Importance sampling with free energies and autoencoders for multimodal probability distributions

Gabriel STOLTZ

(CERMICS, Ecole des Ponts & MATHERIALS team, Inria Paris)

Joint work with Z. Belkacemi (Sanofi & ENPC), P. Gkeka, M. Bianciotto, H. Minoux (Sanofi), T. Pigeon (Inria & IFPEN), Wei Zhang (FU Berlin) and T. Lelièvre (ENPC & Inria)

Mostly Monte Carlo seminar, May 2024

Outline

- **Free energy importance sampling**
  - Multimodal probability measures in Bayesian statistics
  - Importance sampling
  - Reaction coordinates and free energy

- **Finding good reaction coordinates using autoencoders**
  - Preliminaries: definitions, training, etc.
  - An interpretation in terms of conditional expectations
  - Iterative learning of free energy and RC

Chopin, T. Lelièvre and G. Stoltz, Free energy methods for efficient exploration of mixture posterior densities, *Stat. Comput.* **22**(4) (2012) 897-916

Z. Belkacemi, P. Gkeka, T. Lelièvre and G. Stoltz, Chasing collective variables using autoencoders and biased trajectories, *J. Chem. Theory Comput.* **18**(1), 59-78 (2022)

Lelièvre, T. Pigeon, G. Stoltz and W. Zhang, Analyzing multimodal probability measures with autoencoders, *J. Phys. Chem. B* **128**(11) 2607-2631 (2024)

Free energy importance sampling

Multimodality in Bayesian inference (1)

- Data points $\{y_i\}_{i=1,\dots,N_{\text{data}}}$
- Elementary likelihood $P(y|\theta)$, with θ parameters of probability measure
- A priori distribution of the parameters p_{prior} (usually not so informative)

Aim

Find the values of the parameters θ describing correctly the data: sample

$$\nu(\theta) \propto p_{\text{prior}}(\theta) \prod_{i=1}^{N_{\text{data}}} P(y_i|\theta)$$

- Example of Gaussian mixture model

Multimodality in Bayesian inference (2)

Elementary likelihood: mixture of K Gaussians

$$P(y \,|\, \theta) = \sum_{k=1}^{K} a_k \sqrt{\frac{\lambda_k}{2\pi}} \exp\left(-\frac{\lambda_k}{2}(y-\mu_k)^2\right)$$

Parameters $\theta = (a_1, \dots, a_{K-1}, \mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K)$ with

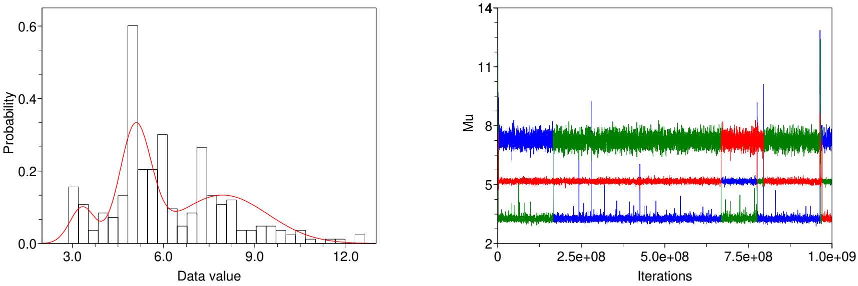
$$\mu_k \in \mathbb{R}, \qquad \lambda_k \geqslant 0, \qquad 0 \leqslant a_k \leqslant 1, \qquad a_1 + \dots + a_K = 1$$

Prior distribution: Random beta model → additional variable

- uniform distribution of the weights a_k
- $\mu_k \sim \mathcal{N}\left(M, R^2/4\right)$ with M = mean of data, $R = \max y_i - \min y_i$
- $\lambda_k \sim \Gamma(\alpha, \beta)$ with $\beta \sim \Gamma(g, h)$, $g = 0.2$ and $h = 100g/\alpha R^2$

S. Richardson and P. J. Green, *J. Roy. Stat. Soc. B*, 1997.
A. Jasra, C. Holmes and D. Stephens, *Statist. Science*, 2005

Multimodality in Bayesian inference (3)

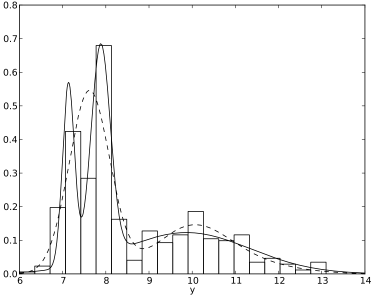


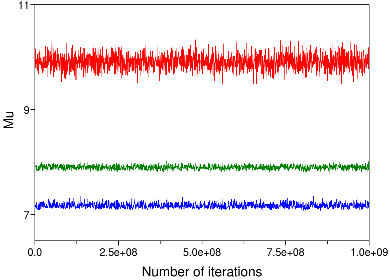
Left: Lengths of snappers ($N_{\rm data} = 256$), and a possible fit for $K = 3$ using the last configuration from the trajectory plotted in the right picture.

Right: Typical sampling trajectory, Metropolis/Gaussian random walk with $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta) = (0.0005, 0.025, 0.05, 0.005)$.

A. J. Izenman and C. J. Sommer, *J. Am. Stat. Assoc.*, 1988.
K. Basford *et al.*, *J. Appl. Stat.*, 1997

Multimodality in Bayesian inference (4)





Left: Thickness of Mexican stamps ("Hidalgo stamp data", $N_{\rm data} = 485$), and two possible fits for $K = 3$ ("genuine multimodality", solid line: dominant mode).

Right: Typical sampling trajectory

D. Titterington *et al.*, *Statistical Analysis of Finite Mixture Distributions*, 1986.
S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, 2006.

Importance sampling

Importance sampling function $\widetilde{V} : \mathbb{R}^D \to \mathbb{R}$

- Target measure $\nu_0(d\theta) = Z_0^{-1}\mathrm{e}^{-V(\theta)}\,d\theta$
- Sample a modified target measure $\nu_{\widetilde{V}}(d\theta) = Z_{\widetilde{V}}^{-1}\mathrm{e}^{-(V+\widetilde{V})(\theta)}\,d\theta$
- Reweight sample points $\theta^n \sim \pi_{\widetilde{V}}$ by $\mathrm{e}^{\widetilde{V}}$

$$\widehat{\varphi}_{N_{\mathrm{iter}},\widetilde{V}} = \frac{\displaystyle\sum_{n=1}^{N_{\mathrm{iter}}} \varphi(\theta^n)\mathrm{e}^{\widetilde{V}(\theta^n)}}{\displaystyle\sum_{n=1}^{N_{\mathrm{iter}}} \mathrm{e}^{\widetilde{V}(\theta^n)}} \xrightarrow[N_{\mathrm{iter}}\to+\infty]{\mathrm{a.s.}} \frac{\displaystyle\int \varphi\,\mathrm{e}^{\widetilde{V}}\,d\nu_{\widetilde{V}}}{\displaystyle\int \mathrm{e}^{\widetilde{V}}\,d\nu_{\widetilde{V}}} = \int \varphi\,d\nu_0$$

A good choice of the importance sampling function can improve the performance of the estimator... but a bad choice can degrade it!

Importance sampling in high dimensions

General strategy:

- low-dimensional (nonlinear) function $\xi(\theta) \in \mathbb{R}^d$ with $d \ll D$, encoding the metastability of the sampling method (**reaction coordinate**)
- bias by the associated free energy: $\widetilde{V}(\theta) = F(\xi(\theta))$ with

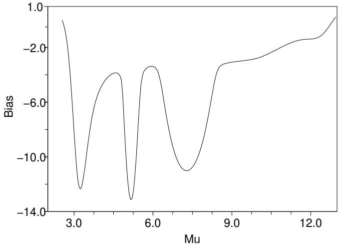
$$\mathrm{e}^{-F(z)} = \int \mathrm{e}^{-V(\theta)}\, \delta_{\xi(\theta)-z}(d\theta)$$

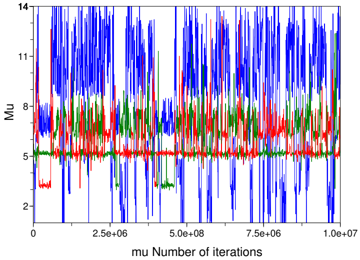
- Simple case: $\xi(\theta) = \theta_1$, for which

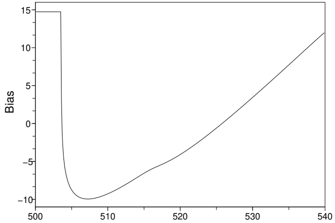
$$F(z) = -\ln\left(\int \mathrm{e}^{-V(z,\theta_2,\dots,\theta_d)}\, d\theta_2 \dots d\theta_d\right)$$

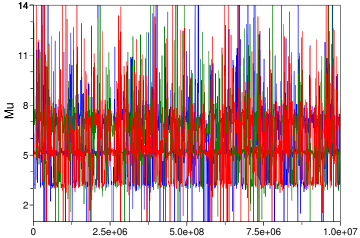
Various methods to compute the free energy: thermodynamic integration, umbrella sampling, adaptive methods, ...

Lelièvre/Rousset/Stoltz, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, 2010)









Choices $\xi(x) = \mu_1$ and $\xi(x) = V(x)$

N. Chopin, T. Lelièvre and G. Stoltz, *Statist. Comput.*, 2012

Machine learning approaches for finding reaction coordinates

(A biased perspective on some) References

- **ML reviews in MD** (biased towards dimensionality reduction, not force fields)
  - A. Gliemlo, B. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, *Chem. Rev.* **121**(16), 9722-9758 (2021)
  - P. Gkeka *et al.*, *J. Chem. Theory Comput.* **16**(8), 4757-4775 (2020)
  - F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, *Annu. Rev. Phys. Chem.* **71**, 361-390 (2020)
  - A.L. Ferguson, *J. Phys.: Condens. Matter* **30**, 04300 (2018)
  - M. Chen, *Eur. Phys. J. B* **94**, 211 (2021)

- **More general ML references**
  - P. Mehta, M. Bukov, C.-H. Wang, A.G.R.Day, C. Richardson, C.K. Fisher, D.J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, *Physics Reports* **810**, 1-124 (2019)
  - I. Goodfellow, Y. Bengio, A. Courville *Deep Learning* (MIT Press, 2016) `http://www.deeplearningbook.org`
  - K.P. Murphy, *Probabilistic Machine Learning: An Introduction* (MIT Press, 2022)

Some representative approaches for finding RC (1)

- Domain knowledge/intuition (log-likelihood, approximate summary statistics, etc)

- **Short list of data-oriented approaches** (depending on the data at hand...)
  - [supervised learning] separate metastable states
  - [unsupervised/static] distinguish linear models (PCA) and nonlinear ones (e.g. based on autoencoders such as MESA[1])
  - [unsupervised/dynamics] operator based approaches (VAC, EDMD, diffusion maps, MSM; incl. tICA and VAMPNets)

(Huge literature! I am not quoting precise references here because the list would be too long)

- Other classifications[2,3] possible, e.g. **slow vs. high variance RC**

[1] W. Chen and A.L. Ferguson, *J. Comput. Chem.* 2018; W. Chen, A.R. Tan, and A.L. Ferguson, *J. Chem. Phys.* 2018

[2] P. Gkeka et al., *J. Chem. Theory Comput.* (2020)

[3] A. Gliemlo *et al.*, *Annu. Rev. Phys. Chem.* (2021)

Some representative approaches for finding CV (2)

Methods for Choosing Collective variables

High-variance CVs

- Principal Components Analysis (PCA)
- Locally Linear Embedding (LLE)
- Independent Component Analysis (ICA)
- Laplacian and Hessian eigenmaps
- Local tangent space alignment
- Kernel PCA
- Nonlinear PCA
- Isomap
- Diffusion maps
- Multidimensional scaling
- Semidefinite embedding/ Maximum variance unfolding

Available tools for CV identification

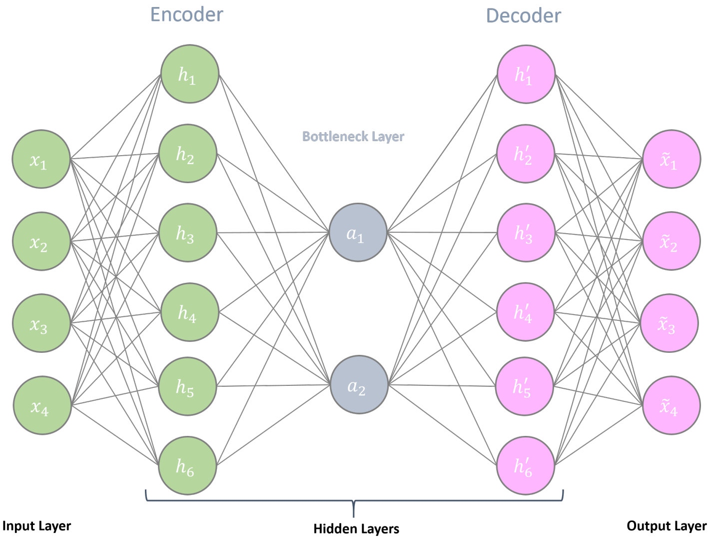
- Diffusion-Map-directed MD (DM-d-MD)
- Intrinsic Map Dynamics (iMapD)
- Smooth and nonlinear datadriven CVs (SandCV)
- Molecular Enhanced Sampling with Autoencoders (MESA)
- Reweighted Autoencoded Variations Bayes for Enhanced Sampling (RAVE)
- REinforcement Learning based on Adaptive samPling (REAP)

Slow CVs

- Variational Approach to Conformational dynamics (VAC)
- (extended) Dynamical Mode Decomposition ((E)DMD)
- Kernel TICA
- Markov State Models (MSM)
- Time-lagged autoencoders (TAEs)
- Time-lagged Independent Component Analysis (TICA)
- Deep Canonical Correlation Analysis (DCCA)
- Variational Dynamics Encoders (VDEs)
- Variational Approach for Markov Processes nets (VAMPnets)
- State-free Reversible VAMPnets (SRV)

Constructing CVs with autoencoders

Bottleneck autoencoders (1)



Bottleneck autoencoders (2)

- Input space $\Theta \subseteq \mathbb{R}^D$, bottleneck space $\mathcal{B} \subseteq \mathbb{R}^d$ with $d < D$

$$f(\theta) = f_{\mathsf{dec}}\Big(f_{\mathsf{enc}}(\theta)\Big)$$

where $f_{\mathsf{enc}} : \Theta \to \mathcal{B}$ and $f_{\mathsf{dec}} : \mathcal{B} \to \Theta$

Collective variable = encoder part

$$\xi = f_{\mathrm{enc}}$$

- Fully connected neural network, symmetrical structure, $2L$ layers

- Parameters $\mathbf{p} = \{p_k\}_{k=1,\dots,K}$ (bias vectors b_ℓ and weights matrices W_ℓ)

$$f_{\mathbf{p}}(\theta) = g_{2L}\left[b_{2L} + W_{2L} \dots g_1(b_1 + W_1\theta)\right],$$

with activation functions g_ℓ

(examples: $\tanh(a)$, ReLU $\max(0,a)$, sigmoid $\sigma(a) = 1/(1+\mathrm{e}^{-a})$, etc)

Training autoencoders

- Theoretically: minimization problem in $\mathcal{P} \subset \mathbb{R}^K$

$$\mathbf{p}_\nu \in \underset{\mathbf{p}\in\mathcal{P}}{\operatorname{argmin}}\, \mathcal{L}(\nu, \mathbf{p}),$$

with cost function

$$\mathcal{L}(\nu, \mathbf{p}) = \mathbb{E}_\nu(\|\theta - f_{\mathbf{p}}(\theta)\|^2) = \int_\Theta \|\theta - f_{\mathbf{p}}(\theta)\|^2 \, \nu(d\theta)$$

- In practice, access only to a sample: **minimization of empirical cost**

$$\mathcal{L}(\hat{\nu}, \mathbf{p}) = \frac{1}{N}\sum_{i=1}^N \|\theta^i - f_{\mathbf{p}}(\theta^i)\|^2, \qquad \hat{\mu} = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i}$$

- Actual training: Adam + early stopping; possibly add some regularization

Some properties of autoencoders

Three viewpoints on the loss function (1/2)

Idealized setting: $f_{\mathrm{enc}} : \Theta \to \mathcal{Z}$ and $f_{\mathrm{dec}} : \mathcal{Z} \to \Theta$ measurable

$$\mathcal{F} = \{ f = f_{\mathrm{dec}} \circ f_{\mathrm{enc}},\ f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}},\ f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}} \}$$

Usual loss: $\inf_{f \in \mathcal{F}} \mathbb{E}\left[\|\theta - f(\theta)\|^2 \right]$

Principal manifold formulation: fix decoder, minimize over encoders

$$\begin{aligned} \inf_{f \in \mathcal{F}} \mathbb{E}\left[\|\theta - f(\theta)\|^2 \right] &= \inf_{f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}}} \left\{ \inf_{f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}}} \mathbb{E}\left[\|\theta - f_{\mathrm{dec}} \circ f_{\mathrm{enc}}(\theta)\|^2 \right] \right\} \\ &= \inf_{f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}}} \mathbb{E}\left[\left\|\theta - f_{\mathrm{dec}} \circ h^{\star}_{f_{\mathrm{dec}}}(\theta)\right\|^2 \right] \end{aligned}$$

with ideal encoder $h^{\star}_{f_{\mathrm{dec}}}(\theta) \in \underset{z \in \mathcal{Z}}{\mathrm{argmin}} \, \|\theta - f_{\mathrm{dec}}(z)\|$

Hastie/Stützle (1986), Tibshirani (1992)
Venturoli/Vanden–Eijnden (2009)
Gerber/Whitaker, *J. Mach. Learn. Res.* (2013); Gerber, *arXiv preprint* **2104.05000**

Three viewpoints on the loss function (2/2)

Formulation with conditional expectation: fix encoder, minimize over decoders

$$\inf_{f \in \mathcal{F}} \mathbb{E}\left[\|\theta - f(\theta)\|^2\right] = \inf_{f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}}} \left\{ \inf_{f_{\mathrm{dec}} \in \mathcal{F}_{\mathrm{dec}}} \mathbb{E}\left[\|\theta - f_{\mathrm{dec}} \circ f_{\mathrm{enc}}(\theta)\|^2\right] \right\}$$
$$= \inf_{f_{\mathrm{enc}} \in \mathcal{F}_{\mathrm{enc}}} \mathbb{E}\left[\left\|\theta - g^\star_{f_{\mathrm{enc}}} \circ f_{\mathrm{enc}}(\theta)\right\|^2\right]$$

with ideal decoder $g^\star_{f_{\mathrm{enc}}}(z) = \mathbb{E}[\,\theta \mid f_{\mathrm{enc}}(\theta) = z]$

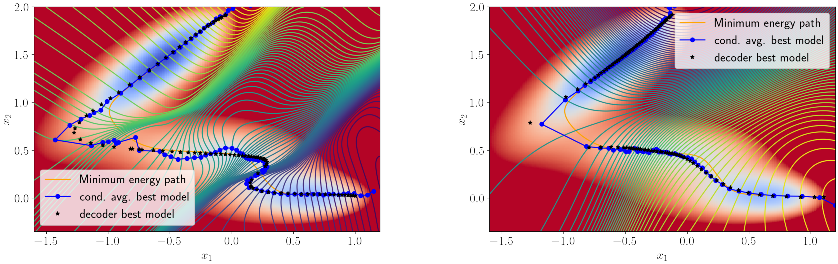
Alternative interpretations of the reconstruction error

$$\mathbb{E}\left[\left\|\theta - g^\star_{f_{\mathrm{enc}}} \circ f_{\mathrm{enc}}(\theta)\right\|^2\right] = \mathrm{Var}(\theta) - \mathrm{Var}\left[\mathbb{E}(\theta | f_{\mathrm{enc}}(\theta))\right]$$
$$= \mathbb{E}\left[\mathrm{Var}(\theta | f_{\mathrm{enc}}(\theta))\right]$$

First equality by developing square, second by conditioning on $f_{\mathrm{enc}}(\theta)$

Practical implication: minimizing reconstruction loss amounts to...

- minimizing intraclass dispersion
  (small spread of data points for $f_{\rm enc}$ given around the mean)
- maximizing interclass dispersion
  (the mean values associated with $f_{\rm enc}$ given should be as spread out as possible)



Left: topology (2, 5, 5, 1, 5, 5, 2). Right: topology (2, 5, 5, 1, 20, 20, 2)

Additional properties

- Necessary condition for critical points of the loss functional on f_{enc}

$$\left[\theta - g^{\star}_{f_{\mathrm{enc}}}(f_{\mathrm{enc}}(\theta))\right]^{\top} \partial_{z_j} g^{\star}_{f_{\mathrm{enc}}}(f_{\mathrm{enc}}(\theta)) = 0, \qquad 1 \leqslant j \leqslant d,\ \theta \in \mathrm{Supp}(\mu)$$

Low temperature limit: $\{g^{\star}_{f_{\mathrm{enc}}}(z)\}_{z \in [z_A, z_B]}$ minimum energy path[4]

- **Formulation with conditional expectations for other models:**
  - clustering
  - PCA (= autoencoders with identity activation functions)

- **Various extensions:**
  - change reference measure to better take transition states into account
  - multiple transition paths: single encoder and several decoders
  - possibly add some regularization terms

[4] Venturoli/Vanden–Eijnden (2009)

Free energy biasing and iterative learning

Training on modified target measures

• Interesting systems are metastable (no spontaneous exploration of phase space)
Sample according to a biased distribution $\widetilde{\nu}$

• Need for reweighting to learn the correct encoding!

$$w(\theta) = \frac{\nu(\theta)}{\widetilde{\nu}(\theta)}$$

• **Minimization problem:** theoretical cost function

$$\mathcal{L}(\nu, \mathbf{p}) = \int_\Theta \|\theta - f_{\mathbf{p}}(\theta)\|^2 \, w(\theta) \, \widetilde{\nu}(d\theta)$$

actual cost function

$$\mathcal{L}(\widehat{\nu}_{\mathsf{wght}}, \mathbf{p}) = \sum_{i=1}^N \widehat{w}_i \|\theta^i - f_{\mathbf{p}}(\theta^i)\|^2, \qquad \widehat{w}_i = \frac{\nu(\theta^i)/\widetilde{\nu}(\theta^i)}{\sum_{j=1}^N \nu(\theta^j)/\widetilde{\nu}(\theta^j)}$$

• Only requires the knowledge of ν and $\widetilde{\nu}$ up to a multiplicative constant

• Stochastic gradients in training: sampling with replacement according to multinomial distribution

Proof of concept (1)

• **Gaussian distributions** $\mu_i = \mathcal{N}(0, \Sigma_i)$ with

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.01 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.01 & 0 \\ 0 & 1 \end{pmatrix}$$

Datasets $\mathcal{D}_i$ of $N = 10^6$ i.i.d. points

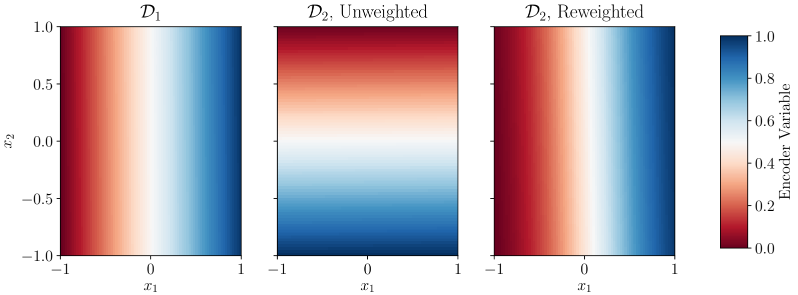
• Autoencoders with 2 layers of resp. 1 and 2 nodes, linear activation functions ($\simeq$ PCA)

• **Training on:**

- $\mathcal{D}_1$
- $\mathcal{D}_2$
- $\mathcal{D}_2$ with reweighting $\widehat{w}_i \propto \mu_1/\mu_2$

Proof of concept (2)

Heat maps of f_{enc}

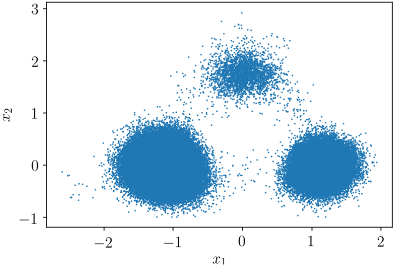


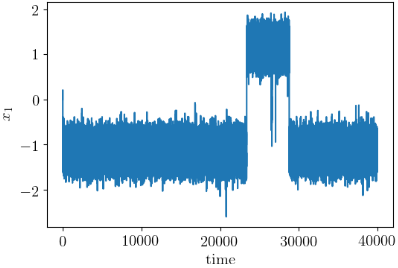
Third encoder very similar to the first one: projection on θ_1
Second encoder projects on a direction close to θ_2

Proof of concept with free energy biasing (1)

Two dimensional potential ("entropic switch")[5]

$$V(\theta_1,\theta_2) = 3\mathrm{e}^{-\theta_1^2}\left(\mathrm{e}^{-(\theta_2-1/3)^2} - \mathrm{e}^{-(\theta_2-5/3)^2}\right) - 5\mathrm{e}^{-\theta_2^2}\left(\mathrm{e}^{-(\theta_1-1)^2} + \mathrm{e}^{-(\theta_1+1)^2}\right) + 0.2\theta_1^4 + 0.2(\theta_2-1/3)^4$$





Trajectory from $\theta^{j+1} = \theta^j - \nabla V(\theta^j)\Delta t + \sqrt{2\beta^{-1}\Delta t}G^j$ for $\beta = 4$ and $\Delta t = 10^{-3}$ $\longrightarrow$ metastability in the θ_1 direction

[5] S. Park, M.K. Sener, D. Lu, and K. Schulten (2003)

Proof of concept with free energy biasing (2)

• **Free energy biasing:** distributions $Z_i^{-1} \exp\left(-V(\theta) + F_i(\xi_i(\theta))\right)$

$$F_1(\theta_1) = -\ln\left(\int_{\mathbb{R}} \mathrm{e}^{-V(\theta_1,\theta_2)}\, d\theta_2\right), \qquad F_2(\theta_2) = -\ln\left(\int_{\mathbb{R}} \dots d\theta_1\right)$$

Three datasets: unbiased trajectory, trajectories biased using F_1 and F_2

(free energy biased trajectories are shorter but same number of data points $N = 10^6$)

• Autoencoders: 2-1-2 topology, activation functions $\tanh$ (so that CV is in $[-1,1]$) then identity

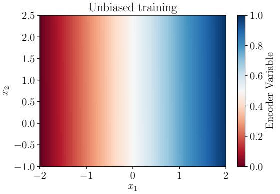
• **Five training scenarios:**

- training on long unbiased trajectory (reference CV)
- ξ_1-biased trajectory, with or without reweighting
- ξ_2-biased trajectory, with or without reweighting

Proof of concept with free energy biasing (3)

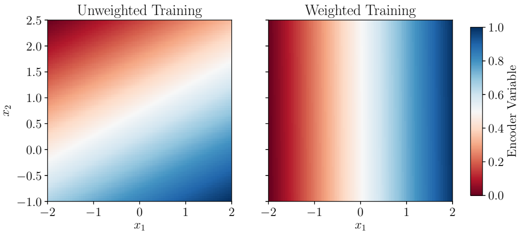
Normalize to compare

$$\xi_{\mathsf{AE}}^{\mathsf{norm}}(\theta) = \frac{\xi_{\mathsf{AE}}(\theta) - \xi_{\mathsf{AE}}^{\min}}{\xi_{\mathsf{AE}}^{\max} - \xi_{\mathsf{AE}}^{\min}}$$

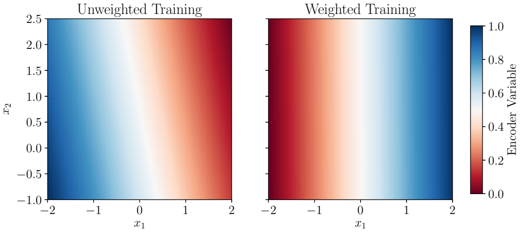


Reference CV

(distinguishes the 3 wells)



θ_1-biased trajectory



θ_2-biased trajectory

Iterative training/exploration

Interesting systems are metastable (no spontaneous exploration of phase space)
Iterate between exploration and update of RC based on new data points

Basic strategy: Metropolis targeting $\widetilde{\nu}$, free energy on a grid

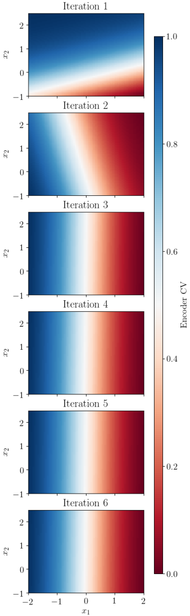
$$\forall z \in [z_i, z_{i+1}], \qquad \mathrm{e}^{-F(z)} \propto \sum_{n=1}^{N_{\mathrm{iter}}} \mathbf{1}_{z_i \leqslant \xi(\theta^n) \leqslant z_{i+1}}$$

More advanced strategies: **adaptive methods**
(Wang–Landau, Self-Healing Umbrella Sampling, well-tempered metadynamics, ...)

Convergence criterion: based on stabilization of RC (up to transformation)

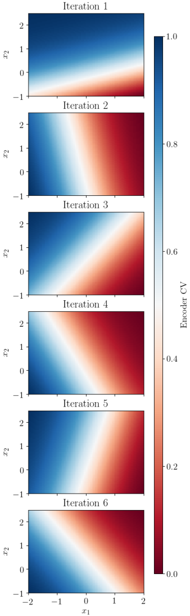
G. Fort, B. Jourdain, T. Lelièvre and G. Stoltz, *J. Stat. Phys.* (2018)
G. Fort, B. Jourdain, T. Lelièvre and G. Stoltz, *Stat. Comput.* (2017)
G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz, *Math. Comput.* (2015)

The iterative algorithm on the toy 2D example



Left: with reweighting
Convergence to CV $\simeq \theta_1$

Right: without reweighting
No convergence
(cycles between two CVs)



Conclusion and perspectives

Conclusion et perspectives

Methodology well established in molecular dynamics

- importance sampling with the free energy associated with a reaction coordinate
- find the reaction coordinate using ML approaches (here, autoencoders)
- iterate between update of the RC/recomputation of the free energy

Test the approach for problems in statistics?