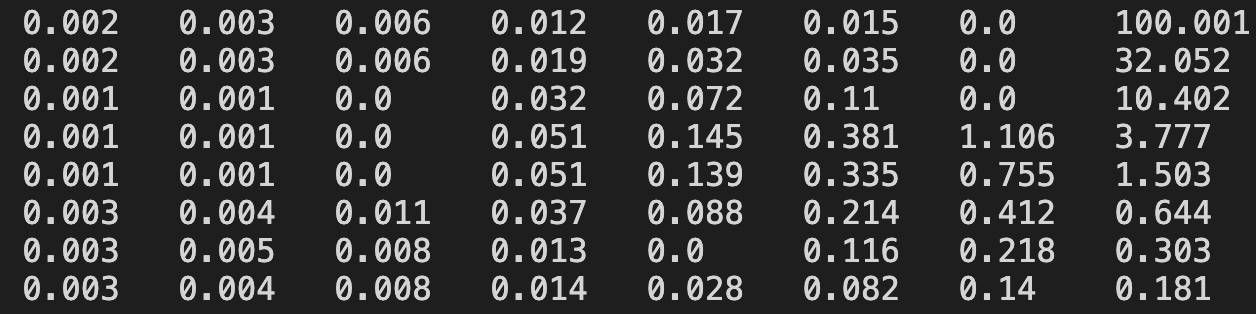
**علی لاریان ۴۰۰۳۰۸۰۴**

**تمرین اول یادگیری تقویتی**

**سوال ۱.**

با توجه به سیاست مطرح شده در سوال، مقادیر ارزیابی شده برای هر حالت به شرح زیر است:

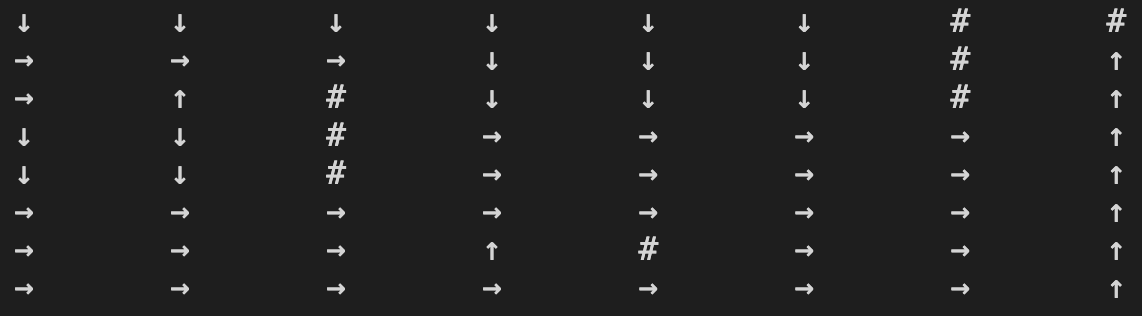
****

در این سوال مقادیر مختلفی برای پارامتر همگرایی یعنی theta بررسی شد. مقادیر theta های کوچکتر از 0.00001 جواب یکسان تولید کردند.

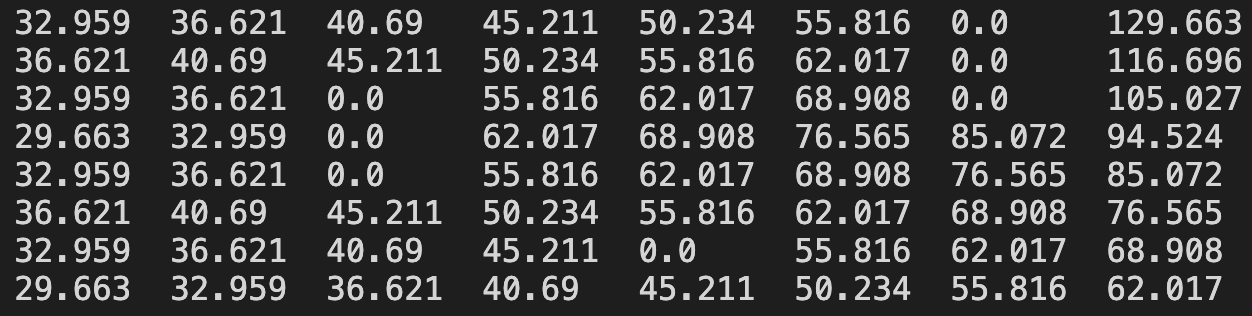
**سوال ۲.**

با استفاده از الگوریتم تکرار سیاست در این سوال، مقادیر اعمال مربوط به سیاست بهینه و ارزش هر حالت به صورت زیر بدست امده است.

اعمال بهینه در هر حالت :



مقادیر بهینه برای هر حالت:

****

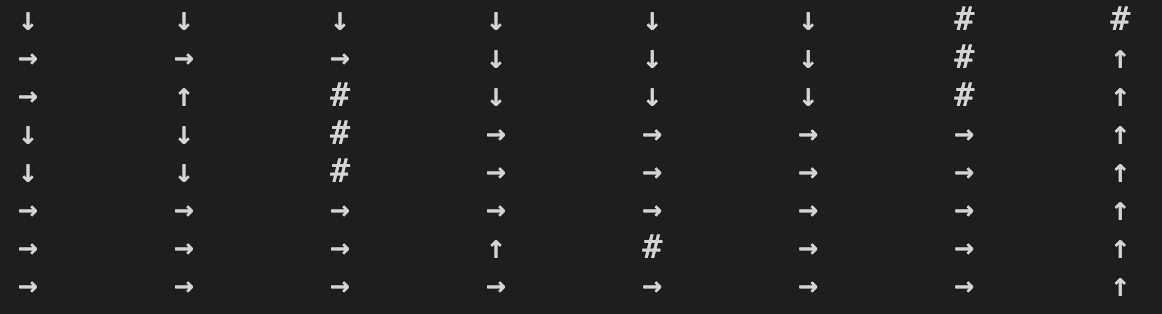
به دلیل این که فرض شده است که مسله مطرح شد یک finite MDP است، این فرایند نه تنها منجر به همگرایی به سیاست بهینه می شود بلکه باعث همگرایی به مقدار بهینه تابع مقدار-حالت نیز می شود.

در این سوال مقادیر مختلفی برای پارامتر همگرایی یعنی theta بررسی شد. مقادیر theta های کوچکتر از 0.00001 جواب یکسان تولید کردند.

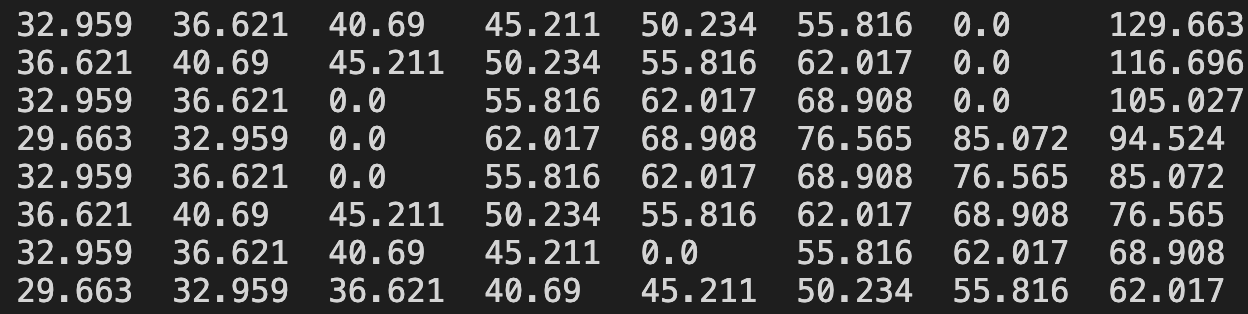
**سوال ۳.**

در این سوال با پیاده سازی الگوریتم تکرار مقدار، مقادیر زیر برای اعمال بهینه و همچنین مقدار بهینه تابع مقدار-حالت بدست امده است.

اعمال بهینه در هر حالت :

****

مقادیر بهینه برای هر حالت:



هر دو الگوریتم تکرار مقدار و تکرار سیاست دقیقا جواب های یکسانی برای سیاست بهینه و همچنین تابع مقدار-حالت تولید کردند.

از لحاظ زمان اجرا بر اساس ثانیه نتایج تجربی در این آزمایش نشان می دهد که الگوریتم تکرار سیاست سریعتر عمل کرده است.

برای انجام این آزمایش هر کدام از الگوریتم ها ۱۰۰۰ بار اجرا شده و زمان اجرای آن ها میانگین گرفته شده است.

زمان اجرا برای الگوریتم تکرار سیاست ۰.۲۰۶ ثانیه بدست آمده است.

زمان اجرا برای الگوریتم تکرار مقدار ۰.۲۹۱ ثانیه بدست آمده است.

**سوال ۴.**