On the problem of repeated supervised learning

My first scientific paper

Andrey Veprikov   Anton Khritankov   Alexander Afanasyev

MIPT
Dolgoprudny, Russia

Spring 2023

Introduction and Related work

In this research paper, we delve into the intricacies of continuous learning artificial intelligence systems as they interact with and influence their environment. Many authors have dealt with the problem from a narrative perspective [2, 7], i.e. so far only the problem of multiple learning has been articulated, but nowhere in the literature is a robust analysis of the process presented.
Repeated supervised learning appears in many machine learning applications, for example in

1. recommendation systems [6]
2. healthcare [1]
3. predictive policing [3]

Contributions of this paper are as follows

1. Develop a mathematical model to examine the process of repeated supervised learning, prediction, and dataset updating in presented cases.
2. Conduct several synthetic experiments based on our findings, hoping to contribute to a better understanding of the behavior of continuous learning AI systems.

Problem statement

The object of our research will be the set **R** of distribution density functions

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

and mappings $\mathrm{D}_t \in \mathbb{D}$ as feedback loop mapping. We consider discrete dynamical system [4]:

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

If $\{\mathrm{D}_t\}_{t=0}^{\infty}$ are independent of time, then (1) takes the form

$$f_{t+1}(x) = \mathrm{D}(f_t)(x), \quad \text{for } \forall x \in \mathbb{R}^n, t \in \mathbb{N} \tag{2}$$

According to Conjecture 1 from [5] the positive feedback loop in a system (1) exists if

$$\forall f, g \in \mathbf{R} \hookrightarrow \rho(R_H(\mathrm{D}(f)), R_H(\mathrm{D}(g))) \leq \lambda \cdot \rho(R_H(f), R_H(g)),$$

where R_H is quality measure of model H, $\rho(\cdot, \cdot)$ is a distance metric and $\lambda \in (0; 1)$.

Problem statement

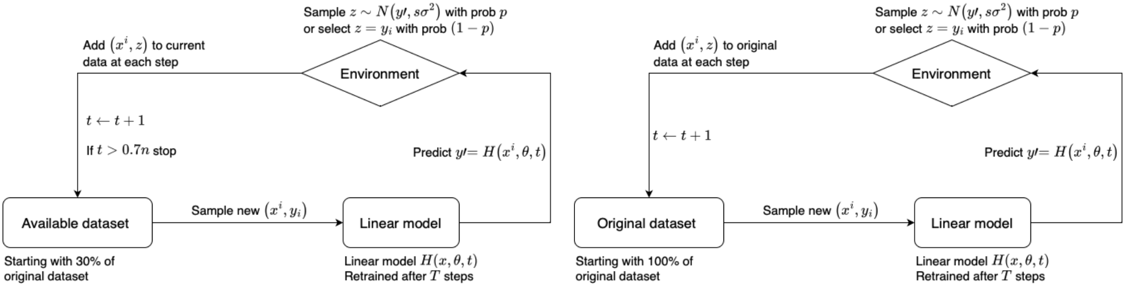


Figure: Simulations of a real systems from [5].

Assumptions for D : $\mathbf{R} \to \mathbf{R}$

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.

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

Limit in a weak sense to δ or zero function

Theorem 3

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

$$\exists \psi : \mathbb{N} \to \mathbb{R} : \psi(t) \underset{t \to +\infty}{\longrightarrow} +\infty \text{ and } \exists g \in L_1(\mathbb{R}) \text{ such that}$$

$$\forall t \in \mathbb{N}\ \forall y \in \mathbb{R} \hookrightarrow f_t\left(\frac{y}{\psi(t)}\right) \leq \psi(t) \cdot |g(y)| \tag{3}$$

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

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

where ϕ is continuous function with compact support

If $\psi(t) \underset{t \to +\infty}{\longrightarrow} 0$, then $f_t(x) \underset{t \to \infty}{\longrightarrow} 0$.

We consider the one-dimensional case because we can compare each operator D to an operator $\widetilde{\mathrm{D}}$, where $\widetilde{\mathrm{D}}$ transforms density distribution functions of random variables $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$

Important example of ψ and g

Important example of operators D_t for which the following is fulfilled $\mathrm{D}_t(f_0)(x) \xrightarrow[t\to\infty]{} \delta(x)$ is as follows

$$\mathrm{D}_t(f_0)(x) = t \cdot f_0(t \cdot x)$$

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

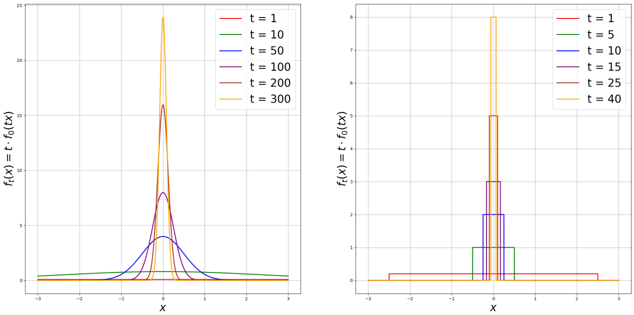


Illustration of weak limit to δ function. $\mathcal{N}(0, 5)$ left, $\mathcal{U}[-2.5, 2.5]$ right.

Consequences of the Theorem 3

Corollary 1 (Experimental ψ determination)

If $\forall x \neq 0 \hookrightarrow f_t(x) \underset{t\to\infty}{\longrightarrow} 0$ and $f_t(0) \underset{t\to\infty}{\longrightarrow} +\infty$, then we can take $\psi(t) = \dfrac{f_t(0)}{|g(0)|}$

In the following we assume that the operators $\mathrm{D}_t \in \mathbb{D}$ have the form

$$\mathrm{D}_t(f_0)(x) = \psi(t) \cdot f_0(\psi(t) \cdot x) \text{ and } \psi(t) \to +\infty \tag{4}$$

Corollary 2 (Autonomy criterion)

If a system satisfies (4), it is autonomous if and only if

$$\psi(\tau + \kappa) = \psi(\tau) \cdot \psi(\kappa) \; \forall \tau, \kappa \in \mathbb{N} \tag{5}$$

Corollary 3 (Decreasing moments)

If D_t have the form (4), then all k-th moments of random variable $\|\mathbf{y} - \mathbf{y}_{\text{pred}}\|$ (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 $\{\nu_k^t\}_{k=1}^{+\infty} \xrightarrow[t\to\infty]{l_1} 0$

Inequality on $\|\mathrm{D}\|_q$

Theorem 4

Consider

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

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 $\mathrm{D}^t(f_A) \in L_q(\mathbb{R}^n)$ is fulfilled that

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

This theorem gives a sufficient condition that the operators D_t wouldn't be a contraction mapping in $\|\cdot\|_q$.

Experiment setup

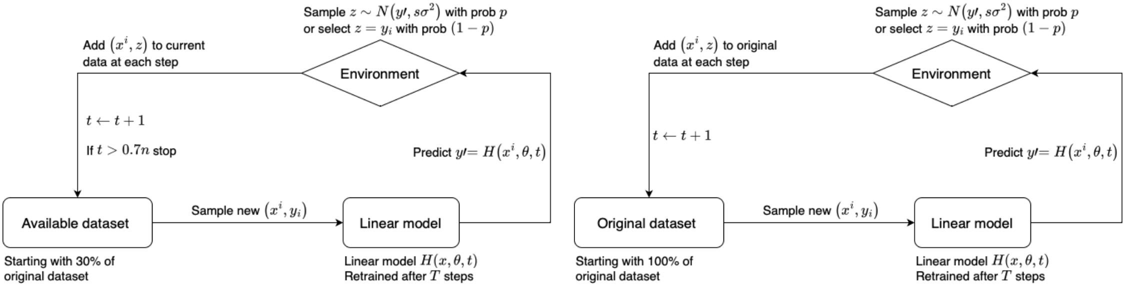


Figure: Sliding window setup (left) and sampling setup (right).

Analysis of deviation

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 $(\mathbf{x}^{\mathbf{i}}, z_i)$ into the active dataset, and *adherence* – the parameter by which we multiply σ^2 when sampling z_i.

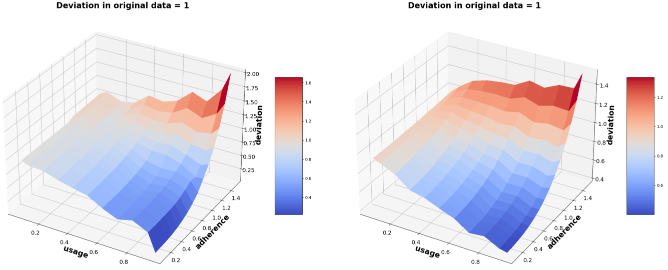
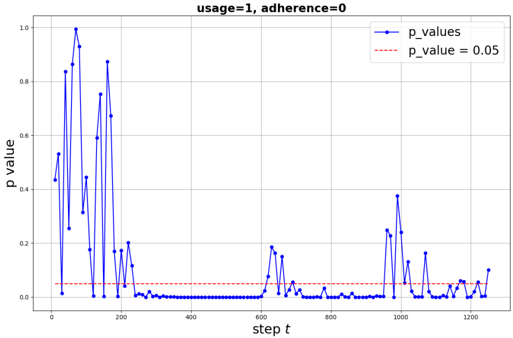


Figure: Analysis of deviation. Sliding window (left), sampling update (right).

As you can see, as a increases *usage* and decreases *adherence* deviation falls, this is because we start adding less noisy data to the active dataset.

Exploring the form of the data distribution over time

We check our data for belonging to a normal distribution by counting p-value from normal test.



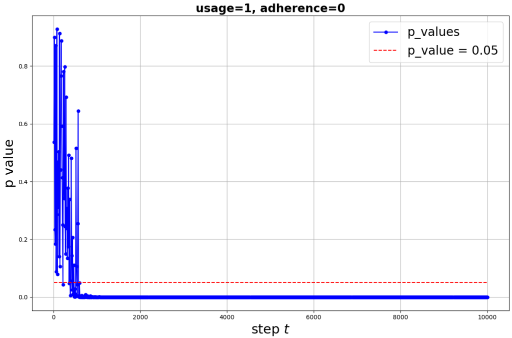


Figure: Analysis of p-value. Sliding window (left), sampling update (right).

As you can see, the normality of the data breaks down as t increases.

Histograms

As you can see, $\mathbf{y} - \mathbf{y}_{\text{pred}}$ is not normally distributed for all experiment types.

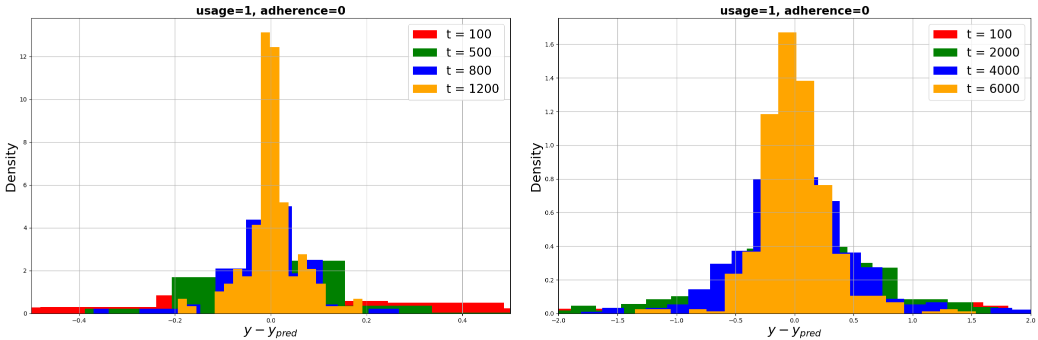


Figure: Histograms for some t. Sliding window (left), sampling update (right).

As you can see $\mathbf{y} - \mathbf{y}_{\text{pred}}$ seems to be a mixture of the two distributions

Limit to delta function

In this experiment we test the conditions from Theorem 3, i.e. we measure $f_t(0)$ and $\int\limits_{-\kappa}^{\kappa} f_t(x)dx$, where κ is sufficiently small.

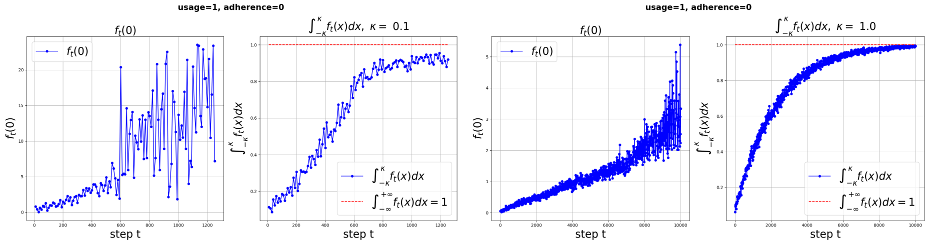


Figure: Limit to delta function. Sliding window (left), sampling update (right).

1. Sliding window: $f_t(0)$ rises initially, then goes into a plateau.
2. Sample update: $f_t(0)$ rises at all t.

Autonomy check

In this experiment we test our system for autonomy, that is, we test the condition (5) in the Corollary 2. According to (5) $\psi(t)$ should be a power function, so we measured $\ln(f_t(0))$ and r2-score of it.

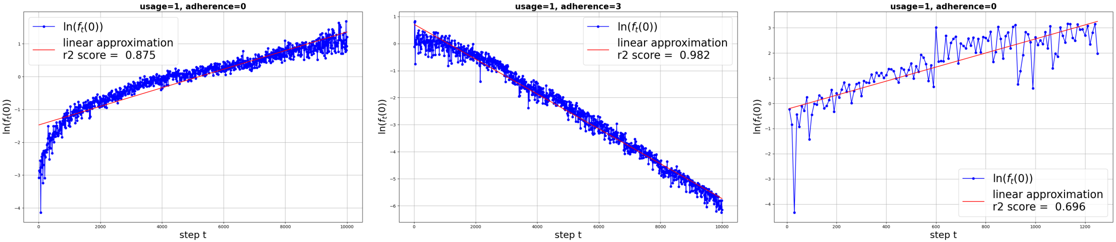


Figure: Autonomy check. Sampling update (left and middle) and sliding window (right).

1. Sample update: r2-score is 0.875 and 0.982, so we can consider this system as autonomic
2. Sliding window: r2-score is 0.696 and points after $t > 600$ don't look like a line, so we can't consider this system as autonomic

Decreasing moments

In this experiment we test the results from Corollary 3. In sampling update experiment we checked the first term of this Lemma, and in sliding window experiment we checked the second term.

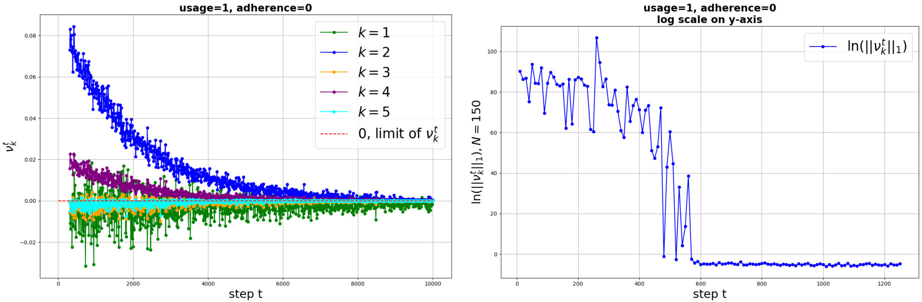


Figure: Decreasing moments. Sampling update (left) and sliding window (right)

As you can see, the moments tend to 0, since the condition that $D^t(f_0)(x) \underset{t\to\infty}{\longrightarrow} \delta(x)$ is fulfilled

Conclusion

1. In this paper, we applied the apparatus of discrete dynamical systems, mathematical analysis and probability theory to build a theoretical framework in the problem of repeated supervised learning
2. All results obtained are strongly proven and tested in a computational experiment
3. We made a simulated environment in the Python language to build a simulation model of repeated supervised learning

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] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

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

[5] 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.

[6] 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.

[7] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.