CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 12

Sharan Vaswani

November 24, 2023

Trust Region Policy Optimization

- Both softmax PG and NPG need to compute the policy gradient for each update to the policy. In scenarios where computing the (approximate) PG is computationally expensive (e.g. involves interaction with a real-world environment or an expensive simulator), these methods can be inefficient.
- PG methods used in practice use the policy gradient to iteratively construct *surrogate functions*, and update the policy parameters in order to maximize these surrogates.
- While forming the surrogate function requires computing the policy gradient, maximizing it and updating the policy parameters does not. Hence, these PG methods can do multiple parameter updates and better re-use the data acquired from the environment.
- Trust Region Policy Optimization (TRPO) is one of the most common PG methods that iteratively constructs such a surrogate function.

Trust Region Policy Optimization

Given a set of feasible policies Π_θ (e.g. those that can be expressed using a model parameterized by θ), TRPO maximizes the following surrogate function (β, δ are parameters) at iteration t:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} h_t(\pi) := \left[v^{\pi_t}(\rho) + \frac{\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma} - \beta \max_s \mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \right] \textbf{(v1)}$$

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} h_t(\pi) := \left[v^{\pi_t}(\rho) + \frac{\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma} \right] \text{ s.t. } \mathbb{E}_{s\sim d^{\pi_t}} \mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \le \delta \textbf{(v2)}$$

- The set Π_θ depends on the policy parameterization, and solving $\max_{\pi\in\Pi_\theta} h_t(\pi)$ by an iterative method such as gradient ascent results in multiple policy updates.
- Theoretical guarantees are proved for **(v1)**, whereas **(v2)** is used in practice (using **(v1)** in practice results in overly conservative updates.)
- β in **(v1)** will be determined theoretically, whereas δ in **(v2)** needs to be tuned empirically.
- For the tabular parameterization, using a linear approximation of $\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)$ and a quadratic approximation of $\mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s))$ leads to a closed-form solution and the resulting update is the same as NPG. (Prove in Assignment 4!).

Trust Region Policy Optimization

- Given exact estimates of the advantage, **(v1)** has monotonic policy improvement guarantees, i.e. $v^{\pi_{t+1}}(\rho) \geq v^{\pi_t}(\rho)$ for all t. Since the function is upper-bounded from above by $\frac{1}{1-\gamma}$, **(v1)** results in convergence to a local maximum.
- Proving monotonic policy improvement relies on the fact that:
  (i) $h_t(\pi)$ is a minorization of $v^{\pi}(\rho)$, meaning that, for all π, $v^{\pi}(\rho) \geq h_t(\pi)$,
  (ii) the inequality is tight at π_t, i.e. $h_t(\pi_t) = v^{\pi_t}(\rho)$.
  (iii) Since π_{t+1} is the maximizer of $h_t(\pi)$, $h_t(\pi_{t+1}) \geq h_t(\pi_t)$.
  Putting these results together,

$$v^{\pi_{t+1}}(\rho) \overset{(i)}{\geq} h_t(\pi_{t+1}) \overset{(iii)}{\geq} h_t(\pi_t) \overset{(ii)}{=} v^{\pi_t}(\rho)$$

  In order to show monotonic policy improvement for **(v1)**, we now show that $v^{\pi}(\rho) \geq h(\pi)$.

Trust Region Policy Optimization

Claim: For any policies π and $\tilde{\pi}$, $\beta = \frac{4\gamma}{(1-\gamma)^3}$,

$$v^{\pi}(\rho) \geq h(\pi) := \left[v^{\tilde{\pi}}(\rho) + \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \beta \max_s \mathrm{KL}(\tilde{\pi}(\cdot|s)||\pi(\cdot|s)) \right].$$

For iteration t of TRPO, $\tilde{\pi} = \pi_t$ and hence $h(\pi) = h_t(\pi)$.

Proof: The proof relies on the following lemma that bounds the difference in the values of arbitrary policies $\pi, \tilde{\pi}$: $v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi}} \mathbb{E}_{a\sim\pi(\cdot|s)}[\mathfrak{a}^{\tilde{\pi}}(s,a)]$ (Prove in Assignment 4!).

$$\begin{aligned}
v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\pi}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \\
&= \frac{1}{1-\gamma} \left[\mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) + \mathbb{E}_{s\sim d^{\pi}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \right] \\
&\qquad \text{(Add/Subtract } \mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a)) \\
&\geq \frac{1}{1-\gamma} \left[\mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \left| \mathbb{E}_{s\sim d^{\pi}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{s\sim d^{\tilde{\pi}}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \right| \right] \\
&\qquad \text{(Since } x \geq -|x|) \qquad \text{4}
\end{aligned}$$

Trust Region Policy Optimization

$$v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{1}{1-\gamma}\left[\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a) - |\mathbb{E}_{s\sim d^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)|\right].$$

$$v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{|\max_s\left\{\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)\right\}|\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1}{1-\gamma}$$

(By Holder's inequality, $|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]| \leq |\max_x f(x)|\,\|P-Q\|_1$)

$$= \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{|\max_s\left\{\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)\right\}|\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1}{1-\gamma}$$

(Since $\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a) = 0$)

$$\geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1 \max_{s,a}|\mathfrak{a}^{\tilde{\pi}}(s,a)|\,\max_s\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1}{(1-\gamma)}$$

(By Holder's inequality, $|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]| \leq |\max_x f(x)|\,\|P-Q\|_1$)

$$\geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{2\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1 \max_s\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1}{(1-\gamma)^2}\ (*) \qquad (\mathfrak{a}^{\pi}(s,a) \leq \tfrac{2}{1-\gamma})$$

Next, we will express $\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1$ in terms of $\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1$, and combine it with (*).

Trust Region Policy Optimization

$$\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s') = \sum_s \Pr^{\pi}(S_{\tau-1} = s)\, \mathbf{P}_{\pi}(s, s') - \sum_s \Pr^{\tilde{\pi}}(S_{\tau-1} = s)\, \mathbf{P}_{\tilde{\pi}}(s, s')$$

$$= \sum_{s,a} \Big[\mathcal{P}(s'|s, a) \Big[\Pr^{\pi}(S_{\tau-1} = s)\, \pi(a|s) - \Pr^{\tilde{\pi}}(S_{\tau-1} = s)\, \tilde{\pi}(a|s) \Big] \Big]$$

$$= \sum_{s,a} \Big[\mathcal{P}(s'|s, a) \Big[\Pr^{\pi}(S_{\tau-1} = s)\, (\pi(a|s) - \tilde{\pi}(a|s)) + \tilde{\pi}(a|s) \Big(\Pr^{\pi}(S_{\tau-1} = s) - \Pr^{\tilde{\pi}}(S_{\tau-1} = s) \Big) \Big] \Big]$$

Taking absolute values, using the triangle inequality and summing over s',

$$\implies \sum_{s'} |\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s')|$$

$$\leq \underbrace{\sum_{s'} \sum_{s,a} \mathcal{P}(s'|s, a) |\Pr^{\pi}(S_{\tau-1} = s)\, (\pi(a|s) - \tilde{\pi}(a|s))\,|}_{(i)}$$

$$+ \underbrace{\sum_{s'} \sum_{s,a} \mathcal{P}(s'|s, a) \tilde{\pi}(a|s)\, |\Pr^{\pi}(S_{\tau-1} = s) - \Pr^{\tilde{\pi}}(S_{\tau-1} = s)|}_{(ii)}$$

Trust Region Policy Optimization

$$
\begin{aligned}
(i) &= \sum_{s,a} |\mathrm{Pr}^{\pi}(S_{\tau-1}=s)\ (\pi(a|s)-\tilde{\pi}(a|s))\,| \sum_{s'} \mathcal{P}(s'|s,a) = \sum_{s,a} |\mathrm{Pr}^{\pi}(S_{\tau-1}=s)\ (\pi(a|s)-\tilde{\pi}(a|s))\,| \\
&= \sum_{s} \mathrm{Pr}^{\pi}(S_{\tau-1}=s) \sum_{a} |\,(\pi(a|s)-\tilde{\pi}(a|s))\,| = \sum_{s} \mathrm{Pr}^{\pi}(S_{\tau-1}=s)\ \|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1 \\
&\leq \max_{s} \|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1 \\
(ii) &= \sum_{s} |\mathrm{Pr}^{\pi}(S_{\tau-1}=s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1}=s)| \sum_{a} \tilde{\pi}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \\
&= \sum_{s} |\mathrm{Pr}^{\pi}(S_{\tau-1}=s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1}=s)|
\end{aligned}
$$

Hence, $\sum_{s'} |\mathrm{Pr}^{\pi}(S_{\tau}=s') - \mathrm{Pr}^{\pi'}(S_{\tau}=s')| \leq \max_s \|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1$ $+ \sum_s |\mathrm{Pr}^{\pi}(S_{\tau-1}=s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1}=s)|$. By recursing over τ, we get that,

$$
\sum_{s'} |\mathrm{Pr}^{\pi}(S_{\tau}=s') - \mathrm{Pr}^{\pi'}(S_{\tau}=s')| \leq \tau \max_{s} \{\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1\}
$$

Trust Region Policy Optimization

Recall that $\sum_{s'} |\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s')| \leq \tau \max_s \left\{ \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \right\}$.

$$[d^{\pi} - d^{\tilde{\pi}}](s')$$

$$= (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^\tau \Pr^{\pi}[S_\tau = s' | S_0 = s_0] - (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^t \Pr^{\tilde{\pi}}[S_\tau = s' | S_0 = s_0]$$

$$\left\| d^{\pi} - d^{\tilde{\pi}} \right\|_1 = \sum_{s'} \left| (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^\tau \left[\Pr^{\pi}[S_\tau = s' | S_0 = s_0] - \Pr^{\tilde{\pi}}[S_\tau = s' | S_0 = s_0] \right] \right|$$

$$\leq (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^\tau \sum_{s'} |\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s')| \qquad \text{(Triangle inequality)}$$

$$\leq (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau \tau \max_s \left\{ \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \right\} = (1-\gamma) \max_s \left\{ \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \right\} \sum_{\tau=0}^{\infty} \gamma^\tau \tau$$

$$\leq \frac{\gamma \max_s \left\{ \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \right\}}{1-\gamma} \qquad \text{(Since } \textstyle\sum_{\tau=0}^{\infty} \gamma^\tau \tau \leq \frac{\gamma}{(1-\gamma)^2}\text{)}$$

Trust Region Policy Optimization

Recalling inequality (*), $v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{2\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1 \max_s\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1}{(1-\gamma)^2}$.

We also know that $\left\|d^{\pi} - d^{\tilde{\pi}}\right\|_1 \leq \frac{\gamma \max_s\left\{\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1\right\}}{1-\gamma}$. Hence,

$$v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{2\gamma\ [\max_s \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1]^2]}{(1-\gamma)^3}$$

$$\implies v^{\pi}(\rho) \geq \left[v^{\tilde{\pi}}(\rho) + \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{4\gamma\ \max_s \mathrm{KL}(\tilde{\pi}(\cdot|s)||\pi(\cdot|s))}{(1-\gamma)^3}\right]$$

(By Pinsker's inequality, $2\,\mathrm{KL}(\tilde{\pi}(\cdot|s)||\pi(\cdot|s)) \geq \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1^2$)

$$\implies v^{\pi}(\rho) \geq h(\pi) \quad \square$$

• For the tabular policy parameterization, a variant of TRPO that uses $\mathrm{KL}(\pi(\cdot|s)||\pi_t(\cdot|s))$ (instead of $\mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s))$) can be shown to converge to the optimal policy at an $O(1/\sqrt{T})$ rate [SEM20, Theorem 16]. However, the rate still involves the distribution mismatch ratio.

Proximal Policy Optimization

- Proximal Policy Optimization (PPO) is an alternative to TRPO. It is computationally more efficient, typically results in better performance, and is hence widely used in practice.
- PPO maximizes the following surrogate function (ϵ is a parameter) at iteration t:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} \left\{ \mathbb{E}_{s \sim d^{\pi_t}} \mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[\mathfrak{a}^{\pi_t}(s,a) \min \left\{ \frac{\pi(a|s)}{\pi_t(a|s)}, \text{clip}\left(\frac{\pi(a|s)}{\pi_t(a|s)}, 1-\epsilon, 1+\epsilon \right) \right\} \right] \right\}$$

  where $\text{clip}(x,a,b) = \min\{\max\{x,a\},b\}$ projects x onto the $[a,b]$ interval.
- Compared to the TRPO surrogate: $\mathbb{E}_{s \sim d^{\pi_t}} \mathbb{E}_{a \sim \pi_t(\cdot|s)} \frac{\pi(a|s)}{\pi_t(a|s)} \mathfrak{a}^{\pi_t}(s,a)$, s.t $\mathbb{E}_{s \sim d^{\pi_t}} \text{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \leq \delta$ which ensures that the importance sampling ratio $\frac{\pi(a|s)}{\pi_t(a|s)}$ does not become too large by controlling $\text{KL}(\pi_t(\cdot|s)||\pi(\cdot|s))$, PPO directly ensures that the importance sampling ratio is bounded by clipping it.
- There is no theoretical justification for the clipped PPO surrogate (even with tabular policy parameterization). In fact, PPO can fail on simple problems [HMDH20].
- Recent literature [EIS+20] suggests that code-level implementation details are responsible for most of PPO's gain over TRPO.

Exploration in Tabular MDPs

Exploration in Tabular MDPs

- In Lectures 7-8, we have seen that Approximate Policy Iteration and Politex resolved the exploration problem by using G experimental design.
- In Lectures 10-11, we saw that Policy Gradient methods that do not handle exploration explicitly can incur a dependence on the concentrability coefficient, especially when using function approximation.
- In Lectures 1-3, we saw systematic ways of handling exploration for bandit problems.
- For the last topic, we will design similar explicit exploration schemes for tabular MDPs.

For today's lecture, we will focus on a simplified case:

- Rewards are deterministic and known, but the transition probabilities are not. It is straightforward to extend to unknown rewards.
- Finite-horizon episodic setting where the agent interacts with the environment in episodes, and each episode has a finite length H. $H = 1$ recovers the bandit setting. Similar techniques work for the infinite-horizon discounted setting, but the analysis more complicated [HZG21].

Finite Horizon MDPs

- The MDP is given as: $M = (\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1})$, where H is the finite horizon.
- The transition probabilities and rewards both depend on the step h. In particular, $\mathcal{P}_h(s'|s,a)$ is the probability of transitioning into state s' when taking action a in state s at step h. Similarly, $r_h(s,a) \in [0,1]$ is the reward obtained when taking action a in state s at step h.
- **Value functions**: For a fixed policy π and state $s \in \mathcal{S}$,

$$v_h^\pi(s) := \mathbb{E}\left[\sum_{i=h}^{H-1} r_h(s_i, a_i) | s_h = s\right] \; ; \; q_h^\pi(s,a) = \mathbb{E}\left[\sum_{i=h}^{H-1} r_h(s_i, a_i) | s_h = s, a_h = a\right],$$

  where the expectation is over the randomness in the reward process induced by policy.
- **Bellman policy evaluation**: For a fixed policy π, the Bellman equation can be written as:

$$q_h^\pi(s,a) := r_h(s,a) + \sum_{s'} \mathcal{P}_h(s'|s,a)\, v_{h+1}^\pi(s') = r_h(s,a) + \langle \mathcal{P}_h(\cdot|s,a), v_{h+1}^\pi \rangle .$$

  We define $q_H^\pi(s,a) = v_H^\pi(s) = 0$ for all policies π and s, a.

Finite Horizon MDPs

- **Objective**: Given a starting state s_0, find a policy π^* that maximizes the value, i.e.

$$\pi^* = \arg\max v^\pi(s_0) := v_0^\pi(s_0)$$

- **Bellman optimality**: q^* is the optimal action-value function iff for all s, a, h,

$$q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot|s,a)} \left[\max_{a'} q_{h+1}^*(s', a') \right]$$

- The optimal policy and value function is given as:

$$\pi_h^*(s) = \arg\max_{a \in \mathcal{A}} q_h^*(s,a) \quad ; \quad v_h^*(s) = \max_{a \in \mathcal{A}} q_h^*(s,a) \quad ; \quad v^*(s_0) := v_0^{\pi^*}(s_0)$$

- Similar to the infinite-horizon discounted case, the optimal policy is deterministic and Markov (given s, h, it does not depend on the history).
- Unlike the infinite-horizon discounted case, the optimal action in state s depends on the step h. Hence, the optimal policy is non-stationary [Chapter 4, PC'23].
- For finite-horizon MDPs, the optimal policy can be found by dynamic programming (does not require solving fixed point equations like in the discounted setting).

Exploration in Tabular MDPs

- The agent episodically interacts with a finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1})$ where the transitions $\{\mathcal{P}_h\}_{h=0}^{H-1}$ are unknown.
- In particular, in each episode $t \in [T]$, the agent chooses a policy $\pi^t := \{\pi_h^t\}_{h=0}^{H-1}$, generates a trajectory $\tau^t = \{s_h^t, a_h^t\}_{h=0}^{H-1}$ and receives the cumulative reward $\sum_{h=0}^{H-1} r(s_h^t, a_h^t)$.
- **Objective**: Minimize the cumulative regret defined as:
  $\text{Regret}(T) := \mathbb{E}\left[\sum_{t=0}^{T-1}\left[v^*(s_0) - \sum_{h=0}^{H-1} r(s_h^t, a_h^t)\right]\right] = \mathbb{E}\left[\sum_{t=0}^{T-1}\left[v^*(s_0) - v_0^{\pi_t}(s_0)\right]\right]$,
  where the expectation is with respect to the randomness in the MDP and the algorithm.
- We will use a *model-based approach* which uses the collected data to build a model of the environment (the MDP in this case). Specifically, in each episode t, we build the approximate MDP: $\hat{M}^t = (\mathcal{S}, \mathcal{A}, \{\hat{\mathcal{P}}_h^t\}_{h=0}^{H-1}, \{\hat{r}_h^t\}_{h=0}^{H-1})$ where

$$N_h^t(s,a,s') := \sum_{i=0}^{t-1} \mathcal{I}\left\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right\} \; ; \; N_h^t(s,a) := \sum_{i=0}^{t-1} \mathcal{I}\left\{(s_h^i, a_h^i) = (s,a)\right\}$$

$$= \hat{\mathcal{P}}_h^t(s'|s,a) := \frac{N_h^t(s,a,s')}{N_h^t(s,a)} \; ; \; \hat{r}_h^t(s,a) = r_h(s,a) + b_h^t(s,a) \; ; \; b_h^t(s,a) = 2H\sqrt{\frac{\ln(SAHT/\delta)}{N_h^t(s,a)}}$$

Exploration in Tabular MDPs

Algorithm UCB-VI

1: **Input**: MDP M, π^0

2: **for** $t = 1 \to T - 1$ **do**

3: Rollout policy π^{t-1} in MDP M and generate the trajectory $\tau^t = \{s_h^t, a_h^t\}_{h=0}^{H-1}$.

4: Use the collected data to build $\hat{M}^t = (\mathcal{S}, \mathcal{A}, \{\hat{\mathcal{P}}_h^t\}_{h=0}^{H-1}, \{\hat{r}_h^t\}_{h=0}^{H-1})$

5: Run value iteration (dynamic programming) in $\hat{M}^t$ i.e. setting $\hat{q}_H^t(s,a) = 0$ and $\hat{v}_H^t(s) = 0$ for all s, a and recursing from $h = (H-1) \to 0$, calculate

$$\hat{q}_h^t(s,a) = \min\{\hat{r}_h^t(s,a) + \sum_{s'} \hat{\mathcal{P}}_h^t(s'|s,a) \max_{a' \in \mathcal{A}} \hat{q}_{h+1}^t(s',a')\rangle, H\}$$

$$\pi_h^t(s) = \arg\max_{a \in \mathcal{A}} \hat{q}_h^t(s,a) \; ; \; \hat{v}_h^t(s) := \hat{v}_h^{\pi_t}(s) = \max_{a \in \mathcal{A}} \hat{q}_h^t(s,a)$$

6: **end for**

- The bonus b_h^t in the rewards $\hat{r}_h^t$ is similar to the UCB term for bandits and results in optimism in the q functions, inducing sufficient exploration.
- For the proof, we will use the fact that $\hat{q}_h^t(s,a) = \min\{\hat{r}_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), \hat{v}_{h+1}^t\rangle, H\}$.

Exploration in Tabular MDPs

Claim: UCB-VI incurs $\text{Regret}(T) = O\left(H^2 S\sqrt{AT}\sqrt{\ln(SAH^2T^2)}\right)$.

- Similar to the bandit setting in Lecture 2, the regret scales as $O(\sqrt{T})$.
- If $S = 1$ and $H = 1$, UCB-VI is similar to UCB and incurs the same $O(\sqrt{AT})$ regret.

Proof: We will do the proof in 4 steps:

(i) **Concentration**: Define a good event $\mathcal{E}$ on which the estimated transitions $\hat{\mathcal{P}}_h^t(\cdot|s,a)$ are "close" to the true transitions $\mathcal{P}_h(\cdot|s,a)$ with high probability for all s, a, t, h.

(ii) **Optimism**: Conditioned on $\mathcal{E}$, prove that for all episodes t, $\hat{v}_0^t(s_0) \geq v_0^*(s_0)$ for starting state s_0, i.e. the value function for policy π_t in $\hat{M}$ is larger than that of π^* in M.

(iii) **Regret decomposition**: Conditioned on $\mathcal{E}$, decompose the regret and prove that $\text{Regret}(T) = O\left(\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} 1/\sqrt{N_h^t(s_h^t,a_h^t)}\right)$.

(iv) **Wrapping up**: Wrap up the proof by bounding $\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} 1/\sqrt{N_h^t(s_h^t,a_h^t)}$ deterministically and use the law of total expectation.

Exploration in Tabular MDPs – Concentration

Define the good event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, where

$$\mathcal{E}_1 := |\langle \mathcal{P}_h(\cdot|s,a) - \hat{\mathcal{P}}_h^t(\cdot|s,a), f \rangle| \leq 8H\sqrt{S \ln(SAHT/\delta)/N_h^t(s,a)} \quad (\forall s,a,h,t \text{ and } f : \mathcal{S} \to [0,H])$$

$$\mathcal{E}_2 := |\langle \mathcal{P}_h(\cdot|s,a) - \hat{\mathcal{P}}_h^t(\cdot|s,a), v_{h+1}^* \rangle| \leq 2H\sqrt{\ln(SAHT/\delta)/N_h^t(s,a)} \quad (\forall s,a,h,t)$$

We use two facts to bound $\Pr[\mathcal{E}]$ (see [AJKS19, Lemmas 7.2, 7.3] for the corresponding proofs)

- **Fact 1**: For $\delta \in (0,1)$, for all t,h,s,a, $\Pr[\mathcal{E}_1] \geq 1-\delta$.
- **Fact 2**: For $\delta \in (0,1)$, for all t,h,s,a, $\Pr[\mathcal{E}_2] \geq 1-\delta$.
- For both statements, since $N_h^t(s,a)$ is itself a random quantity, we cannot directly use the Hoeffding bound. We need to define an appropriate martingale difference sequence followed by the use of the Azuma-Hoeffding inequality.
- Fact 1 is concerned with bounding the inner-product for any f, including those that are random and depend on the collected samples. The proof involves a covering argument followed by a union bound. This results in an additional S dependence.

By a union bound, $\Pr[\mathcal{E}] \geq 1 - 2\delta$.

Exploration in Tabular MDPs – Optimism

Claim: Conditioned on $\mathcal{E}$, for all episodes t, $\hat{v}_0^t(s_0) \geq v_0^*(s_0)$ for starting state s_0.

Proof: We will prove that $\forall s$, $\hat{v}_h^t(s) \geq v_h^*(s)$ by backward induction on h from $h = H$ to $h = 0$.

Base case: $\forall s, t$, $\hat{v}_H^t(s) = v_H^*(s) = 0$, and hence, $\hat{v}_H^t(s) \geq v_H^*(s)$.

Inductive case: Assuming that $\hat{v}_{h+1}^t(s) \geq v_{h+1}^*(s)$, we want to prove that $\hat{v}_h^t(s) \geq v_h^*(s)$.
Case (a): Recall that $\hat{q}_h^t(s,a) = \min\{\hat{r}_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), \hat{v}_{h+1}^t \rangle, H\}$. If for any t, s, a, $\hat{q}_h^t(s,a) = H$, then, $\hat{q}_h^t(s,a) \geq q_h^*(s,a)$, $\hat{v}_h^t(s) = \max_{a \in \mathcal{A}} \hat{q}_h^t(s,a) \geq \max_{a \in \mathcal{A}} q_h^*(s,a) = v_h^*(s)$.
Case (b): If $\hat{q}_h^t(s,a) = \hat{r}_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), \hat{v}_{h+1}^t \rangle = r_h^t(s,a) + b_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), \hat{v}_{h+1}^t \rangle$,

$$\begin{aligned}
\hat{q}_h^t(s,a) - q_h^*(s,a) &= b_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), \hat{v}_{h+1}^t \rangle - \langle \mathcal{P}_h(\cdot|s,a), v_{h+1}^* \rangle \\
&\geq b_h^t(s,a) + \langle \hat{\mathcal{P}}_h^t(\cdot|s,a), v_{h+1}^* \rangle - \langle \mathcal{P}_h(\cdot|s,a), v_{h+1}^* \rangle \quad \text{(Inductive hypothesis)} \\
&\geq b_h^t(s,a) - 2H\sqrt{\ln(SAHT/\delta)/N_h^t(s,a)} \quad \text{(Since we are conditioning on } \mathcal{E}\text{)} \\
\implies \hat{q}_h^t(s,a) &\geq q_h^*(s,a) \quad \text{(Since } b_h^t(s,a) = 2H\sqrt{\ln(SAHT/\delta)/N_h^t(s,a)}\text{)}
\end{aligned}$$

Hence, $\hat{v}_h^t(s) = \max_{a \in \mathcal{A}} \hat{q}_h^t(s,a) \geq \max_{a \in \mathcal{A}} q_h^*(s,a) = v_h^*(s) \implies \hat{v}_0^t(s_0) \geq v_0^*(s_0)$ □

Exploration in Tabular MDPs – Regret Decomposition

Claim: Conditioned on $\mathcal{E}$, $\text{Regret}(T) \leq 10\, H \sqrt{\frac{S \ln(SAHT/\delta)}{N_h^t(s,a)}}\, \mathbb{E} \sum_{t=0}^{T-1} \sum_{h=0}^{H=1} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}}$.

Proof: Consider episode t. By using the optimism result from the previous slide, $v_0^*(s_0) - v_0^{\pi_t}(s_0) \leq \hat{v}_0^t(s_0) - v_0^{\pi_t}(s_0) = \hat{v}_0^{\pi_t}(s_0) - v^{\pi_t}(s_0)$. Hence, we need to bound the difference in the value functions of the same policy but on different MDPs.

Claim: For the same deterministic policy π on $M = (\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1})$ and $\tilde{M} = (\mathcal{S}, \mathcal{A}, \{\tilde{\mathcal{P}}_h\}_{h=0}^{H-1}, \{\tilde{r}_h\}_{h=0}^{H-1})$, for starting state s_0,

$$v_0^{\pi,M}(s_0) - v_0^{\pi,\tilde{M}}(s_0) = \mathbb{E}_{\substack{s_{h+1} \sim \tilde{\mathcal{P}}_h(\cdot|s_h a_h) \\ a_h = \pi_h(s_h)}} \sum_{h=0}^{H-1} \left[[r_h(s_h, a_h) - \tilde{r}_h(s_h, a_h)] + \langle \mathcal{P}_h(\cdot|s_h, a_h) - \tilde{\mathcal{P}}_h(\cdot|s_h, a_h), v_{h+1} \rangle \right]$$

Prove in Assignment 4! Using the above lemma with $M = \hat{M}$, $\tilde{M} = M$, $\pi = \pi_t$ and since $v_0^*(s_0) \leq \hat{v}^{\pi_t}(s_0)$,

$$v_0^*(s_0) - v_0^{\pi_t}(s_0) \leq \mathbb{E} \sum_{h=0}^{H-1} \left[b_h^t(s_h^t, a_h^t) + |\langle \mathcal{P}_h(\cdot|(s_h^t, a_h^t) - \hat{\mathcal{P}}_h^t(\cdot|(s_h^t, a_h^t), \hat{v}_{h+1}^t \rangle| \right],$$

where the expectation is w.r.t the trajectory generated by policy π_t.

Exploration in Tabular MDPs – Regret Decomposition

Recall that $v_0^*(s_0) - v_0^{\pi_t}(s_0) \leq \mathbb{E} \sum_{h=0}^{H-1} \left[b_h^t(s_h^t, a_h^t) + |\langle \mathcal{P}_h(\cdot|(s_h^t, a_h^t) - \hat{\mathcal{P}}_h^t(\cdot|(s_h^t, a_h^t), \hat{v}_{h+1}^t \rangle| \right]$. Since we are conditioning on $\mathcal{E}$, $|\langle \mathcal{P}_h(\cdot|(s_h^t, a_h^t) - \hat{\mathcal{P}}_h^t(\cdot|(s_h^t, a_h^t), \hat{v}_{h+1}^t \rangle| \leq 8H \sqrt{\frac{S \ln(SAHT/\delta)}{N_h^t(s,a)}}$. Hence,

$$
\begin{aligned}
v_0^*(s_0) - v_0^{\pi_t}(s_0) &\leq \mathbb{E} \sum_{h=0}^{H-1} \left[b_h^t(s_h^t, a_h^t) + 8H \sqrt{\frac{S \ln(SAHT/\delta)}{N_h^t(s_h^t, a_h^t)}} \right] \\
&\leq 10\, H \sqrt{S \ln(SAHT/\delta)} \left[\mathbb{E} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} \right] \\
&\qquad \text{(Since } b_h^t(s,a) = 2H \sqrt{\tfrac{\ln(SAHT/\delta)}{N_h^t(s,a)}}\text{)}
\end{aligned}
$$

Summing from $t = 0$ to $T - 1$,

$$
\implies \text{Regret}(T) = \sum_{t=0}^{T-1} [v_0^*(s_0) - v_0^{\pi_t}(s_0)] \leq 10\, H \sqrt{S \ln(SAHT/\delta)} \left[\mathbb{E} \sum_{t=0}^{T-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} \right]
$$

Exploration in Tabular MDPs – Wrapping up

$$\text{Regret}(T) \leq 10\, H \sqrt{\frac{S \ln(SAHT/\delta)}{N_h^t(s_h^t, a_h^t)}} \left[\mathbb{E} \sum_{t=0}^{T-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} \right]$$

Fact 3: $\sum_{t=0}^{T-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} \leq 2H\sqrt{SAT}$. (see [AJKS19, Lemma 7.5] for a proof)

Putting everything together and using the law of total expectation,

$$\begin{aligned}
&\mathbb{E}[\text{Regret}(T)] = \mathbb{E}[\text{Regret}(T)|\mathcal{E}]\ \Pr[\mathcal{E}] + \mathbb{E}[\text{Regret}(T)|\mathcal{E}^c]\ \Pr[\mathcal{E}^c] \\
&\leq \mathbb{E}[\text{Regret}(T)|\mathcal{E}] + T\,H\ \Pr[\mathcal{E}^c] \leq 20\, H^2\, S\sqrt{AT}\, \sqrt{\ln(SAHT/\delta)} + 2\delta\, T\, H \\
&\leq 20\, H^2\, S\sqrt{AT}\, \sqrt{\ln(SA\, H^2\, T^2)} + 2 = O\left(H^2\, S\sqrt{AT}\, \sqrt{\ln(SA\, H^2\, T^2)} \right) \qquad (\text{Setting } \delta = \tfrac{1}{TH})
\end{aligned}$$

- By designing better bonuses, the regret for UCB-VI can be improved to $\Theta(H^{3/2}\sqrt{SAT})$.
- UCB-VI uses a standard planning algorithm (VI) but on a model of the MDP. Similarly, we can use policy optimization such as NPG with a model of the MDP [SERM20, CYJW20] and handle exploration in a systematic manner.
- Given a set of features Φ and under appropriate linearity assumptions about the transitions and rewards, LSVI-UCB [JYWJ20] can attain an $O(d^{3/2}\, H^2\, \sqrt{T})$ regret bound.

Wrapping up

What we covered

- Handling the exploration-exploitation trade-off in (linear) bandit problems.
- Markov Decision Processes and the Fundamental Theorem
- Given a known MDP, value Iteration, policy Iteration, linear programming for planning
- When the MDP is not known, Monte-Carlo estimation and Temporal difference learning to estimate a policy's value.
- When the MDP is not known, approximate policy iteration and Politex for learning good policies under linear function approximation.
- (Natural) policy gradient methods and global convergence to optimal policies, TRPO/PPO
- Systematically handling exploration in MDPs

What we could not cover

- Sample complexity for computing the optimal policy with access to a generative model (simulator) (see [AKY20] for a nice related work section and near-optimal bounds)
- Actor-Critic Methods and their analysis (see [XWL20] for theoretical bounds)
- Q-learning and its analysis (see [JAZBJ18] for theoretical bounds)

Other important topics in sequential decision-making

- RL with constraints [GYD+22] and multiple objectives [HRB+22]
- Batch/Offline RL [LKTF20]
- Continual learning [WZSZ23]
- Imitation learning [ZVZ+21]

References i

- Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun, *Reinforcement learning: Theory and algorithms*, CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep **32** (2019).
- Alekh Agarwal, Sham Kakade, and Lin F Yang, *Model-based reinforcement learning with a generative model is minimax optimal*, Conference on Learning Theory, PMLR, 2020, pp. 67–83.
- Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang, *Provably efficient exploration in policy optimization*, International Conference on Machine Learning, PMLR, 2020, pp. 1283–1294.
- Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry, *Implementation matters in deep policy gradients: A case study on ppo and trpo*, arXiv preprint arXiv:2005.12729 (2020).

References ii

- Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll, *A review of safe reinforcement learning: Methods, theory and applications*, arXiv preprint arXiv:2205.10330 (2022).
- Chloe Ching-Yun Hsu, Celestine Mendler-Dünner, and Moritz Hardt, *Revisiting design choices in proximal policy optimization*, arXiv preprint arXiv:2009.10897 (2020).
- Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al., *A practical guide to multi-objective reinforcement learning and planning*, Autonomous Agents and Multi-Agent Systems **36** (2022), no. 1, 26.
- Jiafan He, Dongruo Zhou, and Quanquan Gu, *Nearly minimax optimal reinforcement learning for discounted mdps*, Advances in Neural Information Processing Systems **34** (2021), 22288–22300.

References iii

- Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan, *Is q-learning provably efficient?*, Advances in neural information processing systems **31** (2018).
- Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan, *Provably efficient reinforcement learning with linear function approximation*, Conference on Learning Theory, PMLR, 2020, pp. 2137–2143.
- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu, *Offline reinforcement learning: Tutorial, review, and perspectives on open problems*, arXiv preprint arXiv:2005.01643 (2020).
- Lior Shani, Yonathan Efroni, and Shie Mannor, *Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 5668–5675.

References iv

- Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor, *Optimistic policy optimization with bandit feedback*, International Conference on Machine Learning, PMLR, 2020, pp. 8604–8613.
- Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu, *A comprehensive survey of continual learning: Theory, method and application*, arXiv preprint arXiv:2302.00487 (2023).
- Tengyu Xu, Zhe Wang, and Yingbin Liang, *Improving sample complexity bounds for (natural) actor-critic algorithms*, Advances in Neural Information Processing Systems **33** (2020), 4358–4369.
- Boyuan Zheng, Sunny Verma, Jianlong Zhou, Ivor Tsang, and Fang Chen, *Imitation learning: Progress, taxonomies and challenges*, arXiv preprint arXiv:2106.12177 (2021).