

# ЧИСЛЕННЫЙ АНАЛИЗ МОДЕЛИ ОПТИМАЛЬНОГО РАСШИРЕНИЯ ПРОИЗВОДСТВА С ФАЗОВЫМ ОГРАНИЧЕНИЕМ И ПРИМЕНЕНИЕМ ТЕХНОЛОГИИ МАШИННОГО ОБУЧЕНИЯ<sup>1</sup>

Жукова А.А., Флёрова А.Ю.

Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия  
zhukova.aa@phystech.edu, a.flerova@mail.ru

Евтухов А.Д.

Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия,  
Московский физико-технический институт (национальный исследовательский университет), Москва, Россия  
evtukhov.ad@phystech.edu

*Аннотация.* В данной работе рассматривается подход к решению задачи о максимизации прибыли компании, имеющей потенциал масштабирования производства в условиях рыночных ограничений, с помощью задачи оптимального управления. Данная задача оптимального управления имеет решение в простом детерминированном случае, когда прогнозируема отдача от вложений в расширение производства. Однако, очевидно, нельзя предполагать этот параметр достоверно предсказуемым, т.к. на эффективность вложений оказывают влияние внешние изменчивые факторы. Такие задачи управления сложны для аналитического исследования. В работе рассмотрен подход к решению задачи о расширении производства в случае непостоянного параметра, описывающего увеличение выручки предприятия в зависимости от вложенных средств, с помощью подхода машинного обучения.

*Ключевые слова:* оптимальное управление, расширение производства, инвестиции.

## Введение

Одним из ключевых показателей эффективности компании в условиях рыночной экономики является прибыль и компания принимает решение о распределении ресурсов таким образом, чтобы прибыль достигала максимума возможного. Задача максимизации прибыли, как правило, является предметом динамического анализа на определенном горизонте планирования. Технически, задача описывается задачей динамической оптимизации или оптимального управления. В данной работе рассматривается подход к решению задачи расширения производства с использованием методов динамического программирования в рамках подхода обучения с подкреплением. Цель работы состоит в выборе стратегии фирмы на среднесрочную перспективу: как использовать ресурсы для максимизации прибыли в течение определенного периода, с учетом внешних и внутренних факторов развития. Такие задачи управления часто встречаются в моделировании проблем менеджмента, и сложны для математического анализа [1]. Чтобы сделать задачу более реалистичной, мы рассматриваем случай рыночного ограничения и нестабильной экономической ситуации. Мы также ставим целью данной работы изучить применимость методов машинного обучения как способа исследования задач оптимального управления, в частности обучения с подкреплением (RL). Это может служить поддержкой для принятия решений на основе математических моделей управления, возникающих в экономике. Чтобы подтвердить или опровергнуть эту возможность, мы сравниваем решение задачи в упрощенной версии модели, когда решение можно найти аналитически, и оценку, полученную с помощью алгоритмов RL.

## 1. Модель

В данной работе рассматривается фирма, которая производит некоторый вид продукции (либо исчисляет показатели выпуска в денежном выражении) и планирует масштабирование производства за определенный период времени. Фирма не привлекает внешних инвестиций, хотя такая постановка тоже представляет отдельный интерес [2], а может использовать часть своей выручки для масштабирования производства, т.е. роста. Цель производителя состоит в максимизации благосостояния владельцев компании, что выражается в интегральном объеме выплат дивидендов. В модель могут быть введены параметры дисконтирования, инфляции и др. важных для инвесторов факторов [3-5]. Но в данной работе делается фокус на базовой постановке и выборе метода

---

<sup>1</sup> Данное исследование выполнено при поддержке гранта Российского научного фонда № 24-21-00494, <https://rscf.ru/project/24-21-00494/>

исследования, дополняя ранее предпринятые попытки [6] новой постановкой задачи с учетом внешнего (фазового) ограничения. Хотя, на первый взгляд, введенное ограничение имеет простую структуру, задачи оптимального управления с фазовыми ограничениями являются технически значительно более трудными для анализа [7].

Обозначим через  $x(t)$  выручку производителя в текущий момент времени. Фирма максимизирует свой доход за период времени  $[0, T]$ . Издержки линейно зависят от выручки, коэффициент издержек обозначим через  $1 - \mu$ . Предполагается, что производитель имеет возможность увеличивать свою выручку путем расширения производства. Обозначим часть выручки, которая инвестируется в расширение производства, через  $u(t)$ . Отдача от инвестиций описывается коэффициентом  $\alpha$ , который в детерминированной версии является положительной константой, а в более интересном случае принимает несколько значений с соответствующими вероятностями.

Формулировка базовой версии модели масштабирования производства дополнена ограничением сверху на выручку  $x(t) \leq x_{max}$ . Можно считать, что версия модели в [6] имела данное ограничение с уровнем  $x_{max}$  выше любого возможного уровня роста выручки. Здесь же мы рассматриваем случаи существенного ограничения сверху, когда при интенсивном росте фирма потенциально не может выйти за возможности потребления рынком, квот торговли, физических ограничений и т.п. ограничений. Таким образом, получаем следующую задачу оптимального управления.

$$\int_0^T (\mu - u(t))x(t)dt \rightarrow \max$$

$$\frac{dx(t)}{dt} = \alpha x(t)u(t), \quad 0 \leq x(t) \leq x_{max}, \quad 0 \leq u(t) \leq \mu, \quad x(0) = x_0.$$

В детерминированной версии модели параметр роста выручки  $\alpha$  постоянен и решение задачи сходно с решением задачи без ограничения  $x(t) \leq x_{max}$ , но только в режиме роста до момента, когда ограничение становится активно. При масштабировании (росте в режиме  $u(t) = \mu$ ) по мере достижения выручкой уровня  $x_{max}$  управление перестает влиять на дальнейший рост и траектория упирается в ограничение. Однако, мы не знаем решение в стохастическом случае, попробуем найти его численно.

## 2. Численный анализ решения

### 2.1. Алгоритм

Для численного решения данной задачи мы используем алгоритм Deep Deterministic Policy Gradient (DDPG) [6]. То есть мы решаем классическую задачу обучения с подкреплением, где агент взаимодействует со средой в дискретном времени. Состояние агента обозначим  $s_t$  и считаем, что оно содержит всю информацию для текущего состояния управления и параметров состояния. В каждый момент времени агент выбирает действие  $a_t$  на основе текущего состояния  $s_t$ , и получает за это награду  $r(s_t, a_t)$  после чего переходит в состояние  $s_{t+1}$ . Суммарная награда за будущие шаги:  $R_t(s_t, a_t) = \sum_{i=t}^T \gamma^{i-t} r(s_i, a_i)$ . Цель агента – максимизация общей награды  $R_0(s_0, a_0)$  выбирая политику  $\pi$ , которая определяет действие для каждого состояния.

Во многих алгоритмах обучения с подкреплением используется функция ценности действия, которая описывает ожидаемый доход после выполнения действия  $a_t$ , получения награды  $r_t$  и перехода в состояние  $s_{t+1}$  и последующее следование выбранной политики  $\pi$ :  $Q^\pi(s_t, a_t) = E_\pi[R_t | s_t, a_t]$ .

Функция ценности действия может быть использована в уравнении Беллмана:  $Q^\pi(s_t, a_t) = E_\pi[r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))]$ .

Алгоритм основан на методе «актёр-критик» [8], то есть алгоритм использует две нейронные сети: актёр в каждом состоянии определяет какое действие стоит выбрать, а критик оценивает выбранное действие. Получается, что идея данного подхода в том, что критик оценивает функцию ценности, а актёр обновляет распределение политики в направлении, предложенном критиком. Критик обновляется с использованием уравнения Беллмана, минимизируя потери. Актёр же обновляется для максимизации ожидаемой общей награды.

В начале обучения политика может быть далека от оптимальной, поэтому для эффективного поиска оптимальной политики во время обучения к действиям может добавляться шум. В данной работе используется нормальный шум с нулевым средним значением.

Стоит отметить, что DDPG – это алгоритм, который может использовать историю предыдущего обучения или другие внешние данные для последующего обучения. Поэтому реализация алгоритма в

данной работе использует буфер воспроизведения для хранения предыдущей истории и её использования в будущих шагах. Также используется идея целевых сетей [8], обеспечивающих сходимость сетей критика.

Так как рассматриваемая задача в данной работе имеет ограничение на допустимую траекторию, то используется метод штрафных функций [2], который усложняет вид целевого функционала.

Теперь давайте перейдём к основным параметрам алгоритма для решения нашей задачи (таблица 1). Сразу будут указаны базовые значения для которых проводились эксперименты.

Таблица 1. Расчетные параметры модели

Название	Обозначение	Значение
Первый коэффициент роста	$\alpha_1$	1.0
Второй коэффициент роста	$\alpha_2$	0.3
Вероятность	$p$	0.5
Доля расходов на производство	$\beta$	0.25
Налоговая ставка	$b$	0.2
Временной горизонт	$T$	10
Шаг по времени	$dt$	1.0
Начальная выручка	$x_0$	8.0
Ограничение на выручку	$x_{max}$	50.0
Коэффициент штрафной функции	$\lambda$	0.5

## 2.2. Проверка качества работы алгоритма в близкой к детерминированной модели

Чтобы посмотреть поведение решение, полученного нашим алгоритмом, в ситуации близкой к детерминированной модели, мы должны устремить параметр вероятности к единице (или к нулю). Тогда с большой вероятностью, коэффициент роста будет постоянным.

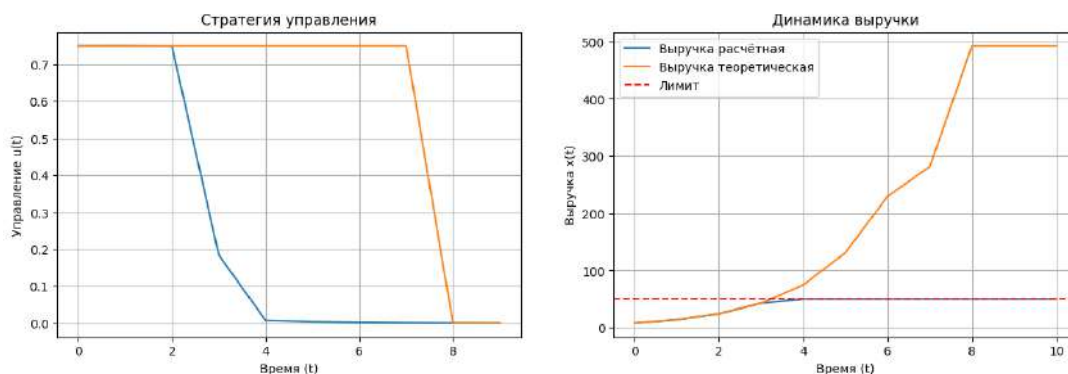


Рис. 1. Зависимость выручки от времени при вероятности  $p = 0.9$

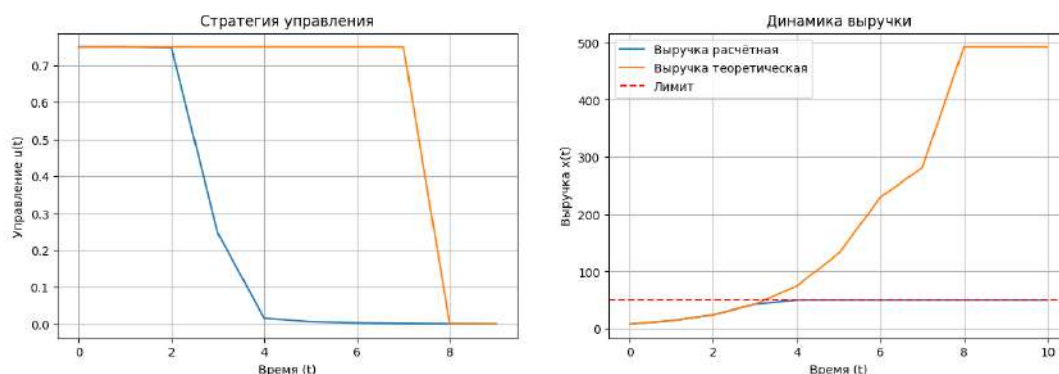


Рис. 2. Зависимость выручки от времени при вероятности  $p = 0.93$

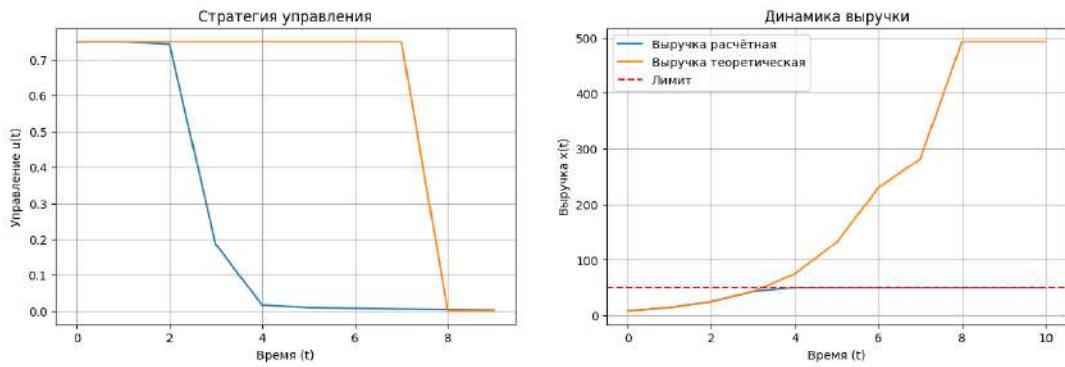


Рис. 3. Зависимость выручки от времени при вероятности  $p = 0.95$

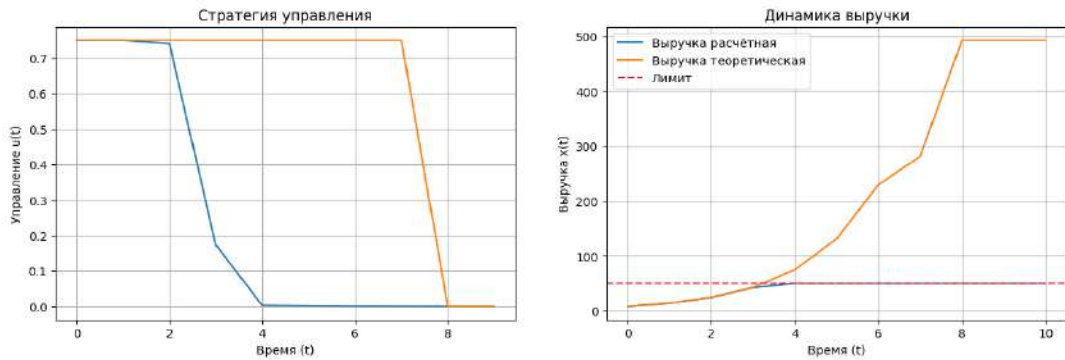


Рис. 4. Зависимость выручки от времени при вероятности  $p = 0.97$

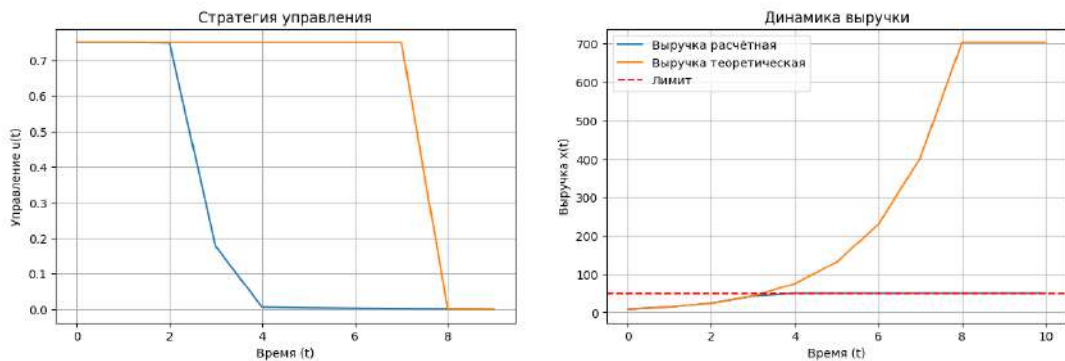


Рис. 5. Зависимость выручки от времени при вероятности  $p = 0.99$

На рисунках 1, 2, 3, 4 и 5 можно видеть, что стратегии управления и траектории выручки слабо изменяются (в пределах погрешности вычислений алгоритма), то есть это говорит о том, что так как с вероятностью, близкой к единице, наша задача стала детерминированной. Также хочется отметить, что подтвердилась гипотеза о том, что расчётная траектория в начале должна совпасть с теоретической, а потом если «упирается» в ограничение, то остаться на нём.

### 2.3. Эксперименты.

#### 2.3.1. Зависимость оптимального управления от ограничения на выручку

Считаем, что во всех дальнейших экспериментах  $p = 0.5$ .

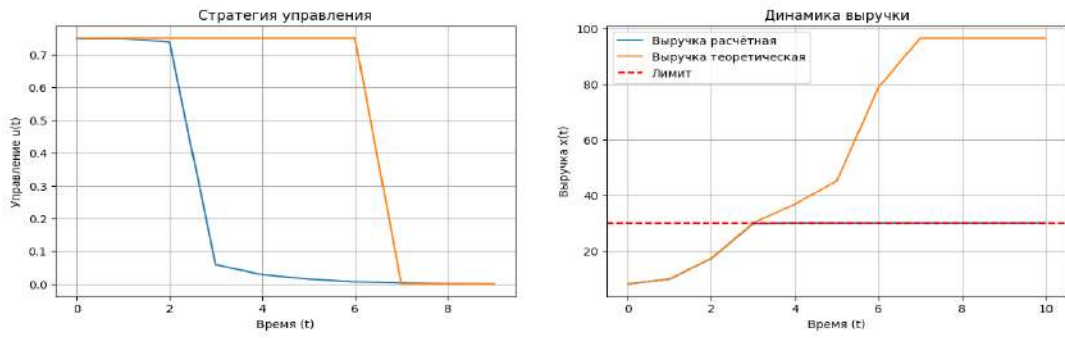


Рис. 6. Зависимость выручки от времени при ограничении на выручку  $x_{max} = 30$

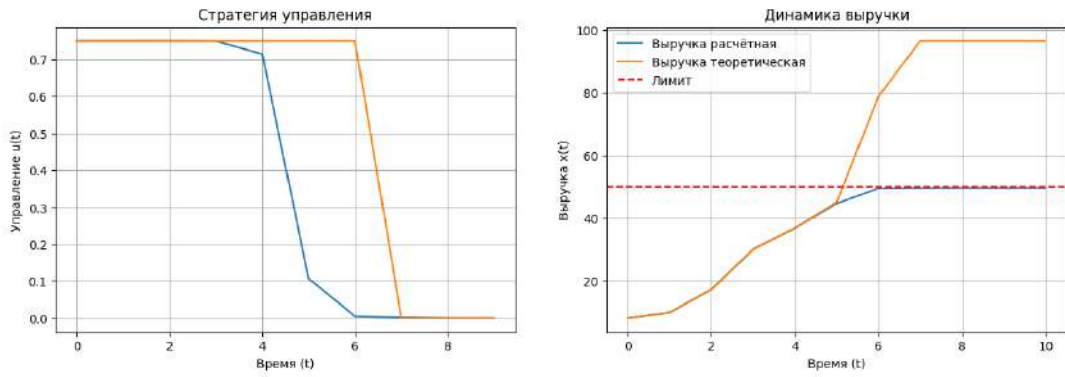


Рис. 7. Зависимость выручки от времени при ограничении на выручку  $x_{max} = 50$

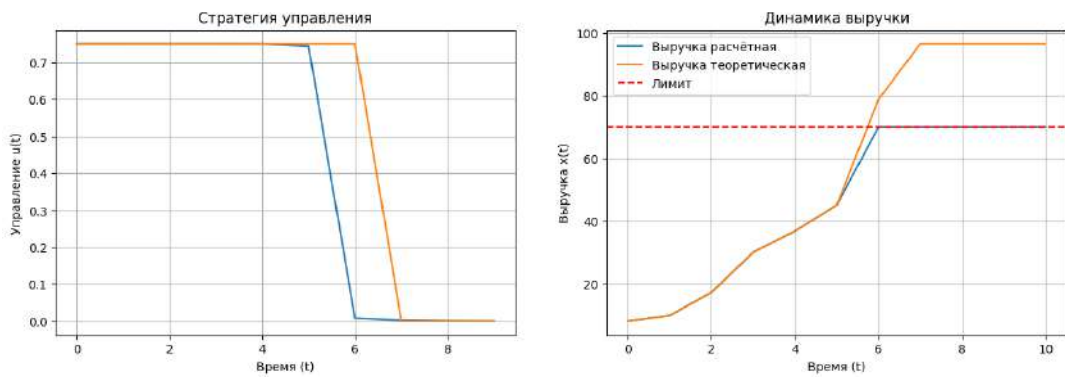


Рис. 8. Зависимость выручки от времени при ограничении на выручку  $x_{max} = 70$

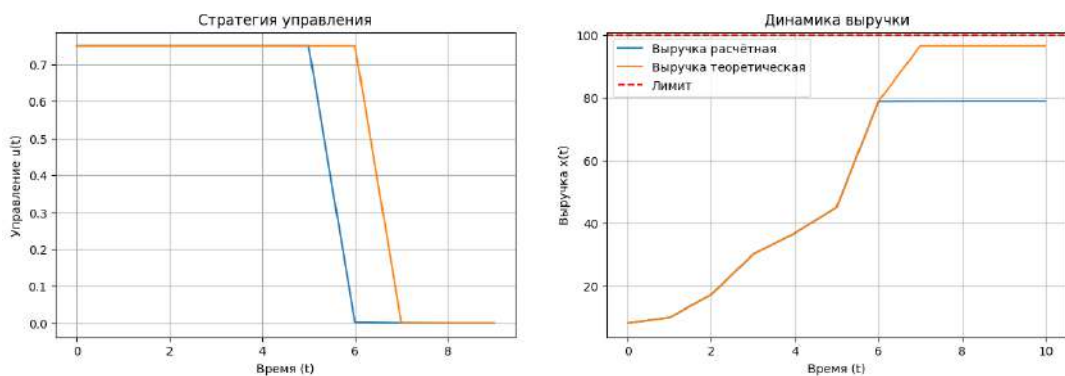


Рис. 9. Зависимость выручки от времени при ограничении на выручку  $x_{max} = 100$

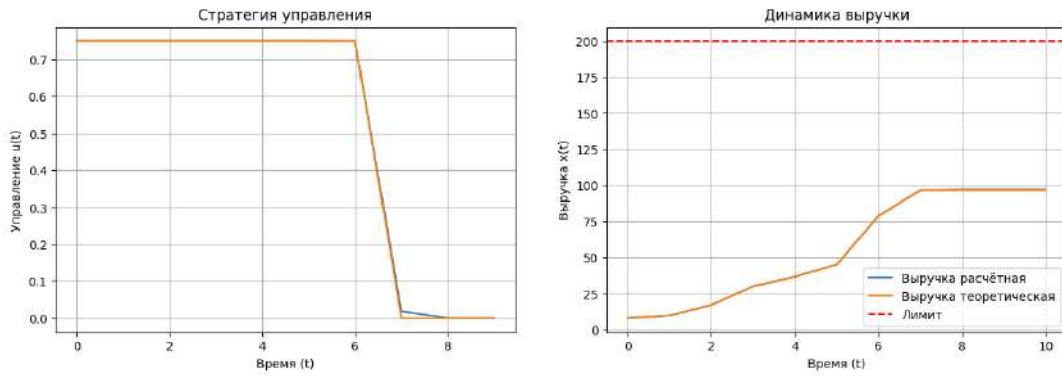


Рис. 10. Зависимость выручки от времени при ограничении на выручку  $x_{max} = 200$

На рисунках 6, 7 и 8 можно видеть, что ограничение на выручку меньше максимального значения выручки в теоретическом расчёте, поэтому расчётная траектория должна «выйти на плато» раньше теоретической, но при этом в начальные моменты времени (до ограничения) они совпадают. Когда же ограничение больше максимального значения, то видим интересный эффект: на рисунке 9 траектория почему-то «выходит на плато» раньше, а на рисунке 10 – траектории полностью совпадают. Есть предположение, что это связано с функцией штрафа и «обрубанием» агента в модели машинного обучения до попадания в нужное множество (то есть проекцией на множество). Поэтому данный эффект стоит подробнее изучить в дальнейшем.

### 2.3.2. Зависимость оптимального управления от ограничения на управление

Считаем, что во всех дальнейших экспериментах  $x_{max} = 50$ .

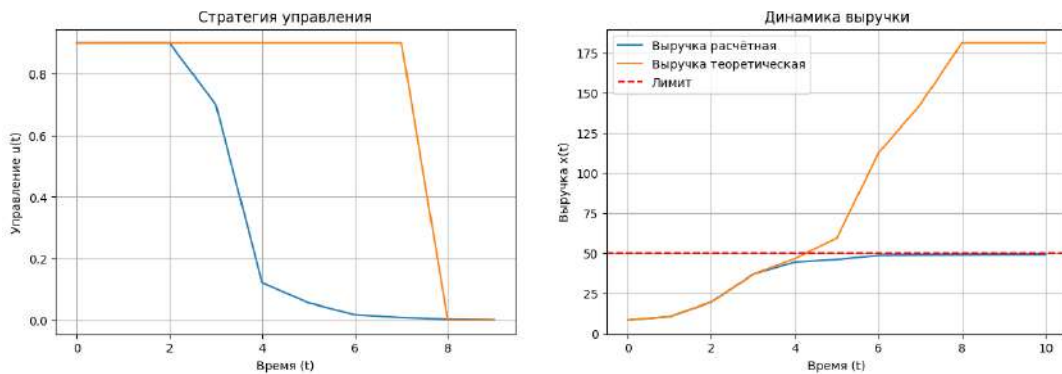


Рис. 11. Зависимость выручки от времени при ограничении на управление  $\beta = 0.1$

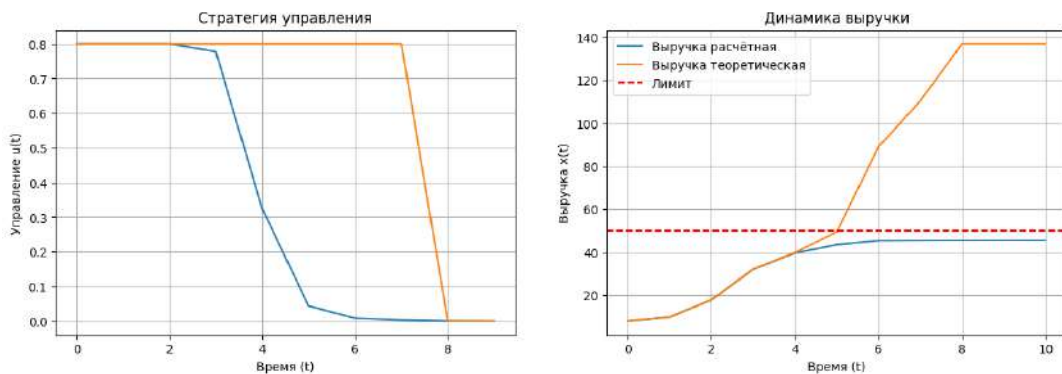


Рис. 12. Зависимость выручки от времени при ограничении на управление  $\beta = 0.2$

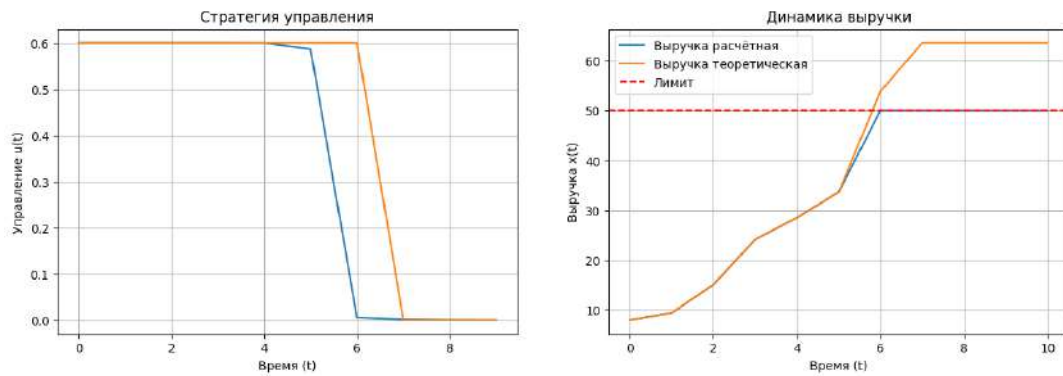


Рис. 13. Зависимость выручки от времени при ограничении на управление  $\beta = 0.4$

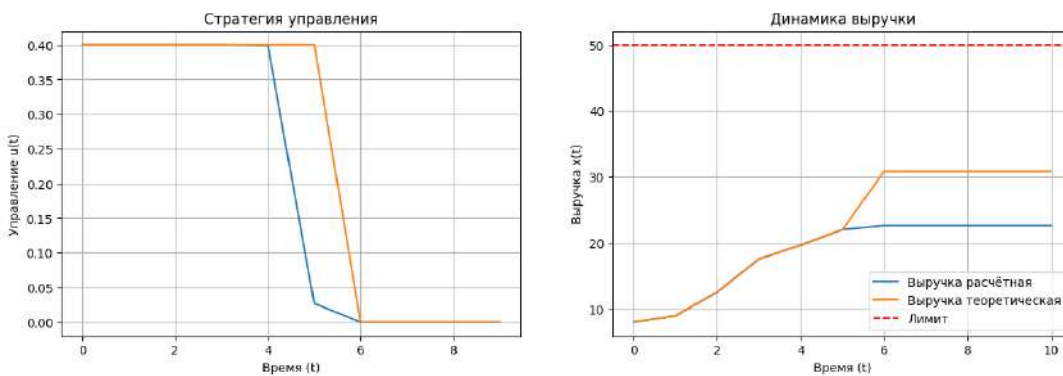


Рис. 14. Зависимость выручки от времени при ограничении на управление  $\beta = 0.6$

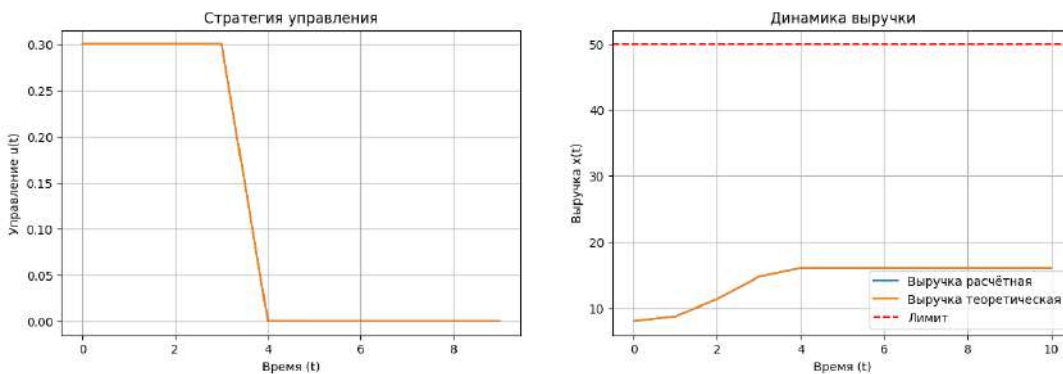


Рис. 15. Зависимость выручки от времени при ограничении на управление  $\beta = 0.7$

На рисунках 11, 12, 13, 14 и 15 можем видеть, что полученные результаты напоминают графики в предыдущем пункте, но стоит обратить внимание на то, что теперь ещё изменяется теоретическая выручка. Появился интересный эффект на рисунке 12: расчётная траектория должна бы «ложиться на плато», но она показывает меньшее значение. Совпадающие ситуации можно наблюдать на рисунках 9 и 14, когда ограничение строго выше теоретического максимума, но расчётная траектория не полностью совпадает с теоретической. На рисунке 15 получили полное совпадение траекторий.

### 3. Заключение

В работе проанализированы возможности применения технологий машинного обучения к задаче масштабирования производства при нестабильных внешних условиях. Реализованный алгоритм продемонстрировал верное поведение в детерминированной задаче, и в стохастических случаях качественно показывает решение почти совпадающее с теоретическим (рост и дальнейшее нахождение на плато). Отличия расчётного и теоретического решения могут быть сокрыты в методе выбора возврата траектории в разрешённое состояние или в параметрах алгоритма машинного обучения, что требует дальнейшего изучения и анализа.

В дальнейшем исследование предполагается усложнить дополнением указанных выше факторов инфляции, дисконтирования будущего, роста издержек и других немаловажных факторов. Следует отметить, что понимание поведения и возможность моделирования решений производственных фирм важны при разработке механизмов стимулирования развития стратегически важных отраслей и мер поддержки бизнеса.

## Литература

1. Zhukova A., Flerova A., Chernov A., Milyuchikhin G. Approaches to numerical analysis of optimal control with linear phase constraints on the example of the assets and liabilities management by a bank // Journal of Computational and Applied Mathematics. – 2025. – V. 453. – P. 116130.
2. Flerova A., Rybkina E., Zhukova A. Numerical analysis of a model of optimal production expansion with external financing // 2024 17th International Conference on Management of Large-Scale System Development (MLSD). – IEEE, 2024. – P. 1–5.
3. Flerova A., Zhukova A. Production expansion in presence of a bank account // Conference Proceedings: 2022 15th International Conference Management of large-scale system development (MLSD). – IEEE, 2022. – P. 1–4.
4. Flerova A., Zhukova A. Analysis of the model of optimal expansion of a firm // Advances in Optimization and Applications. OPTIMA 2022. Communications in Computer and Information Science. – V. 1739. – Cham: Springer Nature Switzerland, 2022. – P. 109–123.
5. Flerova A., Zhukova A. The Role of Inflation and Time Discounting in Production Expansion // 2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), – IEEE, 2022. – C. 245–250.
6. Delev A., Flerova A., Zhukova A. Application of Machine Learning in the producer's optimal control problem with non-stable demand // 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT). – IEEE, 2022. – P. 867–871.
7. Chertovskih R., Karamzin D., Khalil N.T., Pereira F.L. An indirect method for regular state-constrained optimal control problems in flow fields // IEEE Transactions on Automatic Control. – 2020. – V. 66. – №. 2. – P. 787–793.
8. Mnih V., Kavukcuoglu K., Silver D., Graves A., Antonoglou I., Wierstra D., & Riedmiller M. 2013. Playing Atari with Deep Reinforcement Learning. <http://arxiv.org/abs/1312.5602>.