ON THE PROBLEM OF REPEATED SUPERVISED LEARNING

A PREPRINT

Andrey S. Veprikov
Department of Intelligent Systems
MIPT
Dolgoprudny, Russia
veprikov.as@phystech.edu

Anton S. Khritankov
HSE University
Moscow, Russia
akhritankov@hse.ru

Alexander P. Afanasyev
IITP
Moscow, Russia

ABSTRACT

In this research paper, we delve into the intricacies of continuous learning artificial intelligence systems as they interact with and influence their environment. We develop a mathematical model to examine the process of repeated and multiple learning, prediction, and dataset updating. Our investigation of this process is based on the principles of functional analysis and probability theory, which is a novel approach to this problem. We aim to conduct several synthetic experiments based on our findings, hoping to contribute to a better understanding of the behavior of continuous learning AI systems.

Keywords Machine learning · Continuous machine learning · Repeated learning

1 Introduction

In this paper we consider the problem of repeated supervised learning (многократное машинное обучение), in which the training sample is not fixed, but is updated depending on the predictions of the trained model on the test sample [8, 7, 4]. In many applications, the use of multiple learning techniques can actually lead to suboptimal results. Our main goal was to present a mathematical theory that explains why combining multiple learning algorithms can sometimes hinder their effectiveness. Repeated supervised learning appears in many machine learning applications, for example in recommendation systems [7, 11], healthcare [1] and predictive policing [3].

The object of our research will be the set $\mathbf{R}$ of distribution density functions

$$\mathbf{R} := \left\{ f : \mathbb{R}^n \to \mathbb{R}_+ \text{ and } \int\limits_{\mathbb{R}^n} f(x)dx = 1 \right\} \tag{1}$$

and a mapping D, a feedback loop mapping, that includes training a model, forming predictions on a test sample and mixing predictions into a training sample, as a result of which the distribution of features changes.

In this paper we propose a mathematical model of the process of repeated learning that is new to the literature. We find sufficient conditions for the operator D to translate $\mathbf{R}$ into $\mathbf{R}$. We find conditions for our data density functions to tend in a weak sense to a delta function under the operator D. We also find sufficient conditions for the operator D to be non-contraction in the norm $\|\cdot\|_q$. The importance of these properties is explained in detail in Section 4.

Structure of this paper are as follows. In Section 2 we compare our article with the works of other authors and show its novelty to the literature. In Section 3 we build a mathematical model for the process of multiple learning, prediction and updating of the sample and outline the main questions, the answers to which we explore in Section 4. In Section 4 we provide our main contributions for the mapping D. In Section 5, based on the results from Section 4, we conduct some synthetic experiments.

2 Related work

The problem we study is somewhat related to the feedback loops [6, 7] – an observable change in the distribution of input data that occurs over time, because of user interaction with the system. According to Conjecture 1 from [6] the positive feedback loop in a system $\mathrm{D} : R \to R$ exists if D is a contraction mapping, but this conjecture was not proved.

Also our problem is connected with dynamical systems, which are studied in [5, 9], but all these works consider continuous time, and in our work it should be discrete.

Some authors analyzed Markov processes from the point of view of dynamical systems [12, 14]. But in Markov chain the future of the process does not depend on the past. Other authors [13, 10] studied stochastic dynamics systems in general.

An important contribution of this paper is that we built a mathematical model for continuous machine learning problem, which is novel to the literature.

3 Problem statement

We consider $\mathbf{R}$ (1) – the set of distribution density functions and a mapping D representing an algorithm for training a model, forming predictions on a test sample and mixing predictions into a training sample, as a result of which the distribution of features changes.

Let's consider an autonomous (time-independent explicitly) discrete dynamical system. There is a dedicated variable - the step number, which increases, at step t and $t+1$ of which the ratio is fulfilled

$$f_{t+1}(x) = \mathrm{D}(f_t)(x), \quad \text{for } \forall x \in \mathbb{R}^n,$$

where D becomes an evolution operator on the space of the specified functions f and the initial function $f_0(x)$ is known. Generally speaking, D can be an arbitrary mapping, not necessarily smooth or continuous.

In the next section, we provide several conditions on the operator D, in order for it to translate $\mathbf{R}$ into $\mathbf{R}$. It is important, because we consider D as operator of changing distribution of our dataset, so it has to transform $\mathbf{R}$ into $\mathbf{R}$.

We also look at the conditions under which the data density functions will tend to a delta function in the weak sense, i.e. $\mathrm{D}^t(f_0)(x) \underset{t\to+\infty}{\longrightarrow} \delta(x)$. This is an important property, because we it can be used to understand when repeated machine learning improves our metrics. Why this is so will be discussed in detail in Section 4.

Another contribution of our paper is a condition, under that operator D wouldn't be a contraction mapping in any norm $\|\cdot\|_q$, i.e. there always would be function $f \in \mathbf{R}$ such that $\|\mathrm{D}(f)\|_q \geq \|f\|_q$. This is also important property, because if D is a contraction mapping, then it converge any start density distribution f_0 to function that is equal to zero in almost every point $x \in \mathbb{R}^n$, i.e. continuous algorithm transform our data to uniform noise.

4 Main results

Notation: In this paper we will use the common notations: the $L_1(\mathbb{R}^n)$-norm of function f:

$$\|f\|_1 := \int\limits_{\mathbb{R}^n} |f(x)|dx \quad \text{and} \quad L_1(\mathbb{R}^n) := \{f : \mathbb{R}^n \to \mathbb{R} \mid \|f\|_1 < +\infty\}$$

The $L_\infty(\mathbb{R}^n)$-norm of function f:

$$\|f\|_\infty := \operatorname*{esssup}_{x\in\mathbb{R}^n}\{f(x)\} := \inf\{C \geq 0 \,|\, |f(x)| \leq C \text{ for almost every } x \in \mathbb{R}^n\} \quad \text{and} \quad L_\infty(\mathbb{R}^n) := \{f : \mathbb{R}^n \to \mathbb{R} \mid \|f\|_\infty < +\infty\}$$

Theorem 1 (Fact)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) \geq 0$ for almost every $x \in \mathbb{R}^n$ and $\|f\|_1 = \int\limits_{\mathbb{R}^n} f(x)dx = 1$, then there exists a random vector ξ, for which f will be a density distribution function.*

Exactly on the basis of Theorem 1 we define $\mathbf{R}$ (1) in this way.

Theorem 2 (Assumptions for $\mathrm{D} : \mathbf{R} \to \mathbf{R}$)**.** *If $\|D\|_1 = 1, \forall f \in \mathbf{R} \hookrightarrow D(f)(x) \geq 0$ for almost every $x \in \mathbb{R}^n$, and exists D^{-1} such that $\|D^{-1}\|_1 \leq 1$, then $D : \mathbf{R} \to \mathbf{R}$.*

Proof. To begin with, let us note that if $\mathrm{D} : \mathbf{R} \to \mathbf{R}$, then $\|\mathrm{D}\|_1 = 1$, because by definition of operator norm:

$$\|\mathrm{D}\|_1 \stackrel{def}{=} \sup_{\|f\|_1=1} \{\|\mathrm{D}(f)\|_1\}$$

And if f such that $\|f\|_1 = 1$ then $|f| \in \mathbf{R}$ and, because $\mathrm{D} : \mathbf{R} \to \mathbf{R}$, $\|\mathrm{D}(|f|)\|_1 = 1$. But $\|\mathrm{D}\|_1 = 1$ is only a necessary but not a sufficient condition.

If $\|\mathrm{D}\|_1 = 1$, then $\forall f \in \mathbf{R} \hookrightarrow \mathrm{D}(f) \leq 1$. If $\exists f_0 \in \mathbf{R}$ such that $\|\mathrm{D}(f_0)\|_1 < 1$, then we get a contradiction because

$$\|D^{-1}\|_1 \stackrel{def}{=} \sup_{\|f\|_1 \neq 0} \left\{ \frac{\|\mathrm{D}^{-1}(f)\|_1}{\|f\|_1} \right\} \geq [f_1 = \mathrm{D}(f_0)] \geq \frac{\|\mathrm{D}^{-1}(f_1)\|_1}{\|f_1\|_1} = \frac{\|\mathrm{D}^{-1}(\mathrm{D}(f_0))\|_1}{\|\mathrm{D}(f_0)\|_1} = \frac{\|f_0\|_1}{\|\mathrm{D}(f_0)\|_1} = \frac{1}{\|\mathrm{D}(f_0)\|_1} > 1$$

But we assume that $\|\mathrm{D}^{-1}\|_1 \leq 1$. So $\forall f \in \mathbf{R} \hookrightarrow \|\mathrm{D}(f)\|_1 = 1$.

But according to Theorem 2 to $\mathrm{D} : \mathbf{R} \to \mathbf{R}$ we also need second assumption: $\forall f \in \mathbf{R} \hookrightarrow \mathrm{D}(f)(x) \geq 0$ for almost every $x \in \mathbb{R}^n$.

□

Discussion of Theorem 2

In experiments it often difficult to calculate D^{-1} and especially it's norm, so we make a different assumptions. We consider D as algorithm for training a model, forming predictions on a test sample and mixing predictions into a training sample. Distribution of our data is approximating by empirical distribution function [2] as follows (for $n = 1$):

$$\hat{F}_N(x) := \frac{\text{number of elements in sample} \leq x}{N} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{X_i \leq x}, \tag{2}$$

where X_i are elements of sample. We assume that $(X_1, X_2, X_3, ..., X_N)$ are independent, identically distributed real random variables with the common cumulative distribution function $F(x)$. If this assumption is fulfilled, then the DKW inequality is satisfied:

$$\mathbb{P}\left\{ \sup_{x \in \mathbb{R}} \left| \hat{F}_N(x) - F(x) \right| > \varepsilon \right\} \leq Ce^{-2N\varepsilon^2} \quad \forall \varepsilon > 0 \tag{3}$$

And we can build interval that contains the true CDF of our data $F(x)$, with probability $1 - \alpha$ as

$$\hat{F}_N(x) - \varepsilon \leq F(x) \leq \hat{F}_N(x) + \varepsilon, \quad \text{where} \quad \varepsilon = \sqrt{\frac{\ln(2/\alpha)}{2N}} \tag{4}$$

In this case operator D transoms our data, i.e. translates one empirical distribution function to another. So $\mathrm{D} : \mathbf{R} \to \mathbf{R}$ by constructing our experiment.

Theorem 3 (Limit in a weak sense to δ function)**.** *If $f_t : \mathbb{R} \to \mathbb{R}$ such that $\forall t \in \mathbb{N} \hookrightarrow \|f_t\|_1 = 1$, $f_t(x) \geq 0$ in almost every point $x \in \mathbb{R}$ and*

$$\exists \psi : \mathbb{N} \to \mathbb{R} : \psi(t) \underset{t \to +\infty}{\longrightarrow} +\infty \text{ and } \exists g \in L_1(\mathbb{R}) \text{ such that } \forall t \in \mathbb{N}\, \forall y \in \mathbb{R} \hookrightarrow f_t\left(\frac{y}{\psi(t)}\right) \leq \psi(t) \cdot |g(y)| \tag{5}$$

Then $f_t(x) \underset{t \to \infty}{\longrightarrow} \delta(x)$ in a weak sense, i.e.

$$\lim_{t \to +\infty} \left(\int_{-\infty}^{+\infty} f_t(x)\phi(x)dx \right) = \phi(0), \tag{6}$$

where ϕ is continuous function with compact support

Proof. Assume a notation $I_t = \int\limits_{-\infty}^{+\infty} f_t(x)\phi(x)dx$. Then it's fulfilled that

$$I_t - \phi(0) = \int\limits_{-\infty}^{+\infty} f_t(x)\phi(x)dx - \phi(0) \cdot \int\limits_{-\infty}^{+\infty} f_t(x)dx = \int\limits_{-\infty}^{+\infty} f_t(x) \cdot [\phi(x) - \phi(0)]dx$$

The first equation is fulfilled because $\|f_t\| = 1$.

Replacing the variable $y = \psi(t) \cdot x, dy = \psi(t) \cdot dx$ we get

$$I_t - \phi(0) = \frac{1}{\psi(t)} \cdot \int\limits_{-\infty}^{+\infty} f_t\left(\frac{y}{\psi(t)}\right) \cdot \left[\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right] dy \tag{7}$$

Split the integral (7) into 3 parts:

$$I_1 := \frac{1}{\psi(t)} \cdot \int\limits_{-\infty}^{-A} f_t\left(\frac{y}{\psi(t)}\right) \cdot \left[\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right] dy$$

$$I_2 := \frac{1}{\psi(t)} \cdot \int\limits_{-A}^{A} f_t\left(\frac{y}{\psi(t)}\right) \cdot \left[\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right] dy$$

$$I_3 := \frac{1}{\psi(t)} \cdot \int\limits_{A}^{+\infty} f_t\left(\frac{y}{\psi(t)}\right) \cdot \left[\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right] dy$$

Consider the integrals I_1. ϕ is continuous function with compact support, so ϕ is bounded by some constant M, i.e. $\forall x \in \mathbb{R} \hookrightarrow |\phi(x)| \leq M$, so

$$|I_1| \leq \int\limits_{-\infty}^{-A} 2M \cdot \frac{1}{\psi(t)} f_t\left(\frac{y}{\psi(t)}\right) dy \leq [(5)] \leq \int\limits_{-\infty}^{-A} 2M \cdot |g(y)|dy$$

Since $g \in L_1(\mathbb{R})$, there exists some constant $A > 0$ such that $|I_1| \leq \varepsilon$. Similarly, it can be shown that $|I_3| \leq \varepsilon$.

Consider the integral I_2. ϕ is continuous and $\psi \to +\infty$ (5), so $\exists\, T \in \mathbb{N}$ such that $\forall y \in [-A; A]\ \forall t \geq T$ it is fulfilled that

$$\left|\frac{y}{\psi(t)} - 0\right| \leq \delta \text{ and so } \left|\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right| \leq \varepsilon$$

So

$$|I_2| \leq \varepsilon \cdot \int\limits_{-A}^{A} \frac{1}{\psi(t)} f_t\left(\frac{y}{\psi(t)}\right) dy = \varepsilon \cdot \int\limits_{-A/\psi(t)}^{A/\psi(t)} f_t\left(x\right) dx \leq \varepsilon \cdot \int\limits_{-\infty}^{+\infty} f_t\left(x\right) dx = \varepsilon$$

Finally we get

$$\forall \varepsilon > 0\ \ \exists\, T \in \mathbb{N} : \forall t \geq T \hookrightarrow |I_t - \phi(0)| \leq |I_1| + |I_2| + |I_3| \leq 3\varepsilon$$

So $f_t(x) \underset{t\to\infty}{\longrightarrow} \delta(x)$ in a weak sense.

□

Discussion of Theorem 3

If in some step t the model $H(x, \theta, t)$ starts to give good predictions on the training and test samples, then, in the probabilistic formulation, this means that the density function of the distribution of the object-sign vectors becomes similar to the delta function, as the components of the random vector $(\mathbf{x}^{\mathbf{i}}, y_i) \subset (\mathbf{X}, \mathbf{y})$ become linearly dependent. Based on these considerations, we can compare each operator D to an operator $\widetilde{\mathrm{D}}$, where $\widetilde{\mathrm{D}}$ transforms density distribution functions of random variables $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$, where $\mathbf{y}_{\text{pred}} = H(\mathbf{X}, \theta, t)$. Then we can apply Theorem 3 to understand, when $\widetilde{\mathrm{D}}(f_0)(x) \underset{t\to\infty}{\longrightarrow} \delta(x)$.

If $\widetilde{\mathrm{D}}$ has a fixed point, then the distribution of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ does not change, so algorithm of repeated learning does not improves quality metrics, but in this case operator D can still have no fixed point. That's why we consider formulas (3) and (4) only for $n = 1$, because in our experiments we will detect a moment, when MSE error of our algorithm will stop decreasing, so the distribution of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ will be a fixed point of $\widetilde{\mathrm{D}}$.

Let's analyze formula (5).

If we take $x = \psi(t) \cdot y$ then (5) takes form

$$\exists g \in L_1(\mathbb{R}) \text{ such that } \forall t \in \mathbb{N}\, \forall x \in \mathbb{R} \hookrightarrow f_t(x) \leq \psi(t) \cdot |g(x \cdot \psi(t))| \tag{8}$$

If $x \neq 0$ then $f_t(x) \underset{t\to\infty}{\longrightarrow} 0$, because $g_1(x) := \psi(t) \cdot |g(x \cdot \psi(t))| \in L_1(\mathbb{R})$ since

$$\int_{-\infty}^{+\infty} g_1(x)dx = \int_{-\infty}^{+\infty} \psi(t) \cdot |g(x \cdot \psi(t))|dx = \int_{-\infty}^{+\infty} g_1(z)dz < +\infty$$

And so, if $t \to \infty$, then $z := x \cdot \psi(t) \to +\infty$ and $g_1(z) \to 0$, because $g_1 \in L_1(\mathbb{R})$. So, if $x \neq 0$ then $f_t(x) \underset{t\to\infty}{\longrightarrow} 0$.

Since $\forall t \in \mathbb{R} \hookrightarrow \|f_t\|_1 = 1$, then $f_t(0) \to +\infty$.

If we substitute $x = 0$ in the (8) then we get $f_t(0) \leq \psi(t) \cdot |g(0)|$, so we can take

$$\psi(t) = \frac{f_t(0)}{|g(0)|} \tag{9}$$

In our experiments we will measure $f_t(0)$ and $\int_{-\kappa}^{\kappa} f_t(x)dx = \hat{F}_t(\kappa) - \hat{F}_t(-\kappa)$, where κ is sufficiently small. So if $f_t(0) \to +\infty$ and $1 \in I_\varepsilon(\hat{F}_t(\kappa) - \hat{F}_t(-\kappa))$, where I_ε is confidence interval from (4), then we can say that operator $\widetilde{\mathrm{D}}$ converge empirical distribution functions $f_t(x)$ of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ to delta-function.

Now we can see importance of equations (3) and (4). If empirical density distributions functions $\widetilde{\mathrm{D}}$ converge to delta functions then according to (3) and (4) we can consider, that true density distribution function $F(x)$ converge to $\delta(x)$, because from (4) we have

$$\hat{F}_t(x) - \varepsilon \leq F(x) \leq \hat{F}_t(x) + \varepsilon, \quad \text{where} \quad \varepsilon = \sqrt{\frac{\ln(2/\alpha)}{2N}} \text{ and } \hat{F}_t(x) \text{ is empirical CFD on step } t$$

So if $\hat{F}_t(x) \to F_\delta(x) = \text{sign}(x)$, then $F(x) \to F_\delta(x)$

Important example of operator D that translates any function from $\mathbf{R}$ into a δ function is as follows

$$\mathrm{D}^t(f_0)(x) = t \cdot f_0(t \cdot x) \tag{10}$$

Here we take $g(x) = f_0(x)$ and $\psi(t) = t$.

In Figure 1 shown how this operator translates density functions of normal distribution $\mathcal{N}(0, 5)$ and continuous uniform distribution $\mathcal{U}[-2.5, 2.5]$ with $t \to +\infty$:

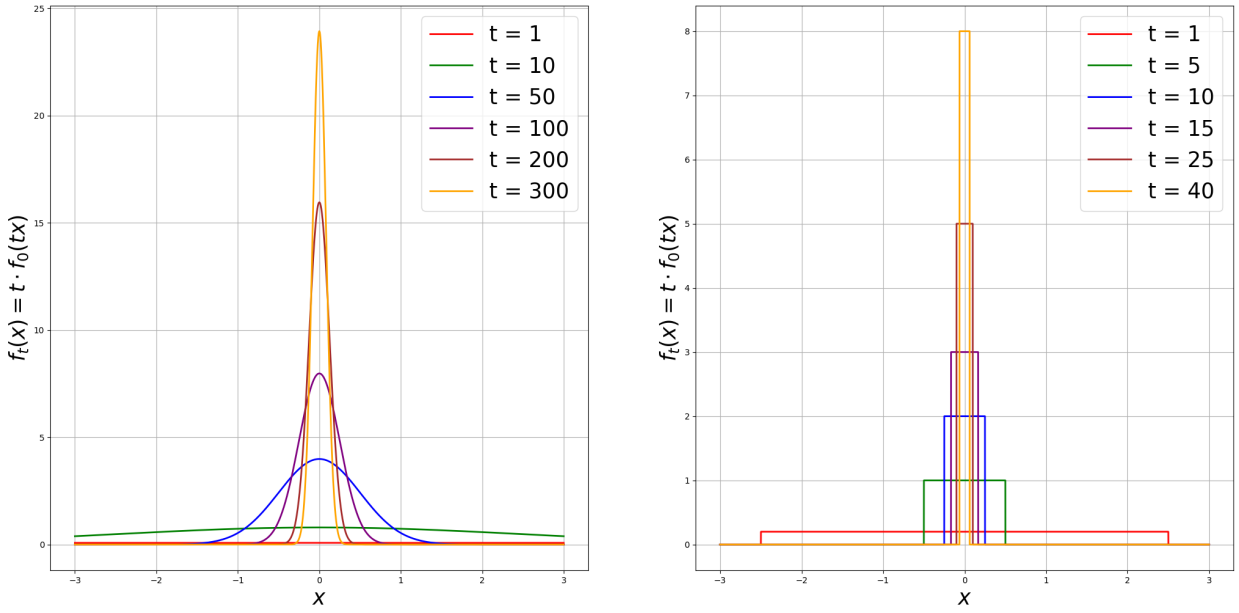

Figure 1: Illustration of weak limit to δ function. $\mathcal{N}(0, 5)$ left, $\mathcal{U}[-2.5, 2.5]$ right.

How we provide several Lemmas based on Theorem 3. In In these lemmas we assume that the operator D has the form

$$\mathrm{D}^t(f_0)(x) = \psi(t) \cdot f_0(\psi(t) \cdot x) \text{ and } \psi(t) \to +\infty \tag{11}$$

Lemma 4 (Conditions on $\{\mathrm{D}^t\}_{t=0}^{+\infty}$ to be a semigroup)**.** *If D has the form* (11)*, then* $\{D^t\}_{t=0}^{+\infty}$ *is a semigroup, i.e.* $(D^\tau \circ D^\kappa)(f) = D^{\tau+\kappa}(f)\ \forall \tau, \kappa \in \mathbb{N}$*, if and only if*

$$\psi(\tau + \kappa) = \psi(\tau) \cdot \psi(\kappa)\ \forall \tau, \kappa \in \mathbb{N} \tag{12}$$

Proof.

$$(\mathrm{D}^\tau \circ \mathrm{D}^\kappa)(f)(x) = \mathrm{D}^\tau(\psi(\kappa) \cdot f(\psi(\kappa) \cdot x)) = \psi(\tau)\psi(\kappa) \cdot f(\psi(\tau)\psi(\kappa) \cdot x)$$

$$\mathrm{D}^{\tau+\kappa}(f)(x) = \psi(\tau + \kappa) \cdot f(\psi(\tau + \kappa) \cdot x)$$

So, $\{\mathrm{D}^t\}_{t=0}^{+\infty}$ is a semigroup $\Leftrightarrow \psi(\tau + \kappa) = \psi(\tau) \cdot \psi(\kappa)\ \forall \tau, \kappa \in \mathbb{N}$ □

Lemma 5 (Decreasing moments)**.** *If D has the form* (11)*, then all k-th moments of random variable* $\|\mathbf{y} - \mathbf{y}_{pred}\|$ *(if they exist) are decreasing with speed* $\psi(t)^{-k}$*, i.e.* $\nu_k^t = \psi(t)^{-k}\nu_k^0$*, where* ν_k^t *is a k-th moment on a step t.*

If $\exists q \in [1; +\infty]$ *such* $\{\nu_k^0\}_{k=1}^{+\infty} \in l_q$*, then* $\{\nu_k^t\}_{k=1}^{+\infty} \in l_1$ *and* $\{\nu_k^t\}_{k=1}^{+\infty} \xrightarrow[t\to\infty]{l_1} 0$

Proof. Let's first prove first term. By the definition of k-moment we have

$$\nu_k^t = \int_{-\infty}^{+\infty} x^k \psi(t) f(\psi(t)x) dx$$

If we make variable substitution $y = \psi(t)x$, when we have

$$\nu_k^t = \int_{-\infty}^{+\infty} \frac{y^k}{\psi(t)^k} f(y) dy = \psi(t)^{-k}\nu_k^0$$

So the first term is proved. Consider the second term.

$$\|\{\nu_k^t\}_{k=1}^{+\infty}\|_1 = \|\{\psi(t)^{-k}\nu_k^0\}_{k=1}^{+\infty}\|_1 \leq \|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_p \cdot \|\{\nu_k^0\}_{k=1}^{+\infty}\|_q$$

The second step follows from Helder's inequality.

How let's calculate $\|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_p$ for $p \in [1; +\infty)$:

$$\|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_p^p = \sum_{k=1}^{+\infty} \psi(t)^{-kp} = \frac{\psi(t)^{-p}}{1-\psi(t)^{-p}} = \frac{1}{\psi(t)^p - 1}$$

The first equality is true only if $\psi(t) > 1$ and the second step follows from sum of infinitely decreasing geometric progression.

So

$$\|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_p = \left(\frac{1}{\psi(t)^p - 1}\right)^{1/p} \underset{t\to+\infty}{\longrightarrow} 0 \quad \forall p \in [1; +\infty)$$

If $p = +\infty$:

$$\|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_\infty = [\text{ if } \psi(t) > 1] = \psi(t)^{-1} \underset{t\to+\infty}{\longrightarrow} 0$$

So, we have

$$\|\{\nu_k^t\}_{k=1}^{+\infty}\|_1 \leq \|\{\psi(t)^{-k}\}_{k=1}^{+\infty}\|_p \cdot \|\{\nu_k^0\}_{k=1}^{+\infty}\|_q \underset{t\to+\infty}{\longrightarrow} 0$$

Because $\|\{\nu_k^0\}_{k=1}^{+\infty}\|_q < +\infty$ as a condition of Lemma.

□

Discussion of Lemmas 4 and 5

If our system is autonomous, i.e. it does not depend on time, then the operators $\{\mathrm{D}^t\}_{t=0}^{+\infty}$ should form a semigroup, since their application should not depend on the number of step t. The condition (12) means that the function $\psi(t)$ is a power function.

The Lemma 5 on moments is interesting from the practical point of view as a condition which is relatively easy to check in the experiment, which we will do in the Section 5.

Theorem 6 (Inequality on $\|\mathrm{D}\|_q$)**.** *Consider*

$$f_A(x) = \frac{1}{\lambda(A)} \cdot \boldsymbol{1}_A(x), \tag{13}$$

where $A \subset \mathbb{R}^n$ is arbitrary set of a non-path measure, $\lambda(A)$ – the measure of a set A.

Then for all $A \subset \mathbb{R}^n : 0 < \lambda(A) < +\infty$ and for all $1 \leq q \leq +\infty$ such that $D(f_A) \in L_q(\mathbb{R}^n)$ is fulfilled that

$$\|D\|_q \geq \int\limits_A D(f_A)(x)dx \tag{14}$$

Proof. First of all let's calculate $\|f_A\|_p$:

$$\|f_A\|_p = \left(\int\limits_{\mathbb{R}^n} \left(\frac{1}{\lambda(A)}\right)^p \cdot \mathbf{1}_A(x)dx\right)^{1/p} = \frac{1}{\lambda(A)} \cdot (\lambda(A))^{1/p} = (\lambda(A))^{-1+1/p}$$

So, $f_A \in L_p(\mathbb{R}^n)$ for all $A \subset \mathbb{R}^n : 0 < \lambda(A) < +\infty$ and $1 \leq p \leq +\infty$.

Now write out a Helder's inequality. For q such that $\frac{1}{p} + \frac{1}{q} = 1$ is fulfilled that

$$\|f_A\|_p \cdot \|\mathrm{D}(f_A)\|_q \geq \|f_A \cdot \mathrm{D}(f_A)\|_1$$

Using common inequality on operators norm $\|\mathrm{D}(f)\|_q \leq \|\mathrm{D}\|_q \cdot \|f\|_q \; \forall f \in L_q(\mathbb{R}^n)$ we get

$$\|f_A\|_p \|f_A\|_q \cdot \|\mathrm{D}\|_q \geq \|f_A \cdot \mathrm{D}(f_R)\|_1$$

Since $\|f_A\|_p \|f_A\|_q = (\lambda(A))^{-1+1/p} \cdot (\lambda(A))^{-1+1/q} = (\lambda(A))^{-2+1/p+1/q} = (\lambda(A))^{-1}$ we get:

$$\|\mathrm{D}\|_q \geq \lambda(A) \cdot \|f_R \cdot \mathrm{D}(f_A)\|_1$$

Let's look at $\|f_A \cdot \mathrm{D}(f_A)\|_1$ in more detail:

$$\|f_A \cdot \mathrm{D}(f_A)\|_1 = \int\limits_{\mathbb{R}^n} \mathrm{D}(f_A)(x) \cdot \frac{1}{\lambda(A)} \mathbf{1}_A(x) dx = \frac{1}{\lambda(A)} \int\limits_A \mathrm{D}(f_A)(x) dx$$

Finally, we get the desired inequality

$$\|\mathrm{D}\|_q \geq \int\limits_A \mathrm{D}(f_A)(x) dx$$

□

Discussion of Theorem 6

To begin with, note that the result of Theorem 6 does not depend in any way on whether D converts $\mathbf{R}$ to $\mathbf{R}$, it is a consequence of Helder's inequality.

If you look at it from a probabilistic point of view (so now $\mathrm{D} : \mathbf{R} \to \mathbf{R}$), then f_A is a distribution density function of vectors uniformly distributed on a set A. So, $\int\limits_A \mathrm{D}(f_A)(x) dx \leq 1$, i.e. from Theorem 6 we can only get that $\|\mathrm{D}\|_q \geq 1$.

But if $\|\mathrm{D}\|_q \geq 1$, so D wouldn't be a contraction mapping in $\|\cdot\|_q$, because there always would be function $f \in \mathbf{R}$ such that $\|\mathrm{D}(f)\|_q \geq \|f\|_q$.

In our experiments we will measure $\hat{F}_t(\sup(A))$ and $\hat{F}_t(\inf(A))$. If $1 \in I_\varepsilon[\hat{F}_t(\sup(A))]$ and $0 \in I_\varepsilon[\hat{F}_t(\inf(A))]$, where I_ε is confidence interval from (4), then we can say that operator D can't be a contraction mapping.

5 Experiments

5.1 Experiment Design

We consider [1]such a formulation of the problem of repeated supervised learning: there is an original data $\mathbf{X}$ and a vector of target variables $\mathbf{y}$. Initially, at round $r = 0$, we select 30% of the original dataset on which our model $H(\mathbf{x}, \theta^0, 0)$ is trained with 80% train size.

Then at each step t we randomly select an element $(\mathbf{x^i}, y_i)$ – the features and target variable of the i-th object from the original dataset, and this element does not lie in the initial dataset. Next, we get the prediction y_i' of our model on the element $(\mathbf{x^i}, y_i)$: $y_i' = H(\mathbf{x^i}, \theta^r, t)$ and sample $z_i \sim \mathcal{N}(y_i', s \cdot \sigma^2)$, where s is an experiment parameter that indicates adherence and σ is the model's mean squared error on held-out data. Then we remove 1 element from the active dataset and, with the probability of p, we add $(\mathbf{x^i}, z_i)$ to it, and with the probability $(1 - p)$ we add the element $(\mathbf{x^i}, y_i)$. We carry out this procedure until we run out of elements that were not initially included in our dataset, so, we make a total of $0.7 \cdot n$ steps, where n is the number of objects in $\mathbf{X}$.

After each T steps round r is increasing: $r = r + 1$ and the machine learning model $H(x, \theta^r, t)$ is retrained with 80% train size on active dataset.

[1]All experiments you can see on our GitHub

The size of active dataset will always be $0.3 \cdot n$, because on each step we remove 1 element and add 1 element to active data. This experiment design is similar to [6] were author was detecting hidden feedback loops in machine learning systems. The scheme of the experiment is shown in Figure 2.

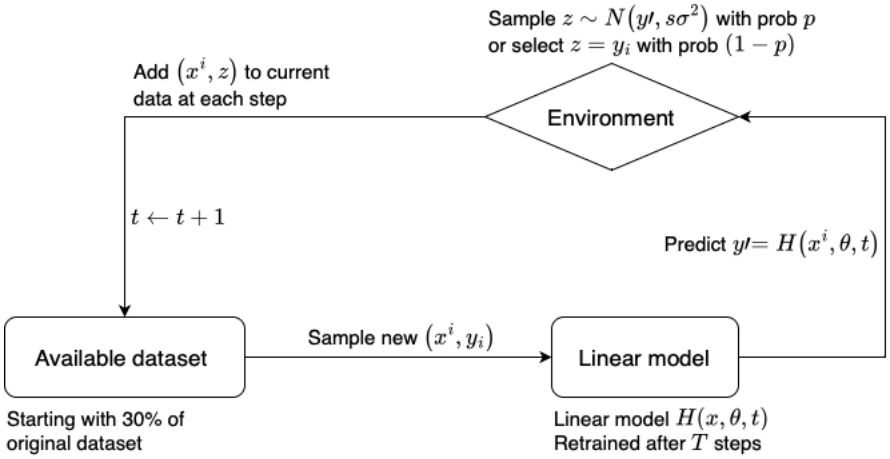

Figure 2: Experiment setup.

In this paper we consider linear model, that is solved as Ridge regression with mean squared error loss function. In the formulation of our experiment, we consider $\mathbf{R}$ (1) as space of density functions of random vectors $(\mathbf{x^i}, y_i)$, where $\mathbf{x^i} \in \mathbf{X}$ and $y_i \in \mathbf{y}$ – the features and target variable of the i-th object. The operator D transforms $\mathbf{R}$ at each step t, and the distribution of features does not change, because at each step we take new $\mathbf{x^i}_t$ from the original set of features $\mathbf{X}$, so only the distribution of the target variable changes. So we can use results from Theorem 3.

5.2 First experiment: analysis of deviation

In almost all distributions, decreasing the variance to 0 means that the distribution function takes the form of a delta function, for example in the normal distribution $\mathcal{N}(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{and} \quad \sigma^2 = \sigma^2$$

And in continuous uniform distribution $\mathcal{U}[a, b]$:

$$f(x) = \frac{1}{b-a} \cdot \mathbf{1}_{[a;b]} \quad \text{and} \quad \sigma^2 = \frac{1}{12} \cdot (b-a)^2$$

For this reason, in our experiments every N steps we measured the standard deviation in the $\mathbf{y} - \mathbf{y}_{\text{pred}}$ array, where $\mathbf{y}_{\text{pred}}$ is the predictions of our model on the active dataset. We measured the standard deviation at different *usage* – the probability with which we take $(\mathbf{x^i}, z_i)$ into the active dataset, and *adherence* – the parameter by which we multiply σ^2 when sampling z_i.

First we consider $\mathbf{X}$ and $\mathbf{y}$ as a regression problem, so $\mathbf{y} = \mathbf{X} \cdot \theta$ for some vector θ. To complicate the model, a normally distributed noise was added to the data. In Figure 3 there are five 3-dimensional plots for different noise in original data, where on X-axis is plotted *usage*, on Y-axis is *adherence* and on Z is deviation of $\mathbf{y} - \mathbf{y}_{\text{pred}}$ for test data samples.

As we can see, the picture for the graphs at noise= $0.3, 1, 3$ and 10 are very similar to each other, they differ only in the deviation values. This is due to the fact that for higher values of noise it becomes more profitable to add new data than to take the original data, since the noises in them are smaller, that is, with increasing *usage* the deviation will fall.

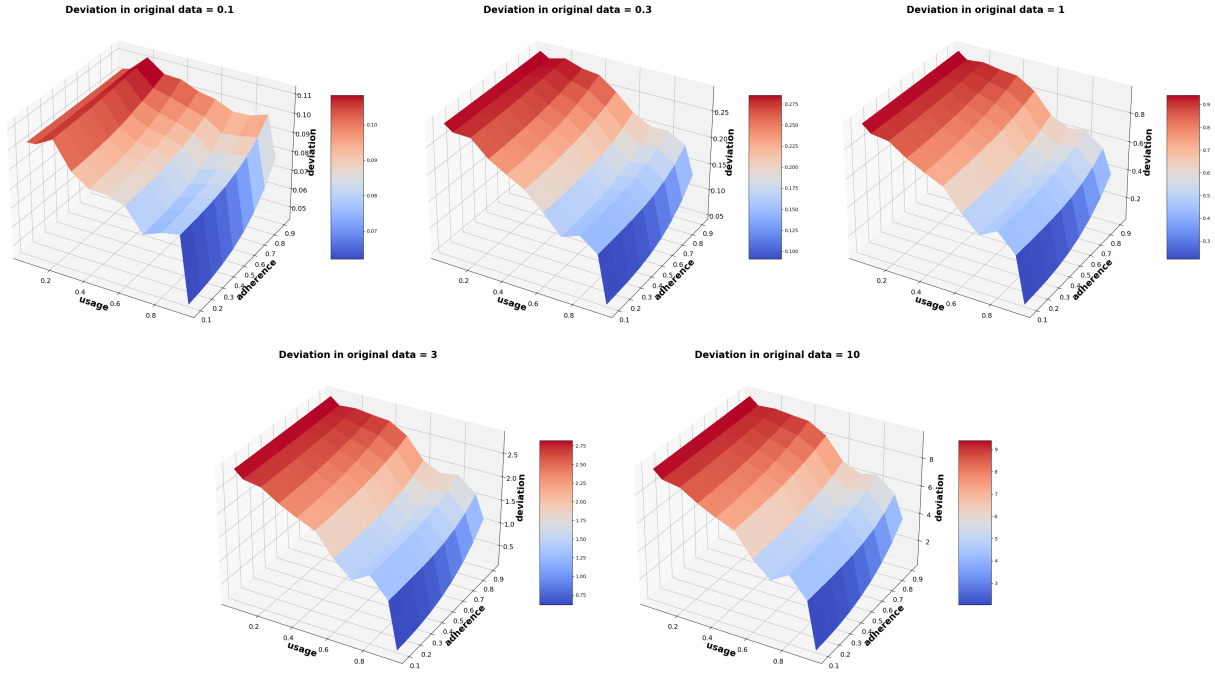

Figure 3: 3D Graphics for different noise in original data

5.3 Second Experiment: analysis of p-value

In this experiment we check our data for belonging to a normal distribution by counting p-value from normal test and we also count the mean and variance of the data $\mathbf{y} - \mathbf{y}_{\text{pred}}$ at each step t. We take *usage* and *adherence* 2 based on the previous experiment 5.2: 1.0 and 0.0, 0.6 and 0.8, 0.1 and 0.9. The results of this experiment are shown in the Figure 4.

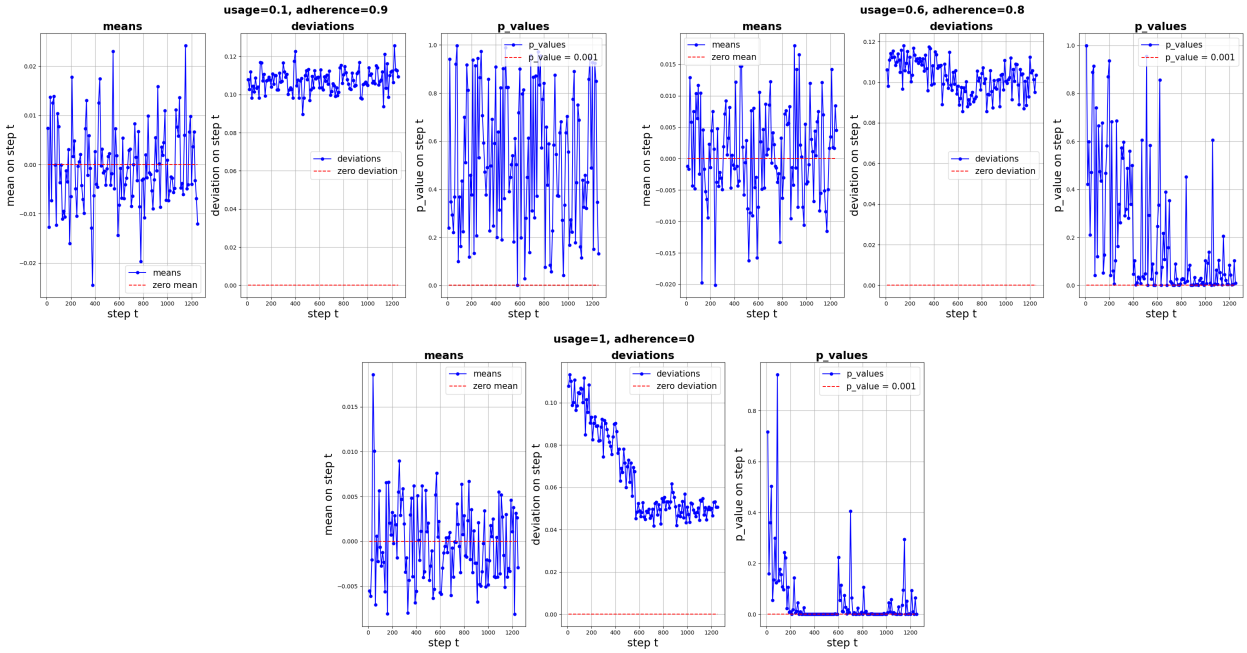

Figure 4: Results of experiment 5.3 *usage* and *adherence* = 0.1 and 0.9 (left), 0.6 and 0.8 (right), 1.0 and 0.0 (bottom).

As we can see for small *usage* the deviation of out data does not decrease, but p-value for all *usage* and *adherence* is bigger than 0.05, so we can assume that our data are normally distributed at each step t.

5.4 Third experiment: limit to delta function

In this experiment we test the conditions from Theorem 3, i.e. we measure $f_t(0)$ and $\int\limits_{-\kappa}^{\kappa} f_t(x)dx = \hat{F}_t(\kappa) - \hat{F}_t(-\kappa)$, where κ is sufficiently small. So if $f_t(0) \to +\infty$ and $1 \in I_\varepsilon(\hat{F}_t(\kappa) - \hat{F}_t(-\kappa))$, where I_ε is confidence interval from (4), then we can say that operator $\widetilde{\mathrm{D}}$ converge empirical distribution functions $f_t(x)$ of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ to delta-function. The results of this experiment are shown in the Figure 5.

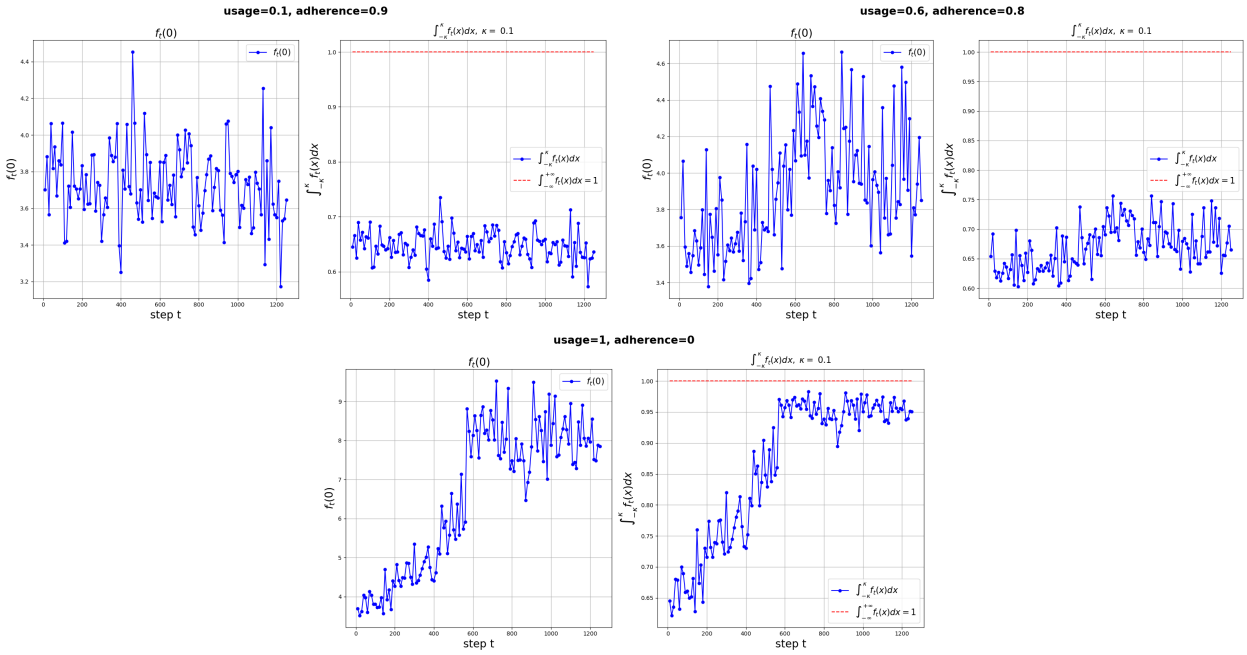

Figure 5: Results of experiment 5.4 *usage* and *adherence* = 0.1 and 0.9 (left), 0.6 and 0.8 (right), 1.0 and 0.0 (bottom).

As we can see, for small values of *usage* and large values of *adherence*, the operator $\widetilde{\mathrm{D}}$ does not translate the distribution density function of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ to delta-function. When *usage* = 1 and *adherence* = 0 we can observe a tendency towards the delta function at small t, and then exit to a plateau. From results of this experiment we can reject the condition $\psi(t) \underset{t \to +\infty}{\longrightarrow} +\infty$ and consider $\psi(t)$ as an arbitrary function that satisfies the inequality (5), but then there will be no limit to the delta function when.

5.5 Fourth experiment: semigroup check

In this experiment we test our system for autonomy, that is, we test the condition (12) in the Lemma 4. As noted in the Discussion of Lemma 4, in order for the $\psi(t)$ function to satisfy the condition (12), it is necessary and sufficient that $\psi(t)$ should be a power function. Therefore, in this experiment we plot $\ln(\psi(t))$, and if we get a straight line, then the system is autonomous, and if not, then no. The $\psi(t)$ function is taken from the Experiment 5.4. The results of this experiment are shown in the Figure 6.

As we can see, our system is not autonomous, since none of the graphs is a straight line.

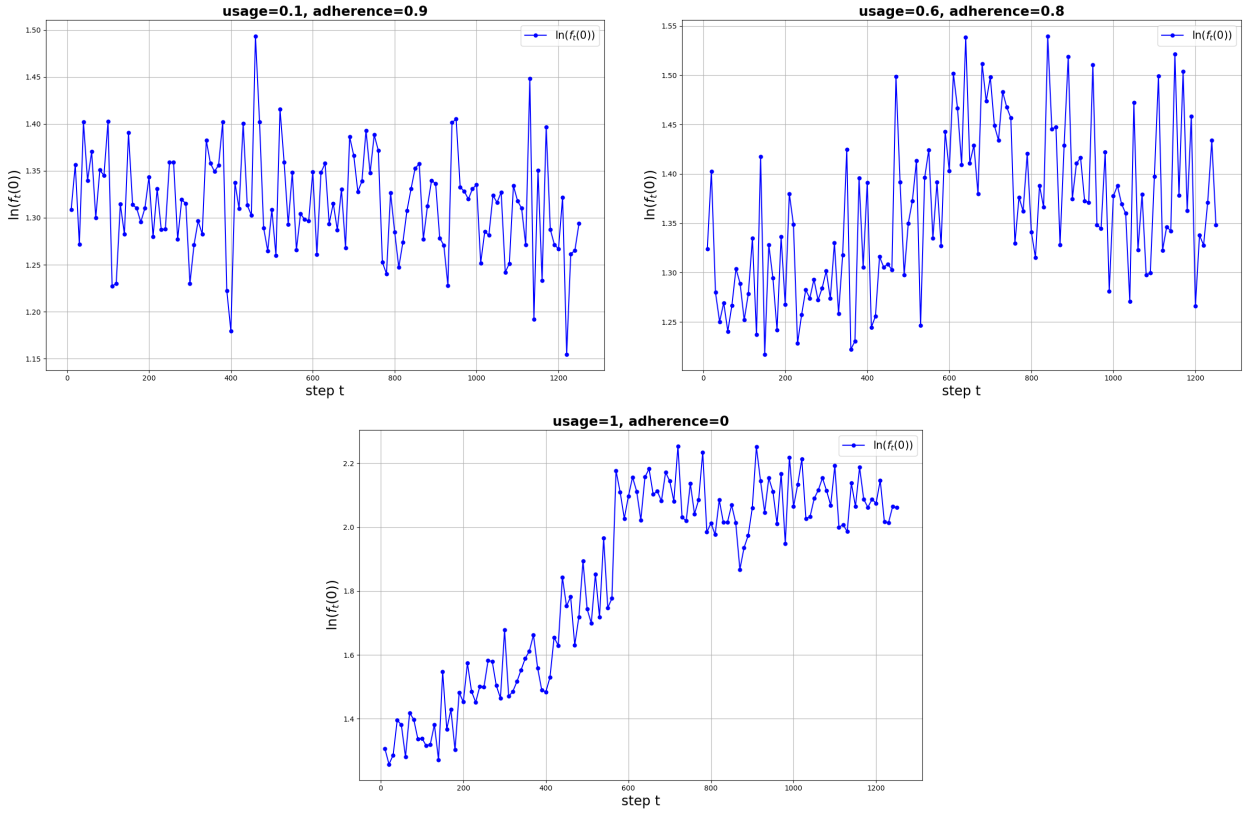

Figure 6: Results of experiment 5.5 *usage* and *adherence* = 0.1 and 0.9 (left), 0.6 and 0.8 (right), 1.0 and 0.0 (bottom).

5.6 Fifth experiment: decreasing in moments

In this experiment we test the results from Lemma 5. Based on the Experiment 5.3 we will approximate our distribution by a normal distribution. Based on Wick's theorem, we can explicitly find the values of moments for the normal distribution if its mathematical expectation is 0:

$$\nu_k = \begin{cases} 0, \text{ if } k = 2n+1 \\ \sigma^k \cdot (k-1)!!, \text{ if } k = 2n \end{cases} \tag{15}$$

The results of this experiment are shown in the Figure 7.

The moments decrease to zero only if *usage* and *adherence* = 1 and 0, because only with this parameters operator D transforms density distribution functions of $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ to delta function.

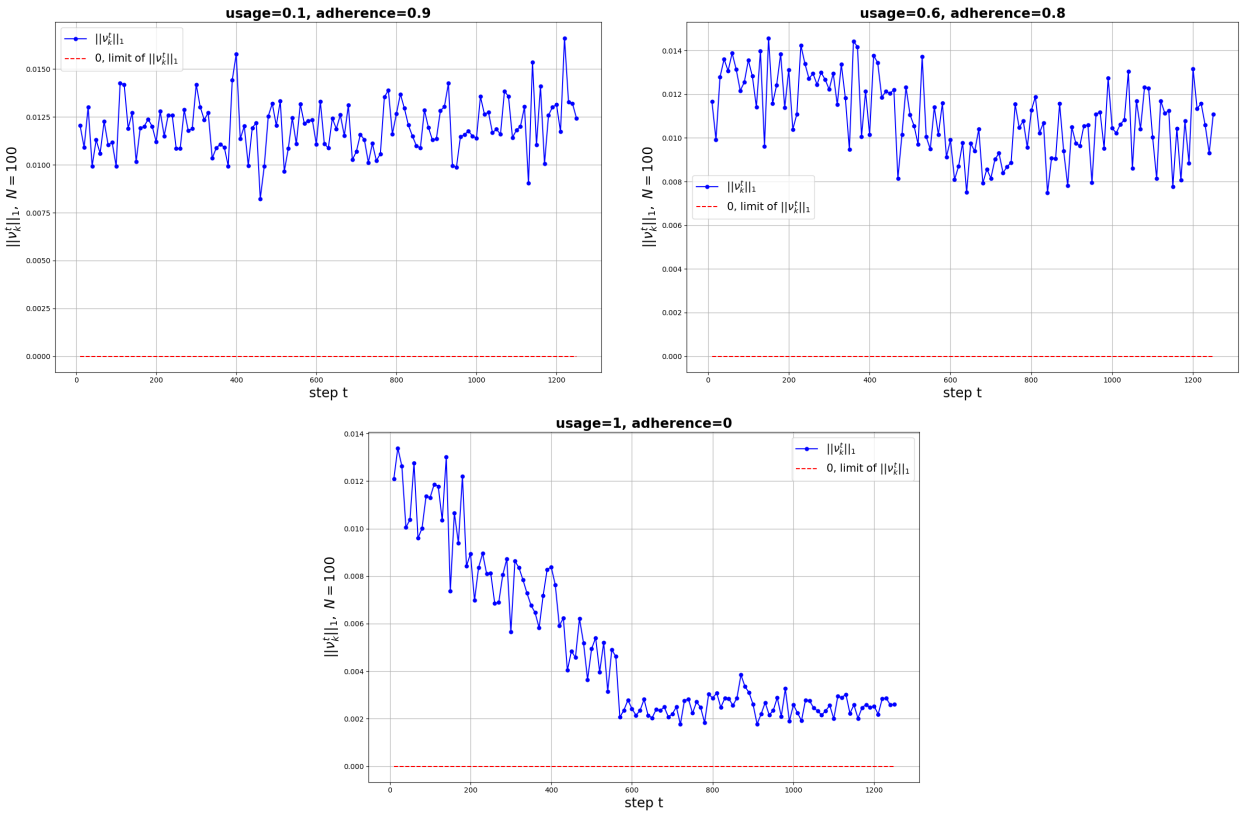

Figure 7: Results of experiment 5.6 *usage* and *adherence* = 0.1 and 0.9 (left), 0.6 and 0.8 (right), 1.0 and 0.0 (bottom).

References

[1] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In *Machine Learning for Healthcare Conference*, pages 710–731. PMLR, 2020.

[2] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[3] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pages 160–171. PMLR, 2018.

[4] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.

[5] Anatole Katok and Boris Hasselblatt. *Introduction to the modern theory of dynamical systems*. Number 54. Cambridge university press, 1995.

[6] Anton Khritankov. Hidden feedback loops in machine learning systems: A simulation model and preliminary results. In *Software Quality: Future Perspectives on Software Engineering Quality: 13th International Conference, SWQD 2021, Vienna, Austria, January 19–21, 2021, Proceedings 13*, pages 54–65. Springer, 2021.

[7] Anton Khritankov and Anton Pilkevich. Existence conditions for hidden feedback loops in online recommender systems. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22*, pages 267–274. Springer, 2021.

[8] Chao Ma, Lei Wu, et al. Machine learning from a continuous viewpoint, i. *Science China Mathematics*, 63(11):2233–2266, 2020.

[9] Viktor Vladimirovich Nemytskii. *Qualitative theory of differential equations*, volume 2083. Princeton University Press, 2015.

[10] E Pap, O Hadžić, and R Mesiar. A fixed point theorem in probabilistic metric spaces and an application. *Journal of Mathematical Analysis and Applications*, 202(2):433–449, 1996.

[11] Ayan Sinha, David F Gleich, and Karthik Ramani. Deconvolving feedback loops in recommender systems. *Advances in neural information processing systems*, 29, 2016.

[12] Dawid Tarłowski. Global convergence of discrete-time inhomogeneous markov processes from dynamical systems perspective. *Journal of Mathematical Analysis and Applications*, 448(2):1489–1512, 2017.

[13] Maylis Varvenne. Rate of convergence to equilibrium for discrete-time stochastic dynamics with memory. 2019.

[14] Anatolii Moiseevich Vershik. What does a typical markov operator look like? *Algebra i Analiz*, 17(5):91–104, 2005.