Efficient Regret Minimizing Strategies for Tabular Average-Reward MDPs

M. Sadegh Talebi
Inria Lille - Nord Europe (SequeL)

Joint work with
Odalric-Ambrym Maillard (Inria) and Hippolyte Bourel (Montanuniversität)

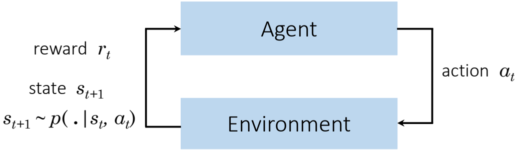
8ème Journée COSMOS, November 2019

Undiscounted RL: MDP Model

We consider reinforcement learning (RL), where the environment is modeled as an undiscounted Markov Decision Process (MDP).

Undiscounted MDP $M = (\mathcal{S}, \mathcal{A}, p, \mu)$:

- State-space $\mathcal{S}$ with cardinality S
- Action-space $\mathcal{A}$ with cardinality A
- Transition kernel p: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$ leads to a transition to s' with probability $p(s'|s,a)$.
- Reward function μ: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$, gives $r(s,a)$ with mean $\mu(s,a)$.



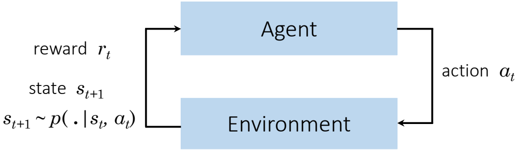
p and μ are unknown, and the goal is to maximize $\sum_{t=1}^{T} r_t$.

Undiscounted RL: MDP Model

We consider reinforcement learning (RL), where the environment is modeled as an undiscounted Markov Decision Process (MDP).

Undiscounted MDP $M = (\mathcal{S}, \mathcal{A}, p, \mu)$:

- State-space $\mathcal{S}$ with cardinality S
- Action-space $\mathcal{A}$ with cardinality A
- Transition kernel p: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$ leads to a transition to s' with probability $p(s'|s,a)$.
- Reward function μ: Selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$, gives $r(s,a)$ with mean $\mu(s,a)$.



p and μ are unknown, and the goal is to maximize $\sum_{t=1}^{T} r_t$.

Undiscounted RL: Objective

Goal: To maximize the collected reward $\sum_{t=1}^{T} r_t$.

- A (Markov deterministic) **policy** π is a mapping from $\mathcal{S}$ to $\mathcal{A}$.
- **Gain** (or long-term average reward) of a policy π is defined as

$$g^{\pi}(s_1) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t, \pi(s_t))\right]$$

- **Assumption:** We consider communicating MDPs in which every state is reachable from any other state by some appropriate policy. For communicating MDPs, g^{π} does not depend on s_1.

Undiscounted RL: Objective

Goal: To maximize the collected reward $\sum_{t=1}^{T} r_t$.

- A (Markov deterministic) **policy** π is a mapping from $\mathcal{S}$ to $\mathcal{A}$.
- **Gain** (or long-term average reward) of a policy π is defined as

$$g^\pi(s_1) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t, \pi(s_t))\right]$$

- **Assumption:** We consider communicating MDPs in which every state is reachable from any other state by some appropriate policy. For communicating MDPs, g^π does not depend on s_1.

Undiscounted RL: Bellman's Equation

Any policy achieving $g^\star := \max_\pi g^\pi$ is called gain-optimal.

Bellman's Optimality Equation (Poisson Equation)

$$g^\star + b^\star(s) = \max_{a \in \mathcal{A}} \Big(\mu(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) b^\star(s') \Big), \quad \forall s$$

where $g^\star$ is called the maximal gain and $b^\star$ is called the optimal bias function.

- In the long run, maximal cumulative reward is achieved by following a gain-optimal policy.
- If MDP is known, one can find $g^\star$ and $b^\star$ by solving Bellman's optimality equation using numerical methods (e.g., Value Iteration).

Undiscounted RL: Bellman's Equation

Any policy achieving $g^\star := \max_\pi g^\pi$ is called gain-optimal.

Bellman's Optimality Equation (Poisson Equation)

$$g^\star + b^\star(s) = \max_{a\in\mathcal{A}}\Big(\mu(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)b^\star(s')\Big), \quad \forall s$$

where $g^\star$ is called the maximal gain and $b^\star$ is called the optimal bias function.

- In the long run, maximal cumulative reward is achieved by following a gain-optimal policy.
- If MDP is known, one can find $g^\star$ and $b^\star$ by solving Bellman's optimality equation using numerical methods (e.g., Value Iteration).

Undiscounted RL: Regret

Goal: To maximize the collected reward $\sum_{t=1}^{T} r_t$.

Regret: Defined as the difference between cumulative reward of the optimal policy $\star$ and that gathered by the decision-maker (in expectation or w.h.p.):

$$\mathrm{Regret}_T := \sum_{t=1}^{T} r_t^\star - \sum_{t=1}^{T} r_t$$

Alternatively, the objective of the decision-maker is to minimize the regret.

By Azuma-Hoeffding's inequality, with probability at least $1-\delta$,

$$\mathrm{Regret}_T := Tg^\star - \sum_{t=1}^{T} r_t + \mathcal{O}(\sqrt{T\log(2/\delta)})$$

So it makes sense to control the following notion of regret:

$$\mathfrak{R}_T := Tg^\star - \sum_{t=1}^{T} r_t$$

Undiscounted RL: Regret

Goal: To maximize the collected reward $\sum_{t=1}^{T} r_t$.

Regret: Defined as the difference between cumulative reward of the optimal policy $\star$ and that gathered by the decision-maker (in expectation or w.h.p.):

$$\mathrm{Regret}_T := \sum_{t=1}^{T} r_t^\star - \sum_{t=1}^{T} r_t$$

Alternatively, the objective of the decision-maker is to minimize the regret.

By Azuma-Hoeffding's inequality, with probability at least $1-\delta$,

$$\mathrm{Regret}_T := Tg^\star - \sum_{t=1}^{T} r_t + \mathcal{O}(\sqrt{T\log(2/\delta)})$$

So it makes sense to control the following notion of regret:

$$\mathfrak{R}_T := Tg^\star - \sum_{t=1}^{T} r_t$$

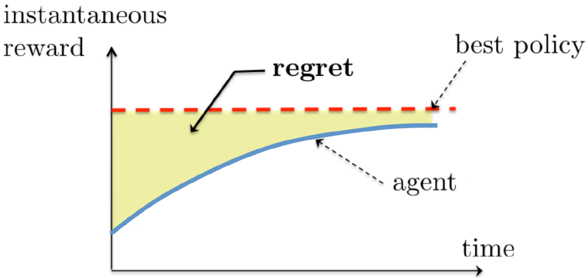
Undiscounted RL: Regret

Alternatively, the objective of the decision-maker is to minimize the regret.

$$\mathfrak{R}_T := Tg^\star - \sum_{t=1}^{T} r_t$$

The key difficulty to do so is to balance *exploration* vs. *exploitation*:

- Play the best action so far, ...
- ... or rather explore a different action?



Outline

Outline

RL Algorithms

Two main approaches in RL:

- Model-Based: Consists in maintaining an approximate MDP model through estimating μ and p, and deriving a value function from the *approximate* MDP.
  - Examples: UCB1, UCRL2.
- Model-Free: Directly learns a value function (without estimating μ and p).
  - Example: Variants of Q-learning.

In this talk we focus on model-based algorithms.

RL Algorithms

Two main approaches in RL:

- Model-Based: Consists in maintaining an approximate MDP model through estimating μ and p, and deriving a value function from the *approximate* MDP.
  - Examples: UCB1, UCRL2.
- Model-Free: Directly learns a value function (without estimating μ and p).
  - Example: Variants of Q-learning.

In this talk we focus on model-based algorithms.

Notations

Under a given algorithm, we define:

- $N_t(s,a)$: number of visits, up to time t, to (s,a).
- $N_t(s,a,s')$: number of visits, up to time t, to (s,a) followed by a visit to s'.
- Empirical estimates of transition probabilities and rewards:

$$\widehat{\mu}_t(s,a) = \frac{\sum_{t'=0}^{t-1} r_{t'}\mathbb{I}\{s_{t'} = s, a_{t'} = a\}}{N_t(s,a)^+}$$

$$\widehat{p}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a)^+}$$

with $N_t(s,a)^+ := \max\{N_t(s,a), 1\}$.

UCRL2

UCRL2 (Jaksch et al., 2010): a model-based algorithm inspired by UCB for stochastic bandits:

- Maintains confidence bounds for μ and p, and chooses an optimistic model that leads to the highest gain g.

Given $\delta \in (0,1)$, UCRL2 defines a set $\mathcal{M}_{t,\delta}$ of plausible MDPs (models) at time t as a collection of candidate MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', p')$ satisfying: For all s, a,

$$\left\|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\right\|_1 \leq \sqrt{\frac{14S}{N_t(s,a)} \log\left(\frac{2At}{\delta}\right)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \sqrt{\frac{7}{2N_t(s,a)} \log\left(\frac{2SAt}{\delta}\right)}$$

$\Rightarrow$ With high probability, $M \in \mathcal{M}_{t,\delta}$.

UCRL2

UCRL2 (Jaksch et al., 2010): a model-based algorithm inspired by UCB for stochastic bandits:

- Maintains confidence bounds for μ and p, and chooses an optimistic model that leads to the highest gain g.

Given $\delta \in (0,1)$, UCRL2 defines a set $\mathcal{M}_{t,\delta}$ of plausible MDPs (models) at time t as a collection of candidate MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', p')$ satisfying: For all s, a,

$$\left\|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\right\|_1 \leq \sqrt{\frac{14S}{N_t(s,a)} \log\left(\frac{2At}{\delta}\right)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \sqrt{\frac{7}{2N_t(s,a)} \log\left(\frac{2SAt}{\delta}\right)}$$

$\Rightarrow$ With high probability, $M \in \mathcal{M}_{t,\delta}$.

UCRL2

UCRL2 (Jaksch et al., 2010): a model-based algorithm inspired by UCB for stochastic bandits:

- Maintains confidence bounds for μ and p, and chooses an optimistic model that leads to the highest gain g.

Given $\delta \in (0,1)$, UCRL2 defines a set $\mathcal{M}_{t,\delta}$ of plausible MDPs (models) at time t as a collection of candidate MDPs $M' = (\mathcal{S}, \mathcal{A}, \mu', p')$ satisfying:
For all s, a,

$$\left\|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\right\|_1 \leq \sqrt{\frac{14S}{N_t(s,a)} \log\left(\frac{2At}{\delta}\right)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \sqrt{\frac{7}{2N_t(s,a)} \log\left(\frac{2SAt}{\delta}\right)}$$

$\Rightarrow$ With high probability, $M \in \mathcal{M}_{t,\delta}$.

UCRL2 (Jaksch et al., 2010)

Algorithm 1 UCRL2

Initialize: For all (s,a), set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state s_1;
for episodes $k \geq 1$ **do**
  Set $t_k = t$;
  Set $N_{t_k}(s,a) = N_{t_{k-1}}(s,a) + v_k(s,a)$ for all (s,a);
  Compute $\widehat{\mu}_{t_k}(s,a)$ and $\widehat{p}_{t_k}(\cdot|s,a)$ for all (s,a);
  Compute $\pi^+_{t_k} = \texttt{EVI}\left(\widehat{\mu}_{t_k}, \widehat{p}_{t_k}, N_{t_k}, \frac{1}{\sqrt{t_k}}, \frac{\delta}{SA}\right)$;
  while $v_k(s_t, \pi^+_{t_k}(s_t)) < \max\{1, N_{t_k}(s_t, \pi^+_{t_k}(s_t))\}$ **do**
    Play $a_t = \pi^+_{t_k}(s_t)$, and observe s_{t+1} and $r_t(s_t, a_t)$;
    Set $v_k(s_t, a_t) = v_k(s_t, a_t) + 1$;
    Set $t = t + 1$;
  end while
end for

UCRL2: `EVI`

`EVI` stands for Extended Value Iteration

Algorithm 2 `EVI`$(\mu, p, N, \varepsilon, \delta)$

Initialize: $u^{(0)} \equiv 0, u^{(-1)} \equiv -\infty$, $n = 0$;
while $\max_s(u^{(n)}(s) - u^{(n-1)}(s)) - \min_s(u^{(n)}(s) - u^{(n-1)}(s)) > \varepsilon$ **do**
  For all (s, a), set $\mu'(s, a) = \mu(s, a) + \beta'_{N(s,a)}(\delta)$;
  For all (s, a), set $p'(\cdot|s, a) \in \mathrm{argmax}_{q \in \mathcal{P}(s,a)} \sum_{x \in \mathcal{S}} q(x) u^{(n)}(x)$ where

$$\mathcal{P}(s, a) := \left\{ q \in \Delta^S : \|q - p(\cdot|s, a)\|_1 \leq \beta_{N(s,a)}(\delta) \right\};$$

  For all s, update $u^{(n+1)}(s) = \max_{a \in \mathcal{A}} \left(\mu'(s, a) + \sum_{x \in \mathcal{S}} p'(x|s, a) u^{(n)}(x) \right)$;
  For all s, update $\pi_{n+1}(s) \in \mathrm{argmax}_{a \in \mathcal{A}} \left(\mu'(s, a) + \sum_{x \in \mathcal{S}} p'(x|s, a) u^{(n)}(x) \right)$;
  Set $n = n + 1$;
end while
Output: π_{n+1}

UCRL2: Regret

Definition (Diameter (Jaksch et al., 2010))

Let $T_\pi(s'|s)$ denote the first hitting time of state s' when following stationary policy π from initial state s. The diameter D of an MDP M is defined as

$$D := \max_{s \neq s'} \min_{\pi} \mathbb{E}[T_\pi(s'|s)].$$

For any communicating MDP, under UCRL2, with probability at least $1-\delta$,

$$\mathfrak{R}_T \leq 34DS\sqrt{AT\log(T/\delta)}$$

Minimax lower bound (Jaksch et al., 2010): $\Omega(\sqrt{DSAT})$

UCRL2: Regret

Definition (Diameter (Jaksch et al., 2010))

Let $T_\pi(s'|s)$ denote the first hitting time of state s' when following stationary policy π from initial state s. The diameter D of an MDP M is defined as

$$D := \max_{s \neq s'} \min_{\pi} \mathbb{E}[T_\pi(s'|s)].$$

For any communicating MDP, under UCRL2, with probability at least $1 - \delta$,

$$\mathfrak{R}_T \leq 34DS\sqrt{AT \log(T/\delta)}$$

Minimax lower bound (Jaksch et al., 2010): $\Omega(\sqrt{DSAT})$

UCRL2: Regret

Definition (Diameter (Jaksch et al., 2010))

Let $T_\pi(s'|s)$ denote the first hitting time of state s' when following stationary policy π from initial state s. The diameter D of an MDP M is defined as

$$D := \max_{s \neq s'} \min_{\pi} \mathbb{E}[T_\pi(s'|s)].$$

For any communicating MDP, under UCRL2, with probability at least $1 - \delta$,

$$\mathfrak{R}_T \leq 34DS\sqrt{AT \log(T/\delta)}$$

Minimax lower bound (Jaksch et al., 2010): $\Omega(\sqrt{DSAT})$

UCRL2

Despite its strong regret guarantee, UCRL2 does not perform well in practice (even in small environments) – In particular, it suffers from a long burn-in phase.

Drawbacks of UCRL2:

(i) Loose and polytopic set of models
(ii) Conservative optimistic policy
(iii) Inefficient stopping criterion for internal episodes

We discuss two variants of UCRL2 aiming to remove (i) and (ii).

UCRL2

Despite its strong regret guarantee, UCRL2 does not perform well in practice (even in small environments) – In particular, it suffers from a long burn-in phase.

Drawbacks of UCRL2:

(i) Loose and polytopic set of models

(ii) Conservative optimistic policy

(iii) Inefficient stopping criterion for internal episodes

We discuss two variants of UCRL2 aiming to remove (i) and (ii).

Outline

UCRL3

UCRL3 is a variant of UCRL2, with the following key differences:

- Uses tight element-wise confidence intervals for p
  - Defined for individual transition probabilities $p(s'|s,a)$, in contrast to UCRL2 that does for $p(\cdot|s,a)$.
  - Intersection of time-uniform Bernstein and Bernoulli concentration for each $p(s'|s,a)$
- Computes a less conservative optimistic policy.

To simplify the presentation, we assume that μ is known.

UCRL3

UCRL3 is a variant of UCRL2, with the following key differences:

- Uses tight element-wise confidence intervals for p
  - Defined for individual transition probabilities $p(s'|s,a)$, in contrast to UCRL2 that does for $p(\cdot|s,a)$.
  - Intersection of time-uniform Bernstein and Bernoulli concentration for each $p(s'|s,a)$
- Computes a less conservative optimistic policy.

To simplify the presentation, we assume that μ is known.

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

- $C^1_{t,\delta}(s',s,a)$ is defined using Bernstein concentration inequality, modified using **a peeling technique**.
- $C^2_{t,\delta}(s',s,a)$ is obtained by applying **the method of mixture** (a.k.a. **the Laplace method**) for sub-Gaussian random variables.

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

$$C^1_{t,\delta}(s',s,a) = \left\{ \lambda : |\widehat{p}_t(s'|s,a) - \lambda| \le \sqrt{\frac{2\lambda(1-\lambda)\ell_{N_t(s,a)}\left(\frac{\delta}{2SA}\right)}{N_t(s,a)}} + \frac{\ell_{N_t(s,a)}\left(\frac{\delta}{2SA}\right)}{3N_t(s,a)} \right\}$$

where $\ell_n(\delta) = \eta \log\left(\frac{\log(n)\log(\eta n)}{\log(\eta^2)\delta}\right)$ with $\eta = 1.12$ (an arbitrary choice).

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

$$C^1_{t,\delta}(s',s,a) = \left\{ \lambda : |\widehat{p}_t(s'|s,a) - \lambda| \le \sqrt{\frac{2\lambda(1-\lambda)\ell_{N_t(s,a)}\left(\frac{\delta}{2SA}\right)}{N_t(s,a)}} + \frac{\ell_{N_t(s,a)}\left(\frac{\delta}{2SA}\right)}{3N_t(s,a)} \right\}$$

where $\ell_n(\delta) = \eta \log\left(\frac{\log(n)\log(\eta n)}{\log(\eta^2)\delta}\right)$ with $\eta = 1.12$ (an arbitrary choice).

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

$$C^2_{t,\delta}(s',s,a) = \Big\{ \lambda \colon -\sqrt{\underline{g}(\lambda)}\beta_{N_t(s,a)}\big(\tfrac{\delta}{2SA}\big) \le \widehat{p}_t(s'|s,a) - \lambda \le \sqrt{g(\lambda)}\beta_{N_t(s,a)}\big(\tfrac{\delta}{2SA}\big) \Big\}$$

where $\beta_n(\delta) := \sqrt{\frac{(1+\frac{1}{n})\log(\sqrt{n+1}/\delta)}{2n}}$, and where

$$\underline{g}(\lambda) = \begin{cases} g(\lambda) & \text{if } \lambda < 1/2 \\ \lambda(1-\lambda) & \text{else} \end{cases}, \text{ and } g(\lambda) = \frac{1/2 - \lambda}{\log(1/\lambda - 1)}.$$

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

$$C^2_{t,\delta}(s',s,a) = \Big\{ \lambda \colon -\sqrt{\underline{g}(\lambda)}\beta_{N_t(s,a)}\big(\tfrac{\delta}{2SA}\big) \le \widehat{p}_t(s'|s,a) - \lambda \le \sqrt{g(\lambda)}\beta_{N_t(s,a)}\big(\tfrac{\delta}{2SA}\big) \Big\}$$

where $\beta_n(\delta) := \sqrt{\frac{(1+\frac{1}{n})\log(\sqrt{n+1}/\delta)}{2n}}$, and where

$$\underline{g}(\lambda) = \begin{cases} g(\lambda) & \text{if } \lambda < 1/2 \\ \lambda(1-\lambda) & \text{else} \end{cases}, \text{ and } g(\lambda) = \frac{1/2 - \lambda}{\log(1/\lambda - 1)}.$$

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

Lemma (Time-uniform confidence bounds)

For any MDP with transition function p, for all $\delta \in (0,1)$, it holds

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, \exists (s,a) \in \mathcal{S} \times \mathcal{A},\ p(\cdot|s,a) \notin \mathcal{C}_{t,\delta}(s,a) \Big) \leq \delta$$

$\Longrightarrow \mathbb{P}\big(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t,\delta}\big) \leq \delta.$

UCRL3: Set of Models

At time t, UCRL3 considers the set $\mathcal{M}_{t,\delta}$ of plausible MDPs

$$\mathcal{M}_{t,\delta} = \Big\{ M' = (\mathcal{S}, \mathcal{A}, p', \mu) : p'(\cdot|s,a) \in \mathcal{C}_{t,\delta}(s,a), \quad \forall s, a, s' \Big\}$$

where for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{C}_{t,\delta}(s,a) := \Big\{ q \in \Delta_{\mathcal{S}} : \forall s', q(s') \in \underbrace{C^1_{t,\delta}(s',s,a)}_{\text{Bernstein}} \cap \underbrace{C^2_{t,\delta}(s',s,a)}_{\text{sub-Gaussian}} \Big\}$$

Lemma (Time-uniform confidence bounds)

For any MDP with transition function p, for all $\delta \in (0,1)$, it holds

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, \exists (s,a) \in \mathcal{S} \times \mathcal{A},\ p(\cdot|s,a) \notin \mathcal{C}_{t,\delta}(s,a) \Big) \le \delta$$

$$\Longrightarrow \mathbb{P}\big(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t,\delta}\big) \le \delta.$$

UCRL3: Revisiting EVI

- To compute an optimistic policy, UCRL2 uses `EVI` as a subroutine, which involves computing

$$p_n^+ : s, a \mapsto \operatorname{argmax}\{p' u_n, p' \in \mathcal{C}_{t,\delta}(s,a)\}$$

  at iteration n of `EVI`.
- `EVI` outputs a conservative policy, in particular when transition function p has a sparse support.
- UCRL3 remedies this issue by combining `EVI` with an adaptive support selection.

Adaptive Support Selection

Given $\widetilde{\mathcal{S}} \subset \mathcal{S}$, a pair (s,a), and a function $f : \mathcal{S} \to \mathbb{R}$, define:

$$\overline{f}_{s,a}(\widetilde{\mathcal{S}}) = \max \left\{ \sum_{s' \in \tilde{\mathcal{S}}} f(s')q(s') : q \text{ s.t. } \forall s' \in \tilde{\mathcal{S}},\, q(s') \in \mathcal{C}_{t,\delta}(s',s,a) \text{ and } \sum_{s' \in \tilde{\mathcal{S}}} q(s') \leq 1 \right\}$$

$$\overline{p}_{s,a} = \operatorname{argmax} \left\{ \sum_{s' \in \tilde{\mathcal{S}}} f(s')q(s') : q \text{ s.t. } \forall s' \in \tilde{\mathcal{S}},\, q(s') \in \mathcal{C}_{t,\delta}(s',s,a) \text{ and } \sum_{s' \in \tilde{\mathcal{S}}} q(s') \leq 1 \right\}$$

Algorithm 3 Adaptive Support Selection (for (s,a))

Input: Target function f, parameter $\kappa \in (0,1)$.

Let $\widetilde{\mathcal{S}} = \operatorname{supp}\big(\widehat{p}_t(\cdot|s,a)\big) \cup \operatorname{argmax}_{s \in \mathcal{S}} f(s)$

while $\overline{f}_{s,a}(\mathcal{S} \setminus \widetilde{\mathcal{S}}) \geq \min(\kappa, \overline{f}_{s,a}(\widetilde{\mathcal{S}}))$ **do**

  Let $\tilde{s} \in \operatorname{argmax}_{s \notin \widetilde{\mathcal{S}}} f(s)$

  Set $\widetilde{\mathcal{S}} = \widetilde{\mathcal{S}} \cup \{\tilde{s}\}$

end while

Output: $\widetilde{\mathcal{S}}$, $\overline{p}_{s,a}$

UCRL3: Revisiting EVI

- Recall that UCRL2 uses `EVI` as a subroutine, which involves computing

$$p_n^+ : s, a \mapsto \mathrm{argmax}\{P' u_n, p' \in \mathcal{C}_{t,\delta}(s,a)\}$$

  at iteration n of `EVI`.
- Now, at iteration n of `EVI`, UCRL3 uses Adaptive Support Selection with $f = u_n - \min_s u_n(s)$.
- To optimize performance, we choose

$$\kappa = \kappa_{t,n}(s,a) = \frac{\mathbb{S}(u_n)|\mathrm{supp}(\hat{p}_t(\cdot|s,a))|}{\max_{s,a} N_t(s,a)^{2/3}}$$

UCRL3: Regret

Theorem

The regret under UCRL3 in any communicating MDP satisfies, uniformly over all $T \geq 1$,

$$\mathfrak{R}_T \leq 24D\sqrt{KSAT\log(\sqrt{T+1}/\delta)} + \widetilde{\mathcal{O}}(DS^{2/3}A^{2/3}T^{1/3})$$

with probability at least $1-2\delta$.

- Improves the regret of UCRL2 by a factor of $\sqrt{S/K}$.
- Holds uniformly over all $T \geq 1$.

UCRL3: Regret

Theorem

The regret under UCRL3 in any communicating MDP satisfies, uniformly over all $T \geq 1$,

$$\mathfrak{R}_T \leq 24D\sqrt{KSAT\log(\sqrt{T+1}/\delta)} + \widetilde{\mathcal{O}}(DS^{2/3}A^{2/3}T^{1/3})$$

with probability at least $1 - 2\delta$.

- Improves the regret of UCRL2 by a factor of $\sqrt{S/K}$.
- Holds uniformly over all $T \geq 1$.

Outline

Variants of UCRL2

There are variants of UCRL2, that mostly differ in the definition of models.

Two approaches to define the set of $\mathcal{M}_{t,\delta}$ of models depending on how uncertainties in p and μ are represented:

- Polytopic uncertainty sets
  - For example, models defined using Weissman's and Hoeffding's inequalities (as in UCRL2).
- Non-polytopic uncertainty sets
  - Smoother sets
  - For example, models defined using KL-divergence and Bernstein's inequality (as in (Burnetas & Katehakis, 1997), **KL-UCRL** (Filippi et al., 2010)).

Polytopic Uncertainty Sets

Polytopic uncertainty models typically provide poor representations (cf. Robust control of MDPs (Nilim & El Ghaoui, 2005) and (Filippi et al., 2010)):

(i) They could lead to inconsistent models by excluding an already observed element of kernel (i.e., $p'(x|s,a) = 0$ even though $\widehat{p}_t(x|s,a) \neq 0$ for some x).

(ii) The maximizer of a linear optimization over L_1 ball could change significantly for a small change in the value function.

Polytopic Uncertainty Sets

Polytopic uncertainty models typically provide poor representations (cf. Robust control of MDPs (Nilim & El Ghaoui, 2005) and (Filippi et al., 2010)):

(i) They could lead to inconsistent models by excluding an already observed element of kernel (i.e., $p'(x|s,a) = 0$ even though $\widehat{p}_t(x|s,a) \neq 0$ for some x).

(ii) The maximizer of a linear optimization over L_1 ball could change significantly for a small change in the value function.

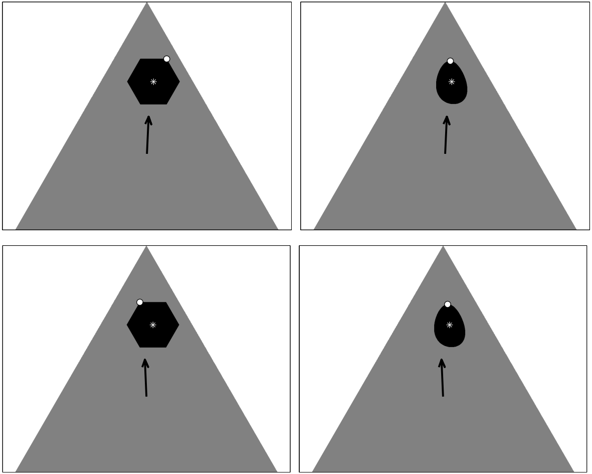
Polytopic Uncertainty Sets

Polytopic uncertainty models typically provide poor representations (cf. Robust control of MDPs (Nilim & El Ghaoui, 2005) and (Filippi et al., 2010)):

(i) They could lead to inconsistent models by excluding an already observed element of kernel (i.e., $p'(x|s,a) = 0$ even though $\widehat{p}_t(x|s,a) \neq 0$ for some x).

(ii) The maximizer of a linear optimization over L_1 ball could change significantly for a small change in the value function.

Linear optimization over L_1-ball (left) vs. KL-ball (right): The vector represents a value function (e.g., in `EVI`).



(Filippi et al., 2010)

KL-UCRL

These shortcomings are avoided by resorting to KL-based confidence bounds (as in KL-UCRL):

$$\mathtt{KL}\Big(\widehat{p}_t(\cdot|s,a), p'(\cdot|s,a)\Big) \leq \frac{\square S \log(\log(T)/\delta)}{N_t(s,a)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \sqrt{\frac{\square \log(\log(T)/\delta)}{N_t(s,a)}}$$

- Numerically, KL-UCRL **outperforms** UCRL2 (uniformly in all environment).
- Yet the best known regret bound for KL-UCRL: $\widetilde{O}(DS\sqrt{AT})$

Our contribution is to investigate the benefit of using KL theoretically.

KL-UCRL

These shortcomings are avoided by resorting to KL-based confidence bounds (as in KL-UCRL):

$$\mathtt{KL}\big(\widehat{p}_t(\cdot|s,a), p'(\cdot|s,a)\big) \leq \frac{\square S \log(\log(T)/\delta)}{N_t(s,a)}$$

$$|\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \sqrt{\frac{\square \log(\log(T)/\delta)}{N_t(s,a)}}$$

- Numerically, KL-UCRL **outperforms** UCRL2 (uniformly in all environment).
- Yet the best known regret bound for KL-UCRL: $\widetilde{\mathcal{O}}(DS\sqrt{AT})$

Our contribution is to investigate the benefit of using KL theoretically.

Variance-Aware Regret Bounds for KL-UCRL

Variance of bias function w.r.t. transition law $p(\cdot|s,a)$:

$$\mathbb{V}_{p(\cdot|s,a)}(b^\star) := \sum_{x\in\mathcal{S}} p(x|s,a)\Big(b^\star(x) - \mathbb{E}_{p(\cdot|s,a)}[b^\star]\Big)^2$$

with $\mathbb{E}_{p(\cdot|s,a)}[b^\star] = \sum_x p(\cdot|s,a)b^\star(x)$.

Theorem

The regret under KL-UCRL in any ergodic MDP satisfies

$$\mathfrak{R}_T \le \Big(31\sqrt{S\textstyle\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)} + 35S\sqrt{A} + 2D\Big)\sqrt{T\log(\log(T)/\delta)} + \widetilde{\mathcal{O}}(\mathrm{polylog}(T))$$

and with probability at least $1-\delta$.

- Improves over the previous bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for KL-UCRL (since $\mathbb{V}_{p(\cdot|s,a)}(b^\star) \le D^2$).
- Proof: Uses novel concentration inequalities

Variance-Aware Regret Bounds for KL-UCRL

Variance of bias function w.r.t. transition law $p(\cdot|s,a)$:

$$\mathbb{V}_{p(\cdot|s,a)}(b^\star) := \sum_{x\in\mathcal{S}} p(x|s,a)\Big(b^\star(x) - \mathbb{E}_{p(\cdot|s,a)}[b^\star]\Big)^2$$

with $\mathbb{E}_{p(\cdot|s,a)}[b^\star] = \sum_x p(\cdot|s,a)b^\star(x)$.

Theorem

The regret under KL-UCRL in any ergodic MDP satisfies

$$\mathfrak{R}_T \le \Big(31\sqrt{S\textstyle\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)} + 35S\sqrt{A} + 2D\Big)\sqrt{T\log(\log(T)/\delta)} + \widetilde{\mathcal{O}}(\mathrm{polylog}(T))$$

and with probability at least $1-\delta$.

- Improves over the previous bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for KL-UCRL (since $\mathbb{V}_{p(\cdot|s,a)}(b^\star) \le D^2$).
- Proof: Uses novel concentration inequalities

Variance-Aware Regret Bounds for KL-UCRL

Variance of bias function w.r.t. transition law $p(\cdot|s,a)$:

$$\mathbb{V}_{p(\cdot|s,a)}(b^\star) := \sum_{x\in\mathcal{S}} p(x|s,a)\Big(b^\star(x) - \mathbb{E}_{p(\cdot|s,a)}[b^\star]\Big)^2$$

with $\mathbb{E}_{p(\cdot|s,a)}[b^\star] = \sum_x p(\cdot|s,a)b^\star(x)$.

Theorem

The regret under KL-UCRL in any ergodic MDP satisfies

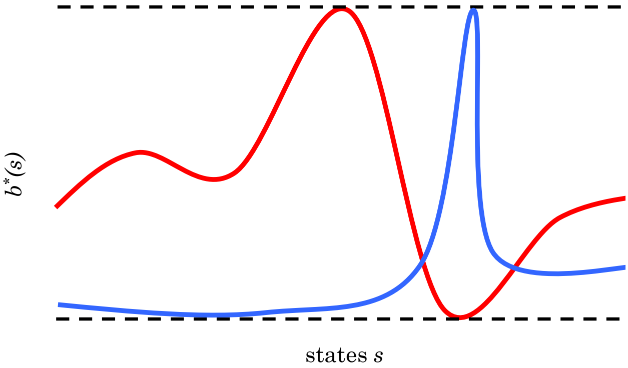
$$\mathfrak{R}_T \le \Big(31\sqrt{S\textstyle\sum_{s,a}\mathbb{V}_{p(\cdot|s,a)}(b^\star)} + 35S\sqrt{A} + 2D\Big)\sqrt{T\log(\log(T)/\delta)} + \widetilde{\mathcal{O}}(\text{polylog}(T))$$

and with probability at least $1-\delta$.

- Improves over the previous bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$ for KL-UCRL (since $\mathbb{V}_{p(\cdot|s,a)}(b^\star) \le D^2$).
- Proof: Uses novel concentration inequalities

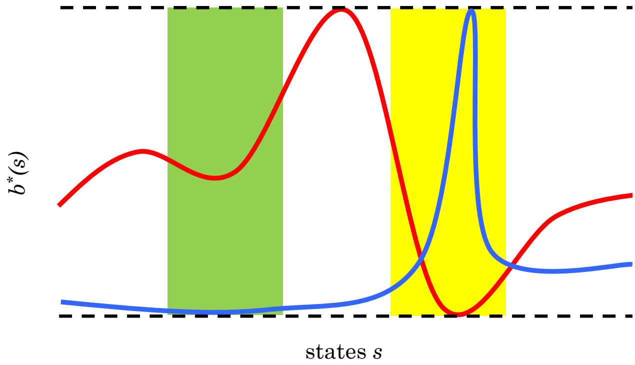
Variance vs. Diameter

In contrast to diameter D (global measures), variance $\mathbb{V}_{p(\cdot|s,a)}(b^\star)$ is a local measure, which is aware of variations of $b^\star$ over state-space.



Variance vs. Diameter

In contrast to diameter D (global measures), variance $\mathbb{V}_{p(\cdot|s,a)}(b^\star)$ is a local measure, which is aware of variations of $b^\star$ over state-space.

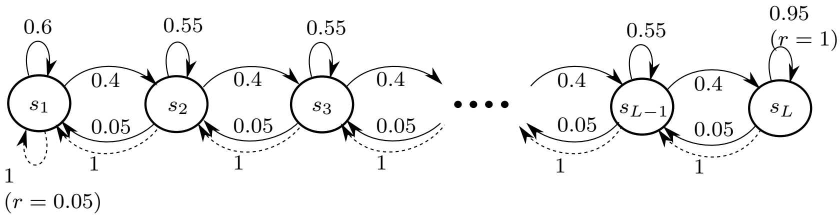


Outline

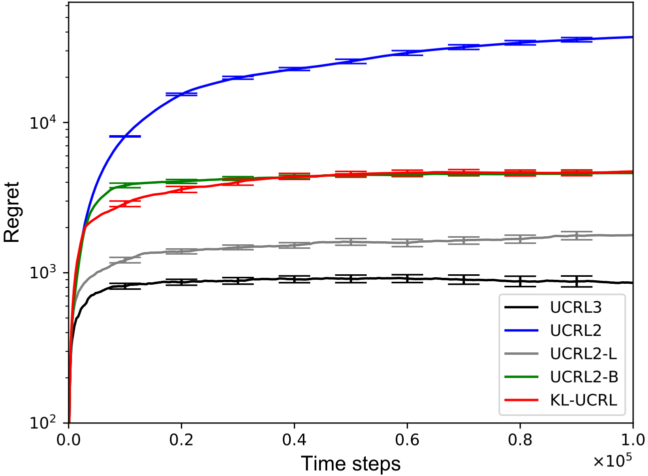
Numerical Experiments

We examine UCRL2, KL-UCRL, UCRL3, **UCRL-L**, and **UCRL-B** on the *RiverSwim* environment (shown below).

- **UCRL-L:** Uses L_1 confidence bounds (as UCRL2) combined with the Laplace method.
- **UCRL-B:** Uses element-wise empirical Bernstein confidence bounds combined with peeling.

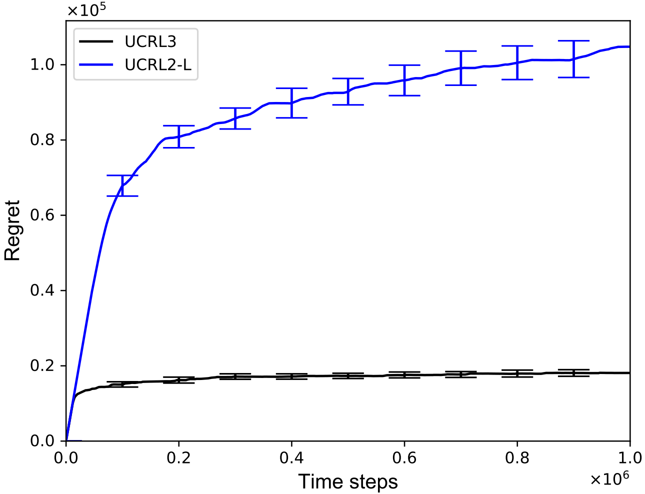


Regret of various algorithms in 6-state RiverSwim:



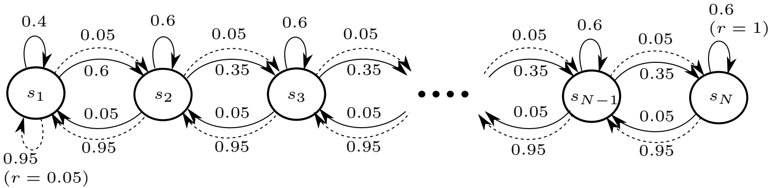
Numerical Experiments

Comparison between UCRL2-L and UCRL3 in 25-state RiverSwim:



Numerical Experiments

Examining the main terms in the regret bounds of KL-UCRL in N-state ergodic *RiverSwim* MDP:



S	D	$\max_{s,a} \mathbb{V}_{p(\cdot\|s,a)}(b^\star)$	$D\sqrt{SA}$	$\sqrt{\sum_{s,a} \mathbb{V}_{p(\cdot\|s,a)}(b^\star)}$
6	6.3	0.6322	21.9	1.8
12	14.9	0.6327	72.9	2.8
20	26.3	0.6327	166.4	3.7
40	54.9	0.6327	490.9	5.3
100	140.6	0.6327	1988.3	8.5

Outline

Old Proof(s)

The regret is decomposed into **per-episode** regret terms.

Consider episode k with **optimistic model** $\tilde{M}_k$ (with kernel $\tilde{p}_k$ and bias function $\tilde{b}_k$), and assume $M \in \mathcal{M}_t$.
The leading term in regret bound for episode k is due to:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) \le \underbrace{\big\|\tilde{p}_k(\cdot|s,a) - p(\cdot|s,a)\big\|_1}_{\mathcal{O}\Big(\sqrt{\frac{S\log(T)}{N_k(s,a)}}\Big)} \underbrace{\|\tilde{b}_k\|_\infty}_{\le D}$$

Summing over episodes k and state-action pairs (s,a), this leads to $\widetilde{\mathcal{O}}(DS\sqrt{AT})$.

Using Cauchy-Schwarz in the above leads to a too conservative bound!

Old Proof(s)

The regret is decomposed into **per-episode** regret terms.

Consider episode k with **optimistic model** $\tilde{M}_k$ (with kernel $\tilde{p}_k$ and bias function $\tilde{b}_k$), and assume $M \in \mathcal{M}_t$.
The leading term in regret bound for episode k is due to:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) \le \underbrace{\big\|\tilde{p}_k(\cdot|s,a) - p(\cdot|s,a)\big\|_1}_{\mathcal{O}\Big(\sqrt{\frac{S\log(T)}{N_k(s,a)}}\Big)} \underbrace{\|\tilde{b}_k\|_\infty}_{\le D}$$

Summing over episodes k and state-action pairs (s,a), this leads to $\widetilde{\mathcal{O}}(DS\sqrt{AT})$.

Using Cauchy-Schwarz in the above leads to a too conservative bound!

Proof Sketch

Decomposition:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) = \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k]}_{\text{transportation cost of } \tilde{b}_k}$$

$$= \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{p(\cdot|s,a)}[b^\star]}_{\text{transportation cost of } b^\star} + \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k - b^\star] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k - b^\star]}_{\text{correction term}}$$

⇒ **Transportation cost of** $b^\star$: using (novel) transportation inequalities

⇒ **Correction term**: using ergodic property of MDP + contraction of induced transition matrices. The total contribution of correction terms (over all (s,a) and k):

$$\widetilde{\mathcal{O}}(S\sqrt{AT})$$

Proof Sketch

Decomposition:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) = \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k]}_{\text{transportation cost of } \tilde{b}_k}$$

$$= \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{p(\cdot|s,a)}[b^\star]}_{\text{transportation cost of } b^\star} + \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k - b^\star] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k - b^\star]}_{\text{correction term}}$$

⇒ **Transportation cost of $b^\star$**: using (novel) transportation inequalities

⇒ **Correction term**: using ergodic property of MDP + contraction of induced transition matrices. The total contribution of correction terms (over all (s,a) and k):

$$\widetilde{\mathcal{O}}(S\sqrt{AT})$$

Proof Sketch

Decomposition:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) = \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k]}_{\text{transportation cost of } \tilde{b}_k}$$

$$= \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{p(\cdot|s,a)}[b^\star]}_{\text{transportation cost of } b^\star} + \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k - b^\star] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k - b^\star]}_{\text{correction term}}$$

$\Rightarrow$ **Transportation cost of $b^\star$**: using (novel) transportation inequalities

$\Rightarrow$ **Correction term**: using ergodic property of MDP + contraction of induced transition matrices. The total contribution of correction terms (over all (s,a) and k):

$$\widetilde{\mathcal{O}}(S\sqrt{AT})$$

Proof Sketch

Decomposition:

$$\sum_x \big(\tilde{p}_k(x|s,a) - p(x|s,a)\big)\tilde{b}_k(x) = \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k]}_{\text{transportation cost of } \tilde{b}_k}$$

$$= \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{p(\cdot|s,a)}[b^\star]}_{\text{transportation cost of } b^\star} + \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[\tilde{b}_k - b^\star] - \mathbb{E}_{p(\cdot|s,a)}[\tilde{b}_k - b^\star]}_{\text{correction term}}$$

- $\Rightarrow$ **Transportation cost of $b^\star$**: using (novel) transportation inequalities
- $\Rightarrow$ **Correction term**: using ergodic property of MDP + contraction of induced transition matrices. The total contribution of correction terms (over all (s,a) and k):

$$\widetilde{\mathcal{O}}(S\sqrt{AT})$$

Transportation Inequalities

Lemma (Transportation Lemma)

For any function f, *introduce* $\varphi_f : \lambda \mapsto \log \mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f]))]$. *Then for all* $Q \ll P$,

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \inf\{x \geq 0 : \varphi_{\star,f}(x) > KL(Q,P)\}$$
$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \geq \sup\{x \leq 0 : \varphi_{\star,f}(x) > KL(Q,P)\}$$

where $\varphi_{\star,f}(x) = \sup_\lambda \lambda x - \varphi_f(\lambda)$.

Lemma (Transportation Inequality I)

For any function f *and distribution* P, *such that* $\mathbb{V}_P(f)$ *and* $\mathbb{S}(f)$ *are finite*

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \sqrt{2\mathbb{V}_P(f) KL(Q,P)} + \frac{2}{3}\mathbb{S}(f) KL(Q,P)$$
$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] \leq \sqrt{2\mathbb{V}_P(f) KL(Q,P)}$$

Transportation Inequalities

Lemma (Transportation Lemma)

For any function f, *introduce* $\varphi_f : \lambda \mapsto \log \mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f]))]$. *Then for all* $Q \ll P$,

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \inf\{x \geq 0 : \varphi_{\star,f}(x) > KL(Q,P)\}$$
$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \geq \sup\{x \leq 0 : \varphi_{\star,f}(x) > KL(Q,P)\}$$

where $\varphi_{\star,f}(x) = \sup_\lambda \lambda x - \varphi_f(\lambda)$.

Lemma (Transportation Inequality I)

For any function f *and distribution* P, *such that* $\mathbb{V}_P(f)$ *and* $\mathbb{S}(f)$ *are finite*

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \sqrt{2\mathbb{V}_P(f) KL(Q,P)} + \frac{2}{3}\mathbb{S}(f) KL(Q,P)$$
$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] \leq \sqrt{2\mathbb{V}_P(f) KL(Q,P)}$$

Transportation Inequalities

A novel refinement of previous transportation inequality:

Lemma (Transportation Inequality II)

For any function f and distributions P, Q defined on a finite alphabet $\mathcal{X}$,

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2KL(P,Q)} + \mathbb{S}(f)KL(P,Q)$$

where $\mathcal{V}_{P,Q}(f) := \sum_{x\in\mathcal{X}:P(x)\geq Q(x)} P(x)(f(x) - \mathbb{E}_P[f])^2$.

The operator $\mathcal{V}_{P,Q}(f)$ is closely related to the local variance of f (under P and Q):

$$\mathcal{V}_{P,Q}(f) \leq \mathbb{V}_P(f)$$

$$\sqrt{\mathcal{V}_{P,Q}(f)} \leq \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathrm{KL}(Q,P)}$$

Proof: Cauchy-Schwarz + local Pinsker's inequalities

Transportation Inequalities

A novel refinement of previous transportation inequality:

Lemma (Transportation Inequality II)

For any function f and distributions P, Q defined on a finite alphabet $\mathcal{X}$,

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2KL(P,Q)} + \mathbb{S}(f)KL(P,Q)$$

where $\mathcal{V}_{P,Q}(f) := \sum_{x\in\mathcal{X}:P(x)\geq Q(x)} P(x)(f(x) - \mathbb{E}_P[f])^2$.

The operator $\mathcal{V}_{P,Q}(f)$ is closely related to the local variance of f (under P and Q):

$$\mathcal{V}_{P,Q}(f) \leq \mathbb{V}_P(f)$$
$$\sqrt{\mathcal{V}_{P,Q}(f)} \leq \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathrm{KL}(Q,P)}$$

Proof: Cauchy-Schwarz + local Pinsker's inequalities

Transportation Inequalities

A novel refinement of previous transportation inequality:

Lemma (Transportation Inequality II)

For any function f and distributions P, Q defined on a finite alphabet $\mathcal{X}$,

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \left(\sqrt{\mathcal{V}_{P,Q}(f)} + \sqrt{\mathcal{V}_{Q,P}(f)}\right)\sqrt{2KL(P,Q)} + \mathbb{S}(f)KL(P,Q)$$

where $\mathcal{V}_{P,Q}(f) := \sum_{x\in\mathcal{X}:P(x)\geq Q(x)} P(x)(f(x) - \mathbb{E}_P[f])^2$.

The operator $\mathcal{V}_{P,Q}(f)$ is closely related to the local variance of f (under P and Q):

$$\mathcal{V}_{P,Q}(f) \leq \mathbb{V}_P(f)$$
$$\sqrt{\mathcal{V}_{P,Q}(f)} \leq \sqrt{2\mathbb{V}_Q(f)} + 3\mathbb{S}(f)\sqrt{|\mathcal{X}|\mathrm{KL}(Q,P)}$$

Proof: Cauchy-Schwarz + local Pinsker's inequalities

Proof Sketch

$$\text{transportation cost of } b^\star = \underbrace{\mathbb{E}_{\tilde{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{\hat{p}_k(\cdot|s,a)}[b^\star]}_{T_1} + \underbrace{\mathbb{E}_{\hat{p}_k(\cdot|s,a)}[b^\star] - \mathbb{E}_{p(\cdot|s,a)}[b^\star]}_{T_2}$$

- ⇒ Term T_1: **Transportation Inequality II** with $P = \tilde{p}_k(\cdot|s,a)$ and $Q = \hat{p}_k(\cdot|s,a)$
- ⇒ Term T_2: **Transportation Inequality I** with $Q = \hat{p}_k(\cdot|s,a)$ and $P = p(\cdot|s,a)$

Combining, and summing over (s,a) and episodes k, the contribution of T_2 terms become

$$\widetilde{\mathcal{O}}\left(\sqrt{S \textstyle\sum_{s,a} \mathbb{V}_{p(\cdot|s,a)}(b^\star) T}\right)$$

Conclusion

Two variants of UCRL2: UCRL3 and KL-UCRL

UCRL3:

- A novel variant of UCRL2 using (i) improved confidence sets, and (ii) novel efficient approach for computing an optimistic policy.
- Beats all existing variants of UCRL2 in practice yet enjoying the same regret guarantees.

KL-UCRL:

- A variant of UCRL2, which uses KL-divergence to define confidence sets.
- We provided improved regret analysis for it in ergodic MDPs, thanks to novel variants of transportation concentration inequalities.

Future Work

- Optimal stopping criterion for UCRL2-style algorithms
- Problem-dependent regret lower and upper bounds for average-reward RL

Acknowledgement

Thank you!