OT Flow Matching for Fast CaloSim & ML Inference Benchmarking

Paul Wollenhaupt
CERN Summer School 2024

June 20^{th}, 2024

Who am I?

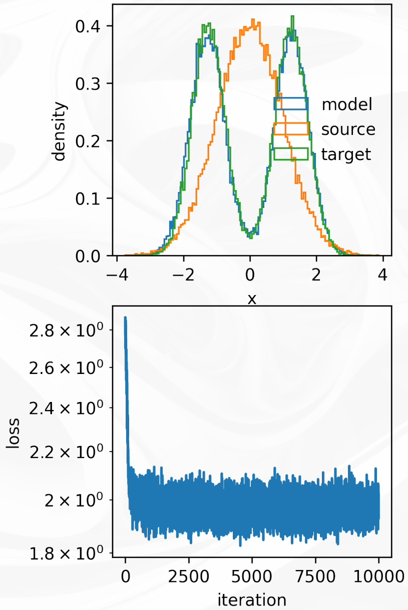
- Paul Anton Maximilian Wollenhaupt
- Mathematics Master in Göttingen
- Statistics/theoretical ML research
- Research at Quadt's, Ecker's & Gipp's Group
- Did lots of STEM competitions, now CP
- Diffusion model projects since 2019
- *Extrapolating Data-MC Disagreements using OT & Normalising Flows in ATLAS*

Inference Benchmark

- Benchmark inference speed in Python
- PyTorch, ONNX runtime, Keras (Tensorflow) and SOFIE
- FastSim VAE Decoder (MLP) on single CPU core
  - ▷ ONNX and Keras fastest SOFIE depends on batch size
- Options to set number of cores seem dubious for ONNX and SOFIE
- Benchmarked memory using memray tracing in native mode
  - ▷ ONNX and SOFIE are fine, PyTorch and TF use weirdly much memory

OT Flow Matching

- OT Flow Matching is the goal
- Computing exact OT is a bottleneck
- There is a great OT library in JAX
  - ▷ Ported I-CFM from `torchcfm` to JAX
  - ▷ Tested on toy data w/ 1d OT



Backup Slides

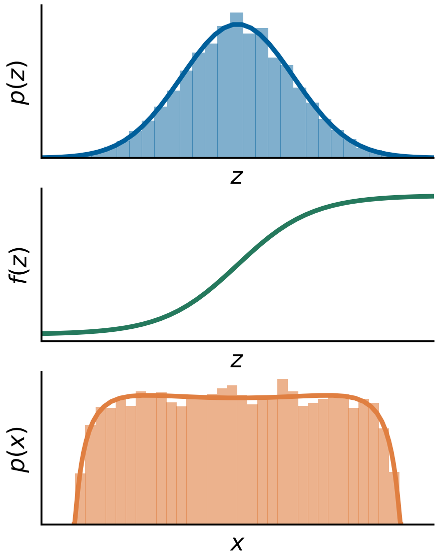
Normalising Flows

- Start with known distribution $z \sim p_z$
- Apply diffeomorphism f_θ to z

$$p_\theta(x) = p_z(f_\theta^{-1}(x)) \cdot \left| \det \frac{\partial f_\theta^{-1}(x)}{\partial x} \right|$$

- Maximize likelihood of data

$$\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(x_i)$$



Rezende and Mohamed 2016

Continuous Normalizing Flows

- Define the transformation as an ODE

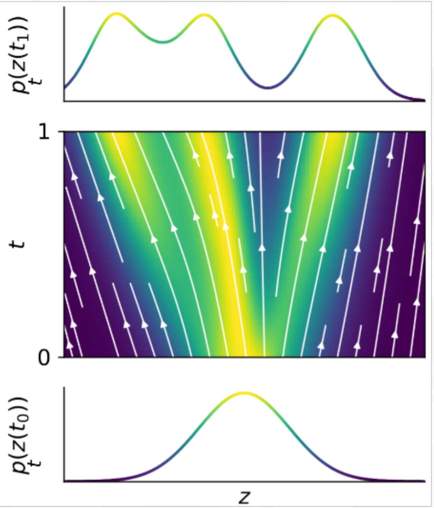
$$x = z(t_1) = \int_{t_0}^{t_1} v_\theta(z(t), t)\,\mathrm{d}t$$

- Instantaneous change of density

$$\frac{\partial \log p_t(z(t))}{\partial t} = -\nabla \cdot v_\theta(z(t), t)$$

- Solve the ODE for $\log p_t(z(t_1))$

$$\log p_t(z(t_0)) - \int_{t_0}^{t_1} \nabla \cdot v_\theta(z(t), t)\,\mathrm{d}t$$



Hutchinson 1990; Grathwohl et al. 2018; Chen et al. 2019

Noise Levels

- Anealed Langevin dynamics levels

$$0 < \sigma_0 < \sigma_1 < \ldots < \sigma_T$$

- Continuous limit $\sigma(t) : [0, 1] \to \mathbb{R}_+$

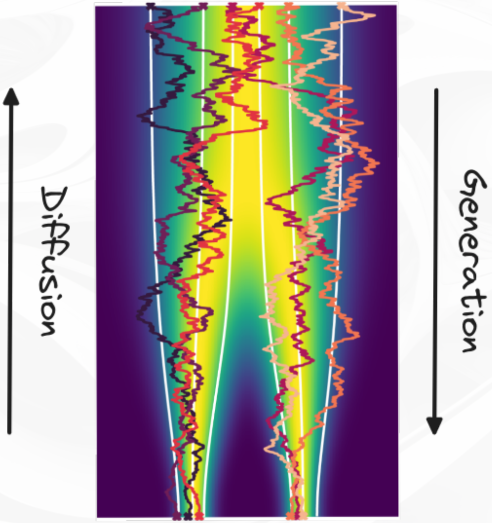
$$\mathrm{d}x = -\sigma(t)^2 \nabla_x \log p_t(x)\, \mathrm{d}t + \sigma(t)\, \mathrm{d}\bar{\omega}$$

- Reverses the *Diffusion* SDE

$$\mathrm{d}x = \sigma(t)\, \mathrm{d}\omega$$

- ODE with same marginal distributions

$$\mathrm{d}x = -\frac{\sigma(t)^2}{2} \nabla_x \log p_t(x)\, \mathrm{d}t$$



Anderson 1982; Song et al. 2021

Flow Matching

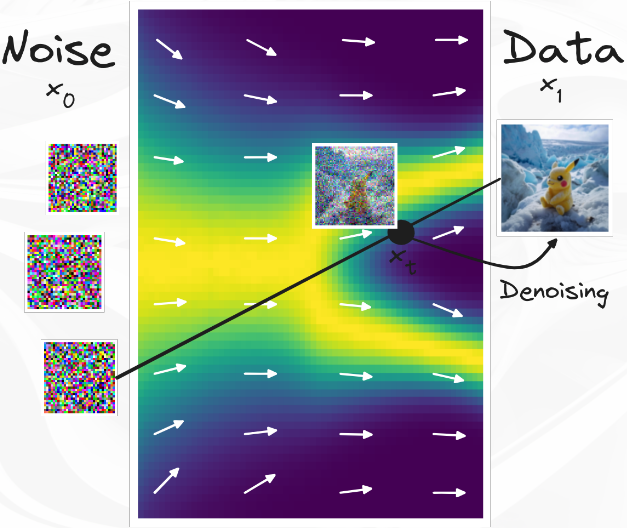
- Sample noise x_0, data x_1
- Interpolate with $t \in [0, 1]$

$$x_t = tx_1 + (1 - t)x_0$$

- Model the denoising direction

$$\mathbb{E}_{x_t, t}\left[x_1 \mid x_t, t\right]$$

- Defines a velocity field v_θ
- v_θ is a sound CNF

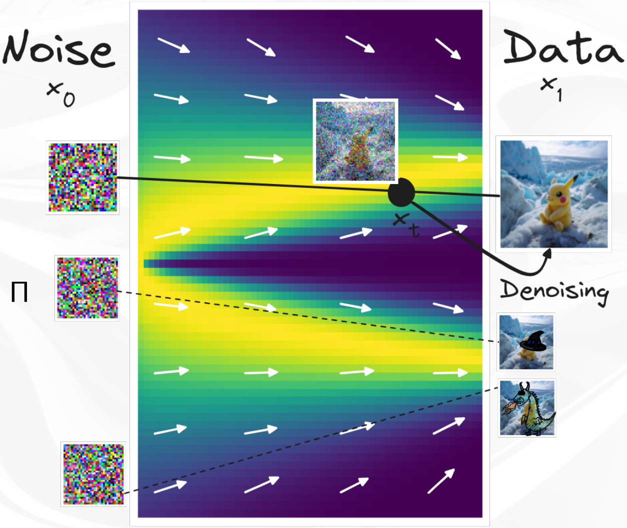


Lipman et al. 2023

Mini Batch OT Flow Matching

- Batch sample $\left\{x_0^{(i)}, x_1^{(i)}\right\}_{i=1}^{n}$
- Compute OT assignments Π
- Construct geodesic points $x_t^{(i)}$

  $x_t = tx_1^{(j)} + (1-t)x_0^{(i)},\ (x_0^{(i)}, x_1^{(j)}) \in \Pi$
- Learn denoising direction
- ODE paths become straight lines, as $n \to \infty$



Tong et al. 2024

References I

Anderson, **Brian D.O.** (1982). "Reverse-time diffusion equation models". In: *Stochastic Processes and their Applications* 12.3, pp. 313–326. ISSN: 0304-4149.

Chen, **Ricky T. Q.** et al. (2019). *Neural Ordinary Differential Equations*. arXiv: 1806.07366 [cs.LG].

Grathwohl, **Will** et al. (2018). *FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models*. arXiv: 1810.01367 [cs.LG].

Hutchinson, **M.F.** (1990). "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines". In: *Communications in Statistics - Simulation and Computation* 19.2, pp. 433–450.

Lipman, **Yaron** et al. (2023). *Flow Matching for Generative Modeling*. arXiv: 2210.02747 [cs.LG].

Rezende, **Danilo Jimenez** and **Shakir** Mohamed (2016). *Variational Inference with Normalizing Flows*. arXiv: 1505.05770 [stat.ML].

Song, **Yang** et al. (2021). *Score-Based Generative Modeling through Stochastic Differential Equations*. arXiv: 2011.13456 [cs.LG].

References II

Tong, **Alexander** et al. (2024). *Improving and generalizing flow-based generative models with minibatch optimal transport*. arXiv: 2302.00482 [cs.LG].