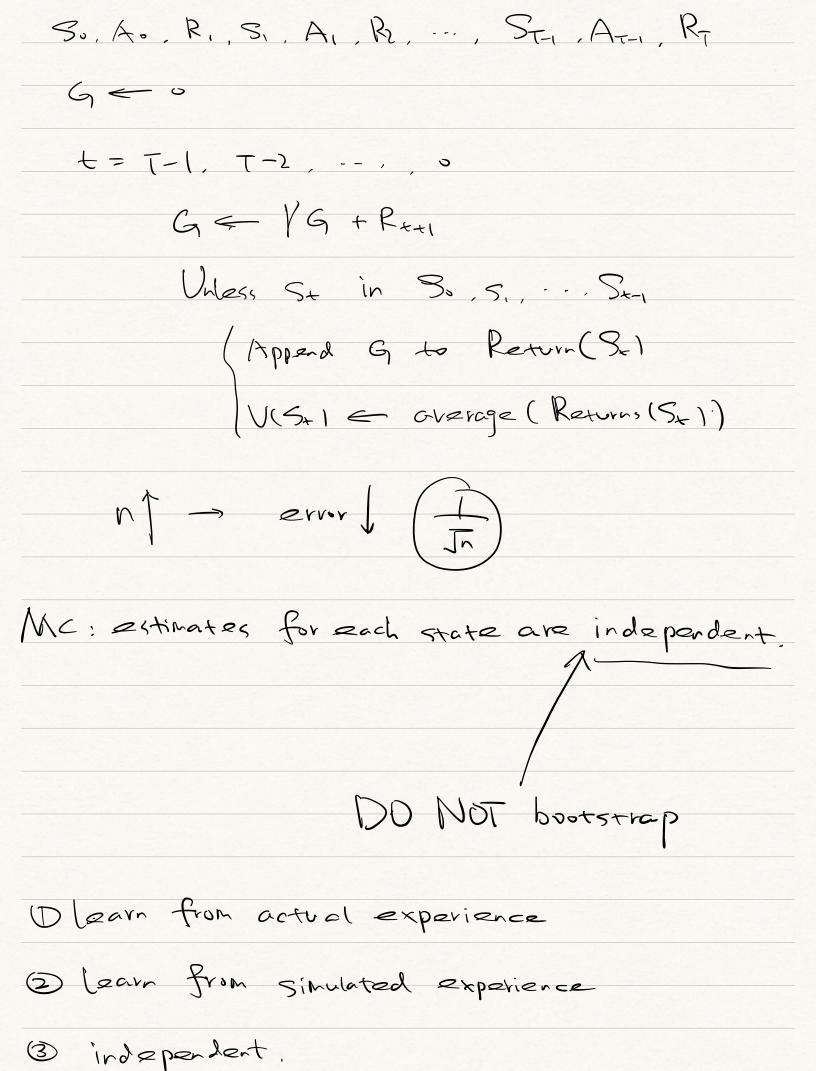
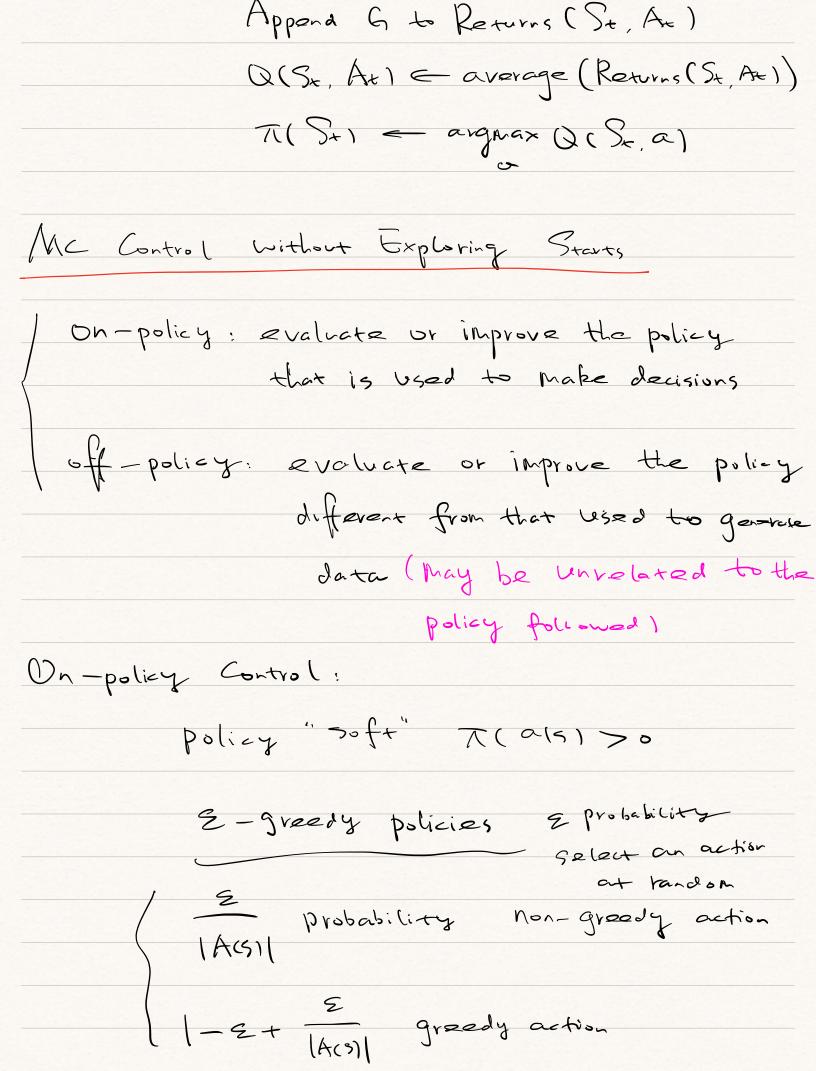
Monte Carlo Methods
(estimate value functions
I discover optimal policies
Monte Carlo: experience sample apisodes
Sample sequences of
States, actions, and remards
from actual or simulated interaction
with an environment
actual experience
simulated experience
Average sample Vetorns"
Experiences -> episodes
1
DP: compute value functions [Mc: Learn value functions
(MC: Learn value functions
Simply average the returns observed after

Visits to that state
"As more returns are observed, the average should converge to the expected value."
A set of apisodes Visit
State 5
first - visit Mc Method
estimate Vx(5) as the average of the Veturns
every-visit Mc method
average all the vecturns following first visit to



Maintaining exploration (Sufficient exploration)
exploring Starts: episodes begin with state-action pairs randomly solected to cover all possibilities action-value function 9
$\pi(s) = avg \max_{\alpha} q(s, \alpha)$
$Q_{\pi_k}(s, \pi_{kn}(s)) = Q_{\pi_k}(s, argmax Q_{\pi_k}(s, a))$
every pair has = Max 9 Tk (5, a) a non-zero probability of being selected as the start 9 Tk (5, Tk (5))
Mc Control with Exploring Start: episode t= T-1, T-2,, o
$G = VG + R_{t+1}$
if St, At in So, Ao, S, A, St., At-1:



Episode: t=T-1, T-2,, 0
G = VG + Rxxx
if St, At in So, Ao, Si, A.,, St., At-1
Append G to Returns (St, At)
Q(St, At) = average (Returns (St, At))
A* = argmax Q(St, a)
For all a E A (S+).
$T(a S_{\star}) = \begin{cases} 1-2+\frac{2}{ A(S_{\star}) } & \text{if } a=A^{\star} \\ \frac{2}{ A(S_{\star}) } & \text{if } a\neq A^{\star} \end{cases}$
(A(Sx)) fatA
Problem: Sufficient exploration
(Maintaining exploration)
with
exploring start exploring start
On-policy Tracks)>0 -) deterministic
1 off - policy

Monte Carlo Control Without Exploring Starts 9/ (s, \(\ta'(s))) = \(\frac{7}{a}\)\(\ta'(\omega) = \(\frac{7 $=\frac{\varepsilon}{|A(S)|} = \frac{\varepsilon}{2} \left(\frac{\varepsilon}{S,\alpha} + (1-\varepsilon) \max_{\alpha} \frac{\varepsilon}{S,\alpha}\right)$ $\frac{2}{|A(S)|} = \frac{2}{a} \frac{1}{|A(S)|} = \frac{2}{|A(S)|} = \frac{2}{|A(S)|$ $=\frac{\varepsilon}{|A(5)|} = \frac{9}{9} (5, \alpha) - \frac{5}{|A(5)|} = \frac{9}{9} (5, \alpha) + \frac{5}{9} \pi (\alpha(5)) (5, \alpha)$ $= \sqrt{\chi(S)}$ $= \sqrt{$ Solect an action at random 2 probability | E for non-greedy actions $\left| \left| -2 + \frac{5}{|A(5)|} \right|$ for gready actions

Off-policy Prediction
On-policy method: learn action values not for a the optimal policy, but for a near-optimal policy that still
2×plov=3
target policy: Lecrned policy, -> optimal policy behavior policy: generate policy -> more explan
Estimate $\sqrt{2}$, $\sqrt{2}$, $\sqrt{5}$ ($5 \neq 7$)
be havior policy
tanger
policy
Converage 7(a/s) >0, b(a/s) >0
Inportance Sampling (ratio)
Weight returns according to the relative probability
of their trajectories occurring under the target

and behavior policies.
State-action tragectory:
At, Stri, Atri,, ST
? {A+, S++1, A++1,, ST (8+, A++, ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~ ~~
= T(Ax15+1P(Ste) St. At) T(Ath Stal)
P (ST) ST-1, AT-1)
T -(
= TT T(AKISK) P(SKH [SK, AK)
TT-1 T (AK (SK) P(SK+1 SK, AK)
1+: T-1= T-1 b(Ak Sk) P(Sk) Sk, Ak)
(1/k2+ 2/ 1/k/)k/)k+(1/)k. Hk/
T-1
= T-1 = T(Ak Sk) = K=+ b(Ak, Sk)
K=+ b(AK,SK)

E[(+: T-1 G+ (5+=5) = Vz (5)

Variance of the importance sampling-scaled veturns:

$$V_{ar}[x] = E[(x - \overline{x})^2]$$

$$= E[x^2 - 2x \overline{x} + \overline{x}^2]$$

$$= E[x^2] - \overline{x}^2$$

Experted square of the improtance sampling

 τ τ_{-1} τ_{-1} τ_{-1}

