Infinite-dimensional Bayesian inference for time evolution PDEs

Based on lectures given by Richard Nickl at Warwick in July 2025

1 Introduction and Setting

1.1 Non-linear dynamics

We work on the torus $\Omega = (0,1]^d$ and define $\Delta = \sum_{i=1}^d \partial_i^2$ and $L^p(\Omega)$ and the periodic Sobolev Hilbert spaces $H^\alpha(\Omega)$ for $\alpha \geq 0$ as usual. The negative Sobolev Hilbert spaces are defined to be the duals $H^{-\alpha}(\Omega) = (H^\alpha(\Omega))'$. Define also

$$L_0^2 := L^2 \cap \{h : \int_\Omega h dx = 0\}$$

a space on which Δ has an inverse, or in the context of the incompressible Navier-Stokes equations (NSE)

$$L_0^2 := L^2 \cap \{h : \int_\Omega h dx = 0 \text{ and } \nabla \cdot h = 0\}$$

which restricts us to consider only incompressible functions. The spaces H_0^α are defined similarly. Lastly, $P : L^2 \to L_0^2$ is the orthogonal projection.

Consider $(u_\theta(t,x) : t \in [0,T], x \in \Omega)$ which solves a particular PDE with initial condition $\theta(x) = u(0,x)$. We study two relevant examples:

$$\frac{\partial u}{\partial t} - \Delta u - f \circ u = 0 \tag{reaction-diffusion}$$

on $[0,T] \times \Omega$, where $f : \mathbb{R} \to \mathbb{R}$ is known and for simplificty in $C_c^\infty(\mathbb{R})$, although can be roughly relaxed to $C_c(\mathbb{R})$. Another more complicated example is

$$\frac{\partial u}{\partial t} - \nu P \Delta u + B(u,u) = 0 \tag{NSE}$$

where $u : \mathbb{R}^2 \to \mathbb{R}^2$,

$$B(u,v) := P(u \cdot \nabla)v$$

$$[(u \cdot \nabla)v]_i = \sum_{j=1}^2 u_j \frac{\partial v_i}{x_j}.$$

Given observations of u_θ throughout $[0,T] \times \Omega$ (to be described in detail shortly), we seek to learn the initial condition θ. Typically we assume $\theta \in H_0^1$.

The following result is standard PDE theory:

Proposition 1.1. Let $\theta \in H_0^1, T > 0, f \in C_c^\infty(\mathbb{R})$ then there is a unique solution $u_\theta \in C([0,T], L^2(\Omega))$ to reaction-diffusion and the NSE.

Thus, given θ, all subsequent values of u_θ are determined.

Remark 1.2. For the NSE, $\theta \in L_0^2$ implies $u_\theta(t,\cdot) \in L_0^2$ for all $t > 0$, i.e. mean-zero and incompressibility are preserved. This isn't the case for reaction-diffusion. Nevertheless, this requirement still be useful statistically going forwards.

1.2 Discrete observations

Consider $(Y_i, t_i, X_i)_{i=1}^N$ which follow a regression model

$$Y_i = u_\theta(t_i, X_i) + \epsilon_i \tag{1.1}$$

We will assume

(i) $\epsilon_i \sim_{iid} N(0, 1)$

(ii) $(t_i, X_i) \sim_{iid} \text{Unif}([0, T] \times \Omega)$

This is simplistic, one expects the noise ϵ_i to depend on (t_i, X_i), but we need to make some decision here to get an initial understanding. The uniform assumption is called probabilistic numerics, and justifications are given in Dioconis (1988). We denote $Z^{(N)} = (Y_i, t_i, X_i)_{i=1}^N$.

This type of problem is called 'data assimilation', where we are given $Z^{(N)}$ and typically wish to predict u_θ at later times or different points. It turns out the best way to do this is to recover the initial condition, which, by Proposition 1.1, allows us to go forwards again. See [1] which treats even the case where the PDE is misspecified, i.e. the data doesn't exactly fit the PDE, although we study the correctly specified case here.

Statistically, this is just a non-parametric regression problem, but we know u_θ solves a particular PDE. Classical frequentist statistical methods construct an estimate $\hat{u}$ from this data, and it is not clear how to use the fact u_θ actually solves this PDE, so while $\hat{u}$ may estimate u_θ well near the observations, it will estimate poorly for larger times. Moreover, the estimate $\hat{u}$ will have zero probability of actually solving the PDE[1], whereas this is something our Bayesian approach will get for free. The joint distribution of $Z^{(N)}$ has probability density

$$P_\theta^N = \prod_{i=1}^N p_\theta(Y_i, t_i, X_i) \tag{1.2}$$

where $p_\theta(y, t, x) \propto \exp(-\frac{1}{2}|y - u_\theta(t, x)|^2)$. The log-likelihood is

$$l_N(\theta) = -\frac{1}{2}\sum_{i=1}^N |Y_i - u_\theta(t_i, X_i)|^2, \quad \theta \in \Theta \tag{1.3}$$

for some parameter space $\Theta \subset H_0^1$. The textbook approach to statistics would try to maximise this over θ. There are several issues with this. For example, the parameter space is infinite dimensional, so we'd probably need to regularise by adding a penalty term. The more fundamental problem is that the map $\theta \to u_\theta$ is the map that sends the initial condition to the solution of a non-linear PDE, hence is non-linear and $-l_N(\theta)$ is not convex, so optimisation is very difficult.

One of the reasons people opt for Bayesian methods is because in Bayesian methods one computes averages (e.g. the posterior mean) instead of optimising a function and this may be more robust to such problems of non-convexity (though to actually prove it is more robust is quite a different story). In principle, given a convex surface, one expects computing its integral may be easier than computing its maximum, and so in order to do this we want to reweight the likelihood and compute the average instead of the mode. To get started, we need to define a prior.

[1] Or prove otherwise...

1.3 Gaussian process priors

The main part of this introduction is to take away some of the fear away from Bayesian statistics, that priors are not anything dangerous. On one hand, they serve the role of regularisation (having a penalty in the likelihood). On the other hand they correspond to a randomisation scheme in our algorithm. The prior is not a subjective belief about whether it will rain tomorrow, rather a tool to introduce randomisation in the class of algorithms we will use. Nevertheless, we must declare what types of priors we will use. Since the influential work of Andrew Stuart (or even Kalman), the use of Gaussian priors has been advocated, for computational reasons. Since we are modelling a parameter in a function space, we need to use Gaussian random fields.

For θ we consider prior Gaussian random fields $(\theta(x) : x \in \Omega)$ over L^2_0 of the form

$$\theta \sim \Pi_\gamma \sim N(0, \rho^2(-\Delta)^{-\gamma})^2$$

where we have used the fact $-\Delta$ is invertible on L^2_0 and in fact positive definite so that fractional powers make sense, and ρ is some positive scaling, and γ models the smoothness of trajectories. We assume $\gamma > 1 + d/2$, so that by the usual Hilbert-Schmidt embedding argument we have $\theta \in H^1$ a.s, or in fact H^β_0 for all $0 < \beta < \gamma - d/2$. These priors have full support in H^1_0 so we will not rule out any particular choices for θ.

If (e_j, λ_j) are the eigenfunctions of the Laplacian so that $\Delta e_j = -\lambda_j e_j$ then one has the series representation

$$\theta = \rho \sum_{i=1}^{\infty} \lambda_j^{-\gamma} g_j e_j \quad L^2\text{-a.s.} \tag{1.4}$$

where $g_j \sim N(0, 1)$. Note that on the torus, the e_j are just the exponentials. We also let $\theta' = \theta/\rho$ be the unscaled parameter.

Remark 1.3. We don't use them for any particular reason other than that they can be numerically implemented and have full support in H^1_0 in a way that we can also quantify which is important in the proofs and we will see later. The usage of these priors has been common in the literature for some time, in machine learning, inverse problems, data assimilation, so this nice family of Gaussian fields have been part of the 'standard toolkit'.

In data assimilation, it is not just important to assign a prior to the initial condition, but also to the regression function (in our case u_θ), which is what the statistician actually needs. We can obtain such a prior by pushing forward θ to u_θ, i.e.

$$\Pi \circ u_\theta^{-1} \text{ in } C([0, T], L^2(\Omega))$$

where u_θ^{-1} denotes the image measure. Because the PDEs we are considering are numerically solvable, this gives us an implementable prior in the space of all trajectories whose draws solve the PDE in question with probability one. However, it is not a Gaussian prior any longer because of the non-linearity, so the main challenge of the statistician will be to deal with the regression model where the prior is not Gaussian, rather the push-forward of a Gaussian through the solution map of a non-linear PDE.

Remark 1.4. This is one of the main reasons why people choose to model the prior at the initial condition because it is not clear how otherwise how to pick a random function that at a given time

[2] Or $-P\Delta$ for the NSE

t solves the NSE. A good problem for a probabilist: give me a prior which describes at time t solves the NSE for some initial condition, without first modelling the initial condition. For the probabilist, modelling the initial condition is fine, but for the statistician by delcaring this to be our statistical parameter, we inherit an inverse problem which is hard (recovering the initial condition θ from observations of u_θ which solve a PDE). We will be able to do it, though it would be much more convenient if we did not have to go back to the initial condition and then solve forward again, but this is how all the algorithms people have ever used have gone. It would be very attractive to have a different projection onto the solution manifold, but this seems computationally hard, so people do this.

1.4 Posterior measure

The extension of Bayes theorem to infinite dimensions is not difficult, but we spell it out here. Suppose the map $(\theta, t, x) \to u_\theta(t, x)$ is jointly measurable for some σ-field over the product space $\Theta \times [0, T] \times \Omega$. Then $P_\theta^N(y, t, x)$ is jointly measurable and if we define the density on $\Theta \times \mathbb{R} \times [0, T] \times \Omega$ by

$$dQ(\theta, y, t, x) = p_\theta(y, t, x) dydtdxd\Pi_\gamma(\theta) \tag{1.5}$$

so by the standard rules for expressions of conditional densities in a product space, we have

$$Z^{(N)}|\theta \sim P_\theta^N \tag{1.6}$$

and the posterior distribution is

$$\begin{aligned} \theta|Z^{(N)} &\sim \Pi(\theta|Z^{(N)}) \\ &\sim \frac{\prod_{i=1}^N p_\theta(Y_i, t_i, X_i) d\Pi_\gamma(\theta)}{\int_\Theta \prod_{i=1}^N p_\theta(Y_i, t_i, X_i) d\Pi_\gamma(\theta)} && (1.7) \\ &\propto e^{l_N(\theta)} d\Pi_\gamma(\theta) && \text{(Gibbs measure form)} \end{aligned}$$

and we have the push-forward posterior

$$u_\theta|Z^{(N)} \sim \text{Law}(u_\theta : \theta \sim \Pi(\theta|Z^{(N)})) \tag{1.8}$$

whose time marginal distributions we denote by $(\widehat{\Pi}_t : t \geq 0)$. For $t^* = \max_{i \leq N} t_i$, we call $\widehat{\Pi}_{t^*}$ the *filtering distribution* because it is the last thing a sequential algorithm like the Kalman filter would target.

The posterior $\Pi(\theta|Z^{(N)})$ is a Gibbs measure with potential l_N which is not concave and has very complicated fluctuations in principle due to the non-linearity, so is certainly not Gaussian on the level of θ. Then, we solve another non-linearity forward, and we will see later that the resulting posterior distributions $\widehat{\Pi}_t$ actually converge to a Gaussian process as $N \to \infty$. In fact, the $\widehat{\Pi}_t$ are much more regular than one may expect and therefore also computable in other ways.

1.5 Posterior computation

Two questions we wish to answer are: how should I compute the posterior? and should I compute the posterior? We have not shown yet that the posterior has any useful statistical guarantees. We will assume a fixed ground truth θ_0 generated the data and show somehow that the output from our algorithm is close to the ground truth dynamical system that we have actually observed, in some distance, and moreover find a convergence rate. Before we do that, we answer the first question.

The algorithms to follow came out fifteen years ago by Andrew Stuart, Gareth Roberts, Martin Hairer, among others, which can at least in principle reliably compute the posterior. The first thing one could try to do is find the θ which maximises the posterior density $\theta \to \Pi(\theta)$ (called *maximum a-priori (MAP) estimates*), which we recall from the Gibbs measure form is somewhat equivalent to minimising the non-convex function l_N. More precisely, one can show the effect of the prior is adding a penalty term, i.e. formally MAP estimates maximise

$$l_N(\theta) + \exp\log d\Pi_\gamma(\theta) \text{ ``$=$'' } -\frac{1}{2}\sum_{i=1}^{N} |Y_i - u_\theta(t_i, X_i)|^2 - \rho^2 \|\theta\|^2_{H_0^\gamma} \tag{1.9}$$

This is what statisticians called penalised least squares or people in inverse problems call a Tychonov regulariser: the sum of least squares is kept small but also the solution cannot be too rough. The point of our discussion is that after one has survived the Bayesian setup (choosing a prior etc), we get something very intuitive. But we have solved nothing becasue the problem is still not convex and so we still cannot compute it.

Instead, let us compute the posterior mean $\tilde{\theta}_N := \mathbb{E}^\Pi(\theta|Z^{(N)})$. Given $\tilde{\theta}_N$, we can estimate the regression function by finding the plug-in $u_{\tilde{\theta}_N}$ which solves the PDE with initial condition $\tilde{\theta}_N$, and for the two examples we have in mind this is well understood numerically. From a Bayesian point of view, one should consider the posterior mean and not the mode. If you go to a decision theory class in statistics, they will tell you the quantity that minimises the Bayes risk is the posterior mean and not the mode, so as a statistician one should go for that anyway and later we can prove it is a good estimate. Computing an integral in infinite dimensions is not necessarily easy, one can ask whether polynomial time algorithms even exist. Computing the mean of the posterior from its Bayes rule definition (1.7) is not something we want to do, it is not even clear how to compute the the normalising constant in the denominator. Instead the big idea is to use Markov Chain Monte Carlo methods: we come up with a Markov chain v_k with invariant measure $\Pi(\cdot|Z^{(N)})$ and use $\frac{1}{K}\sum_{k=1}^{K} v_k$ to estimate $\tilde{\theta}_N$. Given the output of such an algorithm, one can forget the whole Bayesian story and try to prove that the plug in $u_{\frac{1}{K}\sum_{k=1}^{K} v_k}$ converges to the ground truth dynamical system.

Example 1.5. (pCN algorithm) Initialise v_0, step size $\delta > 0$, and compute the proposals as follows:

(i) Given $v_0, \ldots, v_{k-1}$, let $p_k = \sqrt{1-2\delta}v_{k-1} + \sqrt{2\delta}\xi$ where $\xi \sim \Pi_\gamma$.

(ii)
$$v_k = \begin{cases} p_k & \text{with probability } \min(1, e^{l_N(p_k) - l_N(p_{k-1})}) \\ v_{k-1} & \text{otherwise} \end{cases} \tag{1.10}$$

(iii) Estimate $\tilde{\theta}_N$ by $\frac{1}{K}\sum_{k=1}^{K} v_k$ with K large.

We always accept the proposal p_k if $l_N(p_k) > l_N(v_{k-1})$, but even if $l_N(p_k) < l_N(v_{k-1})$ we have some probability of accepting. This is necessary for the Markov chain to explore the whole space (c.f. Metropolis-Hastings). When this is actually implemented on a computer, one usually takes a large finite dimensional subspace $R^D \subset \Theta$.

Proposition 1.6. v_k from the pCN algorithm has invariant measure $\Pi(\cdot|Z^{(N)})$ for Gaussian priors

Proof. This is a straightforward calculation that we wont do.

□

Example 1.7. (Unadjusted/Metropolis-adjusted Langevin) For $\Theta = \mathbb{R}^D$ with D large, initialise v_0 and compute the iterates

$$v_{k+1} = v_k - \delta\nabla \log \Pi(v_k|Z^{(N)}) + \delta\xi_k \tag{1.11}$$

where $\xi_k \sim N(0,1)$.

One shows this also has invariant measure $\Pi(\cdot|Z^{(N)})$, even for non-Gaussian priors. The iterates resemble gradient descent, except we keep δ fixed not tending to zero so the Markov chain explores the whole space.

Remark 1.8. This approach is popular for non-linear problems because it doesn't involve any optimisation. In summary, nowadays we don't think of Bayesian statistics as some school of philosophy, we think of it as a class of algorithms that instead of optimisation use averaging.

2 Posterior consistency for data assimilation

So far we have laid out our algorithm, and we return to the question of whether it actually works. Suppose θ_0 is the ground truth initial condtion. Can we prove $u_{\tilde{\theta}}$ or $u_{\frac{1}{k}\sum_{k=1}^{K} v_k}$ is close to u_{θ_0} under the law $P_{\theta_0}^N$? We quantify this as follows: We seek to prove (suppressing γ),

$$\Pi(\theta \in \Theta : \|\theta - \theta_0\|_{L^2} > M\delta_N|Z^{(N)}) \to 0 \text{ in } P_{\theta_0}^{\mathbb{N}} \text{ probability} \tag{2.1}$$

This sequence of probabilities is random because $Z^{(N)}$ is random. Here, δ_N is the *contraction rate* and tends to zero. M is some constant. The L^2 norm could be replaced by some other norm or metric.

Remark 2.1. The rate of convergence of these probabilities to zero is not important in the theory so far. δ_N cannot be asymptotically smaller than the parametric rate $N^{-1/2}$, but we will need it to be sufficiently small for later results.

2.1 Hellinger distance

Define the *Hellinger distance*

$$h(p_\theta, p_{\theta_0})^2 = \int (\sqrt{p_\theta} - \sqrt{p_{\theta_0}})^2 dydtdx \tag{2.2}$$

where p_θ and p_{θ_0} are the Gaussian densities from before.

Remark 2.2. This definition is specific to our statistical model, but it can be extended to other models similarly and some of the results we give in this section remain true.

For reasons we will discuss shortly, the Hellinger distance is the natural distance in which one would first prove a posterior contraction result. In our regression model, one shows the Hellinger distance is equivalent to the Bochner L_T^2 norm:

Lemma 2.3. Suppose $\Theta_0 \subset \Theta$ containing θ_0 such that for some upper bound U,

$$\sup_{\theta\in\Theta_0} \sup_{0<t<T} \|u_\theta(t)\|_\infty \leq U < \infty \tag{2.3}$$

Then there exists a constant C_U such that for all $\theta \in \Theta_0$

$$C_U \|u_\theta - u_{\theta_0}\|_{L^2_T} \le h(p_\theta, p_{\theta_0})^2 \le \frac{1}{4}\|u_\theta - u_{\theta_0}\|^2_{L^2_T} \tag{2.4}$$

where

$$\|H\|^2_{L^2_T} = \int_0^T \|H(t,\cdot)\|_{L^2(\Omega)} dt \tag{2.5}$$

and $C_U = (1 - e^{-U^2/2})/2U^2$.

Proof. Proposition 1.3.1 in [2]

□

Remark 2.4. The assumption (2.3) is not unexpected, one can only expect to do regression when the regression functions are not completely wild. If the regression function can have lots of poles everywhere and we only have discrete measurements, there is little hope we will be able to reconstruct anything from it, and so we need the assumption to hold on an appropriate subset.

Remark 2.5. This result is specific to our regression model and perhaps one of the reasons why it is a good idea to use probabilistic numerics in the way we have set it up. It is convenient in the proofs to follow that the natural statistical/information theoretic distance for results we are trying to prove is equivalent to the Bochner L^2_T norm which not only has a very clear interpretation but shows up naturally in PDE analysis. If one chooses a different measurement model, life becomes significantly more complicated here, this is one of the things we harvest from our uniform sampling regime.

We will show that if one is able to construct a statistical test/decision rule $\psi_N = \psi_N(Z^{(N)})$ (i.e. an indicator function 1_{A_N}) that distinguishes between θ_0 and θ, then posterior contraction holds. The fundamental result from non-parametric statistcs we will use due to Le Cam is that it is possible to construct such a test using the Hellinger distance in complete universitality without any regularity assumptions other than on the metric complexity of the space of densities:

Proposition 2.6. There exists a test $\psi_N = \psi_N(Z^{(N)}) = 1_{A_N}$ such that

$$\underbrace{\mathbb{E}_{\theta_0}\psi_N}_{\text{Type 1 error}} + \sup_{\theta\in\Theta_N} \underbrace{\mathbb{E}_\theta(1-\psi_N)}_{\text{Type 2 error}} \le e^{-cM^2N\delta_N^2} \tag{2.6}$$

where c is some constant, Θ_N is a suitable alternative hypothesis space of which we will not go into the details but is infinite dimensional, provided $\forall\epsilon > 0$

$$\log N(\Theta_N, h, \epsilon) \le N\epsilon^2 \tag{2.7}$$

where $N(\Theta_N, h, \epsilon)$ is the number of radius ϵ Hellinger balls required to cover Θ_N.

Proof. See Section 7.1 in [3]. □

Remark 2.7. The metric entropy condition (2.7) measures the compactness of a space.

Remark 2.8. In words, the type 1 and type 2 errors are exponentially small uniformly in the alternative Θ_N. δ_N decays slower than $N^{-1/2}$, so $N\delta_N^2 \to \infty$.

We now sketch the idea of posterior contraction. From (1.7),

$$d\Pi(\theta|Z^{(N)}) = \frac{e^{l_N(\theta)}d\Pi(\theta)}{\int_\Theta e^{l_N(\theta)}d\Pi(\theta)} \tag{2.8}$$

To bound posterior probabilities, we need to control the normalising constant in the denominator. More precisely, we need it to not be too close to zero. One shows that with high probability,

$$\int_\Theta e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta) \geq \frac{e^{-aN\delta_N^2}}{\Pi(\|u_\theta - u_{\theta_0}\|_{L_T^2} < \delta_N)} \tag{2.9}$$

see Lemma 1.3.3 in [2] for details. Note that $\delta_N \gtrsim N^{-1/2}$ so $N\delta_N^2 \to \infty$ so the RHS numerator converges to zero, so for this inequality to be meaningful we need $\Pi(\|u_\theta - u_{\theta_0}\|_{L_T^2} < \delta_N)$ to also tend to zero exponentially fast. This is a statement about how 'good' the prior is, which we will discuss in the next section.

Hence, letting $\overline{\Theta} = \{\theta \in \Theta_N : \|\theta - \theta_0\|_{L^2} < M\delta_N\} \subset \Theta$ for convenience, for any b we have

$$\begin{aligned} P_{\theta_0}^N(\Pi(\overline{\Theta}|Z^{(N)}) \geq e^{-bN\delta_N^2}) &= P_{\theta_0}^N\left(\frac{\int_{\overline{\Theta}^c} e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta)}{\int_\Theta e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta)} \geq e^{-bN\delta_N^2}\right) \\ &= P_{\theta_0}^N\left(\int_{\overline{\Theta}^c} e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta) \geq e^{-(b+a)N\delta_N^2}\right) + o(1) \end{aligned}$$

Applying Markov's inequality,

$$\begin{aligned} e^{(b+a)N\delta_N^2}\mathbb{E}_{\theta_0}\int_{\overline{\Theta}^c} e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta) &= e^{(b+a)N\delta_N^2}\mathbb{E}_{\theta_0}\int_{\overline{\Theta}^c} e^{l_N(\theta)-l_N(\theta_0)}d\Pi(\theta)(1-\psi_N) + o(1) \\ &= e^{(b+a)N\delta_N^2}\int_{\overline{\Theta}^c} \mathbb{E}_\theta e^{l_N(\theta)-l_N(\theta_0)}(1-\psi_N)d\Pi(\theta) + o(1) \end{aligned}$$

where we have used (2.6) in the first equality and Fubini plus a change of measure argument to change $\mathbb{E}_{\theta_0}$ into $\mathbb{E}_\theta$, which we will not justify here. Finally, we see by (2.6) again that the RHS tends to zero provided $b + a$ is sufficiently small compared to M. For the details, see Theorem 1.3.2 in [2].

Remark 2.9. What we learn from this proof is that posterior contraction will always hold if we can write down a prior which is not bad in the sense of (2.9) at a contraction rate that is as good as the best test we can construct. We have seen that for the Hellinger distance, such tests can always be constructed under the metric entropy assumption.

2.2 The prior on $C([0,T], L^2(\Omega))$

The result to follow shows that the prior Π_γ is 'good'.

Lemma 2.10. Let $\theta_0 \in H^\gamma$, $\Pi = \Pi_\gamma \sim N(0, \rho^2(1-\Delta)^{-\gamma})$ where $\rho = \rho_N = 1/(N\delta_N)$, $\delta_N = N^{-\gamma/(2\gamma+d)}$, $\gamma > 1 + d/2$. Then $\exists A, c$ such that for reaction-diffusion (with $f \in C_c^\infty$) or the 2D NSE we have

$$\Pi(\theta : \sup_{0\leq t\leq T}\|u_\theta\|_\infty < U, \|u_\theta - u_{\theta_0}\|_{L_T^2} < \delta_N) \geq e^{-AN\delta_N^2} \tag{I}$$

and for $0 < \beta < \gamma - d/2$ and M sufficiently large

$$\Pi(\theta : \|\theta\|_{H^\beta} \leq M, \theta = \theta_1 + \theta_2, \|\theta_1\|_{H^\gamma} < M, \|\theta_2\|_{L^2} \leq M\delta_N) \geq 1 - e^{-c_M N\delta_N^2} \tag{II}$$

Remark 2.11. Here, (I) says that the prior assigns enough mass near the ground truth; we know already that the prior has full support on H^1 but we need to quantify this. (II) says that draws from the prior are likely to be 'regular' in the sense it can be decomposed into a sum of H^γ and L^2 functions. Note that the prior Π depends on N via ρ.

Proof. Straightforward PDE regularity arguments give $u_{\theta_0} \in L^\infty([0,T], H^\gamma)$. It suffices to prove (I) with $\|u_\theta - u_{\theta_0}\|_\infty$ instead of L^2_T. Assuming $\|\theta\|_{H^\xi} + \|\theta_0\|_{H^\xi} \leq U$ for some $\xi > \max\{d/2, 1\}$, standard arguments for dissipative PDEs give the regularity estimates

$$\|u_\theta - u_{\theta_0}\|_{L^2_T} \leq C_U \|\theta - \theta_0\|_{L^2} \tag{2.10}$$

and

$$\|u_\theta - u_{\theta_0}\|_{L^\infty_T} \leq C_U \|\theta - \theta_0\|_{H^\xi} \tag{2.11}$$

Given these results about the PDE in question, the inverse problem is solved for our purposes and the remainder of the proof is clever deviation theory for Gaussian processes. Bounding the RHS of (I),

$$\begin{aligned}&\Pi(\theta : \sup_{0\leq t\leq T} \|u_\theta\|_\infty < U, \|u_\theta - u_{\theta_0}\|_{L^2_T} < \delta_N)\\ &\qquad\qquad \geq \Pi(\theta : \|\theta - \theta_0\|_{H^\xi} \leq U/C_U, \|\theta - \theta_0\|_{L^2} \leq \delta_N/C_U)\end{aligned}$$

by the regularity estimates. There are two issues with bounding this probability. Firstly, we can usually only compute small deviation asymptotics for Gaussian processes centered at zero. Secondly, it is an intersection of two events. We will use two standard tricks to deal with these. To shift back to the origin, the price we pay is the RKHS norm of the shift $\|\theta_0\|_{H^\gamma}$, and this can be proven using the Cameron-Martin theorem plus convexity:

$$\geq e^{-N\delta_N^2 \|\theta_0\|^2_{H^\gamma}} \Pi(\theta : \|\theta'\|_{H^\xi} \leq \sqrt{N}\delta_N U/C_U, \|\theta'\|_{L^2} \leq \sqrt{N}\delta_N^2/C_U)$$

The Gaussian correlation inequality (proven only ten years ago) tells us that the probability of the intersection of two symmetric events centered at the origin is lower bounded by the worst case which is the independent case i.e. the product of the probabilites

$$\geq e^{-cN\delta_N^2} \Pi(\theta : \|\theta'\|_{H^\xi} \leq \sqrt{N}\delta_N U/C_U) \Pi(\theta : \|\theta'\|_{L^2} \leq \sqrt{N}\delta_N^2/C_U)$$

It remains to bound the probability that the norm of a Gaussian process is small. For this we use a results of Lindner and Li from the Annals of Probability 1999. One can compute this probability in terms of the metric entropy of the RKHS of the Gaussian process, and it works out.

$$\geq e^{-AN\delta_N^2}$$

For (II) we give an even rougher sketch. Because $\theta = \frac{1}{\sqrt{N}\delta_N}\theta'$ and $\theta' \in H^\beta$ almost surely and we know that the norm of a Gaussian vector in infinite dimensions is sub-Gaussian, so after rescaling this is the usual Fernique's theorem which tells us these two norms have sub-Gaussian tails, then everything is in the $\sqrt{N}\delta_N$ scale. The other bit is the Borel's isoperimetric inequality for Gaussian processes with RKHS $\sqrt{N}\delta_N H^\gamma$ (where the scaling means the norm is scaled, not the RKHS) : one can always find a large support set tells you that somehow the prior will live in some sort of RKHS ball, not in an almost sure sense, rather it can be enlarged in the ambient Banach norm L^2 and still have overwhelming mass. □

From the results of this section and the previous section, one can deduce the event that regularises survives in the posterior:

$$\Pi(\theta : \|\theta\|_{H^\beta} \leq M, \|u_\theta - u_{\theta_0}\|_{L^2_T} \leq M\delta_N | Z^{(N)}) \to 1 \quad \text{in } P^N_{\theta_0} \text{ probability.} \tag{2.12}$$

This is our first consistency result which tells us that what we've done to infer our dynamical system does work, at least in this parabolic L^2_T norm which averages in space and time. Also by a uniform integrability argument, one can further deduce convergence of plug-in estimates using the posterior mean:

$$\|u_{\tilde{\theta}} - u_{\theta_0}\|_{L^2_T} = O_P(\delta_N) \tag{2.13}$$

where $\tilde{\theta} = E^{\Pi}(\theta | Z^{(N)})$ is the posterior mean.

Remark 2.12. We do compute the posterior mean at the level of the initial condition because there we have a linear space and so it makes sense to average. One could in principle take the expectation of u_θ, but this is not what is done in MCMC, where we want a linear space to iterate but the solution space of a non-linear PDE is not a linear space. We don't talk about the posterior mean of the dynamical system, rather the plug-in estimate using the posterior mean of the initial condition, and we have shown in this case we have convergence at the same rate as posterior contraction.

2.3 Stability estimates

We conclude from the previous two sections that the posterior contracts on sets with $\|u_\theta - u_{\theta_0}\|_{L^2_T}$ small. We have not yet seen that this implies $\|\theta - \theta_0\|$ is small. Until about ten years ago, people in Bayesian non-parametric statistics just stopped here. But from a PDE perspective, one expects that for well-posed problems, if u_θ and u_{θ_0} are close and we have some regularity bound on both, then θ and θ_0 themselves should be close. If the map $\theta \to u_\theta$ were injective with Lipschitz inverse then what we have done here should already be enough.

For the 2D NSE one obtains a differential inequality for $\Phi(t) = \frac{\|\nabla w(t)\|^2_{L^2}}{\|w(t)\|^2_{L^2}}$ where $w = u_\theta - u_{\theta_0}$ to show

Theorem 2.13. For $\|\theta\|_{H^1} + \|\theta_0\|_{H^1} \leq U$ and $u_\theta(t), u_{\theta_0}(t)$ solutions to the 2D NSE. Then $\exists C_{U,T}$ such that for all $t \geq 0$

$$\|\theta - \theta_0\|_{L^2} \lesssim C_{U,T} \left(\log \frac{C_{U,T}}{\|u_\theta(t) - u_{\theta_0}(t)\|_{L^2}} \right)^{-1/2} \tag{2.14}$$

and

$$\|u_\theta(t) - u_{\theta_0}(t)\|_{L^2} \lesssim C_{U,T} \left(\log \frac{C_{U,T}}{\|u_\theta - u_{\theta_0}\|_{L^2_T}} \right)^{-1/2}. \tag{2.15}$$

Corollary 2.14. (Posterior consistency)

$$\Pi\left(\theta : \sup_{0 \leq t \leq T} \|u_\theta(t) - u_{\theta_0}(t)\|_{L^2} \geq \frac{M}{\sqrt{\log N}} | Z^{(N)} \right) \xrightarrow{P^{\mathbb{N}}_{\theta_0}} 0 \tag{2.16}$$

Remark 2.15. When there is no forcing, one shows there is an initial condition such that the logarithmic bound (2.14) is sharp for $t > 0$, so polynomial rates are not possible at this level of generality. More specifically, initial conditions can be chosen so that the NSE takes the form of a heat equation, see Theorem 2 in [4].

We now explore the analogous result for the Reaction-Diffusion equation, which has a more difficult but more intuitive proof than the NSE.

Theorem 2.16. (Reaction-Diffusion) Assume $\|\theta\|_{H^1} + \|\theta_0\|_{H^1} \leq U$. Then,

$$\int_0^T \|u_\theta(t) - u_{\theta_0}(t)\|^2_{L^2(\Omega)} dt \geq C_{U,T} \|\theta - \theta_0\|_{H^{-1}} \tag{2.17}$$

Proof. Take $w = u_\theta - u_{\theta_0}$ so that

$$\begin{aligned} \left(\frac{d}{dt} - \Delta \right) w &= f(u_\theta) - f(u_{\theta_0}) \\ &= f'(\tilde{u}) w \\ &= \tilde{v}(t) w \end{aligned}$$

by the mean value theorem and defining a new function $\tilde{v}(t)$. Thus w solves a linear time-dependent equation with initial condition $\theta - \theta_0$. The story so far suggests ‘all’ of the information of the PDE is contained at $t = 0$, so we proceed by considering the perturbed time-independent equation:

$$\left(\frac{d}{dt} - \Delta \right) w_\epsilon = v_\epsilon w_\epsilon \tag{2.18}$$

where $v_\epsilon = f'(\tilde{u}(0))$ for $t \in [0, \epsilon]$ (and take a piecewise linear continuation for later t, or whatever, as long as a solution w_ϵ exists). Then v_ϵ is time-independent on $[0, \epsilon]$ so considering the LHS of (2.17), we can roughly say (cheating with the ϵ) ,

$$\int_0^T \|u_\theta(t) - u_{\theta_0}(t)\|^2_{L^2(\Omega)} dt \gtrsim \int_0^\epsilon \|w_\epsilon\|^2_{L^2} dt$$

Solutions to the perturbed time-independent equation (a heat equation) are well understood: we have for $t \in [0, \epsilon]$

$$w_\epsilon(t) = \sum_{j=1}^\infty e^{-t\lambda_j} \langle e_j, \theta - \theta_0 \rangle e_j \tag{2.19}$$

where (e_j, λ_j) are eigenpairs such that

$$(\Delta - v_\epsilon) e_j = -\lambda_j e_j. \tag{2.20}$$

Since these form an orthonormal basis, we have by Parseval,

$$\int_0^\epsilon \|w_\epsilon\|^2_{L^2} dt = \int_0^\epsilon \sum_{j=1}^\infty e^{-2t\lambda_j} \langle e_j, \theta - \theta_0 \rangle^2_{L^2} dt \tag{2.21}$$

$$= \sum_{j=1}^\infty \frac{1}{2\lambda_j} (1 - e^{-\epsilon\lambda_j}) \langle e_j, \theta - \theta_0 \rangle^2_{L^2} dt \tag{2.22}$$

$$\gtrsim \|\theta - \theta_0\|^2_{H^{-1}} \tag{2.23}$$

where we have used that $\sum_j \frac{1}{\lambda_j} \langle e_j, \theta - \theta_0 \rangle^2_{L^2}$ is like the H^{-1} norm (bookkeeping).

□

Interpolating L^2 between H^β and H^{-1} and applying Lemma (2.10) we have the important corollary,

$$\Pi\left(\theta : \|\theta\|_{H^\beta} \leq M, \|\theta - \theta_0\|_{L^2} \leq M\tilde{\delta_N}|Z^{(N)}\right) \xrightarrow{P_{\theta_0}^{\mathbb{N}}} 1 \tag{2.24}$$

where $\tilde{\delta_N} = N^{-\frac{\gamma}{2\gamma+d}\frac{\beta}{\beta-1}} = \delta_N^{\frac{\beta}{\beta-1}}$ comes from interpolation.

Remark 2.17. If we choose a more regular model by. increasing γ so that both the prior and the ground truth get smoother, then β can also increase (see Lemma 2.10) and this exponent will approximate $-1/2$ which is the $\sqrt{n}$ parametric rate. One may ask whether the rates in Theorem 2.16 are optimal, and in fact they are not, it is possible to get the $\sqrt{n}$ rate which will come from the Bernstein von Mises theorem to follow. The proof will involve Theorem 2.16, which allows us to localise the posterior of the non-linear model, so we cannot skip this step.

3 Bernstein von Mises Theorems

If $\Theta = \mathbb{R}^D$ with D fixed and the prior density $d\Pi$ is positive on $\mathbb{R}^D$ and the Fisher information of the model $I_N(\theta_0)$ (to be defined shortly) is invertible, then any posterior obtained from an i.i.d sample will be approximated by a normal distribution on $\mathbb{R}^d$ in the sense

$$\|\Pi(\cdot|Z^{(N)}) - N_{\mathbb{R}^D}(\tilde{\theta}, \frac{1}{N}I_N(\theta_0)^{-1})\|_{TV} \xrightarrow{P_{\theta_0}^{N}} 0 \tag{3.1}$$

where $\tilde{\theta}$ is the posterior mean or maximum likelihood estimator and TV is the total variation norm, which is half the L^1 distance between the densities. This is the Bernstein von Mises theorem in finite dimensions and one can ask even without the PDE context in mind whether this result generalises to infinite dimensional models. Friedmann (1999) showed BvM (3.1) fails in infinite dimensions for the ℓ^2 norm instead of TV with a very simple counterexample, which was a famous negative result and meant for a long time people didn't even investigate Bayesian methods in infinite dimensions. Castillo/Nickl (2013) showed that BvM (3.1) holds in Friedmann's example with the weaker H^{-k} norm with $k > d/2$.

3.1 Main result

The laws $\widehat{\Pi}_N$ are (random) Borel probability measures on the separable Banach space

$$\mathcal{C} = C([t_{\min}, t_{\max}], C(\Omega))$$

with the $\|\cdot\|_\infty$ norm induced topology and $0 < t_{\min} < t_{\max}$. We will want to say that these posterior laws converge to some limiting law on $\mathcal{C}$ and for this we need to introduce a metric on the space of probability measures. We define the *Wasserstein distance* by

$$W_{1,\mathcal{C}}(\mu, \nu) = \sup_{\substack{H:\mathcal{C}\to\mathbb{R} \\ \|H\|_{\mathrm{Lip}\leq 1}}} \left|\int_{\mathcal{C}} H(x)(d\mu(x) - d\nu(x))\right| \tag{3.2}$$

where $\|\cdot\|_{\mathrm{Lip}}$ is the Lipschitz constant.

Theorem 3.1. Let $\theta_0 \in H_0^\gamma$ with γ sufficiently large. Then

$$W_{1,\mathcal{C}}\left(\mathrm{Law}(\sqrt{N}(u_\theta - u_{\tilde{\theta}_N})|Z^{(N)}), \mathrm{Law}(U)\right) \xrightarrow{P_{\theta_0}^{\mathbb{N}}} 0 \tag{3.3}$$

where u_θ is drawn from the posterior and $u_{\bar{\theta}_N}$ is the posterior mean. Moreover,

$$\sqrt{N}(u_{\bar{\theta}_N} - u_{\theta_0}) \to_{\text{dist}} \text{Law}(U) \tag{3.4}$$

where convergence in distribution is the usual convergence in distribution on $\mathcal{C}$ and U is the Gaussian random field in $\mathcal{C}$ solving the linear PDE

$$\frac{d}{dt}U - \Delta U = f'(u_{\theta_0})U \tag{3.5}$$

in the case of reaction-diffusion or

$$(\frac{d}{dt} - \Delta)U + B(u_{\theta_0}, U) + B(U, u_{\theta_0}) = 0 \tag{3.6}$$

in the case of the NSE with (random) initial condition $u(0, \cdot) \sim X \sim N_{\theta_0} =: N(0, (\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0})^{-1})$ defining a Gaussian Borel probability measure on H^{-k} for $k > 1 + d/2$.

Remark 3.2. This theorem says the centering of BvM does converge to the right thing, and so in particular once we've proven that the limit is a Gaussian random function, shows that fluctuations around the ground truth $u_{\bar{\theta}_N} - u_{\theta_0}$ are at the $\sqrt{N}$ rate of the central limit theorem even though the model is infinite dimensional. This is uncommon in non-parametric statistics and comes from the parabolic structure of the PDEs. The most delicate part of the proof is the construction of the limiting process U.

3.2 Information operators

For $h \in L^2$, define the *information operator* $\mathbb{I}_{\theta_0}(h) = U_h$ where U_h is the solution to the linearised PDE (3.5) with initial condition h. One shows the information operator is the linearisation of the solution map $\theta \to u_\theta$ in the sense

$$\|u_{\theta_0+h} - u_{\theta_0} - \mathbb{I}_{\theta_0}(h)\|_{L^2} = O(\|h\|^2_{L^2}) \tag{3.7}$$

If $\mathbb{I}_{\theta_0} : L^2_0 \to L^2_T$ has a Hilbert space adjoint $\mathbb{I}^*_{\theta_0} : L^2_0 \to L^2_T$ then the *Fisher information operator* is $\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0} : L^2_0 \to L^2_0$. In general, difficulties in our theory will arise from this adjoint, whereas in the finite dimensional case the adjoint is simply the transpose. In a standard regression model, the Fisher information is just the identity operator, but when the regression function is non-linear the local identifiability of the statistical model is encoded in the Fisher information.

Once one has a consitent estimator, the non-linear structure itself doesn't matter much any more, rather just the Fisher information which drives the local convexity of the expected log likelihood function at the ground truth tells us everything – in particular in infinite dimensions it's mapping properties such as invertability. Studying this object in the context of PDEs is not only difficult, but we quickly arrive at problems that people in PDEs haven't studied yet. For the reaction-diffusion equation and NSE, we will be able to do it, but there are non-trivial examples such as the Darcy flow problem (the standard example every applied mathematician uses when they study non-linear Bayesian inverse problems) where the Fisher information is not invertible, even when restricted on the space of smooth functions. So not only is our goal difficult, it might even be wrong, and in this caseBvM theorem does not hold true. It may be that the fluctations of the posterior are not Gaussian, and this depends on whether the Fisher information can be inverted on suitable domains. Currently we have several positive/negative examples but we do not yet have a clear intuition on when the Fisher information is invertible.

The adjoint usually heavily interacts with the type of measurement model chosen as it maps into L^2_T– we recall L^2_T is not arbitrary, it comes from the Hellinger distance which was forced upon us by the measurement model. Therefore when we compute the adjoint, the inner product contains all the information of how we made our measurments; dealing with this can be mathematically nasty. However, we can prove for data assimilation problems with a laplacian and a non-linearity which is a lower order than the Laplacian, one can show the Fisher information operator is a nice homeomorphism and up to a compact perturbation of the identity, the inverse Laplacian. We now prove this for the reaction-diffusion equation.

Theorem 3.3. For $\eta > 0$, the operator $\Delta \mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}$ is a homeomorphism of $H^\eta_0 =: H^\eta \cap L^2_0$.

Proof. We omit the subscript θ_0. Without the potential in (3.5), we have the standard heat equation: let $I(h) = U_h$ where

$$\begin{cases} \frac{dU}{dt} - \Delta U = 0 \\ U(0,\cdot) = h \end{cases} \tag{3.8}$$

Thus the operator I is just the forward solution of the standard heat equation and I^* is the backward solution integrated over time, and because this is reversible, going backwards then forwards is the same as going twice forwards. When one integrates the eigenfunctions which are just the exponentials, we get the recipocals of the eigenvalues (like we did in the injectivity estimate Theorem 2.16), so applying the Laplacian after integrating $I^*I(h)$ cancels them and we are essentially left with the identity– philosophically why the result holds in this case.

In general it is not quite as easy as we have a time-dependent potential. The idea is to use Fredholm theory: a compact perturbation of the identity is surjective if it is injective. If we can show first $\mathbb{I}^*\mathbb{I}$ is injective (which is necessary for invertibility anwyay), then perhaps we can show the difference $\Delta(\mathbb{I}^*\mathbb{I} - I^*I)$ is compact on H^η_0, hence $\mathbb{I}^*\mathbb{I}$ is a compact perturbation of the identity which is injective hence surjective by Fredholm theory. These turn out to be true though we wont prove them here, we summarise that once we have injectivity, the mapping properties of the linearised flows with time dependency are similar to those of the standard heat equation where we can compute things explicity.

We first do the computation for the standard heat equation:

$$\begin{aligned}(\Delta I^* I)(h) &= \Delta \int_0^T I^*_t I_t(h)dt \\ &= \Delta \sum_{j=1}^\infty \int_0^T e^{-2t\lambda_j}\langle e_j, h\rangle e_j\end{aligned}$$

where we have solved the standard heat equation explicitly in terms of the eigenfunctions of the Laplacian

$$\begin{aligned}&= \sum_{j=1}^\infty (-\lambda_j)\frac{1}{-2\lambda_j}(e^{-2T\lambda_j} - 1)\langle e_j, h\rangle e_j \\ &= -\frac{1}{2}Id + U_h(2T)\end{aligned}$$

where $h \to U_h(2T)$ is evaluation at time $2T$, not multiplication, hence a compact operator. We think of this as saying $(I^*I)^{-1}$ is essentially $-2\Delta^{-1}$ (plus a compact thing).

To show $\mathbb{I}$ is injective, assume $\Delta \mathbb{I}^* \mathbb{I} h = 0$. Then by injectivity of Δ on L_0^2 we have $\mathbb{I}^* \mathbb{I} h = 0$ on L_0^2. Thus,

$$\begin{aligned} \langle \mathbb{I}^* \mathbb{I} h, h \rangle_{L^2} &= 0 \\ &= \|\mathbb{I} h\|_{L^2}^2 \\ &\gtrsim \|h\|_{H^{-1}} \end{aligned} \tag{3.9}$$

where the inequality is a straightforward stability estimate for the linearised Schrodinger equation which $\mathbb{I}$ solves (which is strictly easier to prove than Theorem 2.16 since we don't need the mean-value perturbation argument). We conclude $h = 0$ hence $\mathbb{I}^* \mathbb{I}$ is injective.

□

Remark 3.4. We think of this theorem as saying the inverse Fisher information is 'basically' the inverse Laplacian at least in terms of its mapping properties (though this is far from being the same as the inverse Laplacian). This explains why we have the condition $k > 1 + d/2$ in Theorem 3.1 because U is the image of the white noise process under the half-Laplacian.

Remark 3.5. The difference $\Delta(\mathbb{I}^* \mathbb{I} - I^* I)$ being compact is not unexpected: the potential in (3.5) (or in the NSE case, the B term in (3.6)) is at most a first order differential operator whereas the Laplacian is second order and there is a 2-smoothing property of solutions to parabolic equations on the source term, so one expects $\Delta \mathbb{I}^* \mathbb{I}$ and $\Delta I^* I$ to be similar. More precisely one shows using standard parabolic regularity estimates that $\mathbb{I}^* \mathbb{I} - I^* I$ maps H_0^η into $H_0^{\eta+3}$, hence $\Delta(\mathbb{I}^* \mathbb{I} - I^* I)$ maps H_0^η into $H_0^{\eta+1}$, thus compact.

Remark 3.6. The computation in the proof for the standard heat equation would be very difficult to do in the general case because of the time dependence: the information operator is not reversible in general and so when we run it backwards in time one runs into serious issues. The main takeaway of the proof is that whatever is written next to the Laplacian, as long as it is lower order, doesn't matter for the mapping properties of the solution.

Remark 3.7. This result is true for the NSE but the proof above doesn't work because the linearisation (3.6) is not a symmetric operator so we cannot use spectral theory, where as in (3.5) $f'(u_{\theta_0})U$ is. In general, for any data assimilation problem with a Laplacian perturbed by lower order terms, one should expect the Fisher information to be approximately the inverse Laplacian.

3.3 BvM for initial condition

We will show that there is a Gaussian approximation to the posterior distribution in a weak sense, in particular some negative Sobolev space topology.

Theorem 3.8. For $k > 2d + 3$,

$$W_{1,H^{-k}}(\mathrm{Law}(\sqrt{N}(\theta - \tilde{\theta_N}) | Z^{(N)}), N_{\theta_0}) \xrightarrow{P_{\theta_0}^{\mathbb{N}}} 0 \tag{3.10}$$

and $\sqrt{N}(\tilde{\theta_N} - \theta_0) \to_{\mathrm{dist}} N_{\theta_0}$.

Remark 3.9. We cannot expect a stronger convergence here than in a negative Sobolev space (except possibly optimising k which we didn't try to do). The negative result of Friedmann tells us this cannot hold in L^2. It is necessary to have $k > 1 + d/2$ because the measure N_{θ_0} is only tight in H^{-k} when $k > 1 + d/2$, we will never have any type of weak convergence of (3.10) unless we enforce this weak topology because we have to have a tight limit.

Proof. Our proof strategy has nothing to do with data assimilation, we could have any likelihood model where we can invert the informational operator.
Idea:

(i) Localise: We already know the posterior contracts around the ground truth. The question is, what is the best way to use that? A clever idea due to LeCam when he first wrote these proofs in finite dimensions is to look at a posterior that comes from a prior which is already concentrated on the set we like. We consider $\Pi^{D_N}(\cdot|Z^{(N)})$ where

$$\Pi^{D_N} = \frac{\Pi(\cdot \cap D_N)}{\Pi(D_N)} \tag{3.11}$$

where

$$D_N = \{\theta : \|\theta\|_{H^\beta} \leq M, \|\theta - \theta_0\|_{L^2} \leq M\tilde{\delta_N}\} \tag{3.12}$$

By virtue of our previous results, we can assume our prior has already guessed that I am in some neighbourhood of the ground truth. The rate $\tilde{\delta_N}$ is strictly slower than $1/\sqrt{N}$ so it will not allow me to grind out the $\sqrt{N}$ in (3.10), but it will modulo some expansions of the log likelihood.

(ii) To prove convergence in H^{-k}, we will look at all the actions on test functions by integration and then we will prove a result that is uniform in $\psi \in H^{-k}$ and from this we can reconstruct the norm of the space. Indeed, given $\psi \in H^k_0$, we define the 'least favourable direction' $\overline{\psi} := (\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0})^{-1}\psi$ which belongs to H^{k-2}_0 by the mapping property result for the information operator we proved. We specifically proved this for reaction-diffusion, but it is conceivable one could prove this for other models.

(iii) Study the Laplace transform of the posterior with argument t. In the next display, the expectation is taken over θ, ψ is fixed, and $\widehat{\psi_N}$ is a convenient centering.

$$\mathbb{E}^{\Pi^{D_N}}\left(e^{t\sqrt{N}(\langle\theta,\psi\rangle - \widehat{\psi_N})}|Z^{(N)}\right) = \frac{\int e^{t\sqrt{N}(\langle\theta,\psi\rangle - \widehat{\psi_N} + l_N(\theta)}d\Pi(\theta)}{\int e^{l_N(\theta)}d\Pi(\theta)} \tag{3.13}$$

$$= \frac{\int e^{t\sqrt{N}(\langle\theta,\psi\rangle - \widehat{\psi_N} + l_N(\theta) + l_N(\theta) - l_N(\theta_{(t)}) + l_N(\theta_{(t)}))}d\Pi(\theta)}{\int e^{l_N(\theta)}d\Pi(\theta)} \tag{3.14}$$

where $\theta_{(t)} = \theta - t/\sqrt{N}\overline{\psi}$ is a perturbation of θ in the least favourable direction. The hard work is in controlling $e^{t\sqrt{N}(l_N(\theta) - l_N(\theta_{(t)})}$ in the previous display uniformly in ψ and N. One needs to not only show that this is small, but tht the first order terms will conveniently cancel and use infinite dimensional central limit theorem techiniques – see Chapter 3 in [3]. After expanding,

$$\widehat{\psi_N} = \langle\theta_0, \psi\rangle - \frac{1}{N}\sum_{i=1}^{N} \epsilon_i \mathbb{I}_{\theta_0}(\overline{\psi})(t_i, x_i) \tag{3.15}$$

and (3.14) using a 'quantitative LAN expansion', everything cancels and is small and we are left with

$$= e^{t^2/2\|\mathbb{I}_{\theta_0}\overline{\psi}\|^2_{L^2} + o_P(1)} \cdot \frac{\int e^{l_N(\theta_{(t)})}d\Pi(\theta)}{\int e^{l_N(\theta)}d\Pi(\theta)} \tag{3.16}$$

The first term looks like a Gaussian moment generating function that we would like to see. The second term looks like it should be close to 1; $\theta_{(t)}$ is a small perturbation of θ for fixed ψ. One can show this by using the Gaussian process prior formula explicitly and the Cameron-Martin theorem. Here, we use not only that the prior is supported everywhere, but we need to be able to quantify how much a shift into the least favourable direction affects the prior. For bad priors, this might be huge and explode but for Gaussian process priors it can be controlled.

(iv) So far, we have convergence for fixed ψ (after some uniform integrability arguments and getting rid of $\widehat{\psi_N}$ and getting the posterior means, and some other technicalities we skip over). Then, from the uniformity in ψ, one does the usual weak convergence argument in function space by decomposing into finite dimensional projections of both the conveging sequence and the limit and use the limit theorem for the finite dimensional projections and control the rest by a tightness estimate which tells us all these measures are sort of flatly concentrated in finite dimensional spaces (this is a standard argument for proving CLT theorems in infinite dimensions). For tightness, one needs to control the supremum of infinitely many ψ, but we have already controlled the MGF by essentially a Gaussian, so for each ψ these are sub-Gaussian variables and taking the supremum is like a Gaussian process. This is why we pay almost no price for going from finite to infinite dimensions in a suitably regular space like H^{-k}. The unit balls in these spaces have nice covering numbers so one can control the suprema of sub-Gaussian process (e.g. Dudley's metric entropy method).

(v) To finish off with convegence of the posterior mean, one uses some technical uniform integrability and convergence of moments arguments, which we omit.

Remark 3.10. Idea (iii) is perhaps not so difficult to come up with, but the real hard work in infintie dimensions is to kill the remainder terms in (3.14). Each l_N is a sum of N terms evaluated at functions which are small, but in the CLT because the functions indexing it are very small, we almost end up in the Poissonian regieme of the CLT and so it is a bit expensive and we cannot automatically assume we have subgaussian tails.

Remark 3.11. The bottleneck is $(\mathbb{I}_{\theta_0}^* \mathbb{I}_{\theta_0})^{-1}$ which may not exist, or it may exist but you don't know anything about its mapping properties which is equally bad. If one proves the information equation has a solution but nothing is known about its regularity, one cannot proceed with this proof because we cannot control the remainder terms or the second term in (3.16) when perturbing in the least favourable direction. For data assimilation problems we have proved it is basically the Laplacian so we know exactly what the mapping properties are. If one wants to write a proof, it is probably necessary to dig deep into concentration of product measure phenomena.

We will prove the theorem for the posterior mean because it is easier to see the idea. The map $\theta \to u_\theta$ is non-linear, so we cannot just use the continuous mapping theorem and we have to deal with the non-linearity of the forward map. We know how to linearise this map,

$$\begin{aligned}\sqrt{N}(u_{\tilde{\theta_N}} - u_{\theta_0}) &= (Du_{\theta_0})(\sqrt{N}(\tilde{\theta_N} - \theta_0) + \sqrt{N}\|\tilde{\theta_N} - \theta_0\|_{L^2}^2) && (3.17)\\ &= \mathbb{I}_{\theta_0}(\sqrt{N}(\tilde{\theta_N} - \theta_0) + o_P(1)) && (3.18)\end{aligned}$$

where we have used (2.24) that $\|\tilde{\theta_N} - \theta_0\|_{L^2}^2$ is like $\tilde{\delta_N}^2$, so that $\sqrt{N}\tilde{\delta_N}^2 \to 0$ because we proved already that $\tilde{\delta_N}$ is better than $N^{-1/4}$ (but worse than $N^{-1/2}$). We see here that proving fast rates first is necessary to prove BVM. For example if we had only logarithmic rates then the remainder term would be large. Lastly, one shows using basic parabolic smoothing arguments to show that $\mathbb{I}_{\theta_0}$ maps H^{-k} into $\mathcal{C}$ or in fact into any Sobolev space; one obtains estimates uniform in time away from zero because away from zero the solution gains one smoothness from the initial condition so one can bootstrap by splitting $\mathbb{R}_+$ into infinitely many intervals. Thus the continuous mapping theorem applies and we conclude

$$\sqrt{N}(u_{\tilde{\theta}_N} - u_{\theta_0}) \to_{\text{dist}} \mathbb{I}_{\theta_0}(X) \tag{3.19}$$

where $X \sim N_{\theta_0}$.

Remark 3.12. The same proof works for the Wasserstein distance (3.10) because $\mathbb{I}_{\theta_0}$ is a bounded and linear hence Lipschitz map and the composition of Lipschitz maps is Lipschitz, and the Wasserstein distance is the supremum over Lipschitz maps. However, work is needed to control the $o_P(1)$ term, to do this one shows all the posterior moments are bounded and this is more techinical and relies on the non-linearity of the PDE is well behaved in the sense that the norms of u_θ cannot grow faster than polynomially in the norm of θ, which is checked explicitly for NSE/reaction-diffusion.

Remark 3.13. This result can be used in statistics. If one has computed $u_{\tilde{\theta}_N}$ and draw $1/\sqrt{N}$ Gaussian fluctuations around it, then this is a confidence band for u_{θ_0}. If one has a certain hypothesis and the estimator $u_{\tilde{\theta}_N}$ doesn't lie in this confidence band, then we can reject it. The error probability is known exactly because we know the quantiles of the limit distribution; although one does not need to compute these because the posterior does this automatically by a bootstrap argument, one can use the posterior quantiles and they will stabilise at the right limit. This is why people use Bayesian methods in applied mathematics because they give you error bars, and we have proven that these confidence regions are not just valid, but actually quite small (the parametric rate) even though the model is non-parametric.

□

3.4 Cramer-Rao lower bounds

We next show that the Bayesian posterior mean estimator is optimal in an information theoretic sense. Recall the classical Gauss-Markov theorem: We have

$$\mathbb{E}(\tilde{\theta} - \theta_0)^2 = \text{Variance} + \text{Bias}^2 \tag{3.20}$$

For consistent estimators, the bias tends to zero, so among consistent estimators one wants to minimise the variance. Gauss-Markov states the least squares estimator minimises variance among all linear estimators. In parametric statistics, this is generalised by the Cramer-Rao lower bound, and in infinite dimensions we can prove something similar at least asymptotically. Any estimation algorithm (e.g. from machine learning or anywhere) must write down an estimator which is a measurable function of the data observed. It will have a risk, and if its not silly it will also be consistent. The next result says that it cannot have asymptotically lower variance than our posterior mean $u_{\tilde{\theta}_N}$:

Theorem 3.14. Let $\theta_0 \in H^\gamma$ be the ground truth. Then,

$$\liminf_{N} \inf_{\hat{\theta}(Z^{(N)})\text{ meas.}} \sup_{h:\|h\|_{H^1}\leq 1} \mathbb{E}_{\theta_0+h/\sqrt{N}} \left(\sqrt{N}\|u_{\hat{\theta}} - u_{\theta_0+h/\sqrt{N}}\|_{\mathcal{C}}\right)^2 \geq \mathbb{E}\|U\|^2_{\mathcal{C}} \tag{3.21}$$

where U is from

We explain this big equation in words, inside out. The expectation is the risk for the estimator $\hat{\theta}$ when the data is generated by the perturbed ground truth $u_{\theta_0+h/\sqrt{N}}$, and then we take the supremum over all such permutations to get the worst case risk near θ_0 for $\hat{\theta}$. We then take the infimum over all estimators $\hat{\theta}$, i.e. minimise the worst case risk ('minimax'). This is at least the norm of our limiting field U from (3.5), and this lower bound is attained by our posterior mean estimator $\tilde{\theta_N}$ because U is precisely the limit of $\sqrt{N}(u_\theta - u_{\tilde{\theta_N}})$ and one can hence show convergence of the second moments.

Remark 3.15. Initially it seemed like a silly idea to try predict the system forward by first estimating the initial condition and then solving forwards (Why would I care about what happened yesterday if I care about what happens tomorrow?), but it turns out that this method is sharp. In summary, when you do data assimilation, you might aswell solve backwards and then forwards. Although, this very much depends on the parabolic nature of the PDE, it might not be true for other non-parabolic dissipative systems or time evolutions without smoothing. But equations of the type we have considered are relevant in data assimilations.

Remark 3.16. We haven't mentioned computation much. Recall that computing the posterior mean is too hard, and instead one computes markov chains which approximate the posterior mean, but we haven't discussed how well they do that. Nickl+Wang (2024 JEMS) shows that under conditions slightly weaker than invertibility of the Fisher information then posteriors in these models can be computed by polynomial time algorithms, both in dimension and the number of samples. Hence one can obtain the same convergence rate and lower bound of $u_{\tilde{\theta_N}}$ for such algorithms, e.g. $u_{\frac{1}{k}\sum_{i=1}^k v_k}$ from Langevin. This isn't that surprising given what we've seen so far. The posteriors which seemed complicated at first are actually Gaussian, so essentially we're just working with Gaussians and everything is easy (though it actually turns out to be easier to prove this by not immediately going to the Gaussian limit rather by doing a log concave approximation of the posterior by a measure which satisfies a log-Sobolev inequality).

References

[1] Data Assimilation: A Mathematical Introduction: 62 (Texts in Applied Mathematics, 62). 2015.

[2] R. Nickl, Bayesian Non-linear Statistical Inverse Problems, European Mathematical Society (EMS) Press, 2023.

[3] R. Nickl, E. Gine. Mathematical foundations of infinite-dimensional statistical models. Cambridge University Press (2016).

[4] On posterior consistency of data assimilation with Gaussian process priors: the 2D Navier-Stokes equations. (with E. Titi), Annals of Statistics 52 (2024), 1825-1844.