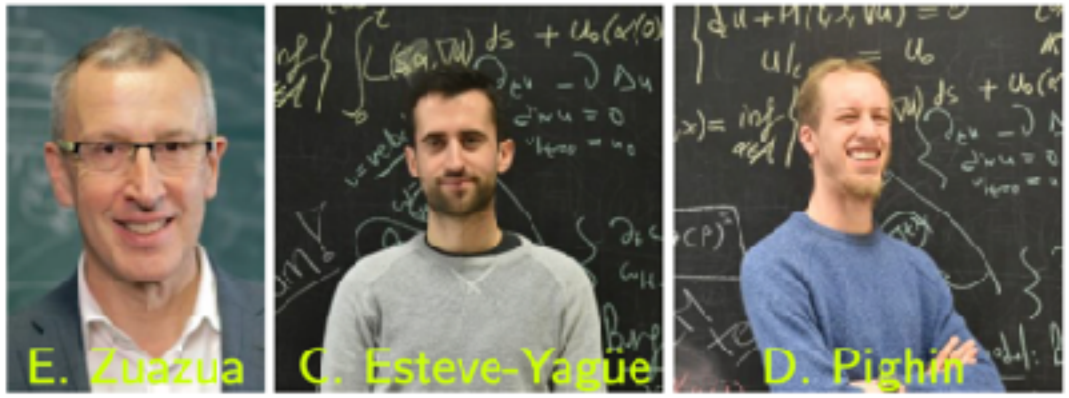
On the intersection of control and machine learning

Borjan[1] Geshkovski
borjang.github.io

IRMA, 29 Novembre, 2022



[1] Bor-*Yann*

Massachusetts Institute of Technology

Supervised learning

Goal: Approximate an unknown function $f : \mathbb{R}^d \to \mathbb{R}^m$ given data

$$\mathscr{D} := \left\{ x^{(i)}, f(x^{(i)}) \right\}_{i \in [n]} \subset \mathbb{R}^d \times \mathbb{R}^m.$$

Supervised learning

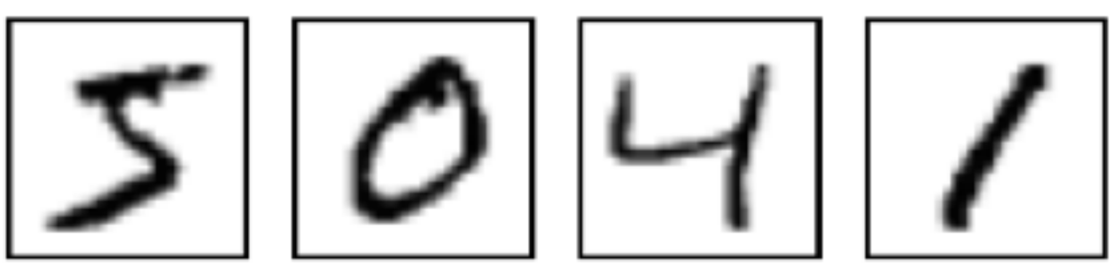
Goal: Approximate an unknown function $f : \mathbb{R}^d \to \mathbb{R}^m$ given data

$$\mathscr{D} := \left\{x^{(i)}, f(x^{(i)})\right\}_{i\in[n]} \subset \mathbb{R}^d \times \mathbb{R}^m.$$

We distinguish:

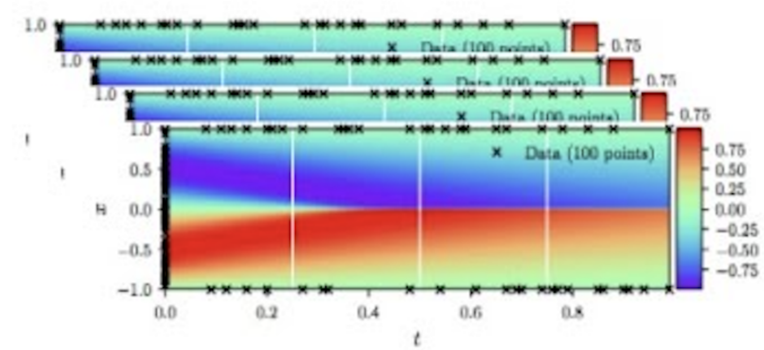
- Classification: $\mathrm{ran}(f)$ is $\{e_j\}_{j\in[m]} \subset \mathbb{R}^m$.
  Image, audio: $d \geqslant 10^3$, $d \gg m$.

$d = 784$
$m = 10$



$x^{(1)} \to \mathbb{R}^{10} \ni e_6$, $x^{(2)} \to e_1$, $x^{(3)} \to e_5$, $x^{(4)} \to e_2$

- Regression: $\mathrm{ran}(f)$ is $\mathbb{R}^m$.
  PDE: fixed initial/boundary data, then learn $(t, \mathrm{x}) \mapsto u(t, \mathrm{x})$



(Feed-forward) neural networks

$$\boxed{f_{\text{approx}}(x) := P\mathrm{x}^{[n_T]}(x)} \qquad \text{for } x \in \mathbb{R}^d.$$

(Feed-forward) neural networks

$$\boxed{f_{\text{approx}}(x) := P\mathrm{x}^{[n_T]}(x)} \qquad \text{for } x \in \mathbb{R}^d.$$

where

- $P \in \mathbb{R}^{m \times d_{n_T}}$ (suppose given)

(Feed-forward) neural networks

$$\boxed{f_{\text{approx}}(x) := P\mathrm{x}^{[n_T]}(x)} \qquad \text{for } x \in \mathbb{R}^d.$$

where

- $P \in \mathbb{R}^{m\times d_{n_T}}$ (suppose given)
- $\mathrm{x}^{[n_T]}(x) \in \mathbb{R}^{d_{n_T}}$ is output of neural net with $n_T \geqslant 1$ *layers*:

$$\begin{cases} \mathrm{x}^{[k+1]} = c^{[k]}\sigma(a^{[k]} \cdot \mathrm{x}^{[k]} + b^{[k]}) & k \in \{0, \ldots, n_T - 1\} \\ \mathrm{x}^{[0]} = x, \end{cases}$$

state $\mathrm{x}^{[k+1]} \in \mathbb{R}^{d_{k+1}}$ and **weights** $c^{[k]} \in \mathbb{R}^{d_{k+1}}, a^{[k]} \in \mathbb{R}^{d_k}$, $b^{[k]} \in \mathbb{R}$

- $\sigma \in C^{0,1}(\mathbb{R})$, typically $\sigma(x) = (x)_+$ or $\sigma(x) = \tanh(x)$
- *widths* d_k given

Residual neural networks (ResNets)

Let $d_k = d$ for all k.

Consider

$$\begin{cases} \mathrm{x}^{[k+1]} = \mathrm{x}^{[k]} + \triangle t c^{[k]} (a^{[k]} \cdot \mathrm{x}^{[k]} + b^{[k]})_+ & k \in \{0, \ldots, n_T - 1\} \\ \mathrm{x}^{[0]} = x, \end{cases}$$

with $c^{[k]}, a^{[k]} \in \mathbb{R}^d$ and $b^{[k]} \in \mathbb{R}$.

Deep **residual** learning for image recognition

K He, X Zhang, S Ren, J Sun - Proceedings of the IEEE …, 2016 - openaccess.thecvf.com

… as learning **residual** functions with … **residual** networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate **residual** …

☆ Save 99 Cite Cited by 142319 Related articles All 72 versions

Neural ODEs

Natural to consider [E '17]:

$$\boxed{\begin{cases} \dot{\mathrm{x}}(t) = c(t)(a(t)\cdot \mathrm{x}(t) + b(t))_+ & t \in (0,T), \\ \mathrm{x}(0) = x \end{cases}}$$

And now $f_{\text{approx}}(x) := P\mathrm{x}(T)$.

Useful in practice:

- Beyond Euler schemes [Chen et al. '18]
- Structure preserving schemes [Schönlieb et al. '20, '22]
- Beyond supervised learning (wait until the end of the talk)
- More compact form for analysis

Approximation theory

Before using the data, a "well-posedness" question can be asked.

Problem (Universal approximation)

Given $f \in \mathscr{H}$ and $\epsilon > 0$, find $(a_\epsilon, b_\epsilon, c_\epsilon) \in L^\infty((0,T);\mathbb{R}^{2d+1})$ such that

$$\|f_{\text{approx},\epsilon} - f\|_{\mathscr{H}} \leqslant \epsilon$$

Approximation theory

Before using the data, a "well-posedness" question can be asked.

Problem (Universal approximation)

Given $f \in \mathscr{H}$ and $\epsilon > 0$, find $(a_\epsilon, b_\epsilon, c_\epsilon) \in L^\infty((0,T);\mathbb{R}^{2d+1})$ such that

$$\|f_{\text{approx},\epsilon} - f\|_{\mathscr{H}} \leqslant \epsilon$$

- Feed-forward nets: [Cybenko '89], [Barron '93] ($n_T = 1$ and $\mathscr{H} = C^0([0,1]^d)$), Pinkus '99 ($n_T \geqslant 1$)
- Neural ODEs: [Li, Lin, Shen '22], [Ruiz-Balet, Zuazua '22] ($\mathscr{H} = L^2((0,1)^d;\mathbb{R}^m)$). ResNets are corollary as controls are piecewise constant
- Strategies generally **non-algorithmic** and suffer from **curse of dimensionality**

Learning is control

- To learn, we **can only use the data** $\mathscr{D}$.

Learning is control

- To learn, we **can only use the data** $\mathscr{D}$.
- So, we consider

Problem

Find controls $(a, b, c) \in L^\infty((0,T); \mathbb{R}^{2d+1})$ such that

$$P\mathrm{x}_i(T) = f(x^{(i)}) \qquad \forall i \in [n]$$

where

$$\begin{cases} \dot{\mathrm{x}}_i(t) = c(t)(a(t) \cdot \mathrm{x}_i(t) + b(t))_+ & t \in (0,T), \\ \mathrm{x}_i(0) = x^{(i)}, \end{cases} \tag{1}$$

and hope predictions of $f(x)$ are good if we take initial data points x outside $\mathscr{D}$ (*generalization*).

Learning is control

- To learn, we **can only use the data** $\mathscr{D}$.
- So, we consider

Problem

Find controls $(a, b, c) \in L^\infty((0, T); \mathbb{R}^{2d+1})$ such that

$$P\mathrm{x}_i(T) = f(x^{(i)}) \qquad \forall i \in [n]$$

where

$$\begin{cases} \dot{\mathrm{x}}_i(t) = c(t)(a(t) \cdot \mathrm{x}_i(t) + b(t))_+ & t \in (0, T), \\ \mathrm{x}_i(0) = x^{(i)}, \end{cases} \tag{1}$$

and hope predictions of $f(x)$ are good if we take initial data points x outside $\mathscr{D}$ (*generalization*).

- It's a simultaneous/ensemble control(lability) problem!
- Nonlinear control-state interaction is necessary

What is done in practice

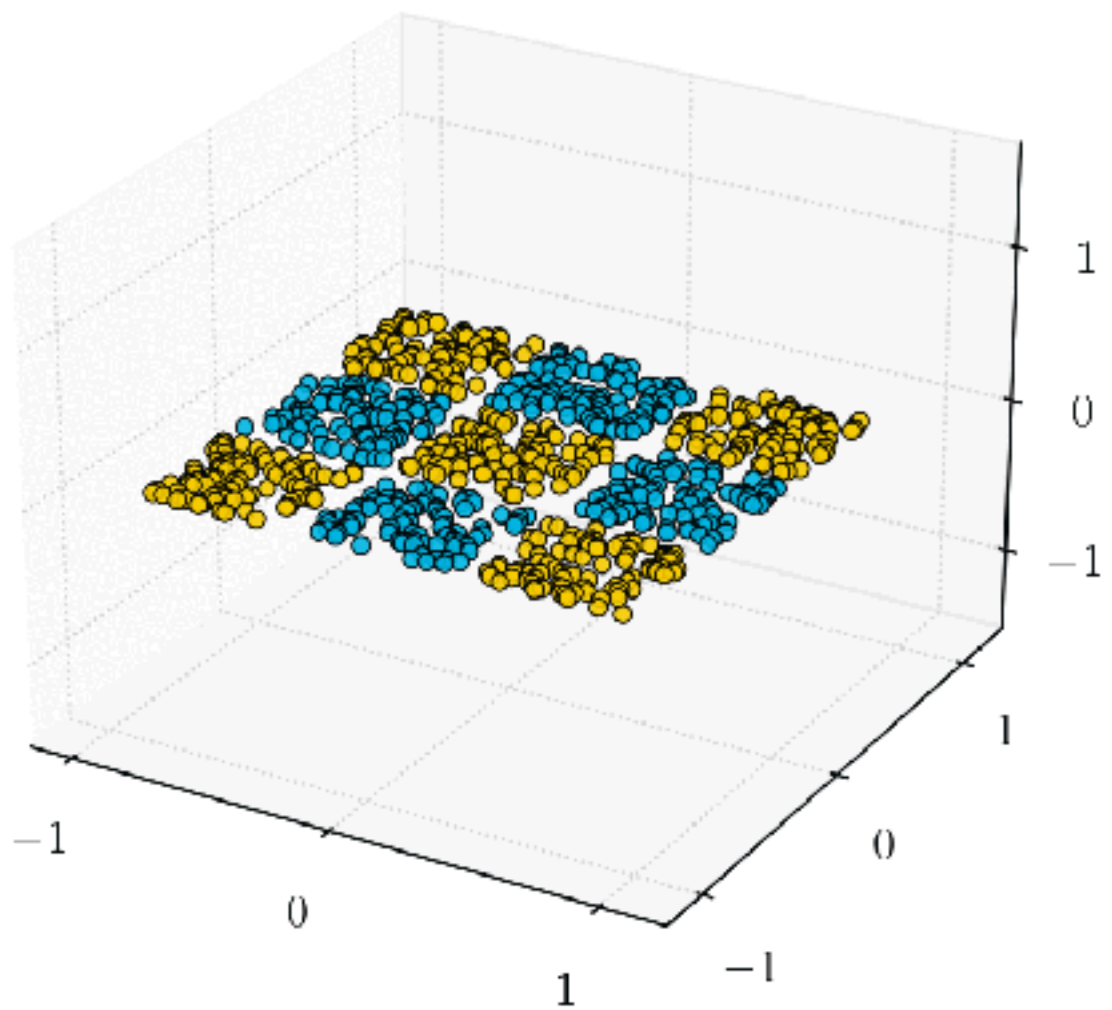
Least squares, with penalty $\lambda > 0$:

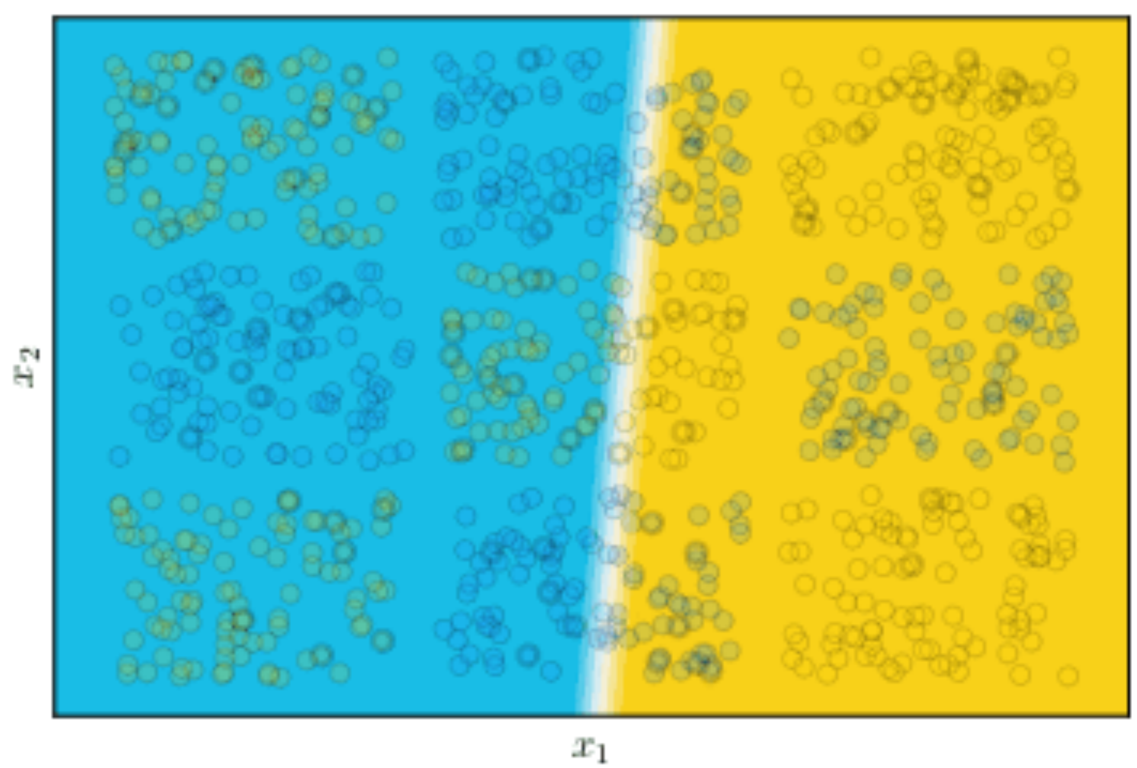
$$\boxed{\min_{\substack{\theta=(a,b,c)\\(a,b)\in H^1((0,T);\mathbb{R}^{d+1})\\c\in L^2((0,T);\mathbb{R}^d)\\\mathrm{x}_i \text{ solves (1)}}} \underbrace{\frac{1}{n}\sum_{i\in[n]}\left|P\mathrm{x}_i(T)-f(x^{(i)})\right|^2}_{:=\mathbf{E}(\boldsymbol{X}(t))} + \lambda\|\theta\|^2_{H^1\times L^2}} \qquad (2)$$

- Empirical risk minimization: $\mathbf{E}(\cdot)$ is the empirical risk.
- H^1 suffices for compactness – if $p\in[1,\infty)$, $\varphi:\mathbb{R}\to\mathbb{R}$ is s.t. $\varphi\circ u_n \rightharpoonup \varphi\circ u^*$ in $L^p(0,1)$ for any $u_n\rightharpoonup u^*$ in $L^p(0,1)$, then φ is affine!
- Can go way beyond squared Euclidean distance (even non-distances such as cross-entropy for classification)

github.com/borjanG/dynamical.systems

$$T = 5, n_T = 16, n = 3000, \lambda = 0.01$$





Optimal control over long time

- In practice n_T can be large (*deep* learning)

Optimal control over long time

- In practice n_T can be large (*deep* learning)
- But $\triangle t = \frac{T}{n_T}$, so n_T large means T large

Optimal control over long time

- In practice n_T can be large (*deep* learning)
- But $\triangle t = \frac{T}{n_T}$, so n_T large means T large

Question

For global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with unique solutions x_i to (1) as columns, **what happens when $T \to \infty$**?

Optimal control over long time

- In practice n_T can be large (*deep* learning)
- But $\triangle t = \frac{T}{n_T}$, so n_T large means T large

Question

For global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with unique solutions x_i to (1) as columns, **what happens when $T \to \infty$**?

Definition (Interpolation)

(1) interpolates $\mathscr{D}$ if $\exists \theta \in H^1((0,1);\mathbb{R}^{d+1}) \times L^2((0,1);\mathbb{S}^{d-1})$ such that

$$P\mathrm{x}_i(1) = f(x^{(i)}) \qquad \forall i \in [n]$$

where $\mathrm{x}_i \in C^0([0,1];\mathbb{R}^d)$ solves (1) with control θ. (I.e., $\mathbf{E}(\boldsymbol{X}(1)) = 0$.)

Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22b]

Fix $\lambda > 0$; c in (2) minimized over $L^2((0,T);\mathbb{S}^{d-1})$; (1) interpolates $\mathscr{D}$. For $T \geqslant 1$, any global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with columns solutions to (1) satisfy:

Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22b]

Fix $\lambda > 0$; c in (2) minimized over $L^2((0,T);\mathbb{S}^{d-1})$; (1) interpolates $\mathscr{D}$. For $T \geqslant 1$, any global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with columns solutions to (1) satisfy:

1. $\exists C(\mathscr{D},\lambda) > 0$,

$$\mathbf{E}(\boldsymbol{X}_T(T)) = \frac{1}{n}\sum_{i\in[n]}\left|P\mathrm{x}_i(T) - f(x^{(i)})\right|^2 \leqslant \frac{C}{T}$$

Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22b]

Fix $\lambda > 0$; c in (2) minimized over $L^2((0,T);\mathbb{S}^{d-1})$; (1) interpolates $\mathscr{D}$. For $T \geqslant 1$, any global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with columns solutions to (1) satisfy:

1. $\exists C(\mathscr{D},\lambda) > 0$,
$$\mathbf{E}(\boldsymbol{X}_T(T)) = \frac{1}{n}\sum_{i\in[n]}\left|P\mathrm{x}_i(T) - f(x^{(i)})\right|^2 \leqslant \frac{C}{T}$$
2. $\boldsymbol{X}_{T_k}(T_k) \to \boldsymbol{X}^*$ for some subsequence $T_k > 0$, $T_k \to \infty$ $(k \to \infty)$ and $\boldsymbol{X}^* \in \mathbb{R}^{d\times n}$ with $\mathbf{E}(\boldsymbol{X}^*) = 0$

Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22b]

Fix $\lambda > 0$; c in (2) minimized over $L^2((0,T);\mathbb{S}^{d-1})$; (1) interpolates $\mathscr{D}$. For $T \geqslant 1$, any global minimizer θ_T for (2) and $\boldsymbol{X}_T \in C^0([0,T];\mathbb{R}^{d\times n})$ matrix with columns solutions to (1) satisfy:

1. $\exists C(\mathscr{D},\lambda) > 0$,

$$\mathbf{E}(\boldsymbol{X}_T(T)) = \frac{1}{n}\sum_{i\in[n]} \left| P\mathrm{x}_i(T) - f(x^{(i)}) \right|^2 \leqslant \frac{C}{T}$$

2. $\boldsymbol{X}_{T_k}(T_k) \to \boldsymbol{X}^*$ for some subsequence $T_k > 0$, $T_k \to \infty$ $(k \to \infty)$ and $\boldsymbol{X}^* \in \mathbb{R}^{d\times n}$ with $\mathbf{E}(\boldsymbol{X}^*) = 0$
3. Set $\theta_k(t) := (T_k a_{T_k}(tT_k), T_k b_{T_k}(tT_k), c_{T_k}(tT_k))$ for $t \in [0,1]$. Then $\theta_k \to \theta^*$ strongly in $H^1 \times L^2$ where θ^* is some solution to

$$\inf_{\substack{\theta=(a,b,c)\\ (a,b)\in H^1((0,1);\mathbb{R}^{d+1})\\ c\in L^2((0,1);\mathbb{S}^{d-1})\\ \mathbf{E}(\boldsymbol{X}(1))=0}} \|\theta\|^2_{H^1\times L^2}$$

1. $\theta_1 = (a_1, b_1, c_1) \ \in H^1((0,1);\mathbb{R}^{d+1}) \times L^2((0,T);\mathbb{S}^{d-1})$ yields x_i^1 solution to (1) on $[0,1]$. Then

$$\theta_T(\cdot) = \left(\frac{1}{T}a_1\left(\frac{\cdot}{T}\right), \frac{1}{T}b_1\left(\frac{\cdot}{T}\right), c_1\left(\frac{\cdot}{T}\right)\right)$$

defined on $[0,T]$, yields solution $\mathrm{x}_i^T(\cdot) \equiv \mathrm{x}_i^1(\frac{\cdot}{T})$ to (1)

1. $\theta_1 = (a_1, b_1, c_1) \in H^1((0,1);\mathbb{R}^{d+1}) \times L^2((0,T);\mathbb{S}^{d-1})$ yields x_i^1 solution to (1) on $[0,1]$. Then

$$\theta_T(\cdot) = \left(\frac{1}{T}a_1\left(\frac{\cdot}{T}\right), \frac{1}{T}b_1\left(\frac{\cdot}{T}\right), c_1\left(\frac{\cdot}{T}\right)\right)$$

defined on $[0,T]$, yields solution $\mathrm{x}_i^T(\cdot) \equiv \mathrm{x}_i^1(\frac{\cdot}{T})$ to (1)

2. In turn,

$$\begin{aligned}\mathbf{E}(\boldsymbol{X}_T(T)) + \lambda \int_0^T |\theta_T(t)|^2 dt \\ = \mathbf{E}(\boldsymbol{X}_1(1)) + \frac{\lambda}{T}\int_0^1 |(a_1(s), b_1(s))|^2 ds + \lambda.\end{aligned}$$

1. $\theta_1 = (a_1, b_1, c_1) \in H^1((0,1);\mathbb{R}^{d+1}) \times L^2((0,T);\mathbb{S}^{d-1})$ yields x_i^1 solution to (1) on $[0,1]$. Then

$$\theta_T(\cdot) = \left(\frac{1}{T}a_1\left(\frac{\cdot}{T}\right), \frac{1}{T}b_1\left(\frac{\cdot}{T}\right), c_1\left(\frac{\cdot}{T}\right)\right)$$

   defined on $[0,T]$, yields solution $\mathrm{x}_i^T(\cdot) \equiv \mathrm{x}_i^1(\frac{\cdot}{T})$ to (1)
2. In turn,

$$\begin{aligned} &\mathbf{E}(\boldsymbol{X}_T(T)) + \lambda \int_0^T |\theta_T(t)|^2 dt \\ &\qquad = \mathbf{E}(\boldsymbol{X}_1(1)) + \frac{\lambda}{T}\int_0^1 |(a_1(s), b_1(s))|^2 ds + \lambda. \end{aligned}$$

3. Take interpolation control (on $(0,1)$), stretch it out to $(0,T)$, and compare with θ_T.

1. $\theta_1 = (a_1, b_1, c_1) \in H^1((0,1);\mathbb{R}^{d+1}) \times L^2((0,T);\mathbb{S}^{d-1})$ yields x_i^1 solution to (1) on $[0,1]$. Then

$$\theta_T(\cdot) = \left(\frac{1}{T}a_1\left(\frac{\cdot}{T}\right), \frac{1}{T}b_1\left(\frac{\cdot}{T}\right), c_1\left(\frac{\cdot}{T}\right)\right)$$

   defined on $[0,T]$, yields solution $\mathrm{x}_i^T(\cdot) \equiv \mathrm{x}_i^1(\frac{\cdot}{T})$ to (1)

2. In turn,

$$\begin{aligned}\mathbf{E}(\boldsymbol{X}_T(T)) + \lambda\int_0^T |\theta_T(t)|^2 dt \\ = \mathbf{E}(\boldsymbol{X}_1(1)) + \frac{\lambda}{T}\int_0^1 |(a_1(s), b_1(s))|^2 ds + \lambda.\end{aligned}$$

3. Take interpolation control (on $(0,1)$), stretch it out to $(0,T)$, and compare with θ_T.

Corollary

In this setting, $T \to \infty$ is equivalent to $\lambda \to 0$.

Interpolation, Controllability

1. *Combinatorics.* For (1): [Li, Lin, Shen '22], [Ruiz-Balet, Zuazua '22]. Distinct targets if $d = m$.
2. *Lie algebra.* For

$$\dot{\mathrm{x}}_i(t) = \theta(t)\sigma(\mathrm{x}_i(t)), \tag{3}$$

with $\sigma \in C^{0,1} \cap C^1(\mathbb{R})$ element-wise: [Agrachev, Sarychev '22]

Interpolation, Controllability

1. *Combinatorics.* For (1): [Li, Lin, Shen '22], [Ruiz-Balet, Zuazua '22]. Distinct targets if $d = m$.
2. *Lie algebra.* For
$$\dot{\mathrm{x}}_i(t) = \theta(t)\sigma(\mathrm{x}_i(t)), \tag{3}$$
with $\sigma \in C^{0,1} \cap C^1(\mathbb{R})$ element-wise: [Agrachev, Sarychev '22]

A digression - little homotopy method inspired by [Coron-Trélat '04]:

Proposition [Esteve-Yagüe, G., Pighin, Zuazua '22b]

Suppose $d \geqslant n$. Fix $\boldsymbol{X}^1 \in \mathbb{R}^{d\times n}$ with

$$\mathrm{span}\{\sigma(\mathrm{x}_1^1), \ldots, \sigma(\mathrm{x}_n^1)\} = \mathbb{R}^d$$

Then $\exists r, C > 0$ such that $\forall \boldsymbol{X}^0 \in B_r(\boldsymbol{X}^1)$, $\exists \theta \in L^\infty((0,1); \mathbb{R}^{d\times d})$ for which the solutions x_i to (3) with $\mathrm{x}_i(0) = \mathrm{x}_i^0$ satisfy $\mathrm{x}_i(1) = \mathrm{x}_i^1$ $\forall i \in [n]$. Moroever

$$\|\theta\|_{L^\infty} \leqslant \frac{C}{T}|\boldsymbol{X}^1 - \boldsymbol{X}^0|.$$

Generalization: a statistical approach

Focus on dynamics (3).

Generalization: a statistical approach

Focus on dynamics (3).

- Look at $\{x^{(i)}, y^{(i)}\}_{i\in[n]} \subset \mathbb{R}^d \times \mathbb{R}^m$ as i.i.d. samples from unknown joint law $\mu \in \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^m)$. Then $f(x) := \mathbb{E}(y|x)$ which minimizes $\mathbb{E}_{(x,y)\sim\mu}|f(x) - y|^2$ over all functions f.

Generalization: a statistical approach

Focus on dynamics (3).

- Look at $\{x^{(i)}, y^{(i)}\}_{i\in[n]} \subset \mathbb{R}^d \times \mathbb{R}^m$ as i.i.d. samples from unknown joint law $\mu \in \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^m)$. Then $f(x) := \mathbb{E}(y|x)$ which minimizes $\mathbb{E}_{(x,y)\sim\mu}|f(x) - y|^2$ over all functions f.
- Associated to (2): population risk minimization

$$\min_{\substack{\theta\in L^2((0,T);\mathbb{R}^{d\times d})\\ \mathrm{x}^\theta \text{ solves (3)}\\ \mathrm{x}^\theta(0)=x}} \mathbb{E}_{(x,y)\sim\mu}|P\mathrm{x}^\theta(T) - y|^2 + \lambda \int_0^T |\theta(t)|^2 dt \tag{4}$$

Generalization: a statistical approach

Focus on dynamics (3).

- Look at $\{x^{(i)}, y^{(i)}\}_{i\in[n]} \subset \mathbb{R}^d \times \mathbb{R}^m$ as i.i.d. samples from unknown joint law $\mu \in \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^m)$. Then $f(x) := \mathbb{E}(y|x)$ which minimizes $\mathbb{E}_{(x,y)\sim\mu}|f(x) - y|^2$ over all functions f.
- Associated to (2): population risk minimization

$$\min_{\substack{\theta\in L^2((0,T);\mathbb{R}^{d\times d})\\ \mathrm{x}^\theta \text{ solves (3)}\\ \mathrm{x}^\theta(0)=x}} \mathbb{E}_{(x,y)\sim\mu}|P\mathrm{x}^\theta(T) - y|^2 + \lambda\int_0^T |\theta(t)|^2 dt \qquad (4)$$

- Generalization: θ_n minimizer of J_n in (2), and θ^* of J in (4), then $\exists\alpha > 0$:

$$\boxed{\mathbb{E}_{(x,y)\sim\mu}\,|P\mathrm{x}^{\theta_n}(T) - y|^2 - \mathbb{E}_{(x,y)\sim\mu}|P\mathrm{x}^{\theta^*}(T) - y|^2 = \mathcal{O}\left(\frac{1}{n^\alpha}\right)}$$

What is known

1. [E, Han, Li '19]:
   - Pontryagin Maximum Principle for both (2) and (4);
   - Hamiltonian $\theta \mapsto H(x,p,\theta) = p \cdot \theta\sigma(x) + \lambda|\theta|^2$ strongly concave for any (x,p)
   - given θ^*, with high probability $\exists\theta_n$ critical point of Hamiltonian for (2) such that

$$\mathbb{E}_{(x,y)\sim\mu}\,|P\mathrm{x}^{\theta_n}(T) - y|^2 - \mathbb{E}_{(x,y)\sim\mu}|P\mathrm{x}^{\theta^*}(T) - y|^2 \leqslant \frac{C(d)}{n^{\frac{1}{2}-\epsilon}}$$

     with high probability, for any $\epsilon > 0$.
   - Ensuring that θ_n is global minimizer: true when $T \ll 1$, so $\lambda \gg 1$!

What is known

1. [E, Han, Li '19]:
   - Pontryagin Maximum Principle for both (2) and (4);
   - Hamiltonian $\theta \mapsto H(x,p,\theta) = p\cdot\theta\sigma(x) + \lambda|\theta|^2$ strongly concave for any (x,p)
   - given θ^*, with high probability $\exists\theta_n$ critical point of Hamiltonian for (2) such that

     $$\mathbb{E}_{(x,y)\sim\mu}\,|P\mathrm{x}^{\theta_n}(T) - y|^2 - \mathbb{E}_{(x,y)\sim\mu}|P\mathrm{x}^{\theta^*}(T) - y|^2 \leqslant \frac{C(d)}{n^{\frac{1}{2}-\epsilon}}$$

     with high probability, for any $\epsilon > 0$.
   - Ensuring that θ_n is global minimizer: true when $T \ll 1$, so $\lambda \gg 1$!
2. [Bonnet et al. '22]:
   - For $\lambda \gg 1$, J_n strongly convex on any L^2 ball
   - Mean-field PMP ... rate $\mathcal{O}\left(\frac{1}{n^{\frac{1}{d}}}\right)$

An observation with C. Letrouit and P. Rigollet

1. Strong convexity on any $B \subset L^2((0,T);\mathbb{R}^{d\times d})$

$$\|\theta_1 - \theta_2\|_{L^2}^2 \lesssim \left|\nabla J_n(\theta_1) - \nabla J_n(\theta_2)\right|, \qquad \forall \theta_1, \theta_2 \in B.$$

An observation with C. Letrouit and P. Rigollet

1. Strong convexity on any $B \subset L^2((0,T);\mathbb{R}^{d\times d})$

$$\|\theta_1 - \theta_2\|_{L^2}^2 \lesssim |\nabla J_n(\theta_1) - \nabla J_n(\theta_2)|, \qquad \forall \theta_1, \theta_2 \in B.$$

2. Then

$$\begin{aligned}\|\theta_n - \theta^*\|_{L^2}^2 &\lesssim |\nabla J_n(\theta_n) - \nabla J_n(\theta^*)| \\ &= |\nabla J(\theta^*) - \nabla J_n(\theta^*)| \\ &= \left|\mathbb{E}_{(x,y)\sim\mu} \nabla_\theta \ell(\mathrm{x}^{\theta^*}(T), y) - \frac{1}{n}\sum_{i\in[n]} \nabla_\theta \ell(\mathrm{x}_i^{\theta^*}(T), y^{(i)})\right|\end{aligned}$$

An observation with C. Letrouit and P. Rigollet

1. Strong convexity on any $B \subset L^2((0,T);\mathbb{R}^{d\times d})$
$$\|\theta_1-\theta_2\|_{L^2}^2 \lesssim |\nabla J_n(\theta_1)-\nabla J_n(\theta_2)|, \qquad \forall \theta_1,\theta_2\in B.$$

2. Then
$$\begin{aligned}\|\theta_n-\theta^*\|_{L^2}^2 &\lesssim |\nabla J_n(\theta_n)-\nabla J_n(\theta^*)| \\ &= |\nabla J(\theta^*)-\nabla J_n(\theta^*)| \\ &= \left|\mathbb{E}_{(x,y)\sim\mu}\nabla_\theta \ell(\mathrm{x}^{\theta^*}(T),y)-\frac{1}{n}\sum_{i\in[n]}\nabla_\theta\ell(\mathrm{x}_i^{\theta^*}(T),y^{(i)})\right|\end{aligned}$$

3. θ^* is fixed, not random. Sum of bounded, independent random variables, compared with expectation $\Rightarrow$ **concentration of measure** (Hoeffding inequality): $\forall\delta>0$, $\exists\kappa>0$ such that $\forall n\geqslant 1$,
$$\mathbb{P}(\|\theta_n-\theta^*\|_{L^2}^2\leqslant \kappa/\sqrt{n})\geqslant 1-\delta.$$
Bound $J(\theta_n)-J(\theta^*)$ from above and get $\mathcal{O}(\frac{1}{\sqrt{n}})$ rate.

Improving $1/T$

We consider

$$\boxed{\inf_{\substack{c\in L^2((0,T);\mathbb{R}^d)\\ \mathrm{x}_i(\cdot)\text{ solves (1)}}} \int_0^T \frac{1}{n}\sum_{i\in[n]} \left|P\mathrm{x}_i(t) - f(x^{(i)})\right|^2 dt + \int_0^T |c(t)|^2 dt} \tag{5}$$

Controls $a \in L^\infty(\mathbb{R}_+;\mathbb{R}^d)$ and $b \in L^\infty(\mathbb{R}_+)$ assumed fixed in (1) (need L^2-penalties and compactness simultaneously)

Can also consider dynamics as (3), or (a, b) can be optimized over $\mathbb{S}^d$.

Two assumptions

Recall, for $\boldsymbol{X} = [\mathrm{x}_1 \cdots \mathrm{x}_n] \in \mathbb{R}^{d\times n}$:

$$\mathbf{E}(\boldsymbol{X}) := \frac{1}{n}\sum_{i\in[n]} \left| P\mathrm{x}_i - f(x^{(i)})\right|^2.$$

Assumption 1.

Set $\mathcal{Z} := \{\boldsymbol{Z} \in \mathbb{R}^{d\times n} : \mathbf{E}(\boldsymbol{Z}) = 0\}$. Then $P \in \mathbb{R}^{m\times d}$ is such that

$$\kappa_1 \mathrm{dist}(\boldsymbol{X}, \mathcal{Z})^2 \leqslant \mathbf{E}(\boldsymbol{X}) \leqslant \kappa_2 \mathrm{dist}(\boldsymbol{X}, \mathcal{Z})^2$$

for some $\kappa_2 \geqslant \kappa_1 > 0$, and $\forall \boldsymbol{X} \in \mathbb{R}^{d\times n}$.

- Lower bound: global Lojasiewicz inequality for analytic functions

Assumption 2.

Fix $\boldsymbol{X}^0 = [\mathrm{x}_1^0 \cdots \mathrm{x}_n^0] \in \mathbb{R}^{d\times n}$. We assume $\exists c \in L^2((0,1);\mathbb{R}^d)$ such that the matrix $\boldsymbol{X} \in C^0([0,1];\mathbb{R}^{d\times n})$ with columns $\mathrm{x}_i(\cdot)$ solutions to (1) with $\mathrm{x}_i(0) = \mathrm{x}_i^0$, satisfies $\boldsymbol{X}(1) \in \mathcal{Z}$.

Moreover, $\exists C(n) > 0$,

$$\int_0^1 |c(t)|^2 dt \leqslant C(n)\,\mathrm{dist}(\boldsymbol{X}^0, \mathcal{Z})^2.$$

Assumption 2.

Fix $\boldsymbol{X}^0 = [\mathrm{x}_1^0 \cdots \mathrm{x}_n^0] \in \mathbb{R}^{d\times n}$. We assume $\exists c \in L^2((0,1);\mathbb{R}^d)$ such that the matrix $\boldsymbol{X} \in C^0([0,1];\mathbb{R}^{d\times n})$ with columns $\mathrm{x}_i(\cdot)$ solutions to (1) with $\mathrm{x}_i(0) = \mathrm{x}_i^0$, satisfies $\boldsymbol{X}(1) \in \mathcal{Z}$.

Moreover, $\exists C(n) > 0$,

$$\int_0^1 |c(t)|^2 dt \leqslant C(n)\,\mathrm{dist}(\boldsymbol{X}^0, \mathcal{Z})^2.$$

- When $d > m$, then $P \in \mathbb{R}^{d\times n}$ is generically surjective and

$$\mathcal{Z} = \left\{ [\mathrm{z}_1 \cdots \mathrm{z}_n] \in \mathbb{R}^{d\times n} \colon \mathrm{z}_i \in P^{-1}\left\{ f(x^{(i)}) \right\} \right\}$$

So $\mathrm{x}_i(1) \in P^{-1}\left\{ f(x^{(i)}) \right\}$ for all $i \in [n]$ and

$$\int_0^1 |c(t)|^2 dt \leqslant C(n) \inf_{\substack{[\mathrm{z}_1 \cdots \mathrm{z}_n] \in \mathbb{R}^{d\times n} \\ \mathrm{z}_i \in P^{-1}\{f(x^{(i)})\}}} \sum_{i\in[n]} \left| \mathrm{x}_i^0 - \mathrm{z}_i \right|^2.$$

Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22a]

Suppose $m = d$ and $P = \mathrm{Id}$. Then $\exists T_*, \omega > 0$, $C \geqslant 1$ such that for $T \geqslant T_*$, any global minimizer $c_T \in L^2((0,T);\mathbb{R}^d)$ to (5) and x_i^T solution to (1) satisfy

$$\sum_{i\in[n]} \left|\mathrm{x}_i^T(t) - f(x^{(i)})\right|^2 + |c_T(t)|^2 \leqslant \left(C \sum_{i\in[n]} \left|x^{(i)} - f(x^{(i)})\right|^2 \right) e^{-\omega t}$$

$\forall t \in [0,T]$.

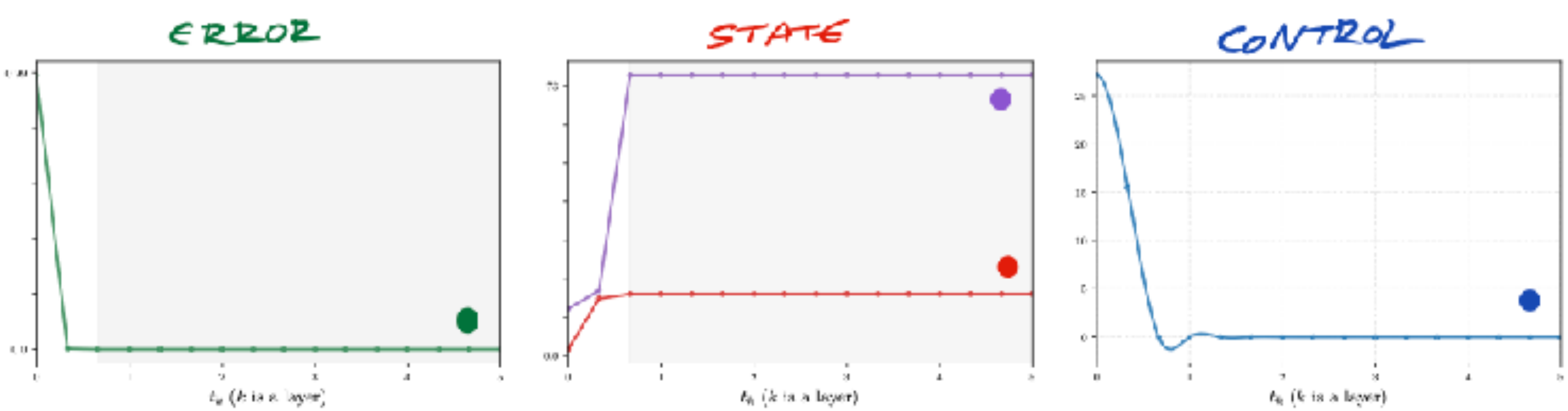
Theorem [Esteve-Yagüe, G., Pighin, Zuazua, '22b]

Replace $(\cdot)_+$ by $\sigma \in L^\infty(\mathbb{R})$ in (1) and let $d > m$. Then $\exists T_*, \omega > 0$, $C \geqslant 1$ such that for $T \geqslant T_*$, any global minimizer $c_T \in L^2((0,T);\mathbb{R}^d)$ to (5) and x_i^T solution to (1) satisfy

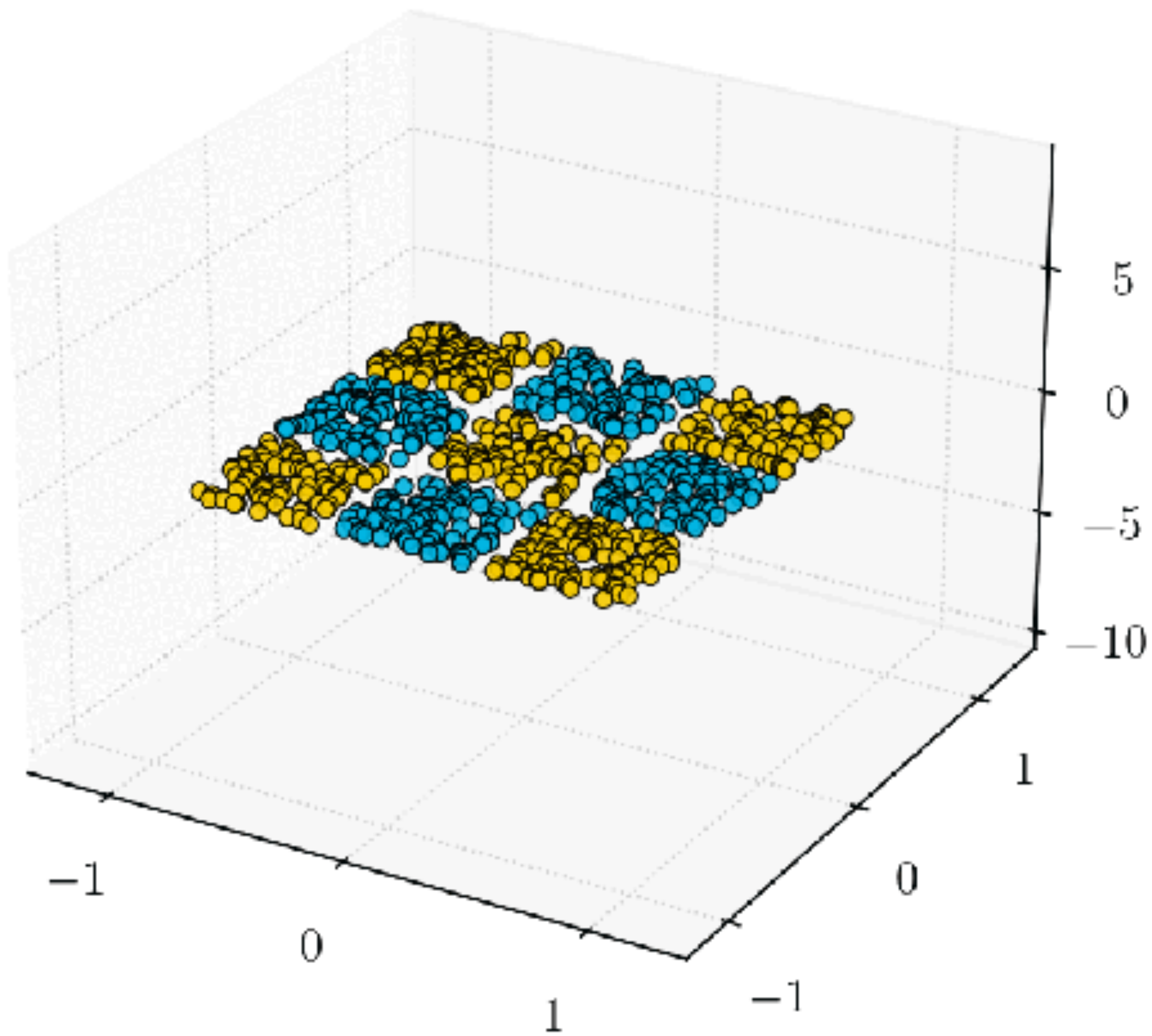
$$\sum_{i\in[n]} \left|P\mathrm{x}_i^T(t) - f(x^{(i)})\right|^2 + \inf_{\substack{[\mathrm{z}_1\cdots\mathrm{z}_n]\in\mathbb{R}^{d\times n}\\ \mathrm{z}_i\in P^{-1}\{f(x^{(i)})\}}} \sum_{i\in[n]} \left|\mathrm{x}_i^T(t) - \mathrm{z}_i\right|^2 + |c_T(t)|^2$$

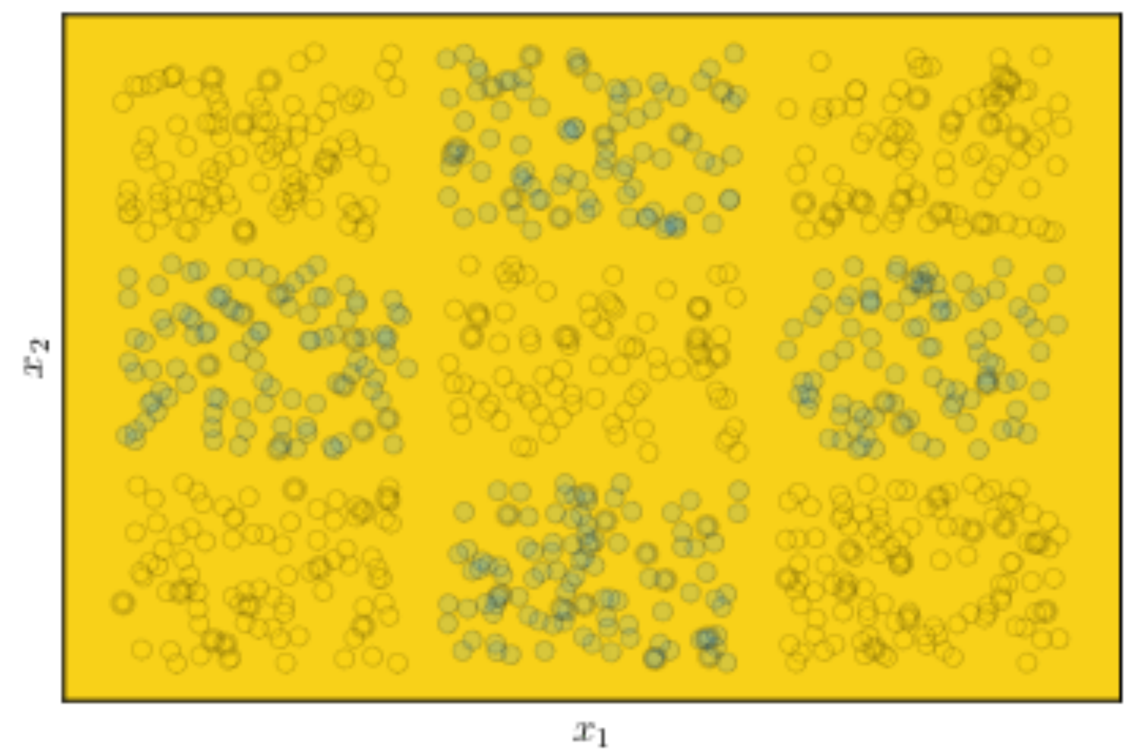
$$\leqslant \left(C \inf_{\substack{[\mathrm{z}_1\cdots\mathrm{z}_n]\in\mathbb{R}^{d\times n}\\ \mathrm{z}_i\in P^{-1}\{f(x^{(i)})\}}} \sum_{i\in[n]} \left|x^{(i)} - \mathrm{z}_i\right|^2 \right) e^{-\omega t}$$

$\forall t \in [0,T]$.



It's faster





Takeaway: when possible, proceed in *model predictive control manner:* start with small T, evaluate error, and proceed by increasing T adaptively.

Important tool

Focus on $d = m$, $P = \mathrm{Id}$.

Lemma

$\exists C_1 > 0$ independent of T, $\forall c \in L^2$ and x_i solution to (1):

$$\sup_{t\in[0,T]} \sum_{i\in[n]} \left|\mathrm{x}_i(t) - f(x^{(i)})\right|^2 \leqslant C_1 \Bigg(\sum_{i\in[n]} \left|x^{(i)} - f(x^{(i)})\right|^2 + \int_0^T \sum_{i\in[n]} \left|\mathrm{x}_i(t) - f(x^{(i)})\right|^2 dt + \int_0^T |c(t)|^2 dt \Bigg)$$

Ingredients

1. $c_{\mathrm{aux}}(t) := c_1(t)1_{[0,1]}(t)$, where c_1 ensures controllability. As c_T is optimal and c_{aux} is not, $\exists C_2 > 0$ independent of T:

$$\int_0^T \sum_{i\in[n]} \left|\mathrm{x}_i^T(t) - f(x^{(i)})\right|^2 dt + \int_0^T |c_T(t)|^2 dt \leqslant C_2 \sum_{i\in[n]} \left|x^{(i)} - f(x^{(i)})\right|^2$$

   Lemma yields pointwise bound uniform in T.

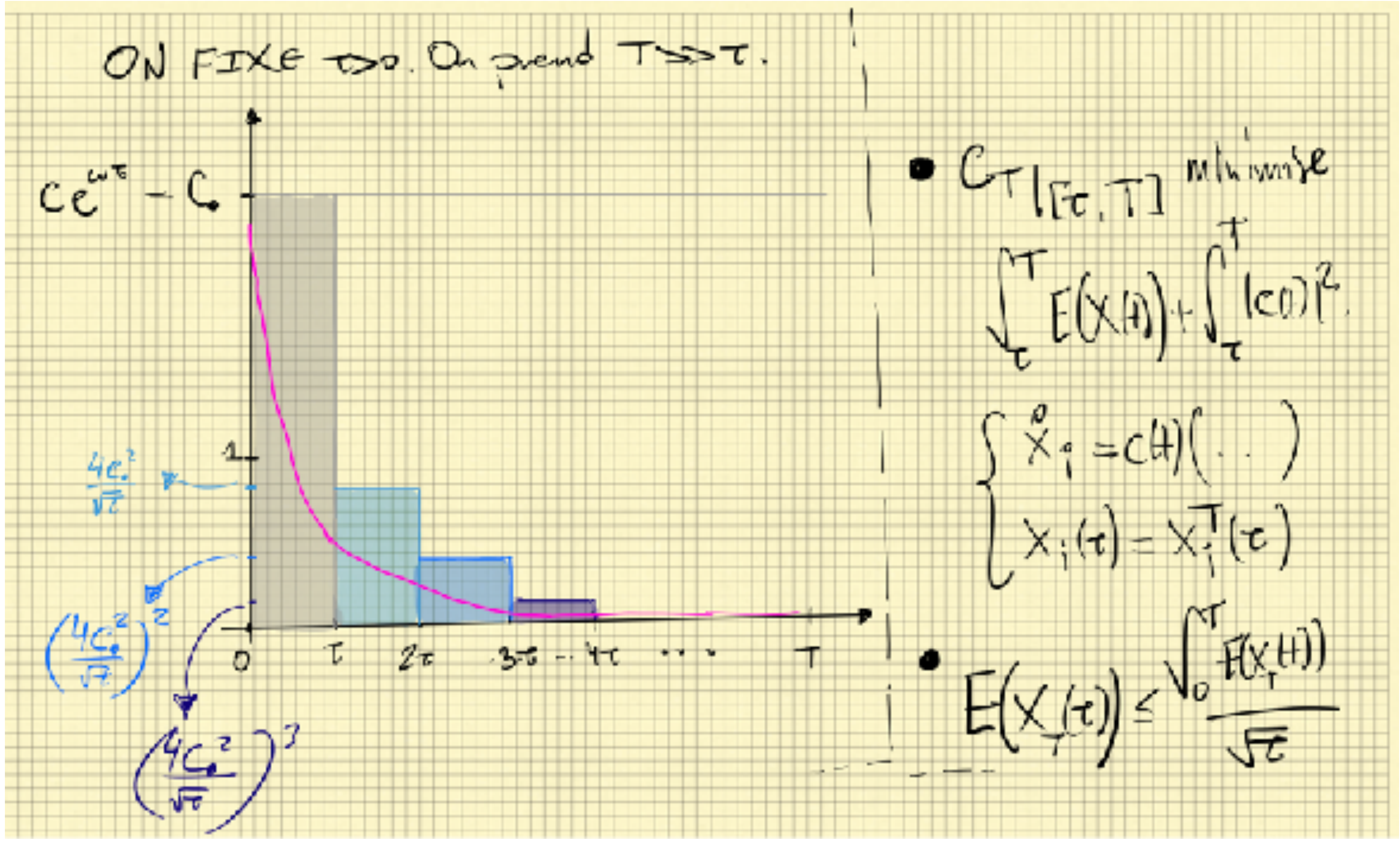
2. Shrink time-intervals:

Ingredients

1. $c_{\text{aux}}(t) := c_1(t)1_{[0,1]}(t)$, where c_1 ensures controllability. As c_T is optimal and c_{aux} is not, $\exists C_2 > 0$ independent of T:

$$\int_0^T \sum_{i\in[n]} \left|\mathrm{x}_i^T(t) - f(x^{(i)})\right|^2 dt + \int_0^T |c_T(t)|^2 dt \leqslant C_2 \sum_{i\in[n]} \left|x^{(i)} - f(x^{(i)})\right|^2$$

Lemma yields pointwise bound uniform in T.

2. 

Extensions

- Using this method, we can't have exponential decay with BV-penalty for (a, b) as norm tracks singularities unlike L^2. We can at most get decay of time-averages of the error and controls.
- **Results are more general.** Per [Esteve-Yagüe, G., Pighin, Zuazua '22a; G., Zuazua '22]: controllable PDE

$$y_t(t,x) - Ay(t,x) + Bu(t,x) = f(y(t,x))$$

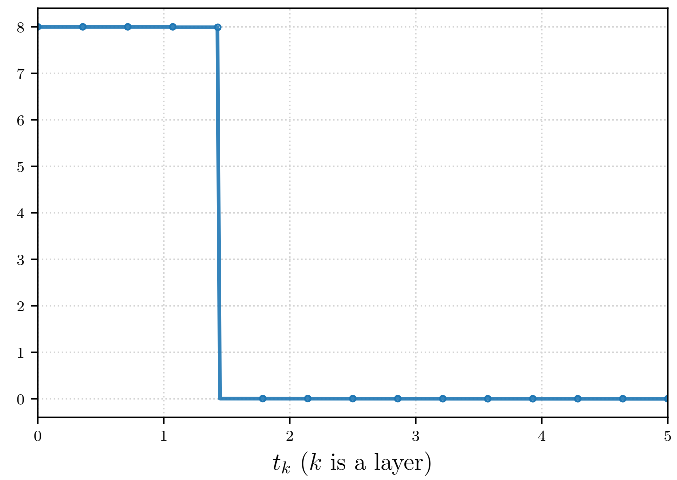
with Lipschitz (possibly non-smooth) nonlinearity f and cost

$$\phi(y(T)) + \int_0^T |y(t) - \bar{y}|^2_{\mathscr{H}} dt + \int_0^T |u(t)|^2_{\mathscr{U}} dt$$

where $\bar{y}$ is any steady state $\Rightarrow$ exponential turnpike without smoothness or smallness assumptions!

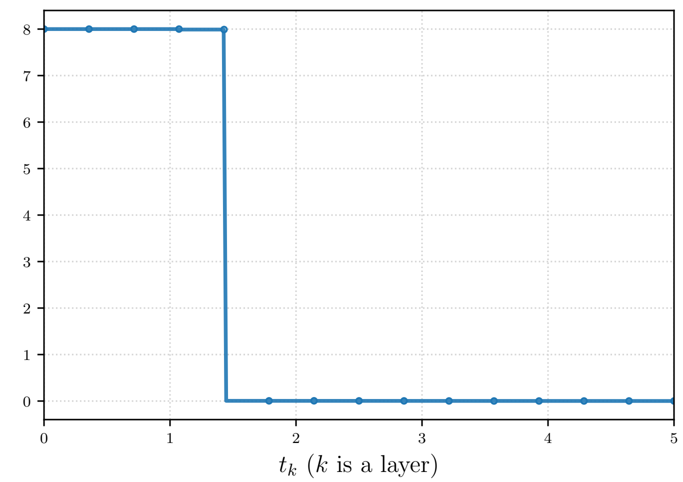
Further comments

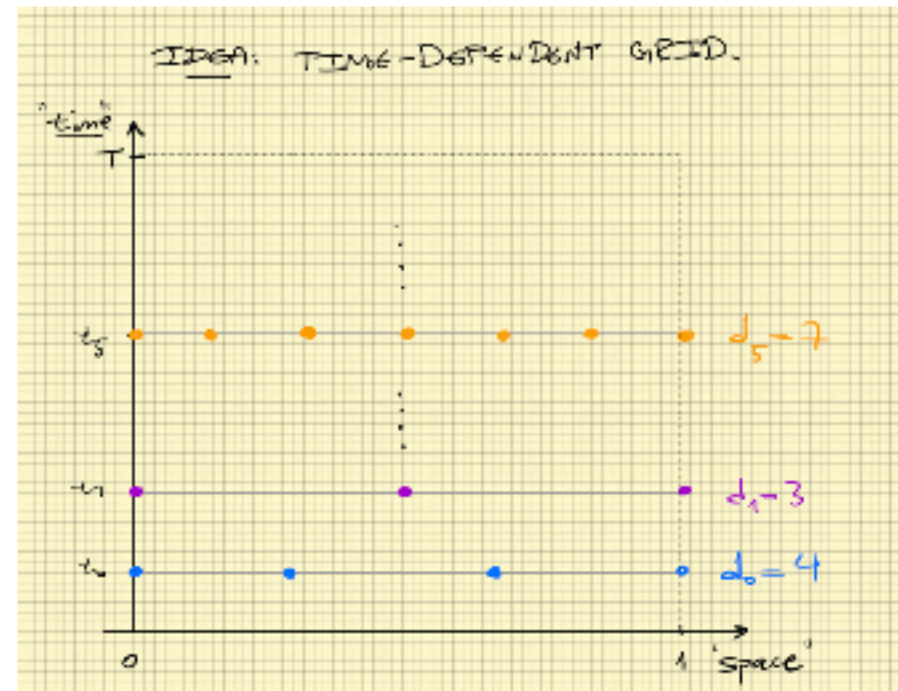
1. $L^1(0,T)$-penalties for (3) [Esteve-Yagüe, G. '22]:



Further comments

1. $L^1(0,T)$-penalties for (3) [Esteve-Yagüe, G. '22]:





2. Variable-width ResNets:

$$\mathrm{x}^{[k+1]} = \Pi^{[k]}\mathrm{x}^{[k]} + c^{[k]}\sigma(a^{[k]}\mathrm{x}^{[k]})$$

where $\Pi^{[k]} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$, $a^{[k]} \in \mathbb{R}^{d_{k+1}\times d_k}$, $c^{[k]} \in \mathbb{R}^{d_{k+1}\times d_{k+1}}$. So,

$$\partial_t \mathrm{x}(t,z) = \int_0^1 c(t,z,\zeta)\sigma(a(t,z,\zeta)\mathrm{x}(t,\zeta))d\zeta \qquad (0,T)\times(0,1)$$

Helpful for structured controls (convolutional neural networks).

Outlook

1. Control
   - Exponential decay/turnpike with BV-penalty for (a, b)?
   - Using control: are feedback controls for $n \gg 1$ trajectories possible, useful?
   - Extrapolating to control: can we get robustness using the lens of many data and statistics?
2. Unsupervised learning/**generative modeling** with normalizing flows (E. Vanden-Eijnden et al.).
   - NF: diffeomorphism $\mathfrak{T} : \mathbb{R}^d \to \mathbb{R}^d$ optimized to transport $\{z^{(i)}\}_{i\in[n]} \subset \mathbb{R}^d$ samples from a known law ρ_0 (Gaussian with unit variance) to unknown target law ρ_1 of which we know samples $\{x^{(i)}\}_{i\in[n]} \subset \mathbb{R}^d$.
   - Parametrizing $\mathfrak{T}$ by the flow of a neural ODE, and then solving an optimal control problem (KL divergence) is quite practical.

Merci pour votre attention!

Bibliographie:

1. [Esteve-Yagüe, G., Pighin, Zuazua '22a]: https://arxiv.org/abs/2011.11091, Nonlinearity
2. [Esteve-Yagüe, G., Pighin, Zuazua '22b]: https://arxiv.org/abs/2008.02491
3. [G., Zuazua '22]: https://arxiv.org/abs/2202.04097, Acta Numerica
4. [Esteve-Yagüe, G. '22]: https://arxiv.org/abs/2102.13566, submitted