Reinforcement Learning

Cheat Sheet

Saar Barak - TAU 2023 B

1 RECAP

Markov's inequality: $\forall$ non-negative random variable X and $a > 0$, it holds that $P[X \geq a] \leq \frac{E[X]}{a}$

Chebyshev's inequality: let X be a random variable with $E[X] = \mu < \infty$ and $0 \neq Var(X) = \sigma^2 < \infty$. It holds that $\forall k > 0 : P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$

Chernoff bound (Hoeffding inequality): let $R_1, ..., R_m$ be m i.i.d samples of a RV $R \in [0, 1]$. Let $\mu = E[R]$, $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m Ri$. For any $\epsilon \in (0, 1)$:

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

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

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

McDiarmid's inequality: let domain X, $f : X^n \to R$, $c_i = \max_{x,x_i,x_i'} |f(x_1, ..., x_i, ..., x_n) - f(x_1, ..., x_i', ..., x_n)|$

then: $P[|f(x) - E[f(x)]| \geq \epsilon] \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}$

Corollary: set $f = \bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$ where $x_i \in [0, 1]$ then $c_i = \frac{1}{n}$ and $P[|\bar{X} - E[\bar{X}]| \geq \epsilon] \leq 2e^{-2\epsilon^2 n}$ **Corollary:** set $f = wavg(x_1, ..., x_n) = \sum_{i=1}^n \beta_i x_i$ where $x_i \in [0, 1]$ then $c_i = \beta_i$ and

$$P[|wavg(X) - E[wavg(X)]| \geq \epsilon] \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n \beta_i^2}}$$

Exponential average:

$$\hat{\mu}_T = \frac{1}{T}\sum_{i=1}^T r_t = \hat{\mu}_{T-1} + \alpha_T(r_T - \hat{\mu}_{T-1}) = \sum_{t=1}^T \beta_t r_r$$

where$\beta_t = \alpha_t \prod_{j=1}^{t-1}(1 - \alpha_j)$

Integration by parts: $\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)dx$

Geom: $\sum_{k=0}^n r^k = \frac{1-r^{n+1}}{1-r}$, $\sum_{k=0}^\infty r^k = \frac{1}{1-r}$

2 Discrete Dynamic Systems

Definition. $t \in [T]$ infinite or finite, S_t is the set of all possible states, A_t is the set of possible control actions, $f_t : S_t \times A_t \to S_{t+1}$ is the state transition function: $s_{t+1} = f_t(s_t, a_t)$, $s_t \in S_t$, $a_t \in A_t$

3 DDP - Finite Horizon

Optimal Control Policies

Algorithm 1 Finite-Horizon DP (value iteration)

1: Initialize the value function $C_T(s) = c_T(s), \forall s \in S_T$
2: Backward recursion:
For $t = T - 1, \ldots, 0$ compute: $\forall s \in S_t$

$$C_t(s) = \min_{a \in A_t} c_t(s, a) + C_{t+1}(f_t(s, a))$$

3: Optimal policy: Choose any $\pi^* = (\pi_t^*)$ that satisfies:$\pi_t^*(s)$

$$\arg\min_{a \in A_t} c_t(s, a) + C_{t+1}(f_t(s, a))$$

Proposition: the following holds for finite-horizon DP value iteration algorithm: $C_0(s) = \min_\pi C_0(\pi; s)$

4 DDP - Average cost

Avg cost criteria: $C_{avg}^\pi = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} c_t(s_t, a_t)$. The aim is to minimize $E[C_{avg}^\pi]$.

Claim. for a deterministic stationary policy, the policy converges to a simple cycle, and the avg cost is the avg cost of the edges on the cycle

Minimum Avg Cost Cycle: Given a directed graph $G(V, E)$,let Ω be the collection of all cycles in $G(V, E)$. For each cycle $\omega = (v_1, \ldots, v_k)$, we define $c(\omega) = \sum_{i=1}^k c(v_i, v_{i+1})$, where (v_i, v_{i+1}) is the i^{th} edge in the cycle ω. Let $\mu(\omega) = \frac{c(\omega)}{k}$. The minimum avg cost cycle is $\mu^* = \min_{\omega\in\Omega} \mu(\omega)$. $(|V| = n)$

Theorem. For any DDP the optimal avg cost is μ^*, and the policy is π_ω that cycles around a simple cycle of avg cost μ^*, where μ^* is the minimum avg cost cycle

Definition. State i is *recurrent* if $\mathbb{P}(X_t = i \text{ for some } t \geq 1|X_0 = i) = 1$. O.W, the state is transient.

Claim. State i is transient if and only if $\sum_{m=1}^\infty \mathbb{P}_i^m < \infty$. Recurrence is a class property. If states i and j are in the same class, then $\mathbb{P}(X_t = j \text{ for some } t \geq 1|X_0 = i) = 1$. Let T_i be the return time to state i (i.e., the number of stages required for (X_t) starting from state i to first return to i). If i is a recurrent state, then $\mathbb{P}(T_i < \infty) = 1$. State i is positive recurrent if $\mathbb{E}(T_i) < \infty$, and null recurrent if $\mathbb{E}(T_i) = \infty$. If the state space X is finite, all recurrent states are positive recurrent.

Theorem. The probability vector $\mu = (\mu_i)$ is an invariant/stationary distribution for the Markov chain if $\mu^\top P = \mu^\top$, namely $\forall j, \mu_j = \sum_i \mu_i p_{ij}$. The probability vector $\mu = (\mu_i)$ is an invariant/stationary distribution for the Markov chain if $\mu^\top P = \mu^\top$, namely $\forall j, \mu_j = \sum_i \mu_i p_{ij}$.

Theorem. Let (X_t) be an irreducible and aperiodic Markov chain over a finite state space X with transition matrix P. Then there is a unique distribution μ such that $\mu^\top P = \mu^\top > 0$. Moreover, for any $j \in X$, we have $\mu_j = \frac{1}{\mathbb{E}[T_j]}$ (on average, state i appears every $\mathbb{E}[T_j] < \infty$ steps). Furthermore, all states are positive recurrent. If the Markov chain is over a countable state space, then all states are either positive/null recurrent or transient

Definition. The ***Total Variation distance*** between distributions D_1 and D_2 is defined as $||D_1 - D_2||_{TV} = \sum_{x\in X} |D_1(x) - D_2(x)|$. The mixing time is the smallest m such that $||s_0^\top P^m - \mu||_{TV} \leq \frac{1}{4}||s_0 - \mu||_{TV}$ where s_0 is the initial state distribution and μ the steady-state distribution. For $k \geq m$, we have $||s_0^\top P^k - \mu||_{TV} \leq \frac{1}{4^k}||s_0 - \mu||_{TV}$.

5 MDP

Markov Decision Process

Controlled Markov chains: $\mathcal{T} = \{0, \ldots, T - 1\}$ (finite horizon) or infinite horizon, $\mathcal{S}$ finite state space, where $S_t \subseteq \mathcal{S}$, $\mathcal{A}$ finite action space where $A_t(s) \subseteq \mathcal{A}$, $s \in S_t$, $\mathcal{P}_t(\cdot|s, a)$ transition probability from state s to state s' given action a (distribution) that is $\mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a) = \mathcal{P}_t(s'|s, a)$. States $s' \in S_{t+1}$.

The Induced Stochastic Process: Let $\mathcal{P}_0(s_0)$ be the initial state distribution, $\pi \in \Pi_{\mathcal{H}_S}$, they induce a probability distribution over any finite state-action sequence $\mathcal{H}_T = (s_0, a_0, \ldots, s_T)$, given by

$$\mathbb{P}(\mathcal{H}_t) = \mathcal{P}_0(s_0)\prod_{t=0}^{T-1} \mathcal{P}_t(s_{t+1}|s_t, a_t)\pi(a_t|\mathcal{H}_t),$$

where $\mathcal{H}_t = (s_0, a_0, \ldots, s_t)$.

Markov policy induces a Markov chain: $\mathbb{P}(s_{t+1} = s'|s_t = s) = \sum_a \mathcal{P}_t(s'|s, a) \cdot \pi_t(a|s)$.

Finite horizon:

$$\mathcal{V}_T^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=0}^T R_t|s_0 = s\right] \equiv \mathbb{E}^{\pi,s}\left[\sum_{t=0}^T R_t\right]$$
$$= \sum_{t=0}^{T-1} \mathbb{E}_\pi[r_t(s_t, a_t)|s_0 = s] + \mathbb{E}_\pi[r_T(s_T)|s_0 = s]$$

Infinite horizon: $\mathcal{V}_\gamma^\pi(s) = \sum_{t=0}^\infty \mathbb{E}_\pi[\gamma^t r_t(s_t, a_t)|s_0 = s]$.

Claim. if $r(s, a) \in [0, 1]$ then $0 \leq V_\gamma^\pi(s) \leq \frac{1}{1-\gamma}$ and the sum of rewards after $t \geq \log_\gamma \varepsilon$ contribute at most $\frac{\varepsilon}{1-\gamma}$

Extended rewards: $R_t = \tilde{r}_t(s_t, a_t, s_{t+1})$. $r_t(s, a) = \mathbb{E}[R_t|s_t = s, a_t = a] = \sum p(s_{t+1}|s_t, a_t) \cdot s_{t+1}\tilde{r}_t(s_t, a_t, s_{t+1})$

Lemma. (finite horizon policy evaluation): let $\pi = (\pi_0, \ldots, \pi_{T-1}) \in \Pi^{MD}$. Define $V_k^\pi(s) = \mathbb{E}^\pi[\sum_{t=k}^T R_t|s_k = s]$ $(V_0^\pi(s) = V^\pi(s))$ then $V_k^\pi(s) = r_k(s, \pi_k(s)) + \sum p_k(s'|s, \pi_k(s))V_{k+1}^\pi(s')$. $s' \in S_{k+1}$, $\forall s \in S_k$ for $k = T - 1, \ldots, 0$ starting with $V_T^\pi(s) = r_T(s)$

Claim. For any policy π, $|V_\gamma^\pi(s)| \leq \frac{R_{\text{Max}}}{1-\gamma}$, where $R_{\text{Max}} \geq |r(s_t, a_t)|$

Lemma For a fixed stationary policy π, the value function $\mathcal{V}_\pi$ satisfies the following set of $|S|$ linear equations:

$$\mathcal{V}_\pi(s) = r(s, \pi(s)) + \gamma\sum_{s'} \mathcal{P}_\pi(s'|s)\mathcal{V}_\pi(s') \text{ for }{}^1$$

Algorithm 2 Finite-Horizon MDP (value iteration)

1: Set $V_T(s) = r_T(s)$ for $s \in S_T$.
2: For $k = T - 1, \ldots, 0$ Compute $V_k(s)$ using the following recursion:

$$V_k(s) = \max_{a\in A_k} r_k(s, a) + \sum_{s'\in S_{k+1}} p_k(s' \mid s, a)V_{k+1}^\pi(s')$$

Where $s \in S_k$. We have $V_k(s) = V_k^*(s)$.

3: Optimal policy: Any Markov policy π^* that satisfies, for $t = 0, \ldots, T - 1$

$$\pi_t^*(s) \in \arg\max_{a\in A_k} r_k(s, a) + \sum_{s'\in S_{k+1}} p_k(s' \mid s, a)V_{k+1}^\pi(s')$$

for all $s \in S_t$, is an optimal policy. Furthermore, π^* maximizes $V_\pi(s_0)$ simultaneously for every initial state $s_0 \in S_0$.

Definition. The Q function:

$$Q^\pi(s, a) = \sum_{s'} p(s' \mid s, a)\left[r(s, a, s') + \gamma V^\pi(s')\right]$$

$$Q_k^*(s, a) = r_k(s, a) + \sum_{s'} p_k(s' \mid s, a)V_{k+1}^*(s').$$

Claim. $V^\pi(s) = \sum_{a\in A} \pi(a \mid s)Q^\pi(s, a)$ for all $s \in S$.

Claim. $\mathcal{V}^*(s) = \max_a Q^*(s, a)$, $\pi_k^*(s) \in \arg\max_{a\in A_k} Q_k^*(s, a)$.

Definition. ***Infinite horizon problems:*** same setting as before (now $\mathbb{E}[R_t] = r(s_t, a_T)$) with discounted return:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \mid s_0 = s\right]$$
$$\equiv \mathbb{E}_{\pi,s}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)\right] \text{ for } \gamma \in (0, 1).$$

Claim. For any policy π, $|V^\pi(s)| \leq \frac{R_{\text{Max}}}{1-\gamma}$, where $R_{\text{Max}} \geq |r(s, a)|$.

Lemma. For $\pi \in \Pi_S^D$, the value function V^π satisfies the following set of $|S|$ linear equations:

$$V^\pi(s) = r(s, \pi(s)) + \gamma\sum_{s'} p(s' \mid s, \pi(s))V^\pi(s'),$$

or in vector form: $V^\pi = r_\pi + \gamma P_\pi V^\pi$.

Lemma. The set of linear equations in the lemma above has a unique solution V^π, which is given by

$$V^\pi = (I - \gamma P_\pi)^{-1}r_\pi.$$

Algorithm 3 Discounted Policy Eval

1: Let $\mathcal{V}_0 = (\mathcal{V}_0(s))_{s\in S}$ be arbitrary.
2: **for** $n = 0, 1, \ldots$
3: $\mathcal{V}_{n+1}(s) = r(s, \pi(s)) + \gamma\sum_{s'} p(s' \mid s, \pi(s))\mathcal{V}_n(s')$ for all $s \in S$. or equivalently $\mathcal{V}_{n+1} = r^\pi + \gamma P\mathcal{V}_n$

Proposition: We have $\lim_{n\to\infty} \frac{\mathcal{V}_n(s)}{\mathcal{V}(s)} = \mathcal{V}_\pi(s) \quad \forall s \in \mathcal{S}$.

Theorem. (Bellman Optimality Equation) For $V^\gamma(s) = \sup_{\pi\in\Pi_{H,S}} V^\pi(s)$: (1) V^* is the unique solution of the following set of (nonlinear) equations:

$$V(s) = \max_a\left\{r(s, a) + \gamma\sum_{s'} p(s' \mid s, a)V(s')\right\}.^1$$

(2) Any stationary policy π^* that satisfies

$$\pi^*(s) \in \arg\max_{a\in A}\left\{r(s, a) + \gamma\sum_{s'} p(s' \mid s, a)V(s')\right\}$$

is an optimal policy (for any initial state $s_0 \in S$).

Definition. For a fixed stationary policy $\pi : S \to A$, define fixed policy DP operator $T_\pi : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ as follows: for all $V = (V(s)) \in \mathbb{R}^{|S|}, \forall s \in S, (T_\pi(V))(s) = r(s, \pi(s)) + \gamma\sum_{s'\in S} p(s'|s, \pi(s))V(s')$. In our column-vector notation: $T_\pi(V) = r_\pi + \gamma P_\pi V$.

Theorem. $T \in \{T^*, T_\pi\}$ is a γ-contraction operator with respect to the max-norm, namely $\|T(\mathcal{V}_1) - T(\mathcal{V}_2)\|_\infty \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$.

Lemma. If $\|\mathcal{V}_{n+1} - \mathcal{V}_n\|_\infty < \frac{\epsilon(1-\gamma)}{2\gamma}$ then $\|\mathcal{V}_{n+1} - \mathcal{V}^*\|_\infty < \frac{\epsilon}{2}$ and $|\mathcal{V}_{\pi_{n+1}} - \mathcal{V}^*| \leq \epsilon$ where π_{n+1} is the greedy policy w.r.t $\mathcal{V}_{n+1}$.

Claim. $T_\pi(\mathcal{V}_\pi) = \mathcal{V}_\pi$, $T_{\pi_n}(\mathcal{V}_n) = T^*(\mathcal{V}_n)$, $T_{\pi_n}(\mathcal{V}_{\pi_n}) \leq T^*(\mathcal{V}_{\pi_n})$, $T_{\pi_{n+1}}(\mathcal{V}_{\pi_n}) = T^*(\mathcal{V}_{\pi_n})$ $(\pi_n, \pi_{n+1}$ greedy w.r.t $\mathcal{V}_n, \mathcal{V}_{n+1})$.

Theorem. We have $\lim_{n\to\infty} \mathcal{V}_n = \mathcal{V}^*$. The rate of convergence is exponential at rate $\mathcal{O}(\gamma^n)$.

Algorithm 4 Policy Iteration

1: choose some stationary policy π_0.
2: Policy Evaluation: computee $\mathcal{V}_{\pi_k}$.
3: Policy Improvement: compute π_{k+1}:

$$\pi_{k+1}(s) \in \arg\max_{a\in A}\{r(s, a) + \gamma\sum_{s'\in S} p(s'|s, a)\mathcal{V}_{\pi_k}(s')\}$$

4: Stop if $\pi_{k+1} = \pi_k$ (or if $\mathcal{V}_{\pi_k}$ satisfied the optimality equation), else continue.

Theorem. The following statements hold: (1)$\mathcal{V}_{\pi_{k+1}} \geq \mathcal{V}_{\pi_k}$, (2) $\mathcal{V}_{\pi_{k+1}} = \mathcal{V}_{\pi_k}$ iff π_k is an optimal policy (3)π_k converges after a finite number of steps since the number of stationary policies is finite.

Theorem. Let $\{V_{I_n}\}$ be the sequence of values created by the VI algorithm (where $V_{I_{n+1}} = T^*(V_n)$) and $\{P_{I_n}\}$ be the sequence of values created by PI algorithm, i.e., $P_{I_n} = \ldots$

6 Contraction Operators

Definition. the operator T is called a **contraction operator** if $\exists\beta \in (0, 1)\forall v_1, v_2 \in \mathbb{R}^d.\|T(v_1)-T(v_2)\| \leq \beta\|v_1-v_2\|$

Theorem. let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a contraction operator. Then (1) The equation $T(v) = v$ has a unique solution $V^* \in \mathbb{R}^d$ (2) $\forall v_0 \in \mathbb{R}^d. \lim_{n\to\infty} T^n(v_0) = V^*$. In fact, $\|T^n(v_0) - V^*\| \leq \mathcal{O}(\beta^n)$

7 Model Based - Off policy

Theorem. Given a discount factor γ, the discounted return in the first $T = \frac{1}{1-\gamma}\log\frac{R_{\text{Max}}}{\epsilon(1-\gamma)}$ time steps, is within ϵ of the total discounted return.

Off policy: Input – sequences of (s, a, r, s') where $r \sim R(s, a)$, $s' \sim p(\cdot|s, a)$. Output – complete MDP model i.e. $r(s, a)$, $p(s'|s, a)$.

Claim. Set $m = \frac{R_{\text{Max}}^2}{2\epsilon^2}\log\frac{2|S|\cdot|A|}{\delta}$. For each (s, a) use samples $R_1(s, a), \ldots, R_m(s, a)$. Set $\hat{r}(s, a) = \frac{1}{m}\sum_{i=1}^m R_i(s, a)$. Then for each (s, a): $P[|\hat{r}(s, a) - r(s, a)| \leq \epsilon] \geq 1 - \frac{\delta}{|S|\cdot|A|}$. Globally: for all (s, a): $P[|\hat{r}(s, a) - r(s, a)| \leq \epsilon] \geq 1 - \delta$.

Influence of reward estimation errors: **Finite horizon:** Assume for every (s, a) and t we have $|r^t(s, a) - \hat{r}^t(s, a)| \leq \epsilon$ and $|r^T(s) - \hat{r}^T(s)| \leq \epsilon$. Define $V_\pi^T(s_0) = E_{\pi,s_0}\left[\sum_{t=0}^T r^t(s_t, a_t) + r^T(s_T)\right]$ and $\hat{V}_\pi^T(s_0) = E_{\pi,s_0}\left[\sum_{t=0}^T \hat{r}^t(s_t, a_t) + \hat{r}^T(s_T)\right]$. Let error$(\pi) = |V_\pi^T(s_0) - \hat{V}_\pi^T(s_0)|$. Then for any $\pi \in \Pi_{MD}$ we have error$(\pi) \leq \epsilon(T + 1)$.

Influence of reward estimation errors
Discounted return: Assume for every (s, a) and t we have $|r^t(s, a) - \hat{r}^t(s, a)| \leq \epsilon$. Define $V_\pi^\gamma(s_0) = E_{\pi,s_0}\left[\sum_{t=0}^T \gamma^t r^t(s_t, a_t)\right]$ and $\hat{V}_\pi^\gamma(s_0) = E_{\pi,s_0}\left[\sum_{t=0}^T \gamma^t\hat{r}^t(s_t, a_t)\right]$. Let error$(\pi) = |V_\pi^\gamma(s_0) - \hat{V}_\pi^\gamma(s_0)|$. Then for any $\pi \in \Pi_{SD}$ we have error$(\pi) \leq \frac{\epsilon}{1-\gamma}$.

Computing approximate optimal policy
Finite horizon: we need to sample $m \geq \frac{R_{\text{Max}}^2}{2\epsilon^2}\log\frac{2|S|\cdot|A|}{\delta}$ for each $RV R^t(s, a)$ $(R^T(s)$ in finite). Given the sample, we compute $\hat{r}^t(s, a)$ ($\hat{r}^T(s)$ in finite). Now compute $\hat{\pi}^*_{\text{optimal policy}}$ w.r.t the estimated rewards.

Theorem. Assume that for $\forall(s, a), t : |r^t(s, a) - \hat{r}^t(s, a)| \leq \epsilon$ $(\forall s : |r^T(s) - \hat{r}^T(s)| \leq \epsilon$ in finite). Then, $V_{\pi^*}^T(s_0) - V_{\hat{\pi}^*}^T(s_0) \leq 2\epsilon(T + 1)$ for finite and $V_{\pi^*}^\gamma(s_0) - V_{\hat{\pi}^*}^\gamma(s_0) \leq \frac{2\epsilon}{1-\gamma}$ for discounted.

Estimate the transitions: $\forall(s, a)$ consider m i.i.d transitions (s, a, s_i') for $i \in [m]$. Then $\hat{p}(s'|s, a) = \frac{|\{i|s_i'=s'\}|}{m}$.

Theorem. Let q_1, q_2 be distributions. Let $f : S \to [0, F_{\text{Max}}]$, then $|E_{s\sim q_1}[f(s)] - E_{s\sim q_2}[f(s)]| \leq F_{\text{Max}}\|q_1 - q_2\|_1$.

Claim. $\|z^T M\|_1 \leq \|z\|_1\|M\|_{\infty,1}$ where $z \in \mathbb{R}^n, M \in \mathbb{R}^{n,n}$ and $\|M\|_{\infty,1} = \max_i \sum_j |M[i, j]|$.

Corollary. For a distribution q and $|M1_{i,j} - M2_{i,j}| \leq \alpha$ $(\|q\|_1 = 1$, $\|M1 - M2\|_{\infty,1} \leq \alpha|S|)$ it holds $\|q^T(M1 - M2)\|_1 \leq \alpha|S|$

Corollary. For a row stochastic M: $\|M\|_{\infty,1} = 1$, $\|z^T M\|_1 \leq \|z\|_1$

Theorem. Consider two Markov chains $M1, M2$ s.t $\forall i, j : |M1_{i,j} - M2_{i,j}| \leq \alpha$. Let q_i^t be the state distribution after t steps i.e. $q_1^t = p_0^T M_1^t, q_2^t = p_0^T M_2^t$, then $\|q_1^t - q_2^t\|_1 \leq \alpha|S|^t$

Definition. Model $\hat{M}$ is an ***α-approx*** of M if: $\forall(s, a).((|\hat{r}(s, a) - r(s, a)| \leq \alpha R_{\text{Max}}) \wedge (\forall s'.|\hat{p}(s'|s, a) - p(s'|s, a)| \leq \alpha))$

Theorem. Let $\hat{M}$ be an ***α-approx*** of M. If $\alpha = O\left(\frac{\varepsilon}{R_{\text{Max}}|S|T^2}\right)$ then $\forall\pi \in MD.|V_\pi^T(s_0; M) - V_\pi^T(s_0; \hat{M})| \leq \varepsilon$ where T is finite

Theorem. Let $\hat{M}$ be an ***α-approx*** of M. If $\alpha = O\left(\frac{\varepsilon(1-\gamma)^2}{R_{\text{Max}}|S|\log_2\left(\frac{R_{\text{Max}}}{\varepsilon(1-\gamma)}\right)}\right)$ then $\forall\pi \in MD.|V_\pi^\gamma(s_0; M) - V_\pi^\gamma(s_0; \hat{M})| \leq \varepsilon$

Theorem. Sample each (s, a) for m times where $m = \frac{1}{\alpha^2}\log\left(\frac{|S^2||A|}{\delta}\right)$ then w.p $(1 - \delta)$ all errors $\leq \alpha$. $m = O\left(\frac{R_{\text{Max}}^2|S|^2T^4}{\varepsilon^2}\log\left(\frac{|S||A|}{\delta}\right)\right)$ for finite horizon, $m = O\left(\frac{R_{\text{Max}}^2|S|^2}{\varepsilon^2(1-\gamma)^4}\log\left(\frac{|S||A|}{\delta}\log_2\left(\frac{R_{\text{Max}}}{\varepsilon(1-\gamma)}\right)\right)\right)$ for discounted Next, build observed MDP $\hat{M}$. Solve for the optimal policy $\hat{\pi}_*$ in $\hat{M}$. This is a 2ε-optimal policy $V_* - V_{\hat{\pi}_*} \leq 2\varepsilon$

Algorithm 5 Approximate V.I

1: Let $V_0 = 0$.
2: **for** $n = 0, \ldots N$:

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

where $s_i' \sim p(\bullet|s, a)$ and $\hat{r} = \frac{1}{m}\sum_{i=1}^m r_i$.

Theorem. For $m = O\left(\frac{R_{\text{Max}}^2}{\varepsilon^2}\log(N|S||A|/\delta)\right)$ w.p $(1 - \delta)$ $\forall n \leq N$ and $(s, a) : |E[\hat{V}_n(s')] - \frac{1}{m}\sum_{i=1}^m \hat{V}_n(s_i')| \leq \varepsilon$ and $|\hat{r}(s, a) - r(s, a)| \leq \varepsilon$.

GAIN: in the dependency of $|S|$. **LOSE:** bound only for optimal policy

8 Model Based - On policy

Definition. Model Based – On policy Learning DDP structure: given observed transition from M: obs $= \{(st, at, rt, st + 1)\}_{t=1,\ldots,T}$, define observed DDP $\hat{M}_T$: $\hat{f}(st, at) = st + 1$, $\hat{r}(st, at) = rt$. For all $(s, a) \notin$ obs set $\hat{f}(s, a) = s$ and $\hat{r}(s, a) = RMax$.

Claim. $V_\pi(s; \hat{M}_T) \geq V_\pi(s; M)$

Algorithm 6 Leaning DDP:

1: at time T: compute $\hat{M}_T$, compute $\hat{\pi}_T^*$ the optimal policy for $\hat{M}_T$
2: Let $a_T = \hat{\pi}_T(s_T)$. Do action a_T and observe r_T and s_{T+1}.

Claim. We change $\hat{M}_T$ at most $|S| \cdot |A|$ times.

[1]forall $s \in S$