Entropy numbers of nonlinear systems

Master's thesis presentation

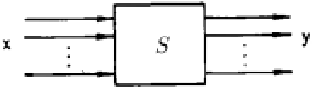
Guillaume Wang

ETHZ MINS

March 15, 2021

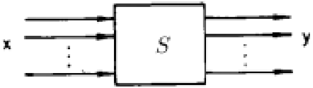
Motivation

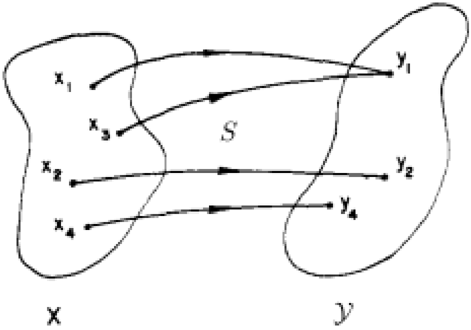
"What is a nonlinear system?"



Motivation

"What is a nonlinear system?"

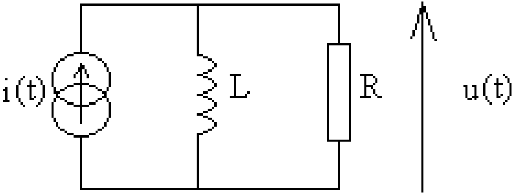




Goal: learn S from observations (x_i, y_i)

Motivation

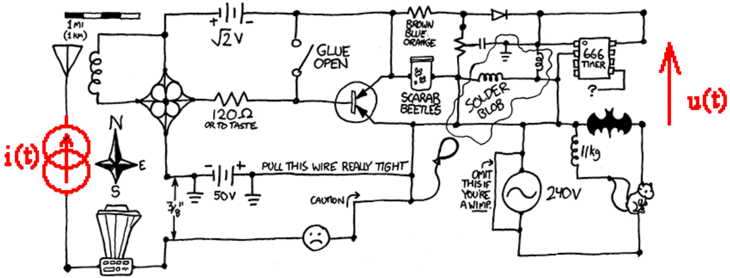
System identification



Input: $i(t)$, output: $u(t)$. $u(t) = S[\ i(t)\ ]$?

Motivation

System identification

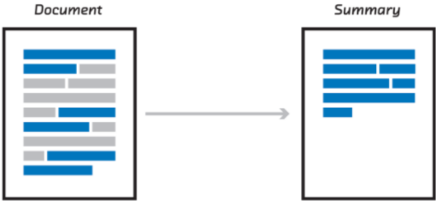


Input: $i(t)$, output: $u(t)$. $u(t) = S[\ i(t)\ ]$?

Motivation

Signal-to-signal tasks in ML

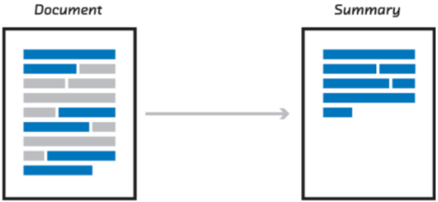
- Text-to-text



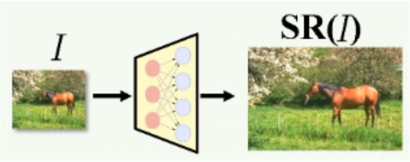
Motivation

Signal-to-signal tasks in ML

- Text-to-text



- Super-resolution imaging



Motivation

- Classical (regression/classification):

$$\text{learn} \qquad f : \mathsf{R}^d \to \mathsf{R} \text{ or } \{0, 1\}$$

- Nonlinear system identification:

$$\text{learn (e.g)} \qquad S : L^1(\mathsf{R}) \to L^1(\mathsf{R})$$

Motivation

- Classical (regression/classification):

$$\text{learn} \qquad f : \mathrm{R}^d \to \mathrm{R} \text{ or } \{0,1\}$$

- Nonlinear system identification:

$$\text{learn (e.g)} \qquad S : L^1(\mathrm{R}) \to L^1(\mathrm{R})$$

«How difficult is it to learn a mapping?»

1. Framework for this thesis
2. "Parametrize": LTI systems case
3. "Parametrize": Volterra series
4. Generalize classical techniques

1. Framework for this thesis

2. "Parametrize": LTI systems case

3. "Parametrize": Volterra series

4. Generalize classical techniques

Metric entropy

« How difficult is it to learn a set of objects $\mathcal{C}$? »

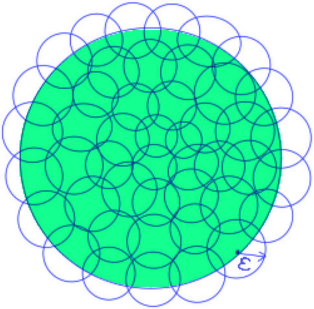
- Learning theory tool: ε-covering number

Metric entropy

« How difficult is it to learn a set of objects $\mathcal{C}$? »

- Learning theory tool: ε-covering number

$$N_\varepsilon(\mathcal{C}; \|\cdot\|) = \min\left\{ n; \quad \exists (p_1, \ldots, p_n) \subset \mathcal{C} \quad \text{s.t} \quad \mathcal{C} \subset \bigcup_i B^{k\,k}_{p_i,\varepsilon} \right\}$$



An ε-covering

Metric entropy

- ε-covering number $N_\varepsilon(\mathcal{C}; \|\cdot\|)$
- Metric entropy $\log_2 N_\varepsilon(\mathcal{C}; \|\cdot\|)$

Metric entropy

- ε-covering number $N_\varepsilon(\mathcal{C}; \|\cdot\|)$
- Metric entropy $\log_2 N_\varepsilon(\mathcal{C}; \|\cdot\|)$

Why metric entropy? Intuition:

Proposition

For any "bitstring length" $\ell \in$ N, consider encoder/decoder scheme

$$E : \mathcal{C} \to \{0,1\}^\ell \qquad D : \{0,1\}^\ell \to \mathcal{C}$$

$\log_2 N_\varepsilon(\mathcal{C}; \|\cdot\|)$ is the minimum ℓ s.t

$$\inf_{E,D} \sup_{c2C} \|c - D(E(c))\| \leq \varepsilon$$

("best-obtainable worst-case error")

Metric entropy

- ε-covering number $N_\varepsilon(\mathcal{C}; \|\cdot\|)$
- Metric entropy $\log_2 N_\varepsilon(\mathcal{C}; \|\cdot\|)$

Why metric entropy? Intuition:

Proposition

For any "bitstring length" $\ell \in \mathrm{N}$, consider encoder/decoder scheme

$$E : \mathcal{C} \to \{0,1\}^\ell \qquad D : \{0,1\}^\ell \to \mathcal{C}$$

$\log_2 N_\varepsilon(\mathcal{C}; \|\cdot\|)$ is the minimum ℓ s.t

$$\inf_{E,D} \sup_{c2C} \|c - D(E(c))\| \;\leq\; \varepsilon$$

("best-obtainable worst-case error")

$\rightarrow$ quantifies "massiveness"

$\rightarrow$ fundamental bound on compressibility / learnability

Entropy numbers

$N_\varepsilon(\mathcal{C}; \|\cdot\|)$ ε-covering number $\mathsf{R}_+ \to \mathsf{N}$

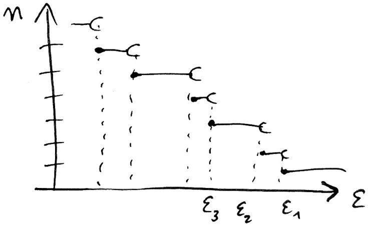
Entropy numbers

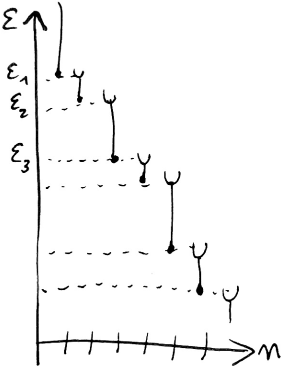
$N_\varepsilon(\mathcal{C}; \|\cdot\|)$	ε-covering number	$\mathsf{R}_+ \to \mathsf{N}$
$\varepsilon_n(\mathcal{C}; \|\cdot\|)$	n-th entropy number	$\mathsf{N} \to \mathsf{R}_+$

Entropy numbers

$N_\varepsilon(\mathcal{C}; \|\cdot\|)$ ε-covering number $\mathbb{R}_+ \to \mathbb{N}$

$\varepsilon_n(\mathcal{C}; \|\cdot\|)$ n-th entropy number $\mathbb{N} \to \mathbb{R}_+$





"Metric entropy" "Entropy number"

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R}), C(\mathrm{R}), ...$)
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R}), C(\mathrm{R}), \ldots)$
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Variants

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R})$, $C(\mathrm{R})$, ...)
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Variants

1. Worst-case error

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_1 = \sup_x \left\| S[x] - \widehat{S}[x] \right\|_Y$$

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R}), C(\mathrm{R}), \ldots$)
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Variants

1. Worst-case error

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_{1} = \sup_{x} \left\| S[x] - \widehat{S}[x] \right\|_{Y}$$

2. Worst-case error over a subset $U \subset \mathcal{X}$

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_{1\ U} = \sup_{x 2 U} \left\| S[x] - \widehat{S}[x] \right\|_{Y}$$

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R}), C(\mathrm{R}), \ldots$)
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Variants

1. Worst-case error

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_{1} = \sup_{x} \left\| S[x] - \widehat{S}[x] \right\|_{Y}$$

2. Worst-case error over a subset $U \subset \mathcal{X}$

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_{1\,U} = \sup_{x2U} \left\| S[x] - \widehat{S}[x] \right\|_{Y}$$

3. Average error over a distribution

$$\mathsf{S} = \{S : (\mathcal{X}, \Sigma, \mathrm{P}) \to \mathcal{Y}\} \quad \left\| S - \widehat{S} \right\|_{L^1_{\mathrm{P}}} = \mathrm{E}_x \left\| S[x] - \widehat{S}[x] \right\|_{Y}$$

Framework

- $\mathcal{X}$ space of input signals $x(t)$ (e.g $L^2(\mathrm{R}), C(\mathrm{R}), ...$)
- $\mathcal{Y}$ space of output signals $y(t)$
- S space of systems S

Variants

1. Worst-case error

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \|S\|_{1} = \sup_{x} \|S[x]\|_{Y}$$

2. Worst-case error over a subset $U \subset \mathcal{X}$

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \|S\|_{1\ U} = \sup_{x2U} \|S[x]\|_{Y}$$

3. Average error over a distribution

$$\mathsf{S} = \{S : (\mathcal{X}, \Sigma, \mathrm{P}) \to \mathcal{Y}\} \qquad \|S\|_{L^1_{\mathrm{P}}} = \mathrm{E}_x \|S[x]\|_{Y}$$

Framework

Worst-case error over a subset $U \subset \mathcal{X}$:

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \|S\|_{1\;U} = \sup_{x2U} \|S[x]]\|_Y$$

Framework

Worst-case error over a subset $U \subset \mathcal{X}$:

$$\mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \|S\|_{1\,U} = \sup_{x 2 U} \|S[x]]\|_Y$$

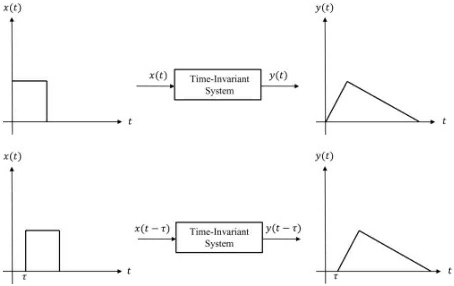
Goal:

$$\text{Given } \mathsf{S}_+ \subset \mathsf{S}, \quad \text{estimate } N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\,U})$$

LTI systems

LTI = Linear Time-Invariant

$$S(\lambda x_1 + x_2) = \lambda S x_1 + S x_2$$



Convolution representation

Theorem (Schwartz kernel theorem)

For all $S : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ *LTI, there exists* $k \in \mathcal{D}'(\mathbb{R})$ *s.t*

$$Sx(t) = \int_{\mathbb{R}} d\tau \; k(t-\tau)x(\tau) = (k * x)(t)$$

Convolution representation

Theorem (Schwartz kernel theorem)

For all $S : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ *LTI, there exists* $k \in \mathcal{D}'(\mathbb{R})$ *s.t*

$$Sx(t) = \int_{\mathbb{R}} d\tau \; k(t-\tau)x(\tau) = (k * x)(t)$$

Proof:

$$\begin{aligned} Sx(t) &= S \int_{\mathbb{R}} d\tau \; x(\tau) \; \delta_\tau(t) \\ &= \int_{\mathbb{R}} d\tau \; x(\tau) \; \underbrace{S\delta_\tau(t)} = \int_{\mathbb{R}} d\tau \; x(\tau) \; k(t,\tau) \end{aligned}$$

Time-invariance $\implies$ $k(t,\tau) = k(t-\tau)$

Convolution representation and norm

$$\mathsf{S} = \{S : L^2(\mathsf{R}) \to L^2(\mathsf{R}) \ \ \mathsf{LTI}\}$$
$$= \{L_k : x \mapsto (k * x), \quad k \in \mathcal{K}\} \simeq \mathcal{K}$$

$(K = D'(\mathsf{R}))$

Convolution representation and norm

$$\mathsf{S} = \left\{ S : L^2(\mathsf{R}) \to L^2(\mathsf{R}) \ \ \mathsf{LTI} \right\}$$
$$= \left\{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K} \right\} \simeq \mathcal{K}$$
$$\mathsf{S}_+ = \left\{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K}_+ \right\} \simeq \mathcal{K}_+$$

$(K = D'(\mathsf{R}))$

Convolution representation and norm

$$\mathsf{S} = \left\{ S : L^2(\mathsf{R}) \to L^2(\mathsf{R}) \;\; \mathsf{LTI} \right\}$$
$$= \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K} \} \simeq \mathcal{K}$$
$$\mathsf{S}_+ = \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K}_+ \} \simeq \mathcal{K}_+$$

Want $(K = D'(\mathsf{R}))$

Convolution representation and norm

$$\begin{aligned} \mathsf{S} &= \left\{ S : L^2(\mathsf{R}) \to L^2(\mathsf{R}) \ \ \mathsf{LTI} \right\} \\ &= \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K} \} \simeq \mathcal{K} \\ \mathsf{S}_+ &= \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K}_+ \} \simeq \mathcal{K}_+ \end{aligned}$$

Want $(K = D'(\mathsf{R}))$

$$\begin{aligned} & \|L_k\|_{1\ U} = \|k\|_K \\ \implies \quad & N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\ U}) = N_\varepsilon(\mathcal{K}_+; \|\cdot\|_K) \end{aligned}$$

Convolution representation and norm

$$\mathsf{S} = \left\{ S : L^2(\mathrm{R}) \to L^2(\mathrm{R}) \ \ \mathrm{LTI} \right\}$$
$$= \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K} \} \simeq \mathcal{K}$$
$$\mathsf{S}_+ = \{ L_k : x \mapsto (k * x), \quad k \in \mathcal{K}_+ \} \simeq \mathcal{K}_+$$

Want $(K = D'(\mathrm{R}))$

$$\|L_k\|_{1\,U} = \|k\|_K$$
$$\Longrightarrow \quad N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\,U}) = N_\varepsilon(\mathcal{K}_+; \|\cdot\|_K)$$

Theorem

Suppose $U = B^{(L^2(\mathrm{R}))}$.

$$\forall k \in L^1(\mathrm{R}), \ \ \|L_k\|_{1\,U} = |||L_k||| = \left\| \hat{k} \right\|_{L^\infty(\mathrm{R})}$$

Convolution representation and norm

$$
\begin{aligned}
\mathrm{S} &= \left\{S : L^2(\mathrm{R}) \to L^2(\mathrm{R}) \ \ \mathrm{LTI}\right\} \\
&= \{L_k : x \mapsto (k * x), \quad k \in \mathcal{K}\} \simeq \mathcal{K} \\
\mathrm{S}_+ &= \{L_k : x \mapsto (k * x), \quad k \in \mathcal{K}_+\} \simeq \mathcal{K}_+
\end{aligned}
$$

Want $(K = D'(\mathrm{R}))$

$$
\begin{aligned}
& \|L_k\|_{1\,U} = \|k\|_K \\
\implies \quad & N_\varepsilon(\mathrm{S}_+; \|\cdot\|_{1\,U}) = N_\varepsilon(\mathcal{K}_+; \|\cdot\|_K)
\end{aligned}
$$

Theorem

Suppose $U = B^{(L^2(\mathrm{R}))}$.

$$
\forall k \in L^1(\mathrm{R}), \ \ \|L_k\|_{1\,U} = |||L_k||| = \left\|\hat{k}\right\|_{L^\infty(\mathrm{R})}
$$

Conclusion: reduced to metric entropy in function space ,

Volterra series

- LTI system:

$k(t)$

$$L_k x = \int_{\mathbb{R}} d\tau \; k(\tau) \; x(t-\tau)$$

Volterra series

- LTI system: $k(t)$

$$L_k x = \int_{\mathrm{R}} d\tau \; k(\tau) \; x(t-\tau)$$

- Volterra series: sum of monomials $k = (k_0, k_1, ...)$ $k_n(t_1, ..., t_n)$

$$V_k[x] = \sum_{n=0}^{1} \int_{\mathrm{R}^n} d\boldsymbol{\tau} \; k_n(\boldsymbol{\tau}) \; x(t-\tau_1)...x(t-\tau_n)$$

Volterra series

- LTI system: $k(t)$

$$L_k x = \int_{\mathrm{R}} d\tau \; k(\tau) \; x(t-\tau)$$

- Volterra series: sum of monomials $k = (k_0, k_1, ...)$ $k_n(t_1, ..., t_n)$

$$V_k[x] = \sum_{n=0}^{1} \int_{\mathrm{R}^n} d\boldsymbol{\tau} \; k_n(\boldsymbol{\tau}) \; x(t-\tau_1)...x(t-\tau_n)$$

Rk: Taylor series $f : \mathrm{R}^d \to \mathrm{R}$,

$$f(x) = \sum_{n=0}^{1} \sum_{\mathrm{i}\, 2\, f1...dg^n} a_{\mathrm{i}} \; x_{i_1}...x_{i_n}$$

Volterra series

- LTI system:

$k(t)$

$$L_k x = \int_{\mathbb{R}} d\tau \; k(\tau) \; x(t-\tau)$$

- Volterra series: sum of monomials

$k = (k_0, k_1, \dots)$

$k_n(t_1, \dots, t_n)$

$$V_k[x] = \sum_{n=0}^{\infty} \int_{\mathbb{R}^n} d\boldsymbol{\tau} \; k_n(\boldsymbol{\tau}) \; x(t-\tau_1) \dots x(t-\tau_n)$$

Rk: Taylor series $f : \mathbb{R}^d \to \mathbb{R}$,

$$f(x) = \sum_{n=0}^{\infty} \sum_{\mathbf{i} \in \{1 \dots d\}^n} a_{\mathbf{i}} \; x_{i_1} \dots x_{i_n}$$

Volterra series

$$V_k[x] = \sum_{n=0}^{1} \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)...x(t-\tau_n)$$

Where do x, k, $V_k[x]$ live? i.e: $\mathcal{X}, \mathcal{K}, \mathcal{Y}$?

Volterra series

$$V_k[x] = \sum_{n=0}^{1} \int_{\mathrm{R}^n} d\boldsymbol{\tau}\; k_n(\boldsymbol{\tau})\; x(t-\tau_1)\ldots x(t-\tau_n)$$

Where do x, k, $V_k[x]$ live? i.e: $\mathcal{X}, \mathcal{K}, \mathcal{Y}$?

Theorem

$$\|V_k[x]\|_Y \leq \sum_{n=0}^{1} \|k_n\|_{K_n} \|x\|_X^n$$

- $\mathcal{X} = L^p(\mathrm{R}),\ \mathcal{K}_n = L^q(\mathrm{R}^n),\ \mathcal{Y} = L^1(\mathrm{R})$ $\quad (1/p + 1/q = 1)$

or

- $\mathcal{X} = C_b(\mathrm{R}),\ \mathcal{K}_n = ba(\mathrm{R}^n),\ \mathcal{Y} = C_b(\mathrm{R})$

Volterra series

$$V_k[x] = \sum_{n=0}^{1} \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)\ldots x(t-\tau_n)$$

Where do x, k, $V_k[x]$ live? i.e: $\mathcal{X}, \mathcal{K}, \mathcal{Y}$?

Theorem

$$\|V_k[x]\|_Y \le \sum_{n=0}^{1} \|k_n\|_{K_n} \|x\|_X^n$$

- $\mathcal{X} = L^p(\mathrm{R}),\ \ \mathcal{K}_n = L^q(\mathrm{R}^n),\ \ \mathcal{Y} = L^1(\mathrm{R})$ $\qquad (1/p + 1/q = 1)$

or

- $\mathcal{X} = C_b(\mathrm{R}),\ \ \mathcal{K}_n = ba(\mathrm{R}^n),\ \ \mathcal{Y} = C_b(\mathrm{R})$

(Convergence of $\sum_{n=0}^{1}$?)

Volterra series

For simplicity assume finite order N

$$V_k[x] = \sum_{n=0}^{N} \int_{\mathbb{R}^n} d\boldsymbol{\tau} \; k_n(\boldsymbol{\tau}) \; x(t-\tau_1) \ldots x(t-\tau_n)$$

Volterra series

For simplicity assume finite order N

$$V_k[x] = \sum_{n=0}^{N} \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)\ldots x(t-\tau_n)$$

$$\|V_k[x]\|_Y \leq \sum_{n=0}^{N} \|k_n\|_{K_n} \|x\|_X^n$$
$$\leq \mathrm{cst}(N, U) \quad \|k\|_K$$

$(K = K_0 \quad K_1 \quad \ldots)$

Volterra series

For simplicity assume finite order N

$$V_k[x] = \sum_{n=0}^{N} \int_{\mathbb{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)\ldots x(t-\tau_n)$$

$$\begin{aligned}\|V_k[x]\|_Y &\leq \sum_{n=0}^{N} \|k_n\|_{K_n} \|x\|_X^n \\ &\leq \mathrm{cst}(N,U) \quad \|k\|_K\end{aligned}$$

$$(K = K_0 \quad K_1 \quad \ldots)$$

$$N_{\mathrm{cst}(N,U)\ \varepsilon}(\mathsf{S}_+; \|\cdot\|_{\mathit{1}\ U}) \leq N_\varepsilon(\mathcal{K}_+; \|\cdot\|_K)$$

$$(\mathsf{S}_+ = f V_k;\ k\ 2\ K_+ g)$$

Volterra series

For simplicity assume finite order N

$$V_k[x] = \sum_{n=0}^{N} \int_{\mathrm{R}^n} d\boldsymbol{\tau} \; k_n(\boldsymbol{\tau}) \; x(t-\tau_1)\ldots x(t-\tau_n)$$

$$\begin{aligned} \|V_k[x]\|_Y &\leq \sum_{n=0}^{N} \|k_n\|_{K_n} \|x\|_X^n \\ &\leq \mathrm{cst}(N, U) \quad \|k\|_K \end{aligned}$$

$$(K = K_0 \quad K_1 \quad \ldots)$$

$$N_{\mathrm{cst}(N,U)\,\varepsilon}(\mathrm{S}_+; \|\cdot\|_{\mathit{1}\ U}) \leq N_\varepsilon(\mathcal{K}_+; \|\cdot\|_K)$$

$$(\mathrm{S}_+ = f V_k; \;\; k \; 2 \; K_+ g)$$

Conclusion: again reduced to metric entropy in function space ,

"Parametrize" path

$$S = \{\Phi_k, \quad k \in \mathcal{K}\}$$

Summary: if $\quad S_+ = \{\Phi_k, \quad k \in \mathcal{K}_+\}$ then ,

$$\|\Phi_k\|_{1\,U} \leq c\,\|k\|_K$$

"Parametrize" path

$$S = \{\Phi_k, \quad k \in \mathcal{K}\}$$

Summary: if $\quad S_+ = \{\Phi_k, \quad k \in \mathcal{K}_+\}$ then ,

$$\|\Phi_k\|_{1\,U} \leq c\,\|k\|_K$$

Problems:

"Parametrize" path

$$S = \{\Phi_k, \quad k \in \mathcal{K}\}$$

Summary: if $S_+ = \{\Phi_k, \quad k \in \mathcal{K}_+\}$ then ,

$$\|\Phi_k\|_{1\ U} \leq c\, \|k\|_K$$

Problems:

- Upper estimate may be too loose

"Parametrize" path

$$\mathsf{S} = \{\Phi_k, \quad k \in \mathcal{K}\}$$

Summary: if $\quad \mathsf{S}_+ = \{\Phi_k, \quad k \in \mathcal{K}_+\} \quad$ then ,

$$\|\Phi_k\|_{1\ U} \leq c\,\|k\|_K$$

Problems:

- Upper estimate may be too loose
- Given S_+ specified differently, how to find $\mathcal{K}_+$?

"Parametrize" path

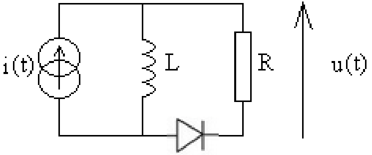
$$S = \{\Phi_k, \quad k \in \mathcal{K}\}$$

Summary: if $\quad S_+ = \{\Phi_k, \quad k \in \mathcal{K}_+\} \quad$ then ,

$$\|\Phi_k\|_{1\ U} \leq c\,\|k\|_K$$

Problems:

- Upper estimate may be too loose
- Given S_+ specified differently, how to find $\mathcal{K}_+$?



$S_+ = f[i(t)\ 7!\ \ u(t)]\,;\ \ R\ 2\ [R_{min}, R_{max}], L\ 2\ [L_{min}, L_{max}]g$ $\qquad K_+$??

Generalize classical techniques

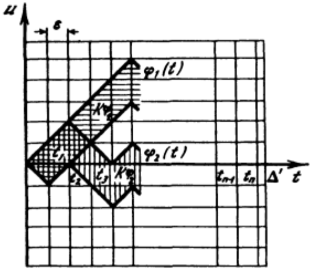
- Different approach: adapt techniques from the case $\mathcal{F}_+ \subset \mathcal{F} = \left\{ f : \mathsf{R}^d \to \mathsf{R} \right\}$

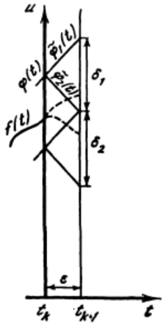
Generalize classical techniques

- Different approach: adapt techniques from the case $\mathcal{F}_+ \subset \mathcal{F} = \{f : \mathrm{R}^d \to \mathrm{R}\}$
- Illustrate the "sample and quantize" technique

Generalize classical techniques

- Different approach: adapt techniques from the case $\mathcal{F}_+ \subset \mathcal{F} = \{f : \mathrm{R}^d \to \mathrm{R}\}$
- Illustrate the "sample and quantize" technique





Proof of metric entropy estimate for the set of lipschitz-continuous functions

(Lipschitz)-continuous functions

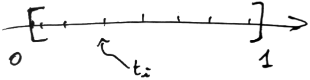
Lipschitz-continuous functions over a compact interval

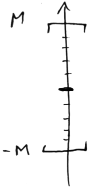
$$\mathcal{F}_+ \subset \left\{ f : [0,1] \to \mathrm{R}; \qquad \forall t, t^0, \quad \left| f(t) - f(t^0) \right| \leq L \left| t - t^0 \right| \right\}$$

(Lipschitz)-continuous functions

Lipschitz-continuous functions over a compact interval

$$\mathcal{F}_+ \subset \left\{ f : [0,1] \to \mathrm{R}; \qquad \forall t, t', \quad \left| f(t) - f(t') \right| \leq L \left| t - t' \right| \right\}$$

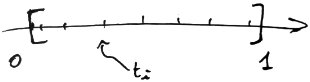


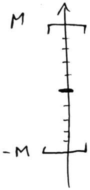


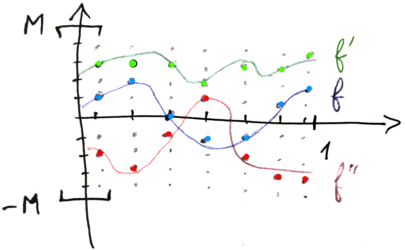
(Lipschitz)-continuous functions

Lipschitz-continuous functions over a compact interval

$$\mathcal{F}_+ \subset \{f : [0,1] \to \mathrm{R}; \qquad \forall t, t^0, \quad |f(t) - f(t^0)| \leq L\,|t - t^0|\}$$







(Lipschitz)-continuous systems

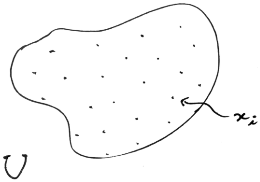
Lipschitz-continuous system over a compact metric space

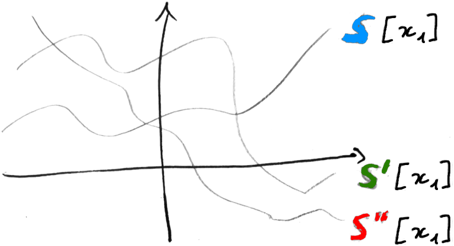
$$\mathsf{S}_+ \subset \left\{ S : (U, d) \to (\mathcal{Y}, \|\cdot\|_Y);\ \forall x, x', \left\| S[x] - S[x'] \right\|_Y \leq Ld(x, x') \right\}$$

(Lipschitz)-continuous systems

Lipschitz-continuous system over a compact metric space

$$\mathsf{S}_+ \subset \left\{ S : (U, d) \to (\mathcal{Y}, \|\cdot\|_Y);\ \forall x, x^0, \left\| S[x] - S[x^0] \right\|_Y \leq L d(x, x^0) \right\}$$



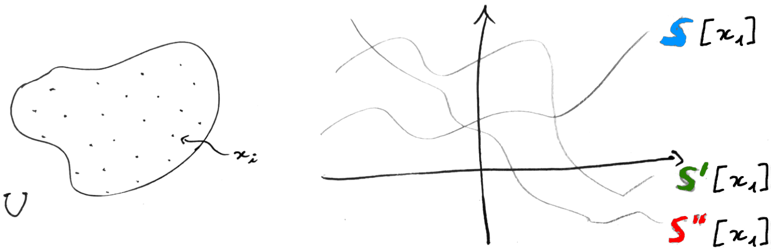


Quantize the output set $\{S[x];\ S \in \mathsf{S}_+, x \in U\}$?

(Lipschitz)-continuous systems

Lipschitz-continuous system over a compact metric space

$$\mathsf{S}_+ \subset \left\{ S : (U, d) \to (\mathcal{Y}, \|\cdot\|_Y);\ \forall x, x^0, \|S[x] - S[x^0]\|_Y \leq L d(x, x^0) \right\}$$



Theorem (Banach-valued Arzela-Ascoli)

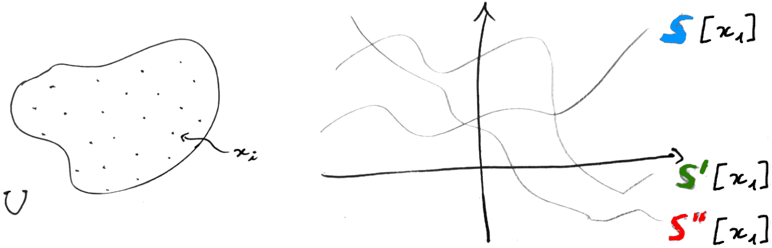
S_+ *relatively compact in* $C(U; \mathcal{Y})$ *i*

- S_+ *equicontinuous* ($\Leftarrow$ *L-lipschitz*)
- S_+ *"equicompact":* $\{S[x];\ S \in \mathsf{S}_+, x \in U\}$ *relatively compact*

(Lipschitz)-continuous systems

Lipschitz-continuous system over a compact metric space

$$\mathsf{S}_+ \subset \left\{ S : (U, d) \to (\mathcal{Y}, \|\cdot\|_Y);\ \forall x, x^0, \left\|S[x] - S[x^0]\right\|_Y \leq L d(x, x^0) \right\}$$



Theorem (Banach-valued Arzela-Ascoli)

S_+ *can be quantized i*

- S_+ *equicontinuous* ($\Leftarrow$ *L-lipschitz*)
- S_+ *"equicompact":* $\{S[x];\ S \in \mathsf{S}_+, x \in U\}$ *can be quantized*

Conclusion

Conclusion

- Framework: want $N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\ U})$ where

$$\mathsf{S}_+ \subset \mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\| S - \widehat{S} \right\|_{1\ U} = \sup_{x 2 U} \left\| S[x] - \widehat{S}[x] \right\|_Y$$

Conclusion

- Framework: want $N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\ U})$ where

$$\mathsf{S}_+ \subset \mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\|S - \widehat{S}\right\|_{1\ U} = \sup_{x 2 U} \left\|S[x] - \widehat{S}[x]\right\|_Y$$

- Two paths:
  - "Parametric"

$$\mathsf{S}_+ = \{\Phi_k, k \in \mathcal{K}_+\}$$
$$N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{\infty U}) \simeq N_\varepsilon(\mathcal{K}_+; \|\cdot\|_{\mathcal{K}})$$

  - Generalize classical techniques

$$f : \mathrm{R}^d \to \mathrm{R} \qquad\qquad S : \mathcal{X} \to \mathcal{Y}$$

Conclusion

- Framework: want $N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{1\ U})$ where

$$\mathsf{S}_+ \subset \mathsf{S} = \{S : \mathcal{X} \to \mathcal{Y}\} \qquad \left\|S - \widehat{S}\right\|_{1\ U} = \sup_{x2U} \left\|S[x] - \widehat{S}[x]\right\|_Y$$

- Two paths:
  - "Parametric"

$$\mathsf{S}_+ = \{\Phi_k, k \in \mathcal{K}_+\}$$
$$N_\varepsilon(\mathsf{S}_+; \|\cdot\|_{\infty U}) \simeq N_\varepsilon(\mathcal{K}_+; \|\cdot\|_{\mathcal{K}})$$

  - Generalize classical techniques

$$f : \mathsf{R}^d \to \mathsf{R} \qquad\qquad S : \mathcal{X} \to \mathcal{Y}$$

- Direction for future work: "non-parametric" path? Kernel methods for nonlinear system identification

Illustrations adapted from

- "Nonlinear system modeling based on the Wiener theory" (Schetzen 1981)
- https://commons.wikimedia.org/wiki/File:Circuit_L_R_parall%C3%A8le_-_courant_en_entr%C3%A9e_et_tension_en_sortie.png
- https://xkcd.com/730/
- https://towardsdatascience.com/8a3fbfdc5e9b
- "'Zero-Shot' Super-Resolution using Deep Internal Learning" (Shocher et al. 2017)
- *Neural Network Theory lecture notes* HS2019 ETHZ
- *Introduction to Digital Communications*, Chapter 3 (Grami 2016)
- "ε-Entropy and ε-Capacity of Sets In Functional Spaces" (Kolmogorov and Tikhomirov 1959)

Appendix

The idea

- Time-invariant system $\rightarrow$ scalar-valued functional (cf report)

The idea

- Time-invariant system → scalar-valued functional (cf report)
- Volterra monomial with fixed n

$$V_{k_n}[x](t) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)...x(t-\tau_n)$$

$$F_{\theta_n}[x] = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau}) x(\tau_1)...x(\tau_n) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau}) x\ \ ^n(\boldsymbol{\tau})$$

The idea

- Time-invariant system $\to$ scalar-valued functional (cf report)
- Volterra monomial with fixed n

$$V_{k_n}[x](t) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)...x(t-\tau_n)$$

$$F_{\theta_n}[x] = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau}) x(\tau_1)...x(\tau_n) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau}) x^{\ n}(\boldsymbol{\tau})$$

- Idea: view as linear combination of feature map

$$F_{\theta_n}[x] = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau}) \phi(x)(\boldsymbol{\tau}) = \langle \phi(x), \theta_n \rangle$$

$$x \in L^p(\mathrm{R}) \quad \mapsto \quad \phi(x) = x^{\ n} \in L^p_{\mathrm{Sym}}(\mathrm{R}^n)$$

The idea

- Time-invariant system $\rightarrow$ scalar-valued functional (cf report)
- Volterra monomial with fixed n

$$V_{k_n}[x](t) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ k_n(\boldsymbol{\tau})\ x(t-\tau_1)...x(t-\tau_n)$$

$$F_{\theta_n}[x] = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau})x(\tau_1)...x(\tau_n) = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau})x^{\ \ n}(\boldsymbol{\tau})$$

- Idea: view as linear combination of feature map

$$F_{\theta_n}[x] = \int_{\mathrm{R}^n} d\boldsymbol{\tau}\ \theta_n(\boldsymbol{\tau})\phi(x)(\boldsymbol{\tau}) = \langle \phi(x), \theta_n \rangle$$

$$x \in L^p(\mathrm{R}) \quad \mapsto \quad \phi(x) = x^{\ \ n} \in L^p_{\mathrm{Sym}}(\mathrm{R}^n)$$

- So $F_{\theta_n} \in$ RKHS with

$$K(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle = \left(\int_{\mathrm{R}} x\ \tilde{x} \right)^n$$

Problem: ill-defined when $x, \tilde{x} \in L^p(\mathrm{R})$...

RKBS

RKBS = Reproducing Kernel *Banach* Space

Definition (Lin et al. 2019)

Pair of RKBS: a tuple $(\mathcal{B}_1, \mathcal{B}_2, \langle \cdot, \cdot \rangle_{B_1 \quad B_2})$ s.t

- $\mathcal{B}_i$ Banach space of (real-valued) functions on Ω_i
- $\langle \cdot, \cdot \rangle_{B_1 \quad B_2}$ continuous bilinear form on $\mathcal{B}_1 \times \mathcal{B}_2$

RKBS

RKBS = Reproducing Kernel *Banach* Space

Definition (Lin et al. 2019)

Pair of RKBS: a tuple $(\mathcal{B}_1, \mathcal{B}_2, \langle \cdot, \cdot \rangle_{B_1 \;\; B_2})$ s.t

- $\mathcal{B}_i$ Banach space of (real-valued) functions on Ω_i
- $\langle \cdot, \cdot \rangle_{B_1 \;\; B_2}$ continuous bilinear form on $\mathcal{B}_1 \times \mathcal{B}_2$
- $\exists K : \Omega_1 \times \Omega_2 \to \mathrm{R}$, called *reproducing kernel*, s.t

$$K(x, \cdot) \in \mathcal{B}_2 \qquad \forall f \in \mathcal{B}_1,\; f(x) = \langle f, K(x, \cdot) \rangle_{B_1 \;\; B_2}$$
$$K(\cdot, y) \in \mathcal{B}_1 \qquad \forall g \in \mathcal{B}_2,\; g(y) = \langle K(\cdot, y), g \rangle_{B_1 \;\; B_2}$$

(K is unique)

RKBS from feature maps

Proposition (Lin et al. 2019)

- $\phi_i : \Omega_i \to \mathcal{W}_i$ Banach $(i = 1, 2)$

RKBS from feature maps

Proposition (Lin et al. 2019)

- $\phi_i : \Omega_i \to \mathcal{W}_i$ Banach $(i = 1, 2)$
- $\langle \cdot, \cdot \rangle_{W_1 \quad W_2}$ continuous bilinear form on $\mathcal{W}_1 \times \mathcal{W}_2$

RKBS from feature maps

Proposition (Lin et al. 2019)

- $\phi_i : \Omega_i \to \mathcal{W}_i$ Banach ($i = 1, 2$)
- $\langle \cdot, \cdot \rangle_{W_1 \ \ W_2}$ continuous bilinear form on $\mathcal{W}_1 \times \mathcal{W}_2$
- "span $\phi_i(\Omega_i)$ are dense":

$$\left\{ v \in \mathcal{W}_2 ;\ \forall x \in \Omega_1, \langle \phi_1(x), v \rangle_{W_1 \ \ W_2} = 0 \right\} = \{0\}$$
$$\left\{ u \in \mathcal{W}_1 ;\ \forall \tilde{x} \in \Omega_2, \langle u, \phi_2(\tilde{x}) \rangle_{W_1 \ \ W_2} = 0 \right\} = \{0\}$$

RKBS from feature maps

Proposition (Lin et al. 2019)

- $\phi_i : \Omega_i \to \mathcal{W}_i$ Banach $(i = 1, 2)$
- $\langle \cdot, \cdot \rangle_{W_1 \quad W_2}$ continuous bilinear form on $\mathcal{W}_1 \times \mathcal{W}_2$
- "span $\phi_i(\Omega_i)$ are dense":

$$\left\{ v \in \mathcal{W}_2;\ \forall x \in \Omega_1, \langle \phi_1(x), v \rangle_{W_1 \quad W_2} = 0 \right\} = \{0\}$$
$$\left\{ u \in \mathcal{W}_1;\ \forall \tilde{x} \in \Omega_2, \langle u, \phi_2(\tilde{x}) \rangle_{W_1 \quad W_2} = 0 \right\} = \{0\}$$

This induces a pair of RKBS

$$\mathcal{B}_1 := \left\{ F_v = \langle \phi_1(\cdot), v \rangle_{W_1 \quad W_2};\ v \in \mathcal{W}_2 \right\} \qquad \|F_v\|_{B_1} := \|v\|_{W_2}$$
$$\mathcal{B}_2 := \left\{ G_u = \langle u, \phi_2(\cdot) \rangle_{W_1 \quad W_2};\ u \in \mathcal{W}_1 \right\} \qquad \|G_u\|_{B_2} := \|u\|_{W_1}$$
$$\langle F_v, G_u \rangle_{B_1 \quad B_2} := \langle u, v \rangle_{W_1 \quad W_2}$$

and $K(x, \tilde{x}) = \langle \phi_1(x), \phi_2(\tilde{x}) \rangle_{W_1 \quad W_2}$.

Volterra series as polynomial RKBS

Proposition

- $\phi_1 : L^p(\mathrm{R}) \to L^p_{\mathrm{Sym}}(\mathrm{R}^n),\ x(t) \mapsto x\ \ {}^n(t)$
  $\phi_2 : L^q(\mathrm{R}) \to L^q_{\mathrm{Sym}}(\mathrm{R}^n),\ \tilde{x}(t) \mapsto \tilde{x}\ \ {}^n(t)$
- $\langle u_n(t), \tilde{u}_n(t)\rangle_{L^p} = \int_{\mathrm{R}^n} dt\ \ u_n(t)\tilde{u}_n(t)$ (Rk: $\equiv$ duality bracket)

Volterra series as polynomial RKBS

Proposition

- $\phi_1 : L^p(\mathbb{R}) \to L^p_{\mathrm{Sym}}(\mathbb{R}^n),\ x(t) \mapsto x^{\otimes n}(t)$
  $\phi_2 : L^q(\mathbb{R}) \to L^q_{\mathrm{Sym}}(\mathbb{R}^n),\ \tilde{x}(t) \mapsto \tilde{x}^{\otimes n}(t)$
- $\langle u_n(t), \tilde{u}_n(t)\rangle_{L^p} = \int_{\mathbb{R}^n} dt\ u_n(t)\tilde{u}_n(t)$ (Rk: $\equiv$ duality bracket)
- "span $\phi_i(\Omega_i)$ are dense": (cf report)

$$\left\{u_n \in L^q_{\mathrm{Sym}}(\mathbb{R}^n);\ \forall x \in L^p(\mathbb{R}), \left\langle x^{\otimes n}, u_n\right\rangle_{L^p} = 0\right\} = \{0\}$$

$$\left\{\tilde{u}_n \in L^p_{\mathrm{Sym}}(\mathbb{R}^n);\ \forall \tilde{x} \in L^q(\mathbb{R}), \left\langle \tilde{u}_n, \tilde{x}^{\otimes n}\right\rangle_{L^p} = 0\right\} = \{0\}$$

Volterra series as polynomial RKBS

Proposition

- $\phi_1 : L^p(\mathrm{R}) \to L^p_{\mathrm{Sym}}(\mathrm{R}^n),\; x(t) \mapsto x^{\otimes n}(t)$
  $\phi_2 : L^q(\mathrm{R}) \to L^q_{\mathrm{Sym}}(\mathrm{R}^n),\; \tilde{x}(t) \mapsto \tilde{x}^{\otimes n}(t)$
- $\langle u_n(t), \tilde{u}_n(t)\rangle_{L^p} = \int_{\mathrm{R}^n} dt\; u_n(t)\tilde{u}_n(t)$ (Rk: $\equiv$ duality bracket)
- "span $\phi_i(\Omega_i)$ are dense": (cf report)

$$\left\{u_n \in L^q_{\mathrm{Sym}}(\mathrm{R}^n);\; \forall x \in L^p(\mathrm{R}), \left\langle x^{\otimes n}, u_n\right\rangle_{L^p} = 0\right\} = \{0\}$$

$$\left\{\tilde{u}_n \in L^p_{\mathrm{Sym}}(\mathrm{R}^n);\; \forall \tilde{x} \in L^q(\mathrm{R}), \left\langle \tilde{u}_n, \tilde{x}^{\otimes n})\right\rangle_{L^p} = 0\right\} = \{0\}$$

This induces $(\mathcal{B}_1, \mathcal{B}_2, \langle \cdot, \cdot\rangle_{B_1 \times B_2})$ and

$$\mathcal{B}_1 = \left\{F_{\theta_n} : \left[x \mapsto \int_{\mathrm{R}^n} d\boldsymbol{\tau}\; \theta_n(\boldsymbol{\tau}) x^{\otimes n}(\boldsymbol{\tau})\right];\; \theta_n \in L^q_{\mathrm{Sym}}(\mathrm{R}^n)\right\}$$

$\|F_{\theta_n}\|_{B_1} = \|\theta_n\|_{L^q_{\mathrm{Sym}}(\mathrm{R}^n)}$ and $K(x, \tilde{x}) = \left(\int_{\mathrm{R}} x\,\tilde{x}\right)^n$.

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS
- Volterra series of finite order: idem (just concatenate)

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS
- Volterra series of finite order: idem (just concatenate)
- Volterra series of infinite order? $\rightarrow$ yes, cf report

$$K(x, \tilde{x}) = \sum_{n=0}^{1} \left(\int_{\mathrm{R}} x \ \tilde{x} \right)^{n}$$

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS
- Volterra series of finite order: idem (just concatenate)
- Volterra series of infinite order? $\rightarrow$ yes, cf report

$$K(x, \tilde{x}) = \sum_{n=0}^{1} \left(\int_{\mathrm{R}} x \ \tilde{x} \right)^n$$

Possible future directions:

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS
- Volterra series of finite order: idem (just concatenate)
- Volterra series of infinite order? $\rightarrow$ yes, cf report

$$K(x, \tilde{x}) = \sum_{n=0}^{1} \left(\int_{\mathrm{R}} x \; \tilde{x} \right)^{n}$$

Possible future directions:

- Learning in RKBS vs. classical Volterra-series-based system identification?

Volterra series as polynomial RKBS

- Volterra monomials are exactly elements of a RKBS
- Volterra series of finite order: idem (just concatenate)
- Volterra series of infinite order? $\rightarrow$ yes, cf report

$$K(x, \tilde{x}) = \sum_{n=0}^{1} \left(\int_{\mathrm{R}} x \; \tilde{x} \right)^{n}$$

Possible future directions:

- Learning in RKBS vs. classical Volterra-series-based system identification?
- What if we want $K(x, \tilde{x}) = \sum_{n=0}^{1} a_n \left(\int_{\mathrm{R}} x \; \tilde{x} \right)^{n}$ for some $a_n \geq 0$?

Entropy numbers

$$N_\varepsilon(\mathcal{C}; \|\cdot\|) = \min\left\{n; \quad \exists(p_1, ..., p_n) \subset \mathcal{C} \quad \text{s.t} \quad \mathcal{C} \subset \bigcup_i B^{k\,k}_{p_i,\varepsilon}\right\}$$

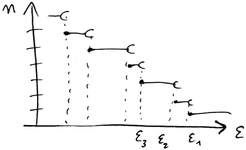
Entropy numbers

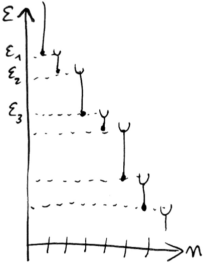
$$N_\varepsilon(\mathcal{C}; \|\cdot\|) = \min \left\{ n; \quad \exists (p_1, \dots, p_n) \subset \mathcal{C} \quad \text{s.t} \quad \mathcal{C} \subset \bigcup_i B^{k\,k}_{p_i,\varepsilon} \right\}$$

$$\varepsilon_n(\mathcal{C}; \|\cdot\|) = \inf \left\{ \varepsilon; \quad \exists (p_1, \dots, p_n) \subset \mathcal{C} \quad \text{s.t} \quad \mathcal{C} \subset \bigcup_i B^{k\,k}_{p_i,\varepsilon} \right\}$$

Entropy numbers

$$N_\varepsilon(\mathcal{C};\|\cdot\|) = \min\left\{n;\quad \exists(p_1,\dots,p_n)\subset\mathcal{C}\quad \text{s.t}\quad \mathcal{C}\subset\bigcup_i B^{k\,k}_{p_i,\varepsilon}\right\}$$
$$\varepsilon_n(\mathcal{C};\|\cdot\|) = \inf\left\{\varepsilon;\quad \exists(p_1,\dots,p_n)\subset\mathcal{C}\quad \text{s.t}\quad \mathcal{C}\subset\bigcup_i B^{k\,k}_{p_i,\varepsilon}\right\}$$





$$N_\varepsilon(\mathcal{C};\|\cdot\|) \le n_0 \iff \varepsilon_n(\mathcal{C};\|\cdot\|) \le \varepsilon_0$$

"Metric entropy" "Entropy number"