A Mathematical Model of the Hidden Feedback Loop Effect in Machine Learning Systems

Andrey Veprikov [1, 2] Alexander Afanasyev [2] Anton Khritankov [1, 2, 3]

[1] Moscow Institute of Physics and Technology [2] IITP RAS [2] HSE University

General Idea

This work solves the problem of mathematical modelling of systems with adaptive control. The system with artificial intelligence corresponds to a discrete dynamic system, the behaviour of which can be used to evaluate the original object. Link to full paper: https://arxiv.org/abs/2405.02726

Problem Statement. We consider a set $\mathbf{F}$ of probability density functions (PDFs), each of which describes the data available to a machine learning system at a given time step t. We then introduce a mapping $\mathrm{D}_t \in \mathbf{D}$ that acts on a given PDF $f_t(x) \in \mathbf{F}$ to produce a new data distribution $f_{t+1}(x)$. A general model of the repeated learning process we are studying can be written as

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

Examples of the repeated learning processes:

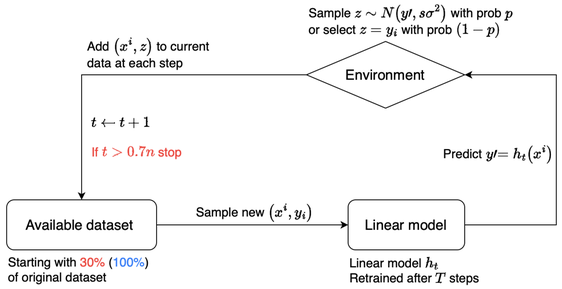

Figure: Two different experiments schemes. Sliding window update setup and sampling update setup.

Setting

- All operators $\mathrm{D}_t \in \mathbf{D}$ are transformations of the set $\mathbf{F}$, that is, for all $f(x) \in \mathbf{F}$ it holds that

$$\mathrm{D}_t(f)(x) \geq 0 \text{ for a.e. } x \in \mathbb{R}^n \text{ and } \int_{\mathbb{R}^n} \mathrm{D}_t(f)(x)dx = 1.$$

- At each step t the operator D_t can be different, we only assume that all D_t belong to some set $\mathbf{D}$.
- The D_t operators are not always computable, i.e., we can only observe samples from the distribution generated by the probability density function $f_{t+1}(x)$

Main Contributions

- We construct a mathematical model of the effect of feedback loops using discrete dynamical systems.
- We obtain results for finding the limit set of the dynamical system, sufficient conditions for the existence of a feedback loop and the autonomy criterion.
- We developed a bench of computational experiments simulating the process of repeated machine learning.

Related Work

Dynamical Systems. An important concept in the theory of dynamical systems is the so called minimal set. Since the set of density functions is compact in the $\|\cdot\|_1$-norm (if the discrete distributions are included there), the considered dynamical system must have at least one minimal set, since it is positively Lagrangian stable. In this paper we find the set of so called ω limit points for the considered system, which includes the minimal set.

Iterated Maps. An area in which the consistent application of different functions is considered is iterated map. The main object of study in this area is compressive mappings, which can be used to find fixed points. However, the restriction that D_t operators are not always computable makes it unfeasible to apply this theory for our problem.

Markov Decision Process. In this area authors consider such objects as Markov kernel, stationary distribution of the Markov chain, time-independent transition matrices. However, the same problems as when considering dynamical systems and iterated maps arise in this subject, since for example, for finding a stationary distribution, ability to computing the Markov kernel is necessary.

Feedback Loop. When there is a high automation bias, that is, when the use of predictions is high and adherence to them is tight, a so-called positive feedback loop occurs. As a result of the loop, the learning algorithm is repeatedly applied to the data containing previous predictions. This repeated learning produces a noticeable unintended shift in the distributions of the input data and the predictions of the system. For example, in systems that recommend products to consumers or forecast market prices and learn from user responses, healthcare decision support systems, and predictive policing and public safety systems that introduce bias in the training data as a result of an unintended feedback loop.

Results for a General System

Theorem 1 (Limit set)

Consider that D_t is a transformation of the set $\mathbf{F}$ for all $t \in \mathbb{N}$ and for any probability density function $f_0(x), x \in \mathbb{R}^n$ and discrete dynamical system (1), if there exists a measurable function $g(x)$ from $L_1(\mathbb{R}^n)$ and a non-negative sequence $\psi_t \geq 0$ such that $f_t(x) := \mathrm{D}_{\overline{1,t}}(f_0) \leq \psi_t^n \cdot |g(\psi_t \cdot x)|$ for all $t \in \mathbb{N}$ and $x \in \mathbb{R}^n$.

- Then, if ψ_t diverges to infinity, the density $f_t(x)$ tends to Dirac's delta function, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \delta(x)$ weakly.
- If ψ_t converges to zero, then the density $f_t(x)$ tends to a zero distribution, $f_t(x) \underset{t \to +\infty}{\longrightarrow} \zeta(x)$ weakly.

For the regression problem when the data have the form $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ Theorem 1 is stated not for the data in the AI system, but for a random vector of model h residuals of the form $\mathbf{y} - h(\mathbf{x})$.

Analysis of Results from Theorem 1

From Theorem 1 we can presume that envelopes of our mappings $\mathrm{D}_{\overline{1,t}}$ are in the form

$$\mathrm{D}_{\overline{1,t}}(f_0)(x) = \psi_t^n \cdot f_0(\psi_t \cdot x) \quad \forall x \in \mathbb{R}^n \text{ and } \forall t \in \mathbb{N}. \qquad (2)$$

When ψ_t converges to a constant $c \in (0, +\infty)$, then according to equation (2) the distribution of our data remains the same, that is the mapping $\mathrm{D}_{\overline{1,t}}$ is an identity mapping after some time step in the process.

If we substitute $x = 0$ into the equation (2), then we can get an expression for ψ_t: $\psi_t = \sqrt[n]{f_t(0)/g(0)}$. Let us take $\kappa > 0$ and consider an integral of the form

$$J_t := \int_{B^n(\kappa)} f_t(x)dx = \int_{B^n(\kappa)} \psi_t^n \cdot f_0(\psi_t \cdot x)dx = \int_{B^n(\kappa \cdot \psi_t)} f_0(y)dy,$$

If ψ_t diverges to infinity, then J_t converges to $\|f_0\|_1 = 1$, and if ψ_t converges to zero, then J_t will also converge to zero. In the experiments we measure $\psi_t \cong f_t(0)$ and $J_t \approx \hat{F}_t(\kappa) - \hat{F}_t(-\kappa)$, where $\kappa > 0$ is sufficiently small.

Corollaries of Theorem 1

Corollary 1 (Convergence rate)

For any $q \geq 1$, under conditions of Theorem 1, if $g(x) \in L_q(\mathbb{R}^n)$ and ψ_t converges to zero, it holds that

$$\|f_t(x) - \zeta(x)\|_q \leq (\psi_t^n)^{1-1/q} \cdot \|g\|_q.$$

Corollary 2 (Decreasing moments)

If a system (1) with $n = 1$ satisfies the conditions of Theorem 1 and ψ_t diverges to infinity, then for all $k \in \mathbb{N}$ it holds that

$$\mathbb{E}_{\xi \sim f_t(x)}\left[\xi^{2k}\right] \leq \psi_t^{-2k} \cdot \mathbb{E}_{\xi \sim f_0(x)}\left[\xi^{2k}\right].$$

Results for an Autonomous System

Theorem 2 (Autonomy criterion)

If the evolution operators D_t of a dynamic system (1) have the form (2), then the system is autonomous if and only if

$$\psi_{\tau+\kappa} = \psi_\tau \cdot \psi_\kappa \ \forall \tau, \kappa \in \mathbb{N}. \qquad (3)$$

This criterion is easy to check in practice, since the condition (3) means that the sequence ψ_t is a power sequence, that is $\psi_t = a^t$ for some $a > 0$. An example of a mapping of the form (2) is given in this work with the name Sampling update setup.

Limit to Delta Function or Zero Distribution

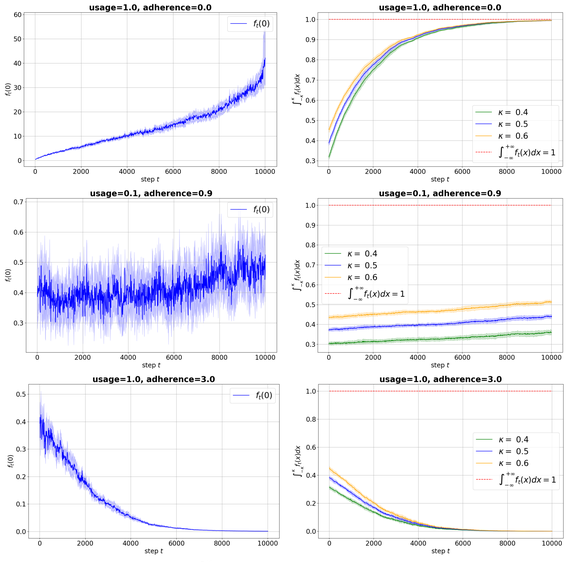

Figure: Counting $f_t(0)$ and $\int_{-\kappa}^{\kappa} f_t(x)dx$ for sampling update setup. We consider such parameters: usage, adherence $= 1, 0$ (first); $0.1, 0.9$ (second); $1, 3$ (third).

As we can see, if usage $p = 1$ and adherence $s = 0$, the limiting probability density of $\mathrm{D}_{\overline{1,t}}(f_0)$, that is the probability density of $\mathbf{y} - h(\mathbf{x})$, is delta function $\delta(x)$.

When usage $p = 0.1$ and adherence $s = 0.9$, the probability density of $\mathbf{y} - h(\mathbf{x})$ remains almost the same, that is $\psi_t \to c \in (0; +\infty)$.

If usage $p = 1$ and adherence $s = 3$ we observe a tendency to the zero distribution $\zeta(x)$.

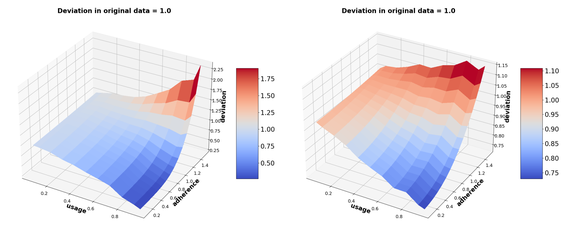

Figure: Change in the standard deviation of the model error for different usage and adherence. Sliding window setup (left), sampling update setup (right).

The graph is almost everywhere either red or blue, hence Theorem 1 is applicable in practice.

Autonomy Check

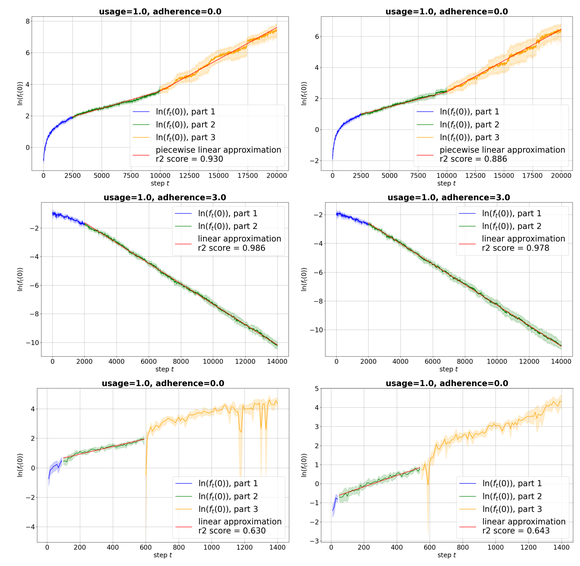

Figure: Testing two designs for autonomy. Sampling update setup (first and second) and sliding window setup (third).

As you can see, in case of the sliding window update we obtain a poor fit on all models and data sets, so the system is not autonomous. The sampling update setup in case of usage $p = 1$ and adherence $s = 3$ is autonomous on all models and data sets, since there is a good fit. In case of usage $p = 1$ and adherence $s = 0$ we observe two linear segments.

Decreasing Moments

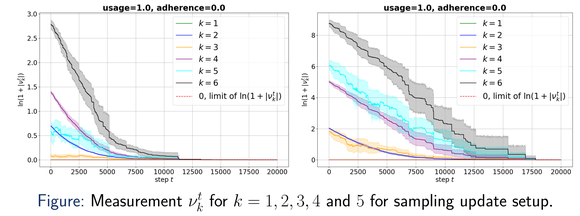

Figure: Measurement ν_k^t for $k = 1, 2, 3, 4$ and 5 for sampling update setup.

As we can see from the measurements, claim of Corollary 2 is satisfied in all observed cases. When usage $p = 1$ and adherence $s = 0$ the limit of mappings $\mathrm{D}_{\overline{1,t}}(f_0)$, and correspondingly of $\mathbf{y} - h(\mathbf{x})$, is the delta function $\delta(x)$.