THE STATISTICAL PHYSICS OF DATA ASSIMILATION AND MACHINE LEARNING

Data assimilation is a hugely important mathematical technique, relevant in fields as diverse as geophysics, data science, and neuroscience. This modern book provides an authoritative treatment of the field as it relates to several scientific disciplines, with a particular emphasis on recent developments from machine learning and its relation to data assimilation. Underlying theory from statistical physics, such as path integrals and Monte Carlo methods, is developed in the text as a basis for data assimilation, and the author then explores examples from current multidisciplinary research such as the modeling of shallow water systems, ocean dynamics, and neuronal dynamics in the avian brain. The theory of data assimilation and machine learning is introduced in an accessible and unified manner, and the book is suitable for undergraduate and graduate students from science and engineering without specialized experience of statistical physics.

HENRY D. I. ABARBANEL has worked in several fields of physics including high energy physics, nonlinear dynamics, and data assimilation in neurobiology. He is the author of two previous books: *Analysis of Observed Chaotic Data* (1996) and *Predicting the Future: Completing Models of Observed Complex Systems* (2013). He is a Distinguished Professor of Physics at the University of California, San Diego (UCSD) and a Distinguished Research Physicist at UCSD's Scripps Institution of Oceanography.

THE STATISTICAL PHYSICS OF DATA ASSIMILATION AND MACHINE LEARNING

HENRY D. I. ABARBANEL
University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781316519639
DOI: 10.1017/9781009024846

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Abarbanel, H. D. I., author.
Title: The statistical physics of data assimilation and machine learning / Henry D.I. Abarbanel.
Description: Cambridge ; New York, NY : Cambridge University Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021025371 (print) | LCCN 2021025372 (ebook) | ISBN 9781316519639 (hardback) | ISBN 9781009024846 (ebook)
Subjects: LCSH: Statistical physics–Data processing. | Stochastic processes. | Supervised learning (Machine learning)–Mathematical models. | Discrete-time systems.
Classification: LCC QC174.8 .A225 2022 (print) | LCC QC174.8 (ebook) | DDC 530.13–dc23
LC record available at https://lccn.loc.gov/2021025371
LC ebook record available at https://lccn.loc.gov/2021025372

ISBN 978-1-316-51963-9 Hardback

Contents

Preface

This book explores methods for performing the tasks in Data Assimilation, a critical practical step in the transfer of information from observed data collected during measurements of a physical or biological dynamical system to a nonlinear dynamical model proposed for the that system.

The name Data Assimilation emerged over the years in the context of Numerical Weather Prediction in meteorology, but has found the same challenge in many fields of study: including in numerical weather prediction and quantitative aspects of neurobiology, as examples, and many other areas of science and technology where one must address this transfer of information as well.

Further, as we proceed through this book we show that the same questions and associated tools for answering them appear in the **equivalent** problem of Supervised Machine Learning.

These seemingly quite different formulations of questions and tools for addressing them all appear in the framework of Statistical Physics. If you have not had experience with Statistical Physics, this volume will introduce you to much of its strength without any undue suffering on your part–that's the plan anyway!

In our discussions these problems are placed into a common path integral formulation (Zinn-Justin (2002); Hochberg et al. (1999); Abarbanel (2013)) which provides a unification of critical questions and a framework in which to view methods developed in various disparate fields as they apply to many others.

What does it mean to transfer information in data to a model of the dynamical system generating those data?

In these problem areas we have dynamical equations (the model) having state variables {voltages, $V(t)$, density or concentrations of chemical constituents, $\rho(x, y, x, t)$, velocities, $\mathbf{v}(x, y, z, t)$, and so forth} as well as time independent parameters such as {viscosities, conductivities, and so forth}, any of which may be *unobserved or unobservable* in the collection of the data.

Figure 0.1 Monsieur Proust's Collection of Books: "Remembrance of Things Past"

Over some observation window in time $[t_0, t_{final}]$ we collect information on the observable variables, then use data assimilation tools to estimate the unobservable state variables and time independent parameters. With estimates of the full set of state variables at t_{final} and all the parameters, we can use these as initial conditions at t_{final} to predict forward $t > t_{final}$ as a test or validation or generalization of the model.

This volume is intended for data scientists, physical scientists, and life scientists who wish to utilize machine learning methods (Goodfellow et al. (2016); Abarbanel et al. (2018)), to simplify or accelerate calculations within their inquiries (Pathak et al. (2018); Ott (2019)).

It is intended for scientists and engineers with experience in methods of statistical physics, typically covered in beginning graduate courses in Physics and Chemistry, for transferring information in observed data to models of the observed processes with the goal of estimating unknown fixed parameters in the models as well as *unobserved state variables* in the models.

This transfer process is called by various names in different fields of study. We adopt the designation **data assimilation** following the terminology used in numerical weather prediction (Ghil and Malanotte-Rizzoli (1991); Pires et al. (1996); Kalnay (2003); Lorenc and Payne (2007); Evensen (2009); Reich and Cotter (2015)) over many decades. It is also widely referred to as state and parameter estimation in many engineering applications. These are just a few: Dochain (2003); Horváth and Manini (2008); Lei et al. (2017).

Collecting the data takes quite skilled personnel.

- The data is always noisy and typically sparse in the sense that only a (usually quite small) subset of the dynamical variables in the system producing the data are observable.
- Formulating a model for the dynamics of that system is not algorithmic. It takes experience and insight into the physical or biological or other mechanisms identified to be operating within the observed system.
- Transferring the information residing in the data to critical aspects of the selected model also takes skill.

This book is primarily about the last of these items, especially when the data are noisy and the model has errors.

An overview of these steps is this:

> In some time interval $[t_0, t_{final}]$ (or many such time intervals) we measure L quantities $\mathbf{y}(\tau_k) = \{y_1(\tau_k), y_2(\tau_k), \ldots, y_L(\tau_k)\}$ at times τ_k: $\{t_0 \le \tau_k \le t_{final}\}$ of physical, biophysical, geophysical, or other subject of interest.
>
> After some contemplation of the forces acting on the system producing these measurements and some consideration of the environment in which the observed nonlinear dynamical system resides, one proposes some D-dimensional $(D \ge L)$ dynamical equations for the state variables $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_D(t)\}$ in the form of nonlinear ordinary differential equations.
>
> $$\frac{dx_a(t)}{dt} = F_a(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}); \quad a = 1, 2, \ldots, D, \tag{0.1}$$
>
> where $\boldsymbol{\theta}$ is a collection of N_p time independent parameters, $\mathbf{u}(t)$ are some time varying quantities, perhaps under the control of the observer, and $\mathbf{F}(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta})$ is called the vector field of the dynamics. $\mathbf{u}(t)$ is a set of "time dependent" parameters for which no dynamical equation is usually given. It is specified outside the observed system, and it can be treated as a sequence of parameters at each time $t_n = t_0 + n\Delta t$ where the state variables $\mathbf{x}(t_n)$ are desired.
>
> How can we estimate the parameters $\boldsymbol{\theta}$ we do not know, as well as the $D - L$ state variables we do not (or cannot) observe, and the 'external' or environmental forces we may not know using the potentially sparse $(D \gg L)$ and certainly noisy measurements we have acquired?

If we were able to accomplish the required estimations, we would be in a good position to ask and answer the additional, critical question:

> With estimates of *all* state variables at the end t_{final} of the observation window $[t_0, t_{final}]$ $\mathbf{x}(t_{final})$, all parameters $\boldsymbol{\theta}$, and all forces $\mathbf{u}(t \geq t_{final})$, can we predict $\mathbf{x}(t > t_{final})$ by solving the initial value problem, Eq. (0.1), with initial conditions now given at t_{final}?

As we have L observed state variables, we can check the consistency of the model output with those observations in and beyond $[t_0, t_{final}]$. In the prediction phase of data assimilation, $t \geq t_{final}$, we must have accurate estimations of the *unobserved* state variables. In this step, called 'generalization' in machine learning, we are testing both our selection of the model Eq. (0.1) and the workings of our data assimilation, information transfer, protocols.

Let's address just a few technical tidbits before proceeding:

1. If the model of the dynamics producing the data is in the form of *partial* differential equations, then one has an infinite number of degrees-of-freedom $(D \to \infty)$! So one puts the fields on a $\mathbf{x}(\mathbf{r}, t)$; $\mathbf{r} = (x, y, z)$ spatial grid (or equivalent) with $N = N_x \times N_y \times N_z$ grid points, resulting in ND ordinary differential equations of the form Eq. (0.1), which brings us back to the same discussion.
2. The 'external' quantity $\mathbf{u}(t)$ may be known, for example, if the forces driving the observed system are known. If they are not known, then in a variational treatment of the overall data assimilation problem, as discussed in Gelfand and Fomin (1963); Kirk (1970), there are equations determining $\mathbf{u}(t)$ from knowledge of the $\mathbf{y}(\tau_k)$.
3. We discuss one observation window in time $[t_0, t_{final}]$; however, if the dynamics is chaotic, there may be a need for a sequence of observation windows. The reason is that we are numerically able to estimate parameters and states to various levels of accuracy; however, such unavoidable errors are amplified by the chaotic dynamics, and one needs to observe again to put the evolving trajectory into the correct region of state space.
4. The number of measurements, called L here, made at each observation time may vary as observations are made.
5. The time Δt separating steps in the utilization of the dynamics need not be uniform across all time windows.

Who wants to do this sort of thing anyway?

The simplest answer is *everyone* working in science and technology. Many have data, many have models describing how those data emerged from observations of some dynamical system, and all need tools to transfer the information in those data to the model. Since there are unknown parameters and surely unobserved state variables, the data alone may be insufficient to provide confidence in the ability of the model to predict beyond t_{final}.

To this quite general discussion, it might be of value to discuss an example briefly. You are tasked with making measurements on a neuron isolated from its working environment and placed, *in vitro*, in a dish with the goal of developing some model dynamical equations that allow you to know (predict) with accuracy how this neuron, or its equivalent back in the working biophysical neural circuit, will respond to currents in its environment and coming through to it via synaptic or other connections to other neurons.

An Illustrative Example – Dynamical Equations for a Neuron

The biophysical equations describing the dynamics of neurons were established by work done mostly in Cambridge, UK before and after World War II. Hodgkin and Huxley (1952); Johnston and Wu (1995); Sterratt et al. (2011) were some of the researchers whose names are prominent in understanding that equations imposing current conservation on the ions flowing into and departing from the neuron body (soma) and equations capturing the voltage dependent permeability of the cell membrane to these ions would provide a quantitative framework for the biophysical description of the neural processes.

In the experiments they performed they considered two ion channels for Na and K ions flowing through proteins penetrating the cell membrane. They also introduced a 'leak' channel describing other aspects of neuron behavior. The nonlinear equations they proposed, and tested, have the form

$$C_m \frac{dV(t)}{dt} = g_{Na} m(t)^3 h(t)[E_{Na} - V(t)] + g_K n(t)^4 [E_K - V(t)] + g_L[E_L - V(t)] + I_{DC} + I_{app}(t). \tag{0.2}$$

The three voltage dependent 'gating' variables $a(t) = \{m(t), h(t), n(t)\}$; $0 \leq a(t) \leq 1$ are taken to satisfy the first order kinetics

$$\frac{da(t)}{dt} = \frac{a_0(V(t)) - a(t)}{\tau_a(V(t))}. \tag{0.3}$$

The parameters $\{g_j\}$ in the voltage equation Eq. (0.2) variables are constants while the gating variables are state variables that are voltage dependent. $a_0(V)$ and $\tau_a(V)$ are voltage dependent functions. The first is dimensionless and sets the scale for the gating variables, and the second is a voltage dependent time scale for the gating variables.

This is a $D = 4$ dimensional dynamical system. It has rich behavior (Hodgkin and Huxley (1952); Johnston and Wu (1995); Sterratt et al. (2011)). In laboratory

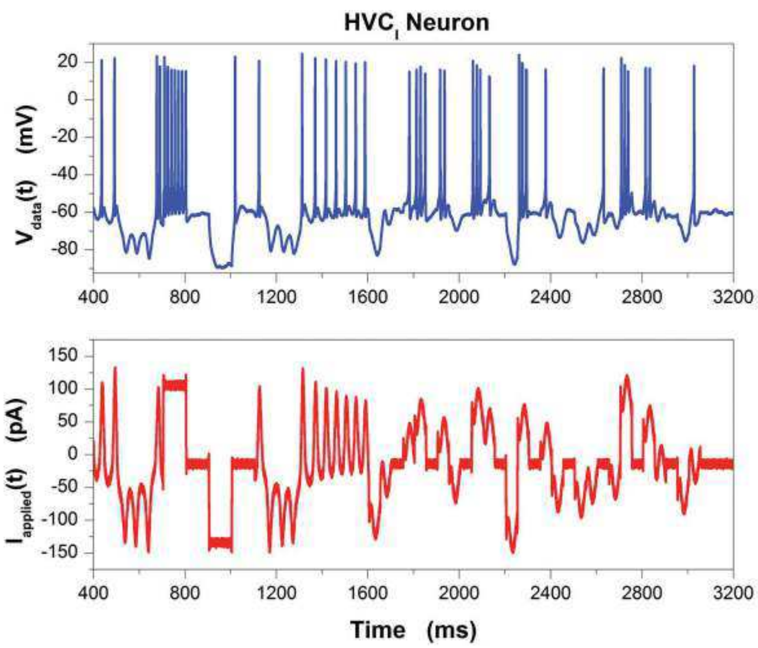

Figure 0.2 An illustrative example of the data assimilation challenge discussed in this book. This figure shows data from Daniel Margoliash of the University of Chicago and Daniel Meliza of the University of Virginia. An interneuron within the nucleus HVC of the avian brain was isolated in a glass dish in the laboratory – an *in vitro* experiment. An electrode was inserted into the body of the neuron and the applied current $I_{app}(t)$ shown in the **bottom panel** was injected into the neuron. The resulting membrane potential response, shown in the **top panel** of the display, was measured using the same electrode. From these data, one is asked to estimate all the parameters, here $N_p = 20$, and the three (unmeasured) gating variables $a(t)$ in the HH equation, Eq. (0.2).

experiments one can directly measure the cross membrane voltage $V(t)$, but no instruments are available (as of July 2020) to observe the gating variables. In the general language used here, $L = 1$, and three state variables are unobserved.

An experiment consists of selecting a current $I_{app}(t)$ (the analog of $\mathbf{u}(t)$ discussed above), and it is typically known. With only $V(t)$ observed, the challenge is to estimate all the fixed parameters in Eq. (0.2) as well as to estimate all of the $a(t)$ and all of the parameters in the $a_0(V)$ and $\tau_a(V)$ appearing in, Eq. (0.3), over $[t_0, t_{final}]$. Validation (or not) of the model associated with the observations comes from solving Eq. (0.2) for $t \geq t_{final}$, using $V(t_{final})$ and the $a(t_{final})$ as initial conditions and the estimated $\boldsymbol{\theta}$ to complete the HH equations.

If you are a 'data scientist,' there is often a directive to find a 'domain expert' with whom to work on the problem just posed. I personally encourage each reader to become a domain expert and a data scientist at the same time and not artificially divide one's self into two or more parts. The problems you wish to solve usually require both of these parts of you, and your appreciation of the data and the modeling will be increased by this strategy. If you collaborate with other domain experts and/or other data scientists to address the issues in your problem, that brings even more experience to the table.

In the instance of neurobiology, for example, I recommend the domain expert's 'manual' by Daniel Durstewitz (2017). It addresses what you need to know about neurons and then provides an easy entry into computational modeling of neurons and networks thereof. The textbook by Sterratt et al. (2011) covers less neuro-data analysis than Durstewitz (2017) but focuses more in modeling networks of neurons.

The other topic we often use as examples in this book arises in geophysics, and your road to domain expertise could be via Pedlosky (1986); Vallis (2017).

Why should we expect this will work? Because the dynamical model is nonlinear in the state variables, the state variables are generically coupled together through the nonlinear model. Information is passed through the observable $V(t)$ and determines the unobserved $a(t)$ and the parameters $\boldsymbol{\theta}$, consistent with the data.

The functions $a_0(V)$ and $\tau_a(V)$ for the neurobiological problem may be estimated by looking at experimental data (Senselab-Yale (2020)) for simulations of each ion channel: Na, K, Ca, ...

This sets the challenge. We'll see how it all works, in detail, as we proceed.

A Bit More Just Before We Start Out

This book is primarily about how one effects this information transfer when the data are noisy (always) and the model has errors (also always). There are many methods for this that have been developed in various fields; and while we will note those developments, our focus here will be on the path integral formulation of the critical questions.

Why path integrals? They sound quite exotic; however, as we will see in the chapters ahead, they are formulated in a natural manner, and they are integral representations of the solutions to equations such as Eq. (0.1) when errors in the model and errors in the data are present. Such representations give us insight into global properties of the dynamical systems we will be analyzing, while the differential equations, such as Eq. (0.1), are focused on local time evolution of those equations.

We will discuss variational methods for both continuous time and discrete time. We will discuss Monte Carlo methods. We will discuss both of these using an annealing method that turns on the magnitude of the model precision, and thus the nonlinearity in an adiabatic fashion.

I will also discuss supervised Machine Learning because, as it turns out, this is mathematically equivalent to DA. So everything we say about data assimilation applies.

This is not something (yet) discussed in other books, and I hope by drawing attention to this equivalence the methods from DA will be utilized in ML, and, hopefully, *vice versa*.

Another topic I will emphasize, again not widely addressed elsewhere, is the question: How many L measurements does one require to accurately estimate the state variables $\mathbf{x}(t)$ as well as the parameters $\boldsymbol{\theta}$? As estimating each of these costs bits of information (Rissanen (1989)), accuracy must depend on how much independent data is available. Furthermore, we'll see that the answer depends on the vector field $\mathbf{F}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{u})$, and the DA efficacy depends on the instabilities of the nonlinear communication protocol connecting the data, $\mathbf{y}(t)$, to the model (Kostuk (2012)).

We will select examples from geosciences and neurobiology as we proceed. As noted, the material here is often encountered by students of Physics, Chemistry, and Geophysics. It is not at all common to see it in a Neurobiology curriculum. I hope this book and the papers presenting research that precede it will become common practice in computational neuroscience as well; we'll see.

Many Thanks Are Owed

The results in this monograph cannot be claimed by me to be mine alone. My many productive interactions with former and present Physics PhD students at UCSD contributed to every word and paragraph. These women and men include Daniel Creveling, Brian Toth, Mark Kostuk, Chris Knowlton, Will Whartenby, Jack Quinn, Uriel Morone, Michael Eldridge, Jason An, Xingxin Ye, Daniel Rey, Nirag Kadakia, Sasha Shirman, and Paul Rozdeba, and I recommend their PhD dissertations to the reader. Those may be found in the University of California's archive **escholarship.org**:

> Creveling (2008); Toth (2011); Kostuk (2012); Shirman (2018); Kadakia (2017); Ye (2016); Quinn (2010); An (2019); Rey (2017); Rozdeba (2017); Knowlton (2014); Eldridge (2016); Morone (2016); Whartenby (2012).

Equally important have been my productive interactions over many years with Daniel Margoliash at the University of Chicago and Ulli Parlitz in Göttingen. This includes as well their postdoctoral fellows Daniel Meliza (now at the University of Virginia) and Arij Daou (now at the American University of Beiruit) and many PhD students, especially Jöchen Bröcker, now at the University of Reading (UK).

To this list I am pleased to add Alain Nogaret at the University of Bath (UK), Eve Armstrong (NYIT), George Michael Fuller (UCSD, Physics) and Philip Gill, Melvin Leok, Michael Holst, and Randy Bank (UCSD, Mathematics), Gert Cauwenberghs, Gabriel Silva (UCSD, Bioengineering) and Tim Gentner (UCSD, Psychology and Neurobiology), Michael Long (NYU, Neuroscience), Bruce Cournelle and Art Miller (UCSD, Scripps Institution of Oceanography), and Alexandre Chorin (UCB, Mathematics).

Finally, I am indebted to my old and reliable friends Robert Sugar (University of California Santa Barbara, Physics) and Jerry Marsden (UC Berkeley and Caltech) who have tolerated and often dissipated my confusion about many topics in statistical Physics for many years. Marsden's youthful death did not impede the importance of his contributions to the ideas we discuss here.

My wife has always encouraged my efforts by noting my results are "obvious" and looking forward to even more.