Balanced Training of Energy-Based Models with Adaptive Flow Sampling

Louis Grenioux, Éric Moulines, Marylou Gabrié

CMAP, École polytechnique, France

Energy-Based Models

Energy-Based Models (EBM) are flexible and powerful **density estimation tools**. EBM models a target distribution using the Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta}\exp(-E_\theta(x)),$$

where E_θ is a neural network with weights θ. The parameters θ can be estimated using maximum likelihood

$$\mathcal{L}(\theta) = -\mathbb{E}_{p^\star}[\log p_\theta(X)],$$

where $p^\star$ is the data distribution. The gradient of $\mathcal{L}$ can be expressed as the difference of two expectations

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{p^\star}[\nabla_\theta E_\theta(X)] - \mathbb{E}_{p_\theta}[\nabla_\theta E_\theta(X)].$$

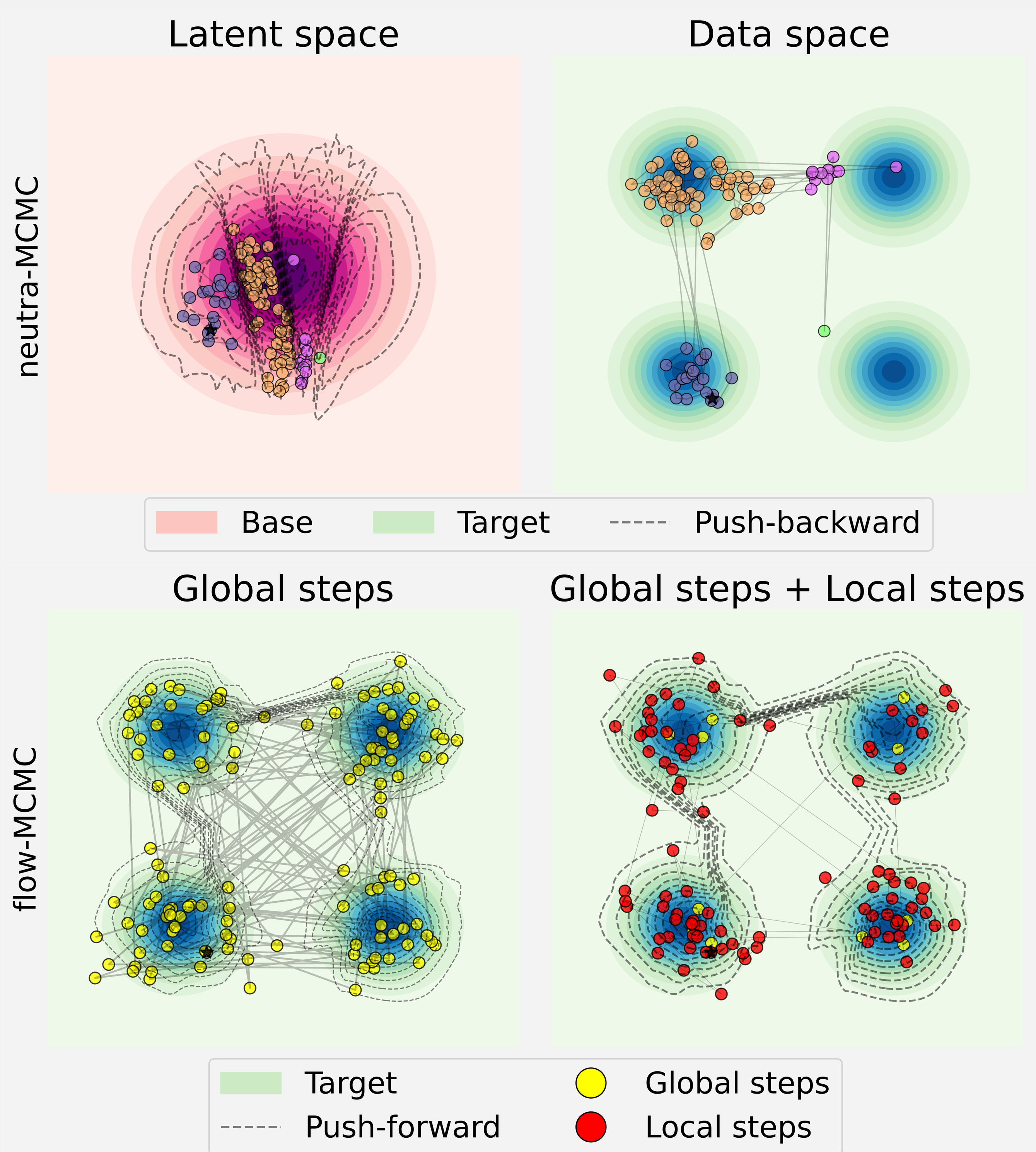
⚠ Sampling p_θ can be **high-dimensional** and very **multimodal** - in which case **sampling is hard**.

Our contribution (arXiv:2306.00684)

Sampling p_θ using **flow-MCMC** [Gabrié et al., 2022] which uses a companion **Normalizing Flows (NF)** as a proposal in a MCMC algorithm.

- In [Nijkamp et al., 2022], authors developed **NT-EBM** which leverages neutra-MCMC to sample p_θ with a pre-trained flow;
- In [Xie et al., 2022], authors developed **CoopFlow** where they use a flow to reset the chains of local MCMC which is closer to flow-MCMC.

Using NF in MCMC



- **Multimodality** flow-MCMC algorithms are able to mix between modes while neutra-MCMC algorithms get stuck in the latent space. This is because NFs can't erase energy barriers in the latent space.
- **Dimensionality** flow-MCMC doesn't scale in high-dimension and require an increasingly good flow.

More details can be found in our recent paper **arXiv/2302.04763** [Grenioux et al., 2023a].

Step of joint training of an EBM p_θ and a flow T_α

θ_k, α_k

Sample p_{θ_k} using `FlowMCMC` leveraging T_{α_k} producing $\{y_i\}_i^{m^-}$

$\{x_i\}_i^{m^+} \sim p^\star$

Update θ_k using gradient descent

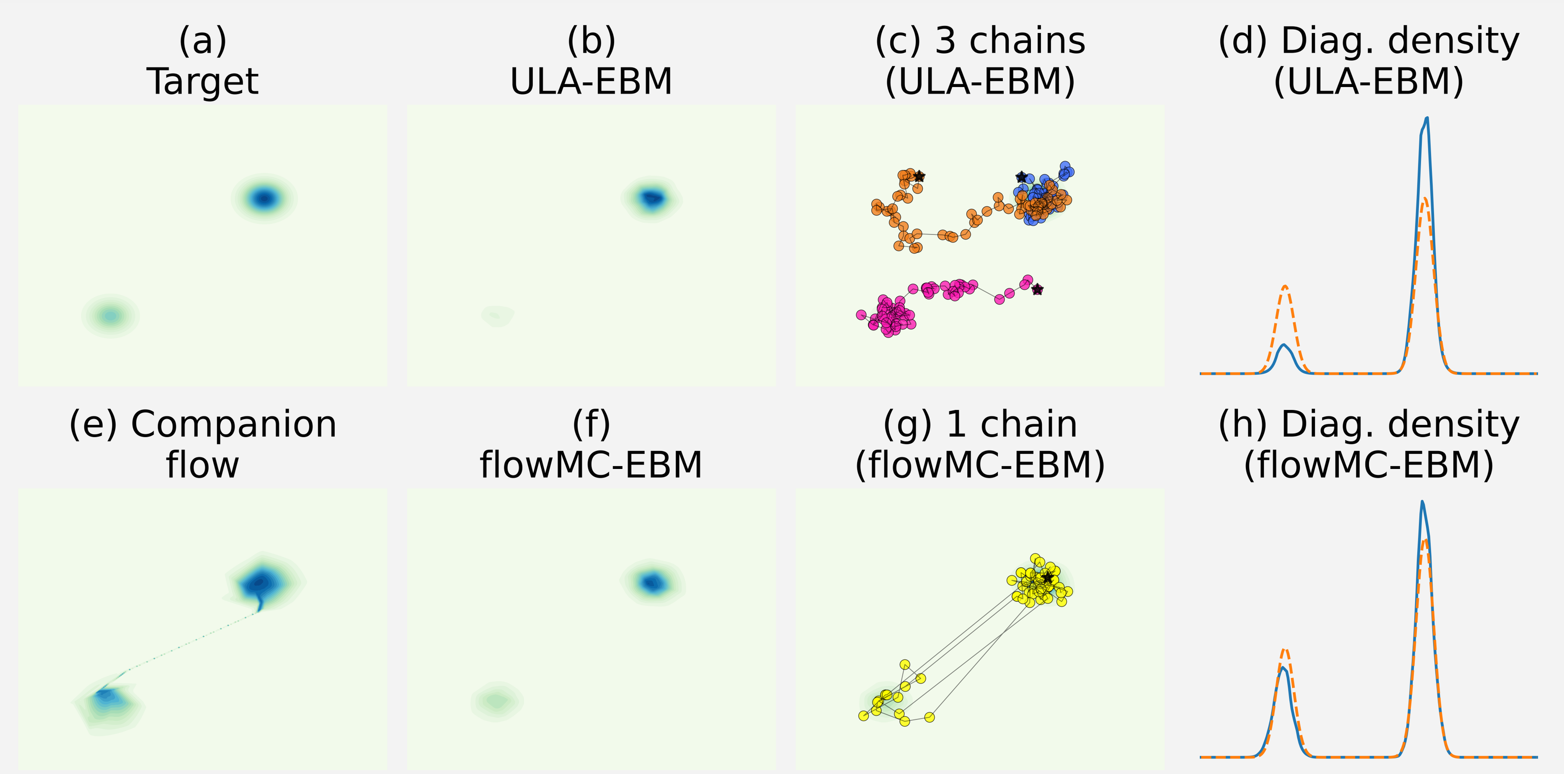
$$\theta_{k+1} \leftarrow \theta_k - \gamma\left(\frac{1}{m^+}\sum_{j=1}^{m^+}\nabla_\theta E_{\theta_k}(x_j) - \frac{1}{m^-}\sum_{i=1}^{m^-}\nabla_\theta E_{\theta_k}(y_i)\right)$$

Update α_k using gradient ascent on the likelihood

$$\alpha_{k+1} \leftarrow \alpha_k + \lambda\frac{1}{m^-}\sum_{i=1}^{m^-}\nabla_\alpha \log q_{\alpha_k}(y_i)$$

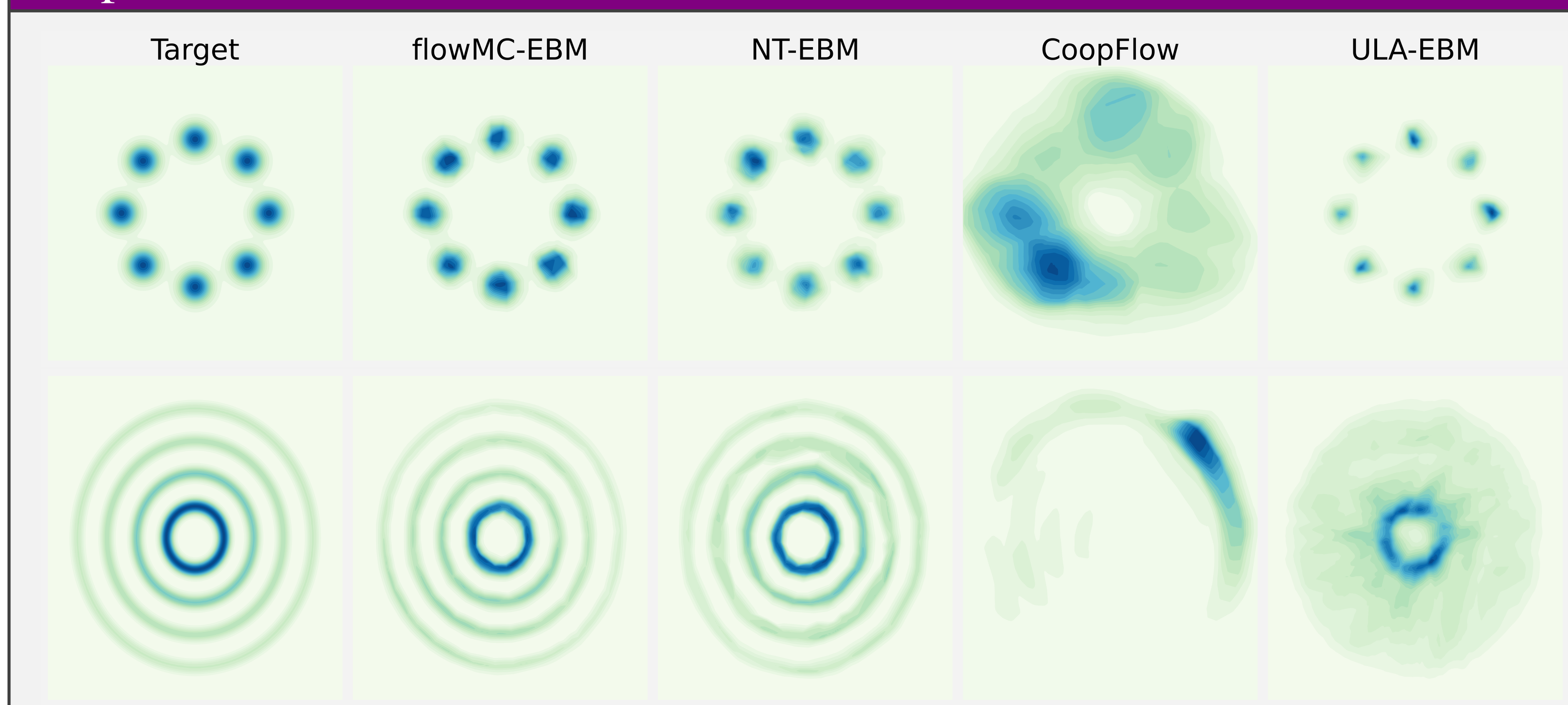
θ_{k+1}
α_{k+1}

Motivating example



The EBM trained with *ULA* in *(b)* doesn't approximate well the target *(a)* because the MCMC chains *(c)* don't mix between modes. Using flow-MCMC *(g)* with its companion flow *(e)* provides a good approximation *(f)*.

Experiments in 2D



References

[Gabrié et al., 2022] Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2022). Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119.

[Grenioux et al., 2023a] Grenioux, L., Durmus, A., Moulines, É., and Gabrié, M. (2023a). On sampling with approximate transport maps. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

[Grenioux et al., 2023b] Grenioux, L., Éric Moulines, and Gabrié, M. (2023b). Balanced training of energy-based models with adaptive flow sampling.

[Nijkamp et al., 2022] Nijkamp, E., Gao, R., Sountsov, P., Vasudevan, S., Pang, B., Zhu, S.-C., and Wu, Y. N. (2022). MCMC should mix: Learning energy-based model with neural transport latent space MCMC. In *International Conference on Learning Representations*.

[Xie et al., 2022] Xie, J., Zhu, Y., Li, J., and Li, P. (2022). A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In *International Conference on Learning Representations*.