Variational approximation of dynamical systems using machine learning

Frank Noé (FU Berlin)
frank.noe@fu-berlin.de
**Aim:** learn model of dynamical system from observations $x_t, x_{t+\tau}$

How do we choose?

- Method / Representation (DMD, EDMD, TICA, VAMPnets, time-Autoencoder, ...)
- Type of Basis Set (polynomial, cos/sin, characteristic functions, ...)
- Number of Basis Functions
- Type of Kernel in a Kernel approach
- Regularization Hyperparameters
- Type of Neurons and Architecture of Neural Net
**Cost or loss function** $C$ quantifies performance of network with parameters $\theta$ to predict observations $X$.

\[
\hat{\theta} = \arg\min_{\theta} C(X, Y, \theta)
\]

Minimizing cost $C \equiv$ maximizing score $-C$. Often we just write $C(\theta)$.

**Standard approach** (general and flexible):
- Parameter optimization: Training data
- Hyperparameter optimization: Validation data
- Quantifying model performance: Test data
- Details: **Statistical Estimator Theory** (e.g., Vapnik and others)
Simulating biological timescales at atomic resolution

Microsecond MD Trajectories
Simulating biological timescales at atomic resolution

Microsecond MD Trajectories
Mathematical theory:
Schütte et al, *J Comp Phys* 1999,
Also: Weber, Deuflhard, Friesecke, Dellnitz …

Developments for high-throughput molecular dynamics:
Noé, Pande, Swope, Hummer (mid 2000’s)
Dynamical operators

Propagator

\[ \rho_{t+\tau} = \mathcal{P}_\tau \rho_t = \int p_\tau(y \mid x) \rho_t(x) \, dx \]

Transfer operator / Perron-Frobenius operator

= propagator for densities \( u(x) = \frac{\rho(x)}{\pi(x)} \) with stationary density \( \pi(x) \).

\[ u_{t+\tau} = \mathcal{T}_\tau u_t = \int \frac{\pi(x)}{\pi(y)} p_\tau(y \mid x) \rho_t(x) \, dx \]

Koopman operator

Adjoint to \( \mathcal{P} \), adjoint to \( \mathcal{T} \) with respect to \( \pi \)

\[ f_{t+\tau} = \mathcal{K}_\tau f_t = \int p_\tau(y \mid x) f_t(y) \, dy = \mathbb{E}[f_{t+\tau}(x)] \]

with detailed balance: \( \pi(x)p_\tau(y \mid x) = \pi(y)p_\tau(x \mid y) \) we have \( \mathcal{K}_\tau \equiv \mathcal{T}_\tau \)

Slow processes (unique equilibrium distribution, detailed balance)

Backward propagator

\[ \rho_\tau = \mathcal{T}(\tau) \rho_0 \]

Spectral decomposition

\[ \rho_\tau = \sum_{i=1}^{\infty} e^{-\tau \kappa_i} \langle \psi_i | \rho_0 \rangle \psi_i + \text{fast part} \]

Eigenvalues / timescales \( \kappa_i^{-1} \)

Processes:

The first $m$ eigenfunctions $\psi_1, \ldots, \psi_m$ are the solution to the problem

$$\max_{f_1, \ldots, f_m} \sum_{i=1}^{m} \mathbb{E} [f_i(x_t) f_i(x_{t+\tau})]$$

s.t. \quad \mathbb{E} [f_i(x_t)^2] = 1

$$\mathbb{E} [f_i(x_t) f_j(x_{t+\tau})] = 0, \text{ for } i \neq j$$

and the maximum value is the sum of $\lambda_1, \ldots, \lambda_m$

Properties:

1. $\psi_i$ and $\psi_j$ are uncorrelated for $i \neq j$.
2. $\psi_i$ are the directions of slow kinetics with maximal autocorrelations $\mathbb{E}_\mu [\psi_i(x_t) \psi_i(x_{t+\tau})] = \lambda_i(\tau)$.
3. Population changes along $\psi_i$ coordinates decay with $\lambda_i(\tau) = e^{-\frac{\tau}{\tau_i}}$.
4. For every other set of functions, the eigenvalues will be underestimated $\hat{\lambda_i}(\tau) \leq \lambda_i(\tau)$.
Method of linear variation

**Ansatz:** Define Basis set $\chi = [\chi_1(\mathbf{x}), \ldots, \chi_n(\mathbf{x})]^T$ and seek the linear expansions:

$$\hat{\psi}_i(\mathbf{x}) = \sum_j r_{ij} \chi_j(\mathbf{x})$$
Variational approach for reversible Markov processes: Estimator

1. Define

$$X_0 = \begin{bmatrix} \chi_1(x_0) & \cdots & \chi_n(x_0) \\ \vdots & \ddots & \vdots \\ \chi_1(x_{T-\tau}) & \cdots & \chi_n(x_{T-\tau}) \end{bmatrix}, \quad X_\tau = \begin{bmatrix} \chi_1(x_\tau) & \cdots & \chi_n(x_\tau) \\ \vdots & \ddots & \vdots \\ \chi_1(x_T) & \cdots & \chi_n(x_T) \end{bmatrix}$$

2. Empirical covariance matrices: $C^0$ and $C^\tau$ with:

$$C^0 = X_0^T X_0, \quad C^\tau = X_0^T X_\tau$$

3. Solve

$$C^\tau r_i = C^0 \hat{\lambda}_i r_i$$

4. The projections

$$\Psi = XR$$

approximate the transfer operator eigenfunctions on the sampled configurations $x_t$.

Comparison between methods


Noé & Nüske *MMS* 11, 635-655 (2013)


also: Noé, Pande, Hummer, Weber, Swope, …

Variational approach for Markov processes (VAMP)

Koopman operator

\[ \mathcal{K}_\tau f(x) = \mathbb{E}[f(x_{t+\tau}) | x_t = x] = \int p_\tau(x, y) f(y) \, dy \]

Wu and Noé, \texttt{arXiv}:1707.04659 (2017)
Variational approach for Markov processes (VAMP)

Koopman operator

\[ \mathcal{K}_\tau f (x) = \mathbb{E} [f (x_{t+\tau}) | x_t = x] \]
\[ = \int p_\tau (x, y) f (y) \, dy \]

Singular value decomposition:

\[ \mathcal{K}_\tau f = \sum_i \sigma_i \langle \phi_i, f \rangle_{\rho_1} \psi_i \]

Variational approach for Markov processes (VAMP)

Koopman operator

\[ K_\tau f(x) = \mathbb{E}[f(x_{t+\tau}) | x_t = x] \]
\[ = \int p_\tau(x,y) f(y) \, dy \]

Singular value decomposition:

\[ K_\tau f = \sum_i \sigma_i \langle \phi_i, f \rangle_{\rho_1} \psi_i \]

- \( \rho_0, \rho_1 \): empirical distribution of \( x_t, x_{t+\tau} \)
- If data are in equilibrium: stationary distribution \( \mu = \rho_0 = \rho_1 \)
- \( \{\phi_i\} \) and \( \{\psi_i\} \) are both orthonormal bases with respect to \( \langle \cdot, \cdot \rangle_{\rho_1} \) and \( \langle \cdot, \cdot \rangle_{\rho_0} \),
- \( \sigma_i \) denotes the \( i \)th largest singular value.

Variational approach for Markov processes (VAMP)

Theorem VAMP variational principle. The $k$ dominant singular components of a Koopman operator are the solution of the following maximization problem:

$$
\sum_{i=1}^{k} \sigma_i^r = \max_{f,g} \mathcal{R}_r [f, g],
\text{s.t. } \langle f_i, f_j \rangle_{\rho_0} = 1_{i=j},
\langle g_i, g_j \rangle_{\rho_1} = 1_{i=j},
$$

(10)

where $r \geq 1$ can be any positive integer. The maximal value is achieved by the singular functions $f_i = \psi_i$ and $g_i = \phi_i$ and

$$
\mathcal{R}_r [f, g] = \sum_{i=1}^{k} \langle f_i, \mathcal{K}_r g_i \rangle_{\rho_0}
$$

(11)

is called the VAMP-$r$ score of $f$ and $g$. 

1. Compute

\[
\begin{align*}
C_{00} &= \frac{1}{T-\tau} X^T X \\
C_{01} &= \frac{1}{T-\tau} X^T Y \\
C_{11} &= \frac{1}{T-\tau} Y^T Y
\end{align*}
\]

with

\[
\begin{align*}
X &= (\chi_0(x_1), \chi_0(x_2), \ldots, \chi_0(x_{T-\tau}))^T \\
Y &= (\chi_1(x_{1+\tau}), \chi_1(x_{2+\tau}), \ldots, \chi_1(x_T))^T
\end{align*}
\]

1. Compute

\[ C_{00} = \frac{1}{T-\tau} X^T X \]

\[ C_{01} = \frac{1}{T-\tau} X^T Y \]

\[ C_{11} = \frac{1}{T-\tau} Y^T Y \]

with

\[ X = (x_0(x_1), x_0(x_2), \ldots, x_0(x_{T-\tau}))^T \]

\[ Y = (x_1(x_{1+\tau}), x_1(x_{2+\tau}), \ldots, x_1(x_T))^T \]

2. Perform the truncated SVD

\[ C_{00}^{-\frac{1}{2}} C_{01} C_{11}^{-\frac{1}{2}} \approx U_k \hat{\Sigma}_k V_k^T \]

3. Output \( \hat{\Sigma}_k \), \( \psi = U_k^T C_{00}^{-\frac{1}{2}} x_0 \) and \( \phi = V_k^T C_{11}^{-\frac{1}{2}} x_1 \)

Implementation: time-lagged canonical covariance analysis (TCCA)

1. Compute

\[
\begin{align*}
C_{00} &= \frac{1}{T-\tau} X^T X \\
C_{01} &= \frac{1}{T-\tau} X^T Y \\
C_{11} &= \frac{1}{T-\tau} Y^T Y
\end{align*}
\]

with

\[
\begin{align*}
X &= (\chi_0(x_1), \chi_0(x_2), \ldots, \chi_0(x_{T-\tau}))^T \\
Y &= (\chi_1(x_{1+\tau}), \chi_1(x_{2+\tau}), \ldots, \chi_1(x_T))^T
\end{align*}
\]

2. Perform the truncated SVD

\[
C_{00}^{-\frac{1}{2}} C_{01} C_{11}^{-\frac{1}{2}} \approx U_k \hat{\Sigma}_k V_k^T
\]

3. Output \( \hat{\Sigma}_k, \psi = U_k^T C_{00}^{-\frac{1}{2}} \chi_0 \) and \( \phi = V_k^T C_{11}^{-\frac{1}{2}} \chi_1 \)

For the choice \( \chi_0 = \chi_1 \), TCCA is consistent with EDMD:

\[
K_{\tau}^T = C_{01}^T C_{00}^{-1}
\]

**Theorem (Koopman approximation error):** For an operator $\hat{K}_\tau$ defined by

$$\hat{K}_\tau h = \sum_i \hat{\sigma}_i \langle g_i, h \rangle_{\rho_1} f_i$$

we have

$$\|\hat{K}_\tau - K_\tau\|^2_{HS} = \text{tr} \left[ \hat{\Sigma} C_{00} \hat{\Sigma} C_{11} - 2 \hat{\Sigma} C_{01} \right] + \sum_i \sigma_i^2$$

where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm, $\hat{\Sigma} = \text{diag} (\hat{\sigma}_1, \hat{\sigma}_2, \ldots)$, and

$$[C_{00}]_{ij} = \mathbb{E}_{\rho_0} \left[ f_i(x(t)) f_j(x(t)) \right]$$

$$[C_{01}]_{ij} = \mathbb{E}_{\rho_0} \left[ f_i(x(t)) g_j(x(t + \tau)) \right]^T$$

$$[C_{11}]_{ij} = \mathbb{E}_{\rho_1} \left[ g_i(x(t)) g_j(x(t)) \right]^T.$$
1D-Example

Markov process with Gaussian noise $w_t$

$$x_{t+1} = \frac{x_t}{2} + \frac{25x_t}{1 + x_t^2} + \sqrt{10}(1.1 + \cos(x_t))w_t$$

(a) Score

{Graph showing score and number of clusters:}

- **Exact**
- **MCV (train)**
- **MCV (test)**

(b.1) Eigenfunctions

10 clusters

(b.2) Eigenfunctions

10 clusters

(c.1) Eigenfunctions

62 clusters

(c.2) Eigenfunctions

62 clusters

(d.1) Eigenfunctions

300 clusters

(d.2) Eigenfunctions

300 clusters

Generalization: VAMP reduces nonequilibrium processes

\[ dx_t = -\pi A \sin(\pi x_t) \cos(\pi y_t) - \varepsilon(2y_t - 1) + \varepsilon dW_{t,1}, \]
\[ dy_t = \pi A \cos(\pi x_t) \sin(\pi y_t) - \varepsilon(2x_t - 3) + \varepsilon dW_{t,2} \]

**VAMP variational principle (subspace version)**

For any two sets of linearly independent functions $\chi_0(x) = (\chi_{01}(x), ..., \chi_{0n}(x))$ and $\chi_1(x) = (\chi_{11}(x), ..., \chi_{1n}(x))$, let us call

$$\hat{R}_2[\chi_0, \chi_1] = \left\|C_{00}^{-\frac{1}{2}} C_{01} C_{11}^{-\frac{1}{2}}\right\|_F^2$$

their VAMP-2 score, where $C_{00}$, $C_{01}$, $C_{11}$ are the feature correlation matrices as defined earlier and $\|\cdot\|_F$ indicates the Frobenius norm. The maximum value of the VAMP-2 score is achieved when the top $n$ left and right Koopman singular functions belong to $\text{span}(\chi_0)$ and $\text{span}(\chi_1)$, respectively.

VAMPnets

Resulting Markov model:

\[ K = C_{00}^{-1} C_{01} . \]

Relaxation timescales:

\[ t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|} , \]

Validate (Chapman-Kolmogorov test):

\[ K(n\tau) = K^n(\tau) , \]

VAMPnets

Alanine dipeptide

Max. transition probability: 41%
Min. transition probability: 0.5%

Validation

Results as a function of the number of states

VAMPnets

NTL9 Protein folding

Validation

Simulating biological timescales at atomic resolution

1000’s of **Microsecond** MD Trajectories

Markov State Model — **Millisecond kinetics**
Simulating biological timescales at atomic resolution

Adaptive Markov State Model — *seconds to hours kinetics*
Sampling biological timescales at atomic resolution

0.1 millisecond binding trajectory

Plattner, Doerr, De Fabritiis, Noé

Generative learning and molecular design

Structure and dynamics from MD Simulation data
Experimental data …

“Latent” variables
(states, kinetics, properties…)

Example:
Generative Adversarial Network
Generative Adversarial Network

Deep Generative Markov State Models

Deep Generative MSM

\[ x_t \]

Encoder \( \chi \)

\[ y_t \]

\[ K(\tau) \] Markov model

\[ y_{t+\tau} \]

Generator \( q \)

\[ \tilde{x}_{t+\tau} \]

noise

Rewiring Trick

\[ y_t \]

\[ q \]

\[ \tilde{y}_{t+\tau} \]

sample expectation

\[ \tilde{K}(\tau) \] Markov model

Deep Generative Markov State Models

Data

Deep MSM, resampled

“good” classical MSM

Wu, Mardt, Pasquali, Noé  
*NIPS* 2018 — arXiv:1805.07601
Deep Generative Markov State Models

Energy distance between distributions:

\[ D_E \left( \mathbb{P}(x), \mathbb{P}(y) \right) = \mathbb{E} \left[ 2 \| x - y \| - \| x - x' \| - \| y - y' \| \right] \]

with

\[ x, x' \sim \mathbb{P}(x) \]
\[ y, y' \sim \mathbb{P}(y) \]

\( D_E = 0 \) only if distributions are equal

\[ \implies \text{Train Generator Network by minimizing Energy Distance} \]

Wu, Mardt, Pasquali, Noé  \textbf{NIPS} 2018 — arXiv:1805.07601
Deep Generative Markov State Models

Learning transition densities

Deep Generative Markov State Models

Deep Generative MSM

Data

Deep MSM, resampled

“good” classical MSM

Deep generative MSM

Wu, Mardt, Pasquali, Noé *NIPS* 2018 — arXiv:1805.07601
Acknowledgements

Collaborations
Cecilia Clementi (Rice University)
Christof Schütte (FU Berlin)
Eric Vanden-Eijsden (Courant Institut NY)
Thomas Weikl (MPI Potsdam)
Edina Rosta (King’s College London)
Vijay Pande (Stanford)
Volker Haucke (FMP Berlin)
Stephan Sigrist (FU Berlin)
Katja, Faelber, Oliver Daumke (MDC)
John Chodera (MSKCC NY)
Gianni de Fabritiis (Barcelona)

Funding

[Logos of ERC, Einstein Foundation, DFG]