Some Topics on Dynamics and Machine Learning

Eduardo Sontag

Northeastern University

Electrical and Computer Engineering & BioEngineering Departments

Affiliate:
NEU: Mathematics & Chemical Engineering
HMS: Laboratory of Systems Pharmacology
MIT: LIDS

Joint work with:

Leilei Cui / Zhong-Ping Jiang

Milad Siami / Arthur Castello B. de Oliveira

Zexiang Liu / Necmiye Ozay

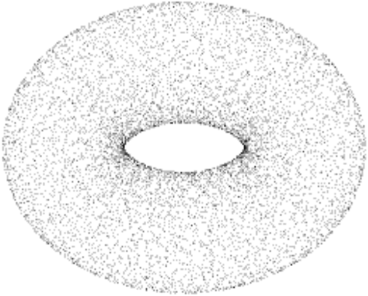
Matthew D. Kvalheim

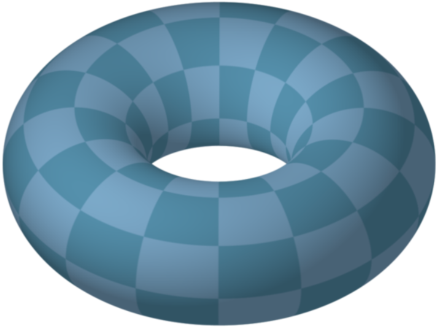
2024-08-21:22:12:59

Outline

- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

Dimensionality reduction of data





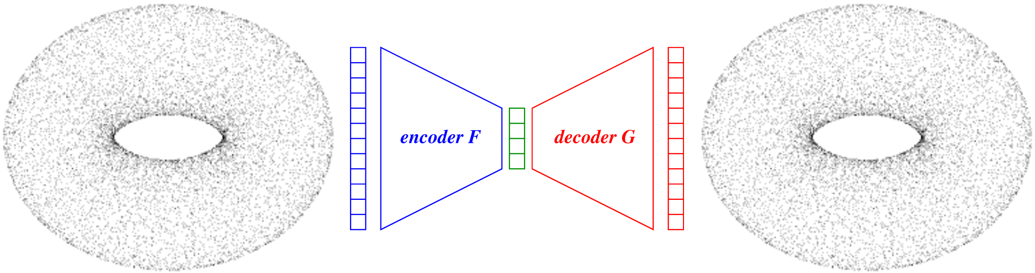
"manifold hypothesis:" data set in $\mathbb{R}^n$ lies on some k-dimensional submanifold $K \subset \mathbb{R}^n$

$\implies$ data can be parametrized *locally* by $k < n$ real numbers

classical approaches like PCA to learn these parameters work well when K is linear

but not when K is nonlinear

Autoencoding as nonlinear dimensionality reduction



$\mathcal{C}^0$ **encoder** $F\colon \mathbb{R}^n \to \mathbb{R}^k$ and **decoder** $G\colon \mathbb{R}^k \to \mathbb{R}^n$ (typically "neural networks")

ideal autoencoder: $G(F(x)) = x \ \forall\, x \in K$ (data manifold)

useful also for "interpolation" of images by "walking along" latent space

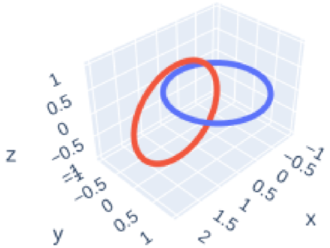
and for encoding vector fields in reduced space

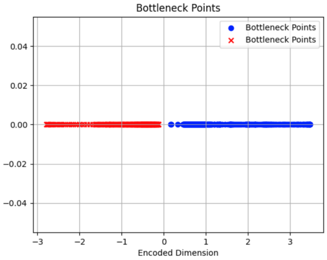
obvious obstruction: $\implies$ K homeomorphic to subset of $\mathbb{R}^k$ (generally false for k-dim K)

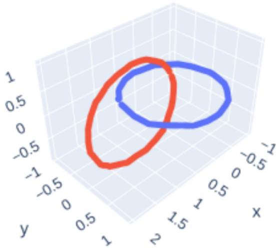
so relax to "approximate" versions:

Numerical example using a “deep” NN

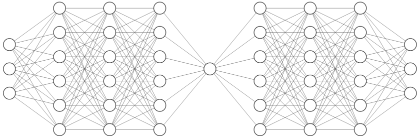
(Python TensorFlow, Adaptive Moment Estimation [Adam] optimizer, L^2 error loss)







K = pair of circles in $\mathbb{R}^3$
after thickening, then deleting small intervals,
diffeo to a pair of disjoint intervals in $\mathbb{R}$
encoder $F\colon \mathbb{R}^3 \to \mathbb{R}$ = any extension of this diffeo
decoder $G\colon \mathbb{R} \to \mathbb{R}^3$ = any extension of inverse diffeo



Semi-global autoencoders always exist (Kvalheim & EDS 2024)

let $\mathcal{F}^{\ell,m}$ be dense in the space of continuous functions $\mathbb{R}^\ell \to \mathbb{R}^m$
(e.g., the collection of possible feedforward neural network I/O functions)

Theorem 1: Let:

- $K \subset \mathbb{R}^n =$ finite union of disjoint compact $\leq k$-dimensional submanifolds (with or without boundary)
- μ, $\partial\mu$ any smooth measures on K, ∂K

Then: $\forall\ \delta > 0$, $\forall$ finite data subset $S \subset K$, $\exists$ a closed set $K_0 \subset K$ s.t.:

- $K_0 \cap S = \varnothing$, $\mu(K_0) < \delta$, $\partial\mu(K_0 \cap \partial K) < \delta$;
- $M \setminus K_0$ is connected for each component M of K;
- for each $\varepsilon > 0$ there are functions $F \in \mathcal{F}^{n,k}$, $G \in \mathcal{F}^{k,n}$ s.t.

$$\sup_{x \in K \setminus K_0} \|G(F(x)) - x\| < \varepsilon.$$

$\Longrightarrow$ data S can be reconstructed with error ε
and generalization error also uniformly smaller than ε (with probability $> 1 - \delta$)

. . . but not *global* (approx) autoencoders (Kvalheim & EDS 2024)

Theorem 2:

Consider k-dimensional compact submanifold without boundary $K \subset \mathbb{R}^n$; $r_K > 0$ its reach.

Then: for any continuous functions $F\colon \mathbb{R}^n \to \mathbb{R}^k$ and $G\colon \mathbb{R}^k \to \mathbb{R}^n$,

$$\sup_{x \in K} \|G(F(x)) - x\| \geq r_K$$

recall:

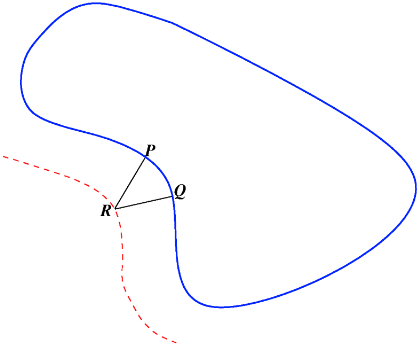
"reach" $r_K > 0$ of K measures "local curvature"

defined as largest r s.t. $\forall\, x \in \mathbb{R}^n$:

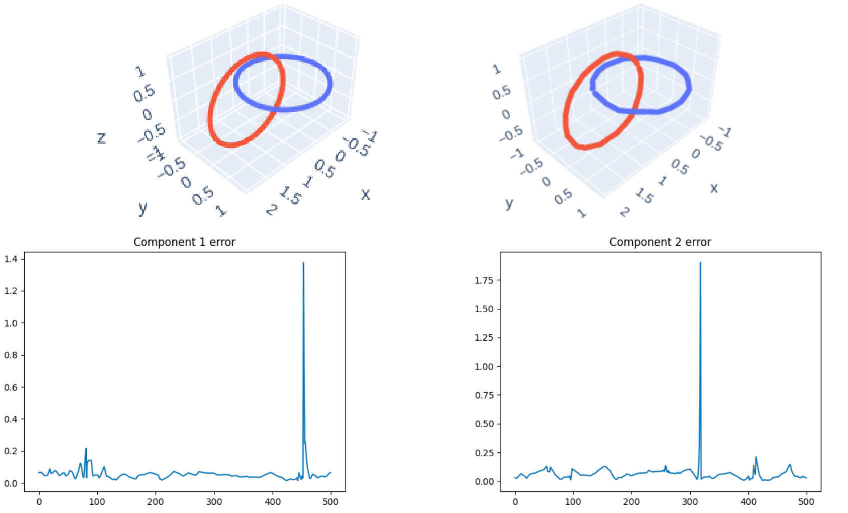
$\mathrm{dist}(x, K) < r \implies x$ has unique projection on K

(nearest point)

(both shown line segments shown have length r_K)



Example: K = union of two unit circles $\Longrightarrow$ L^∞ error $> r_K = 1$


Component 1 error

Component 2 error

errors $\|G(F(x)) - x\|$ vs k-th evenly-spaced point on respective circle

Summary and interpretation for autoencoder training error

Theorem 1 $\implies$ $(\forall\, \varepsilon > 0)$ $(\exists\, F, G)$ making $L^p(\mu)$ loss, $p \in (1, \infty)$, small:

$$\|G(F(x)) - x\|_{p.\mu} \;<\; \varepsilon$$

Theorem 2 $\implies$ (for K w/o boundary, $\forall\, F, G$)

$$\|G(F(x)) - x\|_{\infty} \geq r_K > 0$$

New Theorem (unpublished):
bottleneck dim must be $\geq k$ (for Lipschitz encoding/decoding)

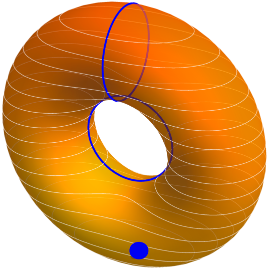
notes: notation means $\left(\int_K \|G(F(x)) - x\|^p d\mu(x)\right)^{1/p}$ for any reasonable measure

p loss is basically $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \|G(F(x_i)) - x_i\|^p$ for training data

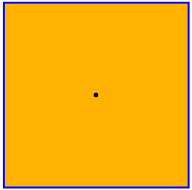
just modify G off $F(K \setminus K_0)$ to make the AE error smaller than some $C_K > 0$ on K_0

Idea behind proof of Theorem 1: negative gradient flow

take any polar Morse function h (e.g. height function below) and $-\nabla h$ flow
consider region of attraction of unique local minimum
remove a fattened version of $\bigcup_q W_s(q)$, $q =$ other critical points ($\mu = 0$)



$\longleftrightarrow$



encoder $F\colon \mathbb{R}^n \to \mathbb{R}^k$ = any extension of this diffeomorphism
decoder $G\colon \mathbb{R}^k \to \mathbb{R}^n$ = any extension of inverse diffeomorphism

can always find such a "codim > 0 set" disjoint from the data

Idea behind proof of Theorem 2

the set $N := \{x \in \mathbb{R}^n \colon \operatorname{dist}(x, K) < r_K\}$

contains line segment from $x \in N$ to unique projection $\rho(x) \in K$ (and ρ continuous)

which implies given

$$\sup_{x \in K} \|G(F(x)) - x\| < r_K$$

that $t \mapsto \rho \circ (tG \circ F|_K + (1-t)\,\mathrm{id}_K)$ is a homotopy of id_K to $\rho \circ G \circ F|_K$

so induced homomorphism on singular homology groups

$$(\rho \circ G \circ F|_K)_* = \rho_* \circ G_* \circ F_* \colon H_k(K) \to H_k(K)$$

(1) equals the identity homomorphism $(\mathrm{id}_K)_*$ induced by id_K, but

(2) this contradicts that this map factors through G_*, which is zero

small print: use $\mathbb{Z}_2$ homology
use factorization through "fattened" $U := G^{-1}(N)$, noncompact manifold, $H_k(U; \mathbb{Z}_2) = 0$
and that $H_k(K; \mathbb{Z}_2) = \mathbb{Z}_2 \neq 0$ when K is a compact (connected) manifold

Extension to vector fields (Kvalheim & EDS, in progress)

Theorem 3 (only partially proved):

Let:

- $K \subset \mathbb{R}^n$ = finite union of disjoint compact $\leq k$-dimensional submanifolds (with or without boundary);
- μ, $\partial\mu$ any smooth measures on K, ∂K;
- f a smooth vector field on $\mathbb{R}^n$ s.t. $f|_K$ is tangent to K.

Then:

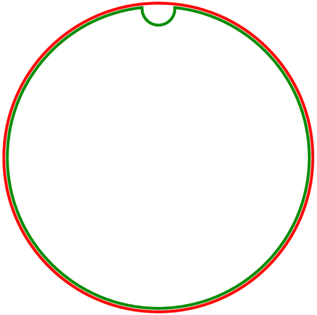
$\forall\ \delta > 0$ and finite set $S \subset K$, $\exists$ closed set $K_0 \subset K$ and smooth vector field g on $\mathbb{R}^k$ s.t.:

- $K_0 \cap S = \varnothing$, $\mu(K_0) < \delta$, $\partial\mu(K_0 \cap \partial K) < \delta$;
- for each ε, $T > 0$ $\exists$ functions $F \in \mathcal{F}^{n,k}$, $G \in \mathcal{F}^{k,n}$ s.t.

$$\sup_{\substack{x \in K \setminus K_0 \\ t \in [-T,T]}} \| G \circ \Phi_g^t \circ F(x) - \Phi_f^t(x) \| < \varepsilon.$$

$\implies$ only obstructions to autoencoding vector field trajectories arise from autoencoding initial conditions

*** True min-max error is often larger than the reach



min max error “reach” is conservative
e.g.: K = green, reach $= \varepsilon \ll 1$
but expect error lower bound much worse

define “**dewrinkled reach of** K” $= 1 - \varepsilon$

witnessed by red circle, which has reach 1,
and projection mapping $T : L \to K$

where $\delta(T) := \max_{y!eL} \|T(y) - y\|$
$=$ maximum deviation of T from the identity

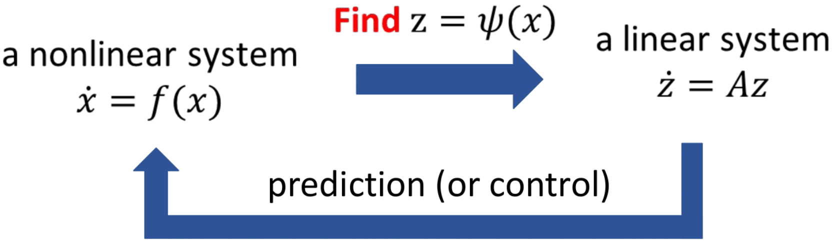
Corollary:
Consider $K \subset \mathbb{R}^n$ be a k-dimensional compact submanifold without boundary.
For any continuous functions $F\colon \mathbb{R}^n \to \mathbb{R}^k$ and $G\colon \mathbb{R}^k \to \mathbb{R}^n$:

$$\sup_{x \in K} \|G(F(x)) - x\| \geq \underbrace{r^*_{K,k}}_{\textbf{dewrinkled reach}} := \sup_{L \in \mathcal{M}_{n,k}, T \in C(L \to K)} \{r_L - \delta(T)\}.$$

Outline

- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

(Global) linearization-based identification and control



a nonlinear system

$\dot{x} = f(x)$

Find $z = \psi(x)$

a linear system

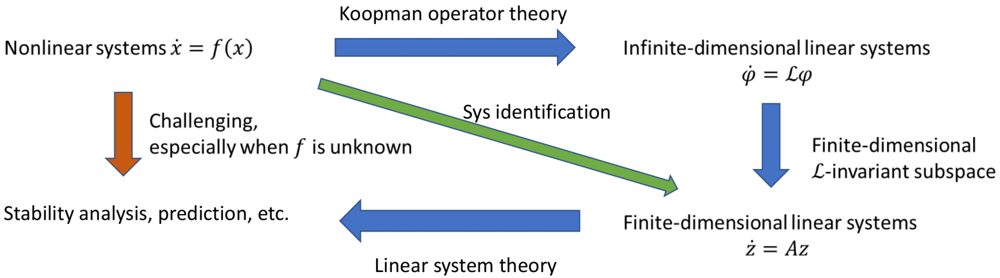
$\dot{z} = Az$

prediction (or control)

for example, $\psi_w(x)$ represented by neural network with weights w

and one attempts to "learn" the network weights and matrix A from data

“Koopman” (& Kalman!) approach


Koopman operator theory
Nonlinear systems $\dot{x} = f(x)$
Infinite-dimensional linear systems
$\dot{\varphi} = \mathcal{L}\varphi$
Challenging,
especially when f is unknown
Sys identification
Finite-dimensional
$\mathcal{L}$-invariant subspace
Stability analysis, prediction, etc.
Finite-dimensional linear systems
$\dot{z} = Az$
Linear system theory

often hard to a linear *(finite-dimensional global, continuous)* representation . . .

- algorithmic problem?
- $\nexists$ representation?

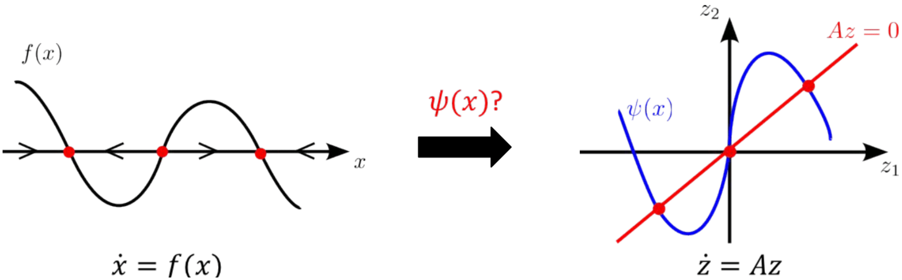
here: show impossibility of *continuous* immersion (one to one) ψ

if there are (more than one) isolated omega limit sets (equilibria, cycles, . . .)

Seems obvious no?

there cannot be multiple isolated equilibria in linear systems . . .

but what about an immersion like this (with singular A)?



need to rule out that this can ever happen (with consistent dynamics)

Setup

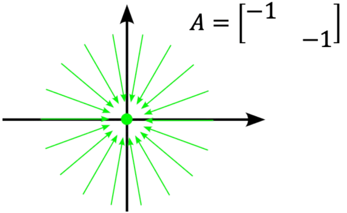
$$\dot{x} = f(x), \;\; x \in \mathcal{X} \;\; \text{(connected subset of manifold } \mathcal{M}\text{)}$$

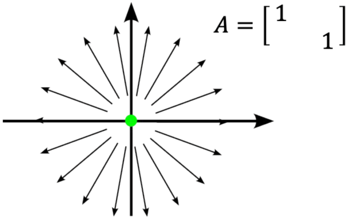
f smooth enough to guarantee uniqueness, completeness, continuous dependence

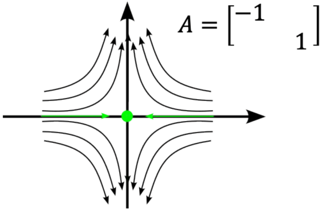
a key property: “closed basins of attraction”

$$\text{for any } \omega\text{-limit set } \Omega, \text{ define: } D^+(\Omega) := \{\xi \in \mathcal{X} \mid \omega^+(\xi) = \Omega\}$$

Observation: for linear systems, and all Ω, $D^+(\Omega)$ closed







Main impossibility result (Liu, Ozay, EDS 2023/4)

$\Psi : \mathcal{X} \to \mathcal{Z}$ continuous, one-to-one, is an *"immersion"* of $\dot{x} = f(x)$ in $\dot{z} = g(z)$ if:

$$\Psi(\varphi_{\mathcal{X}}(t,\xi)) = \varphi_{\mathcal{Z}}(t,\Psi(\xi)) \quad \forall \text{ initial states } \xi \in \mathcal{X} \text{ and times } t \geq 0$$

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\varphi_{\mathcal{X}}"] \arrow[d, "\Psi"'] & \mathcal{X} \arrow[d, "\Psi"] \\
\mathcal{Z} \arrow[r, "\varphi_{\mathcal{Z}}"] & \mathcal{Z}
\end{tikzcd}$$

Theorem: Suppose:

- trajectories of $\dot{x} = f(x)$ on $\mathcal{X}$ are precompact in $\mathcal{X}$, and
- there is more than one but at most a countable number of ω-limit sets.

Then: $\dot{x} = f(x)$ on $\mathcal{X}$ cannot be immersed in a system with closed basins

$\Longrightarrow$ impossible to immerse in linear systems

Our conditions are “necessary” in a sense

no obstructions if weakening:

- uncountably many ω-limit sets (obvious: even linear)
- local immersions
- “approximate” immersions
- unbounded trajectories (Arathoon & Kvalheim)
- discontinuous (or into hybrid system)

Implications to learning from data

recall immersion condition:

$$\Psi(\varphi_{\mathcal{X}}(t,\xi)) = \varphi_{\mathcal{Z}}(t,\Psi(\xi)) \quad \forall \text{ initial states } \xi \in \mathcal{X} \quad \text{and times } t \geq 0$$

given sampling time τ and pairs $\{(x_\ell, x_\ell^+)\}_{\ell=1}^{\infty}$ s.t. $x_\ell^+ = \varphi(\tau, x_\ell)$

continuous linear immersions Ψ **that match first N samples** belong to:

$$\mathcal{F}(\tau, N) \; = \; \{\Psi \mid \exists\, A \in \mathbb{R}^{m\times m}, \Psi(x_\ell^+) = e^{A\tau}\Psi(x_\ell), \forall l = 1, \cdots, N\}$$

Corollary: Suppose

- $\{x_\ell\}_{\ell=1}^{\infty}$ is a dense subset of $\mathcal{X}$ (e.g. random sampling),
- trajectories of $\dot{x} = f(x)$ on $\mathcal{X}$ are precompact in $\mathcal{X}$, and
- Ψ distinguishes between at least two ω-limit sets.

Then for all small enough sampling times τ and large enough N, $\Psi \notin \mathcal{F}(\tau, N)$.

Intuitively: immersion candidates F that can distinguish at least two ω-limit sets will be ruled out as more data is collected (if sampling time small enough)

*** *Discontinuous immersions much easier*

discontinuous less interesting, as they destroy global dynamics

Minor result: 1-d isolated equilibria embeddadble in 2-d linear

1D example:

Consider the 1-d system with **three isolated equilibria**

$$\dot{x} = \frac{x(1-x^2)}{1+x^2}.$$

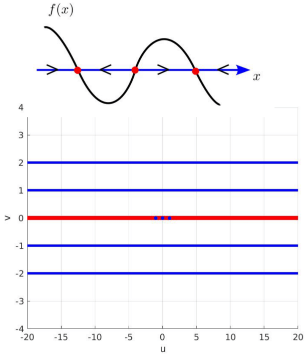
Equilibrium points are $\{-1, 0, 1\}$.

Then,

$$\dot{x} = \frac{x(1-x^2)}{1+x^2} \xrightarrow[\psi(x)]{} \begin{cases} \dot{u} = v \\ \dot{v} = 0 \end{cases}$$

where

$$\psi(x) = \begin{cases} (x, 0) & f(x) = 0 \\ \left(\ln\left|\frac{3x}{-2+2x^2}\right|, 1\right) & x < -1 \\ \left(2\ln\left|\frac{3x}{-2+2x^2}\right|, 2\right) & -1 < x < 0 \\ \left(-\ln\left|\frac{3x}{-2+2x^2}\right|, -1\right) & 0 < x < 1 \\ \left(-2\ln\left|\frac{3x}{-2+2x^2}\right|, -2\right) & x > 1. \end{cases}$$



*** Observe that limit cycles per se are not an obstruction!

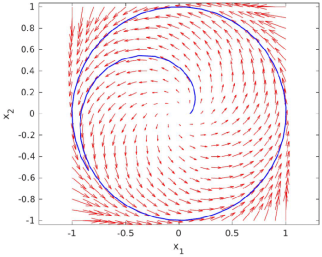
even though a linear system cannot have isolated limit cycles,

it is nonetheless possible to immerse a nonlinear system with an isolated limit cycle

into a linear system via a continuous and one to one mapping:

Consider the 2-d system with one **isolated limit cycle**

$$\begin{cases} \dot{x}_1 = x_1 - x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 = x_1 + x_2 - x_2(x_1^2 + x_2^2) \end{cases}$$



The ω-limit sets are

the origin $\{0\}$ and the unit circle $\{x \mid \|x\|_2 = 1\}$.

Let $\mathcal{X} = \mathbb{R}^2/\{0\}$. Then,

$$\begin{cases} \dot{x}_1 = x_1 - x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 = x_1 + x_2 - x_2(x_1^2 + x_2^2) \end{cases} \Longrightarrow \begin{cases} \dot{u} = -v \\ \dot{v} = u \\ \dot{w} = -2w \end{cases}$$

$$\psi(x) = \left(\frac{x_1}{\|x\|_2}, \frac{x_2}{\|x\|_2}, \|x\|_2^{-2} - 1\right)$$

continuous and **one-to-one**

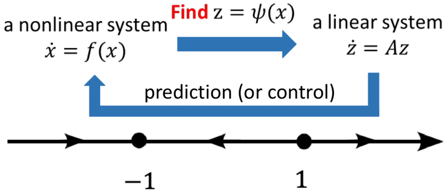
*** Example of local

Consider the 1-d system

$$\dot{x} = x^2 - 1.$$

The ω-limit sets are $x = 1$ and $x = -1$.

Let $\mathcal{X} = (-\infty, 1)$. Then,



a nonlinear system $\dot{x} = f(x)$

Find $z = \psi(x)$

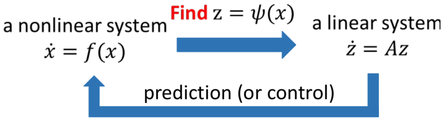
a linear system $\dot{z} = Az$

prediction (or control)

$$\dot{x} = x^2 - 1 \xrightarrow[\psi(x) = \frac{x+1}{x-1}]{} \dot{z} = -2z$$

If we extend $\mathcal{X}$ by a point to $\mathcal{X}' = (-\infty, 1]$, ψ is not an immersion anymore as $\psi(1)$ is undefined.

*** Similarly, for limit cycle in previous example



a nonlinear system $\dot{x} = f(x)$

Find $z = \psi(x)$

a linear system $\dot{z} = Az$

prediction (or control)

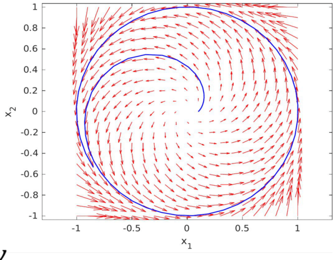
Consider the previous 2-d system

$$\begin{cases} \dot{x}_1 = x_1 - x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 = x_1 + x_2 - x_2(x_1^2 + x_2^2) \end{cases}$$

The ω-limit sets are:

- the origin $\{0\}$ and the unit circle $\{x \mid \|x\|_2 = 1\}$.

Let $\mathcal{X} = \mathbb{R}^2/\{0\}$. Then,

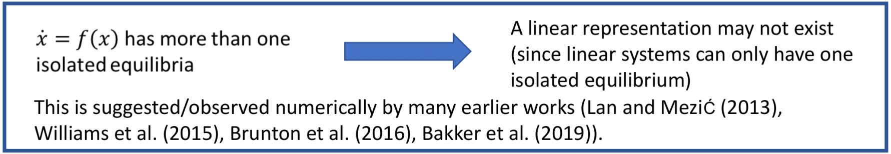


$$\begin{cases} \dot{x}_1 = x_1 - x_2 - x_1(x_1^2 + x_2^2) \\ \dot{x}_2 = x_1 + x_2 - x_2(x_1^2 + x_2^2) \end{cases} \longrightarrow \begin{cases} \dot{u} = -v \\ \dot{v} = u \\ \dot{w} = -2w \end{cases}$$

$$\psi(x) = \left(\frac{x_1}{\|x\|_2}, \frac{x_2}{\|x\|_2}, \|x\|_2^{-2} - 1 \right)$$

If we extend $\mathcal{X}$ by a point to $\mathbb{R}^2$,ψ is not an immersion anymore as $\psi(0)$ is undefined.

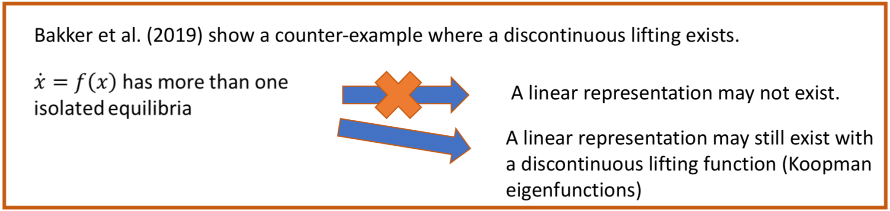
*** Some previous work



$\dot{x} = f(x)$ has more than one isolated equilibria

A linear representation may not exist (since linear systems can only have one isolated equilibrium)

This is suggested/observed numerically by many earlier works (Lan and MeziĆ (2013), Williams et al. (2015), Brunton et al. (2016), Bakker et al. (2019)).



Bakker et al. (2019) show a counter-example where a discontinuous lifting exists.

$\dot{x} = f(x)$ has more than one isolated equilibria

A linear representation may not exist.

A linear representation may still exist with a discontinuous lifting function (Koopman eigenfunctions)

Outline

- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

Iterations and flows in optimization

consider a continous time (flow) or a discrete time (iteration) system:

$$\begin{aligned} \dot{x}(t) &= f(x(t)) \\ x(t+1) &= f(x(t)) \end{aligned}$$

for which it is desired that $x(t) \to \xi$, where ξ solves an optimization problem

e.g. if $\mathcal{L}(x)$ is a (generally non-convex) loss function, one may look at gradient flow:

$$\dot{x}(t) = -\eta \nabla \mathcal{L}(x(t))^\top$$

of natural (Riemannian) flows, Newton or Quasi-Newton, etc ...

similarly can study (discrete) steepest descent version (= Euler of gradient flow):

$$x(t+1) = x(t) - \eta \nabla \mathcal{L}(x(t))^\top$$

Review of general convergence theory

assume $\mathcal{L}$ is continuously differentiable and has a minimum value 0

target and critical sets:

$$\begin{aligned} \mathcal{T}_{\mathcal{L}} &:= \{x \,|\, \quad \mathcal{L}(x) = 0\} \\ \mathcal{C}_{\mathcal{L}} &:= \{x \,|\, \nabla\mathcal{L}(x) = 0\} \supseteq \mathcal{T}_{\mathcal{L}} \end{aligned}$$

$$\dot{x}(t) = -\nabla\mathcal{L}(x(t))^\top \qquad \text{(for simplicity, take } \eta = 1\text{)}$$

- precompact trajectories approach $\mathcal{C}_{\mathcal{L}}$ (Krasovskii-LaSalle):

  $$\dot{\mathcal{L}} = -\|\nabla\mathcal{L}\|^2 \leq 0$$

  ($\implies$ convergence to $\mathcal{T}_{\mathcal{L}}$ if $\mathcal{C}_{\mathcal{L}} = \mathcal{T}_{\mathcal{L}}$)

- for analytic $\mathcal{L}$, all ω-limit sets ($\subseteq \mathcal{C}_{\mathcal{L}}$) are single equilibria (Łojasiewicz)

- generically, precompact trajectories converge to $\mathcal{C}_{\mathcal{L}} \setminus$ strict saddles

  (“strict” $=$ linearization has at least one positive eigenvalue)

*** A Theorem

Suppose that:

- $\mathcal{L}$ is a real-analytic (loss) function;
- $\mathcal{C}_{\mathcal{L}} = \mathcal{T}_{\mathcal{L}} \cup S$, where S consists of strict saddles for $\dot{x} = -\nabla \mathcal{L}(x)$; and
- every trajectory of the gradient flow dynamics is pre-compact.

Then, except for a set of measure zero, all trajectories converge to points in $\mathcal{T}_{\mathcal{L}}$.

Convergence to $\mathcal{T}_{\mathcal{L}}$ under gradient dominance conditions

if $\exists\,\lambda > 0$ s.t.: $\|\nabla\mathcal{L}(x)\|^2 \;\geq\; \lambda\,\mathcal{L}(x)$ (*global* Polyak-Łojasiewicz Inequality)

then: $\dot{\mathcal{L}} = -\|\nabla\mathcal{L}\|^2 \leq -\lambda\,\mathcal{L}(t) \implies \mathcal{L}(t) \;\leq\; e^{-\lambda t}\mathcal{L}(0)$

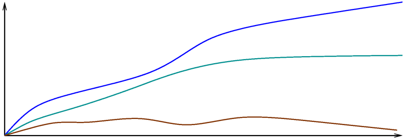
several weaker versions also guarantee (not necessarily exponential) convergence

(but more useful for subsequent "ISS" discussion):

$\|\nabla\mathcal{L}(x)\|^2 \;\geq\; \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{K}_\infty$

$\|\nabla\mathcal{L}(x)\|^2 \;\geq\; \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{K}$

$\|\nabla\mathcal{L}(x)\|^2 \;\geq\; \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{PD}$

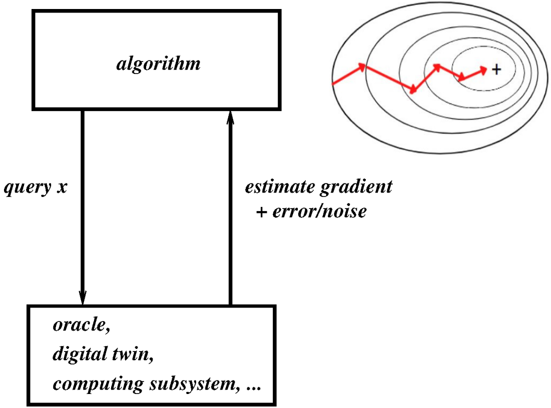


some notes:

- if $\|\nabla\mathcal{L}(x)\|^2 \geq \lambda_c\mathcal{L}(x)$, where λ_c's depend on sublevel set $\mathcal{L}(x) \leq c$, then $\exists\;\alpha \in \mathcal{PD}$
- $\mathcal{K}$ condition is a global "Kurdyka–Łojasiewicz Inequality"
- strict convexity ⇒ PŁI, but convexity not needed

But: $\mathcal{L}$ and/or its gradient $\nabla\mathcal{L}$ might be imprecisely known

- adversarial attack
- errors in evaluation of $\mathcal{L}$ by “oracle”
- early stopping of a simulation
- inacurate and very approximate digital twin
- stochastic computations (“reproducibility”!)
- learning by sampling from limited data



model by “input” or “disturbance” $u(t)$:

$$\dot{x}(t) \;=\; -\eta\,\nabla\mathcal{L}(x(t))^\top + u(t)$$

or more generally

$$\dot{x}(t) \;=\; f(x(t), u(t))$$

natural questions: graceful degradation if $\|u\|$ “small” (sup norm, integral, ...)?

- is $\operatorname{dist}(x(t), \mathcal{T}_{\mathcal{L}})$ (or $\mathcal{L}(x(t))$) ≈ 0 for large t? (asymptotic question); and rate?
- is $\operatorname{dist}(x(t), \mathcal{T}_{\mathcal{L}})$ (or $\mathcal{L}(x(t))$) not “too large” for intermediate computational times t?

ISS-like perturbation theory of gradient descent

$$\dot{x}(t) \; = \; -\eta \nabla \mathcal{L}(x(t))^T \; + \; B(x(t))\, u(t)$$

(“learning rate” $\eta > 0$; $B : \mathbb{X} \to \mathbb{R}^{n \times m}$ bounded locally Lipschitz)

[integral] input to state stability (ISS [iISS]): how do inputs affect dynamics?

if input $u(\cdot)$ is bounded

(small, “eventually” small, convergent)

then solutions inherit properties

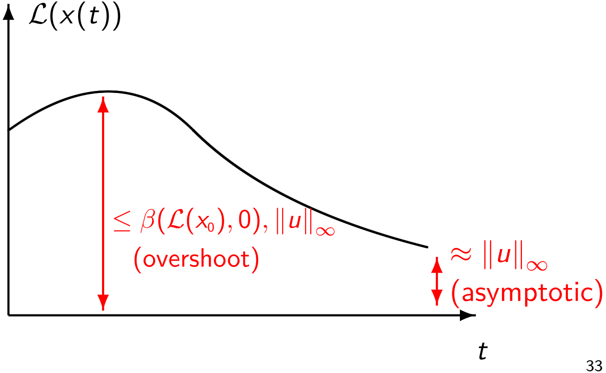
& well-controlled transient behavior:

$\exists\ \beta \in \mathcal{KL}$, $\gamma \in \mathcal{K}_\infty$ s.t. $\forall\ x_0, u$:

$$\mathcal{L}(x(t)) \;\leq\; \max\left\{\beta(\mathcal{L}(x_0), t)\,,\, \gamma\left(\|u\|_\infty\right)\right\}$$

$$\left[\int_0^t \gamma(|u(s)|) ds \text{ for integral ISS}\right]$$

think of $\beta(r,t) = \alpha_1(e^{-\lambda t}\alpha_2(r))$



Key relationships (+ technical details!)

- $\|\nabla \mathcal{L}(x)\|^2 \geq \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{K}_\infty$ $\iff$ ISS
- $\|\nabla \mathcal{L}(x)\|^2 \geq \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{K}$ $\iff$ small-input ISS
- $\|\nabla \mathcal{L}(x)\|^2 \geq \alpha(\mathcal{L}(x))$, for some $\alpha \in \mathcal{PD}$ $\iff$ iISS

*** Recall: [i]ISS is natural generalization of linear stability

for linear $\dot{x} = Ax + Bu$ (Hurwitz A), typical estimates of stability for operators

$$\left\{L^2, L^\infty\right\} \to \left\{L^2, L^\infty\right\} \; : \; (x_0, u) \mapsto x(\cdot)$$

are:

$$\begin{array}{rcll} |x(t,x_0,u)| & \leq & c_1\,|x_0|\,e^{-\lambda t} + c_2 \sup\limits_{s\in[0,t]} |u(s)| & (L^\infty \to L^\infty) \\ |x(t,x_0,u)| & \leq & c_1\,|x_0|\,e^{-\lambda t} + c_2 \displaystyle\int_0^t |u(s)|^2\,ds & (L^2 \to L^\infty) \\ \displaystyle\int_0^t |x(s,x_0,u)|^2\,ds & \leq & c_1\,|x_0| + c_2 \displaystyle\int_0^t |u(s)|^2\,ds & (L^2 \to L^2) \end{array}$$

for linear systems, all equivalent (with different constants)

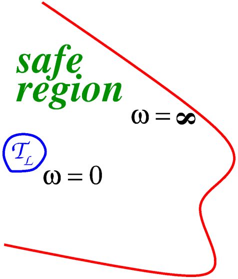
changing (nonlinear) coordinates $x(t) = T(z(t)) \rightsquigarrow$ ISS, iISS, and (again) ISS

*** [i] ISS with respect to $\mathcal{T}_{\mathcal{L}} \subset \mathbb{X} \subseteq \mathbb{R}^n$ (open set)

definition: $\omega : \mathbb{X} \to \mathbb{R}$ is a *size function for* $(\mathbb{X}, \mathcal{T}_{\mathcal{L}})$ if:

continuous, positive definite wrt $\mathcal{T}_{\mathcal{L}}$, and proper [coercive]

(meaning $\omega(x) \to \infty$ as $x \to \partial\mathbb{X}$ or $|x| \to \infty$)



assume $\mathcal{L}$ has strict minimum $\overline{\mathcal{L}}$ on $\mathcal{T}_{\mathcal{L}}$, and

$$\|\nabla\mathcal{L}(x)\|^2 \;\geq\; \alpha(\mathcal{L}(x) - \overline{\mathcal{L}}) \quad \forall\, x \in \mathbb{X}$$

with $\alpha \in \mathcal{K}_\infty$, $\mathcal{K}$, $\mathcal{PD}$

conclude ISS, "small-input" ISS, iISS properties for

$$\dot{x}(t) = -\eta\,\nabla\mathcal{L}(x(t))^T + B(x(t))u(t)$$

$$\omega(x(t, x_0, u)) \;\leq\; \max\{\beta((\omega(x_0), t)\,,\ \gamma\,(\|u\|_\infty)\} \quad \left[\int_0^t \gamma(|u(s)|)ds \text{ for integral ISS}\right]$$

(EDS, SCL 2022)

*** Tool: [i]ISS-Lyapunov (dissipation) inequalities

$\mathcal{C}^1$ function $\mathcal{L} : \mathbb{X} \to \mathbb{R}$ is *[i]ISS-Lyapunov function* for $\dot{x} = f(x, u)$ wrt $(\mathbb{X}, \mathcal{A})$ if

- $(\exists \overline{\mathcal{L}})$ $\mathcal{L} - \overline{\mathcal{L}}$ is a size function for $(\mathbb{X}, \mathcal{A})$
- $\exists\, \alpha, \gamma \in \mathcal{K}_\infty$ s.t. $\dot{\mathcal{L}}(x, u) \ \leq\ -\alpha(\mathcal{L}(x) - \overline{\mathcal{L}}) + \gamma(|u|) \quad \forall\, (x, u) \in \mathbb{X} \times \mathbb{R}^m$

  [for iISS, ask only α positive definite]

(where $\dot{\mathcal{L}}(x, u) := \nabla\mathcal{L}(x) \cdot f(x, u)$, i.e. $d\mathcal{L}(x(t))/dt = \dot{\mathcal{L}}(x(t), u(t))$)

Theorem. $\exists$ [i]ISS Lf $\Longleftrightarrow$ system is [i]ISS

Theorem. Similarly for $\alpha \in \mathcal{K}$ and “small-input ISS”

*** Steepest descent w/ line search

steepest descent algorithm: given guess x^k,

perform line minimization search in negative gradient direction:

$$\lambda^k := \arg\min_{\lambda \geq 0} V(x^k - \lambda \nabla \mathcal{L}(x^k)^T)$$

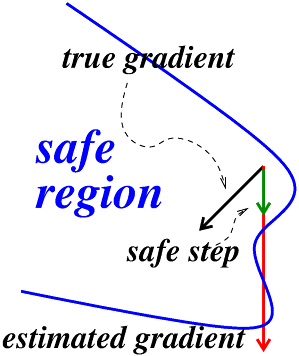
and define $x^{k+1} := x^k - \lambda^k \nabla \mathcal{L}(x^k)^T$

but noisy on gradient estimation, so really:

$$x^{k+1} = x^k - \lambda^k \left[\nabla \mathcal{L}(x^k)^T + B(x^k) d^k \right]$$

Theorem: if $\mathcal{L}$ is $\mathcal{K}_\infty$ loss function,

then iteration is (DT) ISS



*** Other works on ISS-like gradient flows

- Cherukuri-Mallada-Low-Cortés 2018, saddle dynamics
  (ISS gradient wrt additive errors, V has a convexity property, $\mathbb{X} = \mathbb{R}^n$)

- Poveda-Krstić 2019/21, fixed-time convergence in extremum seeking
  (gradient flow, "D-ISS" property wrt a time-varying uncertainty, $\mathbb{X} = \mathbb{R}^n$)

- Bianchin-Poveda-Dall'Anese, 2020 switched LTI systems
  (ISS gradient flow wrt unknown disturbances acting on plant, $\mathbb{X} = \mathbb{R}^n$)

- Suttner-Dashkovskiy 2021, extremum seeking
  (ISS gradient flow for kinematic unicycle, $\mathbb{X}$ closed submanifold of $\mathbb{R}^n$)

- Cunis-Kolmanovsky 2022, bilevel optimization
  (ISS gradient flow; errors arise from "inner loop" incomplete optimization)

- Pang-Tian-Jiang 2021/2022, LQ
  (Kleinman's policy iteration, "small-input" ISS)

Outline

- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

A motivating application

consider the most classical linear control problem: LQR for $\dot{x} = Ax + Bu$,

$$\mathcal{J}(x_0, u(\cdot)) \;:=\; \int_0^\infty x^T(t)Qx(t) \;+\; u^T(t)Ru(t)\,dt$$

$\implies\; u(t) = -K_{\text{opt}}x(t),\; K_{\text{opt}} = R^{-1}B^T\Pi$, where $\Pi > 0$ solves ARE

reformulate as: $\min\; V(K) := \mathbb{E}[\mathcal{J}(x_0, Kx(\cdot))]$ (average over initial states)

over open set $\mathcal{K}(A,B) := \{K \mid A - BK \text{ Hurwitz}\} \subset \mathbb{R}^{m\times n}$

but: solving problem requires precise knowledge of system/cost matrices

alternative (e.g. in RL): “direct” or “model free” approach, where controllers designed by numerically estimating cost (loss) function, using plant or digital twin

$\rightsquigarrow$ *“direct policy update”* approach: optimize gain K

renewed interest in an old approach! (Levine-Athans 1970)

Main result (Cui, Jiang, EDS 2024)

LQR for $\dot{x} = Ax + Bu$, $\mathcal{J}(x_0, u(\cdot)) := \int_0^\infty x^T(t)Qx(t) + u^T(t)Ru(t)\,dt$

$\Rightarrow u(t) = -K_{\text{opt}}x(t)$, $K_{\text{opt}} = R^{-1}B^T\Pi$, where $\Pi > 0$ solves ARE $\rightsquigarrow$ min $V(K)$

Theorem.

V is a $\mathcal{K}$ loss function.

Corollary.

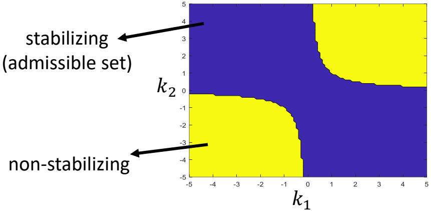
Gradient system is small-input ISS wrt gradient noise/errors.

- greatly generalizes known results (previously only $\mathcal{PD}$, so only iISS)
- key: $\|\nabla V(K)\|_F^2 \geq \alpha(V(K) - V(K_{\text{opt}}))$, some $\alpha \in \mathcal{K}$ $\left(\text{can pick } \alpha(r) = \dfrac{ar}{b + cr}\right)$
- **also:** (1) Newton flow, (2) natural gradient (over appropriate Riemannian metric)

Why is the problem challenging?

- for LQR, problem generally **non-convex**

  e.g.: $A = 0_{2\times 2}$, $B = -I_{2\times 2}$, $Q = I_{2\times 2}$, $R = I_{2\times 2}$, $K = \begin{bmatrix} -1 & k_1 \\ k_2 & -1 \end{bmatrix}$



stabilizing (admissible set)

non-stabilizing

- perturbed gradient flow is **nonlinear** dynamical system evolving in **matrix space**

$$\frac{\mathrm{d}K(s)}{\mathrm{d}s} = -2\eta(RK(s) - B^T P(s))Y(s) + \Delta(s)$$

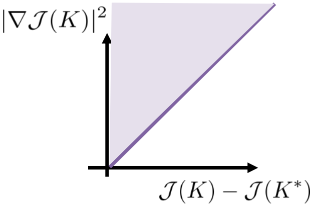
$$(A - BK(s))^T P(s) + P(s)(A - BK(s)) + Q + K(s)^T RK(s) = 0$$

$$(A - BK(s))Y(s) + Y(s)(A - BK(s))^T + I_n = 0$$

Ensure robustness by Polyak-Łojasiewicz (PL) condition?

$\boxed{|\nabla \mathcal{J}(K)|^2 \geq \alpha(\mathcal{J}(K) - \mathcal{J}(K^*))}$ where α is a **constant** [Polyak, Łojasiewicz, 1963]

"gradient flow is fast far from the optimal function value"



PL Condition

$$\mathcal{J}(K(s)) - \mathcal{J}(K^*) \leq \underbrace{e^{-\lambda s}(\mathcal{J}(K(0)) - \mathcal{J}(K^*))}_{\text{exponentially stable}} + \underbrace{\gamma(\|\Delta\|_\infty)}_{\text{robust}}$$

Exponential input-to-state stability

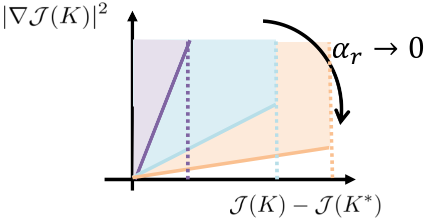
$$\mathcal{V}(K) = \mathcal{J}(K) - \mathcal{J}(K^*)$$

$$\frac{\mathrm{d}\mathcal{V}(K(s))}{\mathrm{d}s} = -\eta|\nabla \mathcal{J}(K(s))|^2 + \nabla \mathcal{J}(K(s))^T \Delta(s) \leq -\frac{\eta\alpha}{2}\mathcal{V}(K(s)) + \frac{1}{2\eta}|\Delta(s)|^2$$

No: for LQR, PL condition only holds over compact sets!

$\boxed{|\nabla \mathcal{J}(K)|^2 \geq \alpha_r(\mathcal{J}(K) - \mathcal{J}(K^*))}$ where α_r tends to zero as r tends to infinity

[Mesbahi UW; Polyak RAS; Jovanovic USC]



PL condition over compact sets

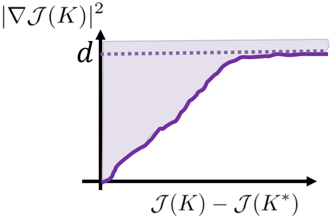
Robustness may vanish.

linear form of PL condition is too strong

can we have a **nonlinear** PL condition?

New: CJS-PL (“comparison just saturated”) condition

$\boxed{|\nabla \mathcal{J}(K)|^2 \geq \alpha(\mathcal{J}(K) - \mathcal{J}(K^*))}$ where α is a $\mathcal{K}$-function.



CJS-PL condition

If $\|\Delta\|_\infty < \frac{\eta\sqrt{d}}{\sqrt{2}}$ small disturbance

$$\mathcal{J}(K(s)) - \mathcal{J}(K^*) \leq \underbrace{\beta(\mathcal{J}(K(0)) - \mathcal{J}(K^*), s)}_{\text{asymptotically stable}} + \underbrace{\gamma(\|\Delta\|_\infty)}_{\text{robust}}$$

Small-disturbance input-to-state stability

$$\mathcal{V}(K) = \mathcal{J}(K) - \mathcal{J}(K^*)$$

$$\frac{\mathrm{d}\mathcal{V}(K(s))}{\mathrm{d}s} = -\eta|\nabla \mathcal{J}(K(s))|^2 + \nabla \mathcal{J}(K(s))^T \Delta(s) \leq -\frac{\eta}{2}\alpha\Big(\mathcal{V}(K(s))\Big) + \frac{1}{2\eta}|\Delta(s)|^2$$

LQR satisfies the CJS-PL condition

- for LQR cost: $|\nabla \mathcal{J}_{LQR}(K)|_F^2 \geq \alpha(\mathcal{J}_{LQR}(K) - \mathcal{J}_{LQR}(K^*))$, where $\alpha(r) = \left(\frac{a\,r}{r+b}\right)^2$
- perturbed gradient flow for the LQR problem is small-disturbance ISS
- perturbed natural gradient flow is small-disturbance ISS:

$$\frac{\mathrm{d}K(s)}{\mathrm{d}s} = -2\eta(RK(s) - B^T P(s)) + \Delta(s)$$

  $2(RK(s) - B^T P(s))$ is the gradient over the Riemannian manifold $(\mathcal{G}, \langle \cdot, \cdot \rangle_{Y_K})$.

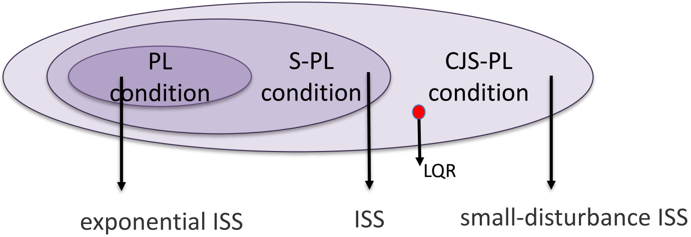
- perturbed Newton's gradient flow is small-disturbance ISS:

$$\frac{\mathrm{d}K(s)}{\mathrm{d}s} = -\eta(K(s) - R^{-1}B^T P(s)) + \Delta(s)$$

  $K(s) - R^{-1}B^T P(s)$ is the Newton's gradient direction.

Comparing gradient dominance conditions

$$\frac{\mathrm{d}K(s)}{\mathrm{d}s} = -\eta\nabla\mathcal{J}(K(s)) + \Delta(s), \quad \boxed{|\nabla\mathcal{J}(K)|^2 \geq \alpha(\mathcal{J}(K) - \mathcal{J}(K^*))}.$$



PL condition

S-PL condition

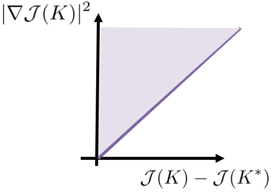
CJS-PL condition

LQR

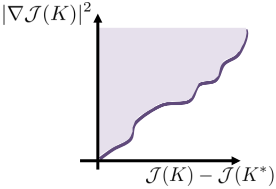
exponential ISS

ISS

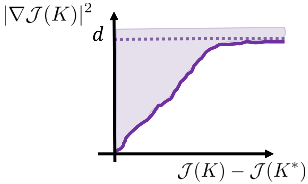
small-disturbance ISS



PL Condition



S-PL condition

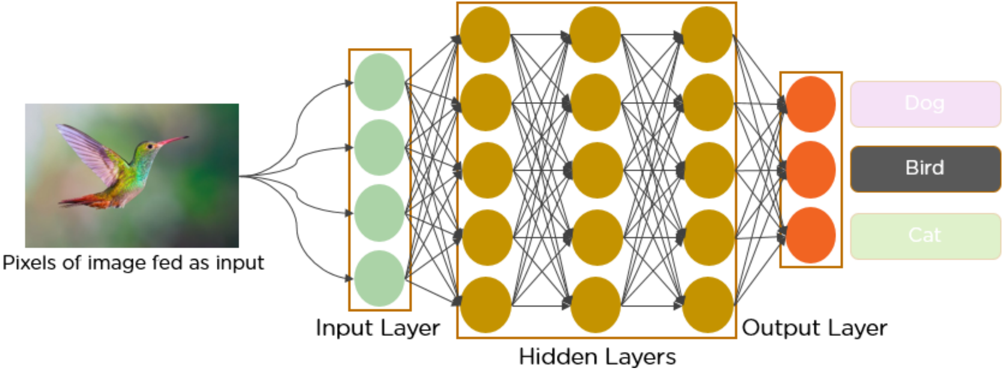


CJS-PL condition

Outline

- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

Motivation: “neural network” learning



typically “trained” from samples using variants of gradient descent on “loss”

(picture from web)

*** e.g. 50+ years ago, in this AI book

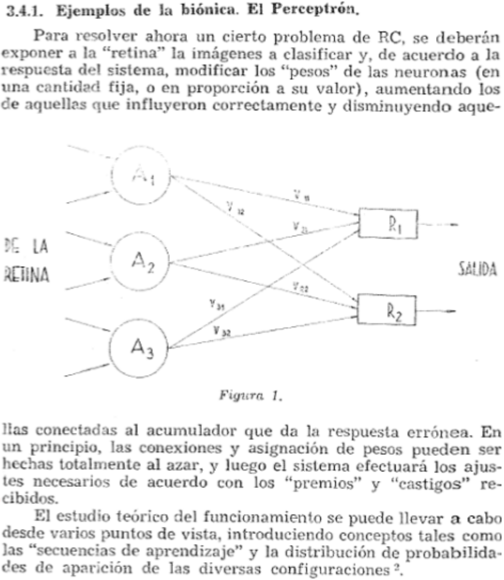
TEMAS DE INTELIGENCIA ARTIFICIAL

Eduardo Daniel Sontag

BUENOS AIRES / ARGENTINA / 1972

neural nets are a very old field, of course

pattern recognition
(reconocimiento de configuraciones)
input features,
neurons,
weights adjustment by stochastic gradient descent,
reinforcement learning ("rewards" and "penalties"),
theory based on probability distributions on inputs



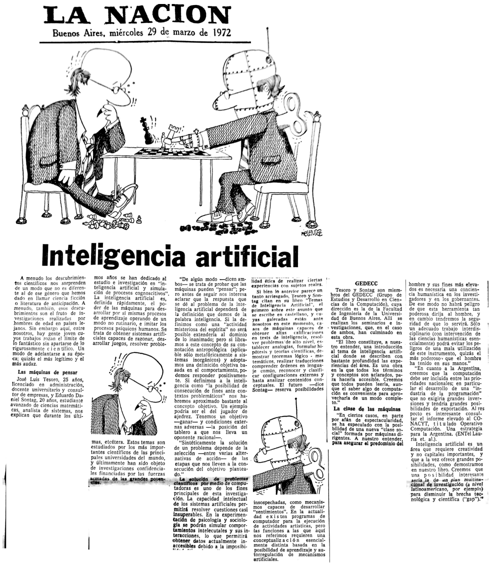
3.4.1. Ejemplos de la biónica. El Perceptrón.

Para resolver ahora un cierto problema de RC, se deberán exponer a la "retina" la imágenes a clasificar y, de acuerdo a la respuesta del sistema, modificar los "pesos" de las neuronas (en una cantidad fija, o en proporción a su valor), aumentando los de aquellas que influyeron correctamente y disminuyendo aquellas conectadas al acumulador que da la respuesta errónea. En un principio, las conexiones y asignación de pesos pueden ser hechas totalmente al azar, y luego el sistema efectuará los ajustes necesarios de acuerdo con los "premios" y "castigos" recibidos.

Figura 1.

El estudio teórico del funcionamiento se puede llevar a cabo desde varios puntos de vista, introduciendo conceptos tales como las "secuencias de aprendizaje" y la distribución de probabilidades de aparición de las diversas configuraciones [2].

*** Old area, always same issues . . . but now hype is not hype



LA NACION

Buenos Aires, miércoles 29 de marzo de 1972

Inteligencia artificial

A menudo los descubrimientos científicos nos sorprenden de un modo que no es diferente al de ese género que hemos dado en llamar ciencia ficción o literatura de anticipación. A menudo, también, esos descubrimientos son el fruto de investigaciones realizadas por hombres de edad en países lejanos. Sin embargo aquí, entre nosotros, hay gente joven cuyos trabajos rozan el límite de lo fantástico sin apartarse de lo rigurosamente científico. Un modo de adelantarse a su época; quizás el más legítimo y el más audaz.

Las máquinas de pensar

José Luis Tesoro, 25 años, licenciado en administración, docente universitario y consultor de empresas, y Eduardo Daniel Sontag, 20 años, estudiante avanzado de ciencias matemáticas, analista de sistemas, nos explican que durante los últimos años se han dedicado al estudio e investigación en "inteligencia artificial y simulación de procesos cognoscitivos". La inteligencia artificial es, definida rápidamente, el poder de las máquinas para desarrollar por sí mismas procesos de aprendizaje operando de un modo no rutinario, e imitar los procesos psíquicos humanos. Se trata de obtener sistemas artificiales capaces de razonar, desarrollar juegos, resolver problemas, etcétera. Estos temas son estudiados por los más importantes científicos de las principales universidades del mundo, y últimamente han sido objeto de investigaciones confidenciales financiadas por las fuerzas armadas de las grandes potencias.

"De algún modo —dicen ambos— se trata de probar que las máquinas pueden "pensar"; pero antes de seguir nos interesa aclarar que la respuesta que se dé al problema de la inteligencia artificial dependerá de la definición que demos de la palabra inteligencia. Si la definimos como una "actividad misteriosa del espíritu" no será posible extenderla al dominio de lo inanimado; pero si libramos a este concepto de su connotación antropológica (aplicable sólo metafóricamente a sistemas inorgánicos) y adoptamos una definición objetiva basada en el comportamiento, podemos responder positivamente. Si definimos a la inteligencia como "la posibilidad de consecución de fines ante contextos problemáticos" nos habremos aproximado bastante al concepto objetivo. Un ejemplo podría ser el del jugador de ajedrez. Tenemos un objetivo —ganar— y condiciones externas adversas —la posición del tablero a que nos lleva un oponente racional—.

"Sintéticamente la solución de un problema depende de la selección —entre varias alternativas de acción— de las etapas que nos lleven a la consecución del objetivo planteado."

La solución de problemas científicos por medio de computadoras es uno de los fines principales de esta investigación. La capacidad intelectual de los sistemas artificiales permitirá resolver cuestiones casi insuperables. En la experimentación de psicología y sociología se podrán simular comportamientos intelecutales y sus interacciones, lo que permitirá obtener datos actualmente inaccesibles debido a la imposibilidad ética de realizar ciertas experiencias con sujetos reales.

Si bien lo anterior parece un tanto arriesgado, Tesoro y Sontag citan en su libro "Temas de Inteligencia Artificial", el primero sobre este asunto que se escribe en castellano, y cuyas galeradas están ante nosotros en este momento, casos de máquinas capaces de obtener altas calificaciones en tests de inteligencia, resolver problemas de alto nivel, establecer analogías, formular hipótesis y teorías científicas, demostrar teoremas lógico - matemáticos, realizar traducciones comprender órdenes en lenguaje común, reconocer y clasificar configuraciones externas y hasta analizar contenidos conceptuales. El futuro —dice Sontag— reserva posibilidades insospechadas, como mecanismos capaces de desarrollar "sentimientos". En la actualidad existen programas de computador para la ejecución de actividades artísticas, pero las funciones a las que aquí nos referimos requieren una conceptualización esencialmente distinta basada en la posibilidad de aprendizaje y autorregulación de mecanismos artificiales.

GEDECC

Tesoro y Sontag son miembros del GEDECC (Grupo de Estudios y Desarrollo en Ciencias de la Computación), cuya dirección es la de la Facultad de Ingeniería de la Universidad de Buenos Aires. Allí se realizan los seminarios e investigaciones, que, en el caso de ambos, han culminado en esta obra.

"El libro constituye, a nuestro entender, una introducción al tema de inteligencia artificial donde se describen con bastante profundidad las experiencias del área. Es una obra en la que todos los términos y conceptos son aclarados, para hacerla accesible. Creemos que todos pueden leerla, aunque el saber algo de computación es conveniente para aprovecharla de un modo completo."

La clase de las máquinas

"En ciertos casos, en parte por afán de espectacularidad, se ha especulado con la posibilidad de una nueva "clase social" formada por máquinas dirigentes. A nuestro entender, para asegurar el predominio del hombre y sus fines más elevados es necesaria una conciencia humanística en los investigadores y en los gobernantes. De ese modo no habrá peligro de que esta herramienta tan poderosa dirija al hombre, y en cambio tendremos la seguridad de que lo servirá. Sólo un adecuado trabajo interdisciplinario (con intervención de las ciencias humanísticas esenccialmente) podrá evitar los peligros de una mala utilización de este instrumento, quizás el más poderoso que el hombre ha tenido en sus manos."

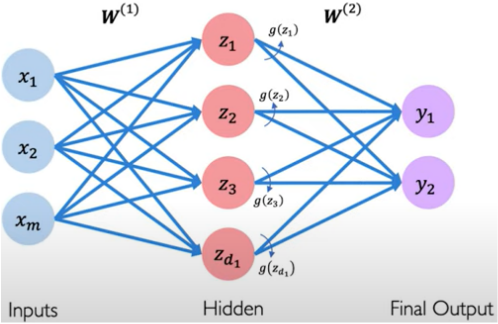
"En cuanto a la Argentina, creemos que la computación debe ser incluida entre las prioridades nacionales; en particular el desarrollo de una "industria de la programación" que no exigiría grandes inversiones y tendría grandes posibilidades de exportación. Al respecto es interesante consultar el informe elevado al CONACYT, titulado Operativo Computación. Una estrategia para la Argentina. (ENTel Lauria et. al.).

Inteligencia artificial es un área que requiere creatividad y no capitales importantes, y que a la vez ofrece grandes posibilidades, como demostramos en nuestro libro. Creemos que una posibilidad interesante sería la de *un plan multinacional de investigación* (a nivel latinoamericano, por ejemplo) para disminuir la brecha tecnológica y científica ("gap")."

our AI research group on B. Aires, 1969-72:

- sentience possible?
- AI will solve most major problems of science
- need "humanists" to help set limits
- "the most powerful technology ever"

Single-hiden layer case for simplicity here



(picture from web)

study special case where "activation function" is the identity: $Y = W_2 W_1 X$

given "data" pairs (X_i, Y_i): minimize $\|Y - W_2 W_1 X\|$ $\quad [X = (X_1, \ldots, X_s),\ Y = (Y_1, \ldots, Y_s)]$

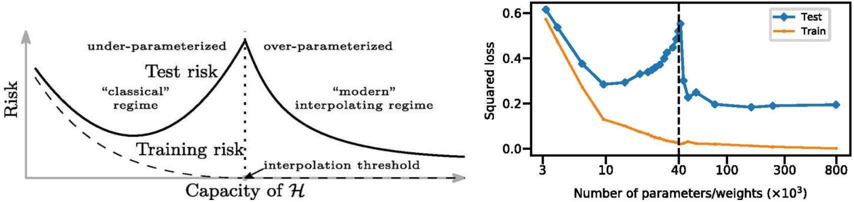
trivial linear regression: first argmin $\|Y - WX\|$; then factorize $W = W_2 W_1$

highly non-unique: $W = (W_2 T)(T^{-1} W_1),\ T \in GL(k)$

Why do people study this “uninteresting” problem?

- as a way to understand & conceptualize convergence of gradient descent
- understanding why “overparametrization” seems to “work”

(if $W_2 W_1 \in \mathbb{R}^{n\times m}$ and middle dim $= k \gg n, m$, have $(n+m)k \gg nm$ parameters)



(figures from Belkin/Hsu/Ma/Mandal, PNAS 2019)

- . . . and even faster convergence in certain cases
- we will study effect of disturbances (in ISS formalism):

think errors in ∇ computation, adversarial attacks, stochastic learning, . . .

Gradient flow associated to problem

for simplicity here, *matrix factorization problem*: $X = I$

loss to be minimized: $\mathcal{L}(P, Q) := \frac{1}{2}\|Y - PQ^\intercal\|_F^2$

"noisy" gradient flow on $(P; Q)$: $\begin{bmatrix}\dot{P} \\ \dot{Q}\end{bmatrix} = \begin{bmatrix}(Y - PQ^\intercal)Q \ + U \\ (Y - PQ^\intercal)^\intercal P + V\end{bmatrix}$

if no disturbances (errors) U, V: a.e. convergence to "target set" $\mathcal{T}_\mathcal{L}$ where $\mathcal{L} = 0$

follows from

- precompactness of trajectories
  (conservation law $P^T P - Q^T Q \equiv$ constant – symmetries and Noether's Theorem)
- analyticity
- critical points not in $\mathcal{T}_\mathcal{L}$ are strict saddles

[Remark: Riemannian gradient flow on fixed-rank matrices w/suitable metric]

[Baldi&Hornik1989, Monzón&Potrie'06, Kawaguchi'16, Panageas&Piliouras'16, Du...'18, Schaeffer&McCalla'20, Eftekhari'20, Bah...'21, Chitour...'23, ...]

The loss function as a candidate ISS-Lyapunov function

Theorem.

$$\dot{\mathcal{L}}(P,Q) \;\leq\; -\mathcal{L}(P,Q)\cdot\Big(\sigma_{\min}^2(Q)+\sigma_{\min}^2(P)\Big) \;+\; \frac{1}{2}\left\|\begin{bmatrix} U \\ V \end{bmatrix}\right\|_F^2$$

unfortunately, the factor $\sigma_{\min}^2(Q)+\sigma_{\min}^2(P)$ is not bounded away from zero,

so we do not have an ISS-Lyapunov function!

must restrict domain. . .

plan: first consider phase space for system with no disturbances, $U = V = 0$

then find invariant sets s.t. flow sufficiently transversal at boundary

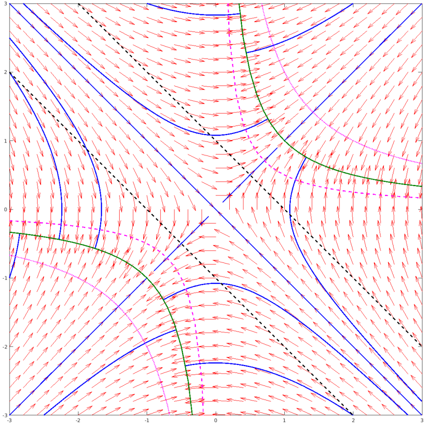
so perturbations allow staying inside and decreasing $\mathcal{L}$

and so that on this set the factor is bounded away

“Vector” case: $m=n=1$ (but latent layer size k arbitrary)

since here $\begin{bmatrix}\dot{P}\\ \dot{Q}\end{bmatrix} = (Y - PQ^\mathsf{T})\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}P\\ Q\end{bmatrix}$ **and $Y - PQ^\mathsf{T}$ is a scalar**,

this is just a scalar multiple (time-reparametrization) of a linear saddle; pic for $k = 1$:



green: target set $\mathcal{T} = \{(p, q) \,|\, pq = 1\}$ (if $Y{=}1$)

(components in 1st/3rd quadrants)

dashed black lines: sets $p + q = \alpha$

magenta: “transversal” sets $\{pq = \alpha\}$

blue: solution trajectories converging to target set

$(q^2 - p^2 \equiv \text{constant})$

Convergence region w/o inputs (de Oliveira, Siami, EDS, 2023)

Theorem.

- linearized stable manifold $\mathcal{S}^-$ at saddle $[P;Q]=0$ is global $\mathcal{W}_{\text{stable}}(0)$
- all other solutions converge to target set $\mathcal{T}_{\mathcal{L}}$

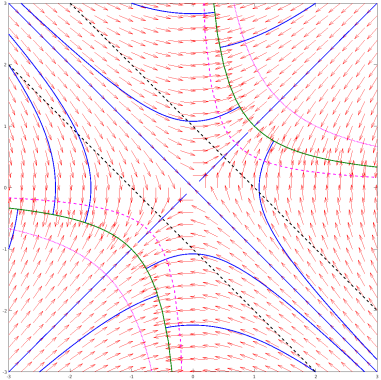
distance from $[P;Q]$ to S^-,
equals norm of projection of $[P;Q]$ into $S^+ := (S^-)^\perp$,
computed as $\|P+Q\|_2$

for any $\alpha > 0$, define:

$$\mathcal{R}_\alpha = \{[P;Q] \in \mathbb{R}^{2k} \mid \|P+Q\|_2^2 \geq \alpha^2\}$$

(delimited by black dashed lines in k=1 plane)

this gives "enough room" for perturbations:



Now with inputs (de Oliveira, Siami, EDS, 2023)

Theorem.
For any $\alpha \in [0, 2\sqrt{Y})$, $\mathcal{R}_\alpha$ is forward-invariant under gradient flow dynamics if

$$\|U\|_2 + \|V\|_2 \leq \frac{1}{\sqrt{2}}\,|\alpha| \left(Y - \frac{\alpha^2}{4} \right)$$

Moreover, if $(P, Q) \in \mathcal{R}_\alpha$, then $PP^\intercal + QQ^\intercal = \sigma(P)^2 + \sigma(Q)^2 \geq \alpha^2/2$.

Corollary.
For solutions in $\mathcal{R}_\alpha$ and with (U, V) constrained as above,

$$\dot{\mathcal{L}}(L, P) \;\leq\; -\mathcal{L}(P, Q) \cdot \frac{\alpha^2}{2} + \frac{1}{2} \left\| \begin{bmatrix} U \\ V \end{bmatrix} \right\|_2^2$$

gives an ISS estimate in that region.

*** A remark on speed of convergence (w/o inputs)

(still for simplicity $n{=}m{=}k{=}1$ and $Y{=}1$)

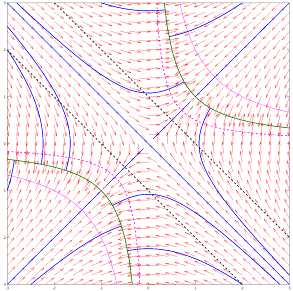
target set $\mathcal{T} = \{(p,q) \mid pq = 1\}$

eigenvalues of linearization at $\mathcal{T}$?

one eigen $= 0$ (tangent to $\mathcal{T}$)

and the other one is $-(p^2+q^2)$,

which is $\gg 1$ as $p \to \infty$ or $q \to \infty$



very fast (at least local) convergence vs gradient descent for non-overparametrized

$$\mathcal{L}(p) := (1/2)(1-p)^2\,, \text{i.e. } \dot{p} = 1 - p$$

which has eigenvalue -1 at stable equilibrium $p = 1$

*** Remarks about general case $\max\{m, n\} > 1$ (no inputs)

- $[P; Q]$ equilibrium **iff** $\exists$ SVDs $Y - PQ^\top = \Psi\Sigma\Phi^\top$, $P = \Psi\Sigma_P\Gamma_P^\top$, $Q = \Phi\Sigma_Q\Gamma_Q^\top$ s.t. $\Sigma\Sigma_Q = 0$ and $\Sigma^\top\Sigma_P = 0$ (note then $(Y - PQ^\top)Q = 0$ and $(Y - PQ^\top)^\top P = 0$)

- at $[P; Q] = 0$: same $\#$ of $+$ and $-$ (and no zero, if Y full rank square) eigenvalues

- target set $\mathcal{T}$ has dimension $(n + m)k - nm$

- at $[P; Q] \in \mathcal{T}$: mn strictly negative eigs, $-$'s of squares of SV's of P and Q

*** A sufficient condition for convergence

Theorem.

For undisturbed dynamics, solutions with $P_0 Q_0^\top - \bar{Y} \geq 0$ (in PSD sense) must converge to target set.

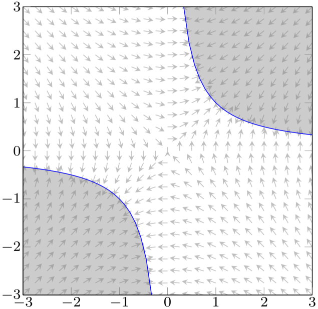


illustration for scalar case: initializing in the gray area

(condition conservative)

*** A necessary condition for convergence

Theorem. (Assuming $m = n$ for simplicity)
For undisturbed dynamics, solutions that converge to target set must have $\operatorname{rank}(\Psi^\top P_0 + \Phi^\top Q_0) = n$.

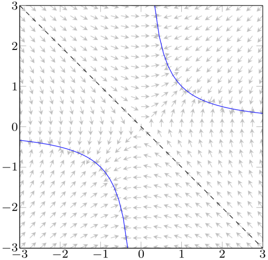


illustration for scalar case: initializing anywhere except $P + Q = 0$

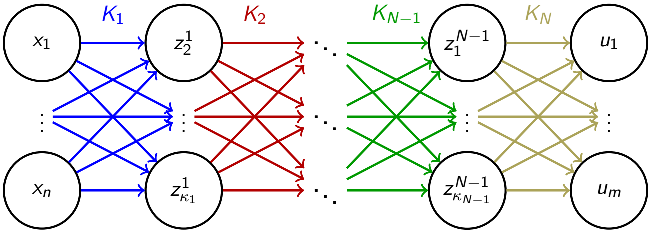
necessary and sufficient in the vector case

$\frac{1}{2}(\Psi^\top P + \Phi^\top Q)$ is projection onto unstable manifold of linearization at origin

Outline

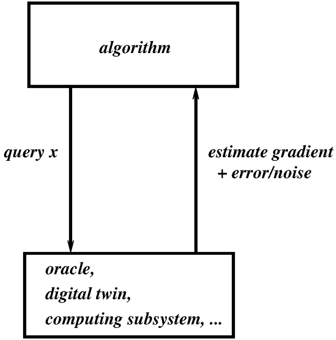
- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

Linear feedforward neural networks for state feedback



$$u = K_N K_{N-1} \dots K_2 K_1 x = \mathbf{K}(x)$$

Problem setup (de Oliveira, Siami, EDS, 2024)



optimization problem:

$$\min_{\mathbf{K}\in\mathcal{K}} \mathcal{J}(\mathbf{K}) = \mathbb{E}_{x_0\sim\mathcal{X}_0}\left[\int_0^\infty x(t)^\top Q x(t) + u(t)^\top R u(t)\, dt\right]$$

with constraint

$$\dot{x} = Ax + Bu$$

$$u = \mathbf{K}(x) = K_N \ldots K_2 K_1 x$$

parameter training via gradient flow:

$$\dot{K}_i = -\nabla_{K_i}\mathcal{J}(\mathbf{K})$$

Theoretical results (de Oliveira, Siami, EDS, 2024)

Theorem (informal statement):

gradient flow for overparametrized LQR always converges (to a finite solution)

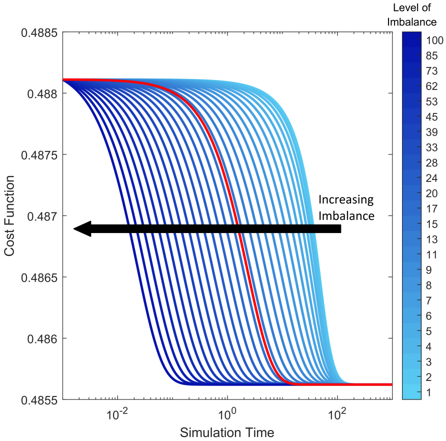
Theorem (informal statement):

when $N = 2$ ("single hidden layer")

gradient flow for overparametrized LQR converges to optimal feedback

for all but a set of measure zero of initializations

Simulations for overparametrized LQR gradient flow

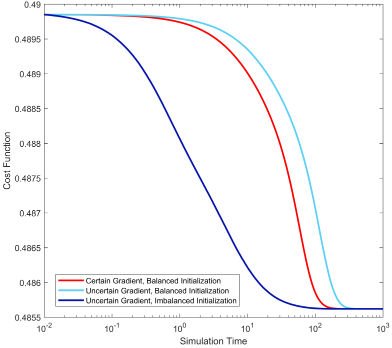


speed of convergence of overparametrized (blue)

slower or quicker than non-overparametrized (red)

depending on parameter initialization

Simulations when using uncertain gradients



red curve: balanced initialization without uncertainty

light blue: balanced initialization with uncertainty

dark blue: imbalanced initialization with uncertainty

accelerated convergence in overparametrized case

Outline

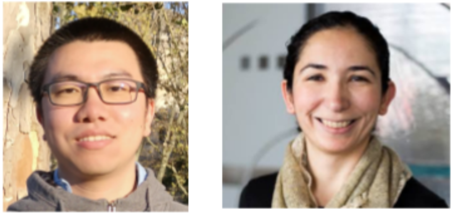
- Autoencoders (w/ Kvalheim)
- Some limitations of linearization-based control (w/ Liu & Ozay)
- (Disturbed) gradient flows
- ISS for LQR direct problem (w/ Cui & Jiang)
- Gradient dynamics for (linear) neural networks (w/ de Oliveira & Siami)
- Putting it all together: NN/overparametrized LQR (w/ de Oliveira & Siami)
- Collaborators

Collaborators

Arthur Castello de Oliveira / Milad Siami, NU

Matthew D. Kvalheim, UMBC

Leilei Cui / Zhong-Ping Jiang, NYU



Zexiang Liu / Necmiye Ozay, Michigan

http://www.sontaglab.org/publications.html