Some open problems for average reward MDPs

Ronald Ortner

Montanuniversitaet Leoben

EWRL MMXVI

Apologies

- This talk is about things I can't do.

Apologies

- This talk is about things I can't do.
- Solving the problems of this talk won't solve RL.

Apologies

- This talk is about things I can't do.
- Solving the problems of this talk won't solve RL.
- Rather: regret bounds and the UCRL algorithm

Outline

1 Introduction

2 Regret Dependence on Number of States

3 The Diameter

4 Optimal Policies with Bounded Bias

5 Digression on Continuous State MDPs

6 Recent Approaches

Setting: Markov Decision Processes

Definition

Markov decision process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, s_1, p, r \rangle$:

$$
\begin{array}{lll}
\mathcal{S} & \ldots & \text{state space} \\
\mathcal{A} & \ldots & \text{a set of actions available in each state}
\end{array}
$$

- Start in initial state s_1.
- When choosing action a in state s:
  - ▷ random reward with mean $r(s, a)$ in $[0, 1]$,
  - ▷ transition to next state according to transition probability distributions probability $p(\cdot|s, a)$.

Policies, their Average Reward and the Diameter

Definition

A *policy* on an MDP $\mathcal{M}$ is a mapping $\pi : \mathcal{S} \to \mathcal{A}$.

Definition

The *average reward* of a policy is

$$\rho(\mathcal{M}, \pi) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(s_t, \pi(s_t)),$$

where s_t is the state at step t.

Definition

The *diameter* of an MDP is the maximal expected time it takes to reach any state from any other state (using an appropriate policy).

Optimal Policies and the Regret

The Learner's Goal(s):

1. Find optimal policy $\pi^* = \arg\max_\pi \rho(\mathcal{M}, \pi)$.
2. Do this online, so that you don't lose too much w.r.t. $\rho^* := \rho(\mathcal{M}, \pi^*)$.

$\rightsquigarrow$ Minimize the *regret*:

Definition

The learner's *regret* after T steps is

$$T\rho^* - \sum_{t=1}^{T} r_t,$$

where r_t is the random reward the learner receives at step t.

Algorithm for the Finite Case: UCRL

UCRL (Auer, Jaksch, Ortner 2008 & 2010)

For episodes $k = 1, 2, \ldots$ do:

1. Maintain UCB-like confidence intervals for rewards and transition probabilities to define set of plausible MDPs $\mathbb{M}$.
2. Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi).$$

3. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled.

Regret of UCRL

Theorem (Jaksch et al., 2010)

In an MDP with S states, A actions, and diameter D with probability of at least $1 - \delta$ the regret of UCRL after T steps is bounded by

$$34 \cdot DS\sqrt{AT \log\left(\frac{T}{\delta}\right)}.$$

Regret of UCRL

Theorem (Jaksch et al., 2010)

In an MDP with S states, A actions, and diameter D with probability of at least $1-\delta$ the regret of UCRL after T steps is bounded by

$$34 \cdot DS\sqrt{AT \log\left(\tfrac{T}{\delta}\right)}.$$

Proof Idea:

$$\tilde{\rho}(\tilde{\mathcal{M}}, \tilde{\pi}) \;\geq\; \rho^* \;=\; \rho(\mathcal{M}, \pi^*) \geq \; \rho(\mathcal{M}, \tilde{\pi}).$$

so that the regret is upper bounded by the sum over the confidence intervals in each step

$$\sum_k \sum_{s,a} v_k(s,a) \cdot \mathrm{conf_k(s,a)} \;\leq\; \tilde{O}(DS\sqrt{AT}).$$

A Lower Bound on the Regret

Theorem (Jaksch et al. 2010)

For any algorithm and any natural numbers T, S, $A > 1$, and $D \geq \log_A S$ there is an MDP $\mathcal{M}$ with S states, A actions, and diameter D, such that for any initial state s the expected regret after T steps is

$$\Omega\big(\sqrt{DSAT}\big).$$

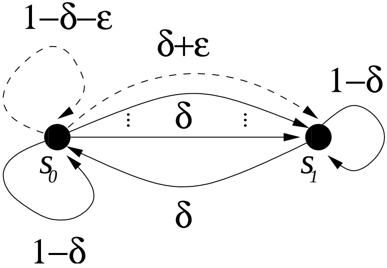
This is close to the upper bound, but there is a gap of $\sqrt{DS}$.

Outline

1. Introduction
2. Regret Dependence on Number of States
3. The Diameter
4. Optimal Policies with Bounded Bias
5. Digression on Continuous State MDPs
6. Recent Approaches

Lower Bound Example

Consider the following MDP:

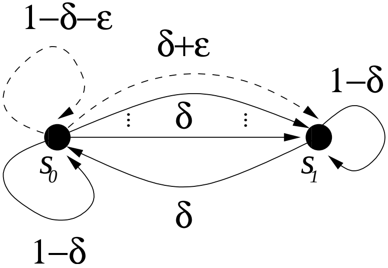


- State s_0 gives 0 reward, state s_1 gives reward 1.
- One action in s_0 has higher transition probability to s_1.
- Learner has to find this action to obtain sublinear regret.

Lower Bound Example

Now consider S copies of the MDP with

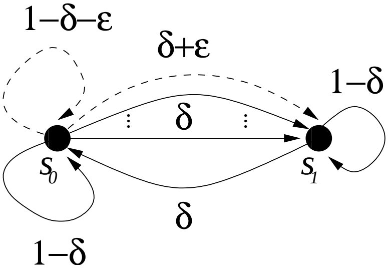
- common state s_0,
- only one copy has the special action.



Lower Bound Example

Notes on the example:

- Each state has only two possible successor states.



Regret of UCRL

Theorem (Jaksch et al., 2010)

In an MDP with S states, A actions, diameter D and transition probability distributions with support $\leq B$ with probability of at least $1-\delta$ the regret of UCRL after T steps is bounded by

$$34 \cdot D\sqrt{SBAT \log\left(\tfrac{T}{\delta}\right)}.$$

$\rightsquigarrow$ Any lower bound example showing that regret is linear in S must have transition probability distributions with support $\Omega(S)$.

Lower Bounds

$\rightsquigarrow$ Any lower bound example showing that regret is linear in S must have transition probability distributions with support $\Omega(S)$.

Lower Bounds

$\rightsquigarrow$ Any lower bound example showing that regret is linear in S must have transition probability distributions with support $\Omega(S)$.

In principle, the error for estimating transition probability distribution from N samples is $\Omega\left(\sqrt{\frac{S}{N}}\right)$.

Yet, it's hard to find an MDP where basically all S^2A transition probabilities matter.

Lower Bounds

$\rightsquigarrow$ Any lower bound example showing that regret is linear in S must have transition probability distributions with support $\Omega(S)$.

In principle, the error for estimating transition probability distribution from N samples is $\Omega\left(\sqrt{\frac{S}{N}}\right)$.

Yet, it's hard to find an MDP where basically all S^2A transition probabilities matter.

OPEN PROBLEM 1

Decide how regret depends on number of states.

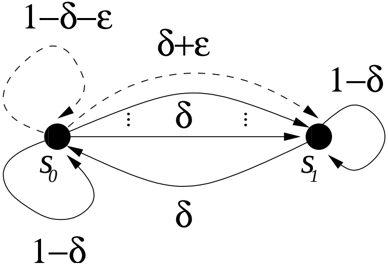
Outline

1. Introduction
2. Regret Dependence on Number of States
3. The Diameter
4. Optimal Policies with Bounded Bias
5. Digression on Continuous State MDPs
6. Recent Approaches

Lower Bound Example

Notes on the example:

- Each state has only two possible successor states.
- Choosing a wrong action once does not seem to cause long-term regret.



Lower Bounds

One can show that an error of ε in the transition probabilities can amount to an error of $D \cdot \varepsilon$ in average reward.

$\rightsquigarrow$ Intuitively for proper ε the regret should be linear in D.

Lower Bounds

One can show that an error of ε in the transition probabilities can amount to an error of $D \cdot \varepsilon$ in average reward.

$\leadsto$ Intuitively for proper ε the regret should be linear in D.

OPEN PROBLEM 2

Decide how regret depends on diameter.

UCRL and the Diameter

Where does the diameter in the regret bound for UCRL come from?

- When bounding

$$T\rho^* - \sum_{t=1}^{T} r_t \leq T\tilde{\rho}(\tilde{\mathcal{M}}, \tilde{\pi}) - \sum_{t=1}^{T} r(s_t, \tilde{\pi}(s_t))$$

UCRL and the Diameter

Where does the diameter in the regret bound for UCRL come from?

- When bounding

$$
\begin{aligned}
T\rho^* - \sum_{t=1}^{T} r_t \;&\leq\; T\tilde{\rho}(\tilde{\mathcal{M}},\tilde{\pi}) - \sum_{t=1}^{T} r(s_t,\tilde{\pi}(s_t)) \\
&\leq\; T\tilde{\rho}(\tilde{\mathcal{M}},\tilde{\pi}) - \sum_{t=1}^{T} \tilde{r}(s_t,\tilde{\pi}(s_t)) + \sum_{t=1}^{T} \big(\tilde{r}(s_t,\tilde{\pi}(s_t)) - r(s_t,\tilde{\pi}(s_t))\big),
\end{aligned}
$$

UCRL and the Diameter

Where does the diameter in the regret bound for UCRL come from?

- When bounding

$$
\begin{aligned}
T\rho^* - \sum_{t=1}^{T} r_t \;&\leq\; T\tilde{\rho}(\tilde{\mathcal{M}},\tilde{\pi}) - \sum_{t=1}^{T} r(s_t,\tilde{\pi}(s_t)) \\
&\leq\; T\tilde{\rho}(\tilde{\mathcal{M}},\tilde{\pi}) - \sum_{t=1}^{T} \tilde{r}(s_t,\tilde{\pi}(s_t)) + \sum_{t=1}^{T} \big(\tilde{r}(s_t,\tilde{\pi}(s_t)) - r(s_t,\tilde{\pi}(s_t))\big),
\end{aligned}
$$

  we have to relate the average reward $\tilde{\rho}(\tilde{\mathcal{M}},\tilde{\pi})$ of the chosen policy to the individual rewards $\tilde{r}(s_t,\tilde{\pi}(s_t))$ earned at each step t.

The Poisson equation

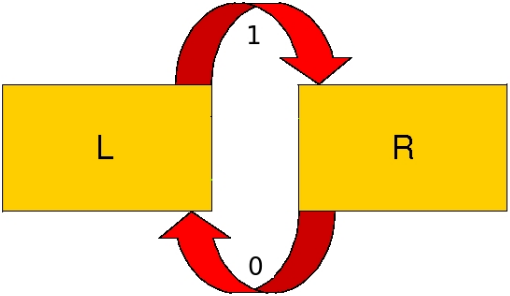
- We have to relate the average reward $\tilde{\rho}(\tilde{\mathcal{M}}, \tilde{\pi})$ of the chosen policy to the individual rewards $\tilde{r}(s_t, \tilde{\pi}(s_t))$ earned at each step t.

Poisson equation

$$\rho(\pi) - r(s, \pi(s)) = \sum_{s'} p(s'|s, \pi(s)) \cdot \lambda_\pi(s') - \lambda_\pi(s),$$

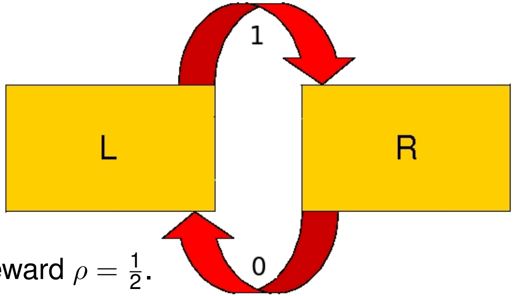
where $\lambda_\pi(s)$ is the bias of state s.

- Intuitively, the bias indicates how much you gain/lose in accumulated rewards w.r.t. average reward when starting in state s.

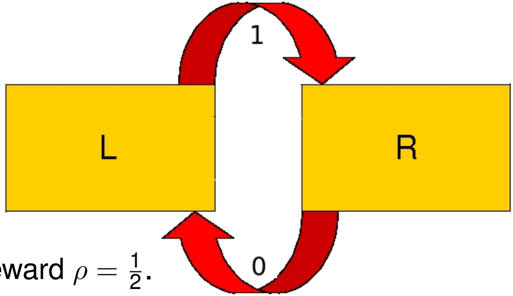


- Average reward $\rho = \frac{1}{2}$.
- Poisson equation:

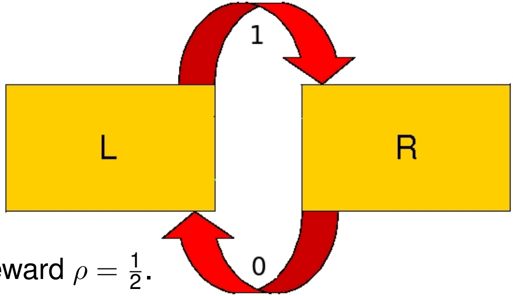
$$
\begin{aligned}
\rho - r(L) &= \lambda(R) - \lambda(L) \\
\rho - r(R) &= \lambda(L) - \lambda(R)
\end{aligned}
$$



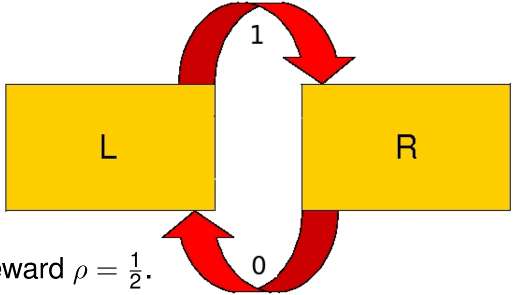
- Average reward $\rho = \frac{1}{2}$.
- Bias $\lambda(L) = \frac{1}{4}$, $\lambda(R) = -\frac{1}{4}$.
- **Interpretation:** Accumulated reward after $t = 1, 2, \ldots$ steps ...
  - ... when starting in L: $1, 1, 2, 2, 3, 3, 4, 4, \ldots$
  - ... when starting in R: $0, 1, 1, 2, 2, 3, 3, 4, \ldots$



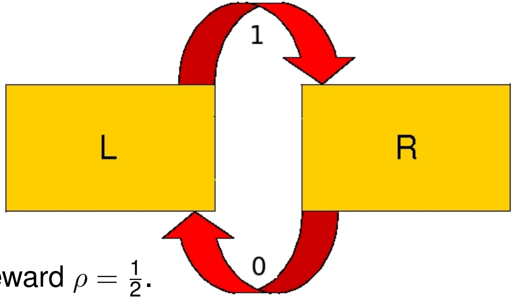
- Average reward $\rho = \frac{1}{2}$.
- Bias $\lambda(L) = \frac{1}{4}$, $\lambda(R) = -\frac{1}{4}$.
- **Interpretation:** Accumulated reward after $t = 1, 2, \ldots$ steps ...
  - ... when starting in L: $1, 1, 2, 2, 3, 3, 4, 4, \ldots$
  - ... when starting in R: $0, 1, 1, 2, 2, 3, 3, 4, \ldots$
  - accum. average reward: $\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \ldots$



- Average reward $\rho = \frac{1}{2}$.
- Bias $\lambda(L) = \frac{1}{4}$, $\lambda(R) = -\frac{1}{4}$.
- **Interpretation:** Accumulated reward after $t = 1, 2, \dots$ steps ...
  - ... when starting in L: $1, 1, 2, 2, 3, 3, 4, 4, \dots$
  - accum. average reward: $\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \dots$
  - $\rightsquigarrow$ diff. sequence for L: $\frac{1}{2}, 0, \frac{1}{2}, 0, \frac{1}{2}, 0, \frac{1}{2}, 0, \dots \to$ on avg. $\frac{1}{4}$



- Average reward $\rho = \frac{1}{2}$.
- Bias $\lambda(L) = \frac{1}{4}$, $\lambda(R) = -\frac{1}{4}$.
- **Interpretation:** Accumulated reward after $t = 1, 2, \ldots$ steps ...
  - ... when starting in R: $0, 1, 1, 2, 2, 3, 3, 4, \ldots$
  - accum. average reward: $\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, \ldots$
  - $\rightsquigarrow$ diff. sequence for R: $-\frac{1}{2}, 0, -\frac{1}{2}, 0, -\frac{1}{2}, 0, \ldots \to$ on avg. $-\frac{1}{4}$



- Average reward $\rho = \frac{1}{2}$.
- Bias $\lambda(L) = \frac{1}{4}$, $\lambda(R) = -\frac{1}{4}$.
- **Interpretation:** Accumulated reward after $t = 1, 2, \ldots$ steps ...
  - ... when starting in L: $1, 1, 2, 2, 3, 3, 4, 4, \ldots$
  - ... when starting in R: $0, 1, 1, 2, 2, 3, 3, 4, \ldots$
  - $\rightsquigarrow$ difference sequence: $1, 0, 1, 0, 1, 0, 1, 0, \ldots$
  - average difference $= \frac{1}{2} = \lambda(L) - \lambda(R)$ "bias span Λ"

The Bias and the Diameter

Definition

The *diameter* of an MDP is the maximal expected time it takes to reach one state from any other state (using an appropriate policy).

- Intuitively, the bias indicates how much you gain/lose in accumulated rewards w.r.t. average reward when starting in state s.
- If the rewards are bounded in $[0, 1]$, the bias span Λ of the optimal policy is bounded by the diameter.

The Bias and the Diameter

- Intuitively, the bias indicates how much you gain/lose in accumulated rewards w.r.t. average reward when starting in state *s*.
- If the rewards are bounded in $[0,1]$, the bias span Λ of the optimal policy is bounded by the diameter.
- For UCRL one can show:

The bias $\tilde{\Lambda}_{\tilde{\pi}}$ of the optimal policy $\tilde{\pi}$ in the optimistic model $\tilde{\mathcal{M}}$ is bounded by the diameter *D* in the true MDP $\mathcal{M}$ w.h.p.

Regret of UCRL

Looking at the bound again, now knowing about the bias ...

Theorem (Jaksch et al., 2010)

In an MDP with S states, A actions, and diameter D with probability of at least $1-\delta$ the regret of UCRL after T steps is bounded by

$$34 \cdot DS\sqrt{AT \log\left(\frac{T}{\delta}\right)}.$$

The REGAL algorithm (Bartlett&Tewari 2009)

... there is an obvious question:

Shouldn't it be the bias span Λ instead of the the diameter?

The REGAL algorithm (Bartlett&Tewari 2009)

... there is an obvious question:

Shouldn't it be the bias span Λ instead of the the diameter?

Yeah, but how do you relate the optimistic bias $\tilde{\Lambda}_{\tilde{\pi}}$ to the real one?

The REGAL algorithm (Bartlett&Tewari 2009)

... there is an obvious question:

Shouldn't it be the bias span Λ instead of the the diameter?

Yeah, but how do you relate the optimistic bias $\tilde{\Lambda}_{\tilde{\pi}}$ to the real one?

Well, you can do the following:

- Look for optimistic model with bias bounded by the real bias.
- If you don't know the bias, try to guess it.

The REGAL algorithm (Bartlett&Tewari 2009)

REGAL.C (Bartlett&Tewari 2009)

For episodes $k = 1, 2, \ldots$ do:

1. Maintain UCB-like confidence intervals for rewards and transition probabilities to define set of plausible MDPs $\mathbb{M}$.
2. Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \bar{\mathbb{M}}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \bar{\mathbb{M}}} \rho(\mathcal{M}, \pi),$$

   where $\bar{\mathbb{M}} := \{\mathcal{M} \in \mathbb{M} \mid \Lambda(\mathcal{M}) \leq \Lambda\}$ is the set of plausible MDPs with bias bounded by the true bias Λ.
3. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled.

Regret of REGAL

Theorem (Bartlett&Tewari 2009)

In an MDP with S states, A actions, and bias span Λ with probability of at least $1-\delta$ the regret of REGAL.C after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(\frac{AT}{\delta}\right)}.$$

Regret of REGAL

Theorem (Bartlett&Tewari 2009)

In an MDP with S states, A actions, and bias span Λ with probability of at least $1-\delta$ the regret of REGAL.C after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(\tfrac{AT}{\delta}\right)}.$$

If Λ is not known, one can use the doubling trick to guess it.

$\rightsquigarrow$ same bound with large additive constant (exponential in Λ)

The REGAL algorithm (Bartlett&Tewari 2009)

REGAL (Bartlett&Tewari 2009)

For episodes $k = 1, 2, \ldots$ do:

1. Maintain UCB-like confidence intervals for rewards and transition probabilities to define set of plausible MDPs $\mathbb{M}$.
2. Calculate optimal **regularized** policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \left\{ \rho(\mathcal{M}, \pi) - c_k \cdot \Lambda(\mathcal{M}) \right\}.$$

3. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled.

Regret of REGAL

Theorem (Bartlett&Tewari 2009)

In an MDP with S states, A actions, and bias span Λ with probability of at least $1 - \delta$ the regret of REGAL for a suitable choice of c_k after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(\frac{AT}{\delta}\right)}.$$

Regret of REGAL

Theorem (Bartlett&Tewari 2009)

In an MDP with S states, A actions, and bias span Λ with probability of at least $1-\delta$ the regret of REGAL for a suitable choice of c_k after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(\frac{AT}{\delta}\right)}.$$

Unfortunately, the suitable choice of c_k depends on the length of episode k, which is not known in advance.

Regret of REGAL

Theorem (Bartlett&Tewari 2009)

In an MDP with S states, A actions, and bias span Λ with probability of at least $1-\delta$ the regret of REGAL for a suitable choice of c_k after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(\tfrac{AT}{\delta}\right)}.$$

Unfortunately, the suitable choice of c_k depends on the length of episode k, which is not known in advance.

$\rightsquigarrow$Using doubling to guess the episode length gives worse dependence $S^{3/2}$ on number of states in regret bound.

Outline

1. Introduction
2. Regret Dependence on Number of States
3. The Diameter
4. Optimal Policies with Bounded Bias
5. Digression on Continuous State MDPs
6. Recent Approaches

How to find the optimistic MDP (UCRL)

Choose optimistic MDP $\tilde{\mathcal{M}} \in \mathbb{M}$ and optimal policy $\tilde{\pi}$ such that

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi).$$

- Set rewards $\tilde{r}$ to the upper confidence bounds.
- For the transition probabilities $\tilde{p}$ one can use an extension of value iteration. That is, for all states s set

$$\begin{aligned} u_0(s) &:= 0, \quad \text{and} \\ u_{n+1}(s) &:= \max_a \left\{ \tilde{r}(s,a) + \max_{p \in \mathcal{P}(s,a)} \left\{ \sum_{s'} p(s') u_n(s') \right\} \right\}, \end{aligned}$$

where $\mathcal{P}(s,a)$ is the set of all plausible transitions from s, a.

How to find the optimal average reward with bounded bias (REGAL.C)

Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \bar{\mathbb{M}}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \bar{\mathbb{M}}} \rho(\mathcal{M}, \pi),$$

where $\bar{\mathbb{M}} := \{\mathcal{M} \in \mathbb{M} \mid \Lambda(\mathcal{M}) \leq \Lambda\}$ is the set of plausible MDPs with bias bounded by the true bias Λ.

How to find the optimal average reward with bounded bias (REGAL.C)

Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \bar{\mathbb{M}}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \bar{\mathbb{M}}} \rho(\mathcal{M}, \pi),$$

where $\bar{\mathbb{M}} := \{\mathcal{M} \in \mathbb{M} \mid \Lambda(\mathcal{M}) \leq \Lambda\}$ is the set of plausible MDPs with bias bounded by the true bias Λ.

This is different from constrained MDP problems usually found in the literature (e.g., E. Altman: *Constrained MDPs*, 1999).

How to find the optimal average reward with bounded bias (REGAL.C)

Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \bar{\mathbb{M}}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \bar{\mathbb{M}}} \rho(\mathcal{M}, \pi),$$

where $\bar{\mathbb{M}} := \{\mathcal{M} \in \mathbb{M} \,|\, \Lambda(\mathcal{M}) \leq \Lambda\}$ is the set of plausible MDPs with bias bounded by the true bias Λ.

This is different from constrained MDP problems usually found in the literature (e.g., E. Altman: *Constrained MDPs*, 1999).

OPEN PROBLEM 3A

Find efficient algorithm to compute optimal average reward under bias constraint $\Lambda \leq C$.

How to find the optimal regularized average reward (REGAL)

Calculate optimal **regularized** policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi) - c_k \cdot \Lambda(\mathcal{M}).$$

How to find the optimal regularized average reward (REGAL)

Calculate optimal **regularized** policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi) - c_k \cdot \Lambda(\mathcal{M}).$$

OPEN PROBLEM 3B

Find efficient algorithm to compute regularized optimal average reward

$$\rho - c \cdot \Lambda.$$

Value Iteration for Average Reward MDPs

Consider finite MDP with finite diameter (communicating).

Value iteration for Average Reward MDPs

For all states s set

$$\begin{aligned} v_0(s) &:= 0, \quad \text{and} \\ v_{n+1}(s) &:= \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}, \end{aligned}$$

Value Iteration for Average Reward MDPs

Consider finite MDP with finite diameter (communicating).

Value iteration for Average Reward MDPs

For all states s set

$$\begin{aligned} v_0(s) &:= 0, \quad \text{and} \\ v_{n+1}(s) &:= \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}, \end{aligned}$$

Convergence of VI:

$$\lim_{n\to\infty} (v_{n+1} - v_n) = \rho^* \cdot \mathbf{1}$$

Value Iteration for Average Reward MDPs

Consider finite MDP with finite diameter (communicating).

Value iteration for Average Reward MDPs

For all states s set

$$\begin{aligned} v_0(s) &:= 0, \quad \text{and} \\ v_{n+1}(s) &:= \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}, \end{aligned}$$

Convergence of VI:

$$\lim_{n\to\infty} (v_{n+1} - v_n) = \rho^* \cdot \mathbf{1}$$

If $\text{span}(v_{n+1} - v_n) < \varepsilon$, then found policy is ε-optimal.

Modified Value Iteration for Bias Constrained MDPs

- Value vector v_n gives maximal n-step reward for each initial state
- span(v_n) converges to bias Λ
- Taking into account error span($v_{n+1} - v_n$),
  one can eliminate actions that violate bias constraint.

Modified Value Iteration for Bias Constrained MDPs

- Value vector v_n gives maximal n-step reward for each initial state
- $\mathrm{span}(v_n)$ converges to bias Λ
- Taking into account error $\mathrm{span}(v_{n+1} - v_n)$, one can eliminate actions that violate bias constraint.

Modified Value Iteration for Bias Constrained MDPs

Input: Bias constraint C.

- Set $v_0(s) := 0$ for all states s.
- For $n = 1, 2, \ldots$:
  - Set $v_{n+1}(s) := \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}$.
  - If $\mathrm{span}(v_n) - \mathrm{span}(v_{n+1} - v_n) > C$:
    - Eliminate maximizing action in state with maximal v_n.
    - Recompute v_n.

Modified Value Iteration for Bias Constrained MDPs

Input: Bias constraint C.

- Set $v_0(s) := 0$ for all states s.
- For $n = 1, 2, \ldots$:
  - Set $v_{n+1}(s) := \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}$.
  - If $\mathrm{span}(v_n) - \mathrm{span}(v_{n+1} - v_n) > C$:
    - Eliminate maximizing action in state with maximal v_n.
    - Recompute v_n.

- Convergence proof missing

Modified Value Iteration for Bias Constrained MDPs

Modified Value Iteration for Bias Constrained MDPs

Input: Bias constraint C.

- Set $v_0(s) := 0$ for all states s.
- For $n = 1, 2, \ldots$:
  - Set $v_{n+1}(s) := \max_a \left\{ r(s,a) \; + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}$.
  - If $\mathrm{span}(v_n) - \mathrm{span}(v_{n+1} - v_n) > C$:
    - Eliminate maximizing action in state with maximal v_n.
    - Recompute v_n.

- Convergence proof missing
- Assumes finite action space, whereas in REGAL action sets are compact (coming from confidence intervals).

Modified Value Iteration for Bias Constrained MDPs

Input: Bias constraint C.

- Set $v_0(s) := 0$ for all states s.
- For $n = 1, 2, \ldots$:
  - Set $v_{n+1}(s) := \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) \cdot v_n(s') \right\}$.
  - If $\mathrm{span}(v_n) - \mathrm{span}(v_{n+1} - v_n) > C$:
    - Eliminate maximizing action in state with maximal v_n.
    - Recompute v_n.

- Convergence proof missing
- Assumes finite action space, whereas in REGAL action sets are compact (coming from confidence intervals).
  $\rightsquigarrow$ discretization is possible

Replacing Diameter With Span in Regret Bounds

OPEN PROBLEM 3C

Find efficient RL algorithm with regret depending on bias instead of diameter.

Outline

1. Introduction
2. Regret Dependence on Number of States
3. The Diameter
4. Optimal Policies with Bounded Bias
5. Digression on Continuous State MDPs
6. Recent Approaches

Continuous State MDPs

Consider MDP with continuous state space where rewards and transition probabilities are Lipschitz or Hölder, that is,

Assumption 1

There are $L, \alpha > 0$ such that for any two states s, s' and all actions a,

$$\begin{aligned} |r(s,a) - r(s',a)| &\leq L|s-s'|^{\alpha}, \\ \left\| p(\cdot|s,a) - p(\cdot|s',a) \right\|_1 &\leq L|s-s'|^{\alpha}. \end{aligned}$$

Then close states behave similarly and discretization is possible.

Continuous State MDPs

Consider MDP with continuous state space where rewards and transition probabilities are Lipschitz or Hölder, that is,

Assumption 1

There are $L, \alpha > 0$ such that for any two states s, s' and all actions a,

$$\begin{aligned} |r(s,a) - r(s',a)| &\leq L|s-s'|^{\alpha}, \\ \big\|p(\cdot|s,a) - p(\cdot|s',a)\big\|_1 &\leq L|s-s'|^{\alpha}. \end{aligned}$$

Discretization

For simplicity, assume $\mathcal{S} = [0, 1]$ and $\alpha = 1$.

- Then consider discretization
  $I_1 = \left[0, \frac{1}{n}\right], I_2 = \left(\frac{1}{n}, \frac{2}{n}\right], \ldots, I_n = \left(\frac{n-1}{n}, 1\right]$.
- States within each interval have (by Lipschitz assumption) close rewards and transition probabilities.

Problems

1. Original state space infinite.
2. $\rightsquigarrow$ The diameter is usually infinite.
3. However, the bias under Lipschitz conditions usually is finite!

$\rightsquigarrow$ Hence, it would be helpful if we had regret bounds with the diameter replaced with the bias!

REGAL-approach to continuous state MDPs

UCCRL (Ortner&Ryabko, 2012, Lakshmanan et al., 2015)

Input: Upper bound Λ on bias span of optimal policy,
Hölder parameters L, α, discretization parameter n

1. Discretize $[0, 1]$ into n intervals $I_1, \ldots, I_n$ of equal size.
2. For episodes $k = 1, 2, \ldots$ do:
   1. Maintain UCB-like confidence intervals $(+\varepsilon := Ln^{-\alpha})$ for rewards and transition probabilities of each interval I_j.
   2. Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$ under constraint that bias span of $\tilde{\pi}$ is upper bounded by H.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}: H(M) \leq H} \rho(\mathcal{M}, \pi).$$

   3. Execute $\tilde{\pi}$ until the visits in some interval-action pair have doubled.

Regret Bounds for UCCRL

Theorem (Ortner&Rybako, NIPS 2012)

Under Assumption 1, with probability $1-\delta$ the regret of UCCRL after T steps is bounded by

$$const \cdot \Lambda n \sqrt{AT \log\left(\frac{T}{\delta}\right)} + const \cdot \frac{\Lambda L T}{n}.$$

Regret Bounds for UCCRL

Theorem (Ortner&Rybako, NIPS 2012)

Under Assumption 1, with probability $1-\delta$ the regret of UCCRL after T steps is bounded by

$$const \cdot \Lambda n \sqrt{AT \log\left(\frac{T}{\delta}\right)} + const \cdot \frac{\Lambda LT}{n}.$$

Choosing $n = T^{\frac{1}{4}}$ gives regret upper bounded by

$$const \cdot \Lambda L T^{\frac{3}{4}} \sqrt{A \log\left(\frac{T}{\delta}\right)}.$$

If Λ is unknown, use e.g. $\log T$ to guess it.

Improved Estimation of Transitions

Assumption 2

The transition functions $p(\cdot|s, a)$ are κ-times smoothly differentiable. That is, there are L, $\alpha > 0$ such that for any state s and all actions a,

$$\left|p^{(\kappa)}(s'|s, a) - p^{(\kappa)}(s''|s, a)\right| \leq L|s' - s''|^{\alpha}.$$

If also Assumption 2 holds, we can compute $\hat{p}(\cdot|s, a)$ using a **kernel density estimator**, assuming that all samples $X_1, \ldots, X_N$ come from the same distribution:

$$\hat{p}_N(x|s, a) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right).$$

One can show respective concentration inequalities for kernel estimation.

Improved Regret Bound

Theorem (Lakshmanan et al., ICML 2015)

Consider an MDP with state space $[0, 1]$, A actions, rewards and transition probabilities satisfying Assumptions 1 and 2, and bias span Λ.

Setting $n = T^{\frac{\beta}{3\beta+2}}$, with probability $1 - \delta$ the regret of UCCRL using a suitable kernel density estimator after T steps is upper bounded by

$$c \cdot \Lambda(C_0 L + C_1')\sqrt{14A \log\left(\frac{2AT^2}{\delta}\right)}\, T^{\frac{2\beta+2}{3\beta+2}}$$

for $\beta := \kappa + 1$ and an independent constant c.

Improved Regret Bound

Theorem (Lakshmanan et al., ICML 2015)

Consider an MDP with state space $[0,1]$, A actions, rewards and transition probabilities satisfying Assumptions 1 and 2, and bias span Λ.

Setting $n = T^{\frac{\beta}{3\beta+2}}$, with probability $1-\delta$ the regret of UCCRL using a suitable kernel density estimator after T steps is upper bounded by

$$c \cdot \Lambda(C_0 L + C_1')\sqrt{14A\log\left(\frac{2AT^2}{\delta}\right)}\, T^{\frac{2\beta+2}{3\beta+2}}$$

for $\beta := \kappa + 1$ and an independent constant c.

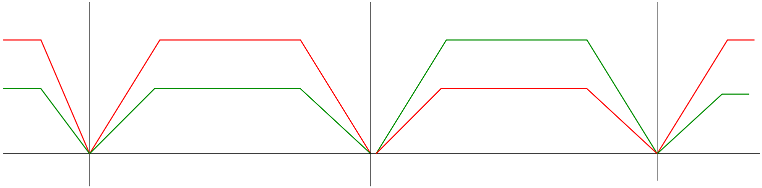
This is an improvement if $\kappa > 1$.

The bound approaches $\tilde{O}(T^{\frac{2}{3}})$ for $\kappa \to \infty$.

Lower Bound on Regret

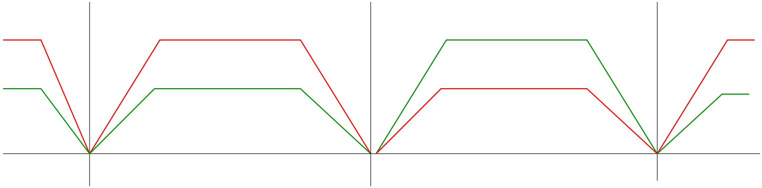
Example for lower bound
(pointed out by P. Auer, special case of bandit problem with side information, cf. Perchet&Rigollet 2013, Audibert&Tsybakov 2007):

- Partition $[0, 1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.



Example for lower bound
(pointed out by P. Auer, special case of bandit problem with side information, cf. Perchet&Rigollet 2013, Audibert&Tsybakov 2007):

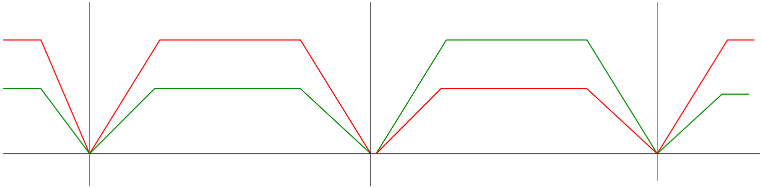
- Partition $[0, 1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.
- All transition probability distributions are uniform.



Lower Bound on Regret

Example for lower bound
(pointed out by P. Auer, special case of bandit problem with side information, cf. Perchet&Rigollet 2013, Audibert&Tsybakov 2007):

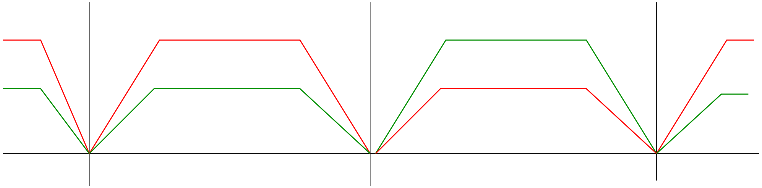
- Partition $[0,1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.
- All transition probability distributions are uniform.
- Rewards are piecewise linear, constant $\frac{1}{2}$ in the middle of each interval, and 0 on the boundary.



Lower Bound on Regret

Example for lower bound
(pointed out by P. Auer, special case of bandit problem with side information, cf. Perchet&Rigollet 2013, Audibert&Tsybakov 2007):

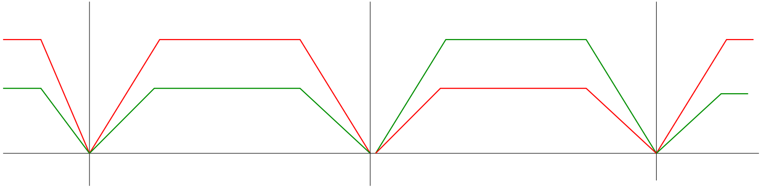
- Partition $[0, 1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.
- All transition probability distributions are uniform.
- Rewards are piecewise linear, constant $\frac{1}{2}$ in the middle of each interval, and 0 on the boundary.
- In each interval, there is one particular action with reward $\frac{1}{2} + T^{-\frac{1}{3}}$ in the middle.



Lower Bound on Regret

Example for lower bound

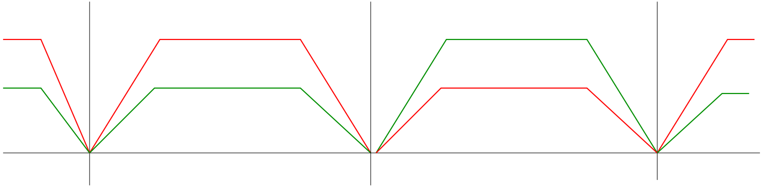
- Partition $[0,1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.
- All transition probability distributions are uniform.
- Rewards are piecewise linear, constant $\frac{1}{2}$ in the middle of each interval, and 0 on the boundary.
- In each interval, there is one particular action with reward $\frac{1}{2}+T^{-\frac{1}{3}}$ in the middle.



$\rightsquigarrow$ Learner needs $\Omega\big(AT^{\frac{2}{3}}\log\frac{1}{\delta}\big)$ examples in each interval.

Example for lower bound

- Partition $[0,1]$ in $T^{\frac{1}{3}}$ intervals of equal size $T^{-\frac{1}{3}}$.
- All transition probability distributions are uniform.
- Rewards are piecewise linear, constant $\frac{1}{2}$ in the middle of each interval, and 0 on the boundary.
- In each interval, there is one particular action with reward $\frac{1}{2} + T^{-\frac{1}{3}}$ in the middle.



$\rightsquigarrow$ Regret is $\Omega\left(AT^{\frac{2}{3}} \log \frac{1}{\delta}\right)$.

Regret in Continuous State Space

Summing up:

- **Nice:**
  have basically best possible bounds on regret wrt T for RL in Lipschitz continuous MDPs

- However, our algorithm is not efficiently computable.

Outline

1. Introduction
2. Regret Dependence on Number of States
3. The Diameter
4. Optimal Policies with Bounded Bias
5. Digression on Continuous State MDPs
6. Recent Approaches

Replacing Diameter With Span in Regret Bounds

OPEN PROBLEM 3C

Find efficient RL algorithm with regret depending on bias instead of diameter.

The Environmental Norm (Maillard et al., 2014)

Maillard, Mann & Mannor (2014) introduce the environmental norm of an MDP, which is always upper bounded by Λ.

For a modification of UCRL with different confidence intervals it is shown:

Theorem (Maillard et al., 2014)

In an MDP with S states, A actions, diameter D, and bias span Λ with probability of at least $1-\delta$ the regret after T steps is bounded by

$$c \cdot \frac{\Lambda D S \sqrt{AT}}{\sqrt{p_0}} \log\left(\tfrac{AST}{\delta}\right),$$

where p_0 is the smallest nonzero transition probability.

The Environmental Norm (Maillard et al., 2014)

Theorem (Maillard et al., 2014)

In an MDP with S states, A actions, diameter D, and bias span Λ with probability of at least $1-\delta$ the regret after T steps is bounded by

$$c \cdot \frac{\Lambda D S \sqrt{AT}}{\sqrt{p_0}} \log\left(\tfrac{AST}{\delta}\right),$$

where p_0 is the smallest nonzero transition probability.

Conjecture: bound is $O\left(\frac{\Lambda\sqrt{SAT}}{\sqrt{p_0}} \log\left(\frac{AST}{\delta}\right)\right)$

Some Recent Research (with P. Auer & C.-K. Chiang)

Optimistic model-free counterpart to UCRL:

Optimistic Q-Learning

Input: Λ
Initialize: Set $\tilde{V} := \mathbf{0}$.
For episodes $k = 1, 2, \ldots$ do:

1. Maintain optimistic Q-function $\tilde{Q}$, i.e., given set $\mathcal{O}_{k-1}$ of of observations of previous episode set

$$\tilde{Q}(s,a) := \tfrac{1}{N(s,a)} \sum_{(s,a,r,s') \in \mathcal{O}_{k-1}} \left(r + \tilde{V}(s') + \Lambda \sqrt{\tfrac{S \log(AT^2)}{N(s,a)}} \right)$$

2. Choose $\tilde{\pi}(s) := \arg\max_a \tilde{Q}(s,a)$.
3. Set $\tilde{V}(s) := \arg\max_a \tilde{Q}(s,a)$.
4. If $\tilde{V}(s) - \min_{s'} \tilde{V}(s') > \Lambda$, reset $\tilde{V}(s) := \min_{s'} \tilde{V}(s') + \Lambda$.
5. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled.

Some Recent Research (with P. Auer & C.-K. Chiang)

Idea: Cut off optimistic value vector so that

- vector is still optimistic (higher than true value)
- satisfies bias constraint

Some Recent Research (with P. Auer & C.-K. Chiang)

Idea: Cut off optimistic value vector so that

- vector is still optimistic (higher than true value)
- satisfies bias constraint

Theorem

In an MDP with S states, A actions, and bias span Λ the expected regret after T steps is bounded by

$$c \cdot \Lambda S \sqrt{AT \log\left(AT^2\right)}.$$

If Λ is not known:

- Could use $\log T$ to guess Λ.

If Λ is not known:

- Could use $\log T$ to guess Λ.

  $\rightsquigarrow$ gives large additive constant

If Λ is not known:

- Could use $\log T$ to guess Λ.
  $\rightsquigarrow$ gives large additive constant

- Use doubling trick to guess Λ:

If Λ is not known:

- Could use log T to guess Λ.
  $\leadsto$ gives large additive constant

- Use doubling trick to guess Λ:
  - Start with guess $\tilde{\Lambda} = 1$.
  - Double $\tilde{\Lambda}$ whenever collected reward does not meet regret bound.

If Λ is not known:

- Could use $\log T$ to guess Λ.

  $\rightsquigarrow$ gives large additive constant

- Use doubling trick to guess Λ:
  - Start with guess $\tilde{\Lambda} = 1$.
  - Double $\tilde{\Lambda}$ whenever collected reward does not meet regret bound.

Problem: Our regret bound only holds in expectation.

$\rightsquigarrow$ need to use something like Chebyshev inequality

If Λ is not known:

- Could use $\log T$ to guess Λ.

  $\rightsquigarrow$ gives large additive constant

- Use doubling trick to guess Λ:
  - Start with guess $\tilde{\Lambda} = 1$.
  - Double $\tilde{\Lambda}$ whenever collected reward does not meet regret bound.

  Problem: Our regret bound only holds in expectation.

  $\rightsquigarrow$ need to use something like Chebyshev inequality

  $\rightsquigarrow$ obtain regret bound as before but with $T^{\frac{1}{2}+\varepsilon}$ instead of $T^{\frac{1}{2}}$

Big Open Problems

- Close gap in regret bounds wrt number of states and diameter.

Big Open Problems

- Close gap in regret bounds wrt number of states and diameter.
- Replace diameter by bias span in regret bounds for *efficiently computable* algorithm.

Big Open Problems

- Close gap in regret bounds wrt number of states and diameter.
- Replace diameter by bias span in regret bounds for *efficiently computable* algorithm.
- Find *efficiently computable* algorithm with sharp regret bounds for continuous state MDPs.

Big Open Problems

- Close gap in regret bounds wrt number of states and diameter.
- Replace diameter by bias span in regret bounds for *efficiently computable* algorithm.
- Find *efficiently computable* algorithm with sharp regret bounds for continuous state MDPs.
- Don't neglect things that are important beyond the RL community:

Big Open Problems

- Close gap in regret bounds wrt number of states and diameter.
- Replace diameter by bias span in regret bounds for *efficiently computable* algorithm.
- Find *efficiently computable* algorithm with sharp regret bounds for continuous state MDPs.
- Don't neglect things that are important beyond the RL community:

Create Heavy Metal Style for Beamer