Reinforcement Learning

Lecture 5:
Model-based Learning

Yishay Mansour, Tel-Aviv University

Yandex Distinguished Machine Learning Lecture

- Speaker: Ivan TItov (University of Edinburgh)
- Time: Wednesday, March 27th, 1pm (after class!)
- Place: The Steinhardt Museum of Natural History
- Title: Inducing and Modelling Discrete Structures in Natural Language Processing

Lecture 5: outline

- Learning off-policy
  - Building Model
- Inaccurate Model
  - Rewards
  - Dynamics
- Approximate Value Iteration
  - Better sampling for OPT
- Learning on-policy
  - Deterministic
  - Rmax

Markov Decision Process: review

- Finite state space:
  - $S = \{1, \dots, n\}$
- Finite action set:
  - $A = \{1, \dots, m\}$
- Transition function:
  - $p(s'|s, a)$

- Rewards:
  - $R(s, a)$
    - can be r.v.
  - $r(s, a) = E[R(s, a)]$
  - Assume:
    $R(s, a) \in [0, R_{max}]$

Finite Horizon Return: review

- Finite horizon return
  - Finite horizon
  - Parameter $T \geq 1$
  - $J_T^{\pi}(s_0) \triangleq E^{\pi}[\sum_{t=0}^{T} r(s_t, a_t) + r_T(s_T) | s_0 = s\ ]$

- Optimal return
  - $J_T^*(s_0) = \sup_{\pi \in HR} J_T^{\pi}(s_0) = \max_{\pi \in MD} J_T^{\pi}(s_0)$

Discounted Return: review

- Discounted return
  - Infinite horizon
  - Parameter $\gamma \in (0,1)$

$$J_\gamma^\pi(s_0) \triangleq E^\pi\left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) | s_0 = s\right]$$

- Optimal return
  - $J_\gamma^*(s_0) = \sup_{\pi \in HR} J_\gamma^\pi(s_0) = \max_{\pi \in SD} J_\gamma^\pi(s_0)$

Reducing discounted to finite

- Effective horizon
  - For any sequence $R_t \in [0, R_{max}]$
  - $\sum_{t=0}^{\infty} \gamma^t R_t - \sum_{t=0}^{T-1} \gamma^t R_t = \sum_{t=T}^{\infty} \gamma^t R_t \leq \frac{\gamma^T}{1-\gamma} R_{max}$
  - For $T \geq \frac{1}{1-\gamma} \log \frac{R_{max}}{\epsilon(1-\gamma)}$
    - The sum is bounded by ϵ
- Claim: The discounted return for finite horizon

$T \geq \frac{1}{1-\gamma} \log \frac{R_{max}}{\epsilon(1-\gamma)}$ is ϵ-optimal

Off-policy Model-based

Off-Policy Model-based learning

- Input
  - Sequence of (s, a, r, s')
    - $r \sim R(s, a)$
    - $s' \sim p(\cdot\,|s, a)$
- Output
  - Complete MDP model
    - $r(s, a)$ and $p(s'|s, a)$

Mean Estimation

- Assume we have a r.v.
  - $R \in [0, R_{max}]$
- Goal:
  - Estimate $E[R]$
- Tool: sampling
  - Get samples $R_1, \ldots, R_m$
  - Estimate using average: $\hat{r} = \frac{1}{m} \sum_{i=1}^{m} R_i$
- How good is the mean?
  - $\hat{r} \rightarrow_{m \rightarrow \infty} E[R]$
  - Law of large numbers
- How many samples do we need?
  - How to choose m?
  - Finite convergence bounds

Concentration bounds

- Let X_i be i.i.d. r.v.
  - $X_i \in [0,1]$
  - $\mu = E[X_i]$
- Observed average:
  - $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i$

- Chernoff/Hoeffding
  - Additive bound

$$\Pr[|\mu - \hat{\mu}| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$

  - Relative bounds:

$$\Pr[\hat{\mu} \leq (1-\varepsilon)\mu\,] \leq e^{-\frac{m\epsilon^2}{2}}$$

$$\Pr[\hat{\mu} \geq (1+\varepsilon)\mu] \leq e^{-\frac{m\epsilon^2}{3}}$$

    - $\varepsilon \in (0,1)$

Estimation: Sample size

- Let X_i be i.i.d. r.v.
  - $X_i \in [0, R_{max}]$
  - $\mu = E[X_i]$
- Observed average:
  - $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i$
- With probability $1 - \delta$
  - We have
    $$|\mu - \hat{\mu}| \leq \epsilon$$
  - For
    $$m \geq \frac{R_{max}^2 \log \frac{2}{\delta}}{2\epsilon^2}$$

Estimating the rewards

- Set $m = \frac{R_{max}^2 \log \frac{2|S|\ |A|}{\delta}}{2\epsilon^2}$
- For each state action (s, a)
  - use samples:
  - $R_1(s,a), \dots, R_m(s,a)$
- Estimate using average:

$$\hat{r}(s,a) = \frac{1}{m}\sum_{i=1}^{m} R_i(s,a)$$

- For each (s, a):

With probability $1 - \frac{\delta}{|S|\ |A|}$

$$|\hat{r}(s,a) - r(s,a)| \leq \epsilon$$

- Globally:
  - with probability $1 - \delta$
  - For every (s, a)
  - $|\hat{r}(s,a) - r(s,a)| \leq \epsilon$

Influence of estimation errors

- Finite Horizon:
- Fix a policy $\pi \in MD$
  - Compute the return under:
    - $r_t(s,a)$ and $\hat{r}_t(s,a)$
  - Would like to bound difference
    - As a function of ϵ and δ
- Assume that $|r_t(s,a) - \hat{r}_t(s,a)| \leq \epsilon$ and $|r_T(s) - \hat{r}_T(s)| \leq \epsilon$
  - W.p. $1-\delta$

Evaluating estimates: Finite Horizon

- Fix policy $\pi \in MD$:
  - $J_T^{\pi}(s_0) = E^{\pi, s_0}[\sum_{t=0}^{T-1} r_t(s_t, a_t) + r_T(S_T)]$
  - $\hat{J}_T^{\pi}(s_0) = E^{\pi, s_0}[\sum_{t=0}^{T-1} \hat{r}_t(s_t, a_t) + \hat{r}_T(S_T)]$

- $error(\pi) = |J_T^{\pi}(s_0) - \hat{J}_T^{\pi}(s_0)|$

Bounding policy error: Finite Horizon

- Fix policy $\pi \in MD$
- For each trajectory $\sigma = (s_0, a_0, \dots, s_{T-1}, a_{T-1}, s_T)$
- $|\sum_{t=0}^{T-1} r_t(s_t, a_t) + r_T(S_T) - \sum_{t=0}^{T-1} \hat{r}_t(s_t, a_t) + \hat{r}_T(S_T)|$
- $= \left|\sum_{t=0}^{T-1}\left(r_t(s_t, a_t) - \hat{r}_t(s_t, a_t)\right) + \left(r_T(S_T) - \hat{r}_T(S_T)\right)\right|$
- $\leq \sum_{t=0}^{T-1}|r_t(s_t, a_t) - \hat{r}_t(s_t, a_t)| + |r_T(S_T) - \hat{r}_T(S_T)|$
- $\leq \epsilon T + \epsilon$
- Therefore,
  - $error(\pi) \leq \epsilon(T + 1)$

Near-optimal policy: Finite Horizon

- For each (s, a) and s
  - Given sample of m rewards
    - where m $\geq \frac{R_{max}^2 \log 2|S||A|T/\delta}{2\epsilon^2}$
  - Compute $\hat{r}_t(s, a)$ and $\hat{r}_T(s)$
- With prob. $1 - \delta$: $|r_t(s, a) - \hat{r}_t(s, a)| \leq \epsilon$; $|\hat{r}_T(s) - \hat{r}_T(s)| \leq \epsilon$
  - implies $error(\pi) \leq \epsilon(T + 1)$ for any $\pi \in MD$
- Compute $\hat{\pi}^*$
  - the optimal policy w.r.t. $\hat{r}_t(s, a)$ and $\hat{r}_T(s)$

Near-optimal policy: Finite Horizon

- Let π^* be the true optimal policy

- Claim: $J_T^{\pi^*}(s_0) - J_T^{\hat{\pi}^*}(s_0) \leq 2\epsilon(T+1)$
- Proof:
  - $J_T^{\pi^*}(s_0)$-$\hat{J}_T^{\pi^*}(s_0) \leq error(\pi^*) \leq \epsilon(T+1)$
  - $\hat{J}_T^{\hat{\pi}^*}(s_0) - J_T^{\hat{\pi}^*}(s_0) \leq error(\hat{\pi}^*) \leq \epsilon(T+1)$
  - $\hat{J}_T^{\pi^*}(s_0) \leq \hat{J}_T^{\hat{\pi}^*}(s_0)$

Evaluating estimates: discounted

- Discounted infinite Horizon
- Fix policy $\pi \in SD:$
  - $J_\gamma^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t)]$
  - $\hat{J}_\gamma^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{\infty} \gamma^t \hat{r}(s_t, a_t)]$

- $error(\pi) = |J_\gamma^\pi(s_0) - \hat{J}_\gamma^\pi(s_0)|$

Bounding policy error: discounted

- Fix policy $\pi \in SD$
- For each trajectory $\sigma = (s_0, a_0, \dots)$
- $|\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - \sum_{t=0}^{\infty} \gamma^t \hat{r}(s_t, a_t)|$
- $= \left|\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \hat{r}(s_t, a_t)\right)\right|$
- $\leq \sum_{t=0}^{\infty} \gamma^t |r(s_t, a_t) - \hat{r}(s_t, a_t)|$
- $\leq \epsilon \sum_{t=0}^{\infty} \gamma^t = \frac{\epsilon}{1-\gamma}$
- Therefore
  - $error(\pi) \leq \frac{\epsilon}{1-\gamma}$

Computing near-optimal policy: discounted

- For each (s, a) and s
  - Given a sample m:
    - where $m = \frac{R_{max}^2 \log|S||A|/\delta}{2\epsilon^2}$
  - Compute $\hat{r}(s, a)$
- With prob. $1 - \delta$: $|r(s, a) - \hat{r}(s, a)| \leq \epsilon$
  - Need $error(\pi) \leq \frac{\epsilon}{1-\gamma}$ for any $\pi \in SD$
- Compute $\hat{\pi}^*$
  - the optimal policy w.r.t. $\hat{r}(s, a)$

Computing near-optimal policy: discounted

- Let π^* be the optimal policy
- Claim: $J_\gamma^{\pi^*}(s_0) - J_\gamma^{\hat{\pi}^*}(s_0) \leq \frac{2\epsilon}{1-\gamma}$
- Proof:
  - $J_\gamma^{\pi^*}(s_0) - \hat{J}_\gamma^{\pi^*}(s_0) \leq error(\pi^*) \leq \frac{\epsilon}{1-\gamma}$
  - $\hat{J}_\gamma^{\hat{\pi}^*}(s_0) - J_\gamma^{\hat{\pi}^*}(s_0) \leq error(\hat{\pi}^*) \leq \frac{\epsilon}{1-\gamma}$
  - $\hat{J}_\gamma^{\pi^*}(s_0) \leq \hat{J}_\gamma^{\hat{\pi}^*}(s_0)$

Lecture 5: outline

- Learning off-policy
  - Building Model
- Inaccurate Model
  - Rewards
  - Dynamics
- Approximate Value Iteration
  - Better sampling for OPT
- Learning on-policy
  - Deterministic
  - Rmax

Estimating the transitions

- For each state-action (s, a)
  - Have a sample (s, a, s_i') for $1 \leq i \leq m$
  - Define a distribution

$$\hat{p}(s_i'|s, a) = \frac{|\{(s,a,s_i')\}|}{m}$$

- How good is this model?

Distribution perturbation: function mean

- Theorem:
  - Let q_1, q_2 ditsributions
  - Let $f: S \to [0, F_{max}]$
  - Then
$$\left|E_{s\sim q_1}[f(s)] - E_{s\sim q_2}[f(s)]\right| \leq F_{max}\left|\left|q_1 - q_2\right|\right|_1$$

- Proof:
  - $\left|E_{s\sim q_1}[f(s)] - E_{s\sim q_2}[f(s)]\right|$
  - $= \left|\sum_s q_1(s)\, f(s) - \sum_s q_2(s) f(s)\right|$
  - $\leq \sum_s f(s)\, |q_1(s) - q_2(s)|$
  - $\leq F_{max} \sum_s |q_1(s) - q_2(s)|$
  - $= F_{max}\, \left|\left|q_1 - q_2\right|\right|_1$
- QED

Distribution perturbation: few simple facts

- Fact:
  - $\left|\left|z^\top M\right|\right|_1 \leq \left|\left|z\right|\right|_1 \left|\left|M\right|\right|_\infty$
    - where $\left|\left|M\right|\right|_\infty = \max_i \sum_j |M[i,j]|$
- Proof:
  - $\left|\left|z^\top M\right|\right|_1 \leq \sum_{i,j} |z_i| \ |M[i,j]|$
  - $= \sum_i |z_i| \sum_j |M[i,j]|$
  - $\leq \sum_i |z_i| \ \left|\left|M\right|\right|_\infty$
  - $= \left|\left|z\right|\right|_1 \left|\left|M\right|\right|_\infty$

- Fact 1:
  - For a distribution q, and $|M_1[i,j] - M_2[i,j]| \leq \alpha$
    - $\left|\left|q\right|\right|_1 = 1; \left|\left|M_1 - M_2\right|\right|_\infty \leq \alpha|S|$
  - $\left|\left|q^\top (M_1 - M_2)\right|\right|_1 \leq \alpha|S|$
- Fact 2:
  - For a row stochastic M
    - $\left|\left|M\right|\right|_\infty = 1$
  - $\left|\left|z^\top M\right|\right|_1 \leq \left|\left|z\right|\right|_1$

Distribution perturbation: states

- Consider two Markov chain
  - M_1 and M_2
  - $\forall i, j$: $|M_1[i,j] - M_2[i,j]| \leq \alpha$
- Let q_i^t state dist after t steps
  - $q_1^t = x_0 M_1^t$; $q_2^t = x_0 M_2^t$
- Theorem:
  - $\left| |q_1^t - q_2^t| \right|_1 \leq \alpha |S| t$

- Proof:
  - $\left| |q_1^t - q_2^t| \right|_1 = \left| |x_0 M_1^t - x_0 M_2^t| \right|_1$
  - $= \left| \left| q_1^{t-1} M_1 - q_2^{t-1} M_2 \right| \right|_1$
  - $\leq \left| \left| q_1^{t-1} (M_1 - M_2) \right| \right|_1 + \left| |z^\top M_2| \right|_1$
    - $z = q_1^{t-1} - q_2^{t-1}$
  - $\leq \alpha |S| + \alpha |S| (t-1)$
    - Fact 1 & Fact 2 with inductive claim

Approximate models: Definition

- Model $\widehat{M}$ is an α-approximation of M if:
- For every (s, a)
  - For rewards:

$$|\hat{r}(s,a) - r(s,a)| \leq \alpha R_{max}$$

  - For transitions
    - For very s'

$$|\hat{p}(s'|s,a) - p(s'|s,a)| \leq \alpha$$

Simulation Lemma: Finite Horizon

- Theorem
  - If model $\widehat{M}$ is an α-approximation of M:
  - Then for any policy $\pi \in MD$
  - $\left|J_T^{\pi}(s_0; M) - J_T^{\pi}(s_0; \widehat{M})\right| \leq \epsilon$
  - For $\alpha = O(\frac{\epsilon}{R_{max}|S|T^2})$
    - where T is the horizon

- Proof:
  - State distribution at time t
    - Differ by at most $\alpha|S|t$
  - Rewards bounded by R_{max}
  - Maximum difference
    - $\sum_t \alpha|S|\, t\, R_{max} \leq \alpha|S|T^2 R_{max}$
  - Rewards difference
    - At most $\alpha R_{max} T$

Simulation Lemma: discounted

- Theorem
  - If model $\widehat{M}$ is an α-approximation of M:
  - For any policy π
  - $\left|J_\gamma^\pi(s_0; M) - J_\gamma^\pi(s_0; \widehat{M})\right| \leq \epsilon$
  - For $\alpha = O\left(\frac{\epsilon(1-\gamma)^2}{R_{max}|S| \log^2 R_{max}/(\epsilon(1-\gamma))}\right)$
    - where γ is the discount factor

- Proof:
  - Reduce to the finite horizon
  - Horizon reduction error $\leq \frac{\epsilon}{2}$
  - Finite horizon error $\leq \frac{\epsilon}{2}$
- QED

Sample Size

- Objective: w. p. $1-\delta$ accuracy α
- Need to sample $\frac{1}{\alpha^2}\log\frac{|S|^2|A|}{\delta}$
- For finite horizon: $m=O(\frac{R_{max}^2|S|^2T^4}{\epsilon^2}\log\frac{|S||A|}{\delta})$
- For discounted: $m=(\frac{R_{max}^2|S|^2}{\epsilon^2(1-\gamma)^4}\log\frac{|S|\,|A|}{\delta}\log^2\frac{R_{max}}{\epsilon(1-\gamma)})$

Putting it together

- Sample each (s, a) for m times:
  - With probability $1 - \delta$ all errors $\leq \alpha$
- Build observed MDP $\widehat{M}$
- Solve for the optimal policy $\hat{\pi}^*$ in $\widehat{M}$
- This is a 2ϵ-optimal policy
  - $\left|V^* - V^{\hat{\pi}^*}\right| \leq 2\epsilon$
    - Proof, similar to the rewards only

Dependency on Parameters

- Error rate
  - Should scale like $\frac{R_{max}^2}{\epsilon^2}$
- Effective horizon
  - Should have a dependency
    - Probably not optimal
- Can we get a better dependence on $|S|$?

Value Iteration: optimal policy

- Recall Value iteration Algorithm:
  - Let $V_0 = \left(V_0(s)\right)_{s \in S}$
  - For $n = 0,1, \ldots$

$$V_{n+1}(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_n(s') \right\}$$

- Theorem: $\lim_{n \to \infty} V_n = V^*$
  - Convergence rate $O\left(\frac{\gamma^n}{1-\gamma} R_{max}\right)$

Approximate Value Iteration

- Run value iteration for $N = \frac{1}{1-\gamma}\log(\frac{R_{max}}{\epsilon(1-\gamma)})$
  - Error rate $= O(\frac{\gamma^N}{1-\gamma}R_{max}) \leq \epsilon$
- For iteration n state-action (s, a) approx:
  - $V_{n+1}(s) = max_a\, r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V_n(s')$
  - $= max_a r(s,a) + \gamma E_{s' \sim p(\cdot|s,a)}[V_n(s')]$
    - Replace expectation by sample
- How many samples we need? accuracy?

Approximate Value iteration

- Approximate Value iteration Algorithm:
  - Let $V_0 = 0$
  - For $n = 0\ to\ N$

$$\hat{V}_{n+1}(s) = \max_{a \in A} \left\{ \hat{r}(s,a) + \gamma \frac{1}{m} \sum_{i=1}^{m} \hat{V}_n(s_i') \right\}$$

    - where $s_i' \sim p(\cdot\,|s,a)$

- Note: we use only a sample to approximate the expectation

Approximate Value iteration

- Analysis: for a fixed (s, a) assume
  - $\left|E\left[\hat{V}_n(s')\right] - \frac{1}{m}\sum_{i=1}^{m} \hat{V}_n\left(s_i'\right)\right| \leq \epsilon$
  - $|\hat{r}(s,a) - r(s,a)| \leq \epsilon$
- If $|\hat{V}_n(s) - V_n(s)| \leq \lambda$ Then
  - $\left|\hat{V}_{n+1}(s) - V_{n+1}(s)\right| \leq \epsilon + \gamma\lambda \leq \lambda$
- Holds for $\lambda \geq \frac{\epsilon}{1-\gamma}$

Approximate Value iteration

- For $m = O\left(\frac{1}{\epsilon^2}\log\left(\frac{N\,|S||A|}{\delta}\right)\right)$
  - With probability $1-\delta$
  - For every $n \leq N$ and (s,a)
    - $\left|E\left[\hat{V}_n(s')\right] - \frac{1}{m}\sum_{i=1}^{m}\hat{V}_n\left(s_i'\right)\right| \leq \epsilon$
    - $|\hat{r}(s,a) - r(s,a)| \leq \epsilon$
- What did we gain? lose?
  - GAIN: in the dependency on $|S|$
  - LOSE: bound only for optimal policy

Lecture 5: outline

- Learning off-policy
  - Building Model
- Inaccurate Model
  - Rewards
  - Dynamics
- Approximate Value Iteration
  - Better sampling for OPT
- Learning on-policy
  - Deterministic
  - Rmax

On-policy Model-Based

Deterministic Decision Process

- Recall:
  - Directed graph $G(V, E)$
    - Strongly connected
  - $V = S,$
    - states are nodes
  - $E = \{(s, s') : \exists a : f(s, a) = s'\}$
  - Rewards as usual

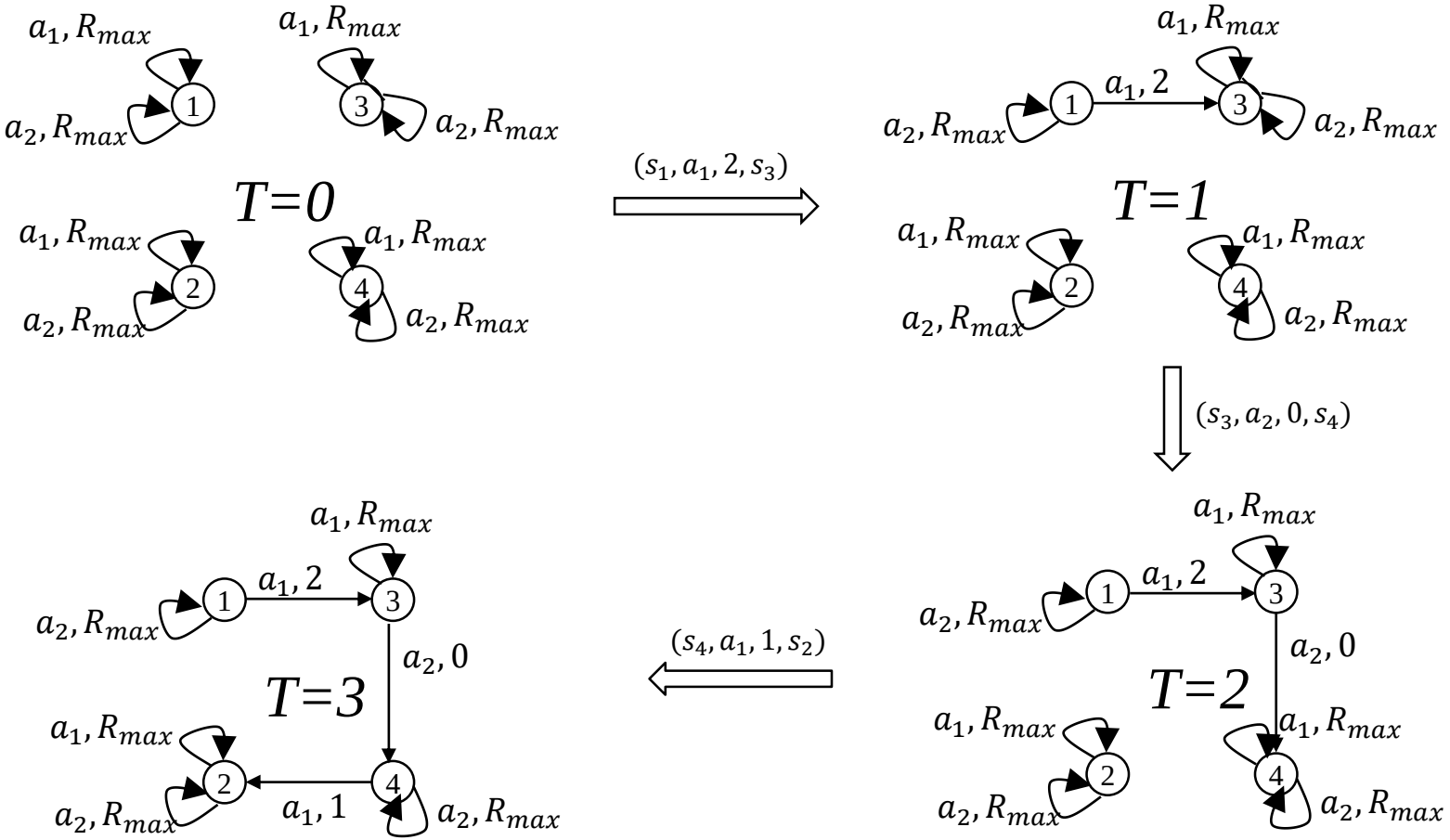
Learning DDP structure

- Given observed transitions from M
  - $Obs = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1,\ldots,T}$
- Define Observed DDP $\widehat{M}_T$
  - $\hat{f}(s_t, a_t) = s_{t+1}, \hat{r}(s_t, a_t) = r_t$
- Complete $\widehat{M}_T$ to a complete model
  - For $(s, a) \notin Obs$
    - set $\hat{f}(s, a) = s$ and $\hat{r}(s, a) = R_{max}$

Learning DDP

- Claim: $V^{\pi}\left(s; \widehat{M}_T\right) \geq V^{\pi}(s; M\ )$
- Actually holds for any trajectory
  - holds point-wise

Learning DDP: Example



Learning algorithm DDP

- At time T:
  - Compute $\widehat{M}_T$
  - Compute $\hat{\pi}_T^*$
    - The optimal policy for $\widehat{M}_T$
  - Let $a_T = \hat{\pi}_T^*(s_T)$
  - Do action a_T and observe r_T and s_{T+1}

Learning DDP: correctness

- Claim:

We change $\widehat{M}_T$ at most $|S| \cdot |A|$ times

- Proof:
  - Each change of $\widehat{M}_T$ discovers a new (s, a)
  - After we have all pairs (s, a) we are done.
- QED

Learning DDP: correctness

- Claim: either:
  - We change $\widehat{M}_T$ in some $t \in [T, T + |S|]$, or
  - We never change $\widehat{M}_T$
- Proof:
  - The process is deterministic
  - If no change for $|S|$ steps $\hat{\pi}_T^*$ closes a cycle
  - Never changes again
- QED

Learning DDP: correctness

- Claim: After $|S|^2|A|$ steps $\hat{\pi}_T^*$ is optimal
- Proof:
  - After $|S|^2|A|$ we converge:
    - Number of changes at most $|S|\,|A|$
    - Time between changes at most $|S|$
  - Once converged:
    - Return similar in $\widehat{M}$ and M
      - Return of $\hat{\pi}_T^*$ is optimal for $\widehat{M}_T$, versus π^* in M
- QED

DDP: return functions

- Generic construction:
  - works for multiple return functions
- Average reward:
  - Discovers all the structure
- Finite Horizon/discounted
  - Might stop early
  - Will compute an optimal policy
  - Uncertainty is in “uninteresting” region.

On-policy Learning MDP

- Similar to DDP
  - Be optimistic
    - Set unknown reward to maximum
- Optimism in face of uncertainty
  - Either
    - Explore: find out actual rewards
    - Exploit: Get near-optimal return

Learning MDP

- Similar to DDP
  - Partition (s, a) pairs:
    - Known
    - Unknown
  - Difference:
    - DDP: single sample enough
    - MDP: need multiple samples
      - Have a threshold when is enough

Learning MDP: Rmax

- Build MDP $\widehat{M}_0$
  - For all (s, a)
    - $p(s|s, a) = 1$,
    - $r(s, a) = R_{max}$
    - Unknown (s, a)
- Time t:
  - Build MDP $\widehat{M}_t$
  - Compute $\hat{\pi}_t^*$
  - Execute $a_t = \hat{\pi}_t^*(s_t)$
  - Observe r_T and s_{T+1}

- MDP $\widehat{M}_t$
  - For each (s, a):
  - If $\#(s, a) = m$
  - Update $\widehat{M}_{t-1}$
    - (s, a) becomes known
    - $r(s, a) = obs\ reward$
    - $p(s'|s, a) = obs\ trans$ition

Rmax: Intuition

- Consider discounted return
- Fix a horizon T
  - Large enough
- Assume we run policy $\hat{\pi}_t^*$,
  - for T steps

- Either:
  - Explore:
    - add new sample
  - Exploit:
    - near optimal
- If the probability of Explore low, we are near optimal return

Rmax: Analysis

- Consider the event

$$W = \{in\ the\ next\ T\ time\ steps\ visit\ an\ unknown\ (s,a)\}$$

- For $\hat{\pi}_t^*$ we have
  - $V^{\hat{\pi}_t^*} \geq V^* - \Pr[W]\frac{R_{max}}{1-\gamma} - \lambda$
    - λ is the error due to the approximation
      - Can be made small by making m large
      - For $m = poly\left(|S|, |A|, R_{max}, \frac{1}{1-\gamma}\right)$ we have $\lambda \leq \epsilon$

Rmax

- Analysis (sketch):
  - IF $\Pr[W] \leq \frac{\epsilon(1-\gamma)}{R_{max}}$ THEN $V^{\hat{\pi}_t^*} \geq V^* - \epsilon - \lambda$
    - Rmax is near optimal in such blocks
  - IF $\Pr[W] > \frac{\epsilon(1-\gamma)}{R_{max}}$ THEN
    - Good probability to visit unknown (s, a)
    - Can happen at most $m|S||A|$
    - Expected number of such blocks: $m|S||A|\frac{R_{max}}{\epsilon(1-\gamma)}$

Rmax

- Performance guarantee:
  - Number of times not ϵ-optimal at most

$$m|S||A|\frac{R_{max}}{\epsilon(1-\gamma)}$$

  - Not necessarily at the beginning.
- More guarantees can be given …

Lecture 5: outline

- Learning off-policy
  - Building Model
- Inaccurate Model
  - Rewards
  - Dynamics
- Approximate Value Iteration
  - Better sampling for OPT
- Learning on-policy
  - Deterministic
  - Rmax