From mathematics through rough paths to data science

https://www.datasig.ac.uk

Terry Lyons

University of Oxford

DataSig

A rough path between mathematics and data science

The Alan Turing Institute

Imperial College London

Engineering and Physical Sciences Research Council

The mathematics of controlled systems - early days

The tools of integration, calculus, and (controlled) differential equations make possible the mathematical modelling of interaction in evolving systems. A massive contribution to science even without the stirring of tea!

> *Newton, Isaac, Methodus fluxionum et serierum infinitarum 1671; The Method of Fluxions and Infinite Series; with Its Application to the Geometry of Curve-lines... Translated from the Author's Latin Original Not Yet Made Publick. To which is Subjoin'd a Perpetual Comment Upon the Whole Work... by J. Colson. 1736.*

Already there in Newton's work and language:

- systems that evolve
  - Position (fluent)
  - Direction (fluxion)
- one solves for the evolution of
  - the fluents
- with a notion of control
  - Relate (dependent)
  - Correlate (independent)

Newtons controlled equations

It is remarkable how close Newton's work is to today.

- Using a power series (and the error function) he solved

$$dy = (1 - 3x + y + x\hat{}2 + xy)\, dx$$

- He used interacting controls and non-formulaic equations

$$\begin{aligned} dp^i &= \frac{Gm_i m_j \left(q^i - q^j\right)}{||q^i - q^j||^3} dt \\ dq^i &= p^i dt \end{aligned}$$

Controlled equations today

Todays language.

- The basic framework connecting the control, a path $x_u \in U$, and the response or system state, a path $y_u \in V$, is an equation

$$dy_u = f(y_u, x_u)\, dx_u$$

where f is a one form on U with values in the space of vector fields on V . The f describes the system.

- By fixing a chart, enhancing the state variable to $z = (x, y)$ and slightly modifying the equation f to $\tilde{f}(z)(\delta x) := (\delta x, f(z)(\delta x))$ one locally replaces the equation with a related one

$$dz_u = \tilde{f}(z_u)\, dx_u$$

where U is a vector space and $\tilde{f}$ is a linear map from U into the space of vector fields on V . Time dependence is dealt with in a similar way. This is the usual model rough path people study.

Reparameterisation invariance

There is a natural time reparameterisation invariance in this formulation

Lemma (Gauge Invariance)

Suppose that $x_u \in U$, and $y_u \in V$ and they solve the differential equation

$$dy_u = f(y_u)\, dx_u \tag{1}$$

and suppose further that that $\phi : \mathbb{R} \to \mathbb{R}$ is smooth, then

$$dy_{\phi(u)} = f\left(y_{\phi(u)}\right) dx_{\phi(u)}.$$

Historically there has been a tendency to think of the derivative $\dot{x}$ of the control x as "the control" and follow Newton and write $\dot{x}dt$ instead of dx; however, this hides the gauge invariance, forces the control to be differentiable when this is not mathematically appropriate, and stops one thinking of the controlled system (1) as a transformation on paths that can be chained.

Impulse controls, jumps, discrete data,

Impulses are events that happen quickly. There is a "jump" in x at t_0

Solution (Naive)

Use ϕ to add some "virtual time" between t_0^- and t_0^+ and linearly interpolate x at that time then solve

$$dy_{\phi(u)} = f\left(y_{\phi(u)}\right) dx_{\phi(u)}.$$

Then speed up the virtual movement and return to the original paramterisation with a jump.

This is canonical and may work well; it may be that something more complicated is dictated by the context.

Fact

There is no issue dealing with jumps in x - there is a modelling question - what happens to the system at a jump! Is the naive solution the correct one. Alternative interpolations, different dynamics at jumps, ...

Streamed Information is everywhere

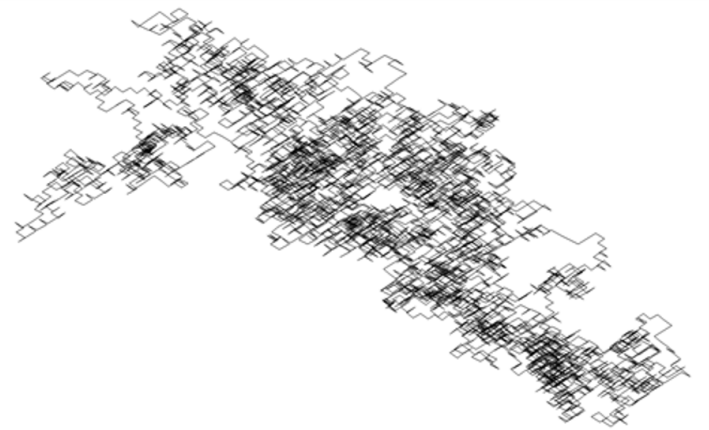
- can be syncronous
  - Wikepedia “A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.”
  - a video
  - fingerwriting on screen of mobile phone
  - evolving facial emotion
- or asynchronous
  - electronic health records
  - financial feed from different stocks
  - social data
- is not usually stationary in time

controlled equations to describing streams

- Unified approach to studying streams x_u that evolve
  - *measure* their effects as controls,
  - on special prototypical equations

$$dS_u = S_u \otimes dx_u, \;\; S_{t_0} = 1$$

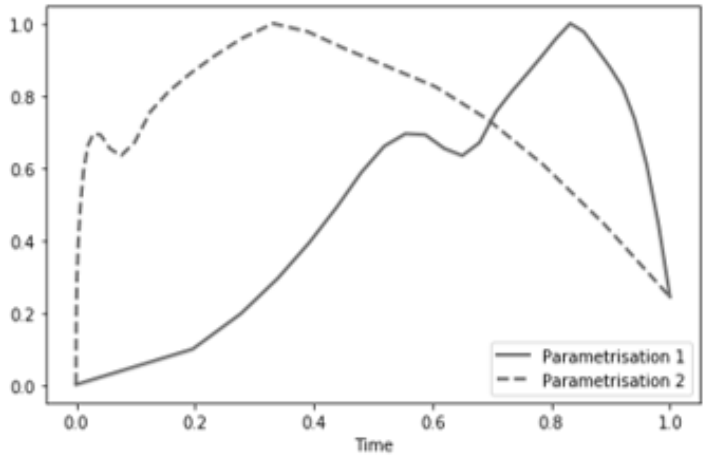
- The signature of the stream
  - get a tensor description $S_{t_1} \in T((U))$ of x over the interval $[t_0, t_1]$.
  - truncating this description to level one tensors would give a chordal description: the frequency of words.

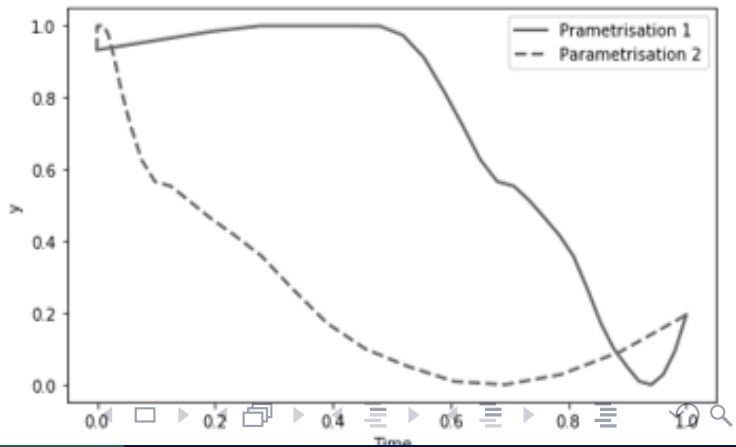


"Words differently arranged have a different meaning, and meanings differently arranged have different effects." Blaise Pascal, Pensées (1670)

unparamaterised paths - a gauge invariance

- The meaning of a path (x_u, u) does not depend on the speed of traversal
  - reparameterisation is a symmetry
  - the equivalence classes are not linear
  - wavelet transforms, fourier series are seriously challenged
- How does one describe a multdimensional signal up to reparameterisation?





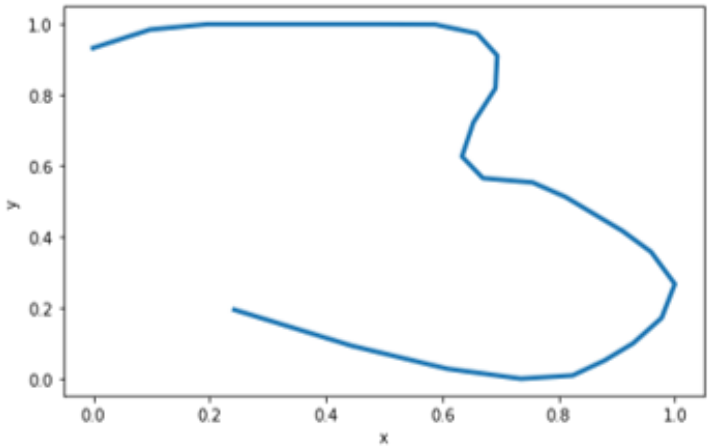
unparamaterised paths - a gauge invariance

- Return to those special prototypical equations

$$dS_u = S_u \otimes dx_u, \;\; S_{t_0} = 1$$

and get a tensor description $S \in T((U))$ of x over the interval $[t_0, t_1]$.

- We call S the signature of the path. It is a transform converting the path into a non-commutative group element described by a collection of definite iterated integrals.



The curve is insensitive to sampling rate. The signature gives a top down description that is insensitive to sampling rate.

Tensor algebras

- Let E be d-dimensional Banach space with basis $\mathcal{E} = \{e_1, \ldots, e_d\}$. Denote by

$$T(E) = \bigoplus_{k=0}^{\infty} E^{\otimes k}$$

and

$$T((E)) = \prod_{k=0}^{\infty} E^{\otimes k}$$

the spaces of formal polynomials, power series in the letters from $\mathcal{E}$.

  - Words with letters in $\mathcal{E}$ are a basis.
  - $T()$ and $T(())$ are functors from the category of vector spaces to that of algebras
  - moreover one can introduce norms, $T(E^*) \subset T((E))^*$ etc..

The signature as a universal feature set for unparameterised paths

- Suppose that $x \in E$ is a finite stream of information defined on $[s,t]$ evolving in the vector space E.
  - The signature is the solution of the universal CDE driven by x

$$dS(x)_{s,u} = S(x)_{s,u} \otimes dx_u\,, \qquad S(x)_{s,s} = \mathbf{1} = (1,0,0,\ldots)$$

$$S(x)_{s,t} = \mathbf{1} + \int_s^t S(x)_{s,u} \otimes dx_u, \qquad S(x)_{s,s} = \mathbf{1} = (1,0,0,\ldots)$$

  - Informs about the stream $x|_{[s,t]}$ through the response $S(x)_{s,t}$ of the exponential nonlinear system. (*meaning without maths*).
  - If E is the formal span of a finite alphabet $A = \{a_1,\ldots,a_n\}$ then $S \in T((E))$ the space of infinite formal linear combinations of words with letters drawn from A. The solution lives in a vector space. S is a feature map on streams!
  - S does not depend on the parameterisation of the path segment. It is a powerful nonlinear filter that removes sampling rate from data and **faithfully** preserves the order of events, the curve. Hambly, Lyons 2010 Chen 1958

The range of the signature is a group

- The tensor algebra is associative algebra containing U
  - Closed under multiplication

$$S(x)_{s,u} \otimes S(x)_{u,v} = S(x)_{s,v}$$

  - Path run backwards gives the inverse
- Defining $[u, v] := u \otimes v - v \otimes u$ makes $T((U))$ a Lie algebra; the smallest Lie subalgebra $L(U)$ containing U is free.
- Grouplike elements (Chen, Magnus, Bourbaki, Reutenauer, ...):

$$G = \exp L(U), \quad L(U) = \log G$$

  - The functions on a group are a linear space with a pointwise multiplication, the points of G are linear functionals on this space
  - Now $T(U^*)$ is dual to $T((U))$, and G is a set of linear functionals on it. Conversely $T(U^*)$,⧢ is a real abelian algebra of functions on G that separates points and contains the constants. We have recognised the functions on unparameterised paths.

The polynomial functions on paths

- Classical results then tell us
  - We can recognise a path through its representation as
    - its signature
    - its logsignature
  - we can recognise a function on paths (Fliess)
    - as a linear functional on the tensor algebra restricted to G
    - a polynomial function on L
- Much interesting work still to be done
  - Scalability can you work out parts of the signature
  - Explanability can you work out which parts of the signature are used
  - local or global (power series or smooth function)
- Measures on paths and expected signatures
- Rough path theory and regularity structures
  - top down descriptions of complex systems, calculus, CDEs for complex signals,....

Multidimensional streamed data is an important class for data.

- In the talks that come you will see a very gritty set of connections, and I hope they will be of interest.
  - The analysis of distributional data where the paths are already expected signatures of measures on paths.
  - The generic application of signature methods to data; does it work as a package for the inexperienced.
  - Given a corpus of streams, can we identify if a new stream is anomolous, with an approach motivated by concentration of measure?
  - Can one use neural methods to identify the best (controlled) differential equation models?
  - Can signatures allow better integration of TDA into the normal machine learning pipelines?
  - Are there kernel methods for unparamterised streamed data and are they useful?
- Computational examples https://www.datasig.ac.uk/examples e.g. landmark-based human action recognition

Kernels

Kernels form a useful tool in machine learning, they offer concrete approaches to resolving the core problem data science problems, and also can be very useful in reducing the dimension and complexity of calculations

- What is a kernel?
  - A set X embedded into E and E^*

$$\begin{aligned} \phi &: X \hookrightarrow E \\ \psi &: X \hookrightarrow E^* \end{aligned}$$

    then one gets $K(x, x') := \langle \phi(x), \psi(x') \rangle$ and conversely.
  - Amari's statistical manifolds of probability measures all have this property

$$\begin{aligned} P_{\preccurlyeq \mu} &\to L_p(\mu) \\ p d\mu &\to p^{1/p} \end{aligned}$$

  - Sometimes a Hilbert space (e.g. $p = 2$ in Amari)

Kernel based regression

Although there are many issues, one cannot deny the convenience of kernels. For every data point one gets a new function on points: $K(x, x') := \langle \phi(x), \psi(x') \rangle$.

- Kernels are useful for regression and machine learning
  - Some real observational data $(x_j, y_j)_{j=1\ldots N}$ then solve this system

$$y_j = \sum_{k=1\ldots N} \lambda_k K(x_j, x_k)$$

    to express
  - the observed function in terms of the kernel functions $\psi(x_i) \in E^*$.

$$F(\cdot) := \sum_{k=1\ldots N} \lambda_k K(\cdot, x_k)$$

  - Crucially this calculation only depends on the N^2 numbers $K(x_j, x_k)$ and does not need the embedding - E can be infinite dimensional.

Inner products

It is not clear that kernels should induce inner products, but when they do, one can use Gaussian tricks. Symmetric kernels are an essential component in algorithms for pattern analysis (Bishop, 1995; Hastie et al., 2001; Scholköpf and Smola, 2002)

- suppose we have a symmetric and universal kernel
  - Consider a gaussian random variable X' on E^* with co-variance given by the inner product on E .

$$\mathbb{E}\left[X'(x_1)X'(x_2)\right] = < x_1, x_2 >, \quad x_1, x_2 \in E$$

  - A measure on functions (images).
  - Can sample from the conditional distribution of X' given the evaluations $(x_j, y_j)_{j=1\ldots N}$.
  - Allows interpolation - get a random function defined everywhere and consistent with the monochrome data.

From alphabets to Hilbert spaces

The basis $\mathcal{E} = \{e_1, \ldots, e_d\} \subset E$ induces structure on E and a path in E is already a path in E^* and a signature in $T((E))$ is already in $T((E^*))$

- A basis for $E^{\otimes n}$ is the words of length n with letters drawn from the alphabet $\mathcal{E}$:

$$\{e_K = e_{k_1} \otimes \ldots \otimes e_{k_n}\}_{K=(k_1,\ldots,k_n)\in\{1,\ldots,d\}^n}$$

- The choice of $\mathcal{E}$ induces an inner product on E, $E^{\otimes n}$, $T(E)$, $T((E))$.

$$\langle e_{i_1} \otimes \ldots \otimes e_{i_n}, e_{j_1} \otimes \ldots \otimes e_{j_n} \rangle = \delta_{i_1,j_1} \ldots \delta_{i_n,j_n}, \qquad \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

- Making the canonical projection $\pi_n : T = (T^0, T^1, \ldots, T^n, \ldots) \to T^n \in E^{\otimes n}$ orthogonal.

- The inner product is
  - Defined for $A, B \in T((E))$ as

$$\langle A, B \rangle = \sum_{n=0}^{\infty} \langle \pi_n(A), \pi_n(B) \rangle$$

Signature Kernels

Signatures exist over any interval and (for rough paths) decay factorially.

- Suppose that x, $\mathbb{X}$ are rough paths/streams of finite 1 and p variation over $[s,t]$ and controlled by w_x, $w_{\mathbb{X}}$;Let $(s_1,s_2) \subset [s,t]$
  - then x, $\mathbb{X}$ have signatures:

$$(s_1,s_2) \mapsto S(x_{s_1,s_2}) = (1, x^1_{s_1,s_2}, \ldots, \ldots, \ldots, x^m_{s_1,s_2}, \ldots) \in T((E))$$

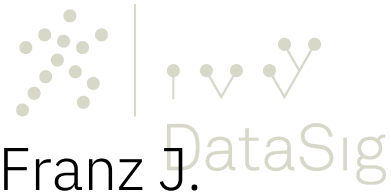
$$(s_1,s_2) \mapsto S(\mathbb{X}_{s_1,s_2}) = (1, \mathbb{X}^1_{s_1,s_2}, \ldots, \mathbb{X}^{\lfloor p \rfloor}_{s_1,s_2}, \ldots, \mathbb{X}^m_{s_1,s_2}, \ldots) \in T((E))$$

  - they are multiplicative functionals
  - and they have factorial decay (neoclassical inequality)

$$\begin{aligned} ||x^k_{s_1,s_2}||_{E^{\otimes k}} &\leq \frac{\omega_x(s_1,s_2)}{k!}, \ \forall (s_1,s_2) \in \Delta_I \\ ||\mathbb{X}^k_{s_1,s_2}||_{E^{\otimes k}} &\leq \frac{\omega_{\mathbb{X}}(s_1,s_2)^{k/p}}{\beta_p(k/p)!}, \ \forall (s_1,s_2) \in \Delta_I \end{aligned}$$

- For any two unparameterised rough paths $\mathbb{X}$ and $\mathbb{Y}$ the signature kernel $K(\mathbb{X},\mathbb{Y}) := \langle S(\mathbb{X}_I), S(\mathbb{Y}_J) \rangle$ is always well defined.

Oberhauser and Kiraly



- In their paper *Kernels for Sequentially Ordered Data* (Franz J. Kiraly, Harald Oberhauser; JMLR 20(31):1-45, 2019) Franz and Harald observed that a kernel on the space carrying the data always implies a kernel on truncated signatures of sequences in the implied linear space, and importantly.
  - they use dynamic programming and low-rank techniques to demonstrate that for the truncated signature kernel and bounded variation paths there were efficient algorithms to compute this truncated kernel.
  - Harald, with his student Csaba, *Bayesian Learning from Sequential Data using Gaussian Processes with Signature Covariances*. (Toth, Csaba, and Harald Oberhauser., ICML in press (2020)) explored the practical ramifications of this kernel in a range of practical contexts and demonstrate its value.

The Goursat PDE

It is easy to see that, unlike the truncated signature, the full signature kernel is universal, but calculating it directly would seem to involve infinite series of integrals with exponentially increasing numbers of terms.

- This part of the talk is based on *Computing the full signature kernel as the solution of a Goursat problem (*Thomas Cass, Terry Lyons, Cristopher Salvi, Weixin Yang https://arxiv.org/abs/2006.14794)
  - For any kernel there is a PDE. evaluating the *full* signature kernel between two unparameterised paths
  - The PDE is well defined, quick to compute, has numerics and solutions for any rough streams. It is an interesting ”S”PDE!

The basic statement

Theorem

Let $I = [u, u']$ and $J = [v, v']$ be two closed time intervals and let $x \in C^1(I, E)$ and $y \in C^1(J, E)$. Consider the bilinear form $k_{x,y} : I \times J \to \mathbb{R}$ defined as follows

$$k_{x,y} : (s, t) \mapsto \langle S(x)_s, S(y)_t \rangle$$

then $k_{x,y}$ is a solution of the following linear hyperbolic PDE

$$\frac{\partial^2 k_{x,y}}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle k_{x,y}$$

with initial conditions $k_{x,y}(u, \cdot) = k_{x,y}(\cdot, v) = 1$ and where $\dot{x}_s = \frac{dx_p}{dp}|_{p=s}$ and $\dot{y}_t = \frac{dx_q}{dq}|_{q=t}$.

The Equation continued

By the fundamental theorem of calculus we can differentiate

$$1 + \int_{p=u}^{s} \int_{q=v}^{t} k_{x,y}(p,q)\langle \dot{x}_p, \dot{y}_q \rangle dpdq$$

firstly with respect to s

$$\frac{\partial k_{x,y}(s,t)}{\partial s} = \int_{q=v}^{t} k_{x,y}(s,q)\langle \dot{x}_s, \dot{y}_q \rangle dq$$

and then with respect to t to obtain the desired linear hyperbolic PDE

$$\frac{\partial^2 k_{x,y}(s,t)}{\partial s \partial t} = \langle \dot{x}_s, \dot{y}_t \rangle k_{x,y}(s,t)$$

Goursat

Theorem (Lees 1960 (Goursat 1916) Theorems 2 & 4)

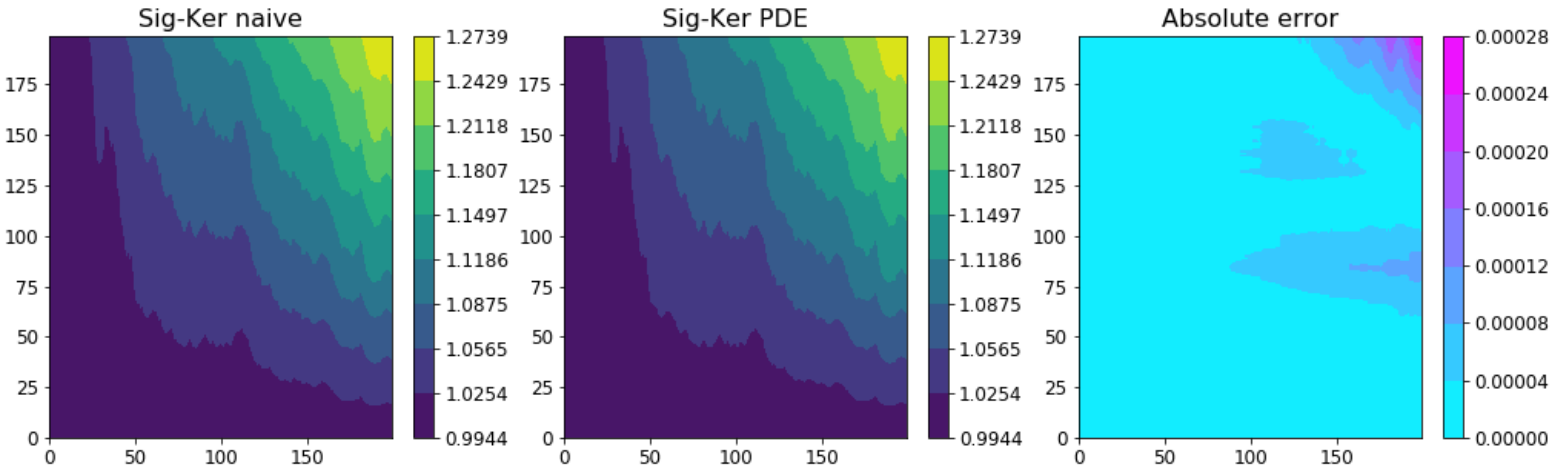
Let $\sigma : I \to \mathbb{R}$ and $\tau : J \to \mathbb{R}$ be two absolutely continuous functions whose first derivatives are square integrable and such that $\sigma(u) = \tau(v)$. Let $C_1, C_2, C_3 : \mathcal{D} \to \mathbb{R}$ be a bounded and measurable over $\mathcal{D}$ and $C_4 : \mathcal{D} \to \mathbb{R}$ be square integrable. Then there exists a unique function $u : \mathcal{D} \to \mathbb{R}$ such that $u(s, v) = \sigma(s), u(u, t) = \tau(t)$ and (almost everywhere on $\mathcal{D}$)

$$\frac{\partial^2 u}{\partial s \partial t} = C_1(s,t)\frac{\partial u}{\partial s} + C_2(s,t)\frac{\partial u}{\partial t} + C_3(s,t)u + C_4(s,t)$$

If in addition $C_i \in C^{p-1}(\mathcal{D})$ $(i = 1, 2, 3, 4)$ and σ and τ are C^p, then the unique solution $u : \mathcal{D} \to \mathbb{R}$ of the Goursat problem is of class C^p.

Set $C_1 = C_2 = C_4 = 0$ and $C_3(s,t) = \langle \dot{x}_s, \dot{y}_t \rangle$ .If the two input paths x, y are C^p then their derivatives will be of class C^{p-1} and therefore the solution $k_{x,y}$ will be of class C^p. Finite difference approximation works.

The numerics



Example of error distribution of $k_{x,y}(s,t)$ on the whole grid $(s,t) \in \mathcal{D}$.

The end

Thank you!