Efficient Bayesian inference for full waveform inversion : an overview of modern approaches

SEISCOPE — UGA Université Grenoble Alpes — SEG | AAPG image International Meeting for Applied Geoscience & Energy

Philippe Marchner[1], Romain Brossier[1], Ludovic Métivier[1-2]
[1]Univ. Grenoble Alpes, ISTerre [2]Univ. Grenoble Alpes, CNRS, LJK

Full waveform inversion

Full waveform inversion attempts to recover a high-resolution image of the subsurface given observed seismic recordings

$$\boxed{d_{\text{obs}} = \mathcal{F}(m) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\Gamma)} \qquad \textit{(inverse problem)}$$

$\mathcal{F}$: parameter-to-observation map, $d_{\text{syn}} := \mathcal{F}(m)$ synthetic data
→ requires **solving a wave equation**
Goal: find a model that minimizes a data misfit (Virieux et al., 2017)

$$\min_{m\in\mathbb{R}^n} C(m), \quad C(m) = \frac{1}{2}\|\mathcal{F}(m) - d_{\text{obs}}\|^2_{L^2(\Gamma)} \qquad (L^2\textit{-misfit/cost})$$

FWI is a challenging inverse problem:

- several models m may explain **equally well** the observations d_{obs}
- the map $\mathcal{F}$ is strongly **nonlinear**
- **parameter dimension** $n \approx 10^9$ in 3D FWI

Uncertainty quantification characterizes the ill-posedness of the inversion

Bayesian approach to inverse problems

Bayes' theorem computes a **probability distribution** of all models

$$\boxed{\pi(m|d_{\text{obs}}) = \frac{\pi(d_{\text{obs}}|m)\,\pi_{\text{prior}}(m)}{\pi(d_{\text{obs}})}} \qquad \textit{(Bayes' theorem)}$$

- $\pi_{\text{prior}}(m)$: prior knowledge of the subsurface ↔ cost regularization
- $\pi(d_{\text{obs}}|m) = \exp(-C(m))$: how likely an observation explains the model. One evaluation requires **solving a wave equation**

How to sample from the posterior $\pi_{\text{post}}(m) := \pi(m|d_{\text{obs}})$?

(*Machine learning*)

Markov Chain Monte Carlo	SVGD, normalizing flows diffusion models, ...	Gaussian approx. at the MAP

← expensive, accurate posterior ? cheap, approximate posterior →

A unified perspective ?

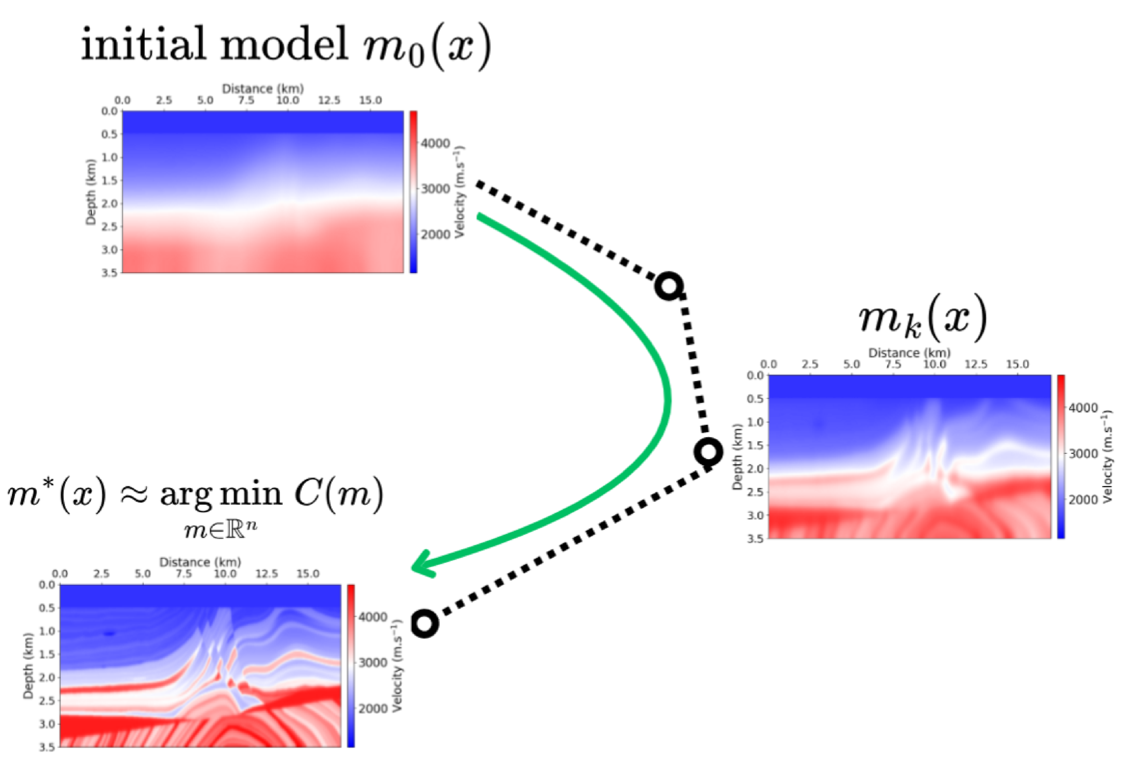
Gradient flows in probability spaces

Standard FWI uses gradient iterative methods of the form

$$m_{k+1} = m_k - \underbrace{\tau_k}_{\text{line search}} \underbrace{\nabla C(m_k)}_{\text{adjoint method}}, \quad \tau_k \text{ step size,}$$

to reach a local minimizer of the misfit C. It can be seen as an explicit Euler scheme in **artificial time**

$$\begin{cases} \dfrac{dm}{dt} = -\nabla C(m), \\ m(0) = m_0 \end{cases} \qquad \textit{(gradient flow in } \mathbb{R}^n\textit{)}$$



Can we define a gradient flow in $\mathcal{P}_2(\mathbb{R}^n)$?

Probabilistic FWI aims to minimize a cost functional over **probability densities**

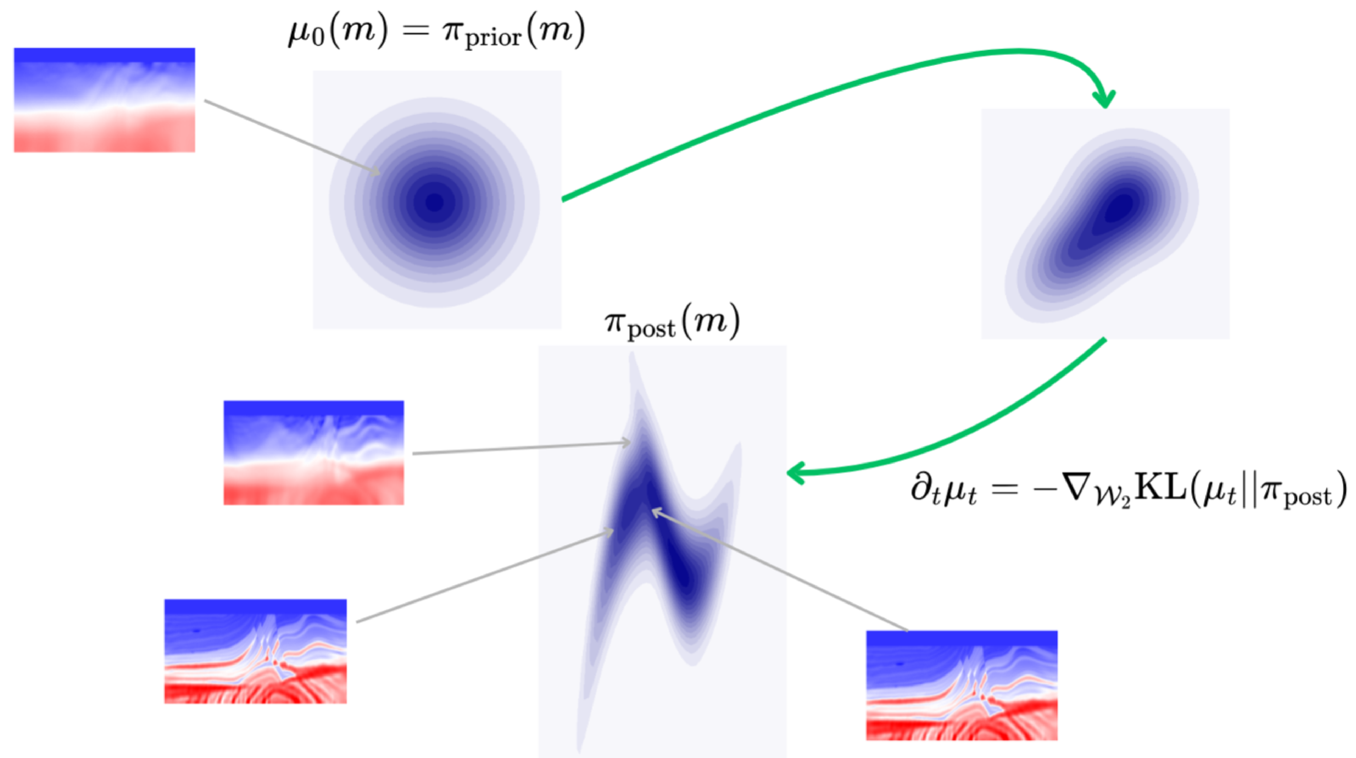
$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^n)} \text{KL}(\mu||\pi_{\text{post}}) := \int \mu \log\left(\frac{\mu}{\pi_{\text{post}}}\right) dm \qquad (\text{KL } \textit{divergence})$$

One choice is the relative information between μ and the posterior π_{post}.
We can take the gradient in $\mathcal{P}_2(\mathbb{R}^n)$ in the **Wasserstein geometry** (Jordan et al., 1998)

$$\frac{\partial \mu_t}{\partial t} = \text{div}\left(\mu_t \nabla\left[\frac{\delta}{\delta\mu}\text{KL}(\mu_t||\pi_{\text{post}})\right]\right), \quad \frac{\delta}{\delta\mu}\text{KL}(\mu||\pi) = \log\frac{\mu}{\pi} + 1,$$

$$= \text{div}(\nabla C \mu_t) + \Delta\mu_t, \quad (\textit{advection} + \textit{diffusion})$$

which is a Fokker-Planck PDE that transports the prior to the posterior



Gradient flow in $\mathcal{P}_2(\mathbb{R}^n)$ is a mass transport problem

How to numerically perform such a gradient descent ?

Particle dynamics

1 - Stochastic particle evolution (SDE) - Langevin dynamics

$$dM_t = -\nabla C(M_t) + \sqrt{2}dB_t, \quad M_t \sim \mu_t$$

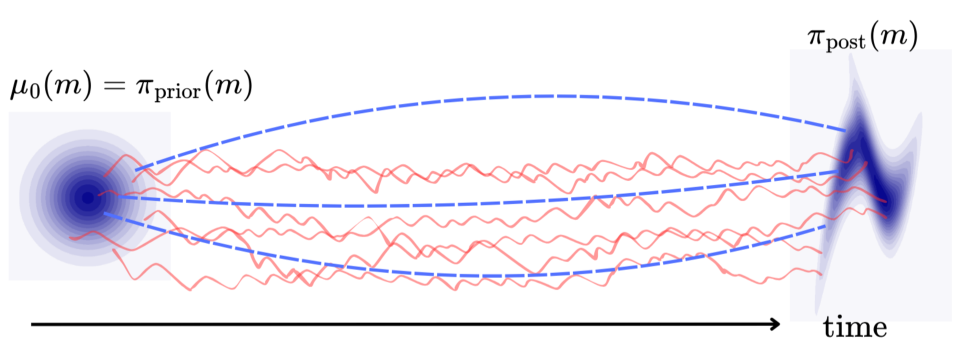
Gradient drift towards a mode + Brownian motion

2 - Deterministic particle evolution (ODE)

$$\frac{dM_t}{dt} = -\nabla\log(\mu_t/\pi_{\text{post}}) = -\nabla C(M_t) - \nabla\log\mu_t, \quad M_t \sim \mu_t$$

but we need an approximation of $\nabla\log\mu_t$ from the particles !

(i) use a kernel density estimation → WGD (Wang et al., 2022)
(ii) constrain gradient in kernelized space → SVGD (Liu and Wang, 2016)



SDE vs **ODE** trajectories

Gaussian variational inference

Density μ_t is intractable → assume it is Gaussian $\mu_t(m) \approx \mathcal{N}(\overline{m}, \Sigma)$
Wasserstein gradient flow found by moment closure (Lambert et al., 2022)

$$\begin{cases} \dfrac{d\overline{m}}{dt} = -\mathbb{E}_{\mu_t}[\nabla C] \quad \left(:= -\int \mu_t(m)\nabla C(m)dm\right) \\ \dfrac{d\Sigma}{dt} = 2I - \Sigma\mathbb{E}_{\mu_t}[\nabla^2 C] - \mathbb{E}_{\mu_t}[\nabla^2 C]\Sigma \end{cases} \qquad \textit{(Gaussian ODE)}$$

More realistic : Gaussian mixture $\mu_t(m) \approx \frac{1}{K}\sum_{i=1}^K \mathcal{N}(\overline{m}_i, \Sigma_i)$

$$\begin{cases} \dfrac{d\overline{m}^{(i)}}{dt} = -\mathbb{E}_{\mu_t^{(i)}}[\nabla\ln(\mu_t/\pi_{\text{post}})] \quad \left(\mu_t^{(i)}(m) := \mathcal{N}(\overline{m}_i, \Sigma_i)\right) \\ \dfrac{d\Sigma^{(i)}}{dt} = -\mathbb{E}_{\mu_t^{(i)}}[\nabla^2\ln(\mu_t/\pi_{\text{post}})]\Sigma^{(i)} - \Sigma^{(i)}\mathbb{E}_{\mu_t^{(i)}}[\nabla^2\ln(\mu_t/\pi_{\text{post}})] \end{cases}$$

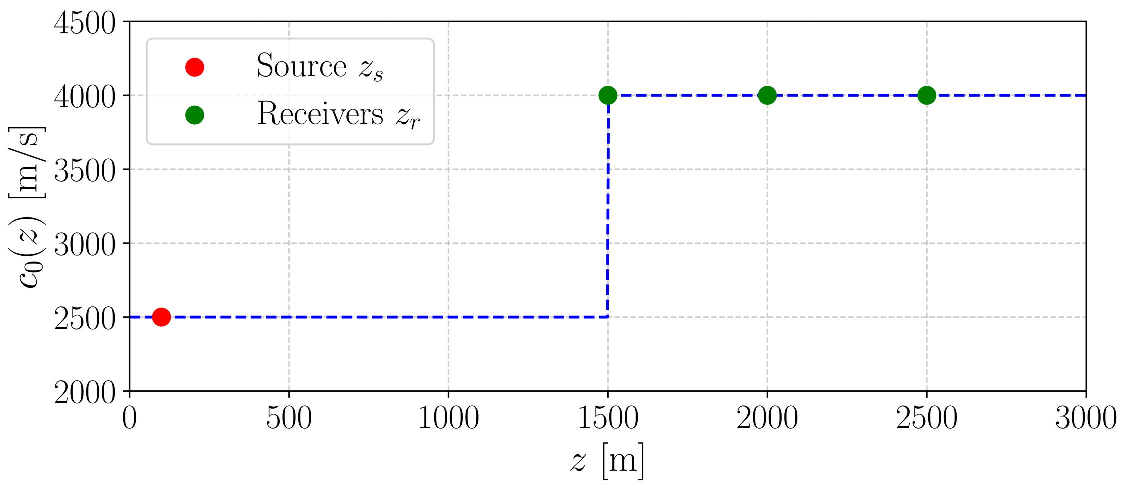
- Integration by parts $\mathbb{E}_\mu[\nabla^2\pi]\Sigma = \mathbb{E}_\mu[\nabla\pi\otimes(m-\overline{m})] \Rightarrow$ Hessian-free
- We need a time-discretization scheme + high-dimensional quadrature

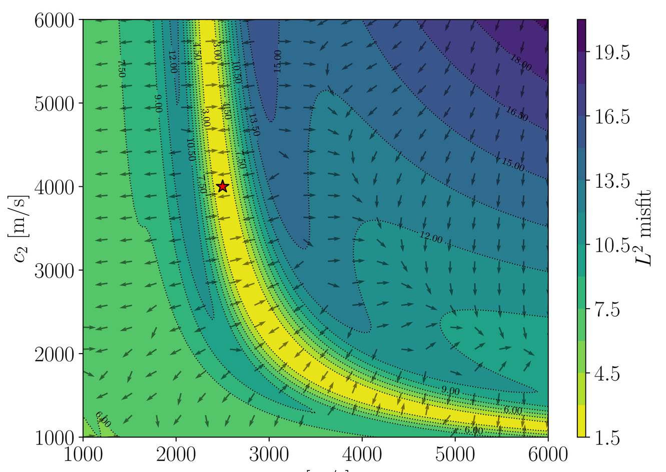
Conclusion: many options for Bayesian inference !

Efficient Bayesian inference for full waveform inversion : an overview of modern approaches

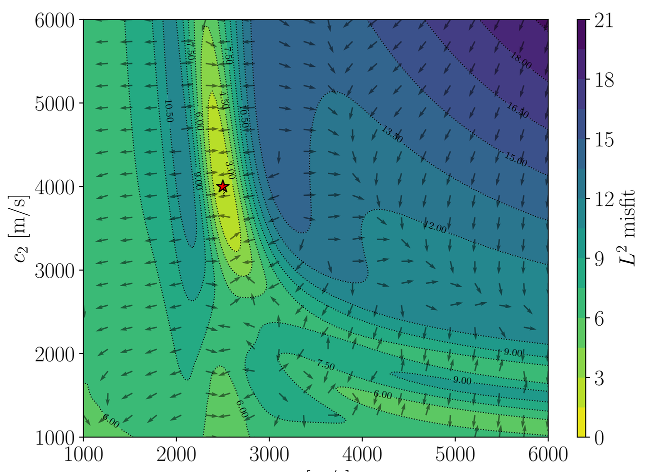
philippe.marchner@univ-grenoble-alpes.fr

Application to full waveform inversion





$C(m)$ - **1 receiver at z_r= 2000m**



$C(m)$ - **3 receivers**

Simple FWI with 2 velocity parameters $m(x) := c_0(z) = (c_1, c_2)$

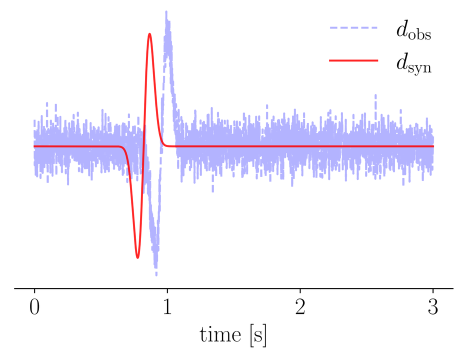
Acoustic Wave Equation (1D)

$$\frac{1}{c_0^2(z)}\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial z^2} = s(z_s, t) \quad \text{in } \Omega \times [0, T], \quad \text{Ricker wavelet at 5 Hz}$$

$$\frac{\partial u}{\partial t} \pm c_0(z)\frac{\partial u}{\partial z} = 0 \quad \text{at } z = \{0, D\} \quad \text{(absorbing BC)}$$

Bayesian inference setting:

- noisy artificial observations d_{obs}
- signal-to-noise ratio $\approx$ 4 dB
- uniform prior $\pi_{\text{prior}}(m) = 1$



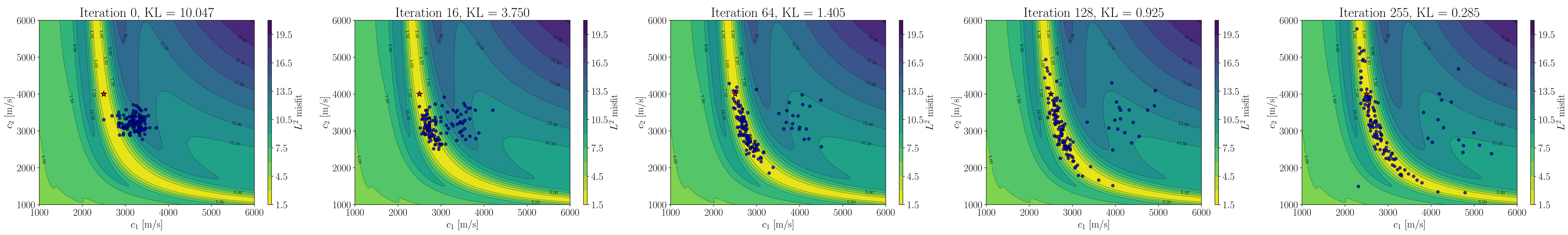
we want to infer the posterior density $\pi_{\text{post}}(m) \propto \exp(-C(m))$

Langevin dynamics

Euler-Maruyama discretization : for each particle $M_k \in \mathbb{R}^n$,

$$M_{k+1} = M_k - \tau_k \nabla C(M_k) + \sqrt{2\tau_k}\,\xi_k, \quad \xi_k \sim \mathcal{N}(0, I)$$

Cost: $n_{\text{particles}}\times$ **gradient evaluations** per iteration
Finite $\tau_k \Rightarrow$ sampling bias, removed by an acceptance step (Roberts and Tweedie, 1996)
Can represent complex posteriors but is not computationally efficient



Simulation with 128 particles, 256 iterations. Cycle-skipping creates an energy barrier for particles.

Gaussian variational inference

Wasserstein gradient flow with Gaussian particles

Means $\overline{m}^{(i)} \in \mathbb{R}^n$ and covariances $\Sigma^{(i)} \in \mathbb{R}^{n\times n}$ updated at each iteration

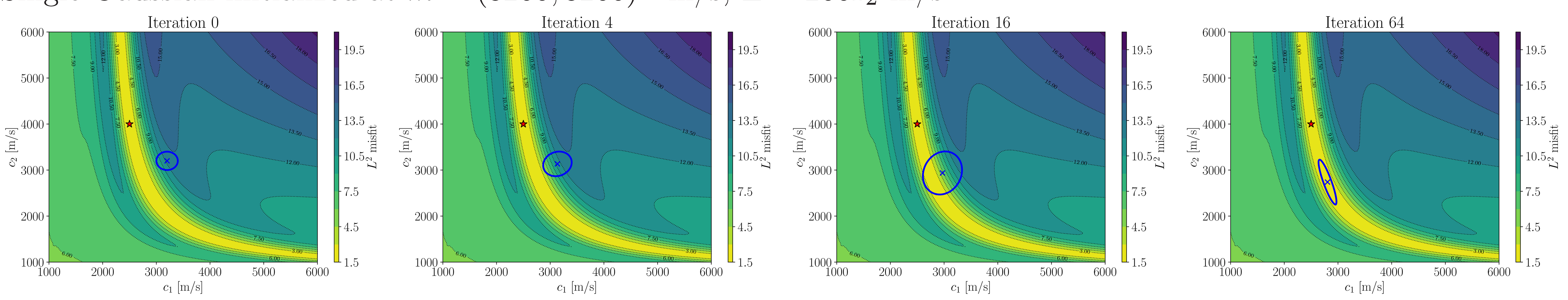
Cost per Gaussian: $n_q\times$ **gradient evaluations** per iteration

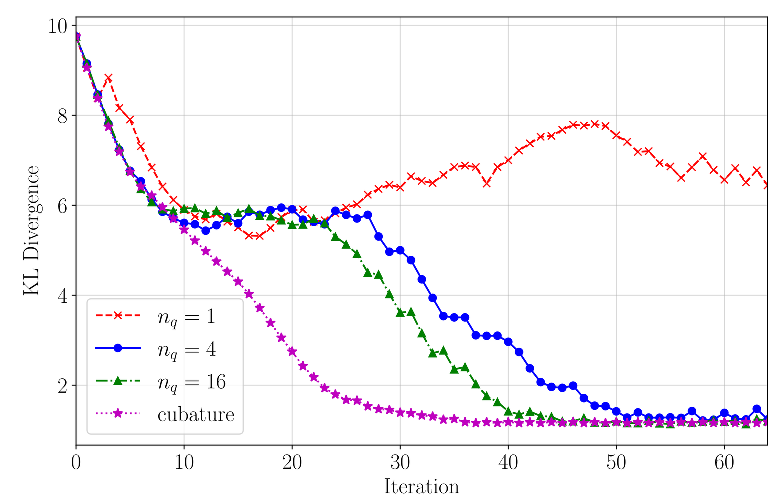
- How many quadrature points n_q ?
- Cycle-skipping $\Rightarrow$ non-concave posterior $\Rightarrow$ good initialization

Numerical resolution of the ODE system

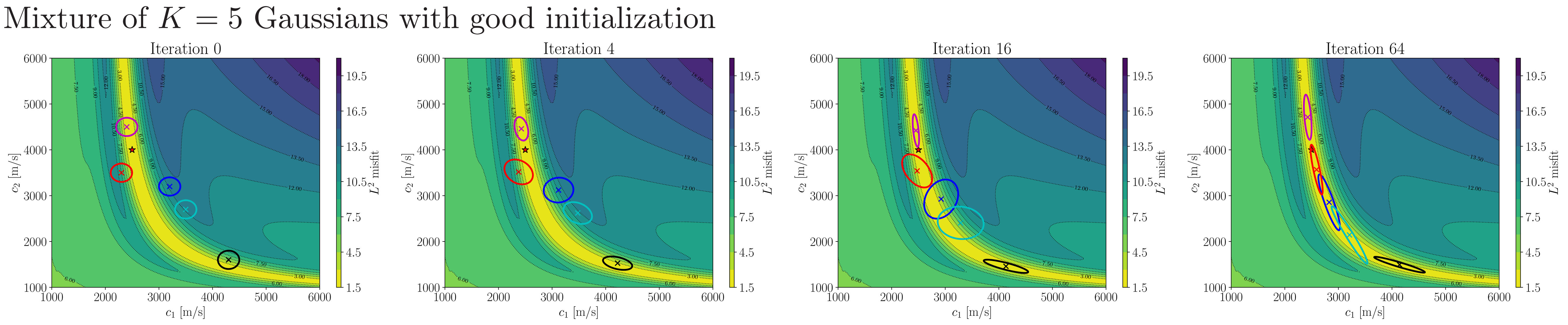
1. Time discretization: *Adam* optimization (Kingma and Ba, 2015)
   $\rightarrow$ Remark - the ODE is stiff
2. Quadrature rule for expectations, for each Gaussian:
   (i) Monte-Carlo: n_q samples
   (ii) Cubature: n_q= $2n$ (increases with parameter dimension)

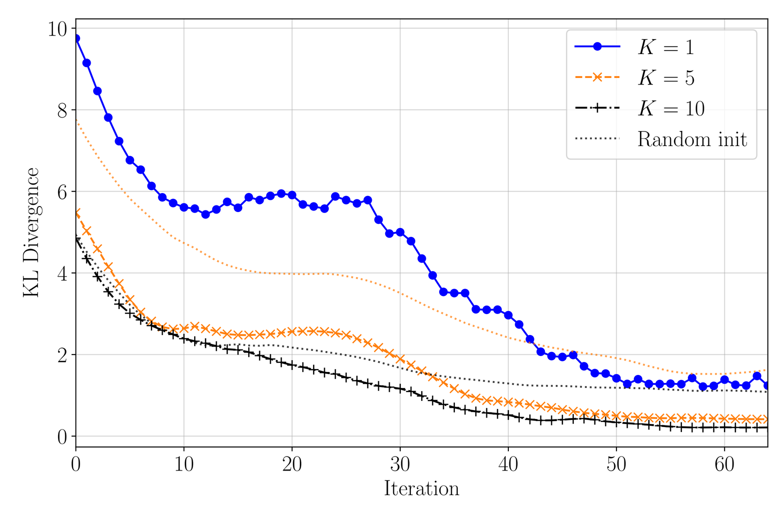
Single Gaussian initialized at $\overline{m} = (3200, 3200)^T$ m/s, $\Sigma = 200 I_2$ m/s





Mixture of $K = 5$ Gaussians with good initialization





Perspectives: efficient numerical schemes + evolve weights of Gaussian mixture to mitigate cycle-skipping (Chen et al., 2024)

Towards high dimensional inference

Most sampling methods face the **curse of dimensionality** when $n \gg 1$

- Langevin dynamics: step size $\tau \sim 1/n$ for controlled accuracy
- Gaussian VI: n^2 parameters + high-dimensional quadrature

Dimensionality reduction: find a data-informed subspace with $r \ll d$
$\Rightarrow$ randomized SVD of gradient log-likelihood (Hessian) (Wang et al., 2022)

Perspective: How large is r for typical FWI applications ?

Conclusion

Bayesian inference for FWI $\Leftrightarrow$ sampling from a high-dimensional probability density. Crucial ingredients are required

- Efficient sampling: **Bayesian inference as a gradient flow in $\mathcal{P}_2$** opens many new methods and numerical schemes
- Dimensionality reduction strategy
- Forward solver surrogate
- Quality metrics for uncertainty assessment