Foundations of RL and Interactive Decision Making

Yuanhao ZHU Xiaoxian DING

Slides adapted from MIT course notes (chapters 5 & 6) by Dylan J. Foster and Alexander Rakhlin

https://doi.org/10.48550/arXiv.2312.16730

Outline

Ch 5. Reinforcement Learning: Basics

Ch 6. General Decision Making

Finite-Horizon Episodic MDP Formulation

A *Markov Decision Process*(MDP) M takes the form

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

where

- $\mathcal{S}$ is the state space
- $\mathcal{A}$ is the action space
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the prob transition kernel at step h
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward distribution at step h
- $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution

Markov property refers

$$\mathbb{P}^M(s_{h+1} = s' | s_h, a_h) = \mathbb{P}^M(s_{h+1} = s' | s_h, a_h, s_{h-1}, a_{h-1}, \ldots, s_1, a_1).$$

Finite-Horizon Episodic MDP Formulation

A *Markov Decision Process*(MDP) M takes the form

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

where

- $\mathcal{S}$ is the state space
- $\mathcal{A}$ is the action space
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the prob transition kernel at step h
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward distribution at step h
- $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution

Markov property refers

$$\mathbb{P}^M(s_{h+1} = s' | s_h, a_h) = \mathbb{P}^M(s_{h+1} = s' | s_h, a_h, s_{h-1}, a_{h-1}, \ldots, s_1, a_1).$$

MDP Episode Protocol

At the beginning of the episode, the learner selects $\pi = (\pi_1, \ldots, \pi_H) \in \Pi_{\text{rns}}$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$.

1. Begin from $s_1 \sim d_1$
2. For $h = 1, \ldots, H$:
   - $a_h \sim \pi_h(s_h)$
   - $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(s_h, a_h)$
3. Deterministic terminal state s_{H+1} for simplicity

MDP Episode Protocol

At the beginning of the episode, the learner selects $\pi = (\pi_1, \dots, \pi_H) \in \Pi_{\text{rns}}$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$.

1. Begin from $s_1 \sim d_1$
2. For $h = 1, \dots, H$:
   - $a_h \sim \pi_h(s_h)$
   - $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(s_h, a_h)$
3. Deterministic terminal state s_{H+1} for simplicity

MDP Episode Protocol

At the beginning of the episode, the learner selects $\pi = (\pi_1, \ldots, \pi_H) \in \Pi_{\mathrm{rns}}$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$.

1. Begin from $s_1 \sim d_1$
2. For $h = 1, \ldots, H$:
   - $a_h \sim \pi_h(s_h)$
   - $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(s_h, a_h)$
3. Deterministic terminal state s_{H+1} for simplicity

- State-action value function:

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a]$$

- State value function: $V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s]$
- Optimal value functions: $Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\mathrm{rns}}} Q_h^{M,\pi}(s,a)$, $V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a)$
- Value for a policy π under M:

$$f^M(\pi) = \mathbb{E}^{M,\pi}[\sum_{h=1}^{H} r_h] = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s \sim d_1}[V_1^{M,\pi}(s)]$$

- Optimal policy: $\pi_M \in \arg\max_{\pi \in \Pi_{\mathrm{rns}}} f^M(\pi)$

- State-action value function:

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a]$$

- State value function: $V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s]$
- Optimal value functions: $Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\mathrm{rns}}} Q_h^{M,\pi}(s,a)$, $V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a)$
- Value for a policy π under M:

$$f^M(\pi) = \mathbb{E}^{M,\pi}[\sum_{h=1}^{H} r_h] = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s \sim d_1}[V_1^{M,\pi}(s)]$$

- Optimal policy: $\pi_M \in \arg\max_{\pi \in \Pi_{\mathrm{rns}}} f^M(\pi)$

- State-action value function:

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a]$$

- State value function: $V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s]$
- Optimal value functions: $Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\mathrm{rns}}} Q_h^{M,\pi}(s,a)$, $V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a)$
- Value for a policy π under M:

$$f^M(\pi) = \mathbb{E}^{M,\pi}[\sum_{h=1}^{H} r_h] = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s \sim d_1}[V_1^{M,\pi}(s)]$$

- Optimal policy: $\pi_M \in \arg\max_{\pi \in \Pi_{\mathrm{rns}}} f^M(\pi)$

- State-action value function:

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a]$$

- State value function: $V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s]$
- Optimal value functions: $Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\text{rns}}} Q_h^{M,\pi}(s,a)$, $V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a)$
- Value for a policy π under M:

$$f^M(\pi) = \mathbb{E}^{M,\pi}[\sum_{h=1}^{H} r_h] = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s \sim d_1}[V_1^{M,\pi}(s)]$$

- Optimal policy: $\pi_M \in \arg\max_{\pi \in \Pi_{\text{rns}}} f^M(\pi)$

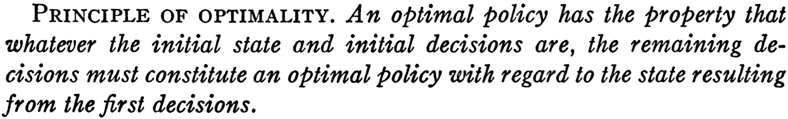
- State-action value function:

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s, a_h = a]$$

- State value function: $V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^{H} r_{h'} | s_h = s]$
- Optimal value functions: $Q_h^{M,\star}(s,a) = \max_{\pi \in \Pi_{\text{rns}}} Q_h^{M,\pi}(s,a)$, $V_h^{M,\star}(s) = \max_a Q_h^{M,\star}(s,a)$
- Value for a policy π under M:

$$f^M(\pi) = \mathbb{E}^{M,\pi}[\sum_{h=1}^{H} r_h] = \mathbb{E}_{s \sim d_1, a \sim \pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s \sim d_1}[V_1^{M,\pi}(s)]$$

- Optimal policy: $\pi_M \in \arg\max_{\pi \in \Pi_{\text{rns}}} f^M(\pi)$



Principle of optimality. *An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.*

- Bellman Optimality, $V_{H+1}^{M,\pi_M}(s) := 0$ and for $h \in [H]$,

$$V_h^{M,\pi_M}(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s,\ a_h = a\right]$$

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}^M\left[r_h + \max_{a' \in \mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1}, a') \mid s_h = s, a_h = a\right]$$

- Value Iteration (VI) and Bellman Operators

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}_{s_{h+1} \sim P_h^M(s,a),\, r_h \sim R_h^M(s,a)}\left[r_h(s,a) + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a')\right]$$

or equivalently

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}^M\left[r_h + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a') \;\middle|\; s_h = s,\ a_h = a\right]$$

- In the language of Bellman operators,

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M Q_{h+1}^{M,\pi_M}$$

PRINCIPLE OF OPTIMALITY. *An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.*

- Bellman Optimality, $V_{H+1}^{M,\pi_M}(s) := 0$ and for $h \in [H]$,

$$V_h^{M,\pi_M}(s) = \max_{a\in\mathcal{A}} \mathbb{E}\big[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s,\ a_h = a\big]$$

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}^M\big[r_h + \max_{a'\in\mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1},a') \mid s_h = s, a_h = a\big]$$

- Value Iteration (VI) and Bellman Operators

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}_{s_{h+1}\sim P_h^M(s,a),\, r_h\sim R_h^M(s,a)}\big[r_h(s,a)+\max_{a'\in\mathcal{A}} Q(s_{h+1},a')\big]$$

or equivalently

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}^M\left[r_h + \max_{a'\in\mathcal{A}} Q(s_{h+1},a') \,\middle|\, s_h = s,\ a_h = a\right]$$

- In the language of Bellman operators,

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M\, Q_{h+1}^{M,\pi_M}$$

PRINCIPLE OF OPTIMALITY. *An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.*

- Bellman Optimality, $V_{H+1}^{M,\pi_M}(s) := 0$ and for $h \in [H]$,

$$V_h^{M,\pi_M}(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s,\ a_h = a\right]$$

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}^M\left[r_h + \max_{a' \in \mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1}, a') \mid s_h = s, a_h = a\right]$$

- Value Iteration (VI) and Bellman Operators

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}_{s_{h+1}\sim P_h^M(s,a),\, r_h \sim R_h^M(s,a)}\left[r_h(s,a) + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a')\right]$$

or equivalently

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}^M\left[r_h + \max_{a' \in \mathcal{A}} Q(s_{h+1}, a') \,\middle|\, s_h = s,\ a_h = a\right]$$

- In the language of Bellman operators,

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M Q_{h+1}^{M,\pi_M}$$

PRINCIPLE OF OPTIMALITY. *An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.*

- Bellman Optimality, $V_{H+1}^{M,\pi_M}(s) := 0$ and for $h \in [H]$,

$$V_h^{M,\pi_M}(s) = \max_{a\in\mathcal{A}} \mathbb{E}\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s,\ a_h = a\right]$$

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}^M\left[r_h + \max_{a^{'}\in\mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1},a^{'}) \mid s_h = s, a_h = a\right]$$

- Value Iteration (VI) and Bellman Operators

$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}_{s_{h+1}\sim P_h^M(s,a),\, r_h\sim R_h^M(s,a)}\left[r_h(s,a) + \max_{a'\in\mathcal{A}} Q(s_{h+1},a')\right]$$

or equivalently

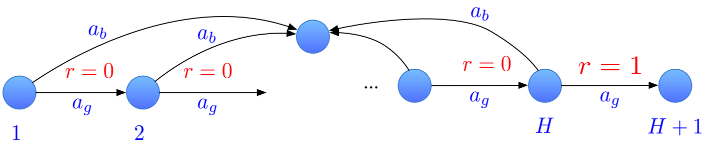
$$[\mathcal{T}_h^M Q](s,a) = \mathbb{E}^M\left[r_h + \max_{a'\in\mathcal{A}} Q(s_{h+1},a') \,\middle|\, s_h = s,\ a_h = a\right]$$

- In the language of Bellman operators,

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M\, Q_{h+1}^{M,\pi_M}$$

Failure of Uniform Exploration

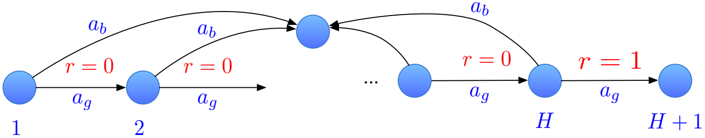
- Planning with a known MDP is straightforward, but minimizing regret in an unknown MDP requires exploration.
- ε-Greedy:
  - Reasonable for bandits and contextual bandits (suboptimal rate: $T^{2/3}$ vs. $\sqrt{T}$).
  - But disastrous in reinforcement learning, e.g. **Combination Lock MDP**.



    - Require selecting a_g for all the H time steps within the episode; otherwise, gain no info.
    - Uniform exploration ⇒ prob. of the correct sequence is 2^{-H} ⇒ need $T = O(2^H)$ to achieve nontrivial regret.

Failure of Uniform Exploration

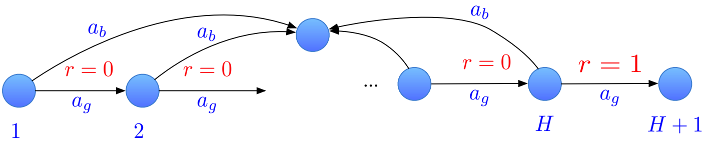
- Planning with a known MDP is straightforward, but minimizing regret in an unknown MDP requires exploration.
- ε-Greedy:
  - Reasonable for bandits and contextual bandits (suboptimal rate: $T^{2/3}$ vs. $\sqrt{T}$).
  - But disastrous in reinforcement learning, e.g. **Combination Lock MDP**.



    - Require selecting a_g for all the H time steps within the episode; otherwise, gain no info.
    - Uniform exploration ⇒ prob. of the correct sequence is 2^{-H} ⇒ need $T = O(2^H)$ to achieve nontrivial regret.

Failure of Uniform Exploration

- Planning with a known MDP is straightforward, but minimizing regret in an unknown MDP requires exploration.
- ε-Greedy:
  - Reasonable for bandits and contextual bandits (suboptimal rate: $T^{2/3}$ vs. $\sqrt{T}$).
  - But disastrous in reinforcement learning, e.g. **Combination Lock MDP**.



    - Require selecting a_g for all the H time steps within the episode; otherwise, gain no info.
    - Uniform exploration $\Rightarrow$ prob. of the correct sequence is 2^{-H} $\Rightarrow$ need $T = O(2^H)$ to achieve nontrivial regret.

Principle of Optimism Succeeds

- Other algorithmic principles?
- **Optimism in the face of uncertainty** succeeds, which implies that *one should act as if the environment is as nice as plausibly possible*.
- An analogue of UCB yields regret polynomial in $|S|$, $|A|$, and H.
- We will introduce standard MDP analysis tools to show this.

Principle of Optimism Succeeds

- Other algorithmic principles?
- **Optimism in the face of uncertainty** succeeds, which implies that *one should act as if the environment is as nice as plausibly possible*.
- An analogue of UCB yields regret polynomial in $|S|$, $|A|$, and H.
- We will introduce standard MDP analysis tools to show this.

Principle of Optimism Succeeds

- Other algorithmic principles?
- **Optimism in the face of uncertainty** succeeds, which implies that *one should act as if the environment is as nice as plausibly possible*.
- An analogue of UCB yields regret polynomial in $|S|$, $|A|$, and H.
- We will introduce standard MDP analysis tools to show this.

Principle of Optimism Succeeds

- Other algorithmic principles?
- **Optimism in the face of uncertainty** succeeds, which implies that *one should act as if the environment is as nice as plausibly possible*.
- An analogue of UCB yields regret polynomial in $|S|$, $|A|$, and H.
- We will introduce standard MDP analysis tools to show this.

Principle of Optimism Succeeds

- Other algorithmic principles?
- **Optimism in the face of uncertainty** succeeds, which implies that *one should act as if the environment is as nice as plausibly possible*.
- An analogue of UCB yields regret polynomial in $|S|$, $|A|$, and H.
- We will introduce standard MDP analysis tools to show this.

Some Standard MDP Analysis Tools

Lemma 1 (Performance Difference)

For any $s \in \mathcal{S}$ and $\pi, \pi' \in \Pi_{rns}$,

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi}\left[Q_h^{M,\pi'}(s_h,\, \pi'(s_h)) - Q_h^{M,\pi'}(s_h,\, a_h)\,\middle|\, s_1 = s\right].$$

Key idea: The difference in values between π' and π in the same MDP can be expressed via the expected advantage of π''s action over π's under state distribution induced by π at each timestep.

Some Standard MDP Analysis Tools

Lemma 1 (Performance Difference)

For any $s \in \mathcal{S}$ and $\pi, \pi' \in \Pi_{rns}$,

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi}\left[Q_h^{M,\pi'}\left(s_h,\, \pi'(s_h)\right) - Q_h^{M,\pi'}\left(s_h,\, a_h\right) \middle| s_1 = s \right].$$

Key idea: The difference in values between π' and π in the same MDP can be expressed via the expected advantage of π''s action over π's under state distribution induced by π at each timestep.

Lemma 2 (Bellman Residual Decomposition)

For any pair of MDPs $M = (P^M, R^M)$ and $\widehat{M} = (P^{\widehat{M}}, R^{\widehat{M}})$, any $s \in \mathcal{S}$, and policies $\pi \in \Pi_{rns}$,

$$V_1^{M,\pi}(s) - V_1^{\widehat{M},\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1}) \middle| s_1 = s\right]$$

In addition, for any M and $Q = (Q_1, \ldots, Q_H, 0)$ (need not to be a value function), letting $\pi_{Q,h}(s) = \arg\max_{a \in \mathcal{A}} Q_h(s, a)$, we have

$$\begin{aligned}&\max_{a \in \mathcal{A}} Q_1(s, a) - V_1^{M,\pi_Q}(s)\\ &= \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\left[Q_h(s_h, a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h, a_h) | s_1 = s\right]\end{aligned}$$

Key idea: The difference in initial value for the same policy under two MDPs decomposes into layer-wise errors.

Lemma 2 (Bellman Residual Decomposition)

For any pair of MDPs $M = (P^M, R^M)$ *and* $\widehat{M} = (P^{\widehat{M}}, R^{\widehat{M}})$*, any* $s \in \mathcal{S}$*, and policies* $\pi \in \Pi_{rns}$*,*

$$V_1^{M,\pi}(s) - V_1^{\widehat{M},\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1}) \middle| s_1 = s \right]$$

In addition, for any M *and* $Q = (Q_1, \ldots, Q_H, 0)$ *(need not to be a value function), letting* $\pi_{Q,h}(s) = \arg\max_{a \in \mathcal{A}} Q_h(s, a)$*, we have*

$$\begin{aligned} &\max_{a \in \mathcal{A}} Q_1(s,a) - V_1^{M,\pi_Q}(s) \\ &= \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\left[Q_h(s_h, a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h, a_h) | s_1 = s \right] \end{aligned}$$

Key idea: The difference in initial value for the same policy under two MDPs decomposes into layer-wise errors.

Lemma 2 (Bellman Residual Decomposition)

For any pair of MDPs $M = (P^M, R^M)$ *and* $\widehat{M} = (P^{\widehat{M}}, R^{\widehat{M}})$*, any* $s \in \mathcal{S}$*, and policies* $\pi \in \Pi_{rns}$*,*

$$V_1^{M,\pi}(s) - V_1^{\widehat{M},\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{\widehat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1}) \middle| s_1 = s \right]$$

In addition, for any M *and* $Q = (Q_1, \ldots, Q_H, 0)$ *(need not to be a value function), letting* $\pi_{Q,h}(s) = \arg\max_{a \in \mathcal{A}} Q_h(s, a)$*, we have*

$$\begin{aligned} &\max_{a \in \mathcal{A}} Q_1(s, a) - V_1^{M,\pi_Q}(s) \\ = &\sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\left[Q_h(s_h, a_h) - [\mathcal{T}_h^M Q_{h+1}](s_h, a_h) | s_1 = s \right] \end{aligned}$$

Key idea: The difference in initial value for the same policy under two MDPs decomposes into layer-wise errors.

Optimism in Unknown MDPs

Key points:

- Construct *optimistic value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ over-estimating $Q^{M,\star}$.
- Use *Bellman residuals* to measure the self-consistency of these optimistic estimates.
- Lemma 3 :
  - Closeness of $\overline{Q}_h$ to $\mathcal{T}_h^M \overline{Q}_{h+1} \implies$ closeness of $\widehat{\pi}$ to π^M in value.
  - On-policy nature: distribution of states s_h is induced by executing $\widehat{\pi}$ in model M (roll-in distribution) instead of π^M.
- Errors do not accumulate exponentially; they remain controlled by H.

Optimism in Unknown MDPs

Key points:

- Construct *optimistic value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ over-estimating $Q^{M,\star}$.
- Use *Bellman residuals* to measure the self-consistency of these optimistic estimates.
- Lemma 3 :
  - Closeness of $\overline{Q}_h$ to $\mathcal{T}_h^M \overline{Q}_{h+1} \implies$ closeness of $\widehat{\pi}$ to π^M in value.
  - On-policy nature: distribution of states s_h is induced by executing $\widehat{\pi}$ in model M (roll-in distribution) instead of π^M.
- Errors do not accumulate exponentially; they remain controlled by H.

Optimism in Unknown MDPs

Key points:

- Construct *optimistic value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ over-estimating $Q^{M,\star}$.
- Use *Bellman residuals* to measure the self-consistency of these optimistic estimates.
- Lemma 3 :
  - Closeness of $\overline{Q}_h$ to $\mathcal{T}_h^M \overline{Q}_{h+1} \implies$ closeness of $\widehat{\pi}$ to π^M in value.
  - On-policy nature: distribution of states s_h is induced by executing $\widehat{\pi}$ in model M (roll-in distribution) instead of π^M.
- Errors do not accumulate exponentially; they remain controlled by H.

Optimism in Unknown MDPs

Key points:

- Construct *optimistic value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ over-estimating $Q^{M,\star}$.
- Use *Bellman residuals* to measure the self-consistency of these optimistic estimates.
- Lemma 3 :
  - Closeness of $\overline{Q}_h$ to $\mathcal{T}_h^M \overline{Q}_{h+1} \implies$ closeness of $\widehat{\pi}$ to π^M in value.
  - On-policy nature: distribution of states s_h is induced by executing $\widehat{\pi}$ in model M (roll-in distribution) instead of π^M.
- Errors do not accumulate exponentially; they remain controlled by H.

Optimism in Unknown MDPs

Key points:

- Construct *optimistic value functions* $\overline{Q}_1, \ldots, \overline{Q}_H$ over-estimating $Q^{M,\star}$.
- Use *Bellman residuals* to measure the self-consistency of these optimistic estimates.
- Lemma 3 :
  - Closeness of $\overline{Q}_h$ to $\mathcal{T}_h^M \overline{Q}_{h+1} \implies$ closeness of $\widehat{\pi}$ to π^M in value.
  - On-policy nature: distribution of states s_h is induced by executing $\widehat{\pi}$ in model M (roll-in distribution) instead of π^M.
- Errors do not accumulate exponentially; they remain controlled by H.

Error Decomposition for Optimistic Policies

Lemma 3

Let $\{\overline{Q}_h\}_{h=1}^{H}$ *be a sequence of optimistic value functions where* $Q_h^{M,\star}(s,a) \leq \overline{Q}_h(s,a)$, $\overline{Q}_{H+1} \equiv 0$, *and* $\widehat{\pi} = (\widehat{\pi}_1, \ldots \widehat{\pi}_H)$ *where* $\widehat{\pi}_h = \arg\max_a \overline{Q}_h(s,a)$, *then*

$$V_1^{M,\star}(s) - V_1^{M,\widehat{\pi}}(s) \leq \sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}}\left[\overline{Q}_h - \left(\mathcal{T}_h^M \overline{Q}_{h+1}\right)(s_h, \widehat{\pi}(s_h))|s_1 = s\right]$$

- If $\overline{Q}_h = Q_h^{M,\star}$, then $Q_h^{M,\star} = \mathcal{T}_h^M Q_{h+1}^{M,\star}$, then the right-hand side is 0.
- Hence, exact Bellman consistency implies no sub-optimality gap.

Error Decomposition for Optimistic Policies

Lemma 3

Let $\{\overline{Q}_h\}_{h=1}^{H}$ *be a sequence of optimistic value functions where* $Q_h^{M,\star}(s,a) \leq \overline{Q}_h(s,a)$, $\overline{Q}_{H+1} \equiv 0$, *and* $\widehat{\pi} = (\widehat{\pi}_1, \dots \widehat{\pi}_H)$ *where* $\widehat{\pi}_h = \arg\max_a \overline{Q}_h(s,a)$, *then*

$$V_1^{M,\star}(s) - V_1^{M,\widehat{\pi}}(s) \leq \sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}}\left[\overline{Q}_h - \left(\mathcal{T}_h^M \overline{Q}_{h+1}\right)(s_h, \widehat{\pi}(s_h)) | s_1 = s\right]$$

- If $\overline{Q}_h = Q_h^{M,\star}$, then $Q_h^{M,\star} = \mathcal{T}_h^M Q_{h+1}^{M,\star}$, then the right-hand side is 0.
- Hence, exact Bellman consistency implies no sub-optimality gap.

UCB-VI for Tabular MDPs: Setup

Assumptions 1.1

- *State and action spaces are small, with* $S = |\mathcal{S}|$ *and* $A = |\mathcal{A}|$
- *For simplicity,* $R_h^M(s,a) = \delta_{r_h}(s,a)$ *for some known* $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ *are known,* $V_1^{M,\star}(s) \in [0,1]$ *for any* $s \in \mathcal{S}$*;*
- *Only transition probabilities* P^M *are unknown.*

Empirical counts:

$$n_h^t(s,a) = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i) = (s,a)\},$$

$$n_h^t(s,a,s') = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

Estimated transitions prob: $\widehat{P}_h^t(s' \mid s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)}$.

UCB-VI for Tabular MDPs: Setup

Assumptions 1.1

- *State and action spaces are small, with* $S = |\mathcal{S}|$ *and* $A = |\mathcal{A}|$
- *For simplicity,* $R_h^M(s,a) = \delta_{r_h}(s,a)$ *for some known* $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ *are known,* $V_1^{M,\star}(s) \in [0,1]$ *for any* $s \in \mathcal{S}$*;*
- *Only transition probabilities* P^M *are unknown.*

Empirical counts:

$$n_h^t(s,a) = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i) = (s,a)\},$$

$$n_h^t(s,a,s') = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

Estimated transitions prob: $\widehat{P}_h^t(s' \mid s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)}$.

UCB-VI for Tabular MDPs: Setup

Assumptions 1.1

- *State and action spaces are small, with* $S = |\mathcal{S}|$ *and* $A = |\mathcal{A}|$
- *For simplicity,* $R_h^M(s,a) = \delta_{r_h}(s,a)$ *for some known* $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ *are known,* $V_1^{M,\star}(s) \in [0,1]$ *for any* $s \in \mathcal{S}$*;*
- *Only transition probabilities* P^M *are unknown.*

Empirical counts:

$$n_h^t(s,a) = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i) = (s,a)\},$$

$$n_h^t(s,a,s') = \sum_{i=1}^{t-1} \mathbb{I}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}.$$

Estimated transitions prob: $\widehat{P}_h^t(s' \mid s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)}$.

UCB-VI Algorithm

Algorithm: UCB-VI

```
for $t = 1$ to $T$ do
    $\overline{V}^t_{H+1} \leftarrow 0$;
    for $h = H$ to $0$ do
        Update $n^h_t(s,a)$,$n^h_t(s,a,s')$ and $b^t_{h,\delta}(s,a)$(defined later);
        $\overline{Q}^t_h(s,a) \leftarrow \Big(r_h(s,a) + \mathbb{E}_{s'\sim\widehat{P}^h_t(\cdot|s,a)}\big[\overline{V}^t_{h+1}(s')\big] + b^t_{h,\delta}(s,a)\Big) \wedge 1$;
        $V^t_h(s) \leftarrow \max_{a\in A} Q^t_h(s,a)$, and $\pi^b_h(s) \leftarrow \arg\max_{a\in A} Q^t_h(s,a)$;
    end
    Collect trajectory $(s^t_1, a^t_1, r^t_1), \ldots, (s^t_H, a^t_H, r^t_H)$;
end
```

Key ideas:

- *Optimism:* Augment rewards with a bonus to ensure high-probability over-estimation.
- *Surrogate Bellman operator:* Use empirical $\widehat{P}^h_t$ in place of the unknown P^M.

UCB-VI Algorithm

Algorithm: UCB-VI

```
for $t = 1$ to $T$ do
    $\overline{V}_{H+1}^t \leftarrow 0$;
    for $h = H$ to $0$ do
        Update $n_t^h(s,a)$,$n_t^h(s,a,s')$ and $b_{h,\delta}^t(s,a)$(defined later);
        $\overline{Q}_h^t(s,a) \leftarrow \left(r_h(s,a) + \mathbb{E}_{s' \sim \widehat{P}_t^h(\cdot|s,a)}\left[\overline{V}_{h+1}^t(s')\right] + b_{h,\delta}^t(s,a)\right) \wedge 1$;
        $V_h^t(s) \leftarrow \max_{a \in A} Q_h^t(s,a)$, and $\pi_h^b(s) \leftarrow \arg\max_{a \in A} Q_h^t(s,a)$;
    end
    Collect trajectory $(s_1^t, a_1^t, r_1^t), \ldots, (s_H^t, a_H^t, r_H^t)$;
end
```

Key ideas:

- *Optimism:* Augment rewards with a bonus to ensure high-probability over-estimation.
- *Surrogate Bellman operator:* Use empirical $\widehat{P}_t^h$ in place of the unknown P^M.

Design Goals for $\overline{Q}_h$ in UCB-VI

1. **Optimism:**
   - With high probability, we require

$$\overline{Q}_h(s,a) \geq Q_h^{M,\star}(s,a)$$

   - Achieved by adding a bonus $b_{h,\delta}^t(s,a)$ to $r_h(s,a)$ (analogous to widening a confidence interval).

2. **Self-Consistency:**
   - $\overline{Q}_h$ should be approximately consistent with the Bellman backup:

$$\overline{Q}_h(s,a) \approx \left[\mathcal{T}_h^M \overline{Q}_{h+1}\right](s,a) \quad \text{(lemma 3).}$$

   - This minimizes the accumulation of errors across stages.

Design Goals for $\overline{Q}_h$ in UCB-VI

1. **Optimism:**
   - With high probability, we require

     $$\overline{Q}_h(s,a) \geq Q_h^{M,\star}(s,a)$$

   - Achieved by adding a bonus $b_{h,\delta}^t(s,a)$ to $r_h(s,a)$ (analogous to widening a confidence interval).
2. **Self-Consistency:**
   - $\overline{Q}_h$ should be approximately consistent with the Bellman backup:

     $$\overline{Q}_h(s,a) \approx \left[\mathcal{T}_h^M \overline{Q}_{h+1}\right](s,a) \quad \text{(lemma 3)}.$$

   - This minimizes the accumulation of errors across stages.

Theorem 4: Regret Bound for UCB-VI

Theorem 4

For any $\delta > 0$, UCB-VI with

$$b_{h,\delta}^{t}(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_h^t(s,a)}}$$

guarantees that with probability at least $1-\delta$,

$$Reg \lesssim HS\sqrt{AT\,\log\left(SAHT/\delta\right)}.$$

Remarks

- *A slight variation (using Freedman's inequality) yields an improved rate of $O\left(H\sqrt{SAT} + poly(H,S,A)\log T\right)$.*
- *The optimal rate is $\Theta(\sqrt{HSAT})$, achievable via a more refined bonus choice and analysis.*

Theorem 4: Regret Bound for UCB-VI

Theorem 4

For any $\delta > 0$, UCB-VI with

$$b^t_{h,\delta}(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n^t_h(s,a)}}$$

guarantees that with probability at least $1-\delta$,

$$Reg \lesssim HS\sqrt{AT\,\log\left(SAHT/\delta\right)}.$$

Remarks

- *A slight variation (using Freedman's inequality) yields an improved rate of $O\left(H\sqrt{SAT} + poly(H,S,A)\log T\right)$.*
- *The optimal rate is $\Theta(\sqrt{HSAT})$, achievable via a more refined bonus choice and analysis.*

Analysis for a Single Episode

We aim to bound $\mathrm{Reg} = \sum_{t=1}^{T} \left[f^M(\pi_{M^\star}) - f^M(\pi_t) \right]$ for UCB-VI. Fix episode t and omit the superscript t for notational simplicity. Define the estimated MDP

$$\widehat{M} = \left\{ \mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}_{h=1}^{H}, \{R_h^M\}_{h=1}^{H}, d_1 \right\},$$

with Bellman operator

$$\mathcal{T}_h^{\widehat{M}} Q(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \widehat{P}_h(\cdot|s,a)} \left[\max_a Q(s',a) \right].$$

Consider $\overline{Q}_{H+1} \equiv 0$, $\overline{Q}_h(s,a) = \left\{ [\mathcal{T}_h^{\widehat{M}} \overline{Q}_{h+1}](s,a) + b_{h,\delta}(s,a) \right\} \wedge 1$ and $\overline{V}_h(s) = \max_a \overline{Q}_h(s,a)$.

Lemma 5

Suppose for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$\left| \sum_{s'} \widehat{P}_h(s' \mid s, a) V_h^{M,\star}(s') - \sum_{s'} P_h^M(s' \mid s, a) V_h^{M,\star}(s') \right| \leq b_{h,\delta}(s, a),$$

then $\overline{Q}_h \geq Q_h^{M,\star}$ and $\overline{V}_h \geq V_h^{M,\star}$.

i.e., sufficiently large $b_{h,\delta}$ bounding transition error ensures $\overline{Q}_h$ optimism.

Lemma 6

Suppose

$$\max_{V \in \{0,1\}^S} \left| \sum_{s'} \widehat{P}_h(s' \mid s, a) V(s') - \sum_{s'} P_h^M(s' \mid s, a) V(s') \right| \leq b'_{h,\delta}(s, a),$$

then $\overline{Q}_h - \mathcal{T}_h^M \overline{Q}_{h+1} \leq (b_{h,\delta} + b'_{h,\delta}) \wedge 1$.

Lemma 5

Suppose for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$\left| \sum_{s'} \widehat{P}_h(s' \mid s, a) V_h^{M,\star}(s') - \sum_{s'} P_h^M(s' \mid s, a) V_h^{M,\star}(s') \right| \leq b_{h,\delta}(s, a),$$

then $\overline{Q}_h \geq Q_h^{M,\star}$ and $\overline{V}_h \geq V_h^{M,\star}$.

i.e., sufficiently large $b_{h,\delta}$ bounding transition error ensures $\overline{Q}_h$ optimism.

Lemma 6

Suppose

$$\max_{V \in \{0,1\}^S} \left| \sum_{s'} \widehat{P}_h(s' \mid s, a) V(s') - \sum_{s'} P_h^M(s' \mid s, a) V(s') \right| \leq b'_{h,\delta}(s, a),$$

then $\overline{Q}_h - \mathcal{T}_h^M \overline{Q}_{h+1} \leq (b_{h,\delta} + b'_{h,\delta}) \wedge 1$.

Overall Regret Analysis

Bring back time index t.

Lemma 7

With probability at least $1 - \delta$*, the functions*

$$b^t_{h,\delta}(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n^t_h(s,a)}}, \text{ and } b'^t_{h,\delta}(s,a) = 8\sqrt{\frac{S\log\left(2SAHT/\delta\right)}{n^t_h(s,a)}}$$

satisfy the assumptions of Lemmas 5 and 6 for all s, a, h, t.

Now put everything together. Under the event in Lemma 7, the optimism of $\overline{Q}_h^t$ satisfies the conditions of Lemma 3 thereby guaranteeing the instantaneous regret on round t,

$$\sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[\underbrace{\left(\overline{Q}_h^t - \mathcal{T}_h^M \overline{Q}_{h+1}^t\right)}_{\leq (b_{h,\delta}+b'_{h,\delta})\wedge 1}(s_h^t, \widehat{\pi}^t(s_h^t))|s_1 = s\right]$$

Summing over t and applying Azuma-Hoeffding gives

$$\mathrm{Reg} \lesssim \sum_{t=1}^{T}\sum_{h=1}^{H}\Big(b_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t))\Big)\wedge 1 + \sqrt{HT\log(1/\delta)}.$$

Substituting the bonus term and summation bounds (details omitted), the regret is ultimately controlled by $O(H\sqrt{SAT})$.

Now put everything together. Under the event in Lemma 7, the optimism of $\overline{Q}_h^t$ satisfies the conditions of Lemma 3 thereby guaranteeing the instantaneous regret on round t,

$$\sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[\underbrace{\left(\overline{Q}_h^t - \mathcal{T}_h^M \overline{Q}_{h+1}^t\right)}_{\leq (b_{h,\delta}+b'_{h,\delta})\wedge 1}(s_h^t, \widehat{\pi}^t(s_h^t))|s_1 = s\right]$$

Summing over t and applying Azuma-Hoeffding gives

$$\text{Reg} \lesssim \sum_{t=1}^{T}\sum_{h=1}^{H}\Big(b_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t))\Big)\wedge 1 + \sqrt{HT\log(1/\delta)}.$$

Substituting the bonus term and summation bounds (details omitted), the regret is ultimately controlled by $O(H\sqrt{SAT})$.

Now put everything together. Under the event in Lemma 7, the optimism of $\overline{Q}_h^t$ satisfies the conditions of Lemma 3 thereby guaranteeing the instantaneous regret on round t,

$$\sum_{h=1}^{H} \mathbb{E}^{M,\widehat{\pi}^t}\left[\underbrace{\left(\overline{Q}_h^t - \mathcal{T}_h^M \overline{Q}_{h+1}^t\right)}_{\leq (b_{h,\delta}+b'_{h,\delta})\wedge 1}(s_h^t, \widehat{\pi}^t(s_h^t))|s_1 = s\right]$$

Summing over t and applying Azuma-Hoeffding gives

$$\text{Reg} \lesssim \sum_{t=1}^{T}\sum_{h=1}^{H}\left(b_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t)) + b'_{h,\delta}(s_h^t, \widehat{\pi}^t(s_h^t))\right)\wedge 1 + \sqrt{HT\log(1/\delta)}.$$

Substituting the bonus term and summation bounds (details omitted), the regret is ultimately controlled by $O(H\sqrt{SAT})$.

Outline

Ch 5. Reinforcement Learning: Basics

Ch 6. General Decision Making

Setting: Decision Making with Structured Observations

The protocol runs for T rounds. For $t = 1, \dots, T$:

1. The learner picks a *decision* $\pi^t \in \Pi$.
2. Nature chooses a *reward* $r^t \in \mathcal{R} \subseteq \mathbb{R}$ and an *observation* $o_t \in \mathcal{O}$ based on π^t with $\mathcal{R}$. Both the reward and observation are then observed by the learner.

Consider a stochastic variant.

Assumptions 2.1 (Stochastic Rewards and Observations)

Rewards and observations are generated independently via

$$(r^t, o^t) \sim M^\star(\cdot \mid \pi^t)$$

where $M^\star : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is the underlying model.

Setting: Decision Making with Structured Observations

The protocol runs for T rounds. For $t = 1, \dots, T$:

1. The learner picks a *decision* $\pi^t \in \Pi$.
2. Nature chooses a *reward* $r^t \in \mathcal{R} \subseteq \mathbb{R}$ and an *observation* $o_t \in \mathcal{O}$ based on π^t with $\mathcal{R}$. Both the reward and observation are then observed by the learner.

Consider a stochastic variant.

Assumptions 2.1 (Stochastic Rewards and Observations)

Rewards and observations are generated independently via

$$(r^t, o^t) \sim M^\star(\cdot \mid \pi^t)$$

where $M^\star : \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is the underlying model.

To facilitate learning and function approximation, the learner has access to a *model class* $\mathcal{M}$ that contains $M^\star$.

Assumptions 2.2 (Realizability)

$\mathcal{M}$ contains the true model $M^\star$.

For any $M \in \mathcal{M}$, define the *mean reward function*

$$f^M(\pi) := \mathbb{E}^{M,\pi}[r]$$

where $\mathbb{E}^{M,\pi}[\cdot]$ denotes the expectation under $r, o \sim M(\pi)$, and let

$$\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$$

be the *optimal decision*. Finally, define the induced class

$$\mathcal{F}_\mathcal{M} := \{f^M \mid M \in \mathcal{M}\}$$

To facilitate learning and function approximation, the learner has access to a *model class* $\mathcal{M}$ that contains $M^\star$.

Assumptions 2.2 (Realizability)

$\mathcal{M}$ contains the true model $M^\star$.

For any $M \in \mathcal{M}$, define the *mean reward function*

$$f^M(\pi) := \mathbb{E}^{M,\pi}[r]$$

where $\mathbb{E}^{M,\pi}[\cdot]$ denotes the expectation under $r, o \sim M(\pi)$, and let

$$\pi_M := \arg\max_{\pi \in \Pi} f^M(\pi)$$

be the *optimal decision*. Finally, define the induced class

$$\mathcal{F}_\mathcal{M} := \{f^M \mid M \in \mathcal{M}\}$$

Performance Measure: Regret

We evaluate the learner's performance in terms of regret to optimal decision for $M^\star$:

$$\mathbf{Reg} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t) \right]$$

where $p^t \in \Delta(\Pi)$ is the learner's distribution over decisions at round t.
Abbreviate $f^\star = f^{M^\star}$ and $\pi^\star = \pi_{M^\star}$ for brevity.

Examples in the DMSO Framework

The DMSO framework is general enough to capture most online decision-making problems.

- **Structured Bandits:** $\mathcal{O} = \{\varnothing\}$.
- **Contextual Bandits:**

$$\begin{aligned} &\text{Select } \pi^t : \mathcal{X} \to [A] \text{ and then observe } x^t \\ \Longleftrightarrow\ &\text{first observe } x^t \text{ and then select } \pi^t(x^t) \in [A] \end{aligned}$$

  $\mathcal{O} = \mathcal{X}, \Pi = \mathcal{X} \to [A], x \sim \mathcal{D}^M, r \sim \mathcal{R}^M(\cdot|x, \pi(x))$.

- **Online Reinforcement Learning**: $\Pi = \Pi_{\text{rns}}$, $r^t = \sum_{h=1}^{H} r_h^t$, and $o_t = \tau_t$.
- **Other Examples**:
  - Partially Observed Markov Decision Processes (POMDPs)
  - Bandits with graph-structured feedback
  - Partial monitoring

Examples in the DMSO Framework

The DMSO framework is general enough to capture most online decision-making problems.

- **Structured Bandits**: $\mathcal{O} = \{\varnothing\}$.
- **Contextual Bandits:**

$$
\begin{aligned}
&\text{Select } \pi^t : \mathcal{X} \to [A] \text{ and then observe } x^t \\
\iff &\text{first observe } x^t \text{ and then select } \pi^t(x^t) \in [A]
\end{aligned}
$$

  $\mathcal{O} = \mathcal{X}, \Pi = \mathcal{X} \to [A], x \sim \mathcal{D}^M, r \sim \mathcal{R}^M(\cdot|x, \pi(x))$.
- **Online Reinforcement Learning**: $\Pi = \Pi_{\text{rns}}, r^t = \sum_{h=1}^{H} r_h^t$, and $o_t = \tau_t$.
- **Other Examples**:
  - Partially Observed Markov Decision Processes (POMDPs)
  - Bandits with graph-structured feedback
  - Partial monitoring

Examples in the DMSO Framework

The DMSO framework is general enough to capture most online decision-making problems.

- **Structured Bandits:** $\mathcal{O} = \{\varnothing\}$.
- **Contextual Bandits:**

$$\begin{aligned} &\text{Select } \pi^t : \mathcal{X} \to [A] \text{ and then observe } x^t \\ \Longleftrightarrow\ &\text{first observe } x^t \text{ and then select } \pi^t(x^t) \in [A] \end{aligned}$$

$$\mathcal{O} = \mathcal{X}, \Pi = \mathcal{X} \to [A], x \sim \mathcal{D}^M, r \sim \mathcal{R}^M(\cdot|x, \pi(x)).$$

- **Online Reinforcement Learning**: $\Pi = \Pi_{\mathrm{rns}}$, $r^t = \sum_{h=1}^{H} r_h^t$, and $o_t = \tau_t$.
- **Other Examples**:
  - Partially Observed Markov Decision Processes (POMDPs)
  - Bandits with graph-structured feedback
  - Partial monitoring

Examples in the DMSO Framework

The DMSO framework is general enough to capture most online decision-making problems.

- **Structured Bandits**: $\mathcal{O} = \{\varnothing\}$.
- **Contextual Bandits:**

$$\begin{aligned} &\text{Select } \pi^t : \mathcal{X} \to [A] \text{ and then observe } x^t \\ \Longleftrightarrow\ &\text{first observe } x^t \text{ and then select } \pi^t(x^t) \in [A] \end{aligned}$$

  $\mathcal{O} = \mathcal{X}, \Pi = \mathcal{X} \to [A], x \sim \mathcal{D}^M, r \sim \mathcal{R}^M(\cdot|x, \pi(x))$.

- **Online Reinforcement Learning**: $\Pi = \Pi_{\text{rns}}$, $r^t = \sum_{h=1}^{H} r_h^t$, and $o_t = \tau_t$.
- Other Examples:
  - Partially Observed Markov Decision Processes (POMDPs)
  - Bandits with graph-structured feedback
  - Partial monitoring

Examples in the DMSO Framework

The DMSO framework is general enough to capture most online decision-making problems.

- **Structured Bandits**: $\mathcal{O} = \{\varnothing\}$.
- **Contextual Bandits:**

$$\begin{aligned} &\text{Select } \pi^t : \mathcal{X} \to [A] \text{ and then observe } x^t \\ \iff &\text{first observe } x^t \text{ and then select } \pi^t(x^t) \in [A] \end{aligned}$$

  $\mathcal{O} = \mathcal{X}, \Pi = \mathcal{X} \to [A], x \sim \mathcal{D}^M, r \sim \mathcal{R}^M(\cdot|x, \pi(x))$.

- **Online Reinforcement Learning**: $\Pi = \Pi_{\text{rns}}$, $r^t = \sum_{h=1}^H r_h^t$, and $o_t = \tau_t$.
- **Other Examples**:
  - Partially Observed Markov Decision Processes (POMDPs)
  - Bandits with graph-structured feedback
  - Partial monitoring

Information-Theoretic Divergences

Csiszár f-divergence: $D_f(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$ if $\mathbb{P} \ll \mathbb{Q}$.

- **Total Variation:** $f(t) = \frac{1}{2}|t-1|$

$$D_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int\left|\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu}\right|d\nu = \sup_{A\in\mathcal{F}}|\mathbb{P}(A) - \mathbb{Q}(A)|.$$

- **Squared Hellinger:** $f(t) = (1-\sqrt{t})^2$

$$D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

- **Kullback-Leibler:** $f(t) = t\log t$

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \int \log\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{P} \quad \text{if } \mathbb{P} \ll \mathbb{Q} \text{ else } +\infty.$$

Information-Theoretic Divergences

Csiszár f-divergence: $D_f(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$ if $\mathbb{P} \ll \mathbb{Q}$.

- **Total Variation:** $f(t) = \frac{1}{2}|t-1|$

$$D_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int\left|\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu}\right|d\nu = \sup_{A\in\mathcal{F}}|\mathbb{P}(A) - \mathbb{Q}(A)|.$$

- **Squared Hellinger:** $f(t) = (1-\sqrt{t})^2$

$$D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

- **Kullback-Leibler:** $f(t) = t\log t$

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \int \log\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{P} \quad \text{if } \mathbb{P} \ll \mathbb{Q} \text{ else } +\infty.$$

Information-Theoretic Divergences

Csiszár f-divergence: $D_f(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$ if $\mathbb{P} \ll \mathbb{Q}$.

- **Total Variation:** $f(t) = \frac{1}{2}|t-1|$

$$D_{\text{TV}}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int\left|\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu}\right|d\nu = \sup_{A\in\mathcal{F}}|\mathbb{P}(A) - \mathbb{Q}(A)|.$$

- **Squared Hellinger:** $f(t) = (1-\sqrt{t})^2$

$$D_{\text{H}}^2(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

- **Kullback-Leibler:** $f(t) = t\log t$

$$D_{\text{KL}}(\mathbb{P}\|\mathbb{Q}) = \int \log\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{P} \quad \text{if } \mathbb{P} \ll \mathbb{Q} \text{ else } +\infty.$$

Information-Theoretic Divergences

Csiszár f-divergence: $D_f(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right]$ if $\mathbb{P} \ll \mathbb{Q}$.

- **Total Variation:** $f(t) = \frac{1}{2}|t-1|$

$$D_{\mathrm{TV}}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int\left|\frac{d\mathbb{P}}{d\nu} - \frac{d\mathbb{Q}}{d\nu}\right|d\nu = \sup_{A\in\mathcal{F}}|\mathbb{P}(A) - \mathbb{Q}(A)|.$$

- **Squared Hellinger:** $f(t) = (1-\sqrt{t})^2$

$$D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) = \int\left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

- **Kullback-Leibler:** $f(t) = t\log t$

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \int \log\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{P} \quad \text{if } \mathbb{P} \ll \mathbb{Q} \text{ else } +\infty.$$

Lemma 8

For all distributions $\mathbb{P}$ *and* $\mathbb{Q}$*,*

$$D_{\mathrm{TV}}^2(\mathbb{P},\mathbb{Q}) \le D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q}) \le D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q})$$

Lemma 9

If $\sup_{F\in\mathcal{F}} \frac{\mathbb{P}(F)}{\mathbb{Q}(F)} \le V$*,*

$$D_{\mathrm{KL}}(\mathbb{P}\,\|\,\mathbb{Q}) \le \Big(2+\log(V)\Big) D_{\mathrm{H}}^2(\mathbb{P},\mathbb{Q})$$

Lemma 8

For all distributions $\mathbb{P}$ *and* $\mathbb{Q}$*,*

$$D_{\mathrm{TV}}^2(\mathbb{P}, \mathbb{Q}) \leq D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q}) \leq D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q})$$

Lemma 9

If $\sup_{F \in \mathcal{F}} \frac{\mathbb{P}(F)}{\mathbb{Q}(F)} \leq V$*,*

$$D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) \leq \Big(2 + \log(V)\Big) D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q})$$

(Offset) Decision-Estimation Coefficient

How to optimally explore/make decisions connects to statistical complexity (e.g. minimax regret for $\mathcal{M}$), requires coverage of

- simple problems (e.g., mean rewards suffice), and
- complex problems (e.g., structured observations provide extra information).

Definition 10

For a model class $\mathcal{M}$, reference model $\widehat{M} \in \mathcal{M}$, and scale parameter $\gamma > 0$, the DEC is defined via

$$\operatorname{dec}_\gamma(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[\underbrace{f^M(\pi_M) - f^M(\pi)}_{\text{reg of decision}} - \gamma \underbrace{D^2_{\mathrm{H}}(M(\pi), \widehat{M}(\pi))}_{\text{info gain for obs}}\Big]$$

and

$$\operatorname{dec}_\gamma(\mathcal{M}) := \sup_{\widehat{M} \in \operatorname{co}(\mathcal{M})} \operatorname{dec}_\gamma(\mathcal{M}, \widehat{M})$$

(Offset) Decision-Estimation Coefficient

How to optimally explore/make decisions connects to statistical complexity (e.g. minimax regret for $\mathcal{M}$), requires coverage of

- simple problems (e.g., mean rewards suffice), and
- complex problems (e.g., structured observations provide extra information).

Definition 10

For a model class $\mathcal{M}$, reference model $\widehat{M} \in \mathcal{M}$, and scale parameter $\gamma > 0$, the DEC is defined via

$$\mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[\underbrace{f^M(\pi_M) - f^M(\pi)}_{\text{reg of decision}} - \gamma \underbrace{D_{\mathrm{H}}^2\big(M(\pi), \widehat{M}(\pi)\big)}_{\text{info gain for obs}}\right]$$

and

$$\mathrm{dec}_\gamma(\mathcal{M}) := \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M})$$

E2D for General Decision Making

Algorithm: Estimation to Decision-Making (E2D) for General Decision Making

```
parameters: Exploration parameter $\gamma > 0$;
for $t = 1$ to $T$ do
    Obtain $\widehat{M}^t$ from the online estimation oracle with
        $\mathcal{H}^{t-1} = \{(\pi^1, r^1, o^1), \dots, (\pi^{t-1}, r^{t-1}, o^{t-1})\};$
    Compute
        $p^t \leftarrow \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}^t(\pi)\right)\right];$
    Sample decision $\pi^t \sim p^t$ and update estimation algorithm with
    $(\pi^t, r^t, o^t)$;
end
```

Regret Bound for E2D

Estimation error for the estimation oracle is defined via

$$\mathbf{Est}_{\mathrm{H}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathrm{H}}\big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\big)\right]$$

Proposition 2.1

E2D with exploration parameter $\gamma > 0$ guarantees that, almost surely,

$$\mathbf{Reg} \leq \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T \; + \; \gamma \cdot \mathbf{Est}_{\mathrm{H}}$$

For any finite class, it is possible to achieve

$$\mathbf{Reg} \leq \mathrm{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \log(|\,\mathcal{M}\,|/\delta)$$

with probability at least $1 - \delta$.

Regret Bound for E2D

Estimation error for the estimation oracle is defined via

$$\mathbf{Est}_{\mathrm{H}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathrm{H}}\big(M^\star(\pi^t), \widehat{M}^t(\pi^t)\big)\right]$$

Proposition 2.1

E2D with exploration parameter $\gamma > 0$ guarantees that, almost surely,

$$\mathbf{Reg} \le \sup_{\widehat{M} \in \widehat{\mathcal{M}}} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) \cdot T \ + \ \gamma \cdot \mathbf{Est}_{\mathrm{H}}$$

For any finite class, it is possible to achieve

$$\mathbf{Reg} \le \mathrm{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \log(|\,\mathcal{M}\,|/\delta)$$

with probability at least $1 - \delta$.

Notions of Optimality

Optimality notions vary; here we focus on *minimax optimality*.

Definition 11 (Minimax Regret)

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{p_1, \dots, p_T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}(T)]$$

where $p^t = p^t(\cdot | \mathcal{H}^{t-1})$

Remarks

We will say that an algorithm is minimax optimal if it achieves minimax regret up to absolute constants that do not depend on $\mathcal{M}$ or T.

Notions of Optimality

Optimality notions vary; here we focus on *minimax optimality*.

Definition 11 (Minimax Regret)

$$\mathfrak{M}(\mathcal{M}, T) = \inf_{p_1,\ldots,p_T} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star, p}[\mathbf{Reg}(T)]$$

where $p^t = p^t(\cdot | \mathcal{H}^{t-1})$

Remarks

We will say that an algorithm is minimax optimal if it achieves minimax regret up to absolute constants that do not depend on $\mathcal{M}$ or T.

Constrained DEC

Definition 12 (Constrained DEC)

For $\varepsilon > 0$, $\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}, \widehat{M})$ is defined as

$$\inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M) - f^{M}(\pi) \right] \,\middle|\, \mathbb{E}_{\pi \sim p}\left[D^2_{\mathrm{H}}\big(M(\pi), \widehat{M}(\pi)\big) \right] \leq \varepsilon^2 \right\},$$

with

$$\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}) := \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}^{c}_{\varepsilon}\left(\mathcal{M} \cup \{\widehat{M}\}, \widehat{M} \right).$$

Proposition 2.2

Define the localized subclass

$$\mathcal{M}_{\alpha}(\widehat{M}) = \{ M \in \mathcal{M} : f^{\widehat{M}}(\pi_{\widehat{M}}) \geq f^{M}(\pi_M) - \alpha \},$$

then for all $\varepsilon > 0$ and $\gamma \geq c_1 \varepsilon^{-1}$,

$$\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}) \leq c_3 \cdot \sup_{\gamma \geq c_1 \varepsilon^{-1}} \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_{\gamma}\left(\mathcal{M}_{\alpha(\varepsilon, \gamma)}(\widehat{M}), \widehat{M} \right)$$

Constrained DEC

Definition 12 (Constrained DEC)

For $\varepsilon > 0$, $\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}, \widehat{M})$ is defined as

$$\inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M) - f^{M}(\pi) \right] \middle| \mathbb{E}_{\pi \sim p}\left[D^2_{\mathrm{H}}\left(M(\pi), \widehat{M}(\pi) \right) \right] \leq \varepsilon^2 \right\},$$

with

$$\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}) := \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}^{c}_{\varepsilon}\left(\mathcal{M} \cup \{\widehat{M}\}, \widehat{M} \right).$$

Proposition 2.2

Define the localized subclass

$$\mathcal{M}_{\alpha}(\widehat{M}) = \{ M \in \mathcal{M} : f^{\widehat{M}}(\pi_{\widehat{M}}) \geq f^{M}(\pi_M) - \alpha \},$$

then for all $\varepsilon > 0$ and $\gamma \geq c_1 \varepsilon^{-1}$,

$$\mathrm{dec}^{c}_{\varepsilon}(\mathcal{M}) \leq c_3 \cdot \sup_{\gamma \geq c_1 \varepsilon^{-1}} \sup_{\widehat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_{\gamma}\left(\mathcal{M}_{\alpha(\varepsilon,\gamma)}(\widehat{M}), \widehat{M} \right)$$

DEC is Necessary and Sufficient

Proposition 2.3 (DEC Lower Bound)

Let $\underline{\varepsilon}_T := \frac{1}{\sqrt{T}}$ for some sufficiently small constant $c > 0$. If $\operatorname{dec}^c_{\underline{\varepsilon}_T}(\mathcal{M}) \geq 10\,\underline{\varepsilon}_T$ for all T, then $\exists M \in \mathcal{M}$ for which

$$\mathbb{E}[\mathbf{Reg}(T)] \gtrsim dec^c_{\underline{\varepsilon}_T}(\mathcal{M}) \cdot T$$

Proposition 2.4 (Upper bound for constrained DEC)

For a finite $\mathcal{M}$ and set $\overline{\varepsilon}_T := c\sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}$ for some sufficiently large constant c. Under some conditions, there exists an algorithm achieving

$$\mathbb{E}[\mathbf{Reg}(T)] \lesssim dec^c_{\overline{\varepsilon}_T}(\mathcal{M}) \cdot T$$

with prob. at least $1 - \delta$.

DEC is Necessary and Sufficient

Proposition 2.3 (DEC Lower Bound)

Let $\underline{\varepsilon}_T := \frac{1}{\sqrt{T}}$ *for some sufficiently small constant* $c > 0$*. If* $\mathrm{dec}^c_{\underline{\varepsilon}_T}(\mathcal{M}) \geq 10\,\underline{\varepsilon}_T$ *for all* T*, then* $\exists M \in \mathcal{M}$ *for which*

$$\mathbb{E}[\mathbf{Reg}(T)] \gtrsim dec^c_{\underline{\varepsilon}_T}(\mathcal{M}) \cdot T$$

Proposition 2.4 (Upper bound for constrained DEC)

For a finite $\mathcal{M}$ *and set* $\overline{\varepsilon}_T := c\sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}$ *for some sufficiently large constant* c*. Under some conditions, there exists an algorithm achieving*

$$\mathbb{E}[\mathbf{Reg}(T)] \lesssim dec^c_{\overline{\varepsilon}_T}(\mathcal{M}) \cdot T$$

with prob. at least $1 - \delta$*.*

Application to Tabular RL

- **Model Class $\mathcal{M}$:** All non-stationary MDPs

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

  with state space $S = [\mathcal{S}]$, action space $A = [\mathcal{A}]$, horizon H and normalized rewards (i.e. $\sum_{h=1}^H r_h \in [0, 1]$ a.s.).

- **Decision Space Π:** $\Pi = \Pi_{\text{rns}}$ — the set of all randomized, non-stationary Markov policies.

- **Occupancy Measures:**

$$d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s), \quad d_h^{M,\pi}(s, a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a).$$

  For all M and policy π, $d^1_{M,\pi}(s) = d_1(s)$.

Application to Tabular RL

- **Model Class $\mathcal{M}$:** All non-stationary MDPs

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

  with state space $S = [\mathcal{S}]$, action space $A = [\mathcal{A}]$, horizon H and normalized rewards (i.e. $\sum_{h=1}^H r_h \in [0, 1]$ a.s.).

- **Decision Space Π:** $\Pi = \Pi_{\text{rns}}$ — the set of all randomized, non-stationary Markov policies.

- **Occupancy Measures:**

$$d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s), \quad d_h^{M,\pi}(s, a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a).$$

  For all M and policy π, $d_{M,\pi}^1(s) = d_1(s)$.

Application to Tabular RL

- **Model Class $\mathcal{M}$:** All non-stationary MDPs

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

  with state space $S = [\mathcal{S}]$, action space $A = [\mathcal{A}]$, horizon H and normalized rewards (i.e. $\sum_{h=1}^H r_h \in [0,1]$ a.s.).
- **Decision Space Π:** $\Pi = \Pi_{\mathrm{rns}}$ — the set of all randomized, non-stationary Markov policies.
- **Occupancy Measures:**

$$d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s), \quad d_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a).$$

  For all M and policy π, $d_{M,\pi}^1(s) = d_1(s)$.

Application to Tabular RL

- **Model Class $\mathcal{M}$:** All non-stationary MDPs

$$M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^n, d_1\}$$

with state space $S = [\mathcal{S}]$, action space $A = [\mathcal{A}]$, horizon H and normalized rewards (i.e. $\sum_{h=1}^H r_h \in [0, 1]$ a.s.).

- **Decision Space Π:** $\Pi = \Pi_{\text{rns}}$ — the set of all randomized, non-stationary Markov policies.
- **Occupancy Measures:**

$$d_h^{M,\pi}(s) = \mathbb{P}^{M,\pi}(s_h = s), \quad d_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a).$$

For all M and policy π, $d_{M,\pi}^1(s) = d_1(s)$.

PC-IGW

Algorithm: Policy Cover Inverse Gap Weighting (PC-IGW)

parameters: Estimated model $\widehat{M}$, Exploration parameter $\eta > 0$;
Define *inverse gap weighted policy cover* $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ via

$$\pi_{h,s,a} \leftarrow \arg\max_{\pi\in\Pi_{\mathrm{rns}}} \frac{d_h^{\widehat{M},\pi}(s,a)}{2HSA + \eta\big(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)\big)};$$

For each $\pi \in \Psi \cup \{\pi_{\widehat{M}}\}$, define $p(\pi) = \frac{1}{\lambda+\eta\left(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi)\right)}$ with $\lambda \in [1, 2HSA]$ chosen s.t. $\sum_\pi p(\pi) = 1$;
return p

Proposition 2.5

For tabular RL setting, PC-IGW with $\eta = \frac{\gamma}{21H^2}$ *and* $\widehat{M} \in \mathcal{M}$ *ensures* $\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma D^2_{\mathrm{H}}\big(M(\pi), \widehat{M}(\pi)\big)\right] \lesssim \frac{H^3SA}{\gamma}$ *and consequently* $\mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) \lesssim \frac{H^3SA}{\gamma}$.

PC-IGW

Algorithm: Policy Cover Inverse Gap Weighting (PC-IGW)

parameters: Estimated model $\widehat{M}$, Exploration parameter $\eta > 0$;
Define *inverse gap weighted policy cover* $\Psi = \{\pi_{h,s,a}\}_{h\in[H],s\in[S],a\in[A]}$ via

$$\pi_{h,s,a} \leftarrow \underset{\pi\in\Pi_{\mathrm{rns}}}{\arg\max} \frac{d_h^{\widehat{M},\pi}(s,a)}{2HSA + \eta\big(f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi)\big)};$$

For each $\pi \in \Psi \cup \{\pi_{\widehat{M}}\}$, define $p(\pi) = \frac{1}{\lambda+\eta\left(f^{\widehat{M}}(\pi_{\widehat{M}})-f^{\widehat{M}}(\pi)\right)}$ with
$\lambda \in [1, 2HSA]$ chosen s.t. $\sum_\pi p(\pi) = 1$;
return p

Proposition 2.5

For tabular RL setting, PC-IGW with $\eta = \frac{\gamma}{21H^2}$ *and* $\widehat{M} \in \mathcal{M}$ *ensures* $\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma\, D^2_{\mathrm{H}}\big(M(\pi), \widehat{M}(\pi)\big)\Big] \lesssim \frac{H^3SA}{\gamma}$ *and consequently* $\mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}) \lesssim \frac{H^3SA}{\gamma}$.

Lower Bound on DEC

- **Obtain proper estimator $\widehat{M} \in \mathcal{M}$ instead of $\mathrm{co}(\mathcal{M})$:**
  - At each t, given $\{(\pi^i, r^i, o^i)_{i=1}^{t-1}\}$, use layerwise estimator $\mathbf{Alg}_{\mathrm{Est};h}$ to get an estimator $\widehat{P}_h^t$ for $P_h^{M^\star}$
  - Measure performance via layer-wise Hellinger error

$$\mathbf{Est}_{\mathrm{H};h} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}^{M^\star, \pi^t}\left[D_{\mathrm{H}}^2\left(P_h^{M^\star}(s_h, a_h), \widehat{P}_h^t(s_h, a_h)\right)\right]$$

  - Obtain an estimator for the full model by taking $\widehat{M}^t$ with $\widehat{P}_h^t$
  - The estimator above has $\mathbf{Est}_{\mathrm{H}} \leq O(\log(H)) \sum_{h=1}^{H} \mathbf{Est}_{\mathrm{H};h}$ and $\widehat{M}^t \in \mathcal{M}$.

Proposition 2.6

For tabular MDPs with $S \geq 2$, $A \geq 2$, and $H \geq 2\log_2(S/2)$,

$$\mathrm{dec}_\varepsilon^{\mathrm{c}}(\mathcal{M}) \gtrsim \varepsilon\sqrt{HSA} \text{ and hence } \mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{HSAT}.$$

Lower Bound on DEC

- **Obtain proper estimator $\widehat{M} \in \mathcal{M}$ instead of $\mathrm{co}(\mathcal{M})$:**
  - At each t, given $\{(\pi^i, r^i, o^i)_{i=1}^{t-1}\}$, use layerwise estimator $\mathbf{Alg}_{\mathrm{Est};h}$ to get an estimator $\widehat{P}_h^t$ for $P_h^{M^\star}$
  - Measure performance via layer-wise Hellinger error

$$\mathbf{Est}_{\mathrm{H};h} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \mathbb{E}^{M^\star, \pi^t} \left[D_{\mathrm{H}}^2 \left(P_h^{M^\star}(s_h, a_h), \widehat{P}_h^t(s_h, a_h) \right) \right]$$

  - Obtain an estimator for the full model by taking $\widehat{M}^t$ with $\widehat{P}_h^t$
  - The estimator above has $\mathbf{Est}_{\mathrm{H}} \leq O(\log(H)) \sum_{h=1}^{H} \mathbf{Est}_{\mathrm{H};h}$ and $\widehat{M}^t \in \mathcal{M}$.

Proposition 2.6

For tabular MDPs with $S \geq 2$, $A \geq 2$, and $H \geq 2\log_2(S/2)$,

$$\mathsf{dec}_\varepsilon^{\mathrm{c}}(\mathcal{M}) \gtrsim \varepsilon\sqrt{HSA} \text{ and hence } \mathbb{E}[\mathbf{Reg}] \gtrsim \sqrt{HSAT}.$$

Guarantees Based on Decision Space Complexity

Key Idea: Low estimation complexity (small bound on $\mathbf{Est}_{\mathrm{H}}$ or $\log|\mathcal{M}|$) is not needed everywhere; focusing on regions critical for distinguishing decision quality suffices.

Proposition 2.7

There exists an algorithm s.t. $\forall\,\delta > 0$, with prob. at least $1-\delta$,

$$\mathbf{Reg} \lesssim \inf_{\gamma>0}\left\{\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M}))\cdot T + \gamma\cdot\log\left(\frac{|\Pi|}{\delta}\right)\right\}.$$

Remarks

- *Replace* $\log|\mathcal{M}|$ *with smaller* $\log|\Pi|$, $\mathrm{dec}_\gamma(\mathcal{M})$ *with the potentially larger* $\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M}))$
- *For convex $\mathcal{M}$ (e.g., multi-armed, linear, convex bandits), this provides strict improvement.*
- *For non-convex ones (e.g., tabular MDPs), the trade-off differs.*

Guarantees Based on Decision Space Complexity

Key Idea: Low estimation complexity (small bound on $\mathbf{Est}_\mathrm{H}$ or $\log|\mathcal{M}|$) is not needed everywhere; focusing on regions critical for distinguishing decision quality suffices.

Proposition 2.7

There exists an algorithm s.t. $\forall\, \delta > 0$, *with prob. at least* $1 - \delta$,

$$\mathbf{Reg} \lesssim \inf_{\gamma>0}\left\{\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M})) \cdot T + \gamma \cdot \log\left(\frac{|\Pi|}{\delta}\right)\right\}.$$

Remarks

- *Replace* $\log|\mathcal{M}|$ *with smaller* $\log|\Pi|$, $\mathrm{dec}_\gamma(\mathcal{M})$ *with the potentially larger* $\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M}))$
- *For convex* $\mathcal{M}$ *(e.g., multi-armed, linear, convex bandits), this provides strict improvement.*
- *For non-convex ones (e.g., tabular MDPs), the trade-off differs.*

Guarantees Based on Decision Space Complexity

Key Idea: Low estimation complexity (small bound on $\mathbf{Est}_{\mathrm{H}}$ or $\log|\mathcal{M}|$) is not needed everywhere; focusing on regions critical for distinguishing decision quality suffices.

Proposition 2.7

There exists an algorithm s.t. $\forall\,\delta > 0$, *with prob. at least* $1-\delta$,

$$\mathbf{Reg} \lesssim \inf_{\gamma>0}\Big\{\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M}))\cdot T + \gamma\cdot\log\big(\frac{|\Pi|}{\delta}\big)\Big\}.$$

Remarks

- *Replace* $\log|\mathcal{M}|$ *with smaller* $\log|\Pi|$, $\mathrm{dec}_\gamma(\mathcal{M})$ *with the potentially larger* $\mathrm{dec}_\gamma(\mathrm{co}(\mathcal{M}))$
- *For convex* $\mathcal{M}$ *(e.g., multi-armed, linear, convex bandits), this provides strict improvement.*
- *For non-convex ones (e.g., tabular MDPs), the trade-off differs.*

General Divergences and Randomized Estimators

Algorithm: E2D for General Divergences and Randomized Estimators

```
parameters: Exploration parameter $\gamma > 0$, divergence $D(\cdot \| \cdot)$;
for $t = 1$ to $T$ do
    Obtain randomized estimate $\nu^t \in \Delta(\mathcal{M})$ from estimation oracle with $\{(\pi^i, r^i, o^i)\}_{i<t}$;
    Compute
        $p^t \leftarrow \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[ f^M(\pi_M) - f^M(\pi) - \gamma \, \mathbb{E}_{\widehat{M} \sim \nu^t}\left[ D^\pi\left( M(\pi) \| \widehat{M}^t(\pi) \right) \right] \right];$
    Sample decision $\pi^t \sim p^t$ and update estimation algorithm with $(\pi^t, r^t, o^t)$;
end
```

- **Motivation:**
  - **Generalized distance**: Beyond squared Hellinger distance, use a general divergence $D_\pi(\cdot\|\cdot)$.
    - $\exists\,\Psi$ and $\psi:\mathcal{M}\to\Psi$, s.t. $D^\pi(M\|M')=D^\pi(\psi(M)\|M')$, $f^M(\pi)=f^{\psi(M)}(\pi)$ and $\pi_M=\pi_{\psi(M)}$ for all M,M'.
    - Can derive bounds on **Est** scaling with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.
  - **Generalized DEC** $\overline{\mathrm{dec}}^D_\gamma(\mathcal{M},\nu)$:

$$\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f_M(\pi_M)-f_M(\pi)-\gamma\cdot\mathbb{E}_{\widehat{M}\sim\nu}\left[D_\pi(\widehat{M}\|M)\right]\right].$$

  - **Randomized Estimators**: Instead of a point estimate, produce a distribution $\nu^t\in\Delta(\mathcal{M})$.

Proposition 2.8 (Regret Guarantee)

Define $\mathbf{Est}_{\mathrm{D}}:=\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t,\widehat{M}\sim\nu^t}D^{\pi^t}(\widehat{M}\|M^\star)$*, we have*

$$\mathbf{Reg}\le\overline{\mathrm{dec}}^D_\gamma(\mathcal{M})\cdot T+\gamma\cdot\mathbf{Est}_{\mathrm{D}}.$$

- **Motivation**:
  - **Generalized distance**: Beyond squared Hellinger distance, use a general divergence $D_\pi(\cdot\|\cdot)$.
    - $\exists\, \Psi$ and $\psi : \mathcal{M} \to \Psi$, s.t. $D^\pi(M\|M') = D^\pi(\psi(M)\|M')$, $f^M(\pi) = f^{\psi(M)}(\pi)$ and $\pi_M = \pi_{\psi(M)}$ for all M, M'.
    - Can derive bounds on **Est** scaling with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.
  - **Generalized DEC** $\overline{\operatorname{dec}}^D_\gamma(\mathcal{M}, \nu)$:

$$\inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[f_M(\pi_M) - f_M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu}\big[D_\pi(\widehat{M}\|M)\big]\Big].$$

  - **Randomized Estimators**: Instead of a point estimate, produce a distribution $\nu^t \in \Delta(\mathcal{M})$.

Proposition 2.8 (Regret Guarantee)

Define $\mathbf{Est}_{\mathrm{D}} := \sum_{t=1}^T \mathbb{E}_{\pi^t \sim p^t, \widehat{M} \sim \nu^t} D^{\pi^t}(\widehat{M}\|M^\star)$, *we have*

$$\mathbf{Reg} \le \overline{\operatorname{dec}}^D_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathrm{D}}.$$

- **Motivation**:
    - **Generalized distance**: Beyond squared Hellinger distance, use a general divergence $D_\pi(\cdot\|\cdot)$.
        - $\exists\,\Psi$ and $\psi : \mathcal{M} \to \Psi$, s.t. $D^\pi(M\|M') = D^\pi(\psi(M)\|M')$, $f^M(\pi) = f^{\psi(M)}(\pi)$ and $\pi_M = \pi_{\psi(M)}$ for all M, M'.
        - Can derive bounds on **Est** scaling with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.
    - **Generalized DEC** $\overline{\mathrm{dec}}^D_\gamma(\mathcal{M},\nu)$:

$$\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f_M(\pi_M) - f_M(\pi) - \gamma\cdot\mathbb{E}_{\widehat{M}\sim\nu}\left[D_\pi(\widehat{M}\|M)\right]\right].$$

    - **Randomized Estimators**: Instead of a point estimate, produce a distribution $\nu^t \in \Delta(\mathcal{M})$.

Proposition 2.8 (Regret Guarantee)

Define $\mathbf{Est}_{\mathrm{D}} := \sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t,\widehat{M}\sim\nu^t} D^{\pi^t}(\widehat{M}\|M^\star)$*, we have*

$$\mathbf{Reg} \le \overline{\mathrm{dec}}^D_\gamma(\mathcal{M})\cdot T + \gamma\cdot\mathbf{Est}_{\mathrm{D}}.$$

- **Motivation**:
  - **Generalized distance**: Beyond squared Hellinger distance, use a general divergence $D_\pi(\cdot\|\cdot)$.
    - $\exists\, \Psi$ and $\psi : \mathcal{M} \to \Psi$, s.t. $D^\pi(M\|M') = D^\pi(\psi(M)\|M')$, $f^M(\pi) = f^{\psi(M)}(\pi)$ and $\pi_M = \pi_{\psi(M)}$ for all M, M'.
    - Can derive bounds on **Est** scaling with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.
  - **Generalized DEC** $\overline{\operatorname{dec}}^D_\gamma(\mathcal{M}, \nu)$:

$$\inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f_M(\pi_M) - f_M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M}\sim\nu}\left[D_\pi(\widehat{M}\|M)\right]\right].$$

  - **Randomized Estimators**: Instead of a point estimate, produce a distribution $\nu^t \in \Delta(\mathcal{M})$.

Proposition 2.8 (Regret Guarantee)

Define $\mathbf{Est}_{\mathrm{D}} := \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t, \widehat{M}\sim\nu^t} D^{\pi^t}(\widehat{M}\|M^\star)$, *we have*

$$\mathbf{Reg} \le \overline{\operatorname{dec}}^D_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathrm{D}}.$$

- **Motivation**:
  - **Generalized distance**: Beyond squared Hellinger distance, use a general divergence $D_\pi(\cdot\|\cdot)$.
    - $\exists\, \Psi$ and $\psi : \mathcal{M} \to \Psi$, s.t. $D^\pi(M\|M') = D^\pi(\psi(M)\|M')$, $f^M(\pi) = f^{\psi(M)}(\pi)$ and $\pi_M = \pi_{\psi(M)}$ for all M, M'.
    - Can derive bounds on **Est** scaling with $\log|\Psi|$ instead of $\log|\mathcal{M}|$.
  - **Generalized DEC** $\overline{\mathrm{dec}}^D_\gamma(\mathcal{M}, \nu)$:

$$\inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f_M(\pi_M) - f_M(\pi) - \gamma \cdot \mathbb{E}_{\widehat{M}\sim\nu}\left[D_\pi(\widehat{M}\|M)\right]\right].$$

  - **Randomized Estimators**: Instead of a point estimate, produce a distribution $\nu^t \in \Delta(\mathcal{M})$.

Proposition 2.8 (Regret Guarantee)

Define $\mathbf{Est}_\mathrm{D} := \sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t, \widehat{M}\sim\nu^t} D^{\pi^t}(\widehat{M}\|M^\star)$, *we have*

$$\mathbf{Reg} \le \overline{\mathrm{dec}}^D_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_\mathrm{D}.$$

Optimistic Estimation and E2D.Opt

- Incorporates a bonus to encourage over-estimate $f^{M^\star}(\pi_{M^\star})$.
- **Optimistic Estimation Error:**

$$\mathbf{OptEst}_\gamma^D = \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t,\, \widehat{M}^t\sim\nu^t}\left[D^{\pi_t}(\widehat{M}^t\|M^\star) + \gamma^{-1}\left(f^{M^\star}(\pi_{M^\star}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t})\right)\right]$$

- **Optimistic DEC:**

$$\text{o-dec}_\gamma^D(\mathcal{M},\nu) = \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p,\, \widehat{M}\sim\nu}\left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f_M(\pi) - \gamma\, D^{\pi}(\widehat{M}\|M)\right].$$

Proposition 2.9

E2D.Opt *ensures that*

$$\mathbf{Reg} \le \text{o-dec}_\gamma^D(\mathcal{M})\cdot T + \gamma\cdot\mathbf{OptEst}_\gamma^D$$

almost surely.

Optimistic Estimation and E2D.Opt

- Incorporates a bonus to encourage over-estimate $f^{M^\star}(\pi_{M^\star})$.
- **Optimistic Estimation Error:**

$$\mathbf{OptEst}_\gamma^D = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t, \widehat{M}^t \sim \nu^t} \left[D^{\pi_t}(\widehat{M}^t \| M^\star) + \gamma^{-1} \left(f^{M^\star}(\pi_{M^\star}) - f^{\widehat{M}^t}(\pi_{\widehat{M}^t}) \right) \right]$$

- **Optimistic DEC:**

$$\text{o-dec}_\gamma^D(\mathcal{M}, \nu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p, \widehat{M} \sim \nu} \left[f^{\widehat{M}}(\pi_{\widehat{M}}) - f_M(\pi) - \gamma D^\pi(\widehat{M} \| M) \right].$$

Proposition 2.9

E2D.Opt *ensures that*

$$\mathbf{Reg} \le \text{o-dec}_\gamma^D(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{OptEst}_\gamma^D$$

almost surely.