Randomness in training algorithms

Josef Teichmann

ETH Zürich

Konstanz, July 2019

... how it started

- Deep Hedging (learn trading strategies): joint project with Hans Bühler, Lukas Gonon and Ben Wood at JP Morgan (2017).
- Deep Calibration (learn model parameters): joint project with Christa Cuchiero and Wahid Khosrawi-Sardroudi.

... how it started

- Deep Hedging (learn trading strategies): joint project with Hans Bühler, Lukas Gonon and Ben Wood at JP Morgan (2017).
- Deep Calibration (learn model parameters): joint project with Christa Cuchiero and Wahid Khosrawi-Sardroudi.

Randomness appears prominently in learning

- random initialization of network weights.
- stochastic gradient descent.
- dropouts.
- randomization of strategies for stochastic control problems or games.
- generic (random) architectures with depth often work surprisingly well.
- most radical appearance of randomness: reservoir computing – just learn the readout!

Goal of this talk is ...

- to present a non-standard perspective on approximation techniques in machine learning.
- interpret learning as target reaching.
- to connect this non-standard perspective to reservoir computing.
- to apply random projections to construct reservoirs and prove generalization results.
- Randomness appears twofold here: it avoids degenerate behavior and therefore creates expressiveness, or it allows to construct low dimensional replica of high dimensional expressive structures.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Goal of this talk is ...

- to present a non-standard perspective on approximation techniques in machine learning.
- interpret learning as target reaching.
- to connect this non-standard perspective to reservoir computing.
- to apply random projections to construct reservoirs and prove generalization results.
- Randomness appears twofold here: it avoids degenerate behavior and therefore creates expressiveness, or it allows to construct low dimensional replica of high dimensional expressive structures.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Goal of this talk is ...

- to present a non-standard perspective on approximation techniques in machine learning.
- interpret learning as target reaching.
- to connect this non-standard perspective to reservoir computing.
- to apply random projections to construct reservoirs and prove generalization results.
- Randomness appears twofold here: it avoids degenerate behavior and therefore creates expressiveness, or it allows to construct low dimensional replica of high dimensional expressive structures.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Goal of this talk is ...

- to present a non-standard perspective on approximation techniques in machine learning.
- interpret learning as target reaching.
- to connect this non-standard perspective to reservoir computing.
- to apply random projections to construct reservoirs and prove generalization results.
- Randomness appears twofold here: it avoids degenerate behavior and therefore creates expressiveness, or it allows to construct low dimensional replica of high dimensional expressive structures.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Goal of this talk is ...

- to present a non-standard perspective on approximation techniques in machine learning.
- interpret learning as target reaching.
- to connect this non-standard perspective to reservoir computing.
- to apply random projections to construct reservoirs and prove generalization results.
- Randomness appears twofold here: it avoids degenerate behavior and therefore creates expressiveness, or it allows to construct low dimensional replica of high dimensional expressive structures.

(joint works with Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Wahid Khosrawi-Sardroudi, Thomas Krabichler, Martin Larsson, and Juan-Pablo Ortega)

Controlled ordinary differential equations

It has been very fruitful in mathematical finance, economics, dynamical systems to consider discrete and continuous time (stochastic) systems under one roof ...

Consider a controlled ODE (CODE)

$$dX_t = V(X_t, \theta(t))du_t\,,\ X_s = x \in E$$

on some state space E and some set of general controls $t \mapsto \theta(t) \in \Theta$ and some linear controls u.

- if $u_t = \sum_{n \geq 1} 1_{\{n \leq t\}}$, then the above differential equation becomes a discrete dynamical system just concatenating (controlled) maps $x \mapsto x + V(x, u(n))$ up to terminal time.
- if $u_t = t$, then we have a standard controlled ordinary differential equation.

Controlled ordinary differential equations

It has been very fruitful in mathematical finance, economics, dynamical systems to consider discrete and continuous time (stochastic) systems under one roof ...

Consider a controlled ODE (CODE)

$$dX_t = V(X_t, \theta(t))du_t\,,\ X_s = x \in E$$

on some state space E and some set of general controls $t \mapsto \theta(t) \in \Theta$ and some linear controls u.

- if $u_t = \sum_{n \geq 1} 1_{\{n \leq t\}}$, then the above differential equation becomes a discrete dynamical system just concatenating (controlled) maps $x \mapsto x + V(x, u(n))$ up to terminal time.
- if $u_t = t$, then we have a standard controlled ordinary differential equation.

Neural networks

Neural networks are nowadays frequently used to approximate functions due to ubiquitous universal approximation properties.

Neural networks are just concatenations of shallow neural networks, which are of the form

$$V(x,u) = \sum_{l=1}^{N} \alpha_l \varphi(\langle \mu_l, x \rangle + \beta_l)$$

for some vectors $\alpha_j, \mu_j \in \mathbb{R}^m$ and numbers β_j, which we can consider controls u, and some activation function φ.

$\rightarrow$ Neural networks are CODEs with a readout layer, i.e. a linear map W acting on the solutions of CODE at terminal time. Depths appears as time.

What is supervised learning for CODE?

We have convinced ourselves that CODE of the form

$$dX_t = \sum_i V_i(X_t) du_t^i \,,\ X_0 = x \in E$$

can provide an abstract setting for deep feed forward neural networks.

Simplicity is a feature: notice how incredibly easy it is to write backpropagation in this setting.

What is supervised learning?

It appears as a target problem, i.e. consider a training data set $(x_l, y_l)_{l \in L}$, i.e. a finite subset of a graph of an $\mathbb{R}$-valued continuous function on some compact set.

Consider the following controlled ODE on $\mathbb{E}^L$

$$dZ_t = \sum_{i=1}^{d} \mathbf{V}_i(Z_t) du^i(t)\,,$$

where

$$(\mathbf{V}_i((x_l)))_l := V_i(x_l)$$

for $l \in L$ and $i = 1, \ldots, d$.

We search to minimize, e.g.,

$$\left\| \mathbf{W} Z_T - \mathbf{y} \right\|_2^2 = \sum_l \left\| W X_T^{x_l} - y_l \right\|_2^2 \to \min$$

over readouts W and controls u.

What is supervised learning?

Lift CODE to a transport equation on state space of smooth functions $f \in C^\infty(E;\mathbb{R})$

$$df_t(x) = \sum_i \langle V_i(x), \nabla f_t(x)\rangle du^i(t)\,,$$

then learning appears as

$$\left\| \operatorname{eval}_{\mathbf{x}}(f_T) - \mathbf{y} \right\|_2^2 =\!\!\rightarrow \min$$

for initial value $f_0 = \langle W, . \rangle$ again searching for optimal W and optimal controls u.

Why could randomness possibly matter for learning?

- In the view of the above light, learning appears as target problem of stacked CODE or of a controlled transport equation.
- There are, at least in finite dimension, criteria which explain when targets can always be reached by CODEs: Chow-Rashevsky theory.
- random vector fields satisy the assumptions of Chow-Rashevsky theory and guarantee reachability.
- However, the theory does not really apply, since we need it very general ...

Why could randomness possibly matter for learning?

- In the view of the above light, learning appears as target problem of stacked CODE or of a controlled transport equation.
- There are, at least in finite dimension, criteria which explain when targets can always be reached by CODEs: Chow-Rashevsky theory.
- random vector fields satisy the assumptions of Chow-Rashevsky theory and guarantee reachability.
- However, the theory does not really apply, since we need it very general ...

Why could randomness possibly matter for learning?

- In the view of the above light, learning appears as target problem of stacked CODE or of a controlled transport equation.
- There are, at least in finite dimension, criteria which explain when targets can always be reached by CODEs: Chow-Rashevsky theory.
- random vector fields satisy the assumptions of Chow-Rashevsky theory and guarantee reachability.
- However, the theory does not really apply, since we need it very general ...

Why could randomness possibly matter for learning?

- In the view of the above light, learning appears as target problem of stacked CODE or of a controlled transport equation.
- There are, at least in finite dimension, criteria which explain when targets can always be reached by CODEs: Chow-Rashevsky theory.
- random vector fields satisy the assumptions of Chow-Rashevsky theory and guarantee reachability.
- However, the theory does not really apply, since we need it very general ...

Chow-Rashevsky revisited

Let E be a convenient space. Consider a controlled ordinary differential equation (CODE), i.e.

$$X_t = x + \sum_{i=1}^{d} \int_s^t V_i(X_r) du^i(r), \tag{1}$$

where $V_i : E \to E$ are some smooth vector fields on E The *control* $u : \mathbb{R} \to \mathbb{R}^d$ is considered a smooth curve with values in $\mathbb{R}^d$. We shall specify an open set of controls $\mathcal{U}$ such that the above equation has solutions for all times.

E & U

For our purposes we shall consider a particular class of vector fields where CODEs admit solutions for all times (compare Filipovic-Teichmann (2003) and Hamilton (1983)):

Theorem

Let E be a convenient vector space, $A_1, \ldots, A_d$ bounded linear generators of (commuting) smooth groups, $W_1, \ldots, W_d$ smooth tempered Banach vector fields, and $u : \mathbb{R} \to \mathbb{R}^d$ a smooth control, then

$$X_t = x + \sum_{i=1}^{d} \int_s^t (A_i X_r + W_i(X_r)) du^i(r) \tag{2}$$

has a unique solution for all times for any initial value.

Evolutions

Additionally the solution map

$$(s, t, x) \mapsto \mathrm{Evol}_{s,t}(x)$$

is a well defined diffeomorphism which satisfies

$$\mathrm{Evol}_{s,t} \circ \mathrm{Evol}_{r,s} = \mathrm{Evol}_{r,t}$$

and $\mathrm{Evol}_{s,s}(x) = x$ for all $r, s, t \in \mathbb{R}$ and $x \in E$.

Lie brackets

Let $V, W : E \to E$ be two smooth vector fields, then

$$[V, W] = dV \bullet W - dW \bullet V$$

is called the Lie bracket.

Definition

Let E be a convenient vector space and let $V_1, \ldots, V_d$ be smooth vector fields, then we denote the subspace of directions obtained by evaluating arbitrary linear combinations of Lie brackets at $x \in E$ by $\mathcal{D}(V_1, \ldots, V_d)(x)$.

Chow-Rashevsky revisited

Theorem (Cuchiero, Larsson, Teichmann (2019))

Let $\pi : E \to \mathbb{R}^m$ be a finite dimensional projection (i.e. surjective and linear) and let Evol *be a smooth solution of Equation* (1)*. Fix $x \in E$ and $s \neq t$ and assume that $\mathcal{D}(V_1, \dots, V_d)(x)$ is dense at $x \in E$ (Hörmander condition), then*

$$u \mapsto \pi \circ \mathsf{Evol}_{s,t}(x)$$

is locally surjective at an open dense set of controls u on $[s, t]$.

Specification of controls

Fix s. We denote by $\mathcal{U}_\omega$ the set of real analytic controls $u \in \mathcal{U} = C^\infty(\mathbb{R}, \mathbb{R}^d)$. We speak of a *random choice of controls* $u \in \mathcal{U}_\omega$ if the coefficients of the expansion at basis point s are chosen with respect to an infinite product of probability densities on $\mathbb{R}^d$.

Random controls classify

Given a fixed family of real numbers $a_{i_1 \dots i_k}$ indexed by $i_j = 1, \dots, d$, $j = 1, \dots, k$ and $k \geq 0$. Assume

$$\sum_{k=0}^{M} \sum_{i_1, \dots, i_k = 1}^{d} a_{i_1 \dots i_k} \int_{s \leq t_1 \leq \cdots \leq t_k \leq r} du^{i_1}(t_1) \cdots du^{i_k}(t_k) = \mathcal{O}\big((r-s)^{M+1}\big)$$

for all $M \geq 0$ and $r \to s$ for a random choice of $u \in \mathcal{U}_\omega$. Then $a_{i_1 \dots i_k} = 0$ identically.

Chow-Rashevsky revisited

Theorem (Cuchiero, Larsson, Teichmann (2019))

Let $\pi : E \to \mathbb{R}^m$ be a finite dimensional projection (i.e. surjective and linear) and let Evol *be a smooth solution operator linear in $x \in E$ of Equation* (1) *with linear vector fields. Fix $s \neq t$ and a non-zero finite dimensional subspace $L \subset E$ and assume that $\mathcal{D}(V_1, \ldots, V_d)(x)$ is dense in E for every $0 \neq x \in L$, then for every control u in a dense subset of controls the set*

$$(x, v) \mapsto \pi \circ \mathrm{Evol}^v_{s,t}(x) - \pi \circ \mathrm{Evol}^u_{s,t}(x)$$

is surjective for v in an open neighborhood of u.

Determine gradients

The first derivative D of the map

$$u \mapsto \pi \circ \mathrm{Evol}_{s,t}^{u}(x)$$

is surjective at a control u if and only if the quadratic form $C(u)$ defined through

$$\lambda \mapsto \sum_{i=1}^{d} \int_{s}^{t} \Big\langle \lambda, \pi\big(J_{s,t}(x) \bullet J_{r,s}(X_r) \bullet V_i(X_r)\big) \Big\rangle^2 dr$$

is positive definite (notice there that $X_s = x$). Assume now that $C(u)$ is positive definite and denote its matrix by $C(u)$, too. Then

$$\frac{d}{dr} a^i(r) = \langle \pi\big(J_{s,t}(x) \bullet J_{r,s}(X_r) \bullet V_i(X_r)\big), C(u)^{-1} v \rangle$$

defines a control such that $D(a) = v$.

Fix a control such that $C(u)$ is invertible. Then at least locally in λ the differential equation

$$\frac{\partial}{\partial \lambda}\frac{d}{dr}u_\lambda(r) = \langle \pi\big(J_{s,t}(x) \bullet J_{r,s}(X_r) \bullet V_i(X_r)\big), C(u_\lambda)^{-1} v\rangle$$

has a solution for $u_0 = u$ and $r \in [s,t]$ (notice that the Jacobians on the right hand side depend on u_λ, too) and we obtain

$$\frac{d}{d\lambda}\pi \circ \mathrm{Evol}_{s,t}^{u_\lambda}(x) = v$$

i.e. $\pi \circ \mathrm{Evol}_{s,t}^{u_\lambda}(x)$ moves with constant speed v away from its value at control u_0. In particular this means that one can explicitly construct a family of controls u_λ which reach a target in a star-shaped neighborhood of $\pi \circ \mathrm{Evol}_{s,t}^{u}(x)$.

An explicit construction for $d = 2$

We consider an explicit construction of $d = 2$ vector fields V_1, V_2 on $\mathbb{R}^m$ such that for a generic choice of distinct points $x_1, \ldots, x_L \in \mathbb{R}^m$ for any $L \geq 1$ the vector fields

$$W_i := \oplus_{j=1}^{L} V_i : \bigoplus_{j=1}^{L} \mathbb{R}^m \to \bigoplus_{j=1}^{L} \mathbb{R}^m$$

satisfy a Hörmander condition on $(x_1, \ldots, x_L) \in \bigoplus_{i=1}^{L} \mathbb{R}^m$.

An explicit construction for $d = 2$

Let σ be a polynomial in one variable of degree larger than 1 and consider two vector fields on $\mathbb{R}^m$

$$V_i(x) = \sigma(A_i x + b_i)$$

for $x \in \mathbb{R}^m$, where the components of the matrices A_i, b_i, $i = 1, 2$ are independently drawn with respect to a probability density on $\mathbb{R}$. Here σ denotes the componentwise application of σ on a vector in $\mathbb{R}^m$. Then for any choice of independently sampled $x_1, \ldots, x_L$, for any $L \geq 1$ with respect to a possibly different probability density, we obtain that W_1, W_2 satisfy a Hörmander condition at $(x_1, \ldots, x_L) \in \bigoplus_{j=1}^{L} \mathbb{R}^m$.

Why does randomness matter?

- Consider a controlled ODE

$$dX_t = \sum_{i=1}^{d} \sigma(A_i X_t + b_i) du_i(t)$$

in $\mathbb{R}^m$, or transport lifted on $C^\infty(\mathbb{R}^m; \mathbb{R})$, with 'generic' A_i and b_i, e.g. sufficiently random, then Chow-Rashevsky applies.

- This does explain why depths, a low amount of controllable parameters in each layer in combination with an overall randomness and a fully trainable readout layer can become very expressive.

Why does randomness matter?

- Consider a controlled ODE

$$dX_t = \sum_{i=1}^{d} \sigma(A_i X_t + b_i) du_i(t)$$

in $\mathbb{R}^m$, or transport lifted on $C^\infty(\mathbb{R}^m; \mathbb{R})$, with 'generic' A_i and b_i, e.g. sufficiently random, then Chow-Rashevsky applies.

- This does explain why depths, a low amount of controllable parameters in each layer in combination with an overall randomness and a fully trainable readout layer can become very expressive.

CODEs: control as input

In our previous view we consider maps

$$x \mapsto W\,\mathrm{Evol}_{s,t}(x)$$

trained by controls to reach targets. We could also fix $xinE$ and consider

$$u \mapsto W\,\mathrm{Evol}_{s,t}(x)$$

and just train the readout and/or the vector fields.

Does this also correspond to classes of networks? Yes: it generalizes rNNs, LSTMs, etc.

Used for time series, predictions, etc.

Reservoir Computing (RC)

... We aim to learn an input-output map on a high- or infinite dimensional input state space. Consider the input as well as the output as dynamic, e.g. a time series,

Paradigm of Reservoir computing (Herbert Jäger, Lyudmila, Grigoryeva, Wolfgang Maas, Juan-Pablo Ortega, et al.)

Split the input-output map into a generic part of generalized rNN-type (the *reservoir*), which is *not* trained and a readout part, which is trained.

Often the readout is chosen linear and the reservoir has random features. The reservoir is usually a numerically very tractable dynamical system.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

Applications of RC

- Often reservoirs can be realized physically, whence ultrafast evaluations are possible. Only the readout map W has to be trained.
- One can learn dynamic phenomena *without* knowing the specific characteristics.
- It works unreasonably well with generalization tasks.

An instance of RC in CODEs

Consider a controlled differential equation

$$dY_t = \sum_{i=1}^{d} V_i(Y_t) du_t^i, \ Y_0 = y \in \mathbb{R}^m$$

for some smooth vector fields $V_i : \mathbb{R}^m \to \mathbb{R}^m$, $i = 1, \ldots, d$ and d independent (Stratonovich) Brownian motions u^i, or finite variation continuous controls, or a rough path. This describes a controlled dynamics on $\mathbb{R}^m$.

We want to learn the dynamics, i.e. the map

(input control u) $\mapsto$ (solution Y).

Obviously a complicated, non-linear map, ...

We introduce some notation for this purpose:

Definition

Let $V : E \to E$ be a smooth vector field, and let $f : E \to \mathbb{R}$ be a smooth function, then we call

$$Vf(x) = df(x) \bullet V(x)$$

the transport operator associated to V, which maps smooth functions to smooth functions and determines V uniquely.

Theorem

Let Evol *be a smooth evolution operator on a convenient vector space* E *which satisfies (again the time derivative is taken with respect to the forward variable* t*) a controlled ordinary differential equation* (1)

$$d\,\mathrm{Evol}_{s,t}(x) = \sum_{i=1}^{d} V_i(\mathrm{Evol}_{s,t}(x))du^i(t)$$

then for any smooth function $f : E \to \mathbb{R}$

$$\begin{aligned} & f\left(\mathrm{Evol}_{s,t}\right) = \\ & = \sum_{k=0}^{M} \sum_{i_1,\ldots,u_k=1}^{d} V_{i_1} \cdots V_{i_k} f(x) \int_{s\le t_1\le\cdots\le t_k\le t} du^{i_1}(t_1)\cdots du^{i_k}(t_k) + \\ & + R_M(s,t,f) \end{aligned}$$

with remainder term

$$R_M(s,t,f) = \\ = \sum_{i_0,\dots,u_M=1}^{d} \int_{s\le t_1\le\cdots\le t_{M+1}\le t} V_{i_0}\cdots V_{i_k} f\big(\mathrm{Evol}_{s,t_0}(x)\big)\, du^{i_0}(t_0)\cdots du^{i_k}(t_M)$$

holds true for all times $s \le t$ and every natural number $M \ge 0$.

A lot of work has been done to understand the analysis, algebra and geometry of this expansion (Kua-Tsai Chen, Gerard Ben-Arous, Terry Lyons). It is a starting point of *rough path analysis* (Terry Lyons, Peter Friz, etc).

Definition

Consider the free algebra $\mathbb{A}_d$ of formal series generated by d non-commutative indeterminates $e_1, \ldots, e_d$. A typical element $a \in \mathbb{A}_d$ is written as

$$a = \sum_{k=0}^{\infty} \sum_{i_1,\ldots,i_k=1}^{d} a_{i_1 \ldots i_k} e_{i_1} \cdots e_{i_k} ,$$

sums and products are defined in the natural way. We consider the complete locally convex topology making all projections $a \mapsto a_{i_1 \ldots i_k}$ continuous on $\mathbb{A}_d$, hence a convenient vector space.

Definition

We define on $\mathbb{A}_d$ smooth vector fields

$$a \mapsto ae_i$$

for $i = 1, \ldots, d$.

Theorem

Let u be a smooth control, then the controlled differential equation

$$d\,\mathrm{Sig}_{s,t}(a) = \sum_{i=1}^{d} \mathrm{Sig}_{s,t}(a) e_i du^i(t)\,,\ \mathrm{Sig}_{s,s}(a) = a \tag{3}$$

has a unique smooth evolution operator, called signature of u and denoted by Sig*, given by*

$$\mathrm{Sig}_{s,t}(a) = a \sum_{k=0}^{\infty} \sum_{i_1,\ldots,u_k=1}^{d} \int_{s\le t_1\le\cdots\le t_k\le t} du^{i_1}(t_1)\cdots du^{i_k}(t_k)\; e_{i_1}\cdots e_{i_k}\,. \tag{4}$$

Theorem (Signature is a reservoir)

Let Evol *be a smooth evolution operator on a convenient vector space E which satisfies (again the time derivative is taken with respect to the forward variable t) a controlled ordinary differential equation* (1)

$$d\,\mathrm{Evol}_{s,t}(x) = \sum_{i=1}^{d} V_i(\mathrm{Evol}_{s,t}(x))du^i(t)\,.$$

Then for any smooth (test) function $f : E \to \mathbb{R}$ and for every $M \geq 0$ there is a time-homogenous linear $W = W(V_1, \ldots, V_d, f, M, x)$ from $\mathbb{A}_d^M$ to the real numbers $\mathbb{R}$ such that

$$f\big(\,\mathrm{Evol}_{s,t}(x)\big) = W\big(\pi_M(\mathrm{Sig}_{s,t}(1))\big) + \mathcal{O}\big((t-s)^{M+1}\big)$$

for $s \leq t$.

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

Signature as reservoir

- This explains that any solution can be represented – up to a linear readout – by universal reservoir, namely signature.
- This is used in many instances of provable machine learning by, e.g., groups in Oxford (Harald Oberhauser, Terry Lyons, etc), and also ...
- ... at JP Morgan, in particular great recent work on 'Nonparametric pricing and hedging of exotic derivatives' by Terry Lyons, Sina Nejad and Imanol Perez Arribas.
- in contrast to reservoir computing: signature is high dimensional (i.e. infinite dimensional) and a precisely defined, non-random object.
- Can we approximate signature by a lower dimensional random object with similar properties?

It is the assertion of the Johnson-Lindenstrauss (JL) Lemma that for every $0 < \epsilon < 1$ an N point set Q in some arbitrary (scalar product) space H, can be embedded into a space $\mathbb{R}^k$, where $k = \frac{24 \log N}{3\epsilon^2 - 2\epsilon^3}$ in an almost isometric manner, i.e. there is a linear map $f : H \to \mathbb{R}^k$ such that

$$(1 - \epsilon)\|v_1 - v_2\|^2 \leq \|f(v_1) - f(v_2)\|^2 \leq (1 + \epsilon)\|v_1 - v_2\|^2$$

for all $v_1, v_2 \in Q$. It is remarkable that f can be chosen randomly from a set of linear projection maps and the choice satisfies the desired requirements with high probability.

The result is due to concentration of measure results in high dimensional spaces and has been discovered in the eighties, for some details see below.

In order to make this program work, we need a definition:

Definition

Let Q be any (finite or infinite) set of elements of norm one in $\mathbb{A}_d^M$. For $v \in \mathbb{A}_d^M$ we define the function

$$\|v\|_Q := \inf\left\{ \sum_j |\lambda_j| \;\middle|\; \sum_j \lambda_j v_j = v \text{ and } v_j \in Q \right\}.$$

We use the convention $\inf \emptyset = +\infty$ since the function is only finite on $\mathrm{span}(Q)$. Actually the function $\|.\|_Q$ behaves precisely like a norm on the span. Additionally $\|v\|_{Q_1} \geq \|v\|_{Q_2}$ for $Q_1 \subset Q_2$ and $\|v\|_Q \geq \|v\|$ for all sets Q of elements of norm one.

Proposition

Fix $M \geq 1$, $\epsilon > 0$ and consider the free nilpotent algebra $\mathbb{A}_d^M$. Let Q be any N point set of vectors with norm one, then there is linear map $f : \mathbb{A}_d^M \to \mathbb{R}^k$ (k being the above JL constant with N), such that

$$\left|\langle v_1, v_2 - (f^* \circ f)(v_2)\rangle\right| \leq \epsilon\,,$$

for all $v_1, v_2 \in Q$. In particular

$$\left|\langle v_1, v_2 - (f^* \circ f)(v_2))\rangle\right| \leq \epsilon \|v_1\|_Q \|v_2\|_Q\,,$$

for $v_1, v_2 \in \mathbb{A}_d^M$.

Theorem (Cuchiero, Gonon, Grigoryeva, Ortega, Teichmann (2019))

Let u be a smooth control and f the previously constructed JL map associated to a spanning N point set Q of norm one. We denote by r-Sig *the smooth evolution of*

$$dZ_t = \sum_{i=1}^{d} f(f^*(Z_t)e_i)du^i(t)\,,\ Z_0 = f^*(1)$$

a controlled differential equation on $\mathbb{R}^k$. Then

$$\begin{aligned}
&\langle u, \mathrm{Sig}_{s,t}(1) - f^*(\text{r-Sig}_{s,t}(1))\rangle \\
&\leq \Big(\left|\langle \Gamma_{\mathrm{Sig}_{s,t}(1)}(u), 1 - f^*(f(1))\rangle\right| + \\
&+ C\epsilon \sum_{i=1}^{d} \int_s^t \|\Gamma_{\mathrm{Sig}_{r,t}(1)}(u)\|_Q \|f^*(Y_r)e_i\|_Q \, dr \Big)\,,
\end{aligned}$$

with constant $C = \sup_{s\leq r\leq t,\, i} \left|\frac{du^i(r)}{dr}\right|$, and for each $u \in Q$.

Corollary

Let u be a smooth control and f the previously constructed JL map associated to a spanning N point set Q of norm one. Assume additionally $1 = f * (f(1))$, then

$$\left\| \mathrm{Sig}_{s,t}(1) - f^*(\text{r-Sig}_{s,t}(1)) \right\| \leq$$

$$\left(\epsilon C \sum_{i=1}^{d} \int_s^t \sup_{\|u\|=1} \|\Gamma_{\mathrm{Sig}_{r,t}(1)}(u)\|_Q \, \|f^*(Y_r)e_i\|_Q \, dr \right).$$

Hence f^*(r-Sig) approximates Sig *up to order ϵ and can be used as a proxy for signature.*

r-Sig is a random dynamical system

It is fascinating that we can actually calculate approximately the vector fields which determine the dynamics of r-Sig, i.e.

$$y \mapsto f(f^*(y)e_i)$$

for each $i = 1, \ldots, d$ for $y \in \mathbb{R}^k$.

Theorem

For $M \to \infty$ the linear vector fields

$$y \mapsto f(f^*(y)e_i)$$

for $i = 1, \ldots, d$, are built from matrices on $\mathbb{R}^k$ with asymptotically normally distributed, independent entries.

Randomness matters

Consider

$$dY_t = \sum_{i=1}^{d} V_i(Y_t) du^i(t), \ Y_0 \in \mathbb{R}^m$$

where we observe *one* trajectory on $[0, T]$ and do not know the characteristics.

Randomized Signature

A random localized signature

- there is a set of hyper-parameters $\theta \in \Theta$, and a dimension M.
- depending on θ choose randomly matrices $A_1, \ldots, A_d$ on $\mathbb{R}^M$ as well as shifts $\beta_1, \ldots, \beta_d$ such that maximal non-integrability holds on a starting point $x \in \mathbb{R}^M$.
- one can tune the hyper-parameters $\theta \in \Theta$ and dimension M such that

$$dX_t = \sum_{i=1}^{d} \sigma(A_i X_t + \beta_i) du^i(t)\,,\; X_0 = x$$

locally (in time) approximates CODE Y via a linear readout W up to arbitrary precision.

What does ML give to MF?

- numerical evaluation of almost any thought experiment becomes feasible: we see 'solutions' of problems numerically which we have never seen before.
- new concepts of modeling.

What does ML give to MF?

- numerical evaluation of almost any thought experiment becomes feasible: we see 'solutions' of problems numerically which we have never seen before.
- new concepts of modeling.

What does MF give to ML?

- it is important to mathematize problems to support understanding.
- roles of randomness.

What does MF give to ML?

- it is important to mathematize problems to support understanding.
- roles of randomness.

Summary and Outlook

- construct universal and randomized reservoirs byond (branched) rough paths, for instance in the realm of regularity structures.
- improve the JL argument and adapt it to algebraic structures.

Summary and Outlook

- construct universal and randomized reservoirs byond (branched) rough paths, for instance in the realm of regularity structures.
- improve the JL argument and adapt it to algebraic structures.

References

- H. Bühler, L. Gonon, J. Teichmann, and B. Wood: *Deep Hedging*, Arxiv, 2018.
- C. Cuchiero, M. Larsson, J. Teichmann: *Controlled neural ordinary differential equations*, working paper, 2019.
- C. Cuchiero, L. Gonon, L. Grigoryeva, J.-P. Ortega, J. Teichmann: *Representation of Dynamics by randomized signatures*, working paper, 2019.
- T. Lyons, *Rough paths, Signatures and the modelling of functions on streams Terry Lyons*, Arxiv, 2014.