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. Kritanlov
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 [7, 6, 3]. 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 [6, 10], healthcare [1] and predictive policing [2].

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.

A range of problems, that we are considering in this paper are conditions under which D translates $\mathbf{R}$ into $\mathbf{R}$, conditions when D is a contraction mapping and its fixed point existence conditions.

Main contributions ...

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 [5, 6] – 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 [5] 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 [4, 8], 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 [11, 13]. But in Markov chain the future of the process does not depend on the past. Other authors [12, 9] studied stochastic dynamics systems in general.

An important contribution of this paper is that ...

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 this paper we consider the following research questions:

1. What are the conditions so that the mapping D becomes a transformation $\mathbf{R}$ into $\mathbf{R}$?
2. Under what conditions the mapping D is a contraction mapping?
3. Under what conditions the mapping D has a fixed point?

Let's analyze the importance of each question. We consider D as operator of changing distribution of our dataset, so it has to transform $\mathbf{R}$ into $\mathbf{R}$. If D is a contraction mapping, then for any starting function f_0 after a sufficiently large T, the result of our algorithm after T steps would be $D^T(f_0) = f^*$, where f^* is a fixed point, so after T steps metrics of our algorithm cannot be improved. If D has a fixed point f, when if $f_0 = f$ and we cannot improve our metrics as well.

4 Main results

In this section we will provide several theorems on operator D to answer questions 1-3.

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:

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

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 1 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 $x \in \mathbb{R}^n$.

□

In our experiment it often difficult to calculate D^{-1} and especially it's norm, so we make a different assumptions. We assume that our data sample describes the distribution function sufficiently well, so approximating the density function by our sample closely enough approaches the true density function of the data distribution. And if we consider D as algorithm for training a model, forming predictions on a test sample and mixing predictions into a training sample, then $\mathrm{D}: \mathbf{R} \to \mathbf{R}$ by constructing our density functions of data distributions.

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*

$$\exists \psi: \mathbb{N} \to \mathbb{R} \ \textit{such that} \ \psi \uparrow +\infty \ \textit{i.e.} \ \psi(t+1) \geq \psi(t) \ \forall t \in \mathbb{N} \ \textit{and} \ \psi(t) \underset{t \to +\infty}{\longrightarrow} +\infty \tag{2}$$

And

$$\exists g \in L_1(\mathbb{R}) \ \textit{such that} \ \forall t \in \mathbb{N} \, \forall y \in \mathbb{R} \hookrightarrow f_t\left(\frac{y}{\psi(t)}\right) \leq \psi(t) \cdot |g(y)| \tag{3}$$

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

$$\lim_{t \to +\infty} \left(\int_{-\infty}^{+\infty} f_t(x)\phi(x)dx \right) = \phi(0), \tag{4}$$

where ϕ is continuous function with compact support

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

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

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

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

$$I_t - \phi(0) = \frac{1}{\psi(t)} \cdot \int\limits_{-\infty}^{+\infty} f_t\left(\frac{y}{\psi(t)}\right) \cdot \left[\phi\left(\frac{y}{\psi(t)}\right) - \phi(0)\right] dy \tag{5}$$

Split the integral (5) 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

$$|I_1| \leq \int\limits_{-\infty}^{-A} 2M \cdot \frac{1}{\psi(t)} f_t\left(\frac{y}{\psi(t)}\right) dy \leq [(3)] \leq \int\limits_{-\infty}^{-A} 2M \cdot |g(y)| dy$$

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 $|I_1| \leq \varepsilon$.

Consider the integral I_2. ϕ is continuous and $\psi \uparrow +\infty$ (2), 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

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

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.

□

Let's analyze results of this theorem. Assume that distribution of our data does not change with t, i.e. operator D changes only the distribution of the target variable y. If in some step t the model 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 $(\mathbf{x^i}, y_i)$ become linearly dependent. So, if density distribution function of $\mathbf{y} - \mathbf{y}_{\text{pred}}$ becomes δ function.

Let's analyze formula (3). If we take $x = \phi(x)$ then (3) takes form

$$\exists g \in L_1(\mathbb{R}) \text{ such that } \forall t \in \mathbb{N}\, \forall x \in \mathbb{R} \hookrightarrow f_t(x) \leq \psi(t) \cdot |g(x \cdot \psi(t))| \tag{6}$$

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 $f_t(0) \uparrow +\infty$.

If we substitute $x = 0$ in the (6) then we get $f_t(0) \leq \psi(t) \cdot |g(0)|$, so we can take

$$\psi(t) = \frac{f_t(0)}{|g(0)|} \tag{7}$$

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

$$\mathrm{D}^t(f_0)(x) = t \cdot f_0(t \cdot x) \tag{8}$$

Here we take $g(x) = f_0(x)$ and $\psi(t) = t$. Let's look how this operator translates density functions of normal distribution $\mathcal{N}(0, 5)$ and continuous uniform distribution $\mathcal{U}[-2.5, 2.5]$:

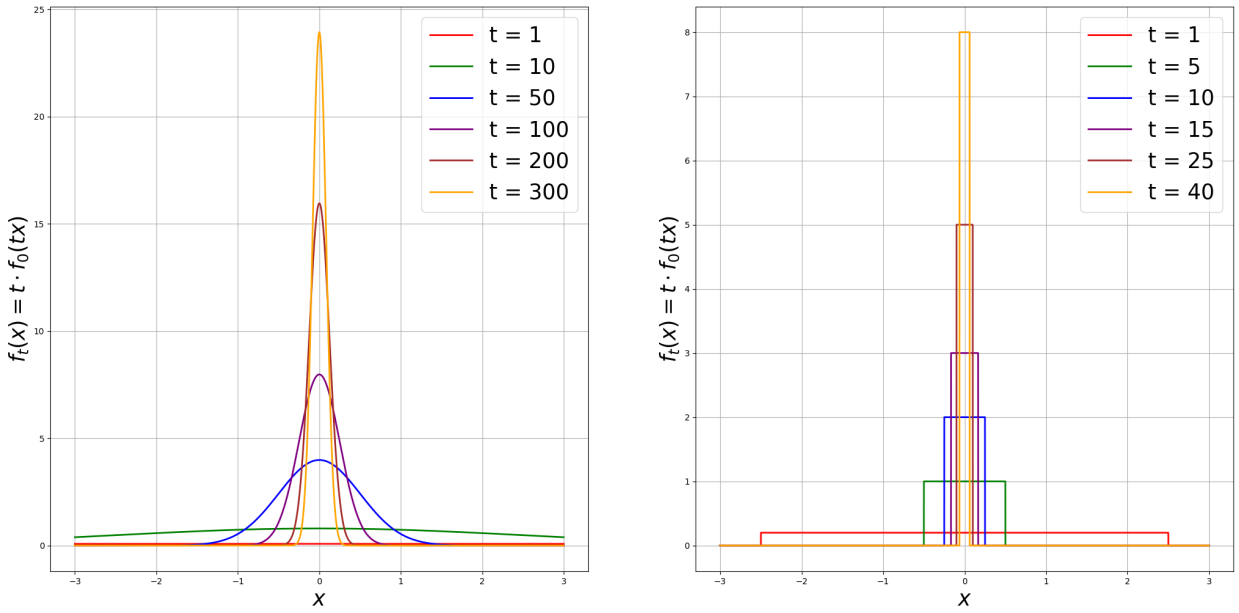

Figure 1: Illustration of weak limit to δ function. $\mathcal{N}(0, 5)$ left, $\mathcal{U}[-2.5, 2.5]$ right.

5 Experiments

5.1 Experiment Design

We consider 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 $f(\mathbf{x}, \theta^0)$ 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 $(\mathbf{x^i}, y_i)$: $y'_i = f(\mathbf{x^i}, \theta^r)$ and sample $z_i \sim \mathcal{N}(y'_i, s \cdot \sigma^2)$, where s is an experiment parameter that indicates adherence. 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 $f(x, \theta)$ is retrained with 80% train size on active dataset.

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 [5] were author was detecting hidden feedback loops in machine learning systems.

In this paper we consider linear model, that is solved as Ridge 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.

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.

5.2 Experiment results

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. This 3-dimensional plot will illustrate our results

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] 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.

[3] 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.

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

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

[9] 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.

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

[11] 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.

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

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