

# Rising Above Chaotic Likelihoods

Hailiang Du<sup>1,2</sup>      Leonard A. Smith<sup>1,2,3</sup>

<sup>1</sup>Centre for the Analysis of Time Series,

London School of Economics, London WC2A 2AE. UK

<sup>2</sup>Center for Robust Decision Making on Climate and Energy Policy,

University of Chicago, Chicago, IL 60637, US

<sup>3</sup>Pembroke College, Oxford, UK

December 30, 2016

## Abstract

Berliner (Likelihood and Bayesian prediction for chaotic systems, J. Am. Stat. Assoc. 1991) identified a number of difficulties in using the likelihood function within the Bayesian paradigm which arise both for state estimation and for parameter estimation of chaotic systems. Even when the equations of the system are given, he demonstrated “chaotic likelihood functions” both of initial conditions and of parameter values in the Logistic Map. Chaotic likelihood functions, while ultimately smooth, have such complicated small scale structure as to cast doubt on the possibility of identifying high likelihood states in practice. In this paper, the challenge of chaotic likelihoods is overcome by embedding the observations in a higher dimensional sequence-space; this allows good state estimation with finite computational power. An importance sampling approach is introduced, where Pseudo-orbit Data Assimilation is employed in the sequence-space, first to identify relevant pseudo-orbits and then relevant trajectories. Estimates are identified with likelihoods orders of magnitude higher than those previously identified in the examples given by Berliner. Pseudo-orbit Data Assimilation importance sampler exploits the information both from the model dynamics

and from the observations. While sampling from the relevant prior (here, the natural measure) will, of course, eventually yield an accountable sample, given the realistic computational resource this traditional approach would provide no high likelihood points at all. While one of the challenges Berliner posed is overcome, his central conclusion is supported. “Chaotic likelihood functions” for parameter estimation still pose a challenge; this fact helps clarify why physical scientists maintain a strong distinction between the initial condition uncertainty and parameter uncertainty.

## 1 Introduction

Nonlinear chaotic systems pose several challenges both for state estimation and for parameter estimation. Chaos as a phenomenon implies sensitive dependence to initial condition: initially nearby states will eventually diverge in the future. The bifurcations of various chaotic systems [30] reveal how the behavior of the system differs as a parameter value changes. One might think that Bayesian analysis should be able to obtain good estimation both of initial conditions and of parameter values without much trouble. Berliner [2] examined the log-likelihood function of estimates of initial conditions and parameter values for the Logistic Map, noting that chaotic systems can lead to “chaotic likelihood functions” and suggesting that Bayesian analysis would require prohibitively intensive computing. The failure of variational approaches, when applied to long window observations of chaotic systems [7, 11, 20], supports his point. Sensitivity to initial condition on the other hand suggests that information in the observations (even over a relatively short window) can lead to good estimates of the initial condition [12]. An importance sampling approach to Berliner’s challenge without “intensive computing” is deployed in this paper. Adopting Pseudo-orbit Data Assimilation (PDA) [7, 11] recasts the problem into a higher dimensional sequence space, where truly high likelihood states are successfully located. The challenges of initial condition estimation and parameter estimation are dissimilar for chaotic systems. PDA does not easily generalize to parameter estimation, as it is unclear how to mathematically define a relevant subspace of the parameter space in which the high likelihood trajectories

might exist. Thus challenges remain in identifying high likelihood parameter values given the initial condition; this asymmetry reflects differences between the initial conditions and parameter values. In terms of estimating initial conditions given the parameter values, however, Berliner’s challenge as originally stated is met and resolved.

## 2 Chaotic Likelihood Function of Initial Conditions

Berliner’s [2] formulation of the problem is adapted here. The Logistic Map is the system; in sections 2-4, the parameter  $a = 4$  is known but the true initial state  $\tilde{x}_0$  is not. In that case, the experiment is said to fall within the perfect model scenario<sup>1</sup>. In the first four sections of this paper, the mathematical system and the model are identical, the word model is not used. Once real systems are considered, either structural model error or parameter inaccuracy requires one to distinguish the system which generated the data from the model(s) employed to analyze it.

The evolution of system states<sup>2</sup>  $x_i \in \mathbb{R}^m$  is then governed by the nonlinear dynamics  $f : x_{i+1} = f(x_i)$ , where for the Logistic Map

$$f(x_i) = ax_i(1 - x_i). \tag{1}$$

Assuming additive observational noise,  $\delta_i$ , yields observations,  $s_i = \tilde{x}_i + \delta_i$  where  $\tilde{x}$  is the true system state (Truth) and the observational noise,  $\delta_i$ , is Independent Normally Distributed (IND),  $\delta_i \sim N(0, \sigma^2)$ . Under this assumption of normality, the log-likelihood function is:

$$LLik(x_0) = - \sum_{i=1}^{n-1} (s_i - f^i(x_0))^2 / 2\sigma^2, \tag{2}$$

where  $f^i(x)$  is the  $i^{th}$  iteration of  $f(x)$ ,  $s_i$  is the  $i^{th}$  observation, and  $n$  is the duration of observations considered.

---

<sup>1</sup>The perfect model scenario, when the system and the model are identical, eases use of a digital computer; one avoids the issue of “round-off error” by consistently use of the same computer and code. This implies, of course, that Equation 1 no longer reflects the perfect model. For discussion, see [1, 27, 28] and references thereof.

<sup>2</sup>For  $m = 1$ , the state  $x_i$  is a scalar.

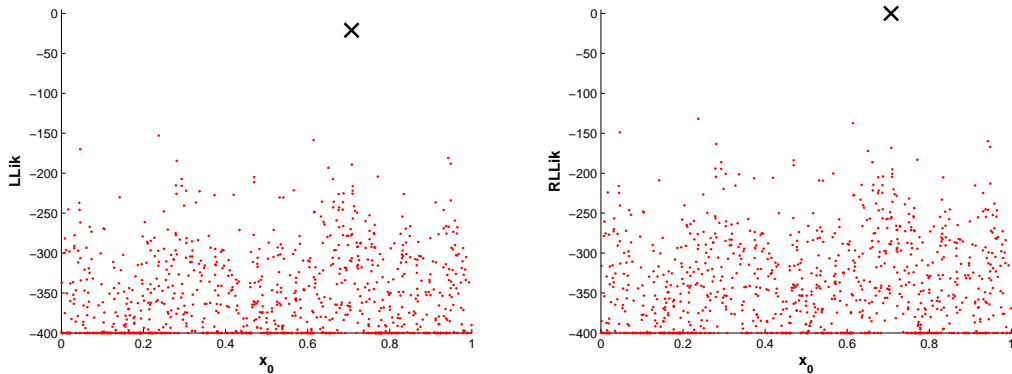


Figure 1: Typical log-likelihood of 1024 states (uniformly distributed on  $[0, 1]$ ) for the Logistic Map. The true initial condition  $\tilde{x}_0 = \sqrt{2}/2$  (denoted by ‘ $\times$ ’),  $\sigma = 0.1$  and  $n = 32$ . a) Log-likelihood function, b) Relative log-likelihood to  $\tilde{x}_0$ . States which have (relative) log-likelihood less than  $-400$  are plotted on the  $-400$  horizontal line. All logarithms are using natural base. The maximum likelihood value found was  $\sim e^{-109}$ , which corresponds to log-likelihood of  $-132$  relative to Truth.

Figure 1 shows the chaotic likelihood structure of 1024 candidate values of  $x_0$  drawn uniformly on the interval zero, one. Panel (a) plots the log-likelihood for each  $x_0$ , this can be contrasted with various panels in Berliner’s [2] Figure 3.<sup>3</sup> Panel (b) shows the log-likelihood relative to that of the true trajectory of the system state<sup>4</sup>. For the convenience of illustration, the same normalization used in panel (b) is applied in Figures 2-4 in this paper. From Figure 1, it is clear that no high likelihood states are identified. This is not a

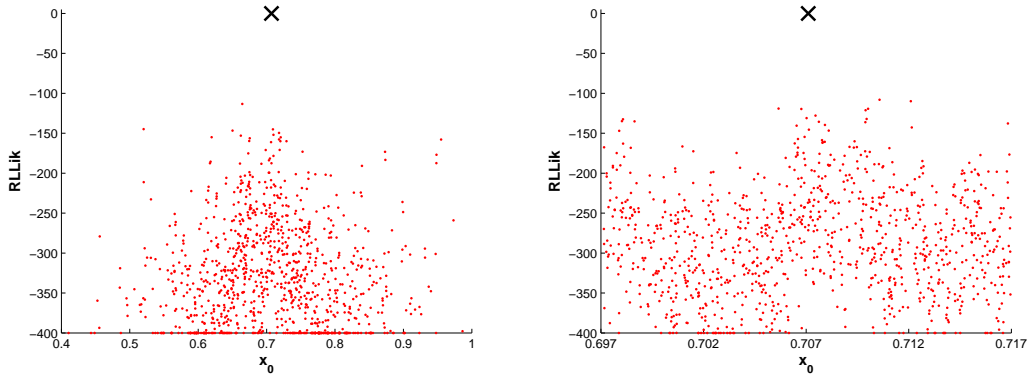


Figure 2: Relative log-likelihood of 1024 states a) sampled from inverse observational noise, b) uniformly sampled from  $[\tilde{x}_0 - \frac{\sigma}{10}, \tilde{x}_0 + \frac{\sigma}{10}]$ . The true initial condition  $\tilde{x}_0 = \sqrt{2}/2$  (denoted by ‘x’),  $\sigma = 0.1$  and  $n = 32$ .

case of equifinality<sup>5</sup>.

Given the observational noise distribution, one can add random draws from the inverse of the observational noise distribution to the observation to obtain candidate estimates of initial condition. Figure 2a shows the relative log-likelihood of 1024 samples from inverse observational noise. No high likelihood states are identified in this way. (The maximum likelihood value found was  $\sim e^{-90}$ .) To illustrate the impact of making much more precise observations, consider a case where  $\tilde{x}_0$  is known to be within a region of radius only  $\sigma/10$ . Figure 2b shows the relative log-likelihood of 1024 uniformly sampled states in the region around Truth with  $\sigma/10$  radius. Yet again, no high likelihood state are identified. (The

---

<sup>3</sup>Here log-likelihood functions based on a sequence of 32 observations are computed because problem becomes more obvious when more observations are used. Berliner examined 15 & 10 observations. Shorter sequences of observations are examined in Section 3.

<sup>4</sup>Note: given that only finite observations are considered the true state of the system is, with probability 1, not the maximum likelihood state.

<sup>5</sup>Equifinality occurs when many potential solutions to a task are each good, making it impossible to identify the true solution given the information in hand. In such cases the sampled likelihood functions are relatively flat. Equidismality arises when the sampled relative likelihood function is flat yet all solutions tested have vanishingly small likelihood given the information. Examining the relative likelihood obscures the difference; fortunately the expected (distribution of) likelihood can be computed from the noise model alone without knowledge of the true initial condition.

maximum likelihood value found increases to  $\sim e^{-85}$ .)

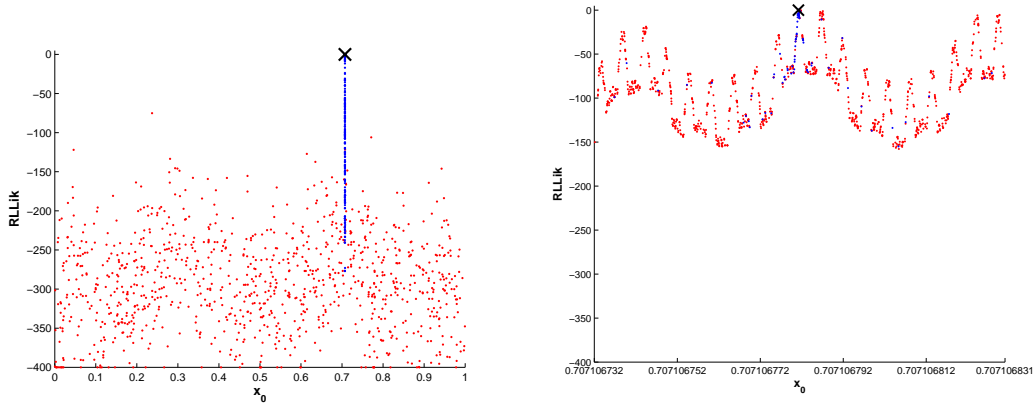


Figure 3: a) Following Figure 1b, but in this case including a set of 1024 states (blue), which are extremely close to  $\tilde{x}_0$ , generated by spiral sampling around the  $\tilde{x}_0$  (see footnote 6); b) zoom in of a).

This difficulty here is not a shortcoming of the likelihood framework, as there are high likelihood states other than Truth. One may demonstrate that such high likelihood states exist by sampling the points on a logarithmic spiral approaching Truth (to machine precision)<sup>6</sup>. Figure 3 demonstrates that high likelihood states other than Truth do exist (i.e. some of the blue points). A smooth curve of the log-likelihood function is only observed within a radius of  $\tilde{x}_0$  smaller than  $\sim 10^{-7}$ , (see Figure 3b).

Without knowing Truth, of course, this approach to identifying high likelihood points is inaccessible. The likelihood function is extremely jagged; as Berliner [2] stressed, finding even one high likelihood state by sampling the state space is prohibitively costly. That said, there is no sense in which “sensitivity to the initial conditions” can be taken to imply that the information in the initial condition is “forgotten” or “lost”. There is sufficient information in the observation segment to distinguish high likelihood initial states. Candidate states with non vanishing log-likelihood relative to Truth can be found by extracting the information from the system dynamics using a relatively new approach to data assimilation:

---

<sup>6</sup>In this experiments 1024 points are generated by  $\tilde{x}_0 + 2^{-(10 + \frac{60i}{1024})} \epsilon_i$ ,  $i = 1, 2, \dots, 1024$  where  $\epsilon_i$  is random drawn from  $U(0, 1)$ .

### 3 Importance Sampling via Pseudo-orbit Data Assimilation

#### 3.1 Methodology

Uniform sampling in state space is not an efficient approach to locating high likelihood states. As the dimension of the system increases, the task becomes even more computationally impractical. PDA importance sampler<sup>7</sup> locates high likelihood states in the trajectory manifold by adopting the Pseudo-orbit Data Assimilation approach [7, 11]. PDA takes advantage of the fact that solutions consistent with the known dynamics will lie along a lower dimensional manifold in the higher dimensional sequence space. A brief introduction of the PDA approach is given in the following paragraph (see [7, 11] for additional details).

Given a dynamical system of dimension  $m$  and a sequence of  $n$  observations<sup>8</sup>, define a sequence space as the  $m \times n$  dimensional space in which a single point can be thought of as a particular series of  $n$  states  $\mathbf{u}_i$ , for  $i = 0, \dots, n - 1$  where  $\mathbf{u}_i$  is an  $m$  dimensional vector. Most points in sequence space do not correspond to a trajectory of the system. Define a *pseudo-orbit*,  $\mathbf{U} \equiv \{\mathbf{u}_0, \dots, \mathbf{u}_{n-2}, \mathbf{u}_{n-1}\}$ , to be a point in the  $m \times n$  dimensional sequence space for which  $\mathbf{u}_{i+1} \neq f(\mathbf{u}_i)$  for one or more components of  $\mathbf{U}$ . Thus a pseudo-orbit corresponds to a sequence of states which is **not** a trajectory of the system. Each sequence of  $n$  observations  $\mathbf{s}_i, i = 0, \dots, n - 1$  defines a pseudo-orbit (a point in the sequence space). Call this an *observed pseudo-orbit*,  $\mathbf{S} \equiv \{\mathbf{s}_0, \dots, \mathbf{s}_{n-2}, \mathbf{s}_{n-1}\}$ , which with probability one will not be a trajectory.

---

<sup>7</sup>In high-dimensional space the sampler targets the relevant lower-dimensional trajectory manifold which is more efficient than sampling a hypersphere suggested by one observation. Even in the one-dimensional Logistic Map this approach succeeds by using PDA to sample the trajectory manifold in the  $n$ -dimensional sequence space.

<sup>8</sup>For the Logistic Map, observations and system states share the same one dimensional space.

Define the *mismatch* to be:

$$e_i = | f(\mathbf{u}_i) - \mathbf{u}_{i+1} | \quad (3)$$

By construction, trajectories have a mismatch of zero. The mismatch cost function is then given by:

$$C(\mathbf{U}) = \sum_{i=0}^{n-1} e_i^2 \quad (4)$$

The most common form of Pseudo-orbit Data Assimilation simply minimizes the mismatch cost function for  $\mathbf{U}$  in the  $m \times n$  dimensional sequence space starting from the observed pseudo-orbit. If a gradient descent (GD) approach is adopted<sup>9</sup>, then a minimum of the mismatch cost function can be obtained by solving the ordinary differential equation

$$\frac{d\mathbf{U}}{d\tau} = -\nabla C(\mathbf{U}), \quad (5)$$

where  $\tau$  denotes algorithmic time.<sup>10</sup> In practice, the minimization is initialized with the observed pseudo-orbit, i.e.  ${}^0\mathbf{U} = \mathbf{S}$  where the pre-super-script zero on  $\mathbf{U}$  denotes the initial algorithmic time ( $\tau = 0$ ) of the GD. Under gradient descent, a point in sequence space moves towards the trajectory manifold under Equation 5. The minimization algorithm requires differentiating the mismatch cost function (Equation 4), which gives:

$$\frac{\partial C(\mathbf{U})}{\partial \mathbf{u}_i} = 2 \times \begin{cases} -(\mathbf{u}_{i+1} - f(\mathbf{u}_i))J(\mathbf{u}_i) & i = 0 \\ -(\mathbf{u}_i - f(\mathbf{u}_{i-1})) + (\mathbf{u}_{i+1} - f(\mathbf{u}_i))J(\mathbf{u}_i) & 0 < i < n - 1 \\ -\mathbf{u}_i - f(\mathbf{u}_{i-1}) & i = n - 1 \end{cases} \quad (6)$$

where  $J(\mathbf{u}_i)$  is the Jacobian of  $f$  at  $\mathbf{u}_i$ . The ordinary differential Equation 5 is solved using the Euler approximation in the examples below.

The mismatch cost function has no local minima other than points on the manifold, for which  $C(\mathbf{U}) = 0$ , (this defines the trajectory manifold<sup>11</sup>) and of course every segment

---

<sup>9</sup>Other methods for this minimization are available, GD is discussed here due to its simplicity and adequacy.

<sup>10</sup>The approach can be generalized to situations where gradient is not known analytically [13]; improving the ability to work without gradient information would widen the application of the approach significantly.

<sup>11</sup>Back-substitution of the solution of  $-\mathbf{u}_{n-1} - f(\mathbf{u}_{n-2}) = 0$  into Equation 6 shows that the only critical points for  $C(\mathbf{U})$  have  $\mathbf{u}_i - f(\mathbf{u}_{i-1}) = 0$  for all  $i$  in  $0 \leq i \leq n - 1$ . All points on the trajectory manifold have zero mismatch (are trajectories) and only points on the trajectory manifold have zero mismatch.



of trajectory lies on this manifold [11]. Denote the result of the GD minimization at algorithmic time  $\tau = \alpha$  as  ${}^\alpha\mathbf{U} \equiv {}^\alpha \mathbf{u}_0, \dots, {}^\alpha \mathbf{u}_{n-1}$ . Here  $\alpha$  indicates *algorithmic time* in GD (i.e. the number of iterations of the GD minimization). As  $\alpha \rightarrow \infty$ , the pseudo-orbit  ${}^\alpha\mathbf{U}$  approaches a trajectory of the model asymptotically. In other words, the GD minimization takes the observed pseudo-orbit towards a model trajectory ( ${}^\infty\mathbf{U}$  is a point in sequence space on the trajectory manifold).

In practice, the GD algorithm is run for a finite time and a trajectory is not obtained. Nevertheless, the values of  ${}^\alpha u_0$  for large  $\alpha$ , define candidates trajectories using information from the observation window  $0 \leq i \leq n - 1$ . For  $i > 0$ , the  $i$ -step preimages of the relevant component of  ${}^\alpha u_i$  provide candidates for the initial state. Complications arise from the fact that the Logistic Map is two-to-one; these complications have nothing to do with chaos per se<sup>12</sup>. In one-to-one chaotic maps the calculation of preimages is straightforward. The Logistic Map is a two-to-one map, and in most cases<sup>13</sup> only one of the two preimages for each  ${}^\alpha u_i$  is relevant to  $\tilde{x}_{i-1}$ . In practice a threshold criteria to discard irrelevant preimages must be defined, a simple example would be to discard (with high probability) those preimages whose distance from the corresponding previous observation exceeds some threshold based on the properties of the observational noise (a  $3\sigma$  criteria is adopted in the following section and shown to be adequate for purpose).

## 3.2 Results

The green points in Figure 4 were located using PDA importance sampler; note that some have relative log-likelihood near zero (maximum 1.3). As expected, the observations do not contain sufficient information to identify the state of the system at the time of the final observation with the same degree of precision as the initial state. This is reflected in the fact that the green points are much less close to the true state at time 31 (Figure 4b) than the corresponding green points at time 0 (Figure 4a).

Two experiments were conducted to test the robustness of the PDA importance sam-

---

<sup>12</sup>Beyond the fact that one-to-one maps in one-dimension cannot display chaotic dynamics, of course.

<sup>13</sup>Not in all cases, however. For discussion of the point see [15].

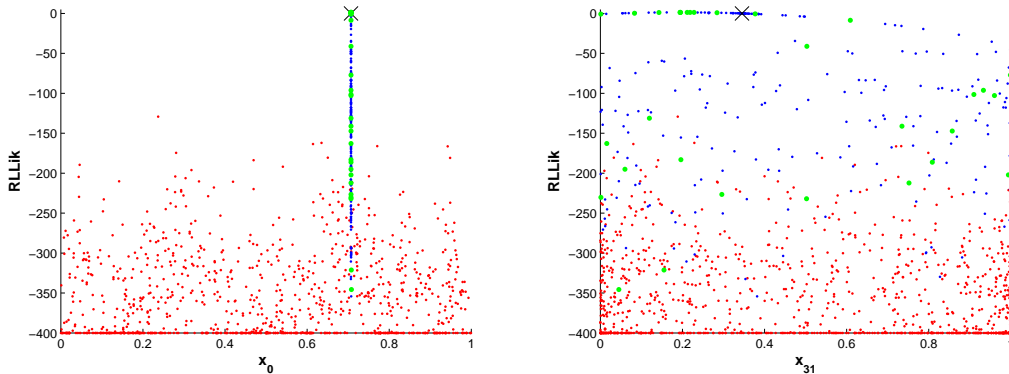


Figure 4: a) Following Figure 3a, but in this case including the states located by PDA importance sampler are plotted in green. b) The forward images of those states plotted in (a) at time 31.

pling approach. The first is based on 2048 different realizations of observations given the same initial condition,  $\tilde{x}_0 = \sqrt{2}/2$ , to examine consistency. The second considers 2048 different initial conditions to examine robustness over different states. Three different observation window lengths were used in each experiment. Table 1 and Table 2 shows the results. States with greater likelihood than the true state are often identified.

Given uncertain observations, one can never identify Truth of a chaotic system unambiguously; this was noted by Berliner & MacEachen [3, 18], then Lalley [15, 16] and later explored by Judd & Smith [11]. Using the PDA approach, high likelihood states are indeed found: states with relative log-likelihood larger than minus one are found in every single experimental run. The fact that some PDA states have greater likelihood than Truth reflects the fact that Truth is not expected to be the most likely model state given the observations.

For each experimental run, the minimum distance between those states obtained by PDA importance sampler (whose relative log-likelihood larger than minus one) and Truth is recorded. The minimum, maximum and median statistical values of the minimum distance from Truth are reported in Table 1 and 2. It is clear that the PDA importance sampler identifies states at higher quality (the minimum distance from Truth decreases) as the observation window length increases. This is expected inasmuch as more information from

the dynamics is available given a longer window. Table 1 shows that the maximum value of the minimum distance among the 2048 different realizations is  $1.49 \times 10^{-10}$  for a window length of 32 and in Table 2 the maximum value of the minimum distance among different true initial conditions is  $2.02 \times 10^{-10}$ . The PDA importance sampler is both robust and efficient in this case.<sup>14</sup>

Window length	# of RLLik > -1			# of RLLik > 0			Minimum distance to $\tilde{x}_0$		
	Min	Max	Median	Min	Max	Median	Min	Max	Median
32	6	15	8	0	14	6	$2.00 \times 10^{-15}$	$1.49 \times 10^{-10}$	$9.98 \times 10^{-12}$
16	2	11	8	0	11	6	$1.57 \times 10^{-10}$	$7.63 \times 10^{-6}$	$5.13 \times 10^{-7}$
8	2	7	7	0	7	5	$8.58 \times 10^{-8}$	$4.55 \times 10^{-2}$	$2.56 \times 10^{-4}$

Table 1: Statistics of high likelihood states located by PDA importance sampler based on 2048 different realizations of observations (of  $\tilde{x}_0 = \sqrt{2}/2$ ) for the Logistic Map, i) statistics of the number of states (whose log-likelihood relative to the true state larger than minus one) ii) statistics of the number of states (whose relative log-likelihood larger than zero) iii) statistics of the minimum distance between the states (whose relative log-likelihood larger than minus one) and Truth.

The experiments above solve by demonstration Berliner’s “impossible” challenge; they demonstrate that truly high likelihood points can be located using dynamical information, easing Berliner’s identification problem of initial condition. Selecting an ensemble from this high likelihood set allows for informative forecasts which do not become useless until after those from the point forecasts illustrated by Berliner [2] become uninformative.

---

<sup>14</sup>Drawing samples uniformly from within a distance of 0.1 of Truth would require  $\sim 10^8$  candidates in order to find a candidate within  $\sim 2 \times 10^{-10}$  of Truth; even then that close candidate need not have high likelihood. The results of Table 1 and 2 were obtained with only 1024 GD minimization iterations in each realization (each and every one of which identified high likelihood states close to Truth).

Window length	# of RLLik > -1			# of RLLik > 0			Minimum distance to $\tilde{x}_0$		
	Min	Max	Median	Min	Max	Median	Min	Max	Median
32	6	360	28	0	338	16	$4.77 \times 10^{-15}$	$2.02 \times 10^{-10}$	$1.50 \times 10^{-11}$
16	3	56	14	0	48	9	$3.04 \times 10^{-10}$	$1.54 \times 10^{-5}$	$9.93 \times 10^{-7}$
8	2	20	7	0	20	7	$6.36 \times 10^{-8}$	$5.60 \times 10^{-3}$	$3.32 \times 10^{-4}$

Table 2: Statistics of high likelihood states located by PDA importance sampler based on 2048 different true initial states for the Logistic Map, i) statistics of the number of states (whose log-likelihood relative to the true state larger than minus one) ii) statistics of the number of states (whose relative log-likelihood larger than zero) iii) statistics of the minimum distance between the states (whose relative log-likelihood larger than minus one) and Truth.

### 3.3 Equifinality and Relative Likelihood

Maximum likelihood estimation has been widely used [25] since its introduction by Fisher [8] in 1922. The “best” estimate is often chosen from a set of candidates, and only the relative likelihood of candidates within that sample is considered. Figure 5 shows the log-likelihood of 1024 states (the same set used in Figure 2b), the red dashed line is the median log-likelihood of those states. In this case, the problem is not one of which estimate to select (equifinality), but one of recognizing that each and every candidate has vanishing likelihood (equidismality). In practice, Truth is unknown therefore it cannot be used as a reference, as it is in Figure 2. Given the observations and the noise model, however, the expected log-likelihood of Truth can be determined<sup>15</sup> and serve as a reference. Figure 5, the log-likelihood of 1024 states are plotted along with the expected log-likelihood of Truth (black dashed line). Figure 5 shows that it is not a case of equifinality but a case of equidismality. In cases where it is observed that all traditional candidate states have vanishingly small log-likelihood relative to the expected log-likelihood of Truth, a PDA based importance

<sup>15</sup>The log-likelihood of Truth is  $-\frac{\sum_{i=0}^{n-1} \delta_i^2}{2\sigma^2}$  (from Equation 2) where  $\delta_i$  (observation noise) is  $IID \sim N(0, \sigma^2)$  distributed. Let  $Z = \frac{\sum_{i=0}^{n-1} \delta_i^2}{\sigma^2}$ ,  $Z$  is a random variable following chi-squared distribution with  $n$  degrees of freedom. Statistics of the log-likelihood of Truth can therefore simply derived from  $Z$ .

sampling might prove of valuable.

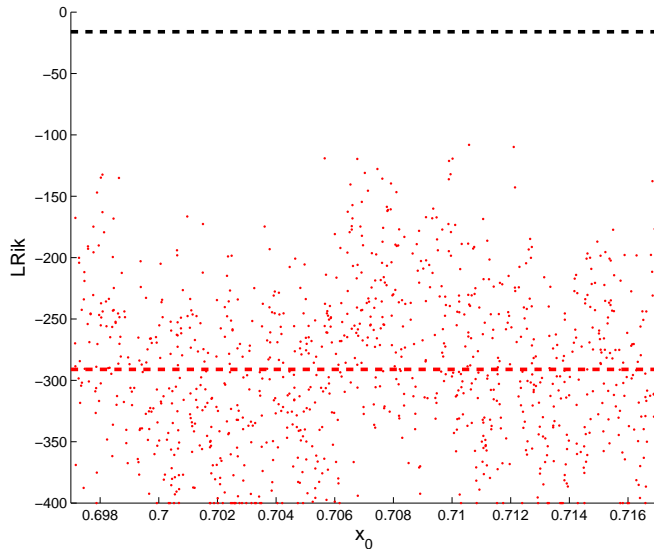


Figure 5: Log-likelihoods of 1024 states uniformly sampled from  $[\tilde{x}_0 - \frac{\sigma}{10}, \tilde{x}_0 + \frac{\sigma}{10}]$ , the red dashed line is the median log-likelihood of those states and the black dashed line is the expected log-likelihood of Truth, following Figure 2b.

## 4 Implications for Quantifying Uncertainty in Practice

Having solved one of Berliner’s “impossible” challenges for nonlinear systems, a reviewer asks for potential implications this solution might hold in practice. Two possibilities stand out. The first regards improved noise reduction on the initial condition of the forecast via a PDA importance sampler; this would indeed be expected to improve forecast skill. Berliner’s problem, however, focuses on identification of the point at the starting (oldest) observation, while weather forecasts are launched near the time of the end (most recent) observation. In operational weather forecasting, a relatively small set of initial conditions ( $\sim 51$ ) are launched as forecasts in a relatively high dimensional space ( $\sim 10^7$ ) in a Monte Carlo fashion every forecast cycle (typically every 6 hours); see [10, 17, 31]. As is long held

in weather forecasting (see [9] for a discussion), uncertainty in the state at the start of the observation window in time collapses much more dramatically than that in the state at the end (present time); this is consistent with results shown in Figure 4. Figure 2 of [9] captures challenge Berliner & MacEachen [3, 18] , Hansen & Smith [9] and Lalley [16] foresaw. High likelihood states at present time lie along the unstable manifold, such sets of states are more efficient and informative, in terms of exploring initial condition uncertainty, than traditional approaches like Singular Vectors [23] or Bred Vectors [32].

A second possibility lies in reanalysis [24]. A reanalysis uses the most modern weather model to revisit and improve estimate of the state of the atmosphere of the distant past; here the opportunity for PDA importance sampler to make a contribution appears greater. A reanalysis takes a modern model and reconsiders past periods of time; reanalysis now common in atmospheric (weather) and oceanic contexts [14, 22, 33]. This task closely resembles that posed by Berliner. The results achieved in Table 1 and Table 2 reflect the aims of reanalysis directly, abet in a perfectly known one-dimensional system. Interpretations in terms of scalar “noise reduction” are limited, as there can be information gained simply by restricting the state estimate to the model manifold.

Reanalysis have many applications, one of significant current interest is index based insurance. In this case, uncertainty quantification of the index would be of significant interest. In short, the complications of evaluating insurance claims based on certifying damages in far flung regions of the world has lead to the insurance policies triggered by a model based meteorological index. [34] To the extent that the methods deployed in this paper better quantify (and reduce) the uncertainty in the wind speeds of a storm within a few days of its passing, the efficiency of index based insurance could be improved.

Note from Table 1 and 2 that the longer the window length the better estimate (noise reduction) is achieved. Noise reduction based upon PDA importance sampler might also help fill in missing historical observations, as well identify model deficiency when identifying model trajectories intended to reflect past observations.

## 5 Distinguishing between initial conditions and parameters

The likelihood functions of initial conditions and that of parameter values have similar features [2]. There are, however, fundamental differences in the information available to address these two distinct estimation problems.

Given the structure of the model class, the model parameter value determines dynamical behaviors of the model (e.g. natural measures) which are not changed by the initial condition. Given the model and “true” parameter value, the search for unknown initial conditions is aided by the restriction the nature measure(s) places on candidate initial conditions in the state space (and, thereby, on trajectories in the sequence space).

It is unclear how to construct similar constraints on unknown parameter values in the parameter space given the “true” initial state (if they exist<sup>16</sup>). Uncertainty in initial state differs from uncertainty in the parameter value. The information in a *measurement* of the initial condition uncertainty will decay with time and eventually become statistically indistinguishable from a random sample of the natural measure, while the information in the initial condition is preserved, arguably forever.

While assuming the parameter value is perfect might initially appear extreme, it appears less nonsensical given that one has already assumed that the model structure is perfect. Assuming the initial state is perfect indicates a noise free observation is possible. Let the model’s parameters be contained in the vector  $\mathbf{a} \in \mathbb{R}^l$ . A set of  $l + 1$  sequential noise free observations  $s_i, s_{i+1}, \dots, s_{i+l}$  would, in general, be sufficient to determine  $\mathbf{a}$  [19]. Thus if one noise free observation is obtainable, obtaining only a few more noise free observations would define the true parameter value precisely. Perhaps a more realistic way to put the problem is to estimate the parameter value(s) given realistic observations, without assuming the “true” initial condition is known. In that case, the goal is to locate high likelihood trajectories

---

<sup>16</sup>Neither is it clear how to construct the set of parameter values whose corresponding invariant measure contains the “true” initial state, or how to exploit this set; what is clear is how to exploit the existence of trajectory manifold given a particular value of the parameter.

(Smith et al. [29] call these shadowing trajectories) for particular sets of parameter values. It is not yet clear how to solve such a problem. In fact, it is not clear if it is possible to constrain the solution in the parameter space in a manner analogous to the constraints in the space of initial condition achieved by using trajectory manifold in the state space.

Given a perfect model structure and knowing the true parameter value(s), the true initial state is a well defined goal of the identification. Inasmuch as structural model errors imply no true parameter value exists [6, 12], it is unclear how one might define “true” initial state and the goal of estimation must be rethought when the structure of the system differs from that of the model.

Despite the importance of model parameters, outside linear systems there is no general method of parameter estimation<sup>17</sup>. Methods have been developed to obtain useful parameter values with some success: McSharry and Smith [19, 26] estimate model parameters by incorporating the global behaviour of the model into the selection criteria; Creveling et al. [4] have exploited synchronization for parameter estimation; Smith et al. [29] focused on the geometric properties of trajectories; Du and Smith [5] select parameter values based on the Ignorance score of ensemble forecasts. Each of these methods, however, require a large set of observations. Challenges remain when only a short sequence of observations is available. Berliner’s second challenge still stands.

## 6 Conclusion

Berliner illustrated that even in the perfect model scenario traditional approaches are unable to provide good estimates of the initial condition for nonlinear chaotic systems. In large part, those failures are due to the inability of traditional approaches to skillfully meld the information in the dynamics of the nonlinear system itself with that in the observations. The importance sampling approach presented here uses Pseudo-orbit Data Assimilation to

---

<sup>17</sup>That is not to dismiss important advances in parameter estimation when attention is restricted to certain particular classes of model structure [21]; our claim here is that no general approach to parameter estimation is known which is applicable to nonlinear models in general.



combine information from both observations and dynamics more effectively, thereby locating high likelihood initial states; this achieves one aim Berliner (1991) argued to be impossible. As discussed in Section 5, this achievement may prove useful in reanalysis studies.

Despite the similarity of state estimation and parameter estimation, there are fundamental differences between uncertainty in the initial state and uncertainty in parameter value. Significant obstacles remain in solving Berliner's second challenge regarding parameter estimation while Pseudo-orbit Data Assimilation overcomes his first challenge regarding initial condition.

## Acknowledgment

This research was supported by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. Additional support for H.D. was also provided by the National Science Foundation Award No. 0951576 "DMUU: Center for Robust Decision Making on Climate and Energy Policy (RDCEP)". L.A.S. gratefully acknowledges the continuing support of Pembroke College, Oxford. We also acknowledge the insightful comments of Michael Stein and Mark Berliner in discussions of this work.

## References

- [1] C. Beck. Effects of roundoff errors on chaotic dynamics. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI - Theories and Applications Proceedings of EUSIPCO-92*, volume 1, page 183186, 1992.
- [2] M.L. Berliner. Likelihood and bayesian prediction for chaotic systems. *J. Am. Stat. Assoc.*, 86:938952, 1991.
- [3] M. L. Berliner and S. N. MacEachern. Examples of inconsistent bayes procedures

- based on observations on dynamical systems. *Statistics and Probability Letters*, 314 17(5):355360, 1993.
- [4] D.R. Creveling, P.E. Gill, and H.D.I. Abarbanel. State and parameter 258 estimation in nonlinear systems as an optimal tracking problem. *Physics Letters A*, 372:26402644, 2008.
- [5] H. Du and L.A. Smith. Parameter estimation through ignorance. *Phys. Rev. E*, 016213, 2012.
- [6] H. Du and L.A. Smith. Pseudo-orbit data assimilation. II: Assimilation with imperfect models. *Journal of the Atmospheric Sciences*, 71:483495, 2014.
- [7] H. Du and L.A. Smith. Pseudo-orbit data assimilation part I: The perfect model scenario. *Journal of the Atmospheric Sciences*, 71:469482, 2014.
- [8] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222:309368, 1922.
- [9] J. A. Hansen and L. A. Smith. Probabilistic noise reduction. *Tellus*, 5(53):585598, 2001.
- [10] Brian Hoskins. The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672):573584, 2013.
- [11] K. Judd and L.A. Smith. Indistinguishable states I: The perfect model scenario. *Physica D*, 151:125141, 2001.
- [12] K. Judd and L.A. Smith. Indistinguishable states II: the imperfect model scenario. *Physica D*, 196:224242, 2004.
- [13] Kevin Judd, Leonard Smith, and Antje Weisheimer. Gradient free descent: shadowing, and state estimation using limited derivative information. *Physica D: Nonlinear Phenomena*, 190(34):153166, April 2004.

- [14] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, Roy Jenne, and Dennis Joseph. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437471, 1996.
- [15] S.P. Lalley. Beneath the noise, chaos. *The Annals of Statistics*, 27(2):461479, 1999.
- [16] S.P. Lalley. Removing the noise from chaos plus noise. In A.I. Mees, editor, *Nonlinear Dynamics and Statistics*, Boston, 2000.
- [17] M. Leutbecher and T. Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227:35153539, 2008.
- [18] S.N. MacEachern and L.M. Berliner. Asymptotic inference for dynamical systems observed with error. *Journal of Statistical Planning and Inference*, 46:277292, 1995.
- [19] P. McSharry and L.A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Phys. Rev. Lett*, 83(21):42854288, 1999.
- [20] R.N. Miller, M. Ghil, and F. Gauthiez. Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.*, 51:10371056, 1994.
- [21] Noel A. C. Cressie and Christopher K. Wikle. *Statistics for spatio-temporal data*. Wiley series in probability and statistics. Hoboken, N.J. Wiley, 2011.
- [22] P.R. Oke, G.B. Brassington, D.A. Griffin, and A. Schiller. The bluelink ocean data assimilation system (bodas). *Ocean Modelling*, 21:4670, 2008.
- [23] T. N. Palmer, R. Buizza, F. Molteni, Y.-Q. Chen, and S. Corti. Singular vectors and the predictability of weather and climate. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 348(1688):459475, 1994.

- [24] W. Parker. Computer simulation, measurement, and data assimilation. *British Journal for the Philosophy of Science*, 2015.
- [25] Y. Pawitan. In all likelihood : statistical modelling and inference 283 using likelihood. Oxford science publications. Clarendon press, Oxford, 2001.
- [26] V.F. Pisarenko and D. Sornette. Statistical methods of parameter estimation for deterministically chaotic time series. *Phys. Rev. E*, 69:036122, 2004.
- [27] L.A. Smith. Lacunarity and Chaos in Nature. PhD thesis, Columbia University, New York, 1987.
- [28] L.A. Smith. What might we learn from climate forecasts? In *Proc. National Acad. Sci.*, volume 4, pages 2487-2492, USA, 2002.
- [29] L.A. Smith, M.C. Cuellar, H. Du, and K. Judd. Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models. *Physics Letters A*, 374:26182623, 2010.
- [30] J.C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, 2003.
- [31] Z. Toth and E. Kalnay. Ensemble forecasting at nmc: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74:23172330, 1993.
- [32] Z. Toth and Eugenia Kalnay. Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):32973319, December 1997.
- [33] S. M. Uppala, P. W. Killberg, A. J. Simmons, U. Andrae, V. Da Costa Bechtold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. Van De Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hlm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, R. Jenne, A. P. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth,

A. Untch, D. Vasiljevic, P. Viterbo, and J. Woollen. The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):29613012, 2005.

[34] Caribbean Catastrophe Risk Insurance Facility. Understanding ccrif: A collection of questions and answers. CCRIF SPC, March 2015.