Разработка робастных методов схвата манипулятором произвольных объектов на основе обучения с подкреплением и методов оптимального управления

Синицкий Максим Игоревич

Центр когнитивного моделирования МФТИ

Научно-исследовательская работа

Мотивация



Мотивация

- Автоматизация производства: Рост потребности в гибких и адаптивных манипуляторах для различных производственных задач.
- Разнообразие объектов
- Эффективность методов
- Интеграция технологий обучения с подкреплением и оптимального управления. Объединение обучения с подкреплением и методов оптимального управления

Цель исследования

- Разработка и внедрение робастных методов схвата для манипуляторов, способных эффективно взаимодействовать с произвольными объектами в условиях неопределённости.
- Объединение алгоритмов обучения с подкреплением и методов оптимального управления для создания гибкой и устойчивой системы управления схватом.
- Проведение сравнительного анализа предложенных методов с существующими подходами для подтверждения преимуществ гибридных методов.

Формальная постановка задачи (POMDP)

- $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle$
  - $\mathcal{S}$: Пространство состояний
    $s_t = [q_1, \dot{q}_1, q_2, \dot{q}_2, F_{\text{вз}}, O_{\text{об}}]$
    - ★ $q_i \in \mathbb{R}^n$: углы в сочленениях i-й руки
    - ★ $\dot{q}_i \in \mathbb{R}^n$: скорости в сочленениях
    - ★ $F_{\text{вз}} \in \mathbb{R}^m$: силы взаимодействия с объектом
    - ★ $O_{\text{об}} \in \mathbb{R}^k$: положение/ориентация объекта
  - $\mathcal{A}$: Пространство действий
    $a_t = [\tau_1, \tau_2]$, где $\tau_i \in \mathbb{R}^n$ — управляющие моменты для i-й руки
  - $\mathcal{P}$: Функция переходов
    $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$, задаётся уравнениями динамики
  - $\mathcal{R}$: Функция вознаграждения

$$r_t = -\alpha \underbrace{\|O_{\text{об}} - O_{\text{цел}}\|^2}_{\text{точность}} - \beta \underbrace{\sum_{i=1}^{2} \|\tau_i\|^2}_{\text{энергия}} + \gamma \underbrace{\mathbb{I}_{\text{grasp}}}_{\text{успех схвата}}$$

  - Ω: Пространство наблюдений
    $o_t = [q + \epsilon_q, \dot{q} + \epsilon_{\dot{q}}, F_{\text{вз}} + \epsilon_F, O_{\text{об}} + \epsilon_O]$
  - $\mathcal{O}$: Функция наблюдений
    $O(o_t|s_t)$ — распределение наблюдений при заданном состоянии
  - γ: Коэффициент дисконтирования

Связь с оптимальным управлением

Динамика манипулятора (задает функцию перехода среды):

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) = \tau$$

- $q \in \mathbb{R}^n$ — вектор обобщённых координат (углы в сочленениях)
- $M(q) \in \mathbb{R}^{n\times n}$ — матрица инерции (симметричная положительно определённая)
- $C(q,\dot{q}) \in \mathbb{R}^{n\times n}$ — матрица Кориолиса и центробежных сил
- $G(q) \in \mathbb{R}^n$ — вектор гравитационных сил
- $\tau \in \mathbb{R}^n$ — вектор управляющих моментов

Целевая функция оптимального управления:

$$\min_{\tau} \int_0^T \left(\underbrace{\|q(t) - q_d(t)\|_Q^2}_{\text{Точность}} + \underbrace{\|\tau(t)\|_R^2}_{\text{Энергия}} \right) dt$$

- $q_d(t)$ — желаемая траектория
- $Q \succeq 0$, $R \succ 0$ — весовые матрицы

Ограничения:

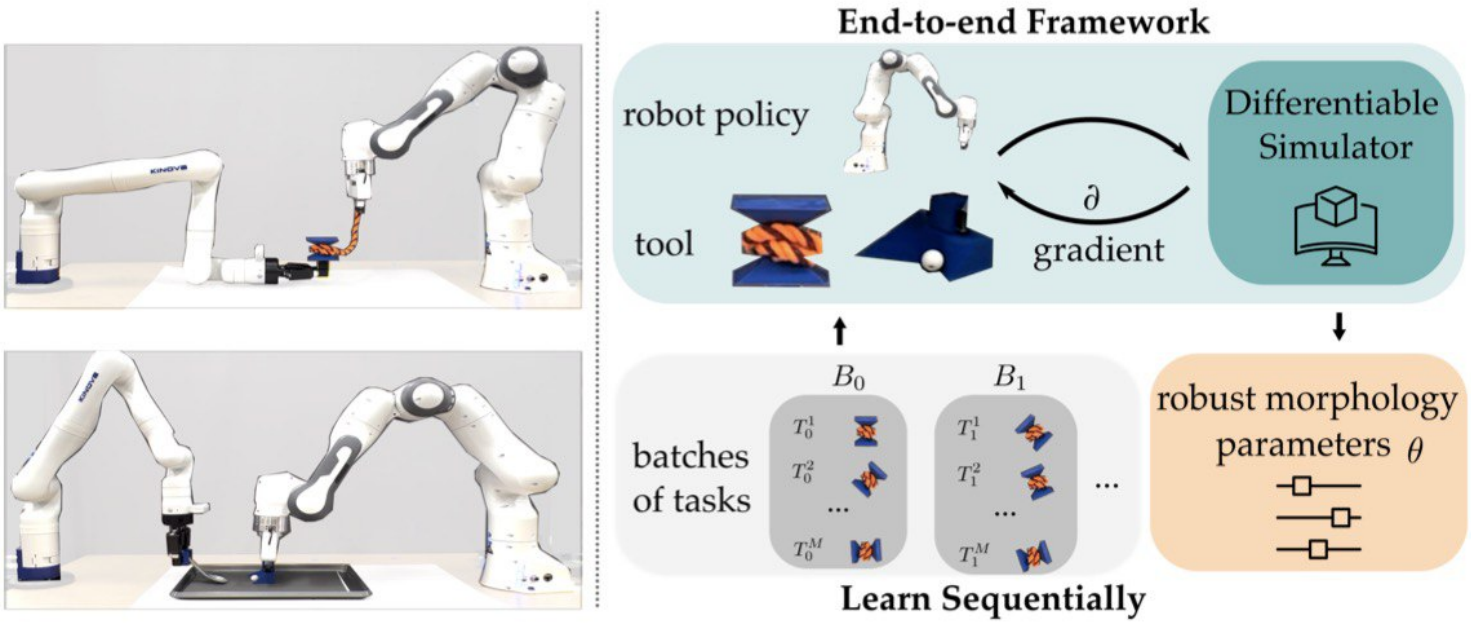
$$F_{\min} \le \|F_{\text{вз}}(t)\| \le F_{\max}$$

Выделенные подзадачи работы

1. Запуск среды и базовых решений, основанных на обучении с подкреплением
   - Подбор параметров обучения.
   - Запуск одноагентный и мультиагентных алгоритмов.
2. Интегрирование методов оптимального управления в методы обучения с подкреплением
3. Симуляции и эксперименты:
   - Проведение экспериментов и анализ полученных данных.
   - Определение метрики качества для задачи.

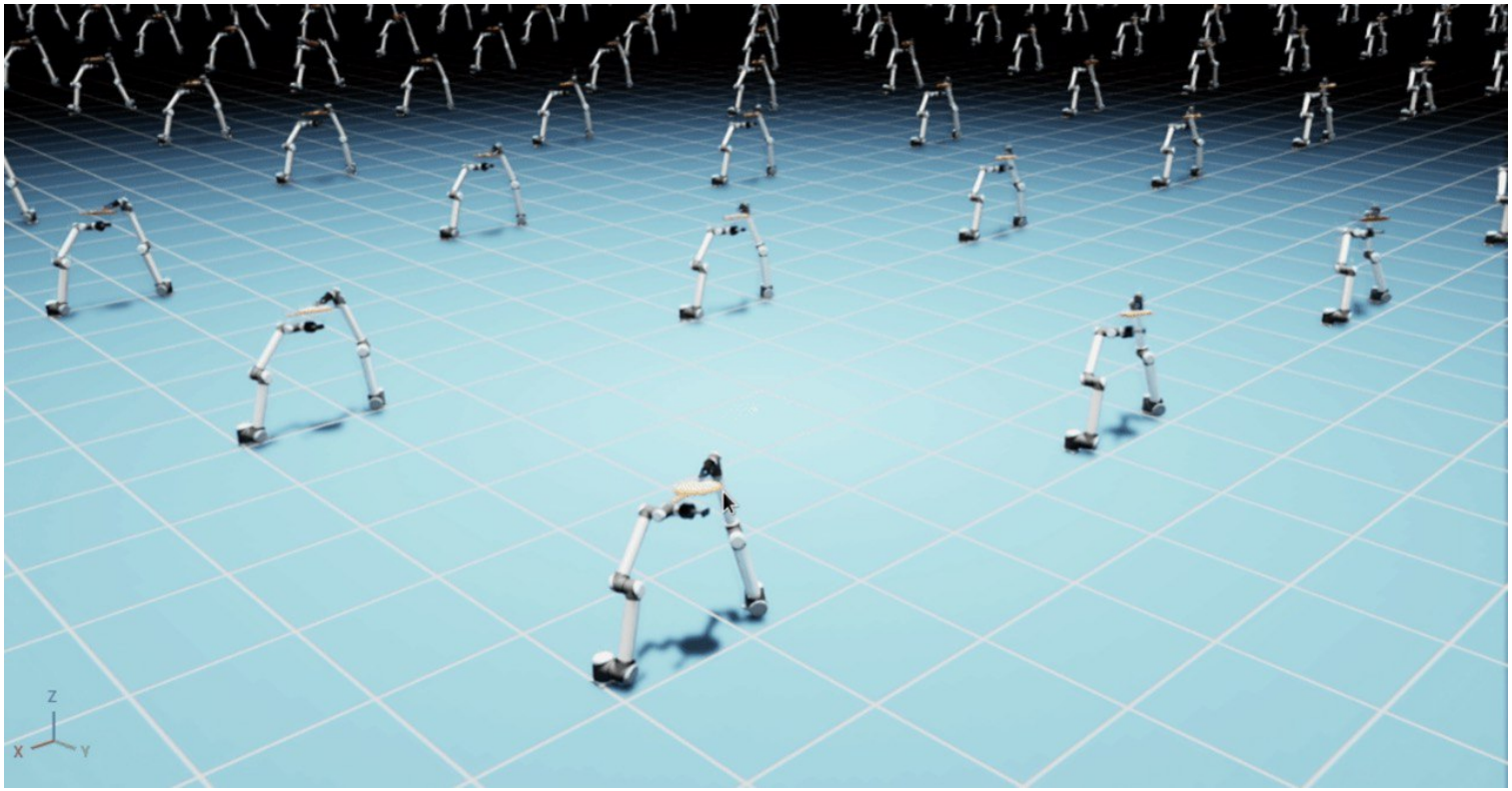
Описание используемого метода

- **Оптимальное управление:**
  - Применение методов, таких как линейно-квадратичное регуляторное управление (LQR), для обеспечения устойчивости.
- **Интеграция подходов:**
  - Сочетание преимуществ RL и оптимального управления для создания гибкой и надёжной системы управления.
  - Использование RL для адаптации к динамическим изменениям, а оптимального управления — для обеспечения стабильности.



Метод

Первые эксперименты/создание бейзлайнов



Среда выполнение

Первые эксперименты/создание бейзлайнов

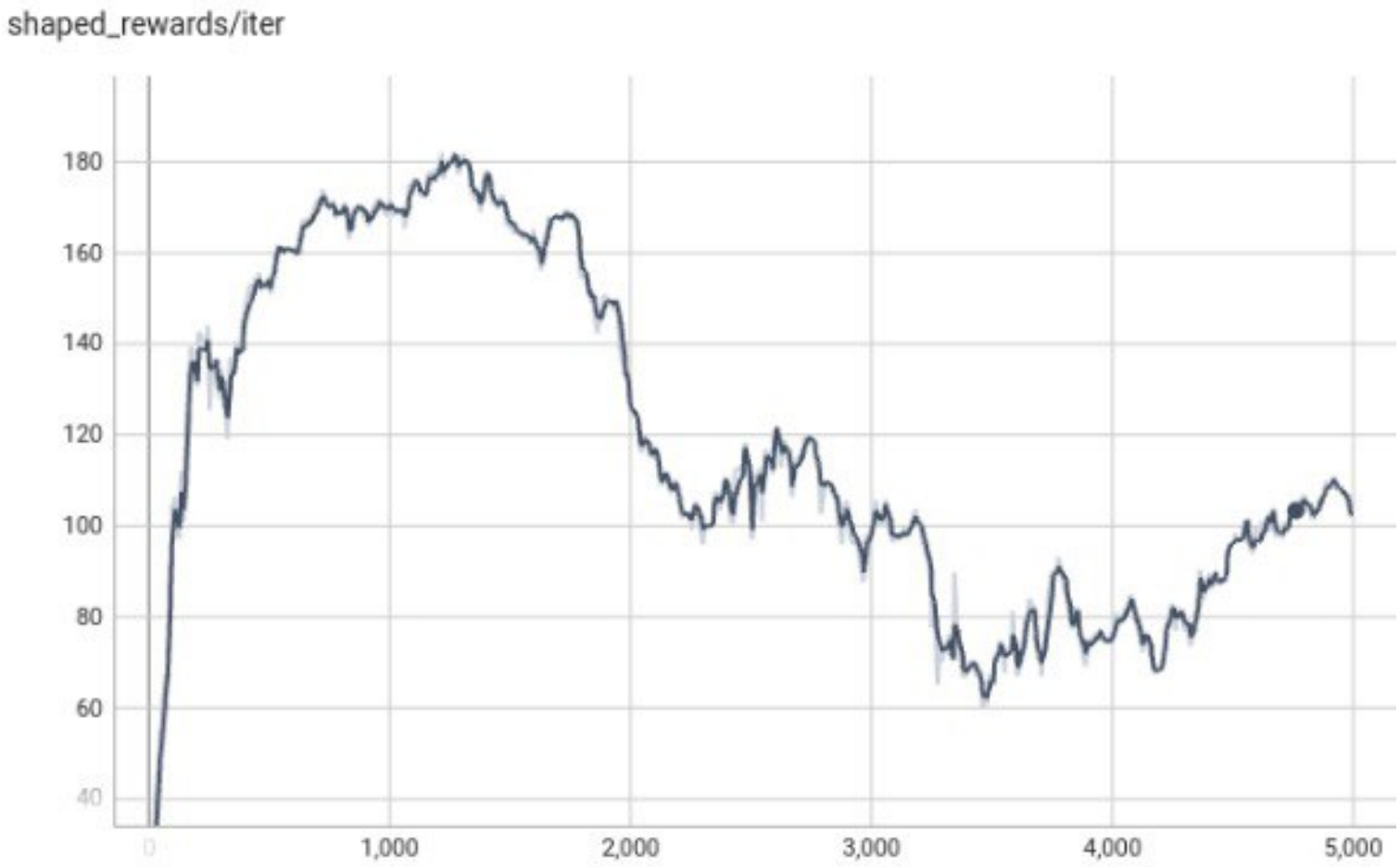


Рисунок 2: График shaped rewards MAPPO в зависимости от итераций

Спасибо за внимание!