

ABSTRACT

Any reinforcement learning algorithm that applies to all MDPs will suffer $\Omega(\sqrt{SAT})$ regret on some MDP, where T is the elapsed time and S is the number of states and A is the number of actions. In many problems S and A are so huge that any regret bounds are totally impractical.

We show that, if the system is known to be a *factored* MDP, it is possible to achieve regret that scales polynomially in the number of *parameters* encoding the factored MDP, which may be exponentially smaller than S or A . We provide two algorithms that satisfy near-optimal regret bounds in this context: PSRL and UCRL-Factored.

PROBLEM FORMULATION

Learn to optimize a random finite horizon MDP M in repeated finite episodes of interaction.



Figure 1: classic reinforcement learning setting

- State space \mathcal{S} , action space \mathcal{A}
- Rewards $r_t \sim R^M(s_t, a_t)$
- Transitions $s_{t+1} \sim P^M(s_t, a_t)$
- Episode length τ , define $t_k := (k-1)\tau + 1$

For MDP M and policy μ , define a value function

$$V_{\mu,i}^M(s) := \mathbb{E}_{M,\mu} \left[\sum_{j=i}^{\tau} \bar{R}^M(s_j, a_j) \mid s_i = s \right],$$

Define the regret in episode k using μ_k on M^*

$$\Delta_k := \sum_S \rho(s) \left(\underbrace{V_{\mu^*,1}^{M^*}(s)}_{\text{optimal value}} - \underbrace{V_{\mu_k,1}^{M^*}(s)}_{\text{actual value}} \right)$$

And finally $\text{Regret}(T, \pi, M^*) := \sum_{k=1}^{\lceil T/\tau \rceil} \Delta_k$.

Naive exploration such as Boltzman or ϵ -greedy can lead to exponential regret. Good performance requires balancing **exploration vs exploitation**. Carefully designed optimism or posterior sampling can learn quickly in factored MDPs.

FACTORED MDPs

MDP with conditional independence structure.

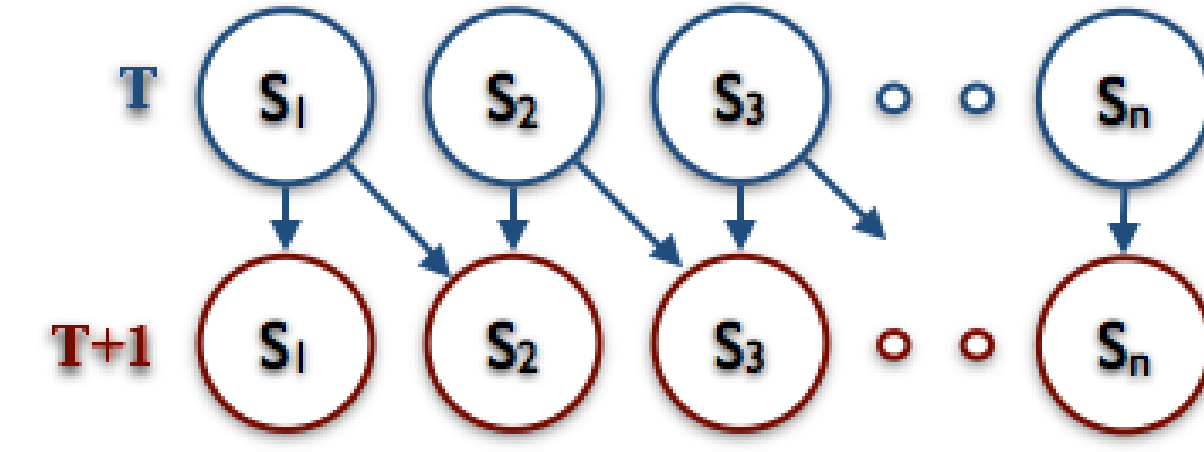


Figure 2: a graphical model for transitions.

Definition 1 (Scope operation for factored sets). For any $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ and $Z \subseteq \{1, 2, \dots, n\}$ define $\mathcal{X}[Z] := \bigotimes_{i \in Z} \mathcal{X}_i$ and elements $x[Z] \in \mathcal{X}[Z]$.

Definition 2 (Factored reward functions). The reward function r is factored over $\mathcal{S} \times \mathcal{A} = \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ with scopes $Z_1, \dots, Z_l \iff$

$$\mathbb{E}[r(x)] = \sum_{i=1}^l \mathbb{E}[r_i(x[Z_i])] \text{ and each } r_i \text{ observed}$$

Definition 3 (Factored transition functions). The transition function P is factored over $\mathcal{S} \times \mathcal{A} = \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ and $\mathcal{S} = \mathcal{S}_1 \times \dots \times \mathcal{S}_m$ with scopes $Z_1, \dots, Z_m \iff$

$$P(s|x) = \prod_{i=1}^m P_i(s[i] \mid x[Z_i])$$

MAIN RESULTS

For M^* factored with known graphical structure as above then for PSRL and UCRL-Factored

$$\text{Regret}(T, M^*) = \tilde{O} \left(\Xi \sum_{j=1}^m \sqrt{|\mathcal{X}[Z_j^P]| |\mathcal{S}_j| T} \right).$$

Here Ξ is a measure of MDP connectedness for each algorithm, expected span $\mathbb{E}[\Psi]$ for PSRL and diameter D for UCRL-Factored.

PSRL's bounds are tighter since $\Psi(M) \leq D(M)$ and may be exponentially smaller. However, UCRL-Factored holds with high probability for any M^* not just in expectation over the prior.

Key point: For m independent components with S states and A actions $= \tilde{O}(mS\sqrt{AT})$ and close to lower bounds of $\underbrace{m\sqrt{SAT}}_{\text{factored MDP}} \ll \underbrace{\sqrt{(SA)^m T}}_{\text{general MDP}}$.

POSTERIOR SAMPLING

For each episode k :

1. Sample an MDP from the posterior distribution for the true MDP: $M_k \sim \phi(\cdot | H_t)$.
2. Use policy $\mu_k \in \arg \max_{\mu, M} V_{\mu}^{M_k}$.

Proof sketch:

$$\begin{aligned} \Delta_k &= V_{*,1}^*(s) - V_{k,1}^*(s) \\ &= \underbrace{(V_{k,1}^k(s) - V_{k,1}^*(s))}_{\text{Imagined - Actual}} + \underbrace{(V_{*,1}^*(s) - V_{k,1}^k(s))}_{\mathbb{E}[\cdot]=0} \end{aligned}$$

We can decompose this into Bellman error:

$$V_{k,1}^k - V_{k,1}^* = \sum_{i=1}^{\tau} \underbrace{(\mathcal{T}_{k,i}^k - \mathcal{T}_{k,i}^*)}_{B := \text{Bellman error}} V_{k,i+1}^k + \sum_{i=1}^{\tau} \underbrace{d_{t_k+1}}_{\mathbb{E}=0 \text{ martingale}}$$

We can now use the Hölder inequality to bound:

$$B \leq \sum_{i=1}^{\tau} \left\{ \underbrace{|\bar{R}^k - \bar{R}^*|}_{\text{reward error}} + \frac{1}{2} \underbrace{\Psi_k}_{\text{MDP span}} \underbrace{\|P^k - P^*\|_1}_{\text{transition error}} \right\}$$

We conclude the proof by upper bounding these deviations by maximum possible within \mathcal{M}_k . Our analysis allows us to build tight \mathcal{M}_k that contain M^* with high probability.

KEY LEMMA

For any P, \tilde{P} factored transition functions we may bound their L1 distance by the sum of the differences of their factorizations:

$$\|P(x) - \tilde{P}(x)\|_1 \leq \sum_{i=1}^m \|P_i(x[Z_i]) - \tilde{P}_i(x[Z_i])\|_1$$

Proof sketch:

For any $\alpha_1, \alpha_2, \beta_1, \beta_2 \in [0, 1]$:

$$|\alpha_1 \alpha_2 - \beta_1 \beta_2| \leq \alpha_2 |\alpha_1 - \beta_1| + \beta_1 |\alpha_2 - \beta_2|.$$

Repeat this argument for desired result.

REFERENCES

Please see the full paper:
<http://arxiv.org/abs/1403.3741>

OPTIMSIM

For each episode k :

1. Form \mathcal{M}_k subset of MDPs M that are statistically plausible given the data.
2. Use policy $\mu_k \in \arg \max_{\mu} \left\{ \max_{M \in \mathcal{M}_k} V_{\mu}^M(s) \right\}$.

Proof sketch:

$$\begin{aligned} \Delta_k &= V_{*,1}^*(s) - V_{k,1}^*(s) \\ &= \underbrace{(V_{k,1}^k(s) - V_{k,1}^*(s))}_{\text{Imagined - Actual}} + \underbrace{(V_{*,1}^*(s) - V_{k,1}^k(s))}_{\leq 0 \text{ by optimism}} \end{aligned}$$

Then follow the analysis per posterior sampling.

EXAMPLE

Production line with 100 machines, each with 3 states and 3 actions. Each machine generates some revenue we want to maximize jointly.



Figure 3: automated production line

This MDP has state $s = (s_1, \dots, s_{100})$ and action $a = (a_1, \dots, a_{100})$. Here $S = A = 3^{100} \simeq 10^{50}$, so even a maximally efficient general-purpose learner would have regret $\Omega(\sqrt{SAT}) \simeq 10^{50} \sqrt{T}$.

If over a single timestep, each machine depends directly only upon its neighbours then this becomes a factored MDP. Now $|\mathcal{X}[Z_j^P]| \leq 3^3$ and $|\mathcal{S}_j| \leq 3$ for each machine j .

We exploit this graphical structure for exponentially smaller regret $\simeq 100\sqrt{3^3 \times 3 \times T} \simeq 10^3 \sqrt{T}$.

KEY TAKEAWAY

Our regret bounds scale with the number of parameters, not the number of states.

CONTACT INFORMATION

Web www.stanford.edu/~iosband
Email iosband@stanford.edu