Gradient flow on control space with rough initial condition

Paul Gassiat

CEREMADE, Université Paris-Dauphine

Finance and Stochastics Seminar
Imperial College London, May 2024

Joint work with Florin Suciu (Paris Dauphine)

Outline

Outline

(Sub-Riemannian type) control problem

Consider the controlled ODE

$$dX_t = \sum_{i=1}^{d} V_i(X_t)u^i(t)dt, \qquad X_0 = x \in \mathbb{R}^n$$

and the problem, for a fixed $y \in \mathbb{R}^n$,

$$\text{Find } u \in L^2([0,1], \mathbb{R}^d) \text{ s.t. } X_1 = y.$$

Under the Hörmander bracket-generating condition,

$$\forall z \in \mathbb{R}^n, \quad \mathrm{Lie}(V_1, \dots, V_d)_{|z} = \mathbb{R}^n,$$

the classical **Chow-Rashevskii theorem** (1938) guarantees the existence of such a control.
(Simplest example : Heisenberg group, i.e. $d = 2$, $n = 3$,
$V_1 = \partial_x - \frac{y}{2}\partial_z$, $V_2 = \partial_y + \frac{x}{2}\partial_z$. Corresponds to finding a planar path with fixed endpoints and prescribed area.)

Gradient flow

$$\text{Find } u \in L^2([0,1], \mathbb{R}^d) \text{ s.t. } X_1 = y.$$

This problem is classical in the (deterministic) control community (**(non-holonomic) motion planning**) with many applications (robotics,...), and many specialized algorithms.

We are interested (see next section for motivation) in a very simple / non-specific gradient flow procedure : consider

$$u \in L^2 \mapsto \mathcal{L}(u) = \|y - X_1^u\|_{\mathbb{R}^n}^2 ,$$

and solve the gradient flow (in $L^2[0,1]$)

$$\frac{d}{ds} u(s) = -\nabla \mathcal{L}(u(s)),$$

hoping that $u(s) \to_{s\to\infty} u_\infty$ a solution of the problem.

(Some gradient methods have already been considered in the control literature, in particular the continuation method by Sussmann '93, Sussmann and Chitour '96).

Gradient flow : first properties

$$u \in L^2 \mapsto \mathcal{L}(u) = \|y - X_1^u\|_{\mathbb{R}^n}^2 ,$$

$$\frac{d}{ds} u(s) = -\nabla \mathcal{L}(u(s)),$$

- **Good news** : no strict local minimum for $\mathcal{L}$ (under bracket-generating condition).

Immediate computation :

$$\nabla \mathcal{L}(u(s)) = (y - X_1^u) \cdot_{\mathbb{R}^n} \nabla X_1^u.$$

- **Bad news** : in general, **saddle points** ! possible at each control u s.t. $d_u X_1 : L^2 \to \mathbb{R}^n$ is not onto. (**singular controls** in sub-Riemannian geometry).
  For instance, if $d < n$, $u = 0$ is always singular.
  ($d_u X_1(0)$ only spans $\{V_1(x), \ldots, V_d(x)\}$.)
- **Other serious problem :** no penalization term on u : $\to u(s)$ may diverge to "infinity".

Stochastic initial condition

The existence of saddle points means we cannot hope for convergence from any starting point.
$\longrightarrow$ what about for random initial condition ?

Singular controls are rare : for instance, one part of Malliavin ('76) 's stochastic proof of Hörmander's theorem relies on the fact that

If $u = \dot{W}$ (white noise), then, a.s. , u is non-singular.

(More recently, rough path generalizations to other Gaussian processes, e.g. Cass-Friz '10 and subsequent literature.)

Q : Does stochasticity / roughness of starting point help for the gradient flow to converge ? (Or at least : to prove it that it does)

Rest of the talk : (very partial) answer to this question.

Outline

1. Problem description
2. Motivation from deep learning
3. Results

Motivation from deep learning

- **Supervised learning** :
  given a map $x \in \mathbb{R}^n \mapsto y(x) \in \mathbb{R}^n$ and probability measure μ, want to find Φ in a certain class s.t.

$$\mathcal{E} = \int \mu(dx) \left|\Phi(x) - y(x)\right|^2$$

  is small. Typically, we only have access to finite $(x_i, y_i = y(x_i))_{i=1,\dots,N}$, and we instead try to minimize the empirical loss

$$\widehat{\mathcal{E}} = \frac{1}{N} \sum_{i=1}^{N} \left|\Phi(x_i) - y_i\right|^2 .$$

- **Deep residual neural networks** :
  $\Phi(x) = X_L$, where

$$X_0 = x, \;\; X_{k+1} = X_k + \delta_k \sigma(X_k, \theta_k),$$

  Can be seen as discretization of ODE

$$x_0 = x, \;\; dX_t = \sigma(X_t, \theta_t) dt$$

  Many papers drawing on this connection.
  (starting with E '17, Haber-Ruthotto '17, Chen et al. '18, ...)

ResNets as Rough / Stochastic dynamics

Several people have suggested that ResNets should be understood via S/RDE and not just classical ODE.

- Cohen, Cont, Rossier, Xu '22 : empirical roughness of layer weights, scaling limits.
- Marion, Fermanian, Biau, Vert '22. Hayou '22 : SDE limits for initialization choices
  $X_{k+1} = X_k + L^{-1/2}\sigma(X_k)W_k$, W Gaussian $\mathcal{N}(0, I_m)$.
- Bayer, Friz, Tapia '22 : (discrete) rough path bounds as a robustness measure for ResNets.

The N-point control problem

Consider σ of the form $\sigma(X_t, \theta_t) = \sum_{i=1}^{d} \sigma_i(X_t)\theta_t^i$.
For the ODE limit :

- The problem of minimizing empirical loss can be written as

$$\text{find } \theta \text{ s.t. } X_1(\theta, x_i) = y_i, \quad i = 1, \dots, N. \qquad (*)$$

  This is in fact a problem of the form introduced in the first section, but in $\mathcal{M} = (\mathbb{R}^n)^N \setminus \Delta$.

- Question studied by control-theoretic methods by several people (Agrachev-Sarychev '21, Scagliotti '22,...)
  In particular, Cuchiero, Larsson, Teichmann '21 : There exist $d = 5$ fixed vector fields s.t. for any arbitrary N, there exists a solution to $(*)$.

Motivating question : training of ResNets via gradient descent

Q : Can we obtain theoretical results guaranteeing convergence of (stochastic) gradient descent for ResNets ? Does stochasticity/ roughness of the initial condition help ? (and what about generalization ?)

Note : we are considering a regime where **depth is large** but **width is fixed**, whereas most results in the ML literature require some relation between width n and data size N.

(when $d = \#$ parameters per layer $< nN = \#$ data dimension
$\approx$ sub-Riemannian control problem.)

(No answers in this talk !)

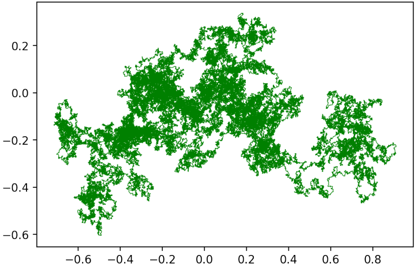
Outline

Irregular controls

We want to consider (replacing u by $z = \int_0^\cdot u_t dt \in C([0,1], \mathbb{R}^d)$) a solution to

$$X_t = x + \int_0^t V(X_s) dz_s \qquad (1)$$

where $z : [0,1] \ \to \mathbb{R}^d$ is irregular (e.g. Brownian motion).



Trajectory of a $2d$ Brownian motion.

Note : if $z = B(\omega)$ is a Brownian path, then a.s. :

z is not absolutely continuous,

z only in $C^{1/2-\epsilon}$.

But one can still make sense of (1) (+regularity of flow, etc) via Itô calculus (1950s), or **rough path theory** (Lyons '98).

Rough path theory

We will formulate everything in the **rough path** (Lyons '98) framework :
For $1/3 < \alpha \leqslant 1/2$, a C^α rough path is the data of

$$\mathrm{z} = \left(\int_s^t dz_u, \int_{s \leqslant u_1 \leqslant u_2 \leqslant t} dz_{u_2} \otimes dz_{u_1} \right)_{s<t}$$

satisfying some algebraic and Hölder-type analytic conditions.
(similar definition for arbitrary $0 < \alpha$ with more iterated integrals :
$z \in C^\alpha \left([0, T], G^{\lfloor \alpha^{-1} \rfloor}(\mathbb{R}^d) \right)$).

For

$$X_t = x + \int_0^t V(X_s) dz_s,$$

the map

$$\mathrm{z} \mapsto X$$

is then continuous (for the corresponding "rough path" topology), under suitable regularity assumptions on the coefficients V.

Rough path translation

In our setting, we will want to consider

$$z = w + h$$

where w is the initial condition (irregular, a C^α rough path), and h is in the tangent space $\mathcal{H} = H^1([0,1], \mathbb{R}^d)$.

Note that for any such w, h, we can define canonically the "sum" $w \oplus h$ by letting

$$\int (w \oplus h) d(w \oplus h) = \int w dw + \int w dh + \int h dw + \int h dh.$$

(This follows from $\mathcal{H} \subset C^{1-var}$).

The map $(w, h) \mapsto w \oplus h$ is then smooth.

The gradient flow setup

We fix :

- $V_1, \ldots, V_d$ smooth, bracket-generating vector fields on $\mathbb{R}^n$.
- **initial condition :** w, a $C^\alpha([0,1],\mathbb{R}^d)$-geometric rough path, $0 < \alpha < 1$.
- **tangent space :** a Hilbert space $\mathcal{H} = H^1([0,1],\mathbb{R}^d)$

and consider the RDE

$$dX_t^{w,h} = \sum_i V^i(X_t) d(w_t \oplus h_t), \quad X_0 = x.$$

For $g = \frac{1}{2}|\cdot - y|^2$, the map

$$h \in \mathcal{H} \mapsto \mathcal{L}(h) := g\left(X_1^{w;h}\right)$$

is smooth. In particular, we can consider the gradient flow trajectory

$$h(0) = 0, \quad \frac{d}{ds} h(s) = -\nabla_{\mathcal{H}} \mathcal{L}(h(s))$$

which defines a trajectory $(h(s))_{s \geqslant 0}$ with values in $\mathcal{H}$.
(Remark : rough path theory is definitely much more convenient than Itô calculus here, even if w is a Brownian motion !)

Some preliminary positive results

We have the following results.

Proposition ("Chow-Rashevskii with rough drift")

Under the bracket-generating condition, for any $x, y \in \mathbb{R}^n$, any fixed w, there exists a smooth path h such that

$$X_1^{w;h}(x) = y.$$

Proposition

Let $\mathbb{P}$ be the law of (enhanced) Brownian motion on $C^\alpha([0,1], \mathbb{R}^d)$. Then

$$\mathbb{P}\left(w : h(s) \to_{s \to \infty} h_\infty \text{ with } \mathcal{L}(h_\infty) = 0\right) > 0.$$

(Brownian motion could be replaced by any non-degenerate Gaussian rough path).

In other words : we do not lose anything from starting from a rough initial condition. Do we gain anything ?

True roughness (Hairer-Pillai '11, Friz-Shekhar ' 12)

Recall that w is a.e. truly β-rough, if, for a.e. s in $[0,1]$,

$$\forall 0 \neq v \in \mathbb{R}^d \limsup_{t \downarrow s} \frac{|w_{s,t} \cdot v|}{|t-s|^{\beta}} = +\infty.$$

Under this assumption, if $\beta < 2\alpha$, then

$$\int_0^{\cdot} \sum_i f_s^i dw_s^i \equiv 0 \quad \Rightarrow \quad f^i \equiv 0.$$

(Most classical stochastic processes, such as (fractional) Brownian motion, satisfy this condition a.s.).

Lemma

Let w be a.e. truly β-rough, and $h \in C^{q-var}$, with $\frac{1}{q} > \beta$, then $w+h$ is a.e. truly β-rough.

In particular, for our gradient flow, if the initial condition is truly rough, so is $w + h(s)$ at any time $s \geqslant 0$.

Expressions for $\nabla_{\mathcal{H}}\mathcal{L}$

Recall that for our gradient flow :

$$\nabla_{\mathcal{H}}\mathcal{L}(w;h) = (X_1^{w;h} - y) \cdot_{\mathbb{R}^n} \nabla_{\mathcal{H}} X_1^{w;h}.$$

A classical computation yields, for $\xi \in \mathbb{R}^n$,

$$\left\| \xi \cdot \nabla_{\mathcal{H}} X_1^{w;h} \right\|_{\mathcal{H}}^2 = \sum_i \int_0^1 \left(J_{t\to 1} V_i(X_t) \cdot_{\mathbb{R}^n} \xi \right)^2 dt$$

where $J_{t\to 1}$ is the Jacobian matrix of the flow $X_t \mapsto X_1$.

In addition, for any vector field W,

$$J_{t\to 1} W(X_t) = W(X_1) - \sum_j \int_t^1 J_{t\to 1}[W, V^j](X_t) d(w+h)_t^i.$$

True roughness ⇒ saddle-points are at infinity

An iteration then implies the following (standard result from Malliavin calculus, cf e.g.Friz-Hairer chap. 11)

Proposition

Under the bracket-generating condition, if w is truly rough, then

$$\xi \in \mathbb{R}^n \setminus \{0\} \;\Rightarrow\; \xi \cdot \nabla_{\mathcal{H}} X_1^{w;0} \neq 0.$$

Combined with the lemma from a previous slide, this means that all the saddle points of $\mathcal{L}$ are now at infinity !

Corollary

Assume that w is truly rough, then if $(h(s))_{s \geqslant 0}$ is bounded in $\mathcal{H}$, it converges to a minimizer of $\mathcal{L}$.

(Remark : a similar result holds for $\mathcal{L}^{\mu}(h) = \int \mu(dx) \left|y(x) - X_1^x(w \oplus h)\right|^2$.)

Global convergence results

We have convergence to a minimum in two simple (but non-trivial) cases.

Theorem

(Elliptic) *Assume that for all* $z \in \mathbb{R}^N$,

$$span\{V_1(z), \dots, V_d(z)\} = \mathbb{R}^n,$$

then for all r.p. w, *for all* x, y,

$$\lim_{s \to \infty} h_s = h_\infty \in \mathcal{H}, \quad \mathcal{L}(h_\infty) = 0. \qquad \text{(ConvMin)}$$

(Step-2 nilpotent) *Assume that (the* V_i *are bracket-generating and)*

$$\forall i, j, k, [V_i, [V_j, V_k]] \equiv 0.$$

Then, with $\mathbb{P}$ *the law of Brownian motion, for* $\mathbb{P}$*-a.e.* w, *for all* x, y, (ConvMin) *holds.*

(Remark : in 2nd case, we could replace BM by fBm with $H < \frac{1}{2}$ but not $H > \frac{1}{2}$!)

Convergence for discrete approximations

The continuity properties of rough path theory allow for simple proofs of convergence of discrete approximations.

For instance, assume that we know that for w a Brownian motion, the g.f. solution $h \to h_\infty$(non-degenerate minimum) a.s.

For fixed N, let $\mathcal{H}_N \sim \mathbb{R}^{Nd}$ the space of piecewise linear controls, linear on $[i/N, (i+1)/N]$. Let h^N be the gradient flow :

$$\frac{d}{ds} h^N(s) = -\nabla_{\mathcal{H}_N} \mathcal{L}(h^N(s)), \quad \dot{h}^{N,j}(0) = \frac{1}{\sqrt{N}} Z_{ij} \text{ on } [i/N, (i+1)/N],$$

where the Z_{ij} are i.i.d. $\mathcal{N}(0,1)$.
Then the convergence for B.M. implies

$$\lim_{N\to\infty} \mathbb{P}\left(h^N(s) \to_{s\to+\infty} h^N_\infty \text{ with } \mathcal{L}(h^N_\infty) = 0\right) = 1.$$

Major ingredient of proof : Łojasiewicz inequality

Consider a function $L : H \to \mathbb{R}_+$ satisfying, for some $c > 0$,

$$\forall x \in H, \quad |(\nabla L)(x)|^2 \geqslant c^2 L(x). \tag{Ł}$$

Then, for the gradient flow $\dot{x}(s) = -\nabla L(x(s))$, it holds that

- $L(x(s)) \leqslant L(x(0))e^{-c^2 s}$ converges to 0.
- More importantly : $x(s) \to_{s\to\infty} x_\infty$, where $L(x_\infty) = 0$.
  Proof : (Łojasiewicz 1960's)

  $$\frac{d}{ds}\left\{2\sqrt{L}(x(s)) + c\int_0^s |\dot{x}(u)|du\right\} \leqslant 0$$

  which implies that the trajectory $(x(s); s \geqslant 0)$ has finite length, and, in particular, converges (to a minimizer).

Local Łojasiewicz inequality

Proposition

Assume that $L : H \to \mathbb{R}_+$ *satisfies,*

$$\forall x \in H, \quad |(\nabla L)(x)|^2 \geqslant c^2(|x|)L(x) \qquad \text{(Łloc)}$$

where $c(\cdot)$ *is decreasing, and satisfies* $\int^{+\infty} c(r)dr = +\infty$.
Then for the gradient flow $\dot{x}(s) = -\nabla L(x(s))$, *it holds that*

$$x(s) \to_{s\to\infty} x_\infty, \text{ where } L(x_\infty) = 0.$$

Proof : (Łojasiewicz's argument again)

$$\frac{d}{ds}\left\{\frac{1}{2}\sqrt{L}(x(s)) + C\left(|x_0| + \int_0^s |\dot{x}(u)|du\right)\right\} \leqslant 0$$

with $C = \int_0^{\cdot} c$. □

For instance, one can have $c(r) = \frac{c}{1+r^\alpha}$, $\alpha \leqslant 1$.

Arguments of proof

In our case, we have,

$$\frac{\|\nabla \mathcal{L}\|_{\mathcal{H}}^2}{\mathcal{L}} \geqslant c(w;h)^2,$$

where

$$\begin{aligned} c(w;h)^2 &= \inf_{|\xi|=1} \|\xi \cdot_{\mathbb{R}^n} \nabla_{\mathcal{H}}(X_1)\|_{\mathcal{H}}^2 \\ &= \inf_{|\xi|=1} \sum_i \int_0^1 \left(J_{t\to 1} V_i(X_t) \cdot_{\mathbb{R}^n} \xi\right)^2 dt \end{aligned}$$

where $J_{t\to 1}$ is the Jacobian matrix of the flow of X between t and 1.

(Familiar object from Malliavin calculus : c is the smallest eigenvalue of the Malliavin matrix at $w+h$ for the functional X_1).

In both cases, we prove

$$c(w;h)^2 \gtrsim \frac{1}{1+\|h\|_{\mathcal{H}}^2}.$$

Proof in the elliptic case

$$
\begin{aligned}
c(w;h)^2 &= \inf_{|\xi|=1} \sum_i \int_0^1 \left(J_{t\to 1} V_i(X_t) \cdot_{\mathbb{R}^n} \xi\right)^2 dt \\
&\geqslant \int_0^1 \left|\lambda_-(J_{t\to 1} J_{t\to 1}^T)\right| dt \\
&\gtrsim \int_0^1 e^{-c\|h\|_{\mathbf{1}-var;[t;\mathbf{1}]}} \gtrsim \frac{1}{1+\|h\|_{H^{\mathbf{1}}}^2}
\end{aligned}
$$

using that (Sobolev embedding) $\|h\|_{1-var;[t;1]} \lesssim \|h\|_{H^{\mathbf{1}}}(1-t)^{1/2}$.

Remark 1 : replacing H^1 by another Sobolev space H^δ, $\delta \in (1/2, 1]$ does not change the exponent appearing in the Łojasiewicz inequality...).

Remark 2 : Sussmann and Chitour '93, '96 proved convergence for their method of continuation using a similar inequality under less restrictive assumptions (but regular controls).

Step-2 nilpotent case

- The nilpotent hypothesis yields (letting $z = X_1$)

$$J_{t,1} V_i(X_t) = V_i(z) - \sum_j [V_j, V_i](z)(w+h)^j_{t,1}.$$

  This yields

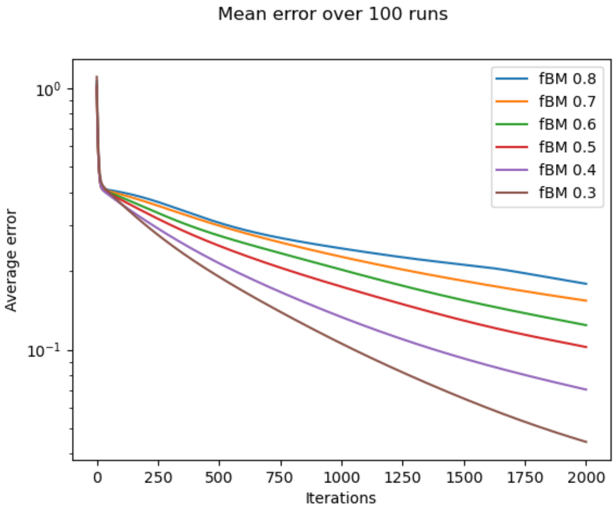
$$c(w;h)^2 \gtrsim \inf_{\sum_{i,j} \xi_{i,j}^2 = 1} \sum_i \left(\int_0^1 dt \left(\xi_{ii} + \sum_j \xi_{ij} (w^j + h^j)_{t,1} \right)^2 \right)$$

- For w B.M.,

$$\|w - h\|_{L^2} \geqslant \frac{C(w)}{1 + \|h\|_{H^1}}.$$

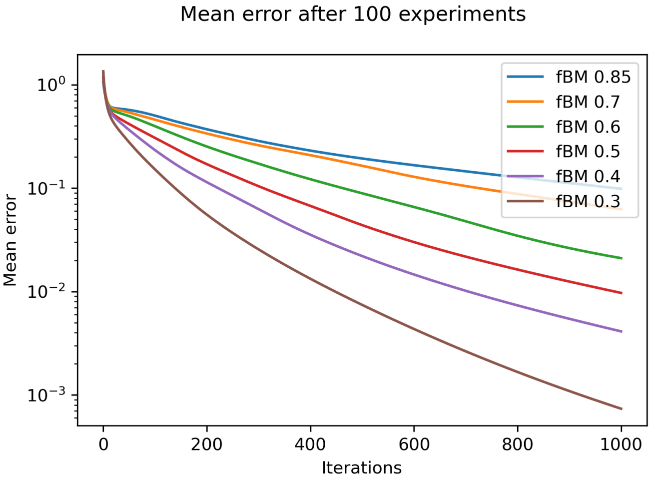
  (This is a similar result to the fact that the norm of w in the Besov space $B^{1/2}_{2,\infty}$ is $\geqslant 1$ a.s.).

Numerical experiment



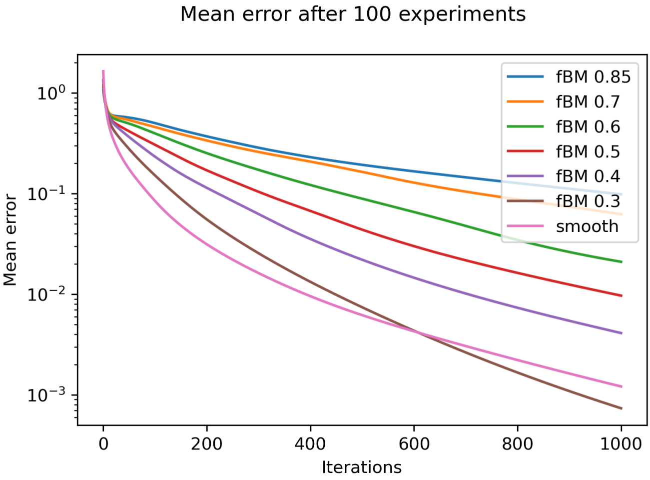
(rank $d = 10$, $n = 55$ (step 2 nilpotent), 100 time points, learning rate$= 0.1$)

Numerical experiment



(rank $d = 2$, step 3 nilpotent ($n = 5$), 100 time points, learning rate$= 0.1$)

Numerical experiment : smooth = rough ?



(rank $d = 2$, step 3 nilpotent ($n = 5$), 100 time points, learning rate= 0.1)

Conclusion : (many) remaining questions

We are able to show convergence of gradient flow for the control problem

$$\inf_h |X_1(h) - y|^2$$

with rough (Brownian) initialization in the simplest non-trivial cases (elliptic, step-2 nilpotent).

Can we do better ?

- Convergence for more general vector fields : Step-3 nilpotent, arbitrary nilpotent, general case ?
- Convergence for discretized problems ? (Quantitative discretized roughness, number of steps vs. number of Lie brackets needed,...)
- Variants of gradient descent ? (stochastic, ...)
- Applications to Deep Learning ?
- Other criteria than roughness ?