Reducing Memorisation in Generative Models via Riemannian Bayesian Inference

Johanna Marie Gegenfurtner*, Albert Kjøller Jacobsen* and Georgios Arvanitidis
Section for Cognitive Systems - Technical University of Denmark
*Equal contribution, listed in arbitrary order

DTU

Motivation

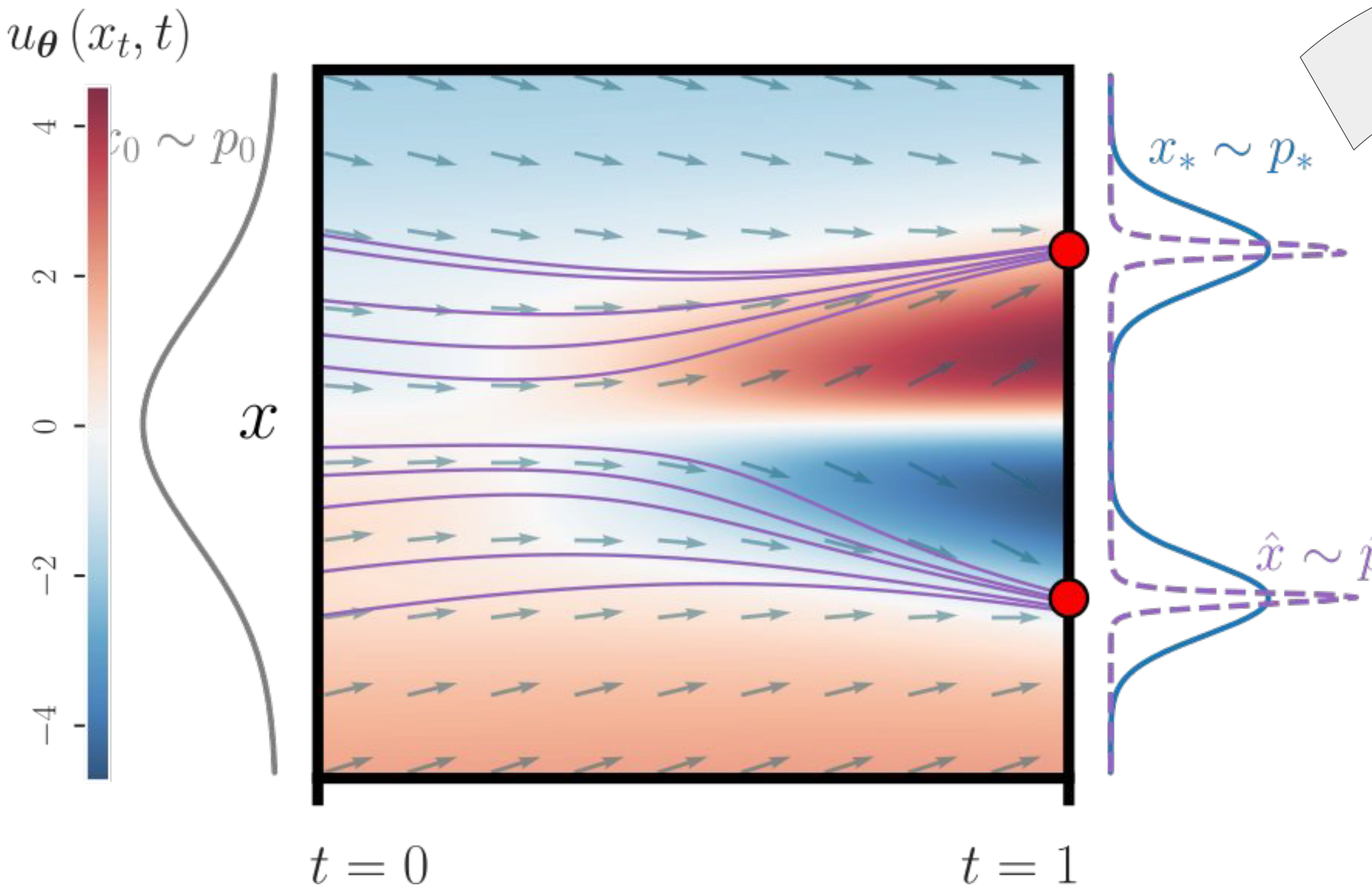
A generative model should **capture the data distribution without memorising specific data samples.**

A **key challenge** is to limit memorisation while preserving the model's ability to generate meaningful samples.

RQ: Can we **reduce memorisation** in generative models **through uncertainty** on the parameters?

TL;DR: Yes! By using a **geometry-informed approximate posterior** distribution over model parameters.

Flow matching & memorisation



A learnt **generative model** maps samples from a **known distribution** to **new samples** that approximately come from the **true data distribution**.

$$\hat{\boldsymbol{x}} = g_{\boldsymbol{\theta}}(\boldsymbol{x}_0), \qquad \boldsymbol{x}_0 \sim p_0 .$$

In **flow matching** [1], the generator's output is the solution to an IVP evaluated at time t = 1:

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \qquad \dot{\boldsymbol{x}}(t) = u_\theta(\boldsymbol{x}(t), t).$$

The flow matching loss is given by

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t \sim \mathcal{U}_{[0,1]}, \boldsymbol{x}_* \sim p_*, \boldsymbol{x}_0 \sim p_0} \left[\| u_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - (\boldsymbol{x}_* - \boldsymbol{x}_0) \|_2^2 \right] .$$

A generated sample is **memorised** [3,4] if it is much closer to one particular training sample than the rest:

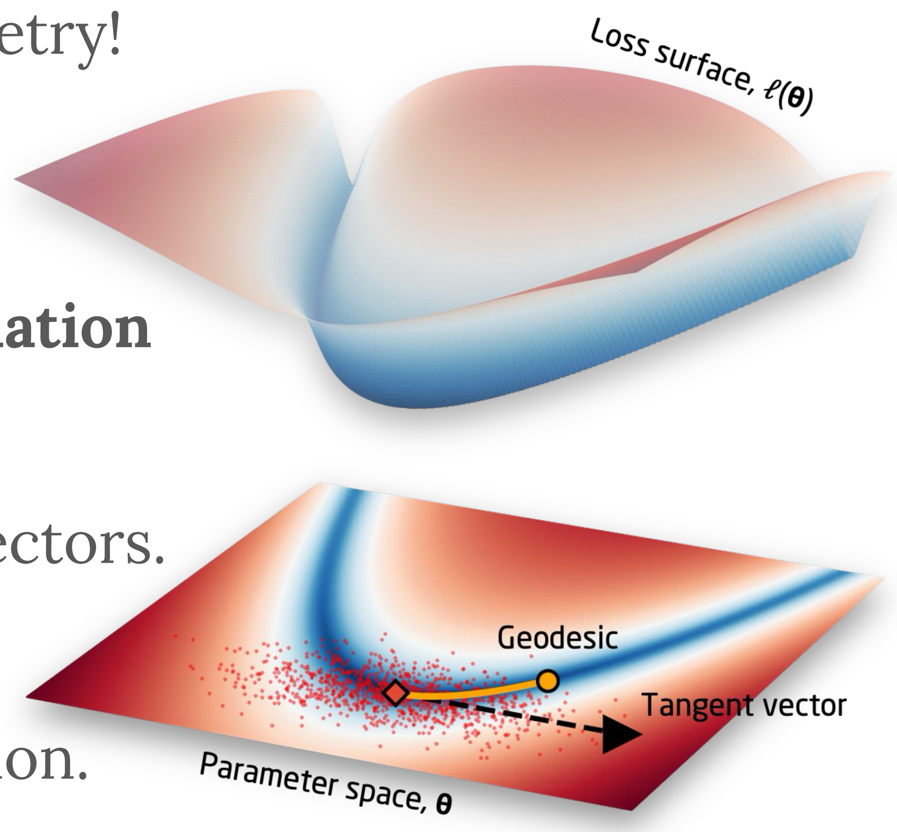
$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^{(1)}(\hat{\boldsymbol{x}})\|^2 \leq c \|\hat{\boldsymbol{x}} - \boldsymbol{x}^{(2)}(\hat{\boldsymbol{x}})\|^2.$$

closest and second closest training samples

What is Riemannian Bayesian inference?

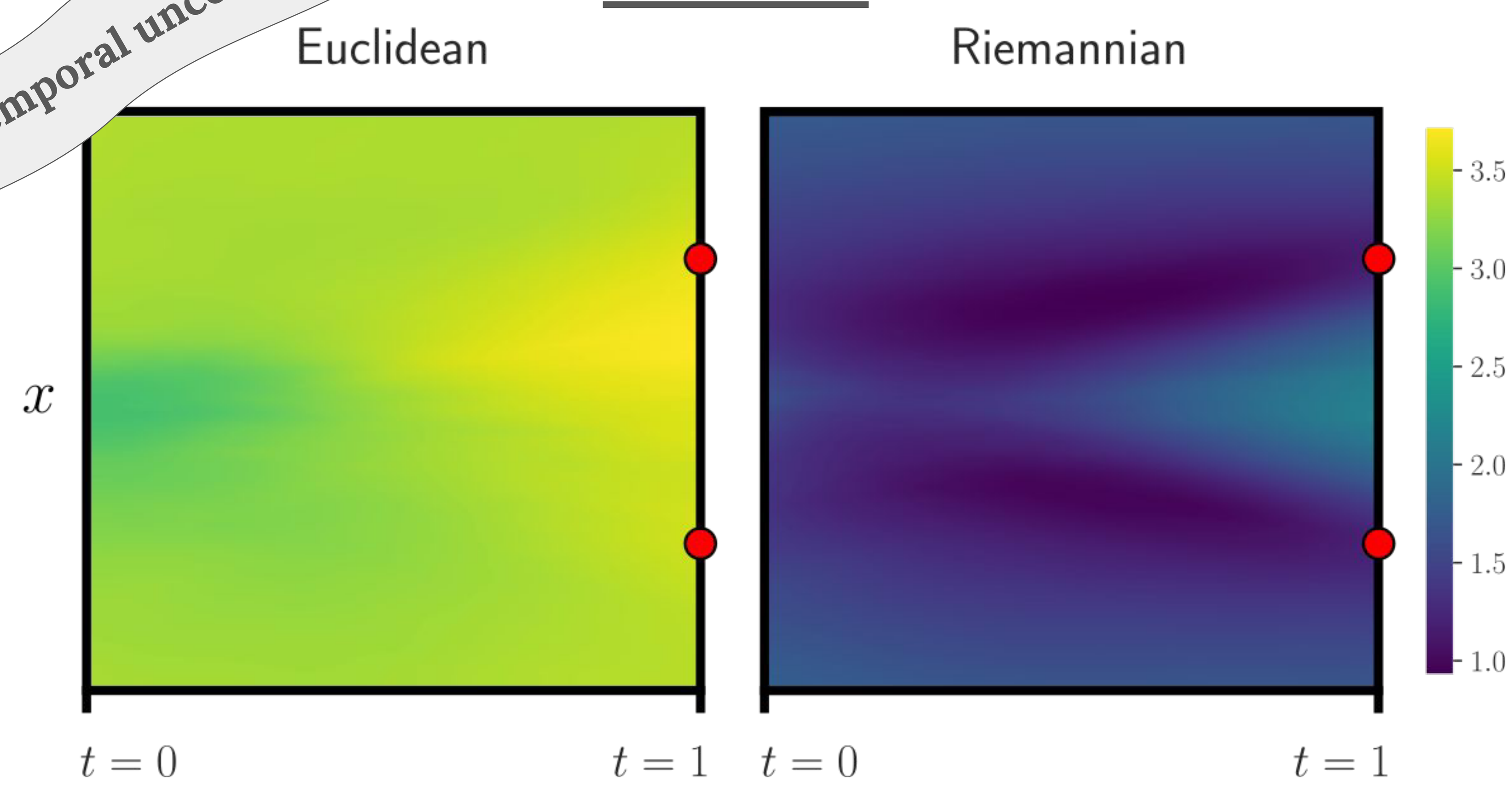
A **flexible** approximate posterior distribution [5] that adapts to the loss geometry!

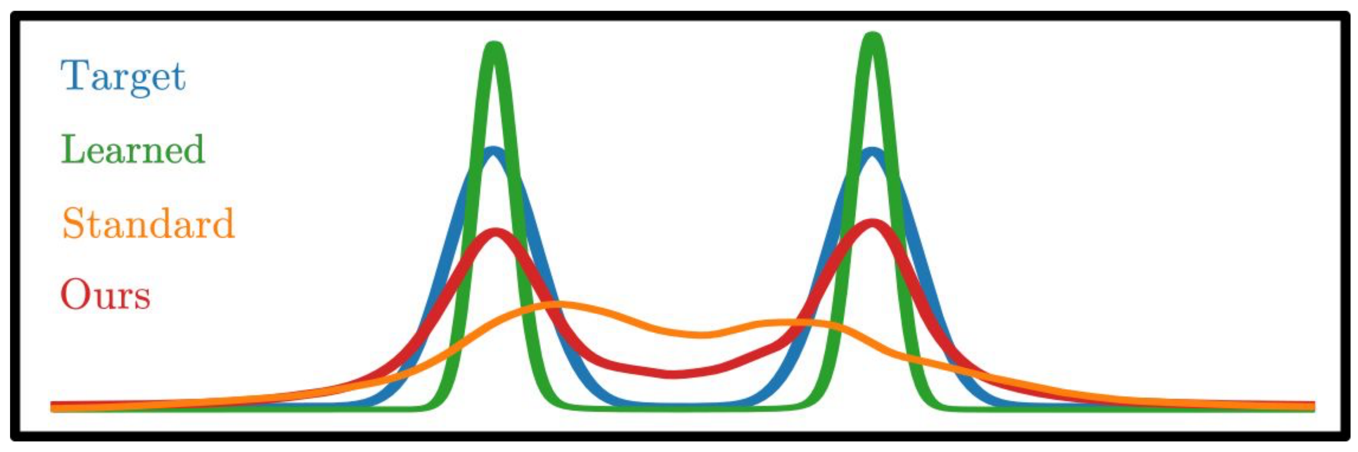
1. Find optimum with **SGD**.
2. Define **Laplace approximation** in the tangent plane.
3. Sample **initial velocity** vectors.
4. Compute **geodesics** using these initial condition.

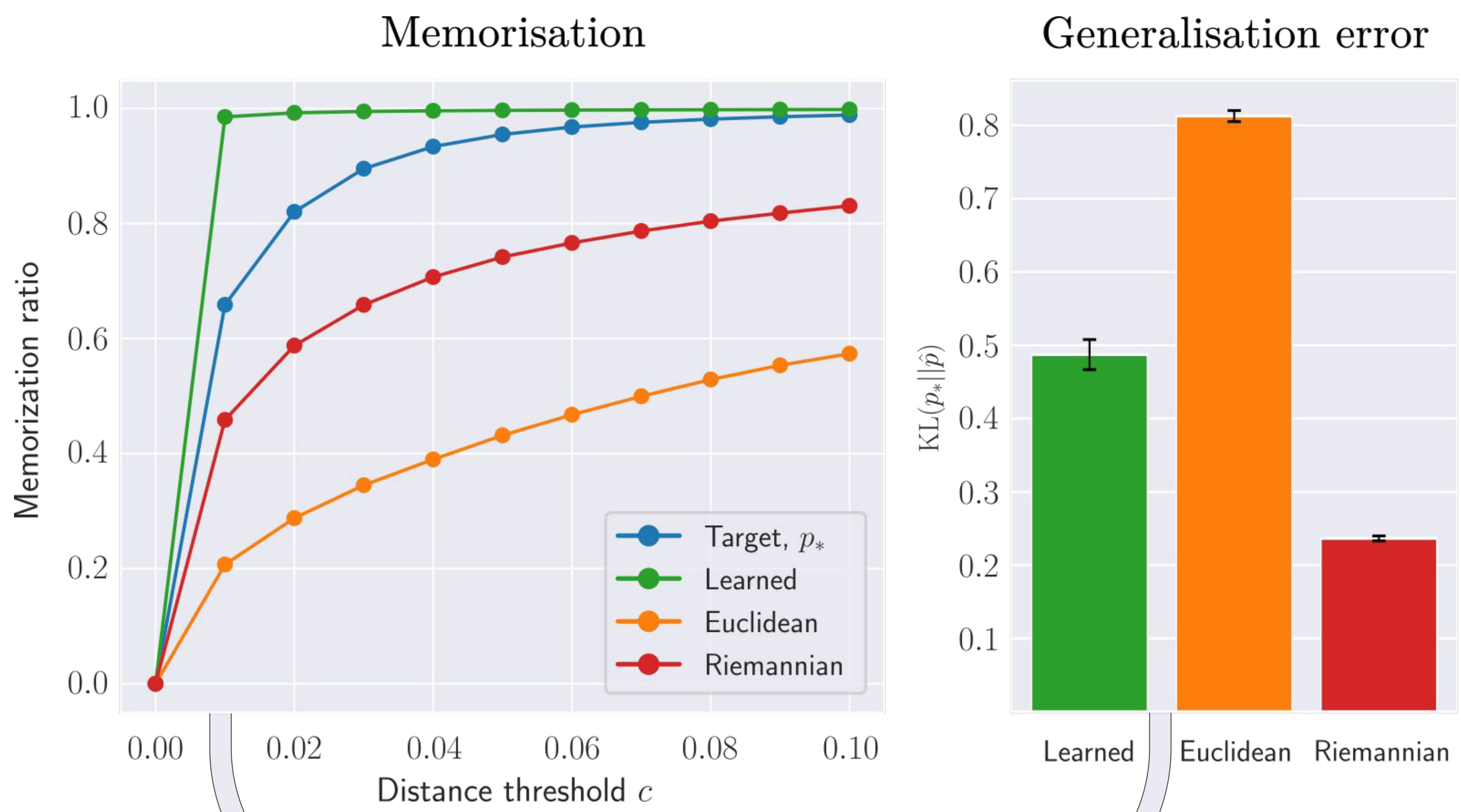


Spatio-temporal uncertainty

Results







It's a trade off!

1. Lipman et al. "*Flow Matching for Generative Modeling*", arXiv preprint 2022.
2. Buchanan et al, "*On the edge of memorization in diffusion models*". NeurIPS 2025
3. Yoon et al. "*Diffusion probabilistic models generalize when they fail to memorize*". SPIGM Workshop 2023 @ ICML
4. Bergamin et al. "*Riemannian Laplace Approximations for Bayesian Neural Networks*", NeurIPS 2023

INDEPENDENT RESEARCH FUND DENMARK
Danish Data Science Academy