Safe Reinforcement Learning

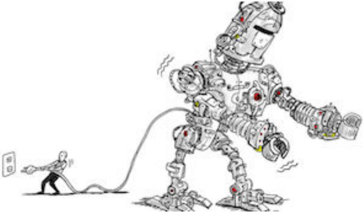
Eloïse BERTHIER & Julien CHHOR

January 29, 2019

Introduction

Safety guarantees are essential to bring RL into real world applications.

There are different definitions of safety, leading to different safe RL algorithms.



Layout

Increasing Performance

Definition

An algorithm is safe if it guarantees an improvement of the policy performance over time.

The performance of a policy is measured by: $J_\mu^\pi = \mathbb{E}_{s\sim\mu}[V^\pi(s)]$.

This is true for exact policy iteration as $V^{\pi_{k+1}} \geq V^{\pi_k}$.

$\rightarrow$ Aim: ensuring a monotonic improvement even if the algorithm makes approximations.

Lower bounds on the performance gap

Lemma ([KL02])

For any policies π and π' and any initial state distribution μ

$$J_\mu^{\pi'} - J_\mu^{\pi} = \int_{\mathcal{S}} d_\mu^{\pi'}(s) A_\pi^{\pi'}(s) \mathrm{d}s$$

Theorem ([PRPC13])

For any π and π' and μ,

$$J_\mu^{\pi'} - J_\mu^{\pi} \geq \int_{\mathcal{S}} d_\mu^{\pi}(s) A_\pi^{\pi'}(s) \mathrm{d}s - \frac{\gamma}{2(1-\gamma)^3} ||\pi' - \pi||_\infty^2$$

This lower bound can be used to design safe algorithms.

Safe Policy Iteration

- Conservative policy update [KL02]: $\pi' = \alpha\bar{\pi} + (1-\alpha)\pi$.
- α is chosen to **maximize the lower bound**.
- It guarantees that [PRPC13]: $J_\mu^{\pi'} - J_\mu^{\pi} \geq \dfrac{(1-\gamma)^2 {\mathbb{A}_{\pi,\mu}^{\bar{\pi}}}^2}{2\gamma\|\bar{\pi}-\pi\|_\infty \Delta A_\pi^{\bar{\pi}}}$.

SPI Algorithm (one step):
select $\bar{\pi}$ maximizing a sample-based version of the Q-function
produce $\hat{\mathbb{A}}_{\pi,\mu}^{\bar{\pi}}$ an estimate of $\mathbb{A}_{\pi,\mu}^{\bar{\pi}}$
if $\hat{\mathbb{A}}_{\pi,\mu}^{\bar{\pi}} \geq \eta$ **then**
  set α maximizing the lower bound $\pi \leftarrow \alpha\bar{\pi} + (1-\alpha)\pi$
else
  stop and return π
end

Adaptive Step-Size for Policy Gradient

Gaussian policies $\pi(.|s,\theta) \sim \mathcal{N}(\theta^\top \phi(s), \sigma^2)$.

- Policy gradient update: $\theta' = \theta + \alpha \nabla_\theta J_\mu(\theta)$.
- α is chosen to **maximize the lower bound**.
- It guarantees that [PRB13]: $J_\mu(\theta') - J_\mu(\theta) \geq \frac{1}{2}\alpha^\star ||\nabla_\theta J_\mu(\theta)||_2^2$

PG Algorithm with adaptive step-size (one step):

estimate $\nabla_\theta J_\mu(\theta)$ with REINFORCE
set α maximizing the lower bound
$\theta' \leftarrow \theta + \alpha \nabla_\theta J_\mu(\theta)$

Layout

Definition

Definition

An algorithm is safe if it ensures a return $\geq J_{-}$ with probability $\geq 1 - \delta$

Moreover: no hyper parameters $\Rightarrow$ further source of safety

Idea: Offline Estimation

We know trajectories already executed: $\mathcal{D} = \{(\tau_i, \theta_i) | i = 1, ..., n\}$

For a new policy π_e, **importance sampling** gives an unbiased estimator of $J(\pi_e)$

$$\hat{J}(\pi_e, \tau_i, \theta_i) = \mathcal{R}(\tau_i) \frac{Pr(\tau_i | \theta)}{Pr(\tau_i | \theta_i)}$$

(Condition: $\forall a, s, \theta_i : \pi(a|s, \theta_i) = 0 \Rightarrow \pi(a|s, \theta) = 0$) Otherwise

the $\hat{J}(\pi_e, \tau_i, \theta_i)$ are underestimated $\Rightarrow$ more conservative

3 ways to compute $1 - \delta$ lower bound:

1. Concentration inequality
2. Asymptotic normality
3. Bootstrapping

1. Concentration Inequality

Lemma

$\boldsymbol{X} = (X_i)_{i=1}^n$ independent, ≥ 0, bounded, with $E(X_i) \leq \mu$.
$c_i \in \mathbb{R}$, $\delta > 0$, and $Y_i := \min(X_i, c_i)$.
Then w.p. $\geq 1 - \delta$

$$\mu \geq \left(\sum_{i=1}^n \frac{1}{c_i}\right)^{-1} \left\{ \sum_{i=1}^n \frac{Y_i}{c_i} - \frac{7 c_i \log(2/\delta)}{3(n-1)} - \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^n \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)} \right\}$$

1. Concentration Inequality

Theorem

Let $J_-(\boldsymbol{X}, \delta, n, c)$ be a $1-\delta$ confidence lower bound on $\overline{x}$ computed thanks to $\boldsymbol{X}$ from the lemma with $c_i = c$, where $\boldsymbol{X}$ contains n random variables. If we had made the computation thanks to a set $\boldsymbol{X}'$ containing m random variables, where $\boldsymbol{X}'$ has the same variance as $\boldsymbol{X}$, then the lower bound would have been:

$$J_-(\boldsymbol{X}, \delta, m, c) = \frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{7c\log(2/\delta)}{3(m-1)}$$

$$-\sqrt{\frac{2\log(2/\delta)}{mn(n-1)}\left(n\left(\sum_{i=1}^{n} Z_i^2\right) - \left(\sum_{i=1}^{n} Z_i\right)^2\right)}$$

where $Z_i = \min(X_i, c)$.

2. Asymptotic normality

"Under mild conditions":
$\frac{1}{n}\sum_{i=1}^{n} X_i$ is asymptotically normally distributed
Hence $1-\delta$ lower bound

$$\hat{X} - \frac{\hat{\sigma}}{\sqrt{m}} t_{1-\delta,m-1}$$

where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{X})$

And we apply that to $X_i = \hat{J}(\pi_e, \tau_i, \theta_i)$

3. Bootstrap

Idea: Estimate the true distribution of the $\hat{J}(\pi_e, \tau_i, \theta_i)$

Using bias corrected and accelerated bootstrapping

Hence: $1 - \delta$ confidence lower bound

Policy Improvement Algorithm

Goal: Find $\pi' = \arg\max_{\text{safe } \pi} \hat{J}(\pi|\mathcal{D})$

We split $\mathcal{D}$ into $\mathcal{D}_{train}$ (20%) and $\mathcal{D}_{test}$ (80%)

PolicyImprovement$(\mathcal{D}_{train}, \mathcal{D}_{test}, \delta, J_{-})$

- Find the best π_c on $\mathcal{D}_{train}$
- Test it on $\mathcal{D}_{test}$: if $\geq J_{-}$ return π_c else NSF

To find π_c on $\mathcal{D}_{train}$ we can use cross-validation

Deadalus

Idea: At every step:

- Threshold $\leftarrow$ best lower bound so far.

Incremental algorithm $\Rightarrow$ brings back to the first definition

Layout

Constrained MDP

An extension to the MDP: $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{R}, \gamma, \mu, (\mathbf{C_j}, \mathbf{d_j})_{\mathbf{1 \leq j \leq m}})$.

Constraint satisfaction is required in expectation:

$$J^{\pi}_{C_j} = \mathbb{E}_{s \sim \mu, a \sim \pi} \Big[\sum_{t=0}^{\infty} \gamma^t C_j(s_t, a_t) | s_0 = s \Big] \leq d_j$$

Solving the MDP:

$$\underset{\pi \ s.t. \ \forall j, J^{\pi}_{C_j} \leq d_j}{\arg\max} \quad J^{\pi}_{\mu}$$

Definition

An algorithm is safe if it provides policies satisfying the constraints.

Algorithms for CMDPs

- Solving the Lagrangian is numerically hard and gives **no safety guarantee during training**.
- Local policy search over a **trust region**:

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} J_\mu^\pi \text{ s.t. } \mathbb{E}_{s \sim \pi_k}[KL(\pi || \pi_k)(s)] \leq \delta \text{ and } \forall j, J_{C_j}^\pi \leq d_j$$

- Results from [KL02] are extended to simultaneously lower-bound the **performance** increase and the **constraint** decrease.

Constrained Policy Optimization

Theorem ([AHTA17])

For any policies π and π':

$$J_{\mu}^{\pi'} - J_{\mu}^{\pi} \geq \mathbb{A}_{\pi,\mu}^{\pi'} - \frac{2\gamma\epsilon^{\pi'}}{(1-\gamma)^2}\int_{\mathcal{A}} \pi'(a|s)\sqrt{\frac{1}{2}\mathbb{E}\left[KL(\pi'||\pi)(s)\right]}\mathrm{d}a$$

$$J_{C_j}^{\pi'} - J_{C_j}^{\pi} \leq \mathbb{A}_{\pi,C_j}^{\pi'} + \frac{2\gamma\epsilon_{C_j}^{\pi'}}{(1-\gamma)^2}\int_{\mathcal{A}} \pi'(a|s)\sqrt{\frac{1}{2}\mathbb{E}\left[KL(\pi'||\pi)(s)\right]}\mathrm{d}a$$

With high probability:

- the constraints are enforced;
- the performance increases at each iteration.

Layout

Discussion

Three approaches (among others):

1. increasing performance
2. surpassing a baseline
3. enforcing constraints

Direct comparisons:

- (2) covers (1) with a sliding baseline;
- (3) can cover (2) for a fixed baseline on performance;
- (3) enables constraints on other criteria.

Generalization: CMDP with varying constraint level during training?

References I

- Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel, *Constrained policy optimization*, International Conference on Machine Learning, 2017, pp. 22–31.
- Sham Kakade and John Langford, *Approximately optimal approximate reinforcement learning*, 2002.
- Matteo Pirotta, Marcello Restelli, and Luca Bascetta, *Adaptive step-size for policy gradient methods*, Advances in Neural Information Processing Systems, 2013, pp. 1394–1402.
- Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello, *Safe policy iteration*, International Conference on Machine Learning, 2013, pp. 307–315.