MAP estimators and 4D-VAR

Jochen Voss

University of Leeds

12 May 2014, Reading-Warwick Data Assimilation Meeting

joint work with Masoumeh Dashti, Kody Law and Andrew Stuart

Outline

Data Assimilation using 4D-VAR

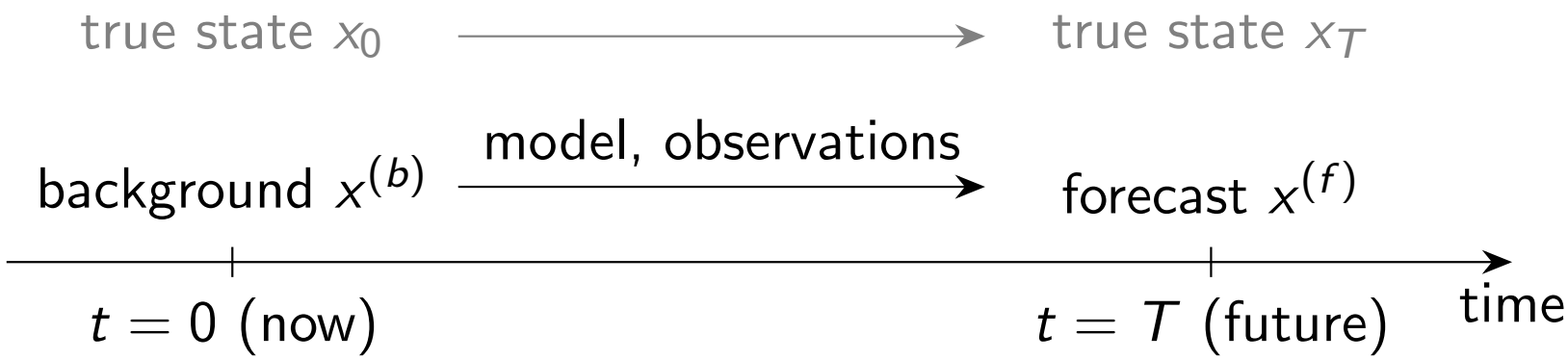
Infinite Dimensional MAP Estimators

Consistency

Conclusions

Data Assimilation using 4D-VAR

We consider the following "caricature" of a forecasting problem:



- We want to forecast the unknown state x_T of a system for a future time T, starting from the current state x_0.
- The current state x_0 is unknown, our "best guess" is $x^{(b)}$.
- For times $0 \leq t_1 \leq t_2 \leq \cdots \leq t_J \leq T$ we have noisy observations $y_j \approx H(x_{t_j})$.

For a Bayesian approach we make the following assumptions:

- $x^{(b)} - x_0 \sim \mathcal{N}(0, C)$, *i.e.* $x_0 \in \mathbb{R}^d$ has density

$$p(x_0) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp\Big(-\frac{1}{2}(x_0 - x^{(b)})^\top C^{-1}(x_0 - x^{(b)})\Big).$$

- The observations are independent and satisfy $y_j \sim \mathcal{N}\big(H(x_{t_j}), R\big)$, *i.e.* $y_j \in \mathbb{R}^m$ has density

$$p(y_j|x_0) = \frac{1}{(2\pi)^{m/2}|R|^{1/2}} \exp\Big(-\frac{1}{2}(y_j - H(x_{t_j}))^\top R^{-1}(y_j - H(x_{t_j}))\Big),$$

  where $x_{t_j} = M_{t_j}(x_0)$ is the system state at time t_j, for $j = 1, \ldots, J$.

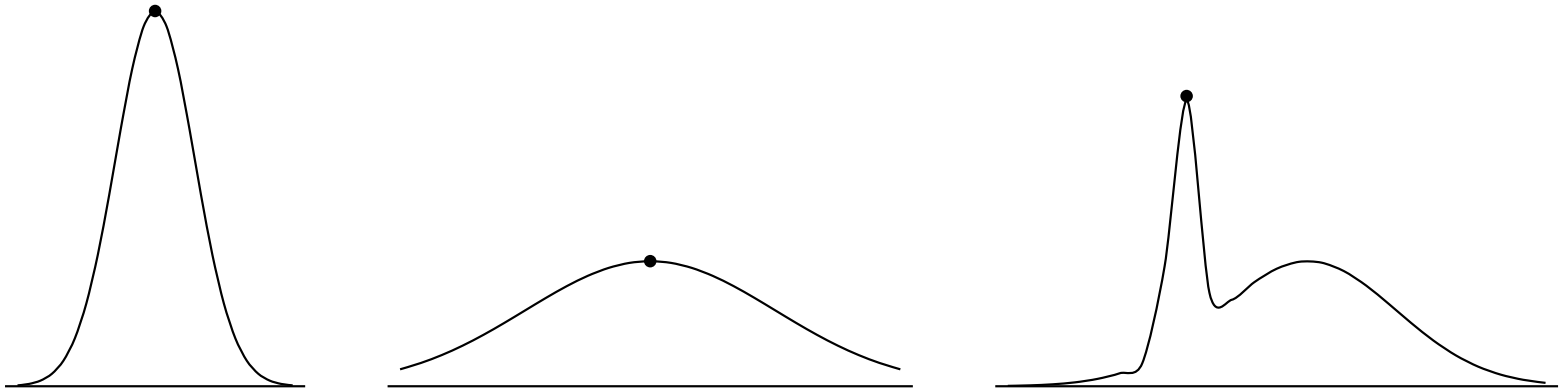
From these assumptions we can find the posterior density of x_0 as

$$p(x_0|y) = \frac{p(y|x_0)p(x_0)}{p(y_j)} \propto \prod_{j=1}^{J} p(y_j|x_0)p(x_0) = \cdots .$$

For high-dimensional models, the full posterior density for x_0 can be difficult to work with and often it is convenient to use a point estimate for x_0 instead. Here we use the **maximum a posteriori (MAP)** estimator $\hat{x}_0$, defined as

$$\hat{x}_0 = \arg\max_{x_0} p(x_0 \,|\, y).$$

This estimator attempts to obtain a "typical" value from the posterior distribution. The MAP estimator works well if the posterior is unimodal and highly concentrated.



For the forecasting problem, the MAP estimator is the $x_0 \in \mathbb{R}^d$ which maximises

$$\begin{aligned} p(x_0|y) &\propto \prod_{j=1}^{J} p(y_j|x_0)p(x_0) \\ &\propto \exp\Big(-\frac{1}{2}\sum_{j=1}^{J}(y_j - H(x_{t_j}))^\top R^{-1}(y_j - H(x_{t_j})) \\ &\qquad\qquad - \frac{1}{2}(x_0 - x^{(b)})^\top C^{-1}(x_0 - x^{(b)})\Big) \\ &=: \exp\big(-I(x_0)\big) \end{aligned}$$

or, equivalently, minimises the "cost function" I. The data assimilation method based on this procedure is called **4D-VAR**.

Summary: the 4D-VAR method minimises

$$I(x) = \Phi(x) + \frac{1}{2}\|x - x^{(b)}\|_E^2$$

where

$$\|x\|_E^2 = x^\top C^{-1} x$$

and

$$\Phi(x) = \frac{1}{2}\sum_{j=1}^{J}(y_j - H(x_{t_j}))^\top R^{-1}(y_j - H(x_{t_j}))$$

where H is the observation map and $x_{t_j} = M_{t_j}(x)$ is obtained by integrating the model, starting with state x at time 0, until time t_j.

By shifting coordinates, we can assume $x^{(b)} = 0$ without loss of generality.

Infinite Dimensional MAP Estimators

In many applications of MAP estimators, including applications in weather forecasting, the system state x is an infinite-dimensional object. Thus, it is natural to ask whether MAP estimators (and the 4D-VAR method) still work in infinite-dimensional spaces:

- If a numerical method does not make sense for the limiting, infinite dimensional object, the method may be ill-behaved for high-dimensional systems.
- Separating issues of discretisation from issues of the estimation method can lead to greater clarity.
- General rule: discretise as late as possible.

We assume that the posterior μ is a probability measure on an infinite dimensional, separable Banach space $(X, \|\cdot\|_X)$.

Problem. The MAP estimator is defined in terms of densities w.r.t. Lebesgue measure, but Lebesgue measure does not exist on infinite dimensional spaces.

Solution 1. We can consider reference measures μ_0 other than Lebesgue measure. Here we assume that μ_0 is a Gaussian measure on X and that μ has density

$$\frac{d\mu}{d\mu_0}(x) \propto \exp\bigl(-\Phi(x)\bigr)$$

w.r.t. μ_0.

In finite dimensions:

$$\begin{aligned}\frac{d\mu}{d\mathrm{Leb}}(x) &= \frac{d\mu}{d\mu_0}(x) \cdot \frac{d\mu_0}{d\mathrm{Leb}}(x) \\ &\propto \exp\bigl(-\Phi(x)\bigr) \cdot \exp\bigl(-\frac{1}{2}x^\top C^{-1}x\bigr) \\ &= \exp\bigl(-\Phi(x) - \frac{1}{2}\|x\|_E^2\bigr)\end{aligned}$$

for all $x \in \mathbb{R}^d$.
Infinite dimensional analogue of the right-hand side:

$$\text{“}\frac{d\mu}{d\mathrm{Leb}}(x)\text{”} \propto \exp\bigl(-\Phi(x) - \frac{1}{2}\|x\|_E^2\bigr)$$

for all $x \in E \subset X$ where $(E, \|\cdot\|_E)$ is the Cameron-Martin space of the Gaussian measure μ_0. Even if the left-hand side does not make sense any more, we can still try to maximise the right-hand side over E.

Example. If μ is the distribution of the solution of the stochastic differential equation (SDE) $dx_t = f(x_t)\,dt + dw_t$ on $X = L^2([0, T], \mathbb{R})$, then we can choose μ_0 to be Wiener measure (*i.e.* the distribution of a Brownian motion). By the Girsanov formula from stochastic analysis, μ has density $\exp\bigl(-\Phi(x)\bigr)$ w.r.t. μ_0, where

$$\Phi(x) = \frac{1}{2}\int_0^T \bigl|f(x_t)\bigr|^2\,dt - \int_0^T f(x_t)\,dx_t,$$

and the Cameron-Martin space of μ_0 is

$$E = \left\{x \in H^1([0, T], \mathbb{R}) \;\middle|\; x_0 = 0, \int_0^T \dot{x}_t^2\,dt < \infty\right\},$$

with norm

$$\|x\|_E^2 = \int_0^T \dot{x}_t^2\,dt$$

for all $x \in E$. The "MAP estimator" minimises $\Phi(x) + \frac{1}{2}\|x\|_E^2$.

Solution 2. Without using densities we can consider small ball probabilities $\mu\big(B(x,\varepsilon)\big)$ and then let $\varepsilon \downarrow 0$.

Definition. $\hat{x} \in X$ is a MAP estimator for μ, if

$$\lim_{\varepsilon \downarrow 0} \frac{\mu\big(B(\hat{x},\varepsilon)\big)}{\sup_{x \in X} \mu\big(B(x,\varepsilon)\big)} = 1.$$

Our main result show that $\hat{x} \in X$ is a MAP estimator for μ, if and only if $\hat{x}$ is a minimiser of the **Onsager-Machlup functional**

$$I(x) = \begin{cases} \Phi(x) + \frac{1}{2}\|x\|_E^2, & \text{if } x \in E, \text{ and} \\ +\infty & \text{otherwise.} \end{cases}$$

In particular this implies that MAP estimators always lie in the Cameron-Martin space E.

Assumptions. The function $\Phi\colon X \to \mathbb{R}$ satisfies the following conditions:

A1 Φ is bounded from below, *i.e.* there is an $M \in \mathbb{R}$, such that for all $x \in X$ we have

$$\Phi(x) \geq M.$$

A2 Φ is locally bounded from above, *i.e.* for every $r > 0$ there exists $K = K(r) > 0$ such that for all $x \in X$ with $\|x\|_X < r$ we have

$$\Phi(x) \leq K.$$

A3 Φ is locally Lipschitz continuous, *i.e.* for every $r > 0$ there exists $L = L(r) > 0$ such that for all $x_1, x_2 \in X$ with $\|x_1\|_X, \|x_2\|_X < r$ we have

$$|\Phi(x_1) - \Phi(x_2)| \leq L\|x_1 - x_2\|_X.$$

Theorem. Assume A1, A2 and A3. Then the following statements hold:

i) Any MAP estimator $\hat{x} \in X$ minimises the Onsager-Machlup functional I. In particular, $\hat{x}$ satisfies $\hat{x} \in E$.

ii) Any $\hat{x} \in E$ which minimises the Onsager-Machlup functional I is a MAP estimator.

The proof of the result is long and technical, but it is based on the following property of the Onsager-Machlup functional: If $x_1, x_2 \in E$, then

$$\lim_{\varepsilon \downarrow 0} \frac{\mu(B(x_2, \varepsilon))}{\mu(B(x_1, \varepsilon))} = \exp\big(I(x_1) - I(x_2)\big).$$

The technical difficulties are caused, among other things, by the following facts:

- A copy of the measure μ shifted by $x \in X$ is absolutely continuous w.r.t. μ, if and only if $x \in E$. Thus working with the probabilities $\mu(B(x, \varepsilon))$ works best if $x \in E$.
- $E \subseteq X$ is dense, but $\mu(E) = 0$.

Consistency

We have seen that the 4D-Var method minimises

$$I(x) = \Phi(x) + \frac{1}{2}\|x\|_E^2$$

where

$$\Phi(x) = \frac{1}{2}\sum_{j=1}^{J}|y_j - \mathcal{G}_j(x)|_R^2$$

and

$$\mathcal{G}_j(x) = H\big(M_{t_j}(x)\big).$$

The observations y_j satisfy $y_j = \mathcal{G}(x^\dagger) + \eta_j$, where $\eta_j \sim \mathcal{N}(0, R)$ are i.i.d. and $x^\dagger \in X$ is the true state.

Question. Does the 4D-VAR estimate converge to $x^\dagger$ as $J \to \infty$? (Answer: no, but . . .)

Large sample size limit. Let $x^\dagger \in X$ and

$$y_j = \mathcal{G}(x^\dagger) + \eta_j$$

where $\eta_j \sim \mathcal{N}(0, R)$ are i.i.d. for $j = 1, \ldots, J$. Then the corresponding Onsager-Machlup functional is

$$I_J(x) := \|x\|_E^2 + \sum_{j=1}^{J} |y_j - \mathcal{G}(x)|_R^2.$$

Theorem. Assume that $\mathcal{G}\colon X \to \mathbb{R}^K$ is locally Lipschitz continuous and $x^\dagger \in E$. For $J \in \mathbb{N}$, let $x_J \in E$ be a minimiser of I_J. Then

$$\lim_{J\to\infty} \mathcal{G}(x_J) = \mathcal{G}(x^\dagger)$$

almost surely.

Small noise limit. Let $x^\dagger \in X$ and

$$y_n = \mathcal{G}(x^\dagger) + \frac{1}{n}\eta_n,$$

where $\eta_j \sim \mathcal{N}(0, R)$ are i.i.d. for $j \in \mathbb{N}$. Then the corresponding Onsager-Machlup functional is

$$I_n(x) := \|x\|_E^2 + n^2 |y_n - \mathcal{G}(x)|_R^2.$$

Theorem. Assume that $\mathcal{G}\colon X \to \mathbb{R}^K$ is locally Lipschitz continuous. For $n \in \mathbb{N}$, let $x_n \in E$ be a minimiser of I_n. Then

$$\lim_{n\to\infty} \mathcal{G}(x_n) = \mathcal{G}(x^\dagger)$$

almost surely.

Example. Consider again the process $x = (x_t)_{t\in[0,T]}$ defined by the SDE

$$dx_t = f(x_t)\,dt + dW, \quad x_0 = a$$

and assume we want to make inference about the path x based on observations

$$y_j = x_{t_j} + \eta_j$$

where $0 \le t_1 < t_2 < \cdots < t_J \le T$ and $\eta_j \sim \mathcal{N}(0, \gamma^2)$.

To apply our results, we choose μ_0 to be the Wiener measure on $X = L^2([0,T],\mathbb{R})$ with Cameron-Martin space

$$E = \left\{ x \in H^1([0,T],\mathbb{R}) \;\middle|\; x_0 = 0, \int_0^T \dot{x}_t^2\,dt < \infty \right\}.$$

The Onsager-Machlup functional is again

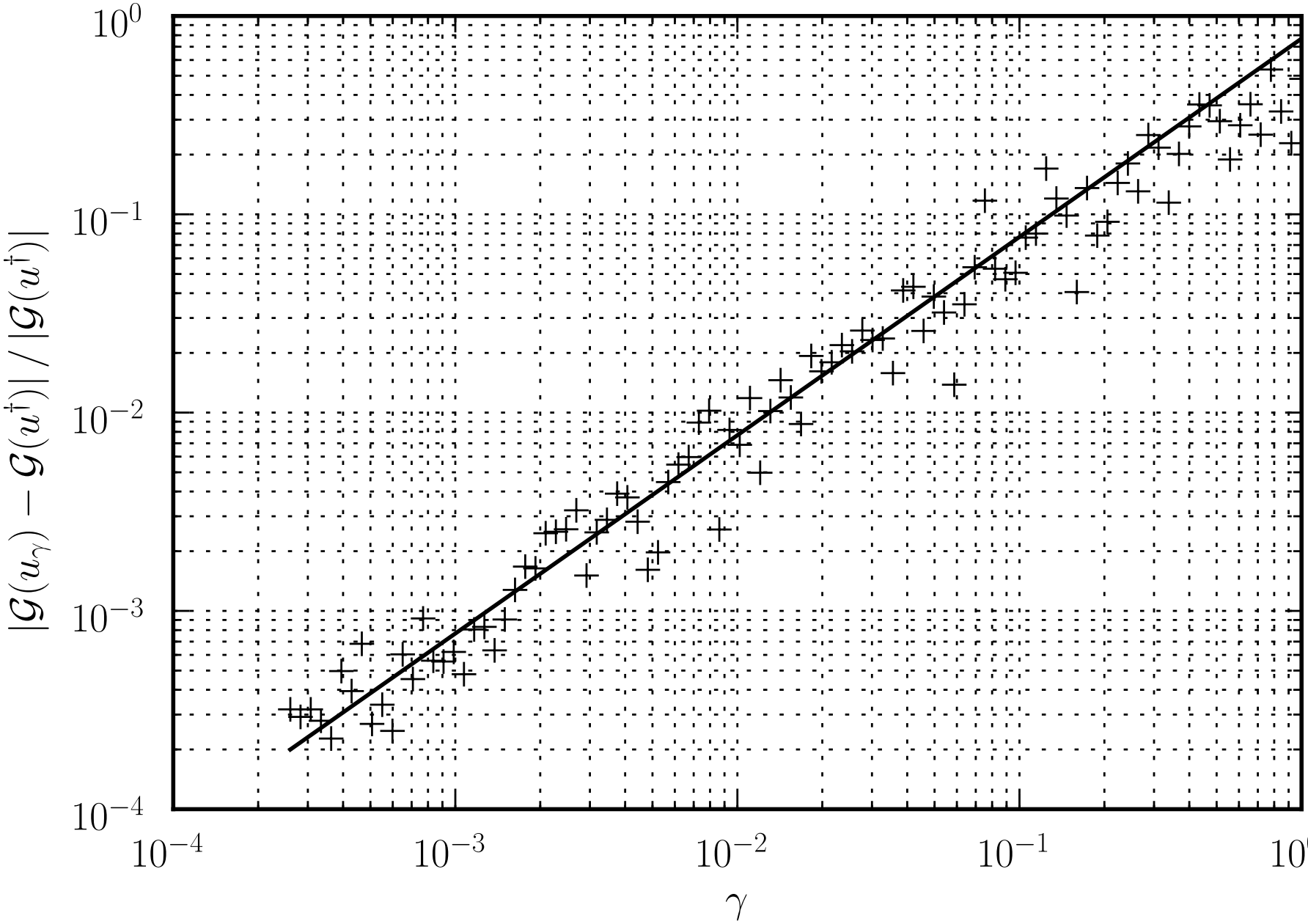
$$I(x) = \Phi(x) + \frac{1}{2}\|x\|_{H^1}^2.$$

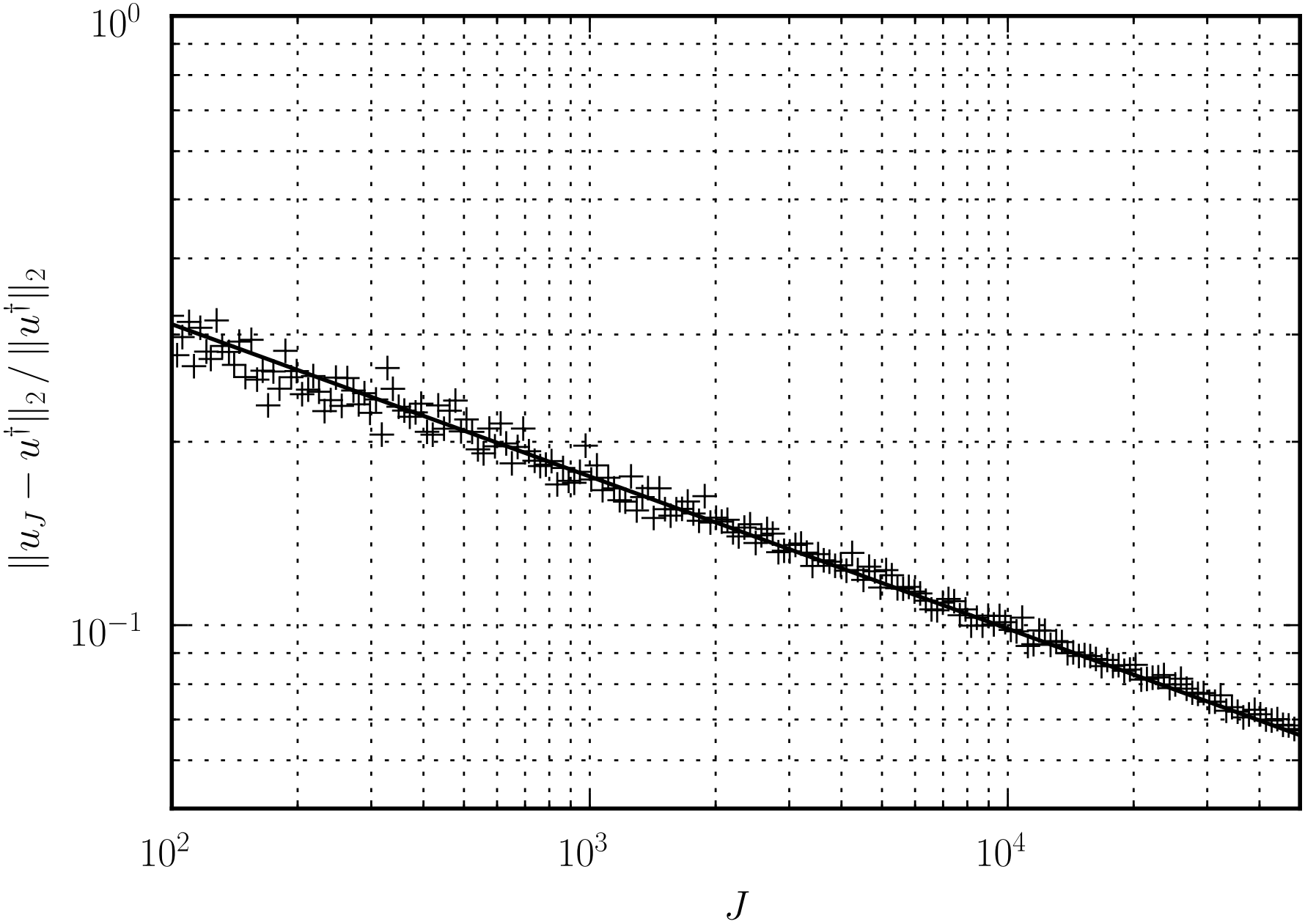
The function Φ incorporates both the density $d\mu/d\mu_0$ (found using the Girsanov formula) and the observations: Assuming that the drift satisfies $f = F'$, we find

$$\Phi(x) = \int_0^T \Psi(x_t)\,dt - F(x_T) + \frac{1}{2\gamma^2}\sum_{j=1}^{J}\left|y_j - x_{t_j}\right|^2$$

where

$$\Psi(x) = \frac{1}{2}\left(|f(x)|^2 + f'(x)\right).$$


10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
$|\mathcal{G}(u_\gamma) - \mathcal{G}(u^\dagger)| \,/\, |\mathcal{G}(u^\dagger)|$
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^0
γ


10^0
10^{-1}
10^2
10^3
10^4
J
$\|u_J - u^\dagger\|_2 \,/\, \|u^\dagger\|_2$

Conclusions

- We have shown that MAP estimators can be used in infinite dimensional problems.
- The 4D-VAR method for data assimilation can be described in this framework.
- The infinite dimensional approach allows for insights into the regularity properties of the problem.

Masoumeh Dashti, Kody J. H. Law, Andrew M. Stuart and Jochen Voss. *MAP Estimators and their Consistency in Bayesian Nonparametric Inverse Problems*. Inverse Problems, vol. 29, 2013.