Title: Doob's Lagrangian: A Sample-Efficient Variational Approach to Transition Path Sampling

Speakers: Kirill Neklyudov

Collection/Series: Special Seminars

Subject: Other

Date: November 29, 2024 - 9:30 AM

URL: https://pirsa.org/24110049

Abstract:

The 3rd talk of a monthly webinar series jointly hosted by Perimeter, IVADO, and Institut Courtois.

Rare event sampling in dynamical systems is a fundamental problem arising in the natural sciences, which poses significant computational challenges due to an exponentially large space of trajectories. For settings where the dynamical system of interest follows a Brownian motion with known drift, the question of conditioning the process to reach a given endpoint or desired rare event is definitively answered by Doob's h-transform. However, the naive estimation of this transform is infeasible, as it requires simulating sufficiently many forward trajectories to estimate rare event probabilities. In this talk, I'll present our recent findings on the variational formulation of Doob's h-transform as an optimization problem over trajectories between a given initial point and the desired ending point. To solve this optimization, we propose a simulation-free training objective with a model parameterization that imposes the desired boundary conditions by design. Our approach significantly reduces the search space over trajectories and avoids expensive trajectory simulation and inefficient importance sampling estimators which are required in existing methods. We demonstrate the ability of our method to find feasible transition paths on real-world molecular simulation and protein folding tasks.

Zoom link

Doob's Lagrangian: A Sample-Efficient Variational Approach to Transition Path Sampling

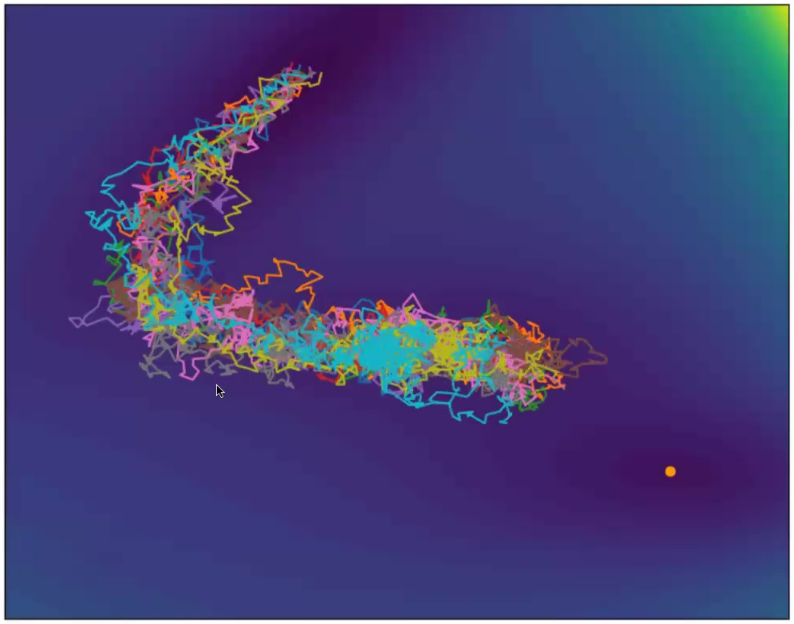
Yuanqi Du*, Michael Plainer*, Rob Brekelmans*,
Chenru Duan, Frank Noé, Carla P. Gomes, Alán Aspuru-Guzik,
Kirill Neklyudov

*equal contribution

Structure

1. Doob’s h-transform
   (Conditioning the Brownian motion on the end-point)
2. Doob’s Lagrangian
   (Variational formulation of Doob’s h-transform)
3. How to solve the Lagrangian formulation
   (Computational challenges and a way to overcome them)
4. Empirical study
   (Simulations for 2d potentials, a molecule, a protein)

Doob's h-transform



Let's say that the following SDE

$$dx_t = b_t(x_t) \cdot dt + \Xi_t dW_t\,, \quad x_0 \sim \rho_0\,,$$

defines the following transition probability

$$\rho_t(y\,|\,x_s = x)$$

Let's define the conditional probability

$$\rho_t(y\,|\,x_s = x, x_T \in \mathcal{B})$$

The corresponding SDE then includes an additional control drift term

$$dx_{t\,|\,T} = \big(b_t(x_{t\,|\,T}) + \underbrace{2G_t}_{\frac{1}{2}\Xi_t\Xi_t^T}\nabla_x \log \underbrace{h_{\mathcal{B}}}_{\rho_T(\mathcal{B}\,|\,x_t = x)}(x_{t\,|\,T}, t)\big) \cdot dt + \Xi_t dW_t\,, \quad x_0 \sim \rho_0$$

Doob’s h-transform

Once again, to condition this SDE

$$dx_t = b_t(x_t) \cdot dt + \Xi_t dW_t\,, \quad x_0 \sim \rho_0\,,$$

We have to add this term

$$dx_{t\,|\,T} = \big(b_t(x_{t\,|\,T}) + 2G_t \nabla_x \log h_{\mathcal{B}}(x_{t\,|\,T}, t)\big) \cdot dt + \Xi_t dW_t\,, \quad x_0 \sim \rho_0$$

However, h is very hard to evaluate, because it’s an integral

$h_{\mathcal{B}}(x,t) = \rho_T(\mathcal{B}\,|\,x_t = x)$ = integral over trajectories that start at x

$$= \int \mathbb{I}[x_T \in \mathcal{B}]\rho(x_{(t:T]}|x_t = x)dx_{(t:T]}$$

Doob's Lagrangian

Evaluating h directly is hard!

$$h_{\mathcal{B}}(x,t) = \rho_T(\mathcal{B}\,|\,x_t = x) = \int \mathbb{I}[x_T \in \mathcal{B}]\rho(x_{(t:T]}|x_t = x)dx_{(t:T]}$$

But why can we do better and don't give up?

We propose the following variational formulation

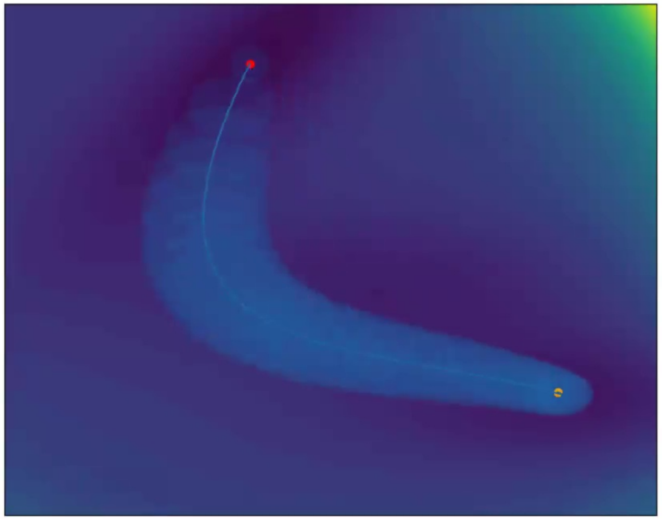
$$\mathcal{S} = \min_{q,v} \int_0^T dt \int dx\; q_{t\,|\,0,T}\big\langle v_{t\,|\,0,T}(x), G_t v_{t\,|\,0,T}(x)\big\rangle\,,$$

$$\text{s.t.}\quad \frac{\partial q_{t\,|\,0,T}}{\partial t} = -\big\langle \nabla_x, q_{t\,|\,0,T}\big(b_t(x) + 2G_t v_{t\,|\,0,T}(x)\big)\big\rangle + \sum_{ij}(G_t)_{ij}\frac{\partial^2}{\partial x_i \partial x_j} q_{t\,|\,0,T}(x)\,,$$

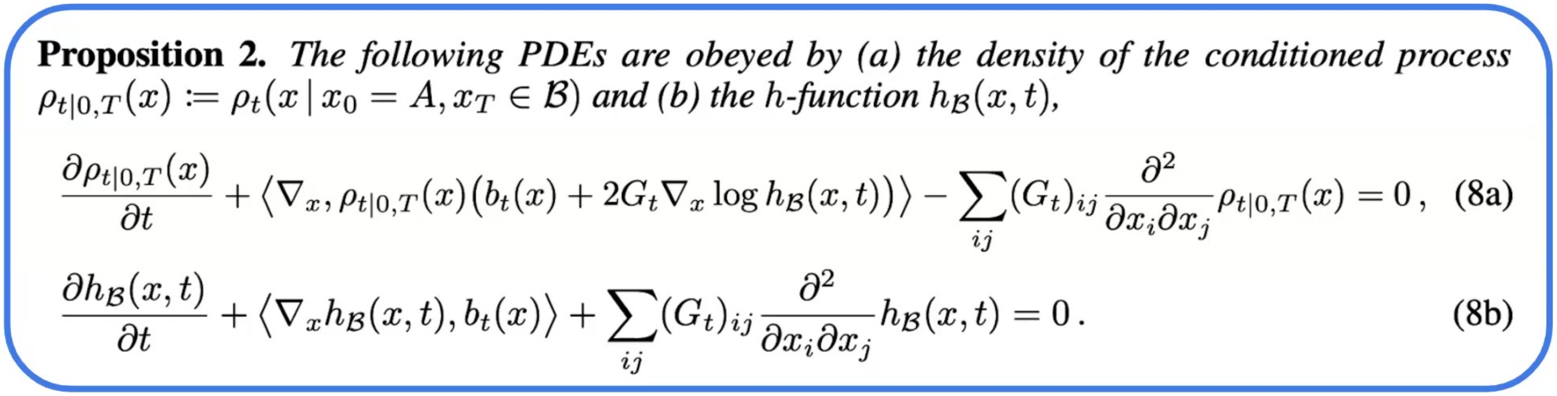
(the density q_t and the vector field v are connected through this)

$$q_0(x) = \delta(x - A)\,,\quad q_T(x) = \delta(x - B)\,.$$

(the density q_t has to satisfy the boundary conditions)



How one could possibly come up with Doob's Lagrangian?



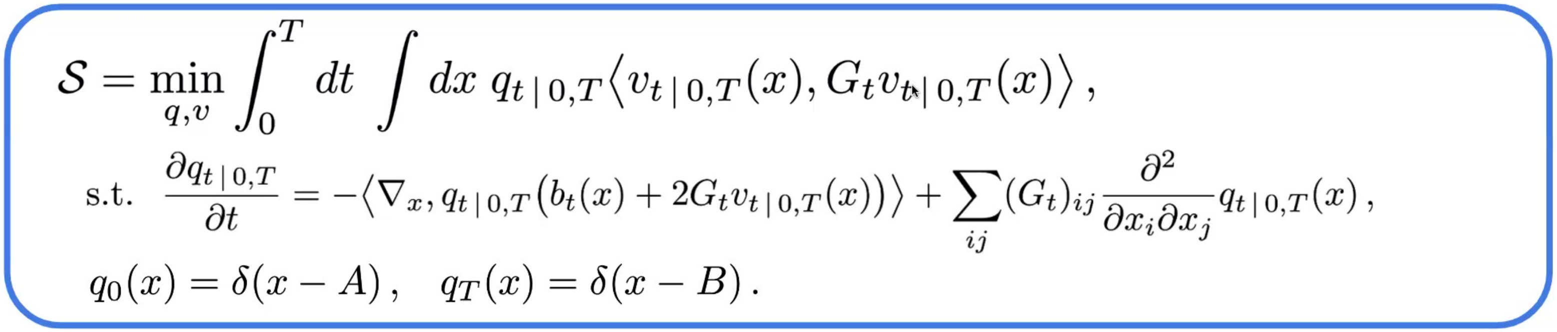
Proposition 2. *The following PDEs are obeyed by (a) the density of the conditioned process* $\rho_{t|0,T}(x) := \rho_t(x \,|\, x_0 = A, x_T \in \mathcal{B})$ *and (b) the h-function* $h_{\mathcal{B}}(x,t)$*,*

$$\frac{\partial \rho_{t|0,T}(x)}{\partial t} + \big\langle \nabla_x, \rho_{t|0,T}(x)\big(b_t(x) + 2G_t \nabla_x \log h_{\mathcal{B}}(x,t)\big)\big\rangle - \sum_{ij} (G_t)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \rho_{t|0,T}(x) = 0\,, \quad \text{(8a)}$$

$$\frac{\partial h_{\mathcal{B}}(x,t)}{\partial t} + \langle \nabla_x h_{\mathcal{B}}(x,t), b_t(x)\rangle + \sum_{ij} (G_t)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} h_{\mathcal{B}}(x,t) = 0\,. \quad \text{(8b)}$$

(Doob's h-transform in the form of PDEs)

We reverse-engineer the objective that has (8a) and (8b) as its optimum



$$\mathcal{S} = \min_{q,v} \int_0^T dt \int dx\; q_{t\,|\,0,T} \big\langle v_{t\,|\,0,T}(x), G_t v_{t\,|\,0,T}(x)\big\rangle\,,$$

$$\text{s.t.} \quad \frac{\partial q_{t\,|\,0,T}}{\partial t} = -\big\langle \nabla_x, q_{t\,|\,0,T}\big(b_t(x) + 2G_t v_{t\,|\,0,T}(x)\big)\big\rangle + \sum_{ij} (G_t)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} q_{t\,|\,0,T}(x)\,,$$

$$q_0(x) = \delta(x - A)\,, \quad q_T(x) = \delta(x - B)\,.$$

(Doob's h-transform in its variational form)

Connection to the Schrödinger Bridge problem

Consider the following optimization problem on path measures

$$\mathcal{S} := \min_{\mathbb{Q}^v_{0:T}} D_{\mathrm{KL}}[\mathbb{Q}^v_{0:T} : \mathbb{P}^{\mathrm{ref}}_{0:T}]$$

$$dx_t = b_t(x_t) \cdot dt + \Xi_t dW_t$$

$$dx_{t\,|\,T} = \big(b_t(x_{t\,|\,T}) + 2G_t v_{t\,|\,0,T}(x_{t\,|\,T}, t)\big) \cdot dt + \Xi_t dW_t$$

And the optimized path measure satisfy boundary constraints!!!

$$\mathbb{Q}^v_{0:T} \quad \text{s.t.} \quad \mathbb{Q}^v_0 = \delta_A \, , \quad \mathbb{Q}^v_0 = \delta_B$$

The MARGINALS of these paths = marginals from the Lagrangian formulation

Structure

1. Doob’s h-transform
   (Conditioning the Brownian motion on the end-point)
2. Doob’s Lagrangian
   (Variational formulation of Doob’s h-transform)
3. How to solve the Lagrangian formulation
   (Computational challenges and a way to overcome them)
4. Empirical study
   (Simulations for 2d potentials, a molecule, a protein)

Computational challenges of the Lagrangian formulation

1. We need samples to estimate the loss

$$\mathcal{S} = \min_{q,v} \int_0^T dt \int dx \; q_{t\,|\,0,T}(x) \big\langle v_{t\,|\,0,T}(x), G_t v_{t\,|\,0,T}(x) \big\rangle \,,$$

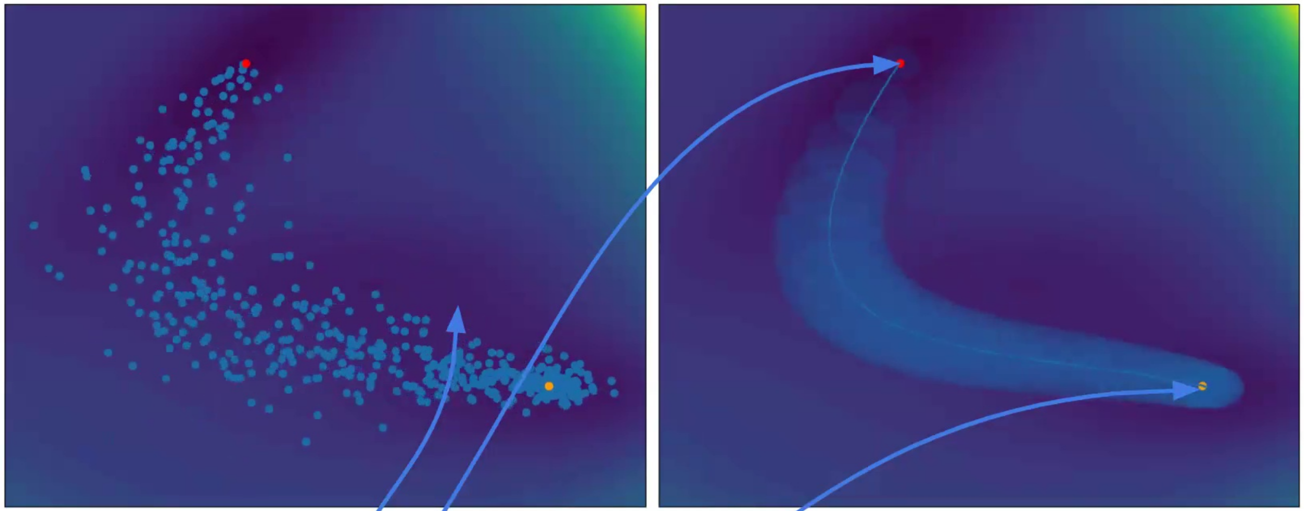
2. We have to satisfy boundary conditions

$$q_0(x) = \delta(x - A)\,, \quad q_T(x) = \delta(x - B)$$

3. We have to satisfy this relation between q and v

$$\text{s.t.} \quad \frac{\partial q_{t\,|\,0,T}}{\partial t} = -\big\langle \nabla_x, q_{t\,|\,0,T}\big(b_t(x) + 2G_t v_{t\,|\,0,T}(x)\big)\big\rangle + \sum_{ij} (G_t)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} q_{t\,|\,0,T}(x)\,,$$

Gaussian parameterization (challenges 1 and 2)



We parameterize every time marginal as a Gaussian

$$x_{t\,|\,0,T} = \mu^{(\theta)}_{t\,|\,0,T} + \Sigma^{(\theta)}_{t\,|\,0,T}\varepsilon, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \mathbb{I}_D)$$

(this is very easy to sample)

$$\mu^{(\theta)}_{t\,|\,0,T} = \left(1 - \frac{t}{T}\right)A + \frac{t}{T}B + \frac{t}{T}\left(1 - \frac{t}{T}\right)\mathrm{NNET}_\theta(t, A, B)_{[:D]}$$

$$\Sigma^{(\theta)}_{t\,|\,0,T} = \frac{t}{T}\left(1 - \frac{t}{T}\right)\mathtt{diag}\left(\mathrm{NNET}_\theta(t, A, B)_{[D:]}\right) + \sigma^2_{\min}\mathbb{I}_D$$

(very easy to guarantee the boundary conditions)

Gaussian parameterization (challenge 3)

(connect the vector field and the marginals)

Finally, taking Gaussian q_t in the Fokker-Planck equation

$$\frac{\partial q_{t\,|\,0,T}(x)}{\partial t} = -\langle \nabla_x, q_{t\,|\,0,T}(x) u_{t\,|\,0,T}(x) \rangle + \sum_{ij} (G_t)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} q_{t\,|\,0,T}(x)$$

The vector field has an analytic formula

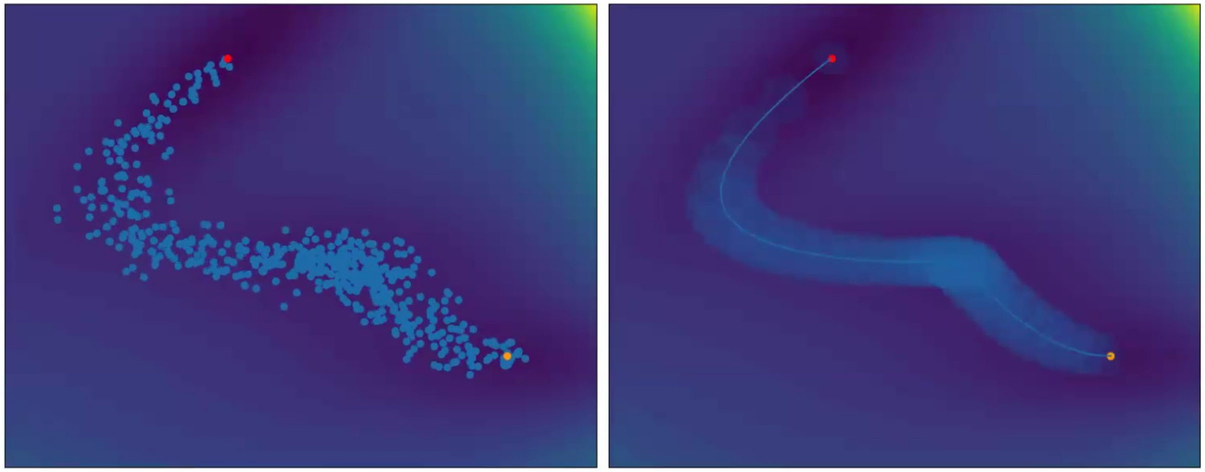
$$u^{(\theta)}_{t\,|\,0,T}(x) \coloneqq \frac{\partial \mu_{t\,|\,0,T}}{\partial t} + \left[\frac{1}{2}\frac{\partial \Sigma_{t\,|\,0,T}}{\partial t}\Sigma^{-1}_{t\,|\,0,T} - G_t \Sigma^{-1}_{t\,|\,0,T}\right](x - \mu_{t\,|\,0,T})$$

Thus, accounting for the drift we have the analytic formula for v_t based on given q_t

$$v^{(\theta)}_{t\,|\,0,T}(x) = \frac{1}{2} G_t^{-1} (\mu^{(\theta)}_{t\,|\,0,T}(x) - b_t(x))$$

The training process (although, there is no dataset)

```
Algorithm 1: Training (Single Gaussian)
  Input: Reference drift $b_t$, diffusion matrix $G_t$
         Conditioning endpoints
  while not converged do
    Sample $t \sim \mathcal{U}(0, T)$
    Sample $x_t \sim q^{(\theta)}_{t|0,T}$
    Calculate $u^{(q,\theta)}_{t|0,T}(x_t)$
    Calculate $v^{(q,\theta)}_{t|0,T}(x_t)$ using $u^{(q,\theta)}_{t|0,T}(x_t)$, $b_t(x_t)$
➡   Calculate $\mathcal{L} = \langle v^{(q,\theta)}_{t|0,T}(x_t), G_t\, v^{(q,\theta)}_{t|0,T}(x_t)\rangle$ (Thm. 1)
    Update $\theta \leftarrow \text{optimizer}(\theta, \nabla_\theta \mathcal{L})$
  end while
  return $\theta$
```



Path parameterization

$$x_{t\,|\,0,T} = \mu^{(\theta)}_{t\,|\,0,T} + \Sigma^{(\theta)}_{t\,|\,0,T}\varepsilon\,, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \mathbb{I}_D)$$

$$\mu^{(\theta)}_{t\,|\,0,T} = \left(1 - \frac{t}{T}\right)A + \frac{t}{T}B + \frac{t}{T}\left(1 - \frac{t}{T}\right)\text{NNET}_\theta(t, A, B)_{[:D]}$$

$$\Sigma^{(\theta)}_{t\,|\,0,T} = \frac{t}{T}\left(1 - \frac{t}{T}\right)\texttt{diag}\big(\text{NNET}_\theta(t, A, B)_{[D:]}\big) + \sigma^2_{\min}\mathbb{I}_D$$

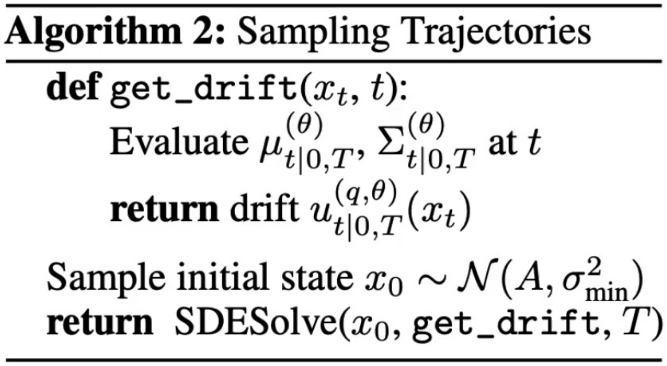
Corresponding Vector Fields

$$u^{(\theta)}_{t\,|\,0,T}(x) := \frac{\partial \mu_{t\,|\,0,T}}{\partial t} + \left[\frac{1}{2}\frac{\partial \Sigma_{t\,|\,0,T}}{\partial t}\Sigma^{-1}_{t\,|\,0,T} - G_t\Sigma^{-1}_{t\,|\,0,T}\right](x - \mu_{t\,|\,0,T})$$

$$v^{(\theta)}_{t\,|\,0,T}(x) = \frac{1}{2}G_t^{-1}(\mu^{(\theta)}_{t\,|\,0,T}(x) - b_t(x))$$

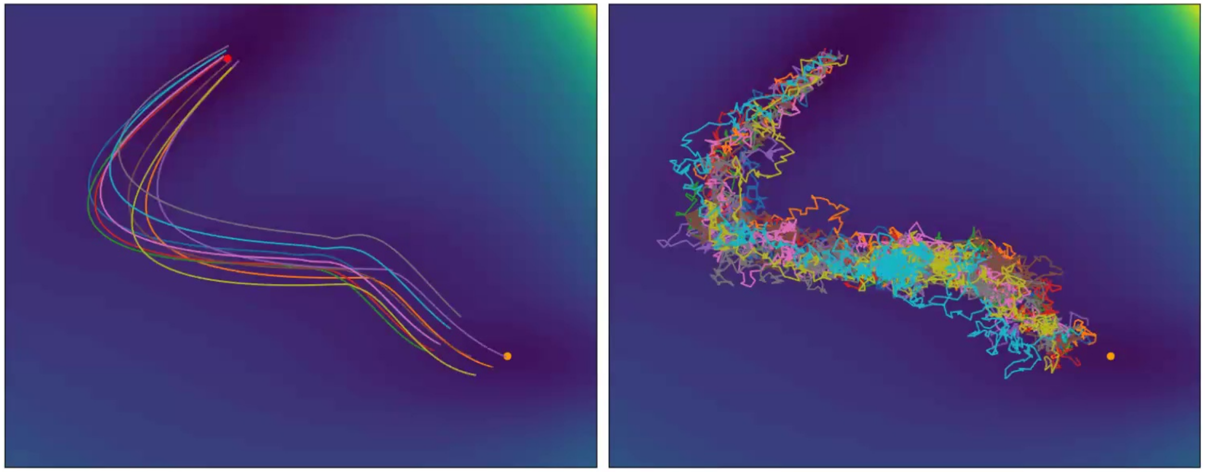
$$\mathcal{S} = \min_{q,v}\int_0^T dt \int dx\; q_{t\,|\,0,T}(x)\big\langle v_{t\,|\,0,T}(x), G_t v_{t\,|\,0,T}(x)\big\rangle$$

The inference process



Algorithm 2: Sampling Trajectories

```
def get_drift($x_t$, $t$):
    Evaluate $\mu^{(\theta)}_{t|0,T}$, $\Sigma^{(\theta)}_{t|0,T}$ at $t$
    return drift $u^{(q,\theta)}_{t|0,T}(x_t)$
Sample initial state $x_0 \sim \mathcal{N}(A, \sigma^2_{\text{min}})$
return SDESolve($x_0$, get_drift, $T$)
```



$$u^{(q,\theta)}_{t|0,T}(x) := \frac{\partial \mu_{t|0,T}}{\partial t} + \left[\frac{1}{2}\frac{\partial \Sigma_{t|0,T}}{\partial t}\Sigma^{-1}_{t|0,T} - G_t\, \Sigma^{-1}_{t|0,T}\right]\left(x - \mu_{t|0,T}\right)$$

(this is it)

Structure

1. Doob’s h-transform
   (Conditioning the Brownian motion on the end-point)
2. Doob’s Lagrangian
   (Variational formulation of Doob’s h-transform)
3. How to solve the Lagrangian formulation
   (Computational challenges and a way to overcome them)
4. Empirical study
   (Simulations for 2d potentials, a molecule, a protein)

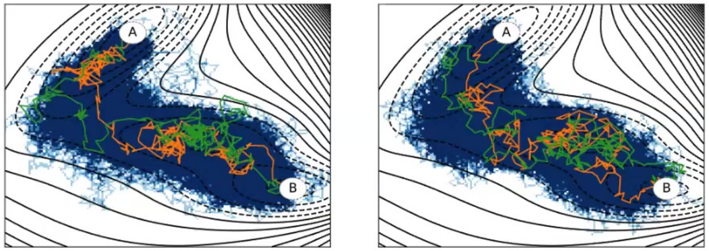
2D potentials

Toy-data analysis:

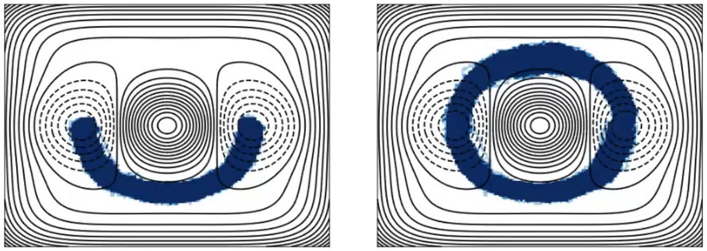
1. Our learned distribution is close to the target distribution generated by MCMC

2. We can sample **multiple paths** by considering **mixtures of Gaussians**

3. We find transition paths with **fewer energy evaluations** than baselines



(a) MCMC (b) Ours

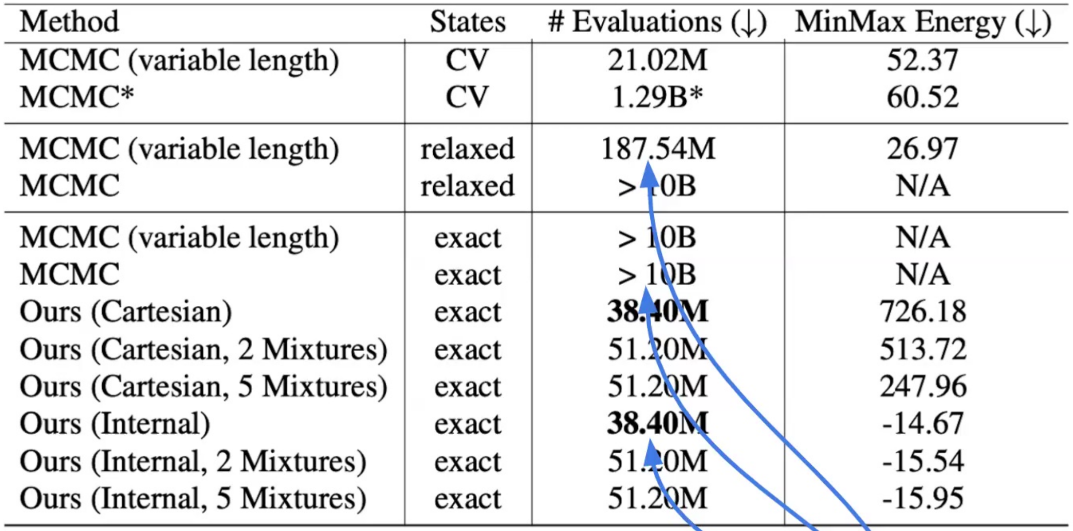


(a) Single Gaussian (b) Mixture of Gaussians

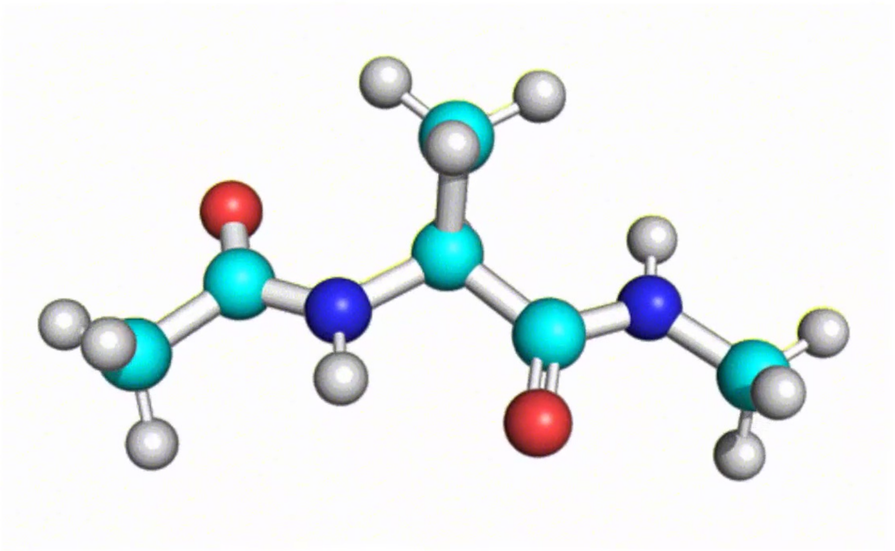
Method	# Evaluations (↓)	Max Energy (↓)	MinMax Energy (↓)	Log-Likelihood (↑)	Max Log-Likelihood (↑)
MCMC (variable)	3.53M	-13.77 ± 16.43	-40.75	-	-
MCMC	1.03B	-17.80 ± 14.77	-40.21	866.56 ± 17.00	907.15
Ours	**1.28M**	-14.81 ± 13.73	-40.56	858.50 ± 17.61	909.74

Alanine Dipeptide molecule

We need fewer energy evaluations under less assumptions!



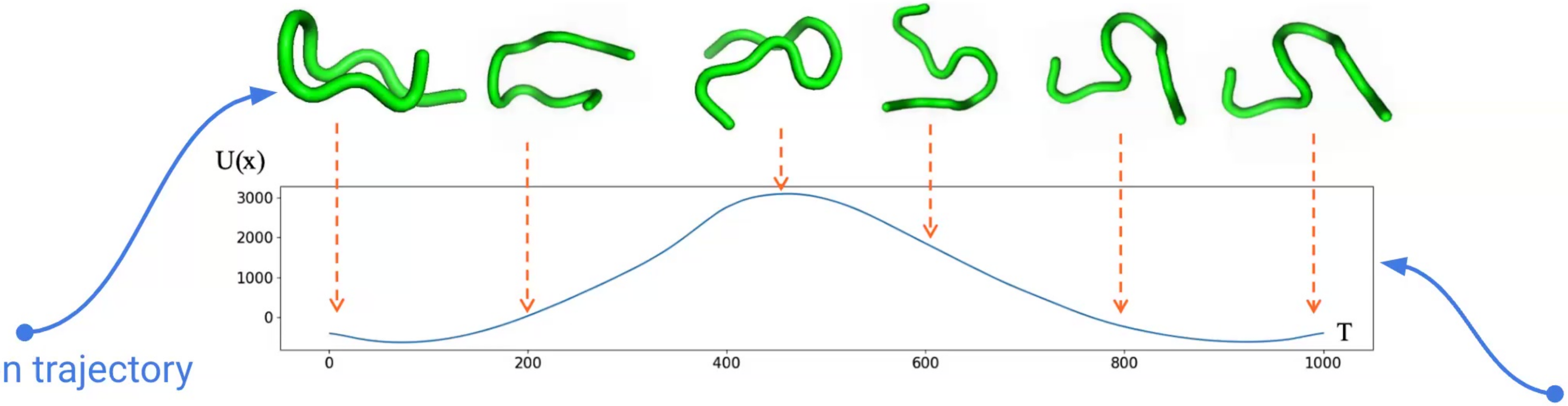
Method	States	# Evaluations (↓)	MinMax Energy (↓)
MCMC (variable length)	CV	21.02M	52.37
MCMC*	CV	1.29B*	60.52
MCMC (variable length)	relaxed	187.54M	26.97
MCMC	relaxed	> 10B	N/A
MCMC (variable length)	exact	> 10B	N/A
MCMC	exact	> 10B	N/A
Ours (Cartesian)	exact	**38.40M**	726.18
Ours (Cartesian, 2 Mixtures)	exact	51.20M	513.72
Ours (Cartesian, 5 Mixtures)	exact	51.20M	247.96
Ours (Internal)	exact	**38.40M**	-14.67
Ours (Internal, 2 Mixtures)	exact	51.20M	-15.54
Ours (Internal, 5 Mixtures)	exact	51.20M	-15.95



(This is sampled without any knowledge of collective variables (CV))

MCMC algorithms cannot find the path even after 10B evaluations. For easier problem they still need 5x more compute

Chignolin small protein



Transition trajectory

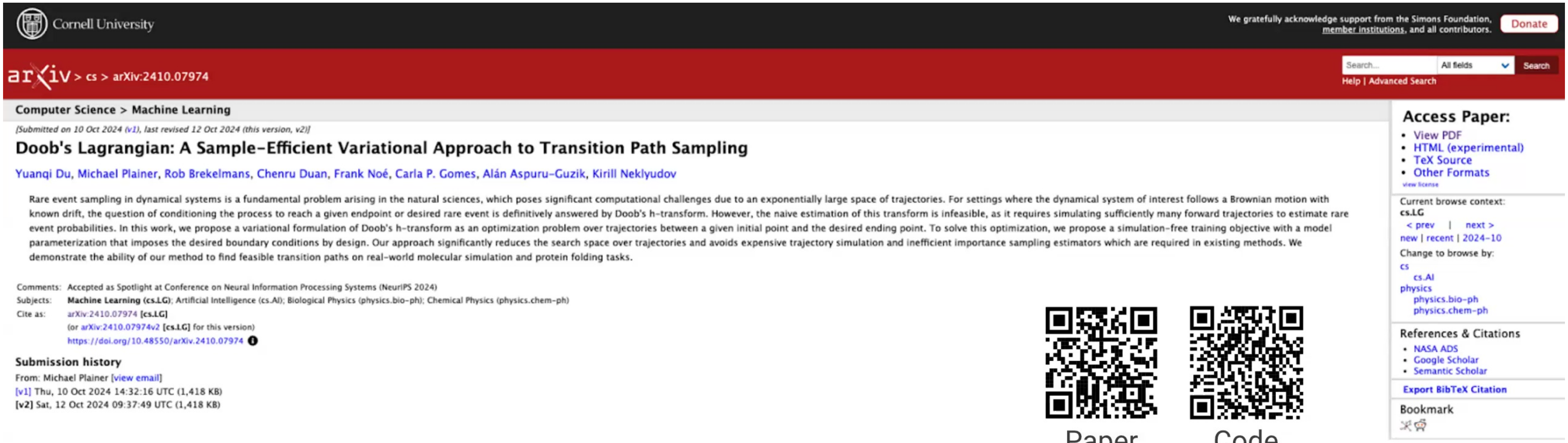
Energy along the trajectory

Our algorithm finds a transition path for the Chignolin protein

1. Chignolin consists of 166 atoms (with the total dimensionality 166x3x1000)

2. We simulate everything in Cartesian Coordinates

3. These are preliminary results and there is a lot of room for improvement



Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv > cs > arXiv:2410.07974

Search... All fields Search
Help | Advanced Search

Computer Science > Machine Learning

[Submitted on 10 Oct 2024 (v1), last revised 12 Oct 2024 (this version, v2)]

Doob's Lagrangian: A Sample-Efficient Variational Approach to Transition Path Sampling

Yuanqi Du, Michael Plainer, Rob Brekelmans, Chenru Duan, Frank Noé, Carla P. Gomes, Alán Aspuru-Guzik, Kirill Neklyudov

Rare event sampling in dynamical systems is a fundamental problem arising in the natural sciences, which poses significant computational challenges due to an exponentially large space of trajectories. For settings where the dynamical system of interest follows a Brownian motion with known drift, the question of conditioning the process to reach a given endpoint or desired rare event is definitively answered by Doob's h-transform. However, the naive estimation of this transform is infeasible, as it requires simulating sufficiently many forward trajectories to estimate rare event probabilities. In this work, we propose a variational formulation of Doob's h-transform as an optimization problem over trajectories between a given initial point and the desired ending point. To solve this optimization, we propose a simulation-free training objective with a model parameterization that imposes the desired boundary conditions by design. Our approach significantly reduces the search space over trajectories and avoids expensive trajectory simulation and inefficient importance sampling estimators which are required in existing methods. We demonstrate the ability of our method to find feasible transition paths on real-world molecular simulation and protein folding tasks.

Comments:	Accepted as Spotlight at Conference on Neural Information Processing Systems (NeurIPS 2024)
Subjects:	**Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Biological Physics (physics.bio-ph); Chemical Physics (physics.chem-ph)
Cite as:	arXiv:2410.07974 **[cs.LG]**
	(or arXiv:2410.07974v2 **[cs.LG]** for this version)
	https://doi.org/10.48550/arXiv.2410.07974

Submission history

From: Michael Plainer [view email]
[v1] Thu, 10 Oct 2024 14:32:16 UTC (1,418 KB)
[v2] Sat, 12 Oct 2024 09:37:49 UTC (1,418 KB)

Access Paper:

- View PDF
- HTML (experimental)
- TeX Source
- Other Formats

view license

Current browse context:
cs.LG
< prev | next >
new | recent | 2024-10
Change to browse by:
cs
cs.AI
physics
physics.bio-ph
physics.chem-ph

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

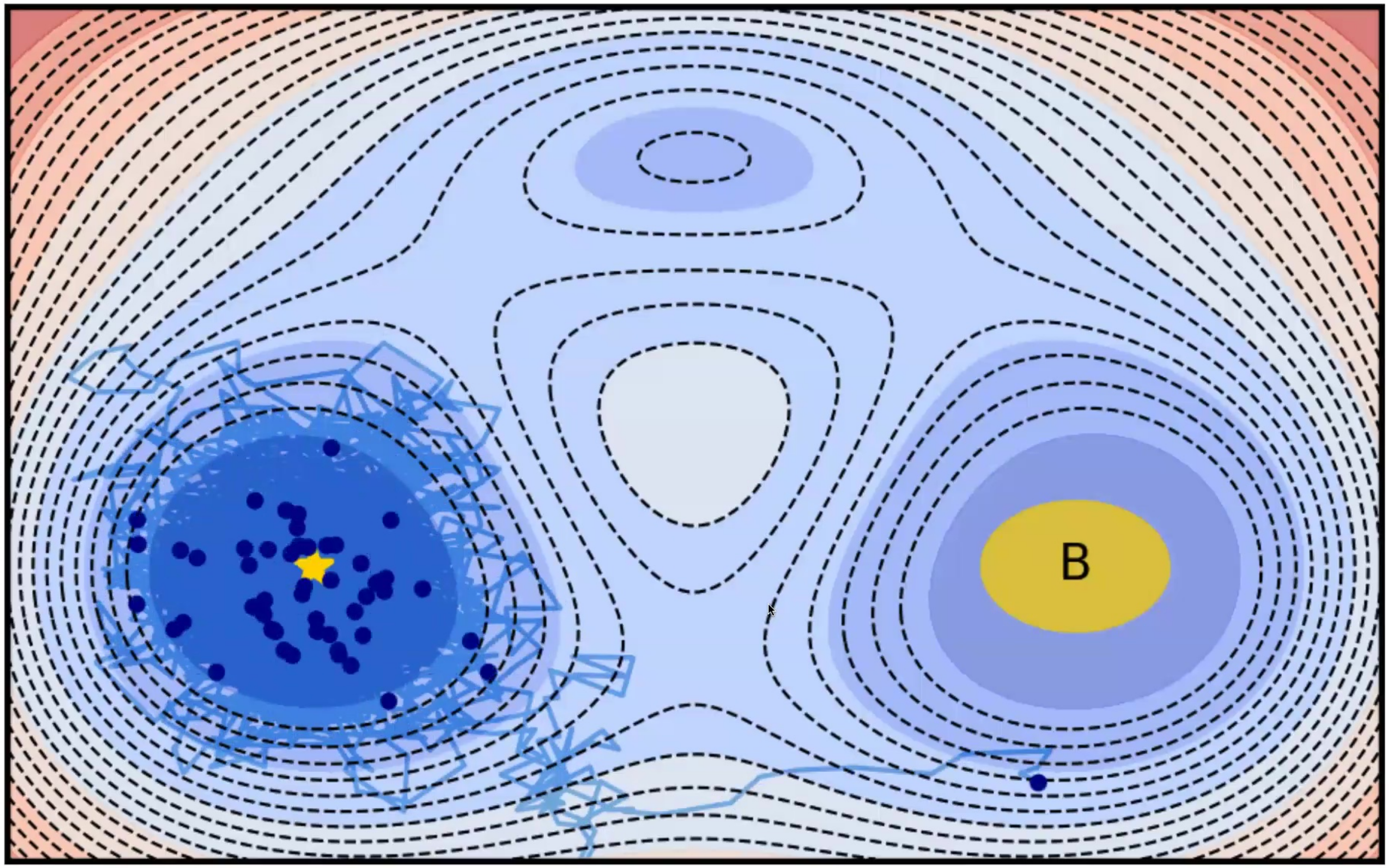
Bookmark

Paper Code

You will find more details in the paper!

- Why simply optimizing splines does not work?
- How to use the mixture of Gaussians and what's the corresponding vector field?
- Finally, we have the code for all experiments polished and open-sourced!

Transition Path Sampling



Transition Path Sampling