Обзор статьи
Safe Model-based Reinforcement Learning with Stability Guarantees

Гаврилов Илья

6 ноября 2017

Содержание

Related work

- Risk-sensitive RL
  - Дискретный MDP
  - Требует точную вероятностную модель
- Model-free policy search
  - Итоговые стратегии специфичны для задачи
  - Требует перезагрузок системы
- Model-based RL
  - Применяемые алгоритмы не имеют теоретических гарантий
  - Подходы с переключением известных безопасных стратегий

Неформальная идея

- Есть неизвестное множество безопасных состояний
- Изначально задаем локально безопасную политику из априорных знаний
- Исследуем текущее безопасное множетсво
- Улучшаем стратегию и расширяем безопасное множество

Background

- $\mathcal{X} \subseteq \mathbb{R}^q$ - пространство состояний,
  $\mathcal{U} \subseteq \mathbb{R}^p$ - пространство действий,
  $t \in \mathbb{N}$ - индекс времени (дискретное)
- $x_{t+1} = f(x_t, u_t) = h(x_t, u_t) + g(x_t, u_t)$
- $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ - true dynamic
- h - известная априорная модель, g - априорно неизвестная ошибка модели
- $\pi : \mathcal{X} \to \mathcal{U}$ - стратегия

Background

- Кодируем эффективность через стоимость (вместо награды)
- Стоимость $r : \mathcal{X} \times \mathcal{U} \to \mathbb{R}_{\geq 0}$,
  $r(0, 0) = 0$,
  $r(x, u) > 0$ в остальных случаях

Предположения

- Функции g и h - Липшецевы с константами L_g и L_h с 1-нормой
  Def: Функция f L_f - Липшицева, если
  $\forall x, y \in Dom(f),\ \rho(f(x) - f(y)) \leq L_f \rho(x - y)$, где ρ - некоторая норма
- Рассматриваются стратегии из множества Π_L - L_π-Липшицевы функции
- Для теоретических выкладок считаем, что распределения нормальные

Предположения

- Пусть μ_n, Σ_n - среднее и ковариационная матрица апостериорного распределения после n наблюдений, $\sigma_n = tr(\Sigma_n^{1/2})$.
  Существует такое $\beta_n > 0$, что с вероятностью $1 - \delta$ имеем
  $\forall n \geq 0, \forall x \in \mathcal{X}, u \in \mathcal{U} \quad \|f(x,u) - \mu_n(x,u)\|_1 \leq \beta_n \sigma_n(x,u)$.
- Это условие необходимо, чтобы в дальнейшем строить интервалы, в которых будем уверены с высокой вероятностью

Функция Ляпунова

- Def: бесконечно дифференцируемая функция $v : \mathcal{X} \to \mathcal{R}_{\geq 0}$, причем $v(0) = 0$ и $v(x) > 0,\ \forall x \in \mathcal{X} \backslash \{0\}$
- Эта функция нужна, чтобы определить безопасную область и затем не выходить за ее пределы в процессе исследования пространства

Теорема 1

- Пусть v - функция Ляпунова, f - Липшицева функция процесса, π - стратегия.
  Если $v(f(x, \pi(x)) < v(x)$, $x \in \mathcal{V}(c) = \{x \in \mathcal{X} \backslash \{0\} | v(x) < c\}$, $c > 0$, то $\mathcal{V}(c)$ - безопасная область, причем $x_0 \in \mathcal{V}(c)$ следует, что $x_t \in \mathcal{V}(c)$, $\forall t > 0$ и $lim_{t \to +\infty} x_t = 0$

Безопасная область

- Хотим: $v(f(x, \pi(x)) - v(x) < 0,\ x \in \mathcal{V}(c)$
- Это сложная задача, поэтому дискретизируем X:
  $X_\tau \subset X :\ \|x - [x]_\tau\|_1 \leq \tau,\ \forall x \in X$
- Теорема: Пусть $L_{\Delta v} = L_v L_f (L_\pi + 1) + L_v$. Если
  $\forall x \in \mathcal{V}(c) \cap X_\tau,\ c > 0$ выполнено $v(f(x, \pi(x))) - v(x) < -L_{\Delta v}\tau$,
  то $v(f(x, \pi(x))) < v(x),\ \forall x \in \mathcal{V}(c)$ с вероятностью не меньше
  $(1 - \delta)$, и $\mathcal{V}(c)$ - безопасная область

Оптимизация стратегии

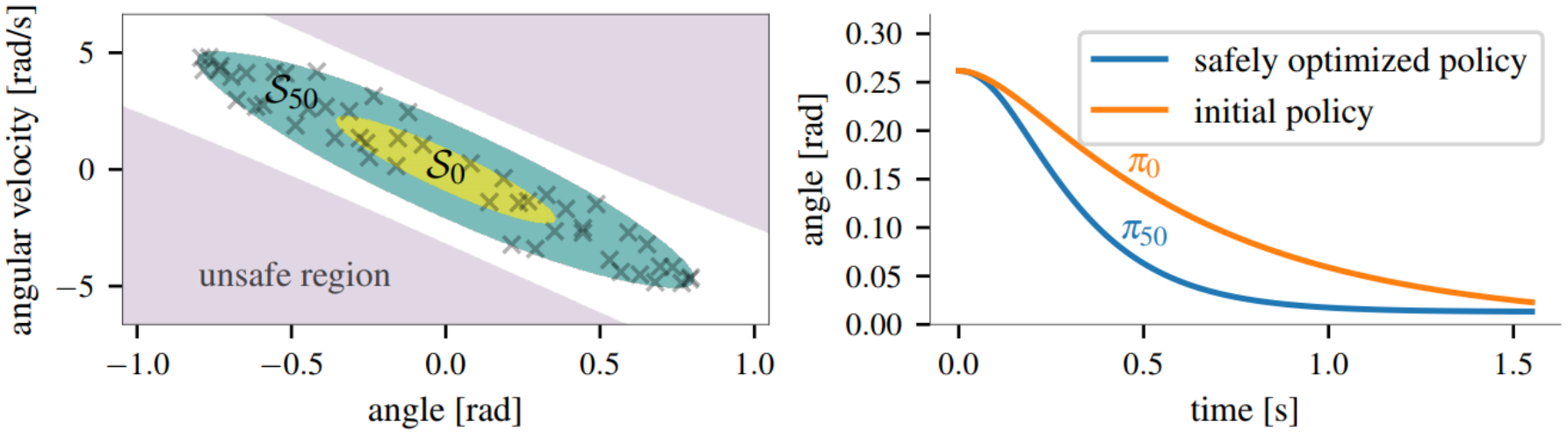
- Цель: максимально расширить множество $\mathcal{V}(c)$
- $\pi_n, c_n = argmax_{\pi \in \Pi_L,\, c \in \mathbb{R}_{>0}} c$, такое что $\forall x \in \mathcal{V} \cap \mathcal{X}_\tau$ $v(f(x, \pi(x))) - v(x) < -L_{\Delta v}\tau$
- Задача сложная, поэтому на практике использовалась эвристика: сначала просто оптимизировалась стратегия (обычными методами), затем вычислялась безопасная область, т.к. точки, по которым обновлялась политика лежат в безопасной область, то ничего не сломалось, и если безопасная область сузилась, то откатываемся к предыдущей политике

Практическая реализация

- π_θ - стратегия с параметрами θ
- J_{π_θ} - cost-to-go function, дисконтированная сумма стоимостей с дисконтом γ.
- $\pi_n = argmin_{\pi_\theta \in \Pi_L} \sum_{x \in \mathcal{X}} r(x, \pi_\theta(x)) + \gamma J_{\pi_\theta}(\mu_{n-1}(x, \pi_\theta(x))) + \lambda(v(f(x, \pi_\theta(x)) - v(x) + L_{\Delta v}\tau)$
- В экспериментах брали $\lambda = 1$ и $v = J$
- Вычисление безопасного множества через функцию Ляпунова

Эксперимент

- Задача: inverted pendulum



(a) Estimated safe set.

(b) State trajectory (lower is better).

Figure 2: Optimization results for an inverted pendulum. Fig. 2(a) shows the initial safe set (yellow) under the policy π_0, while the green region represents the estimated region of attraction under the optimized neural network policy. It is contained within the true region of attraction (white). Fig. 2(b) shows the improved performance of the safely learned policy over the policy for the prior model.

Литература

- Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, Andreas Krause, Safe Model-based Reinforcement Learning with Stability Guarantees, arXiv ,2017