Stochastic system consideration for EDMD

Yuanchao Xu

October 18, 2024

1 Notation introduction

A list of all the notation and relevant definition:

1. (Autonomous) SDE in the dynamical system (Ω, μ):

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t$$

2. Φ^t: Transformation map/flow

3. $\Psi = \{\psi_i\}_{i=1}^N$: Dictionary or set of basis functions
   Each $\psi_i \in \mathcal{F} := L^2(\Omega, \mu)$

4. $\mathcal{P}_N$: Projection onto the subspace $\mathcal{F}_N := \operatorname{span}\{\psi_1, \dots, \psi_N\} \subset \mathcal{F}$

5. $\mathcal{P}_{\mathcal{D}_N}$: Projection of $\mathcal{D}$ onto $\mathcal{F}_N$ using the inner product on $\mathcal{D}$

6. m: Number of (i.i.d.) data points

7. N: Dictionary size

8. $n, t = \frac{1}{n}$: sampling frequency and time step between two data snapshots

9. Data(i.i.d.) snapshots:

$$X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_m \\ | & & | \end{bmatrix}, \quad Y = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_m \\ | & & | \end{bmatrix}$$

10. (Row)Vector of basis functions:

$$\boldsymbol{\Psi}_N(x) := [\psi_1(x), \dots, \psi_N(x)]$$

11. Matrices of dictionary $\{\psi_j\}_{j=1}^N$ evaluated on data $\{x_i, y_i\}_{i=1}^m$:

$$\Psi_X = \begin{bmatrix} \rule{1em}{0.4pt} & \boldsymbol{\Psi}_N(x_1) & \rule{1em}{0.4pt} \\ & \vdots & \\ \rule{1em}{0.4pt} & \boldsymbol{\Psi}_N(x_m) & \rule{1em}{0.4pt} \end{bmatrix} = \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_N(x_1) \\ \vdots & \ddots & \vdots \\ \psi_1(x_m) & \cdots & \psi_N(x_m) \end{bmatrix}$$

$$\Psi_Y = \begin{bmatrix} \rule[0.5ex]{1em}{0.4pt} & \mathbf{\Psi}_N(y_1) & \rule[0.5ex]{1em}{0.4pt} \\ & \vdots & \\ \rule[0.5ex]{1em}{0.4pt} & \mathbf{\Psi}_N(y_m) & \rule[0.5ex]{1em}{0.4pt} \end{bmatrix} = \begin{bmatrix} \psi_1(y_1) & \cdots & \psi_N(y_1) \\ \vdots & \ddots & \vdots \\ \psi_1(y_m) & \cdots & \psi_N(y_m) \end{bmatrix}$$

12. Definition of Koopman operator in deterministic system:

$$\mathcal{K}f = f \circ \Phi$$

Definition of stochastic Koopman operator(**semigroup**):

$$\mathcal{K}^t f = \mathbb{E}[f \circ \Phi^t]$$

Remark: In the deterministic system, we can still use $\mathcal{K}^t, \Phi^t$ where t is the (fixed) time interval between two sampling snapshots.

13. Definition of Koopman generator:

$$\begin{aligned} \mathcal{L}f(x) &:= \lim_{t \to 0} \frac{[\mathcal{K}^t f(x)] - f(x)}{t} \\ &= \lim_{t \to 0} \frac{\mathbb{E}[f \circ \Phi^t(x)] - f(x)}{t} \end{aligned} \tag{1}$$

EDMD framework review

Let $f(x) = \mathbf{\Psi}_N(x)a$. Then

$$\mathcal{K}f(x) = f(y) = \mathbf{\Psi}_N(y)a = \mathbf{\Psi}_N(x)Ka + r(x)$$

where $r(x)$ is the residual. Notice that the span of $\{\psi_1, \ldots, \psi_N\}$ is not necessarily invariant under $\mathcal{K}$.

In the EDMD framework, given a dataset $\{(x_i, y_i)\}_{i=1}^M$, we want to minimize the residual over the finite data set:

$$J := \sum_{i=1}^M |r(x_i)|^2 = \sum_{i=1}^M \left| (\mathbf{\Psi}_N(y_i) - \mathbf{\Psi}_N(x_i)K)\, \mathbf{a} \right|^2,$$

so, this is equivalent to $\min_K \|\Psi_Y - \Psi_X K\|_F$ and the minimal K is

$$K = \Psi_X^\dagger \Psi_Y = (\Psi_X^T \Psi_X)^\dagger (\Psi_X^T \Psi_Y) = G^\dagger A$$

where $\dagger$ is the pseudoinverse and $G = \frac{1}{m}\Psi_X^T \Psi_X, A = \frac{1}{m}\Psi_X^T \Psi_Y$.

Remark. *Regularization through truncated SVD or by adding a small perturbation is typically applied.*

S-EDMD framework

First, we can expand the Koopman operator(semigroup) $\mathcal{K}(t)$ similar as in EDMD:

$$\begin{aligned} \mathcal{K}(t)f(x) &= \mathbb{E}[f(y)] \\ &= \mathbf{\Psi}_N(x)K(t)\mathbf{a} + r(x) \end{aligned} \tag{2}$$

Alternatively, we approximate the Koopman operator(semigroup) $\mathcal{K}(t)$ analogous to Taylor expansion using the definition of generator(1):

$$\begin{aligned} \mathcal{K}(t)f(x) &= \mathbb{E}[f(y)] \\ &= \mathbb{E}[\mathbf{\Psi}_N(y)]\mathbf{a} \\ &= \left[\mathbf{\Psi}_N(x) + t \cdot \mathcal{L}\mathbf{\Psi}_N(x) + O(t^2)\right]\mathbf{a} \end{aligned} \tag{3}$$

where $\mathcal{L}\mathbf{\Psi}_N(x) = [\mathcal{L}\psi_1(x), \dots, \mathcal{L}\psi_N(x)]$ and

$$\mathcal{L}\psi_j(x) = b(x) \cdot \nabla\psi_j(x) + \frac{1}{2}\sigma^2(x) \colon \nabla^2\psi_j(x) \tag{4}$$

Remark. *We can get explicit higher order term $O(t^2)$ in (3) by applying Itô's formula recursively. See A for more details.*

Remark. *In order to obtain (4), we can use Itô's formula. From Itô's formula, we know that given $f \in C_b^2(\Omega)$ and $X_0 = x$,*

$$df(X_t) = \left(b(X_t) \cdot \nabla f(X_t) + \frac{1}{2}\sigma^2(x) \colon \nabla^2 f(X_t)\right)dt + \sigma(X_t) \cdot \nabla f(X_t)dW_t$$

Taking expectation on its integral form, we have:

$$\mathbb{E}[f(X_t)|X_0 = x] - f(x) = \int_0^t \left(b(X_s) \cdot \nabla f(X_s) + \frac{1}{2}\sigma^2(X_s) \colon \nabla^2 f(X_s)\right)ds$$

Let $t \to 0$, we have

$$\begin{aligned} \frac{d}{dt}f = \mathcal{L}f(x) &\coloneqq \lim_{t \to 0} \frac{\mathbb{E}[f(X_t) - f(x)|X_0 = x]}{t} \\ &= b(X_t) \cdot \nabla f(X_t) + \frac{1}{2}\sigma^2(x) \colon \nabla^2 f(X_t) \end{aligned}$$

In our case, combning (3) and (2), we can minimize the following:

$$J \coloneqq \sum_{i=1}^m |r(x_i)|^2 = \sum_{i=1}^m \left|\left[\mathbf{\Psi}_N(x_i) + t \cdot \mathcal{L}\mathbf{\Psi}_N(x_i) + O(t^2) - \mathbf{\Psi}_N(x_i)K\right]\mathbf{a}\right|^2$$

which is equivalent to minimize

$$\boxed{J = \min_K \|\Psi_X + t \cdot \mathcal{L}\Psi_X + O(t^2) - \Psi_X K\|_F^2} \tag{5}$$

and the minimal $K(t)$ is

$$K(t) = \Psi_X^\dagger (\Psi_X + t \cdot \mathcal{L}\Psi_X + O(t^2)) \tag{6}$$

where the generator(time differential) matrix approximation $[\mathcal{L}\Psi_X]_{ij} = \mathcal{L}\psi_j(x_i)$.

Remark. *$O(t^2)$ in (7), (3) and (5) are different. They are scalar, vector and matrix respectively. For example, each element of $O(t^2)$ in (3) and (5) is a scalar $O(t^2)$ in (7).*

Empirically, if t is very small, we update the Koopman matrix approximation in the following way with $O(t^2)$ omitted:

$$\widehat{K}(t) = \widehat{\Psi}_X^\dagger (\widehat{\Psi}_X + t \cdot \widehat{\mathcal{L}\Psi}_X)$$

Remark. *For the case of trained basis by NN, we replace $\Psi_X + t \cdot \mathcal{L}\Psi_X + O(t^2)$ by Ψ_Y in (5),*

Now, we will show that the computed $K_{N,n,m}$ by our S-EDMD method converges not only in large data m and large dictionary size N, but also in zero-limit of sampling time $t = 1/n$.

2 Convergence in the limit of large data

(1) i.i.d. data:

$$P\left(\|\widehat{K}_{N,n,m} - K_{N,n,m}\|_F > \epsilon\right)$$

$$= P\left(\|\widehat{\Psi}_X^\dagger(\widehat{\Psi}_X + t\widehat{\mathcal{L}\Psi_X}) - \Psi_X^\dagger\mathbb{E}[\Psi_Y]\|_F > \epsilon\right)$$

$$= P\left(\|t\widehat{\Psi}_X^\dagger\widehat{\mathcal{L}\Psi_X} - \Psi_X^\dagger(\Psi_X + t\mathcal{L}\Psi_X + O(t^2))\|_F > \epsilon\right)$$

$$= P\left(\|t\widehat{\Psi}_X^\dagger\widehat{\mathcal{L}\Psi_X} - t\Psi_X^\dagger\mathcal{L}\Psi_X + O(t^2)\|_F > \epsilon\right)$$

$$\leq P\left(t\|\widehat{\Psi}_X^\dagger\widehat{\mathcal{L}\Psi_X} - \Psi_X^\dagger\mathcal{L}\Psi_X\|_F + O(t^2) > \epsilon\right)$$

$$= P\left(t\|\widehat{G}^{-1}\widehat{A} - G^{-1}A\|_F + O(t^2) > \epsilon\right)$$

where $G = \frac{1}{m}\Psi_X^*\Psi_X$ and $A = \frac{1}{m}\Psi_X^*\mathcal{L}\Psi_X$.

(From Wiki)

A function $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfies the *bounded differences property* if substituting the value of the ith coordinate x_i changes the value of f by at most c_i. More formally, if there are constants $c_1, c_2, \ldots, c_n$ such that for all $i \in [n]$, and all $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_n$,

$$\sup_{x_i' \in \mathcal{X}_i} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

Lemma 1 (McDiarmid's Inequality). *Let* $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ *satisfy the bounded differences property with bounds* $c_1, c_2, \ldots, c_n$. *Consider independent random variables* $X_1, X_2, \ldots, X_n$ *where* $X_i \in \mathcal{X}_i$ *for all* i. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}(|f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots, X_n)]| \geq \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. Omitted. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Lemma 2 (Hoeffding's Inequality). *Assume* $X_i \in [a_i, b_i]$. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

and

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right| \geq \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Proof. The McDiarmid's Inequality(1) directly implies Hoeffding's inequality by taking $f(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

Theorem 3. *Let $|\psi_i(X_k)| \leq C$ $\mathbb{P} - a.e.$ for all $1 \leq i \leq N$ and $1 \leq k \leq m$. Suppose $\|\mathcal{L}_{N,n,m}\| \leq L$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left\|\widehat{G} - G\right\|_F \geq \epsilon\right) \leq 2N^2 \exp\left(-\frac{m\epsilon^2}{8N^2C^4}\right),$$

$$\mathbb{P}\left(\left\|\widehat{A} - A\right\|_F \geq \epsilon\right) \leq 2N^2 \exp\left(-\frac{m\epsilon^2}{8N^2C^4L^2}\right).$$

Proof. Define the random matrix $\eta(X) := \Psi(X)^T\Psi(X) \in \mathbb{R}^{N \times N}$, i.e., $\eta_{ij}(X) = \psi_i(X)\psi_j(X)$. Let $G = \mathbb{E}[\eta(X_1)], \widehat{G} = \frac{1}{m}\sum_{k=1}^m \eta(X_k)$. Then,

$$\|\widehat{G} - G\|_F^2 = \sum_{i=1}^N \sum_{j=1}^N \left|\widehat{G}_{ij} - G_{ij}\right|^2$$

$$= \sum_{i=1}^N \sum_{j=1}^N \left|\frac{1}{m}\sum_{k=1}^m \eta_{ij}(X_k) - \mathbb{E}\left[\eta_{ij}(X_k)\right]\right|^2$$

where $\tilde{\eta}(X) := \eta(X) - \mathbb{E}\left[\eta(X_1)\right]$ and thus $|\tilde{\eta}_{ij}(X)| \leq 2C^2$ for all $1 \leq i, j \leq N$. Next, applying Hoeffding's inequality (2) and union bound, we have

$$\mathbb{P}\left(\left\|\widehat{G} - G\right\|_F \geq \epsilon\right) = \mathbb{P}\left(\left\|\widehat{G} - G\right\|_F^2 \geq \epsilon^2\right)$$

$$\leq \mathbb{P}\left(\sum_{i=1}^N \sum_{j=1}^N \left|\frac{1}{m}\sum_{k=1}^m \tilde{\eta}_{ij}(X_k)\right|^2 \geq \epsilon^2\right)$$

$$\leq N^2\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^m \tilde{\eta}_{ij}(X_k)\right|^2 \geq \epsilon^2/N^2\right)$$

$$= N^2\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^m \tilde{\eta}_{ij}(X_k)\right| \geq \epsilon/N\right)$$

$$\leq 2N^2 \exp\left(-\frac{2(m\epsilon/N)^2}{m(4C^2)^2}\right)$$

$$= 2N^2 \exp\left(-\frac{m\epsilon^2}{8N^2C^4}\right)$$

Similarly, define $\xi(X) := \Psi(X)^T\mathcal{L}_{N,n,m}\Psi(X)$. Since

$$|\psi_i(X)\mathcal{L}\psi_j(X)| \leq C^2\|\mathcal{L}_{N,n,m}\| \leq C^2L,$$

we have

$$\mathbb{P}\left(\|\widehat{A} - A\|_F \geq \epsilon\right) \leq 2N^2 \exp\left(-\frac{m\epsilon^2}{8N^2C^4L^2}\right),$$

$\square$

Lemma 4 (Philipp, Lemma C.5). *Let $G, A \in \mathbb{R}^{N \times N}$ be such that G is invertible and $A \neq 0$. Let $\widehat{G}, \widehat{A} \in \mathbb{R}^{N \times N}$ be random matrices such that $\widehat{G}$ is invertible a.s. Then for any sub-multiplicative matrix norm $\| \cdot \|$ on $\mathbb{R}^{N \times N}$ and any $\epsilon > 0$ we have*

$$\mathbb{P}\left(\| G^{-1}A - \widehat{G}^{-1}\widehat{A} \| > \epsilon \right) \leq \mathbb{P}\left(\| A - \widehat{A} \| > \frac{\epsilon}{\tau}\|A\| \right) + \mathbb{P}\left(\| G - \widehat{G} \| > \frac{\epsilon}{\tau}\|G^{-1}\|^{-1} \right),$$

where $\tau = 2\|G^{-1}\|\|A\| + \epsilon$.

Theorem 5. *Let $\epsilon > 0$. Define $\epsilon_t = \left(\epsilon - O(t^2) \right)/t$. Then, with same conditions defined in Theorem 3 and Lemma 4, we have*

$$P\left(\|\widehat{K}_{N,n,m} - K_{N,n,m}\|_F > \epsilon \right) \leq 2N^2 \left[\exp\left(-\frac{m}{8}\left(\frac{\epsilon_t}{NC^2\tau\|G^{-1}\|} \right)^2 \right) + \exp\left(-\frac{m}{8}\left(\frac{\epsilon_t\|A\|}{NC^2\tau L} \right)^2 \right) \right]$$

Proof.

$$
\begin{aligned}
P\left(\|\widehat{K}_{N,n,m} - K_{N,n,m}\|_F > \epsilon \right) &= P\left(t\|\widehat{G}^{-1}\widehat{A} - G^{-1}A\|_F + O(t^2) > \epsilon \right) \\
&= P\left(\|\widehat{G}^{-1}\widehat{A} - G^{-1}A\|_F > \epsilon_t \right) \\
&\leq \mathbb{P}\left(\| G - \widehat{G} \| > \frac{\epsilon_t}{\tau}\|G^{-1}\|^{-1} \right) + \mathbb{P}\left(\| A - \widehat{A} \| > \frac{\epsilon_t}{\tau}\|A\| \right) \\
&\leq 2N^2 \exp\left(-\frac{m(\frac{\epsilon_t}{\tau}\|G^{-1}\|^{-1})^2}{8N^2C^4} \right) + 2N^2 \exp\left(-\frac{m(\frac{\epsilon_t}{\tau}\|A\|)^2}{8N^2C^4L^2} \right) \\
&= 2N^2 \left[\exp\left(-\frac{m}{8}\left(\frac{\epsilon_t\|G^{-1}\|^{-1}}{NC^2\tau} \right)^2 \right) + \exp\left(-\frac{m}{8}\left(\frac{\epsilon_t\|A\|}{NC^2\tau L} \right)^2 \right) \right]
\end{aligned}
$$

$\square$

3 Convergence in zero-limit of sampling time

Now we have Koopman matrix approximation at discrete time $K_{N,n}(1/n)$. We will use it to construct a sequence of generator approximant $\{A_{N,n}\}_{n \geq 1}$ and find its limit as $n \to \infty$, if the limit exists in some sense. Once we find the limit, we can create the continuous time N-dimensional Koopman semigroup.

First, for each $N > 0$, construct the following sequence of matrices:

$$A_{N,n} := \frac{K_{N,n}\left(\frac{1}{n}\right) - I}{\frac{1}{n}}$$

$$= \frac{\left[\Psi_X^\dagger \left(\Psi_X + \frac{1}{n} \cdot \mathcal{L}\Psi_X + O(\frac{1}{n^2})\right)\right] - I}{\frac{1}{n}}$$

$$= \frac{\left[I + \frac{1}{n}\Psi_X^\dagger \mathcal{L}\Psi_X + O(\frac{1}{n^2})\right] - I}{\frac{1}{n}}$$

$$= \Psi_X^\dagger \mathcal{L}\Psi_X + O(\frac{1}{n})$$

(Here the first term is not dependent on n)

Next, choose $f = \boldsymbol{\Psi}_N \mathbf{a} \in C_b^2(\Omega)$, we have

$$\|\lim_{n \to \infty} \boldsymbol{\Psi}_N A_{N,n}\mathbf{a}\|$$

$$= \|\lim_{n \to \infty} \boldsymbol{\Psi}_N \left(\Psi_X^\dagger \mathcal{L}\Psi_X + O(\frac{1}{n})\right)\mathbf{a}\|$$

$$= \|\boldsymbol{\Psi}_N \left(\Psi_X^\dagger \mathcal{L}\Psi_X\right)\mathbf{a} + \lim_{n \to \infty} \boldsymbol{\Psi}_N O(\frac{1}{n})\mathbf{a}\| \quad \text{(Here the } O(\frac{1}{n}) \text{ is a matrix)}$$

Then,

$$\|\lim_{n \to \infty} \boldsymbol{\Psi}_N O(\frac{1}{n})\mathbf{a}\|$$

$$= \|\lim_{n \to \infty} \begin{bmatrix} \psi_1 & \cdots & \psi_N \end{bmatrix} \begin{bmatrix} O(\frac{1}{n}) & \cdots & O(\frac{1}{n}) \\ \vdots & \ddots & \vdots \\ O(\frac{1}{n}) & \cdots & O(\frac{1}{n}) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}\|$$

$$= \|\lim_{n \to \infty} \begin{bmatrix} \psi_1 & \cdots & \psi_N \end{bmatrix} \begin{bmatrix} a_1' \\ \vdots \\ a_N' \end{bmatrix}\|$$

$$\text{(where } a_i' = \sum_{j=1}^N [O(\frac{1}{n})]_{ij} a_j)$$

$$= \|\lim_{n \to \infty} \sum_{i=1}^N \psi_i a_i'\|$$

$$= \|\lim_{n \to \infty} \sum_{i=1}^N \psi_i \left(\sum_{j=1}^N [O(\frac{1}{n})]_{ij} a_j\right)\|$$

$$\leq \| \lim_{n\to\infty} \sum_{i=1}^{N} \psi_i \left((N \cdot O(\tfrac{1}{n})) \sum_{j=1}^{N} a_j \right) \|$$

$$= \| \lim_{n\to\infty} (N \cdot O(\tfrac{1}{n})) \sum_{i,j=1}^{N} \psi_i \left(\sum_{i,j=1}^{N} a_j \right) \| = 0$$

Therefore, for $f = \boldsymbol{\Psi}_N \mathbf{a}$, we have

$$\lim_{n\to\infty} \boldsymbol{\Psi}_N A_{N,n} \mathbf{a} = \boldsymbol{\Psi}_N A_N \mathbf{a}$$

where $A_N = \Psi_X^\dagger \mathcal{L} \Psi_X$. In other words,

$$\lim_{n\to\infty} \mathcal{A}_{N,n} f = \mathcal{A}_N f$$

where $\mathcal{A}_{N,n}, \mathcal{A}_N$ are the operators whose matrix representation are $A_{N,n}, A_N$ respectively.(Not sure if it is appropriate and necessary to mention $\mathcal{A}_{N,n}, \mathcal{A}_N$ here. If yes, do we need to define the domain for these operators?)

Next, by Trotter-Kato Approximation theorem, we have

$$\lim_{n\to\infty} \boldsymbol{\Psi}_N K_{N,n} \mathbf{a} = \boldsymbol{\Psi}_N K_N \mathbf{a}$$

where $K_{N,n}(t) = e^{tA_{N,n}}, K_N(t) = e^{tA_N}$. In other words, we have

$$\lim_{n\to\infty} \mathcal{K}_{N,n}(t) f = \mathcal{K}_N(t) f$$

where $\mathcal{K}_{N,n}(t) := e^{t\mathcal{A}_{N,n}}, \mathcal{K}_N(t) := e^{t\mathcal{A}_N}$.

First Trotter–Kato Approximation Theorem.

(Trotter 1958, Kato 1959). Let $(T(t))_{t\geq0}$ and $(T_n(t))_{t\geq0}$, $n \in \mathbb{N}$, be strongly continuous semigroups on X with generators $\mathcal{A}$ and $\mathcal{A}_n$, respectively, and assume that they satisfy the estimate

$$\|T(t)\|, \|T_n(t)\| \leq Me^{wt} \quad \text{for all } t \geq 0, n \in \mathbb{N},$$

and some constants $M \geq 1$, $w \in \mathbb{R}$. Take $\mathcal{D}$ to be a core for $\mathcal{A}$ and consider the following assertions.

(a) $\mathcal{D} \subset \mathcal{D}(\mathcal{A}_n)$ for all $n \in \mathbb{N}$ and $\mathcal{A}_n x \to \mathcal{A}x$ for all $x \in \mathcal{D}$.

(b) For each $x \in \mathcal{D}$, there exists $x_n \in \mathcal{D}(\mathcal{A}_n)$ such that $x_n \to x$ and $\mathcal{A}_n x_n \to \mathcal{A}x$.

(c) $R(\lambda, \mathcal{A}_n)x \to R(\lambda, \mathcal{A})x$ for all $x \in X$ and some/all $\lambda > w$.

(d) $T_n(t)x \to T(t)x$ for all $x \in X$, uniformly for t in compact intervals.

Then the implications

$$(a) \implies (b) \iff (c) \iff (d)$$

hold, while (b) does not imply (a).

4 Convergence in large dictionary size N

In this section, we want to show that $\mathcal{P}_N \mathcal{A}|_{\mathcal{F}_N} = \mathcal{A}_N \to \mathcal{A}$ in strong operator topology as $N \to \infty$, under a similar mathematical framework established in gEDMD analysis paper . In addition, if we can assume the $\{\mathcal{K}_N(t)\}_{N \geq 1}, \mathcal{K}(t)$ is exponentially bounded for all $t \geq 0$, then we can use the Trotter-Kato Approximation theorem to obtain $\mathcal{P}_N \mathcal{K}(t)|_{\mathcal{F}_N} = \mathcal{K}_N(t) \to \mathcal{K}(t)$ in strong operator topology as $N \to \infty$, uniformly for t in a compact interval.

4.1 Convergence of $\mathcal{A}_N$ to $\mathcal{A}$ (in the L^2 norm)

Definition 6. $\mathcal{A}$: *operator of interest* $\mathcal{D} := \{f \in \mathcal{F} : \mathcal{A}f \in \mathcal{F}\}$
   $(\mathbb{X}, \mathcal{B}, \mu)$: *state space* $\mathcal{F} := L^2(\mathbb{X}, \mu)$
   Graph of $\mathcal{A}$: $\{(f, \mathcal{A}f) \in \mathcal{F} \times \mathcal{F} : f \in \mathcal{D}\}$
   $\langle f, g \rangle_{\mathcal{D}} := \langle f, g \rangle_{\mathcal{F}} + \langle \mathcal{A}f, \mathcal{A}g \rangle_{\mathcal{F}}$ $\|f\|^2 := \langle f, f \rangle_{\mathcal{D}}, \quad \forall f, g \in \mathcal{D}$
   $\|\mathcal{A}\| \leq 1$ $\|f\|_{\mathcal{F}} \leq \|f\|_{\mathcal{D}}$
   If $\mathcal{A}$ *is continuous on* $\mathcal{F}$, *then* $\mathcal{D} = \mathcal{F}$
   $\mathcal{D} = H^k(\mathbb{X}, \mu)$: *weighted Sobolev space* $\|\psi\|_{H^k(\mathbb{X}, \mu)} = \sum_{|\alpha| \leq k} \left(\int_{\mathbb{X}} |D^\alpha \psi|^2 d\mu \right)^{1/2}$
   If $\mathcal{A}$ *is generator of a stochastic process:* $\mathcal{D} = H^2(\mathbb{X}, \mu)$ *If system is deterministic:* $\mathcal{D} = H^1(\mathbb{X}, \mu)$

We denote the operator of interest by A, which can represent various operators, such as K, K^t, L, or L^*, in discrete or continuous time. We consider the state space $(X, \mathcal{B}, \mu)$ and let $\mathcal{F} := L^2(X, \mu)$. The observables on which A acts define the domain of A, i.e.,

$$D := \{f \in \mathcal{F} : Af \in \mathcal{F}\}.$$

The operator A is said to be a *closed operator* if the graph of A,

$$\{(f, Af) \in \mathcal{F} \times \mathcal{F} : f \in D\},$$

is a closed subspace of $\mathcal{F} \times \mathcal{F}$. If A is closed, then D is a Hilbert space with the inner product

$$\langle f, g \rangle_D := \langle f, g \rangle_{\mathcal{F}} + \langle Af, Ag \rangle_{\mathcal{F}}, \quad \forall f, g \in D,$$

and norm

$$\|f\|_D^2 := \langle f, f \rangle_D.$$

A direct consequence of this definition is that $\|f\|_{\mathcal{F}} \leq \|f\|_D$ and that $A : D \to \mathcal{F}$ is a continuous operator with $\|A\| \leq 1$.

In practice, there are two main cases. If A is continuous on $\mathcal{F}$, then $D = \mathcal{F}$ and the norms $\|\cdot\|_D$ and $\|\cdot\|_{\mathcal{F}}$ are equivalent. Otherwise, A is typically the infinitesimal generator of a strongly continuous semigroup, and D will be dense in $\mathcal{F}$.

A common example is where D is a weighted Sobolev space $H^k(X, \mu)$ containing all measurable functions $\psi : X \to \mathbb{R}$ with finite norm:

$$\|\psi\|_{H^k(X,\mu)} = \left(\sum_{|\alpha| \leq k} \int_X |D^\alpha \psi|^2 \, d\mu \right)^{1/2},$$

where D^α is the weak derivative of order α. For example, if A is the generator of a stochastic process, then $D = H^2(X, \mu)$. If the system is deterministic, then $D = H^1(X, \mu)$.

Assumption 1. *We assume the following:*

1. *The basis functions $\Psi = \{\psi_1, \ldots, \psi_N\} \subset \mathcal{D}$ are linearly independent.*

2. *The functions $\{\psi_i, \mathcal{A}\psi_i\}_{i=1}^N$ are continuous μ-a.e..*

3. *The points $\{x_i\}_{i=1}^m \subset \Omega$ are i.i.d. samples from μ.*

Assumption 2. *Assumption 1 holds and*

$$\lim_{N \to \infty} \|\mathcal{P}_N \phi - \phi\|_{\mathcal{F}} = 0, \quad \forall \phi \in \mathcal{F},$$

where $\mathcal{P}_N$ is the projection of $\mathcal{F}$ onto $\mathcal{F}_N$ using the inner product on $\mathcal{F}$.

Assumption 3. *Assumption 1 holds, $\mathcal{A}$ is a closed operator, and*

$$\lim_{N \to \infty} \|\mathcal{P}_{\mathcal{D}_N} f - f\|_{\mathcal{D}} = 0, \quad \forall f \in \mathcal{D}.$$

Here, $\mathcal{P}_{\mathcal{D}_N}$ is the projection of $\mathcal{D}$ onto $\mathcal{F}_N$ using the inner product on $\mathcal{D}$.

Theorem 7. *Let Ψ satisfy Assumption 2 and 3, then*

$$\lim_{N \to \infty} \|\mathcal{A}_N \mathcal{P}_{\mathcal{D}_N} f - \mathcal{A} f\|_{\mathcal{F}} = 0, \quad \forall f \in \mathcal{D}.$$

Proof. By definition $\mathcal{A}_N = \mathcal{P}_N \mathcal{A}|_{\mathcal{F}_N}$, we obtain

$$\begin{aligned}
\mathcal{A}_N \mathcal{P}_{\mathcal{D}_N} - \mathcal{A} &= (\mathcal{A}_N - \mathcal{A})\mathcal{P}_{\mathcal{D}_N} + \mathcal{A}\mathcal{P}_{\mathcal{D}_N} - \mathcal{A} \\
&= (\mathcal{P}_N - \mathrm{Id})\mathcal{A}\mathcal{P}_{\mathcal{D}_N} + \mathcal{A}(\mathcal{P}_{\mathcal{D}_N} - \mathrm{Id}) \\
&= (\mathcal{P}_N - \mathrm{Id})\mathcal{A} + (\mathcal{P}_N - \mathrm{Id})\mathcal{A}(\mathcal{P}_{\mathcal{D}_N} - \mathrm{Id}) + \mathcal{A}(\mathcal{P}_{\mathcal{D}_N} - \mathrm{Id}).
\end{aligned}$$

Consider now $f \in \mathcal{D}$. By Assumptions 2, 3 and the fact that $\mathcal{A}$ is continuous on its domain, we have

$$\begin{aligned}
\|\mathcal{A}_N \mathcal{P}_{\mathcal{D}_N} f - \mathcal{A} f\|_{\mathcal{F}} &\leq \|(\mathcal{P}_N - \mathrm{Id})\mathcal{A} f\|_{\mathcal{F}} + \|(\mathcal{P}_N - \mathrm{Id})\mathcal{A}\| \|\mathcal{P}_{\mathcal{D}_N} f - f\|_{\mathcal{D}} \\
&\quad + \|\mathcal{A}\| \|\mathcal{P}_{\mathcal{D}_N} f - f\|_{\mathcal{D}} \to 0.
\end{aligned}$$

$\square$

4.2 Convergence of $\mathcal{K}_N(t)$ to $\mathcal{K}(t)$ (in the L^2 norm)

Outline: make suitable assumptions(Koopman operators are exponentially bounded) and use Trotter-Kato Approximation theorem to show the convergence.

A "Taylor" expansion of stochastic Koopman operator

By applying Itô's formula to both $f(X_t)$ and $\mathcal{A}f(X_t)$, we can derive a "Taylor expansion" for $\mathbb{E}[f(X_t)]$ as in (3).

First, we apply Itô's formula to $f(X_t)$:

$$f(X_t) = f(x) + \int_0^t (\mathcal{A}f)(X_s)\,ds + \int_0^t f'(X_s)\,\sigma(X_s)\,dW_s.$$

Next, we treat $\mathcal{A}f$ as a function and apply Itô's formula to $\mathcal{A}f(X_s)$:

$$(\mathcal{A}f)(X_t) = (\mathcal{A}f)(x) + \int_0^t [\mathcal{A}(\mathcal{A}f)](X_s)\,ds + \int_0^t (\mathcal{A}f')(X_s)\,\sigma(X_s)\,dW_s.$$

Then, we substitute this expression for $\mathcal{A}f(X_t)$ back into the formula for $f(X_t)$:

$$
\begin{aligned}
f(X_t) &= f(x) + \int_0^t (\mathcal{A}f)(X_s)\,ds + \int_0^t f'(X_s)\,\sigma(X_s)\,dW_s \\
&= f(x) + \int_0^t \left[(\mathcal{A}f)(x) + \int_0^s [\mathcal{A}(\mathcal{A}f)](X_u)\,du + \int_0^s (\mathcal{A}f')(X_u)\,\sigma(X_u)\,dW_u \right] ds \\
&\quad + \int_0^t f'(X_s)\,\sigma(X_s)\,dW_s.
\end{aligned}
$$

After rearranging terms, we have:

$$
\begin{aligned}
f(X_t) &= f(x) + (\mathcal{A}f)(x)\,t + \int_0^t \int_0^s [\mathcal{A}(\mathcal{A}f)](X_u)\,du\,ds \\
&\quad + \int_0^t \int_0^s (\mathcal{A}f')(X_u)\,\sigma(X_u)\,dW_u\,ds + \int_0^t f'(X_s)\,\sigma(X_s)\,dW_s.
\end{aligned}
$$

Taking expectations on both sides and noting that the stochastic integrals have zero mean (assuming appropriate integrability conditions), we get:

$$\mathbb{E}[f(X_t)] = f(x) + (\mathcal{A}f)(x)\,t + \int_0^t \int_0^s \mathbb{E}\left[\mathcal{A}(\mathcal{A}f)(X_u)\right] du\,ds.$$

The double integral term represents the accumulated effect of the higher-order derivatives of f over time. To understand its order, consider that if $\mathbb{E}\left[\mathcal{A}(\mathcal{A}f)(X_u)\right]$ is bounded by some constant M, then:

$$\left| \int_0^t \int_0^s \mathbb{E}\left[\mathcal{A}(\mathcal{A}f)(X_u)\right] du\,ds \right| \le M \int_0^t \int_0^s du\,ds = M \int_0^t s\,ds = M\frac{t^2}{2}.$$

This shows that the double integral is of order $O(t^2)$ when t is small.

Therefore, for small t, the expected value simplifies to:

$$\mathbb{E}[f(X_t)] = f(x) + (\mathcal{A}f)(x)\,t + O(t^2). \tag{7}$$