Исследование алгоритмов оптимального управления, основанных на обучении с подкреплением

Студент: Динь Нгок Туан, группа R34372

Научный руководитель:
Перегудин Алексей Алексеевич

Санкт-Петербург 2024

ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

- Обзор литературы
  - Концепция и компоненты обучения с подкреплением
  - Основные алгоритмы обучения с подкреплением
- Математическая модель регулятора на основе обучения с подкреплением
  - Линейная система
  - Нелинейная система
- Моделирование и анализ эффективности регуляторов
- Применение алгоритмов к физическим системам

Оптимальное управление

Оптимальное управление

- Уравнения Гамильтона Якоби Беллмана
- Функция ценности (HJB)
Зачастую, невозможно найти
- Оптимальная стратегия управления невозможно вычислить, если функция ценности или динамика объекта неизвестны

Численные решения

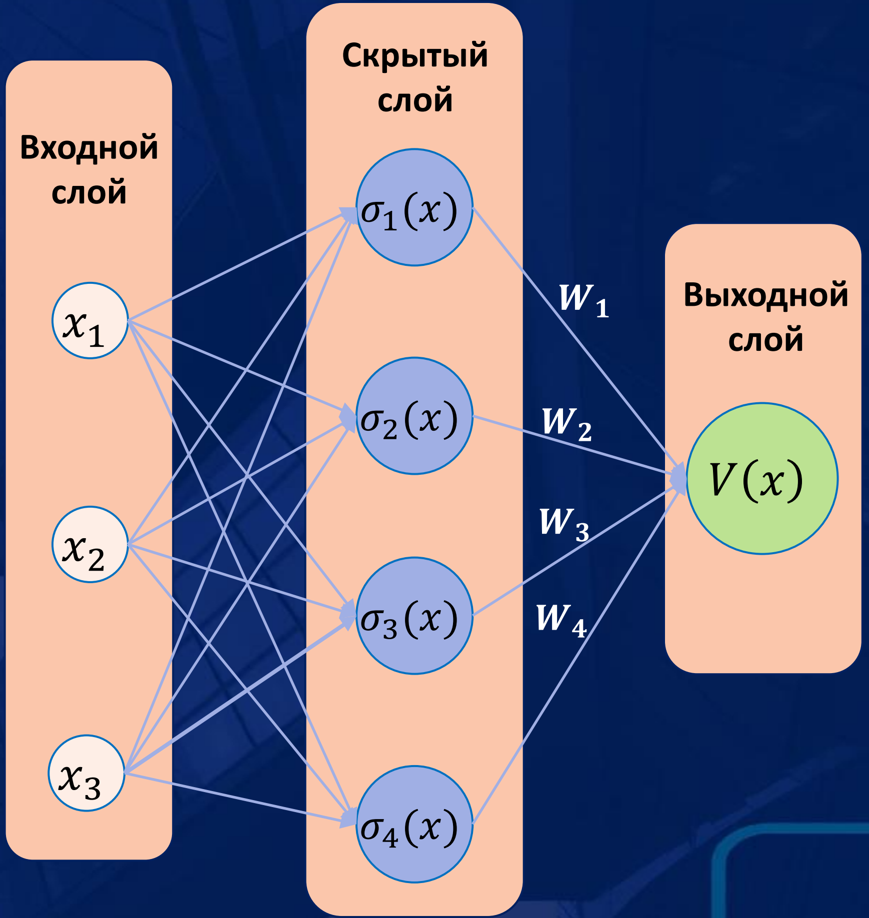
- Как правильно, в обратном времени
- Обычно решаются в автономном режиме
- Высокая сложность вычислений
- Требуется знание модели
- Плохо приспособлены к изменению параметров объекта.

Приближительное динамическое программирование (ADP)

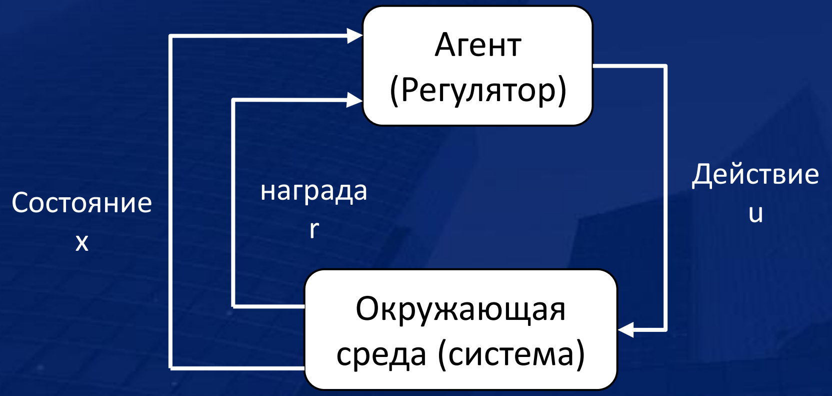
Динамическое программирование
+ В обратном времени
+ Необходимо знание модели

Обучение с подкреплением
+ Обучение через взаимодействия
+ Компромисс между исследованием и использованием

ADP
+ в прямом времени
+ Нейронные сети



Оптимальное управление с использованием метода обучения с подкреплением



Задача оптимального управления

Поиск

Оптимальное значение функции ценности

$$V^*(x) = \min_{\pi} V^{\pi}(x)$$

Оптимальная стратегия управления

$$u^* = \underset{u}{\operatorname{argmin}} \sum_{x'} P_{xx'}^{u}\left[R_{xx'}^{u} + \gamma V^{\pi}(x')\right]$$

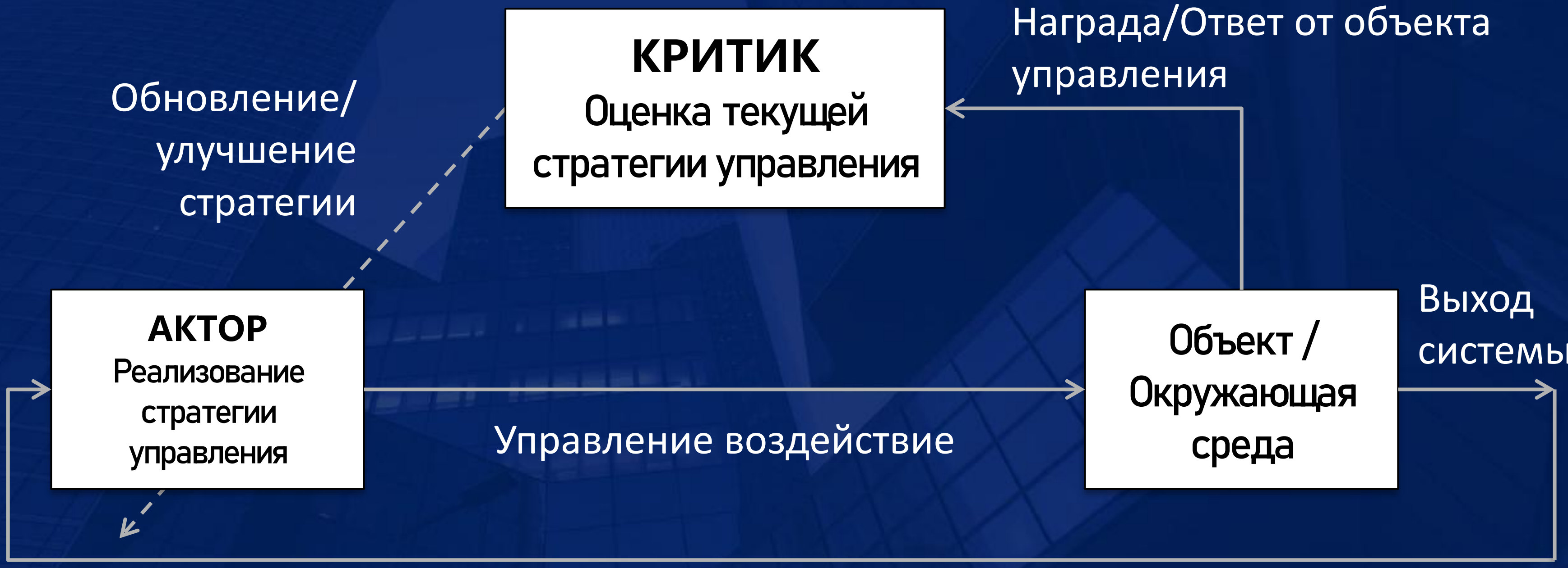
Функция ценности в виде уравнения Беллмана

$$V^{\pi}(x) = \sum_{u} \pi(x,u) \sum_{x'} P_{xx'}^{u}\left[R_{xx'}^{u} + \gamma V^{\pi}(x')\right]$$

Решение

- Оценка стратегии
- Улучшение стратегии

Оценка и улучшение стратегии



Постановка задачи для линейных систем

Рассмотрим линейную динамическую систему

$$\dot{x} = Ax + Bu$$

$$A \ - \text{неизвестная матрица}$$

Закон управления с обратной связью

$$u = -Kx(t)$$

Цель: Синтезировать регулятор, который обеспечивает выполнение условия

$$\lim_{t\to\infty} |x(t)| = 0$$

а также минимизацию квадратичного функционала качества

$$J(x,u) = \int_0^\infty \big(x^T(t)Qx(t) + u^T(t)Ru(t)\big)dt$$

Обучение с интегральным подкреплением (IRL)

Функции ценности:

$$V(x(t)) = \int_t^{\infty} \big(x^T(t)Qx(t) + u^T(t)Ru(t)\big)dt$$

IRL Уравнение Беллмана:

$$V\big(x(t)\big) = \int_t^{t+T} (x^T(\tau)(Q + K^T RK)x(\tau))d\tau + V\big(x(t+T)\big)$$

Форма интегрального обучения с подкреплением:

$$\rho(x(t), t, T) = \int_t^{t+T} (x^T(\tau)(Q + K^T RK)x(\tau))d\tau$$

Итерации по стратегии

Этап оценки стратегии

$$x_t^T P_k x_t = \int_t^{t+T} \big(x_\tau^T Q x_\tau + K_k^T R K_k\big) x_\tau d\tau \ + x_{t+T} P_k x_{t+T}$$

Этап улучшения стратегии

$$K_{k+1} = R^{-1} B^T P_k$$

Итерации по критерию

Этап оценки стратегии

$$x_t^T P_{k+1} x_t = \int_t^{t+T} \big(x_\tau^T Q x_\tau + K_k^T R K_k\big) x_\tau d\tau \ + x_{t+T} P_k x_{t+T}$$

Этап улучшения стратегии

$$K_{k+1} = R^{-1} B^T P_{k+1}$$

Моделирование

Данная система

$$\dot{x}(t) = Ax(t) + Bu(t), A = \begin{bmatrix} 0 & 1 \\ 10 & -10 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Функция стоимости

$$J(x,u) = \int_0^\infty \left(x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau)\right)d\tau$$

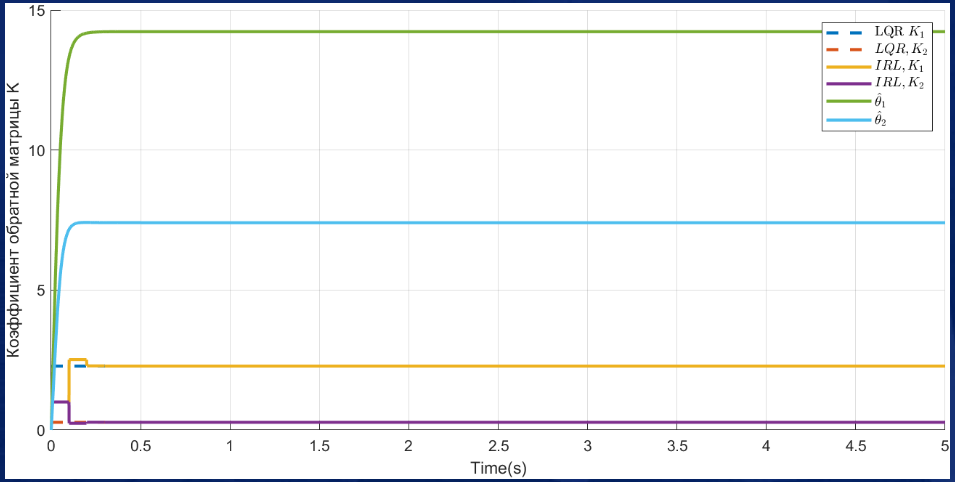
$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = 1$$

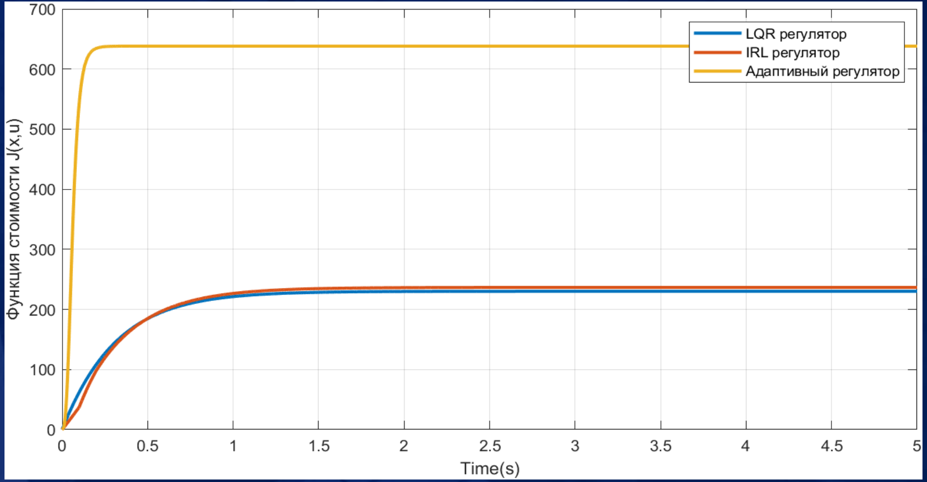
Решение уравнения Риккати

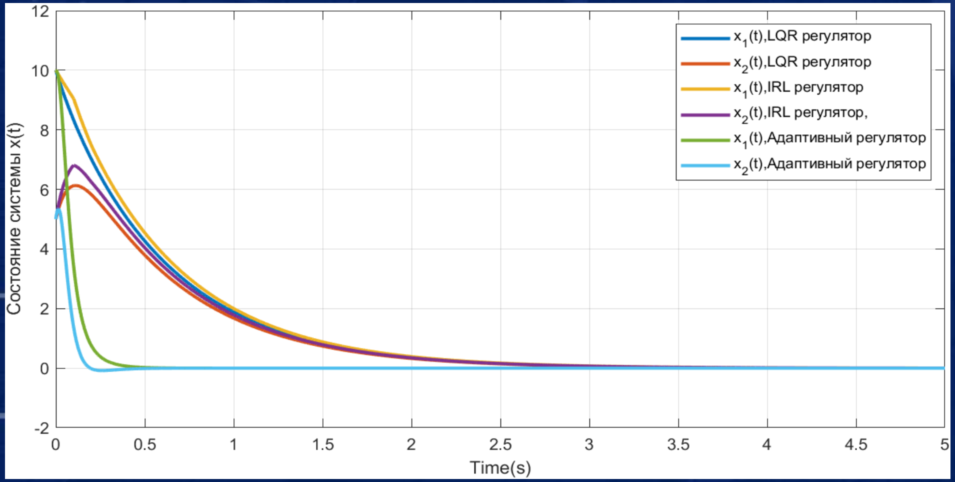
$$P = \begin{bmatrix} 2.0739 & 0.211 \\ 0.211 & 0.0672 \end{bmatrix}, K = R^{-1}B^TP = [2.284 \quad 0.278]$$

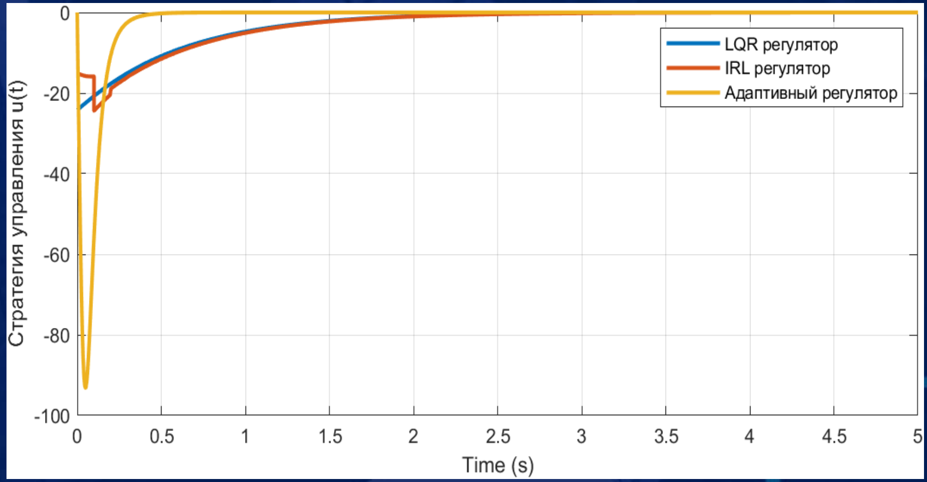
$$T = 0.025$$

Результаты моделирования

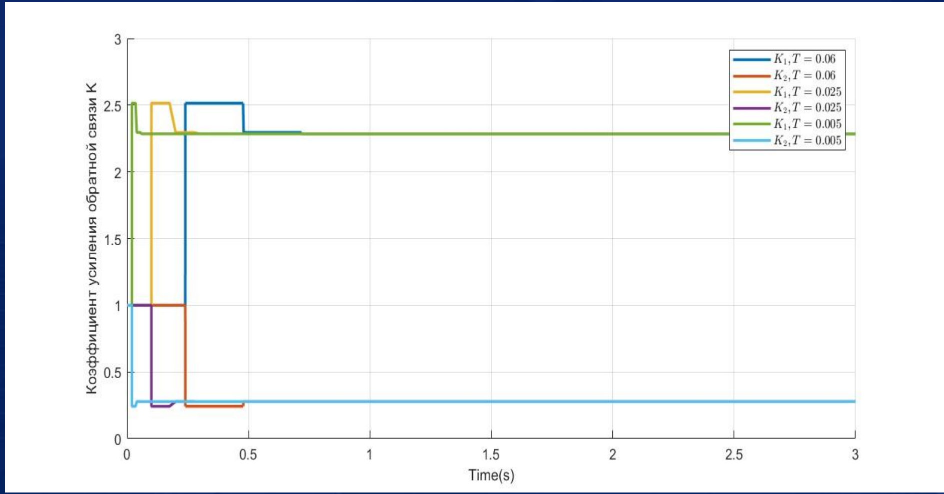


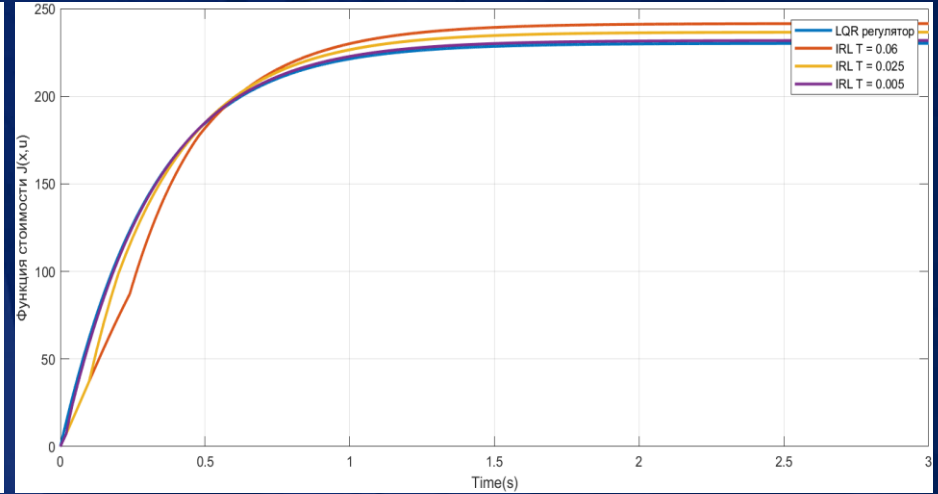


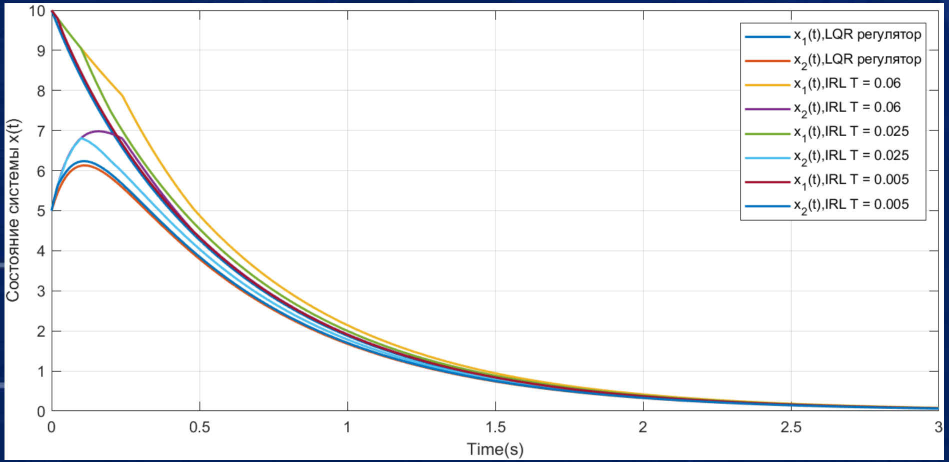


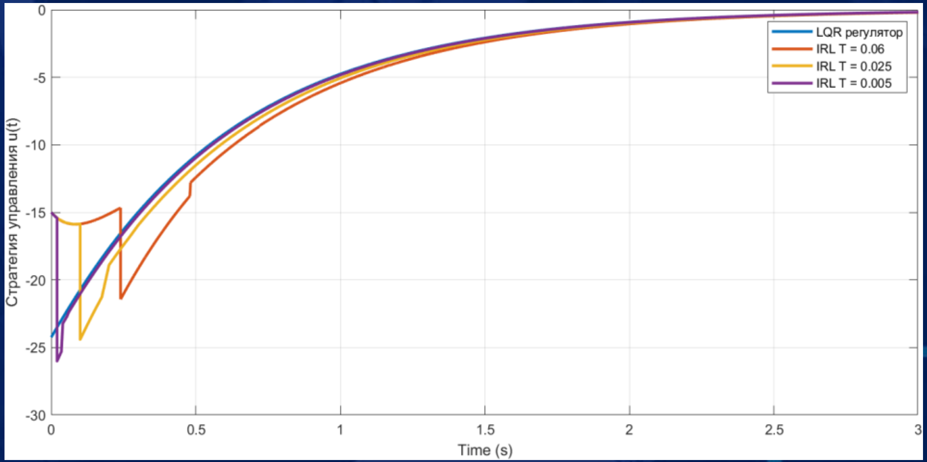


Влияние времени выборки данных T









Регулятор для перевернутого маятника и тележки

Объект управления

$$\begin{bmatrix}\dot\theta\\ \ddot\theta\\ \dot x\\ \ddot{x}\end{bmatrix}=\begin{bmatrix}0 & 1 & 0 & 0\\ \frac{(M+m)g}{ML} & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ -\frac{mg}{M} & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\theta\\ \dot\theta\\ x\\ \dot{x}\end{bmatrix}+\begin{bmatrix}0\\ -\frac{1}{ML}\\ 0\\ \frac{1}{M}\end{bmatrix}u$$

$$M=2.4\text{ кг}, m=0.23\text{ кг}, g=9.81\frac{\text{м}}{\text{с}^2}, L=0.46\text{ м}$$

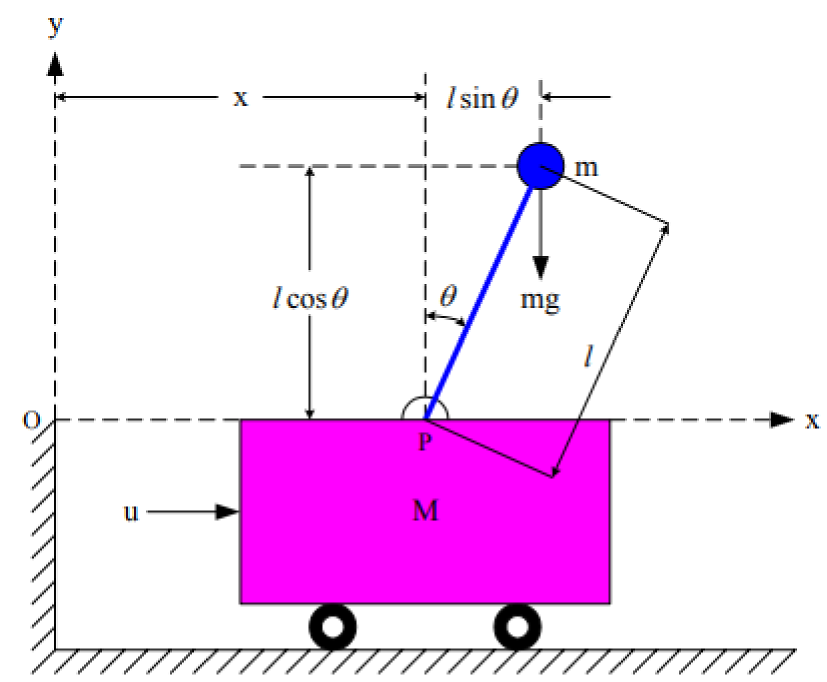
Цель: Стабилизировать систему и минимизировать функционал качества

Весовые матрицы функционала качества:

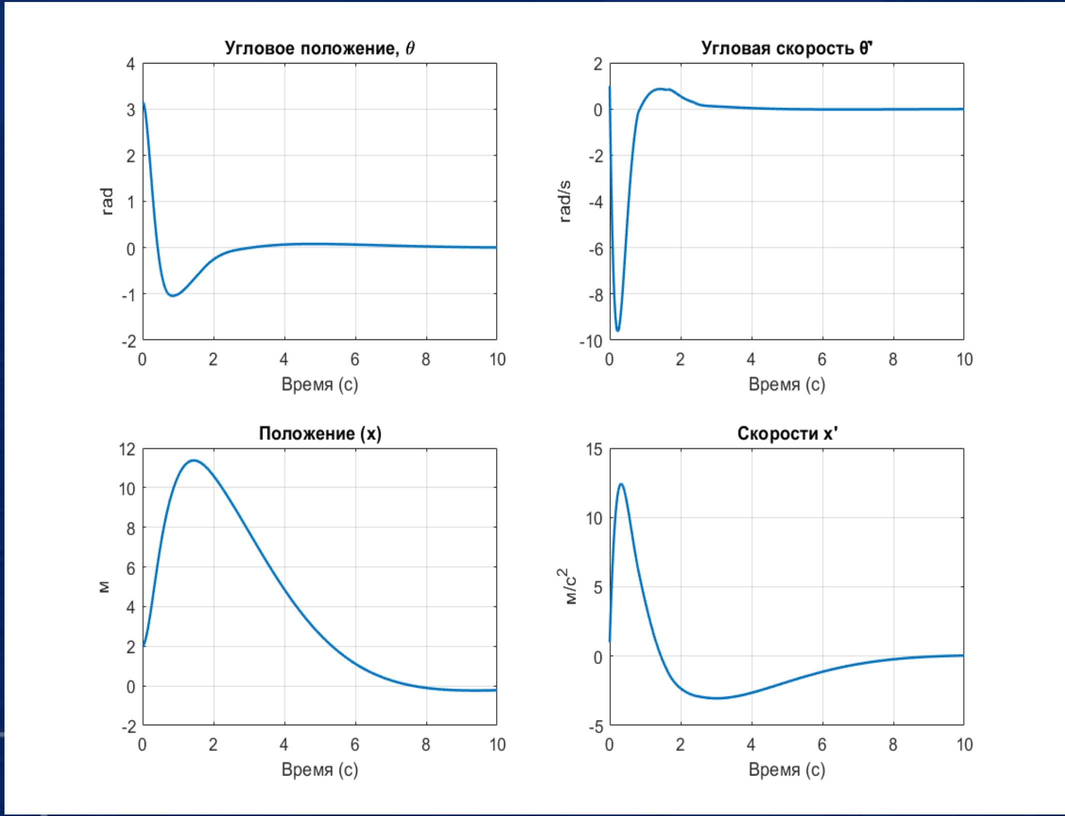
$$Q=\begin{bmatrix}1 & 0.1 & 0 & 0\\ 0.1 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.1\\ 0 & 0 & 0.1 & 1\end{bmatrix}, R=1$$

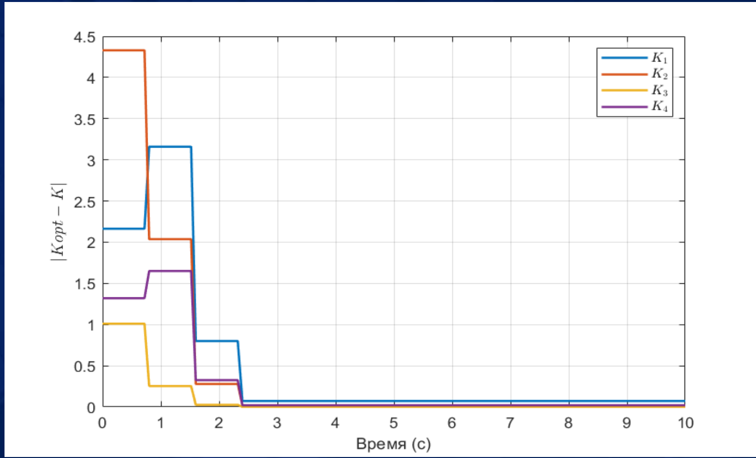
Матрица оптимального регулятора

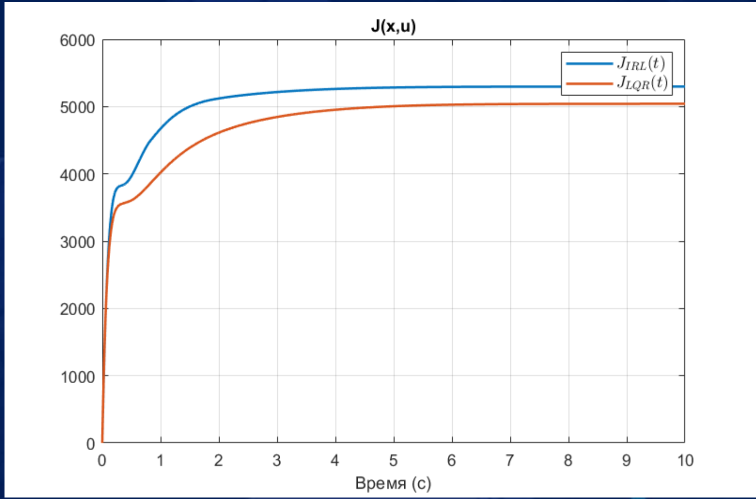
$$\text{K}=[-62.7 \quad -13.1 \quad -1.0 \quad -2.92]$$



Регулятор для перевернутого маятника и тележки







Постановка задачи для нелинейных систем

Динамическая система

Рассмотрим управление нелинейной динамической системой

$$\dot{x} = f(x) + g(x)u$$

Цель управления

Цель - разработать регулятор, который минимизирует функцию ценности

$$J(x,u) = \int_0^{\infty} \left(x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau)\right)d\tau$$

Точное решение

- Функция оптимальной ценности

$$V^*(x) = \min_u \int_0^{\infty} \left(x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau)\right)d\tau$$

Обучение с интегральным подкреплением (IRL)

Функции ценности:

$$V\big(x(t)\big) = \int_t^{\infty} r\big(x(\tau), u(\tau)\big) d\tau$$

$$r(x,u) = Q(x) + u^T R u$$

IRL Уравнение Беллмана:

$$V^u\big(x(t)\big) = \int_t^{t+T} (r(x(\tau), u(\tau))d\tau \; + V^u\big(x(t+T)\big)$$

Интегральное армирование:

$$\rho(x(t), t, T) = \int_t^{t+T} (r(x(\tau), u(\tau))d\tau$$

Итерации по стратегии

Этап оценки стратегии

$$V_{j+1}\big(x(t)\big) = \int_t^{t+T} r\left(x(s), u_j\big(x(s)\big)\right) ds \; + V_{j+1}\big(x(t+T)\big)$$

$$V_{j+1}(0) = 0$$

Этап улучшения стратегии

$$u_{j+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V_{j+1}$$

Применение ADP

$$V(x) = \widehat{W}^T \phi(x)$$

Этап оценки стратегии

$$\widehat{W}_{j+1}^T \left[\phi\big(x(t)\big) - \phi\big(x(t+T)\big)\right] = \int_t^{t+T} r\left(x(s), u_j\big(x(s)\big)\right) ds$$

Этап улучшения стратегии

$$u_{j+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \big(\nabla_x \phi(x)\big)^T \widehat{W}_{j+1}$$

Моделирование для простой нелинейной системы

Объект управления

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix} + \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix} u$$

Функция ценности:

$$J = \int_0^\infty (x^T Q x + u^T R u) dt \qquad (4.4)$$

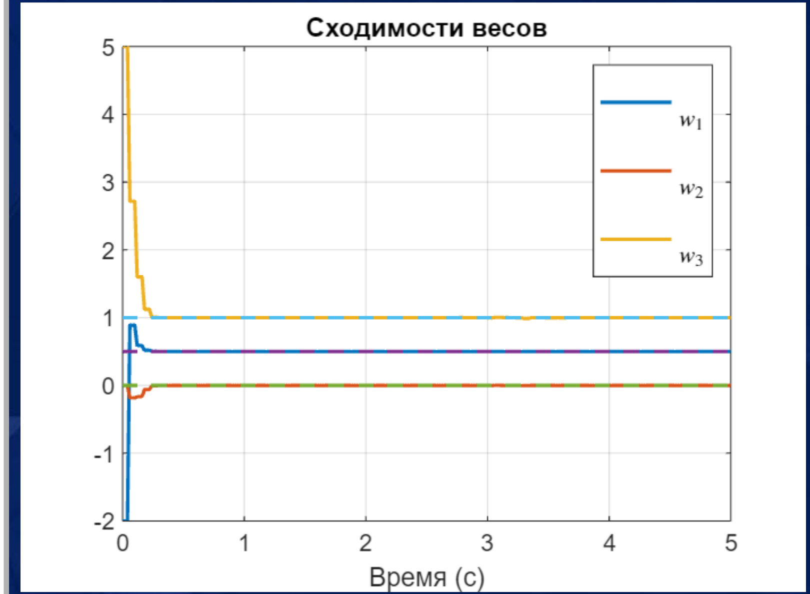
$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = 1$$

Оптимальное значение

$$V = W^T \phi(x)$$

$$V^*(x) = 0.5x_1^2 + x_2^2$$

$$u^*(x) = -(\cos(2x_1) + 2)x_2^2$$



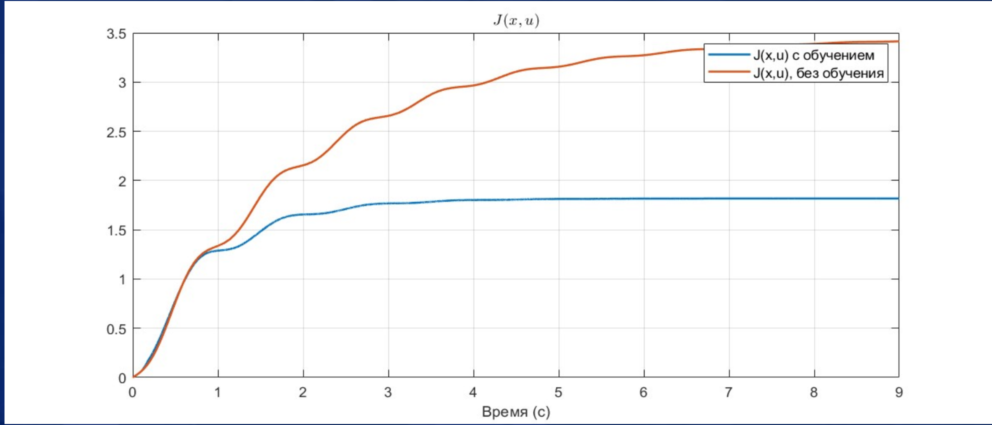
Оптимальный адаптивный регулятор для маятника

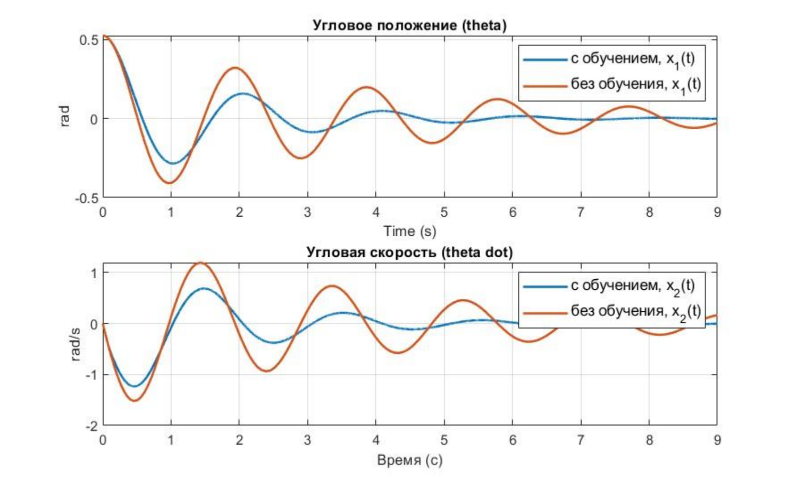
Объект управления:

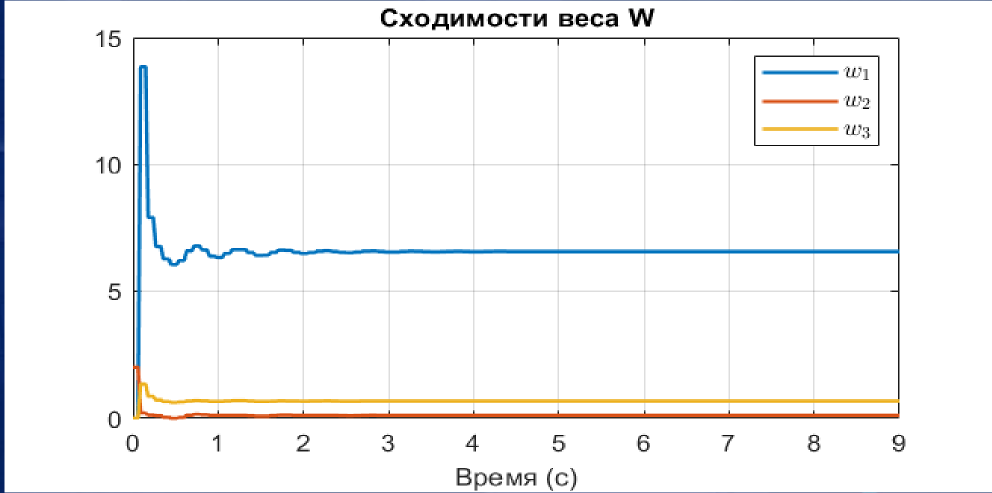
$$\dot{x} = \begin{bmatrix} \dot{\theta} \\ \dfrac{-mglsin(theta) - B\dot{\theta}}{ml^2} \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{ml^2} \end{bmatrix} u$$

Цель: Минимизация функционала качества [здесь система и так была устойчива]

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = 1$$







Заключение

- Рассмотрены классические методы управления, концепция обучения с подкреплением и его основные алгоритмы.
- Математические модели регуляторов на основе обучения с подкреплением разработаны для линейных и нелинейных систем.
- Для линейной системы, моделирование регулятора LQR, адаптивного регулятора и регулятора IRL реализованы и сравнены их эффективности. Было изучено влияние различных параметров T на эффективности регулятора на основе обучении с подкреплением.
- Успешно применил регулятор в линейных системах с перевернутым маятником и тележкой.
- Моделирование регулятора на основе обучения с подкреплением для нелинейных систем. Показана сходимость к оптимальному значению управления. Регулятор применяется для минимизации функции качества нелинейной маятниковой системы

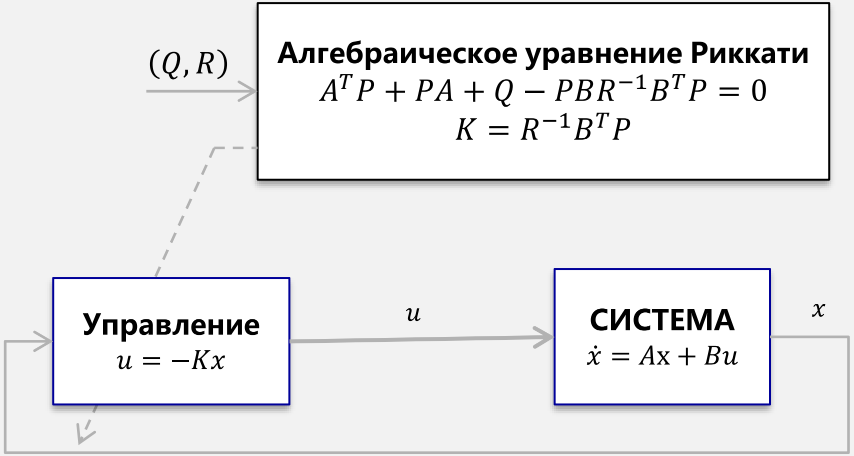
Спасибо
за внимание!

ITsMOre than a UNIVERSITY

dinhngoctuan6789@gmail.com

LQR и адаптивный регулятор

LQR



Цель:

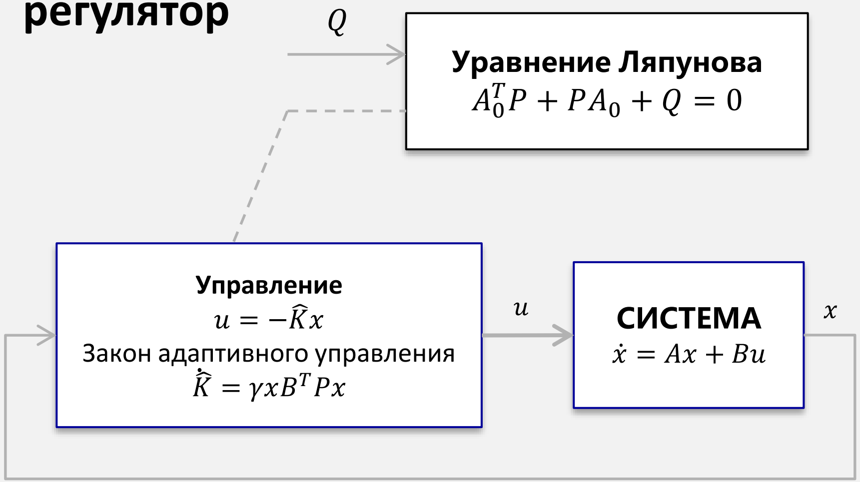
Минимизация

$$J(x) = \int_0^{\infty} \left(x^T(t)Qx(t) + u^T(t)Ru(t)\right)dt$$

Цель: xem xét xóa slide này

$$\lim_{t\to\infty} |x(t)| = 0$$

Адаптивный регулятор



Постановка задачи

Уравнение Гамильтона Якоби Беллмана (HJB)

Уравнение Гамильтона Якоби Беллмана (HJB)

$$0 = (\nabla V_x^*)^T(x)(f(x) + g(x)u) + x^T Q x + u^*(x)^T R u^*(x)$$

Оптимальное управление

Оптимальный регулятор - Из решения уравнения HJB

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V_x^* \qquad (1)$$

- Невозможно решить HJB аналитически.
- Аппроксимация функции значения (V*)
    - Нейронные сети