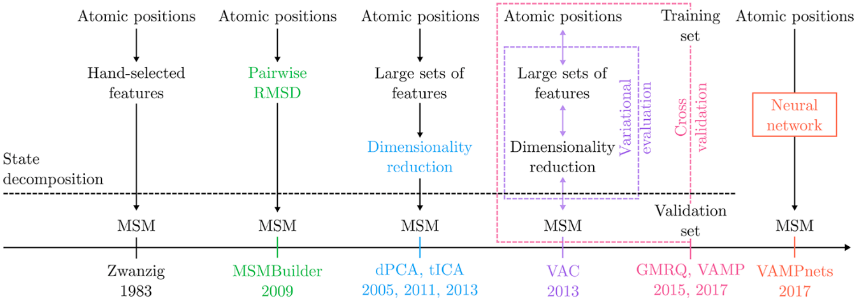
Deep Architectures for sampling ~~macro-~~molecules

Grégoire Sergeant-Perthuis
LCQB Sorbonne Université
3-6-24

- Sampling? In what sense?
  $\hookrightarrow$ sampling conformations !

- Configuration of atoms of a molecule encoded as $x \in \mathbb{R}^D$
- Total Interaction potential $V : \mathbb{R}^D \to \mathbb{R}$
  - $\hookrightarrow$ prescribes dynamics of the molecule
- Ideally, $t >> 1 \rightsquigarrow$ equilibrium Boltzmann distribution $e^{-\beta V}$
- Sampling? $\rightsquigarrow (x_i, i = 1, \ldots, n)$ such that $\frac{1}{n}\sum_{i=1}^n f(x_i) \approx \mathbb{E}_Q[f]$
- How sample?
  - $\hookrightarrow$ Markov chain $(X_t, t \geq 1)$ such that $\frac{1}{n}\sum_t f(X_t) \approx \mathbb{E}_Q[f]$
- In this course we are interested in sampling:
  - $\hookrightarrow$ significant conformations of molecules
  - $\hookrightarrow$ one molecule at a time.

- $(X_t \in \mathbb{R}^d, t \in \mathbb{R}_+)$ is a Markov chain:
  $\mathbb{P}(X_{t_n+\delta t} \mid X_{t_n}, \dots, X_{t_1}) = \mathbb{P}(X_{\delta t} \mid X_0)$
- Grouping together times and states:
  $Y_n = \phi(X_{[n\tau,(n+1)\tau]}) \in \{0, \dots, K\}$
  !!!! $Y_1, \dots, Y_n$ are not, in general, a Markov chain
- Markov State Model: appropriate ϕ (state decomposition) and lags τ
  - $\hookrightarrow$ with $Y_1, \dots, Y_n$ that define a Markov chain
  - $\hookrightarrow$ with $\{0, \dots, K\}$ characteristic configurations of molecules
  - $\hookrightarrow$ $\mathbb{P}(Y_{n+1} = j \mid Y_n = i)$ characteristic transition
  - $\hookrightarrow$ for modeling longer time dynamics of molecules



Timeline of MSM research: reproduction of Figure 2 [HP18]

Variational Approach to Conformational Dynamics (VAC)

- Discrete Markov chain $\pi(y \mid x), y, x \in E$, with stationary law μ, is reversible:
  - $\hookrightarrow$ $\pi(y \mid x)\mu(x) = \pi(x \mid y)\mu(y)$
- $P : \mathbb{R}^{|E|} \to \mathbb{R}^{|E|}$ defined as $P(f)(x) = \sum_{y \in E} f(y)\pi(y \mid x)$
- Scalar product $\langle f, g \rangle = \sum_{x \in E} \mu(x) f(x) \overline{g(x)}$
- Reversibility is equivalent to $P^\dagger = P$
  - $\hookrightarrow$ Eigenvalues of P are positive.

Variational Approach to Conformational Dynamics (VAC)

- $\pi(X_{t+\tau} = y \mid X_t = x) = e^{tM}_{y,x}$
- $p_{t+\tau}(y) = \sum_{x \in E} \pi(X_{t+\tau} = y \mid X_t = x) p_t(x)$
- When π is reversible (Equation 2 [WNP+16]):
  - $\hookrightarrow$ $e^{-\tau/t_i}$ is an eigenvalue of P.
  - $\hookrightarrow$ $\psi_i, i \in I$ are the eigenfunctions.
  - $\hookrightarrow$ $p_{t+\tau}(x) = \sum_i e^{-\tau/t_i} \mu(x) \psi_i(x) \langle \psi_i, p_t \rangle$
- The m dominant eigenfunctions $\psi_1, \ldots, \psi_m \rightsquigarrow m$ slow collective variables
  - $\hookrightarrow$ Characterizes the behavior of a molecule on time scales $\tau \gg t_{m+1}$

Variational Approach to Conformational Dynamics (VAC)

- Dimensionality reduction:
  - $\hookrightarrow$ Projection on eigenfunctions ordered by time scale
  - $\hookrightarrow$ Approximation of eigenfunctions
- How? Optimize $\rightsquigarrow$ $\max\limits_{\substack{f_1,\ldots,f_m \\ \|f_i\|=1 \\ \langle f_i, f_j\rangle = 0,\, i\neq j}} \sum_{i=1}^m \langle f_i, Af_i\rangle$

$$
\begin{aligned}
R_m = \max_{f_1,\ldots,f_m} \; & \sum_{i=1}^{m} \mathbb{E}_\mu \left[f_i(\mathbf{x}_t) f_i(\mathbf{x}_{t+\tau}) \right], \\
\text{s.t.} \quad & \mathbb{E}_\mu \left[f_i(\mathbf{x}_t)^2 \right] = 1, \\
& \mathbb{E}_\mu \left[f_i(\mathbf{x}_t) f_j(\mathbf{x}_t) \right] = 0, \; for\; i \neq j,
\end{aligned}
\tag{3}
$$

Variational Approach to Conformational Dynamics (VAC) [WNP$^+$16]

Linear VAC

- A priori choice of functions $\chi : \Omega \to \mathbb{R}^M$, coordinates $\chi_i, i = 1, \dots, M$
- Find optimal linear combination:
  $\rightarrow \; f(x) = \sum_{j=1}^{M} b_j \chi_j(x)$
  Pose,
$$C(0)_{i,j} = \mathbb{E}_\mu[\chi_i(X_t)\chi_j(X_t)]$$
$$C(\tau)_{i,j} = \mathbb{E}_\mu[\chi_i(X_t)\chi_j(X_{t+\tau})]$$
  Then, optimal $(b^1, b^2, \dots)$ are given by the eigenvectors of the matrix $K = C(0)^{-1}C(\tau)$,
$$C(\tau)B = C(0)B\Lambda$$
  with $B_{i,j} = b^i_j$ and $\Lambda_{i,j} = \lambda_i \delta_{i,j}$ being the associated eigenvalues.

Linear VAC, two examples: TICA and MSM

- Time-lagged independent component analysis (TICA)
  $\hookrightarrow$ $\chi(x) = x - \mu$

- Markov Sate Model (MSM):
  $\hookrightarrow$ $\Omega = \bigsqcup_i A_i$
  $\hookrightarrow$ $\chi_i(x) = 1[x \in A_i]$

VAC Limitations!

$\rightarrow$ Not applicable to non-reversible processes

  $\hookrightarrow$ Eigenvalues can be complex

$\rightarrow$ In linear-VAC, χ is given a priori and not learned

$\rightarrow$ Most $\chi : \mathbb{R}^{D} \rightarrow \mathbb{R}^{M}$ do not allow for generation of configurations.

VAC Limitations!

$\rightarrow$ Not applicable to non-reversible processes

- $\hookrightarrow$ Eigenvalues can be complex
- $\hookrightarrow$ Variational principle for Markov processes (VAMP)

$\rightarrow$ In linear-VAC, χ is given a priori and not learned

$\rightarrow$ Most $\chi : \mathbb{R}^D \rightarrow \mathbb{R}^M$ do not allow for generation of configurations.

Variational principle for Markov processes (VAMP)

- Idea: approximate the time evolution on observables:
  - $\hookrightarrow$ Approximate: $K_\tau(f) = \mathbb{E}_\mu[f(X_{t+\tau})|X_t] \rightsquigarrow$ Koopman operator
  - $\rightarrow$ Finding an optimal orthogonal basis $f_1, ...f_k \in L^2(\Omega, \mu)$ and $g_1, ...g_k \in L^2(\Omega, \mu)$ such $K_\tau \approx F \Lambda G^\dagger$ (Singular Value Decomposition)

Theorem 3. VAMP variational principle. *The k dominant singular components of a Koopman operator are the solution of the following maximization problem:*

$$\begin{aligned} \sum_{i=1}^{k} \sigma_i^r &= \max_{\mathbf{f},\mathbf{g}} \mathcal{R}_r\left[\mathbf{f},\mathbf{g}\right], \\ s.t. \left\langle f_i, f_j \right\rangle_{\rho_0} &= 1_{i=j}, \\ \left\langle g_i, g_j \right\rangle_{\rho_1} &= 1_{i=j}, \end{aligned} \tag{11}$$

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

$$\mathcal{R}_r\left[\mathbf{f},\mathbf{g}\right] = \sum_{i=1}^{k} \left\langle f_i, \mathcal{K}_\tau g_i \right\rangle_{\rho_0}^r \tag{12}$$

is called the VAMP-r score of $\mathbf{f}$ *and* $\mathbf{g}$.

Variational principle for Markov processes (VAMP) [WN17]

- Pose $f = \chi_0$ and $g = \chi_1$

$$\mathbf{C}_{00} = \mathbb{E}_t \left[\boldsymbol{\chi}_0(\mathbf{x}_t) \boldsymbol{\chi}_0(\mathbf{x}_t)^\top \right] \tag{2}$$

$$\mathbf{C}_{01} = \mathbb{E}_t \left[\boldsymbol{\chi}_0(\mathbf{x}_t) \boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})^\top \right] \tag{3}$$

$$\mathbf{C}_{11} = \mathbb{E}_{t+\tau} \left[\boldsymbol{\chi}_1(\mathbf{x}_{t+\tau}) \boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})^\top \right] \tag{4}$$

***VAMP variational principle**: For any two sets of linearly independent functions* $\boldsymbol{\chi}_0(\mathbf{x})$ and $\boldsymbol{\chi}_1(\mathbf{x})$, *let us call*

$$\hat{R}_2[\boldsymbol{\chi}_0, \boldsymbol{\chi}_1] = \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2$$

their VAMP-2 score, where $\mathbf{C}_{00}, \mathbf{C}_{01}, \mathbf{C}_{11}$ *are defined by Eqs. (2-4) and* $\|\mathbf{A}\|_F^2 = n^{-1} \sum_{i,j} A_{ij}^2$ *is the Frobenius norm of* $n \times n$ *matrix* $\mathbf{A}$. *The maximum value of VAMP-2 score is achieved when the top* m *left and right Koopman singular functions belong to* $\mathrm{span}(\chi_{01}, ..., \chi_{0m})$ *and* $\mathrm{span}(\chi_{11}, ..., \chi_{1m})$ *respectively.*

VAMP reformulation [MPWN17]

- Samples (e.g. short time MD)

$$\mathbf{X} = \begin{pmatrix} \chi_1(\mathbf{x}_1) & \cdots & \chi_m(\mathbf{x}_1) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_{T-\tau}) & \cdots & \chi_m(\mathbf{x}_{T-\tau}) \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} \chi_1(\mathbf{x}_{\tau+1}) & \cdots & \chi_m(\mathbf{x}_{\tau+1}) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_T) & \cdots & \chi_m(\mathbf{x}_T) \end{pmatrix},$$

- $\hat{C}(0) = \frac{1}{N} X^T X$, $\hat{C}(\tau) = \frac{1}{N} X^T Y$
- **VAMPnet** $\rightsquigarrow \chi_0 = \chi_1$ learned with a neural network.

VAC Limitations!

$\rightarrow$ Not applicable to non-reversible processes

$\hookrightarrow$ Eigenvalues can be complex

$\hookrightarrow$ VAMP

$\rightarrow$ In linear-VAC, χ is given a priori and not learned

$\hookrightarrow$ use VAMPnet

$\rightarrow$ Most $\chi : \mathbb{R}^D \rightarrow \mathbb{R}^M$ do not allow for generation of configurations.

$\hookrightarrow$ Generative model

Generative model

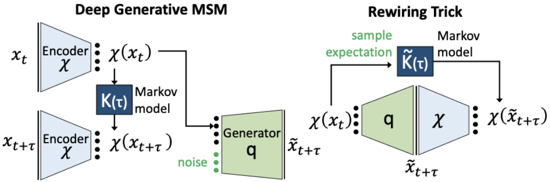


Figure 1: Schematic of Deep Generative Markov State Models (DeepGenMSMs) and the rewiring trick. The function χ, here represented by neural networks, maps the time-lagged input configurations to metastable states whose dynamics are governed by a transition probability matrix $\mathbf{K}$. The generator samples the distribution $x_{t+\tau} \sim \mathbf{q}$ by employing a generative network that can produce novel configurations (or by resampling $x_{t+\tau}$ in DeepResampleMSMs). The rewiring trick consists of reconnecting the probabilistic networks $\mathbf{q}$ and $\boldsymbol{\chi}$ such that the time propagation in latent space can be sampled: From the latent state $\chi(x_t)$, we generate a time-lagged configuration $x_{t+\tau}$ using $\mathbf{q}$, and then transform it back to the latent space, $\chi(x_{t+\tau})$. Each application of the rewired network samples the latent space transitions, thus providing the statistics to estimate the Markov model transition matrix $\mathbf{K}(\tau)$, which is needed for analysis. This trick allows $\mathbf{K}(\tau)$ to be estimated with desired constraints, such as detailed balance.

Generative methodology [WMPN18]

$$\mathbb{P}(x_{t+\tau} = y | x_t = x) = \boldsymbol{\chi}(x)^\top \mathbf{q}(y;\tau) = \sum_{i=1}^{m} \chi_i(x) q_i(y;\tau). \tag{1}$$

Here, $\boldsymbol{\chi}(x)^\top = [\chi_1(x), ..., \chi_m(x)]$ represent the probability of configuration x to be in a metastable (long-lived) state i

$$\chi_i(x) = \mathbb{P}(x_t \in \text{state } i \mid x_t = x).$$

with $q_i(y;\tau) = P(x_{t+\tau} = y | x_t \in \text{ state } i)$. γ_i up to normalizing is q_i.

$\rightarrow$ Trained with maximum likelihood

$\rightarrow$ Better results than VAMP

Now we can optimize χ_i and γ_i by maximizing the likelihood (ML) of generating the pairs $(x_t, x_{t+\tau})$ observed in the data. The log-likelihood is given by:

$$LL = \sum_{t=1}^{T-\tau} \ln \left(\sum_{i=1}^{m} \chi_i(x_t) \bar{\gamma}_i^{-1} \gamma_i(x_{t+\tau}) \right), \tag{8}$$

Thank you very much for your attention

Thank you very much for your attention!

References I

- Brooke E. Husic and Vijay S. Pande, Markov state models: From an art to a science, Journal of the American Chemical Society **140** (2018), no. 7, 2386–2396, PMID: 29323881.
- Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé, Vampnets for deep learning of molecular kinetics, Nature Communications **9** (2017).
- Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noé, Deep generative markov state models, ArXiv **abs/1805.07601** (2018).
- Hao Wu and Frank No'e, Variational approach for learning markov processes from time series data, Journal of Nonlinear Science **30** (2017), 23 – 66.

Hao Wu, Feliks Nüske, Fabian Paul, Stefan Klus, Péter Koltai, and Frank Noé, Variational koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations., The Journal of chemical physics **146 15** (2016), 154104.