Consistent Long-Term Forecasting of geometrically ergodic dynamical systems

linear algebra tools in the service of statistical machine learning

Karim Lounici

Pietro Novelli

Prune Inzerili

Massimiliano Pontil

iit CSML
ISTITUTO ITALIANO DI TECNOLOGIA

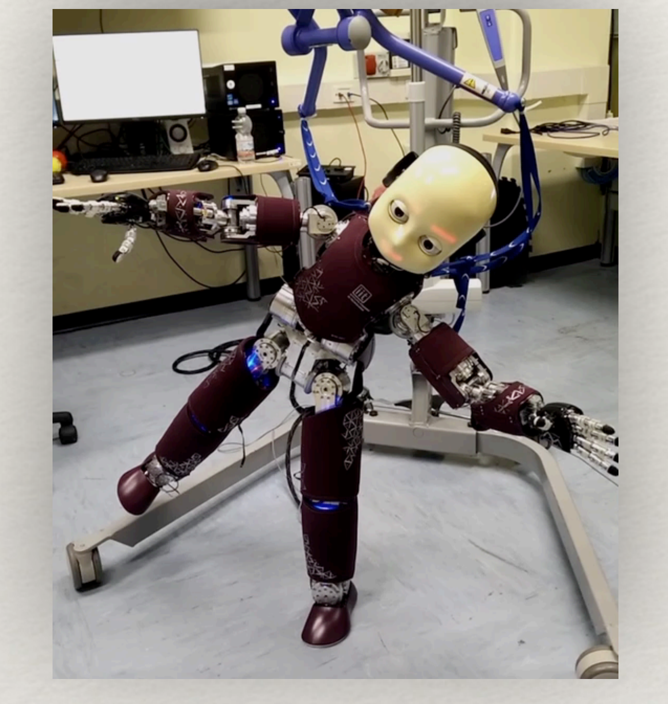
l'X ÉCOLE POLYTECHNIQUE

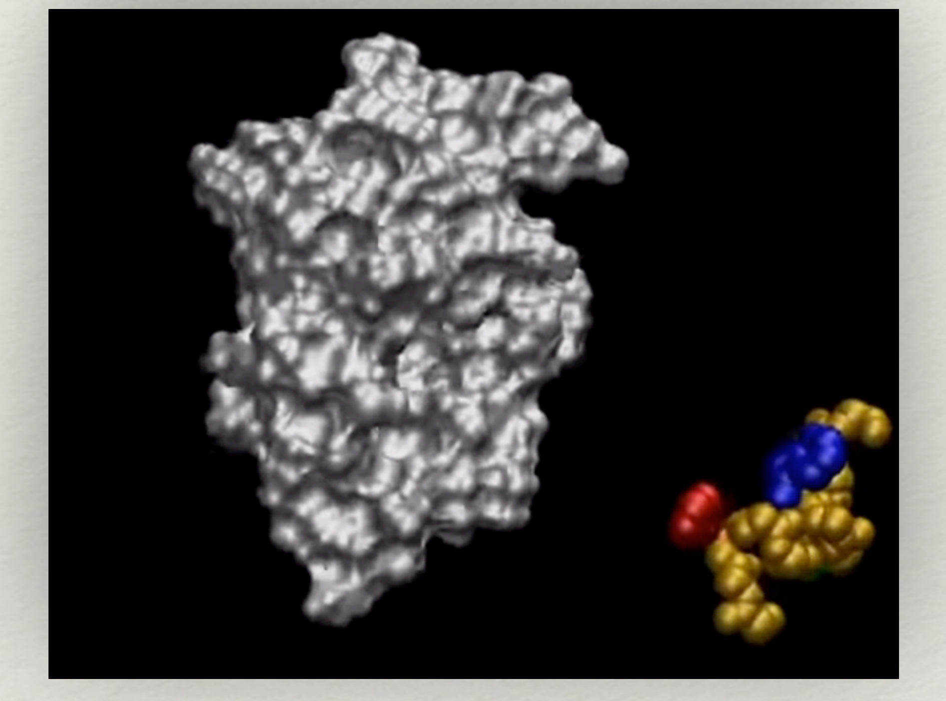
UCL

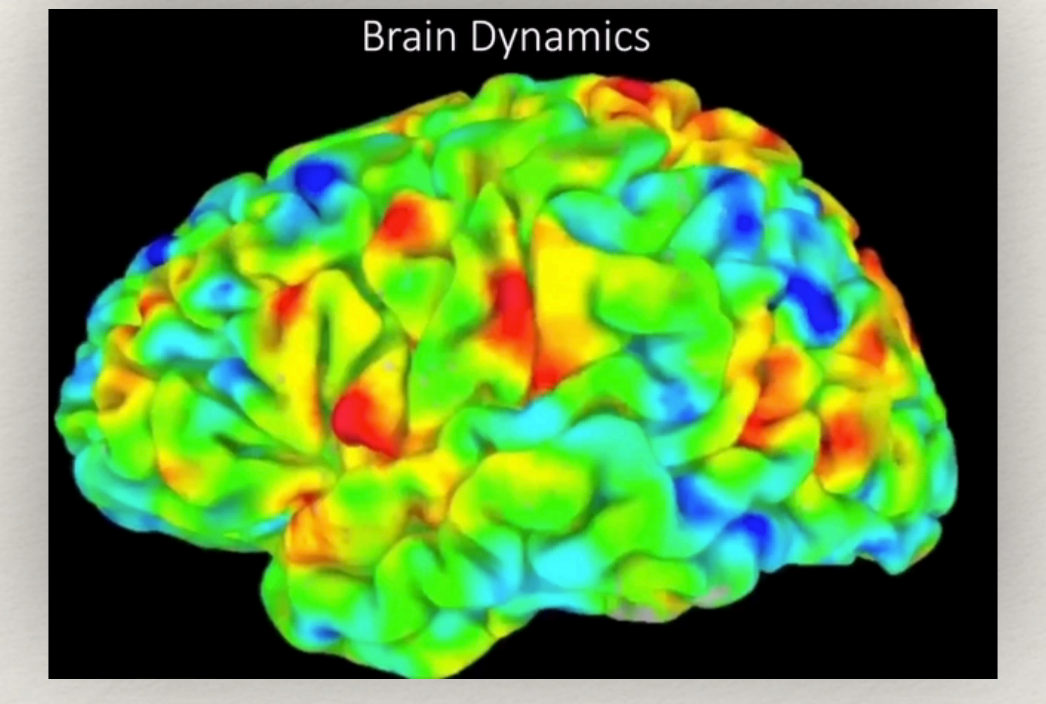
DMI DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Dynamical Systems & ML

DS are backbone mathematical models of temporally evolving phenomena

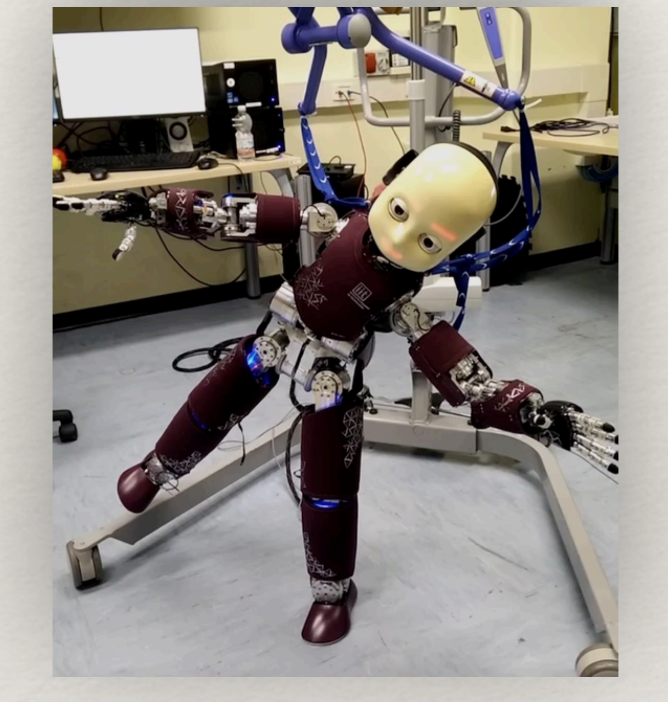


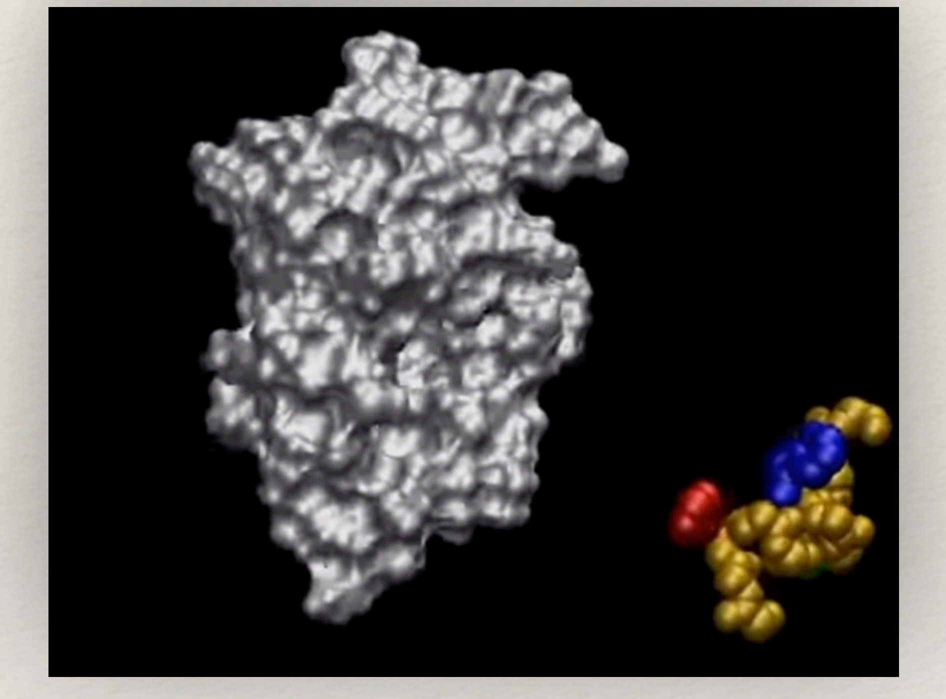


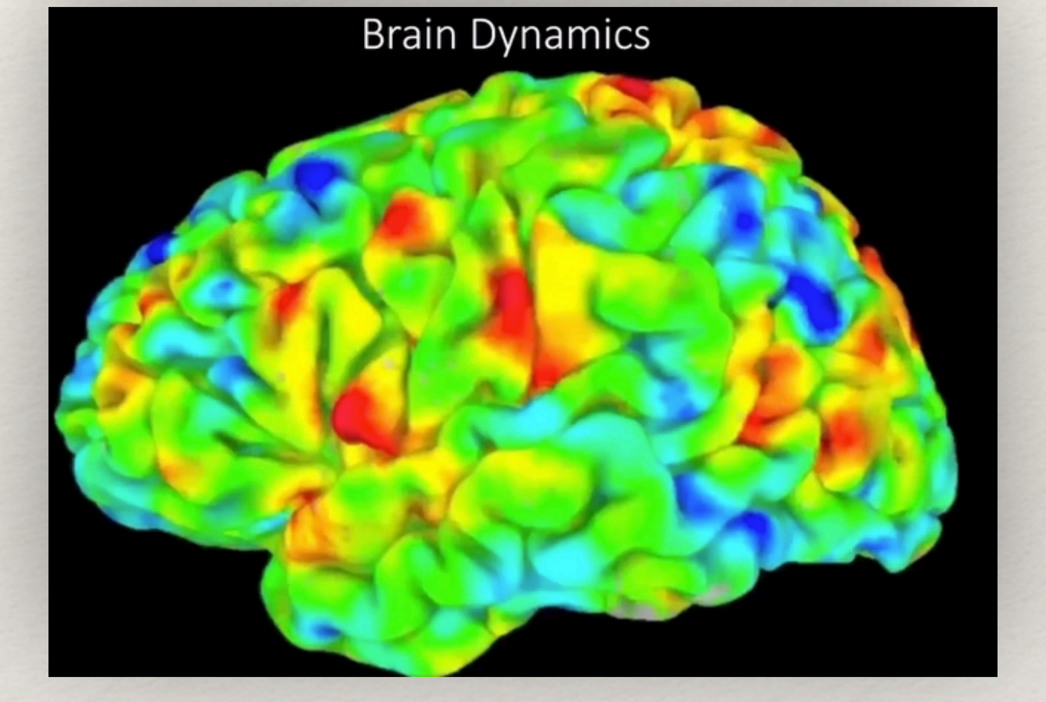


Dynamical Systems & ML

DS are backbone mathematical models of temporally evolving phenomena





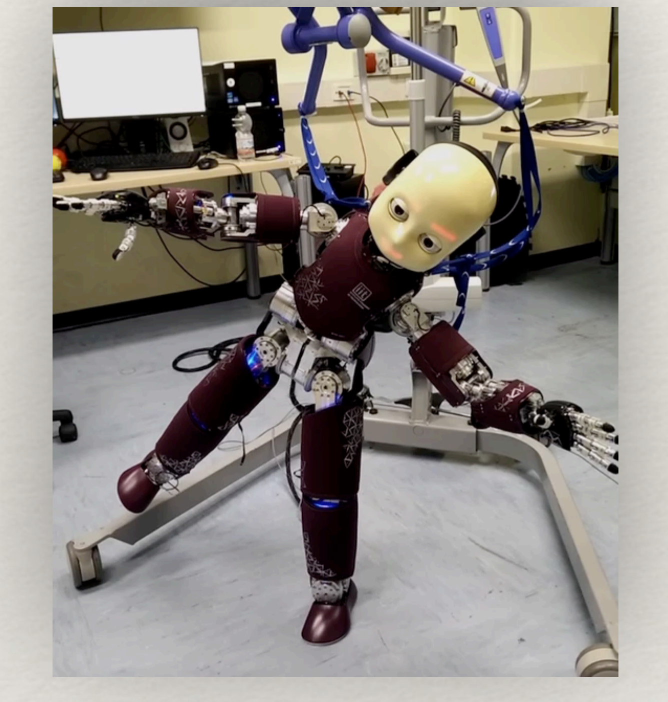


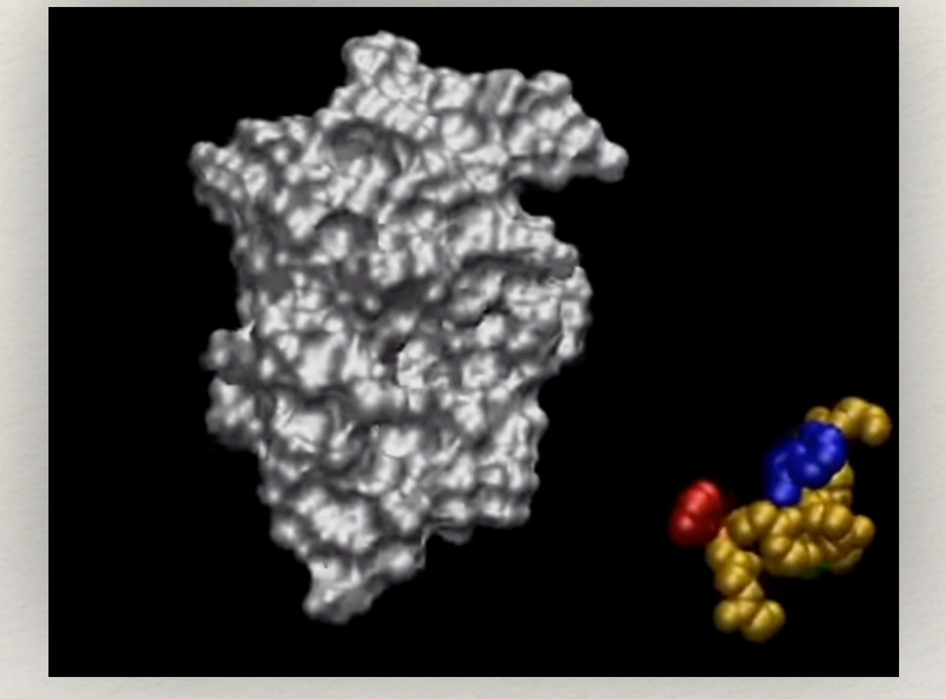
Paradigm shift in Sci & Eng:

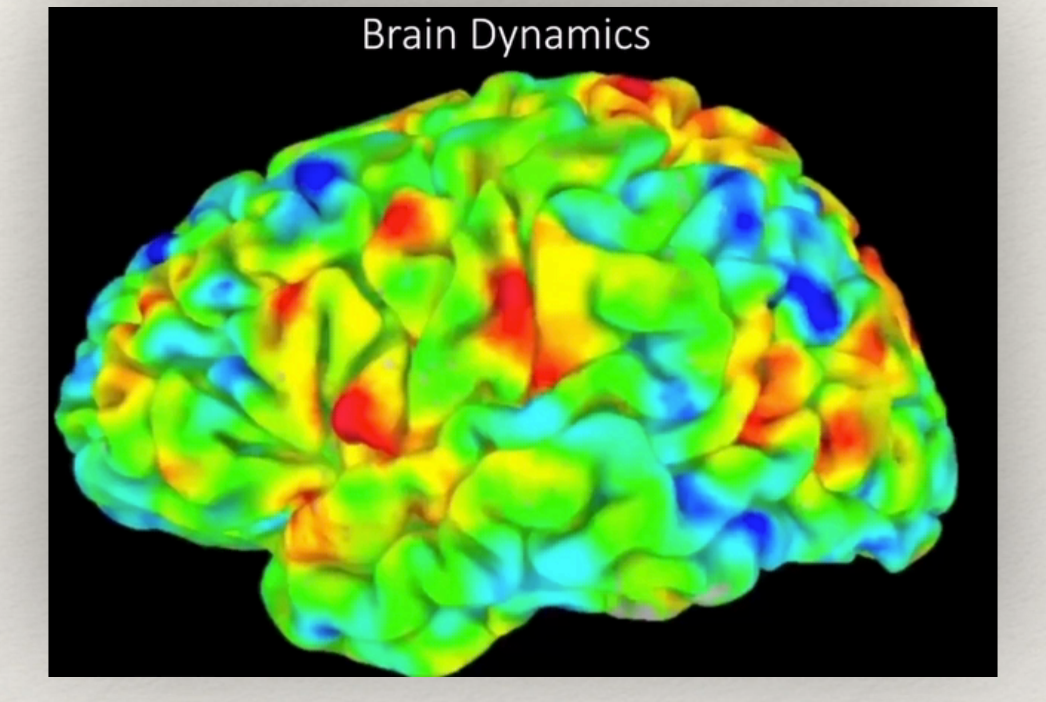
- Classical approach: ODE / PDE / SDE models + parameter fitting
- ML approach: Can we build dynamical models purely from the observed data?

Dynamical Systems & ML

DS are backbone mathematical models of temporally evolving phenomena







Paradigm shift in Sci & Eng:

- Classical approach: ODE/PDE/SDE models + parameter fitting
- ML approach: Can we build dynamical models purely from the observed data?

This is remarkably elegant via transfer operators theory!

The Problem of Long-Term Forecasting

The Problem of Long-Term Forecasting

- We focus on discrete time homogenous Markov process:

$$(X_t)_{t \geq 0} \subseteq \mathcal{X}, \quad X_t \sim \mu_t, \quad \mathbb{P}\left[X_{t+1} \,|\, X_1, \ldots, X_t\right] = \mathbb{P}\left[X_{t+1} \,|\, X_t\right] \text{ independent of } t$$

The Problem of Long-Term Forecasting

- We focus on discrete time homogenous Markov process:

$$(X_t)_{t\geq 0} \subseteq \mathcal{X}, \quad X_t \sim \mu_t, \quad \mathbb{P}\left[X_{t+1} \,|\, X_1, \ldots, X_t\right] = \mathbb{P}\left[X_{t+1} \,|\, X_t\right] \text{ independent of } t$$

- Question:

Given the trajectory data $\mathcal{D}_n = (x_i)_{i\in[n]}$ from one realisation of the process, and given a sample $\mathcal{D}^0_{n_0} = (z_i)_{i\in[n_0]}$ from some arbitrary μ_0 can we find the learning algorithm that produces $\widehat{\mu}_t$ s.t. $\|\mu_t - \widehat{\mu}_t\| \leq \varepsilon(n)$ w.h.p independently of $t \in \mathbb{N}$?

The Problem of Long-Term Forecasting

- We focus on discrete time homogenous Markov process:

$$(X_t)_{t\geq 0} \subseteq \mathcal{X}, \quad X_t \sim \mu_t, \quad \mathbb{P}\left[X_{t+1} \,|\, X_1, \ldots, X_t\right] = \mathbb{P}\left[X_{t+1} \,|\, X_t\right] \text{ independent of } t$$

- Question:

Given the trajectory data $\mathcal{D}_n = (x_i)_{i\in[n]}$ from one realisation of the process, and given a sample $\mathcal{D}^0_{n_0} = (z_i)_{i\in[n_0]}$ from some arbitrary μ_0 can we find the learning algorithm that produces $\hat{\mu}_t$ s.t. $\|\mu_t - \hat{\mu}_t\| \leq \varepsilon(n)$ w.h.p independently of $t \in \mathbb{N}$?

- Spoiler Alert:

For geometrically ergodic processes and MMD norm the answer is YES!

The Transfer Operator Perspective

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge \ \mu_0 = \pi \quad \Longrightarrow \quad \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \to \pi}$$

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge \ \mu_0 = \pi \quad \Longrightarrow \quad \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \ \to \pi}$$

- The forward **transfer operator** evolves observables $f \colon \mathcal{X} \to \mathbb{R}$:

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x \,]}$$

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge\ \mu_0 = \pi \quad \Longrightarrow \quad \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \ \to \pi}$$

- The forward **transfer operator** evolves observables $f\colon \mathcal{X} \to \mathbb{R}$:

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- The backward **transfer operator** evolves distributions $q_t := d\mu_t/d\pi \in L^2_\pi(\mathcal{X})$

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge \ \mu_0 = \pi \quad \Longrightarrow \quad \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \ \to \pi}$$

- The forward **transfer operator** evolves observables $f\colon\ \mathcal{X} \to \mathbb{R}$:

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- The backward **transfer operator** evolves distributions $q_t := d\mu_t/d\pi \in L^2_\pi(\mathcal{X})$

$$\langle q_t , f\rangle_{L^2_\pi(\mathcal{X})} = \mathbb{E}[f(X_t)] = \mathbb{E}[\mathbb{E}[f(X_t)\,|\,X_{t-1}]] = \langle q_{t-1}, A_\pi f\rangle_{L^2_\pi(\mathcal{X})} = \langle A_\pi^* q_{t-1} , f\rangle_{L^2_\pi(\mathcal{X})}$$

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge \ \mu_0 = \pi \quad \Longrightarrow \quad \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \ \to \pi}$$

- The forward **transfer operator** evolves observables $f\colon \mathcal{X} \to \mathbb{R}$:

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- The backward **transfer operator** evolves distributions $q_t := d\mu_t / d\pi \in L^2_\pi(\mathcal{X})$

$$\langle q_t , f \rangle_{L^2_\pi(\mathcal{X})} = \mathbb{E}[f(X_t)] = \mathbb{E}[\mathbb{E}[f(X_t) \,|\, X_{t-1}]] = \langle q_{t-1}, A_\pi f \rangle_{L^2_\pi(\mathcal{X})} = \langle A^*_\pi q_{t-1} , f \rangle_{L^2_\pi(\mathcal{X})}$$

$$\boxed{q_t = A^*_\pi q_{t-1} = (A^*_\pi)^t q_0 \quad \text{Autonomous Linear Dynamical System}}$$

The Transfer Operator Perspective

- Consider stable stochastic dynamics with the invariant measure π

$$\boxed{X_t \sim \mu_t \ \wedge \ \mu_0 = \pi \implies \mu_t = \pi, t \in \mathbb{N}} \quad \boxed{\forall \mu_0 \quad \mu_t \to \pi}$$

- The forward **transfer operator** evolves observables $f\colon \mathcal{X} \to \mathbb{R}$:

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- The backward **transfer operator** evolves distributions $q_t := d\mu_t / d\pi \in L^2_\pi(\mathcal{X})$

$$\langle q_t, f\rangle_{L^2_\pi(\mathcal{X})} = \mathbb{E}[f(X_t)] = \mathbb{E}[\mathbb{E}[f(X_t) \mid X_{t-1}]] = \langle q_{t-1}, A_\pi f\rangle_{L^2_\pi(\mathcal{X})} = \langle A^*_\pi q_{t-1}, f\rangle_{L^2_\pi(\mathcal{X})}$$

$$\boxed{q_t = A^*_\pi q_{t-1} = (A^*_\pi)^t q_0 \quad \text{Autonomous Linear Dynamical System}}$$

LA

Spectral Decomposition

$$\boxed{\mathbb{P}\left[X_{t+1} \mid X_t = \cdot\,\right] \ll \pi \quad \Longrightarrow \quad A_\pi \quad \text{is compact}}$$

Spectral Decomposition

$$\mathbb{P}\left[X_{t+1} \,|\, X_t = \cdot\,\right] \ll \pi \quad \Longrightarrow \quad A_\pi \quad \text{is compact}$$

$$A_\pi = \sum_{i=1}^{\infty} \lambda_i\, f_i \otimes \overline{g_i} \qquad A_\pi f_i = \lambda_i\, f_i, \quad A_\pi^* g_i = \overline{\lambda}_i\, g_i, \quad \langle f_i, \overline{g_j} \rangle_{L^2_\pi(\mathcal{X})} = \delta_{i,j}$$

scalars $\lambda_i \in \mathbb{C}$ are the eigenvalues and and functions f_i and g_i are left and right eigenfunctions

Spectral Decomposition

$$\mathbb{P}\left[X_{t+1} \,|\, X_t = \cdot\,\right] \ll \pi \implies A_\pi \quad \text{is compact}$$

$$A_\pi = \sum_{i=1}^{\infty} \lambda_i \, f_i \otimes \overline{g_i} \qquad A_\pi f_i = \lambda_i \, f_i, \quad A_\pi^* g_i = \overline{\lambda}_i \, g_i, \quad \langle f_i, \overline{g_j} \rangle_{L^2_\pi(\mathcal{X})} = \delta_{i,j}$$

scalars $\lambda_i \in \mathbb{C}$ are the eigenvalues and and functions f_i and g_i are left and right eigenfunctions

$$\mathbb{E}[\, f(X_t) \,|\, X_0 = x\,] = (A_\pi^t f)(x) = \sum_j \boxed{\lambda_j^t} \, \boxed{f_j(x) \, \langle \overline{g_j}, f \rangle_{L^2_\pi(\mathcal{X})}}$$

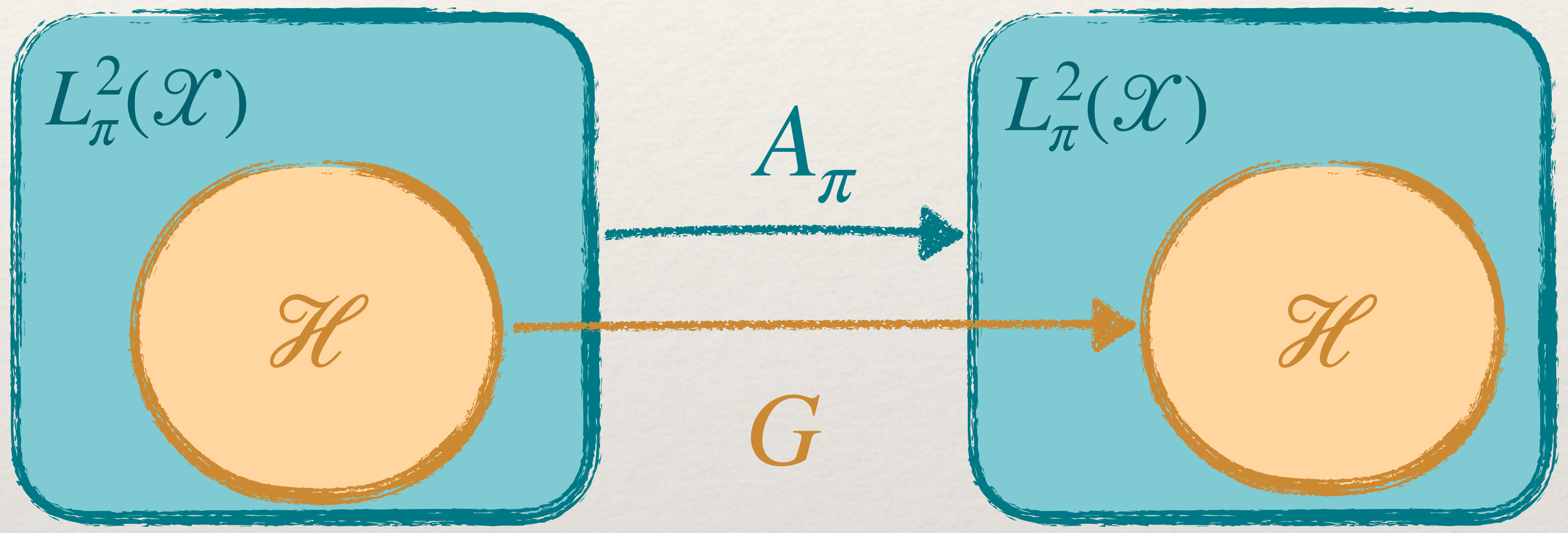
LA

the expectation of an observable is disentangled into **temporal** and **static** components

Learning the operator and its spectra

Learning the operator and its spectra

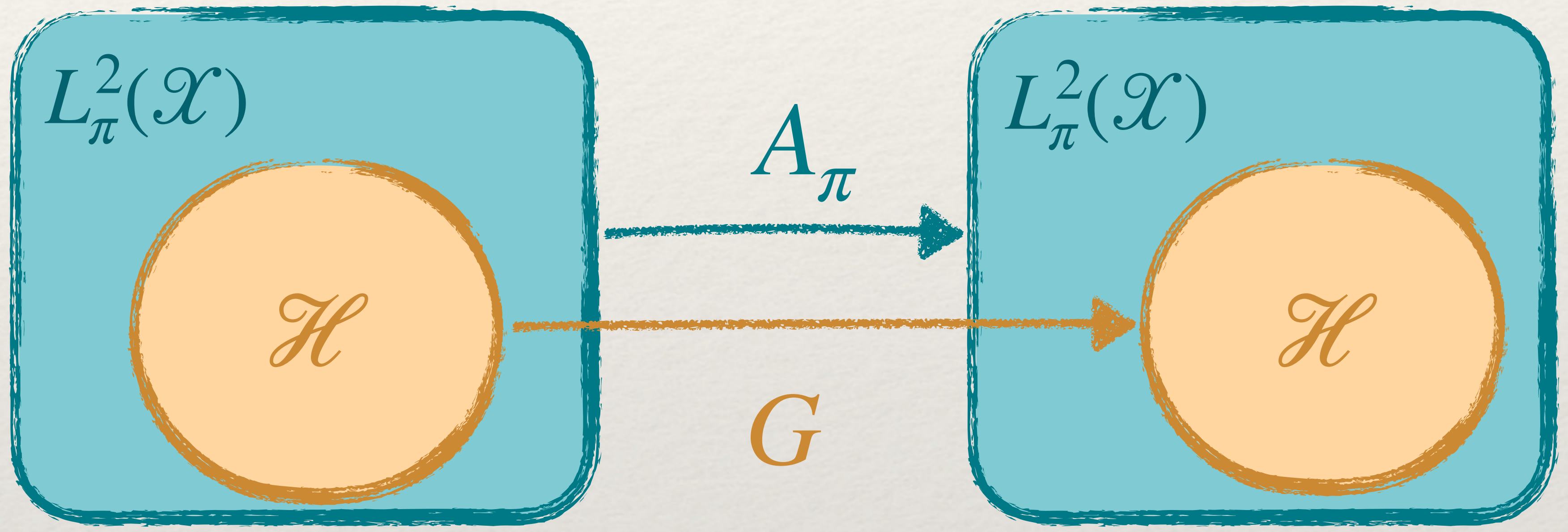
- Since we don't know $L^2_\pi(\mathscr{X})$ we restrict A_π to a chosen RKHS $\mathscr{H}$ and look for an operator $G : \mathscr{H} \to \mathscr{H}$ such that $A_\pi \langle w, \phi(\cdot) \rangle \approx \langle Gw, \phi(\cdot) \rangle$, that is



Learning the operator and its spectra

- Since we don't know $L^2_\pi(\mathcal{X})$ we restrict A_π to a chosen RKHS $\mathcal{H}$ and look for an operator $G : \mathcal{H} \to \mathcal{H}$ such that $A_\pi\langle w, \phi(\,\cdot\,)\rangle \approx \langle Gw, \phi(\,\cdot\,)\rangle$, that is

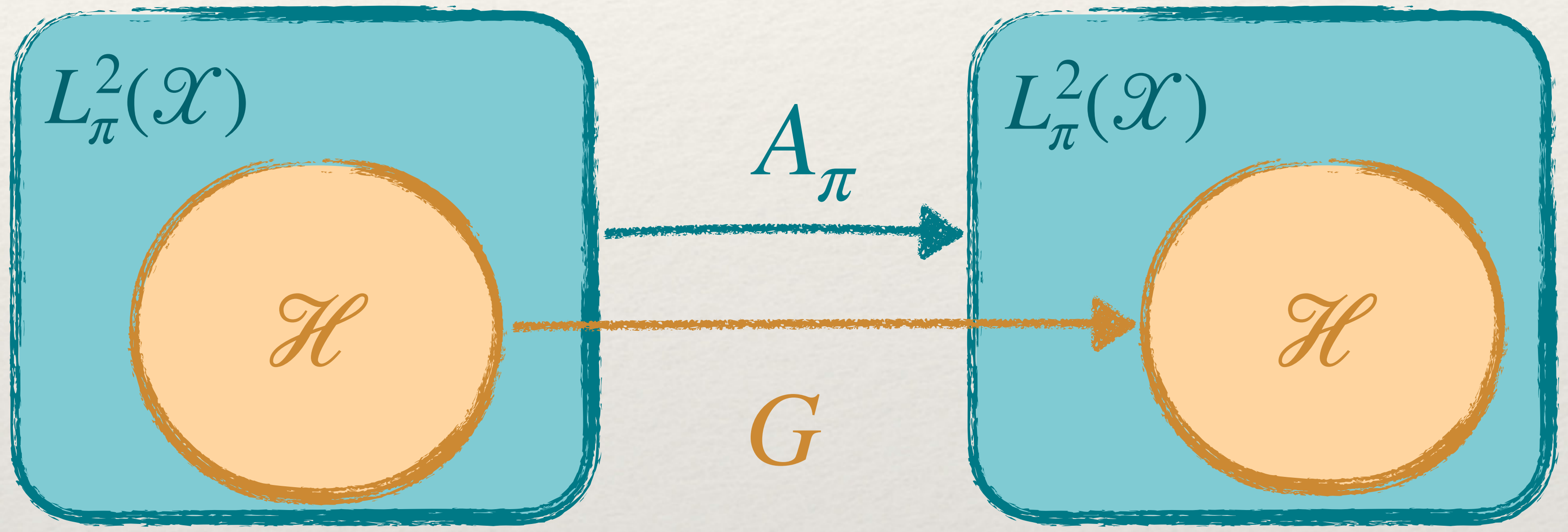
$$\boxed{\mathcal{R}(G) = \mathbb{E}_{X_t \sim \pi} \|\phi(X_{t+1}) - G^* \phi(X_t)\|^2}$$



Learning the operator and its spectra

- Since we don't know $L^2_\pi(\mathcal{X})$ we restrict A_π to a chosen RKHS $\mathcal{H}$ and look for an operator $G : \mathcal{H} \to \mathcal{H}$ such that $A_\pi\langle w, \phi(\,\cdot\,)\rangle \approx \langle Gw, \phi(\,\cdot\,)\rangle$, that is

$$\boxed{\mathcal{R}(G) = \mathbb{E}_{X_t\sim\pi}\|\phi(X_{t+1}) - G^*\phi(X_t)\|^2}$$



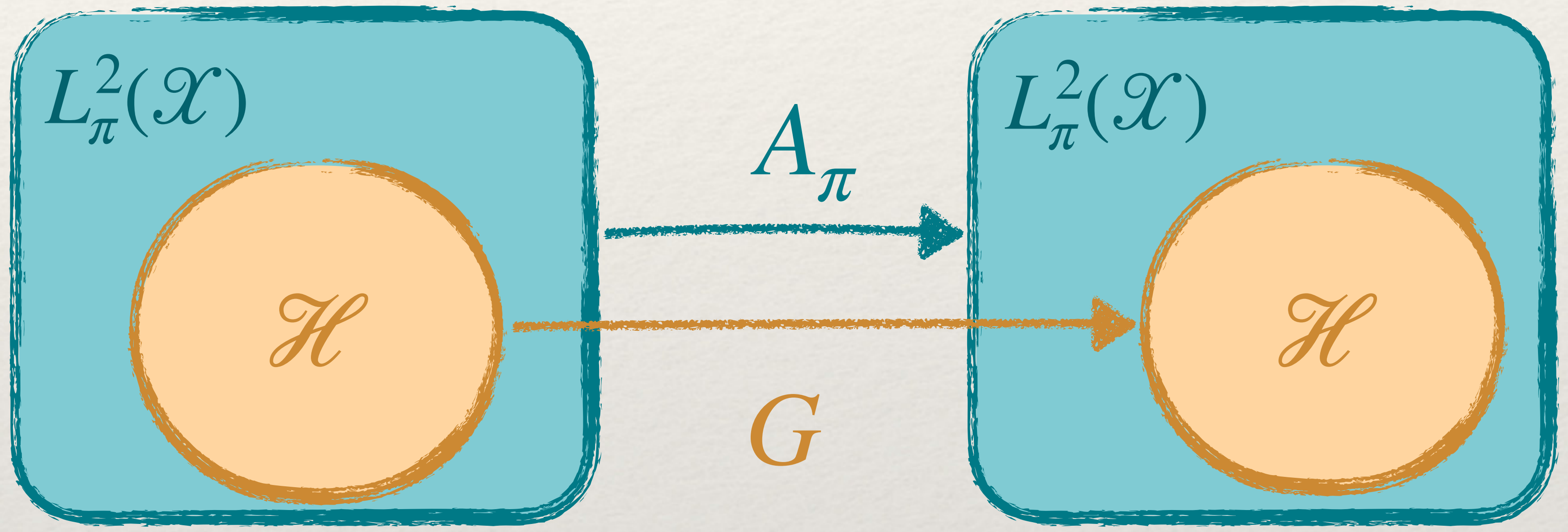
$$Gh_i = \lambda_i h_i \;\Rightarrow\; \|(\lambda_i\, I - A_\pi)^{-1}\|^{-1} \le \|A_\pi h_i - \lambda_i h_i\|_{L^2_\pi(\mathcal{X})} \le \underbrace{\|A_\pi|_{\mathcal{H}} - G\|_{\mathcal{H}\to L^2_\pi(\mathcal{X})}}_{\mathcal{E}(G)} \; \underbrace{\frac{\|h_i\|_{\mathcal{H}}}{\|h_i\|_{L^2_\pi(\mathcal{X})}}}$$

operator norm error **metric distortion**

Learning the operator and its spectra

- Since we don't know $L^2_\pi(\mathscr{X})$ we restrict A_π to a chosen RKHS $\mathscr{H}$ and look for an operator $G : \mathscr{H} \to \mathscr{H}$ such that $A_\pi\langle w, \phi(\,\cdot\,)\rangle \approx \langle Gw, \phi(\,\cdot\,)\rangle$, that is

$$\mathscr{R}(G) = \mathbb{E}_{X_t\sim\pi}\|\phi(X_{t+1}) - G^*\phi(X_t)\|^2$$



$$Gh_i = \lambda_i h_i \;\Rightarrow\; \|(\lambda_i\, I - A_\pi)^{-1}\|^{-1} \le \|A_\pi h_i - \lambda_i h_i\|_{L^2_\pi(\mathscr{X})} \le \underbrace{\|A_\pi|_{\mathscr{H}} - G\|_{\mathscr{H}\to L^2_\pi(\mathscr{X})}}_{\substack{\mathscr{E}(G) \\ \textbf{operator norm error}}} \; \underbrace{\frac{\|h_i\|_{\mathscr{H}}}{\|h_i\|_{L^2_\pi(\mathscr{X})}}}_{\substack{\textbf{metric} \\ \textbf{distorsion}}}$$

$$\|\mathbb{E}[h(X_{t+1}) \mid X_t = \cdot\,] - Gh\| \le \mathscr{E}(G)\|h\|_{\mathscr{H}}$$

one-step ahead prediction

How to generalise beyond one-step ahead?

How to generalise beyond one-step ahead?

1. Review some Linear Algebra tools on understanding linear dynamics
2. Based on these ideas develop Deflate-Learn-Inflate (DLI) approach
3. Use error decomposition techniques and concentration inequalities

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{\,|\lambda| : \lambda \in \Lambda(A)\}$ determines asymptotic behaviour:

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{\,|\lambda| :\ \lambda \in \Lambda(A)\}$ determines asymptotic behaviour:

$$\boxed{\rho(A) < 1 \quad \Longrightarrow \quad \lim_{t\to\infty} A^t = 0 \quad \wedge \quad \lim_{t\to\infty} \ln \|A^t\| / t = \rho(A)}$$

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{\,|\lambda| : \lambda \in \Lambda(A)\}$ determines asymptotic behaviour:

$$\boxed{\rho(A) < 1 \quad \Longrightarrow \quad \lim_{t\to\infty} A^t = 0 \quad \wedge \quad \lim_{t\to\infty} \ln \|A^t\| \,/\, t = \rho(A)}$$

$$\boxed{\rho(A) \le \|A\|}$$

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{ |\lambda| : \lambda \in \Lambda(A) \}$ determines asymptotic behaviour:

$$\boxed{\rho(A) < 1 \quad \Longrightarrow \quad \lim_{t\to\infty} A^t = 0 \quad \wedge \quad \lim_{t\to\infty} \ln \|A^t\| \,/\, t = \rho(A)}$$

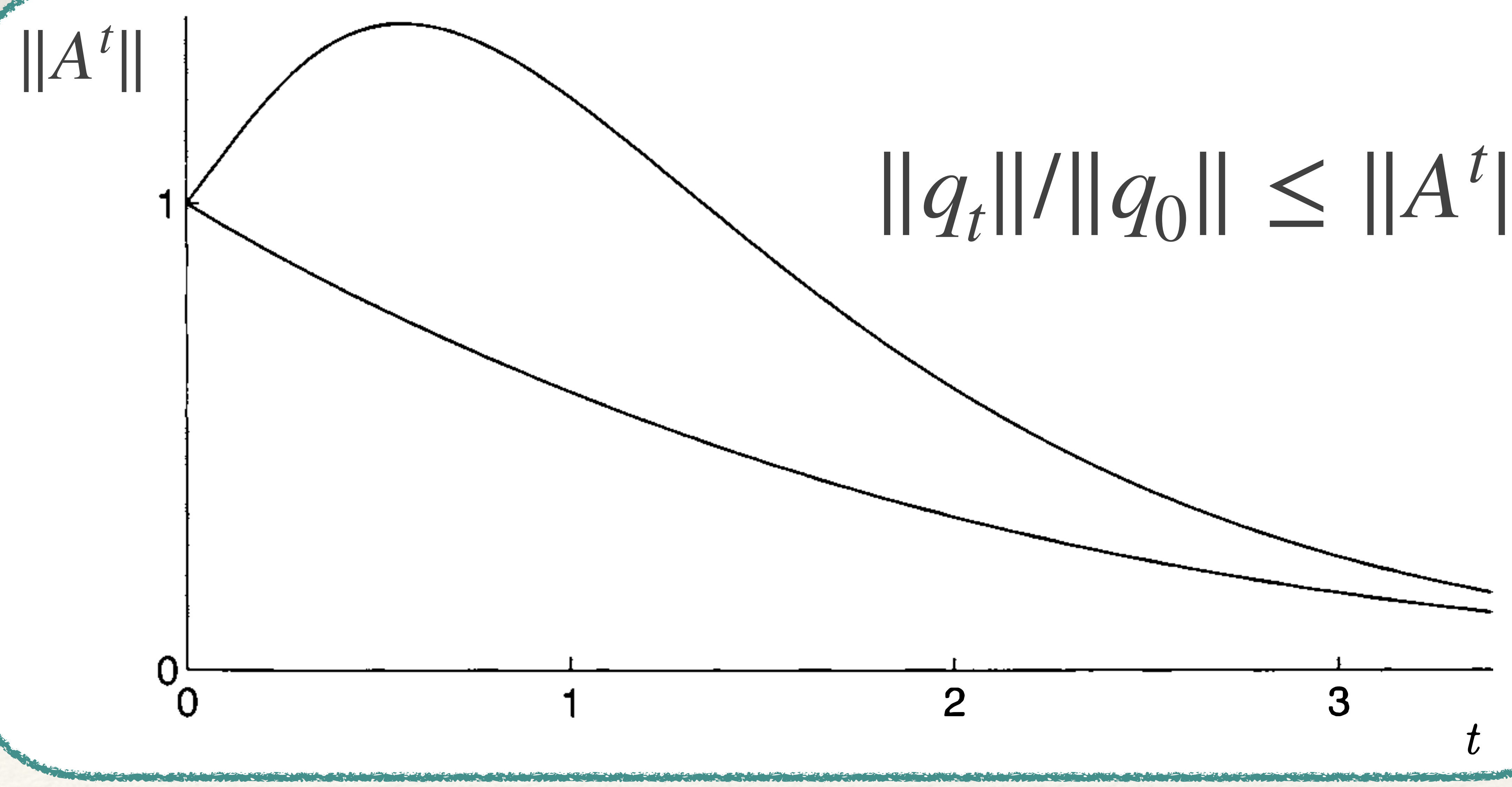
$$\boxed{\rho(A) \leq \|A\|}$$

- if $\|A\| < 1$ then geometric decay $\|A^t\| \leq \|A\|^t$

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{ |\lambda| : \lambda \in \Lambda(A) \}$ determines asymptotic behaviour:

$$\boxed{\rho(A) < 1 \implies \lim_{t\to\infty} A^t = 0 \quad \wedge \quad \lim_{t\to\infty} \ln \|A^t\| / t = \rho(A)}$$



$$\|q_t\|/\|q_0\| \leq \|A^t\|$$

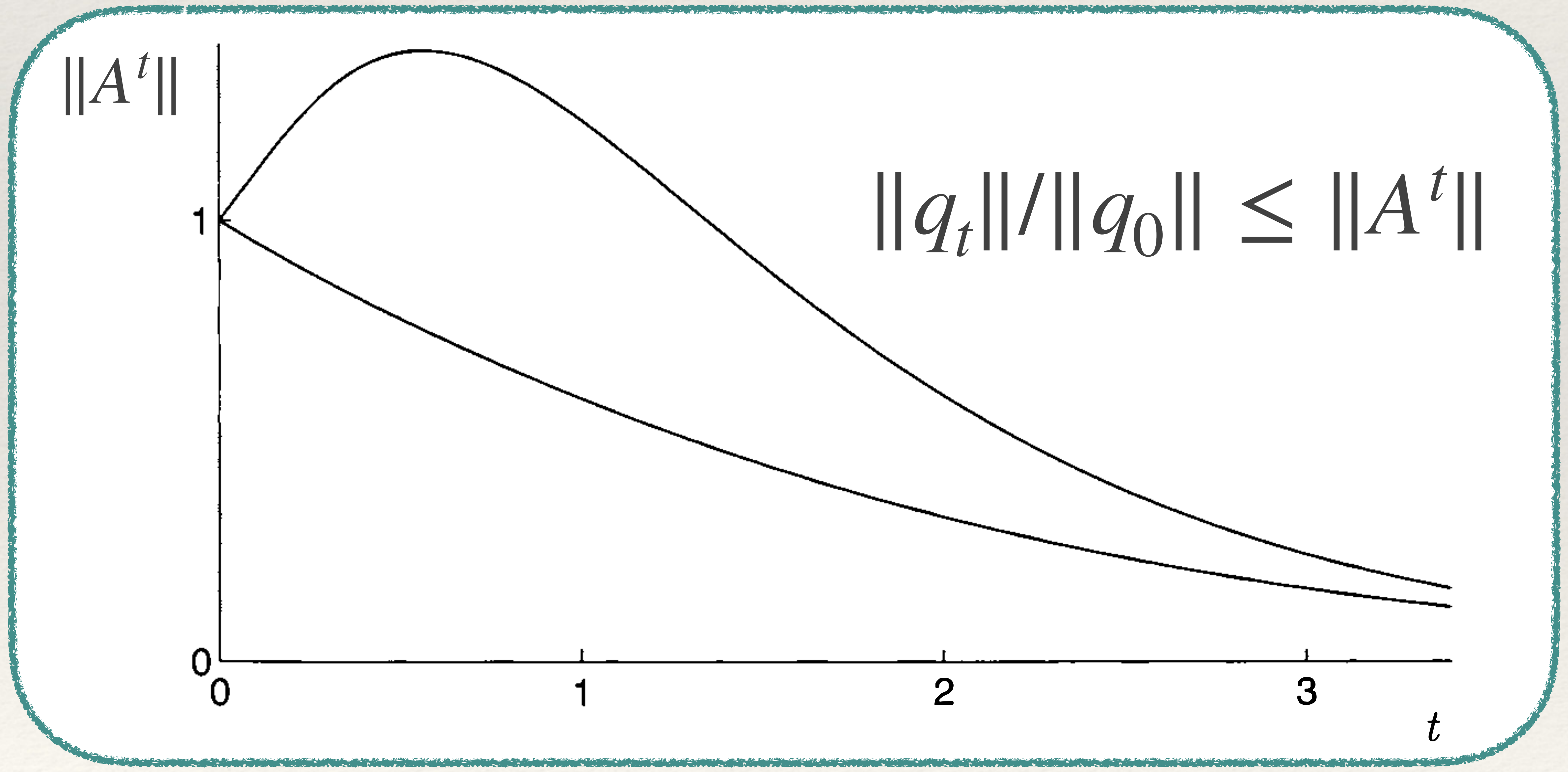
$$\boxed{\rho(A) \leq \|A\|}$$

- if $\|A\| < 1$ then geometric decay $\|A^t\| \leq \|A\|^t$
- if $\|A\| > 1$ then transient growth $\sup_{t\in\mathbb{N}_0} \|A^t\| > 1$

Transient Behaviour of Asymptotically Stable LDS

- Spectral radius $\rho(A) := \{\,|\lambda| :\ \lambda \in \Lambda(A)\}$ determines asymptotic behaviour:

$$\boxed{\rho(A) < 1 \quad \Longrightarrow \quad \lim_{t\to\infty} A^t = 0 \quad \wedge \quad \lim_{t\to\infty} \ln \|A^t\| \,/\, t = \rho(A)}$$



$$\boxed{\rho(A) \leq \|A\|}$$

- if $\|A\| < 1$ then geometric decay $\|A^t\| \leq \|A\|^t$
- if $\|A\| > 1$ then transient growth $\sup_{t\in\mathbb{N}_0} \|A^t\| > 1$
- $\|A\| \gg \rho(A)$ dynamics is highly non-normal

L. N. Trefethen, M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton Uni. Press 2005.

Transient Behaviour of Asymptotically Stable LDS

- Pseudospectrum describes transient behaviour

$$\boxed{\Lambda_\varepsilon(A) := \bigcup_{\|E\|\leq\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI-A)^{-1}\|^{-1} \leq \varepsilon\}}$$

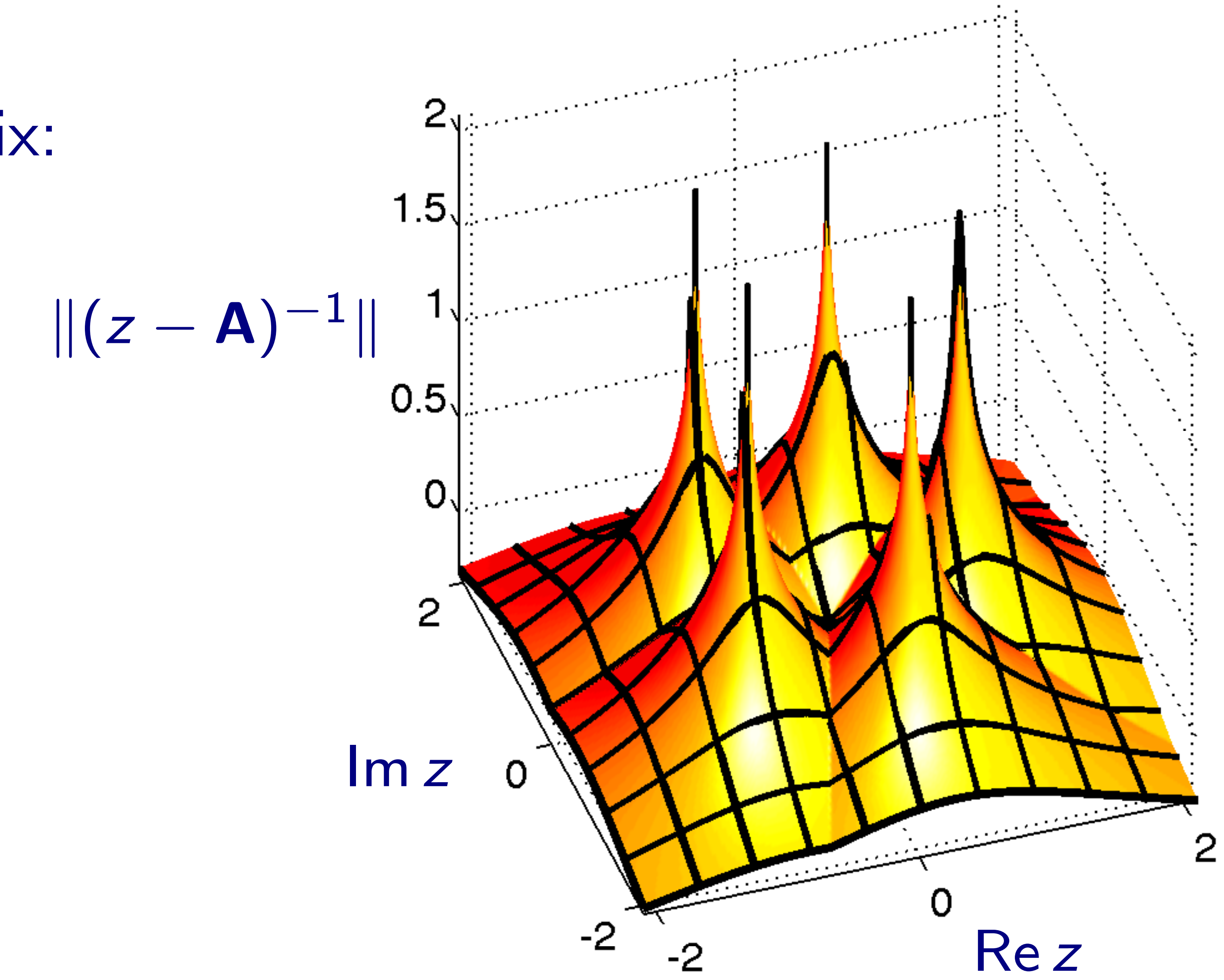
Transient Behaviour of Asymptotically Stable LDS

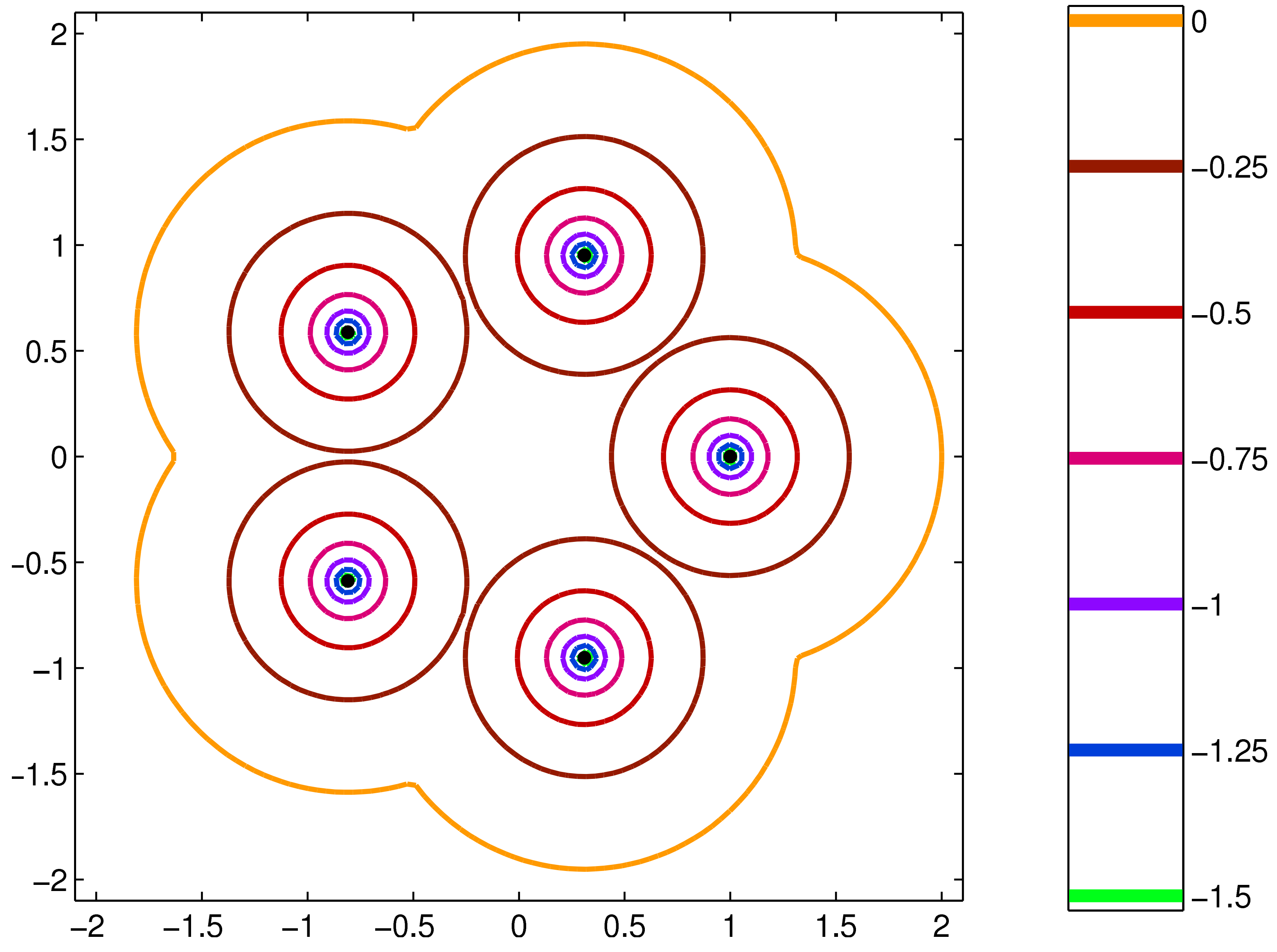
- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\le\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI - A)^{-1}\|^{-1} \le \varepsilon\}$$

A circulant (hence normal) matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$





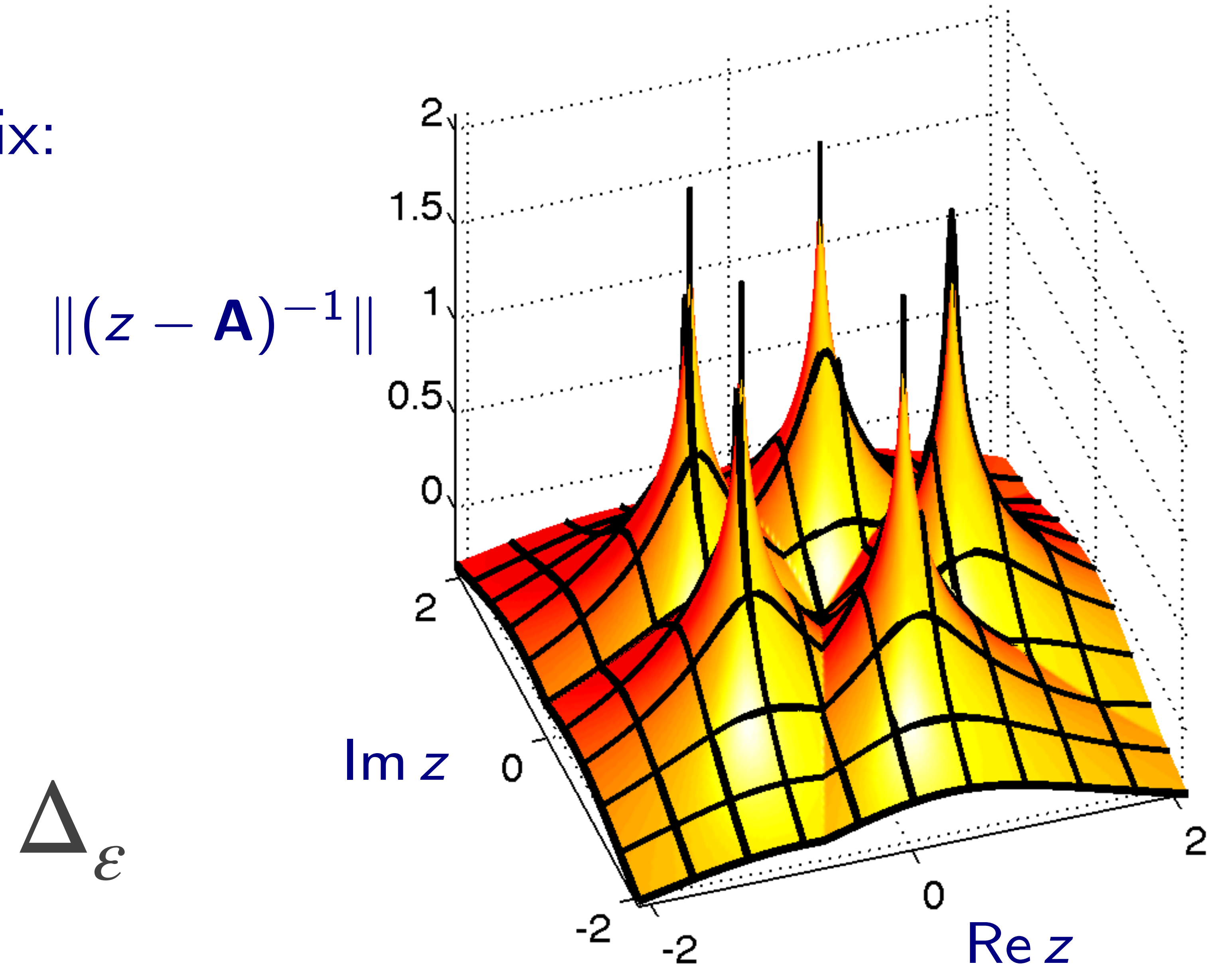
Transient Behaviour of Asymptotically Stable LDS

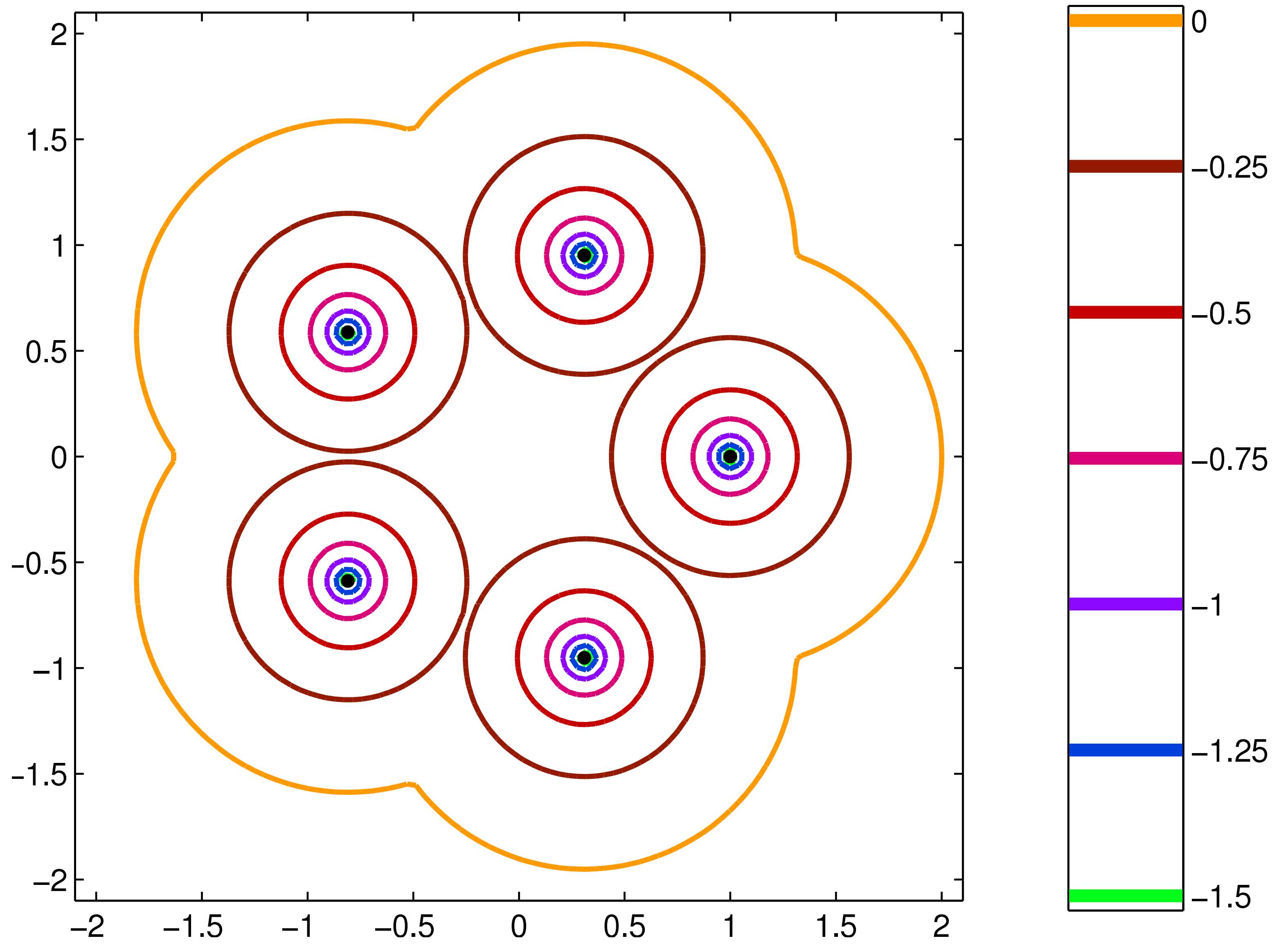
- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\le\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI - A)^{-1}\|^{-1} \le \varepsilon\}$$

A circulant (hence normal) matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$





- normal iff $\Lambda_\varepsilon(A) = \Lambda(A) + \Delta_\varepsilon$

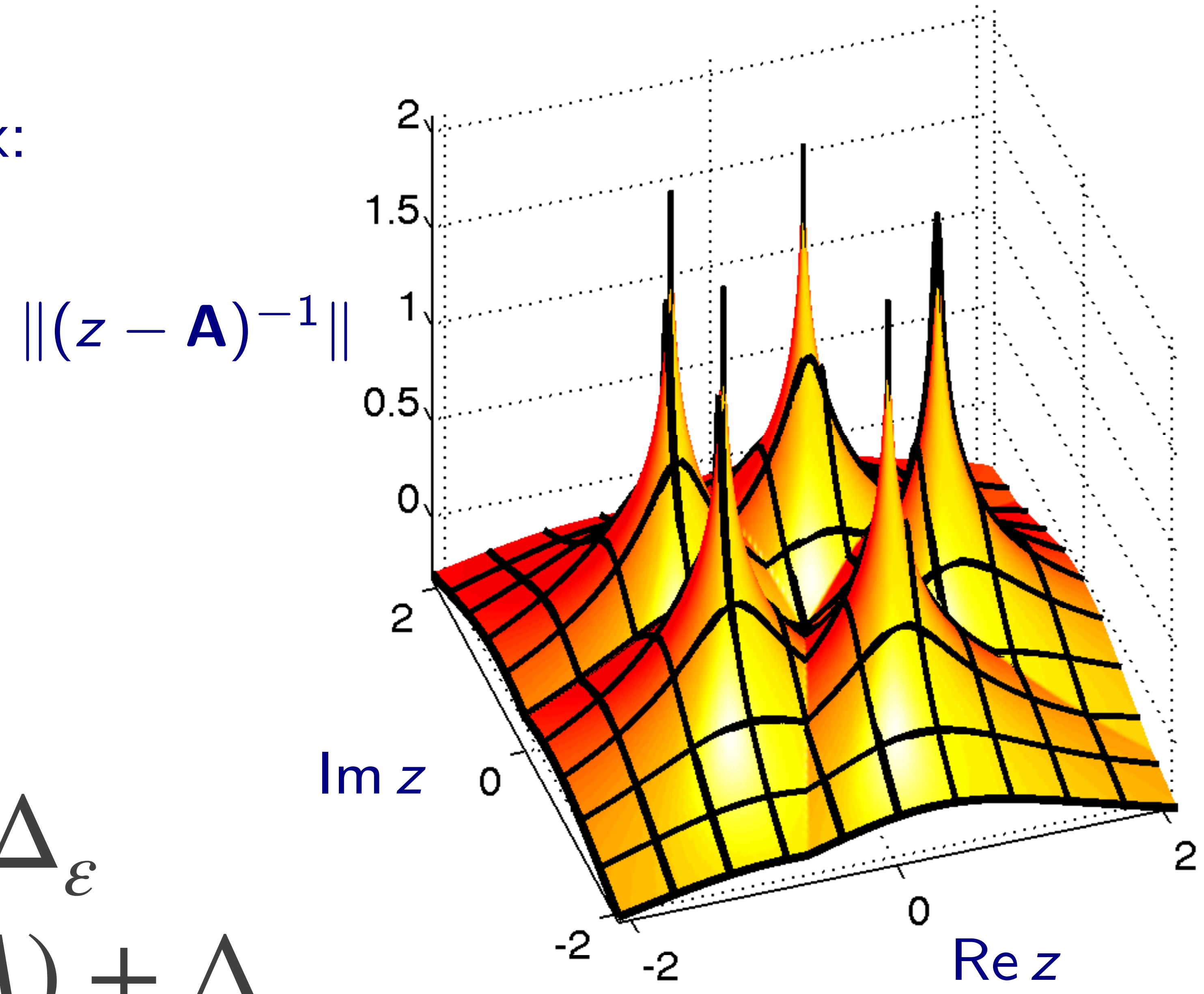
Transient Behaviour of Asymptotically Stable LDS

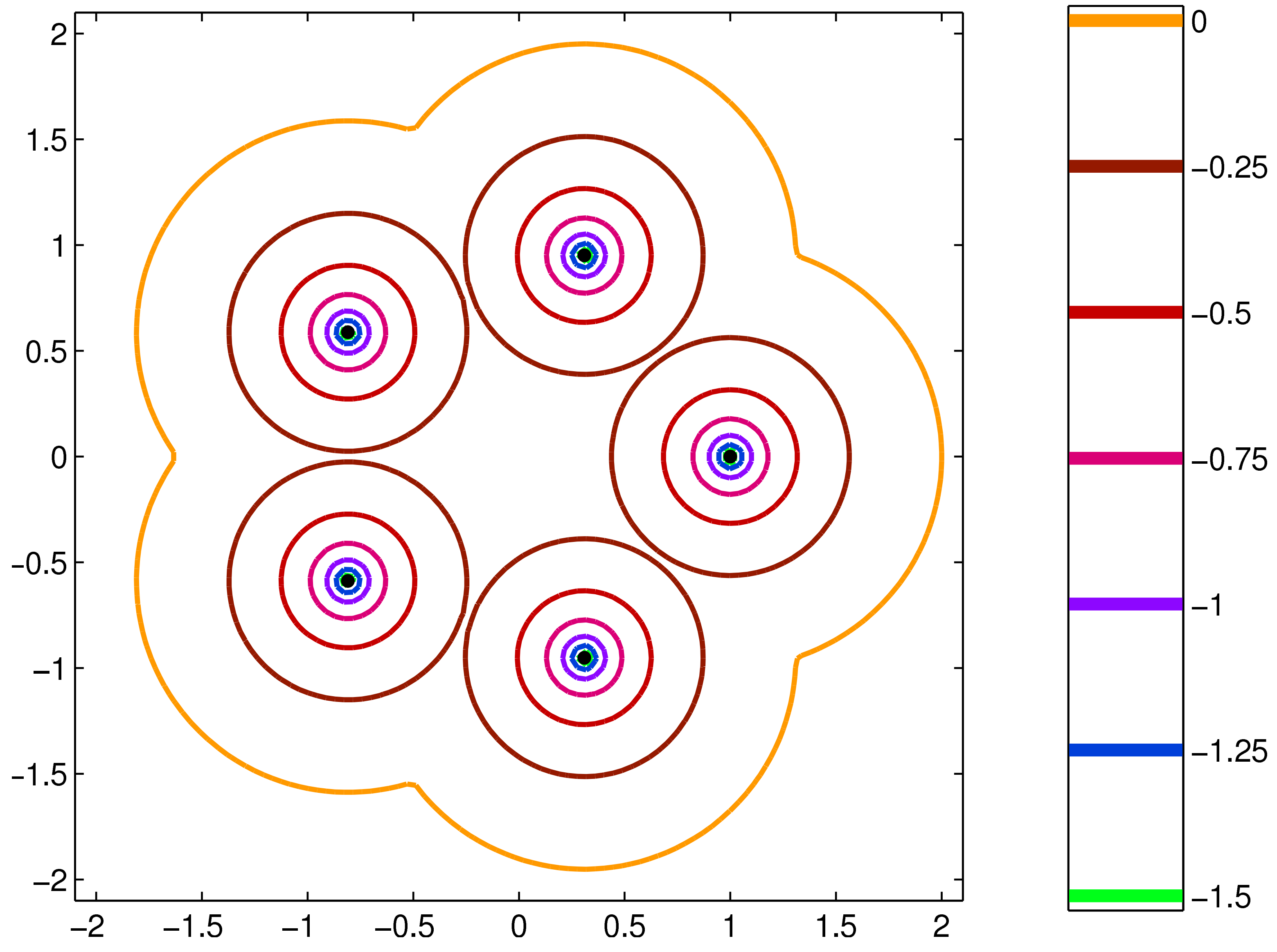
- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\leq\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \|(zI - A)^{-1}\|^{-1} \leq \varepsilon\}$$

A circulant (hence normal) matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$





- normal iff $\Lambda_\varepsilon(A) = \Lambda(A) + \Delta_\varepsilon$
- non-normal iff $\Lambda_\varepsilon(A) \supsetneq \Lambda(A) + \Delta_\varepsilon$

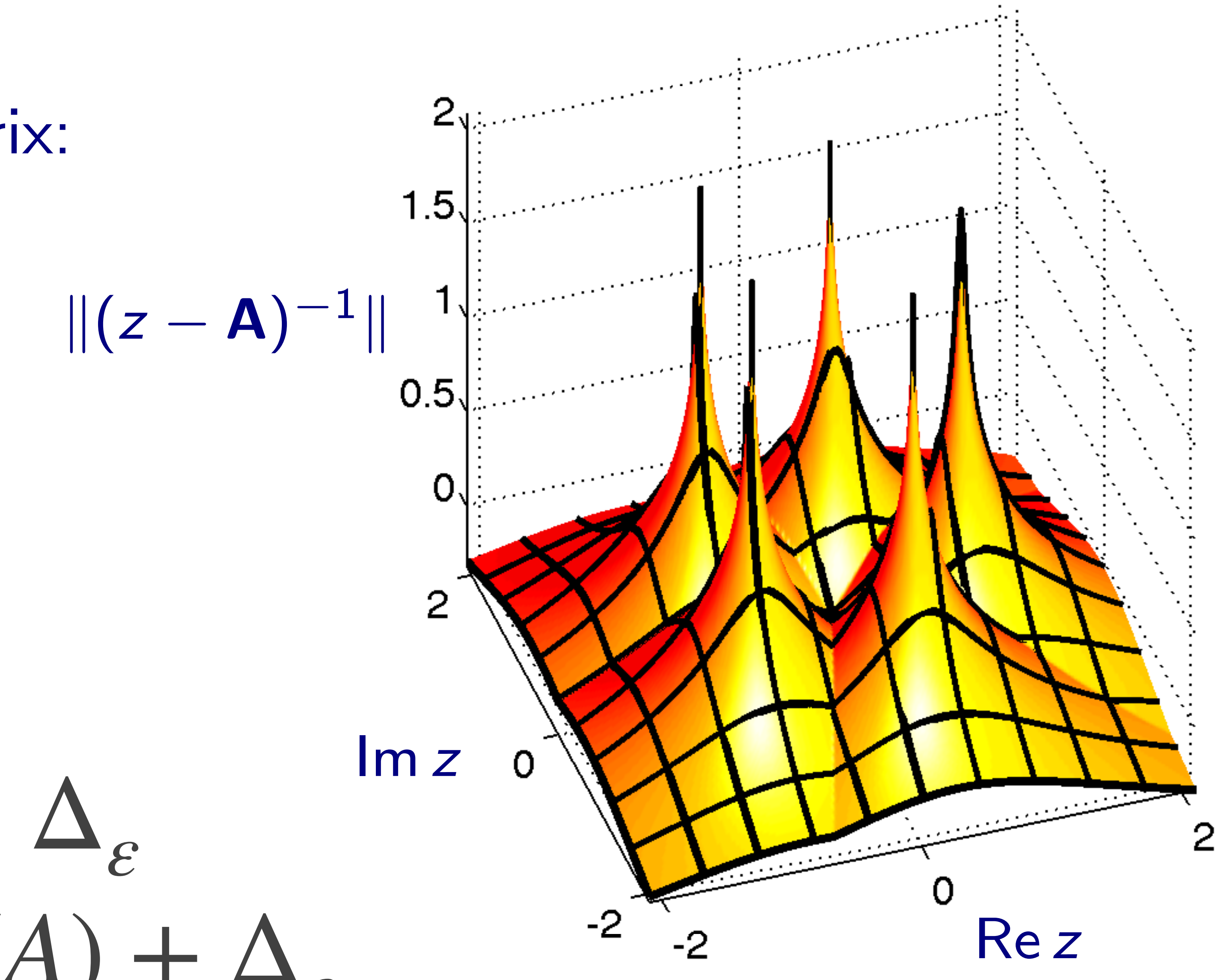
Transient Behaviour of Asymptotically Stable LDS

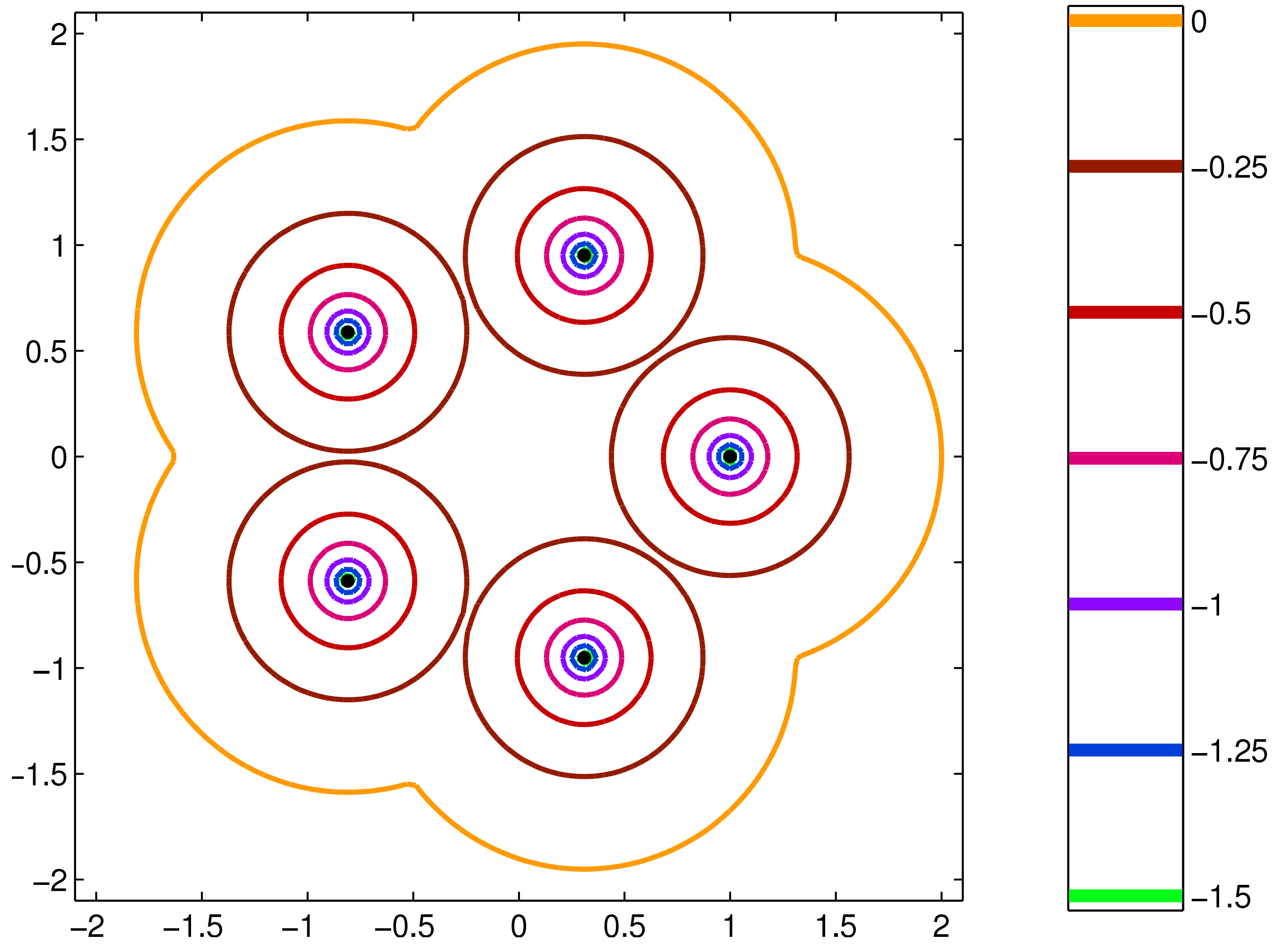
- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\leq\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI - A)^{-1}\|^{-1} \leq \varepsilon\}$$

A circulant (hence normal) matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$





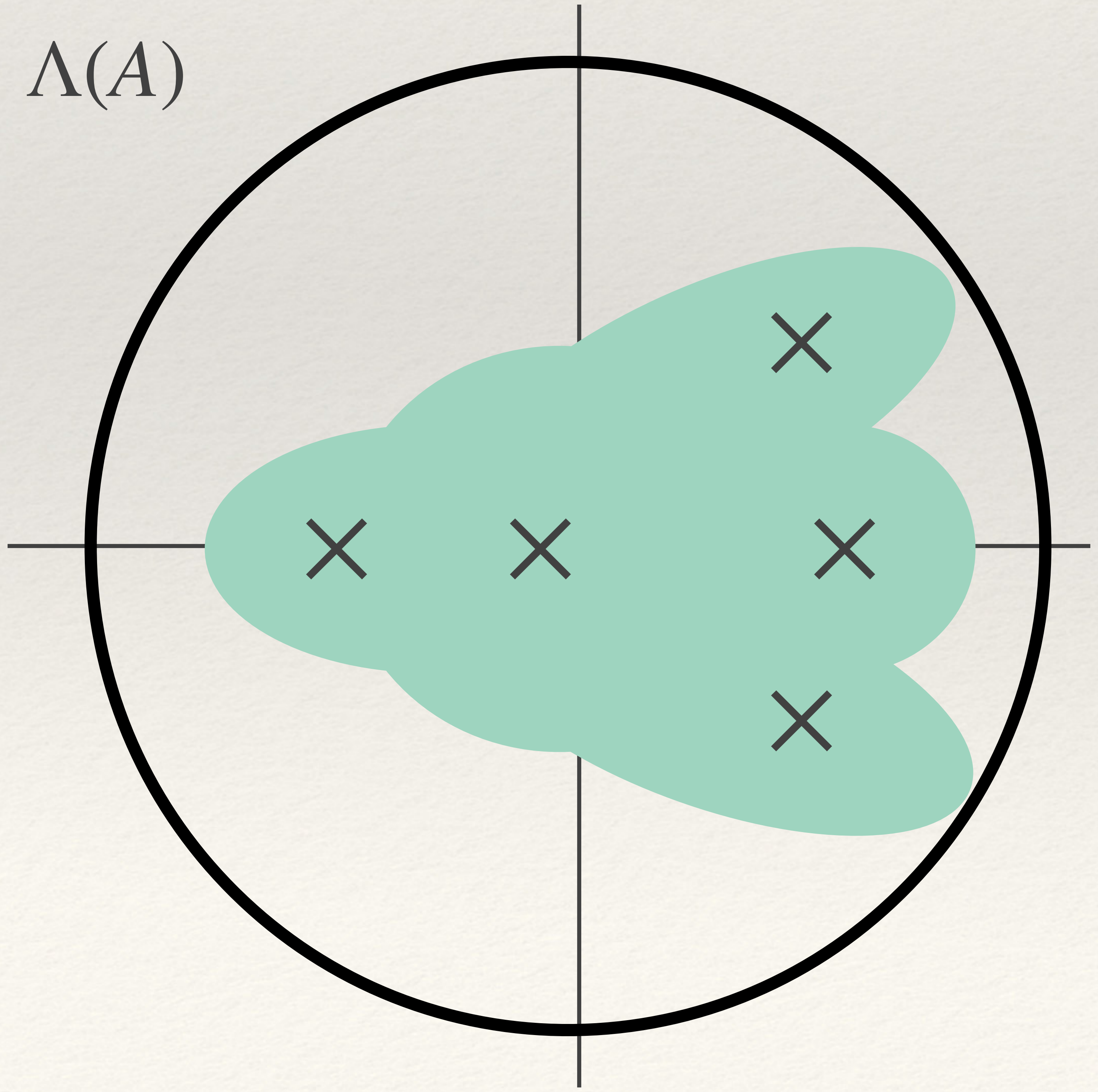
- normal iff $\Lambda_\varepsilon(A) = \Lambda(A) + \Delta_\varepsilon$
- non-normal iff $\Lambda_\varepsilon(A) \supsetneq \Lambda(A) + \Delta_\varepsilon$

A is normal iff $A = QDQ^*$ (unitary diagonalisable) iff $AA^* = A^*A$

Transient Behaviour of Asymptotically Stable LDS

- Pseudospectrum describes transient behaviour

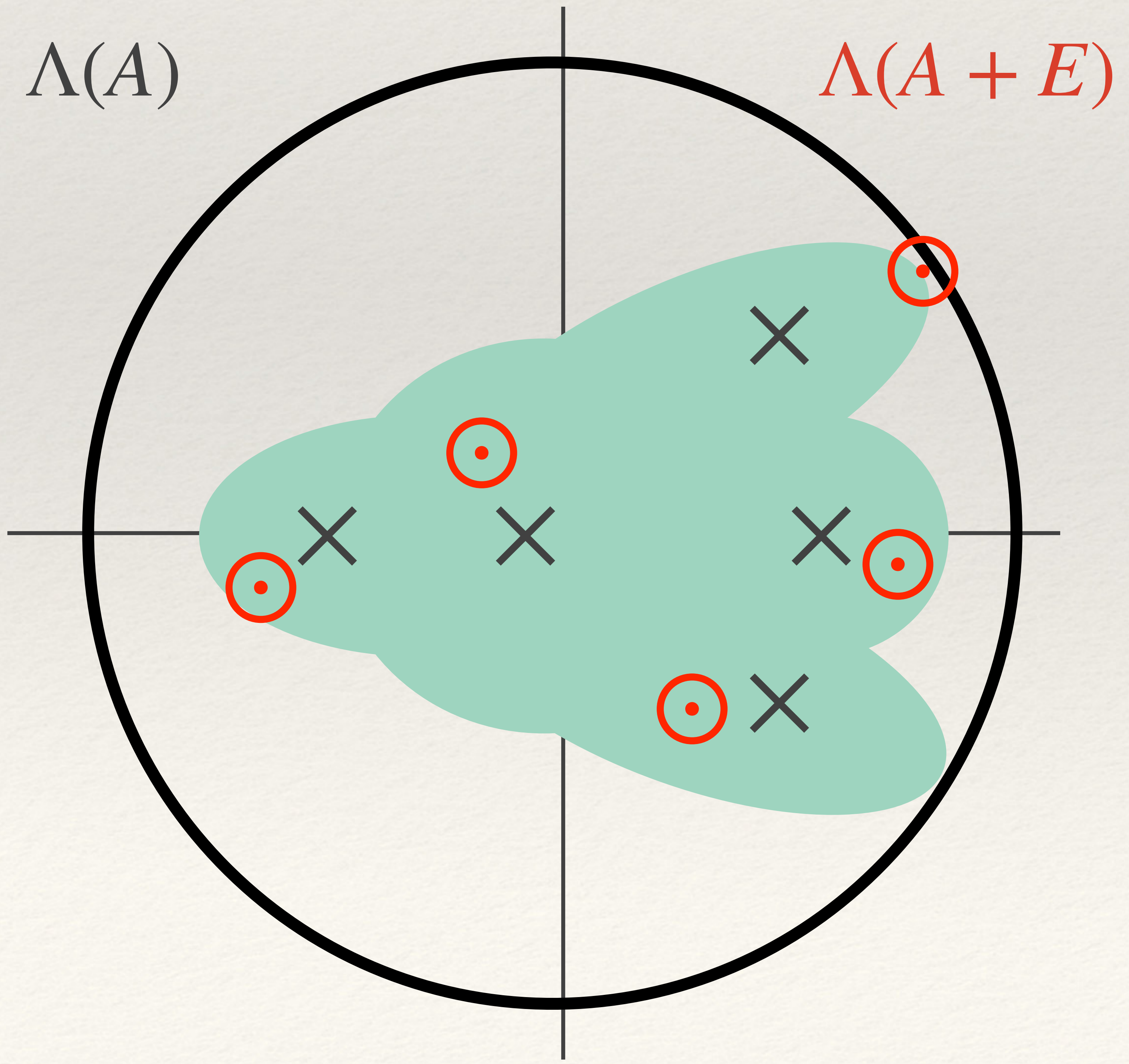
$$\boxed{\Lambda_\varepsilon(A) := \bigcup_{\|E\|\leq\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI - A)^{-1}\|^{-1} \leq \varepsilon\}}$$



Transient Behaviour of Asymptotically Stable LDS

- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\leq\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \|(zI - A)^{-1}\|^{-1} \leq \varepsilon\}$$



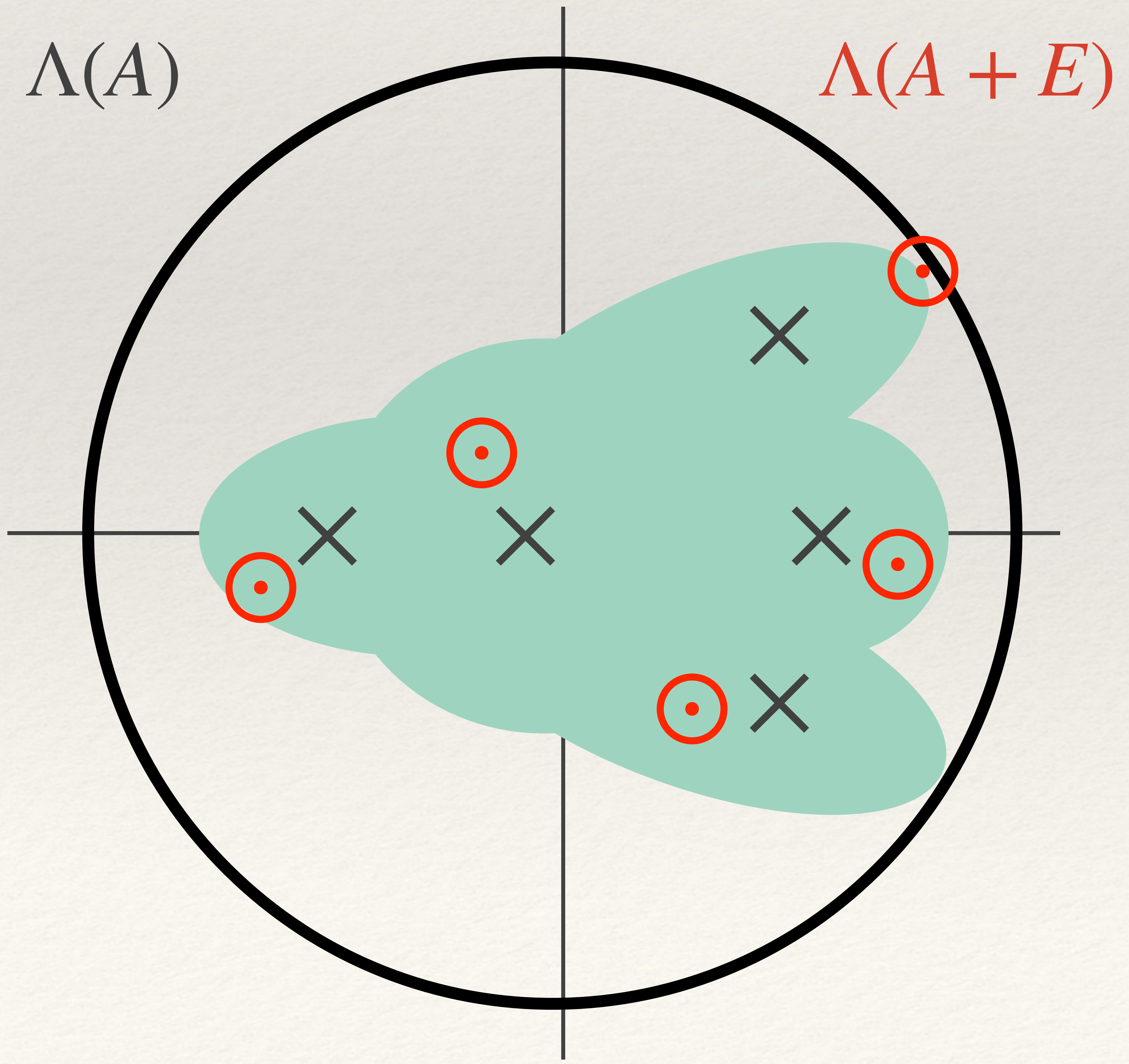
- pseudospectral radius

$$\rho_\varepsilon(A) := \{\, |\lambda| : \lambda \in \Lambda_\varepsilon(A)\}$$

Transient Behaviour of Asymptotically Stable LDS

- Pseudospectrum describes transient behaviour

$$\Lambda_\varepsilon(A) := \bigcup_{\|E\|\le\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \|(zI-A)^{-1}\|^{-1} \le \varepsilon\}$$



- pseudospectral radius

$$\rho_\varepsilon(A) := \{\, |\lambda| : \lambda \in \Lambda_\varepsilon(A)\}$$

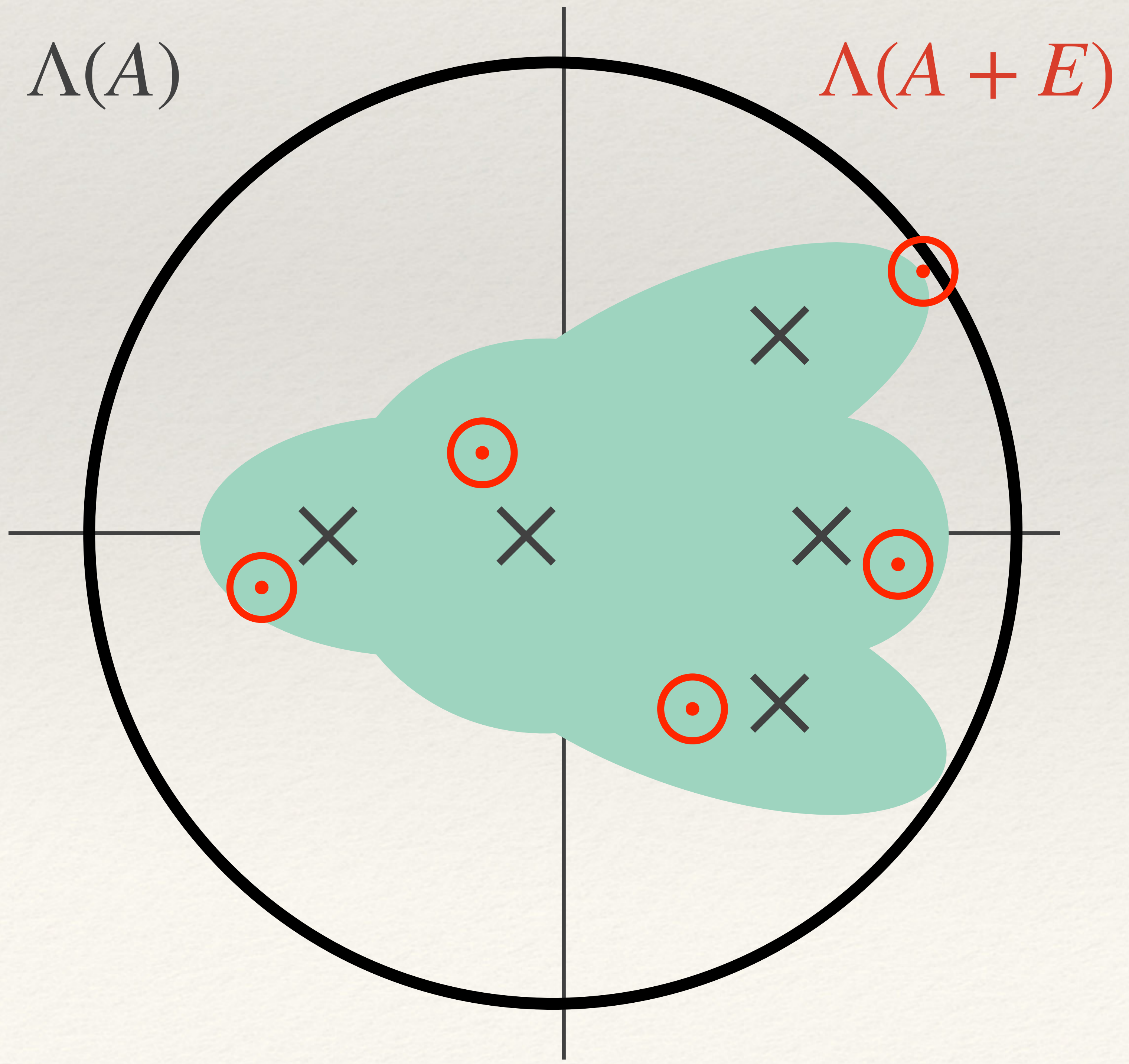
- distance to instability

$$d(A) := \sup_{\rho_\varepsilon(A)\le 1} \varepsilon = \inf_{z\,:\,|z|=1} \|(zI-A)^{-1}\|^{-1}$$

Transient Behaviour of Asymptotically Stable LDS

- Pseudospectrum describes transient behaviour

$$\boxed{\Lambda_\varepsilon(A) := \bigcup_{\|E\|\le\varepsilon} \Lambda(A+E) = \{z \in \mathbb{C} : \ \|(zI-A)^{-1}\|^{-1} \le \varepsilon\}}$$



- pseudospectral radius

$$\boxed{\rho_\varepsilon(A) := \{\,|\lambda| : \ \lambda \in \Lambda_\varepsilon(A)\}}$$

- distance to instability

$$\boxed{d(A) := \sup_{\rho_\varepsilon(A)\le 1} \varepsilon = \inf_{z\,:\,|z|=1} \|(zI-A)^{-1}\|^{-1}}$$

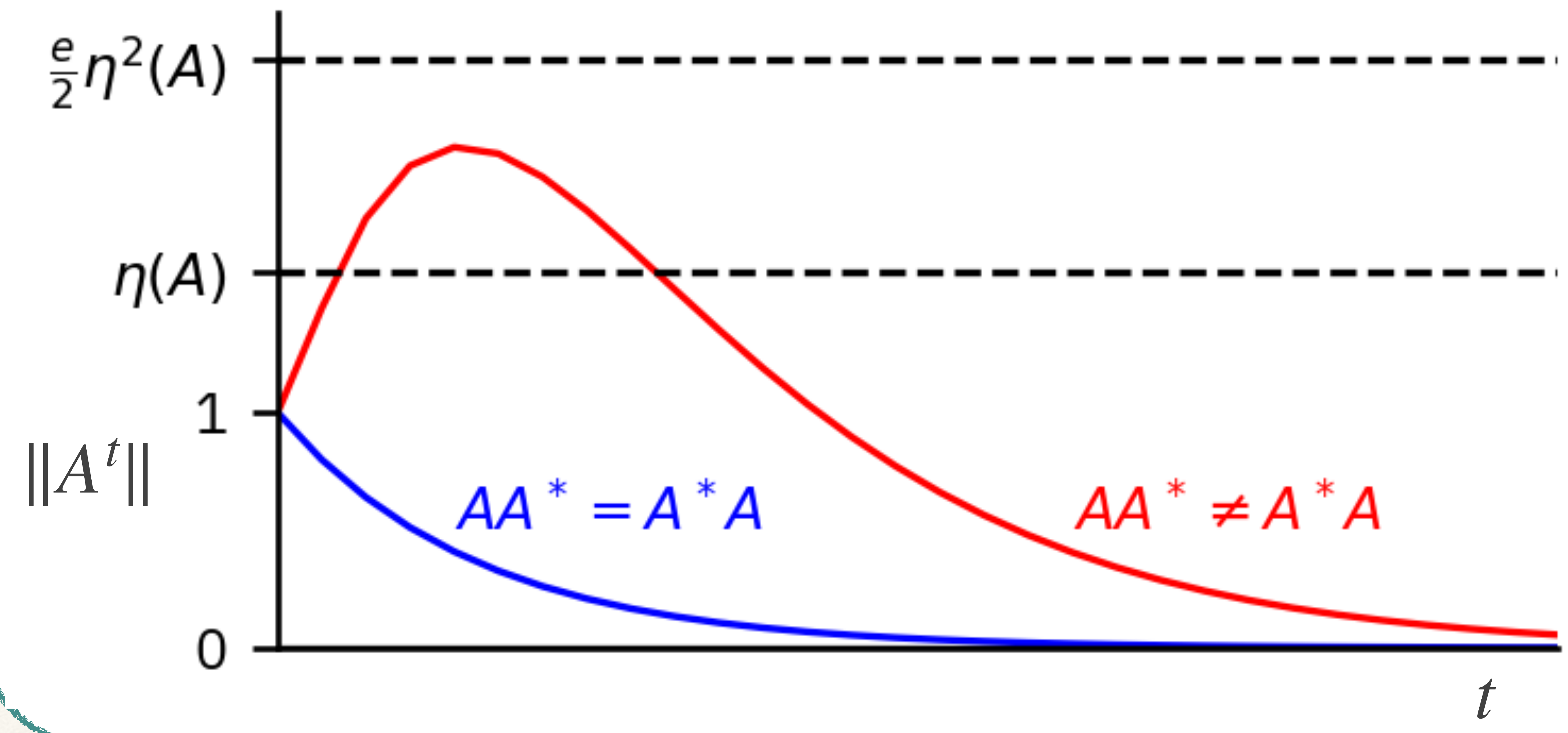
- Kreiss constant

$$\boxed{\eta(A) := \sup_{\varepsilon\,:\,\rho_\varepsilon>1} \frac{\rho_\varepsilon(A)-1}{\varepsilon} = \sup_{z\,:\,|z|>1} (|z|-1)\|(zI-A)^{-1}\|}$$

Transient Behaviour of Asymptotically Stable LDS

- Kreiss constant bounds transient behaviour:

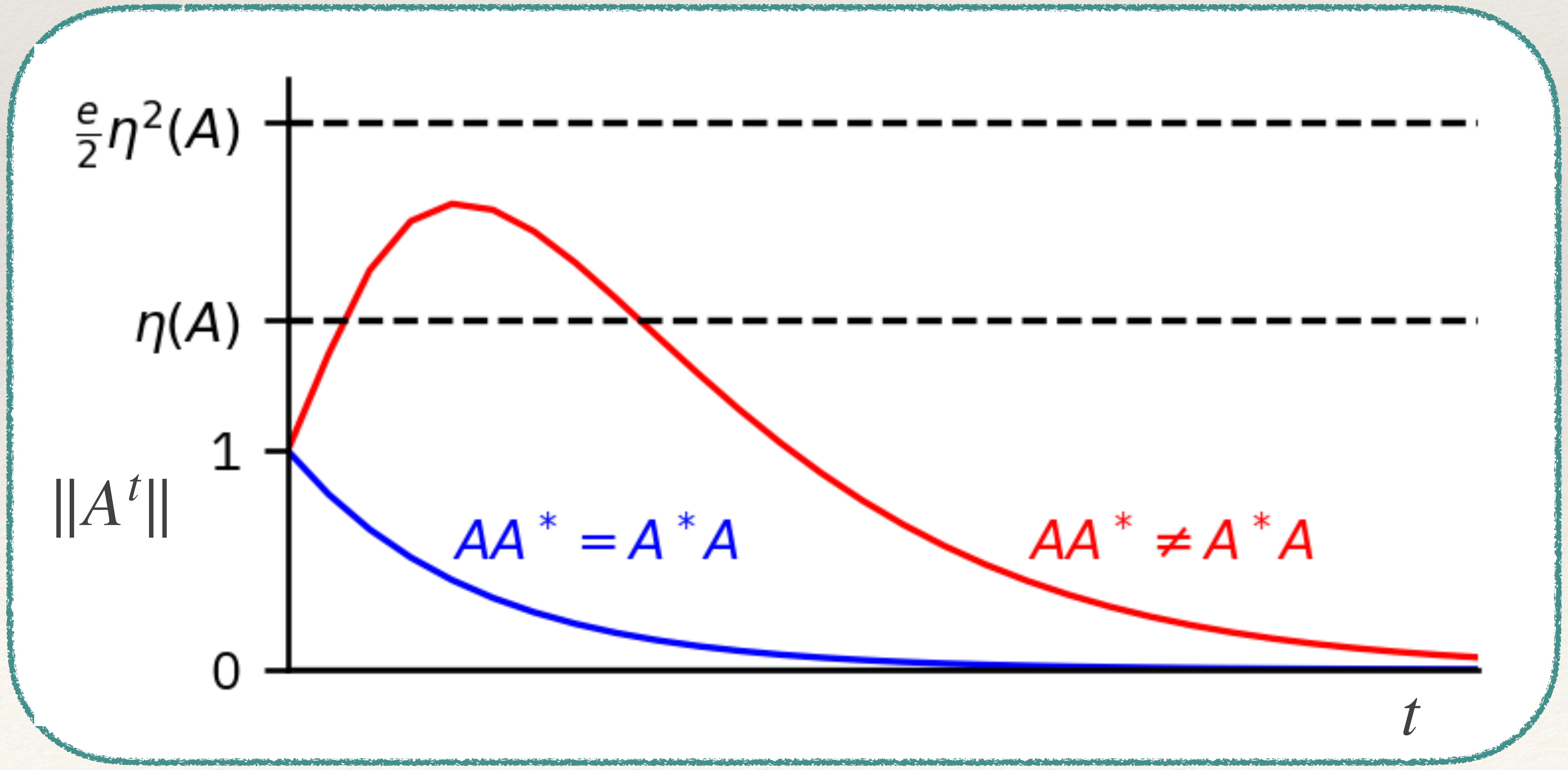
$$p(A) := \sup_{t\in\mathbb{N}_0} \|A^t\| \quad \Longrightarrow \quad \eta(A) \leq p(A) \leq (e/2)\,\eta^2(A)$$



Transient Behaviour of Asymptotically Stable LDS

- Kreiss constant bounds transient behaviour:

$$p(A) := \sup_{t\in\mathbb{N}_0} \|A^t\| \quad \Longrightarrow \quad \eta(A) \le p(A) \le (e/2)\,\eta^2(A)$$



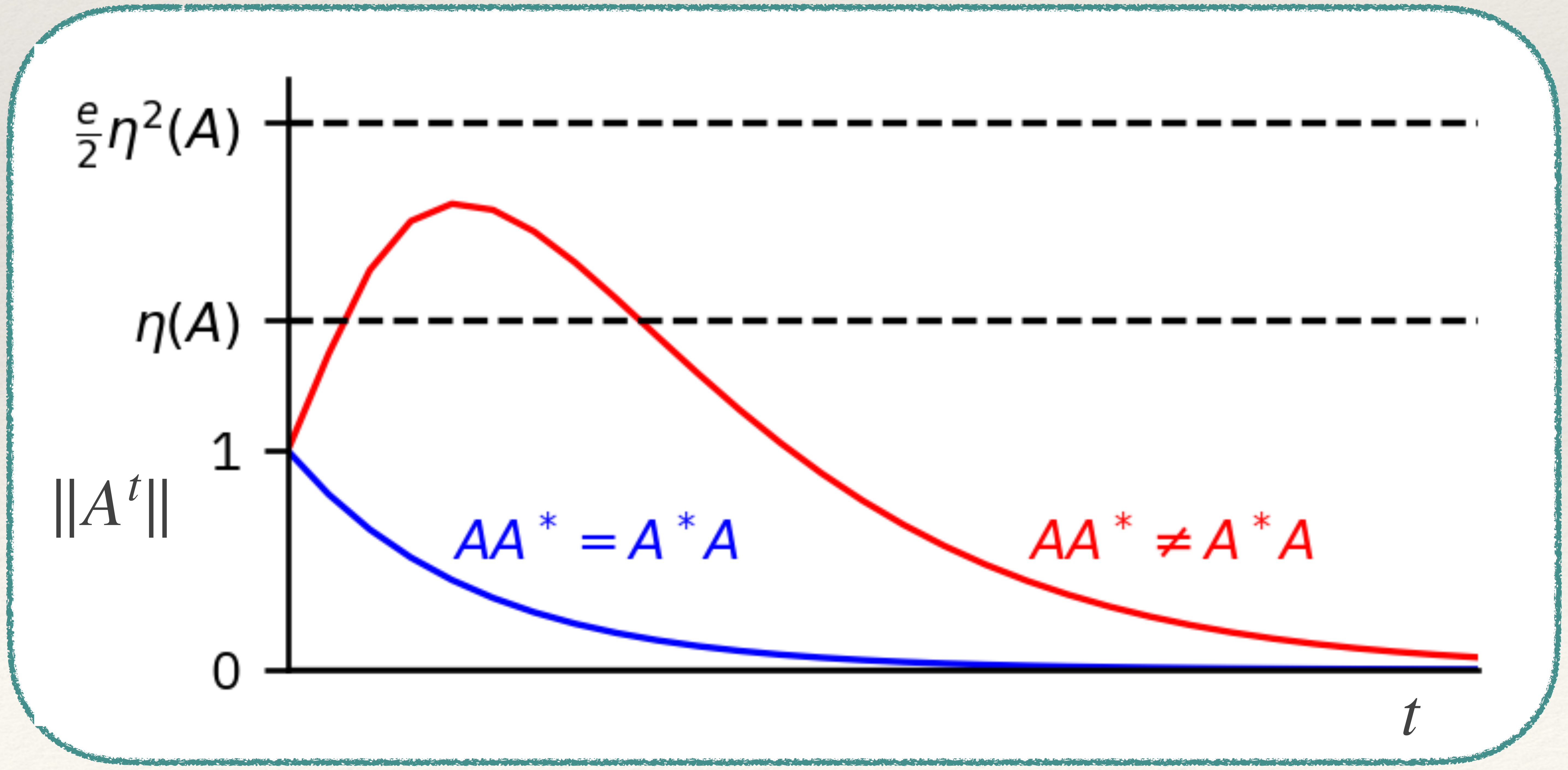
- Matrix case (Spijker's lemma)

$$\eta(A) \le p(A) \le ed\,\eta(A)$$

Transient Behaviour of Asymptotically Stable LDS

- Kreiss constant bounds transient behaviour:

$$\boxed{p(A) := \sup_{t\in\mathbb{N}_0}\|A^t\| \quad\Longrightarrow\quad \eta(A) \le p(A) \le (e/2)\,\eta^2(A)}$$



- Matrix case (Spijker's lemma)

$$\boxed{\eta(A) \le p(A) \le ed\,\eta(A)}$$

- Cumulative behaviour

$$\boxed{s(A) := \sum_{t\in\mathbb{N}_0}\|A^t\| < \infty}$$

Deflate-Learn-Inflate

Deflate-Learn-Inflate

- Recalling the transfer operator

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

Deflate-Learn-Inflate

- Recalling the transfer operator

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- Process is **geometrically ergodic** iff trivial leading eigeinvalue is **simple**

$$\boxed{A_\pi 1_\pi = A_\pi^* 1_\pi = 1_\pi \qquad \rho(A_\pi) = \|A_\pi\| = 1}$$

Deflate-Learn-Inflate

- Recalling the transfer operator

$$\boxed{A_\pi : L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X}) \quad (A_\pi f)(x) = \mathbb{E}[\, f(X_{t+1}) \mid X_t = x\,]}$$

- Process is **geometrically ergodic** iff trivial leading eigeinvalue is **simple**

$$\boxed{A_\pi 1_\pi = A_\pi^* 1_\pi = 1_\pi \qquad \rho(A_\pi) = \|A_\pi\| = 1}$$

- Remove (**deflate**) the trivial spectral component while keeping the rest untacked

$$\boxed{\mathbf{A}_\pi := A_\pi - 1_\pi \otimes 1_\pi \quad \Longrightarrow \quad \rho(\mathbf{A}_\pi) < 1 \quad \wedge \quad q_t - 1_\pi = \mathbf{A}_\pi^*(q_{t-1} - 1_\pi)}$$

Deflate-Learn-Inflate

Deflate-Learn-Inflate

- Learn deflated transfer operator

$$\boxed{\mathbf{A}_\pi f := \mathbb{E}[f(X_{t+1} \,|\, X_t = \cdot\,] - \mathbb{E}_{X\sim\pi} f(X)}$$

Deflate-Learn-Inflate

- Learn deflated transfer operator

$$\boxed{\mathbf{A}_\pi f := \mathbb{E}[f(X_{t+1} \mid X_t = \cdot\,] - \mathbb{E}_{X\sim\pi} f(X)}$$

- Notion of kernel mean embedding (KME)

$$\boxed{\mathbb{E}_{X\sim\mu}\,\phi(X) = \mathbb{E}_{X\sim\mu}\,k(X,\cdot\,) = k_\mu \qquad \forall h \in \mathscr{H}\ \langle k_\mu, h\rangle = \mathbb{E}_{X\sim\mu}\,h(X)}$$

Deflate-Learn-Inflate

- Learn deflated transfer operator

$$\boxed{\mathbf{A}_\pi f := \mathbb{E}[f(X_{t+1} \mid X_t = \cdot\,] - \mathbb{E}_{X\sim\pi} f(X)}$$

- Notion of kernel mean embedding (KME)

$$\boxed{\mathbb{E}_{X\sim\mu}\, \phi(X) = \mathbb{E}_{X\sim\mu}\, k(X, \cdot\,) = k_\mu \qquad \forall h \in \mathcal{H}\ \langle k_\mu, h\rangle = \mathbb{E}_{X\sim\mu}\, h(X)}$$

leads to approximating the flow in $\mathcal{H}$, i.e. for all $h \in \mathcal{H}$

Deflate-Learn-Inflate

- Learn deflated transfer operator

$$\boxed{\mathbf{A}_\pi f := \mathbb{E}[f(X_{t+1} \,|\, X_t = \cdot\,] - \mathbb{E}_{X\sim\pi} f(X)}$$

- Notion of kernel mean embedding (KME)

$$\boxed{\mathbb{E}_{X\sim\mu}\,\phi(X) = \mathbb{E}_{X\sim\mu}\,k(X,\cdot\,) = k_\mu \qquad \forall h \in \mathscr{H}\;\; \langle k_\mu, h\rangle = \mathbb{E}_{X\sim\mu}\,h(X)}$$

leads to approximating the flow in $\mathscr{H}$, i.e. for all $h \in \mathscr{H}$

$$\langle k_{\mu_t}, h\rangle = \mathbb{E}[h(X_t)] = \mathbb{E}[\mathbb{E}[h(X_t) \,|\, X_{t-1}]] \approx \langle k_{\mu_{t-1}}, Gh\rangle = \langle G^* k_{\mu_{t-1}}, h\rangle$$

Deflate-Learn-Inflate

- Learn deflated transfer operator

$$\boxed{\mathbf{A}_\pi f := \mathbb{E}[f(X_{t+1}\,|\,X_t = \cdot\,] - \mathbb{E}_{X\sim\pi} f(X)}$$

- Notion of kernel mean embedding (KME)

$$\boxed{\mathbb{E}_{X\sim\mu}\,\phi(X) = \mathbb{E}_{X\sim\mu}\,k(X,\cdot\,) = k_\mu \qquad \forall h \in \mathscr{H}\;\; \langle k_\mu, h\rangle = \mathbb{E}_{X\sim\mu}\,h(X)}$$

leads to approximating the flow in $\mathscr{H}$, i.e. for all $h \in \mathscr{H}$

$$\langle k_{\mu_t}, h\rangle = \mathbb{E}[h(X_t)] = \mathbb{E}[\mathbb{E}[h(X_t)\,|\,X_{t-1}]] \approx \langle k_{\mu_{t-1}}, Gh\rangle = \langle G^* k_{\mu_{t-1}}, h\rangle$$

deflation results in

$$\boxed{\mathscr{R}(G) = \mathbb{E}_{X_t\sim\pi}\,\|[\phi(X_{t+1}) - k_\pi] - G^*[\phi(X_t) - k_\pi]\|^2}$$

features centering

Deflate-Learn-Inflate

- Given a sample $(x_i, y_i := x_{i+1})_{i=1}^n$ we learn $\hat{G}: \mathscr{H} \to \mathscr{H}$ via the empirical risk:

$$\boxed{\hat{\mathscr{R}}(\hat{G}) = \frac{1}{n}\sum_{i=1}^n \|[\phi(y_i) - k_{\hat{\pi}_y}] - \hat{G}^*[\phi(x_i) - k_{\hat{\pi}_x}]\|^2 + \gamma\|\hat{G}\|_{\mathrm{HS}}^2}$$

Deflate-Learn-Inflate

- Given a sample $(x_i, y_i := x_{i+1})_{i=1}^n$ we learn $\hat{G}\colon \mathscr{H} \to \mathscr{H}$ via the empirical risk:

$$\boxed{\hat{\mathscr{R}}(\hat{G}) = \frac{1}{n}\sum_{i=1}^n \|[\phi(y_i) - k_{\hat{\pi}_y}] - \hat{G}^*[\phi(x_i) - k_{\hat{\pi}_x}]\|^2 + \gamma\|\hat{G}\|_{\mathrm{HS}}^2}$$

- Typical estimators:
  - Kernel ridge minimizes the regularized empirical risk $\boxed{\hat{G} = (\hat{\mathrm{C}} + \gamma I)^{-1}\hat{\mathrm{T}}}$
  - RRR minimizes the empirical risk with a rank constraint $\boxed{\hat{G} = \hat{\mathrm{C}}_\gamma^{-1/2}[\![\hat{\mathrm{C}}_\gamma^{-1/2}\hat{\mathrm{T}}]\!]_r}$

Deflate-Learn-Inflate

- Given a sample $(x_i, y_i := x_{i+1})_{i=1}^n$ we learn $\hat{G}: \mathscr{H} \to \mathscr{H}$ via the empirical risk:

$$\boxed{\hat{\mathscr{R}}(\hat{G}) = \frac{1}{n}\sum_{i=1}^n \|[\phi(y_i) - k_{\hat{\pi}_y}] - \hat{G}^*[\phi(x_i) - k_{\hat{\pi}_x}]\|^2 + \gamma\|\hat{G}\|_{\mathrm{HS}}^2}$$

- Typical estimators:
  - Kernel ridge minimizes the regularized empirical risk $\boxed{\hat{G} = (\hat{\mathbf{C}} + \gamma I)^{-1}\hat{\mathbf{T}}}$
  - RRR minimizes the empirical risk with a rank constraint $\boxed{\hat{G} = \hat{\mathbf{C}}_\gamma^{-1/2}[\![\hat{\mathbf{C}}_\gamma^{-1/2}\hat{\mathbf{T}}]\!]_r}$

$$\boxed{\mathbf{C} = \frac{1}{n}\sum_{i=1}^n [\phi(x_i) - k_{\hat{\pi}_x}] \otimes [\phi(x_i) - k_{\hat{\pi}_x}] \quad \mathbf{T} = \frac{1}{n}\sum_{i=1}^n [\phi(x_i) - k_{\hat{\pi}_x}] \otimes [\phi(y_i) - k_{\hat{\pi}_y}]}$$

Deflate-Learn-Inflate

- Recalling the flow $q_t - 1_\pi = (\mathbf{A}_\pi^*)^t (q_0 - 1_\pi)$

Deflate-Learn-Inflate

- Recalling the flow $q_t - 1_\pi = (\mathbf{A}_\pi^*)^t (q_0 - 1_\pi)$

$$\boxed{\langle k_{\mu_t} - k_\pi, h \rangle_{\mathscr{H}} = \langle q_0 - 1_\pi, \mathbf{A}_\pi^t h \rangle_{L_\pi^2} \approx \langle k_{\hat{\mu}_0} - k_{\hat{\pi}_x}, \hat{G}^t h \rangle_{\mathscr{H}}}$$

Deflate-Learn-Inflate

- Recalling the flow $q_t - 1_\pi = (\mathbf{A}_\pi^*)^t (q_0 - 1_\pi)$

$$\boxed{\langle k_{\mu_t} - k_\pi, h \rangle_{\mathscr{H}} = \langle q_0 - 1_\pi, \mathbf{A}_\pi^t h \rangle_{L^2_\pi} \approx \langle k_{\hat{\mu}_0} - k_{\hat{\pi}_x}, \hat{G}^t h \rangle_{\mathscr{H}}}$$

- we inject removed eigen-triple

$$\boxed{k_{\hat{\mu}_t} = k_{\hat{\pi}_y} + [\hat{G}^*]^t [k_{\hat{\mu}_0} - k_{\hat{\pi}_x}]}$$

Deflate-Learn-Inflate

- Recalling the flow $q_t - 1_\pi = (\mathbf{A}_\pi^*)^t (q_0 - 1_\pi)$

$$\boxed{\langle k_{\mu_t} - k_\pi, h \rangle_{\mathscr{H}} = \langle q_0 - 1_\pi, \mathbf{A}_\pi^t h \rangle_{L^2_\pi} \approx \langle k_{\hat{\mu}_0} - k_{\hat{\pi}_x}, \hat{G}^t h \rangle_{\mathscr{H}}}$$

- we inject removed eigen-triple

$$\boxed{k_{\hat{\mu}_t} = k_{\hat{\pi}_y} + [\hat{G}^*]^t [k_{\hat{\mu}_0} - k_{\hat{\pi}_x}]}$$

- and incur the multi-step-ahead error

$$\boxed{\mathscr{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathscr{H} \to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathscr{H} \to L^2_\pi}}$$

Uniform bounds for MMD norm

$$\boxed{\mathcal{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathcal{H}\to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathcal{H}\to L^2_\pi}}$$

Uniform bounds for MMD norm

$$\mathcal{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathcal{H}\to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathcal{H}\to L^2_\pi}$$

- for the choice of universal kernels, we analyse one step ahead error with vector-valued regression analysis

Relationships $\mathcal{H} \sim \mathbf{A}_\pi$ and $\mathcal{H} \sim L^2_\pi(\mathcal{X})$ are captured by $\alpha \in [1,2]$ and $\beta \in [0,1]$ we have

$$\varepsilon_n = n_{eff}^{-\frac{\alpha}{2(\alpha+\beta)}}$$

Uniform bounds for MMD norm

$$\mathcal{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathcal{H}\to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathcal{H}\to L^2_\pi}$$

- for the choice of universal kernels, we analyse one step ahead error with vector-valued regression analysis

Relationships $\mathcal{H} \sim \mathbf{A}_\pi$ and $\mathcal{H} \sim L^2_\pi(\mathcal{X})$ are captured by $\alpha \in [1,2]$ and $\beta \in [0,1]$ we have

$$\varepsilon_n = n_{eff}^{-\frac{\alpha}{2(\alpha+\beta)}}$$

With probability at least $1-\delta$ in the observed training data the estimation error is bounded by

$$\mathcal{E}(\hat{G}) \lesssim \varepsilon_n \ln(\delta^{-1})$$

Uniform bounds for MMD norm

$$\mathcal{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathcal{H}\to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathcal{H}\to L^2_\pi}$$

- for the choice of universal kernels, we analyse one step ahead error with vector-valued regression analysis
- using perturbation bounds we concentrate the Kreiss constant of the estimator

Relationships $\mathcal{H} \sim \mathbf{A}_\pi$ and $\mathcal{H} \sim L^2_\pi(\mathcal{X})$ are captured by $\alpha \in [1,2]$ and $\beta \in [0,1]$ we have

$$\varepsilon_n = n_{eff}^{-\frac{\alpha}{2(\alpha+\beta)}}$$

With probability at least $1-\delta$ in the observed training data the estimation error is bounded by

$$\mathcal{E}(\hat{G}) \lesssim \varepsilon_n \ln(\delta^{-1})$$

Uniform bounds for MMD norm

$$\mathscr{E}_t(\hat{G}) := \|\mathbf{A}_\pi^t - \hat{G}^t\|_{\mathscr{H}\to L^2_\pi} \leq s(\mathbf{A}_\pi)\, p(\hat{G})\, \|\mathbf{A}_\pi - \hat{G}\|_{\mathscr{H}\to L^2_\pi}$$

- for the choice of universal kernels, we analyse one step ahead error with vector-valued regression analysis
- using perturbation bounds we concentrate the Kreiss constant of the estimator
- we additionally concentrate KMEs and obtain maximum mean discrepancy (MMD) error bound

$$\left\| \mu_t - \hat{\mu}_t \right\|_{\mathscr{H}^*} = \left\| k_{\mu_t} - k_{\hat{\mu}_t} \right\|_{\mathscr{H}} \leq C \frac{\log \delta^{-1}}{n_0 \wedge n_{eff}^{\frac{\alpha}{2(\alpha+\beta)}}}$$

Relationships $\mathscr{H} \sim \mathbf{A}_\pi$ and $\mathscr{H} \sim L^2_\pi(\mathscr{X})$ are captured by $\alpha \in [1,2]$ and $\beta \in [0,1]$ we have

$$\varepsilon_n = n_{eff}^{-\frac{\alpha}{2(\alpha+\beta)}}$$

With probability at least $1-\delta$ in the observed training data the estimation error is bounded by

$$\mathscr{E}(\hat{G}) \lesssim \varepsilon_n \ln(\delta^{-1})$$

Empirical results

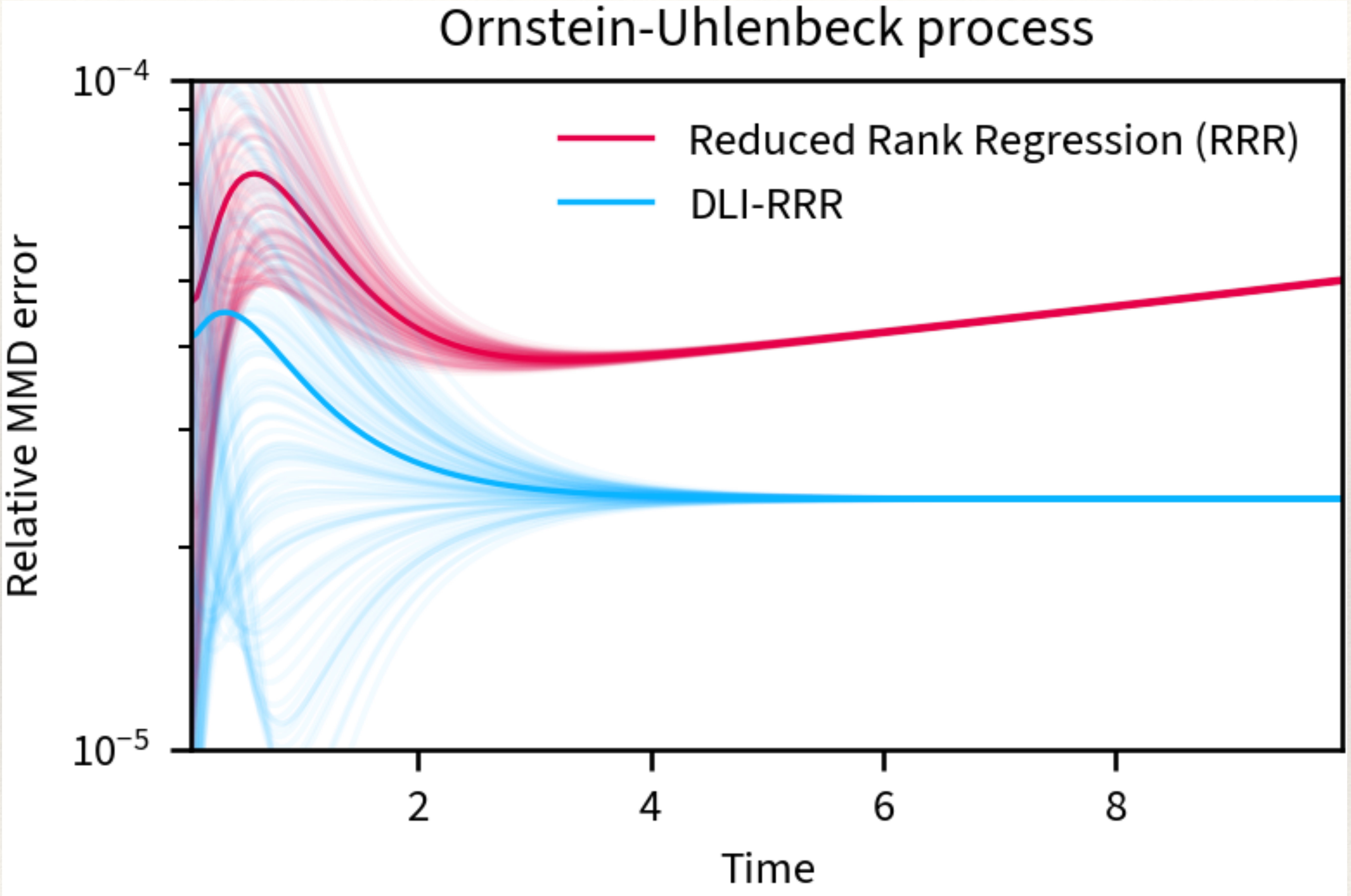
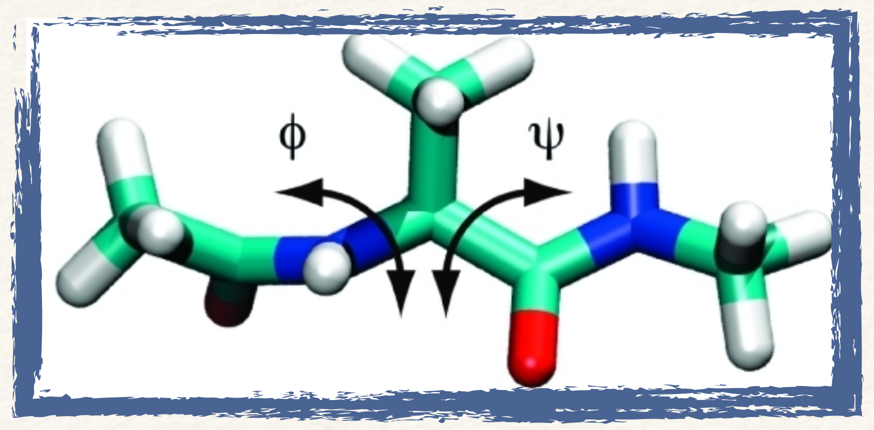


Figure 3. Distribution forecasting: Relative MMD error for the OU process for 100 independent experiments (thin lines).

Observable	RRR	**DLI-RRR**
$\mathbb{E}\left[r_t \mid r_0 = \cdot\right]$	0.0691 ± 0.0333	$\mathbf{0.0673 \pm 0.0328}$
$\mathbb{V}\left[r_t \mid r_0 = \cdot\right]$	0.0470 ± 0.0413	$\mathbf{0.0124 \pm 0.0051}$

Table 1. RMSE in estimating conditional expectation and variance of the CIR model (100 independent training datasets).

Empirical results



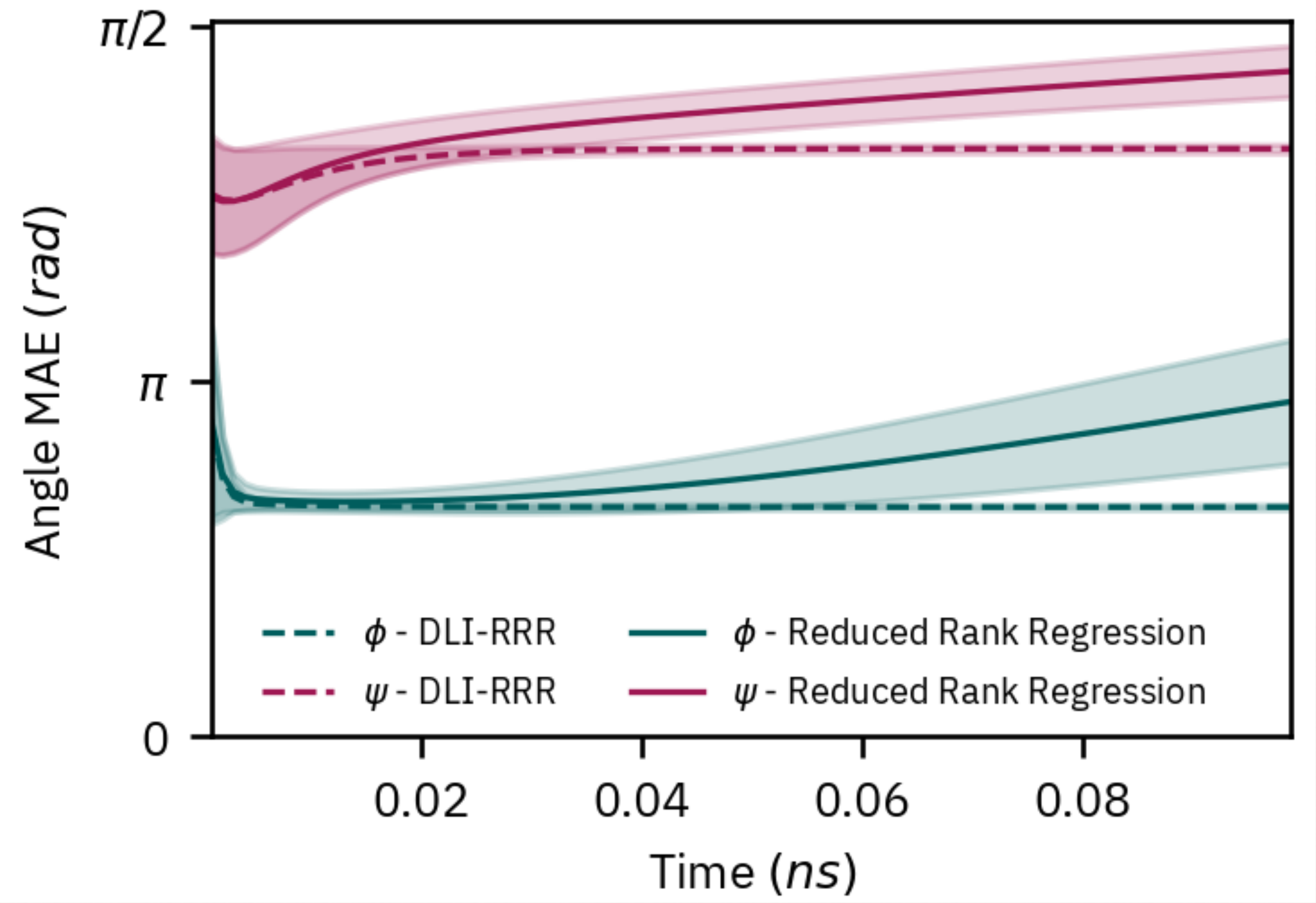


Figure 1. Mean Absolute Error (MAE) in forecasting the backbone dihedral angles of Alanine Dipeptide. Data points are 10^{-3} ns apart.

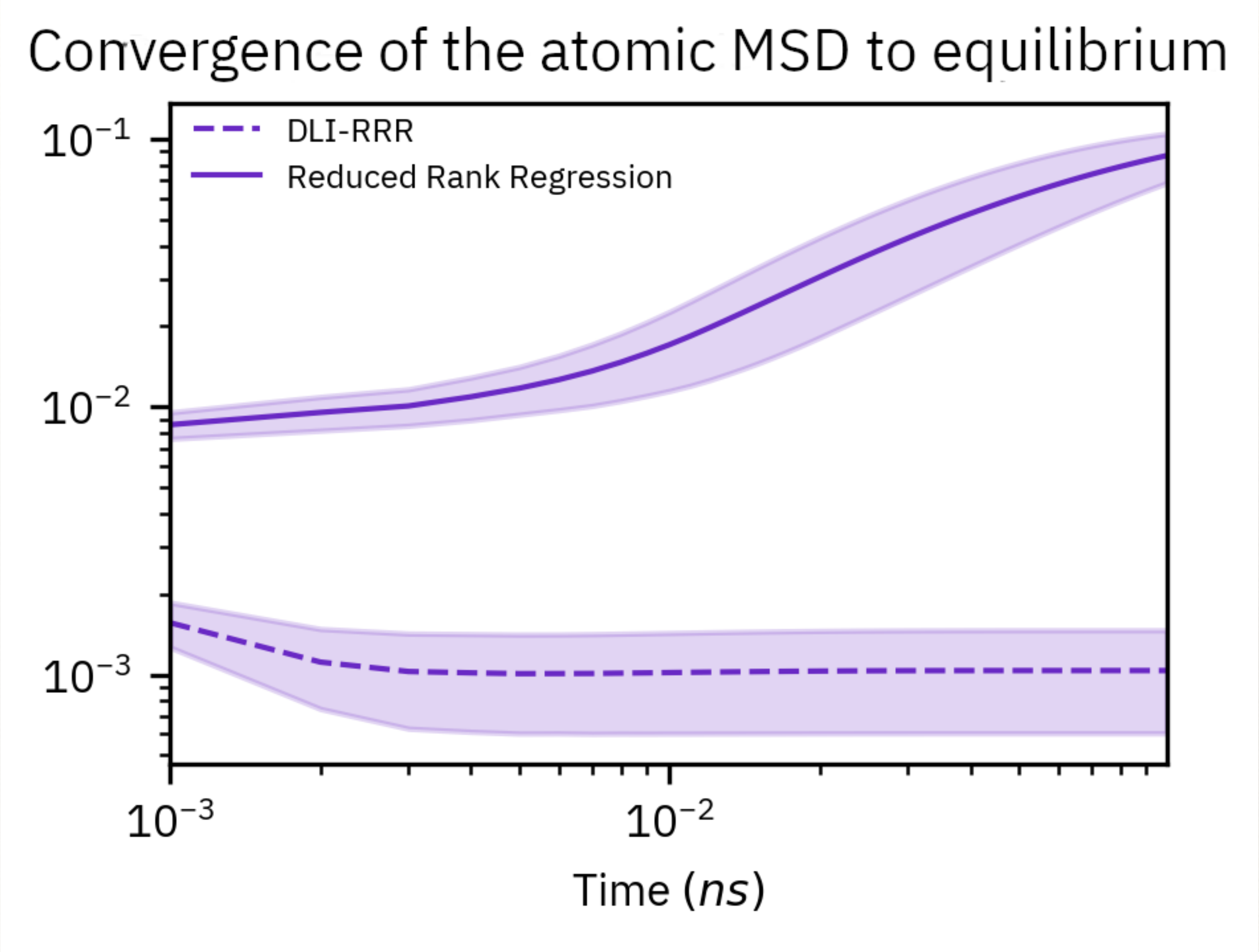


Figure 4. Forecasting the Mean Square Deviation (MSD) of atomic positions in Alanine Dipeptide. Notice the log-log scale.

References and Code

- V. Kostic, P. Novelli, A. Maurer, C. Ciliberto, L. Rosasco, M. Pontil. Learning dynamical systems via Koopman operator regression in reproducing kernel hilbert spaces. NeurIPS 2022.
- V. Kostic, K. Lounici, P. Novelli, M. Pontil. Koopman operator learning: sharp spectral rates and spurious eigenvalues. NeurIPS 2023.
- G. Meanti, A. Chatalic, V. Kostic, P. Novelli, M. Pontil, L. Rosasco. Estimating Koopman operators with sketching to provably learn large scale dynamical systems. NeurIPS 2023.
- V. Kostic, P. Novelli, R. Grazzi, K. Lounici, M. Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. ICLR 2024.
- V. Kostic, K. Lounici, P. Inzerilli, P. Novelli., M. Pontil. Consistent long-term forecasting of ergodic dynamical systems. ICML 2024.
- G. Turri, V. Kostic, P. Novelli, M. Pontil. A randomized algorithm to solve reduced rank operator regression. Submitted 2024.
- K. Lounici, V Kostic, G. Pacreau, G. Turri, P. Novelli, M. Pontil Neural Conditional Probability for Statistical Inference, Submitted 2024.
- V. Kostic, K. Lounici, H. Halconruy, T. Devergne, M. Pontil. Learning the infinitesimal generator of stochastic diffusion processes, Submitted 2024
- T. Devergne, V. Kostic, M. Parrinello, M. Pontil. From biased to unbiased dynamics: an infinitesimal generator approach. Submitted 2024

Code: https://github.com/Machine-Learning-Dynamical-Systems/kooplearn

THANK YOU!