Transport-based Generative Models & Sampling

Cargese – August 05, 2023

Statistical physics and machine learning back together again

Marylou Gabrié

CMAP, École Polytechnique

Based on works with:

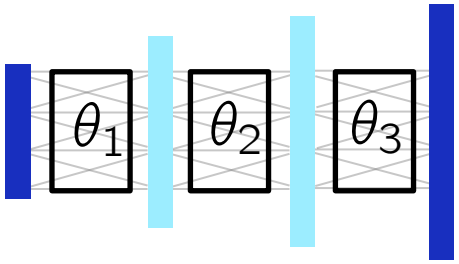
Grant Rotskoff (Stanford), Eric Vanden-Eijnden (Courant Institute, NYU)

Louis Grenioux, Alain Oliviero Durmus & Éric Moulines (École Polytechnique)

Deep generative models based on transport

▷ Deep latent generative models

$$T_\theta : \mathbb{R}^b \to \mathbb{R}^d$$

$z \sim \rho_B(z)$ 

$x = T_\theta(z) \quad \sim \rho_\theta(x)$ "push-forward" distribution

New dog picture for each new base variable!

Song et al. *ICLR* 2021

▷ Invertible map/transport based models: $T_\theta : \Omega \mapsto \Omega$

(Normalizing flows (discrete time), Neural ODEs, Score-based diffusion models, etc.)

$$\rho_\theta(x) = \rho_B(T_\theta^{-1}(x)) \det \left| \nabla_x T_\theta^{-1} \right|$$

▷ Training

- From data samples $\{x_i\}_{i=1}^N$ assumed to be i.i.d
  - Maximum likelihood: minimize $L[\rho_\theta] = -\sum_{i=1}^{N} \log \rho_\theta(x_i)$
  - Score based objectives (cf Eric's talk)

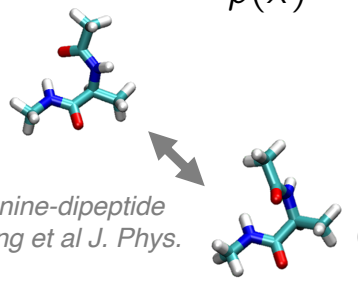
["Normalizing flows" Tabak & V.-E. *Commun. Math. Sci.* 2010, Dinh et al *ICLR* 2017, Papamakarios et al. *JMLR* 2021, "Score based diffusion models" Song et al. *ICLR 2021,* "Stochastic interpolants [..]" Albergo et al. arxiv:2303.08797 *etc.*]

High-dimensional inference

- A distribution of interest known up to normalization (posterior/Boltzmann)

ex: molecular configurations

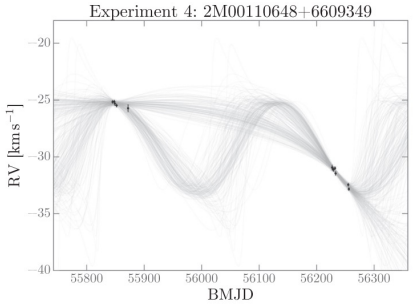
$$\rho(x) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta U(x)}$$



Alanine-dipeptide
Jiang et al J. Phys.

ex: Bayesian model parameters

$$\rho(x|D) = \frac{1}{\mathcal{Z}_D} \rho(D|x)\rho(x)$$



Price-Whelan et al. *The Astrophysical Journal* 2017

- Direct inference is intractable $\mathbb{E}_\rho[f(x)] = \int_\Omega f(x)\rho(x)\mathrm{d}x$, costs $O(e^d)$ for $x \in \mathbb{R}^d$
- Monte Carlo methods rely on samples: $\mathbb{E}_\rho[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ for $x_i \sim \rho(x)$
- Sampling itself can be challenging (dimensionality, geometry, mutlimodality)
- **This talk:** Can we use deep generative models to speed up sampling?

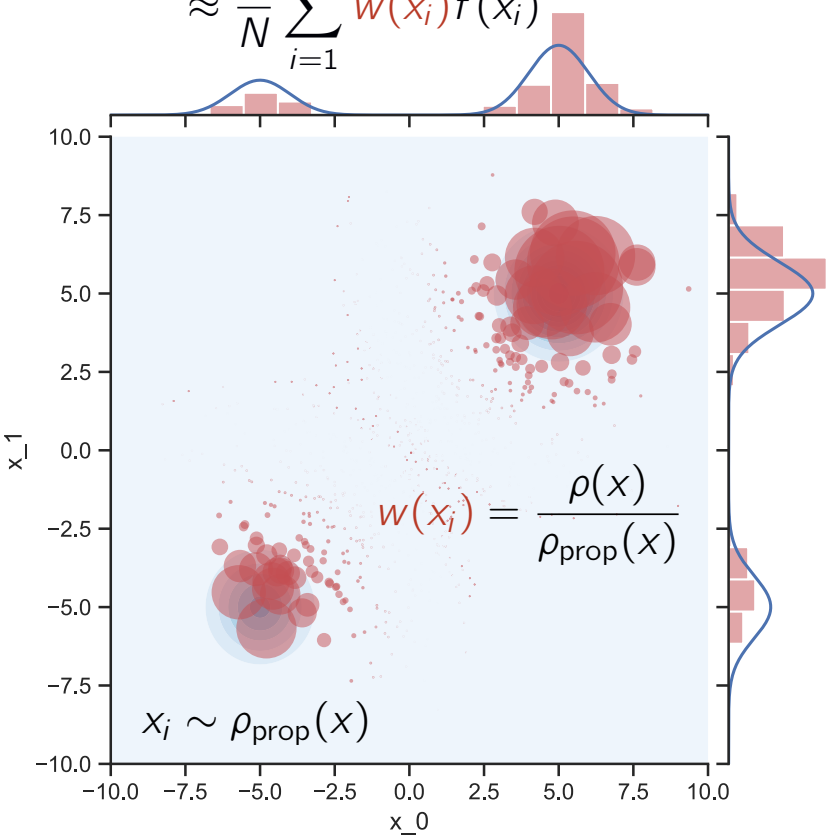
Why is sampling hard? (simple samplers)

- Importance sampling

  rely on tractable proposal

$$\mathbb{E}_\rho[f(x)] = \int_\Omega f(x)\rho(x)\mathrm{d}x$$

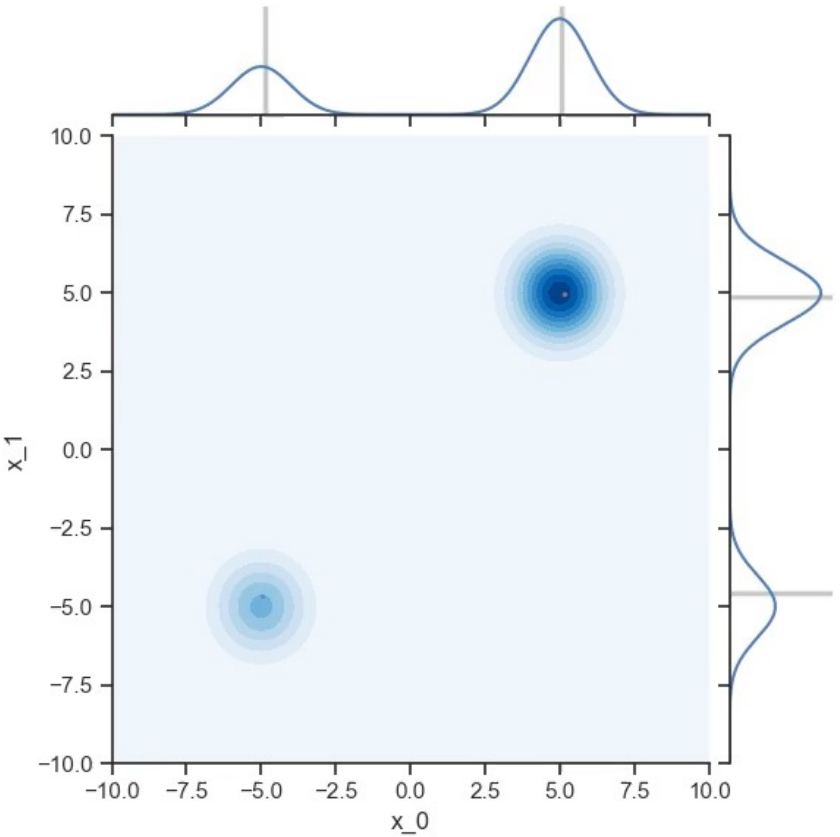
$$\approx \frac{1}{N}\sum_{i=1}^{N} w(x_i) f(x_i)$$



- Markov Chain Monte Carlo:
  e.g. Metropolis Hastings

  rely on local proposal

$$\rho_{\text{prop}}(x_{t+1}|x_t) = \mathcal{N}(x_t - dt\nabla U(x), \sqrt{2dt}I_d)$$

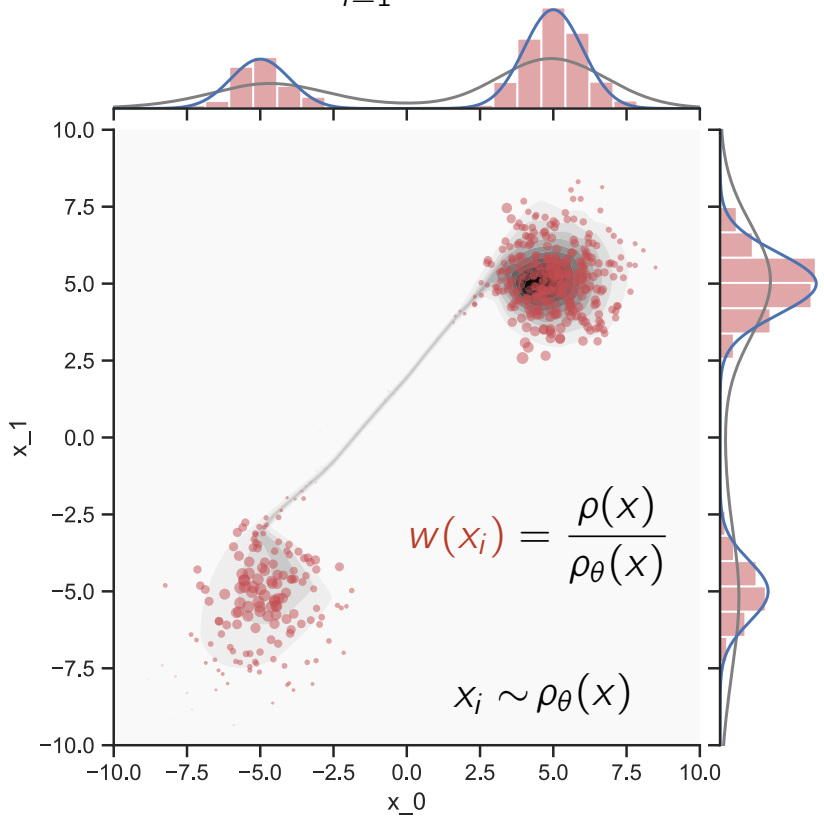


Suppose you can train a model $\rho_\theta(x) \approx \rho_*(x)$, what do you gain?

▷ Importance sampling

*rely on **adpated** tractable proposal!*

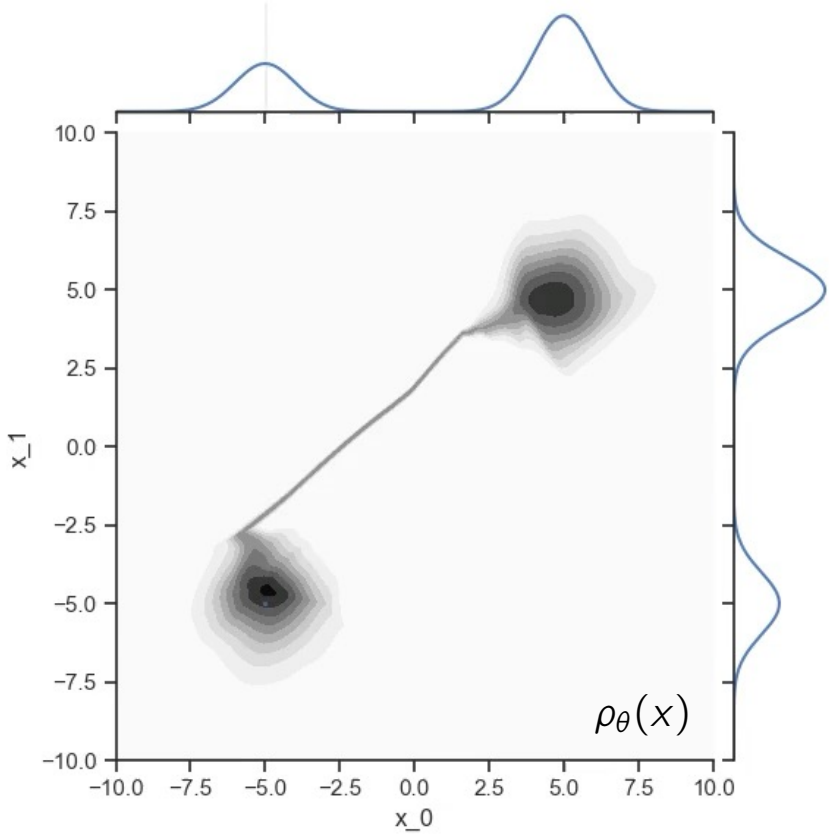
$$\mathbb{E}_\rho[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N} w(x_i) f(x_i)$$



▷ Markov Chain Monte Carlo:
e.g. Metropolis Hastings

*rely on **global** proposal!*

$$\rho_{\text{prop}}(x_{t+1}|x_t) = \rho_\theta(x_{t+1})$$



Suppose you can train a model $\rho_\theta(x) \approx \rho_*(x)$, what do you gain?

$\triangleright$ A lot!

Disclaimer: Not to say that more sophisticated samplers do not exist!

- ▷ Use geometric/gradient information: Hybrid/Hamiltonian MC
  - Hamiltonian MC [Duane et al, Hybrid Monte Carlo. 1987 …]
  - No U-Turn Sampler (NUTS) [Hoffman & Gelman. 2014]
- ▷ Adaptive biasing technics [e.g. Comer at al. *J. Phys. Chem. B* 2015]
- ▷ Gradually approach the target: MCMC/IS within a larger scheme
  - Parallel tempering/Replica exchange [Swendsen & Wang 1986, Geyer 1991 ...]
  - Annealed Importance Sampling [R. Neal 2001 …]
  - Sequential Monte Carlo [Del Moral, Doucet & Jasra. 2006 …]
- ▷ Non-exhaustive

Suppose you can train a model $\rho_\theta(x) \approx \rho_*(x)$, what do you gain?

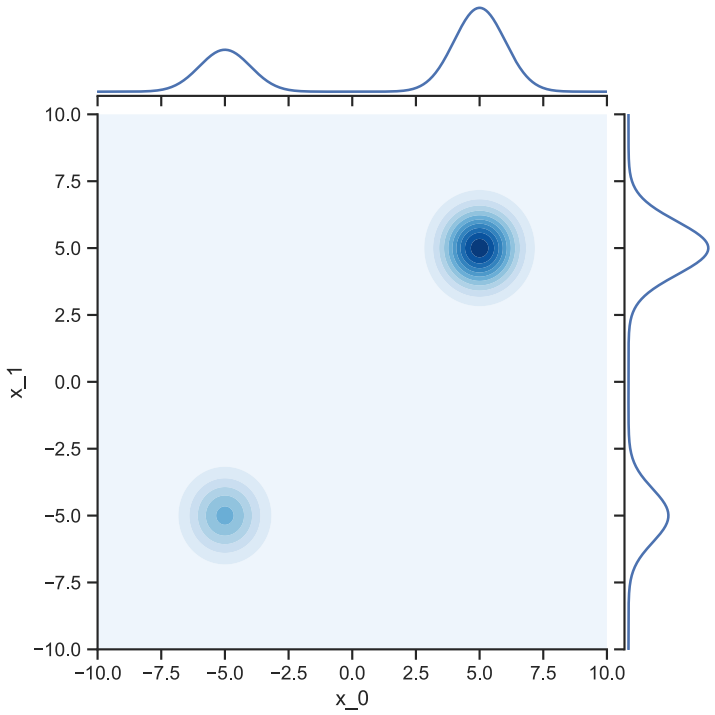
▷ A lot!

Variational inference “on steroids”:

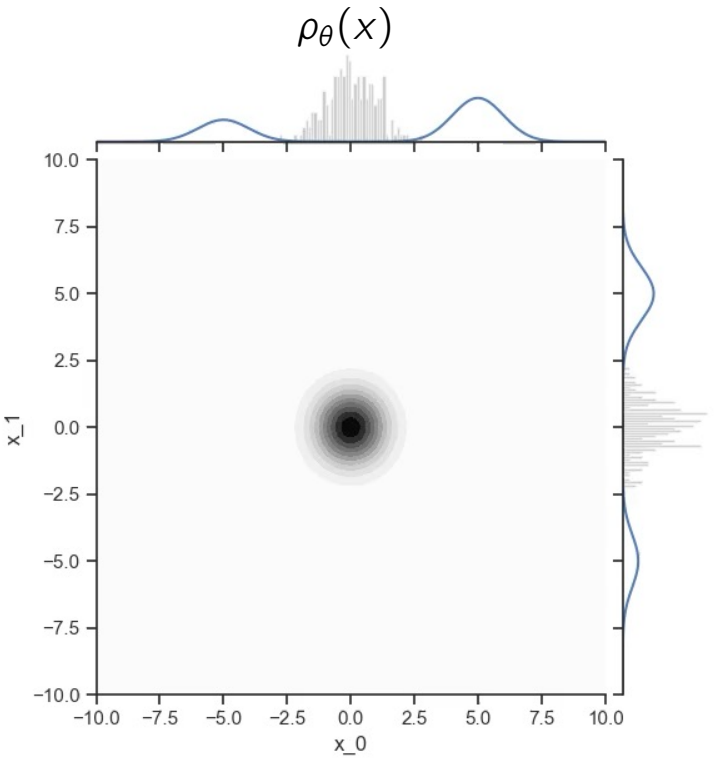
▷ A data-free learning objective: the (reverse) Kullback-Leibler divergence $D_{\mathrm{KL}}(\rho_\theta \| \rho_*)$

$$D_{\mathrm{KL}}(\rho_\theta \| \rho_*) = \int \log \frac{\rho_\theta(x)}{\rho_*(x)} \rho_\theta(x) \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{\rho_B(z_i) \det |\nabla_{z_i} T_\theta|}{\rho_*(T_\theta(z_i))} \qquad z_i \sim \rho_B(z)$$

$\rho_*(x)$



$\rho_\theta(x)$



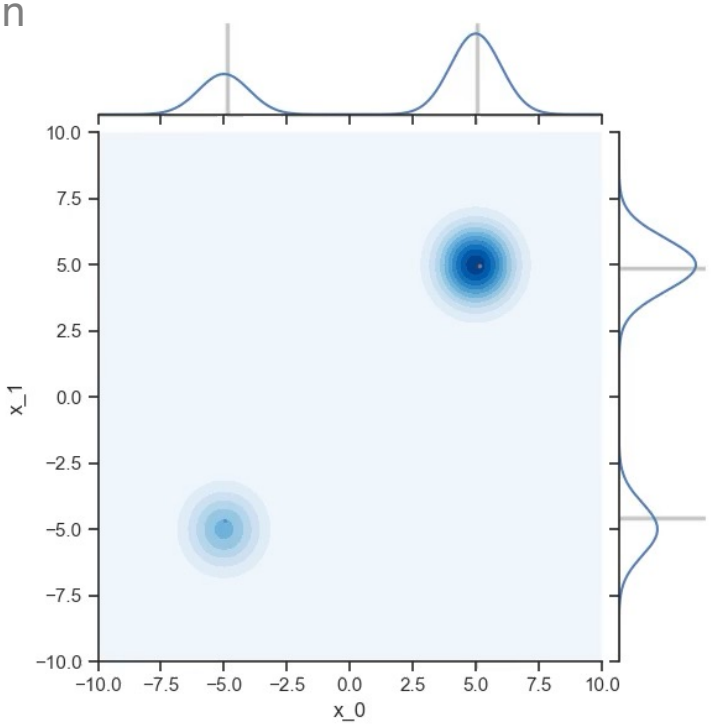
Weiss, P. (1907). L’hypothèse du champ moléculaire et la propriété ferromagnétique.
Wainwright, M. J., & Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference.
Rezende & Mohamed, (2015). Variational inference with normalizing flows
Wu et al. (2019). Solving Statistical Mechanics Using Variational Autoregressive Networks.
Albergo et al (2019). Flow-based generative models for Markov chain Monte Carlo in lattice field theory.
Li, Shuo-Hui, and Lei Wang. “Neural Network Renormalization Group.” *PRL 2018*

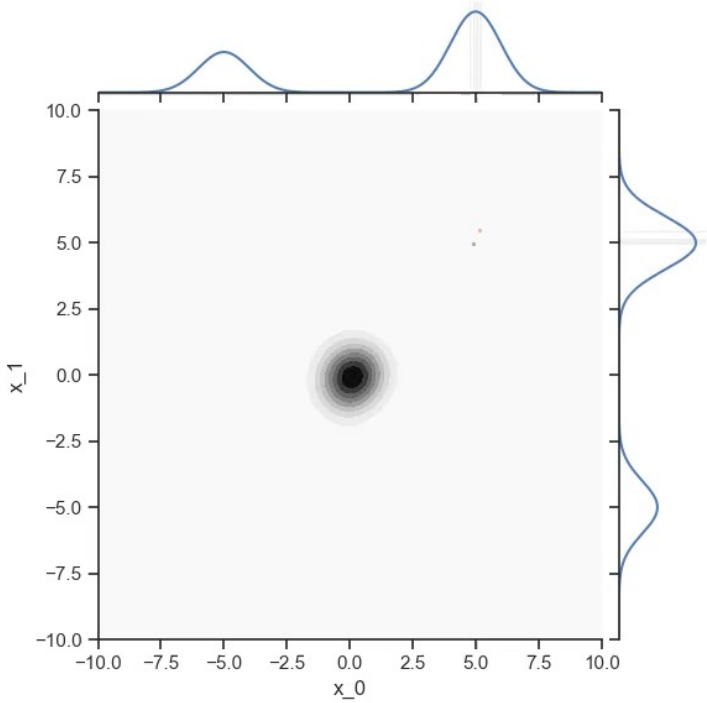
Adaptive Markov Chain Monte Carlo: simultaneous convergence of training & sampling

- ▷ Maximum likelihood objective (a.k.a forward KL)

$$D_{\mathrm{KL}}(\rho_* \| \rho_\theta) = \int \log \frac{\rho_\theta(x)}{\rho_*(x)} \rho_*(x) \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} \log \rho_\theta(x_i) + \mathrm{Cst} \qquad x \sim \rho_*(x_i)$$

- ▷ Use an auxillary simple sampler to create data $x^i_{t+1} \sim \pi_{\mathrm{local}}(x^i_{t+1} | x^i_t)$
- ▷ Train





- ▷ Add non-local steps relying on flow proposals $x \sim \rho_\theta(x_i)$
- ▷ Repeat and converge

No free lunch!

[Parno & Marzouk 2018 SIAM Journal of Uncertainty; Naesseth et al. Neurips 2020; MG, Rotskoff, Vanden-Eijnden PNAS 2022]

Adaptive Markov Chain Monte Carlo: simultaneous convergence of training & sampling

▷ Adaptive MCMCs [Haario at al. *Bernoulli* 2001, Jasra et al Statistics and Computing, 2007, Andrieu et al. *Bernoulli* 2011, Sejdinovic et al *ICML* 2014 …]

▷ Algorithm: Metropolis-Hastings with **non-local** generative model proposal

Initialize: $x_0^i \quad i = 1 \cdots N$

Loop:

Loop over parallel chains: $i = 1 \cdots N$

Metropolis-Hastings with NF

Draw from generative model $x_{t+1}^i \sim \rho_\theta(x)$

Accept-reject $\mathrm{acc}(x_{t+1}^i | x_t^i) = \min\left[1, \frac{\boldsymbol{\rho}_*(x_{t+1}^i)\rho_\theta(x_t^i)}{\boldsymbol{\rho}_*(x_t^i)\rho_\theta(x_{t+1}^i)}\right]$

Local resampling $x_{t+1}^i \sim \pi_{\text{local}}(x_{t+1}^i | x_t^i)$

MCMC steps converging to the target

Update NF parameters $\theta \leftarrow \theta + \eta \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log \rho_\theta(x_{t+1}^i)$

Maximum likelihood GD

Improve surrogate as you go

▷ Local + Mode jumping: [Tjelmeland & Hegstad *Scandinavian J. of Statistics* 2001, Sminchisescu & Welling *AISTAT* 2017, Pompe et al. *Ann. Stat* 2020, Sbailò et al. *J. Chem. Phys.* 2021, Tawn, Moores & Roberts 2021 …]

[Parno & Marzouk 2018 SIAM Journal of Uncertainty; Naesseth et al. Neurips 2020; MG, Rotskoff, Vanden-Eijnden PNAS 2022]

Variational inference

$$\min_\theta D_{\mathrm{KL}}(\rho_\theta \| \rho_*)$$

Adaptive MCMCs

$$\max_\theta \log \rho_\theta(x^t)$$

Suppose you can train a model $\rho_\theta(x) \approx \rho_*(x)$, what do you gain?

▷ A lot!

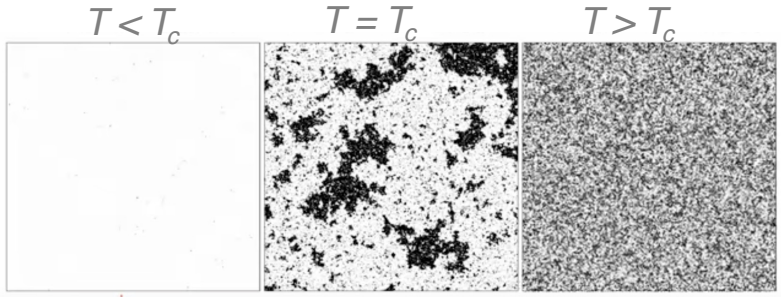
How does the performance depend on the quality of the approximation?

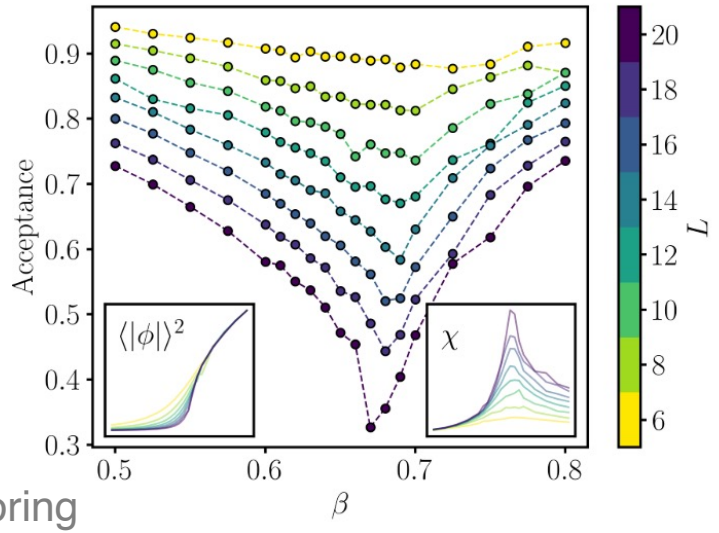
- Ideal case: $\rho_\theta(x) = \rho_*(x)$, then done!

- In practice $\rho_\theta(x) \approx \rho_*(x)$ due to estimation & approxiamtion errors: if the distributions are too different, $\rho_\theta(x)$ is "useless"

- A few papers have tried non-trivial stat-mech models with little success

  - Del Debbio et al. 2021: Φ^4 model scalability





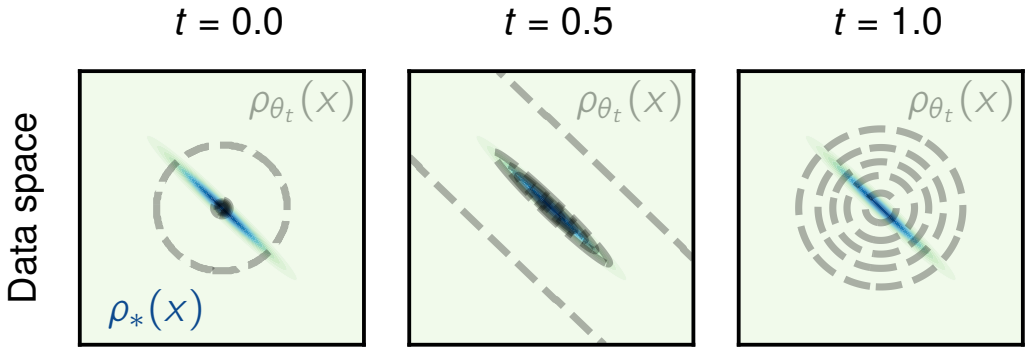
  - Trinquier, Zamponi et al. 2023: Random graph coloring

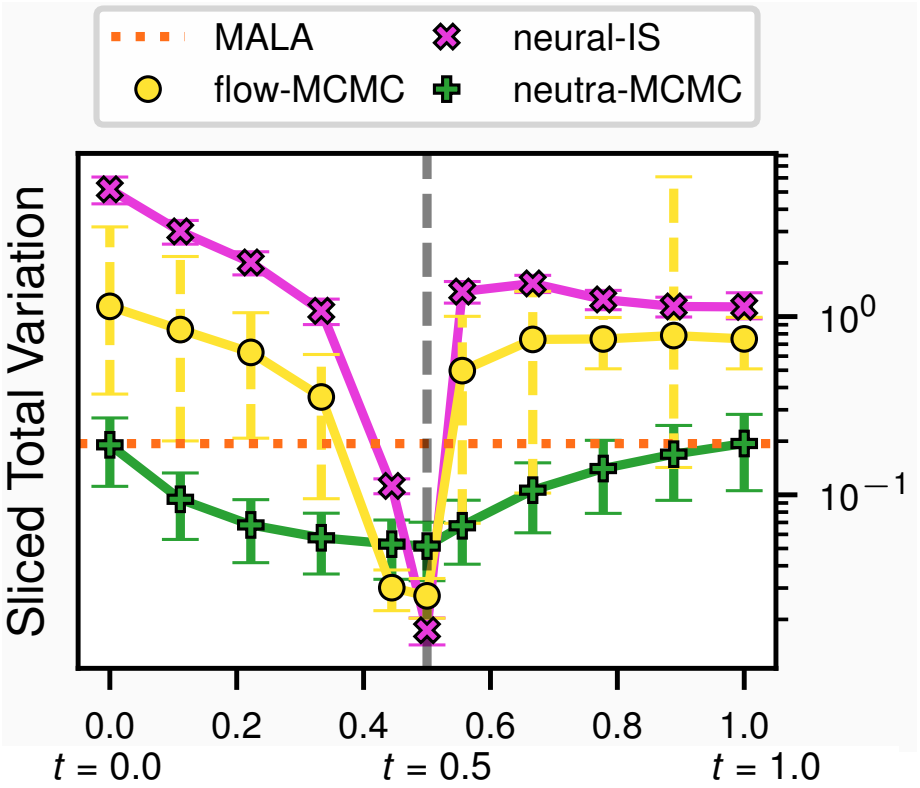
Del Debbio et al. "Efficient Modelling of Trivializing Maps for Lattice ϕ4 Theory Using Normalizing Flows [..]" *PRD* 2021
Ciarella et al. "Machine-Learning-Assisted Monte Carlo Fails at Sampling Computationally Hard Problems." *MLST 2023*
Davide Ghio's talk

On Sampling with Approximate Transport Maps
L. Grenioux, A. Durmus, E. Moulines, MG
ICML 2023

Sensitivity of ML-enhanced samplers to approximation $\rho_\theta(x) \approx \rho_*(x)$

- ▷ Ill-conditioned Gaussian target
- ▷ Imperfect flow partially matching

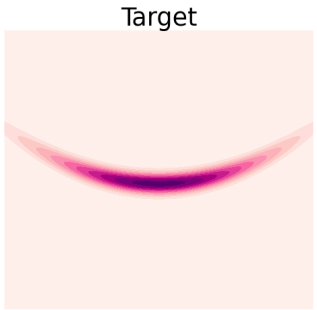




[Grenioux, Durmus, Moulines & MG, *ICML 2023*]

Another strategy for sampling: Exploit transport maps as reparametrization/pre-conditioning

T_θ



T_θ^{-1}

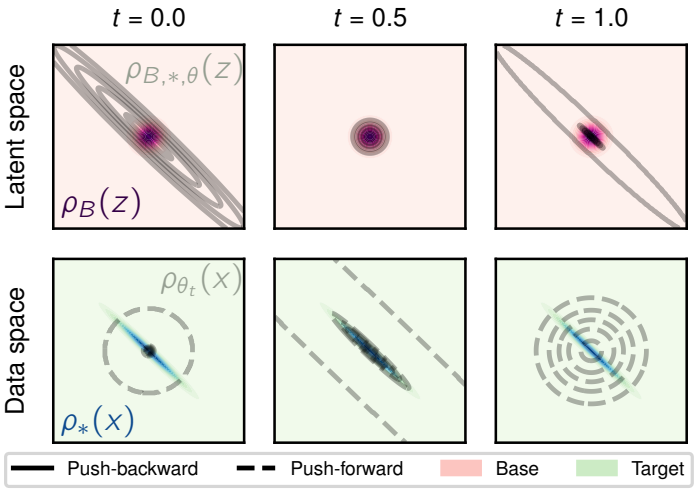
$$\rho_{B,*,\theta}(z) = \rho_{\mathrm{B}}(z) \det |\nabla_z T_\theta(z)|$$

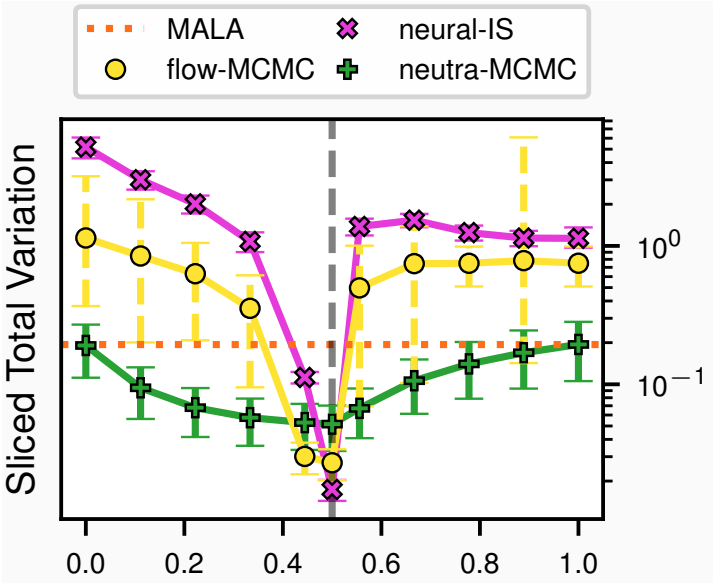
- ▷ "Neutra-MCMC": Run a local sampler in the latent space
- ▷ Intuition:
  - Local samplers are more robust in high-dimension
  - Local samplers performance on log-concave distribution is tied to conditioning

[Li & Wang *PRL 2018,* Wu et al. *PRL 2019,* Noé et al. *Science 2019,* Hoffman et al *arXiv:1903.03704]*

Preconditioned local sampling “Neutra-MCMC” is typically more robust

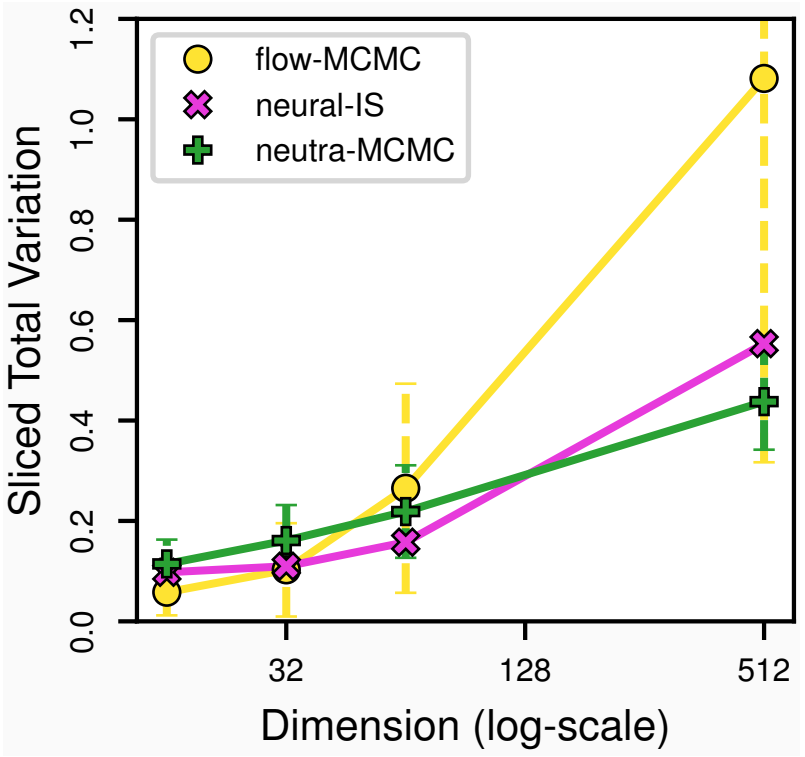
▷ With poorly trained maps





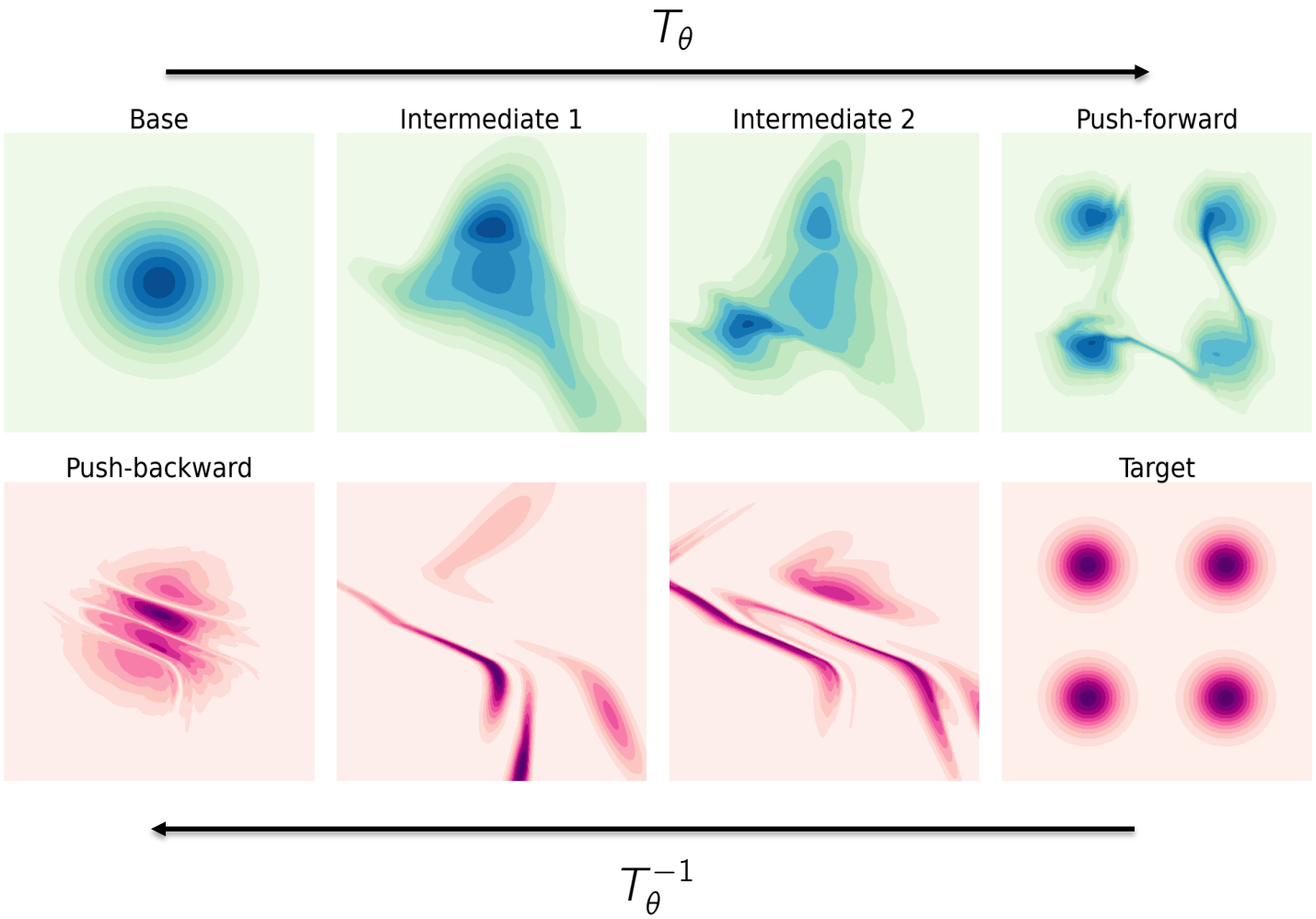
▷ With increasing dimension

“Banana distribution”



[Grenioux, Durmus, Moulines & MG, *ICML 2023*]

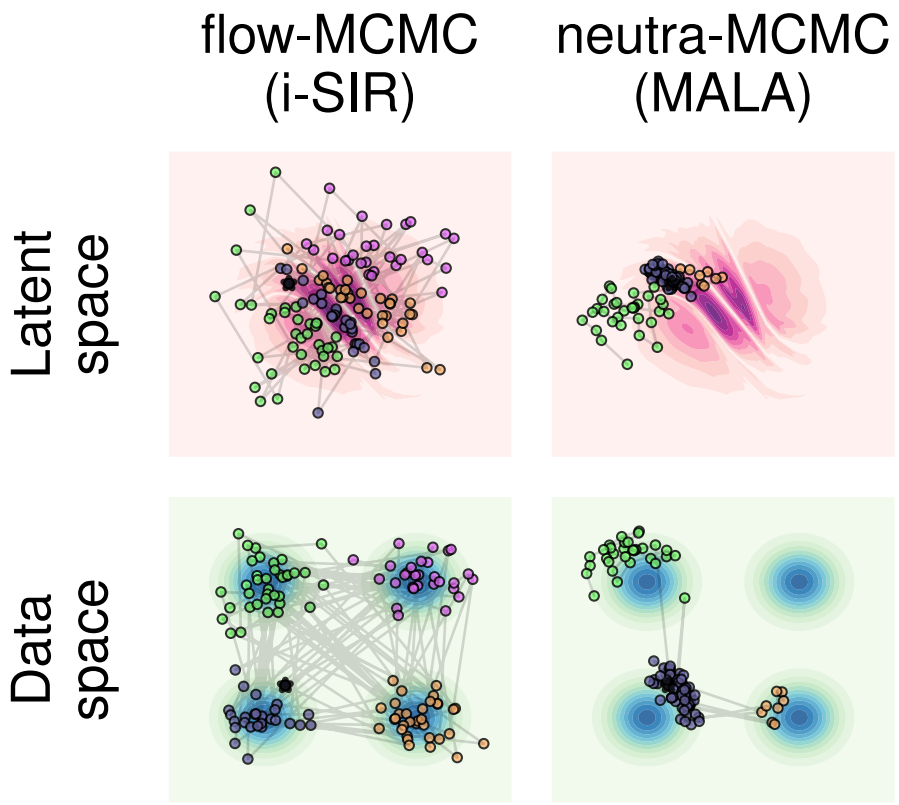
Yet learned transport does not erase mutlimodality

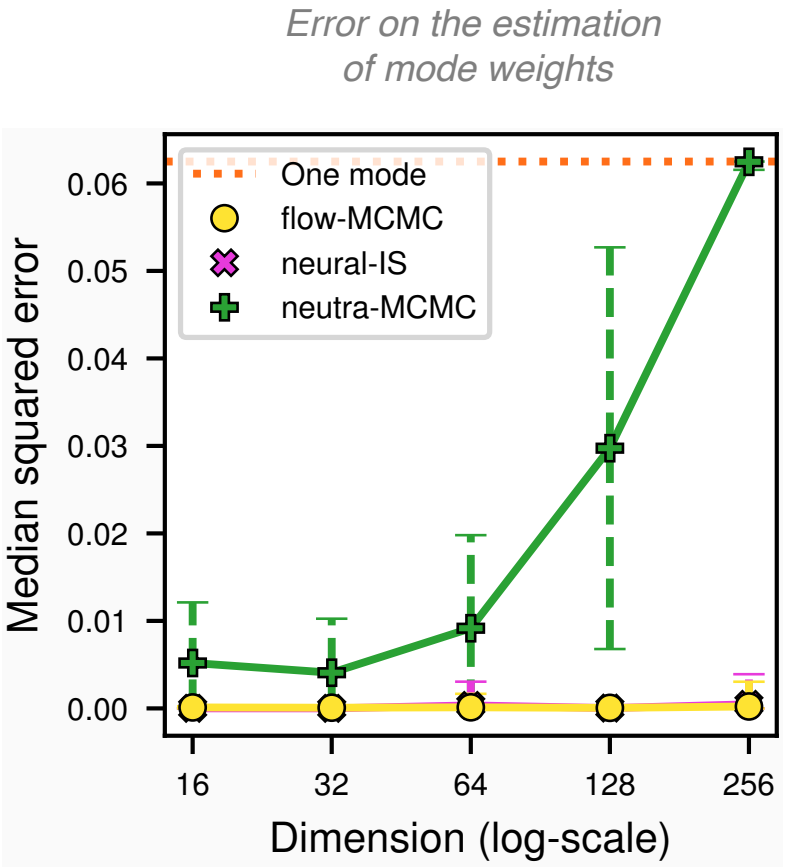


- Transporting a unimodal base to a multimodal target requires learning a very steep/almost flat transformation

Cornish et al. ICML 2020. Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows
Salmona et al. Neurips 2022. Can Push-forward Generative Models Fit Multimodal Distributions?

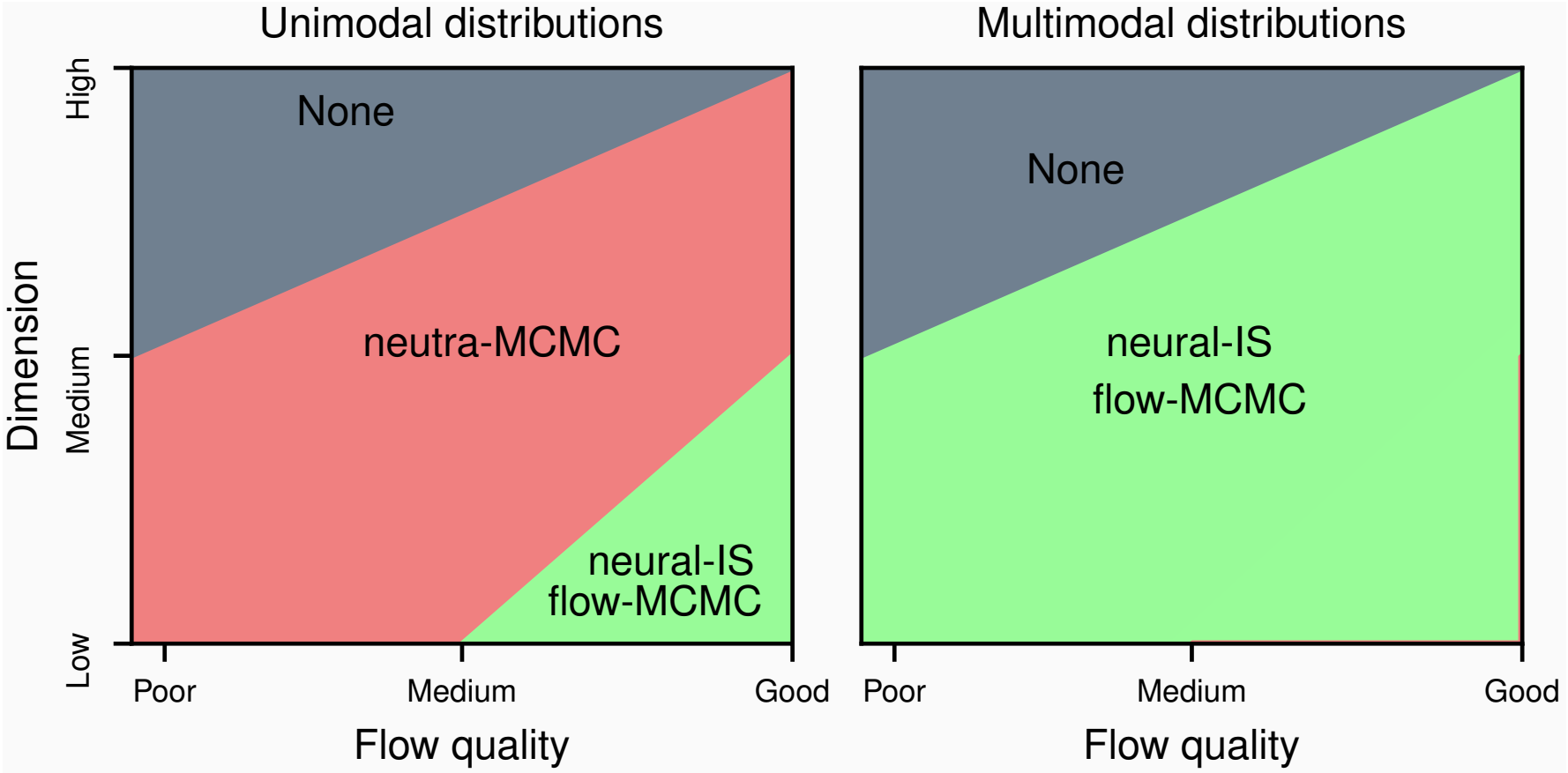
Imperfect learning hinders preconditioned MCMC mixing between modes





[Grenioux, Durmus, Moulines & MG, *ICML 2023*]

Insights from emprical study for $\rho_\theta(x) \approx \rho_*(x)$

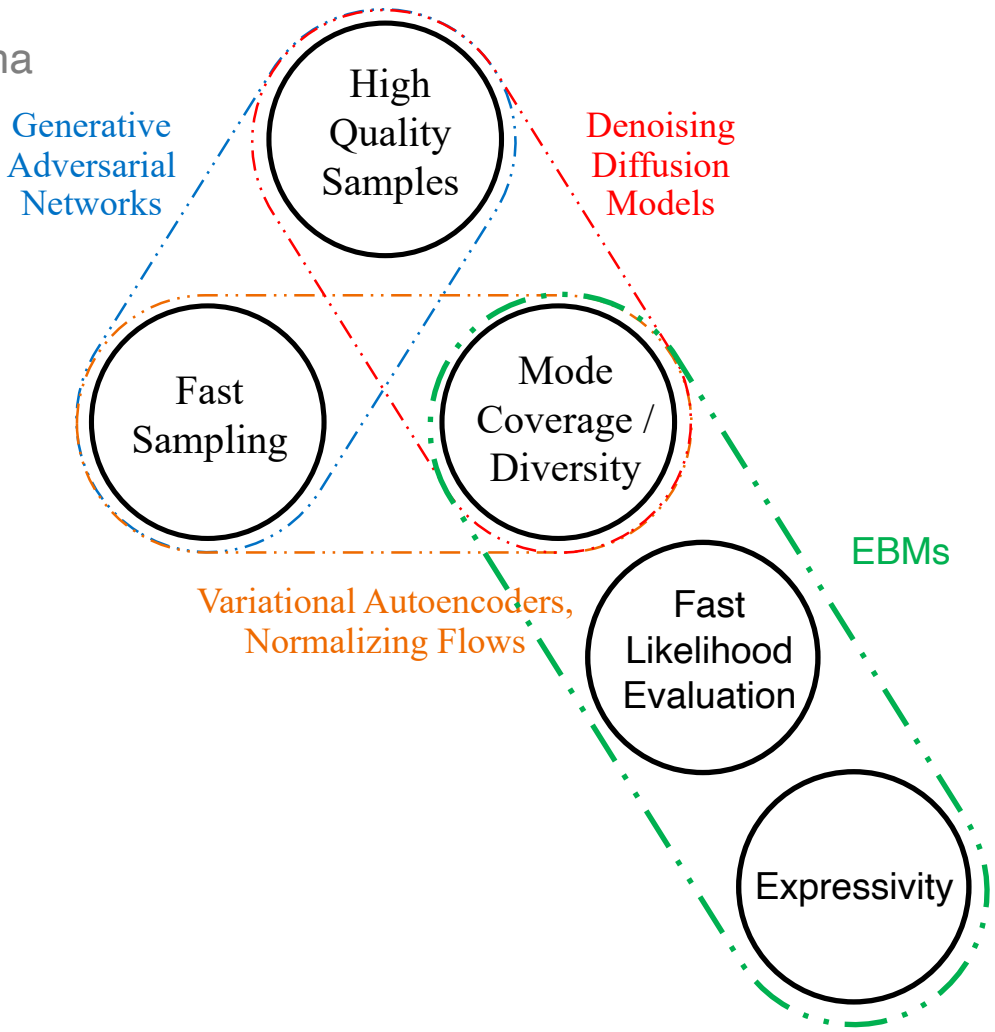


[Grenioux, Durmus, Moulines & MG, *ICML 2023*]

An application of adaptive learned samplers: Energy Based Model learning

- **Context:** The generative model trilemma (push-forward generative models)

  [Xiao et al ICLR 2022]



- Energy based models (EBM)
  - Highly flexible $\rho_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}$
  - Hard to sample (and train)

Song & Kingma. "How to Train Your Energy-Based Models." arxiv:2101.03288.

Training Energy Based Models

- Maximum likelihood gradients

$$\nabla_\theta \ell(\theta) = \mathbb{E}_{\rho_\theta}\left[\nabla_\theta E_\theta(x^-)\right] - \mathbb{E}_{\text{data}}\left[\nabla_\theta E_\theta(x^+)\right]$$

  - In practice (very) approximate sampling only
  - e.g. Contrastive Divergence: short non-converged chains

- Score based method, e.g. score-matching minimizing Fisher divergence

$$\begin{aligned} D_F(\rho_{\text{data}}||\rho_\theta) &= \mathbb{E}_{\text{data}}\left[\|\nabla_x \log \rho_{\text{data}}(x) - \nabla_x \log \rho_\theta(x)\|^2\right] \\ &= \mathbb{E}_{\text{data}}\left[\frac{1}{2}\sum_{i=1}^{d} \frac{\partial E_\theta(x)}{\partial x_i}^2 + \frac{\partial^2 E_\theta(x)}{\partial x_i^2}\right] + C \end{aligned}$$

- **Neither method is guaranteed to yield a statistically accurate model in multimodal cases**

- **Proper sampling** of the EBM during learning appears to be a gold standard for a statistically accurate EBM

Song & Kingma. “How to Train Your Energy-Based Models.” arxiv:2101.03288.

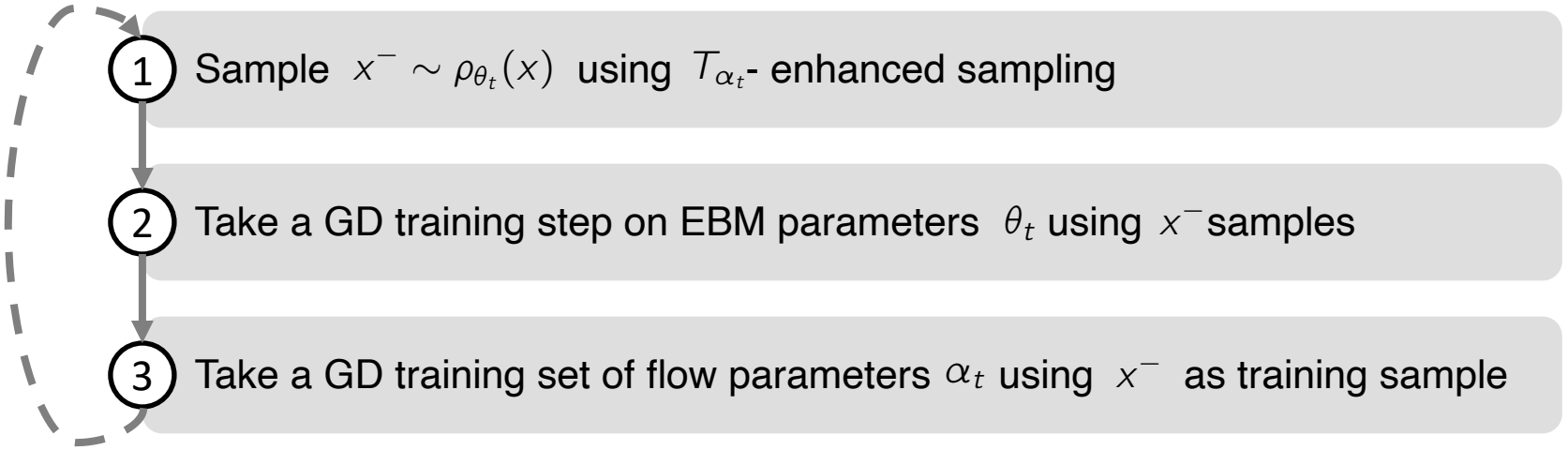
Training an EBM with a companion NF sampler

Idea:

Jointly train an EBM $\rho_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}$ & a Normalizing Flow $\rho_\alpha(x) = \rho_B(T_\alpha^{-1}(x)) \det \left|\nabla_x T_\alpha^{-1}\right|$

such that $\rho_{\theta_t}(x) \approx \rho_{\alpha_t}(x)$

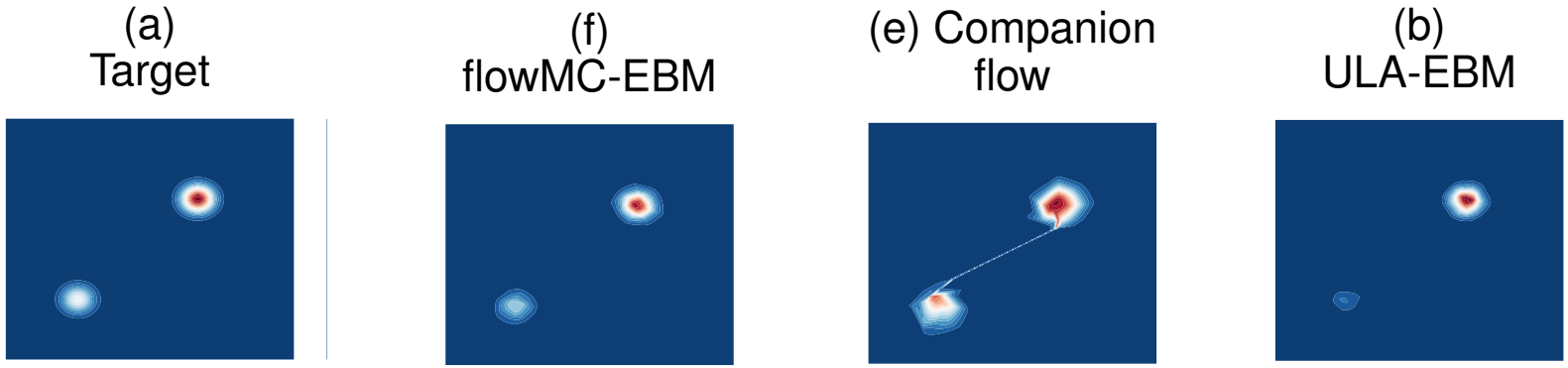
Sketch of algorithm:

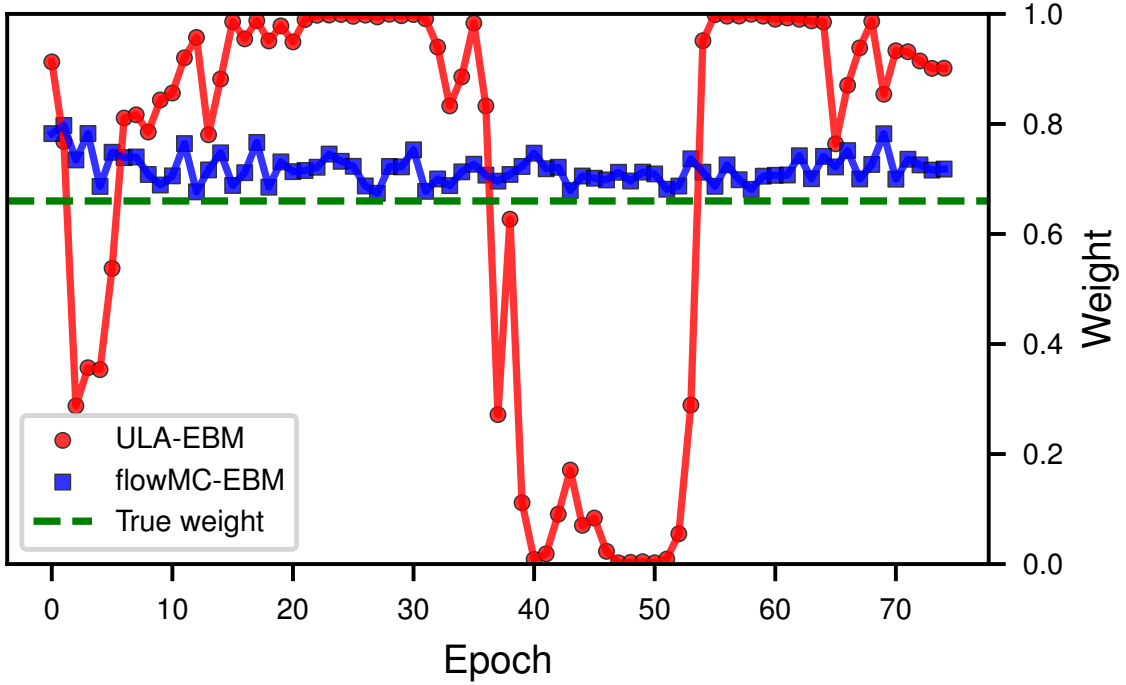


- ▷ Trade-off in training two models but get the best of some of the worlds
- ▷ Previous variants, not strictly using calibrated MCMC

[Xiao et al. arxiv:2006.08100, Nijkamp et al. ICLR 2022, “NT-EBM”, Xie et al ICLR 2022 “CoopFlow”]

Preliminary results (moderate dimensions)





[Grenioux, Moulines & MG, *ICML SPIGM Workshop 2023*]

Prespectives

- Progess in generative modelling suggests a road to machine learning enhance samplers.

  Groups in lattice QCD (P. Shananhan, MIT) and molecular dynamics (F. Noé, MSR) are actively developing/adapting these methods.

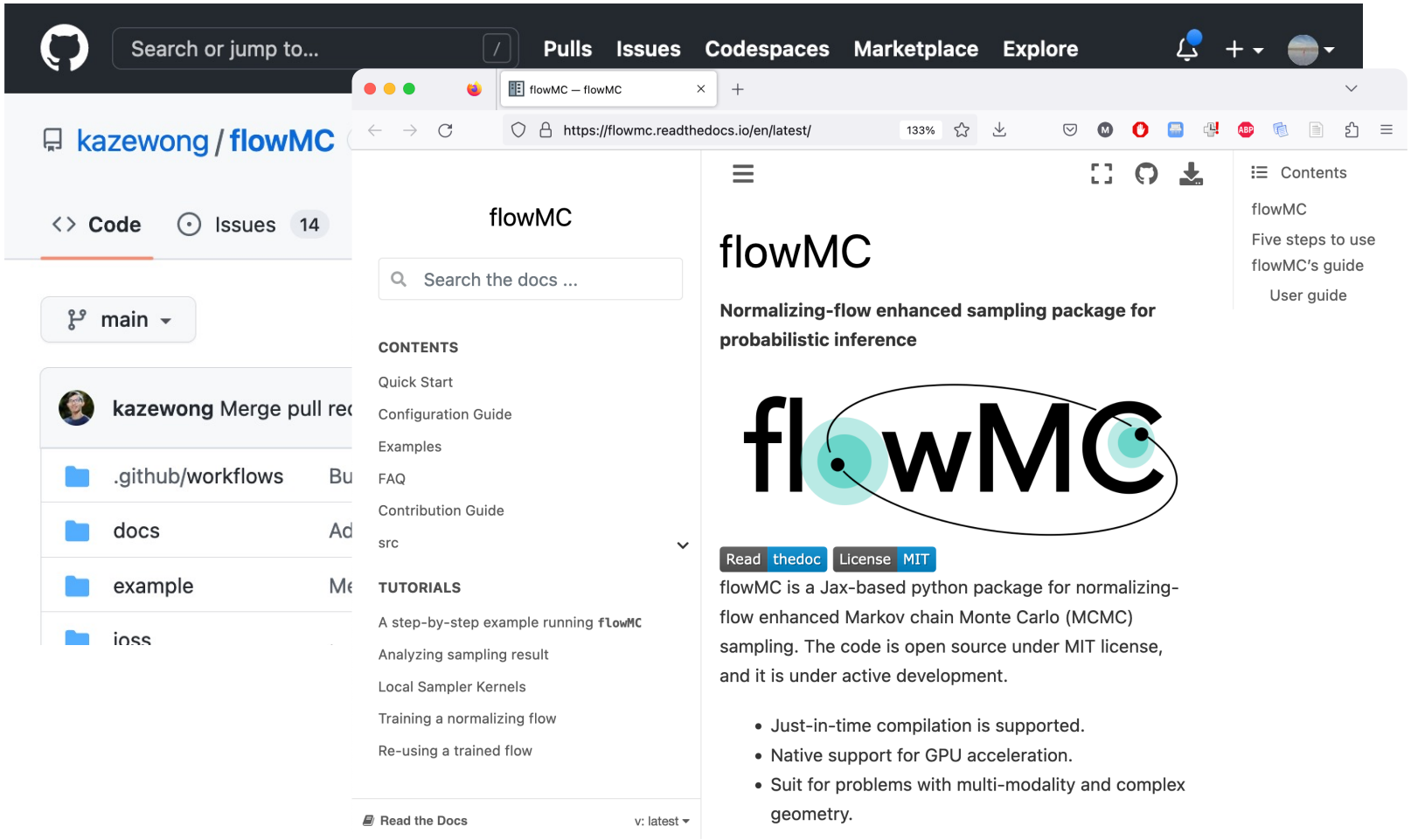
- Reaching the level of training accuracy required is not trivial, and physics-inspired architecture will most likely play an important role.

- These methods appear to be well-suited for Bayesian inference problems in moderate dimension!

- Another example of application of ML-enhanced samplers is Energy Based Models learning.

A sampling python package under-development


```
> pip install flowmc
```



▷Thank you

Collaborators:
Grant Rotskoff (Stanford), Éric Vanden-Eijnden (Courant Institute, NYU)
Éric Moulines & Louis Grenioux (École Polytechnique), Sergey Samonov (HSE University)
James Brofos & Roy Lederman (Yale University), Marcus Brubacker (York university)
Pilar Cossio (Flatiron, CCM), Olga Lopez Acevedo & Ana Molina Taborda (Universidad de Antioquia)
Kaze Wong & Dan Foreman-Mackey (Flatiron, CCA)

References

- M. Gabrié, G. M. Rotskof, and E. Vanden-Eijnden, ‘Efficient Bayesian Sampling Using Normalizing Flows to Assist Markov Chain Monte Carlo Methods’, *ICML workshop 2021*
- M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden, ‘Adaptive Monte Carlo augmented with normalizing flows’, *PNAS 2022*
- J. A. Brofos, M. Gabrié, M. A. Brubaker, and R. R. Lederman, ‘Adaptation of the Independent Metropolis-Hastings Sampler with Normalizing Flow Proposals’. *AISTAT 2022*
- S. Samsonov, E. Lagutin, M. Gabrié, A. Durmus, A. Naumov, and E. Moulines, ‘Local-Global MCMC kernels: the best of both worlds’, in *Neural Information Processing Systems*, 2022.
- K. W. K. Wong, M. Gabrié, and D. Foreman-Mackey, ‘flowMC: Normalizing-flow enhanced sampling package for probabilistic inference in Jax’. Accepted at *Journal of Open Science Software 2023*
- L. Grenioux, A. Durmus, É. Moulines, and M. Gabrié, ‘On Sampling with Approximate Transport Maps’. ICML 2023
- L. Grenioux, É. Moulines, and M. Gabrié, ‘Balanced Training of Energy Based Models’, SPIGM Workshop, ICML 2023