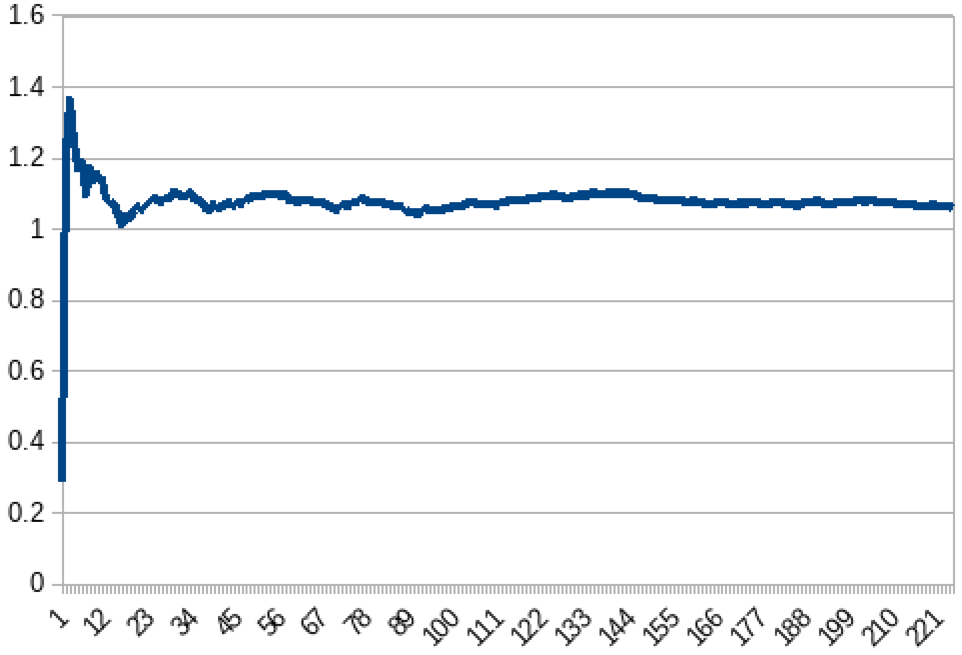
1. 假设 $G_1,G_2,G_3,\ldots,G_t$ 是 $V(s)$ 的采样：
   1) 假设 $E(G_i)=\mu$ ，则 $E(V)=\frac{1}{t}\sum_{i=1}^{t}E(G_i)=\frac{1}{t}t\mu=\mu$
   2) $var(V)=E(G_i^2)-(E(G_i))^2$ ，由于 G_i 的误差是被平方放大的，所以随着样本数量增加，估计误差应该呈平方反比衰减。
   3) $E(V)=\sum_{i=0}(1-\alpha)^{t-i}\alpha^i E(G_i)=1E(G_i)=\mu$
2. 因为在一个确定性策略中，重要性采样率往往是大于0的，例如 $w=\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$ ，分子在确定性策略中为1，或者0，分母往往为小于1的实数，所以重要性采样会导致更新权重增加，引入更大的方差。
3. 首次拜访

```
repeat  $k=1,2,3,\ldots$
    使用策略 $\mu$ 采样第k条轨迹， $S_1,A_1,S_2,A_2,\ldots,S_T$
    $\hat{C}=C,\hat{Q}=Q,G=0,W=1$
    for  $t=T-1,T-2,\ldots 0$  do
        $G=\gamma G+R_{t+1}$
        $\hat{C}(S_t,A_t)=C(S_t,A_t)+W$
        $\hat{Q}(S_t,A_t)=Q(S_t,A_t)+\frac{W}{\hat{C}(S_t,A_t)}(G-Q(S_t,A_t))$
        $W=W\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$
        如果W=0推出循环
    end for
    $C=\hat{C},Q=\hat{Q}$
until 收敛
```

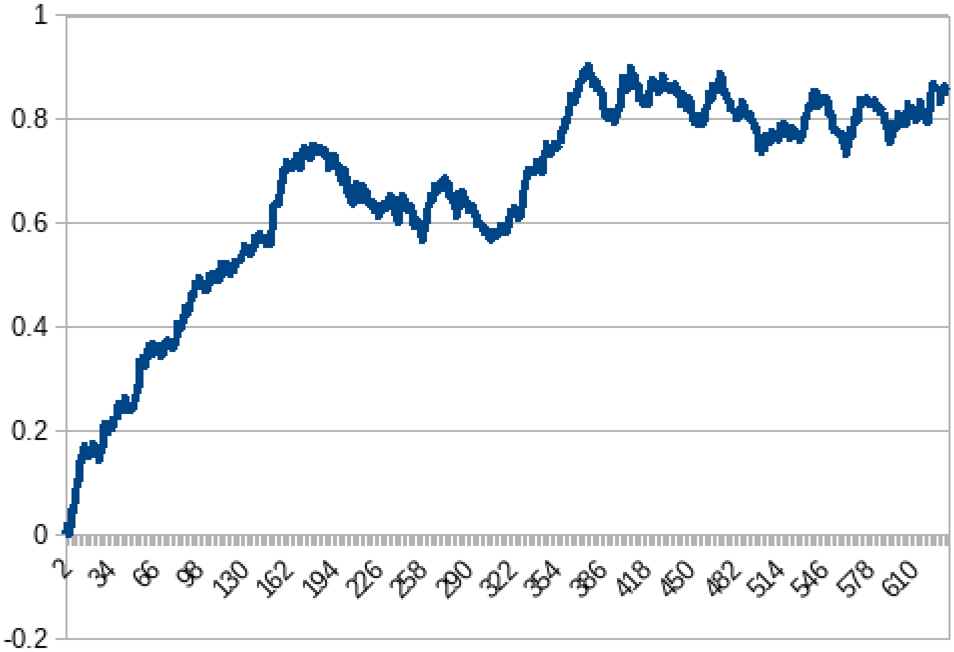
4. 编程作业
   1) 初始化全为0的情况下，收敛曲线如下：



值如下：

3.04634997564735	8.71955434733304	4.45818439900392	5.35693105542727	2.35200508032313
1.32614375715851	2.82913497326059	2.2498735712252	2.16105159659388	1.02217869640418
0.080978575014017	0.622677677448487	0.660405292033089	0.511203427550124	-0.14726285311398
0.094364522942562	-0.245163356225806	-0.251582310734253	-0.527817374204291	-1.09557571409284
-1.11984288204221	-0.750330751290055	-1.11914821406506	-1.51878657412185	-1.45528980226634

2) 固定步长更新，初始值全为 0，收敛曲线如下：



值如下：

3.24131006910427	8.65906376211387	4.57062850422882	5.32085828014935	1.60932137782774
1.28635199164241	2.4143482901023	2.16885398939022	2.14360211349217	0.712100684536799
-0.111057778076717	0.697749737004357	0.60329531087445	0.403067383469183	-0.511980366912051
-1.01420449019762	-0.617806055657609	-0.444779458743911	-0.528550366870863	-1.08400495083697
-1.8726910672582	-1.51554039978956	-1.40573248079615	-1.35383729761303	-2.00521421431674

固定步长更新在初始值偏离较大的情况下收敛较快，但是收敛值不稳定。

3) 最优策略为：

3	0	2	3	2
3	0	0	2	2
3	0	2	0	2
0	0	2	2	2
0	0	2	2	0

最优值函数为：

18.1613864620216	20.2718261058549	18.8556968936593	16.8254743110867	15.2348559634193
15.4802208904371	18.0186245437172	16.3134174557142	15.1015675156634	13.2031124476355
13.5228206292625	15.6716350620038	13.7816723876023	12.8963922443277	11.988363014435
11.7649792196089	13.619825422762	12.2153841448541	11.8434831437241	10.3787026518968
10.6886301564739	11.7689615849423	10.3396823079757	9.51155755789362	9.14774987687646

4) 不能，因为不是所有行为都被更新，导致 Q 函数估计不准确。

5) 最优策略为：

3	0	2	0	2
3	0	0	0	2
0	0	2	0	2
0	0	2	0	0
3	0	2	3	0

6)

最优值函数为：

21.1105412651199	21.7298942673537	19.1073221159073	16.7852923822887	13.5324058618211
17.5758259986523	19.2397278263228	16.9205107605134	14.9083330473701	12.8573367590128
15.6182412619093	16.9493451501234	16.54288017789	13.2547595862463	11.8618652356224
13.9451976892963	14.9144304481817	12.4405093114957	12.4977186497899	11.2270668092037
11.2455145317168	13.2129818753664	12.0422667686216	9.27904553063915	10.3025247995148