Ensemble Inference Methods for Models with Noisy and Expensive Likelihoods

Marie-Therese Wolfram

in collaboration with O.R.A. Dunbar, A. Duncan and A.M. Stuart

Algorithms Seminar
Warwick Stats

Focus of this talk:

Propose, discuss and analyse ensemble methods for noisy and expensive likelihood.

Focus of this talk:

Propose, discuss and analyse ensemble methods for noisy and expensive likelihood.

Today's plan:

- ▷ Bayesian inverse problems
- ▷ Multi-scale forward problems
- ▷ Interacting particle methods aka ensemble sampling
- ▷ Formal multi-scale analysis in case of noisy likelihood functions
- ▷ Computational experiments

Bayesian inverse problems

The inverse problem: Given observations $y \in \mathbb{R}^K$ infer $x \in \mathbb{R}^d$ based on evaluations of $G_0(x)$ polluted by noise ξ:

$$y = G_0(x) + \xi.$$

- ▷ Assumption: noise $\xi \sim \mathcal{N}(0, \Gamma)$, with strictly positive-definite covariance $\Gamma \in \mathbb{R}^{K \times K}$.

Bayes rule: Imposing a Gaussian prior $x \sim \mathcal{N}(m, \Sigma)$, the posterior distribution is given by

$$\pi_0(x) \propto e^{-V_0(x)},$$
$$V_0(x) := \frac{1}{2}|y - G_0(x)|_\Gamma^2 + \frac{1}{2}|x - m|_\Sigma^2.$$

Multi-scale problem

Assume that the forward problem has the following multi-scale structure:

$$G_\epsilon(x) = G_0(x) + G_1(x/\epsilon).$$

where $\epsilon \ll 1$, $G_0(\cdot)$ is expensive to evaluate and only noisy observations of G_ϵ are available.

- ▷ G_1 can be random or periodic
- ▷ Multi-scale nature arises for example when using time-averaged statistics as data from chaotic systems, ...

Goal:

Solve the inverse problem defined by G_0, using only evaluations of G_ϵ, not of G_0.

The associated multi-scale posterior is $\pi_\epsilon(x) \propto e^{-V_\epsilon}(x)$ with multi-scale potential

$$V_\epsilon(x) := \frac{1}{2}|y - G_\epsilon(x)|_\Gamma^2 + \frac{1}{2}|x - m|_\Sigma^2.$$

Multi-scale forward problems when using time-averaged data

Consider the following parameter-dependent dynamical system:

$$\frac{du}{ds} = F(u;\theta), \quad u(0) = u_0,$$

which we assume to be ergodic and mixing.

Goal: identify θ from data y computed from finite time-averages of a function $\varphi(\cdot)$ over time-interval T:

$$y = \mathcal{G}_\epsilon(\theta) + \xi_{obs}, \quad \text{where } \mathcal{G}_\epsilon(\theta) = \frac{1}{T}\int_0^T \varphi\big(u(s;\theta)\big)\,ds,$$

and $\xi_{obs} \sim \mathcal{N}(0, \Delta_{obs})$ is the observational noise.

For ergodic, mixing dynamical systems a central limit theorem may hold; then

$$\mathcal{G}_\epsilon(\theta) \approx \mathcal{G}_0(\theta) + \mathcal{G}_1(\theta),$$
$$\mathcal{G}_1(\theta) \sim \mathcal{N}\big(0, T^{-1}\Delta(\theta)\big),$$

- ▷ $\mathcal{G}_0$ is the infinite time-average, which is independent of the initial condition u_0;
- ▷ Noise induced by the unknown initial condition u_0 only in $\mathcal{G}_1$.

Multi-scale forward problems when using time-averaged data

- We approximate $T^{-1}\Delta(\theta)$ by a constant covariance Δ_{model} estimated from a single long-run of the (ergodic and mixing) model at a fixed parameter $\theta^\dagger$ and batched into windows of length T.

- If $\xi_{model} \sim \mathcal{N}(0, \Delta_{model})$ and if ξ_{obs} is independent of the initial condition u_0, then we can rewrite the inverse problem as

$$y = \mathcal{G}_0(\theta) + \xi,$$

where $\xi = \xi_{obs} + \xi_{model} \sim \mathcal{N}(0, \Delta_{obs} + \Delta_{model})$.

Interacting particle methods aka ensemble sampling

The setting: Consider N interacting particles X_t^i, $i = 1, \dots N$, which explore the data landscape (aka posterior or target) satisfying a stochastic differential equation.

Objective:

- ▷ Sampling: generate approximate samples from the log-posterior distribution; particle ensemble should approximate target as $t \to \infty$
- ▷ Optimisation: find a minimiser of the target particle ensemble collapses in the minimum; no quantification of uncertainty.

Gradient-based vs. gradient-free approaches:

- Gradient-based methods: ensemble Langevin sampler (ELS), Metropolis Adjusted Langevin Algorithm (MALA)....
  Often derived from over-damped Langevin equations

$$dX_t^i = -K\nabla V(X_t)dt + \sqrt{2K}dW_t$$

  where K is symmetric and pos. definite.

- Gradient-free methods: ensemble Kalman sampler (EKS), consensus based optimisation),...

Ensemble Kalman sampler (EKS)

EKS comprises N coupled SDEs in $\mathbb{R}^d$, for X_t^i given by

$$dX_t^i = -\left(\frac{1}{N}\sum_{n=1}^{N}\langle G_\epsilon(X_t^n) - \overline{G}_{\epsilon,t}, G_\epsilon(X_t^i) - y\rangle_\Gamma X_t^n\right) dt - C_t\Sigma^{-1}(X_t^i - m)\, dt$$
$$+ \frac{d+1}{N}(X_t^i - \overline{X}_t)\, dt + \sqrt{2C_t}\, dW_t^i;$$

here the W^i are standard independent Brownian motions in $\mathbb{R}^d$ and

$$\overline{X}_t = \frac{1}{N}\sum_{n=1}^{N} X_t^n, \qquad \overline{G}_{\epsilon,t} = \frac{1}{N}\sum_{n=1}^{N} G_\epsilon(X_t^n),$$
$$C_t = \frac{1}{N}\sum_{n=1}^{N}\left(X_t^n - \overline{X}_t\right)\otimes\left(X_t^n - \overline{X}_t\right).$$

- ▷ Derivation is based on the assumption that all probability distributions involved are Gaussians.
- ▷ Gradient free.
- ▷ Extremely robust.

Ensemble Langevin methods (ELS)

The ELS is given by

$$dX^i{}_t = -C(X_t)\nabla V_\epsilon(X_t^i)\,dt + \nabla_{x^i}\cdot C(X_t)\,dt + \sqrt{2C(X_t)}dW_t^i.$$

Here $C : \mathbb{R}^{Nd} \to \mathbb{R}^{d\times d}$ denotes the empirical covariance function of arbitrary collection of N vectors $\{x^i\}_{i=1}^N$ in $\mathbb{R}^d$ and $X_t = \{X_t^i\}_{i=1}^N$.

- ▷ If G_ϵ is linear, the SDEs defining the ELS and EKS are the same.
- ▷ Performance deteriorates for noisy potentials as fluctuations dominate.

Gaussian process regression

Gaussian process:[a] A Gaussian process is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions.

A Gaussian process is specified by its mean function $m(x)$ and covariance function $k(x, x')$:

$$f \sim GP(m(x), k(x, x'))$$

Example of kernel/covariance functions: $k(x, y; \lambda, l) = \lambda e^{-\frac{\|x-y\|^2}{2l^2}}$, where $\lambda > 0$.

Gaussian process regression: Given noisy observations of the function f

$$y = f(x) + \sigma\xi,$$

where $\xi \sim \mathcal{N}(0, \sigma_n)$, then the Gaussian process posterior f^* is given by

$$f_*|X, Y, X_* \sim \mathcal{N}\Big(k(X_*, X)\left[k(X, X) + \sigma^2 I\right]^{-1} f,$$
$$k(X_*, X_*) - k(X_*, X)\left[k(X, X) + \sigma^2 I\right]^{-1} k(X, X_*)\Big)$$

where X is the matrix of training input $\{x_i\}_{i=1}^n$ and X_* the matrix of test inputs.

[a] Rasmussen & Williams, Gaussian Processes for Machine Learning, MIT Press 2006

Ensemble Gaussian sampler (EGS)

Assumption: data misfit term

$$V_L(x) = \frac{1}{2}\langle y - G(x), \Gamma^{-1}(y - G(x))\rangle \text{ is a Gaussian process.}$$

Given (noisy) evaluations of the potential at $X_t = (X_t^1, \ldots, X_t^N) \in \mathbb{R}^{N\times d}$ we seek a function f such that, for some $\sigma > 0$

$$V_L(X_t^i) = f(X_t^i) + \sigma\xi^i, \quad \xi = (\xi^1, \cdots, \xi^N) \sim \mathcal{N}(0, I).$$

The corresponding Gaussian process posterior for f has mean function

$$\widehat{V_L}(x; \sigma, \lambda, I) = \sum_{i,j=1}^{N} k(x, X_t^i; \lambda, I) K(X; \sigma, \lambda, I)_{ij}^{-1} V_L(X_t^j), \quad x \in \mathbb{R}^d$$

and covariance function

$$\gamma(x, y; \sigma, \lambda, I) = K(x, y; \sigma, \lambda, I) - \sum_{i,j=1}^{N} k(x, X_t^i; \lambda, I) K(X; \sigma, \lambda, I)_{ij}^{-1} k(X_t^j, y; \lambda, I).$$

Here $K(X)_{i,j} = \sigma^2\delta_{i,j} + k(X_t^i, X_t^j)$.

Ensemble Gaussian sampler

Gradient of the posterior mean

$$\nabla \widehat{V_L}(x;\sigma,\lambda,l) = \sum_{i,j=1}^{N} \nabla_x k(x, X_t^i \lambda, l) K(X;\sigma,\lambda,l)_{ij}^{-1} V_L(X_t^j).$$

Ensemble particles evolve according to over-damped Langevin dynamics

$$dX_t^i = -\nabla \widehat{V_L}(X_t^i;\sigma,\lambda,l)\,dt - \Sigma^{-1} X_t^i\,dt + \sqrt{2}\,dW_t.$$

Approximate gradient $\nabla \widehat{V_L}$ depends on the hyper-parameters (σ,λ,l), which have to be trained as the density evolves.

Updating the GP hyperparameters

Priors:

- ▷ log-normal priors on the amplitude λ and the noise's standard deviation σ, and
- ▷ Gamma prior on the lengthscale l.

Update (σ, λ, l) by maximising the log marginal posterior

$$MLP(\sigma, \lambda, l; X) \propto \frac{1}{2} \log \sum_{i,j=1}^{N} \widehat{V_L}(X_t^i; \sigma, \lambda, l) K(X; \sigma, \lambda, l)^{-1} \widehat{V_L}(X_t^j; \sigma, \lambda, l)$$
$$- \frac{1}{2} \log \det K(X; \sigma, \lambda, l) + \log p_0(\sigma, \lambda, l),$$

where p_0 denotes the prior density over the hyperparameters.

Putting it all together

Euler-Maruyama discretisation of the SDE coupled with a gradient descent scheme for adaptively selecting the hyperparameters.

Let $X_n = (X_n^1, \ldots, X_n^N) \in \mathbb{R}^{N \times d}$ denote the particle ensemble at time-step n. Then

- ▷ For $i = 1, \ldots, N$:
  - Set $X_{n+1}^i = X_n^i - \Delta t \nabla \widehat{V_L}(X_n^i; \sigma_n, \lambda_n, l_n) - \Delta t \Sigma^{-1} X_n^i + \sqrt{2\Delta t}\, \xi_n$, where $\xi \sim \mathcal{N}(0,1)$ iid.
- ▷ Update $(\sigma_{n+1}, \lambda_{n+1}, l_{n+1}) = (\sigma_n, \lambda_n, l_n) + \delta t \nabla_{(\sigma,\lambda,l)} MLP(\sigma_n, \lambda_n, l_n; X_{n+1})$.

Here Δt and δt are step-sizes for the Langevin updates and the hyperparameter gradient descent, respectively.

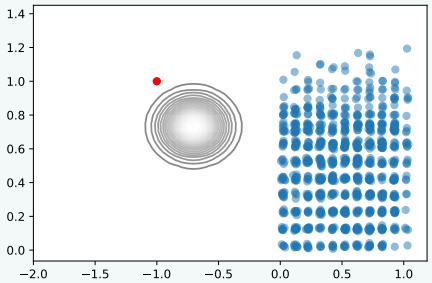
Ensemble methods

ELS

Gradient based

Calculate gradient of log-posterior for every particle and update the particle positions.

Performance deteriorates as $\varepsilon \to 0$.

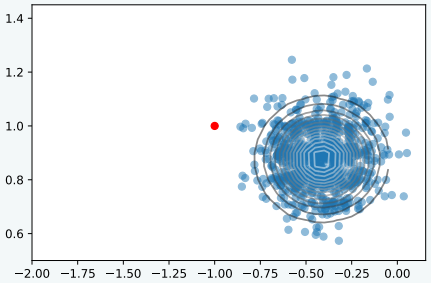


EKS

Gradient free

Approximate gradient of log-posterior under the assumption that all probabilities are Gaussians.

Robust to roughness of posterior landscape

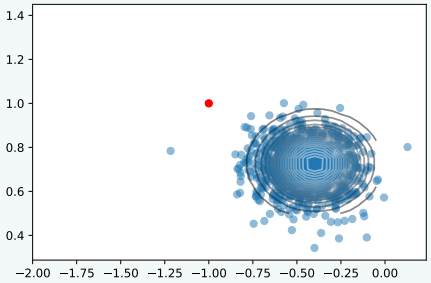


EGPS

Gradient free

Approximate gradient of the log-posterior and assume that it's a Gaussian process.

Robust to roughness of posterior landscape



Performance of the different sampling methods for rough posteriors

- Use formal multi-scale analysis to analyse the behaviour of the interacting particle methods in the case of rapid fluctuations, that is

$$\varepsilon \ll 1.$$

- Do the limiting solutions of the methods converge to the correct (unperturbed) equilibrium distribution as $t \to \infty$?

Recall: we wish to sample from π_0 NOT π_ϵ!

Performance of the different sampling methods for rough posteriors

- Use formal multi-scale analysis to analyse the behaviour of the interacting particle methods in the case of rapid fluctuations, that is

$$\varepsilon \ll 1.$$

- Do the limiting solutions of the methods converge to the correct (unperturbed) equilibrium distribution as $t \to \infty$?

Recall: we wish to sample from π_0 NOT π_ϵ!

Assumption:

The forward model satisfies

$$G_\epsilon(x) = G_0(x) + G_1(x/\epsilon),$$

$G_0 \in C^1(\mathbb{R}^d, \mathbb{R}^K)$, $G_1 \in C^1(\mathbb{T}^d, \mathbb{R}^K)$ and $\int_{\mathbb{T}^d} G_1(y)\, dy = 0$. G_1 is a 1-periodic function in every dimension.

EKS: mean-field limit

The mean field limit is given by

$$dx_t = -\mathcal{F}(x_t, \rho)\, dt - \mathcal{C}(\rho)\Sigma^{-1} x_t\, dt + \sqrt{2\mathcal{C}(\rho)}\, dW_t,$$

where W is a standard Brownian motion in $\mathbb{R}^d$ and,

$$\overline{\mathcal{X}}(\pi) = \int_{\mathbb{R}^d} X'\pi(X')dX', \quad \overline{\mathcal{G}}(\pi) = \int_{\mathbb{R}^d} G_\epsilon(X')\pi(X')dX',$$

$$\mathcal{C}(\pi) = \int_{\mathbb{R}^d} \left(X' - \overline{\mathcal{X}}(\pi)\right) \otimes \left(X' - \overline{\mathcal{X}}(\pi)\right) \pi(X')dX',$$

$$\mathcal{F}(x, \pi) = \Big(\int_{\mathbb{R}^d} \langle G_\epsilon(X') - \overline{\mathcal{G}}(\pi), G_\epsilon(x) - y\rangle_\Gamma X'\pi(X')dX'\Big).$$

The time-dependent density of the process ρ satisfies the nonlinear Fokker-Planck equation

$$\partial_t \rho = \nabla_x \cdot \Big(\nabla_x \cdot \left(\mathcal{C}(\rho)\rho\right) + \mathcal{F}(x, \rho)\rho\Big). \tag{EKS}$$

EKS: formal multi-scale analysis

Mean field limit equations of the unperturbed problem:

$$dx_t = -\mathcal{F}_0(x_t, \rho_0)\, dt - \mathcal{C}(\rho_0)\Sigma^{-1} x_t\, dt + \sqrt{2\mathcal{C}(\rho_0)}\, dW_t,$$

with

$$\overline{\mathcal{G}}_0(\pi) = \int_{\mathbb{R}^d} G_0(X')\pi(X')dX',$$
$$\mathcal{F}_0(x,\pi) = \int_{\mathbb{R}^d} \langle G_0(X') - \overline{\mathcal{G}_0}(\pi), G_0(x) - y\rangle_\Gamma X'\pi(X')dX'.$$

The time dependent density $\rho_0(x,t) \in C\big((0,\infty); L^1(\mathbb{R}^d;\mathbb{R}^+)\big)$ of this process satisfies

$$\partial_t \rho_0 = \nabla_x \cdot \Big(\nabla_x \cdot \big(\mathcal{C}(\rho_0)\rho_0\big) + \mathcal{F}_0(x,\rho_0)\rho_0\Big). \qquad (\mathrm{EKS}_0)$$

EKS: Performance as $\varepsilon \to 0$

Let Assumption (A1) hold with $0 < \epsilon \ll 1$. If ρ satisfying (EKS) is of the form

$$\rho = \rho_0 + \epsilon\rho_1 + \epsilon^2\rho_2 + \cdots ,$$

then formal multi-scale analysis demonstrates that ρ_0 satisfies (EKS_0) (FPE relating to the unperturbed process).

- ▷ As $\varepsilon \to 0$ the limiting mean field PDE for the density of the process (EKS_0) corresponds to the nonlinear FPE of the unperturbed process.

  $\Rightarrow$ EKS behaves as if $G_1 \equiv 0$, and ignores the rapid $\mathcal{O}(1)$ fluctuations.

- ▷ Formal perturbation result confirms empirically observed robustness of the EKS for very noisy problems.

- ▷ Rigorous results: tedious, since the main technical difficulty would be to derive bounds from below for the covariance operator.

ELS: mean field limit

The mean field limit of the ELS is given by

$$dx_t = -\mathcal{C}(\rho_t)\nabla V_\epsilon(x_t) + \sqrt{2\mathcal{C}(\rho)}dW_t,$$

where function $\mathcal{C}(\cdot)$ on densities is defined as for the mean-field equation for the EKS.

The associated non-linear Fokker-Planck equation for the time-dependent density of the process $\rho \in C((0,\infty); L^1(\mathbb{R}^d; \mathbb{R}^+))$ is given by

$$\partial_t \rho = \nabla_x \cdot \big(\mathcal{C}(\rho)\,(\nabla_x \rho + \nabla_x V_\epsilon \rho)\,\big). \tag{ELS}$$

- ▷ Carrillo and Vaes established stability estimates in the Wasserstein distance for solutions in the case of linear G.

ELS: formal multi-scale analysis

If the solution ρ to (ELS) is expanded in the form $\rho = \rho_0 + \epsilon\rho_1 + \epsilon^2\rho_2 + \cdots$, then the formal multi-scale analysis demonstrates that ρ_0 satisfies

$$\partial_t \rho_0 = \nabla_x \cdot \Big(\mathcal{D}(\rho_0) \Big(\nabla_x \rho_0 + \nabla_x \overline{V} \rho_0 \Big) \Big), \tag{ELS_0}$$

where $\overline{V} = V_0 - \log Z(x)$, $Z(x) = \int_{\mathbb{T}^d} e^{-V_1(x,z)}\, dz$ and

$$\mathcal{D}(\rho_0) = \frac{1}{Z(x)} \int_{\mathbb{T}^d} \big(I + \nabla_z \chi(x,z)\big)^\top \mathcal{C}(\rho_0) \big(I + \nabla_z \chi(x,z)\big) e^{-V}\, dz.$$

Here $\chi : \mathbb{R}^d \times \mathbb{T}^d \to \mathbb{R}^d$ solves the following second order PDE in z (parameterized by x):

$$\nabla_z \cdot \Big(\mathcal{C}(\rho_0) e^{-V(x)} (\nabla_z \chi(x,z) + I) \Big) = 0, \quad (x,z) \in \mathbb{R}^d \times \mathbb{T}^d.$$

Furthermore, for arbitrary $\zeta \in \mathbb{R}^d$,

$$\zeta^\top \mathcal{D}(\rho_0) \zeta \leq \zeta^\top \mathcal{C}(\rho_0) \zeta.$$

ELS: performance as $\varepsilon \to 0$

- For $\varepsilon \to 0$ the function ρ_0 satisfying (ELS_0) the unique invariant distribution is given by

$$\bar{\pi}(x) \propto \pi_0 Z(x)$$

No 'averaging out' of fluctuations

- Perturbations slow down convergence: the effective diffusion $\mathcal{D}(\rho_0, x)$ is given by

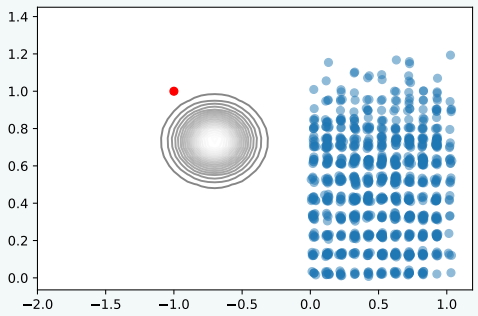
$$\begin{aligned}\mathcal{D}(\rho_0, x) &= \frac{1}{Z(x)} \int_{\mathbb{T}^d} \mathcal{C}(\rho_0) e^{-V_1} (I + \nabla_z \chi)\, dz \\ &= \mathcal{C}(\rho_0) - \int_{\mathbb{T}^d} \nabla_z \chi^\top \mathcal{C}(\rho_0) \nabla_z \chi e^{-V_1(x,z)}\, dz.\end{aligned}$$

Linear forward model

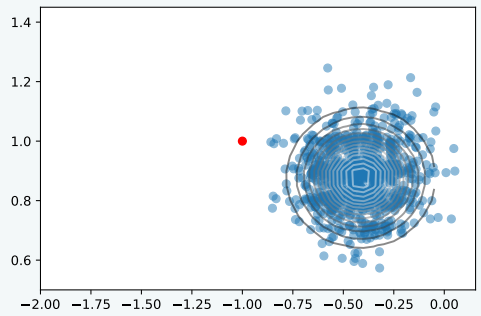
Forward map G_ϵ of the form, for $x = (x_1, x_2)$,

$$G_\epsilon(x) = G_0(x) + G_1(x/\epsilon),$$

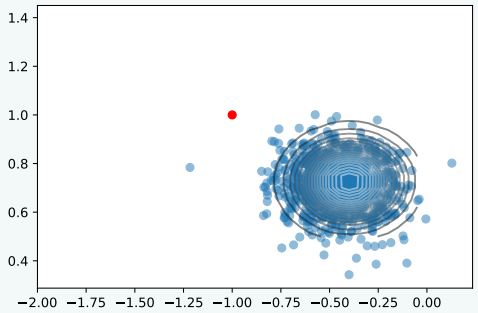
$$G_0(x) = Ax, \quad G_1(x) = \left[\sin(2\pi x_1), \sin(2\pi x_2)\right]^\top, \quad \text{with } A = \begin{pmatrix} -1 & 0 \\ 0 & 2 \end{pmatrix}.$$



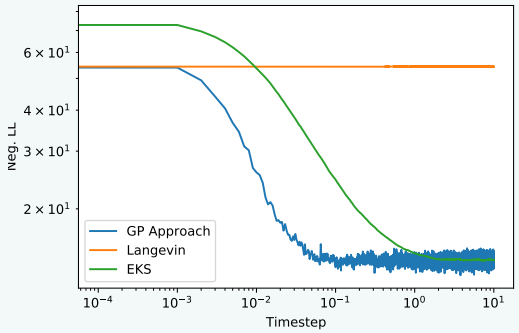
(a) Ensemble Langevin Samples



(b) Ensemble Kalman Sampler



(c) Ensemble GP sampler



(d) Negative Log-Likelihood

Lorenz 63 equations

Consider the 3-dimensional Lorenz 63 equations:

$$\begin{aligned}\dot{x}_1 &= \sigma(x_2 - x_1)\\ \dot{x}_2 &= rx_1 - x_2 - x_1x_3\\ \dot{x}_3 &= x_1x_2 - bx_3,\end{aligned}$$

with parameters σ, r, $b \in \mathbb{R}_+$.

- ▷ Fix $\sigma = 10$ and focus on the inverse problem of identifying r and b from time-averaged data.
- ▷ Impose multivariate log-normal prior on $\theta = (r, b)$ with mean $m = (3.3, 1.2)$ and covariance $\Sigma = \mathrm{diag}(0.15^2, 0.5^2)$.
- ▷ We take $T = 10$ and define $\varphi\colon \mathbb{R}^3 \to \mathbb{R}^9$

$$\varphi(x) = (x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3);$$

Data generation

- Data is generated for $(\sigma, r^\dagger, b^\dagger) = (10, 28, \frac{8}{3})$ (chaotic behaviour) using a single evaluation of the random (with respect to initial condition) function $\mathcal{G}_\epsilon$. Furthermore $\Delta_{obs} \equiv 0$.
- We set Δ_{model} to be the empirical covariance of $\mathcal{G}_\epsilon(\theta^\dagger)$ over windows of size 10 from a single trajectory with $\theta^\dagger = (r^\dagger, b^\dagger)$, over 360 time units.
- Negative LL function $V_L(\theta) := \frac{1}{2}\langle (y - \mathcal{G}_\epsilon(\theta)), \Delta_{model}^{-1}(y - \mathcal{G}_\epsilon(\theta))\rangle$.

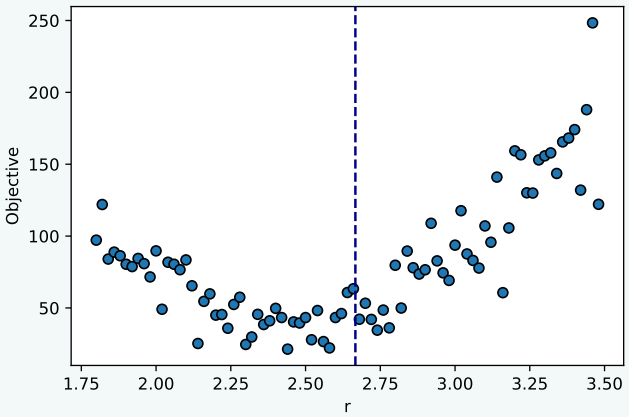
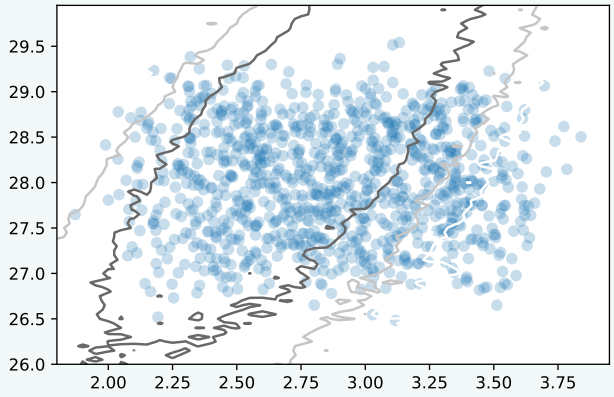
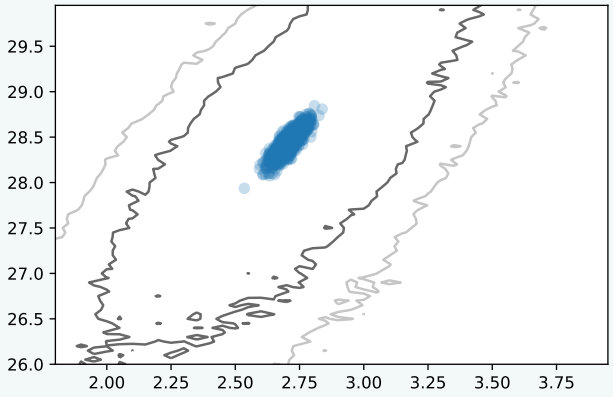


Figure: Profile of the noisy negative log-likelihood over r for b fixed at optimal value. The blue dashed line denotes the 'true' value $r = 8/3$.

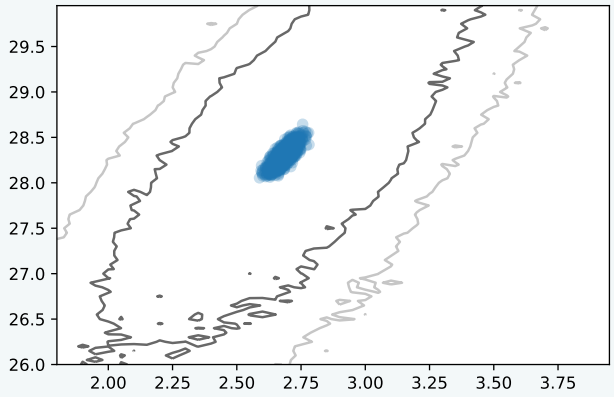
Lorenz 63



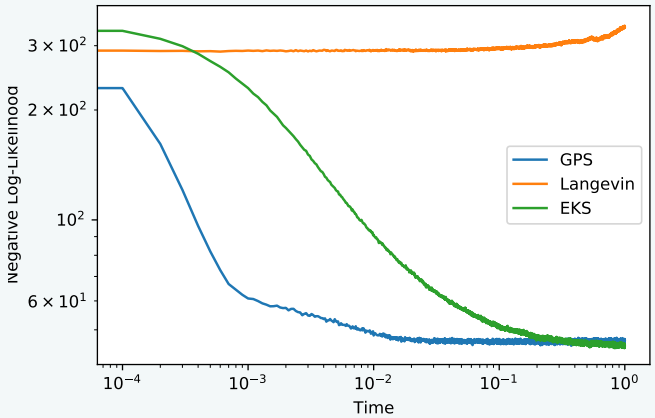
(a) ELS



(b) EKS



(c) EGPS



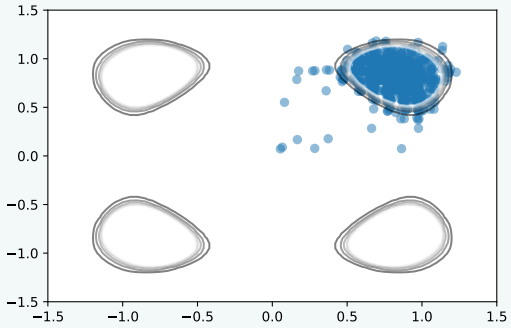
(d) Negative Log-Likelihood

Multi-modal posteriors

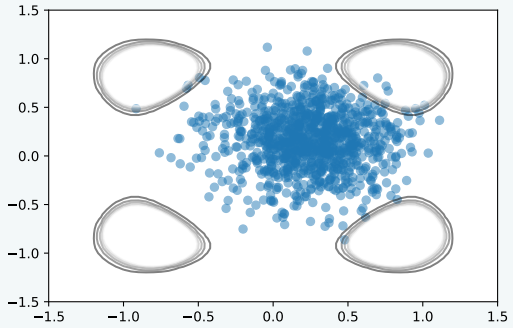
We consider a forward map for $x = (x_1, x_2)$ which is defined by

$$G_\epsilon(x) = G_0(x) + G_1(x/\epsilon),$$
$$G_0(x) = (x_1^2 - 1)^2 + (x_2^2 - 1)^2, \quad G_1(x) = \nu(\sin(2\pi x_1) + \sin(2\pi x_2)),$$

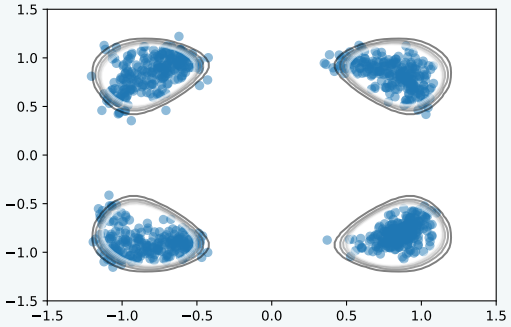
and where $\Gamma = \gamma^2 I$.



(e) ELS



(f) EKS



(g) EGPS

Conclusion & Outlook

- Discussed and analysed different ensemble methods for solving inverse problems with noisy and expensive likelihoods.
- Used formal multi-scale approach to understand the influence of rapid fluctuations, when trying to identify the large-scale smoothly varying underlying structure of the posterior.
- EKS is robust with respect to noisy and periodic fluctuations, while the ELS is significantly impacted by it.
- Propose a new class of ensemble Gaussian process samplers, which are robust to fluctuations but still employ gradient information.
- Open problems:
  - Understanding the tuning of the hyper-parameters for the Gaussian process;
  - Making the formal multi-scale argument rigorous;
  - Long time behaviour of the EGPS;
  - ....

Conclusion & Outlook

- Discussed and analysed different ensemble methods for solving inverse problems with noisy and expensive likelihoods.
- Used formal multi-scale approach to understand the influence of rapid fluctuations, when trying to identify the large-scale smoothly varying underlying structure of the posterior.
- EKS is robust with respect to noisy and periodic fluctuations, while the ELS is significantly impacted by it.
- Propose a new class of ensemble Gaussian process samplers, which are robust to fluctuations but still employ gradient information.
- Open problems:
  - Understanding the tuning of the hyper-parameters for the Gaussian process;
  - Making the formal multi-scale argument rigorous;
  - Long time behaviour of the EGPS;
  - ....

Thank you very much for your attention!

References:

- O.R.A. Dunbar, A.B. Duncan, A.M. Stuart and MTW, *Ensemble Inference Methods for Models with Noisy and Expensive Likelihoods*, to appear in SIADS, 2022
- J.A. Carrillo and U. Vaes, *Wasserstein stability estimates for covariance Fokker-Planck equations*, Nonlinearity, 2020
- A. Garbuno-Inigo, F. Hoffmann, W. Li and A.M. Stuart, *Interacting Langevin diffusions: Gradient Structure and ensemble Kalman sampler*, SIADS 19, 2020
- V. Araujo, I. Melbourne and P. Varandas, *Rapid mixing for Lorenz attractor and statistical limit laws for their time-1 maps*, Communications in Mathematical Physics 340, 2015.
- C.E. Rasmussen and C.K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006