2018

گزارش تکلیف شماره ۱-۳

یادگیری ماشین

سید محمد حسینی - ۸۱۰۱۹۴۵۴۱

قسمت الف:

فضای حالت:

در این مسئله، فضای حالت، میزان سرمایه شخص است. یعنی یک مجموعه به شکل زیر است:

$$s \in \{1, 2, 3, \dots, 99\}$$

دو حالت ۰ و ۱۰۰ ترمینیت ما و خاتمه شبیه سازی هستند.

فضای اعمال:

در این مسئله اعمال میزان سرمایه گذاری در هر مرحله میباشند. به این صورت که در هر مرحله، چون سرمایه شخص در مرحله مشخص میشود، پس اعمالی که دارند از ۰ تا s میباشند. در هر مرحله ما میخواهیم که به هدف یعنی ۱۰۰ تومان برسیم. برای همین در مرحله ۷۰ لازم نیست که بیش از ۳۰ تومان سرمایه گذاری کنیم. به همین دلیل در مرحله ۷۰ تعداد اعمال را از ۰ تا ۳۰ محدود میکنیم. پس فضای اعمال به شکل زیر تشکیل میشود:

$$a \in \{0\ 1, 2, \dots, \min(s, 100 - s)\}$$

خواهدی بود.

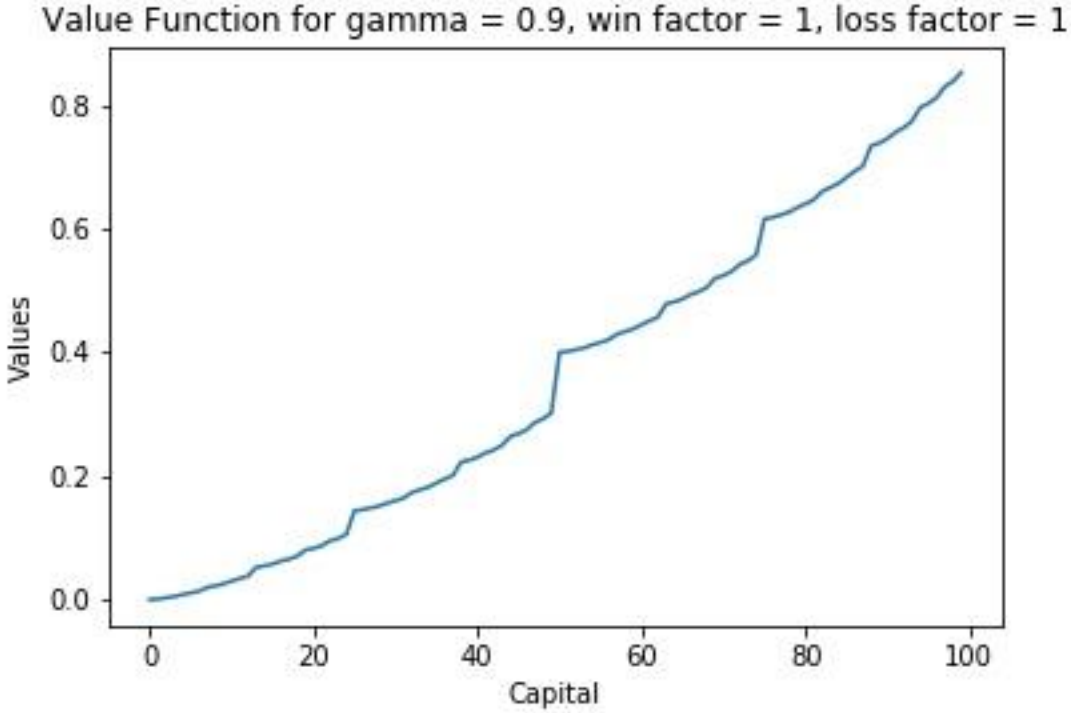
تابع پاداش:

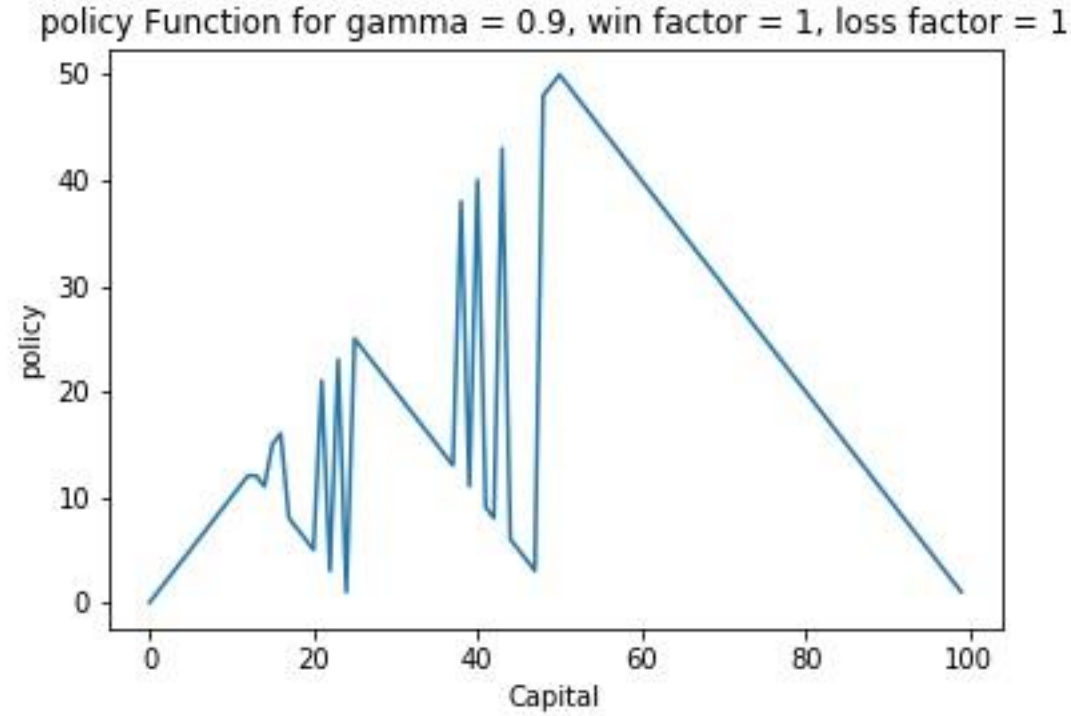
تابع پاداشرابه این صورت تعریف میکنیم که در تمام اعمال ۰ باشد، مگر آنهایی که به هدف کاربر یعنی سرمایه ۱۰۰ تومان برسد. پس تابع پاداش به شکل زیر میشود:

$$r(s) = \begin{cases} 0, & other\ wise \\ 1, & rech\ goal \end{cases}$$

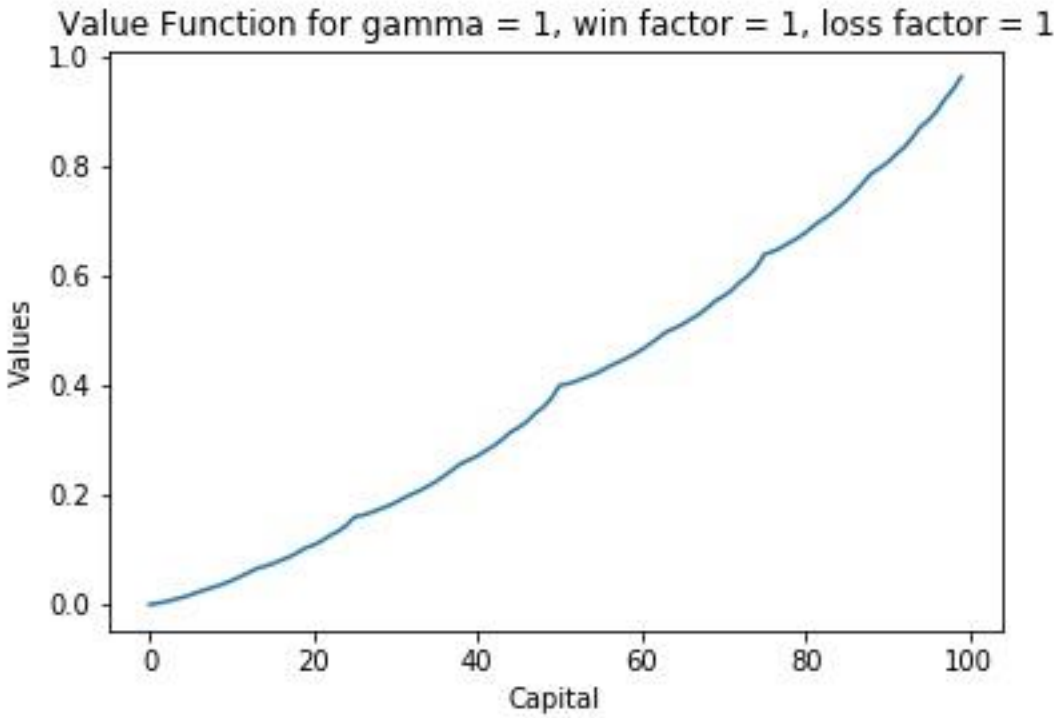
قسمت ب:

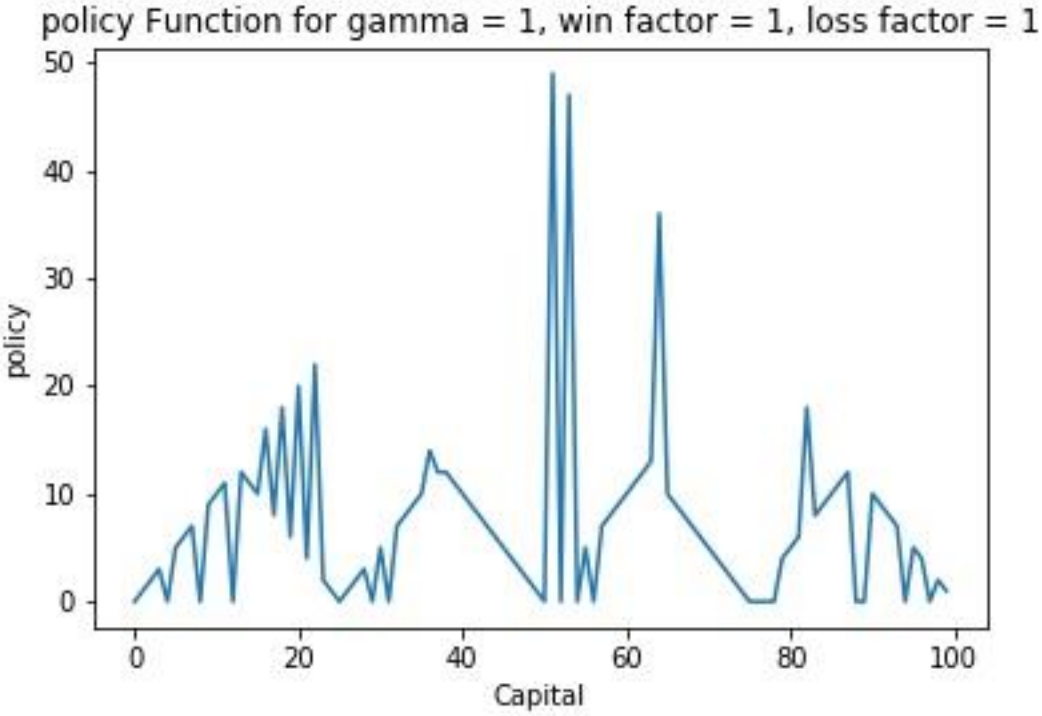
۱. در این قسمت برای gamma = 0.9 داریم:


Value Function for gamma = 0.9, win factor = 1, loss factor = 1
Values
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
Capital


policy Function for gamma = 0.9, win factor = 1, loss factor = 1
policy
0
10
20
30
40
50
0
20
40
60
80
100
Capital

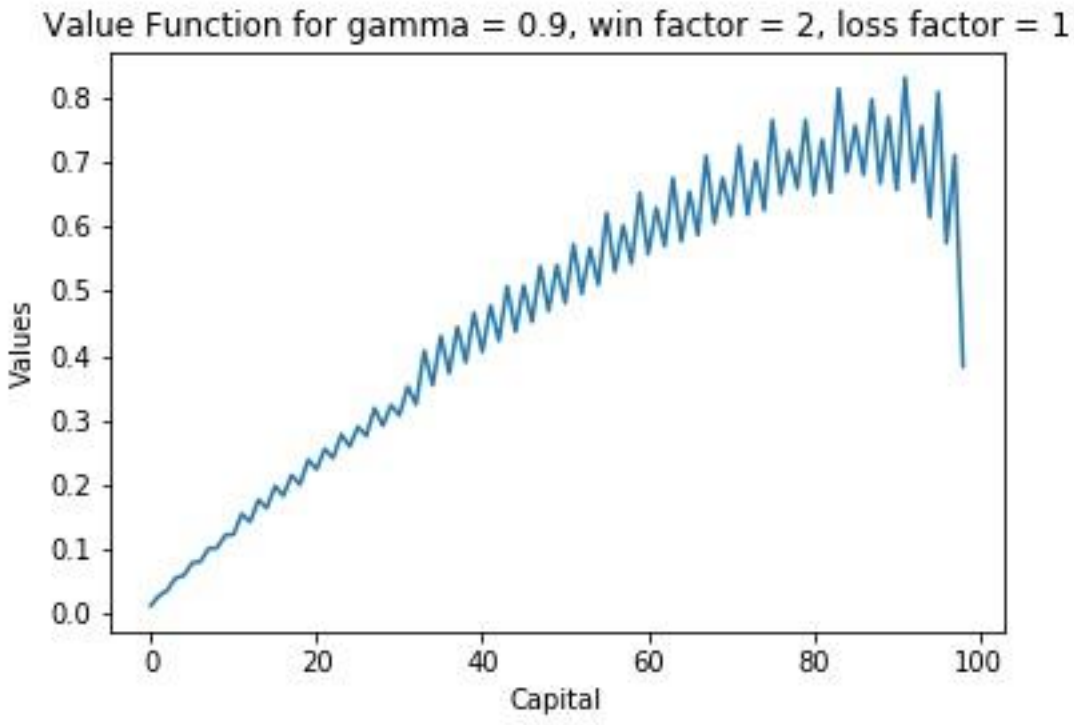
برای gamma = 1 داریم:

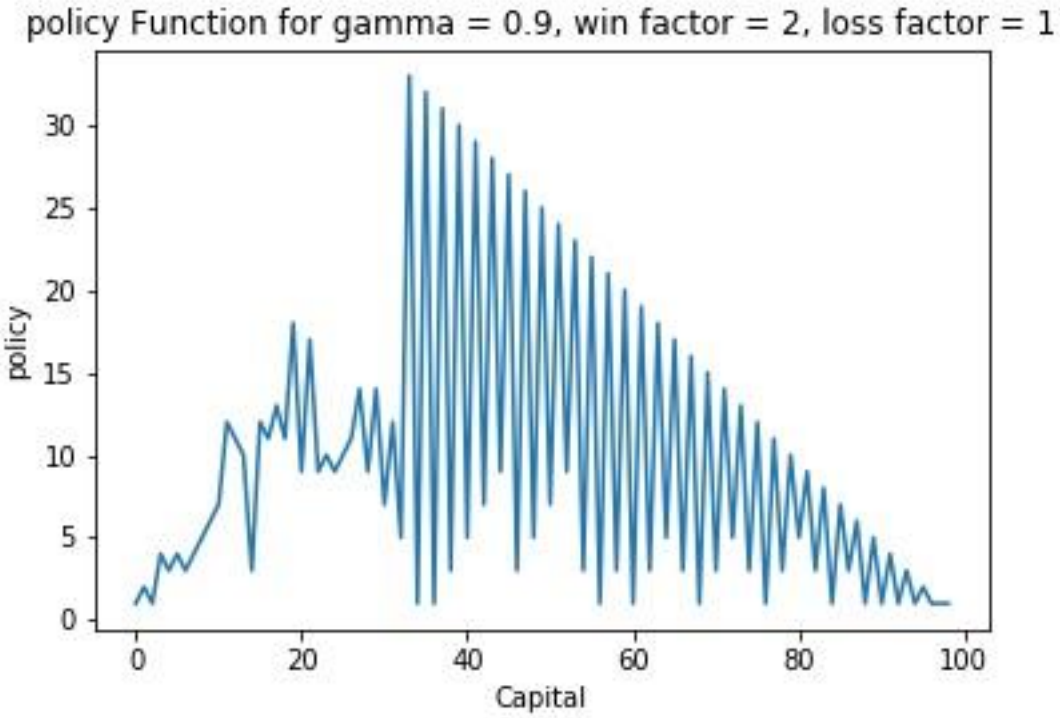




دلیل اختلاف این نمودار های سیاست به صورت واضح گاما میباشد. گاما میزان وابستگی و در نظر گرفتن پاداش های آینده را نشان میدهد. وقتی که $\gamma < 1$ در این حالت، میزان در نظر گرفتن پاداش هایی که در آینده که agent دریافت میکند کوچک میشود. چون پاداش های آینده به صورت geometric میباشند. ولی وقتی $\gamma = 1$ میشود، پاداش های آینده و ضرر ها به اندازه ی پاداش و ضرر اکنون اهمیت پیدا میکنند. به همین دلیل وقتی میزان دارایی کاربر از نصف بیشتر شد، ضرر های آینده را در حالت اول کمتر میبینند و میزان سرمایه گذاری با ریسک بیشتری انجام میدهد. اما وقتی میزان سود وضرر آینده به اندازه حال مهم شد، میزان ریسک پذیری کاهش پیدا میکند و agent با دقت بیشتری سرمایه گذاری میکند. همانطور که در جلوتر کامل توضیح خواهد شد، دلیل دیگر آن نحوه ی سرمایه گذاری باید به صورت حریصانه باشد. چون اگر در هر مرحله به اندازه یک تومان سرمایه گذاری کنیم، تعداد سرمایه گذاری ها به سمت بی نهایت میل میکند و احتمال رسیدن به ۱۰۰ تومان ۰,۴ خواهد بود. برای همین این روش اشتباه است و باید در هر مرحله تمام سرمایه خود را برای رسیدن به مرحله نهایی سرمایه گذاری کنیم.

۲. در این قسمت برای پاداش را دو برابر میکنیم:



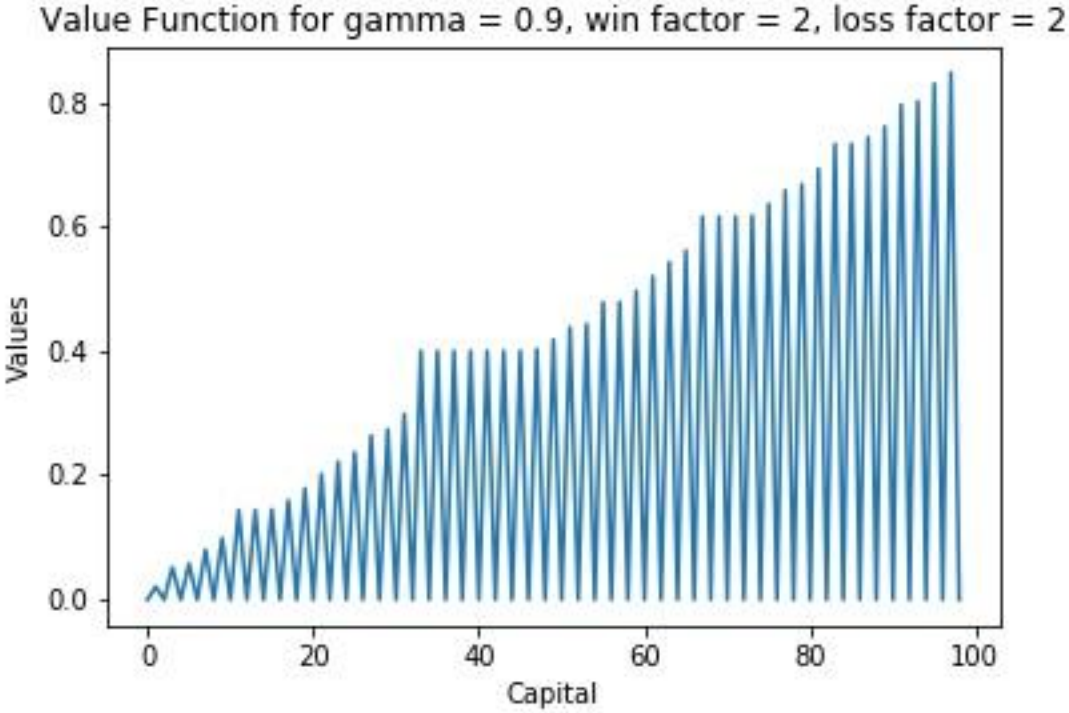


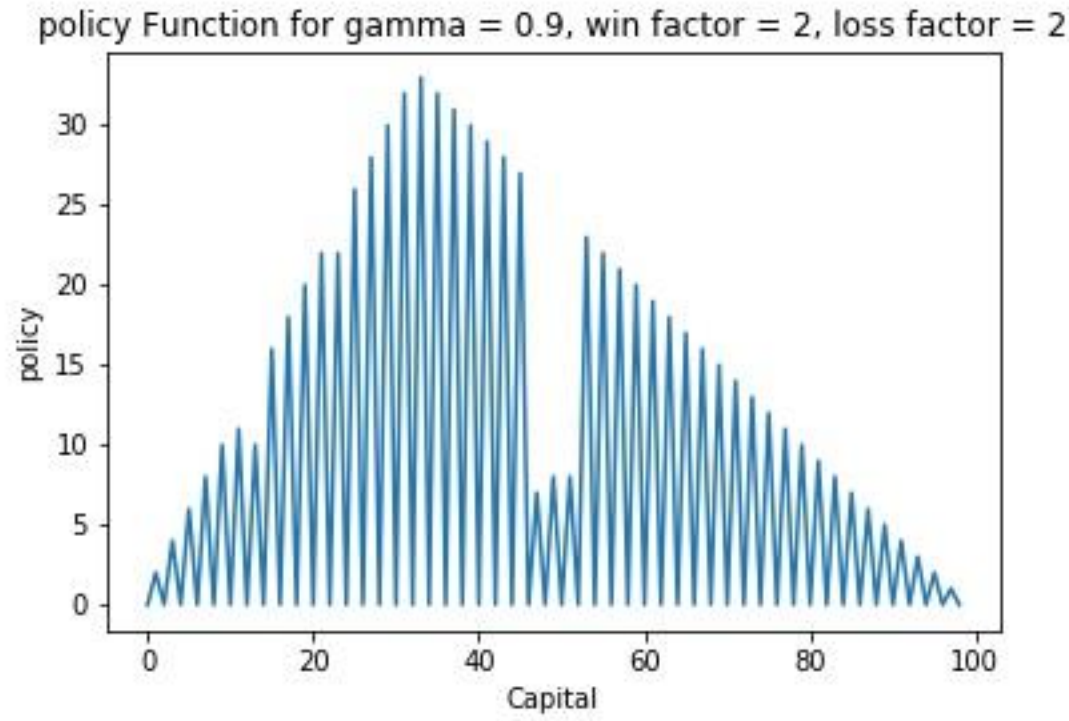
دلیل پرش و سوزنی بودن شکل ها به این دلیل است که در انتخاب بیشینه مقدار sigma در هنگام محاسبه V، کتابخانه اولین مقدار بیشینه را انتخاب میکند و برای میزان پاداش هم اولین بزرگترین مقدار محاسباتی را انتخاب میکند. برای همین بعضی از مقادیر زمانی ایجاد میشوند که مقدار پاداش در حالت دیگری باشد.(خیلی سخت بود توضیحش خدایی!)

دلیل مثلثی بودن سیاست هم این است که طبق توضیحات دوستان، اگر سیاست به صورت خیلی risk averse باشد و هر سری فقط ۱ تومان سرمایه گذاری کند، تعداد سرمایه گذاری ها به سمت عدد بزرگی میرود. در این صورت احتمال رسیدن به مرحله آخر یعنی ۱۰۰ تومان ۰,۴ خواهد بود که هرگز نخواهیم رسید. برای همین بهترین سیاست سرمایه گذاری همه ی سرمایه در هر مرحله(در مراحل پایانی میزان سرمایه ای برای رسیدن به مقدار نهایی) خواهد بود. به همین منظور شکل سیاست به شکل بالا بدست میاید.

دلیل اینکه میزان ماکزیمم روی ۳۴ قرار دارد، این است که میزان پاداش در این حالت دوبرابر است. برای همین وقتی ۳۴ تومان داریم و ۳۳ تومان سرمایه گذاری کنیم سیاست برای رسیدن به مقدار نهایی است(طبق توضیحات داده شده در بالا)

۳. پاداش و ضرر دوبرابر





در این حالت، شکل تابع سیاست خیلی شبیه قسمت قبلی است. اما نکته ای که هست در میزان regret است. طبق محاسبات، در حالت دوم regret بیشتر از قسمت قبل خواهد بود. چون در این قسمت میزان ضرر بیشتر از میزان قبل است. اما برای تاتبع سیاست گفته های قسمت قبل برقرار است.