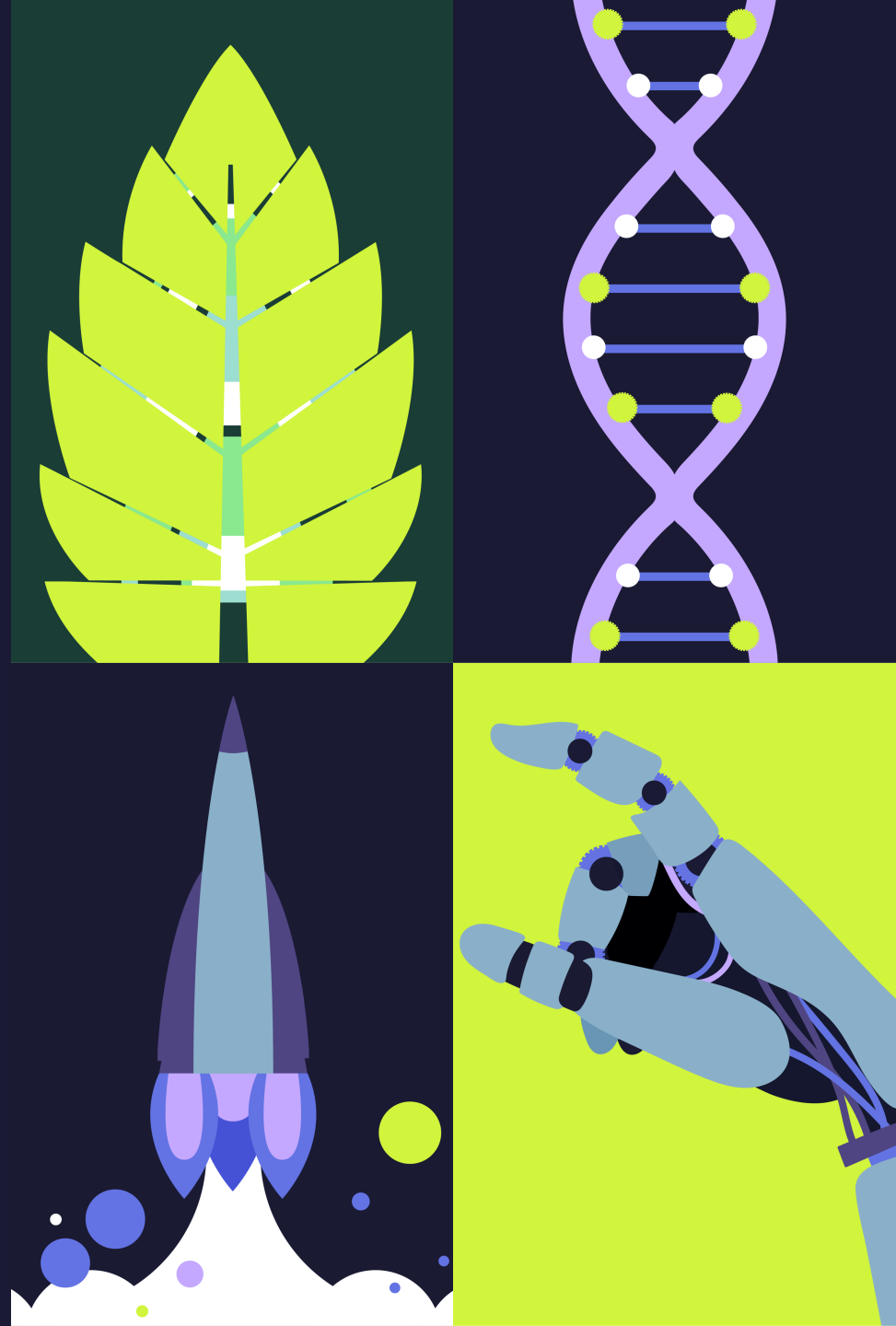
MARS

Conditioning Gaussian Processes on (Almost) Anything

Chris Nemeth
Lancaster University

with Henry Moss, Lachlan Astfalck, Thomas Cowperthwaite, Colin Doumont, Sam Willis, Philipp Hennig and Andrew Zammit-Mangion



UKRI Research England

School of Mathematical Sciences | Lancaster University

MOTIVATION

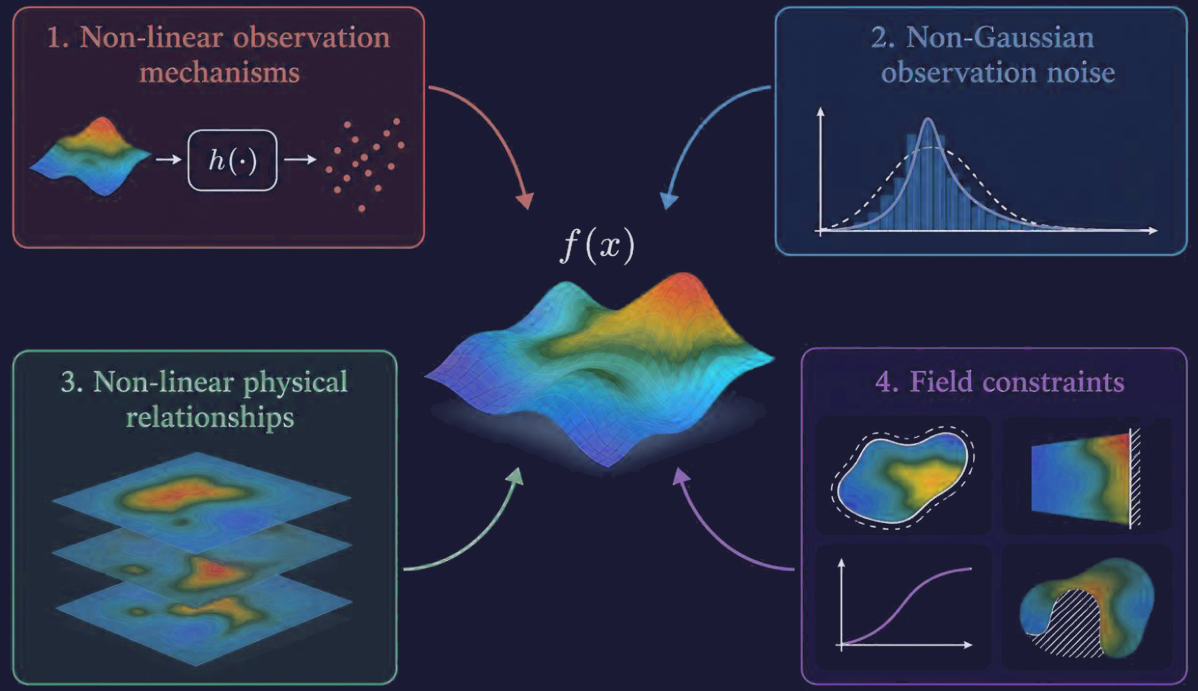
Why GP conditioning?

GPs condition in closed form when observations are linear and Gaussian:

$$y = f_0 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Gamma)$$

Most of what we care about isn't:

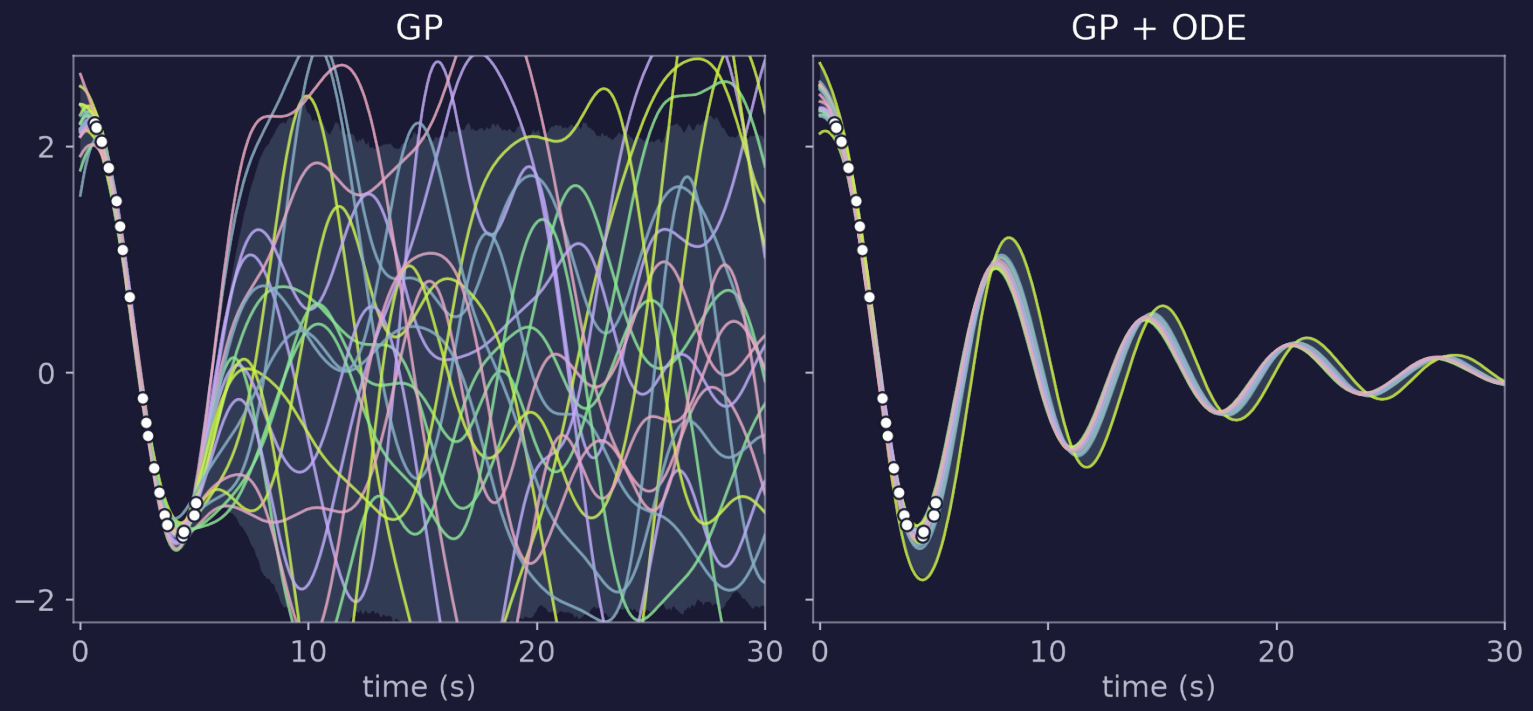
- non-linear observation mechanisms
- non-Gaussian noise: counts, classes, preferences
- physical laws: non-linear ODEs and PDEs
- shape constraints: monotone, bounded, convex
- qualitative descriptions, in words



MOTIVATION

FlowGP – GPs + Conditioning

GP alone vs GP + ODE for a damped pendulum



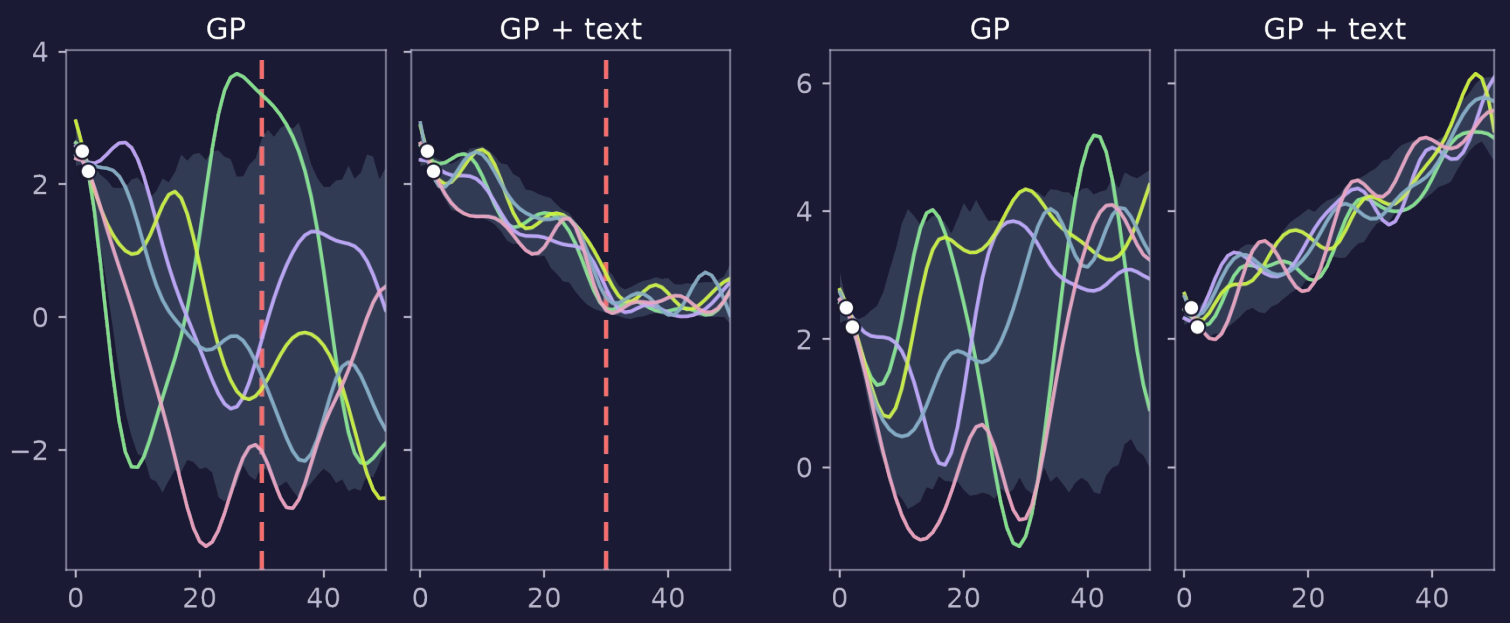
$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \sin\theta + b\,\frac{\mathrm{d}\theta}{\mathrm{d}t} = 0$$

The same GP, but on the right, it is constrained by the pendulum equation.

MOTIVATION

FlowGP – GPs + Conditioning

GP alone vs GP + LLM, for two descriptions of a stock price



"A small UK company's daily stock price in GBP over 50 trading days. On day 30 the company enters compulsory liquidation and is permanently delisted…"

"A small UK company's daily stock price in GBP over 50 trading days. The company has no notable news… but performs well. Ending with a value over 5 GBP per share."

A new approach to GP conditioning:

- Specify a likelihood $p(\boldsymbol{C} \mid f)$
- No bespoke kernel or inference
- No neural network to train

BACKGROUND

Diffusion models and guidance

$$\text{data } \boldsymbol{f}_0 \quad \underset{\text{generate}}{\overset{\text{add noise}}{\rightleftarrows}} \quad \text{noise } \boldsymbol{z}$$

Learn to undo a noising process to generate data; steer that process at test time to condition on new information.

MARS

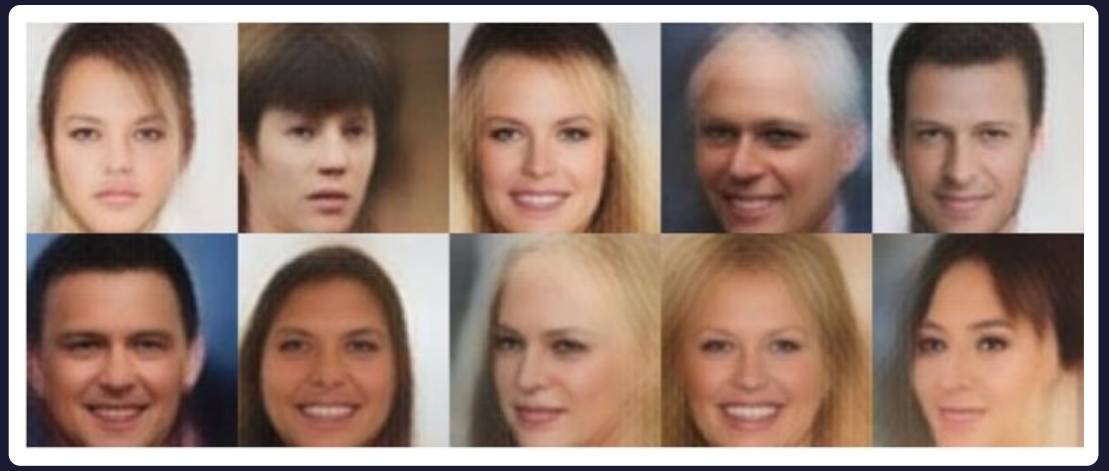
BACKGROUND

Generative modelling: one step or many

Given samples from a data distribution, learn to draw new ones by transforming simple noise z ~ N(0, I).

ONE STEP: VAEs, GANs, NORMALISING FLOWS

$$\boldsymbol{f}_0 = g_\theta(\boldsymbol{z}), \qquad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$



VAE samples: one network evaluation.

MANY STEPS: DIFFUSION MODELS

$$\boldsymbol{z} = \boldsymbol{f}_1 \rightarrow \boldsymbol{f}_{1-h} \rightarrow \cdots \rightarrow \boldsymbol{f}_0$$



Diffusion samples: hundreds of small denoising steps.

Diffusion breaks one hard transformation into many easy ones.

Sohl-Dickstein et al. (2015); Ho, Jain & Abbeel (2020); Song et al. (2021)

The forward process: turn data into noise

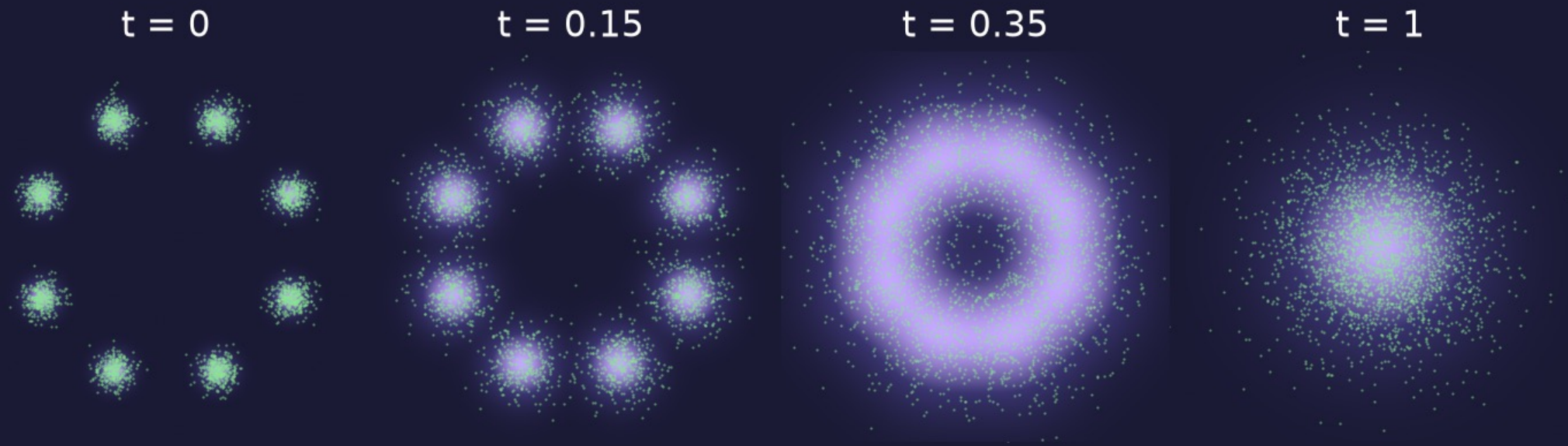
Noise the data, from t = 0 to t = 1:

$$\boldsymbol{f}_t = \alpha(t)\,\boldsymbol{f}_0 + \sqrt{1-\alpha(t)^2}\,\boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

same marginals as the variance-preserving SDE:

$$\mathrm{d}\boldsymbol{f}_t = -\tfrac{1}{2}\beta(t)\,\boldsymbol{f}_t\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\boldsymbol{W}_t$$

with $\alpha(t) = \exp\!\left(-\tfrac{1}{2}\int_0^t \beta(s)\,\mathrm{d}s\right)$



A toy data distribution, a ring of eight Gaussians, noised to t = 0, 0.15, 0.35 and 1. Points: samples of f_t; shading: the density p_t.

p_t is the data blurred by Gaussian noise, in closed form at every t.

BACKGROUND

The reverse process: turn noise into data

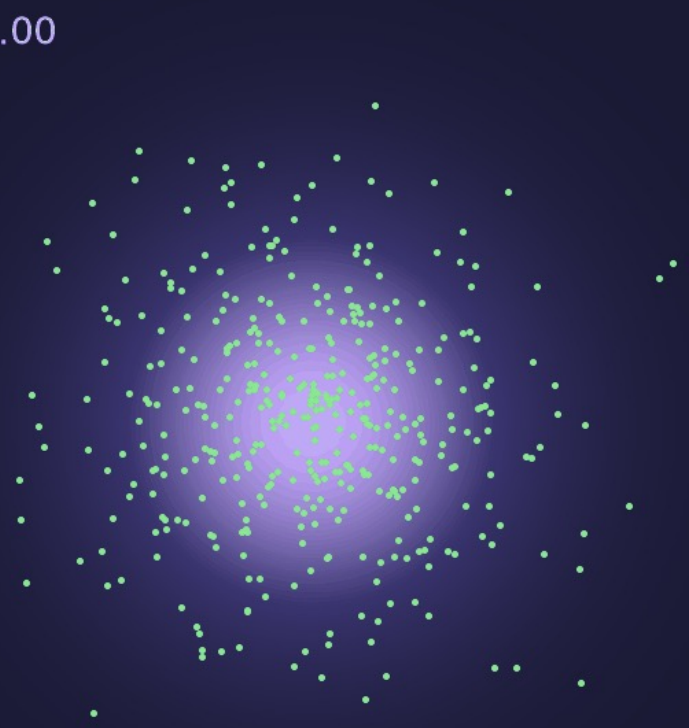
The probability-flow ODE has the same marginals p_t; integrate it from noise (t = 1) to data (t = 0):

$$\frac{\mathrm{d}\boldsymbol{f}_t}{\mathrm{d}t} = -\tfrac{1}{2}\beta(t)\Big(\boldsymbol{f}_t + \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t)\Big)$$

The **score** $\nabla \log p_t$ is unknown so we learn it with a neural network by denoising score matching:

$$\min_\theta \; \mathbb{E}_{t,\,\boldsymbol{f}_0,\,\boldsymbol{z}}\left\| s_\theta(\boldsymbol{f}_t, t) + \frac{\boldsymbol{z}}{\sqrt{1-\alpha(t)^2}}\right\|^2$$

For images, the score network is large (e.g., a U-Net or transformer). In this toy example, the score is derived analytically.



Toy: 400 samples flowing from noise to data along the probability-flow ODE; shading is p_t.

Song et al. (2021); denoising score matching: Vincent (2011); reverse-time SDE: Anderson (1982).

BACKGROUND

Bayesian conditioning: guidance

Bayes' rule splits the conditional score:

$$\nabla \log p_t(\boldsymbol{f}_t \mid \mathcal{C}) = \underbrace{\nabla \log p_t(\boldsymbol{f}_t)}_{\text{learned score}} + \underbrace{\nabla \log p_t(\mathcal{C} \mid \boldsymbol{f}_t)}_{\text{guidance}}$$

Classifier guidance — Dhariwal & Nichol (2021)

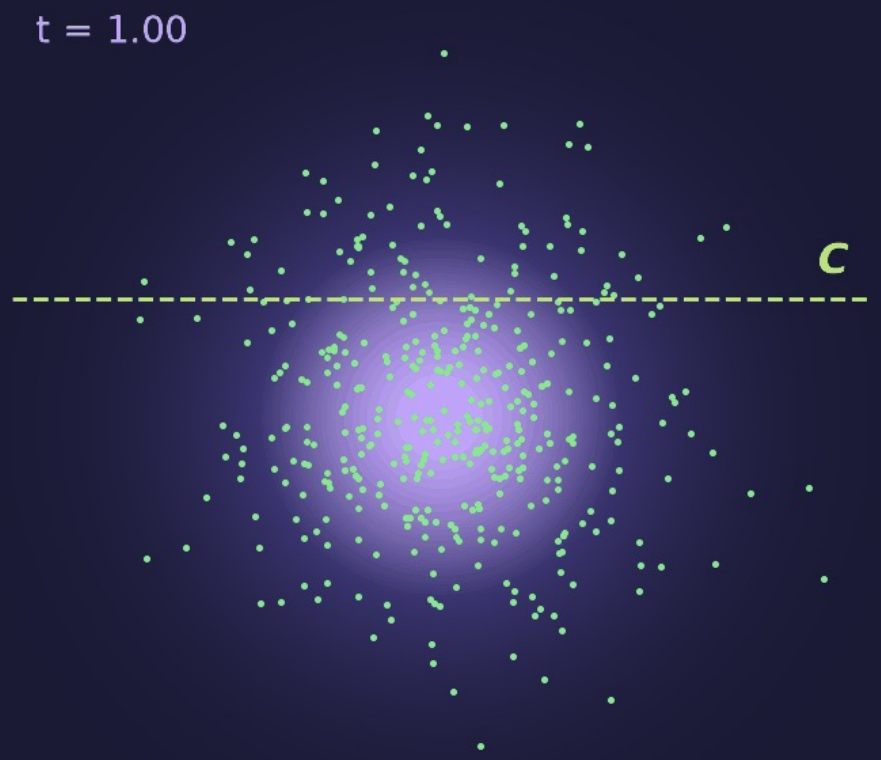
Train a classifier for C on noisy inputs f_t.

Classifier-free guidance — Ho & Salimans (2022)

Train conditional and unconditional scores together.

Training-free guidance — Chung et al. (2023); He et al. (2024)

Use a known likelihood p(C | f_0) on clean data, through the denoiser (next slide). No retraining.



Toy: condition on C = {second coordinate > 1} (dashed line) with exact guidance; shading is the unconditional p_t.

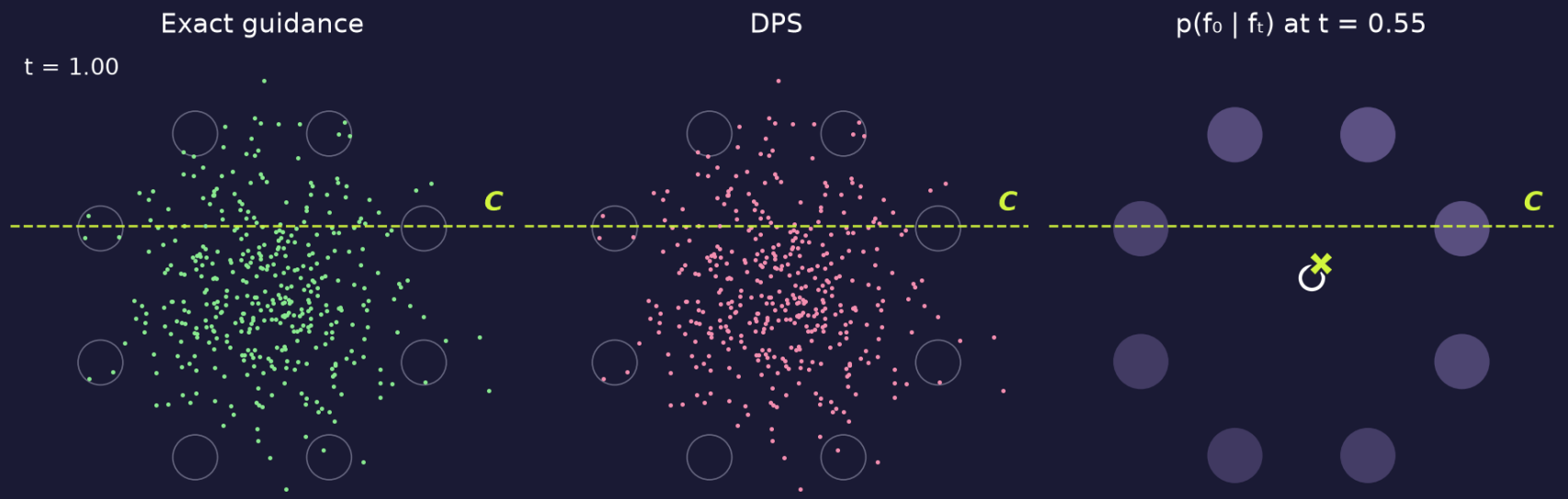
BACKGROUND

Training-free guidance

$$p(\mathcal{C} \mid \boldsymbol{f}_t) = \int p(\mathcal{C} \mid \boldsymbol{f}_0)\, \underbrace{p(\boldsymbol{f}_0 \mid \boldsymbol{f}_t)}_{\text{unknown for a neural network}}\, \mathrm{d}\boldsymbol{f}_0$$

DPS: $p(\mathcal{C} \mid \boldsymbol{f}_t) \approx p\big(\mathcal{C} \mid \mathbb{E}[\boldsymbol{f}_0 \mid \boldsymbol{f}_t]\big)$

Tweedie: $\mathbb{E}[\boldsymbol{f}_0 \mid \boldsymbol{f}_t] = \big(\boldsymbol{f}_t + (1-\alpha^2)\,\nabla \log p_t(\boldsymbol{f}_t)\big)/\alpha$



Left: exact guidance recovers the target (35% on each top mode, 15% on each borderline one); DPS loses the borderline modes and is pushed off the data. Right: between modes, the denoiser's mean (green ×) violates C although half its mass satisfies it.

GPS ARE DIFFUSION MODELS

Linear-Gaussian observations

$$\boldsymbol{y} = \boldsymbol{f}_0 + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$$

Observations of f_0 with Gaussian noise: everything stays Gaussian and in closed form.

MARS

A Gaussian process primer

A GP on m points is a Gaussian vector:

$$f_0 \sim \mathcal{N}(\mathbf{m}_*, \mathbf{K}_{**})$$

Gaussian data $y = f_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$, keep it Gaussian:

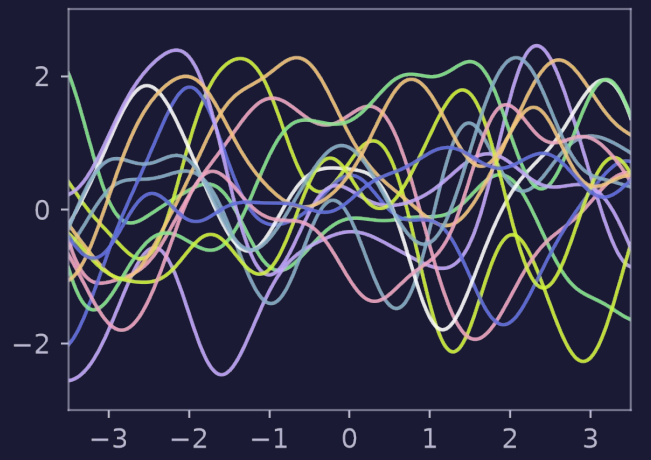
$$\mathbf{m}_{*|y} = \mathbf{m}_* + \mathbf{K}_{**}\left(\mathbf{K}_{**} + \mathbf{\Gamma}\right)^{-1}(y - \mathbf{m}_*)$$

$$\mathbf{K}_{**|y} = \mathbf{K}_{**} - \mathbf{K}_{**}\left(\mathbf{K}_{**} + \mathbf{\Gamma}\right)^{-1}\mathbf{K}_{**}$$

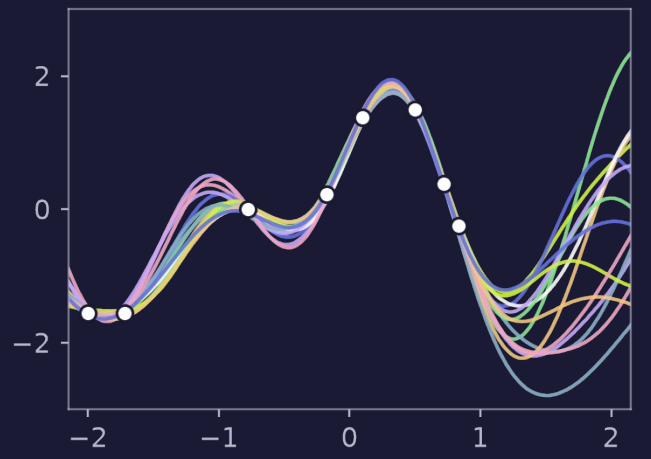
$$\Longrightarrow \quad f_0 \mid y \sim \mathcal{N}\left(\mathbf{m}_{*|y}, \mathbf{K}_{**|y}\right)$$

Exact, but only for linear-Gaussian information.

Prior samples



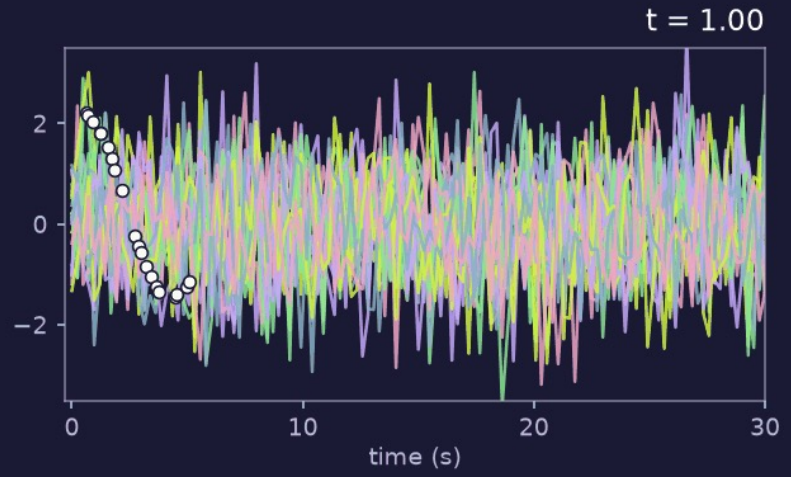
Conditioned on data



Sampling a GP: one step vs many steps

ONE STEP

$$\boldsymbol{f}_0 \sim \mathcal{N}(\mathbf{m}_*, \mathbf{K}_{**})$$
$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$$
$$\boldsymbol{f}_0 = \mathbf{m}_* + \mathbf{K}_{**}^{1/2}\,\boldsymbol{z}$$



Here the GP is also conditioned on a few observations: the same flow with $\mathbf{m}_{*|y}$, $\mathbf{K}_{**|y}$ in place of $\mathbf{m}_*$, $\mathbf{K}_{**}$.

MANY STEPS: NOISE (t = 1) TO GP (t = 0)

Noise the GP sample as done with diffusion models:

$$\boldsymbol{f}_t \sim \mathcal{N}\big(\mathbf{b}(t), \mathbf{A}(t)\big)$$
$$\mathbf{A}(t) = \alpha(t)^2\mathbf{K}_{**} + \big(1-\alpha(t)^2\big)\mathbf{I}_m, \qquad \mathbf{b}(t) = \alpha(t)\,\mathbf{m}_*$$
$$\Longrightarrow \quad \boldsymbol{f}_t = \Phi(t, \boldsymbol{z}) := \mathbf{b}(t) + \mathbf{A}(t)^{1/2}\,\boldsymbol{z}$$

at $t = 1$, $\boldsymbol{f}_1 \approx \boldsymbol{z}$ (noise); at $t = 0$, $\boldsymbol{f}_0 = \mathbf{m}_* + \mathbf{K}_{**}^{1/2}\boldsymbol{z}$, the one-step sample

A LINEAR ODE

$$\frac{\mathrm{d}\boldsymbol{f}_t}{\mathrm{d}t} = -\tfrac{1}{2}\beta(t)\,\mathbf{A}(t)^{-1}\mathbf{b}(t) - \tfrac{1}{2}\beta(t)\big(\mathbf{I}_m - \mathbf{A}(t)^{-1}\big)\boldsymbol{f}_t$$

A silly way to sample a GP? Same O(m³) cost, but with a thousand steps instead of one.

GPs are diffusion models

The linear ODE is the probability-flow ODE of the VP-SDE (Song et al., 2021):

AN EXACT SCORE

$$\frac{\mathrm{d}\boldsymbol{f}_t}{\mathrm{d}t} = -\tfrac{1}{2}\beta(t)\Big(\boldsymbol{f}_t + \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t)\Big)$$

$$\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t) = -\mathbf{A}(t)^{-1}\big(\boldsymbol{f}_t - \mathbf{b}(t)\big)$$

AN EXACT ONE-STEP DENOISER

$$p(\mathcal{C} \mid \boldsymbol{f}_t) = \int p(\mathcal{C} \mid \boldsymbol{f}_0)\, \underbrace{p(\boldsymbol{f}_0 \mid \boldsymbol{f}_t)}_{\text{exact Gaussian for a GP}}\, \mathrm{d}\boldsymbol{f}_0$$

$$\boldsymbol{f}_0 \mid \boldsymbol{f}_t \sim \mathcal{N}\Big(\mathbf{m}_* + \alpha\mathbf{K}_{**}\mathbf{A}^{-1}(\boldsymbol{f}_t - \alpha\mathbf{m}_*),\ \mathbf{K}_{**} - \alpha^2\mathbf{K}_{**}\mathbf{A}^{-1}\mathbf{K}_{**}\Big)$$

Error source	Neural score	GP
1. Score	learned, approximate	**closed form, exact**
2. Guidance	point estimate of $p(f_0 \mid f_t)$	**Monte Carlo over the exact $p(f_0 \mid f_t)$**
3. ODE solver	discretisation error	discretisation error (bounded later)

Guided diffusion applies with no training: exact score, exact denoiser.

FLOWGP

The general model

$$\boldsymbol{y} \mid \boldsymbol{f}_0 \sim p(\boldsymbol{y} \mid \boldsymbol{f}_0)$$

$$p(\boldsymbol{f}_0 \mid \boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{f}_0; \mathbf{m}_*, \mathbf{K}_{**})\, p(\boldsymbol{y} \mid \boldsymbol{f}_0)$$

A likelihood that need not be linear in f_0, or Gaussian: anything we can evaluate pointwise, from physical laws to constraints to text.

MARS

Beyond closed-form: Laplace, EP, VI and MCMC

A non-Gaussian likelihood breaks conjugacy: $p(\boldsymbol{f}_0 \mid \boldsymbol{y}) \propto \mathcal{N}(\boldsymbol{f}_0; \mathbf{m}_*, \mathbf{K}_{**}) \prod_{i=1}^{n} p(y_i \mid [\boldsymbol{f}_0]_i)$

Laplace

Gaussian at the posterior mode

$$q(\boldsymbol{f}_0) = \mathcal{N}(\hat{\boldsymbol{f}}, (\mathbf{K}^{-1} + \mathbf{W})^{-1})$$
$$\mathbf{W} = -\nabla^2 \log p(\boldsymbol{y} \mid \hat{\boldsymbol{f}})$$

+ Fast: a few Newton steps

− Local: misses skew and multiple modes

Williams & Barber (1998); Rue et al. (2009)

EP

Gaussian sites, moment-matched to the cavity

$$p(y_i \mid f_i) \approx \tilde{Z}_i \mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2)$$

+ Often the most accurate for classification

− Can fail to converge; needs a factorising likelihood

Minka (2001); Kuss & Rasmussen (2005)

VI

The best Gaussian under KL(q ‖ p)

$$\max_{q=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{f}_0)\big] - \mathrm{KL}\big(q \,\|\, p(\boldsymbol{f}_0)\big)$$

+ Scales: inducing points, minibatches

− Underestimates variance; tied to the Gaussian family

Opper & Archambeau (2009); Hensman et al. (2013, 2015)

MCMC

Asymptotically exact, e.g. elliptical slice sampling

$$\tilde{\boldsymbol{f}}' = \tilde{\boldsymbol{f}} \cos\theta + \boldsymbol{\nu} \sin\theta$$
$$\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad \tilde{\boldsymbol{f}} = \boldsymbol{f} - \mathbf{m}_*$$

+ Exact in the limit, for any likelihood

− Slow mixing when f is strongly correlated

Neal (1999); Murray, Adams & MacKay (2010)

Gaussian approximations sacrifice exactness; MCMC is slow.

FlowGP: condition on anything

Split what we condition on into two parts, D and C:

$$p(\boldsymbol{f}_0 \mid \mathcal{D}, \mathcal{C}) \;\propto\; \underbrace{p(\boldsymbol{f}_0)\, p(\mathcal{D} \mid \boldsymbol{f}_0)}_{\propto\, p(\boldsymbol{f}_0 \mid \mathcal{D})}\; p(\mathcal{C} \mid \boldsymbol{f}_0)$$

- ***D*** linear-Gaussian information (data): handled in closed form
- ***C*** everything else, as a pointwise likelihood $p(\mathcal{C} \mid \boldsymbol{f}_0)$

Bayes' rule again:

$$\nabla_{\boldsymbol{f}_t} \log p(\boldsymbol{f}_t \mid \mathcal{D}, \mathcal{C}) = \underbrace{\nabla_{\boldsymbol{f}_t} \log p(\boldsymbol{f}_t \mid \mathcal{D})}_{\text{closed form}} + \underbrace{\nabla_{\boldsymbol{f}_t} \log p(\mathcal{C} \mid \boldsymbol{f}_t, \mathcal{D})}_{\text{non-linear guidance}}$$

$$\frac{\mathrm{d}\boldsymbol{f}_t}{\mathrm{d}t} = \boldsymbol{v}(\boldsymbol{f}_t, t \mid \mathcal{D}) - \tfrac{1}{2}\beta(t)\, \nabla_{\boldsymbol{f}_t} \log p(\mathcal{C} \mid \boldsymbol{f}_t, \mathcal{D})$$

Start from white noise $f_1 \sim N(0, I)$ and integrate to $t = 0$.

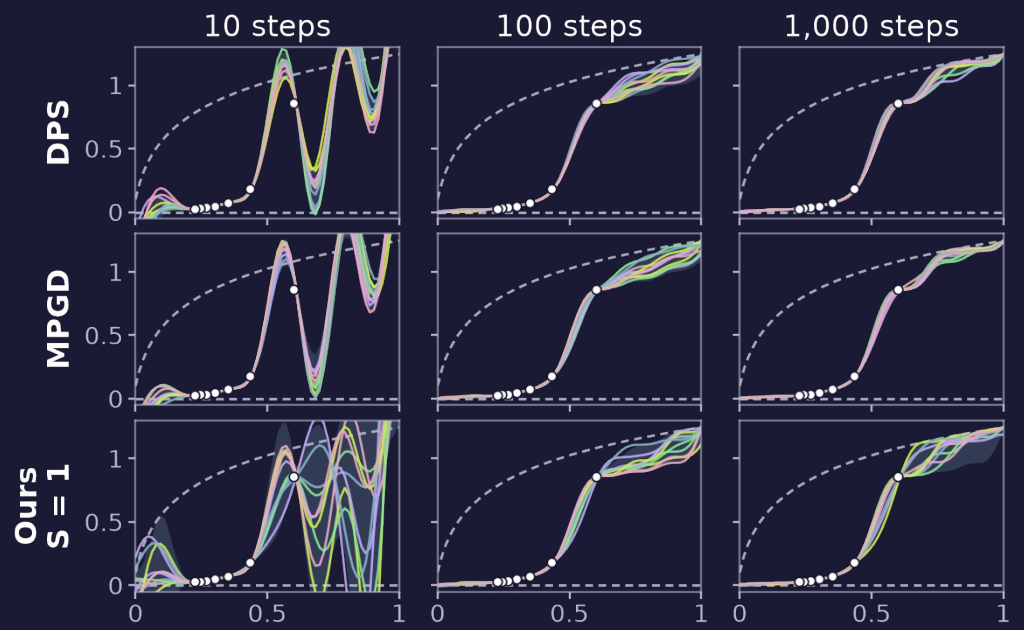
Monte Carlo approximation for the guidance

$$\nabla_{\boldsymbol{f}_t} \log p(\mathcal{C} \mid \boldsymbol{f}_t, \mathcal{D}) = \nabla_{\boldsymbol{f}_t} \log \int \underbrace{p(\boldsymbol{f}_0 \mid \boldsymbol{f}_t, \mathcal{D})}_{\text{Gaussian, closed form}} \, p(\mathcal{C} \mid \boldsymbol{f}_0) \, \mathrm{d}\boldsymbol{f}_0$$

$$\approx \; \alpha(t)\, \mathbf{K}_{**|\boldsymbol{y}} \mathbf{A}_{|\boldsymbol{y}}(t)^{-1} \sum_{i=1}^{S} \bar{w}^{(i)} \, \nabla_{\boldsymbol{f}_0} \log p(\mathcal{C} \mid \boldsymbol{f}_0^{(i)})$$

$$\bar{w}^{(i)} \propto p(\mathcal{C} \mid \boldsymbol{f}_0^{(i)}), \qquad \boldsymbol{f}_0^{(i)} \sim p(\boldsymbol{f}_0 \mid \boldsymbol{f}_t, \mathcal{D})$$

- **Self-normalised importance sampling** from a known Gaussian
- Needs only $p(\mathcal{C} \mid f_0)$ and its gradient
- Broad proposal early on: a tempered likelihood that sharpens as $t \to 0$



DPS and MPGD plug in a point estimate, and their samples collapse. Even S = 1 keeps the spread.

DPS estimates the guidance term with a first-order approximation, $\nabla_{\boldsymbol{f}_t} \log p(\mathcal{C} \mid \mathbb{E}[\boldsymbol{f}_0 \mid \boldsymbol{f}_t])$, where $\mathbb{E}[\boldsymbol{f}_0 \mid \boldsymbol{f}_t] = \alpha(t)^{-1}\big(\boldsymbol{f}_t + (1-\alpha(t)^2)\, s(\boldsymbol{f}_t, t)\big)$.

MPGD uses $\nabla_{\boldsymbol{f}_0} \log p(\mathcal{C} \mid \mathbb{E}[\boldsymbol{f}_0 \mid \boldsymbol{f}_t])$, avoiding differentiation through the denoising process.

Example: monotone and bounded regression

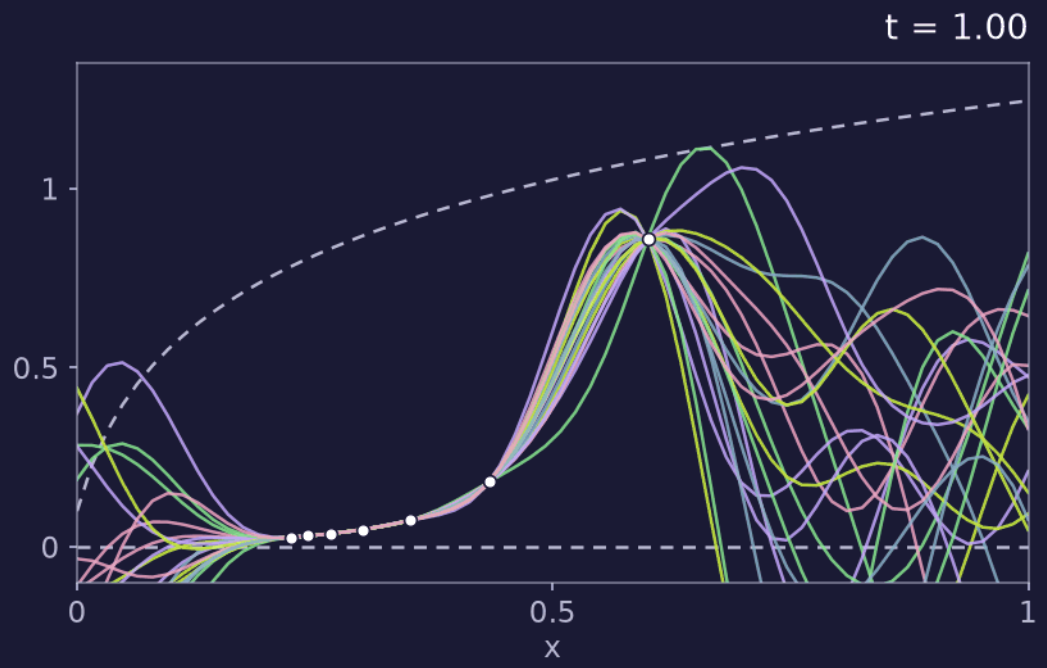
We want samples that are increasing and bounded:

$$\mathcal{C} = \{\, f \text{ increasing},\ 0 < f(x) < u(x) \,\}$$

Hard constraints become smooth likelihoods (probit, small bandwidth v):

$$\mathbb{I}\{a > 0\} \approx \Phi(a/\nu), \qquad a \in \{\, f_{i+1} - f_i,\ f_i,\ u_i - f_i \,\}$$

- Equalities become Gaussians with small variance
- The hard constraint is recovered as v → 0
- Matches bespoke monotone-GP methods on a standard benchmark



Seven observations (white), bounds dashed. In whitened coordinates the flow starts from GP samples through the data, and the constraint bends them into shape.

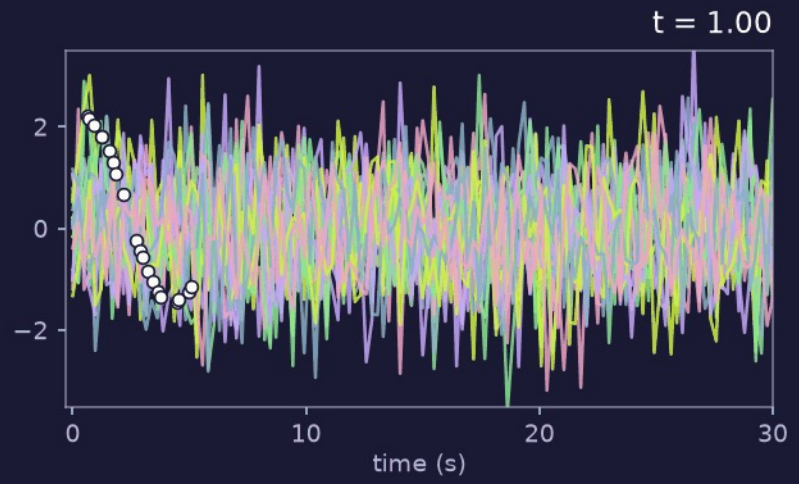
Whitening: simplifying the Gaussian part

$$\hat{f} = \mathbf{K}_{**|y}^{-1/2}(f - \mathbf{m}_{*|y}) \quad \Longrightarrow \quad \hat{\mathbf{A}}(t) = \mathbf{I}, \;\; \hat{\mathbf{b}}(t) = \mathbf{0}$$

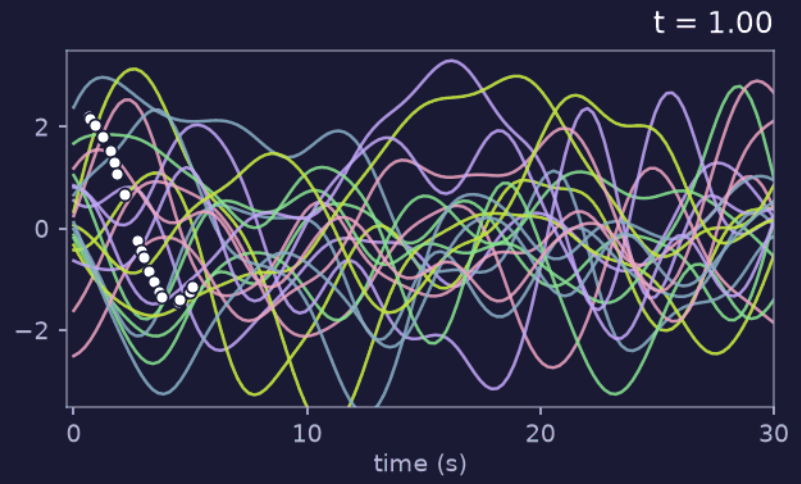
$$\frac{\mathrm{d}\hat{\boldsymbol{f}}_t}{\mathrm{d}t} = \cancelto{0}{\boldsymbol{v}(\hat{\boldsymbol{f}}_t, t \mid \mathcal{D})} - \tfrac{1}{2}\beta(t)\,\nabla_{\hat{\boldsymbol{f}}_t} \log p(\mathcal{C} \mid \hat{\boldsymbol{f}}_t, \mathcal{D})$$

Only the non-Gaussian dynamics remain; the Gaussian transport is done exactly by the linear map.

Original coordinates



Whitened (shown in original coordinates)



- Smooth at every t
- No stiffness from the GP
- Larger steps

Right animation: whitened with respect to the prior, so the data enter through the guidance term.

FlowGP algorithm

1. **Initial samples** white noise at t = 1, in whitened coordinates

$$\hat{f}_1^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m), \qquad j = 1, \ldots, n$$

2. **Whitening** one O(m³) Cholesky handles ***D*** exactly

$$\mathbf{K}_{**|y} = \mathbf{L}\mathbf{L}^\top, \qquad f = \mathbf{m}_{*|y} + \mathbf{L}\,\hat{f}$$

3. **Velocity field** Monte Carlo guidance; the Gaussian part is zero

$$\hat{v}(\hat{f}_t, t) = -\tfrac{1}{2}\beta(t)\,\alpha(t) \sum_{i=1}^{S} \bar{w}^{(i)}\, \nabla_{\hat{f}_0} \log p\big(\mathcal{C} \mid f_0^{(i)}\big)$$

$$\hat{f}_0^{(i)} \sim \mathcal{N}\big(\alpha\,\hat{f}_t,\, (1-\alpha^2)\,\mathbf{I}\big), \quad \bar{w}^{(i)} \propto p\big(\mathcal{C} \mid f_0^{(i)}\big), \quad \alpha = \alpha(t)$$

4. **Conditional ODE** Euler integration

$$\frac{\mathrm{d}\hat{f}_t}{\mathrm{d}t} = \hat{v}(\hat{f}_t, t), \qquad \text{from } t = 1 \text{ to } t = 0$$

5. **Output** independent samples, up to Monte Carlo and solver error

$$f_0 = \mathbf{m}_{*|y} + \mathbf{L}\,\hat{f}_0 \;\sim\; p(f_0 \mid \mathcal{D}, \mathcal{C})$$

FlowGP vs MCMC, Laplace, EP, VI

- \+ Any pointwise likelihood *p(**C** | f_0)*, nothing to derive
- \+ No Gaussian family: keeps skew and multiple modes
- \+ No burn-in or mixing: a fixed cost per sample, and independent samples run in parallel
- \+ Error controlled by S and step size h, with a bound
- − Biased at finite S and h; MCMC is exact as the chain runs longer
- − Needs ∇ log *p(**C** | f_0)*; hard constraints are smoothed
- − Still O(m³); VI scales with inducing points
- − Importance weights degenerate in high dimensions

THEORY

Some results on FlowGP

$$\frac{\mathrm{d}\boldsymbol{f}_t}{\mathrm{d}t} = \boldsymbol{v}(\boldsymbol{f}_t, t \mid \mathcal{D}) - \tfrac{1}{2}\beta(t)\,\nabla_{\boldsymbol{f}_t} \log p(\mathcal{C} \mid \boldsymbol{f}_t, \mathcal{D})$$

Three results about this ODE: stiffness is the condition number, whitening minimises the transport bound, and the error sources.

MARS

THEORY

Stiffness is the condition number

Unwhitened, the Gaussian part of the flow has Jacobian

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{f}_t} = -\tfrac{1}{2}\beta(t)\big(\mathbf{I} - \mathbf{A}(t)^{-1}\big), \qquad \mathbf{A}(t) = \alpha^2\mathbf{K} + (1-\alpha^2)\mathbf{I}$$

THEOREM

Suppose $\lambda_{\max}(\mathbf{K}_n)$ stays below 1 and $\lambda_{\min}(\mathbf{K}_n) \to 0$, e.g. more data in a fixed domain. Then the stiffness ratio satisfies

$$S_n = \kappa(\mathbf{K}_n)\,\frac{1-\lambda_{n,\min}}{1-\lambda_{n,\max}} \asymp \kappa(\mathbf{K}_n)$$

Why it matters

- The condition number that forces jitter into a Cholesky also sets the ODE step size
- Stiffness peaks near t = 0, so an SNR-uniform time grid puts steps there
- Whitening sets A(t) = I and removes it

THEORY

Whitening minimises the transport bound

Benamou–Brenier: any velocity field that carries p_1 to p_0 bounds the transport cost

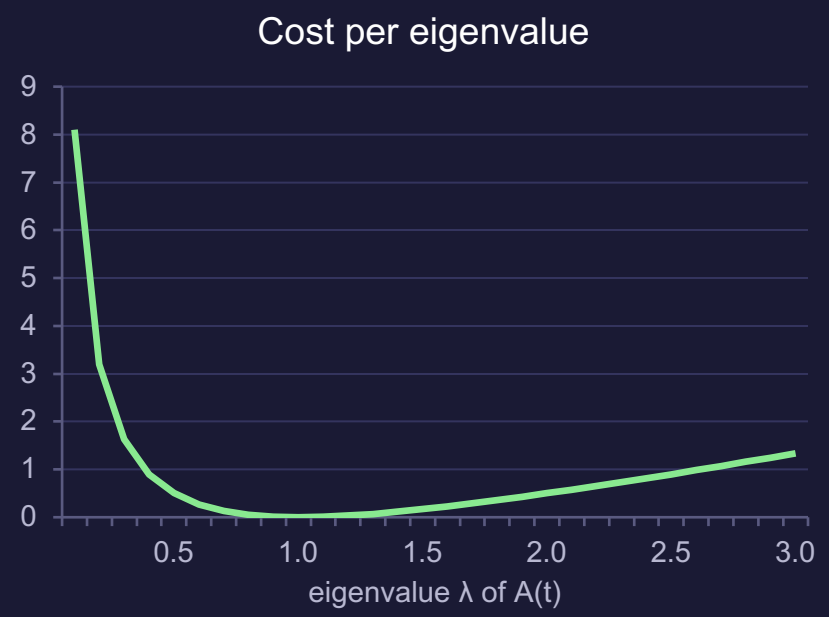
$$W_2^2(p_0, p_1) \;\le\; \int_0^1 \int \|\boldsymbol{v}(\boldsymbol{f}, t)\|^2 \, p_t(\boldsymbol{f}) \, \mathrm{d}\boldsymbol{f} \, \mathrm{d}t$$

THEOREM

For the GP flow (no guidance), with $\lambda_i(t)$ the eigenvalues of A(t):

$$W_2^2(p_0, p_1) \;\le\; \frac{1}{4} \int_0^1 \beta(t)^2 \sum_{i=1}^{m} \frac{\big(\lambda_i(t) - 1\big)^2}{\lambda_i(t)} \, \mathrm{d}t$$

For any schedule β, the bound is minimised, and equals zero, iff A(t) = I for all t: whitening.



Zero only at λ = 1. Small eigenvalues, i.e. an ill-conditioned K, are expensive.

Whitening does the Gaussian transport exactly; the flow only moves mass for C.

THEORY

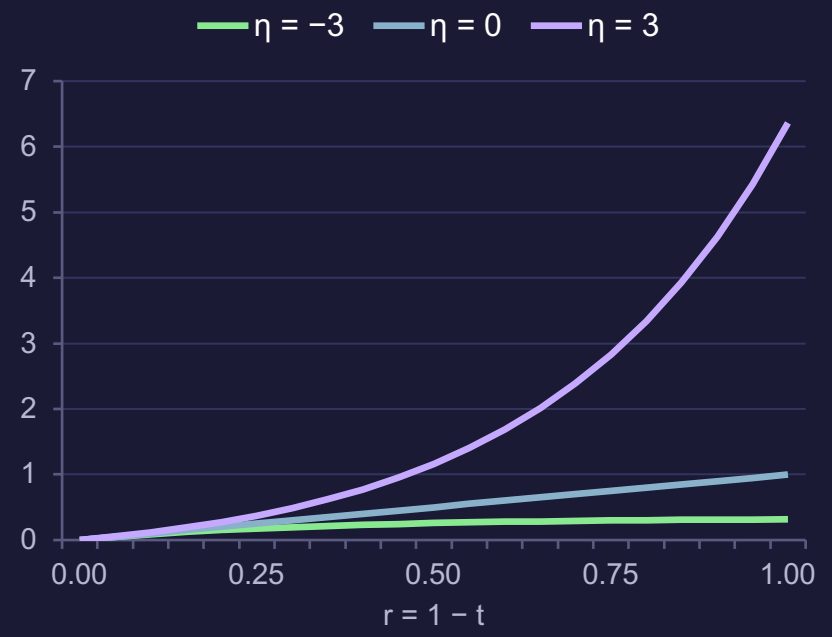
GP error bound

THEOREM

$$\left\| \boldsymbol{f}_{t_n} - \hat{\boldsymbol{f}}^{h}_{t_n} \right\| \leq \underbrace{\tfrac{1}{2} B \varepsilon \Psi_\eta(1 - t_n)}_{\text{guidance error}} + \underbrace{C h^q}_{\text{ODE solver}}$$

ε: Monte Carlo guidance error (shrinks with S) η: one-sided Lipschitz constant of the drift B = sup |β| solver of order q, step h

- **Log-concave** likelihoods (classification, Poisson, convex constraints): η ≤ 0, error bounded or linear
- **Non-log-concave** (mixtures, LLMs): exponential worst case, apparently loose in practice
- No mixing time (cf. MCMC), no variational family to choose (cf. VI): a fixed cost per sample



$$\Psi_\eta(r) = \frac{e^{\eta r} - 1}{\eta}$$

EXAMPLES

Physics and language models

$$p(\mathcal{C} \mid \boldsymbol{f}_0) = \mathcal{N}\big(\mathcal{R}(\boldsymbol{f}_0);\, \mathbf{0},\, \sigma^2\mathbf{I}\big)$$

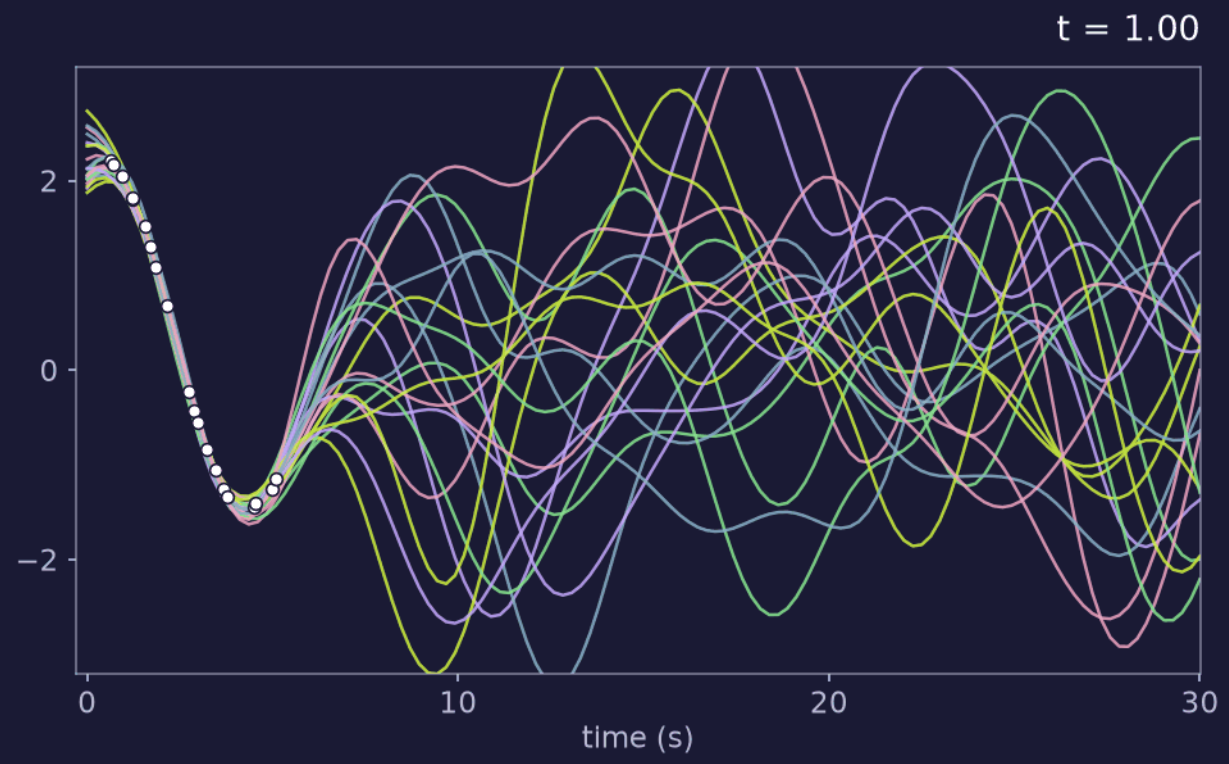
$$p(\mathcal{C} \mid \boldsymbol{f}_0) \;\propto\; q_{\mathrm{LLM}}(\boldsymbol{f}_0 \mid \mathcal{C})$$

C from physics: the residual of a differential equation. C from language: an LLM's density over f_0 given a description.

MARS

EXAMPLES

Physics: a non-linear pendulum



RBF GP + observations + the pendulum ODE, flowing from t = 1 to t = 0.

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \sin\theta + b\,\frac{\mathrm{d}\theta}{\mathrm{d}t} = 0$$

Likelihood: the ODE residual at the grid points, by finite differences

$$p(\mathcal{C} \mid \boldsymbol{f}_0) = \mathcal{N}\big(\mathcal{R}(\boldsymbol{f}_0);\, \mathbf{0},\, \sigma^2\mathbf{I}\big)$$

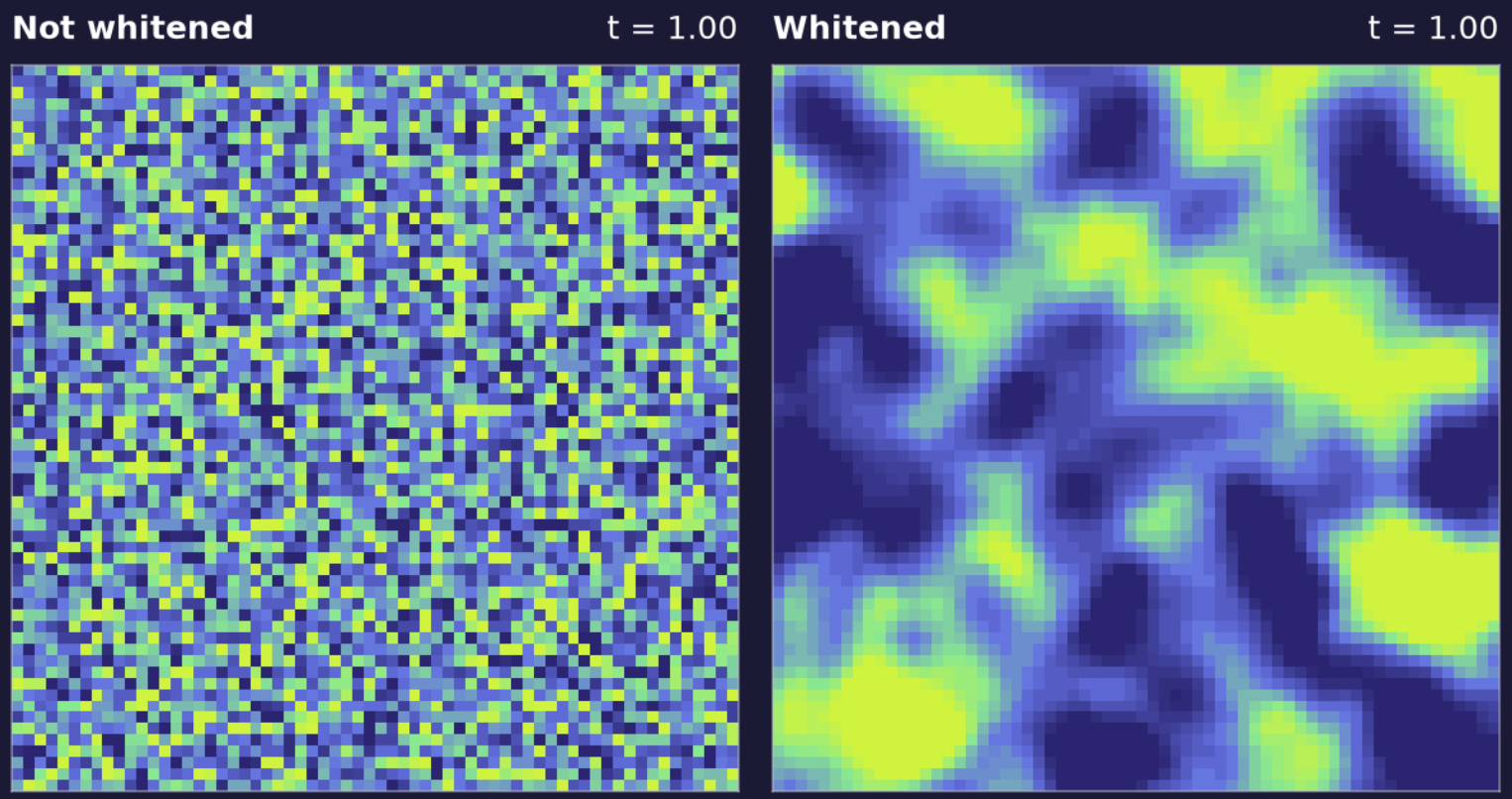
Method	RMSE	NLPD	Time (s)
PHYSS	0.05	−0.38	100
AutoIP	0.05	−0.08	100
FlowGP	0.05	−0.68	4
FlowGP, whitened	**0.05**	**−1.05**	**4**

Same accuracy, 25× faster, better calibrated. Baselines as reported in their papers.

EXAMPLES

Physics: Allen–Cahn with whitening

Stationary Allen–Cahn on a 2D grid: $af - bf^3 + c\Delta f = 0$



Time-dependent Allen–Cahn benchmark

Method	RMSE	NLPD	Time (s)
PHYSS	0.17	1.69	5,000
AutoIP	0.17	−0.29	10,000
FlowGP	0.38	3.20	4
+ whitening	**0.13**	**−0.83**	**4**

1,250–2,500× faster than the baselines. Without whitening the flow struggles; with it, best RMSE and NLPD.

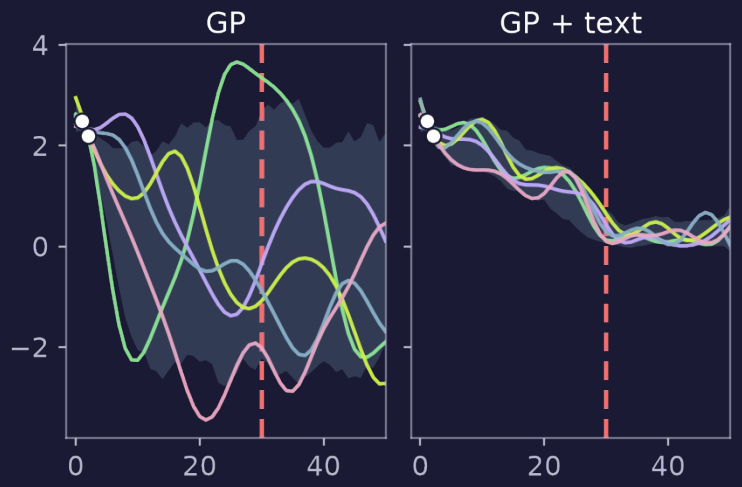
EXAMPLES

Conditioning on text with an LLM

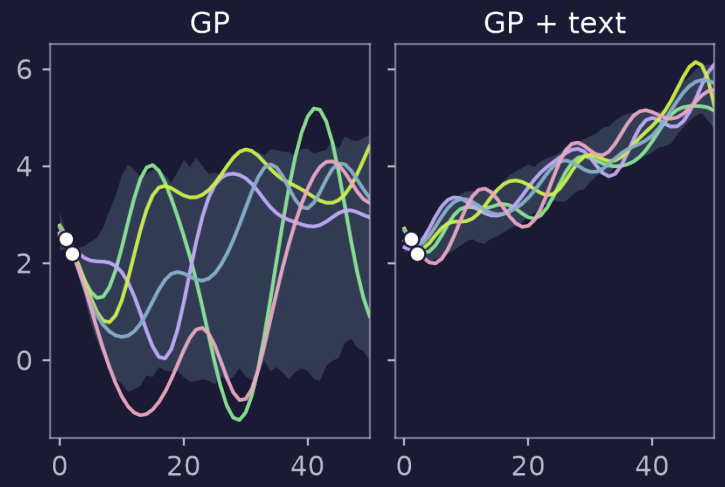
$$\pi(\boldsymbol{f}_0 \mid \mathcal{D}, \mathcal{C}) \;\propto\; p(\boldsymbol{f}_0 \mid \mathcal{D})\; q_{\mathrm{LLM}}(\boldsymbol{f}_0 \mid \mathcal{C})$$

q: an LLM's density over function values given a text description (Qwen3.5, following Biggs & Willis, 2026)

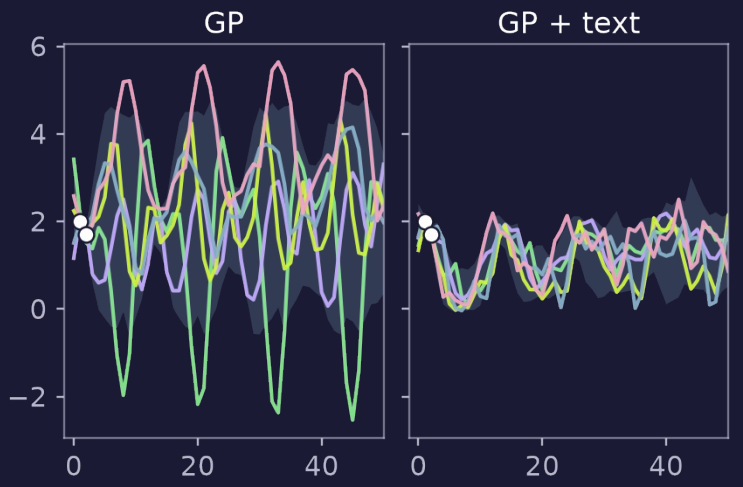
Same Monte Carlo guidance, no backpropagation through the LLM, under 2 s for 100 samples.



"A small UK company's daily stock price in GBP over 50 trading days. On day 30 the company enters compulsory liquidation and is permanently delisted…"



"A small UK company's daily stock price in GBP over 50 trading days. The company has no notable news… but performs well. Ending with a value over 5 GBP per share."



"Monthly average precipitation in San Diego in inches over 50 months, starting in January."

SUMMARY

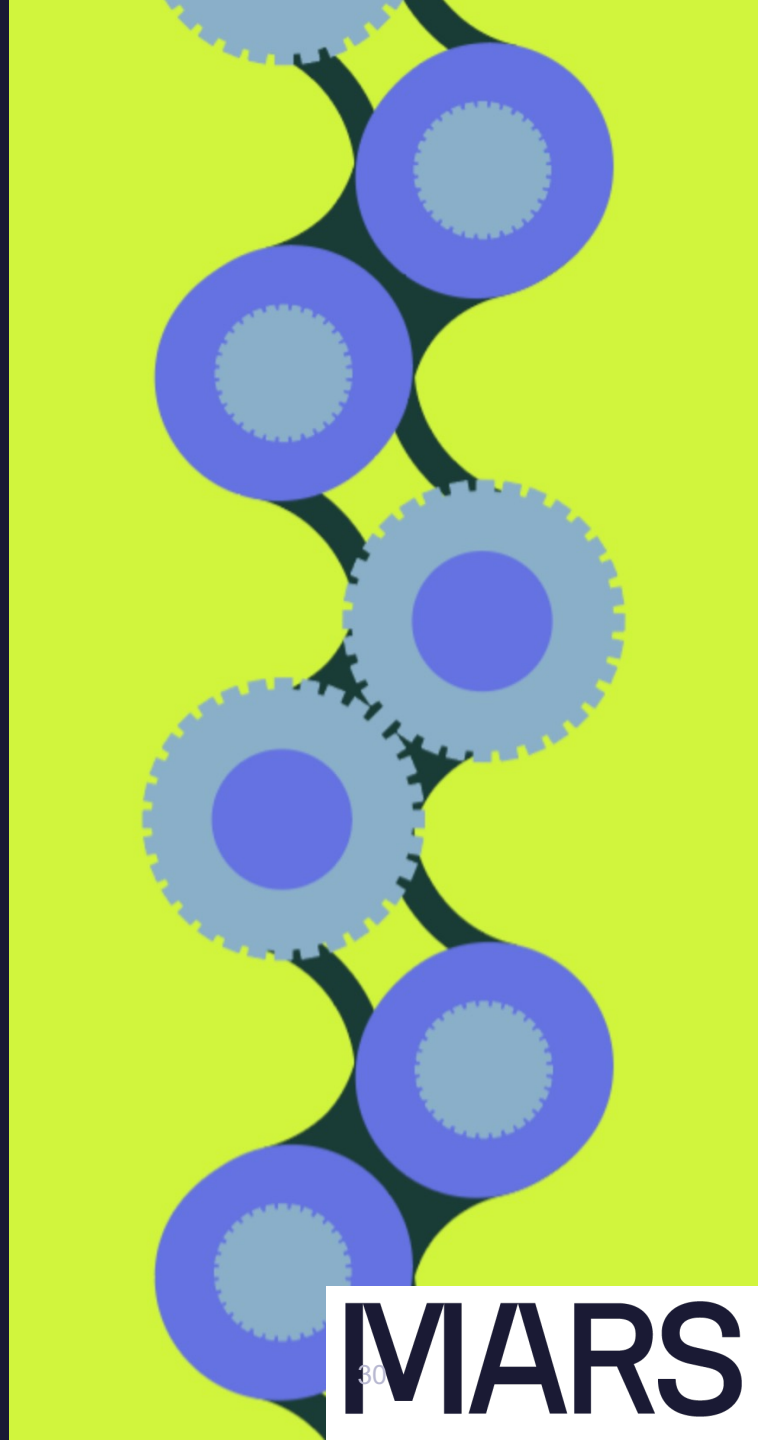
Key take aways

1. GPs are exact, closed-form diffusion models: no score network, and an exact denoiser.

2. Conditioning is guidance. Monte Carlo guidance needs only a pointwise likelihood $p(\boldsymbol{c} \mid f_0)$.

3. Whitening removes the Gaussian dynamics, minimising the transport bound and removing GP-induced stiffness.

4. A simple approach for maintaining UQ in a range of challenging physics and text problems.



OUTLOOK

Open problems

Testing diffusion methods

Which guidance, schedules and distillation methods work? GP flows have exact scores, denoisers and targets to check them against.

Scalability

$O(m^3)$ per solve. Sparse and feature-space approximations (Cowperthwaite et al., 2026) are a start.

Beyond log-concavity

Only worst-case exponential bounds, yet LLM guidance behaves well. Does Gaussian smoothing regularise?

Weight collapse

Self-normalised IS in high dimensions. Adaptive or SMC-style remedies: resampling, twisting?

Hyperparameters

Fitted on $p(f_0 \mid D)$ alone, which is biased when C is informative. Learn them through the ODE?

Beyond GPs

Guide other stochastic processes the same way, e.g. LatentFlow (Sharrock, Astfalck & Moss).

Thank you · c.nemeth@lancaster.ac.uk · Moss et al., Conditioning Gaussian Processes on Almost Anything. NeurIPS 2026.