Online regret in reinforcement learning

Peter Auer

University of Leoben, Austria

Tübingen, 31 July 2007

Undiscounted online regret

- I am interested in the difference (in rewards *during learning*) between an optimal policy and a reinforcement learner:

$$R_T = \sum_{t=0}^{T} r(s_t^*, a_t^*) - \sum_{t=0}^{T} r(s_t, a_t),$$

  where $s_0^*, s_1^*, \ldots$ is the sequence of states visited by an optimal policy choosing actions a_t^*, and $s_0, s_1, \ldots$ is the sequence of states visited by the learner choosing actions a_t.

Regret for discounted RL

- Naive:

$$\sum_{t=0}^{\infty} \gamma^t r(s_t^*, a_t^*) - \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) = O(1),$$

- Counting non-optimal actions (Kakade, 2003):

$$\#\{s_t : a_t \neq a_t^*\}$$

- Using the value function of the optimal policy (Strehl, Littman, 2005):

$$\sum_{t=0}^{T} V^*(s_t) - \sum_{t=0}^{T} \sum_{\tau=t}^{T} \gamma^{\tau-t} r_\tau = \sum_{t=0}^{T} V^*(s_t) - \sum_{\tau=0}^{T} \frac{1-\gamma^{\tau+1}}{1-\gamma} r_\tau$$

Undiscounted regret bounds: $\log T$ vs. $\sqrt{T}$ bounds

Consider the bandit problem first: $|S| = 1$.

- Logarithmic regret bounds for average rewards r_a, $a \in A$:

$$\mathbb{E}\left[R_T\right] = O\left(\sum_{a \neq a^*} \frac{\log T}{r_{a^*} - r_a}\right)$$

- Logarithmic regret depends on a gap between the best action and the other actions.
- A bound independent of the size of the gap:

$$\mathbb{E}\left[R_T\right] = O\left(\sqrt{|A| T \log |A|}\right)$$

- This bound holds even for varying r_a when regret is calculated in respect to the single best action.
- Both bounds are essentially tight.

Theoretical bounds for RL

- PAC-like bounds by Fiechter (1994)
  - Assumes a *reset* action and learns an ϵ-optimal policy of fixed length T in $poly(1/\epsilon, T)$ time.
- E^3 by Kearns and Singh (1998)
  - Learns an ϵ-optimal policy in $poly(1/\epsilon, |S|, |A|, T^{\epsilon}_{mix})$ steps.
- Analysis of Rmax by Kakade (2003)
  - Bounds the number of actions which are not ϵ-optimal:

$$\#\{t : a_t \neq a_t^{\epsilon}\} = \tilde{O}\left(|S|^2|A|(T^{\epsilon}_{mix}/\epsilon)^3\right)$$

- T^{ϵ}_{mix} is the number of steps such that for *any* policy π its actual average reward is ϵ-close to the expected average reward.

log T regret for irreducible MDPs

- Burnetas, Katehakis, 1997:

$$\mathbb{E}\left[R_T\right] = O\left(\frac{|S||A|(T^*_{hit})^2}{\Phi}\log T\right)$$

- $T^*_{hit} = \max_{s,s'} \mathbb{E}\left[T^{\pi^*}_{s,s'}\right]$
- $T^{\pi}_{s,s'} = \min\{t \geq 0 : s_t = s' | s_0 = s, \pi\}$
- Φ measures the distance (in expected future rewards) between the best and second best action at a state.
- But holds only for $T \geq |A|^{|S|}$.

- Related result by Strehl and Littman: The MBIE algorithm, ICML 2006.

log T regret for irreducible MDPs

- Auer, Ortner, 2006:

$$\mathbb{E}\left[R_T\right] = O\left(\frac{|S|^5|A|(T_{hit}^{\max})^3}{\Delta^2}\log T\right)$$

for any T.

- $T_{hit}^{\max} = \max_\pi \max_{s,s'} T_{s,s'}^\pi$
- $\Delta = \rho^* - \max\{\rho(\pi) : \rho(\pi) < \rho^*\}$
- $\rho(\pi) = \lim_T \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r_t|\pi\right]$

$T^{2/3}$ regret for irreducible MDPs

- Analogously the regret in respect to an ϵ-optimal policy can be bounded by

$$\mathbb{E}\left[R_T^{\epsilon}\right] = O\left(\frac{\log T}{\epsilon^2}\right).$$

- For $\epsilon = \sqrt[3]{(\log T)/T}$ this gives

$$\mathbb{E}\left[R_T\right] = O\left(\frac{\log T}{\epsilon^2}\right) + \epsilon T = O\left(T^{2/3}(\log T)^{1/3}\right).$$

The algorithm: Optimistic reinforcement learning

- Let $\mathcal{M}_t$ be the set of plausible MDPs in respect to experience $\mathcal{E}_t$.
- Choose an optimal policy for the most optimistic MDP in $\mathcal{M}_t$,

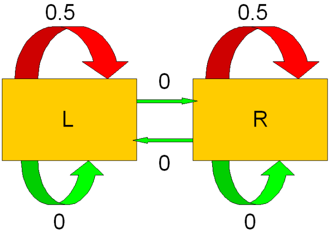
$$\tilde{\pi}_t := \arg\max_{\pi} \max_{M \in \mathcal{M}_t} \rho(\pi | M')$$

- For simplicity we assume that the rewards are known.
- Thus $M \in \mathcal{M}_t$ if for all s, a, s',

$$|p(s'|s,a) - \hat{p}_t(s'|s,a)|_1 \leq \sqrt{\frac{3 \log t}{N_t(s,a)}}.$$

- Hence $M^* \in \mathcal{M}_t$ with probability at least $1 - t^{-3}$.

We cannot change policy too often



- Both states give optimal reward 0.5 under action *red*.
- $\tilde{\pi}_t$ would switch states often to balance the number of visits (as the confidence bounds for the reward is larger for the less frequently visited state).
- Moving from one state to the other is costly, so that always following $\tilde{\pi}_t$ gives large (linear) regret.

We cannot change policy too often

- We change policy only if the numbers of uses of a state/action pair $N_t(s,a)$ has doubled. Let t_k be these times, when a new policy is calculated.

- Thus the number of policy changes in T steps is bounded by
$$|S||A| \log_2 \frac{T}{|S||A|}.$$

Accurate Estimates Imply Optimality of $\tilde{\pi}$

- Let $\tilde{\pi}_k$ be the optimal policy chosen at time $t = t_k$ for corresponding MDP $\tilde{M} \in \mathcal{M}_t$.
- If for all s, s',

$$|\tilde{p}(s'|s) - p^*(s'|s)| < \frac{\Delta}{2T_{hit}^{\max}|S|^2}$$

  then

$$\rho(\tilde{\pi}|M^*) > \rho(\tilde{\pi}|\tilde{M}) - \Delta \geq \rho(\pi^*|M^*) - \Delta.$$

- Thus $\tilde{\pi}$ is also optimal (since the distance between best and second best policy is Δ).

Analysis

- By executing a suboptimal policy for τ steps, we may lose reward at most τ.
- Making τ steps with a fixed policy, each state is visited (on average) $\lfloor \tau / T_{hit}^{\max} \rfloor$ times.
- For $\tilde{\pi}$ being non-optimal, there is (s, a, s'), $\tilde{\pi}(s) = a$, with

$$|\tilde{p}(s'|s,a) - p^*(s'|s,a)| \geq \frac{\Delta}{2 T_{hit}^{\max} |S|^2}$$

  and thus

$$N_t(s,a) \leq \frac{12 (T_{hit}^{\max})^2 |S|^4 \log t}{\Delta^2}.$$

- Hence the number of non-optimal steps is at most

$$\frac{24 (T_{hit}^{\max})^3 |S|^5 |A| \log T}{\Delta^2} + T_{hit}^{\max} |S||A| \log_2 \frac{T}{|S||A|}.$$

Alternative analysis

- We want a bound in terms of

$$T_{hit}^{\min} = \max_{s,s'} \min_{\pi} T_{s,s'}^{\pi}.$$

- We use the *bias equation*:
  Let P be the transition matrix of policy π and $\boldsymbol{r}$ be the vector of rewards. Then

$$\boldsymbol{\lambda} = \rho \boldsymbol{e} - \boldsymbol{r} + P\boldsymbol{\lambda}$$

  for some bias vector $\boldsymbol{\lambda}$ iff ρ is the average reward of π.
- Intuition about $\lambda_s - \lambda_{s'}$: It is the difference in total cumulative reward between starting in state s' and state s.

Bounding the bias

- For any MDP there is an optimal policy with

  $$\boldsymbol{\lambda} = \rho \boldsymbol{e} - \boldsymbol{r} + P\boldsymbol{\lambda}$$

  and

  $$\lambda_s - \lambda_{s'} \leq T_{hit}^{\min}$$

  for any states s, s'.
- Intuition: If $\lambda_s - \lambda_{s'} > T_{hit}^{\min}$ then we could modify the optimal policy to quickly move from s to s'. The number of steps for this is bounded by $T_{hit}^{\min}$. Thus s would be at most $T_{hit}^{\min}$ worse than s'.
- We choose $\min \lambda_s = 0$ such that

  $$0 \leq \lambda_s \leq T_{hit}^{\min}.$$

A general bound on the regret

- Let
$$\boldsymbol{\lambda} = \rho \boldsymbol{e} - \boldsymbol{r} + P\boldsymbol{\lambda}, \quad 0 \le \lambda_s \le T_{hit}^{\min}$$
be the bias equation for an optimal policy π^*.
- Then the expected regret can be bounded by
$$\begin{aligned}\mathbb{E}\left[R_T\right] \quad \le \quad & \sum_{s,a} \mathbb{E}\left[N_T(s,a)\right] \\ & [r(s,\pi^*(s)) - r(s,a) \\ & -(p(\cdot|s,\pi^*(a)) - p(\cdot|s,a))\boldsymbol{\lambda}] + O(1)\end{aligned}$$
- Assuming that $r(s,a) = r(s,a')$ for all a, a', we get
$$\mathbb{E}\left[R_T\right] \le \sum_{s,a} \mathbb{E}\left[N_T(s,a)\right] 2T_{hit}^{\min} ||(p(\cdot|s,\pi^*(s)) - p(\cdot|s,a))||_1$$

Bounding the regret of our algorithm (1)

- Let $\tilde{\pi} = \tilde{\pi}_k$ be the optimal policy chosen by our algorithm at time t_k for a plausible MDP $\tilde{M}$.
- Let $\tilde{\rho} = \rho(\tilde{\pi}|\tilde{M})$, $\tilde{\rho}^* = \rho(\tilde{\pi}|M^*)$, and $\rho^* = \rho(\pi^*|M^*)$ be the average rewards of $\tilde{\pi}$ in the MDP $\tilde{M}$, of $\tilde{\pi}$ in the true MDP M^*, and the optimal average reward in the true MDP.
- Then

$$
\begin{aligned}
&(t_{k+1} - t_k)\rho^* - \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t)|M^*, \tilde{\pi}\right] \\
&= \quad (t_{k+1} - t_k)\rho^* - \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t)|\tilde{M}, \tilde{\pi}\right] \\
&\quad +\mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t)|\tilde{M}, \tilde{\pi}\right] - \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t)|M^*, \tilde{\pi}\right].
\end{aligned}
$$

Bounding the regret of our algorithm (2)

It can be shown that

$$\begin{aligned} T\rho^* - \mathbb{E}\left[\sum_{k=0}^{K} \sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t) | \tilde{M}, \tilde{\pi}\right] &\leq O\left((K+1) T_{hit}^{\min}\right) \\ &= O\left(|S||A| T_{hit}^{\min} \log T\right) \end{aligned}$$

with $t_0 = 0$ and $t_{K+1} = T$.

Bounding the regret of our algorithm (3)

Now we construct an MDP M' with actions a and $\tilde{a}$, $a \in A$, such that a represents an action in the original MDP M^* and $\tilde{a}$ represents the corresponding action in $\tilde{M}^*$. Thus

$$
\begin{aligned}
r(s, \tilde{a}|M') &= r(s, a|M') = r(s, a|M^*) = r(s, a|\tilde{M}^*) \\
p(\cdot|M', s, a) &= p(\cdot|M^*, s, a) \\
p(\cdot|M', s, \tilde{a}) &= p(\cdot|\tilde{M}^*, s, a).
\end{aligned}
$$

Then π' with $\pi'(s) = \tilde{a}$ for $\tilde{\pi}(s) = a$ is an optimal policy for M' and π^* is a suboptimal policy in M'.

Bounding the regret of our algorithm (4)

Thus we can use the general regret bound and get

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t) | \tilde{M}, \tilde{\pi}\right] - \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} r(s_t, a_t) | M^*, \tilde{\pi}\right] \\
\leq \quad & \sum_{s,a} \mathbb{E}\left[N_{k+1}(s,a) - N_k(s,a)\right] \\
& \cdot 2 T_{hit}^{\min} ||(p(\cdot|\tilde{M}^*, s, \tilde{\pi}(s)) - p(\cdot|M^*, s, \tilde{\pi}(s))||_1 \\
\leq \quad & 2 T_{hit}^{\min} |S| \sqrt{3 \log T} \sum_s \mathbb{E}\left[\frac{N_{k+1}(s, \tilde{\pi}(s)) - N_k(s, \tilde{\pi}(s))}{\sqrt{N_k(s, \tilde{\pi}(s)}}\right].
\end{aligned}
$$

Bounding the regret of our algorithm (5)

Since $N_{k+1}(s,a) \le 2N_k(s,a)$ and $\sum_{s,a} N_{K+1}(s,a) = T$, summing over k gives

$$
\begin{aligned}
& 2T_{hit}^{\min}|S|\sqrt{3\log T}\sum_k\sum_s \frac{N_{k+1}(s,\tilde{\pi}_k(s)) - N_k(s,\tilde{\pi}_k(s))}{\sqrt{N_k(s,\tilde{\pi}_k(s))}} \\
= \quad & O\left(T_{hit}^{\min}|S|\sqrt{\log T}\sum_{k,s,a}\frac{N_{k+1}(s,a) - N_k(s,a)}{\sqrt{N_k(s,a)}}\right) \\
= \quad & O\left(T_{hit}^{\min}|S|\sqrt{\log T}\sum_{s,a}\sqrt{N_{K+1}(s,a)}\right) \\
= \quad & O\left(T_{hit}^{\min}|S|\sqrt{|S||A|T\log T}\right).
\end{aligned}
$$