

降阶模型笔记

Shun Li

2025 年 9 月

参考 科学机器学习-黄政宇 <http://faculty.bicmr.pku.edu.cn/~huangdz/teaching.html>,
参考教材 Athanasios C. Antoulas, “Approximation of Large-Scale Dynamical Systems”。

目录

第一章 基于投影的降阶模型	3
1.1 解的低维特性	3
1.2 低维投影	3
1.3 降阶模型	4
1.4 误差与稳定性	5
第二章 低维子空间的构造	7
2.1 本征正交分解 (POD)	7
2.2 平衡截断法和矩匹配方法	8
2.2.1 线性动力系统	9
2.2.2 状态分类	10
2.2.3 平衡截断法	13
2.2.4 矩匹配	16
第三章 针对非线性系统的降阶模型	18
3.1 非线性动力系统	18
3.2 轨迹分段线性方法	18
3.3 离散经验插值方法 (DEIM)	19
第四章 数据驱动的降阶模型	21
4.1 谱分解	21
4.2 Koopman 算子	21
4.2.1 连续时间系统	21
4.2.2 离散时间系统	23
4.3 Koopman 模态	25
4.4 Koopman 算子的有限维近似	26
4.5 动力学模态分解 (DMD)	27
附录 A 数值线性代数基础	29
A.1 奇异值分解	29
A.2 关于 SVD 低秩截断	30

第一章 基于投影的降阶模型

1.1 解的低维特性

考虑常微分方程

$$\frac{d\mathbf{u}}{dt} = f(\mathbf{u}(t), t, x)$$

其中 $\mathbf{u}(t) \in \mathbb{C}^N$ 是高维状态变量, $\mathbf{u}(0) = \mathbf{u}_0$ 是初始条件, 输出变量 $y(t) = g(\mathbf{u}(t), t, x) \in \mathbb{C}^q$ (通常 $q \ll N$)。

在理想情况下, 状态变量可以被精确地表示为一系列向量的线性组合, 这些向量形成了一个低维线性空间的基。定义一个低维线性空间 $\mathcal{S} \in \mathbb{C}^N$

$$\mathcal{S} = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k), \quad k = \dim(\mathcal{S})$$

那么状态变量可以进行分解

$$\mathbf{u}(t) = \sum_{i=1}^k q_i \mathbf{v}_i = V_s \mathbf{q}(t), \quad V = [\mathbf{v}_1 \quad \dots \quad \mathbf{v}_k]$$

$\mathbf{q}(t)$ 是 $\mathbf{u}(t)$ 在 $\mathcal{S}$ 中的广义坐标, 也是降阶状态变量。

更实际的情况是, 低维线性空间 $\mathcal{S}$ 以及基底 V_s 都是未知的, 不妨记试验基底为 $V \in \mathbb{C}^{N \times k}$ (k 可能也不是精确的)。假设有近似 $\mathbf{u}(t) \approx V \mathbf{q}(t)$, 代入原方程可得

$$\frac{d}{dt}(V \mathbf{q}(t)) = f(V \mathbf{q}(t), t, x) + r(t), \quad y(t) \approx g(V \mathbf{q}(t), t).$$

其中 $r(t)$ 是低维子空间逼近所产生的残差。由于基底 V 并不依赖于时间, 可以提出来得到

$$V \frac{d}{dt}(\mathbf{q}(t)) = f(V \mathbf{q}(t), t, x) + r(t), \quad y(t) \approx g(V \mathbf{q}(t), t).$$

这里得到的是 N 个微分方程组成的方程组, 但是只有 $k \ll N$ 个未知函数, 因此是一个超定系统。

1.2 低维投影

Definition 1.2.1. 称矩阵 $\Pi \in \mathbb{C}^{N \times N}$ 是 $\mathbb{C}^N$ 中的投影矩阵, 如果 $\Pi^2 = \Pi$ 。

Proposition 1.2.1. 投影矩阵 Π 有如下性质:

1. Π 的特征值只有 0 和 1;
2. $\text{Im}(\Pi) \oplus \text{Ker}(\Pi) = \mathbb{C}^N$;

3. Π 可以相似对角化

$$\Pi = P \begin{bmatrix} I_m & \\ & 0_{N-m} \end{bmatrix} P^{-1}$$

如果定义

$$\mathcal{S}_1 = \text{Im}(\Pi), \quad \mathcal{S}_2 = \text{Ker}(\Pi).$$

那么, 实际上 Π 定义了沿着 $\mathcal{S}_2$ 方向到 $\mathcal{S}_1$ 的投影。如果满足 $\mathcal{S}_1 = \mathcal{S}_2^\perp$, 那么 Π 就是正交投影矩阵, 因为余量 $v - \Pi v \in \mathcal{S}_2$ 满足

$$(v - \Pi v) \perp \mathcal{S}_1, \quad \forall v$$

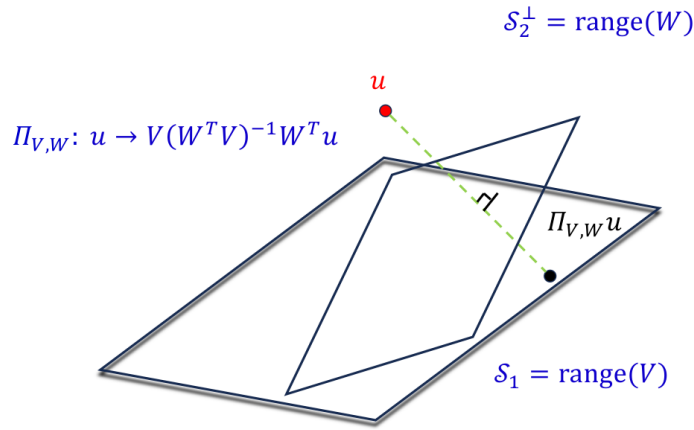
其他情况下, $\mathcal{S}_1 \neq \mathcal{S}_2^\perp$, Π 是斜投影矩阵。可以给出正交投影/斜投影矩阵的一般性构造:

1. 对于列满秩矩阵 $V \in \mathbb{C}^{N \times k}$, 它的列向量张成 $\mathcal{S}_1$, 那么可以定义正交投影矩阵

$$\Pi_{V,V} = V(V^H V)^{-1} V^H$$

2. 考虑两个列满秩矩阵 $V, W \in \mathbb{C}^{N \times k}$, 要求 $W^H V$ 可逆, 它们的列向量分别张成了 $\mathcal{S}_1$ 和 $\mathcal{S}_2^\perp$, 那么可以定义斜投影矩阵

$$\Pi_{V,W} = V(W^H V)^{-1} W^H$$



1.3 降阶模型

对于高维模型 (High Dimensional Model)

$$\frac{d\mathbf{u}}{dt} = f(\mathbf{u}(t), t, x)$$

其中 $\mathbf{u}(t) \in \mathbb{C}^N$ 是高维状态变量, $\mathbf{u}(0) = \mathbf{u}_0$ 是初始条件, 输出变量 $y(t) = g(\mathbf{u}(t), t, x) \in \mathbb{C}^q$ (通常 $q \ll N$)。目标是设计基于投影的降阶模型为 (Projection based Reduced Order Model)

$$\frac{d\mathbf{q}}{dt} = f_r(\mathbf{q}(t), t, x)$$

其中 $\mathbf{q}(t)$ 是降阶的低维状态变量。

考虑投影近似 $\mathbf{u}(t) \approx V\mathbf{q}(t)$, 其中 $V \in \mathbb{C}^{N \times k}$ 是列满秩矩阵, 可以得到

$$V \frac{d\mathbf{q}}{dt} = f(V\mathbf{q}(t), t, x) + r(t)$$

构造列满秩矩阵 $W \in \mathbb{C}^{N \times k}$ (要求保证 $W^H V$ 可逆), 使得残差满足正交关系

$$W^H r(t) = 0$$

可以得到

$$W^H V \frac{d\mathbf{q}}{dt} = W^H f(V\mathbf{q}(t), t, x)$$

整理可得基于 Petrov-Galerkin 投影的降阶模型

$$\frac{d\mathbf{q}}{dt} = (W^H V)^{-1} W^H f(V\mathbf{q}(t), t, x), \quad y(t) \approx g(V\mathbf{q}(t), t)$$

如果特别地取 $W = VQ$ (其中 Q 是一个可逆变换矩阵, 表明 $\mathcal{S}_2^\perp = \mathcal{S}_1$), 可以得到基于 Galerkin 投影的降阶模型

$$\frac{d\mathbf{q}}{dt} = (V^H V)^{-1} V^H f(V\mathbf{q}(t), t, x), \quad y(t) \approx g(V\mathbf{q}(t), t)$$

Remark 1.3.1. 上文构造的降阶模型与子空间基底的选取无关, 例如考虑 $\tilde{W} = WQ_1$, $\tilde{V} = VQ_2$, 其中 $Q_1, Q_2 \in \mathbb{C}^{k \times k}$ 是可逆变换矩阵。变换后的降阶模型满足

$$\frac{d}{dt}(\tilde{V}\tilde{\mathbf{q}}(t)) = \tilde{V}(\tilde{W}^H \tilde{V})^{-1} \tilde{W}^H f(\tilde{V}\tilde{\mathbf{q}}(t), t, x) = V(W^H V)^{-1} W^H f(\tilde{V}\tilde{\mathbf{q}}(t), t, x) = \frac{d}{dt}(V\mathbf{q}(t))$$

考虑初值条件 $\mathbf{u}(0) = \mathbf{u}_0 \in \mathbb{C}^N$ 的处理, 此时第一种做法是对初值进行投影

$$\mathbf{q}(0) = (W^H V)^{-1} W^H \mathbf{u}_0 \in \mathbb{C}^k.$$

此时存在初值投影误差

$$\mathbf{u}(0) - V\mathbf{q}(0) = (I - V(W^H V)^{-1} W^H) \mathbf{u}_0.$$

第二种做法是把初值减掉, 考虑对增量进行低维近似

$$\mathbf{u}(t) = \mathbf{u}_0 + V\mathbf{q}(t), \quad \mathbf{q}(0) = 0.$$

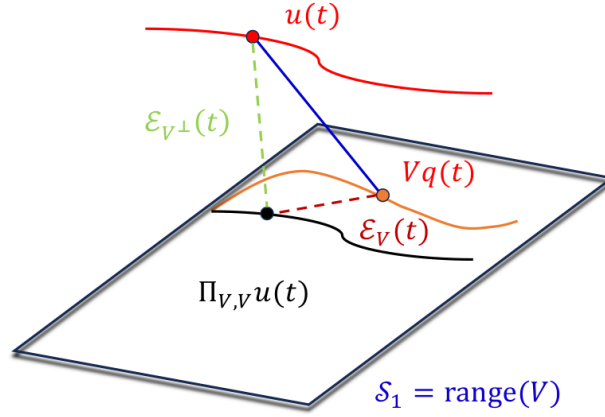
1.4 误差与稳定性

考虑基于 Galerkin 投影的降阶模型的误差, 以下取标准正交基 $V^H V = I$, 因此 $\Pi_{V,V} = VV^H$ 。

$$\begin{aligned} \mathcal{E} &:= \mathbf{u}(t) - V\mathbf{q}(t) = \mathbf{u}(t) - \Pi_{V,V} \mathbf{u}(t) + \Pi_{V,V} \mathbf{u}(t) - V\mathbf{q}(t) \\ &= (I - \Pi_{V,V}) \mathbf{u}(t) + V(V^H \mathbf{u}(t) - \mathbf{q}(t)) \\ &=: \mathcal{E}_{V^\perp}(t) + \mathcal{E}_V(t) \end{aligned}$$

误差可以拆分为两个部分:

1. 投影误差: $\mathcal{E}_{V^\perp}(t) = (I - \Pi_{V,V}) \mathbf{u}(t)$;
2. 模型误差: $\mathcal{E}_V(t) = V(V^H \mathbf{u}(t) - \mathbf{q}(t))$ 。



Theorem 1.4.1. 对于线性模型

$$\frac{d\mathbf{u}(t)}{dt} = A\mathbf{u}(t),$$

考虑基于 *Galerkin* 投影的降阶模型, 假设 $\mathbf{u}(0) = V\mathbf{q}(0)$, 那么

$$\|\mathcal{E}(t)\|_{L^2([0,T])} \leq (\|F(T, V^H A V)\|_2 \|V^H A V^\perp\|_2 + 1) \|\mathcal{E}_{V^\perp}(t)\|_{L^2([0,T])}$$

其中的范数定义为

$$\|h(t)\|_{L^2([0,T])} = \left(\int_0^T |h(t)|^2 dt \right)^{\frac{1}{2}}.$$

$V^\perp$ 的列向量是 V 的正交补空间的一组标准正交基。对 $M \in \mathbb{C}^{k \times k}$, 定义 $F(T, M) : L^2([0, T], \mathbb{C}^k) \rightarrow L^2([0, T], \mathbb{C}^k)$ 满足

$$(F(T, M)h)(t) = \int_0^t e^{M(t-\tau)} h(\tau) d\tau$$

Theorem 1.4.2. 对于稳定线性模型

$$\frac{d\mathbf{u}(t)}{dt} = A\mathbf{u}(t), \quad \text{Re}(\lambda(A)) \leq 0.$$

如果 A 对称, 那么基于 *Galerkin* 投影的降阶模型仍然稳定; 如果 A 不对称, 那么基于 *Galerkin* 投影的降阶模型可能不稳定。(这里只考虑实矩阵)

第二章 低维子空间的构造

2.1 本征正交分解 (POD)

现在关注如何构造降阶模型的基底 V ，一个最基础的做法是本征正交分解 (Proper Orthogonal Decomposition, POD)：固定初始条件 $\mathbf{u}(0) = \mathbf{u}_0$ ，考虑 $t \in [0, T]$ 的解 $\mathbf{u}(t)$ ，本征正交分解寻求一个固定秩为 k 的正交投影 $\Pi_{V,V}$ ，使得如下积分投影误差最小化

$$J(\Pi_{V,V}) = \int_0^T \|\mathcal{E}_{V^\perp}(t)\|_2^2 dt = \int_0^T \|\mathbf{u}(t) - \Pi_{V,V}\mathbf{u}(t)\|_2^2 dt$$

Remark 2.1.1. 这个方法在多元统计学中被称为主成分分析 (Principal Component Analysis, PCA)，都是基于 SVD 分解进行数据降维。

Theorem 2.1.1. 定义 Hermitian 半正定矩阵 $\hat{K} \in \mathbb{C}^{N \times N}$

$$\hat{K} = \int_0^T \mathbf{u}(t)\mathbf{u}(t)^H dt$$

记 $\hat{K}$ 的特征值以及对应的特征向量为

$$\hat{K}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i, \quad i = 1, \dots, N$$

其中 $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_N \geq 0$ 。对于取定的 $k \in \mathbb{N}$ ，假设 $\hat{\lambda}_k > \hat{\lambda}_{k+1}$ ，那么使得 $J(\Pi_{V,V})$ 达到最小值的子空间 $\mathcal{V} = \text{range}(\hat{V})$ 是矩阵 $\hat{K}$ 关于特征值 $\hat{\lambda}_1, \dots, \hat{\lambda}_k$ 的不变子空间。

在实践中通常不会直接计算 $\hat{K}$ 并求解它的特征系统，因为矩阵 $\hat{K}$ 的规模太大，通常是稠密矩阵，并且状态数据通常以离散的快照形式 (snapshot) 提供： $\{\mathbf{u}(t_i)\}_{i=1}^{N_{snap}}$ ，此时可以通过数值积分近似计算

$$\hat{K} = \int_0^T \mathbf{u}(t)\mathbf{u}(t)^H dt \approx \sum_{i=1}^{N_{snap}} \alpha_i \mathbf{u}(t_i)\mathbf{u}(t_i)^H =: K$$

其中 $\{\alpha_i\}$ 是对应的 (非负) 数值积分权重。定义快照矩阵 $S \in \mathbb{C}^{N \times N_{snap}}$

$$S = \begin{bmatrix} \sqrt{\alpha_1}\mathbf{u}(t_1) & \sqrt{\alpha_2}\mathbf{u}(t_2) & \cdots & \sqrt{\alpha_{N_{snap}}}\mathbf{u}(t_{N_{snap}}) \end{bmatrix}$$

那么 $K = SS^H$ 。现在的目标是求解 K 的前 k 个特征向量以及对应的特征值。

利用 SVD 分解的性质可得：如果 S 的奇异值分解为 $S = U_r \Sigma_r V_r^H$ ，那么 K 的特征分解形如

$$K = SS^H = U_r \Sigma_r^2 U_r^H.$$

因此可以取 V 为 U_r 的前 k 列。

Remark 2.1.2. 当然也可以直接求解 $K = SS^H$ 的特征值和特征向量, 复杂度一样, 但是问题的条件数更大: $\text{cond}(SS^H) \approx (\text{cond}(S))^2$ 。

现在考虑如何确定子空间的维度 k : 记 U_k 为 U_r 的前 k 列, 定义

$$\mathcal{E}_{POD}(k) = \frac{\|\Pi_{U_k, U_k} S\|_F^2}{\|S\|_F^2} = \frac{\sum_{i=1}^k \sigma_i^2(S)}{\sum_{i=1}^r \sigma_i^2(S)},$$

代表前 k 个本征正交基向量所捕获的快照能量占比, 可以要求选取 k 使得丢失的能量占比小于给定阈值 $\varepsilon > 0$

$$1 - \mathcal{E}_{POD}(k) \leq \varepsilon.$$

例如取 $\varepsilon = 0.1\%$ 。

现在考虑快照矩阵的构造, 首先考虑含参数但不含时间的偏微分方程问题 $f(u, \mu) = 0$, 其中参数 $\mu \in D$, 我们需要生成快照数据

$$\{u(\mu_1), u(\mu_2), \dots, u(\mu_{N_{snap}})\}$$

或

$$\{u(\mu_1), \nabla_\mu u(\mu_1), u(\mu_2), \nabla_\mu u(\mu_2), \dots, u(\mu_{N_{snap}}), \nabla_\mu u(\mu_{N_{snap}})\}$$

一种常用的做法是贪心选取:

1. 首先, 随机选取参数 μ_1 ;
2. 求解高维模型 $f(u(\mu_1), \mu_1) = 0$ 得到解 $u(\mu_1)$;
3. 构建快照矩阵 $S = [u(\mu_1)]$ 以及对应的基底 V ;
4. 循环 $i = 2, 3, \dots$
 - (a) 求解降阶模型, 从参数范围 D 中贪心选取后验误差最大的参数 (表明现有的基对于这个参数的近似效果最差)

$$\mu_i = \arg \max_{\mu \in D} \|r(\mu)\| = \arg \max_{\mu \in D} \|f(Vq(\mu), \mu)\|$$

- (b) 求解高维模型 $f(u(\mu_i), \mu_i) = 0$ 得到解 $u(\mu_i)$;

- (c) 扩充快照矩阵 $S \leftarrow [S \quad u(\mu_i)]$, 更新对应的基底 V 。

对于含参数和时间的偏微分方程问题

$$\frac{du(t, \mu)}{dt} = f(u(t, \mu), t, \mu)$$

其中参数 $\mu \in D$, 我们需要生成快照数据

$$\{u(t_i, \mu_j)\}_{i,j}^{N_{snap}}, \quad \text{or} \quad \{u(t_i, \mu_j), \nabla_\mu u(t_i, \mu_j)\}_{i,j}^{N_{snap}}$$

应用贪心法时, 对应的后验误差改为

$$\mu_i = \arg \max_{\mu \in D} \|r(\mu)\| = \arg \max_{\mu \in D} \int_0^T \left\| V \frac{d}{dt} q(t, \mu) - f(Vq(t, \mu), t, \mu) \right\|^2 dt$$

 **Note 2.1.1.** 贪心算法的关键是计算或估计后验误差。在降阶模型中, 有时会单独使用贪心算法, 有时会如上文中这样将 POD 和贪心算法结合使用。 

2.2 平衡截断法和矩匹配方法

平衡截断法和矩匹配的部分所讨论的线性系统模型属于现代控制理论中的内容，和我目前关心的偏微分方程的降阶模型有点差距。这两种降阶方法完全不依赖数据，直接基于系统中已知的大规模矩阵进行降阶。

2.2.1 线性动力系统

考虑线性模型（线性时不变系统，LTI）

$$\begin{aligned}\frac{d}{dt}\mathbf{u}(t) &= A\mathbf{u}(t) + Bw(t), \\ y(t) &= C\mathbf{u}(t) + Dw(t)\end{aligned}$$

其中输入函数 $w: t \mapsto w(t) \in \mathbb{C}^q$ ，状态 $\mathbf{u}(t) \in \mathbb{C}^N$ 以及初值 $\mathbf{u}(t_0) = \mathbf{u}_0$ ，输出函数 $y: t \mapsto y(t) \in \mathbb{C}^p$ 。易得四个矩阵的尺寸依次为

$$A \in \mathbb{C}^{N \times N}, B \in \mathbb{C}^{N \times q}, C \in \mathbb{C}^{p \times N}, D \in \mathbb{C}^{p \times q}$$

下文中会使用四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

来代指一个具体的线性模型。

线性模型的状态 $\mathbf{u}(t)$ 的维度 N 可能很大，对线性模型的降阶目标是得到一个简化系统

$$\begin{aligned}\frac{d}{dt}\mathbf{u}_r(t) &= A_r\mathbf{u}_r(t) + B_rw(t), \\ y_r(t) &= C_r\mathbf{u}_r(t) + D_rw(t)\end{aligned}$$

其中状态 $\mathbf{u}_r$ 的维度 r 大幅降低，输入输出函数维度保持不变，使得 $y_r(t) \approx y(t)$ 。

利用矩阵指数函数可以得到线性模型的解析解

$$\begin{aligned}\mathbf{u}(t, w; t_0, \mathbf{u}_0) &= e^{A(t-t_0)}\mathbf{u}_0 + \int_{t_0}^t e^{A(t-\tau)}Bw(\tau) d\tau \\ y(t) &= Ce^{A(t-t_0)}\mathbf{u}_0 + C \int_{t_0}^t e^{A(t-\tau)}Bw(\tau) d\tau + Dw(t) \\ &:= \phi(t, w; t_0, \mathbf{u}_0).\end{aligned}$$

假设 A 的特征值实部均为负，且输入使下列积分收敛。取 $t_0 \rightarrow -\infty$ ，考虑零初始状态响应，可以得到

$$\begin{aligned}\mathbf{u}(t, w; -\infty, \mathbf{0}) &= \int_{-\infty}^t e^{A(t-\tau)}Bw(\tau) d\tau \\ y(t) &= C \int_{-\infty}^t e^{A(t-\tau)}Bw(\tau) d\tau + Dw(t)\end{aligned}$$

引入核函数 $h(t) \in \mathbb{C}^{p \times q}$

$$h(t) = \begin{cases} Ce^{At}B + \delta(t)D, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

其中 $\delta(t)$ 为 Dirac 函数，可以整理得到

$$y(t) = \int_{-\infty}^{+\infty} h(t-\tau)w(\tau) d\tau = (h * w)(t).$$

这里的卷积关系通过 Laplace 变换就会变成乘积关系

$$(\mathcal{L}y)(\xi) = (\mathcal{L}h)(\xi)(\mathcal{L}w)(\xi)$$

 **Note 2.2.1.** Laplace 变换

$$\mathcal{L}h(\xi) = \int_0^\infty h(\tau)e^{-\xi\tau} d\tau$$

◇

Definition 2.2.1. 对于一般的核函数 $h(t)$ ($t > 0$) 假设有分解形式

$$h(t) = h_0 \delta(t) + h_a(t)$$

其中 $h_0 \delta(t)$ 是脉冲响应部分, $h_a(t)$ 是平滑核函数部分, 将其展开为

$$h_a(t) = h_1 + h_2 \frac{t}{1} + h_3 \frac{t^2}{2!} + \dots \in \mathbb{C}^{p \times q}.$$

参数组 $\Sigma = (h_0, h_1, h_2, \dots)$ 称为 Markov 参数。

Definition 2.2.2 (传递函数). 核函数 $h(t)$ 的 Laplace 变换

$$H(\xi) = (\mathcal{L}h)(\xi) = \int_0^\infty h(t)e^{-\xi t} dt.$$

被称为传递函数。计算可得

$$H(\xi) = (\mathcal{L}h)(\xi) = h_0 + h_1 \xi^{-1} + h_2 \xi^{-2} + \dots$$

回到线性模型中, 易得

$$\Sigma = (D, CB, CAB, CA^2B, \dots, CA^{k-1}B, \dots)$$

这些参数与坐标变换无关, 例如考虑坐标变换 $\mathbf{u} = T\tilde{\mathbf{u}}$ ($\det(T) \neq 0$), 可以得到

$$\frac{d}{dt}\tilde{\mathbf{u}}(t) = T^{-1}AT\tilde{\mathbf{u}}(t) + T^{-1}Bw(t), \quad y(t) = CT\tilde{\mathbf{u}}(t) + Dw(t).$$

此时

$$\tilde{\Sigma} = (D, (CT)(T^{-1}B), (CT)(T^{-1}AT)(T^{-1}B), \dots) = \Sigma.$$

2.2.2 状态分类

考虑四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

所对应的线性模型, 对系统涉及到的状态 $\mathbf{u} \in \mathbb{C}^N$ 进行分类定义。

Definition 2.2.3 (可达状态). 对于线性模型, 称状态 $\tilde{\mathbf{u}} \in \mathbb{C}^N$ 是可达状态, 如果存在一个有限能量的输入函数 w 以及一个有限时间 $\bar{T}$, 使得在该输入作用下且初始状态为零的情况

下, 系统状态能够变为 $\bar{\mathbf{u}}$, 即

$$\bar{\mathbf{u}} = \mathbf{u}(\bar{T}, w; t_0, \mathbf{0}).$$

Definition 2.2.4. 定义所有可达状态组成的集合为 $X^{reach} \subset \mathbb{C}^N$. 称该系统是完全可达的, 如果 $X^{reach} = \mathbb{C}^N$.

Theorem 2.2.1 (可达状态的判定). 给定四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

定义可达矩阵

$$\mathcal{R}(A, B) = \begin{bmatrix} B & AB & A^2B & \cdots & A^{N-1}B \end{bmatrix}$$

以及有限可达 Gramians 矩阵

$$\mathcal{P}(t) = \int_0^t e^{A\tau} B B^H e^{A^H \tau} d\tau \in \mathbb{C}^{N \times N}.$$

那么

$$X^{reach} = \text{Range}(\mathcal{P}(t)) = \text{Range}(\mathcal{R}(A, B)), \quad \forall t > 0$$

Remark 2.2.1. 可达矩阵可以定义为所有 $A^k B$ 按列拼接的结果, 但是根据 Cayley-Hamilton 定理, 第 N 项之后的项没有必要添加。

Theorem 2.2.2 (可达状态的最小能量). 给定四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

假设系统完全可达, 因而 $\mathcal{P}(\bar{T})$ 在 $\bar{T} > 0$ 时正定. 对于状态 $\bar{\mathbf{u}} \in \mathbb{C}^N$, 如果输入函数 $\bar{w}(t)$ 使系统从时刻 0 的零状态在 $\bar{T}$ 时刻到达 $\bar{\mathbf{u}}$, 那么输入函数的能量有下界

$$\|\bar{w}\|^2 = \int_0^{\bar{T}} \bar{w}(t)^H \bar{w}(t) dt \geq \bar{\mathbf{u}}^H \mathcal{P}(\bar{T})^{-1} \bar{\mathbf{u}}.$$

取等的能量最小的输入函数为

$$\bar{w}(t) = B^H e^{A^H(\bar{T}-t)} \mathcal{P}(\bar{T})^{-1} \bar{\mathbf{u}}$$

Remark 2.2.2. 这里涉及到的矩阵内积运算实际应该使用共轭转置, 但是我在笔记中全部简化为共轭。

Remark 2.2.3. 可达状态的定义保证了 $\bar{w}(t)$ 的存在性, 这里直接构造了最优的输入函数。

Definition 2.2.5 (可控状态). 对于线性模型, 称状态 $\bar{\mathbf{u}} \in \mathbb{C}^N$ 是可控状态, 如果存在一个有限能量的输入函数 w 以及一个有限时间 $\bar{T}$, 使得在该输入作用下且初始状态为 $\bar{\mathbf{u}}$ 的情况

下，系统状态能够变为 $\mathbf{0}$ ，即

$$\mathbf{0} = \mathbf{u}(\bar{T}, w; t_0, \bar{\mathbf{u}}).$$

Definition 2.2.6. 定义所有可控状态组成的集合为 $X^{contr} \subset \mathbb{C}^N$ 。称该系统是完全可控的，如果 $X^{contr} = \mathbb{C}^N$ 。

Definition 2.2.7 (不可观测状态). 对于线性模型，称状态 $\bar{\mathbf{u}} \in \mathbb{C}^N$ 是不可观测状态，如果以 $\bar{\mathbf{u}}$ 为初始状态，使用 $w(t) = 0$ 作为输入函数，系统输出的观测值永远为 0，即

$$y(t) = C\mathbf{u}(t, 0; t_0, \bar{\mathbf{u}}) = Ce^{A(t-t_0)}\bar{\mathbf{u}} = 0$$

反之则称 $\bar{\mathbf{u}}$ 为可观测状态。

Definition 2.2.8. 定义所有不可观测状态组成的集合为 $X^{unobs} \subset \mathbb{C}^N$ 。称该系统是完全可观测的，如果 $X^{unobs} = \{\mathbf{0}\}$ 。

Theorem 2.2.3 (可观测状态的判定). 给定四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

定义可观测矩阵

$$\mathcal{O}(A, C) = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{N-1} \end{bmatrix}$$

以及有限可观测 Gramians 矩阵

$$\mathcal{Q}(t) = \int_0^t e^{A^H \tau} C^H C e^{A \tau} d\tau \in \mathbb{C}^{N \times N}.$$

那么

$$X^{unobs} = \text{Ker}(\mathcal{O}(A, C)) = \text{Ker}(\mathcal{Q}(t)), \quad \forall t > 0$$

Theorem 2.2.4. 给定四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

取可观测状态 $\bar{\mathbf{u}} \in \mathbb{C}^N$ 为初始状态，采用 $w(t) = 0$ 作为输入函数，系统输出的观测值 $y(t)$ 的能量为

$$\|y\|^2 = \int_0^{\bar{T}} y(t)^H y(t) dt = \bar{\mathbf{u}}^H \mathcal{Q}(\bar{T}) \bar{\mathbf{u}}.$$

Theorem 2.2.5 (可达性可观测性规范分解 (Canonical decomposition)). 给定四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

存在坐标变换使得变换后的四元组形如

$$\begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix} = \begin{pmatrix} A_{r\bar{o}} & * & * & * & B_{r\bar{o}} \\ & A_{ro} & & * & B_{ro} \\ & & A_{\bar{r}\bar{o}} & * & \\ & & & A_{\bar{r}o} & \\ C_{ro} & & & C_{\bar{r}o} & \tilde{D} \end{pmatrix}$$

使得:

1. $\begin{pmatrix} A_{r\bar{o}} & B_{r\bar{o}} \\ & * \end{pmatrix}$ 是完全可达但不可观测的;
2. $\begin{pmatrix} A_{ro} & B_{ro} \\ C_{ro} & * \end{pmatrix}$ 是完全可达且可观测的;
3. $\begin{pmatrix} A_{\bar{r}\bar{o}} & B_{\bar{r}\bar{o}} \\ C_{\bar{r}\bar{o}} & * \end{pmatrix}$ 是不可达且不可观测的;
4. $\begin{pmatrix} A_{\bar{r}o} \\ C_{\bar{r}o} & * \end{pmatrix}$ 是不可达但完全可观测的。

Remark 2.2.4. 在上述定理中, 等于号的含义是将四元组整体视作一个大矩阵, 将其重新拆分为子矩阵。

2.2.3 平衡截断法

本节假设 A 为 Hurwitz 矩阵 (所有特征值实部均为负), 且系统完全可达、完全可观测。因此下述无限时间 Gramians 收敛且正定, 可以进行求逆和 Cholesky 分解。

考虑可达、可观测 Gramians 矩阵

$$\begin{aligned} \mathcal{P}(t) &= \int_0^t e^{A\tau} B B^H e^{A^H \tau} d\tau \in \mathbb{C}^{N \times N}. \\ \mathcal{Q}(t) &= \int_0^t e^{A^H \tau} C^H C e^{A\tau} d\tau \in \mathbb{C}^{N \times N}. \end{aligned}$$

显然有如下性质: 对于 $t_1 < t_2$

$$\mathcal{P}(t_1) \leq \mathcal{P}(t_2), \quad \mathcal{Q}(t_1) \leq \mathcal{Q}(t_2),$$

取 $t \rightarrow \infty$ 可以得到无限时间的 Gramians 矩阵

$$\begin{aligned} \mathcal{P} &= \mathcal{P}(+\infty) = \int_0^{+\infty} e^{A\tau} B B^H e^{A^H \tau} d\tau \in \mathbb{C}^{N \times N}. \\ \mathcal{Q} &= \mathcal{Q}(+\infty) = \int_0^{+\infty} e^{A^H \tau} C^H C e^{A\tau} d\tau \in \mathbb{C}^{N \times N}. \end{aligned}$$

可以推出 $\mathcal{P}, \mathcal{Q}$ 分别满足 Lyapunov 方程

$$\begin{aligned} A\mathcal{P} + \mathcal{P}A^H + BB^H &= 0, \\ A^H\mathcal{Q} + \mathcal{Q}A + C^HC &= 0. \end{aligned}$$

可以利用 $\mathcal{P}, \mathcal{Q}$ 进行估计:

1. 基于 $\mathcal{P}^{-1}$ 的内积表征了将状态从 $\mathbf{0}$ 转换为可达状态 $\bar{\mathbf{u}}$ 所需的最小能量

$$\|\bar{\mathbf{w}}\|^2 = \int_0^{\bar{T}} \bar{\mathbf{w}}(t)^H \bar{\mathbf{w}}(t) dt \geq \bar{\mathbf{u}}^H \mathcal{P}(\bar{T})^{-1} \bar{\mathbf{u}} \geq \mathbf{u}^H \mathcal{P}^{-1} \mathbf{u} =: \|\mathbf{u}\|_{\mathcal{P}^{-1}}^2.$$

2. 基于 $\mathcal{Q}$ 的内积表示在无输入条件下, 初始的可观测状态 $\mathbf{u}_0$ 相对应的系统的输出所能产生的最大能量

$$\|y\|^2 = \int_0^{\bar{T}} y(t)^H y(t) dt = \bar{\mathbf{u}}^H \mathcal{Q}(\bar{T}) \bar{\mathbf{u}} \leq \mathbf{u}_0^H \mathcal{Q} \mathbf{u}_0 =: \|\mathbf{u}_0\|_{\mathcal{Q}}^2.$$

从状态的重要性角度来理解, 下面两类状态是不重要的:

1. 如果状态 $\mathbf{u}$ 对应的 $\|\mathbf{u}\|_{\mathcal{P}^{-1}}^2$ 很大, 表明这个状态难以到达;
2. 如果状态 $\mathbf{u}$ 对应的 $\|\mathbf{u}\|_{\mathcal{Q}}^2$ 很小, 表明这个状态难以观测;

下面两类状态则是重要的:

1. 最小能量可达状态

$$\arg \min \frac{\mathbf{u}^H \tilde{\mathcal{P}}^{-1} \mathbf{u}}{\mathbf{u}^H \mathbf{u}} = \arg \max \frac{\mathbf{u}^H \tilde{\mathcal{P}} \mathbf{u}}{\mathbf{u}^H \mathbf{u}}$$

2. 最大观测能量状态

$$\arg \max \frac{\mathbf{u}^H \tilde{\mathcal{Q}} \mathbf{u}}{\mathbf{u}^H \mathbf{u}}$$

我们希望舍弃两类不重要的状态, 保留两类重要的状态, 以达到模型降阶的目的。

这些状态的定义和基底的选取相关, 因此我们希望找到一组基, 在这组基下, 可达性和可观测性这两个概念是等价的, 此时称系统处于平衡状态。

考虑坐标变换 $\tilde{\mathbf{u}} = T\mathbf{u}$, 可以得到新的四元组

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix} = \begin{pmatrix} TAT^{-1} & TB \\ CT^{-1} & D \end{pmatrix}$$

还有

$$\tilde{\mathcal{P}} = T\mathcal{P}T^H, \quad \tilde{\mathcal{Q}} = T^{-H}\mathcal{Q}T^{-1}$$

Definition 2.2.9. 称一个坐标变换 T_{bal} 为平衡变换, 如果满足

$$\tilde{\mathcal{P}} = \tilde{\mathcal{Q}}$$

直接推导平衡变换所满足的条件

$$\tilde{\mathcal{P}} = T_{bal}\mathcal{P}T_{bal}^H = T_{bal}^{-H}\mathcal{Q}T_{bal}^{-1} = \tilde{\mathcal{Q}}$$

对 $\mathcal{P}$ 进行 Cholesky 分解

$$\mathcal{P} = LL^H$$

对 $L^H Q L$ 进行特征值分解（这是对称矩阵, 因此可以要求 $K^H K = K K^H = I$ ）

$$L^H Q L = K \Sigma^2 K^H$$

其中 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ 按降序排列。那么, 可以取

$$T_{bal} = \Sigma^{1/2} K^H L^{-1}, \quad T_{bal}^{-1} = L K \Sigma^{-1/2}.$$

此时

$$\tilde{\mathcal{P}} = T_{bal} \mathcal{P} T_{bal}^H = \Sigma^{1/2} K^H L^{-1} L L^H L^{-H} K \Sigma^{1/2} = \Sigma,$$

$$\tilde{\mathcal{Q}} = T_{bal}^{-H} \mathcal{Q} T_{bal}^{-1} = \Sigma^{-1/2} K^H L^H Q L K \Sigma^{-1/2} = \Sigma.$$

记 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_N)$, 称 σ_i 为系统的 Hankel 奇异值。

基于平衡变换 $\tilde{\mathbf{u}} = T_{bal} \mathbf{u}$, 将四元组转化为

$$\begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix} = \begin{pmatrix} T_{bal} A T_{bal}^{-1} & T_{bal} B \\ C T_{bal}^{-1} & D \end{pmatrix}$$

可以根据 Σ 对角元的分布挑选前 k 个分量, 进一步拆分

$$\begin{pmatrix} T_{bal} A T_{bal}^{-1} & T_{bal} B \\ C T_{bal}^{-1} & D \end{pmatrix} = \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_1 \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{B}_2 \\ \tilde{C}_1 & \tilde{C}_2 & \tilde{D} \end{pmatrix}$$

其中下标 1 所对应的块对应最可达和可观测的状态分量, $\tilde{A}_{11} \in \mathbb{C}^{k \times k}$ 。选取低维子空间（这里实际上是掩码矩阵）

$$\tilde{V} = \begin{bmatrix} e_1 & e_2 & \cdots & e_k \end{bmatrix}, \quad \tilde{\mathbf{u}} = \tilde{V} q.$$

得到降阶模型

$$\begin{aligned} \frac{d}{dt} q(t) &= \tilde{A}_{11} q(t) + \tilde{B}_1 w(t), \\ y(t) &= \tilde{C}_1 q(t) + \tilde{D} w(t) \end{aligned}$$

或者使用更标准的斜投影表示

$$\begin{aligned} \frac{d}{dt} q(t) &= (W^H V)^{-1} W^H (A V q(t) + B w(t)), \\ y(t) &\approx C V q(t) + D w(t) \end{aligned}$$

其中

$$W = T_{bal}^H[:, 1:k], \quad V = T_{bal}^{-1}[:, 1:k].$$

Theorem 2.2.6. 对于给定的完全可达可观测的稳定线性系统

$$M = \begin{pmatrix} A & B \\ C & \end{pmatrix}$$

基于平衡截断法得到的降阶模型记作

$$M_r = \begin{pmatrix} A_r & B_r \\ C_r & \end{pmatrix}$$

那么：

1. 模型 M_r 稳定；
2. 记系统 M 的 *Hankel* 奇异值为 $\{\sigma_i\}_{i=1}^m$ (对应重数 η_i)，那么系统 M_r 的 *Hankel* 奇异值为 $\{\sigma_i^r\}_{i=1}^k$ (对应重数 η_i^r)；
3. 两个模型在使用相同的 $\mathbf{u}_0 = \mathbf{0}$ 作为初始值，使用 w 作为输入，得到的观测差异记作 $\Delta y(t; w)$ ，有

$$\sup_w \frac{\sqrt{\int_0^\infty \|\Delta y(t; w)\|^2 dt}}{\sqrt{\int_0^\infty \|w(t)\|^2 dt}} \leq 2(\sigma_{k+1} + \cdots + \sigma_m)$$

2.2.4 矩匹配

仍然考虑线性模型

$$\begin{aligned} \frac{d}{dt} \mathbf{u}(t) &= A\mathbf{u}(t) + Bw(t), \\ y(t) &= C\mathbf{u}(t) + Dw(t) \end{aligned}$$

其中输入函数 $w: t \mapsto w(t) \in \mathbb{C}^q$ ，状态 $\mathbf{u}(t) \in \mathbb{C}^N$ 以及初值 $\mathbf{u}(t_0) = \mathbf{u}_0$ ，输出函数 $y: t \mapsto y(t) \in \mathbb{C}^p$ 。我们考虑一般的 Petrov-Galerkin 投影，可以得到降阶模型

$$\begin{aligned} \frac{d}{dt} q(t) &= (W^H V)^{-1} W^H (AVq(t) + Bw(t)), \\ y(t) &\approx CVq(t) + Dw(t) \end{aligned}$$

其中 V, W 的构造待定。降阶模型的四元组记作

$$\begin{pmatrix} A_r & B_r \\ C_r & D_r \end{pmatrix} = \begin{pmatrix} (W^H V)^{-1} W^H AV & (W^H V)^{-1} W^H B \\ CV & D \end{pmatrix}$$

矩匹配方法的核心思路：构造 V, W 使得降阶模型和原模型的前几个 Markov 参数 (h_i ，也称为矩，记作 $\eta_i(\infty)$) 保持一致。

考虑单输入-单输出模型 ($B = b \in \mathbb{C}^N$, $C^H = c^H \in \mathbb{C}^N$)，那么原始模型的矩为

$$\eta_0(\infty) = D, \quad \eta_m(\infty) = CA^{m-1}b, \quad m = 1, 2, \dots$$

降阶模型的矩为

$$\eta'_0(\infty) = D, \quad \eta'_m(\infty) = CV((W^H V)^{-1} W^H AV)^{m-1} (W^H V)^{-1} W^H b, \quad m = 1, 2, \dots$$

因此，目标是选取 V, W ，使得

$$\eta_m(\infty) = \eta'_m(\infty)$$

这里可以明显看到 Krylov 空间的身影。

Theorem 2.2.7. 设 $V \in \mathbb{C}^{N \times k}$ 满足

$$\text{Range}(V) = \mathcal{K}_k(A, b) = \text{span}(b, Ab, \dots, A^{k-1}b)$$

取 $W \in \mathbb{C}^{N \times k}$ 满足 $W^H V = I$, 那么, 利用 V, W 构造的降阶模型满足

$$\eta_m(\infty) = \eta'_m(\infty), \quad m = 0, 1, \dots, k-1.$$

Proof. 验证一下即可: 对 $m = 0$ 显然, 对 $m > 0$

$$\begin{aligned} \eta'_i(\infty) &= CV((W^H V)^{-1} W^H A V)^{i-1} (W^H V)^{-1} W^H b \\ &= CV(W^H A V)^{m-1} W^H b = C A^{m-1} b. \end{aligned}$$

□

Remark 2.2.5. 这里还有一些变体, 但都是 Krylov 子空间理论和各种算法的应用, 因此不再赘述。

第三章 针对非线性系统的降阶模型

3.1 非线性动力系统

考虑非线性模型

$$\frac{d}{dt}\mathbf{u}(t) = A\mathbf{u}(t) + f(\mathbf{u}(t))$$

采用基于 Galerkin 投影的降阶模型，记状态近似 $\mathbf{u} \approx V\mathbf{q}(t)$ ，其中 $V \in \mathbb{C}^{N \times k}$ ，可以得到

$$\frac{d}{dt}\mathbf{q}(t) = V^H A V \mathbf{q}(t) + V^H f(V\mathbf{q}(t))$$

其中线性部分 $A_r = V^H A V \in \mathbb{C}^{k \times k}$ ，非线性部分 $V^H f(V\mathbf{q}(t))$ 的计算量仍然与 N 成正比，需要额外的一层近似来确保简化后的非线性方程组计算量与 N 无关，常见的方法包括线性化方法和超级缩减方法（Hyper-reduction approaches）。

3.2 轨迹分段线性方法

最简单的做法是对非线性函数进行线性近似

$$f(\mathbf{u}(t)) \approx f(\mathbf{u}_1) + \nabla f(\mathbf{u}_1)(\mathbf{u}(t) - \mathbf{u}_1)$$

我们可以选取多个点进行线性近似，然后进行加权组合

$$f(\mathbf{u}(t)) \approx \sum_{i=1}^s w_i(\mathbf{u}_i) (f(\mathbf{u}_i) + \nabla f(\mathbf{u}_i)(\mathbf{u}(t) - \mathbf{u}_i))$$

其中权重 $\sum_{i=1}^s w_i(\mathbf{u}_i) = 1$ ，就可以得到轨迹分段线性方法 (Trajectory Piecewise Linear Method)：

$$\frac{d}{dt}\mathbf{u}(t) = A\mathbf{u}(t) + f(\mathbf{u}(t)) \approx A\mathbf{u}(t) + \sum_{i=1}^s w_i(\mathbf{u}_i) (f(\mathbf{u}_i) + \nabla f(\mathbf{u}_i)(\mathbf{u}(t) - \mathbf{u}_i))$$

此时降阶模型即

$$\frac{d}{dt}\mathbf{q}(t) = V^H A V \mathbf{q}(t) + \sum_{i=1}^s \tilde{w}_i(\mathbf{q}(t)) (V^H f(\mathbf{u}_i) + V^H \nabla f(\mathbf{u}_i)(V\mathbf{q}(t) - \mathbf{u}_i))$$

对应权重函数满足

$$\sum_{i=1}^s \tilde{w}_i(\mathbf{q}(t)) = 1$$

这里涉及到几个实现细节：

1. 线性化点的选取：基于高维模型的轨迹选取，点的间隔足够远；
2. 低维子空间的基底 V 的选取：POD 或者其他方式；

3. 权重函数的选取，一种参考做法为：

$$\tilde{w}_i(\mathbf{q}(t)) = \frac{\exp\left(-\frac{\beta d_i^2}{d^2}\right)}{\sum_{i=1}^s \exp\left(-\frac{\beta d_i^2}{d^2}\right)},$$

其中 $d_i = \|\mathbf{q}(t) - (V^H V)^{-1} V^H \mathbf{u}_i\|_2$, $d = \min\{d_i\}$ 。

Remark 3.2.1. 这种做法虽然简单，但是有很多实现细节需要调整，例如线性化点的选取非常重要，而且要求可以从高维模型中获取雅可比矩阵。

3.3 离散经验插值方法 (DEIM)

回到非线性项的近似问题

$$\underbrace{V^H}_{k \times N} \underbrace{f(V\mathbf{q}(t))}_{N \times 1}$$

计算复杂度 $\mathcal{O}(2kN)$ 。我们可以挑选若干索引 $\varphi_1, \dots, \varphi_m$ 进行近似

$$\underbrace{V^H}_{k \times N} \underbrace{f(V\mathbf{q}(t))}_{N \times 1} \approx \underbrace{V[\varphi_1, \dots, \varphi_m]^H}_{k \times m} \underbrace{f(V\mathbf{q}(t))[\varphi_1, \dots, \varphi_m]}_{m \times 1}$$

这样可以把计算复杂度降低为 $\mathcal{O}(2km)$ 。定义掩码矩阵

$$P := \begin{bmatrix} e_{\varphi_1} & \dots & e_{\varphi_m} \end{bmatrix} \in \mathbb{C}^{N \times m}.$$

可得

$$V^H f(V\mathbf{q}(t)) \approx V^H P P^H f(V\mathbf{q}(t))$$

其中 $P^H f(V\mathbf{q}(t))$ 实际只需要计算 f 的 m 个分量，如果 $f(\mathbf{u})$ 的计算是逐个分量进行的，那么还可以整理为

$$P^H f(V\mathbf{q}(t)) = f(P^H V\mathbf{q}(t))$$

此时只需要计算 $V\mathbf{q}(t)$ 的 m 个分量。

现在的问题是如何选取掩码矩阵 P ：假设 f 也可以在一个低维空间中近似，选取基底

$$V_f = \begin{bmatrix} u_1 & \dots & u_{m'} \end{bmatrix} \in \mathbb{C}^{N \times m'}$$

对应的低维近似为

$$f(V\mathbf{q}(t)) \approx V_f f_r(t), \quad f_r(t) \in \mathbb{C}^{m'}$$

低维近似 $f_r(t)$ 可以通过投影获得：

1. 最直接的做法是正交投影

$$V_f^H (f - V_f f_r) = 0$$

可以解得

$$f_r = (V_f^H V_f)^{-1} V_f^H f$$

2. 也可以通过基于掩码矩阵的斜投影获得：给定 $P \in \mathbb{C}^{N \times m}$ 以及任意的 $f \in \mathbb{C}^N$ ，要求

$$P^H f = P^H V_f f_r,$$

可以在广义逆的意义下解得（ m 和 m' 可能未必相等，例如 $m \geq m'$ 使用更多的索引）

$$f_r = (P^H V_f)^\dagger P^H f$$

虽然正交投影可以到达最优的 L^2 投影误差，但是基于掩码矩阵的斜投影具有计算量小的独特优势。我们希望选取掩码矩阵 P ，使得斜投影误差

$$f - V_f f_r = f - V_f (P^H V_f)^\dagger P^H f$$

尽可能小。基于贪心算法的 DEIM 算法流程见 1（取 $m = m'$ ）。

Algorithm 1: Discrete empirical interpolation method (DEIM)

Input: $\{u_\ell\}_{\ell=1}^m \subset \mathbb{C}^n$ linearly independent.

Output: $\vec{\varphi} = [\varphi_1, \dots, \varphi_m]^\top$. $\varphi_i \in \{1, \dots, n\}$

1 $[|\rho|, \varphi_1] = \max\{|u_1|\}$;

2 $V_1 = [u_1]$, $P_1 = [e_{\varphi_1}]$;

3 **for** $\ell = 2$ **to** m **do**

4 Solve $(P_{\ell-1}^H V_{\ell-1})\alpha = P_{\ell-1}^H u_\ell$ for $\alpha \in \mathbb{C}^m$;

5 Compute the residual r_ℓ :

$$r_\ell = u_\ell - V_{\ell-1}\alpha = u_\ell - V_{\ell-1}(P_{\ell-1}^H V_{\ell-1})^{-1} P_{\ell-1}^H u_\ell$$

6 $[|\rho_\ell|, \varphi_\ell] = \max\{|r_\ell|\}$;

7 $V_\ell = [V_{\ell-1} \ u_\ell]$, $P_\ell = [P_{\ell-1} \ e_{\varphi_\ell}]$;

8 **end**

Theorem 3.3.1. 对于任意向量 $f \in \mathbb{C}^N$ 以及正交基底 $V_f \in \mathbb{C}^{N \times m}$ ，使用 DEIM 算法选取掩码矩阵 P ，可以得到

$$\hat{f} = V_f (P^H V_f)^{-1} P^H f.$$

近似误差满足

$$\|f - \hat{f}\|_2 \leq \|(P^H V_f)^{-1}\|_2 \|(I - V_f V_f^H)f\|_2$$

其中 $\|(I - V_f V_f^H)f\|_2$ 是最优的 L^2 投影误差，系数 $\|(P^H V_f)^{-1}\|_2$ 有上界

$$\|(P^H V_f)^{-1}\|_2 \leq (1 + \sqrt{2N})^{m-1} \|u_1\|_\infty^{-1}$$

第四章 数据驱动的降阶模型

4.1 谱分解

考虑线性常微分方程组

$$\frac{d}{dt}\mathbf{u}(t) = A\mathbf{u}(t).$$

假设方阵 $A \in \mathbb{C}^{N \times N}$ 有完备的特征系统

$$Av_k = \lambda_k v_k, \quad k = 1, 2, \dots, N$$

那么可以展开 $\mathbf{u}(t)$

$$\mathbf{u}(t) = \sum_{k=1}^N a_k(t) v_k.$$

代入可得

$$\frac{d}{dt}a_k(t) = \lambda_k a_k(t), \quad \implies \quad a_k(t) = a_k(0)e^{\lambda_k t}.$$

考虑离散时间的线性常微分方程

$$\mathbf{u}(m+1) = A\mathbf{u}(m)$$

同理可得

$$\mathbf{u}(m) = \sum_{k=1}^N a_k(m) v_k$$

代入可得

$$a_k(m+1) = \lambda_k a_k(m), \quad \implies \quad a_k(m) = a_k(0)\lambda_k^m.$$

4.2 Koopman 算子

Koopman 算子部分参考 Brunton, Steven L. “Notes on koopman operator theory”. Universität von Washington, Department of Mechanical Engineering, Zugriff 30 (2019).

4.2.1 连续时间系统

考虑连续时间的非线性模型

$$\frac{d}{dt}\mathbf{u} = f(\mathbf{u}), \quad \mathbf{u} \in \mathcal{M} \subset \mathbb{C}^N$$

其中 $\mathcal{M}$ 是 $\mathbb{C}^N$ 空间中的一个流形。

定义算子（参数 $t \geq 0$ ）

$$F_t : \mathbf{u}(t_0) \mapsto \mathbf{u}(t_0 + t) = \mathbf{u}(t_0) + \int_{t_0}^{t_0+t} f(\mathbf{u}(\tau)) d\tau$$

显然算子的复合满足如下关系

$$F_{t+s} = F_t \circ F_s = F_s \circ F_t$$

因此，集合 $\{F_t\}_{t \geq 0}$ 在算子复合运算下构成半群。

考虑由观测函数 $g : \mathcal{M} \rightarrow \mathbb{C}$ 全体构成的空间 $\mathfrak{S}$ ，定义（连续时间系统）Koopman 算子

$$\mathcal{K}_t : \mathfrak{S} \rightarrow \mathfrak{S}, \quad \mathcal{K}_t g := g \circ F_t$$

其中 $\circ$ 是算子复合，那么

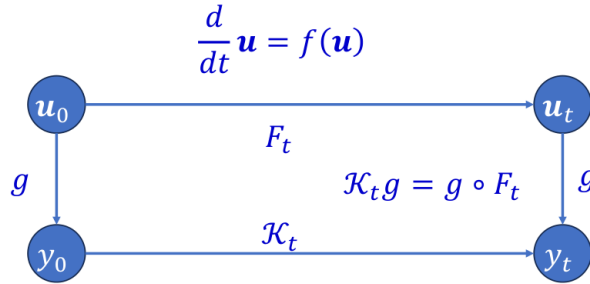
$$g(\mathbf{u}(t_0 + t)) = g \circ F_t(\mathbf{u}(t_0)) = \mathcal{K}_t(g(\mathbf{u}(t_0)))$$

显然 Koopman 算子是线性算子

$$\mathcal{K}_t(c_1 g + c_2 h) = (c_1 g + c_2 h) \circ F_t = c_1 g \circ F_t + c_2 h \circ F_t = c_1 \mathcal{K}_t g + c_2 \mathcal{K}_t h$$

并且构成半群

$$\mathcal{K}_{s+t} g = g \circ F_{t+s} = g \circ (F_t \circ F_s) = (g \circ F_t) \circ F_s = \mathcal{K}_t(g \circ F_s) = \mathcal{K}_s \mathcal{K}_t g$$



Note 4.2.1. Koopman 算子是无限维线性算子，即使 f 是非线性的，这也是 Koopman 算子理论的核心思想：把非线性系统转换为线性系统，代价是问题从有限维变成了无限维。 $\diamond$

对 $\mathcal{K}_t g$ 在 $t = 0$ 求导以定义算子 $\mathcal{K}$

$$\mathcal{K} : \mathfrak{S} \rightarrow \mathfrak{S}, \quad \mathcal{K} g := \left. \frac{d\mathcal{K}_t g}{dt} \right|_{t=0} = \lim_{t \rightarrow 0} \frac{\mathcal{K}_t g - g}{t} = \lim_{t \rightarrow 0} \frac{g \circ F_t - g}{t}.$$

可以得到 g 所对应的方程

$$\left. \frac{d}{dt} g(\mathbf{u}(t_0 + t)) \right|_{t=0} = \left. \frac{d(\mathcal{K}_t g(\mathbf{u}(t_0)))}{dt} \right|_{t=0} = \mathcal{K}(g(\mathbf{u}(t_0)))$$

$\mathcal{K}$ 被称为 $\{\mathcal{K}_t\}$ 的无穷小生成元，简写为

$$\frac{d}{dt} g = \mathcal{K} g.$$

定义满足如下方程的非零观测函数 φ 和复数 λ 为（连续时间系统）Koopman 特征函数和特征值

$$\frac{d}{dt} \varphi(\mathbf{u}) = \mathcal{K}(\varphi(\mathbf{u})) = \lambda \varphi(\mathbf{u})$$

根据链式法则可得

$$\frac{d}{dt}\varphi(\mathbf{u}) = \nabla\varphi(\mathbf{u}) \cdot \frac{d\mathbf{u}}{dt} = \nabla\varphi(\mathbf{u}) \cdot f(\mathbf{u})$$

因此（连续时间系统）Koopman 特征函数 φ 和特征值 λ 的获取需要求解如下方程

$$\nabla\varphi(\mathbf{u}) \cdot f(\mathbf{u}) = \lambda\varphi(\mathbf{u})$$

Proposition 4.2.1. 对于 Koopman 特征函数 φ 和特征值 λ ，有如下性质

$$\mathcal{K}_t\varphi(\mathbf{u}(t_0)) = \varphi(\mathbf{u}(t_0 + t)) = e^{\lambda t}\varphi(\mathbf{u}(t_0)).$$

Proposition 4.2.2. 考虑连续时间系统的 Koopman 算子 $\mathcal{K}_t$ ，如果 (λ_1, φ_1) 和 (λ_2, φ_2) 是两个 Koopman 特征函数以及对应特征值，那么 $(\lambda_1 + \lambda_2, \varphi_1\varphi_2)$ 也是 Koopman 特征函数以及对应特征值。

Proof. 验证即可

$$\mathcal{K}(\varphi_1(u)\varphi_2(u)) = \varphi_1(Fu)\varphi_2(Fu) = e^{\lambda_1 t}e^{\lambda_2 t}\varphi_1(u)\varphi_2(u) = e^{(\lambda_1 + \lambda_2)t}\varphi_1(u)\varphi_2(u)$$

□

Example 4.2.1. 考虑非线性系统

$$\frac{du}{dt} = f(u) = -u - u^3, \quad u \in \mathbb{C}.$$

Koopman 算子

$$\mathcal{K}_t(g(u(t_0))) = g \circ F_t(u(t_0)) = g\left(u(t_0) + \int_{t_0}^{t_0+t} f(u(\tau))d\tau\right)$$

Koopman 特征函数和特征值满足

$$\frac{d\varphi}{dt} = \varphi'(u) \cdot f(u) = \varphi'(u) \cdot (-u - u^3) = \lambda\varphi(u).$$

可以解得

$$\varphi_k(u) = \left(\frac{u}{\sqrt{1+u^2}}\right)^k, \quad \lambda_k = -k, \quad k = 1, 2, \dots$$

4.2.2 离散时间系统

考虑离散时间的非线性模型

$$\mathbf{u}(m+1) = F(\mathbf{u}(m)), \quad \mathbf{u} \in \mathcal{M} \in \mathbb{C}^N.$$

记算子复合为 $F^m(\mathbf{u}) = F(F^{m-1}(\mathbf{u}))$ ，显然有 $\mathbf{u}(m) = F^m(\mathbf{u}(0))$ 。

考虑由观测函数 $g: \mathcal{M} \rightarrow \mathbb{C}$ 全体构成的空间 $\mathfrak{S}$ ，定义（离散时间系统）Koopman 算子

$$\mathcal{K}_{\Delta t}: \mathfrak{S} \rightarrow \mathfrak{S}, \quad \mathcal{K}_{\Delta t}g := g \circ F$$

其中 $\circ$ 是算子复合，那么

$$g(\mathbf{u}(k+m)) = g \circ F^m(\mathbf{u}(k)) = \mathcal{K}_{\Delta t}^m(g(\mathbf{u}(k)))$$

定义满足如下方程的非零观测函数 φ 和复数 λ 为（离散时间系统）Koopman 特征函数和特征值

$$\varphi(\mathbf{u}(k+1)) = \mathcal{K}_{\Delta t}(\varphi(\mathbf{u}(k))) = \lambda \varphi(\mathbf{u}(k))$$

（离散时间系统）Koopman 特征函数 φ 和特征值 λ 的获取需要求解如下方程

$$\mathcal{K}_{\Delta t}(\varphi(\mathbf{u})) = \lambda \varphi(\mathbf{u})$$

Proposition 4.2.3. 对于 Koopman 特征函数 φ 和特征值 λ ，有如下性质

$$\mathcal{K}_{\Delta t}^k \varphi(\mathbf{u}(m)) = \varphi(\mathbf{u}(m+k)) = \lambda^k \varphi(\mathbf{u}(m)).$$

Proposition 4.2.4. 考虑离散时间系统的 Koopman 算子 $\mathcal{K}_{\Delta t}$ ，如果 (λ_1, φ_1) 和 (λ_2, φ_2) 是两个 Koopman 特征函数以及对应特征值，那么 $(\lambda_1 \lambda_2, \varphi_1 \varphi_2)$ 也是 Koopman 特征函数以及对应特征值。

Proof. 验证即可

$$\mathcal{K}_{\Delta t}(\varphi_1(u)\varphi_2(u)) = \varphi_1(Fu)\varphi_2(Fu) = (\lambda_1 \lambda_2) \varphi_1(u)\varphi_2(u)$$

□

Example 4.2.2. 考虑线性模型

$$\mathbf{u}(k+1) = F(\mathbf{u}(k)) = A\mathbf{u}(k), \quad \mathbf{u}(k) \in \mathbb{C}^N.$$

Koopman 算子

$$\mathcal{K}_{\Delta t}(g(\mathbf{u}(k))) = g(F(\mathbf{u}(k))) = g(A\mathbf{u}(k)).$$

Koopman 特征函数和特征值满足

$$\mathcal{K}_{\Delta t} \varphi(\mathbf{u}) = \varphi(A\mathbf{u}) = \lambda \varphi(\mathbf{u})$$

$A \in \mathbb{C}^{N \times N}$ 的特征值和左特征向量满足

$$\mathbf{w}_j^H A = \lambda_j \mathbf{w}_j^H.$$

记 $\varphi_j(\mathbf{u}) = \mathbf{w}_j^H \mathbf{u}$ ，代入可得

$$\mathcal{K}_{\Delta t} \varphi_j(\mathbf{u}) = \varphi_j(A\mathbf{u}) = \mathbf{w}_j^H A\mathbf{u} = \lambda_j \mathbf{w}_j^H \mathbf{u} = \lambda_j \varphi_j(\mathbf{u}).$$

因此，Koopman 特征函数和特征函数包括

$$(\lambda_j, \varphi_j(\mathbf{u}) := \mathbf{w}_j^H \mathbf{u})$$

根据前一个命题，还有其他的 Koopman 特征函数，例如

$$(\lambda_i \lambda_j, \varphi_{ij}(\mathbf{u}) := (\mathbf{w}_i^H \mathbf{u})(\mathbf{w}_j^H \mathbf{u}))$$

4.3 Koopman 模态

给定 Koopman 算子的特征函数 $\{\varphi_j\}_{j=1}^{\infty}$ 以及对应的特征值 $\{\lambda_j\}_{j=1}^{\infty}$ ，如果观测函数 g 可以展开为

$$g = \sum_{j=1}^{\infty} v_j(g) \varphi_j(u),$$

其中的系数 $v_j(g) \in \mathbb{C}$ 被称为与观测函数 g 相关的 Koopman 模态。由于 Koopman 算子是线性算子，利用观测函数 g 的展开式和 Koopman 特征函数的性质可得：

1. 对于连续时间系统，由于 $\mathcal{K}_t \varphi_j = e^{\lambda_j t} \varphi_j$ ，有

$$\mathcal{K}_t g = \sum_{j=1}^{\infty} e^{\lambda_j t} \varphi_j v_j(g).$$

2. 对于离散时间系统，由于 $\mathcal{K}_{\Delta t}^m \varphi_j = \lambda_j^m \varphi_j$ ，有

$$\mathcal{K}_{\Delta t}^m g = \sum_{j=1}^{\infty} \lambda_j^m \varphi_j v_j(g).$$

前面讨论的观测函数都是标量函数 $g: \mathcal{M} \rightarrow \mathbb{C}$ ，我们可以取

$$g^{(i)}(\mathbf{u}) = u_i$$

直接提取状态的分量，进而组合得到恒等观测函数 $\text{id}(\mathbf{u}) = \mathbf{u}$ 。记恒等观测函数 id 的 Koopman 模态为 $\mathbf{v}_j(\text{id})$ （向量值），可得

1. 对于连续时间系统有

$$\mathbf{u}(t) = \mathcal{K}_t \text{id}(\mathbf{u}(0)) = \sum_{j=1}^{\infty} e^{\lambda_j t} \varphi_j(\mathbf{u}(0)) \mathbf{v}_j(\text{id}).$$

2. 对于离散时间系统有

$$\mathbf{u}(m) = \mathcal{K}_{\Delta t}^m \text{id}(\mathbf{u}(0)) = \sum_{j=1}^{\infty} \lambda_j^m \varphi_j(\mathbf{u}(0)) \mathbf{v}_j(\text{id}).$$

Example 4.3.1. 考虑非线性系统

$$\frac{d}{dt} \mathbf{u} = F(\mathbf{u}) = \begin{bmatrix} \mu_1 u_1 \\ \mu_2 (u_2 - u_1^2) \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \in \mathbb{C}^2.$$

Koopman 算子的无穷小生成元

$$\mathcal{K} = \mu_1 u_1 \frac{\partial}{\partial u_1} + \mu_2 (u_2 - u_1^2) \frac{\partial}{\partial u_2}$$

包括如下 Koopman 特征函数和特征值

$$\begin{aligned} \lambda_1 &= \mu_1, & \varphi_1(\mathbf{u}) &= u_1 \\ \lambda_2 &= \mu_2, & \varphi_2(\mathbf{u}) &= u_2 + \frac{\mu_2}{2\mu_1 - \mu_2} u_1^2 \\ \lambda_3 &= 2\mu_1, & \varphi_3(\mathbf{u}) &= u_1^2 \end{aligned}$$

可以得到恒等算子的分解如下

$$\text{id} = \varphi_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \varphi_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \varphi_3 \begin{bmatrix} 0 \\ -\frac{\mu_2}{2\mu_1 - \mu_2} \end{bmatrix}$$

因此

$$\begin{aligned} \mathbf{u}(t) &= e^{\lambda_1 t} \varphi_1(\mathbf{u}(0)) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + e^{\lambda_2 t} \varphi_2(\mathbf{u}(0)) \begin{bmatrix} 0 \\ 1 \end{bmatrix} + e^{\lambda_3 t} \varphi_3(\mathbf{u}(0)) \begin{bmatrix} 0 \\ -\frac{\mu_2}{2\mu_1 - \mu_2} \end{bmatrix} \\ &= e^{\mu_1 t} u_1(0) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + e^{\mu_2 t} \left(u_2(0) + \frac{\mu_2}{2\mu_1 - \mu_2} u_1(0)^2 \right) \begin{bmatrix} 0 \\ 1 \end{bmatrix} + e^{2\mu_1 t} u_1(0)^2 \begin{bmatrix} 0 \\ -\frac{\mu_2}{2\mu_1 - \mu_2} \end{bmatrix} \\ &= \begin{bmatrix} e^{\mu_1 t} u_1(0) \\ e^{\mu_2 t} \left(u_2(0) + \frac{\mu_2}{2\mu_1 - \mu_2} u_1(0)^2 \right) - e^{2\mu_1 t} u_1(0)^2 \frac{\mu_2}{2\mu_1 - \mu_2} \end{bmatrix} \end{aligned}$$

Example 4.3.2. 考虑线性模型

$$\mathbf{u}(m+1) = F(\mathbf{u}(m)) = A\mathbf{u}(m), \quad \mathbf{u}(m) \in \mathbb{C}^N.$$

记 A 的特征值和左右特征向量为

$$\mathbf{w}_j^H A = \lambda_j \mathbf{w}_j^H, \quad A \mathbf{v}_j = \lambda_j \mathbf{v}_j,$$

不妨假设满足 $\mathbf{w}_j^H \mathbf{v}_j = \delta_{ij}$ 。Koopman 算子的特征值和特征函数包括

$$\lambda_j, \quad \varphi_j(\mathbf{u}) = \mathbf{w}_j^H \mathbf{u}.$$

恒等算子可以分解为

$$\text{id} = \sum_{j=1}^N \mathbf{v}_j \varphi_j.$$

因此

$$\mathbf{u}(m) = \mathcal{K}_{\Delta x}^m \text{id}(\mathbf{u}(0)) = \sum_{j=1}^N \lambda_j^m \mathbf{v}_j \varphi_j(\mathbf{u}(0)) = \sum_{j=1}^N \lambda_j^m \mathbf{v}_j \mathbf{w}_j^H \mathbf{u}(0).$$

4.4 Koopman 算子的有限维近似

考虑观测函数 g 组成的函数空间 $\mathfrak{S}$ 的 N 维子空间 $\mathfrak{S}_N \subset \mathfrak{S}$, 不妨设基底为

$$\Psi = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_N \end{bmatrix}, \quad \mathfrak{S}_N = \text{span}(\Psi)$$

定义到 $\mathfrak{S}_N$ 的投影映射

$$\Pi: \mathfrak{S} \rightarrow \mathfrak{S}_N, \quad \Pi(g) := \Psi \Gamma(g)$$

其中 $\Gamma(g) \in \mathbb{C}^N$ 表示提取投影在每一个基上的系数。

对于 Koopman 算子 $\mathcal{K}$, 可以自然地限制在 $\mathfrak{S}_N$ 上, 定义

$$\mathcal{K}_N = \Pi \mathcal{K}|_{\mathfrak{S}_N}: \mathfrak{S}_N \rightarrow \mathfrak{S}_N, \quad \mathcal{K}_N g := \Pi \mathcal{K} \Pi g = \Psi \Gamma \mathcal{K} \Psi \Gamma g$$

可以用 $\mathcal{K}_N$ 来近似 $\mathcal{K}$ 。定义

$$K : \mathbb{C}^N \rightarrow \mathbb{C}^N, \quad Ka := \Gamma \mathcal{K} \Psi a$$

矩阵 $K \in \mathbb{C}^{N \times N}$ 是 $\mathcal{K}_N$ 的矩阵表示形式，被称为系统的 Koopman 矩阵， $\mathbf{a} \in \mathbb{C}^N$ 是观测函数在给定基底下的系数表示。Koopman 矩阵的含义为 Koopman 算子在观测函数有限子空间 $\mathfrak{S}_N$ 的基底 Ψ 下对观测函数的作用

$$\mathbf{a}(m+1) = K\mathbf{a}(m)$$

4.5 动力学模态分解 (DMD)

动力学模态分解 (DMD) 是一种数据驱动的方法，从时空数据集中提取动态模式，典型应用场景是流体力学中的非定常流场模态分析。

假设已有大量的均匀时间序列对应的状态数据

$$\mathbf{u}(t_0), \mathbf{u}(t_1), \mathbf{u}(t_2), \dots, \mathbf{u}(t_n)$$

其中时间 $t_m = m\Delta t$ ，状态 $\mathbf{u} \in \mathbb{C}^N$ 为 N 维向量（例如在所有网格点上的取值）。记 $\mathbf{u}(m) = \mathbf{u}(t_m)$ 将其视作离散系统，Koopman 算子为

$$\mathcal{K}g(\mathbf{u}(m)) = g(\mathbf{u}(m+1))$$

定义获取 $\mathbf{u}$ 的第 j 个分量的观测函数

$$\psi_j : \mathcal{M} \rightarrow \mathbb{C}, \quad \psi_j(\mathbf{u}) := u_j.$$

记分量取值函数全体张成的有限维空间为 $\mathfrak{S}_N$ (N 为分量个数)，显然分量取值函数全体可以构造 $\mathfrak{S}_N$ 的一个基底。（显然有 $id \in \mathfrak{S}_N$ ）

动力学模态分解 (Dynamic model decomposition) 的核心思想为：希望获取近似的 Koopman 矩阵 $K \in \mathbb{C}^{N \times N}$ ，使得

$$\mathbf{u}(m+1) \approx K\mathbf{u}(m), \quad m = 1, 2, \dots, n-1$$

定义矩阵

$$X = \begin{bmatrix} \mathbf{u}(0) & \mathbf{u}(1) & \cdots & \mathbf{u}(n-1) \end{bmatrix} \in \mathbb{C}^{N \times n}, \quad Y = \begin{bmatrix} \mathbf{u}(1) & \mathbf{u}(2) & \cdots & \mathbf{u}(n) \end{bmatrix} \in \mathbb{C}^{N \times n}.$$

问题归结为矩阵方程的近似求解：寻找“最合适的”矩阵 $K \in \mathbb{C}^{N \times N}$ ，使得

$$Y \approx KX$$

这个问题其实仍然是不明确的，因为“最合适”并没有一个明确的含义。如果默认最小二乘意义下，寻找矩阵 $K \in \mathbb{R}^{N \times N}$ ，使得

$$K = \arg \min_{B \in \mathbb{R}^{N \times N}} \|Y - BX\|_F^2$$

可以使用 Moore-Penrose 广义逆给出解

$$K = YX^\dagger = YV\Sigma^{-1}U^H \in \mathbb{C}^{N \times N}, \quad (X = U\Sigma V^H)$$

这里需要的 SVD 分解代价与数据规模有关，但是只需要计算一次即可，实际上我们还需要完整地求解 K 的特征系统。

Remark 4.5.1. 这里和语言大模型对数据的处理很类似，拟合的模型都是基于当前状态预测下一个状态。

假设已经得到 Koopman 矩阵 K ，那么相当于得到近似的线性系统

$$\mathbf{u}(m+1) = K\mathbf{u}(m)$$

可以得到显式解

$$\mathbf{u}(m) = \sum_{j=1}^N \lambda_j^m \mathbf{v}_j \mathbf{w}_j^H \mathbf{u}(0)$$

其中 λ_j 和 $\mathbf{w}_j, \mathbf{v}_j$ 是 K 的特征值和左右特征向量。（实践中可以舍弃不重要的部分，不需要进行完整的 N 项求和）

这里描述的 DMD 没有涉及到数据的降维操作，实践中直接计算并不合适，算法不够稳定，可以加上一些截断和投影操作，舍弃不重要的部分。例如对 X 的 SVD 分解加上低秩截断，得到

$$X = U\Sigma V^H \approx U_r \Sigma_r V_r^H =: X_r.$$

也可以解释为向 $\text{span}(U_r)$ 投影，对 Y 也可以投影到 $\text{span}(U_r)$ ，得到

$$Y \approx U_r U_r^H Y =: Y_r.$$

那么

$$K \approx K_r := Y_r X_r^\dagger = U_r U_r^H Y V_r \Sigma_r^{-1} U_r^H$$

可以利用基底 U_r 缩小线性系统的规模（ $\mathbf{u} \approx U_r \tilde{\mathbf{u}}$ ）

$$\tilde{\mathbf{u}}(m+1) = \tilde{K}_r \tilde{\mathbf{u}}(m), \quad \tilde{K}_r = U_r^H Y V_r \Sigma_r^{-1} \in \mathbb{C}^{r \times r}$$

这里得到的是离散时间推进矩阵。若快照间隔为 Δt ，在选定适当对数分支且矩阵对数存在时，连续生成元可写为 $\log(\tilde{K}_r)/\Delta t$ ；相应的连续增长率为 $\log(\lambda)/\Delta t$ 。这里还涉及到对特征值和特征向量的具体计算，还有一些近似的变体，这里不再赘述。

动力学模态分解适合处理的是近似线性系统，不适合处理强非线性系统。可以进行一些修改，调整观测函数子空间的基函数，例如引入非线性的傅里叶函数，多项式函数等，进而得到扩展的动力学模态分解。相应地，需要修改 X 和 Y 的定义：

$$X = [\mathbf{a}(0) \quad \mathbf{a}(1) \quad \cdots \quad \mathbf{a}(n-1)] \in \mathbb{C}^{N' \times n}, \quad Y = [\mathbf{a}(1) \quad \mathbf{a}(2) \quad \cdots \quad \mathbf{a}(n)] \in \mathbb{C}^{N' \times n}.$$

其中 $\mathbf{a}(m) \in \mathbb{C}^{N'}$ 是 $\mathbf{u}(m)$ 在观测函数子空间 $\mathfrak{S}_{N'}$ 的固定基底下的系数表示。假设已经得到 Koopman 矩阵 K ，那么相当于得到近似的线性系统

$$\mathbf{a}(m+1) = K\mathbf{a}(m)$$

可以得到显式解

$$\mathbf{a}(m) = \sum_{j=1}^N \lambda_j^m \mathbf{v}_j \mathbf{w}_j^H \mathbf{a}(0)$$

其中 λ_j 和 $\mathbf{w}_j, \mathbf{v}_j$ 是 K 的特征值和左右特征向量。

 **Note 4.5.1.** 通过 Koopman 算子理论来解释 DMD 实在是太绕了，如果不考虑 Koopman 算子理论，完全可以用更简单直接的线性代数语言叙述。 

附录 A 数值线性代数基础

A.1 奇异值分解

Theorem A.1.1 (Singular Value Decomposition (SVD)). 给定矩阵 $A \in \mathbb{C}^{N \times M}$, 存在两个酉方阵 $U \in \mathbb{C}^{N \times N}$, $V \in \mathbb{C}^{M \times M}$, 使得

$$A = U \Sigma V^H$$

其中 $\Sigma \in \mathbb{R}^{N \times M}$ 是对角矩阵, 对角线元素为

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(N,M)} \geq 0.$$

称 $\{\sigma_i\}$ 为矩阵 A 的奇异值, U, V 的列向量分别为 A 的左奇异向量和右奇异向量

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_M \end{bmatrix}.$$

如果记 $\text{rank}(A) = r$, 那么显然有 $\sigma_r > 0$ 以及 $\sigma_{r+1} = \cdots = \sigma_N = 0$ 。定义

$$U_r = \begin{bmatrix} u_1 & u_2 & \cdots & u_r \end{bmatrix} \in \mathbb{C}^{N \times r}, \quad V_r = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix} \in \mathbb{C}^{M \times r}.$$

以及 $\Sigma_r = \text{diag}(\sigma_1, \dots, \sigma_r)$, 可以得到精简的 SVD 表示

$$A = U_r \Sigma_r V_r^H.$$

也可以把矩阵 A 展开为 r 个秩一矩阵的和

$$A = \sum_{i=1}^r \sigma_i u_i v_i^H.$$

Proposition A.1.2. 对于矩阵 $A \in \mathbb{C}^{N \times M}$, 以及 $k < r = \text{rank}(A)$, 如果 $\sigma_k(A) > \sigma_{k+1}(A)$, 那么下面两个最优化问题

$$\min_{\text{rank}(X) \leq k} \|A - X\|_2, \quad \min_{\text{rank}(X) \leq k} \|A - X\|_F.$$

均在 $X = A_k = \sum_{i=1}^k \sigma_i u_i v_i^H$ 取到最优解:

$$\min_{\text{rank}(X) \leq k} \|A - X\|_2 = \|A - A_k\|_2 = \sigma_{k+1}(A),$$

$$\min_{\text{rank}(X) \leq k} \|A - X\|_F = \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^r \sigma_i(A)^2}.$$

如果 $\sigma_k(A) = \sigma_{k+1}(A)$, 那么最优解不唯一。

A.2 关于 SVD 低秩截断

对于矩阵的 SVD 截断是一个常见的低秩近似操作：设矩阵 $A \in \mathbb{R}^{m \times n}$ 的精简 SVD 分解为

$$A = U_r \Sigma_r V_r^T.$$

取前 k 个主要奇异值可以得到低秩近似

$$A \approx A_k = U_k \Sigma_k V_k^T.$$

可以从几个不同的角度来理解 SVD 截断操作的含义：

1. 如果把 A 理解为矩阵整体，一个直接的问题就是低秩矩阵近似

$$\min\{\|A - B\|_F \mid \text{rank}(B) \leq k\}.$$

可以证明 $A_k = U_k \Sigma_k V_k^T$ 是这个问题的解

$$\|A - U_k \Sigma_k V_k^T\|_F = \min\{\|A - B\|_F \mid \text{rank}(B) \leq k\}$$

如果 $\sigma_k > \sigma_{k+1}$ ，解还是唯一的。

2. 如果把 A 理解为向量组 $A = [a_1 \ a_2 \ \cdots \ a_n]$ ，一个直接的问题就是寻找 k 维子空间 $\mathcal{V}_k$ ，使得所有向量都差不多落在这个子空间中（投影误差和最小）

$$\min\{\|A - UU^T A\|_F \mid U^T U = I_k\}$$

可以证明 $U = U_k$ 是这个问题的解：因为 $\text{rank}(UU^T A) \leq k$ ，化归到上面讨论的问题

$$\|A - UU^T A\|_F \geq \|A - U_k \Sigma_k V_k^T\|_F$$

令 $U = U_k$ 即可取到等号

$$UU^T A = U_k U_k^T A = U_k \Sigma_k V_k^T.$$

Remark. 其实还有一些变体，例如求子空间使得保留的能量最大化

$$\max\{\|Q^T A\|_F \mid Q^T Q = I_k\}$$

此时的解仍然可以令 $Q = U_k$ 取到。