CIFAR

amii

Temporal Difference Methods for Prediction

Rupam Mahmood

February 24, 2020

RL & AI

UNIVERSITY OF ALBERTA

Prediction as estimating value functions

Predictions are building blocks for many control methods

The usefulness of predictions goes beyond control

Forming a predictive question: **How many times will you get honked at today?**

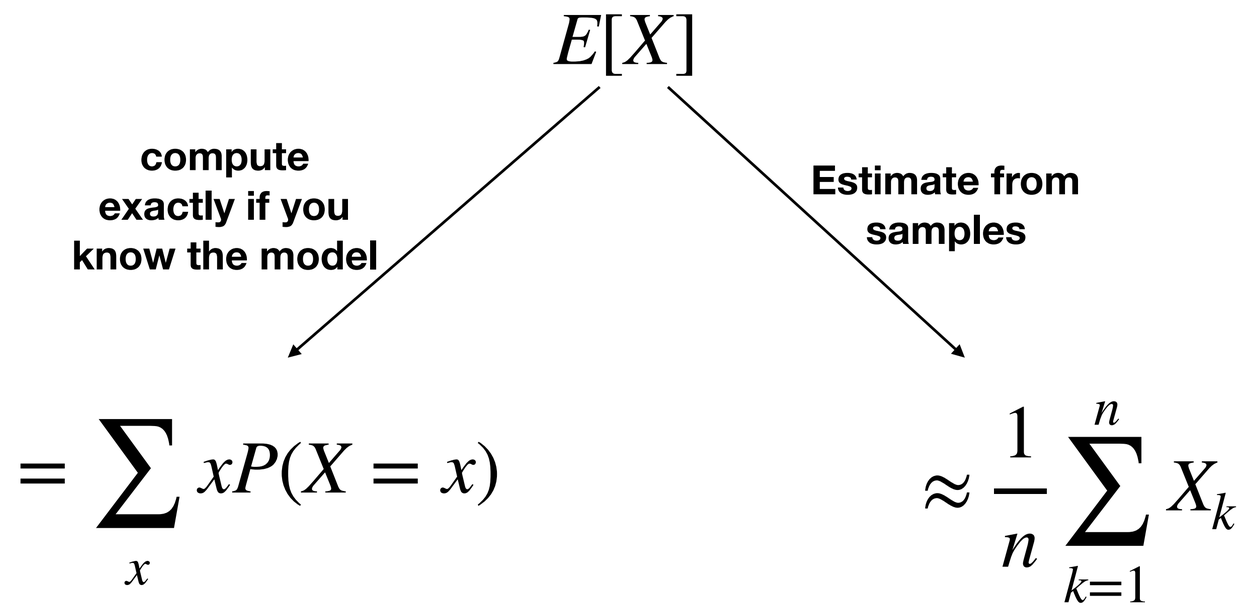
(Pseudo-) reward: +1 for each honk

Termination of episode: end of the day

Behavior: the way you drive (think of your average speed, frequency of changing lanes, etc.)

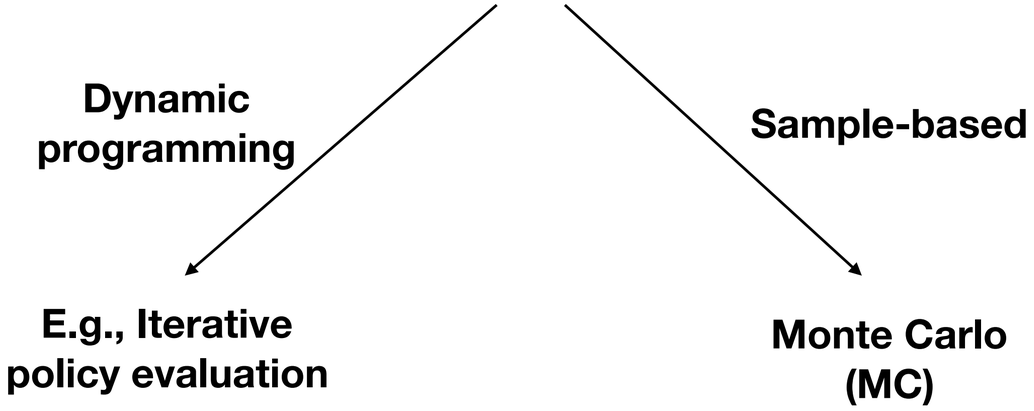
Can be answered by estimating $v_\pi(s) \doteq E_\pi[G_t \,|\, S_t = s]$

Much of prediction is about estimating expected values



Much of prediction is about estimating expected values

$$v_\pi(s) \doteq E_\pi[G_t \mid S_t = s]$$



From MC to TD(0)

Monte Carlo estimator for on-policy prediction: $V(s) \doteq \dfrac{\sum_{t \in \mathcal{T}(s)} G_t}{|\mathcal{T}(s)|}$

Incremental Monte Carlo estimator: $V(S_t) \leftarrow V(S_t) + \dfrac{1}{N(S_t)} \left[G_t - V(S_t) \right]$

Constant-α MC: $V(S_t) \leftarrow V(S_t) + \alpha \left[G_t - V(S_t) \right]$

TD(0): $V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t) \right]$

Unlike Monte Carlo, TD(0) works online

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

Say an oracle gives us return G from future at each step. Replace $R + \gamma V(S')$ with G.

This is an online but acausal Monte Carlo method. Will it be first-visit or every-visit?

From Monte Carlo error to TD error

Constant-α MC: $V(S_t) \leftarrow V(S_t) + \alpha \underbrace{\left[G_t - V(S_t)\right]}_{\textbf{MC error: } \Delta_t}$

TD(0): $V(S_t) \leftarrow V(S_t) + \alpha \underbrace{\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]}_{\textbf{TD error: } \delta_t}$

$$\Delta_t = \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k$$

Based on this equivalence, can you think of an implementation of Monte Carlo method that computes partial updates in an online manner and completes the full update at the end of the episode?