Применение методов машинного обучения с подкреплением в оптимизации миграционной и экономической политики государства

Марков Игорь

Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра оптимального управления

Москва, 2022

Описание модели

Рассмотрим экономическую модель взаимодействия N стран, производящих продукт Y_i при помощи капитала K_i и трудовых ресурсов L_i.

$$Y_i = A_i K_i^{\alpha_i} L_i^{1-\alpha_i}$$

Введем обозначения:

w_i	$\frac{\partial Y_i}{\partial L_i} = A_i (1-\alpha_i) \left(\frac{K_i}{L_i}\right)^{\alpha_i}$	уровень заработных плат
r_i	$\frac{\partial Y_i}{\partial K_i} = A_i \alpha_i \left(\frac{K_i}{L_i}\right)^{(1-\alpha_i)}$	процентная ставка капитала

Описание модели

$$\dot{K}_i = s_i Y_i - \delta_i K_i + \sum_{j=1, j\neq i}^{N} \tau_{ij} [r_i - r_j]^+ K_j - \sum_{j=1, j\neq i}^{N} \tau_{ji} [r_j - r_i]^+ K_i$$

$$\dot{L}_i = n_i L_i + \sum_{j=1, j\neq i}^{N} \sigma_{ij} \frac{[w_i - w_j]^+}{w_j} L_j - \sum_{j=1, j\neq i}^{N} \sigma_{ji} \frac{[w_j - w_i]^+}{w_i} L_i$$

- s_i — инвестиционная ставка
- δ_i — коэффициент амортизации капитала
- n_i — коэффициент роста населения
- τ_{ij} — коэффициент стремительности перетока капитала
- σ_{ij} — коэффициент стремительности перетока трудовых ресурсов

Описание модели

Введем относительные функции $k_i(t) = \frac{K_i(t)}{K_i(0)}$ и $l_i(t) = \frac{L_i(t)}{L_i(0)}$, а также $\lambda_i^0 = \frac{Y_i(0)}{L_i(0)}$, $\rho_{ij}^0 = \frac{L_i(0)}{L_j(0)}$, $\kappa_i^0 = \frac{Y_i(0)}{K_i(0)}$ и $\pi_{ij}^0 = \frac{K_i(0)}{K_j(0)}$

$$\dot{k_i} = s_i \kappa_i^0 y_i - \delta_i k_i + \sum_{j=1, j\neq i}^{N} \tau_{ij} [r_i - r_j]^+ \pi_{ij}^0 k_j - \sum_{j=1, j\neq i}^{N} \tau_{ji} [r_j - r_i]^+ k_i$$

$$\dot{l_i} = n_i l_i + \sum_{j=1, j\neq i}^{N} \sigma_{ij} \frac{[w_i - w_j]^+}{w_j} \rho_{ij}^0 l_j - \sum_{j=1, j\neq i}^{N} \sigma_{ji} \frac{[w_j - w_i]^+}{w_i} l_i$$

- $y_i(t) = \frac{Y_i(t)}{Y_i(0)} = k_i(t)^{\alpha_i} l_i(t)^{1-\alpha_i}$
- $w_i(t) = \lambda_i^0 (1 - \alpha_i) \left(\frac{k_i(t)}{l_i(t)} \right)^{\alpha_i}$
- $r_i(t) = \kappa_i^0 \alpha_i \left(\frac{l_i(t)}{k_i(t)} \right)^{1-\alpha_i}$
- $[x]^+ = \frac{x + |x|}{2}$

Вместе с начальными условиями $k_i(0) = 1$ и $l_i(0) = 1$ данные соотношения описывают динамику исследуемой системы.

Функция полезности

Рассмотрим функцию полезности потребления:

$$U_i = \int_0^{\infty} e^{-d_i t} L_i(t) \ln\left(\frac{(1-s_i)Y_i}{L_i}\right) dt =$$

$$= L_i(0) \int_0^{\infty} e^{-d_i t} l_i(t) \ln\left(\frac{(1-s_i)\lambda_i^0 y_i}{l_i}\right) dt$$

- d_i — коэффициент дисконтирования

i-ая страна стремится максимизировать функционал U_i, управляя параметрами $\tau_{ij}, \tau_{ji}, \sigma_{ij}, \sigma_{ji}$.
Введем также следующий функционал:

$$\hat{U}_{i,t_0} = \int_{t_0}^{t_0+1} l_i(t) \ln\left(\frac{(1-s_i)\lambda_i^0 y_i}{l_i}\right) dt$$

Постановка задачи обучения с подкреплением

Определим среду:

- Агенты — страны
- K шагов. Далее симуляция прекращается. На каждом шаге пересчитывается t_0, а также $I(0)$ и $k(0)$
- $\tau_{ij} = \frac{\tau_{ij}^i + \tau_{ij}^j}{2}$
  $\sigma_{ij} = \frac{\sigma_{ij}^i + \sigma_{ij}^j}{2}$
- Состояние — набор всех меняющихся параметров: $\{t_0, k(0), I(0), \tau, \sigma\}$
- Действие i-ого агента — изменение параметров $\tau_{ij}^i, \tau_{ji}^i, \sigma_{ij}^i, \sigma_{ji}^i$ не более, чем на один шаг по дискретной сетке значений (-1, 0, либо 1 для каждого параметра)
- Награда i-го агента в момент времени t_0: $\frac{\hat{U}_{i,t_0} - \hat{U}_{i,t_0-1}}{\hat{U}_{i,t_0}}$

Метод решения

При решении задачи использовался алгоритм Deep Q-Learning с использованием Replay Buffer и Target Network.
Каждый агент использовал следующую архитектуру нейронной сети:

$$Observation \to Linear(StateDim, 64) \to ReLU() \to$$

$$\to Linear(64, 128) \to ReLU() \to$$

$$\to Linear(128, NumberOfActions) \to QValues$$

В качестве базовых сценарий для сравнения были также реализованы агенты, придерживающиеся наивных стратегий: безусловного открытия границ (далее, OBA — Opened Borders Agent), безусловного закрытия границ (далее, CBA — Closed Borders Agent) и случайной стратегии (далее, RA — Random Agent).

Результаты. Обучение обоих агентов

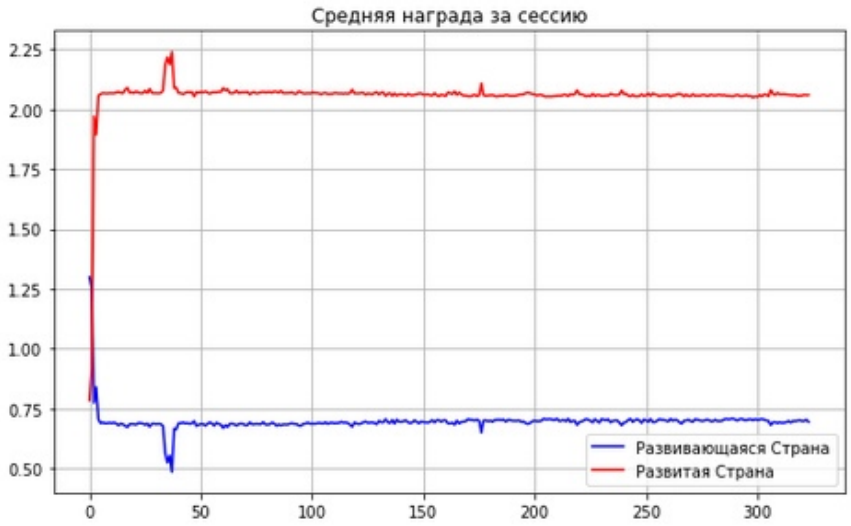


Рис.: Оба агента быстро сходятся к оптимальным политикам

Результаты. Обучение только развивающейся страны

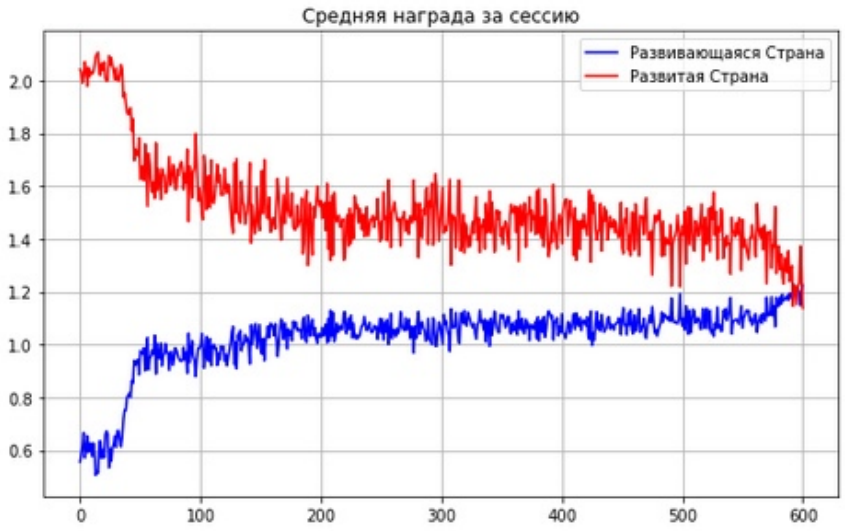


Рис.: Развитой стране удается добиться существенного улучшения

Результаты. Обучение только развитой страны

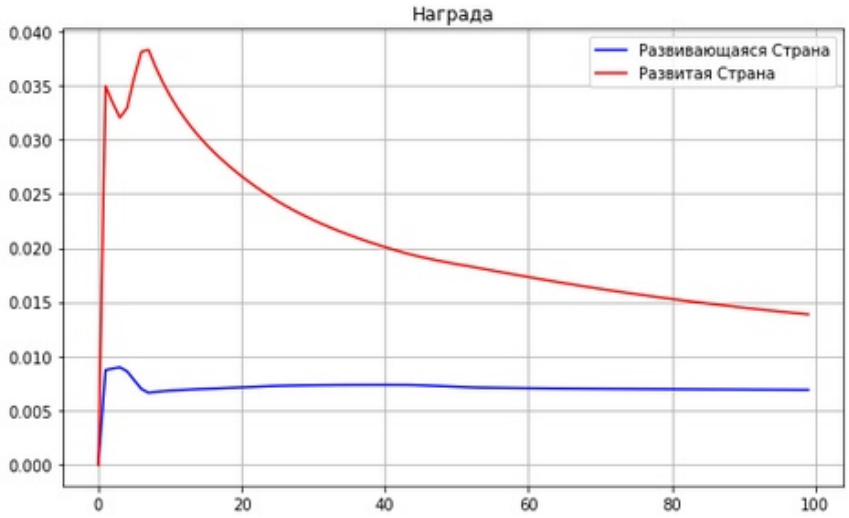
Рис.: Развитая страна полностью доминирует над развивающейся

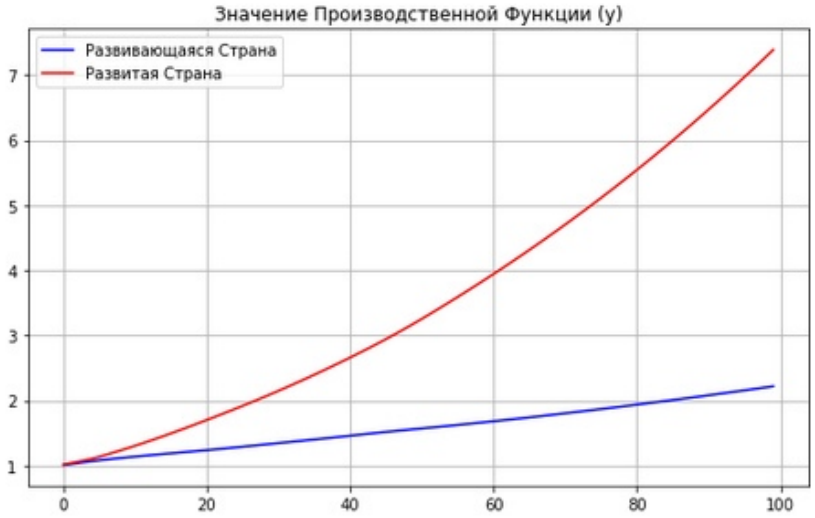
Результаты. Валидация

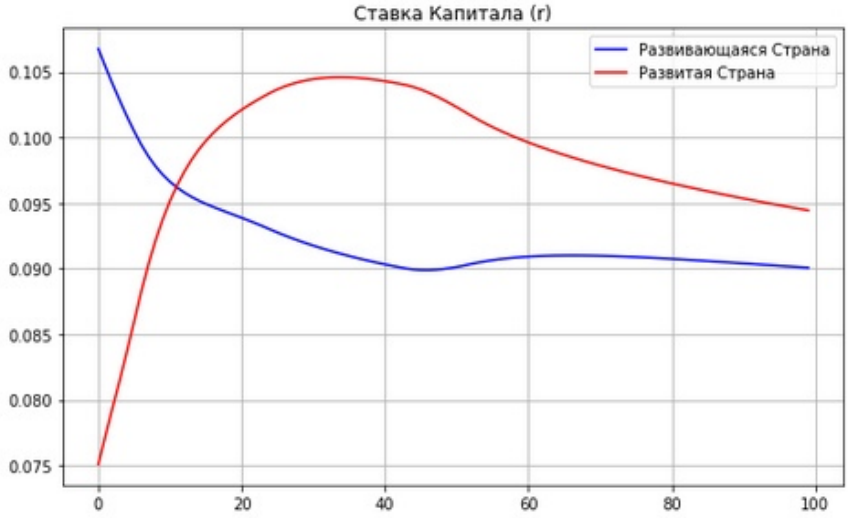
Ниже приведены результаты валидации в виде таблицы.
Строка — агент-развивающаяся страна, столбец — агент-развитая страна.
Значения в ячейках — средняя награда первого и второго агента соответственно, полученная за 100 симуляций среды, нормированная на количество шагов в каждой симуляции, умноженная на 100.

	RA	OBA	CBA	DQNA
RA	0.660 2.047	0.363 2.228	1.041 1.629	0.363 2.229
OBA	0.375 2.223	0.16 2.333	0.687 2.067	0.012 2.335
CBA	1.027 1.669	0.683 2.065	1.391 0.503	0.688 2.069
DQNA	1.033 1.651	0.688 2.073	1.392 0.502	0.700 2.074

Результаты







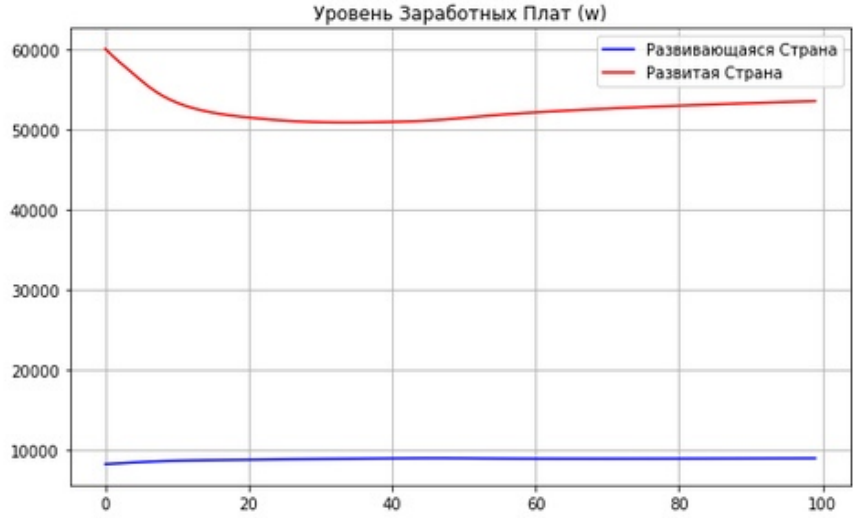
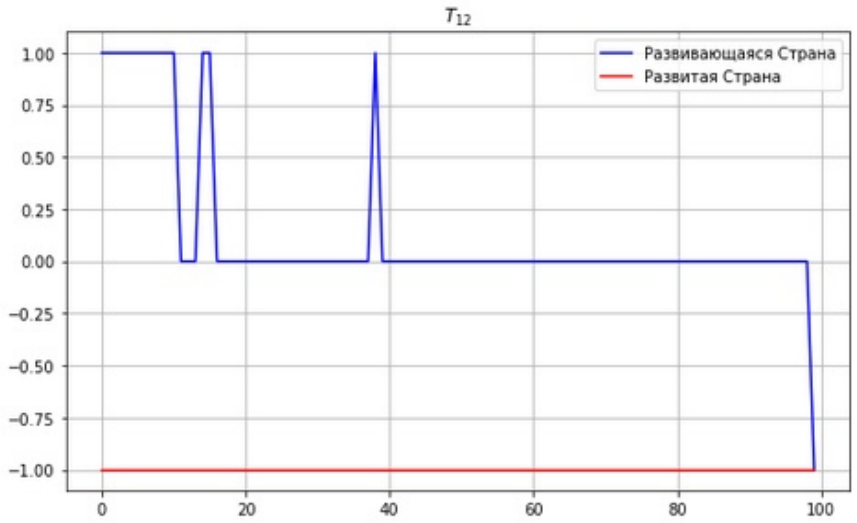
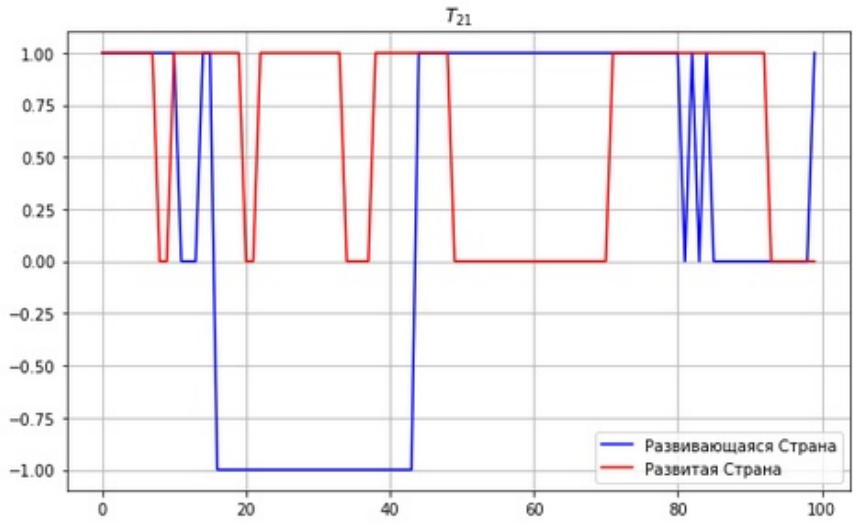
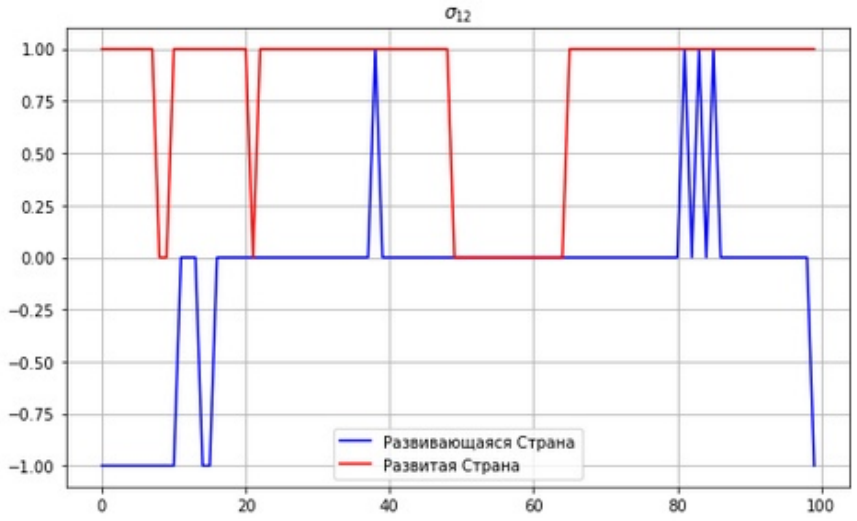


Рис.: Значение награды и экономических показателей от времени

Результаты







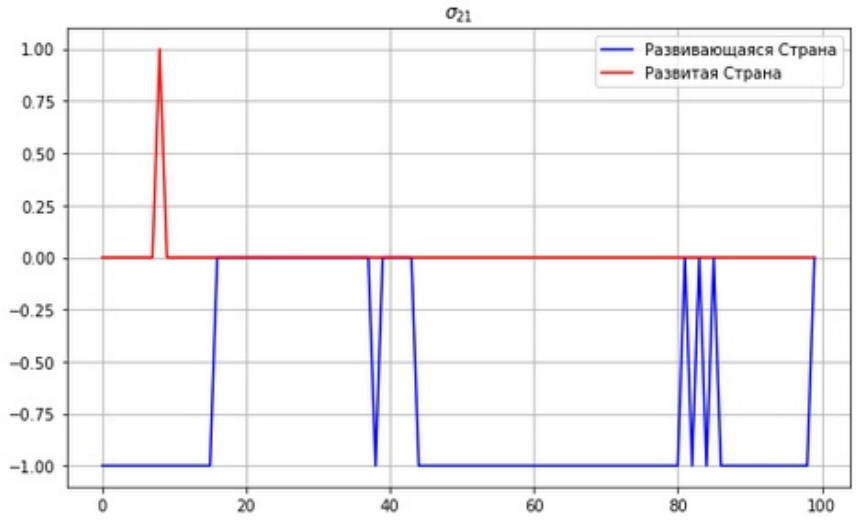


Рис.: Действия обученных агентов по каждому параметру от времени

Результаты

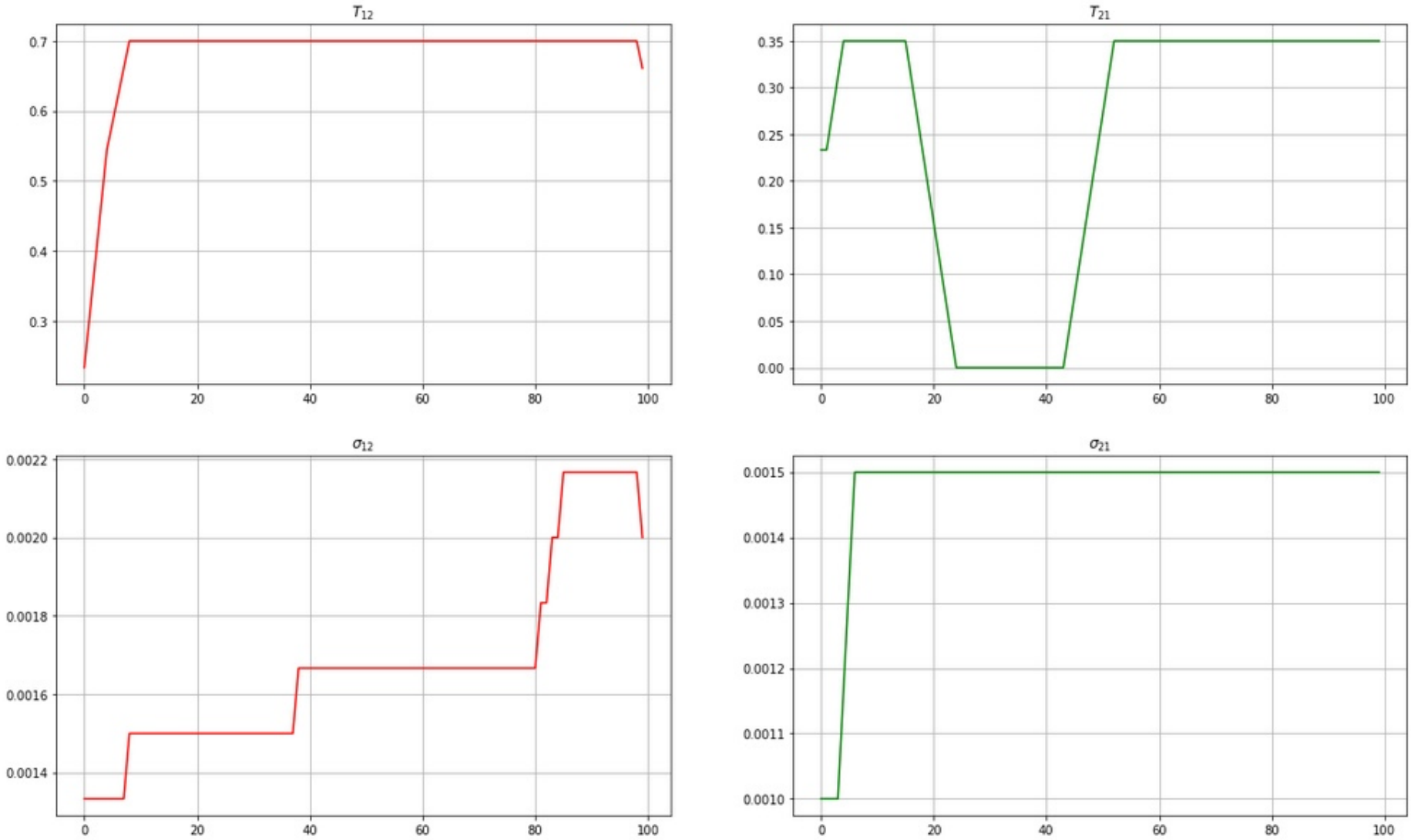


Рис.: Значения управляемых параметров от времени

Выводы

- Обучение происходит успешно. Политики агентов сходятся и показывают лучший результат, чем наивные стратегии (безусловное открытие/закрытие границ, случайное изменение параметров)
- Рост награды одного агента имеет отрицательную корреляцию с ростом награды второго
- Развивающаяся страна предпочитает закрывать свои границы, а развитая — открывать
- Полезность потребления и производимый продукт обеих стран растут со временем

Дальнейшие направления исследований

- Симуляция среды с $N > 2$ странами
- Варьирование стационарных параметров среды
- Введение в награду штрафа за изменение параметров
- Реализация «универсального» агента (все переменные среды изменчивы)
- Введение в среду социального планировщика — агента, оптимизирующего общее благосостояние системы

Спасибо за внимание!