Control Variates with Monte Carlo Methods

Vincent Antaki

McGill University

Control Variates

Definition : Technique that aims to control, with the general aim to reduce, the variance of a policy during training.
Ex.

- Baseline : Set a realist expected return and substract it from the obtained return.
- Importance sampling : A general technique for estimating expected values under one distribution given sample from another. Allows us to do off-policy MC.

First-visit MC vs. Every-visit MC

Let G_i be the discounted return at timestep i.
First-visit : computed until the end of episode

$$G_i = \sum_{j=0}^{T-i} \gamma^j R_{j+i}$$

Every-visit : computed until we come back to the state

Off-policy MC - importance sampling

- a way to reduce variance even when $\gamma = 1$.
- given μ a behavior policy and π a target policy, the importance sampling ratio from iteration t to iteration $T - 1$ is given by :

$$\rho_t^T = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{\mu(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_{T-1}|S_{T-1})}{\mu(A_{T-1}|S_{T-1})}$$

Algoritms

Here we compare :

- On-Policy first-visit Monte Carlo
- On-Policy every-visit Monte Carlo
- Off-Policy every-visit Monte Carlo with importance sampling

Environment

We instanciate 50 random mdp with 10 states and 2 actions.

- One starting state, one ending state.
- Every state returns a fixed reward $r \in U(-1, 1)$
- Every pair state-action can only lead to two state (one with prob p and the other $p - 1$ with $p \in U(0, 1)$)
- Maximum 50 iterations by episodes.
- No penality for reaching maximum number of iteration, no special reward for reaching end state.

N.B.

- $\gamma = 0.9$
- In such context, the best policy is often to take actions that pushes the agent towards a certain cycle of states-actions with an expected positive reward and to avoid actions that could lead to the final state.

On-Policy MC settings

- π is ϵ-soft w.r.t Q with $\epsilon = 0.2$
- Softmax temperature of 1

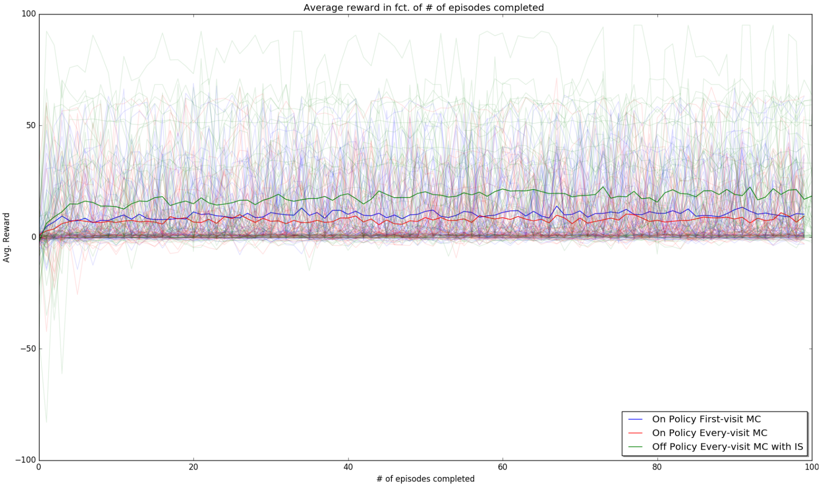
Off-Policy MC settings

- μ is ϵ-soft w.r.t Q $\epsilon = 0.2$
- π is ϵ-soft w.r.t. Q with $\epsilon = 0.025$, evaluated after every training episodes.
- Softmax temperature of 1
- Importance sampling : (for $t = T - 1, \ldots T - 2$)

$$W \leftarrow W \frac{\pi(a_t | s_t)}{\mu(a_t | s_t)}$$

Results*

*No avering applied.



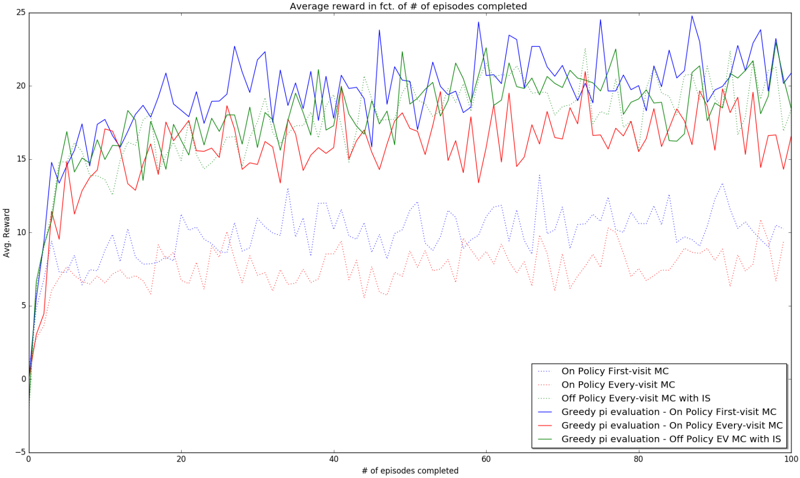
Discussion

- On-policy first-visit MC seems to perform better than on policy every-visit MC. This could be explained by the sudden interruption induced by reaching the maximum number of iteration allowed (100 in our case).
- Off-policy seems to performs better than both on-policy methods.

- Could there be an unfair advantage for the off-policy method in the previous graph since its π has a very low epsilon?
- In the end, our objective is to have good qvalues.
- In the next graph, we have for each algorithm the evaluation at step of the training of an episode under the greedy policy induces by their qvalues.

Results

*Lines in previous graph are now dotted.



- It seems like the on policy first-visit slightly overperforms our off policy every-visit MC.

Let's evaluate, for 100 episodes of each of the 50 MDPs, the average performance of greedy policy derived from the learned Q for each algorithm after training with 100 episodes.

Algorithm	Average reward	Std. reward	Avg. nb. iter.
On-policy first-visit	21.62	23.43	49.61
On-policy every-visit	17.18	20.89	44.00
Off-policy every-visit	19.72	22.32	47.74

Conclusion

- On-policy MC are sensitive to samples variation
- Off-policy allows to separate exploration and exploitation in two different policy, importance sampling allows us to compensate the discrepancy between both policy when updating Q.

Conclusion

In our environment :

- On-policy MC are sensitive to samples variation but still learn quite good qvalues.
- First-visit MC seems to perform better than every-visit MC.
- Importance sampling does not seem to bring significantly better performance.

Conclusion

Other approaches to control variates :

- Average return as baseline
- Value function as baseline
- Return-specific importance sampling

The End

Thank you!