ECE276B: Planning & Learning in Robotics
Lecture 11: Model-Free Prediction

Nikolay Atanasov

natanasov@ucsd.edu

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Outline

Model-Free Policy Evaluation

Monte Carlo Policy Evaluation

Temporal Difference Policy Evaluation

From Optimal Control To Reinforcement Learning

- **Stochastic Optimal Control**: MDP with known motion model $p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u})$ and cost function $\ell(\mathbf{x}, \mathbf{u})$
  - **Model-Based Prediction**: compute value function V^π of given policy π
    - Policy Evaluation Theorem
  - **Model-Based Control**: optimize value function V^π to get improved policy π'
    - Policy Improvement Theorem
- **Reinforcement Learning**: MDP with unknown motion model $p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u})$ and cost function $\ell(\mathbf{x}, \mathbf{u})$ but access to samples $\{(\mathbf{x}'_i, \mathbf{x}_i, \mathbf{u}_i, \ell_i)\}_i$ of system transitions and incurred costs
  - **Model-Free Prediction**: estimate value function V^π of given policy π:
    - Monte-Carlo (MC) Prediction
    - Temporal-Difference (TD) Prediction
  - **Model-Free Control**: optimize value function V^π to get improved policy π':
    - On-policy MC Control: ϵ-greedy
    - On-policy TD Control: SARSA
    - Off-policy MC Control: Importance Sampling
    - Off-policy TD Control: Q-Learning

Bellman Operators

- **Hamiltonian**:

$$H[\mathbf{x}, \mathbf{u}, V] = \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\mathbf{u})} [V(\mathbf{x}')]$$

- Operators for policy value functions:
  - **Policy Evaluation Operator**:

$$\mathcal{B}_\pi[V](\mathbf{x}) := \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\pi(\mathbf{x}))} \left[V(\mathbf{x}')\right] = H[\mathbf{x}, \pi(\mathbf{x}), V(\cdot)]$$

  - **Policy Q-Evaluation Operator**:

$$\mathcal{B}_\pi[Q](\mathbf{x}, \mathbf{u}) := \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\mathbf{u})} \left[Q(\mathbf{x}', \pi(\mathbf{x}'))\right] = H[\mathbf{x}, \mathbf{u}, Q(\cdot, \pi(\cdot))]$$

- Operators for optimal value functions:
  - **Value Operator**:

$$\mathcal{B}_*[V](\mathbf{x}) := \min_{\mathbf{u} \in \mathcal{U}} \left\{\ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\mathbf{u})} \left[V(\mathbf{x}')\right]\right\} = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V(\cdot)]$$

  - **Q-Value Operator**:

$$\mathcal{B}_*[Q](\mathbf{x}, \mathbf{u}) := \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot|\mathbf{x},\mathbf{u})} \left[\min_{\mathbf{u}' \in \mathcal{U}} Q(\mathbf{x}', \mathbf{u}')\right] = H[\mathbf{x}, \mathbf{u}, \min_{\mathbf{u}' \in \mathcal{U}} Q(\cdot, \mathbf{u}')]$$

Model-Free Prediction

- Objective: estimate value function V^π of given policy π
- Approach: approximate Policy Evaluation operators $\mathcal{B}_\pi[V]$ and $\mathcal{B}_\pi[Q]$ using samples $\{(\mathbf{x}'_i, \mathbf{x}_i, \mathbf{u}_i, \ell_i)\}_i$ instead of computing the expectation over $\mathbf{x}'$ exactly:
  - Monte-Carlo (MC) methods:
    - expected long-term cost approximated by sample average over whole system trajectories (applies to First-Exit and Finite-Horizon settings only)
  - Temporal-Difference (TD) methods:
    - expected long-term cost approximated by a sample average over few system transitions and an estimate of the expected long-term cost at the reached state (bootstrapping)

- **Sampling**: value estimates $V^\pi(\mathbf{x})$ rely on samples $\{(\mathbf{x}'_i, \mathbf{x}_i, \mathbf{u}_i, \ell_i)\}_i$:
  - DP does not sample
  - MC samples
  - TD samples

- **Bootstrapping**: value estimates $V^\pi(\mathbf{x})$ rely on other value estimates $V^\pi(\mathbf{x}')$:
  - DP bootstraps
  - MC does not bootstrap
  - TD bootstraps

Outline

Model-Free Policy Evaluation

Monte Carlo Policy Evaluation

Temporal Difference Policy Evaluation

Monte-Carlo Policy Evaluation

- **Assumption**: MC policy evaluation applies to the First-Exit problem
- **Episode**: a sequence ρ_τ of states and controls from initial state x_τ at initial time τ, following the stochastic system transitions under policy π:

$$\rho_\tau := \mathbf{x}_\tau, \mathbf{u}_\tau, \mathbf{x}_{\tau+1}, \mathbf{u}_{\tau+1}, \ldots, \mathbf{x}_{T-1}, \mathbf{u}_{T-1}, \mathbf{x}_T \sim \pi$$

- **Long-Term Cost** of episode ρ_τ:

$$L_\tau(\rho_\tau) := \gamma^{T-\tau}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau}^{T-1} \gamma^{t-\tau}\ell(\mathbf{x}_t, \mathbf{u}_t)$$

- **Goal**: approximate $V^\pi(\mathbf{x})$ from several episodes $\rho_\tau^{(k)} \sim \pi$, $k = 1, \ldots, K$
- **MC Policy Evaluation**: uses the empirical mean of the long-term costs of the episodes $\rho_\tau^{(k)}$ to approximate the value of π:

$$V^\pi(\mathbf{x}) = \mathbb{E}_{\rho\sim\pi}[L_\tau(\rho) \mid \mathbf{x}_\tau = \mathbf{x}] \approx \frac{1}{K}\sum_{k=1}^{K} L_\tau(\rho_\tau^{(k)})$$

Monte-Carlo Policy Evaluation

- **Goal**: approximate $V^\pi(\mathbf{x})$ from episodes $\rho^{(k)} \sim \pi$
- **First-Visit MC Policy Evaluation**:
  - for each state $\mathbf{x}$ and episode $\rho^{(k)}$, find the **first** time step t that state $\mathbf{x}$ is visited in $\rho^{(k)}$ and increment:
    - the number of visits to $\mathbf{x}$: $N(\mathbf{x}) \leftarrow N(\mathbf{x}) + 1$
    - the long-term cost starting from $\mathbf{x}$: $C(\mathbf{x}) \leftarrow C(\mathbf{x}) + L_t(\rho^{(k)})$
  - Approximate the value function of π: $V^\pi(\mathbf{x}) \approx \frac{C(\mathbf{x})}{N(\mathbf{x})}$
- **Every-Visit MC Policy Evaluation**: same approach but the long-term costs are accumulated following **every** time step t that state $\mathbf{x}$ is visited in $\rho^{(k)}$

Monte-Carlo Policy Evaluation

Algorithm First-Visit MC Policy Evaluation

Initialize $\pi(\mathbf{x})$
$C(\mathbf{x}) \leftarrow 0$ for all $\mathbf{x}$, $N(\mathbf{x}) \leftarrow 0$ for all $\mathbf{x}$
loop
    Generate $\rho = \mathbf{x}_0, \mathbf{u}_0, \mathbf{x}_1, \mathbf{u}_1, \ldots, \mathbf{x}_{T-1}, \mathbf{u}_{T-1}, \mathbf{x}_T$ from π
    for $\mathbf{x} \in \rho$ **do**
        $L \leftarrow$ return following first appearance of $\mathbf{x}$ in ρ
        $N(\mathbf{x}) \leftarrow N(\mathbf{x}) + 1$
        $C(\mathbf{x}) \leftarrow C(\mathbf{x}) + L$
return $V^{\pi}(\mathbf{x}) \leftarrow \frac{C(\mathbf{x})}{N(\mathbf{x})}$

- Every-Visit MC adds to $C(\mathbf{x})$ not a single return L but the returns $\{L\}$ following all appearances of $\mathbf{x}$ in ρ

Running Sample Average

- Consider a sequence $x_1, x_2, \ldots,$ of samples from a random variable
- **Sample average**:

$$\mu_{k+1} = \frac{1}{k+1} \sum_{j=1}^{k+1} x_j$$

- **Running average**:

$$\mu_{k+1} = \frac{1}{k+1} \sum_{j=1}^{k+1} x_j = \frac{1}{k+1} \left(x_{k+1} + \sum_{j=1}^{k} x_j \right) = \frac{1}{k+1} (x_{k+1} + k\mu_k)$$
$$= \mu_k + \frac{1}{k+1}(x_{k+1} - \mu_k)$$

- **Weighted running average**: update μ_k using a step-size $\alpha_{k+1} \neq \frac{1}{k+1}$:

$$\mu_{k+1} = \mu_k + \alpha_{k+1}(x_{k+1} - \mu_k)$$

- **Robbins-Monro step size**: convergence to the true mean is guaranteed almost surely under the following conditions:

$$\left(\begin{array}{c}\text{independence from}\\ \text{initial conditions}\end{array}\right) \sum_{k=1}^{\infty} \alpha_k = \infty \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty \text{ (ensures convergence)}$$

First-Visit MC Policy Evaluation (Running Average)

Algorithm First-Visit MC Policy Evaluation (Running Average)

```
Initialize π(x)
V^π(x) ← 0 for all x
loop
    Generate ρ = x_0, u_0, x_1, u_1, ..., x_{T-1}, u_{T-1}, x_T from π
    for x ∈ ρ do
        L ← return following first appearance of x in ρ
        V^π(x) ← V^π(x) + α(L − V^π(x))        ▷ usual choice: α := 1/(N(x)+1)
```

Outline

Temporal-Difference Policy Evaluation

- **Bootstrapping**: the estimate of $V^\pi(\mathbf{x})$ at state $\mathbf{x}$ relies on the estimate $V^\pi(\mathbf{x}')$ at another state

- TD combines the sampling of MC with the bootstrapping of DP:

$$
\begin{aligned}
V^\pi(\mathbf{x}) &= \mathbb{E}_{\rho\sim\pi}[L_\tau(\rho) \mid \mathbf{x}_\tau = \mathbf{x}] \\
&= \mathbb{E}_{\rho\sim\pi}\left[\gamma^{T-\tau}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau}^{T-1}\gamma^{t-\tau}\ell(\mathbf{x}_t,\mathbf{u}_t) \mid \mathbf{x}_\tau = \mathbf{x}\right] \\
&= \mathbb{E}_{\rho\sim\pi}\left[\ell(\mathbf{x}_\tau,\mathbf{u}_\tau) + \gamma\left(\gamma^{T-\tau-1}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau+1}^{T-1}\gamma^{t-\tau-1}\ell(\mathbf{x}_t,\mathbf{u}_t)\right) \mid \mathbf{x}_\tau = \mathbf{x}\right] \\
&\underset{\text{bootstrap}}{\overset{TD(0)}{=\!=\!=}} \mathbb{E}_{\rho\sim\pi}\left[\ell(\mathbf{x}_\tau,\mathbf{u}_\tau) + \gamma V^\pi(\mathbf{x}_{\tau+1}) \mid \mathbf{x}_\tau = \mathbf{x}\right] \\
&\underset{\text{bootstrap}}{\overset{TD(n)}{=\!=\!=}} \mathbb{E}_{\rho\sim\pi}\left[\sum_{t=\tau}^{\tau+n}\gamma^{t-\tau}\ell(\mathbf{x}_t,\mathbf{u}_t) + \gamma^{n+1}V^\pi(\mathbf{x}_{\tau+n+1}) \mid \mathbf{x}_\tau = \mathbf{x}\right] \\
&\overset{MC}{\approx} \frac{1}{K}\sum_{k=1}^{K}\left[\sum_{t=\tau}^{\tau+n}\gamma^{t-\tau}\ell(\mathbf{x}_t^{(k)},\mathbf{u}_t^{(k)}) + \gamma^{n+1}V^\pi(\mathbf{x}_{\tau+n+1}^{(k)})\right]
\end{aligned}
$$

Temporal-Difference Policy Evaluation

- **Goal**: approximate $V^\pi(\mathbf{x})$ from episodes $\rho \sim \pi$
- **MC Policy Evaluation**: updates the value estimate $V^\pi(\mathbf{x}_t)$ towards the long-term cost $L_t(\rho_t)$:

$$V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha(L_t(\rho_t) - V^\pi(\mathbf{x}_t))$$

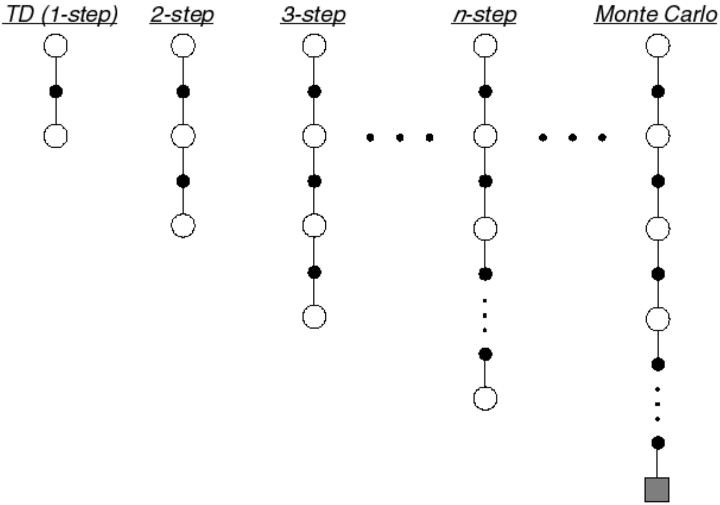
- **TD(0) Policy Evaluation**: updates the value estimate $V^\pi(\mathbf{x}_t)$ towards an *estimated* long-term cost $\ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1})$:

$$V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha(\ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1}) - V^\pi(\mathbf{x}_t))$$

- **TD(n) Policy Evaluation**: updates the value estimate $V^\pi(\mathbf{x}_t)$ towards an *estimated* long-term cost $\sum_{\tau=t}^{t+n} \gamma^{\tau-t}\ell(\mathbf{x}_\tau, \mathbf{u}_\tau) + \gamma^{n+1} V^\pi(\mathbf{x}_{t+n+1})$:

$$V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha\left(\sum_{\tau=t}^{t+n} \gamma^{\tau-t}\ell(\mathbf{x}_\tau, \mathbf{u}_\tau) + \gamma^{n+1} V^\pi(\mathbf{x}_{t+n+1}) - V^\pi(\mathbf{x}_t)\right)$$

TD(n) Policy Evaluation



MC and TD Errors

- **TD error**: measures the difference between the estimated value $V^\pi(\mathbf{x}_t)$ and the improved estimate $\ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1})$:

$$\delta_t := \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1}) - V^\pi(\mathbf{x}_t)$$

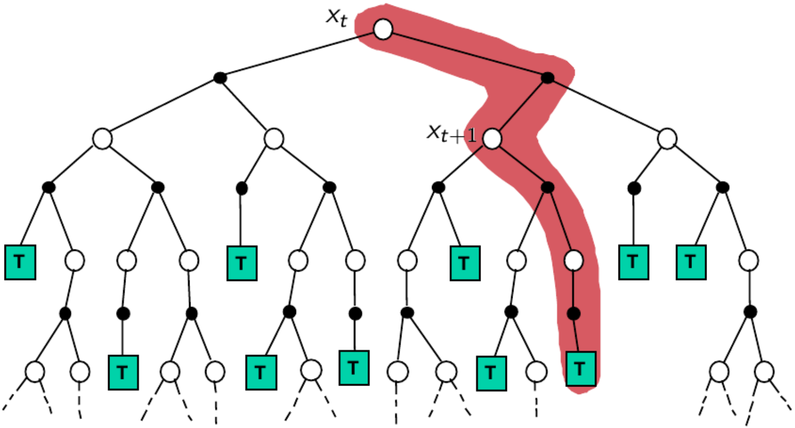
- **MC error**: a sum of TD errors:

$$\begin{aligned}
L_t(\rho_t) - V^\pi(\mathbf{x}_t) &= \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma L_{t+1}(\rho_{t+1}) - V^\pi(\mathbf{x}_t) \\
&= \delta_t + \gamma \left(L_{t+1}(\rho_{t+1}) - V^\pi(\mathbf{x}_{t+1}) \right) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 \left(L_{t+2}(\rho_{t+2}) - V^\pi(\mathbf{x}_{t+2}) \right) \\
&= \sum_{n=0}^{T-t-1} \gamma^n \delta_{t+n}
\end{aligned}$$

- **MC and TD converge**: $V^\pi(\mathbf{x})$ approaches the true value function of π as the number of sampled episodes $\to \infty$ as long as α_k is a Robbins-Monro sequence and $\mathcal{X}$ is finite (needed for TD convergence)

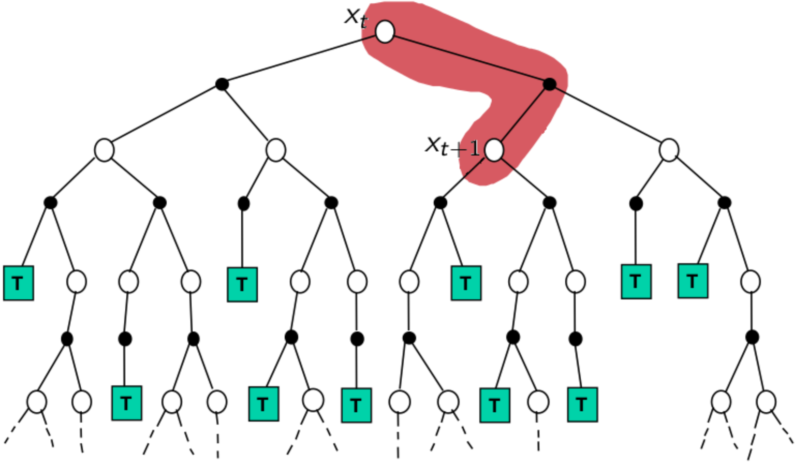
Monte-Carlo Backup

$$V^{\pi}(\mathbf{x}_t) \leftarrow V^{\pi}(\mathbf{x}_t) + \alpha(L_t(\rho_t) - V^{\pi}(\mathbf{x}_t))$$



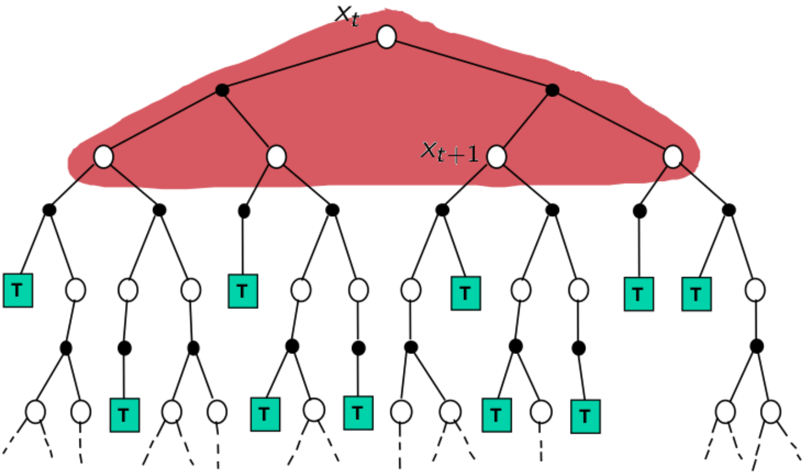
Temporal-Difference Backup

$$V^{\pi}(\mathbf{x}_t) \leftarrow V^{\pi}(\mathbf{x}_t) + \alpha(\ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^{\pi}(\mathbf{x}_{t+1}) - V^{\pi}(\mathbf{x}_t))$$

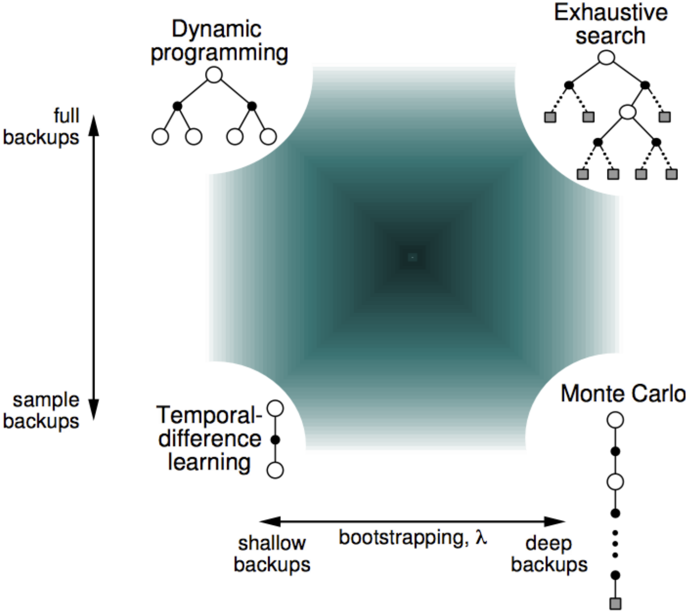


Dynamic-Programming Backup

$$V^{\pi}(\mathbf{x}_t) \leftarrow \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma \mathbb{E}_{\mathbf{x}_{t+1} \sim p_f(\cdot|\mathbf{x}_t, \mathbf{u}_t)} \left[V^{\pi}(\mathbf{x}_{t+1}) \right]$$



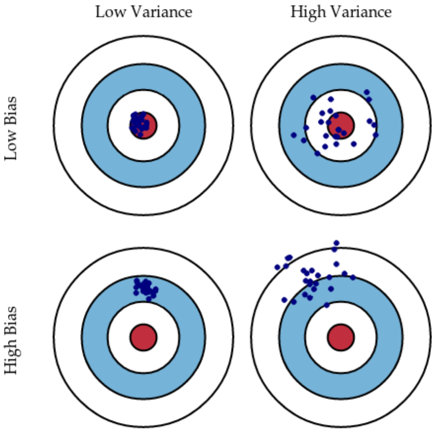
Comparison of Policy Evaluation Methods



MC vs TD Policy Evaluation

- MC:
  - Must wait until the end of an episode before updating $V^{\pi}(\mathbf{x})$
  - Value estimates are **zero bias but high variance** (long-term cost depends on *many* random transitions)
  - Not sensitive to initialization
  - Has good convergence properties even with function approximation (infinite state space)
- TD:
  - Can update $V^{\pi}(\mathbf{x})$ without complete episodes and hence can learn online after each transition
  - Value estimates are **biased but low variance** (the TD(0) target depends on *one* random transition but has bias from bootstrapping)
  - More sensitive to initialization than MC
  - May not converge with function approximation (infinite state space)

Bias-Variance Trade-off



Batch MC and TD Policy Evaluation

- **Batch setting**: given set of episodes $\{\rho^{(k)}\}_{k=1}^{K}$
  - Accumulate value function updates according to MC or TD for $k = 1, \ldots, K$
  - Update the value estimates **only** after a complete pass through all data
  - Repeat until the value function estimate converges

- **Batch MC**: converges to V^π that best fits the observed costs:

$$V^\pi(\mathbf{x}) \in \arg\min_V \sum_{k=1}^{K} \sum_{t=0}^{T_k} \left(L_t(\rho^{(k)}) - V \right)^2 \mathbb{1}\{\mathbf{x}_t^{(k)} = \mathbf{x}\}$$

- **Batch TD(0)**: converges to V^π of the maximum likelihood MDP model that best fits the observed data

$$\hat{p}_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) = \frac{1}{N(\mathbf{x}, \mathbf{u})} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \mathbb{1}\{\mathbf{x}_t^{(k)} = \mathbf{x}, \mathbf{u}_t^{(k)} = \mathbf{u}, \mathbf{x}_{t+1}^{(k)} = \mathbf{x}'\}$$

$$\hat{\ell}(\mathbf{x}, \mathbf{u}) = \frac{1}{N(\mathbf{x}, \mathbf{u})} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \mathbb{1}\{\mathbf{x}_t^{(k)} = \mathbf{x}, \mathbf{u}_t^{(k)} = \mathbf{u}\} \ell(\mathbf{x}_t^{(k)}, \mathbf{u}_t^{(k)})$$

Averaged-Return TD

- Define the *n*-step return:

$$
\begin{aligned}
L_t^{(n)}(\rho) &:= \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma \ell(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}) + \ldots + \gamma^n \ell(\mathbf{x}_{t+n}, \mathbf{u}_{t+n}) + \gamma^{n+1} V^\pi(\mathbf{x}_{t+n+1}) && \color{red}{TD(n)} \\
L_t^{(0)}(\rho) &= \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1}) && \color{red}{TD(0)} \\
L_t^{(1)}(\rho) &= \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma \ell(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}) + \gamma^2 V^\pi(\mathbf{x}_{t+2}) && \color{red}{TD(1)} \\
&\vdots \\
L_t^{(\infty)}(\rho) &= \ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma \ell(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}) + \ldots + \gamma^{T-t-1} \ell(\mathbf{x}_{T-1}, \mathbf{u}_{T-1}) + \gamma^{T-t} \mathfrak{q}(\mathbf{x}_T) && \color{red}{MC}
\end{aligned}
$$

- **TD(n)**:

$$
V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha(\color{red}{L_t^{(n)}(\rho)} - V^\pi(\mathbf{x}_t))
$$

- **Averaged-Return TD**: combines bootstrapping from several states:

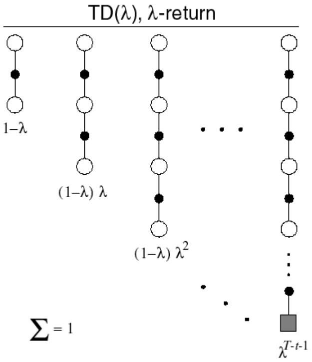
$$
V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha \left(\color{red}{\frac{1}{2} L_t^{(2)}(\rho) + \frac{1}{2} L_t^{(4)}(\rho)} - V^\pi(\mathbf{x}_t) \right)
$$

- Can we combine the information from all time-steps?

Forward-View TD(λ)

- **λ-return**: combines all n-step returns:

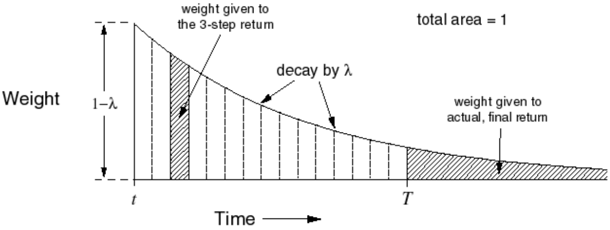
$$L_t^{\lambda}(\rho) = (1-\lambda) \sum_{n=0}^{T-t-2} \lambda^n L_t^{(n)}(\rho) + \lambda^{T-t-1} L_t^{(\infty)}(\rho)$$



- **Forward-View TD(λ)**:

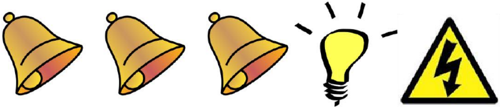
$$V^{\pi}(\mathbf{x}_t) \leftarrow V^{\pi}(\mathbf{x}_t) + \alpha \left(L_t^{\lambda}(\rho) - V^{\pi}(\mathbf{x}_t) \right)$$

- Like MC, the L_t^{λ} return can only be computed from complete episodes



Backward-View TD(λ)

- Forward-View TD(λ) is equivalent to TD(0) for $\lambda = 0$ and to every-visit MC for $\lambda = 1$
- Backward-View TD(λ) allows online updates from incomplete episodes
- **Credit assignment problem**: did the bell or the light cause the shock?



  - **Frequency heuristic**: assigns credit to the most frequent states
  - **Recency heuristic**: assigns credit to the most recent states
  - **Eligibility trace**: combines both heuristics

$$e_t(\mathbf{x}) = \gamma \lambda e_{t-1}(\mathbf{x}) + \mathbb{1}\{\mathbf{x} = \mathbf{x}_t\}$$

- **Backward-View TD(λ)**: updates in proportion to the **TD error** δ_t and the **eligibility trace** $e_t(\mathbf{x})$:

$$V^\pi(\mathbf{x}_t) \leftarrow V^\pi(\mathbf{x}_t) + \alpha \left(\ell(\mathbf{x}_t, \mathbf{u}_t) + \gamma V^\pi(\mathbf{x}_{t+1}) - V^\pi(\mathbf{x}_t) \right) e_t(\mathbf{x}_t)$$