Математическая модель эффекта обратной связи в системах искусственного интеллекта

Андрей Сергеевич Веприков
Научный руководитель: д.ф.-м.н. А. С. Хританков

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.04.01 Прикладные математика и физика

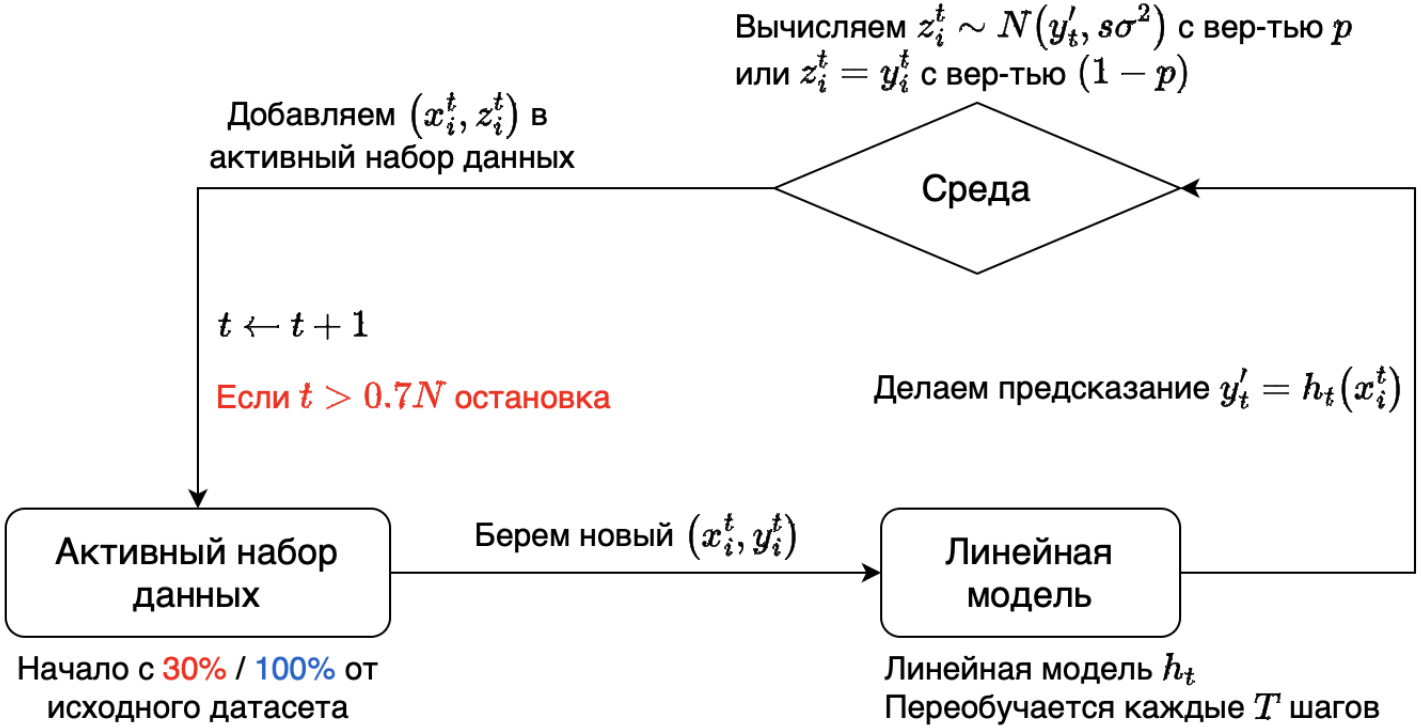
2024

Примеры эффектов петель обратной связи

1. Самоисполняющееся пророчество (self-fulfilling prophecy)
2. Вынужденное смещение данных (data drift) в рекомендательных системах [Khritankov, 2023]
3. Усиление ошибок (error amplification) со временем в задаче медицинского прогнозирования [Adam et al., 2022]
4. Дрейф данных в системах предиктивного полицейского контроля [Ensign et al., 2018]
5. Пузыри фильтров (filter bubbles) [Davies, 2018]

Математическая модель эффекта обратной связи в системах искусственного интеллекта

В данной работе решается задача математического моделирования систем с адаптивным управлением. Системе с ИИ ставится в соответствие дискретная динамическая система, по поведению которой можно судить об исходном объекте.



Две постановки эксперимента. Постановки скользящее окно и обновление выборки.

Постановка задачи и теорема о предельном множестве

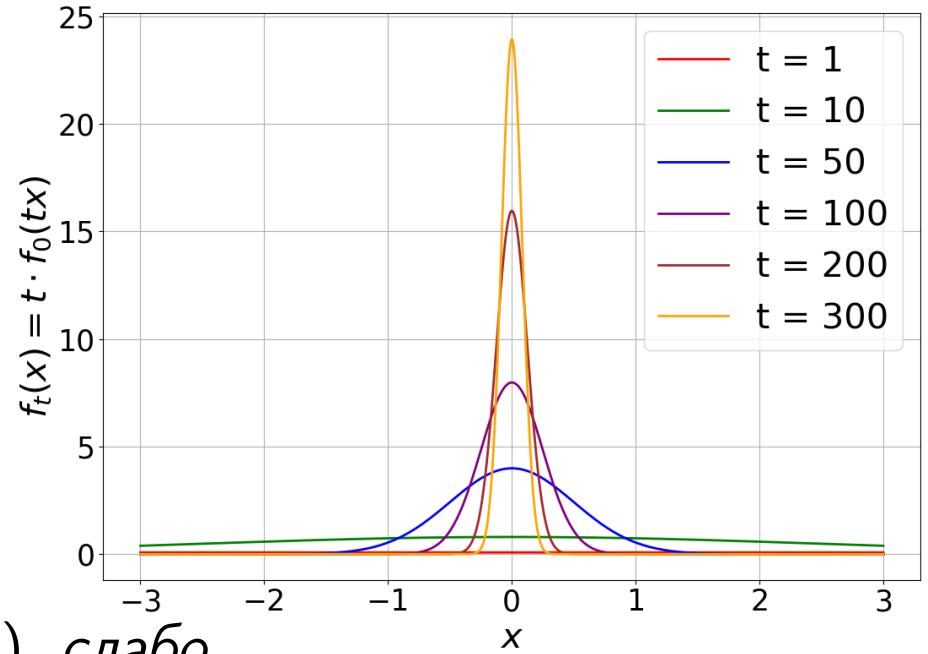
Определим дискретную динамическую систему:

$$f_{t+1}(x) = \mathrm{D}_t(f_t)(x) \quad \text{для} \quad \forall x \in \mathbb{R}^n, t \in \mathbb{N} \quad \text{и} \quad \mathrm{D}_t \in \mathbf{D}, \qquad (1)$$

где D_t – оператор эволюции, x – вектор данных в моделируемой системе, $f_t(x)$ – функции плотности.

Теорема 1 (Веприков, 2023. Предельное множество системы (1))

Для любой функции плотности f_0 и дискретной динамической системы (1), пусть существуют $g \in L_1(\mathbb{R}^n)$ и $\psi_t \geq 0$ такие, что $f_t(x) \leq \psi_t^n \cdot |g(\psi_t \cdot x)|$. Тогда, если $\psi_t \to \infty$, плотности $f_t(x)$ стремятся к дельта-функции, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \delta(x)$, слабо.



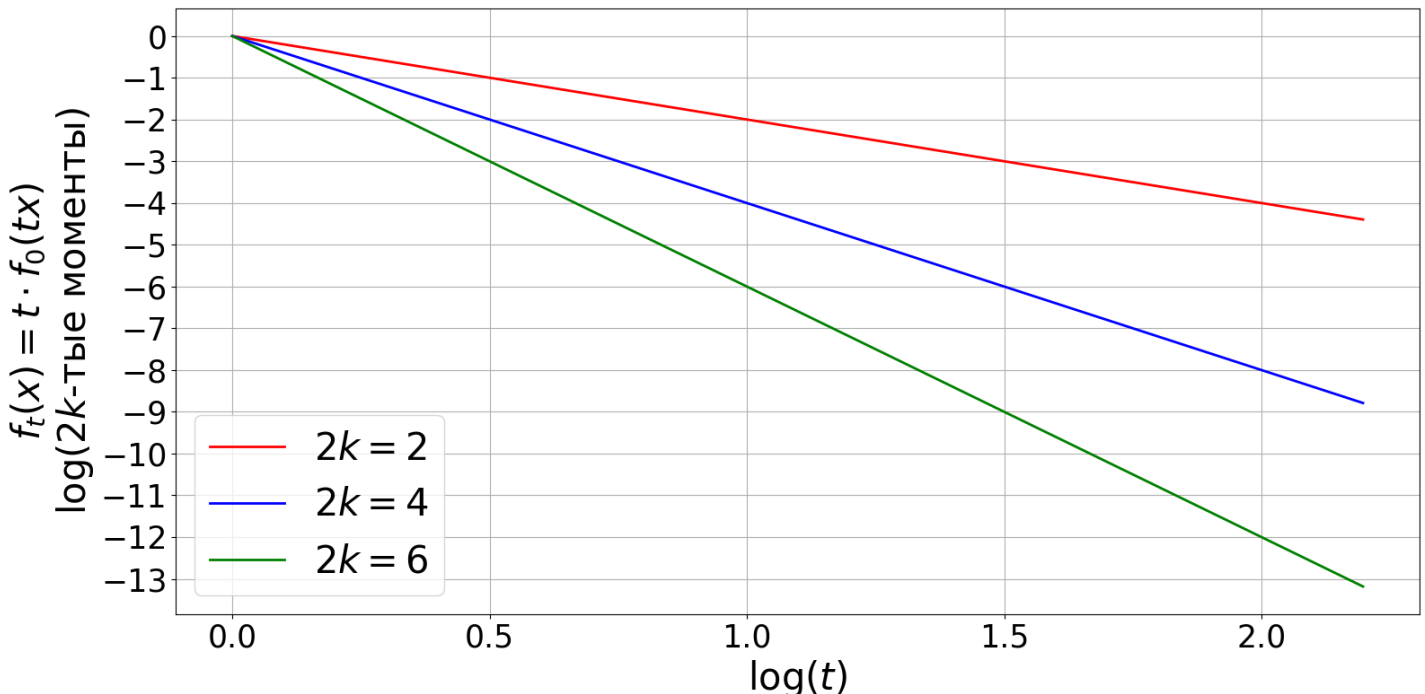
Если $\psi_t \to 0$, тогда плотности $f_t(x)$ сходятся к нулевому распределению, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \zeta(x)$, слабо.

Анализ условий существования петель обратной связи

Для задачи регрессии, когда данные имеют вид $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$
Теорема 1 записывается не для данных в системе ИИ, а для случайного вектора невязок модели h вида $\mathbf{y} - h(\mathbf{x})$.

Лемма 1 (Веприков, 2023. Стремление моментов к нулю)

Если система (1) *с* $n = 1$ *удовлетворяет условиям Теоремы 1 и* $\psi_t \to \infty$, *тогда все* $2k$*-тые моменты невязок* $\boldsymbol{y} - h(\boldsymbol{x})$ *убывают со скоростью как минимум* ψ_t^{-2k}.

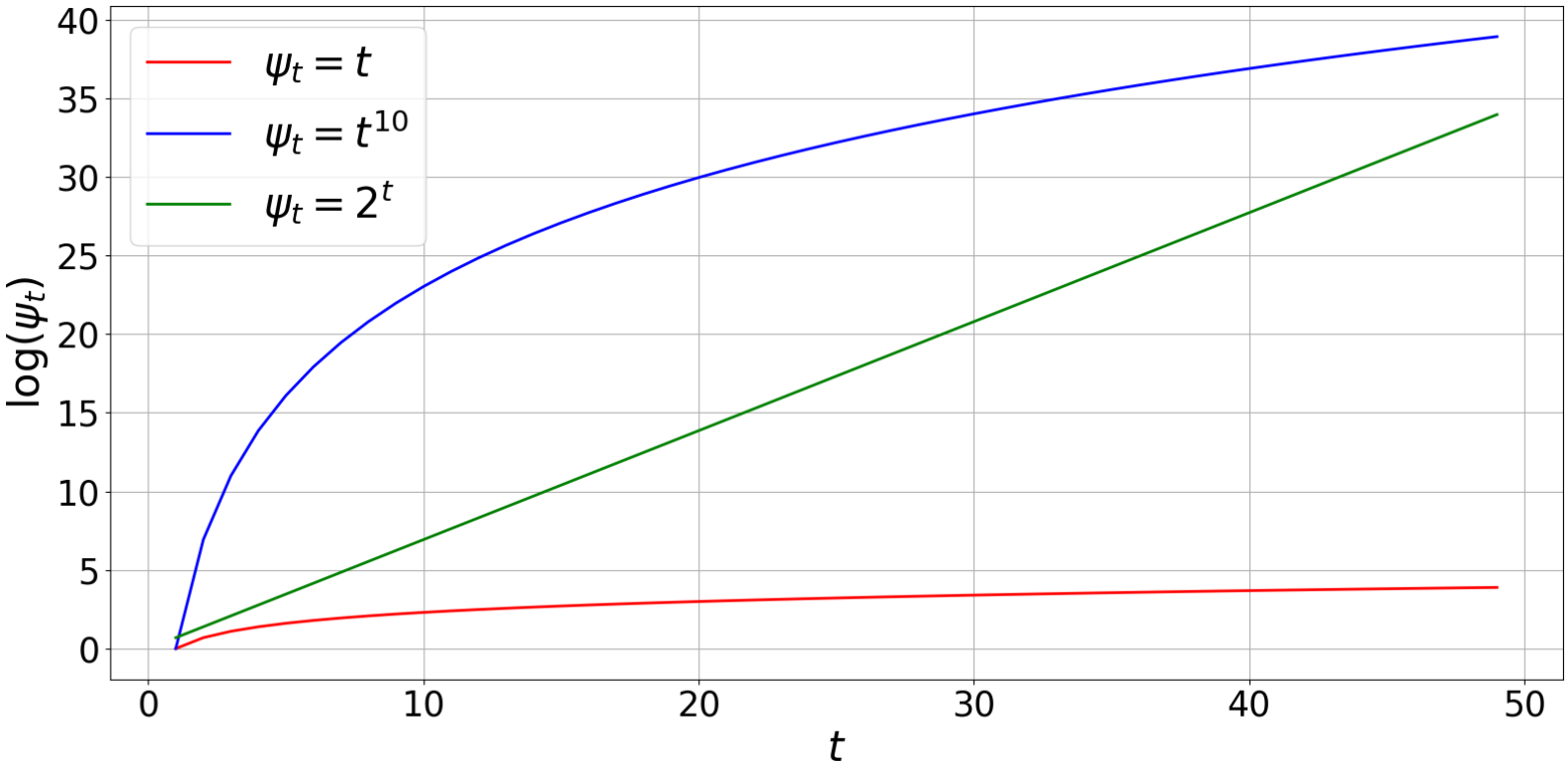


Пример использования Леммы 1.

Анализ автономности системы (1)

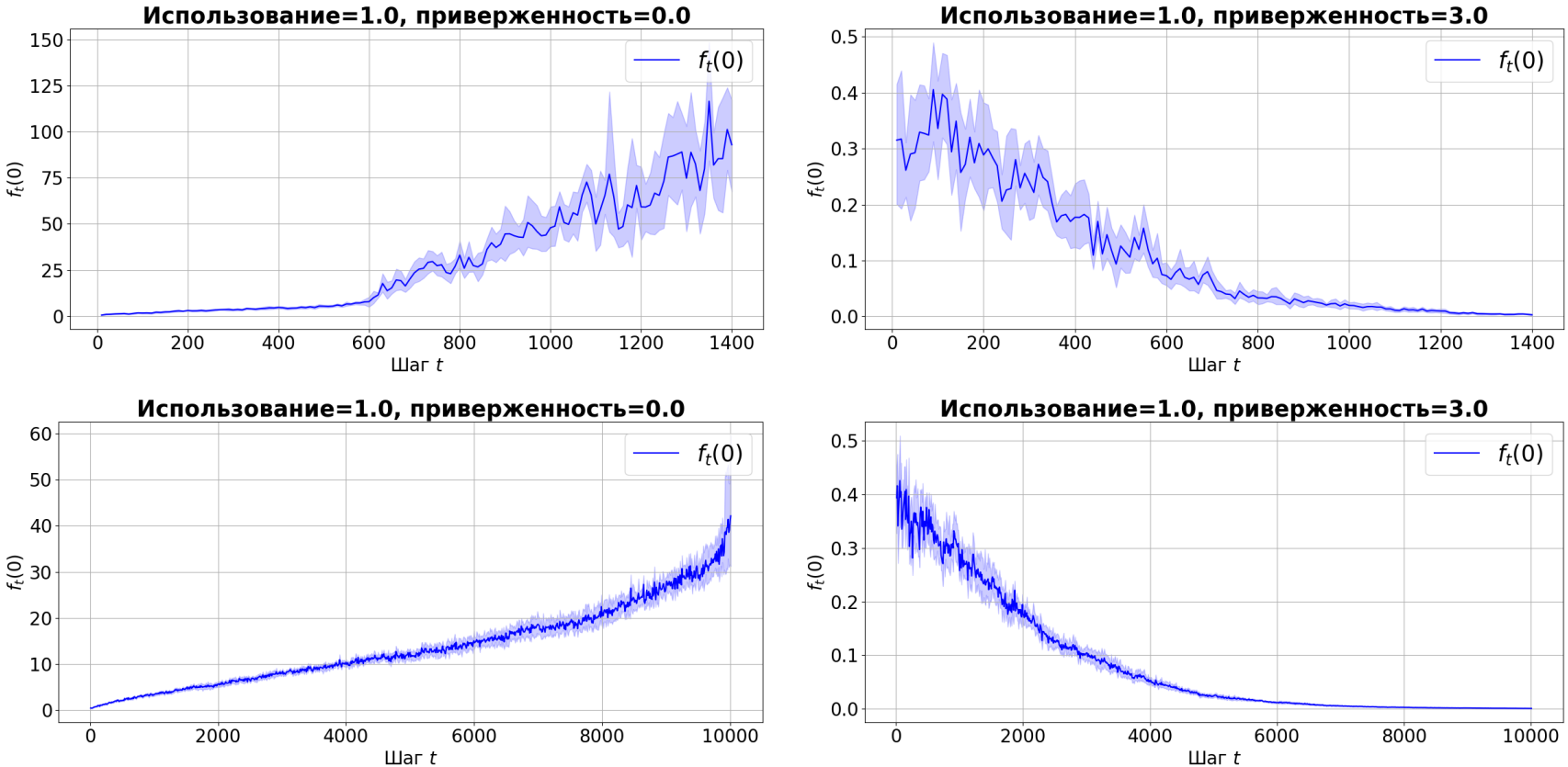
Теорема 2 (Веприков, 2023. Критерий автономности)

Если операторы эволюции D_t динамической системы (1) *имеют вид* $D_{\overline{1,t}}(f_0)(x) = \psi_t^n \cdot f_0(\psi_t \cdot x)$, *тогда система* (1) *автономна тогда и только когда, когда* $\psi_{\tau+\kappa} = \psi_\tau \cdot \psi_\kappa \ \forall \tau, \kappa \in \mathbb{N}$.



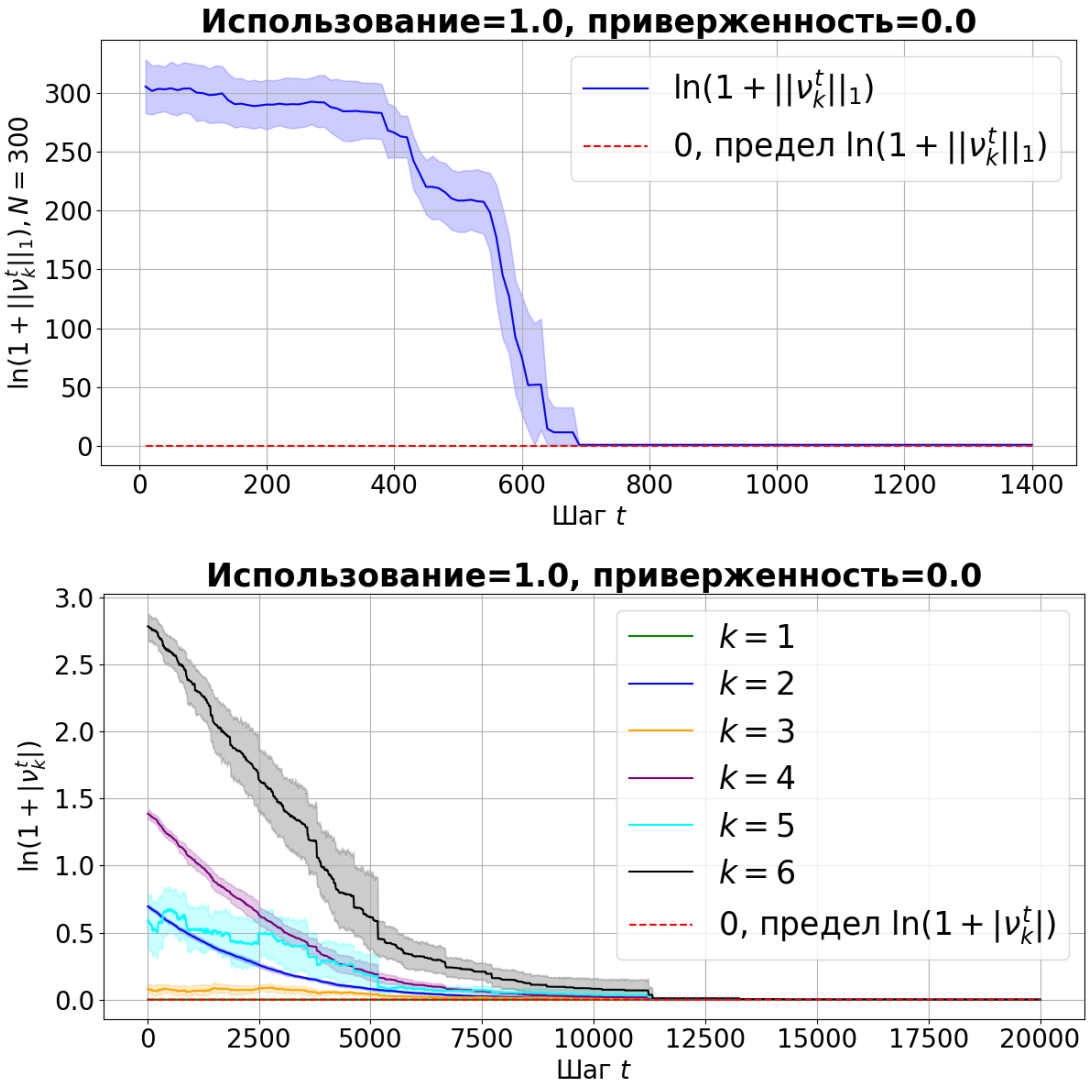
Пример использования Теоремы 2.

Предел к дельта-функции или нулевому распределению



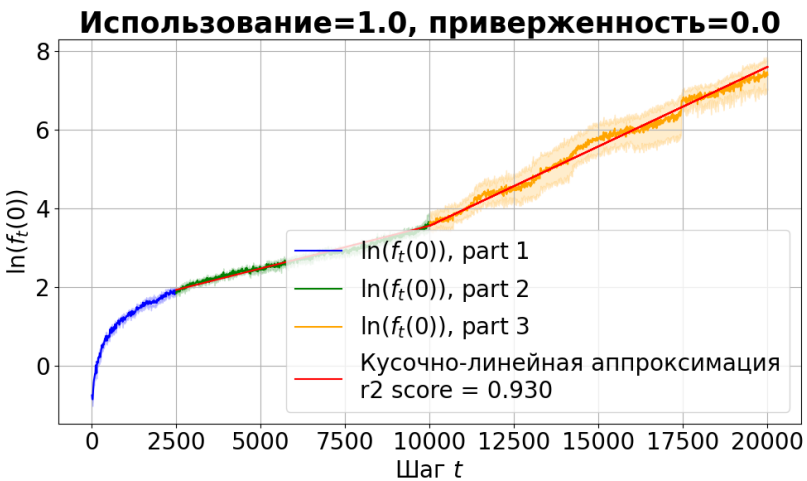
Постановка скользящее окно (сверху), обновление выборки (снизу).

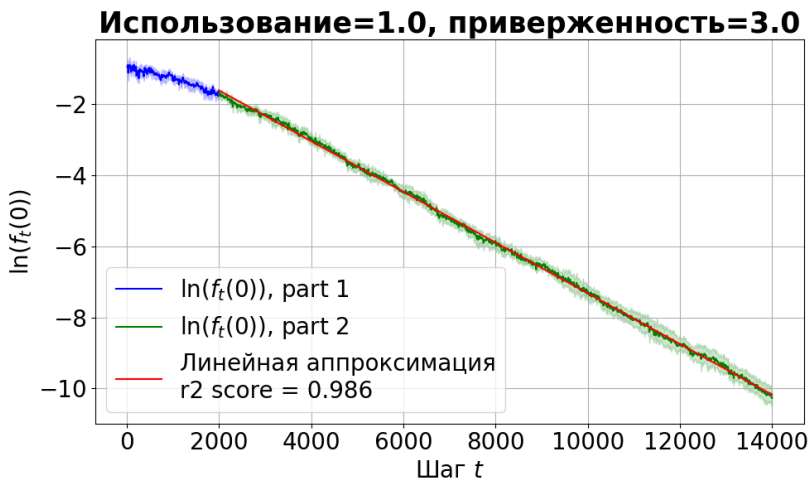
Стремление моментов к нулю



Постановка скользящее окно (сверху), обновление выборки (снизу).

Исследование систем на автономность





Постановка скользящее окно (сверху), обновление выборки(снизу).

Выносится на защиту

1. Построена математическая модель эффекта петель обратной связи с использованием дискретных динамических систем
2. Получены результаты для определения предельного множества динамической системы, достаточных условий существования петли обратной связи и критерий автономности
3. Разработан стенд проведения вычислительных экспериментов, симулирующий процесс многократного машинного обучения

Публикации

1. Подана статья в Q1 журнал «Journal of Machine Learning Research» (JMLR). Preprint: Veprikov A., Afanasiev A., Khritankov A. A Mathematical Model of the Hidden Feedback Loop Effect in Machine Learning Systems // arXiv preprint `https://arxiv.org/abs/2405.02726`
2. Веприков А. С., Афанасьев А. П., Хританков А.С. Математическая модель эффекта обратной связи в системах искусственного интеллекта // Сборник тезисов 21-й Всероссийской конференции Математические методы распознавания образов (ММРО-21). – Российская академия наук, 2023. – С. 35-37
3. Подана статья в журнал «Искусственный интеллект и принятие решений» (ИИПР). Веприков А. С., Хританков А. С. Преобразования Признакового Пространства в Модели Процесса Многократного Машинного Обучения

Список литературы

Adam, G. A., Chang, C.-H. K., Haibe-Kains, B., and Goldenberg, A. (2022).

Error amplification when updating deployed machine learning models.

In *Machine Learning for Healthcare Conference*, pages 715–740. PMLR.

Davies, H. C. (2018).

Redefining filter bubbles as (escapable) socio-technical recursion.

Sociological Research Online, 23(3):637–654.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018).

Runaway feedback loops in predictive policing.

In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR.

Khritankov, A. (2023).

Positive feedback loops lead to concept drift in machine learning systems.

Applied Intelligence, pages 1–19.