Model reduction: past and present

Jiaxi Zhao

23th Oct, 2024

Part I: Existing methods

Classical methods:

* Balanced Proper Orthogonal Decomposition (POD)
* Galerkin projection
* Discrete empirical interpolation method (DEIM)
* Koopman operator inspired methods
* Tensor-based methods

Modern methods:

* Nonlinear ROM
* Non-intrusive ROM via operator inference
* Temporal coarsening

Words at the beginning

We will focus on **scientific** time series modeling. Model reduction is related to lots of other terminologies such as modal analysis, reduced-order modeling, etc.

The key feature of time series modeling:

* Stability issue

* Extrapolation or interpolation?

$$\mathbf{X}_1 \to \mathbf{X}_2 \to \cdots \mathbf{X}_n,$$

$$t_1 \to \mathbf{X}_{t_1}, t_2 \to \mathbf{X}_{t_2}, \cdots$$

Reduced-order modeling: Past

* Balanced Proper Orthogonal Decomposition (POD)
* Galerkin projection
* Discrete empirical interpolation method (DEIM)
* Koopman operator inspired methods
* Tensor-based methods[1]

[1] Benner, Peter, et al., eds. Model reduction and approximation: theory and algorithms. Society for Industrial and Applied Mathematics, 2017.

Modal analysis: POD

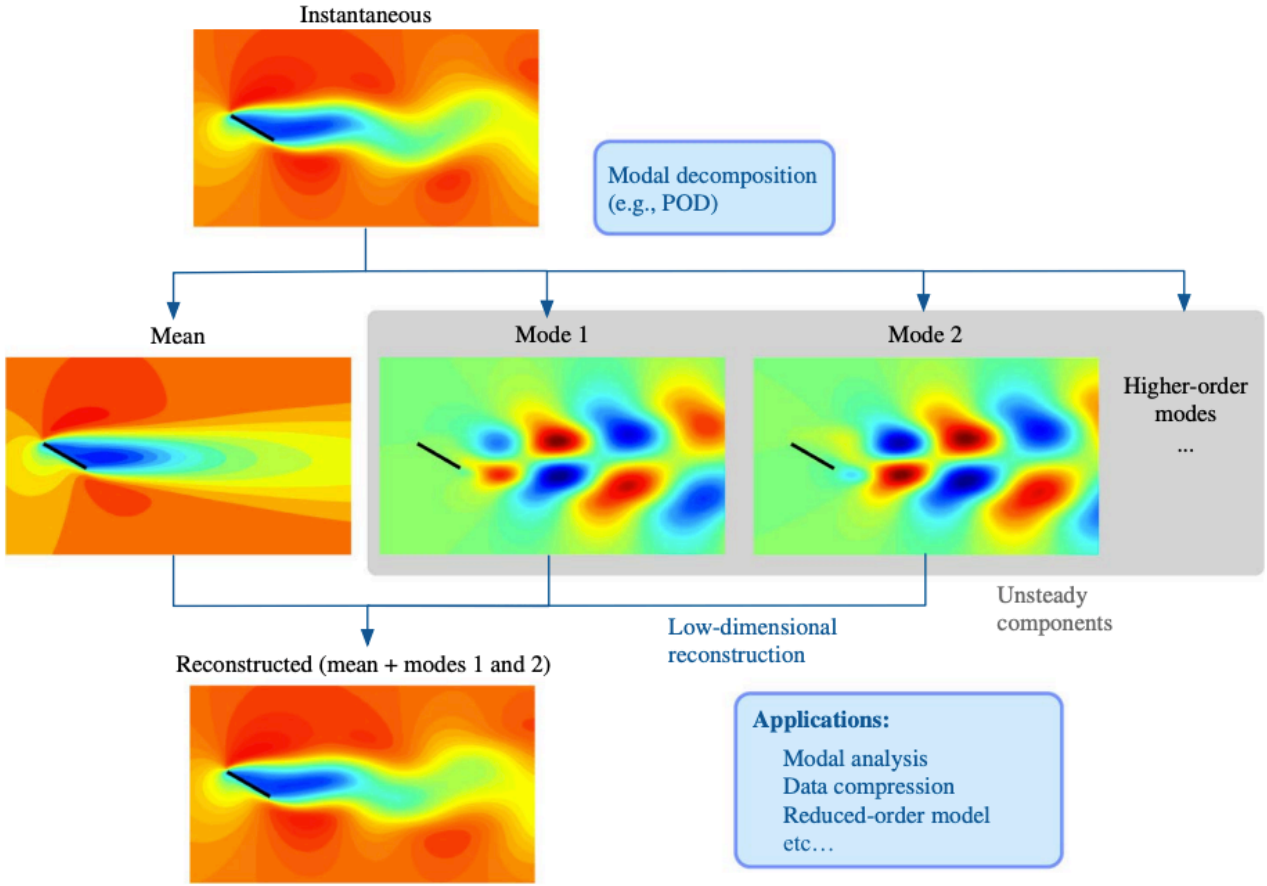

Figure: Modal decomposition of two-dimensional incompressible flow over a flat-plate wing $Re = 100, \alpha = 30$. This example shows complex nonlinear separated flow being well represented by only two POD modes and the mean flowfield. Visualized are the streamwise velocity profiles.[2]

Balanced Transformation

Let us consider the following control system

$$\frac{d}{dt}\mathbf{x}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \quad \mathbf{y}(t) = C\mathbf{x}(t).$$

The key observation is that any invertible transformation $\widetilde{\mathbf{x}} = V\mathbf{x}$ will result in an equivalent system with different POD basis. For this system, the controllability and observability Grammians are defined as

$$W_c = \int_0^\infty e^{At}BB^T e^{A^T t}dt, \quad W_o = \int_0^\infty e^{A^T t}C^T Ce^{At}dt.$$

Balanced transformation V is chosen so that the W_c, W_o are diagonal and equal. [3]

[3] Willcox, Karen, and Jaime Peraire. "Balanced model reduction via the proper orthogonal decomposition." AIAA journal 40.11 (2002): 2323-2330.

Balanced POD

Under the transformation V, two Grammians will transform according to

$$\widetilde{W}_c = V^{-1} W_c V^{-T}, \quad \widetilde{W}_o = V^T W_o V.$$

Then their product transforms as

$$\widetilde{W}_c W_o = V^{-1} W_c W_o V.$$

Projection-based ROM

We consider two types of problem as follows:

$$\frac{d}{dt}\mathbf{x}(t) = A\mathbf{x}(t) + N(\mathbf{x}(t)),$$
$$0 = A_\mu \mathbf{x}(\mu) + N_\mu(\mathbf{x}(\mu)), \quad \mathbf{x} \in \mathbb{R}^{n\times n}.$$

In both systems, $N(\cdot)$ represents the nonlinearity. Given any reduced basis functions of order k, orthogonal projection operator onto this basis is denoted as V_k with reduced system

$$\frac{d}{dt}\widetilde{\mathbf{x}}(t) = V_k^T A V_k \widetilde{\mathbf{x}}(t) + V_k^T N(V_k \widetilde{\mathbf{x}}(t)),$$
$$0 = V_k^T A_\mu V_k \widetilde{\mathbf{x}}(\mu) + V_k^T N_\mu(V_k \widetilde{\mathbf{x}}(\mu)), \quad \widetilde{\mathbf{x}} \in \mathbb{R}^{k\times n}.$$

DEIM

The nonlinear term still remains huge amount of computation:

$$V_k^T N(V_k \widetilde{\mathbf{x}}(t)), \quad \widetilde{J}_N(\mathbf{x}(\mu)) = V_k^T J_F(V_k \widetilde{\mathbf{x}}(\mu)) V_k.$$

The idea is to project this nonlinear term further onto a low-dimensional subspace spanned by $\{\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_m\}$ which is obtained by applying POD to the nonlinear snapshots obtained from the original full-order system.

$$N(V_k \widetilde{\mathbf{x}}(t)) = \mathbf{U} c(t).$$

Interpolation method

4

[4] Amsallem, David, and Charbel Farhat. "Interpolation method for adapting reduced-order models and application to aeroelasticity." AIAA journal 46.7 (2008): 1803-1813.

Difficulties of model reduction

* Nonlinearity, e.g. convection

* Transient modeling and unsteady, especially for long time prediction and turbulence

Draw-back of linear-subspace ROM

In particular, linear-subspace ROMs can be expected to produce low-dimensional models with high accuracy[5]only if the problem admits a fast decaying Kolmogorov n-width (e.g., diffusion-dominated problems).

$$d_n(\mathcal{M}) := \inf_{\mathcal{S}_n} \sup_{f} \inf_{g \in \mathcal{S}_n} \|f - g\|.$$

Unfortunately, many problems of interest exhibit a slowly decaying Kolmogorov n-width (e.g., advection-dominated problems).

[5]Binev, Peter, et al. "Convergence rates for greedy algorithms in reduced basis methods." SIAM journal on mathematical analysis 43.3 (2011): 1457-1472.

Koopman operator

Methods related to the Koopman operator are related to the dynamics of the operator, which is also approximated via a linear dynamics

* Extended Dynamical Model Decomposition (EDMD)
* EDMD-DL
* parametric Koopman

ROM: Present

* Nonlinear ROM
* Non-intrusive ROM via operator inference
* Temporal coarsening

Nonlinear trial manifold: learn the reduced basis

Nonlinear trial manifold[6]

$$\widetilde{\mathbf{x}}(t;\mu) = \mathbf{x}_{ref}(\mu) + g(\widehat{\mathbf{x}}(t;\mu)),$$

where $\mathbf{x}_{ref}(\mu)$ denotes the parametrized reference state specified according to the initial condition and $g : \mathbb{R}^p \to \mathbb{R}^n$ denotes the nonlinear parameterization function refered to as *decoder*. The reduced dynamics can be obtained via chain rule:

$$\frac{d}{dt}\widetilde{\mathbf{x}}(t;\mu) = J_g(\widehat{\mathbf{x}}(t;\mu))\frac{d}{dt}\widehat{\mathbf{x}}(t;\mu).$$

[6]Lee, Kookjin, and Kevin T. Carlberg. "Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders." Journal of Computational Physics 404 (2020): 108973.

Time-continuous residual minimization

The model can be written using the residue function

$$\mathbf{r}(\mathbf{v}, \mathbf{x}, t, \mu) = \mathbf{v} - f(\mathbf{x}, t, \mu).$$

Based on this, we can define the equation for the reduced model as

$$\frac{d}{dt}\widehat{\mathbf{x}}(t;\mu) = \arg\min_{\mathbf{v}\in\mathbb{R}^p} \|\mathbf{r}(J_g(\widehat{\mathbf{x}}(t;\mu))\mathbf{v}, \mathbf{x}_{ref}(\mu) + g(\widehat{\mathbf{x}}(t;\mu)), t, \mu)\|$$

Based on this, the truncation error analysis of the ROM can also be performed using approximation theory of the function spaces.

Operator inference: Learn the reduced operator

Temporal coarsening

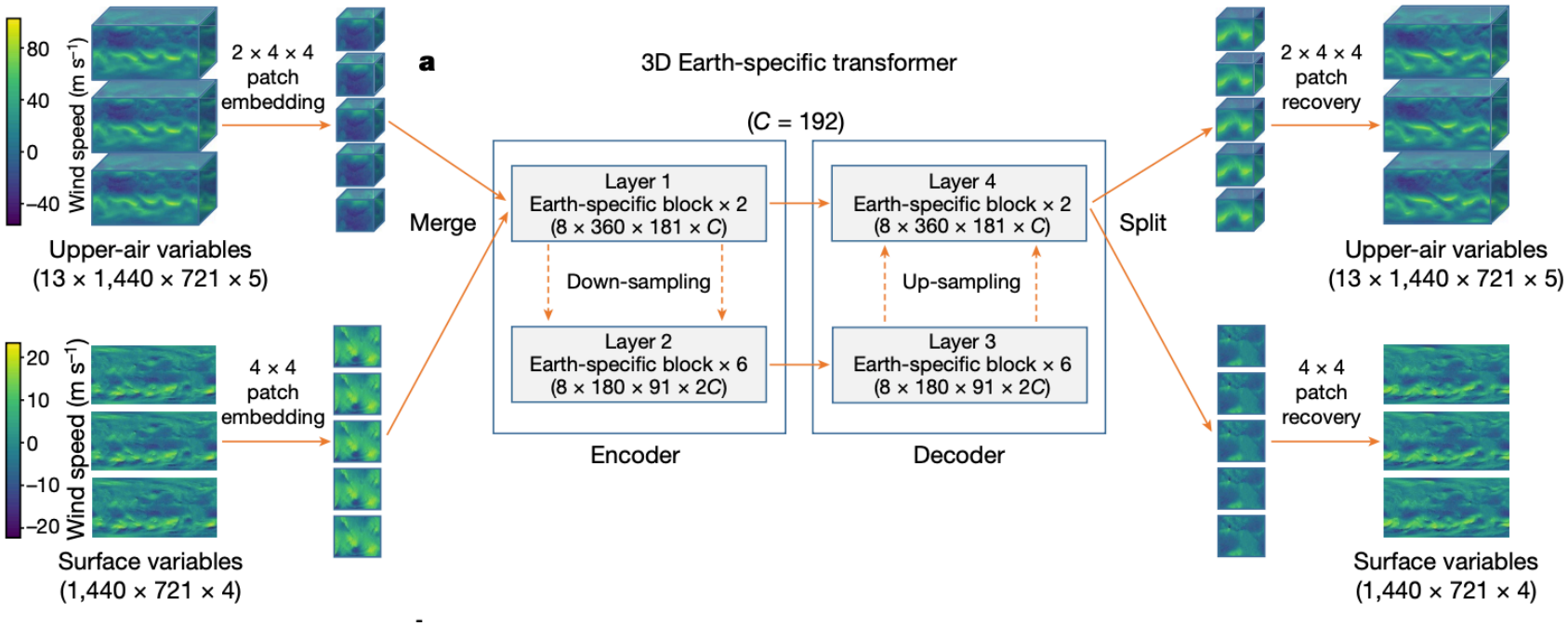

Figure: 3DEST architecture. Based on the standard encoder–decoder design of vision transformers, we adjusted the shifted-window mechanism and applied an Earth-specific positional bias.[7]

[7] Bi, Kaifeng, et al. "Accurate medium-range global weather forecasting with 3D neural networks." Nature 619.7970 (2023): 533-538.

How to do long time prediction?

One of the bottleneck for ROM is the long time prediction accuracy: e.g. for weather forecasting, most data-driven models outperform numerical weather prediction over the 0-7 days regime but quickly

Several methods to perform time series prediction:

* Hierarchical temporal aggregation
* Manifold regularization
* Nonlinear stability issue, especially compared with classical numerical stability

Operator inference ROM

Mesh-based $\Longrightarrow$ Mesh-free

Another kind of nonlinear ROM is based on operator inference. A heuristic: Classical mesh-based solver amounts to solve the high dimesional mapping between the discretization on the huge mesh, e.g. $\mathbb{R}^{N\times N\times N} \to \mathbb{R}^{N\times N\times N}$, how about considering directly $\mathbb{R}^3 \to \mathbb{R}$, which is usually a nonlinear map[8].

Can be viewed as learning the reduced basis and operator simultaneously

[8] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

Operator inference ROM

More over, the parameter can also be fitted into this framework by encoding it as a latent vector[9]

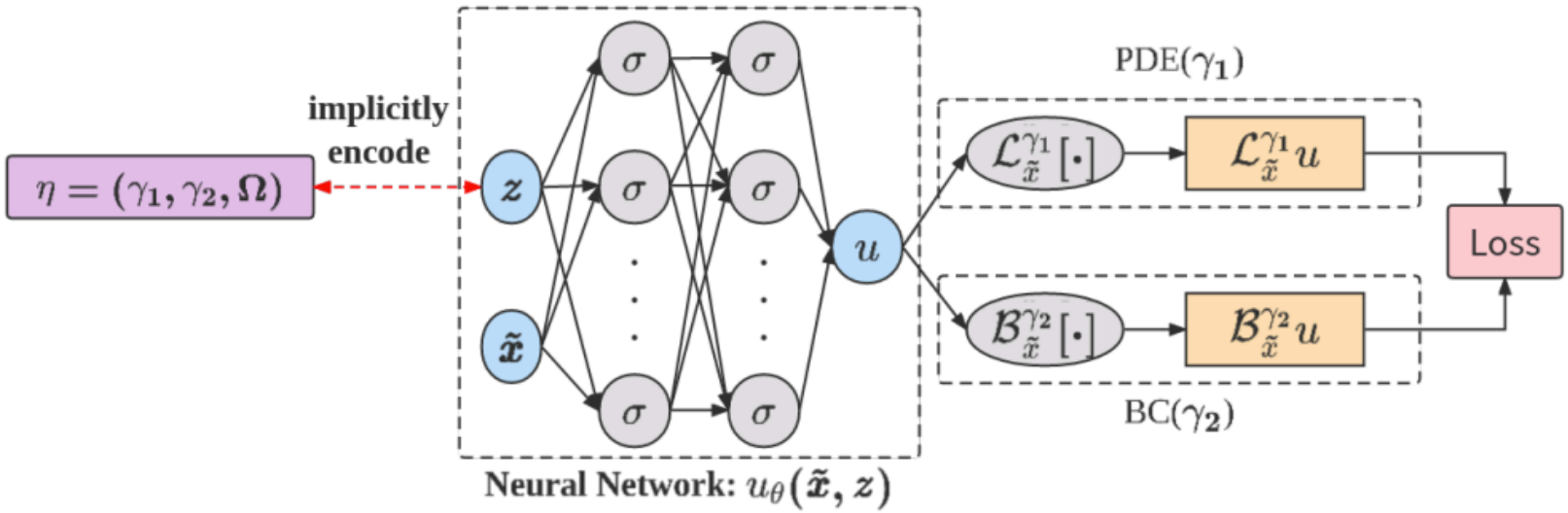

Figure: Architecture of Meta-Auto-Decoder..[10]

[10] Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Relation to the sequence modeling

Given that present ROM are more and more similar to the sequence modeling in lots of CS application, i.e. non-intrusive method, similar transformer network. I personally think it worth to think carefully about their relationship.

* Seq2Seq seems still not prevalent in scientific time series modeling.

* Stability and out-of-distribution issue

Part II:

Consider the parametrized PDE of the following form:

$$\boldsymbol{F}(\mathbf{u};\boldsymbol{\mu}) = 0.$$

A reduction operator $\mathcal{A}$ leads to the reduced field $\overline{\mathbf{u}} = \mathcal{A}(\mathbf{u})$:.
Then, the closure problem amounts to resolve the term

$$\boldsymbol{C}(\boldsymbol{u},\overline{\mathbf{u}};\boldsymbol{\mu}) = \overline{\boldsymbol{F}}(\mathbf{u};\boldsymbol{\mu}) - \boldsymbol{F}(\overline{\mathbf{u}};\boldsymbol{\mu})$$

Caveat: in reduced order modeling, people are "close the equation" instead of learning the closure model. This is different as closure model means that to increase the dimension of the descriptive variable so that with the closure variable the dynamics can be completely determined by the new state variable. While closing the equation just means to make sure there is no residue of the equation.

Part III: A "case" study on collective variable (CV)[11]

* Transfer operator framework
* Effective dynamics
* Algorithms

[10] Zhang, Wei, and Christof Schütte. "On finding optimal collective variables for complex systems by minimizing the deviation between effective and full dynamics." arXiv preprint arXiv:2405.02001 (2024).

Basic settings

Consider a stationary, ergodic, and discrete in time Markov chain $X_n \in \mathbb{R}^d$ with transition probability $p(x, y)$ and invariant measure $\pi(dx)$.

The transfer operator associated with the dynamics is defined as

$$(\mathcal{T}f)(x) = \int_{\mathbb{R}^d} f(y)p(x, y)dy.$$

In deterministic dynamics, the transfer operator coincides with the Koopman operator.

The Dirichlet form is defined as

$$\mathcal{E}(f, g) = \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, y)(f(x) - f(y))(g(x) - g(y))\pi(x)dxdy.$$

If the dynamics is reversible, the transfer operator is self-adjoint and the Dirichlet form is symmetric.

Spectrum and timescale

The spectrum of the transfer operator is related to the timescale of the dynamics and is an important quantity to be studied.

Theorem

If $p(x,y) > 0, \pi(x) > 0$, $\mathcal{T}$ *is self-adjoint, and further*

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x,y)p(y,x)dxdy < \infty,$$

Then, we have that $\mathcal{T}$ *is a compact (Hilbert-Schmidt) operator whose non-zero spectrum are all eigenvalues. Moreover, any real eigenvalue belongs to* $[-1,1]$ *and the only eigenfunction of* $\lambda = 1$ *is the constant function.*

This is the functional analysis generalization for the Perron-Frobenius theorem for stochastic matrices.

Collective variable

A commonly adopted strategy is to utilize the fact that the dynamics of high-dimensional systems, e.g. molecular systems in MD and materials science, can often be well characterized using only a few observables, often called collective variables (CVs) or reaction coordinates, of the systems. For example, temperature, pressure, entropy, Gibbs free energy is used in the thermodynamical description of the ideal gas.

CV can be viewed from the perspective of dimension reduction. PCA, one of the most important dimension reduction method, choose map from the first several singular (eigen-) vectors. So a primary choice is to chose CV as the eigenfunction of the transfer operator.

Transfer operator eigenfunction

The eigenfunction of the transfer operator can be characterized by the following method:

Theorem

The solution of the following optimization problem

$$\min_{f_1,\ldots,f_m} \sum_{i=1}^{m} \omega_i \mathcal{E}(f_i)$$

under the constraint that the basis $f_1, \ldots, f_m$ are orthognormal is given by the first m eigenfunctions of the transfer operator.

However, it is hard to numerically solve this optimization problem if the dimension of the underlying space is high and to enforce the orthognormality constraint become hard to implement.

Design the CV map: I

The Dirichlet form can be estimated according to the sample from the invariant distribution as:

Theorem

The CV map defined by the first eigenfunction of the transfer operator minimizes the averaged of quadratic variations:

$$\min_f \mathcal{E}(f) \approx \min_f \lim_{N\to\infty} \frac{1}{2N} \sum_{n=0}^{2N-1} \|f(X_{i+1}) - f(X_i)\|_\omega^2 ,$$

among all the orthognormal basis f.

The weights ω are introduced in order to remove the non-uniqueness of the minimizer of due to permutation.
Interpretation: ξ should be a slow variable such that the observations $\xi(X_0), \xi(X_1), \ldots$ vary slowly (in most of the time unless essential transitions occur) as the dynamics X_n evolves.

Transition rate

Commitor function is one of the well-known object to be studied in the rare event simulation. It is defined as

$$q(x) = \mathcal{P}(\text{starting from } x,\ X_n \text{ enters } B \text{ before it enters } A.). \quad (1)$$

There is another quantity called the transition rate k_{AB}, which is formally defined as the ratio of the time spent in transition between A and B to the whole trajectory.

Theorem

Commitor function is the minimizer of the Dirichlet energy under the constraint that

$$f\big|_A = 0, \quad f\big|_B = 1. \quad (2)$$

The minimum value is given by the transition rate k_{AB}.

Effective dynamics

Suppose we have a CV map $\xi : \mathbb{R}^d \to \mathbb{R}^k$, define the level set $\Sigma_z := \{x \in \mathbb{R}^d, \xi(x) = z\}$. $\mu_z(dx) \in \mathcal{P}(\Sigma_z)$ is the conditional distribution given z. $\widetilde{\mu} \in \mathcal{P}(\mathbb{R}^k)$ is the pushforward of the measure μ by ξ.

Definition (Effective dynamics)

The effective dynamics of X_n associated to the CV map $\xi : \mathbb{R}^d \to \mathbb{R}^k$ is defined as the Markoc process Z_n in $\mathbb{R}^k$ with transition probability $\widetilde{p}(z, \widehat{z})$:

$$\begin{aligned}\widetilde{p}(z,\widehat{z}) &= \int_{\Sigma_z} \left[\int_{\mathbb{R}^d} p(x,y)\delta(\xi(y) - \widehat{z})dy \right] \mu_z(dx) \\ &= \frac{1}{Q(z)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x,y)\delta(\xi(x) - z)\delta(\xi(y) - \widehat{z})\pi(dx)dxdy.\end{aligned}$$

Variational characterization

The transition probability of the effective dynamics can be viewed as an approximation of the original dynamics in the following sense:

$$p(x, y) = p(\widehat{z}|x)p(y|\widehat{z}, x),$$
$$\overline{p}(x, y) = p_\theta(\widehat{z}|\xi(x))p_\varphi(y|\widehat{z}),$$

Theorem (Variational formulation)

The minimizer of the following problem

$$\arg\min_{\overline{p}\in\Theta^\xi} \mathbb{E}_{x\sim\mu}\left[\mathrm{KL}(p(x,\cdot)||\overline{p}(x,\cdot))\right]$$

is given by $p_\theta = \widetilde{p}, p_\varphi = p_z$ *where* $p_z = \frac{d\mu_z}{d\sigma_{\Sigma_z}}$.

Design the CV map: II

Theorem

Choose the CV map ξ to minimize the KL divergence as

$$\min_{\xi, \overline{p} \in \Theta^{\xi}} \mathbb{E}_{x \sim \mu} \left[\mathrm{KL}(p(x, \cdot) || \overline{p}(x, \cdot)) \right]$$

Personal comment: this is somehow similar to the latent variable generative model such as VAE with ELBO loss.

It would also be inspiring to work out a toy case where we can exactly identify the minimizer of the problem, e.g. μ is Gaussian.

Error estimate

Theorem (Transition rates)

Assume that the sets $A, B \subset \mathbb{R}^d$ are related to the sets $\widetilde{A}, \widetilde{B} \subset \mathbb{R}^k$ throught the CV map ξ via $A = \xi^{-1}(\widetilde{A}), B = \xi^{-1}(\widetilde{B})$. Let q, k_{AB} denote the commitor and the transition rate of X_n from A to B respectively.

$$\widetilde{k}_{\widetilde{A}\widetilde{B}} = k_{AB} + \mathcal{E}(q - \widetilde{q} \circ \xi).$$

How should define a good CV?

Do we really need to achieve a low reconstruction error or a low KL-divergence for a good CV? The answer is not clear. For example, as a CV, the commitor function distinguishs between two attractor basin even if it is only an one-dimensional CV.

How can we characterize the classical CV such as the pressure, temperature, entropy, etc. Are they really minimizing the reconstruction error or the KL-divergence?

Another approach is to start from one well-known CV such as density for the study of water and trying to ask what will be a good extra CV to be used together with the water.