Research Statement

Shashank Kumar Roy
PhD Research Scholar, International Center for Theoretical Sciences, Bangalore

January 2023

Overview

I have pursued my research work in the area of data assimilation before which I finished my M.Sc and B.Sc Honors in physics. During my coursework, I had the chance to work on interesting projects starting with analytical solution of Stokes's flow in spherical geometry using vector spherical harmonics, numerical solution to stochastic forced burgers equation, 1-d visco-eslastic pde model, kuramoto-sivashinsky equation and ensemble kalman filter(EnKF) for Lorenz-96, a 40-dimensional chaotic dynamical system using partial and noisy observations.

I am generally fond of computational problems in prediction and inference in context of complex systems where well defined mathematical notions and ideas of probability, statistics and physics provide insight and can be converted to code and increase my understanding through simulations. Other endeavours orthogonal to my own research work which I enjoy is participation in hackathons where I often stumble across new problems and fascinating progress in deeplearning methods and their applications. I have taken a few courses in machine learning and deeplearning from neuromatch academy and nvidia workshops. Most of my coding experience in machine learning with real datasets was gained during hackathons organized from scientific communities such as NOAA and Ecole Polytechnique, such as writing an LSTM model for prediction of spatial time series data of weather variable and deep generative models for learning distribution of sea surface temperature from historical data using a generative adversarial network(GAN).

My thesis research work concentrates around data assimilation for chaotic dynamical system using EnKF[1], a general sequential state estimation algorithm which computes the best estimate of the state with associated uncertainty. Jointly with others, I have worked on demonstrating numerical filter stability, a crucial property of a filter using Sinkhorn distance, a distances between probability distribution. In another work, I am looking at instability properties of a dynamical system such as lyapunov vectors which are important in improving the existing techniques in prediction and estimation of a dynamical system in general. I now briefly talk about them below, starting with filter stability.

Numerical filter stability

In high-dimensional chaotic system such as atmosphere and ocean, it is not possible to track and predict such a system for long using only the model [2]. Data assimilation uses numerical models of the physical system representing our knowledge of the governing dynamics and combines them with the noisy and sparse observations from the system weighted by their respective uncertainties, in order to produces improved statistical estimates of the true state of the system[3]. In the setting of a deterministic dynamical system, with observations operators h and measurement noise ϵ, we have

$$\mathbf{x}_k = \mathbf{M}(\mathbf{x}_{k-1}), \qquad \mathbf{y}_k = h(\mathbf{x}_k) + \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, \Sigma), \text{ where } \mathbf{x}_k \in R^d, \mathbf{y}_k \in R^p \text{and } p < d. \tag{1}$$

Bayesian filtering and data assimilation Bayesian filtering is defined as the sequential estimation of the conditional distribution in phase space of the state of a physical system coming from an assumed model taking into account the likelihood of new information arriving from the observations [4]. Any filtering algorithm begins with an initial distribution $\rho(x_0)$ of the state at $k = 0$, which could be far from the true unknown initial state and for each time k when the observation arrives, using bayes theorem, one can write,

$$\rho\left(\mathbf{x}_{k+1} | \mathcal{Y}_{k+1}\right) = \frac{\rho\left(\mathbf{y}_{k+1} | \mathbf{x}_k + 1\right) \rho\left(\mathbf{x}_{k+1} | \mathcal{Y}_k\right)}{\rho\left(\mathbf{y}_{k+1} | \mathcal{Y}_k\right)} \tag{2}$$

A numerical filter such as ensemble kalman filter and particle filter numerically approximate the conditional distribution of the state starting with observation $y_{0:K}$ and a initial distribution of the true state at $k = 0$, sequentially over time using ensemble approach. Different filtering algorithms are based upon different approximations which make the conditional posterior distribution computationally tractable. The flow-dependent uncertainty is provided by the governing dynamical equations of the system between the observations times[4].

Stability of a filtering algorithm. The true state is unknown and the choice of $\rho(\mathbf{x_0})$ may be far from the truth, hence all filtering algorithms making their arbitrary choice in order to run the assimilation system. It becomes crucial that the conditional distribution of the state become independent of the initial choice used in initializing the filter so that quantities estimates using the posterior are eventually independent of our arbitrary and often wrong choice of $\rho(\mathbf{x_0})$. Filter stability is the property that the conditional posterior distribution computed sequentially over long time is independent of the choice of the distribution at $k = 0$ used to initialize the filtering algorithm. The question we ask is how to numerically check if a filter is stable for different dynamical systems? What we need is a distance D on the space of probability distributions $P(\mathcal{R}^d)$ which, for two different initial distributions ν_1, ν_2 for any initial conditionx_0 with $\hat{\pi}_n(\nu_1)$ and $\hat{\pi}_n(\nu_2)$ being the posterior obtained after assimilating all observation $y_{1:n}$ at time n, converges with increasing n. Mathematically, for filter stability to hold, we have

$$\lim_{n\to\infty} \mathbf{E}[D(\hat{\pi}_n(\nu_1), \hat{\pi}_n(\nu_2))] = 0 \tag{3}$$

where, D is a distance on $P(\mathcal{R}^d)$, the space of probability measures on $\mathcal{R}^d$ and the expectation is over observation noise accounting for different noise realizations [5]. In our work, we have chosen Sinkhorn distance[6], a type of distance on the space of probability distributions, which has several merits over other distances such as total variation, apart from being computationally cheaper[7]. This distance utilizes two samples from their respective distributions in order to compute the distance between them upto the sampling errors. To finally demonstrate and computationally show filter stability for enkf, I used Lorenz-96 in 40-dimensions as my model to assimilate partial observations consisting of 20 components. With two different initial distribution for the filter, I studied the exponential rate of convergence for the distance over time by numerically computing the Sinkhorn distance [8] between the conditional distribution over time . Using different observation gap and observation covariance, I showed that stability for enkf is quite robust to the above two parameters. I also studied numerically the relationship between the rmse, a measure of filter accuracy with the filter stability.

Future research directions Below are some of problems where application of the above ideas of filter stability may be practically useful.

1. A particularly interesting idea is to see how probabilistic machine learning models proposedly performing filtering, satisfy this criteria improving them.
2. The study can be generalized to different filters in context of model and parameterization errors of unmodelled quantities respectively.
3. Relating filter stability to the chaotic properties of the underlying system, since the interplay between the instability and the informative observations lead to eventually capture the conditional distribution.

Computation of Covariant Lyapunov vectors

Instability properties of a chaotic nonlinear dynamical system are characterized globally by lyapunov exponent and locally by lyapunov vectors(LVs), which correspond to the directions associated with those rates. Different kind of unstable vectors [2] and subspace of a system has been shown to improve forecasts in ocean-atmosphere coupled models when the assimilation takes into account [9]. In the setting of data assimilation with sparse and noisy observations, one can use filtering to find the best estimate of the true state over time. This comes with a caveat that the filter estimates or the analysis mean over time is not a dynamical trajectory of the system, i.e. there is no initial condition which when integrated leads to the filter estimate over time. The true trajectory lies near the analysis mean which is quantifies by the error in the estimates themselves. I formulated this problem into using the filter estimate over time as a proxy of the true trajectory perturbed by the error statistics to recover the lyapunov vectors. This approach to led to studying how sensitive are different lyapunov vectors such as backward and covariant lyapunov vectors even

before one can use analysis trajectory obtained from a filter.
Sensitivity of Lyapunov vectors The specific questions which I address about lyapunov vectors and the exponents in context of the filtered trajectory are as follows:

- How sensitive are the backward and covariant lyapunov vectors and the corresponding exponents to perturbations in the underlying trajectory?
- Under what conditions can one recover them from a filter estimated trajectory instead of the true trajectory of the dynamical system?
- How robust is the unstable subspace to the perturbation strength σ, and are they more robust than the individual vectors themselves?

In order to start answering the above questions, I studied the effect on the LVs to small perturbations added to the underlying trajectory by systematically adding noise of strength σ following a gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$, where d is the dimension of the state. I computed the the CLVs around the true and the perturbed trajectory using Ginelli's algorithm[10] which also gives BLVs as an intermediate step. I used the angles between the respective LV obatined from the true and the perturbed trajectory to understand the limitations of such vectors obtained from the numerical state estimates of the filter. In small dimension, I used Lorenz-63, where visualization is straight forward to interpret. 1.

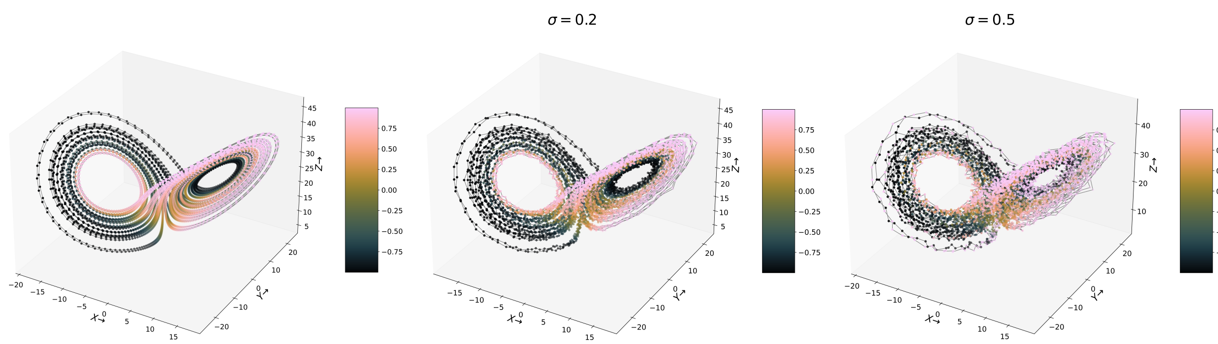

Figure 1: The angle between the first two clvs have been shown to predict regime change in L63 system [11]. We plot the cosine for both true trajectory and perturbed trajectory in phase space. The left-most picture corresponds to the true trajectory.

In high-dimensions, I studied Lorenz-96 in order to explore dimensional dependence added to the sensitivity problem. Another interesting directions is using principle angles [12], which summarize the angle between two different sub-spaces, and seeing how they change with σ, which I found to be more robust than the individual vectors themselves. Such analysis is useful in context of problems where sub-spaces are more important than the individual vectors themselves. It also allows one to understand and interpret how much information can be reconstructed using lyapunov vectors from a filter generated trajectory using partial and noisy observations and a model.
Possible directions for research A set of directions for future work which can be directly extended from the current work are as follows:

1. Combining the ideas of assimilation in unstable subspace where the sub-spaces computed from the historical data can be employed.
2. Studying the effect of model errors themselves on the computed lyapunov vectors.
3. The degree of similarity in the lyapunov vectors between two nearby points on the attractor in phase space which may be important machine learning methods trained on vectors computed once along a trajectory be used to predict vectors at neaby points in the phase space.
4. Studying structure of CLVs for discretized PDE systems such as Kuramoto-sivashinksy equation which has a finite dimensional attractor to shed light on CLV localization problem.

Other research interests

Problems at the interface of climate and data science is something I am interested to work, the topic on which my interest from different workshops and discussions where I believe that the skills which I have acquired in the context of data assimilation are useful. I am interested in understanding ways to incorporate uncertainty and dynamical knowledge together to a general machine learning techniques for modelling and inference of large dynamical systems of practical importance where the limited data combined with physical constraints and conservation laws can balance for the sparsity and scarcity of available data. Such problems are of high interest in climate modeling and related data science problems. Another important topic which I find fascinating to explore further is ideas from optimal transport which I gained some exposure to while working on the filter stability problem and would like implement it for different data-driven problems. Understanding how new developments such as diffusion models in latent space for generative modelling can be used to design effective filtering and probabilistic machine learning algorithms are another of my interests.

References

[1] G. Evensen, "The ensemble kalman filter: theoretical formulation and practical implementation," *Ocean Dynamics*, 2003. [Online]. Available: https://doi.org/10.1007/s10236-003-0036-9

[2] T. N. Palmer and L. Zanna, "Singular vectors, predictability and ensemble forecasting for weather and climate," *Journal of Physics A: Mathematical and Theoretical*, vol. 46, no. 25, p. 254018, jun 2013. [Online]. Available: https://doi.org/10.1088/1751-8113/46/25/254018

[3] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview of methods, issues, and perspectives," *WIREs Clim Change*, vol. 75, no. 1B, pp. 257–288, 2018. [Online]. Available: https://doi.org/10.1002/wcc.535

[4] S. E. Cohn, "An introduction to estimation theory (gtspecial issueltdata assimilation in meteology and oceanography: Theory and practice)," *Journal of the Meteorological Society of Japan. Ser. II*, vol. 75, no. 1B, pp. 257–288, 1997.

[5] P. Mandal, S. K. Roy, and A. Apte, "Stability of nonlinear filters-numerical explorations of particle and ensemble kalman filters," in *2021 Seventh Indian Control Conference (ICC)*. IEEE, 2021, pp. 307–312.

[6] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré, "Interpolating between optimal transport and mmd using sinkhorn divergences," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2681–2690.

[7] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer, "Convergence of entropic schemes for optimal transport and gradient flows," *SIAM Journal on Mathematical Analysis*, vol. 49, no. 2, pp. 1385–1418, 2017.

[8] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, "Overrelaxed sinkhorn–knopp algorithm for regularized optimal transport," *Algorithms*, vol. 14, no. 5, p. 143, 2021.

[9] S. Vannitsem and V. Lucarini, "Statistical and dynamical properties of covariant lyapunov vectors in a coupled atmosphere-ocean model-multiscale effects, geometric degeneracy, and error dynamics," *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 22, p. 224001, may 2016. [Online]. Available: https://doi.org/10.1088/1751-8113/49/22/224001

[10] F. Ginelli, H. Chaté, R. Livi, and A. Politi, "Covariant lyapunov vectors," *Journal of Physics A: Mathematical and Theoretical*, vol. 46, no. 25, p. 254005, jun 2013. [Online]. Available: https://doi.org/10.1088/1751-8113/46/25/254005

[11] E. L. Brugnago, J. A. C. Gallas, and M. W. Beims, "Predicting regime changes and durations in lorenz's atmospheric convection model," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 10, p. 103109, 2020. [Online]. Available: https://doi.org/10.1063/5.0013253

[12] G. H. Golub and H. Z. ha, "The canonical correlations of matrix pairs and their numerical computation," 1992.