمرین سری ششم یادگیری تقویتی در کنترل با 3 سوال از مبحث کنترل بهینه مواجه هستیم، که در هر فصل به سوال و یا سوالات مطرح شده پاسخ داده شده است.

واژه‌های کلیدی: یادگیری تقویتی، کنترل بهینه

# فهرست مطالب

عنوان صفحه

[فهرست مطالب ‌ب](#_Toc185707989)

[فهرست تصاویر و نمودارها ‌ج](#_Toc185707990)

[فصل 1:کنترل بهینه 1](#_Toc185707991)

[1‌.1‌ مقدمه 1](#_Toc185707992)

[1‌.2‌ سوال اول........ 1](#_Toc185707993)

[1‌.3‌ سوال دوم........ 13](#_Toc185707994)

[1‌.4‌سوال سوم......... 15](#_Toc185707995)

# فهرست تصاویر و نمودارها

عنوان صفحه

[شکل 1: سیگنال u و متغیرهای حالت پس از اعمال پالیسی حاصل از idare 2](#_Toc185884383)

[شکل 2: بهبود سیاست در PI 4](#_Toc185884384)

[شکل 3: نمودار همگرایی در PI 5](#_Toc185884385)

[شکل 4: روند تغییرات K در VI 7](#_Toc185884386)

[شکل 5: همگرایی در VI 8](#_Toc185884387)

[شکل 6: مقادیر K در لاندای 0.5 10](#_Toc185884388)

[شکل 7: مقادیر K در لاندای 0.9 11](#_Toc185884389)

[شکل 8: همگرایی در لاندای 0.5 11](#_Toc185884390)

[شکل 9: همگرایی در لاندای 0.9 12](#_Toc185884391)

[شکل 10: تایع مقدار 16](#_Toc185884392)

[شکل 11: تابع پاداش 16](#_Toc185884393)

[شکل 12: تابع ارزش 16](#_Toc185884394)

[شکل 13: تابع ارزش 16](#_Toc185884395)

[شکل 14: جواب استاندارد مسئله LQT 17](#_Toc185884396)

[شکل 15: سیگنال کنترلی 17](#_Toc185884397)

[شکل 16: فرم کوادراتیک تابع ارزش 17](#_Toc185884398)

[شکل 17: بردار حالت افزوده 17](#_Toc185884399)

[شکل 18: فضای حالت سیستم 17](#_Toc185884400)

[شکل 19: دینامیک ردیابی 18](#_Toc185884401)

[شکل 20: سیستم افزوده 18](#_Toc185884402)

[شکل 21: فرم کوادراتیک تابع ارزش 18](#_Toc185884403)

[شکل 22: ساده سازی 18](#_Toc185884404)

[شکل 23: سیگنال کنترلی 18](#_Toc185884405)

[شکل 24: معادله لیاپانوف 19](#_Toc185884406)

[شکل 25: ضریب سیگنال کنترلی 19](#_Toc185884407)

[شکل 26: همیلتونین 19](#_Toc185884408)

[شکل 27: مشتق همیلتونین 19](#_Toc185884409)

[شکل 28: جواب مسئله 19](#_Toc185884410)

[شکل 29: معادله ریکاتی 19](#_Toc185884411)

[شکل 30: تابع ارزش 20](#_Toc185884412)

[شکل 31: تابع ارزش حالت عمل 20](#_Toc185884413)

[شکل 32: تابع ارزش حالت عمل 20](#_Toc185884414)

[شکل 33: تابع ارزش حالت عمل 20](#_Toc185884415)

[شکل 34: فرم نهایی تابع ارزش حالت عمل 20](#_Toc185884416)

[شکل 35: ماتریس H 21](#_Toc185884417)

[شکل 36: درایه‌های ماتریس H 21](#_Toc185884418)

[شکل 37: نعریف جدید برای تبدیل به least square 21](#_Toc185884419)

[شکل 38: تابع ارزش حالت عمل 21](#_Toc185884420)

[شکل 39: تعریف جدید تابع ارزش حالت عمل 21](#_Toc185884421)

[شکل 40: معادله نهایی 22](#_Toc185884422)

[شکل 41: بهبود سیاست 22](#_Toc185884423)

[شکل 42: ارزیابی سیاست 22](#_Toc185884424)

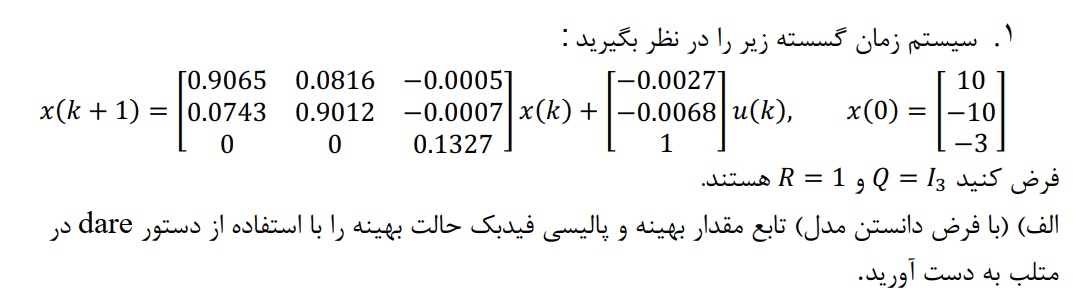
[شکل 43: رابطه کلی 22](#_Toc185884425)

# کنترل بهینه

## مقدمه

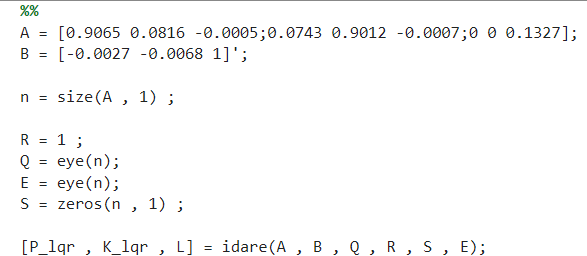
در این فصل به 3 سوال مربوط به این فصل پاسخ داده می‌شود.

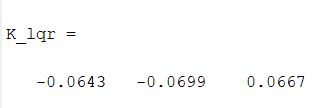
## سوال اول

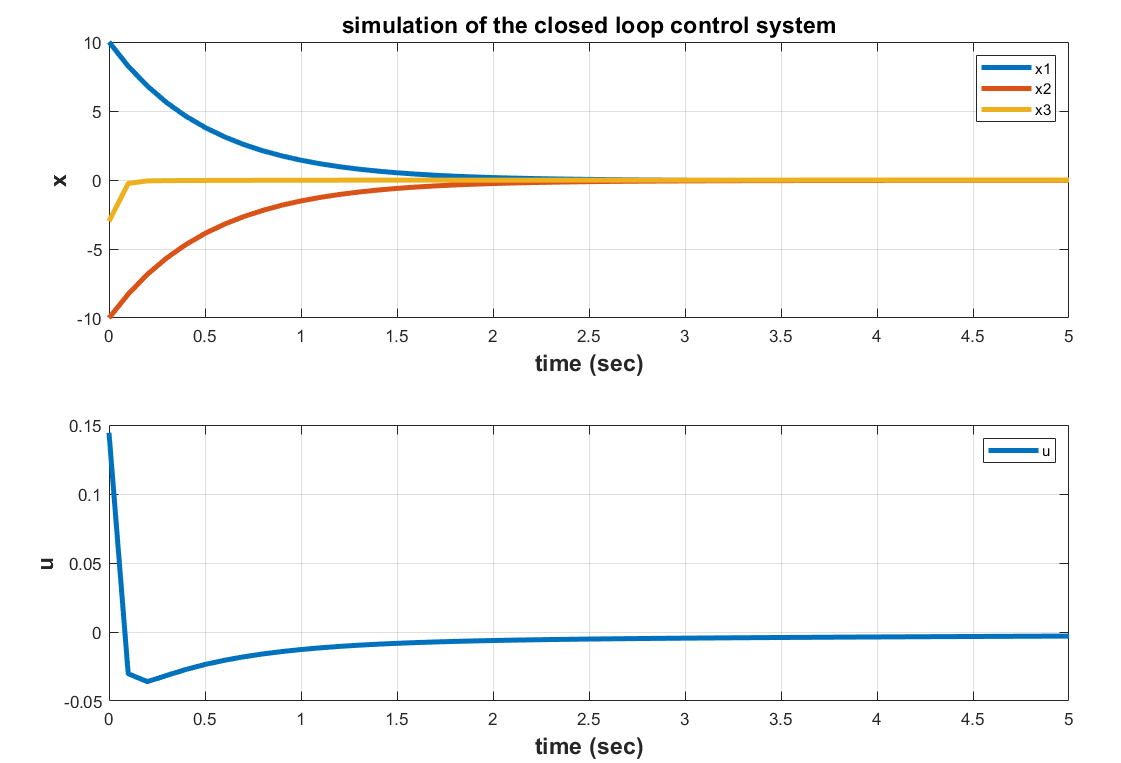
صورت سوال قسمت الف:

پاسخ:

در این قسمت با توجه به پیشنهاد خود وبسایت MATLAB، از دستور idare بجای dare استفاده کردیم.

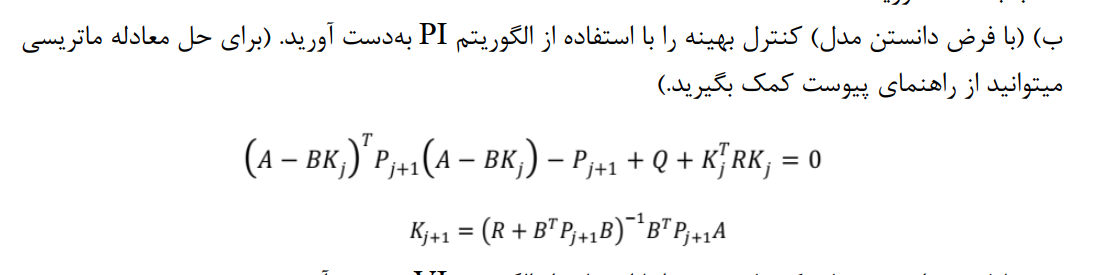
برای حل این سوال ابتدا دینامیک مسئله مطابق زیر تعریف گردید:

در انتهای کد نیز با دستور مذکور مقدار K و P حاصل بدست آمد که K حاصل از LQR برابر است با:

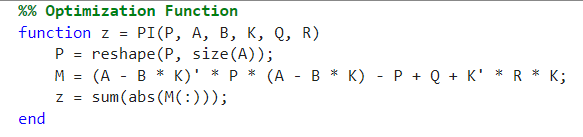
در ادامه‌ی کد سوال اول بخش الف، سیگنال‌ها و متغیرهای حالت را درحضور این پالیسی بدست آمده رسم کردیم که به صورت زیر است:

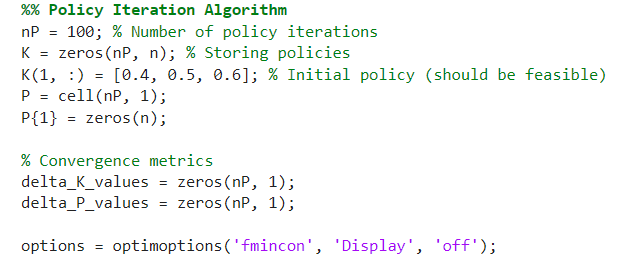
شکل 1: سیگنال u و متغیرهای حالت پس از اعمال پالیسی حاصل از idare

صورت سوال قسمت ب:

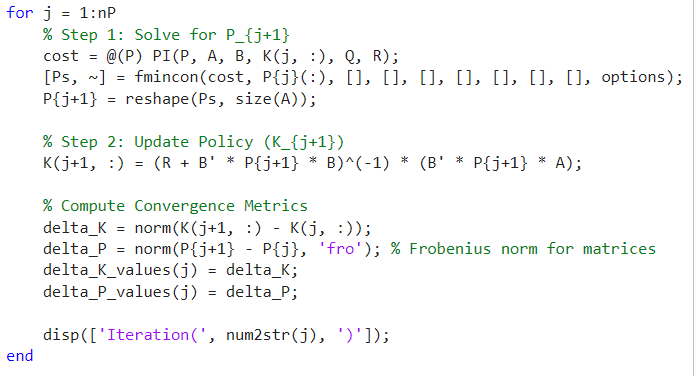


پاسخ:

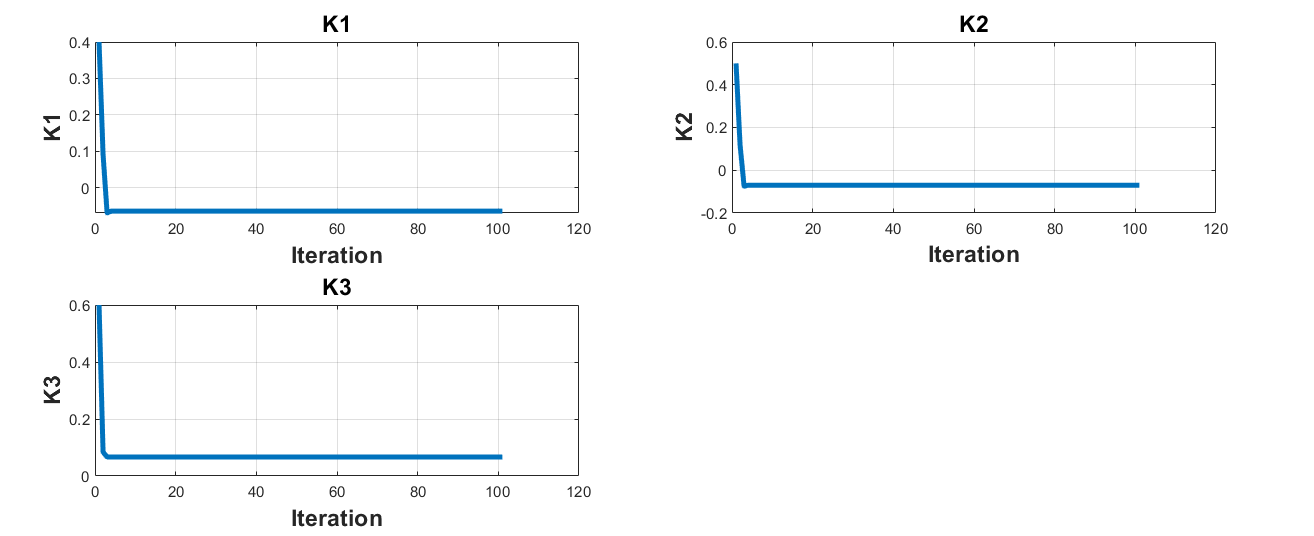
برای حل این قسمت ابتدا تابعی تعریف کردیم که با استفاده از آن مقدار P را بدست آوریم. در اینجا از بهینه سازی استفاده می‌کنیم. فرض می‌کنیم معادله‌‌ای که قرار است از روی آن P محاسبه شود (معادله ریکاتی صورت سوال)، برابر است با M و لازم است این M برابر با صفر باشد فلذا یک P باید انتخاب شود که این M را کمینه کند. پیاده سازی این تابع به صورت زیر است:

حال با لحاظ دینامیک سیستم، تنظیمات الگوریتم را به صورت زیر قرار می‌دهیم:

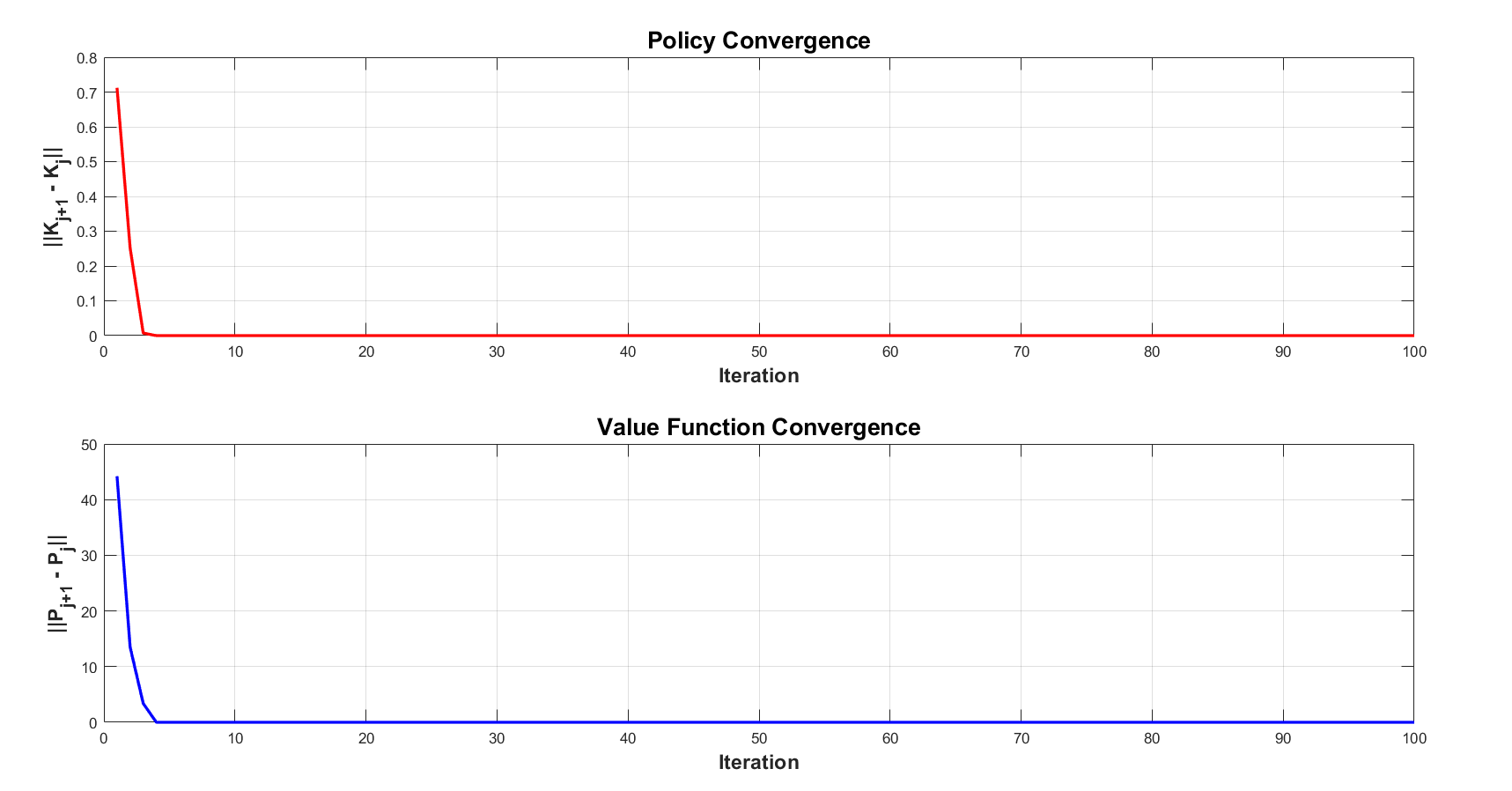
می‌خواهیم در 100 تکرار مسئله حل شود، تمامی سیاست‌های بدست آمده را نیز می‌خواهیم در K ذخیره کنیم. برای اینکه الگوریتم استفاده شود، یک سیاست اولیه نیز مطابق کد بالا تعریف می‌کنیم. فقط باید حواسمان باشد که این سیاست اولیه شدنی باشد. در نهایت برای ذخیره سازی Pها نیز متغیر P که قرار است سلولی باشد و هر سلول آن شامل 9 پارامتر باشد تعریف می‌گردد. برای رسم نموداری که در بخش آخر سوال اول خواسته شده نیز لازم است متغیرهایی تعریف کنیم. تنظیمات بهینه سازی را نیز باید لحاظ کنیم. این تنظیمات مربوط به fmincon است که برای حل تابع بهینه سازی که نوشتیم به کار می‌رود.

حلقه الگوریتم نیز به صورت زیر است، ابتدا در هر تکرار P را بدست می‌آوریم و سپس سیاست را بهبود می‌دهیم.

نتیجه به صورت زیر است:

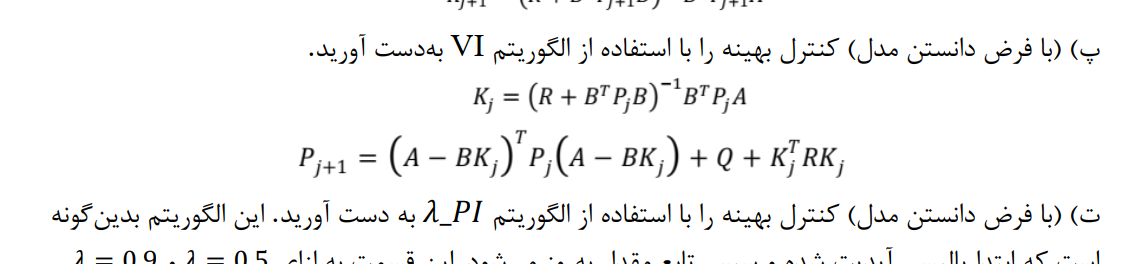
در طی تکرارهای مختلف نمودار سیاست به صورت زیر خواهد بود:

شکل 2: بهبود سیاست در PI

نمودارهای خواسته شده در بخش آخر این سوال نیز به صورت زیر می‌گردد:

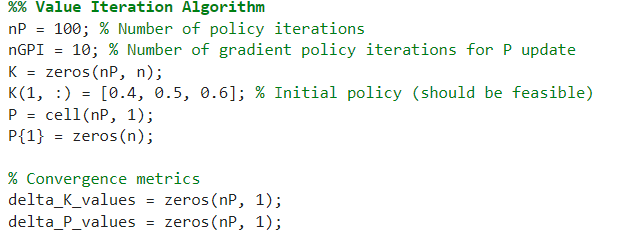
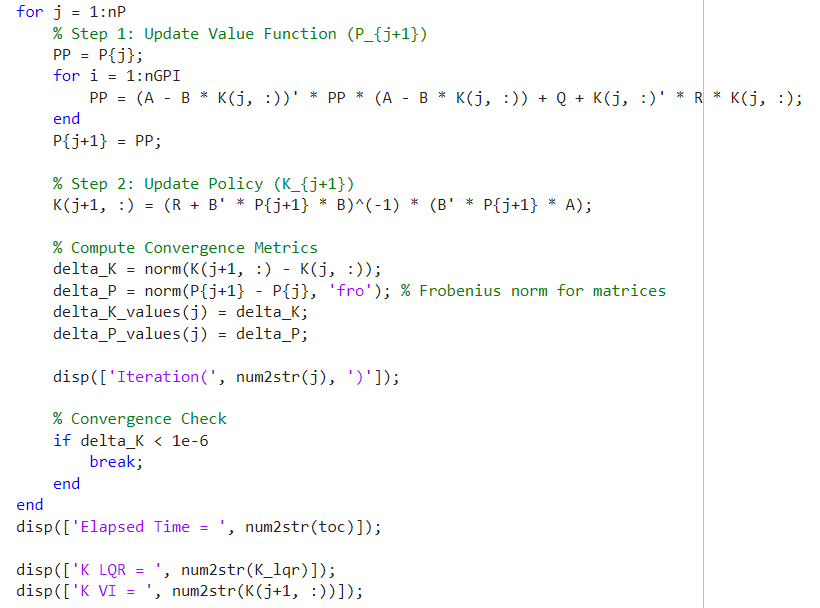
شکل 3: نمودار همگرایی در PI

نکته قابل توجه این است که نیاز نیست حتما 100 تکرار انجام شود، الگوریتم در تعداد تکرارهای کمتر نیز به همگرایی می‌رسد.

صورت سوال قسمت پ:

پاسخ:

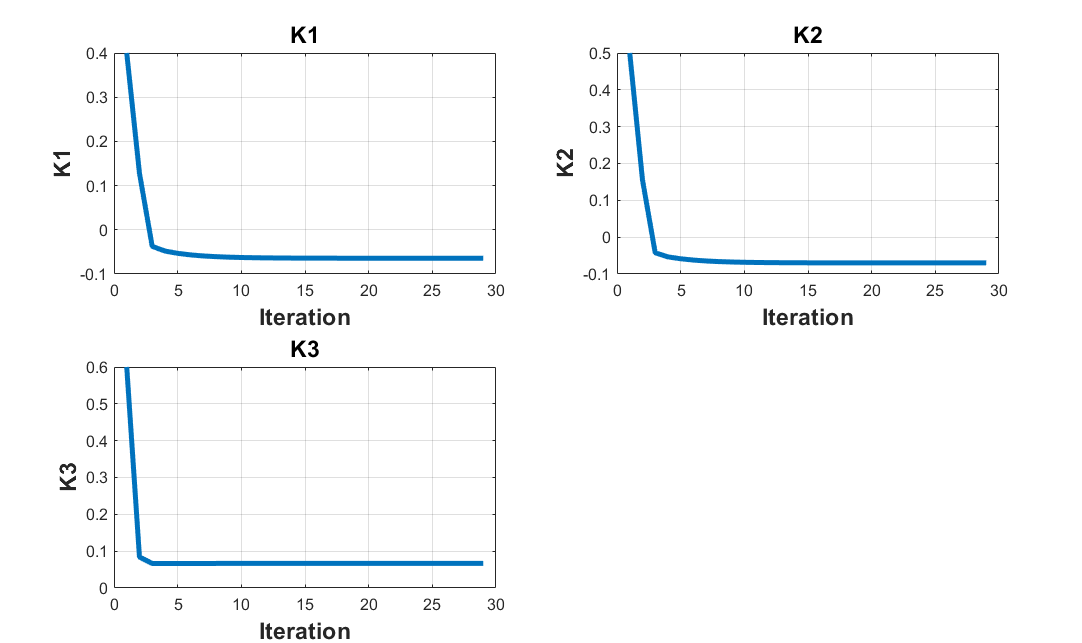
در این بخش نیز کد مشابه قبل است فقط تنظیمات به صورت زیر تغییر می‌کند، در اینجا برای به روز رسانی P از تکرار سیاست تعمیم یافته نیز استفاده می‌کنیم، سایر تنظیمات مشابه قبل است.

در ادامه حلقه‌ی تکرار به صورت زیر است:

با استفاده از GPI، تابع ارزش را بروز کرده و P محاسبه می‌گردد، و از این P برای بروز رسانی سیاست استفاده می‌شود. در اینجا برخلاف کد قسمت قبل، شرط توقف را نیز لحاظ کردیم، که دیگر نیازی به انجام تکرارهای بیشتر در صورت همگرایی نباشد.

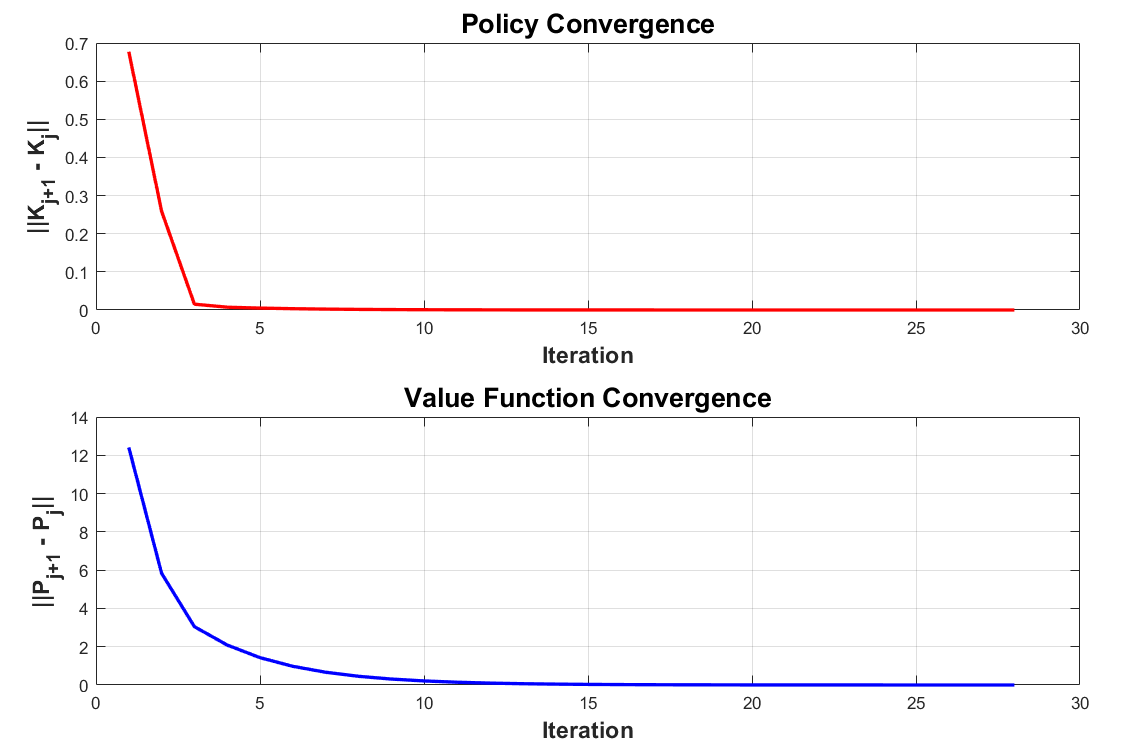
بقیه قسمت‌ها کد توضیح خاصی ندارد و صرفا رسم نمودار است.

سیاست بدست آمده به صورت زیر است:

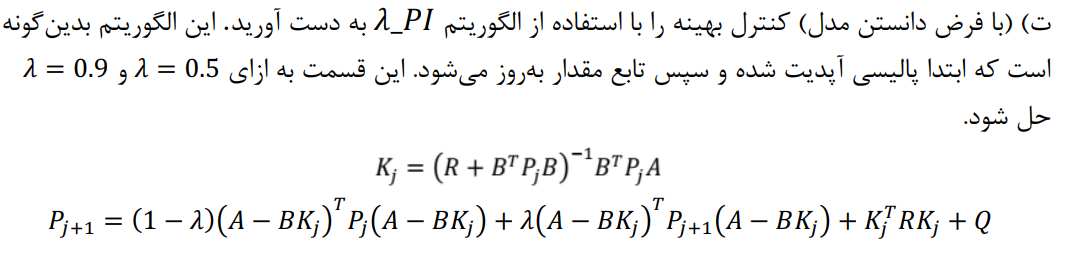
نمودار مقادیر K نیز به صورت زیر خواهد بود، برخلاف قسمت قبلی اینجا تا 28 تکرار انجام شده است.

شکل 4: روند تغییرات K در VI

نمودار خواسته شده در بخش آخر سوال نیز در شکل 5 قابل مشاهده است. در اینجا نیز با استفاده از VI به مقادیر بهینه که توسط idare بدست آمده بود همگرا شدیم.

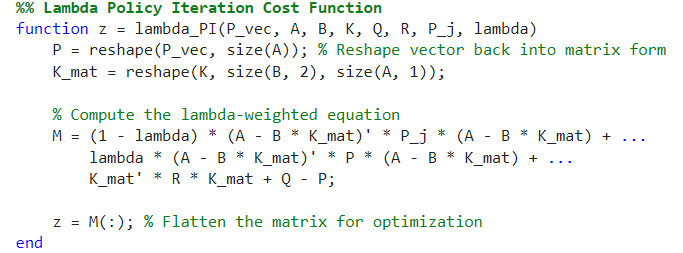
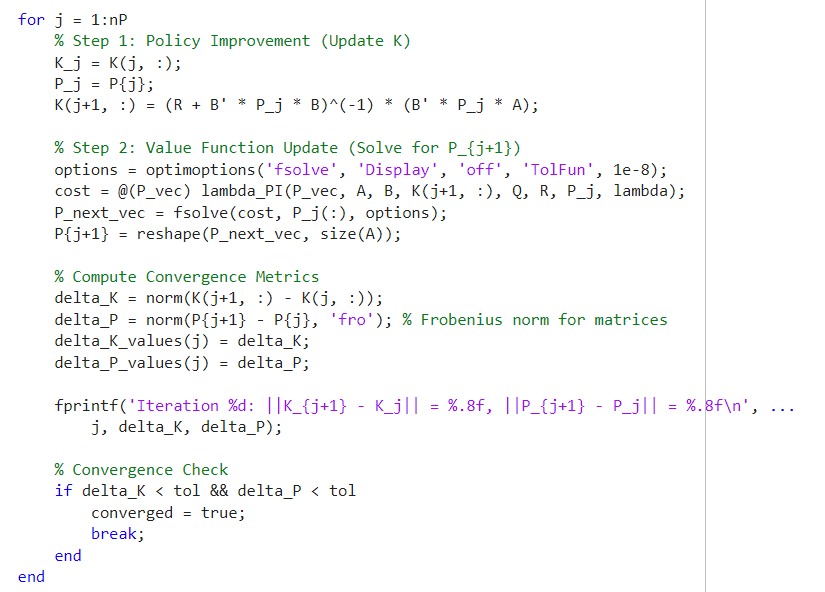
****

شکل 5: همگرایی در VI

**صورت سوال قسمت ت:**

پاسخ:

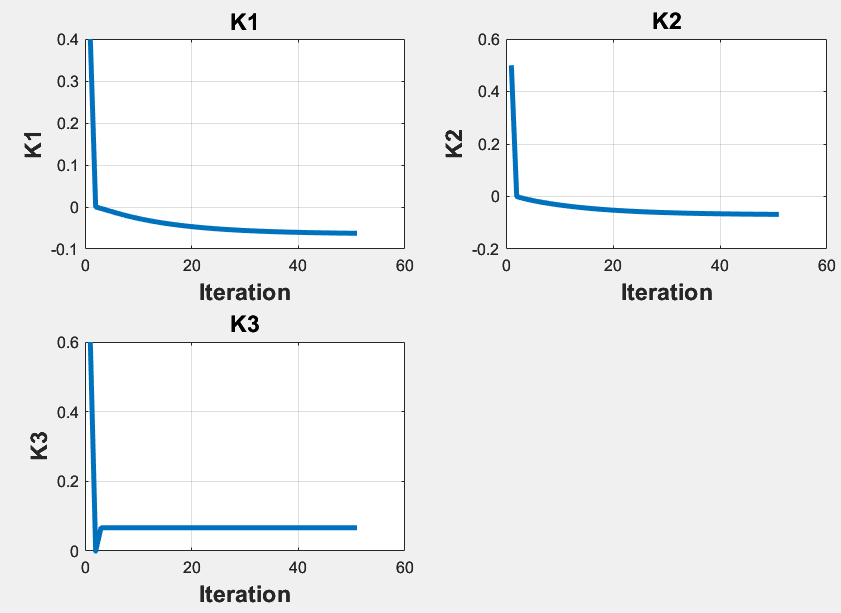
برای پاسخ به این بخش نیز مشابه الگوریتم PI از تابع بهینه سازی استفاده کردیم، فقط متناسب با تغییرات صورت سوال در معادله‌ی ریکاتی تغییرات لازم را ایجاد کردیم که به صورت زیر است:

سایر بخش‌های کد مانند قبل است و فقط حلقه تکرار به صورت زیر می‌گردد:

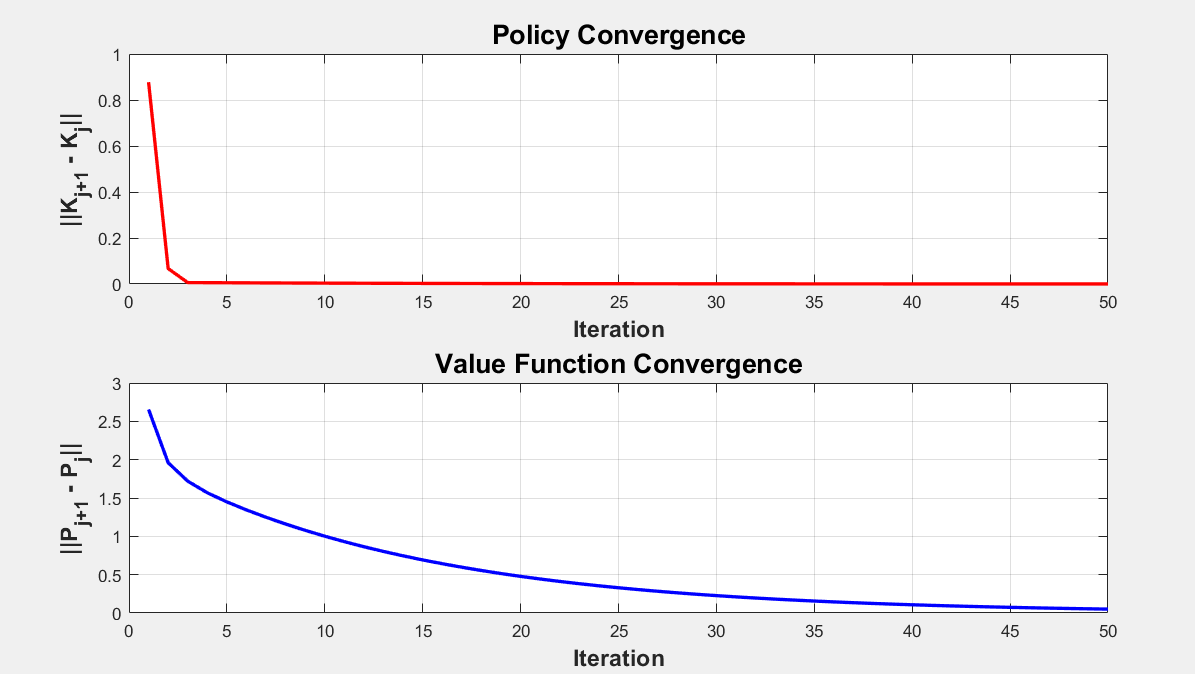
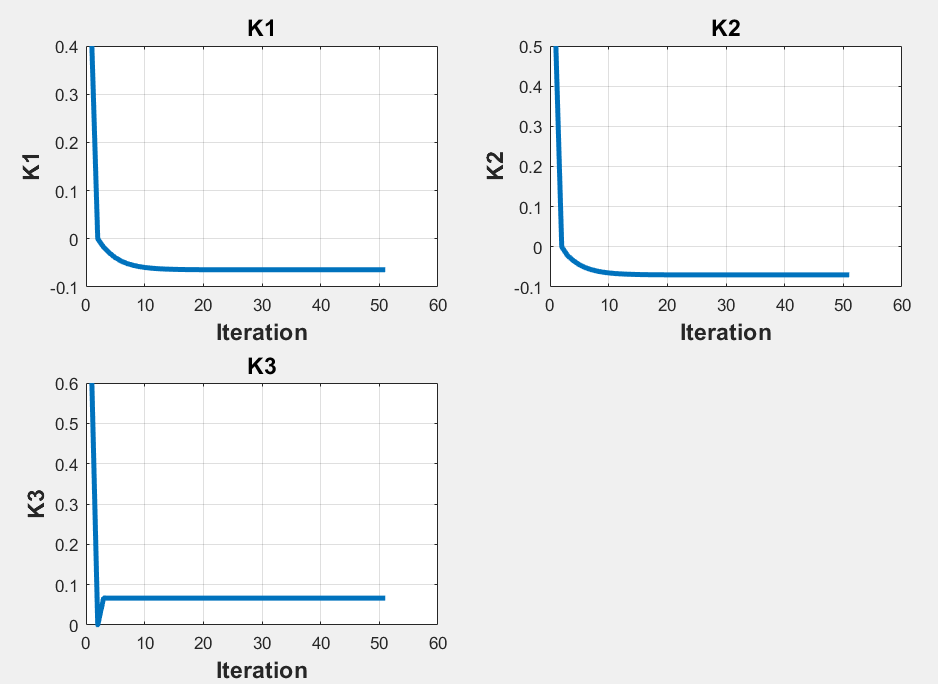
حاصل به ازای مقدار lambda برابر با 0.5 و در 50 تکرار به صورت زیر است:

و برای lambda با مقدار 0.9 می‌شود:

که نشان می‌دهد با این مقدار بهتر به جواب بهینه همگرا می‌شویم تا با اعمال مقدار 0.5.

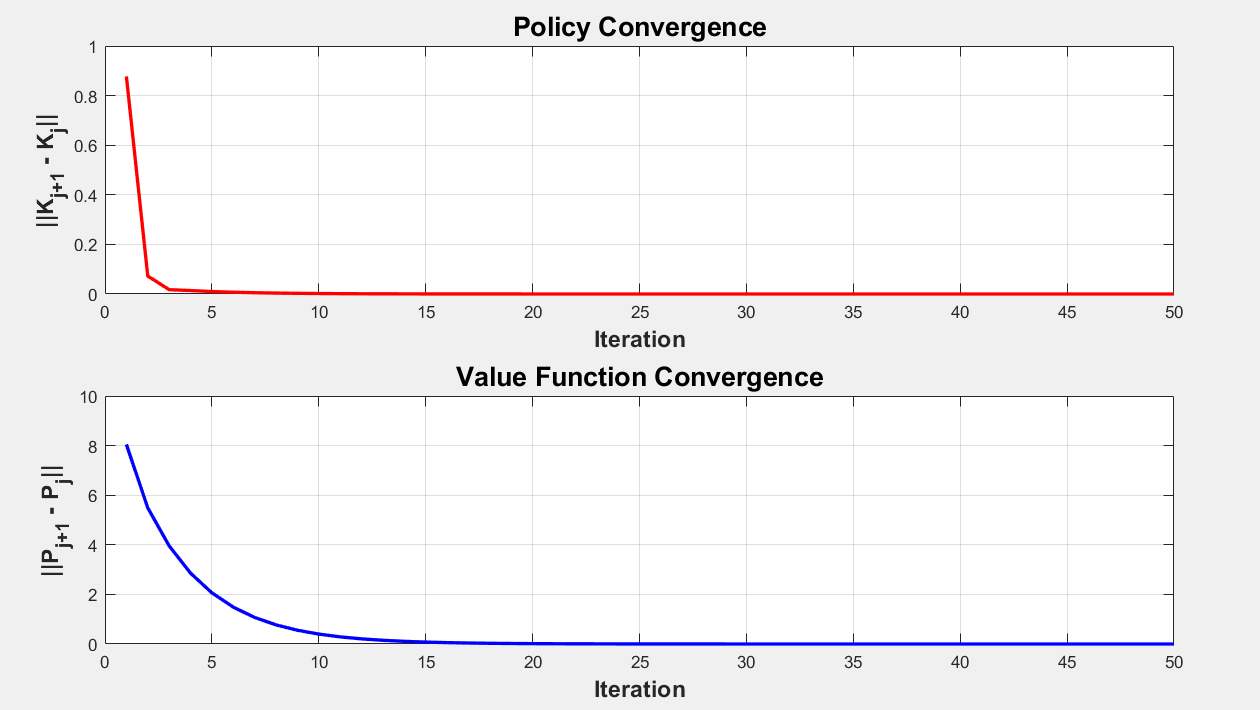
نمودارها نیز به صورت زیر است، برای مقدار لاندای 0.5:

شکل 6: مقادیر K در لاندای 0.5

و سپس برای لاندای 0.9:

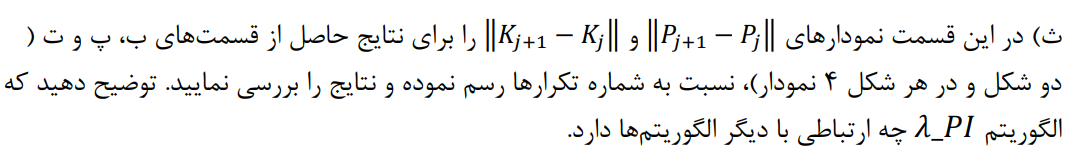
شکل 7: مقادیر K در لاندای 0.9

شکل 8: همگرایی در لاندای 0.5

و برای همگرایی داریم:

شکل 9: همگرایی در لاندای 0.9

همانطور که مشاهده می‌گردد در حالتی که لاندا برابر با 0.9 است، وضعیت همگرایی الگوریتم بهتر است.

**صورت سوال قسمت ث:**

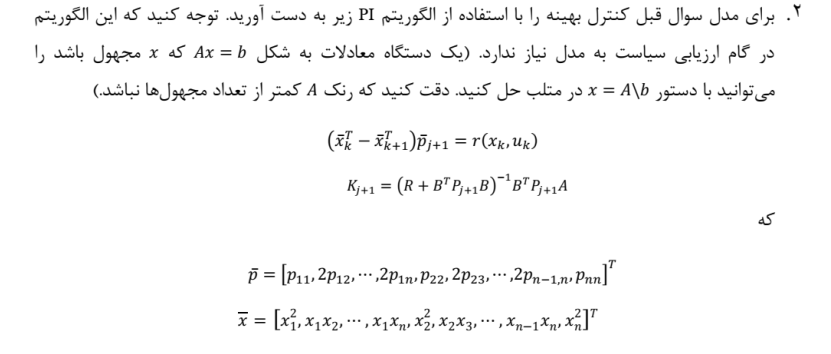
پاسخ:

نمودارهای خواسته شده در بخش‌های قبلی رسم گردید.

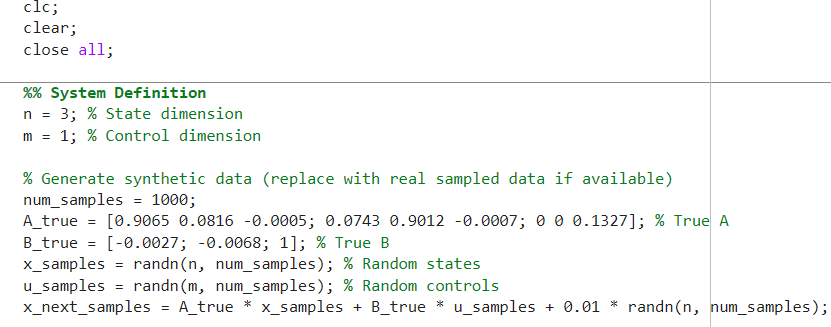
روش lambda-PI با ارائه‌ی پارامتری به نام lambda باعث می‌شود که VI یا PI به شکلی تعمیم یافته تبدیل شوند. که تعیین می‌کند چه مقدار از تابع ارزش فعلی و بعدی در ارزیابی سیاست استفاده شود.

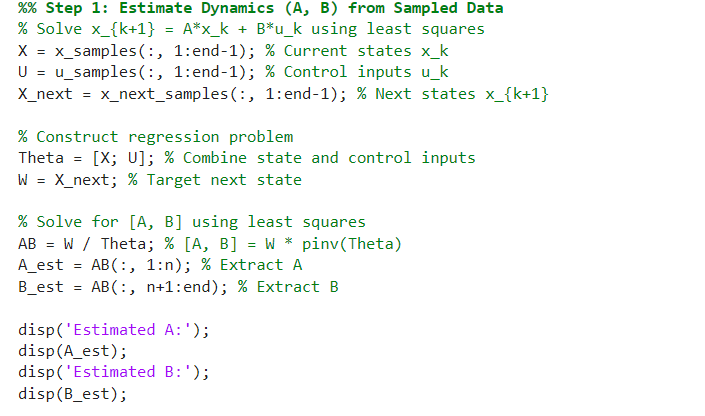
هرچقدر مقدار لاندا کمتر باشد و به صفر نزدیک شود، به سمت VI و هرچقدر مقدار لاندا افزایش یابد و به یک نزدیک شود، الگوریتم به سمت PI، می‌رود. که در نتایج نیز مشهود است. انتظار می‌رود PI همگرایی سریعتر داشته باشد، در اینجا نیز وقتی مقدار لاندا را با 0.9 جایگزین کردیم، همگرایی سریعتر و بهتر شد.

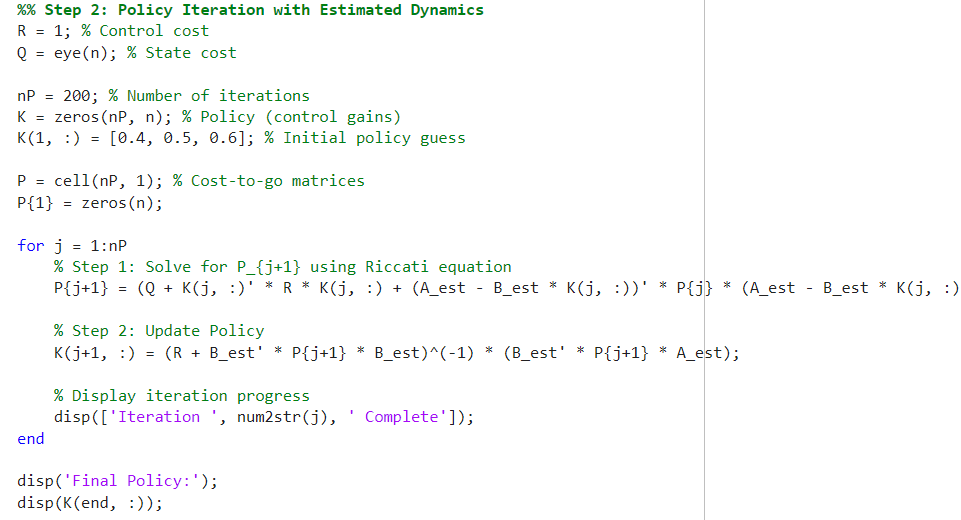
## سوال دوم

صورت سوال:

پاسخ:

در این سوال، برای حل مسئله نیاز داریم داده تولید کنیم. به همین ترتیب با استفاده از مدل اصلی سیستم و کد زیر داده تولید می‌کنیم:

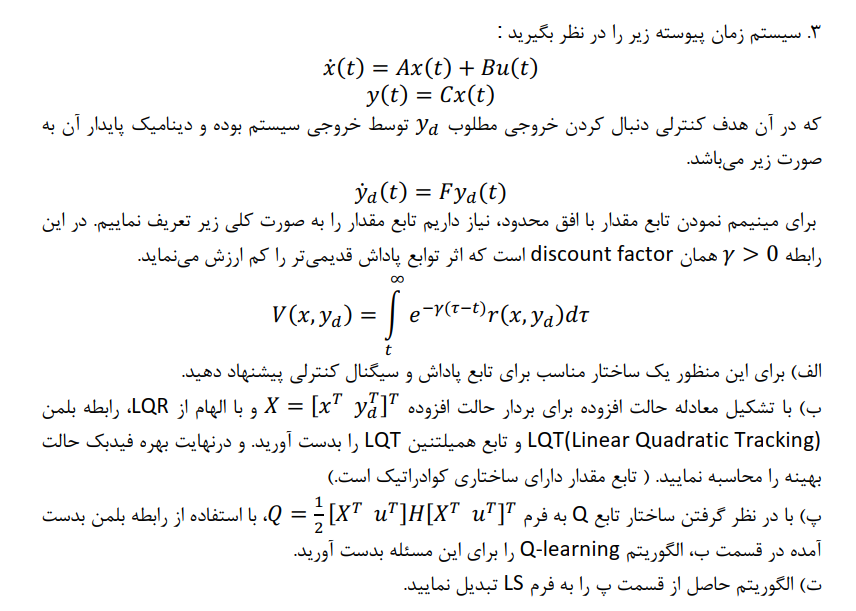
سپس با استفاده از کد بالا، و همانطور که در صورت سوال خواسته شده بود معادله را حل می‌کنیم.

و سپس با استفاده از حل بالا حلقه تکرار را به صورت زیر می‌نویسیم، که در آن از دینامیک تخمین زده شده استفاده می‌گردد:

حاصل پس از 50 تکرار به صورت زیر می‌شود:

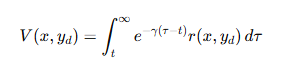


## سوال سوم

**صورت سوال:**

**قسمت الف:**

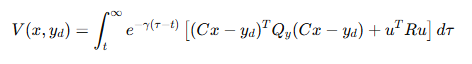
با فرض رابطه مقدار که در صورت سوال آمده است و مطالعه مقاله‌ی پیوست که خودمان یافتیم، داریم:



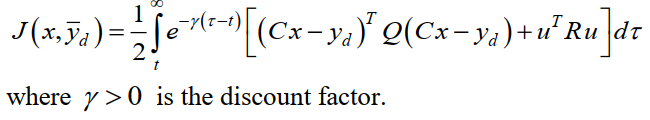
شکل 10: تایع مقدار

می‌توانیم تابع پاداش را با توجه به میزان تلاش کنترلی و خطای ردیابی به صورت زیر تعریف کنیم:

شکل 11: تابع پاداش

**که در آن مثبت معین هستند، Cx نیز خروجی سیستم و خروجی مطلوب است. به همین ترتیب داریم:

شکل 12: تابع ارزش

*البته در مقاله‌ای که یافتیم نیز تعریف زیر موجود است:*

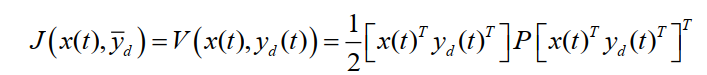
شکل 13: تابع ارزش

*جواب استاندارد مسئله LQT به صورت زیر است:*

شکل 14: جواب استاندارد مسئله LQT

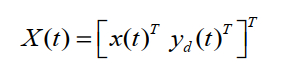
*ولی برای اینکه بتوانیم با لحاظ تابع ارزشی که به آن رسیدیم، این مسئله را به فرم quadratic در بیاوریم سیاست را به صورت زیر تعریف می‌کنیم:*

شکل 15: سیگنال کنترلی

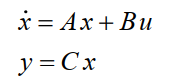
*که این سیگنال admissible fixed control policy است. به همین ترتیب با اثبات ارائه شده در مقاله می‌توان تابع ارزش را به صورت زیر بازنویسی نمود:*

شکل 16: فرم کوادراتیک تابع ارزش

**قسمت پ:**

با تعریف زیر:

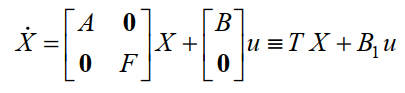
شکل 17: بردار حالت افزوده

و فضای حالت زیر:

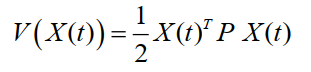
شکل 18: فضای حالت سیستم

*و همچنین:*

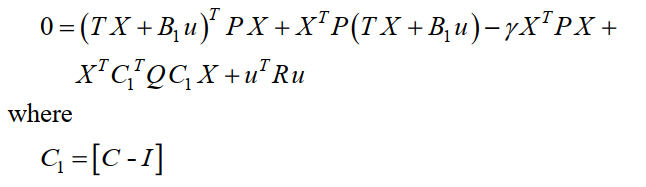
شکل 19: دینامیک ردیابی

*خواهیم داشت:*

شکل 20: سیستم افزوده

*با فرض رابطه کوادراتیک تابع ارزش به صورت زیر:*

شکل 21: فرم کوادراتیک تابع ارزش

*با استفاده از رابطه شکل 21 و قرار دادن آن در سمت چپ رابطه شکل 16، و لحاظ رابطه شکل 20 داریم:*

شکل 22: ساده سازی

*سیگنال کنترلی را نیز به صورت زیر اگر لحاظ کنیم:*

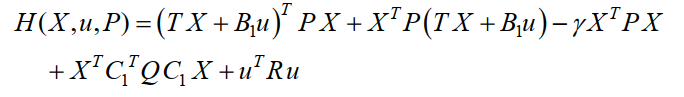
شکل 23: سیگنال کنترلی

*که در آن:*

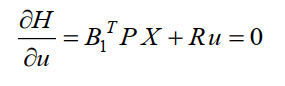
*با قرار دادن تابع شکل 21 و سیگنال 23 در رابطه 22، رابطه بلمن LQT معادله لیاپانوف LQT افزوده را نتیجه می‌دهد:*

شکل 24: معادله لیاپانوف

شکل 25: ضریب سیگنال کنترلی

*براساس معادله شکل 22، رابطه همیلتونین به صورت زیر تعریف می‌شود:*

شکل 26: همیلتونین

*با فرض شکل 23 و مشتق گیری از همیلتونین شکل 26 داریم:*

شکل 27: مشتق همیلتونین

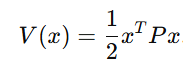
*پاسخ زیر بدست می‌آید:*

شکل 28: جواب مسئله

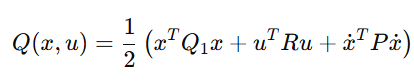
*که P در رابطه زیر صدق می‌کند:*

شکل 29: معادله ریکاتی

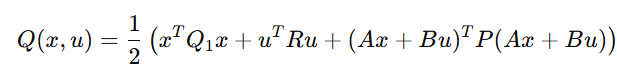
**قسمت پ و ت:**

*مطابق فرضی که داشتیم یعنی:*

شکل 30: تابع ارزش

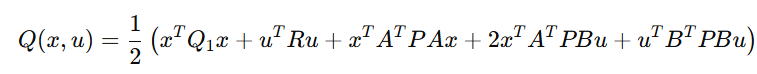
*برای تابع ارزش حالت عمل نیز خواهیم داشت:*

شکل 31: تابع ارزش حالت عمل

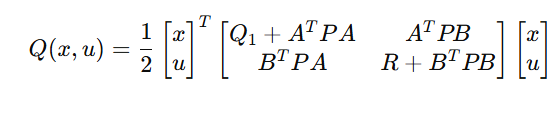
*با لحاظ فضای حالت داریم:*

شکل 32: تابع ارزش حالت عمل

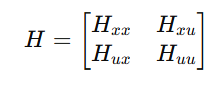
*که رابطه فوق را گسترش می‌دهیم:*

**

شکل 33: تابع ارزش حالت عمل

و سپس داریم:

شکل 34: فرم نهایی تابع ارزش حالت عمل

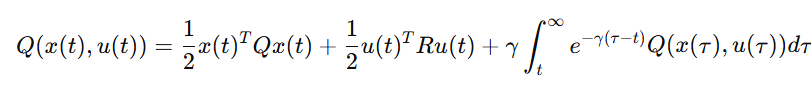
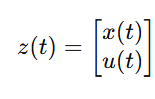
که در آن:

شکل 35: ماتریس H

و درایه‌های آن برابرند با:

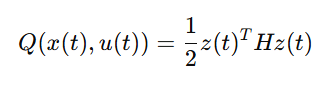
شکل 36: درایه‌های ماتریس H

با فرض زیر:

و تعریف زیر:

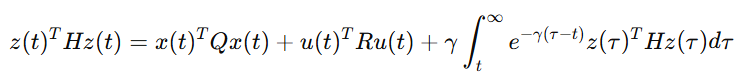
شکل 37: نعریف جدید برای تبدیل به least square

شکل 38: تابع ارزش حالت عمل

خواهیم داشت:

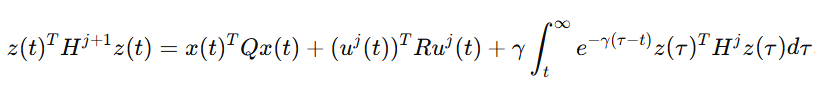
شکل 39: تعریف جدید تابع ارزش حالت عمل

بر اساس تعاریف بالا داریم:



شکل 40: معادله نهایی

حال براساس روابطی که رسیدیم بخش ارزیابی سیاست به صورت زیر تعریف می‌شود که براساس LS قابل حل است:

و بخش بهبود سیاست به صورت زیر تعریف می‌شود:

شکل 41: بهبود سیاست

شکل 42: ارزیابی سیاست

بحث LS نیز به این صورت رفع می‌شود، که این الگوریتم از تکرار سیاست با استفاده از تابع Q به‌صورت آنلاین استفاده می‌کند و می‌تواند بدون نیاز به دانش دینامیک سیستم افزوده اجرا شود. این روش بر اساس استفاده از حداقل مربعات (LS) و داده‌های جمع‌آوری‌شده از مسیرهای سیستم اجرا می‌شود.

با فرض رابطه کلی زیر:

شکل 43: رابطه کلی

با اندازه گیری براساس رابطه بالا و zها می‌توان بخش ارزیابی سیاست که معادله آن در شکل 42 معرفی گردید، را بر اساس LS نیز حل نمود.

## منابع

مقاله معرفی شده در سامانه LMS:

Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics

*مقاله دوم (اصلی):*

Linear Quadratic Tracking Control of Partially-Unknown Continuous-time Systems using Reinforcement Learning