UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Chaos, observability and symplectic structure in optimal estimation**

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics with a Specialization in Computational Science

by

Daniel Rey

Committee in charge:

Professor Henry D.I. Abarbanel, Chair
Professor Daniel Arovas
Professor Philip Gill
Professor Michael Holst
Professor Melvin Leok
Professor Ramamohan Paturi

2017

The dissertation of Daniel Rey is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

*To Grace and Fiona*

*— may the sunshine follow us wherever we go*

EPIGRAPH

*Physics is what physicists do late at night...*

—Attributed to Enrico Fermi

TABLE OF CONTENTS


Signature Page . . . iii

Dedication . . . iv

Epigraph . . . v

Table of Contents . . . vi

List of Figures . . . ix

Acknowledgements . . . xi

Vita . . . xiii

Abstract of the Dissertation . . . xiv

1 Introduction . . . 1
- 1.1 Predicting the outcome of a coin toss . . . 2
  - 1.1.1 Filtered estimates . . . 3
  - 1.1.2 Smoothed estimates . . . 5
  - 1.1.3 Modeling coin physics . . . 6
  - 1.1.4 The observation model . . . 8
  - 1.1.5 The estimation action . . . 8
  - 1.1.6 Evaluating the path integral . . . 9
  - 1.1.7 The effective action . . . 10
  - 1.1.8 Observability . . . 11
  - 1.1.9 Predictability and complexity . . . 12
- 1.2 An historical perspective . . . 13
  - 1.2.1 Pre-twentieth century . . . 13
  - 1.2.2 The Wiener-Komogorov filter . . . 14
  - 1.2.3 Stochastic processes, information, and statistics . . . 14
  - 1.2.4 Kalman, Bellman, and Pontryagin . . . 15
  - 1.2.5 Exact finite-dimensional nonlinear filters . . . 15
  - 1.2.6 Variational methods . . . 16
  - 1.2.7 Ensemble and particle methods . . . 17
  - 1.2.8 Moving forward . . . 17
- 1.3 Overview of this thesis . . . 18

2 The canonical structure of optimal estimation . . . 19
- 2.1 The Hamiltonian structure of fixed-interval smoothing . . . 19
  - 2.1.1 Canonical momentum as a Lagrange multiplier . . . 22
  - 2.1.2 Discrete time . . . 23
  - 2.1.3 The search for minimizing paths . . . 25
- 2.2 Optimal solutions for linear models . . . 28
  - 2.2.1 The simplest tracking problem . . . 29
  - 2.2.2 Continuous time solutions . . . 30
  - 2.2.3 The Kalman-Bucy filter . . . 32
  - 2.2.4 Smoothed estimates . . . 34
  - 2.2.5 The smoothed covariance . . . 35
  - 2.2.6 Discrete time . . . 36

2.2.7 The simplest tracking problem revisited . . . 37
2.3 Nonlinear extensions . . . 40
2.3.1 Two-pass smoothing and Newton's method . . . 40
2.3.2 The second variation and Hamilton's equations . . . 42
2.3.3 The Riccati transformation . . . 42
2.3.4 The filtered solution . . . 43
2.3.5 The backwards pass . . . 45
2.3.6 The Hamilton-Jacobi solution . . . 46
2.4 Discrete time solutions . . . 47
2.4.1 The one-step action . . . 48
2.4.2 Two-pass smoothing, Newton's method, and canonical structure . . . 49
2.4.3 The second variation in discrete time . . . 52
2.4.4 The discrete Hamilton-Jacobi solution . . . 55
2.5 Canonical transformations for two-point boundary value problems . . . 57
2.5.1 Generating functions of canonical transformations . . . 58
2.5.2 Application to fixed interval smoothing . . . 60
2.6 Symplectic structure and optimal stability . . . 64

3 Observability and conditioning in dynamical inverse problems . . . 66
3.1 Conditioning, regularization and numerical stability . . . 68
3.1.1 Conditioning . . . 68
3.1.2 Regularization . . . 69
3.1.3 Stability, consistency and convergence . . . 71
3.1.4 Dynamical regularization . . . 73
3.2 Methodology . . . 74
3.2.1 Lorenz 96 and extensive chaos . . . 75
3.2.2 The average dimension of the unstable subspace . . . 75
3.2.3 Twin experiments . . . 76
3.2.4 Observations and observability . . . 77
3.2.5 Critical values . . . 78
3.2.6 Two criteria for success . . . 79
3.3 The number of observations required for synchronization . . . 80
3.3.1 Kalman filtering, 3DVar, and synchronization . . . 81
3.3.2 The minimum observability threshold $L_c$ . . . 84
3.3.3 Synchronization, phase transitions, and the role of the unstable subspace . . . 86
3.3.4 Approximate bounds on $L_c$ . . . 88
3.3.5 Assimilation in the unstable subspace (AUS) . . . 89
3.3.6 The collapse of the error covariance . . . 89
3.3.7 Filtering observation noise through adaptive annealing of $\boldsymbol{R}_f$ . . . 90
3.4 Conditioning and observability in 4DVar methods . . . 92
3.4.1 Three formulations of 4DVar . . . 92
3.4.2 The Gauss-Newton method . . . 95
3.4.3 Implementation details . . . 97
3.4.4 The collapse of the solution basin . . . 98
3.4.5 Convergence and uniqueness balls for the Gauss-Newton method . . . 100
3.4.6 The inverse condition number of the Hessian . . . 102
3.4.7 Weak constraints, synchronization, and stability . . . 103
3.4.8 Critical observability limits . . . 105
3.4.9 Annealing $\boldsymbol{R}_f$ . . . 106
3.5 The uniqueness ball for the Gauss-Newton method . . . 108
3.6 The Lipschitz constant $\Lambda$ . . . 110

4 Moving horizon estimation for poorly observable systems . . . 122
4.1 Extracting information from the time series of observations . . . 123
4.1.1 Time delay embedding and attractor reconstruction . . . 124
4.1.2 Time delay feedback control synchronization . . . 125
4.1.3 Moving horizon estimation . . . 128
4.1.4 The extended Kalman Filter . . . 129
4.1.5 Strong constraint 4DVar . . . 129
4.1.6 Weak constraint 4DVar . . . 130
4.2 Time delay methods for reducing observability limits . . . 131
4.2.1 Time delay synchronization and the EKF . . . 131
4.2.2 The choice of $\tau$ . . . 133
4.2.3 Overlapping estimation windows . . . 133
4.2.4 Moving horizon estimation with 4DVar . . . 135
4.3 Optimizations and extensions . . . 136
4.3.1 Rank considerations for the pseudoinverse . . . 136
4.3.2 Tikhonov regularization . . . 137
4.3.3 A parallelized adjoint formulation of the Gauss-Newton method . . . 138
4.3.4 Optimizing the embedding . . . 141

5 Conclusions and future work . . . 147
5.1 The canonical structure of optimal estimation . . . 147
5.2 Observability and conditioning in dynamical inverse problems . . . 149
5.2.1 Filtering methods . . . 150
5.2.2 Fixed-interval smoothing . . . 151
5.2.3 Limitations and future work . . . 153
5.3 Moving horizon estimation for poorly observable systems . . . 156
5.4 A personal view on the future of estimation . . . 158

Bibliography . . . 160

Figure 3.1: Chaotic properties of Lorenz 96 model. The Kaplan-Yorke fractal dimension $D_a$ of the attractor, and the average dimension of the unstable subspace $D_u$, both scale linearly with the model dimension $D$. The magnitude of the maximum global Lyapunov exponents $\lambda_{\max}$ exhibits transient behavior for $D < 20$. . . . . 113
Figure 3.2: The magnitude of the 'true' error can be estimated using $\lambda_{\max}$ and measuring the time $\Delta$ it takes for the observed RMS deviations in the prediction to stabilize. This provides a way to quantify success with experimental data. . . . . 113
Figure 3.3: The critical minimum number of observations $L_c$ required to construct a successful estimate, for three related filtering methods. . . . . 114
Figure 3.4: Maximum conditional Lyapunov exponent, and fractional dimension of the unstable subspace $D_u/D$, plotted as a function of observation density $L/D$. 3DVar-FB crosses the synchronization threshold at $L_c \approx 0.4D$, while the EKF is more efficient at $L_c \approx 0.1D$. $D_u$ does not change much with the model resolution $D$. . . . . 115
Figure 3.5: The critical rank of an adaptive observation operator with $L = D$ that targets and stabilizes only the largest $r$ components of the unstable subspace. . . . . 115
Figure 3.6: The critical number of observations using assimilation in the unstable subspace (AUS), where the control perturbations are projected down onto the largest $r$ components of the unstable manifold. . . . . 116
Figure 3.7: Collapse of the estimated error covariance for the EKF with $\boldsymbol{R}_f^{-1} = 0$. Fig. (3.7a) plots the singular values of $\boldsymbol{R}_n^{-1}$ as a function of time for $D = 20$. Fig. (3.7b) plots the asymptotic rank $r_\infty$ of $\boldsymbol{R}_n^{-1}$ as a function of $D$. The latter agrees well with the average dimension of the unstable subspace $D_u$. . . . . 116
Figure 3.8: Annealing the model error covariance $\boldsymbol{R}_f^{-1}$ places more weight on the model as the EKF converges. This improves its ability to filter observation noise, and reduces $L_c$ compared with Figs. (3.3b) and (3.3d) . . . . . 117
Figure 3.9: Critical radius $\rho_c$ at which the estimation no longer succeeds, plotted as a function of the estimation window length $T$. Dots are sampled estimates, and the lines plot the inverse condition number of the Hessian $\nabla^2 A$. Note that WC-4DVar-X method is stable in the long $T$ limit, but SC-4DVar and WC-4DVar-U are not. . . . . 118
Figure 3.10: Cross-sections of the SC-4DVar objective function for various values $T$ plotted as a function of the initial error between the estimate and the optimal solution $|\boldsymbol{x}_0^* - \boldsymbol{x}_0|$. As the length of the estimation window $T$ grows, the basin around the global minimum collapses, making it virtually impossible to find. . . . . 119
Figure 3.11: Largest and smallest eigenvalues of the inverse Hessian $\nabla^2 A(\boldsymbol{\xi}^*)$ as a function of estimation window length $T$. With partial observations, the smallest tends toward zero as $T \to 0$ for all three methods. In the long $T$ limit, the largest grows exponentially for SC-4DVar and WC-4DVar-U. For WC-4DVar-X, it is asymptotically stable. . . . 119
Figure 3.12: Cross-sections of the SC-4DVar objective function, analogous to Fig. (3.10) but including a coupling term in the dynamical constraints. When $L > L_c$, dynamical instabilities in the forecast model are suppressed, and the objective function is regular. 120
Figure 3.13: Calculated values of the critical minimum number of observations $L_c$ required to construct a successful estimate, for three fixed-interval 4DVar methods with $T = 0.5/\lambda_{\max}$. . . . . 120
Figure 3.14: Model error annealing for WC-4DVar-X. Dashed lines and large dots indicate values of $|\boldsymbol{R}_f|$ where the action is optimally conditioned. . . . . 121

Figure 4.1: The critical minimum number of observations $L_c$ for time delay estimation methods, with $\tau = 10\,dt$. . . . . 144

Figure 4.2: The critical minimum number of *total* observations $L_c M_c$, for local initial conditions with no observation noise. The trend for TDVar-FB follows $D_u$ until roughly $D \approx 40$, above which additional 'physical' measurements are required. By contrast, the TD-EKF requires only $L = 1$ until $D > 80$. . . . . 144
Figure 4.3: The effect of the time delay is examined by repeating the experiments from Fig. (4.1) with $\tau = dt$. . . . . 145
Figure 4.4: Singular value spectra of $\nabla \mathcal{H}$ rescaled so that $\sigma_1 = 1$, for different values of $\tau$. Points are sampled across the attractor, with $D = M = 20$, $L = 1$. . . . . 145
Figure 4.5: Repeating Fig. (4.1) using disjoint estimation windows. The vast majority of results find $M_c = 1$ as the critical threshold, indicating that overlapping windows are needed for enhanced observational efficiency. . . . . 145
Figure 4.6: Critical observability limits of moving horizon estimation methods, using the three 4DVar techniques from Chap. (3). . . . . 146

ACKNOWLEDGEMENTS

I have gone through somewhat of an existential crisis while writing this thesis. The subject matter is not typically considered under the realm of physics, although arguably it has been an essential part of it since the beginning. Nevertheless, to enter a field that is so broadly applicable, and at the same time so dominated by mathematicians, statisticians, and engineers, one cannot help but wonder what it is that a physicist can contribute. Or perhaps more generally, what is the role of a modern physicist working outside traditional disciplinary boundaries?

In discussing these concerns with my adviser, he mentioned the quote on the preceding page. He has attributed it to Fermi, although I have found no record of that. Others say it is from Feynman, but perhaps it was Fermi who he heard it from first. At any rate, given that much of this thesis was written in the wee hours of the morning, I thought it appropriate. More to the point however, it sums up quite concisely my intentions for this work.

This is a thesis of ideas. Nothing will be formally proven, but rather justified in true physics style with hand-waving derivations. I now see the benefits to this way of thinking, and after spending the better part of a year wading through a mountain of literature, I have come to believe that this field is perhaps in need of a physicists touch, to distill the ideas down into their most fundamental form, and present them in the simplest way possible. Although realizing this goal in full goes well beyond what I could ever dream of accomplishing during a Ph.D. tenure, I hope that the ideas presented here will help provide part of a new foundation for future scientists, who like myself, have become inexorably fascinated with the truth and beauty that is the estimation problem.

With that, I would like to start by thanking my advisor Henry D.I. Abarbanel, without whom not of this would have been possible. I would also like to thank my committee members, in particular Mike Holst, Melvin Leok, and Philip Gill for their support and tolerance of my constant harassment. Thanks also to the U.S. Department of Energy Computational Science Graduate Fellowship program, for providing the resources and freedom to pursue my own research interests. Without their support, this thesis would have been much different (and perhaps much shorter). I would also like to thank Mark Kostuk, Nirag Kadakia, Paul Rozdeba, and Michael Eldridge for helpful conversations. Mark, in particular, was integral to my success early on in the program. Thanks as well to my parents and sisters.

Finally, thanks to my wife Grace and daughter Fiona. I would not have made it through without your unwavering love and support. I can only hope that wherever life takes us, we will always look back fondly on our time here — in our little home by the Pacific.

Chapter 2, in part is being prepared for submission for publication of the material. Rey, Daniel. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part is being prepared for submission for publication of the material. Rey, Daniel; Abarbanel, Henry DI. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part is being prepared for submission for publication of the material. Rey, Daniel; Abarbanel, Henry DI. The dissertation author was the primary investigator and author of this material.

VITA

| | |
|---|---|
| 2008 | B.S.E. in Engineering Physics, University of Illinois at Urbana-Champaign |
| 2014 | M.S. in Physics, University of California, San Diego |
| 2017 | Ph.D. in Physics with a Specialization in Computational Science, University of California, San Diego |

PUBLICATIONS

Abarbanel, Henry DI, Eve Armstrong, Daniel Breen, Nirag Kadakia, Daniel Rey, Sasha Sherman, and Daniel Margoliash. "A Unifying View of Synchronization for Data Assimilation in Complex Nonlinear Networks." *Chaos* (2017).

Kadakia, Nirag, Daniel Rey, Jingxin Ye, and Henry DI Abarbanel. "Symplectic structure of statistical variational data assimilation." *Quarterly Journal of the Royal Meteorological Society*. (2017).

An, Zhe, Daniel Rey, Jingxin Ye, and Henry DI Abarbanel. "Estimating the State of a Geophysical System with Sparse Observations: Time Delay Methods to Achieve Accurate Initial States for Prediction." *Nonlinear Processes in Geophysics* (2017).

Ye, Jingxin, Daniel Rey, Nirag Kadakia, Michael Eldridge, Uriel Morone, Paul Rozdeba, Henry DI Abarbanel, and John C. Quinn. "Systematic variational method for statistical nonlinear state and parameter estimation." *Physical Review E* 92, no. 5 (2015).

Schumann-Bischoff, Jan, Ulrich Parlitz, Henry DI Abarbanel, Mark Kostuk, Daniel Rey, Michael Eldridge, and Stefan Luther. "Basin structure of optimization based state and parameter estimation." *Chaos* 25, no. 5 (2015).

Rey, Daniel, Michael Eldridge, Mark Kostuk, Henry DI Abarbanel, Jan Schumann-Bischoff, and Ulrich Parlitz. "Accurate state and parameter estimation in nonlinear systems with sparse observations." *Physics Letters A* 378, no. 11 (2014): 869-873.

Rey, Daniel, Michael Eldridge, Uriel Morone, Henry DI Abarbanel, Ulrich Parlitz, and Jan Schumann-Bischoff. "Using waveform information in nonlinear data assimilation." *Physical Review E* 90, no. 6 (2014).

ABSTRACT OF THE DISSERTATION

**Chaos, observability and symplectic structure in optimal estimation**

by

Daniel Rey

Doctor of Philosophy in Physics with a Specialization in Computational Science

University of California, San Diego, 2017

Professor Henry D.I. Abarbanel, Chair

Observation, estimation and prediction are universal challenges that become especially difficult when the system under consideration is dynamical and chaotic. Chaos injects dynamical noise into the estimation process that must be suppressed to satisfy the necessary conditions for success: namely, synchronization of the estimate and the observed data. The ability to control the growth of errors is constrained by the spatiotemporal resolution of the observations, and often exhibits critical thresholds below which the probability of success becomes effectively zero. This thesis examines the connections between these limits and basic issues of complexity, conditioning, and instability in the observation and forecast models. The results suggest several new ideas to improve the collaborative design of combined observation, analysis, and forecast systems. Among these, the most notable is perhaps the fundamental role that symplectic structure plays in the remarkable observational efficiency of Kalman-based estimation methods.

# 1 Introduction

Our ability to model and understand complex dynamical behavior has greatly improved over the last century. As our models become more refined and our computational and data collection resources continues to grow, it brings to the forefront a fundamental problem. Using a model to identify, predict or control a systems behavior requires knowledge of the model's dynamical states, static parameters, and exogenous inputs. For instance, accurate predictions require accurate estimates to initialize the model forecasts. Typically however, many of these variables cannot be directly measured, and must therefore instead be inferred from observational data.

At its core, the estimation process is an inverse problem. It involves the fusion of observational data into estimates of the states and parameters of a computational model. These estimates may then be fed back into the model to generate forecasts, which among other things may be used to validate the model or make policy decisions. Alternatively, the estimates may also be used as input to a feedback control system, designed to push the system towards some target trajectory.

The state and parameter estimation problem constitutes one of the grandest and most universal scientific challenges. Its overall scope spans a wide range of disciplines and boasts virtually endless list of practical applications. From guidance of aircraft and satellites, to magnetic resonance imaging, to forecasts of climate, stock markets and neural activity — it is nearly impossible to identify an area of science or engineering that does not have to confront it.

This broad applicability, coupled with the unprecedented availability of data and computational power, has generated an explosion of algorithmic techniques for solving this problem. While the majority of these approaches are all based in one way or another on the framework of Bayesian inference, the distinction between them generally depends on the goals and constraints of the problem under consideration.

Among these constraints, the suitability of any given technique is largely determined by the availability of prior information. It is worth emphasizing however, that 'prior' information comes in many

forms. The most literal interpretation comes from the Bayesian description as an *a-priori* estimate of the model variables. Here, it is often used to initialize and regularize the search for the optimal values. In another sense however, prior information may also be viewed as the confidence one has in the mathematical model. After all, these models are typically the result of hundreds of years of scientific and intellectual effort. In this way, the availability of this type of prior information determines the underlying confidence that one has in the model. And this, together with the goals and constraints of the problem being addressed, determines the suitability of a given estimation technique.

For example, models based on physical principles, such as the transport processes underlying earth systems, obey rather strict constraints, such as conservation of mass, energy and momentum. As such, they are considerably more accurate and reliable than economic or social models, which must are largely based on perceived patterns in the data and must account for the inherent uncertainty of human behavior.

High-confidence models are not always available of course. When they are not, statistical and data driven models attempt to capture features and trends beyond the limits of first principles. When they are available however, models based on fundamentals tend to offer more reliable forecasts. And when these models are dynamical, their combination with time-series data provides access to a wealth of information hidden in the system dynamics.

Dynamical information serves as a regularizing constraint, to break the degeneracy inherent in the fact that the full state of the system simply cannot be directly observed. The best estimation methods harness this information, to improve accuracy and reliability of the estimates and forecasts, enhance their robustness to errors in the observations and the model, and reduce their dependence on prior information.

This thesis will describe how this is done, both in theory and through numerical simulations. It will also introduce a general framework for comparing the observational efficiency of among observation techniques. To begin however, the theoretical foundations for estimation will be laid out in the context of a simple example.

## 1.1 Predicting the outcome of a coin toss

Consider a simple game where one is asked to predicting the outcome of a series of coin tosses. The ability to model and predict the result inherently depends on the rules of the game, or more precisely on the availability of information.

Suppose for instance that one must make the call before the toss, but is allowed to directly observe the result of the coin falling directly on the floor. In this case, one might choose to model the outcome

probabilistically. If the coin is fair (*i.e.*, has equal likelihood of landing heads or tails) there is no edge on prediction, and this is the best that can be done given the circumstances.

If, on the other hand, one does not with to assume the coin is fair (or perhaps there is reason to suspect the coin might be weighted to give preference to one side over another) then one may instead try to infer the relative probability of the coin from the data. The problem is then to estimate the parameter $\theta$ (representing the probability of the coin landing heads up) given the result of $N$ tosses $\mathcal{Y}_{1:N} := \{y_n\}_{n=1}^{N}$, where $y_n$ is the result of the $n^{th}$ toss ($y_n = 1$ if heads, $y_n = 0$ if tails). More specifically, the goal is to construct a series of estimates $\theta_{n|n}$, where $\theta_{n|n}$ represents the best estimate given the data up to time $\boldsymbol{y}_n$. This allows the estimate to be updated during the course of the game, with the hope of improving one's odds of success in later rounds.

### 1.1.1 Filtered estimates

In the language of estimation theory, the estimate $\theta_{n|n}$ is known as a *filtered estimate*. Its calculation involves a standard application of Bayesian inference, where the goal is to evaluate the conditional distribution $p(\theta_n|\mathcal{Y}_{1:n})$. A recursive estimate can be constructed using Bayes' rule

$$\begin{aligned} p(\theta_n|\mathcal{Y}_{1:n}) &= \int d\theta_{n-1}\, \frac{p(\theta_n, \theta_{n-1}, y_n, \mathcal{Y}_{1:n-1})}{p(\mathcal{Y}_{1:n})} \\ &= \int d\theta_{n-1}\, \frac{p(y_n|\theta_n, \theta_{n-1}, \mathcal{Y}_{1:n-1})\, p(\theta_n|\theta_{n-1}, \mathcal{Y}_{1:n-1})\, p(\theta_{n-1}, \mathcal{Y}_{1:n-1})}{p(\mathcal{Y}_{1:n})} \\ &= \frac{p(y_n|\theta_n)}{p(y_n)} \int d\theta_{n-1}\, p(\theta_n|\theta_{n-1})\, p(\theta_{n-1}|\mathcal{Y}_{1:n-1}). \end{aligned}$$

Three assumptions were made in the third line:

1. mutual independence of measurements

$$p(\mathcal{Y}_{1:n}) \to p(\mathcal{Y}_{1:n-1})\, p(y_n)$$

2. conditional independence of measurements

$$p(y_n|\theta_n, \theta_{n-1}, \mathcal{Y}_{1:n-1}) \to p(y_n|\theta_n)$$

3. the Markov transition property

$$p(\theta_n|\theta_{n-1}, \mathcal{Y}_{1:n-1}) \to p(\theta_n|\theta_{n-1})$$

Since $y_n$ appears outside the integral, the recursion can be split into two steps: a time update

$$p(\theta_n|\mathcal{Y}_{1:n-1}) = \int d\theta_{n-1}\, p(\theta_n|\theta_{n-1})\, p(\theta_{n-1}|\mathcal{Y}_{1:n-1})$$

and a measurement update

$$p(\theta_n|\mathcal{Y}_{1:n}) = \frac{p(y_n|\theta_n)\, p(\theta_n|\mathcal{Y}_{1:n-1})}{p(y_n)}.$$

If one assumes the coin does not change between trials, this is a static parameter estimation problem. The transition probability is therefore a delta function $p(\theta_n|\theta_{n-1}) \to \delta(\theta_n - \theta_{n-1})$, and the time update step can effectively be ignored since $p(\theta_n|\mathcal{Y}_{1:n-1}) = p(\theta_{n-1}|\mathcal{Y}_{1:n-1})$. Also, the conditional probability is known

$$p(y_n|\theta_n) = \theta_n\, \delta(1 - y_n) + (1 - \theta_n)\, \delta(y_n).$$

But the denominator is not. And it is not possible to determine overall probability $p(y_n)$ of a given toss. This is not a problem however, if its value does not change between trials. In this case, it acts as a normalizing constant to ensure the probability sums to unity. The calculation can therefore be performed without knowledge of $p(y_n)$, provided the resulting distribution is then normalized.

To initialize the recursion, a 'prior' distribution is also needed. Any distribution can be selected. But the simplest choice is perhaps a delta function $p(\theta_1|\mathcal{Y}_{1:0}) = \delta(\theta_1 - \theta_0)$. With this choice, the conditional probability $p(\theta_n|\mathcal{Y}_{1:n})$ of the parameter $\theta_n$ after observing $m$ heads in $n$ tosses is a beta distribution

$$p(\theta_n|\mathcal{Y}_{1:n}) = \frac{\Gamma[2+n]}{\Gamma[1+m]\,\Gamma[1-m+n]}\, \theta_n^m\, (1-\theta_n)^{n-m}\, \delta(\theta_n - \theta_0). \tag{1.1}$$

Note also that here the order in which the results occur does not matter.

This distribution can be used to calculate any desired statistic of the parameter $\theta_n$. But since one must choose a specific value for the next outcome, the 'optimal' filtered estimate of $\theta_{n|n} \equiv \theta_n^*$ is chosen to be the one that maximizes likelihood of success, given the preceding data $\mathcal{Y}_{1:n}$. In other words, the optimal parameter $\theta_n^*$ corresponds to the peak (or conditional mode) of the distribution $p(\theta_n|\mathcal{Y}_{1:n})$. This value can be found by maximizing the distribution, although in practice this often involves very large (or small)

numbers. So it is typically better to maximize its log instead, which is also known as its log-likelihood. For the purpose of later discussion, an equivalent formulation is used, which instead minimizes the negative log

$$\theta_n^* = -\operatorname*{argmin}_\theta \log[p(\theta_n|\mathcal{Y}_{1:n})] = -\operatorname*{argmin}_\theta \Big( m\log[\theta] + (n-m)\log[1-\theta] \Big) = \frac{m}{n} =: \theta_{n|n}.$$

This implies rather intuitively that the most likely outcome of the next toss is the one that has appeared most frequently thus far.[1]

### 1.1.2 Smoothed estimates

There is also another type of estimate that can be performed. Suppose after observing the outcome of $N$ trials one were to ask whether a better prediction of $y_n$ could have been made, given knowledge the remaining $n-N$ results. The corresponding estimate $\theta_{n|N}$ (*i.e.*, given the $\mathcal{Y}_{1:N}$) is called the *smoothed estimate*. While this type of 'reanalysis' does not have any benefits for predicting the $N+1$ result, it is useful nonetheless: *e.g.*, for model validation purposes.

There are two predominant ways of constructing smoothed estimates. The first is a recursive approach takes the result of the filtered estimate and propagates it backwards (*i.e.*, from $n \to n-1$). This again involves rewriting the conditional distribution using Bayes' rule ([9]),

$$\begin{aligned} p(\theta_n|\mathcal{Y}_{1:N}) &= \int d\theta_{n+1}\, \frac{p(\theta_n, \theta_{n+1}, \mathcal{Y}_{1:N})}{p(\mathcal{Y}_{1:N})} \\ &= \int d\theta_{n+1}\, p(\theta_{n+1}|\mathcal{Y}_{1:N})\, p(\theta_n|\theta_{n+1}, \mathcal{Y}_{1:N}) \\ &= \int d\theta_{n+1}\, p(\theta_{n+1}|\mathcal{Y}_{1:N})\, p(\theta_n|\theta_{n+1}, \mathcal{Y}_{1:n}) \\ &= p(\theta_n|\mathcal{Y}_{1:n}) \int d\theta_{n+1}\, \frac{p(\theta_{n+1}|\mathcal{Y}_{1:N})\, p(\theta_{n+1}|\theta_n)}{p(\theta_{n+1}|\mathcal{Y}_{1:n})}. \end{aligned}$$

Here, the Markov transition property implies that $p(\theta_n|\theta_{n+1}, \mathcal{Y}_{1:N}) \to p(\theta_n|\theta_{n+1}, \mathcal{Y}_{1:n})$. Note that computing this 'backwards pass' also requires the filtered result. Together, they form the basis for two-pass recursive smoothing algorithms, which also called forward-backward, or sweep methods. These methods will be discussed in more detail in Chap. (2).

Alternatively, one may eschew the recursive formulation altogether. Expanding the conditional density gives

$$p(\Theta_{1:N}|\mathcal{Y}_{1:N}) = \prod_{n=1}^{N} \exp[\mathrm{CMI}(\theta_n, y_n|y_{1:n-1})]\, p(\theta_n|\theta_{n-1})\, p(\theta_0), \tag{1.2}$$

[1] Note also that the normalization constant does not affect this solution.

where

$$\mathrm{CMI}(\theta_n, y_n|y_{1:n-1}) := \log\left[\frac{p(\theta_n, y_n|y_{1:n-1})}{p(y_n|y_{1:n-1})\,p(\theta_n|y_{1:n-1})}\right] \to \log\left[\frac{p(y_n|\theta_n)}{p(y_n)}\right]$$

$$p(\boldsymbol{\Theta}_{1:N}|\mathcal{Y}_{1:N}) = \prod_{n=1}^{N} \exp[\mathrm{CMI}(\theta_n, y_n|y_{1:n-1})]\, p(\theta_n|\theta_{n-1})\, p(\theta_0),$$

is the *conditional mutual information* between $y_n$ and $\theta_n$ ([41]), and the right hand limit holds when the measurements are mutually, and conditionally independent. This form clearly distinguishes among the influence of the prior $p(\theta_0)$, the transition density $p(\theta_n|\theta_{n-1})$, and the information provided by the observations. It also has a direct interpretation as a statistical path integral, which will be discussed momentarily.

Both approaches give the same estimate for the parameter $\theta_n$. In this case, the conditional distribution is the same as Eqn. (1.1) above, except now with $M$ heads obtained over $N$ trials. The resulting maximum-likelihood estimate for $\theta_{n|N} = M/N$. While this is different than the filtered result, the overall strategy remains the same.

If the coin is indeed fair, this type of stochastic model provides no benefits in terms of prediction accuracy. This is related to the fact that this example is a static noiseless parameter estimation problem, so $\theta_{n+1} = \theta_n$, and that the order of events does not matter. It is possible to do better however.

An alternative situation will now be described, which involves modeling the coin physically, and estimating its time-varying states. This requires modifying the rules of the game. But doing so can provide a considerable edge in predictability.

### 1.1.3 Modeling coin physics

Consider a slight change in the rules where one is now allowed to decide the outcome of the toss while the coin is in the air. To simplify the situation further, suppose that the coin is not allowed to drop on the floor, but rather falls onto a pillow to minimize its bounce. Also, let there be no technological restrictions. So one may use a camera to record the toss, and a computer to analyze the data before making a decision. The question is then: how best to use this information to make better predictions?

A number of possible strategies come to mind. For instance, one might use the camera to determine which side is facing up the instant before the coin hits the pillow. After many trials, it may be possible build a statistical model that uses this information to an advantage. However, this approach does not take into account important factors such as its center of mass position, orientation, and linear/angular momentum. It seems likely that knowledge of these quantities will help improve the predictions, and

perhaps even allow one to determine the outcome earlier in the trajectory, with enough time to call the result before it lands. After all, once the coin is in the air, its trajectory is determined by rigid body mechanics. However, it is also clear that these physical quantities are not known or directly observed (by a camera at least). Thus, to adopt this approach, they must be inferred from the available data.

This estimation problem proceeds analogously to the one described above, with one important difference. The relative weight of the coin was assumed to be a static parameter, whereas these physical quantities are clearly time-varying. How then does one estimate these dynamical quantities? The answer is to incorporate this dynamical information into the transition probability, through the development of a state-space model.

Physical models for coin-toss mechanics were examined by [98, 45]. These models use rigid body mechanics to develop a set of ordinary differential equations (ODEs) that describes the time evolution of its dynamical variables, which include center of mass position and momentum, orientation, and angular momentum. These variables may all be combined into a single, time-dependent, *state vector* $\boldsymbol{x}(t) \equiv \boldsymbol{x}_t$. The resulting ODE may generally be represented in continuous time as

$$\dot{\boldsymbol{x}}_t := \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}_t, t) =: \boldsymbol{f}_t(\boldsymbol{x}_t),$$

where $\boldsymbol{f}_t$ is typically called a *state-space model* and incorporates all the physics of the problem.[2] In practice, the equations $\boldsymbol{f}_t$ are often nonlinear, and do not have a closed-form analytical solution. Thus, the approximately trajectory must be solved for computationally, using a suitable method of which there are many. This effectively transforms the continuous time problem into discrete time, by breaking it into a series of steps $\boldsymbol{x}_n \equiv \boldsymbol{x}(t_n)$ with $\{t_n\}|_{n=1}^{N}$ at an increment $dt$, which here is assumed uniform. The model may then be expressed in discrete time as,

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}_n(\boldsymbol{x}_n),$$

where $\boldsymbol{F}_n$ is the discretized version of the continuous model $\boldsymbol{f}_t$. In what follows, it will often be assumed that the model is autonomous (*i.e.*, not explicitly time-dependent), so the time index is neglected. This assumption is not required, but helps simplify the expressions.

In addition to the dynamic states $\boldsymbol{x}_n$, the model $\boldsymbol{F}_n$ may also include static parameters $\boldsymbol{\theta}$ that

[2] Note that this definition of a state is similar but distinct from a thermodynamic state, in which a number of particles may occupy the same energy state but have different positions and momenta.

also need to be estimated. These parameters may be simply treated as additional states in the model, with $d\boldsymbol{\theta}_t/dt = 0$. The estimation of dynamic parameters (also called model inputs or forcing variables) is possible as well. But this is typically more difficult as the problem is often highly underdetermined.

### 1.1.4 The observation model

In addition to the dynamical forecast model $\boldsymbol{F}$, an *observation model* is needed to relate the model states $\boldsymbol{x}_n$ to the observed data $\boldsymbol{y}_n$. This model may be generally written as

$$\boldsymbol{y}_n = \boldsymbol{h}_n(\boldsymbol{x}_n).$$

In the present context, $\boldsymbol{y}_n$ is the image from the camera at time $t_n$. In contrast to the previous example, where $y_n$ denoted the outcome of the $n^{th}$ trial, here the data is a vector, of say pixel values, taken from single toss. Together, $\boldsymbol{f}$ and $\boldsymbol{h}$ comprise the *state-space estimation model*.

The development of a suitable observation model is perhaps the most difficult part of this application. One could for instance, take a direct approach that uses ray-tracing to simulate the camera input. This would require knowledge of certain camera specifications, such as its field of view and aperture ratio, as well as additional preprocessing to remove the background from the image, which the coin model knows nothing about.

An alternative would be to use a computer vision algorithm to track the coin and directly estimate the relative position and velocity of its center of mass, along with other states such as its orientation and rate of rotation. This requires solving an accessory estimation problem to preprocess the data, by converting pixel values into estimates of the coin states. Computer vision methods are typically built on the same framework of Bayesian inference, but involve heuristic or phenomenological models instead of physical ones. This is not a problem however. The Bayesian framework does not care what type of model is used, and provides a way to interface the two methodologies.

### 1.1.5 The estimation action

With these definitions, Eqn. (1.2) can be rewritten as

$$p(\mathcal{X}_{1:N}|\mathcal{Y}_{1:N}) = \prod_{n=1}^{N} \exp[\mathrm{CMI}(\boldsymbol{x}_n, \boldsymbol{y}_n|\boldsymbol{y}_{1:n-1})]\, p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})\, p(\boldsymbol{x}_0), \tag{1.3}$$

or equivalently as

$$p(\mathcal{X}_{1:N}|\mathcal{Y}_{1:N}) \propto \exp[-A(\mathcal{X}_{1:N})]$$

where

$$A(\mathcal{X}_{1:N},\mathcal{Y}_{1:N}) \propto -\sum_{n=1}^{N} \underbrace{\mathrm{CMI}(\boldsymbol{x}_n,\boldsymbol{y}_n|\boldsymbol{y}_{1:n-1})}_{\text{Measurement error}} \quad - \quad \sum_{n=1}^{N} \underbrace{\log[p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})]}_{\text{Model error}} \quad - \underbrace{\log[p(\boldsymbol{x}_0)]}_{\text{Prior uncertainty}} \tag{1.4}$$

separates the respective contributions from the model, the observations, and the prior. The function $A(\mathcal{X}_{1:N},\mathcal{Y}_{1:N}) \equiv A(\mathcal{X}_{1:N})$[3] will be called the *estimation action*, in analogy to the action of a statistical path integral ([3]).

The path integral provides a way to compute expected values of any statistical function of the path $\phi(\mathcal{X})$

$$\langle\phi(\mathcal{X})\rangle = \frac{\int d\mathcal{X}\,\phi(\mathcal{X})\exp[-A(\mathcal{X})]}{\int d\mathcal{X}\exp[-A(\mathcal{X})]}.$$

But this requires evaluating a high-dimensional integral, which varies all states $\mathcal{X}$ at all times along the path. There are two general approaches available for evaluating such integrals: Monte Carlo sampling, and the Laplace approximation.

### 1.1.6 Evaluating the path integral

The Monte Carlo approach performs what is essentially stochastic descent on the action. In its simplest form, it randomly chooses a new path at each iteration. If the action is lower than the previous value, the new path is accepted. If not, it is accepted with a probability proportional to the difference between the two action values. Allowing for relatively small increases in the action value helps mitigate the problem of becoming stuck in local minima. While this method is guaranteed to converge eventually to the global minimum, it may take an undesirably long time. Accelerating the convergence requires a good choice of the proposal distribution from which the new path is selected, but this is not easy since the distribution should ideally approximate the unknown density. This is the basic issue with Monte Carlo methods — one often needs a good estimate to get a good estimate ([36]).

The Laplace method is an alternative approach, which makes an asymptotic approximation of the integral by assuming most of its value is consolidated around its local minima. Therefore, it is primarily concerned with finding minimizers $\mathcal{X}^*$ where $|\nabla A(\mathcal{X}^*)| = 0$ and $\nabla^2 A(\mathcal{X}^*)$ is positive definite.

[3] Since the data $\mathcal{Y}_{1:N}$ is fixed for the estimation, the functional dependence will be suppressed.

The minimizers corresponding to the lowest values $A(\mathcal{X}^*)$ asymptotically dominate the contribution to the integral. The maximum likelihood estimate is the state $\mathcal{X}^*$ corresponding to the global minimum.

Both methods may be considered numerical optimization techniques that minimize the action as an objective function. The Laplace approximation directly reformulates the path integral as an optimization problem, for which a number of general purpose techniques exist. These techniques can utilize local gradient and curvature information to speed up convergence, which may be a benefit or drawback, depending on how difficult it is to implement the derivatives.

### 1.1.7 The effective action

Regardless of which approach is chosen, to actually perform the search requires specifying the functional form for the distribution of measurement and model errors. This information may or may not be available, depending on the problem at hand. If these distributions are unknown, then one must make some assumptions both regarding the functional form and the relative magnitude of the errors. In the limit that both models are without error, the measurement and model error terms in Eqn. (1.4) become $\delta$ functions

$$\exp[\mathrm{CMI}(\boldsymbol{x}_n, \boldsymbol{y}_n | \boldsymbol{y}_{1:n-1})] \rightarrow \delta\big(\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)\big) \qquad\qquad p(\boldsymbol{x}_{n+1}|\boldsymbol{x}_n) \rightarrow \delta\big(\boldsymbol{x}_{n+1} - \boldsymbol{F}_n(\boldsymbol{x}_n)\big).$$

But from an optimization perspective, these functions are typically difficult to work with. Alternatively, it is often assumed that the errors are Gaussian distributed. This may be viewed as a 'broadening' of the delta function, by introduce a small amount of (Gaussian) noise into the otherwise deterministic dynamics. The result is a more tractable optimization problem, as the action becomes a (nonlinear) least squares objective function

$$A(\mathcal{X}) \propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=1}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} |\boldsymbol{x}_{n+1} - \boldsymbol{F}_n(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_f}.$$

The terms $\boldsymbol{R}_m$, $\boldsymbol{R}_f$ and $\boldsymbol{R}_b$ are positive definite (inverse) covariance matrices corresponding to measurement, model, and prior errors. The prior error is also assumed to be specified relative to some known 'background' state $\boldsymbol{x}_b$. Actual values for these parameters may not be available, so in practice the choices are often tuned to a particular problem.

These assumptions give rise to an *effective action*, which may or may not reflect the underlying 'truth' behind how the errors are generated. But the framework is general in the sense that if one has reason to believe that the errors of any of these components obey some other distribution (*e.g.*, Poisson statistics), then that distribution can be simply substituted. Gaussian statistics often work fairly well

in practice, especially for situations where the forecast model is physical and thus of relatively high confidence. However, given no additional information about these errors, these decisions are ultimately left up to the modeler, and inherently depend on the problem at hand.

### 1.1.8 Observability

The above framework provides a way of blending information from the observations and a physical model into an estimate for the dynamical state of the coin. The goal is to balance these two factors in a way that gives the best predictive performance. An estimate that relies too much on the data is said to be *overfit*, and is susceptible to noise in the observations. On the other hand, relying too much on the model can make the estimation process difficult and at times intractable (this will be discussed in more detail later). In practice, the model and the measurements will are be error free. The optimal weighting of the two terms is necessarily a judgment call, although there are empirical ways to assess whether the result is overfit, such as the L-curve ([70]).

The success of this endeavor critically depends on the availability of information, such as the number and type of observations. Any dynamical states that are not directly observed must instead be inferred from the combination of the observation and forecast models. The issue of whether the given observations are sufficient in this regard is called *observability*.

Even the simplest systems exhibit critical thresholds in the density of observations, below which the estimation process routinely fails. But these thresholds are not well-understood, as they depend in a complex manner on the three core components of the problem: the model, the data, and the estimation algorithm. While optimal control theory offers technical conditions under which the process is guaranteed to fail (*e.g.*, [182]), these are not particularly useful for large complex problems. In practice, many such systems are far too ill-conditioned to admit accurate and reliable solutions. There is therefore a need for developing methods to analyze these critical thresholds empirically, as a function of various parameters and constraints of the problem.

A substantial portion of this thesis is dedicated to this task. For the current example, the issue may be posed as follows. Do cameras alone provide enough information for this task? How many cameras are needed? What resolution is required? And cameras alone are not sufficient, what additional observations are needed/available to meet these basic requirements? Questions like these may be addressed initially using simulated data, by systematically testing the assumptions and constraints of the combined forecast, observation, and analysis systems. If the chosen techniques do not work with simulated data, there is little

chance they will succeed with real data.

### 1.1.9 Predictability and complexity

Once the estimates are obtained they may be fed back into the dynamical model to forecast the state trajectory. The fidelity of these predictions inherently depends on the model. A deterministic model will provide a deterministic prediction, while a stochastic model offers a distribution of forecasts. The presence of stochasticity or unmodeled dynamics will shorten the prediction horizon: the length of time in which predictions are expected to be accurate. This is not all however, as even a perfect and deterministic model may exhibit chaotic behavior. This causes errors in the forecast to diverge exponentially in time, at roughly a rate of $\exp(\lambda_{\max} t)$, where $\lambda_{\max}$ is the maximum global Lyapunov exponent of the system ([2]). The errors will eventually asymptote to a value that is roughly the size of the attractor, which effectively puts an upper limit on predictability that depends on the accuracy of the estimate.

Returning to the problem at hand, supposing one is able to accurately estimate and predict the coin's trajectory, these estimates still must be translated into a prediction of the outcome. Such a model for the outcome seems like it might be difficult to implement, especially if the coin is allowed to bounce, which is why initially it was initially suggested the coin be allowed to fall on a pillow. A bouncing coin is chaotic, in the sense that small changes to initial conditions produce widely different results. This uncertainty is largely due unstable minima in its potential energy function, since a coin dropped on its edge carries more uncertainty than a coin landing on its side. But as the outcome is binary, the process of settling into a stable minimum may be viewed as a simple realization of spontaneous symmetry breaking.

Developing an understanding of the origins of uncertainty and limits of our ability to model certain phenomena is one of the core goals in the field of complexity theory. It may be that certain systems are beyond our ability to model from first principles. Consider for instance the problem of predicting the outcome of a coin that is caught by hand. While the observation analysis system may be able guess the result from the relative orientation between the coin and the palm, there is as of yet no conceivable way of predicting when the tosser will begin to move to catch the coin.

On the other hand, given repeated trials it may be possible to build a statistical model that predicts the catching patterns of a particular person. And this in conjunction with the estimated trajectory of the coin may indeed offer an edge on prediction. The trouble is that the resulting model will likely have to be recalibrated if conditions change: such as the person tossing the coin. Thus, while models built on data provide a way of predicting what is beyond our ability to comprehend, they are nonetheless fragile.

Because when the process producing the data changes, the results may no longer be reliable.

## 1.2 An historical perspective

Estimation is one of the most ubiquitous and fundamental problems of existence; one that even the most primitive organisms have evolved to solve with relative ease and efficiency. Consider an insect for example. At each instant, it collects sensory information about its external, which it analyses against some internal 'model', and then uses the result to inform its decisions and movements. To us, this internal process of fusing sensory information with prior knowledge is largely second nature.

While the mechanisms underlying the brain's execution of this task are still very much a mystery, the estimation problem has long been the subject of intellectual and scientific effort. The resulting theory has connections to a remarkably broad range of mathematical disciplines. From statistics, to data assimilation, numerical optimization, electrical computational and control engineering, and even physics — each field has developed its own algorithms, techniques and solutions to this problem.

This section attempts to give a brief account of their historical development. It is not intended to be comprehensive; such a treatise could fill an entire book (or perhaps several). The goal is to provide a more global perspective on the problem that will lay the groundwork for the rest of this thesis.

### 1.2.1 Pre-twentieth century

It is difficult to pinpoint the exact origins of the 'theory' of estimation. Early curve and surface fitting efforts date back to the dawn of civilization, with the ancient Babylonians, Egyptians, and Greeks ([142, 140]). But it was apparently Galileo who in 1632 made the first published attempts to minimize various functions of observed errors ([87]). This served as the predecessor for the discovery of the method of least squares early $19^th$ century. Although first published by Legendre in 1805, Gauss claimed to have knowledge of it as early as a teenager in 1795. These claims are supported by the fact that he famously used it to predict the orbit of the planet Ceres in 1801. Least squares minimization has since become a cornerstone of estimation theory.

In 1697 (not long after Galileo) Johann Bernoulli (1667-1748) published the solution to the brachistochrone problem. From this solution, the calculus of variations was born. The theory was influenced by some of history's greatest mathematical minds — such as Issac Newton (1642-1726), Leonhard Euler (1707-1783), Ludovico Lagrange (1736-1813), Pierre Laplace (1749-1827), Adrien Legendre (1752-1833),

Carl Jacobi (1743-1819), William Hamilton (1805-1865), Karl Weierstrass (1815-1897), Henry Poincaré (1854-1912), David Hilbert (1862-1943), Constantin Carathéodory (1873-1950), Adolph Mayer (1839-1907) and Oskar Bolza (1857-1942). This work laid the foundations for the subsequent development of the modern theory of optimal estimation and control ([186]).

The minister Thomas Bayes (1701-1761) also deserves to be mentioned. Bayes' theorem on conditional probability could arguably be the single most important contribution to estimation theory, although this was not realized until much later.

### 1.2.2 The Wiener-Komogorov filter

The early $20^{th}$ century witnessed the birth of least squares estimation of stochastic processes. The first studies were made by Norbert Wiener (1894-1964), Andrey Kolmogorov (1903-1987), Mark Krein (1907-1989). Kolmogorov gave a comprehensive treatment of the prediction problem for discrete-time stationary processes. Krein extended these results to continuous time, and noted connections to earlier work by Szegö regarding orthogonal polynomials on the unit circle.

Wiener independently formulated and solved the continuous-time linear predictor problem, and also considered the construction of filtered estimates of static (*i.e.*, linear time invariant) processes with observation noise. The solutions may be represented as integral equations, which are either Fredholm equations of the second kind (for smoothed estimates) or the Wiener-Hopf equation (for filtered estimates). Both equations were known at the time, although interestingly the latter first arose in 1894 while studying the radiation equilibrium of a star ([109]). For an in-depth survey of linear filtering and full bibliography see [87].

### 1.2.3 Stochastic processes, information, and statistics

The work of Wiener and Kolmogorov along with earlier work by Andrey Markov (1855-1922) led to the development of the rigorous theory of stochastic processes in the 1940s by Kiyoshi Itô (1915-2008) and Rusian Stratonovich (1930-1997). This, along with the work of Cameron-Martin ([31]) (and later Girsonov [62]) had a profound impact on the measure theoretic formulation of stochastic estimation. It led directly to the development of the Kusher-Stratonovich equation ([104]) and the Duncan-Mortensen-Zakai equation ([215]), which are essentially the Kolmogorov-forward equations for the normalized and unnormalized distributions respectively.

Estimation theory was also heavily influenced by the roughly concurrent discovery of information

theory and statistics. For the former, the innovations technique introduced by Shannon and Bode ([26]), and later presented in its modern form by Kailath ([86, 88]), was used extensively for several decades. Regarding the latter, the notion of a sufficient statistic — initially by Fisher and later generalized to exponential distributions by Pitman-Koopman-Darmois ([160]) — along with the Cramér-Rao lower bound ([162]), helped formalize the fundamental limits of what one should expect from an optimal estimator.

### 1.2.4 Kalman, Bellman, and Pontryagin

The greatest achievements came in the 1950s. Richard Bellman (1920-1984) introduced the Bellman equation for dynamic programming in 1952,[4] and Lev Pontryagin (1908-1988) formulated the maximum/minimum principle in 1956. These are perhaps the two most influential contributions to the deterministic theory of optimal control and estimation.

The decade then culminated with the discovery of the single most important estimation algorithm to date: the Kalman filter (1960) by Rudolf Kálmán (1930-2016). One of Kalman's main innovations was the use of state space models to generalize Wiener's earlier results to the time-varying models. Interestingly, his second was apparently rejected by a prominant electrical engineering journal, due to a reviewer who purportedly stated "it cannot possible be true" ([77]). Nonetheless, it found an enormous number of engineering applications during this time period, including guidance navigation and control of ballistic missiles and space shuttles. It also spurred the subsequent development of Kalman smoothing techniques (*e.g.*, [164]), along with the identification of the inherent ties between the Kalman solution and the Bayesian framework of least squares maximum likelihood estimation.

### 1.2.5 Exact finite-dimensional nonlinear filters

The success of the Kalman filter revolutionized the theory of estimation and ushered in a new era of research that involved trying to extend it to nonlinear forecast and observation models. These attempts may be generally classified as either exact or approximate. Exact approaches seek solutions to the Duncan-Mortensen-Zakai equation, particularly ones where the dimension of the filter remains finite in time. A few cases of these *exact finite-dimensional filters* were identified, for certain models. The first was given by [17] for gradient models $\boldsymbol{f} = \nabla\phi$, and was subsequently extended by [44] to the exponential family of distributions.

In 1981 Brockett ([28]) and Mitter ([130]) independently introduced the concept of an *estimation*

[4] Interestingly, this had immediate connections to computational methods in radiative transfer ([16]).

*algebra*, and the conjecture that an exact finite-dimensional filter exists if and only if the estimation algebra admits a finite representation. These conditions were further elaborated in a series of papers by Stephen Yau ([212]), who showed that the structure of the antisymmetric tensor $\boldsymbol{\Omega} := \nabla \boldsymbol{f} - \nabla^{\dagger} \boldsymbol{f}$ introduced by [208] plays a crucial role in their existence. All this culminated in the filtering algorithm given by [211], which generalizes the Kalman-Bucy and Benes filters. Ultimately however, exact methods are only applicable to very specific models. For most practical purposes, approximations are needed.

### 1.2.6 Variational methods

As mentioned, approximate techniques tend to fall in one of two categories: Monte Carlo methods, and variational methods — although hybrid approaches also exist. Variational methods are based on the Laplace approximation to the path integral, and effectively reduce the estimation process to the recursive solution of a nonlinear optimization problem. This optimization based approach to state estimation has roots in dynamical systems ([54]) and optimal control theory ([22]), with collocation methods and direct/indirect single/multiple shooting. For a thorough overview of these techniques, see [20].

With the variational approximation, one also has to decide whether to try to optimize the full trajectory at once, or break it into smaller parts. The former is more tractable when the estimation window is long. But the result depends inherently on how the splitting is done.

Consider the data assimilation method known as 4DVar, which has become popular in operational numerical weather prediction since the mid 1990s. These methods use a sliding estimation window, in which new observations are analyzed each day, using the previous days' forecast as a regularization term ([82]). The main benefit of this approach is computational efficiency however. The use of adjoint methods avoids having to store and manipulate the error covariance, which reduces the memory complexity from $O(D^2)$ to $O(D)$ (see *e.g.*, [40]). For linear models, the resulting estimate is equivalent to the Kalman filter, as the length of the estimation window becomes asymptotically large ([115]).

The 4DVar approach may be viewed more generally within the context of moving horizon estimation ([8]), which is the estimation dual to model predictive control. The use of a moving window improves the accuracy and robustness of the filter, a fact which has apparently been rediscovered several times in different contexts. An alternative perspective, which connects it to the theory of attractor reconstruction ([187]), will be presented in Chap. (4).

### 1.2.7 Ensemble and particle methods

The most direct nonlinear extension of the Kalman filter is the 'extended Kalman filter', which uses the tangent linear approximation to the forecast model to compute the estimated error covariance. This method has a number of drawbacks however, not the least of which is that it tends to underestimate the error covariance. More recently, the extended Kalman filter has been replaced by a variety of sampling methods, such as the unscented Kalman filter ([84]) and related approaches such as the ensemble Kalman filter ([50]). These methods seek to improve accuracy and robustness by using ensemble of model states to approximate the time evolution of the estimated error covariance. As an added benefit, the derivatives of the model are not required, which facilitates their implementation.

On the other end of the spectrum, are particle methods. Introduced as a 'chainless' alternative to Markov Chain Monte Carlo, these methods approximate the conditional distribution $p(\mathcal{X}|\mathcal{Y})$ as a sum of delta functions. In contrast to ensemble filtering, where the ensemble describes the mean and covariance of a single Gaussian, particle methods can describe an arbitrary distribution. There are also approaches that interpolate between these two extremes, by representing the target distribution as a mixture of Gaussians ([102]). Connections to the variational approximation have also been recently established ([36]).

### 1.2.8 Moving forward

The recent growth in computation and data collection capability has made the estimation problem important and popular than ever. A vast number of solutions have been proposed, each with their own strengths and weaknesses. But it is often difficult to tell *a-priori* which methods are most suitable for a given problem. Moreover, there is substantial overlap between many of these algorithms, which at times are nearly indistinguishable. This is in part due to the fact that its historical development touches such a wide variety of mathematical disciplines that it has become increasingly difficult to get a broad overview of both the problem and the available solutions.

There is therefore a need, moving forward, to unify both the language and shared ideas among the various fields interested in this problem. This involves extending the focus, to emphasize not just the novelty of a particular algorithm or approach, but also how it fits within the wider context of the existing theory. The latter, as stated, can be difficult. It requires reaching beyond disciplinary boundaries, to connect new ideas to the larger body of work. But the results will provide both a deeper understanding of the fundamental limits of observation, estimation and prediction, as well as a broader foundation for innovation, from which the next generation of algorithms will emerge.

## 1.3 Overview of this thesis

This thesis seeks to begin the process of bringing together some of the less commonly known developments in optimal estimation theory. It will focus mainly on the deterministic interpretation of the state estimation problem. As mentioned, static parameter estimation may be viewed as a subset of this problem, by considering the parameters as additional states, with trivial dynamics $d\theta/dt = 0$. Dynamic parameter estimation — *i.e.*, inferring the dynamical inputs or forcing functions — is a more difficult task, which will not be considered here.

Chap. (2) reexamines the original Kalman filtering/smoothing theory from this perspective, and develops the links between optimal estimation, dynamic programming, and Hamiltonian mechanics. A number of distinct but closely related techniques will be compared, which are all equivalent to the Kalman filter in the linear Gaussian case, but perform quite differently in the nonlinear case. This study provides evidence that symplectic structure plays an important role in the optimality of these methods, although this is not formally proven.

Chap. (3) discusses the issues of observability and conditioning of dynamical inverse problems. In particular, it aims to address the question: how many observations are needed to guarantee the accuracy and reliability of the estimate and the resulting forecasts? A computational framework will be introduced to assess the observational efficiency of filtering and smoothing methods, and examine how these factors vary as a function of the three main components of the problem: the forecast and observation models, the data, and the estimation algorithm. An instructive example will be given using the chaotic Lorenz 1996 model ([118]). The results show that Kalman methods contain an inherent mechanism for adaptively targeting and controlling the unstable subspace of the forecasting model. This is a necessary condition for synchronizing the model and the data into an accurate estimate.

Chap. (4) then introduces a new technique for improving the observational efficiency of existing estimation algorithms. The method has its roots in the theory of attractor reconstruction in nonlinear dynamics, as well as inherent connections to 4DVar methods from data assimilation, and moving horizon estimation from optimal control. Relevant examples will be given to demonstrate that the method is capable of effectively reducing the observational requirements of the problem, albeit at increased computational cost.

Finally, Chap. (5) summarizes the main results, and motivates directions for future research.

# 2 The canonical structure of optimal estimation

The history outlined in Chap. (1) highlights the ties between estimation and related ideas from statistics, optimization, and physics. This chapter further investigates these connections, by revisiting the classical theory of Kalman filtering and smoothing from a deterministic point of view. The solutions to this problem are based (in one way or another) on Newton's method (or variants thereof), and are all equivalent for linear problems. However, the same cannot be said of their nonlinear extensions, and this chapter begins the process of documenting the rather subtle differences between these approaches.

Connections to physics are also explored, with the goal of reconciling some of the shared ideas between estimation and Hamiltonian mechanics. While both fields have shared roots in the classical theory of calculus of variations, physics has not adopted many of the 'modern' extensions developed for the theory of optimal control and estimation. On the other hand, certain foundational concepts from physics are notably absent from estimation — such as the notions of symplectic structure and canonical transformations. The estimation analogs of these ideas will be discussed.

This chapter offers a new perspective on a longstanding problem, which brings together shared ideas from estimation, optimization, and physics. While still largely incomplete, the intention is that this will lead to a broader understanding of the inherent connections between these fields, and serve as a source of innovation and creativity for the development of the next generation of estimation algorithms.

## 2.1 The Hamiltonian structure of fixed-interval smoothing

As discussed in Chap. (1), the estimation of a trajectory $\mathcal{X}$ given a set of measurements $\mathcal{Y}$ may be viewed as the statistical evaluation of expected values of an arbitrary function $\boldsymbol{g}(\mathcal{X})$ of the path. These

averages are taken with respect to the unnormalized conditional distribution $\rho(\mathcal{X}|\mathcal{Y})$ and may be evaluated as a statistical path integral

$$\langle \boldsymbol{g}(\mathcal{X})\rangle = \int \delta\mathcal{X}\, \boldsymbol{g}(\mathcal{X}) \exp[-A(\mathcal{X})], \tag{2.1}$$

where the integral is taken with respect to 'all possible' paths, and

$$A(\mathcal{X}) \propto -\log[\rho(\mathcal{X}|\mathcal{Y})]$$

is the *estimation action* or conditional log-likelihood of the path given the observations. Although its value depends on the observations $\mathcal{Y}$, these are treated as fixed within a particular instance of the problem. So the action is written as $A(\mathcal{X})$, suppressing the explicit dependence on $\mathcal{Y}$. Only estimates of the path are considered here, so $\boldsymbol{g}(\mathcal{X}) = \mathcal{X}$, and Laplace's approximation further reduces the problem to one of finding minimizing paths of the action. Of these paths, only the global minimizers are of interest, which is equivalent to choosing the maximum *a posteriori* (MAP) estimate, or conditional mode of the distribution. With these assumptions, it becomes an optimization problem $\min_{\mathcal{X}} A(\mathcal{X},\mathcal{Y})$. Specifying the noise in the measurements and the model to be Gaussian and temporally uncorrelated the action can be written (in continuous time) as the time integral of a Lagrangian,

$$\begin{aligned} A(\mathcal{X}) &\propto \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \int_0^T dt\, \mathcal{H}_t(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_t) \\ \mathcal{H}_t(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_t) &:= \frac{1}{2}|\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_t)|^2_{\boldsymbol{R}_m} + \frac{1}{2}|\dot{\boldsymbol{x}}_t - \boldsymbol{f}(\boldsymbol{x}_t)|^2_{\boldsymbol{R}_f}. \end{aligned} \tag{2.2}$$

where the velocity $\dot{\boldsymbol{x}}_t := d\boldsymbol{x}_t/dt$ is treated as an independent variable in the minimization. The matrices $\boldsymbol{R}_0,\ \boldsymbol{R}_m,\ \boldsymbol{R}_f$ are respectively the (inverse) error covariance matrices of the prior, the measurements and the model, assumed here to be time-independent and positive definite. The problem can thus be treated deterministically using the techniques of calculus of variations underlying classical physics and optimal control. The latter introduces a control variable $\dot{\boldsymbol{x}}_t \to \boldsymbol{f}(\boldsymbol{x}_t) + \boldsymbol{u}_t$

$$\mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{u}_t) := \frac{1}{2}|\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_t)|^2_{\boldsymbol{R}_m} + \frac{1}{2}|\boldsymbol{u}_t|^2_{\boldsymbol{R}_f}.$$

and can be written in Hamiltonian form by incorporating the dynamical constraints with Lagrange multipliers or 'co-states' $\boldsymbol{p}_t$ into an *augmented action*

$$\begin{aligned} A(\mathcal{X},\mathcal{U},\mathcal{P}) &\propto \frac{1}{2}|\boldsymbol{x}_0-\boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0}+\int_0^T dt\,\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{u}_t)+\langle \boldsymbol{p}_t,\dot{\boldsymbol{x}}_t-\boldsymbol{f}(\boldsymbol{x}_t)-\boldsymbol{u}_t\rangle \\ &= \frac{1}{2}|\boldsymbol{x}_0-\boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0}+\int_0^T dt\,\langle \boldsymbol{p}_t,\dot{\boldsymbol{x}}_t\rangle-\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{u}_t,\boldsymbol{p}_t) \end{aligned}$$

where

$$\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{u}_t,\boldsymbol{p}_t) := \langle \boldsymbol{p}_t,\boldsymbol{f}(\boldsymbol{x}_t)+\boldsymbol{u}_t\rangle-\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{u}_t). \tag{2.3}$$

The necessary conditions for optimality are given by considering the variations $\delta A=0$. This produces Hamilton's equations $\dot{\boldsymbol{x}}_t=\nabla_p\mathcal{H}_t,\ -\dot{\boldsymbol{p}}_t=\nabla_x\mathcal{H}_t$ along with an auxiliary equation for the control $\nabla_u\mathcal{H}_t=0$, which are derived in the standard fashion by taking variations with respect to $\boldsymbol{x}_t,\ \boldsymbol{p}_t,\ \boldsymbol{u}_t$ equal to zero and performing integration by parts on $\delta\dot{\boldsymbol{x}}_t$. Also, the equation $\nabla_u\mathcal{H}_t=0$ gives $\boldsymbol{u}_t=\boldsymbol{R}_f\cdot\boldsymbol{p}_t$, which here may be inverted to eliminate the control $\boldsymbol{u}_t$ in terms of the Lagrange multiplier

$$\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{p}_t) := \frac{1}{2}|\boldsymbol{p}_t|^2_{\boldsymbol{R}_f^{-1}}+\langle \boldsymbol{p}_t,\boldsymbol{f}(\boldsymbol{x}_t)\rangle-\frac{1}{2}|\boldsymbol{y}_t-\boldsymbol{h}(\boldsymbol{x}_t)|^2_{\boldsymbol{R}_m}. \tag{2.4}$$

The same result can be obtained from the standard approach in classical Hamiltonian mechanics, which defines the *canonical momentum* $\boldsymbol{p}_t=\nabla_{\dot{x}}\mathcal{H}_t$ and the Hamiltonian via the Legendre transform $\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{p}_t)=\langle \boldsymbol{p},\dot{\boldsymbol{x}}_t\rangle-\mathcal{H}_t(\boldsymbol{x}_t,\dot{\boldsymbol{x}}_t)$. However, it is worth pointing out the rather important (yet often tacit) assumption made in Hamiltonian mechanics regarding the invertibility of the Legendre transform. In order to use the Hamiltonian formalism, the velocity dependence must be eliminated by solving $\boldsymbol{p}_t=\nabla_{\dot{x}}\mathcal{H}_t$ to express $\dot{\boldsymbol{x}}_t$ explicitly in terms of $\boldsymbol{p}_t$. This assumption does not always hold however. For instance, in optimal control the model is often constrained to actuate on a subset of its total degrees of freedom, making the Legendre transform is singular. It was with this in mind that Pontryagin developed his maximum principle as generalization of Hamiltonian mechanics that allows for non-invertibility of the canonical momentum, by writing the Hamiltonian in terms of $\mathcal{H}(\boldsymbol{x}_t,\boldsymbol{p}_t,\boldsymbol{u}_t)$. And while the physics approach can be salvaged using the aforementioned extended coordinate description $\mathcal{H}(\boldsymbol{x}_t,\boldsymbol{p}_t,\boldsymbol{v}_t)$, this is not well-known. But here, since the state of the estimate is not subject to any physical constraints, the assumptions of additive model error together with the invertibility of $\boldsymbol{R}_f$ allow the Hamiltonian to be treated simply as $\mathcal{H}(\boldsymbol{x}_t,\boldsymbol{p}_t)$.

Thus, while it is possible to treat estimation as a classical mechanics problem, it is important to keep in mind however one fundamental difference: estimation is not an initial value problem. Unlike

mechanics problems, where one knows both the initial position and initial momentum, in estimation both quantities are not known at the same time. In other words Hamilton's equations must be solved with with separated boundary conditions

$$\begin{aligned}
\dot{\boldsymbol{x}}_t &= \nabla_p \mathcal{H}_t = \boldsymbol{f}(\boldsymbol{x}_t) + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_t \\
-\dot{\boldsymbol{p}}_t &= \nabla_x \mathcal{H}_t = \nabla^\dagger \boldsymbol{f} \cdot \boldsymbol{p}_t + \nabla^\dagger \boldsymbol{h} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_t)) \\
\boldsymbol{p}_0 &= \boldsymbol{R}_0 \cdot (\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\
\boldsymbol{p}_T &= \boldsymbol{0}.
\end{aligned} \tag{2.5}$$

It is therefore a *two-point boundary value problem*, which by construction lacks the local uniqueness properties inherent to initial value problems. Sometimes it admits many solution, and other times none. In this case, it appears to be rather difficult to construct explicit solutions by direct integration of Hamilton's equations—the first order necessary conditions for optimality. The reason for this will become more clear in Sec. (2.2). But suffice to say, these solutions are inherently unstable.

### 2.1.1 Canonical momentum as a Lagrange multiplier

The control formalism provides an additional interpretation of the rather basic role of the canonical momentum — it is in fact, also a Lagrange multiplier. While this interpretation of momentum as a Lagrange multiplier is well-known in optimization with the Karush-Kuhn-Tucker conditions as a generalization to Hamilton's equations, in physics it is far less familiar ([207]). The closest idea perhaps comes from Lagrangian mechanics, where the Lagrange multipliers are used explicitly to incorporate a generalized forces of constraint, but it is not often discussed in the Hamiltonian context despite some interesting ramifications regarding *secondary constraints* [46]. It is apparent though that our conceptual understanding of momentum must be expanded to reconcile these two fundamental points of view.

Its role as a Lagrange multiplier can be shown directly by considering the augmented action

$$A(\mathcal{X}, \mathcal{V}, \mathcal{P}) \propto \int_0^T dt\, \mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{v}_t) + \langle \boldsymbol{p}_t, \dot{\boldsymbol{x}}_t - \boldsymbol{v}_t \rangle = \int_0^T dt\, \langle \boldsymbol{p}_t, \dot{\boldsymbol{x}}_t \rangle - \mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{p}_t, \boldsymbol{v}_t).$$

Variations with respect to $\boldsymbol{v}_t$ provides $\boldsymbol{p}_t = \nabla_v H_t$, and with respect to $\boldsymbol{p}_t$ simply gives $\dot{\boldsymbol{x}}_t = \boldsymbol{v}_t$. The view suggests an interpretation of momentum as the cost associated with violating $\dot{\boldsymbol{x}}_t = \boldsymbol{v}_t$, although the physical meaning of this is unclear. It is also evident that its choice of sign is just a matter of convention. Hamilton's equations are invariant to defining $\boldsymbol{p}_t$ as positive or negative, although physically the direction of time

implies a natural choice. There is however, also the choice of sign in the definition of the Hamiltonian, as is well known in optimal control where there are two competing conventions $\mathcal{H}_t = \langle \boldsymbol{p}_t, \dot{\boldsymbol{x}}_t \rangle \pm \mathcal{H}_t$ leading to both maximum and minimum principles. The '$-$' convention maximizes the Hamiltonian and makes the Legendre transform an involution, whereas the '$+$' convention maintains the interpretation of minimizing a cost function. The arbitrary nature of this choice seems to imply an inherent parity to the concept of momentum that is reflected as a type of gauge symmetry in its equations. For the record, '$-$' convention is used here to make the connections to classical mechanics as explicit and unambiguous as possible.

### 2.1.2 Discrete time

In practice, measurements are not continuous, and the time evolution of the forecast model is typically computed in discrete time steps. Let $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n)$ be the discrete time update of the model, which in general may be implicit although this will not be considered here. Let $\mathcal{X}^{(d)} := \{x_n\}|_{n=0}^{N}$ be the discrete time path. The action integral in Eqn. (2.2) is typically interpreted as a sum

$$\begin{aligned} A(\mathcal{X}^{(d)}) &\propto \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \frac{1}{2}|\boldsymbol{y}_N - \boldsymbol{h}(\boldsymbol{x}_N)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}) \\ \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}) &:= \frac{1}{2}|\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \frac{1}{2}|\boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_f}. \end{aligned} \tag{2.6}$$

The discrete time description is more precise than its continuous counterpart. It arises directly from the recursive application of Bayes' rule, and from a stochastic point of view the time derivative $\dot{\boldsymbol{x}}_t$ in Eqn. (2.2) is not well-defined as the paths have different forwards and backwards derivatives. The continuous description should thus be viewed as an approximation of the discrete ([93]). While the Duncan-Mortensen-Zakai equation ([215]) provides a rigorous mathematical framework for continuous time filtering, its connection to the underlying variational principle (via the Laplace approximation of the path integral) are still not well-established.

Yet this is not an issue from a deterministic point of view, in which the model and measurement noise processes are not stochastic but evolve smoothly in time. This perspective suggests treating the connection between the continuous and discrete time action principles using the techniques of discrete mechanics ([119]). This typically proceeds by first discretizing the continuous time action functional (*e.g.*, by using a quadrature rule to form a discrete Lagrangian of the form $\mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})$) and applying the discrete time version of Lagrange's equations $\nabla_{x_n} \mathcal{H}_n = \nabla_{x_{n+1}} \mathcal{H}_n = 0$.

However, the inclusion of the final measurement $\frac{1}{2}|\boldsymbol{y}_N - \boldsymbol{h}(\boldsymbol{x}_N)|^2_{\boldsymbol{R}_m}$ in Eqn. (2.6) makes Eqn. (2.6)

inconsistent with a direct discretization of the continuous action integral. But if this term is omitted (and it often is), the integral may be interpreted as a left Riemann sum discretization. This choice of discretization appears to be tied to the choice of stochastic integration convention, with the left Riemann sum providing Ito's interpretation, which is most commonly adopted in stochastic filtering. Nonetheless, if these more subtle stochastic issues are ignored, much of the machinery developed for discrete mechanics is applicable here.

There is also some ambiguity regarding the measurement term. Since the discrete Lagrangian only needs to be a function of $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$, it is unclear whether the observation should be included as a function of $\boldsymbol{x}_n$ or $\boldsymbol{x}_{n+1}$. But this choice evidently leads to different equations of motion. In this case, the more consistent approach is given by the optimal control formulation, where the Lagrangian is a function of $\boldsymbol{x}_n$ and the control $\boldsymbol{u}_n$. Specifically, it can be written as

$$\begin{aligned} A(\mathcal{X}^{(d)}, \mathcal{U}^{(d)}) &\propto \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \frac{1}{2}|\boldsymbol{y}_N - \boldsymbol{h}(\boldsymbol{x}_N)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n) \\ \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n) &:= \frac{1}{2}|\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \frac{1}{2}|\boldsymbol{u}_n|^2_{\boldsymbol{R}_f} \end{aligned} \tag{2.7}$$

and subject to the constraints $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n$. Substituting the constraint for $\boldsymbol{u}_n$ gives back Eqn. (2.6), but this form leaves no ambiguity that the observation term is imposed at time $t_n$. The implication of this choice on the stochastic interpretation of the problem is not immediately clear however.

As in the continuous case, the constraints can be included by introducing a Lagrange multiplier

$$\begin{aligned} A(\mathcal{X}^{(d)}, \mathcal{U}^{(d)}, \mathcal{P}^{(d)}) \propto \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \frac{1}{2}|\boldsymbol{y}_N - \boldsymbol{h}(\boldsymbol{x}_N)|^2_{\boldsymbol{R}_m} + \\ \sum_{n=0}^{N-1} \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n) + \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n) - \boldsymbol{u}_n \rangle \end{aligned}$$

and written as a control Hamiltonian

$$\begin{aligned} A(\mathcal{X}^{(d)}, \mathcal{U}^{(d)}, \mathcal{P}^{(d)}) &= \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \frac{1}{2}|\boldsymbol{y}_N - \boldsymbol{h}(\boldsymbol{x}_N)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n, \boldsymbol{p}_{n+1}) \\ \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n, \boldsymbol{p}_{n+1}) &:= \langle \boldsymbol{p}_{n+1}, \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{u}_n). \end{aligned}$$

The variation $\nabla_{\boldsymbol{u}_n} A = 0$ provides $\boldsymbol{u}_n = \boldsymbol{R}_f \cdot \boldsymbol{p}_{n+1}$, which can be inverted to eliminate $\boldsymbol{u}_n$ in terms of $\boldsymbol{p}_{n+1}$ to give

$$\mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{p}_{n+1}) := \frac{1}{2}|\boldsymbol{p}_{n+1}|^2_{\boldsymbol{R}_f^{-1}} + \langle \boldsymbol{p}_{n+1}, \boldsymbol{F}(\boldsymbol{x}_n) \rangle - \frac{1}{2}|\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m}. \tag{2.8}$$

This Hamiltonian can also be derived directly from the discrete Lagrangian Eqn. (2.6), using discrete Hamiltonian mechanics ([110]), provided the measurement term is evaluated at $\boldsymbol{x}_n$ and not $\boldsymbol{x}_{n+1}$. Also, increasing the sum index by one incorporates the measurement term at the final time due to the terminal boundary condition $\boldsymbol{p}_{N+1} = 0$. This allows the action to be written more succinctly as

$$A(\mathcal{X}^{(d)}, \mathcal{P}^{(d)}) \propto \frac{1}{2}|\boldsymbol{x}_0 - \boldsymbol{\mu}_0|^2_{\boldsymbol{R}_0} + \sum_{n=0}^{N} \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{p}_{n+1}). \tag{2.9}$$

The discrete Hamilton's equations are then given by considering the variations $\nabla_{p_{n+1}} A = \nabla_{x_n} A = 0$ and reindexing the sum — the discrete analog for integration by parts (note there is no sign change in discrete time). This leads to the discrete two-point boundary value problem

$$\begin{aligned}
\boldsymbol{x}_{n+1} &= \nabla_p^\dagger \mathcal{H}_n = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_{n+1} \\
\boldsymbol{p}_n &= \nabla_x^\dagger \mathcal{H}_n = \nabla^\dagger \boldsymbol{F}_n \cdot \boldsymbol{p}_{n+1} + \nabla^\dagger \boldsymbol{h}_n \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)) \\
\boldsymbol{p}_1 &= \boldsymbol{R}_0 \cdot (\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\
\boldsymbol{p}_{N+1} &= 0.
\end{aligned} \tag{2.10}$$

Note that these equations only hold when the dynamical model is explicit. For implicit models $\boldsymbol{F}(\boldsymbol{x}_n) \to \boldsymbol{F}(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})$, Hamilton's equations are also implicit and therefore more complicated. It is also important for the momentum to be specified at $t_{n+1}$ rather than $t_n$, otherwise the equations will be incorrect. However in discrete Hamiltonian mechanics there is an alternative formulation that instead uses $\boldsymbol{p}_n$ and $\boldsymbol{x}_{n+1}$ ([110]), and produces a state update that goes *backwards* in time. The implications of this alternate choice on optimal control and estimation has to my knowledge not yet been investigated.

### 2.1.3 The search for minimizing paths

With the Laplace approximation to the statistical path integral Eqn. (2.1) the goal is to find the paths that locally minimize the action. Of these paths, only the lowest minimizers are of interest, as they exponentially dominate the contribution to the path integral. The difficulty then is finding these paths in a stable and efficient manner. Given the vast generality of the problem, an enormous number of techniques are available for this purpose, each with its own computational trade-offs. However, as exact solutions exist only for a few special cases, the overwhelming majority of these methods are iterative in the sense that given an approximate solution $\mathcal{X}^{(i)}$ at iteration $i$, they produce a more refined solution $\mathcal{X}^{(i+1)}$, and the process proceeds until some measure of convergence is met.

At its core, the problem is simply one of numerical optimization. Thus, the most straightforward approach is perhaps to take a discretized approximation of the optimal path $\mathcal{X}^{(d)}$ and directly minimize the action functional Eqn. (2.6), say using Newton's method (or a variation thereof)

$$\mathcal{X}^{(d,i+1)} - \mathcal{X}^{(d,i)} = -[\nabla^2_{\mathcal{X}\mathcal{X}} A(\mathcal{X}^{(d,i)})]^{-1} \cdot \nabla_{\mathcal{X}} A(\mathcal{X}^{(d,i)}). \tag{2.11}$$

Techniques such as this, which optimize a discretized version of the full path trajectory by constructing a series of estimates $A(\mathcal{X}^{(d,1)}) < A(\mathcal{X}^{(d,2)}) < \ldots < A(\mathcal{X}^{(d,\infty)})$, fall under the broad category of 'direct' methods in optimal control and estimation. They are considered typically part of the larger family of multiple shooting and collocation methods for solving two-point boundary value problems, although as discussed by [20], such methods can also be indirect. Indirect methods by contrast, attempt to locate solutions that satisfy the necessary conditions for optimality by restricting the feasible set of solutions to those satisfying the variational principle $\delta A = 0$. For instance, enforcing the constraint $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n$ in Eqn. (2.7) reduces the search space from $\{\mathcal{X}^{(d)}, \mathcal{U}^{(d)}\}$ to $\{\boldsymbol{x}_0, \mathcal{U}^{(d)}\}$. So one only needs to provide a guess for the initial state $\boldsymbol{x}_0$ and an initial or 'nominal' control law $\mathcal{U}^{(d)}$. In other words, indirect methods typically require the solution to be a valid trajectory ate each iteration, while for direct methods the solution is only approximate and converges to a valid path as $i \to \infty$.

Each approach comes with its own trade-offs. Indirect methods tend to be more accurate given that they can produce exact trajectories of the model, while direct methods typically require implicit integration methods, which inherently limit their accuracy. Whether this is an issue in practice however depends on how accurately the model equations need to be enforced. On the other hand, indirect methods typically either need to be initialized close to the desired solution or require a good *a-priori* guess for the control sequence $\mathcal{U}^{(d)}$. This stems from the fact that its solutions are always a 'true' trajectory of the system $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n$ and thus more susceptible to becoming trapped in local minima. Direct methods by contrast tend to have a considerably large basin of attraction around its global minima, where it is in fact often better to start with a random path, rather than an exact trajectory [174]. The stability and size of this solution basin is closely tied to the overall conditioning of the problem, as Hamilton's equations for this system can be highly ill-conditioned. A numerical examination of this will be given in Chap. (3).

In terms of computational efficiency, since the direct Lagrangian formulation in Eqn. (2.6) is a nonlinear least-squares problem it can be written as $A(\mathcal{X}^{(d,i)}) = |\boldsymbol{\psi}(\mathcal{X}^{(d,i)})|^2$ and minimized with the

Gauss-Newton method

$$\mathcal{X}^{(d,i+1)} - \mathcal{X}^{(d,i)} = -[\nabla_{\mathcal{X}}\psi(\mathcal{X}^{(d,i)}) \cdot \nabla_{\mathcal{X}}\psi(\mathcal{X}^{(d,i)})]^{-1} \cdot \nabla_{\mathcal{X}}\psi(\mathcal{X}^{(d,i)}).$$

This is essentially the same as Newton's method Eqn. (2.11), but uses an approximate Hessian that ignores the terms $\nabla^2_{\mathcal{X}\mathcal{X}}\psi(\mathcal{X}^{(d,i)})$ to avoid computing $D^3$ derivative tensors. If this matrix were dense, it would require $O(N^2D^2)$ storage, which would quickly get out of hand. However, the fact that $\boldsymbol{x}_n$ depends only on $\boldsymbol{x}_{n-1}$ and $\boldsymbol{x}_{n+1}$ gives this matrix block tri-diagonal structure, which requires only $O(ND^2)$ storage and can be exploited to reduce the time required to construct its inverse from $O(N^3D^3)$ to $O(ND^3)$. An example of this will be shown below. However, even this can be to much for large systems, such as ones used for numerical weather prediction, which can require upwards of $O(10^9)$ degrees of freedom. Memory scaling is crucial for these applications, and implicit methods are are efficient in this regard, utilizing adjoint models to compute the descent directions $-\nabla_{\mathcal{X}}\psi(\mathcal{X}^{(d,i)})$ in $O(ND)$. Direct methods by contrast do not require adjoint equations, so they are easier to implement, although recent advances in automatic differentiation have mitigated this to some extent. Also, convergence of derivative free and gradient descent methods tends to slower than those using an approximate Hessian.

These techniques are also not mutually exclusive. Direct and indirect approaches may be blended into hybrid methods by expanding the search space to include $\mathcal{U}^{(d)}$ and $\mathcal{P}^{(d)}$ as independent variables. For instance, one can directly optimize Eqn. (2.7) in $\{\mathcal{X}^{(d)}, \mathcal{U}^{(d)}\}$ coordinates, provided the constraint $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n$ is satisfied at the solution. Enforcing these constraints requires the introduction of Lagrange multipliers, which can either be treated explicitly, by performing the minimization directly in $\{\mathcal{X}^{(d)}, \mathcal{U}^{(d)}, \mathcal{P}^{(d)}\}$ space, or implicitly by eliminating one or the other using the variational equations $\delta A = 0$. A subset of these choices were explored numerically by [85], and showed limited benefit to performing the optimization with constraints in $\{\mathcal{X}^{(d)}, \mathcal{P}^{(d)}\}$ space versus the unconstrained approach in $\mathcal{X}^{(d)}$ space, especially given the added computational cost of evaluating the generally nonlinear constraints. Part of the issue is related to the fact that many constrained nonlinear optimization algorithms operate by first finding a feasible point satisfying the constraints and then minimizing the objective function in a way that maintains feasibility of the solution. However, in this case the constraints are only intended to hold in the neighborhood of the solution, so enforcing feasibility too early often causes the routine to become trapped in local minima with considerably higher action levels than those identified by the unconstrained approach.

This concludes the exposition of the basic structure of the estimation problem and its connections

to the variational principles of optimal control and classical mechanics. It is apparent that given the above assumptions the estimation problem strongly resembles one of classical mechanics, with the main difference being that estimation has separated boundary conditions. This turns out to be extremely important, as simply integrating Hamilton's equations forward in time from some approximate initial guess for $\boldsymbol{z}_0 := \{\boldsymbol{x}_0,\, \boldsymbol{p}_0\}$ produces solutions that become rapidly unbounded. Yet when the models $\boldsymbol{f}(\cdot)$ and $\boldsymbol{h}(\cdot)$ are linear, the Kalman equations provide a forward stable solution, at the cost of increasing the state representation from $2D$ canonical coordinates to $D\,(1+D)$ variables. The question then is what makes these solutions inherently stable given that Hamilton's equations are inherently not? To examine this question, it seems appropriate to start by reexamining the solutions to the linear problem.

## 2.2 Optimal solutions for linear models

Solutions to the linear problem were first given in continuous time by [29]. They showed that the optimal path could be could be computed as a linear combination of two solutions, one integrated forward in time and the other backward. This approach was originally developed for optimal control problems, and goes by various names including 'sweep methods', 'two-pass filters' and the 'method of successive approximations'. While there are an infinite number of ways to construct these solutions, one of which simply involves integrating Hamilton's equations forwards and backwards in time, the forward pass is most appropriately given by the Kalman Bucy filter ([89]). This is due to its inherent stability properties that stem from its formulation as the optimal feedback control. Their backward pass solution however, was unstable and inefficient, and thus was never used in any practical applications [123]. This led to the proposal of several solutions over the next decade or so, with the most notable given by Rauch Tung and Striebel ([164]), Mayne ([121]) and Fraser ([57, 56]).

All of these solutions in one way or another reference the stochastic version of the problem. As a main goal here is to derive these solutions from a purely deterministic perspective the original solution by Bryson and Frazier will now be reexamined from this point of view. While it will not be possible to completely dispense with the underlying stochastic interpretation, the only distinction between the two approaches arises in the two-point boundary value problem for the covariance. Namely, when the underlying system is stochastic, additional inhomogeneous forcing terms in these equations arise from its diffusive properties.

Furthermore, as the stability of the solutions is also of particular interest here, it will be explicitly shown how the necessary conditions for optimality are not suitable for constructing *explicit* solutions, as

Hamilton's equations are inherently unstable. Furthermore the Kalman filter, which satisfies the necessary and sufficient conditions given by the Hamilton-Jacobi(-Bellman) equations, provides optimal stability. To fully appreciate these results it is useful to have an illustrative example, which will now be described.

### 2.2.1 The simplest tracking problem

Consider the filter as an observer, with no model or measurement error, and make a change to 'error' coordinates $\boldsymbol{q}_t := \boldsymbol{x}_t^* - \boldsymbol{x}_t$ where $\boldsymbol{x}_t^*$ is the state of the 'true' system satisfying $\dot{\boldsymbol{x}}_t^* = \boldsymbol{F}^* \cdot \boldsymbol{x}_t^*$, $\boldsymbol{y}_t = \boldsymbol{H}^* \cdot \boldsymbol{x}_t^*$. Assuming also that the models are perfect $\boldsymbol{F} = \boldsymbol{F}^*$ and $\boldsymbol{H} = \boldsymbol{H}^*$, the action Eqn. (2.2) becomes the linear quadratic regulator

$$A \propto \frac{1}{2}|\boldsymbol{q}_0|^2_{\boldsymbol{R}_0} + \frac{1}{2}\int_0^T dt\, |\boldsymbol{H}\cdot\boldsymbol{q}_t|^2_{\boldsymbol{R}_m} + |\dot{\boldsymbol{q}}_t - \boldsymbol{F}\cdot\boldsymbol{q}_t|^2_{\boldsymbol{R}_f},$$

which simplifies the problem by removing the time-dependent measurements $\boldsymbol{y}_t$. Note that for the estimation to be successful, the optimal solution needs to produce a contraction on the error, so that $\boldsymbol{q}_t \to 0$ as $t \to \infty$.

Consider now the problem of tracking the position of a particle moving along a constant linear trajectory with no prior information, so $\boldsymbol{R}_0 = \boldsymbol{F} = 0$ and $\boldsymbol{H} = \boldsymbol{I}$. The latter assumption is needed to make the system observable, as the model does not share any information in the directions transverse to the equations of motion. Without loss of generality the problem may be reduced to $D = 1$. Given these assumptions, Hamilton's equations Eqn. (2.4) become

$$\begin{bmatrix} \dot{q}_t \\ \dot{p}_t \end{bmatrix} = \begin{bmatrix} 0 & R_f^{-1} \\ R_m & 0 \end{bmatrix} \cdot \begin{bmatrix} q_t \\ p_t \end{bmatrix}. \tag{2.12}$$

Or in terms of second derivatives

$$\ddot{q}_t = R_f^{-1} R_m = \omega^2 q_t$$
$$\ddot{p}_t = R_m R_f^{-1} = \omega^2 p_t$$

where $\omega := \sqrt{R_m R_f^{-1}}$. These equations have exponential solutions

$$q_t = \omega\sqrt{R_f^{-1}}\left(c^{(+)}e^{+\omega t} + c^{(-)}e^{-\omega t}\right)$$
$$p_t = \omega\sqrt{R_m}\left(c^{(+)}e^{+\omega t} - c^{(-)}e^{-\omega t}\right)$$

with constants $c^{(\pm)}$ to be determined based on the boundary conditions.

The solution $q_t = c^{(\pm)} = 0$ is clearly trivial. But attempting to solve it as a two-point boundary value problem — *e.g.*, using a 'shooting approach' that determines $c^{(\pm)}$ by integrating Hamilton's equation's forward from $0 \to T$ from approximate initial conditions — immediately runs into issues. While the $c^{(-)}\,e^{-\omega t}$ term is forward stable and will approach $q_t \to 0$ as $t \to \infty$, the $c^{(+)}\,e^{+\omega t}$ term quickly becomes unbounded making the problem of determining the constants $c^{(\pm)}$ highly ill-conditioned. While this issue is often avoided in practice by breaking up the trajectories into small segments (*i.e.*, using a 'multiple shooting approach'), connecting them requires running a numerical optimization algorithm that considers the entire path (*i.e.*, all time segments) simultaneously, which becomes increasingly difficult as the estimation window grows.

On the other hand, the optimal two-pass solution avoids these issues altogether by transforming Hamilton's equations to an 'equivalent' form that preserves its underlying symplectic structure but is inherently stable. For the simple example described here, this amounts to splitting the solution into forward and backward stable modes. This feature will be shown explicitly in a moment, after discussing the general solution.

### 2.2.2 Continuous time solutions

The results of [29] will now be reexamined, paying close attention to the stability of the solutions and the distinguishing features between the stochastic and deterministic interpretations of the problem. Consider the linear estimation system described by the deterministic models $\dot{\boldsymbol{x}}_t = \boldsymbol{F}_t \cdot \boldsymbol{x}_t$, $\boldsymbol{y}_t = \boldsymbol{H}_t \cdot \boldsymbol{x}_t$ without stochastic disturbances. From Eqn. (2.4), Hamilton's equations are

$$\underbrace{\begin{bmatrix}\dot{\boldsymbol{x}}_t \\ \dot{\boldsymbol{p}}_t\end{bmatrix}}_{\dot{\boldsymbol{z}}_t} = \underbrace{\begin{bmatrix}\boldsymbol{0} & +\boldsymbol{I} \\ -\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}}_{\boldsymbol{J}} \cdot \underbrace{\begin{bmatrix}-\boldsymbol{H}_t^{\dagger}\cdot\boldsymbol{R}_m\cdot\boldsymbol{H}_t & \boldsymbol{F}_t^{\dagger} \\ \boldsymbol{F}_t & \boldsymbol{R}_f^{-1}\end{bmatrix}}_{\nabla^2_{zz}\mathcal{H}_t} \cdot \underbrace{\begin{bmatrix}\boldsymbol{x}_t \\ \boldsymbol{p}_t\end{bmatrix}}_{\boldsymbol{z}_t} + \underbrace{\begin{bmatrix}\boldsymbol{0} \\ -\boldsymbol{H}^{\dagger}\cdot\boldsymbol{R}_m\cdot\boldsymbol{y}_t\end{bmatrix}}_{\boldsymbol{\xi}_t} \tag{2.13}$$

with boundary conditions $\boldsymbol{p}_0 = \boldsymbol{R}_0 \cdot (\boldsymbol{x}_0 - \boldsymbol{\mu}_0)$, $\boldsymbol{p}_T = 0$. With the above definitions, these equations may also be written more concisely as

$$\dot{\boldsymbol{z}}_t = \nabla_z \mathcal{H}_t = \boldsymbol{J} \cdot \nabla^2_{zz}\mathcal{H}_t \cdot \boldsymbol{z}_t + \boldsymbol{\xi}_t$$

where $\boldsymbol{z}_t := \{\boldsymbol{x}_t, \boldsymbol{p}_t\}$. The differential equation is linear so its general solution can be written as a superposition of homogeneous and particular solutions $\boldsymbol{z}_t = \boldsymbol{z}_t^{(h)} + \boldsymbol{z}_t^{(p)}$. Setting $\boldsymbol{\xi}_t \to 0$, the homogeneous solution can be determined in a few ways. For instance, introducing a $2D \times 2D$ transition matrix $\boldsymbol{\Phi}_{t|t'}$ such that

$$\boldsymbol{z}_t^{(h)} = \boldsymbol{\Phi}_{t|t'} \cdot \boldsymbol{z}_{t'}^{(h)} = \begin{bmatrix} \boldsymbol{\Phi}_{t|t'}^{(xx)} & \boldsymbol{\Phi}_{t|t'}^{(xp)} \\ \boldsymbol{\Phi}_{t|t'}^{(px)} & \boldsymbol{\Phi}_{t|t'}^{(pp)} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{t'}^{(h)} \\ \boldsymbol{p}_{t'}^{(h)} \end{bmatrix}.$$

Its time evolution is therefore

$$\frac{d}{dt}\left(\boldsymbol{\Phi}_{t|t'} \cdot \boldsymbol{z}_{t'}^{(h)}\right) - \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}_t \cdot \boldsymbol{\Phi}_{t|t'} \cdot \boldsymbol{z}_{t'}^{(h)} = 0$$
$$\dot{\boldsymbol{\Phi}}_{t|t'} = \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}_t \cdot \boldsymbol{\Phi}_{t|t'}.$$

It can also be used as an integrating factor. Left-multiplying by the its inverse $\boldsymbol{\Phi}_{t|t'}^{-1} \equiv \boldsymbol{\Phi}_{t|t'}$ gives

$$\boldsymbol{\Phi}_{t'|t} \cdot (\dot{\boldsymbol{z}}_t^{(h)} - \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}_t \cdot \boldsymbol{z}_t^{(h)}) = \frac{d}{dt}\left(\boldsymbol{\Phi}_{t'|t} \cdot \boldsymbol{z}_t^{(h)}\right) = 0$$
$$-\dot{\boldsymbol{\Phi}}_{t'|t} = \boldsymbol{\Phi}_{t'|t} \cdot \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}_t$$

with

$$\boldsymbol{\Phi}_{t|t} = \boldsymbol{\Phi}_{t'|t} \cdot \boldsymbol{\Phi}_{t|t'} = \boldsymbol{I}.$$

These matrices are the fundamental solutions to the homogeneous differential equation and respectively describe its propagation forward and backward in time. They are related to the Green's function, allowing the inhomogeneous solution $\boldsymbol{z}_t^{(p)}$ can be written as

$$\frac{d}{dt}\left(\boldsymbol{\Phi}_{t'|t} \cdot \boldsymbol{z}_t^{(h)}\right) = \boldsymbol{\xi}_t \quad \Longrightarrow \quad \boldsymbol{z}_t^{(p)} = \boldsymbol{\Phi}_{t|t'} \cdot \boldsymbol{z}_{t'}^{(p)} + \int_{t'}^{t} ds\, \boldsymbol{\Phi}_{t|s} \cdot \boldsymbol{\xi}_s.$$

Alternatively, [29] introduces a $2D \times D$ matrix $\boldsymbol{\Psi}_t$, such that

$$\boldsymbol{\Psi}_t := \begin{bmatrix} \boldsymbol{\Psi}_t^{(x)} \\ \boldsymbol{\Psi}_t^{(p)} \end{bmatrix} \qquad \dot{\boldsymbol{\Psi}}_t = \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}_t \cdot \boldsymbol{\Psi}_t \qquad \boldsymbol{\Psi}_0 = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{I} \end{bmatrix}$$

Note that $\boldsymbol{\Psi}_t$ reduces the total number of degrees of freedom in half from $4D^2$ to $2D^2$, and has different algebraic properties than $\boldsymbol{\Phi}_{t|t'}$ since it is not a square matrix. It plays an important role in the optimal stability of the Kalman solution.

The fact that the boundary conditions have to be applied at the ends of the estimation window

provides some freedom in how these solutions are constructed. This is in contrast to initial-value problems where conditions are matched at a the initial time. Here, in absence of an explicit closed form solution, Eqn. (2.13) is solved from somewhat arbitrary 'initial' conditions, which may generally be specified at any point in the time interval. The choice given by [29] integrates Eqn. (2.13) forward from $t = 0 \to T$ to construct the particular solution with $\boldsymbol{z}_0^{(p)} = \{\boldsymbol{\mu}_0, \mathbf{0}\}$. To match the boundary condition $\boldsymbol{p}_0 = \boldsymbol{R}_0 \cdot (\boldsymbol{x}_0 - \boldsymbol{\mu}_0)$, they write the homogeneous solution as

$$\boldsymbol{z}_t^{(h)} = \boldsymbol{\Phi}_{t|0} \cdot \begin{bmatrix} \mathbf{0} & \boldsymbol{R}_0^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \cdot \boldsymbol{z}_0 = \begin{bmatrix} \boldsymbol{\Psi}_t^{(x)} \cdot \boldsymbol{R}_0^{-1} \\ \boldsymbol{\Psi}_t^{(p)} \end{bmatrix} \cdot \boldsymbol{p}_0.$$

Note that this satisfies the boundary condition at $t = 0$ regardless of the value of $\boldsymbol{p}_0$, and amounts to taking

$$\boldsymbol{\Psi}_t = \begin{bmatrix} \boldsymbol{\Psi}_t^{(x)} \\ \boldsymbol{\Psi}_t^{(p)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{t|0}^{(xx)} + \boldsymbol{\Phi}_{t|0}^{(xp)} \cdot \boldsymbol{R}_0 \\ \boldsymbol{\Phi}_{t|0}^{(px)} \cdot \boldsymbol{R}_0^{-1} + \boldsymbol{\Phi}_{t|0}^{(pp)} \end{bmatrix}.$$

The terminal condition $\boldsymbol{p}_T = 0$ is then enforced by taking $\boldsymbol{z}_0 \to \boldsymbol{\Phi}_{0|T} \cdot \boldsymbol{z}_T$, and the solution is

$$\boldsymbol{z}_t = \boldsymbol{z}_t^{(p)} - \boldsymbol{\Phi}_{t|0} \cdot \begin{bmatrix} \mathbf{0} & \boldsymbol{R}_0^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \cdot \boldsymbol{\Phi}_{0|T} \cdot \boldsymbol{z}_T^{(p)} = \boldsymbol{z}_t^{(p)} - \begin{bmatrix} \boldsymbol{\Psi}_t^{(x)} \cdot \boldsymbol{R}_0^{-1} \\ \boldsymbol{\Psi}_t^{(p)} \end{bmatrix} \cdot [\boldsymbol{\Psi}_T^{(p)}]^{-1} \cdot \boldsymbol{p}_T^{(p)}. \tag{2.14}$$

Although the solution itself is unique [1], it has an infinite number of possible representations that depend on the choice of initial conditions for the particular solution, and the choice of functional form for the homogeneous solution. However, these representations are often unstable. This is evident in the simple tracking problem for instance, where the analytical solutions to Hamilton's equations contain terms that grow exponentially in time. And although that case is simple enough to admit an analytical solution, it makes this technique numerically intractable.

### 2.2.3 The Kalman-Bucy filter

On the other hand, given suitable observability assumptions the estimate of $\boldsymbol{z}_{T|T}$ given by the Kalman-Bucy filter is *guaranteed* to be stable. Starting from the previous solution Eqn. (2.14), [29] showed

[1] The linear action Eqn. (2.4) is a quadratic functional, with a single global minimum.

that it can be derived by considering

$$\boldsymbol{z}_{t|t} = \left(\boldsymbol{I} - \boldsymbol{R}_{t|t}^{-1}\right) \cdot \boldsymbol{z}_t^{(p)}$$

where

$$\boldsymbol{R}_{t|t}^{-1} = \begin{bmatrix} [\boldsymbol{R}_{t|t}^{(x)}]^{-1} \\ [\boldsymbol{R}_{t|t}^{(p)}]^{-1} \end{bmatrix} := \begin{bmatrix} \boldsymbol{\Psi}_t^{(x)} \cdot \boldsymbol{R}_0^{-1} \\ \boldsymbol{\Psi}_t^{(p)} \end{bmatrix} \cdot [\boldsymbol{\Psi}_t^{(p)}]^{-1}$$

and for the time being $\boldsymbol{R}_0 \to \boldsymbol{R}_0^{(x)}$. It is useful here to define a blockwise Hadamard product $\odot$ such that

$$\begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix} \odot \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{D} \end{bmatrix} := \begin{bmatrix} \boldsymbol{A} \cdot \boldsymbol{B} \\ \boldsymbol{C} \cdot \boldsymbol{D} \end{bmatrix},$$

so $\boldsymbol{R}_{t|t}^{-1}$ can be written as

$$\boldsymbol{R}_{t|t}^{-1} = \boldsymbol{\Psi}_t \odot \begin{bmatrix} \boldsymbol{R}_0^{-1} \\ \boldsymbol{I} \end{bmatrix} \cdot [\boldsymbol{\Psi}_t^{(p)}]^{-1} =: \boldsymbol{\Psi}_t \odot \tilde{\boldsymbol{\Psi}}_t^{-1}.$$

The time derivative of $\boldsymbol{R}_{t|t}^{-1}$ can then be written as

$$\begin{aligned} \dot{\boldsymbol{R}}_{t|t}^{-1} &= \dot{\boldsymbol{\Psi}}_t \odot \tilde{\boldsymbol{\Psi}}_t - \boldsymbol{R}_{t|t}^{-1} \odot \dot{\tilde{\boldsymbol{\Psi}}}_t \odot \tilde{\boldsymbol{\Psi}}_t^{-1} \\ &= \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H} \cdot \boldsymbol{R}_{t|t}^{-1} - \boldsymbol{R}_{t|t}^{-1} \odot \left( \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H} \cdot \boldsymbol{R}_{t|t}^{-1} \right) \end{aligned}$$

Substituting for $\boldsymbol{J} \cdot \nabla_{zz}^2 \mathcal{H}$ gives the Riccati equations for the covariance

$$\dot{\boldsymbol{R}}_{t|t}^{-1} = \begin{bmatrix} \boldsymbol{F}_t \cdot [\boldsymbol{R}_{t|t}^{(x)}]^{-1} + [\boldsymbol{R}_{t|t}^{(x)}]^{-1} \cdot \boldsymbol{F}_t^{\dagger} + \boldsymbol{R}_f^{-1} \cdot [\boldsymbol{R}_{t|t}^{(p)}]^{-1} - [\boldsymbol{R}_{t|t}^{(x)}]^{-1} \cdot \boldsymbol{H}_t^{\dagger} \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_t \cdot [\boldsymbol{R}_{t|t}^{(x)}]^{-1} \\ (\boldsymbol{I} - [\boldsymbol{R}_{t|t}^{(p)}]^{-1}) \cdot (\boldsymbol{R}_m \cdot [\boldsymbol{R}_{t|t}^{(x)}]^{-1} - \boldsymbol{F}_t^{\dagger} \cdot [\boldsymbol{R}_{t|t}^{(p)}]^{-1}) \end{bmatrix}$$

when $[\boldsymbol{R}_{t|t}^{(p)}]^{-1} = \boldsymbol{I}$ and therefore $[\dot{\boldsymbol{R}}_{t|t}^{(p)}]^{-1} = \boldsymbol{0}$. The Kalman-Bucy equations may be derived from the time derivative of $\boldsymbol{z}_t$

$$\dot{\boldsymbol{z}}_{t|t} = (\boldsymbol{I} - \boldsymbol{R}_{t|t}^{-1}) \cdot \dot{\boldsymbol{z}}_t^{(p)} + \dot{\boldsymbol{R}}_t^{-1} \cdot \boldsymbol{z}_t^{(p)} = \begin{bmatrix} \boldsymbol{F}_t \cdot \boldsymbol{x}_t + [\boldsymbol{R}_{t|t}^{(x)}]^{-1} \cdot \boldsymbol{H}_t^{\dagger} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{H}_t \cdot \boldsymbol{x}_t) \\ -(\boldsymbol{I} - [\boldsymbol{R}_{t|t}^{(p)}]^{-1}) \cdot \left(\boldsymbol{F}_t^{\dagger} \cdot \boldsymbol{p}_t + \boldsymbol{H}_t^{\dagger} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{H}_t \cdot \boldsymbol{x}_t)\right) \end{bmatrix}.$$

Note the cancellation of terms required to produce this result. Also, $[\boldsymbol{R}_{t|t}^{(p)}]^{-1} = \boldsymbol{I}$ implies $\dot{\boldsymbol{p}}_t = 0$, so the canonical momentum is a constant of the motion. This fact will reappear frequently throughout the remainder of this chapter.

### 2.2.4 Smoothed estimates

The Kalman-Bucy algorithm provides the filtered estimate for the state $\boldsymbol{z}_{T|T}$ and covariance $\boldsymbol{R}_{T|T}^{-1}$ at the final time. To get an estimate for the optimal path $\boldsymbol{z}_{t|T}$ requires a backward pass that propagates the observation information backwards in time. As suggested by [29], this can be accomplished by simply integrating Hamilton's equations backwards in time starting from $\boldsymbol{z}_{T|T} = \{\boldsymbol{x}_{T|T}, \boldsymbol{0}\}$. Once again however, the stability of Hamilton's equations are questionable *both forward and backward* in time. Some alternatives, discussed by [123], are based on choosing to enforce

$$\boldsymbol{R}_{t|t} \cdot \boldsymbol{p}_{t|T} = \boldsymbol{x}_{t|T} - \boldsymbol{x}_{t|t}$$

as an invariant of Hamilton's equations. For instance, using this relation to eliminate $\boldsymbol{p}_{t|T}$ from Hamilton's equations gives the continuous time analog of the Rauch-Tung-Striebel smoother ([164]),

$$\dot{\boldsymbol{x}}_{t|T} = \boldsymbol{F}_t \cdot \boldsymbol{x}_{t|T} + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{R}_{t|t}^{-1} \cdot (\boldsymbol{x}_{t|T} - \boldsymbol{x}_{t|t}). \tag{2.15}$$

Similarly, one can eliminate $\boldsymbol{x}_{t|T}$ from the $\dot{\boldsymbol{p}}_{t|T}$ equation to obtain the 'dual' smoothing equations

$$-\dot{\boldsymbol{p}}_{t|T} = \boldsymbol{F}_t^{\dagger} \cdot \boldsymbol{p}_{t|T} + \boldsymbol{H}_t^{\dagger} \cdot \boldsymbol{R}_m \cdot \big(\boldsymbol{y}_t - \boldsymbol{H}_t \cdot (\boldsymbol{x}_{t|t} - \boldsymbol{R}_{t|t} \cdot \boldsymbol{p}_{t|T})\big).$$

Alternatively, one can use the approach given by [121, 57, 56], which performs the backward pass by running the Kalman filter equations backward. The optimal path is then constructed from a linear combination of the two passes. Namely,

$$\boldsymbol{x}_{t|T} = -(\boldsymbol{R}_t^{+} + \boldsymbol{R}_t^{-})^{-1} \cdot (\boldsymbol{x}_t^{+} + \boldsymbol{x}_t^{-})$$

where $\boldsymbol{x}_t^{+}$, $\boldsymbol{x}_t^{-}$ and $\boldsymbol{R}_t^{+}$, $\boldsymbol{R}_t^{-}$ are respectively the state and (inverse) covariance estimates for the forward and backward paths. It will be shown later how these methods are related to minimizing the Eqn. (2.4) using Newton's method and linearizing around the filtered path.

### 2.2.5 The smoothed covariance

As for the smoothed covariance, [29] considered the covariance matrix

$$\boldsymbol{P}_t = \begin{bmatrix} \boldsymbol{P}_t^{(xx)} & \boldsymbol{P}_t^{(xp)} \\ \boldsymbol{P}_t^{(px)} & \boldsymbol{P}_t^{(pp)} \end{bmatrix} := \begin{bmatrix} \delta\boldsymbol{x}_t \\ \delta\boldsymbol{p}_t \end{bmatrix} \cdot \begin{bmatrix} \delta\boldsymbol{x}_t \\ \delta\boldsymbol{p}_t \end{bmatrix}^\dagger .$$

of fluctuations $\delta\boldsymbol{z}_t = \boldsymbol{J} \cdot \nabla^2 \mathcal{H}_t \cdot \delta\boldsymbol{z}_t$ about the optimal path. Its time evolution is given by

$$\dot{\boldsymbol{P}}_t = \mathcal{F}_t \cdot \boldsymbol{P}_t + \boldsymbol{P}_t \cdot \mathcal{F}_t^\dagger + \boldsymbol{\Xi}_t$$

with boundary conditions

$$\boldsymbol{P}_0^{(xx)} - \boldsymbol{R}_0^{-1} \cdot \boldsymbol{P}_0^{(px)} - \boldsymbol{P}_0^{(xp)} \cdot \boldsymbol{R}_0^{-1} + \boldsymbol{R}_0^{-1} \cdot \boldsymbol{P}_0^{(pp)} \cdot \boldsymbol{R}_0^{-1} = \boldsymbol{R}_0^{-1}$$
$$\boldsymbol{P}_T^{(pp)} = \boldsymbol{P}_T^{(xp)} = \boldsymbol{P}_T^{(px)} = \boldsymbol{0}$$

where

$$\mathcal{F}_t := \boldsymbol{J} \cdot \nabla^2_{zz} \mathcal{H}_t = \begin{bmatrix} \boldsymbol{F}_t & \boldsymbol{R}_f^{-1} \\ \boldsymbol{R}_m & -\boldsymbol{F}_t^\dagger \end{bmatrix} \qquad \boldsymbol{\Xi}_t := \begin{bmatrix} \boldsymbol{R}_f^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_t^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_t \end{bmatrix}.$$

Thus, the optimal smoothed covariance is also the solution to a two-point boundary value problem. Note that the inhomogeneous term $\boldsymbol{\Xi}_t$ is purely a result of the probabilistic interpretation. Specifically, defining $\boldsymbol{P}_t = \mathbb{E}[\delta\boldsymbol{z}_t \cdot \delta\boldsymbol{z}_t^\dagger]$ as an expectation value, it arises from stochastic fluctuations in the model and measurements

$$\mathbb{E}[\delta\boldsymbol{u}_t \cdot \delta\boldsymbol{u}_{t'}^\dagger] = \boldsymbol{R}_f^{-1}\,\delta(t-t') \qquad \mathbb{E}[\delta\boldsymbol{y}_t \cdot \delta\boldsymbol{y}_{t'}^\dagger] = \boldsymbol{R}_m^{-1}\,\delta(t-t') \qquad \mathbb{E}[\delta\boldsymbol{y}_t \cdot \delta\boldsymbol{u}_{t'}^\dagger] = \boldsymbol{0}.$$

This is, to my knowledge, the *only* departure between the deterministic and stochastic interpretations of the linear problem. And while the optimal filtered estimate is the same regardless of the choice of interpretation (though the value of the optimal cost function is not [132]) it is not clear whether this is also true for the smoothed solution. While this has likely been addressed somewhere in the literature, I have not been able to locate it.

The solution proposed by [29] is to run the Kalman-Bucy filter forward to construct the filtered covariance estimate, and then integrate Hamilton's equations backward to get the smoothed covariance. As previously noted however, this has numerical stability issues. On the other hand, the alternatives mentioned

above give stable algorithms for the smoothed covariance. See [123] for some interesting connections between the various approaches.

### 2.2.6 Discrete time

Solutions to the discrete problem Eqn. (2.9) can be constructed in similar fashion using the discrete analogs of $\boldsymbol{\Psi}_n,\ \boldsymbol{\Phi}_n$. The main departure is that in discrete time, the two-point boundary value problem Eqn. (2.10) mixes the time indices between $\boldsymbol{x}_{n+1},\ \boldsymbol{p}_n$, rather than having both $\dot{\boldsymbol{x}}_t,\ \dot{\boldsymbol{p}}_t$ defined at the same time as in the continuous case. This means that constructing the transition matrix $\boldsymbol{\Phi}_n$ by direct integration of (the homogeneous) Hamilton's equations requires the discrete model to be invertible. Namely, solving for $\boldsymbol{p}_n$ in terms of $\boldsymbol{p}_{n+1}$ gives

$$\begin{bmatrix} \boldsymbol{x}_{n+1} \\ \boldsymbol{p}_{n+1} \end{bmatrix} = \quad = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{R}_f^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{F}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}_n^{-\dagger} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{R}_m & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{p}_n \end{bmatrix}$$

plus inhomogeneous terms that depend on $\boldsymbol{y}_n$. This Schur factorization in terms of upper and lower triangular matrices will show up again later. While perhaps not a serious constraint if the $\boldsymbol{F}_n$ is a numerical discretization of a continuous model $\boldsymbol{f}_n(\boldsymbol{x}_n)$, the existence of $\boldsymbol{F}_n^{-\dagger}$ is nonetheless not required by the Kalman solution ([90]) or its smoothing counterparts ([164, 121, 57, 56]). This has likely contributed to why (to my knowledge) the discrete time analog to Bryson and Frazier's solution has never been published. The approaches given by [90, 164, 121, 57] all to some extent use the statistical structure of the problem to obtain their result. Mayne's solutions ([121]) rely on these properties the least, as they are primarily based on the discrete Hamilton-Jacobi-Bellman equations. These solutions will be discussed in more detail in Sec. (2.3).

Also, recently [10] showed that all these methods may be derived from direct solution of $\nabla A(\mathcal{X}^{(d)}) = 0$ from Eqn. (2.6), by exploiting the block tridiagonal structure of $\nabla^2 A(\mathcal{X}^{(d)})$. Specifically, sequential blockwise Gauss-Jordan elimination on the lower diagonal blocks $\nabla^2 A(\mathcal{X}^{(d)})$ can be directly interpreted as a modified form of the Kalman filter. The resulting system then is upper block tridiagonal, which can then be solved by simple back-substitution to obtain the smoothed solution. This technique can be applied forward, backward, and both forward and backward, respectively giving generalizations of the Rauch-Tung-Striebel ([164]), Mayne ([121]), and Mayne-Fraser two-filter solutions ([121, 57, 56]). This result connects Kalman filtering/smoothing with Newton's method ([14, 14, 77]). It will be discussed in more detail in Sec. (2.3), and generalized to nonlinear models, and connect with the Gauss-Newton method.

It is worth pointing out however, that while all these solutions are indeed equivalent in the linear case (since the action Eqn. (2.6) has an unique global minimum), this is not true in general. In particular, it will be shown numerically in Chap. (3) that methods that fully exploit the sequential and recursive properties of the Kalman solution can perform considerably better, especially when the observations are spatially sparse.

### 2.2.7 The simplest tracking problem revisited

This section now concludes by revisiting the simplest tracking problem described above to show that the Kalman smoothing solution is indeed stable. In order to treat the case where $R_0 = 0$ the solution must be modified to use its *information form*, which instead propagates the information state $\tilde{\boldsymbol{q}}_t := R_t\, q_t$. It also uses the Riccati equation for the (inverse) covariance. Namely, in this case,

$$-\dot{R}_t = R_t\, R_f^{-1}\, R_t - R_m.$$

This Riccati equation can be solved by introducing the change of variables $R_t \to P_t\, Q_t^{-1}$, where $Q,\ P$ are $D \times D$ matrices. The above equation becomes

$$-\dot{R}_t = P_t\, Q_t^{-1}\, \dot{Q}_t\, Q_t^{-1} - \dot{P}_t\, Q_t^{-1} = P_t\, Q_t^{-1}\, R_f^{-1}\, P_t\, Q_t^{-1} - R_m,$$

where even though $D = 1$, these expressions have not been simplified so as to facilitate its generalization to higher dimensions. Equating terms provides the well-known linearization of the Riccati equation

$$\begin{bmatrix} \dot{Q}_t \\ \dot{P}_t \end{bmatrix} = \begin{bmatrix} 0 & R_f^{-1} \\ R_m & 0 \end{bmatrix} \cdot \begin{bmatrix} Q_t \\ P_t \end{bmatrix},$$

which is equivalent to its vector-valued version given above by Eqn. (2.12). Its solutions are also described by exponentials,

$$\begin{aligned} Q_t &= \omega \sqrt{R_f^{-1}} \left( C^{(+)}\, e^{+\omega t} + C^{(-)}\, e^{-\omega t} \right) \\ P_t &= \omega \sqrt{R_m} \left( C^{(+)}\, e^{+\omega t} - C^{(-)}\, e^{-\omega t} \right). \end{aligned}$$

The Kalman filter solution requires $R_0 = 0$, but there are an infinite number of ways to satisfy this. The simplest choice is arguably $P_0 = 0,\ Q_0 = 1$, giving $C^{(+)} = C^{(-)} = (2\,\omega\sqrt{R_f^{-1}})^{-1}$, and therefore

$$\begin{aligned} Q_t &= \cosh(\omega t) \\ P_t &= \sqrt{Rm R_f}\,\sinh(\omega t) \\ R_t &= \sqrt{Rm R_f}\tanh(\omega t). \end{aligned}$$

With this, the Kalman filter equation for $\dot{x}_t$ becomes

$$\dot{q}_t = \dot{x}_t^* - \dot{x}_t = -R_t^{-1} R_m\, q_t = -\frac{\omega q_t}{\tanh(\omega t)} \quad \Longrightarrow \quad q_t = q_0\,\mathrm{csch}(\omega t),$$

which is forward stable, but is undefined at $t = 0$. As mentioned above, this issue can be avoided using the information form $\tilde{\boldsymbol{q}}_t = R_t\, q_t$. Multiplying both sides by $R_t$ gives

$$\dot{\tilde{\boldsymbol{q}}}_t = (\dot{R}_t - R_m)\, q_t = -R_t\, R_f^{-1}\, \tilde{\boldsymbol{q}}_t = -\omega\tanh(\omega t) \quad \Longrightarrow \quad \tilde{\boldsymbol{q}}_t = \tilde{\boldsymbol{q}}_0\,\mathrm{sech}(\omega t),$$

which is both forward stable and well-defined at $t = 0$. However, the fact that

$$R_t \Longrightarrow \tilde{\boldsymbol{q}}_0 = 0 \Longrightarrow q_t = 0$$

makes this case less interesting, as there is no need for a backwards pass since the true solution $q_t = 0$ is found automatically. Also, the 'classic' information filter scheme only really applies when the model $\boldsymbol{f}(\cdot)$ is linear, otherwise one has to linearize the dynamics to determine the evolution for the information state. [2]

Consider now a slightly more illuminating case where one still has no idea of the initial state (so $q_0$ is chosen arbitrarily) but decides to set $R_0 = 1$ so that the Kalman filter can be run without singularity issues. The solution to the Riccati equation for the covariance is

$$R_t^{-1} = \frac{\tanh(\omega t + \phi_0)}{\sqrt{R_m R_f}} \qquad \phi_0 := \tanh^{-1}(\sqrt{R_m R_f})$$

and thus

$$\dot{q}_t = -R_t^{-1}\, R_m\, q_t = \omega\tanh(\omega t + \phi_0)\, q_t \quad \Longrightarrow \quad q_t = q_0\,\mathrm{sech}(\omega t + \phi_0).$$

[2] But see [33, 143, 203] for modern examples of nonlinear information filters.

Despite the incorrect choice of $R_0$, the Kalman filter still asymptotically approaches the correct solution for all possible values of $q_0$, and is forward stable.

As for the backwards pass, directly integrating Hamilton's equations backwards in time is still unstable. By contrast, the Rauch-Tung-Striebel solution Eqn. (2.15) in this case gives

$$-\dot{q}_{t|T} = R_f^{-1} R_{t|t}^{-1} (q_{t|t} - q_{t|T}).$$

It has an even more complicated closed form solution, which is backwards stable but not explicitly given here. This solution may however be viewed as *backwards synchronization* with of the smoothing solution with the filtered solution. Moreover, it is also evident that the two-filter solution (*i.e.*, running a Kalman filter backwards from $q_{t|t}$) will continue to refine the estimate. So one does not have to take $t \to \infty$ to achieve $q_t \to 0$, but can rather iterate forwards and backwards passes until convergence is achieved.

Thus, although this is possible the most trivial estimation problem; one where the answer is already known. It nonetheless has a nontrivial solution that provides insight into the structure of the solutions and highlights some rather important features. For instance, the Hamiltonian system that defines the two-point boundary-value-problem is inherently unstable, and cannot be explicitly integrated. This more or less rules out any type of single shooting method for matching the boundary conditions, as the solutions quickly become numerically unstable. Rather, one must use multiple shooting techniques, in which the estimation window is broken up into small subintervals and then solved together as one large optimization problem.

Alternatively, one could use the Kalman technique to *transforms* the problem into an equivalent one that is numerically stable and can be explicitly integrated forwards and backwards in time to achieve a highly accurate estimate. In this case, the problem has the structure of an harmonic oscillator, except for a sign difference. Its solutions are hyperbolic transcendental functions that exponentially decay at a rate of $\omega^{-1}$ determined by the choice $R_m,\ R_f$. Despite the statistical foundations of the problem, only the *relative ratio* $R_m/R_f$ matters. When $\omega$ is large, the solutions decay quickly, as the solutions tend towards the data, which in this case is perfect. Likewise, when $\omega$ is small, they decay slowly, as the solutions prefer to follow the model. In the more realistic situation, where the model and the data both have errors, these values can be tuned to optimally dampen this noise.

Finally, at an informal level it is interesting to note that the Kalman filter involves the solution of two Hamiltonian systems. One of these systems is the standard vector-valued Hamilton's equations, operating on the gradient of the Hamiltonian $\dot{\boldsymbol{z}}_t = \boldsymbol{J} \cdot \nabla_z \mathcal{H}_t$, with the observations providing the time-

dependent forcing. The other is matrix-valued, $\dot{\boldsymbol{\Phi}}_t = \boldsymbol{J} \cdot \nabla^2_{zz} \mathcal{H}_t \cdot \boldsymbol{\Phi}_t$. It describes the behavior of neighboring extremals around the optimal solution, and linearizes the Riccati equation for the covariance by way of a linear fractional transformation. By themselves, these two systems are inherently unstable, and cannot be explicitly integrated for an appreciable amount of time without becoming numerically unstable. The Kalman solution transforms these two systems, by embedding one inside the other in a way that stabilizes the resulting solution through a rather fortuitous cancellation of terms. The remarkable structure of its solutions has far-reaching implications, and a rather direct interpretation in terms of Newton's method for minimizing the action. These ideas will now be explored in the context of its extension to nonlinear systems.

## 2.3 Nonlinear extensions

The example given in the previous section outlined the difficulty associated with applying the necessary conditions for optimality to the solution of two-point boundary value problems. Namely, Hamilton's equations appear to be inherently unstable, even for the simplest problems in optimal estimation. Kalman's solution on the other hand is inherently stable given some rather basic assumptions. It may be viewed as a nonlinear transformation of Hamilton's equations that increases the number of independent variables from $2D$ to $D(D+1)$, and inherits some truly remarkable properties that make it widespread across a broad range of topics in applied mathematics.

This section explores the nonlinear extension of these solutions and their connections to second variation methods in calculus of variations, and the technique of differential dynamic programming in optimal control. Both approaches use local linearizations of the models $\boldsymbol{f}(\cdot)$, $\boldsymbol{h}(\cdot)$ to refine an initial approximate trajectory, and may be viewed as working in a local 'tube' around the proposed solution. While these techniques have been used rather extensively in optimal control, its estimation analog does not seem to have been reported in the literature. The results will provide deeper insight into the structure of the problem, including its inherent connections to Newton's method in optimization, as well as the role the canonical momentum plays in the optimality of its solutions.

### 2.3.1 Two-pass smoothing and Newton's method

The term differential dynamic programming was given to techniques developed by Jacobson ([79]) and Mayne ([120]) in the mid 1960s for computing approximate solutions to nonlinear optimal

control problems. The techniques were developed as nonlinear extensions to the seminal work of Bellman ([15]), which avoid some of the issues regarding the 'curse of dimensionality' associated with the numerical solution of PDEs. A closely related solution was also proposed Mitter ([133]), but derived from second variation method in classical calculus of variations.

While these solutions are all equivalent for linear problems, subtle but important differences exist in the nonlinear case. These distinctions are mainly due to how the linearization is performed and how the iterations are initialized. At the time, the results of [80] appeared to be the most accurate. Yet a recent resurgence of 'new' but closely related methods ([114, 191]) suggests the book may not be entirely closed on the subject. At the very least, these new methods should be reinterpreted in Jacobson's framework, which appears to be the most general of these results.

The core algorithm is based on Newton's method, but utilizes the Hamiltonian structure of the problem to restrict the set of solutions in a way that provides inherent stability properties based on locally optimal feedback control. In the control theoretic formulation, the Hamiltonian action Eqn. (2.3) is expanded at each iteration $i$ around a nominal trajectory $\mathcal{X}^{(i)}$ generated by the current best guess for the initial state $\boldsymbol{x}_0^{(i)}$ and the dynamical controls $\mathcal{U}^{(i)}$. While the linearization is often performed around $\mathcal{X}^{(i)}$, [79] showed this can be generalized by splitting the perturbation $\mathcal{U}^{(i+1)} = \mathcal{U}^{(i)} + \delta\mathcal{U}^{(i,*)} + \delta\mathcal{U}$ and $\mathcal{X}^{(i+1)} = \mathcal{X}^{(i)} + \delta\mathcal{X}^{(i)}$, where $\delta\mathcal{U}^{(i,*)}$ is the 'optimal' perturbation at iteration $i$ given $\mathcal{X}^{(i)}$, and $\delta\mathcal{X}^{(i)}, \delta\mathcal{U}^{(i)}$ are corrections computed by linearizing about the 'optimal' trajectory, computed with $\mathcal{U}^{(i)} + \delta\mathcal{U}^{(i,*)}$. The role of the Lagrange multipliers $\mathcal{P}^{(i)}$ and their perturbations $\delta\mathcal{P}^{(i)}$ are mostly ignored however in these treatments.

For estimation, the linear solution implies a natural splitting of the perturbation $\delta\mathcal{X}^{(i)} = \mathcal{X}^{(i,-)} - \mathcal{X}^{(i,+)}$, into filtered $\mathcal{X}^{(i,+)} := \{\boldsymbol{x}_{t|t}^{(i)}\}|_{t=0}^{T}$ and smoothed $\delta\mathcal{X}^{(i,-)} = \{\boldsymbol{x}_{t|T}^{(i)}\}|_{t=T}^{0}$ components. The filtered solution is generated by running the extended extended Kalman filter forward in time from initial conditions, given by the previous iteration. Once $\boldsymbol{x}_{t|t}^{(i)}$ is known, the problem is to determine $\boldsymbol{p}_{t|T}^{(i)}$ given the boundary conditions $\boldsymbol{p}_{T|T} = 0$. This is solved by propagating the perturbation $\delta\boldsymbol{x}_t^{(i)} = \boldsymbol{p}_{t|T}^{(i)}$ backwards in time, in a process that is equivalent to Newton's method ([83]). The result of each iteration is a new initial estimate $\boldsymbol{x}_0^{(i+1)} := \boldsymbol{x}_{0|T}^{(i)}$, and $\boldsymbol{R}_0^{(i+1)} := \boldsymbol{R}_{0|T}^{(i)}$.

Thus, the role of the canonical momentum is thus limited to the smoothing step. Indeed, it will be shown that the Kalman solution requires $\boldsymbol{p}_{t|t}^{(i)} = 0$ everywhere. This is a somewhat unexpected revelation, which has been largely ignored in the filtering literature. While this statement will not be proven formally, several justifying arguments will be provided. A more intuitive interpretation of this fact, in terms of the

optimality of the filtered estimate, will also be provided.

### 2.3.2 The second variation and Hamilton's equations

The derivation begins by writing the action in $2D$ canonical coordinates $\mathcal{Z} = \{\mathcal{X}, \mathcal{P}\}$, $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{p}\}$ and expanding to

$$A(\mathcal{Z}^{(i,-)}, \delta\mathcal{Z}^{(i)}) \approx A(\mathcal{Z}^{(i,-)}) + \delta A^{(i,-)} + \frac{1}{2}\delta^2 A^{(i,-)}.$$

Suppressing the iteration index $i$, the first and second variations are given by

$$\begin{aligned}\delta A &= \langle \nabla_{\boldsymbol{\delta}\mathcal{Z}} A, \delta\mathcal{Z}\rangle \\ &= \langle \delta\boldsymbol{x}_T, \boldsymbol{p}_{t|T}\rangle - \langle \delta\boldsymbol{x}_0, \boldsymbol{p}_{0|T} - \boldsymbol{R}_{0|0}\cdot(\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0})\rangle + \int_0^T dt \left\langle \delta\boldsymbol{z}_{t|T}, \boldsymbol{J}^\dagger \cdot \dot{\boldsymbol{z}}_{t|T} - \nabla_z \mathcal{H}_{t|T} \right\rangle\end{aligned}$$

$$\begin{aligned}\delta^2 A &= \langle \boldsymbol{\delta}\mathcal{Z}, \nabla^2_{\boldsymbol{\delta}\mathcal{Z}^2} A \cdot \boldsymbol{\delta}\mathcal{Z}\rangle \\ &= \langle \delta\boldsymbol{x}_T, \delta\boldsymbol{p}_T\rangle - \langle \delta\boldsymbol{x}_0, \delta\boldsymbol{p}_0 - \boldsymbol{R}_{0|0}\cdot\delta\boldsymbol{x}_0\rangle + \int_0^T dt \left\langle \delta\boldsymbol{z}_t, \boldsymbol{J}^\dagger\cdot\delta\dot{\boldsymbol{z}}_t - \nabla^2_{zz}\mathcal{H}_{t|T}\cdot\delta\boldsymbol{z}_t \right\rangle.\end{aligned}$$

Note the anti-symmetry property $\boldsymbol{J} = -\boldsymbol{J}^\dagger$ gives

$$\frac{1}{2}\int dt\, \nabla_{\delta\boldsymbol{z}} \langle \delta\boldsymbol{z}_t, \boldsymbol{J}^\dagger\cdot\delta\dot{\boldsymbol{z}}_t\rangle = \frac{1}{2}\int dt \left( \boldsymbol{J}^\dagger\cdot\delta\dot{\boldsymbol{z}}_t + \nabla_{\delta\boldsymbol{z}} \frac{d}{dt} |\delta\boldsymbol{z}_t|^2_{\boldsymbol{J}^\dagger} - \boldsymbol{J}\cdot\delta\dot{\boldsymbol{z}}_t \right) = \int dt\, \boldsymbol{J}^\dagger\cdot\delta\dot{\boldsymbol{z}}_t.$$

Enforcing the variational principle $\delta A = \delta^2 A = 0$ thus results in a pair of nested two-point boundary value problems

$$\begin{aligned}&\boldsymbol{J}^\dagger\cdot\dot{\boldsymbol{z}}_{t|T} = \nabla_z \mathcal{H}_{t|T} && \boldsymbol{p}_{0|T} = \boldsymbol{R}_{0|0}\cdot(\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0}) && \boldsymbol{p}_{t|T} = 0 \\ &\boldsymbol{J}^\dagger\cdot\delta\dot{\boldsymbol{z}}_t = \nabla^2_{\boldsymbol{zz}}\mathcal{H}_{t|T} && \delta\boldsymbol{p}_0 = \boldsymbol{R}_{0|0}\cdot\delta\boldsymbol{x}_0 && \delta\boldsymbol{p}_T = 0.\end{aligned}$$

These are Hamilton's equations (and its linearization) evaluated along the smoothed solution $\boldsymbol{z}_{t|T} = \boldsymbol{z}_{t|t} + \delta\boldsymbol{z}_t$.

### 2.3.3 The Riccati transformation

The linearized two-point boundary value problem implied by $\delta^2 A$ may be solved by making the Riccati transformation $\delta\boldsymbol{p} = \boldsymbol{R}_{t|t}\cdot\delta\boldsymbol{x}_t + \delta\boldsymbol{r}_t$, which converts it to an initial value problem at the expense of an extra system of $D\times D$ ordinary differential equations. The additional variation $\delta\boldsymbol{r}_t$ is not standard, but introduced here to maintain the dimension of the problem and to imply that it should be small. Letting

$\delta\tilde{\boldsymbol{z}}_t := \{\delta\boldsymbol{x}_t, \delta\boldsymbol{r}_t\}$, $\delta^2 A$ becomes

$$\delta^2 A = \langle \delta\boldsymbol{x}_T, \boldsymbol{R}_{t|t} \cdot \delta\boldsymbol{x}_T - \delta\boldsymbol{r}_T \rangle + \langle \delta\boldsymbol{x}_0, \delta\boldsymbol{r}_0 \rangle + \int_0^T dt \, \langle \delta\tilde{\boldsymbol{z}}_t, \boldsymbol{J}^\dagger \cdot \delta\dot{\tilde{\boldsymbol{z}}}_t \rangle - |\delta\tilde{\boldsymbol{z}}_t|^2_{\boldsymbol{\Xi}_t}$$

where

$$\boldsymbol{\Xi}_t := \left[\begin{array}{c|c} \dot{\boldsymbol{R}}_{t|t} + \boldsymbol{\Theta}_t(\boldsymbol{x}_{t|T}, \boldsymbol{R}_{t|t}) & \nabla^2_{xp}\mathcal{H}_{t|T} + \boldsymbol{R}_{t|t} \cdot \nabla^2_{pp}\mathcal{H}_{t|T} \\ \nabla^2_{px}\mathcal{H}_{t|T} + \nabla^2_{pp}\mathcal{H}_{t|T} \cdot \boldsymbol{R}_{t|t} & \nabla^2_{pp}\mathcal{H}_{t|T} \end{array}\right]$$

where

$$\boldsymbol{\Theta}_t(\boldsymbol{x}_{t|T}, \boldsymbol{R}_{t|t}) := \nabla^2_{xx}\mathcal{H}_{t|T} + \boldsymbol{R}_{t|t} \cdot \nabla^2_{px}\mathcal{H}_{t|T} + \nabla^2_{xp}\mathcal{H}_{t|T} \cdot \boldsymbol{R}_{t|t} + \boldsymbol{R}_{t|t} \cdot \nabla^2_{pp}\mathcal{H}_{t|T} \cdot \boldsymbol{R}_{t|t}. \tag{2.16}$$

Setting the derivative of $\delta^2 A$ with respect to $\delta\tilde{\boldsymbol{z}}_t$ equal to zero gives the Riccati equation for the inverse covariance propagation in the upper left block of $\boldsymbol{\Xi}_t$ with $\mathcal{H}_t$ evaluated along the smoothed solution. The filtered estimate $\boldsymbol{R}_{t|t}$ can be obtained by expanding the derivatives of $\nabla^2\mathcal{H}_t$ about the filtered solution and ignoring all third order terms, so that effectively $\nabla^2\mathcal{H}_{t|T} \to \nabla^2\mathcal{H}_{t|t}$. The lower left and upper right blocks give equations for $\delta\dot{\boldsymbol{x}}_t$ and $-\delta\dot{\boldsymbol{r}}_t$ respectively. The lower right block gives $\boldsymbol{R}_f^{-1} \cdot \delta\boldsymbol{r}_t = 0$, which implies that $\delta\boldsymbol{r}_t = 0$ everywhere. This observation is further supported by the fact that the $\delta\dot{\boldsymbol{r}}$ equation is linear and homogeneous and has a boundary condition $\delta\boldsymbol{r}_0 = 0$.

### 2.3.4 The filtered solution

Introducing the Riccati transformation into the first variation $\delta A$ gives

$$\left\langle \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{R}_{t|t} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \delta\boldsymbol{x}_t \\ \delta\boldsymbol{r}_t \end{bmatrix}, \boldsymbol{J}^\dagger \cdot \dot{\boldsymbol{z}}_{t|T} - \nabla_z\mathcal{H}_{t|T} \right\rangle$$

inside the integral. Taking the variation with respect to $\delta\boldsymbol{x}_t$ and $\delta\boldsymbol{r}_t$ results in a linear combination of Hamilton's equations

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{R}_{t|t} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \left(\boldsymbol{J}^\dagger \cdot \dot{\boldsymbol{z}}_{t|T} - \nabla_z\mathcal{H}_{t|T}\right) = \begin{bmatrix} \boldsymbol{R}_{t|t} \cdot (\dot{\boldsymbol{x}}_{t|T} - \nabla_p\mathcal{H}_{t|T}) - (\dot{\boldsymbol{p}}_{t|T} + \nabla_x\mathcal{H}_{t|T}) \\ \dot{\boldsymbol{x}}_{t|T} - \nabla_p\mathcal{H}_{t|T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}.$$

At first glance, this appears to imply Hamilton's equations must hold along the smoothed path

$$\dot{\boldsymbol{x}}_{t|T} = \nabla_p\mathcal{H}_{t|T} \Longrightarrow -\dot{\boldsymbol{p}}_{t|T} = \nabla_x\mathcal{H}_{t|T}.$$

But since the bottom row is the variation with respect to $\delta \boldsymbol{r}$, which is zero everywhere, the condition $\dot{\boldsymbol{x}}_{t|T} = \nabla_p \mathcal{H}_{t|T}$ does not have to be met. This provides some freedom to shape the equations of motion as desired, while still satisfying the variational principle provided the following condition is met

$$\boldsymbol{R}_{t|t} \cdot (\dot{\boldsymbol{x}}_{t|T} - \nabla_p \mathcal{H}_{t|T}) - (\dot{\boldsymbol{p}}_{t|T} + \nabla_x \mathcal{H}_{t|T}) = 0. \tag{2.17}$$

Substituting expressions for the derivatives of the Hamiltonian gives

$$\boldsymbol{R}_{t|t} \cdot \left(\dot{\boldsymbol{x}}_{t|T} - \boldsymbol{f}(\boldsymbol{x}_{t|T}) - \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_{t|T}\right) - \left(\dot{\boldsymbol{p}}_{t|T} + \nabla^\dagger \boldsymbol{f}_{t|T} \cdot \boldsymbol{p}_{t|T} + \nabla^\dagger \boldsymbol{h}_{t|T} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_{t|T}))\right) = 0.$$

Linearizing this equation around the filtered estimate $\boldsymbol{z}_{t|T} \to \boldsymbol{z}_{t|t} + \delta \boldsymbol{z}_t$ and rearranging terms, the zeroth order approximation is simply

$$\boldsymbol{R}_{t|t} \cdot \left(\dot{\boldsymbol{x}}_{t|t} - \boldsymbol{f}(\boldsymbol{x}_{t|t}) - \boldsymbol{R}_{t|t}^{-1} \cdot \nabla^\dagger \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_{t|t}))\right) - \left(\dot{\boldsymbol{p}}_{t|t} + (\nabla^\dagger \boldsymbol{f}_{t|t} + \boldsymbol{R}_{t|t} \cdot \boldsymbol{R}_f^{-1}) \cdot \boldsymbol{p}_{t|t}\right) = 0.$$

Choosing these equations to hold independently gives

$$\begin{aligned} \dot{\boldsymbol{x}}_{t|t} &= \boldsymbol{f}(\boldsymbol{x}_{t|t}) + \boldsymbol{R}_{t|t}^{-1} \cdot \nabla^\dagger \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_{t|t})) \\ -\dot{\boldsymbol{p}}_{t|t} &= (\nabla^\dagger \boldsymbol{f}_{t|t} + \boldsymbol{R}_{t|t} \cdot \boldsymbol{R}_f^{-1}) \cdot \boldsymbol{p}_{t|t}. \end{aligned} \tag{2.18}$$

The $\dot{\boldsymbol{x}}_t$ is immediately recognized as the equation for the extended Kalman-Bucy filter ([89]). Note that the filtered estimate evolves according to the *full nonlinear models* $\boldsymbol{f}$ and $\boldsymbol{h}$, not their linearization. The $\dot{\boldsymbol{p}}_t$ equation has the same form as $\delta \dot{\boldsymbol{r}}_t$ above. The boundary conditions

$$\boldsymbol{p}_{0|T} = \boldsymbol{p}_{0|0} + \boldsymbol{R}_{0|0} \cdot \delta \boldsymbol{x}_0 = \boldsymbol{R}_{0|0} \cdot (\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0}) \Longrightarrow \boldsymbol{p}_0 = \boldsymbol{p}_T = 0$$

imply that the Kalman filter solution is special in the sense that it enforces $\boldsymbol{p}_{t|t} = \dot{\boldsymbol{p}}_{t|t} = 0$ everywhere. More evidence for this will be presented shortly.

### 2.3.5 The backwards pass

The smoothed solution requires computing the perturbation $\delta \boldsymbol{x}_t$. Proceeding as before by collecting the first order terms gives

$$\boldsymbol{R}_{t|t} \cdot \left(\delta \dot{\boldsymbol{x}}_t - \nabla \boldsymbol{f}_{t|t} \cdot \delta \boldsymbol{x}_t - \boldsymbol{R}_f^{-1} \cdot \delta \boldsymbol{p}_t\right) - \left(\delta \dot{\boldsymbol{p}}_t + \nabla^\dagger \boldsymbol{f}_{t|t} \cdot \delta \boldsymbol{p}_t - \nabla^\dagger \boldsymbol{h}_{t|T} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_t \cdot \delta \boldsymbol{x}_t\right) = 0.$$

Substituting $\delta \boldsymbol{p}_t \to \boldsymbol{R}_{t|t} \cdot \delta \boldsymbol{x}_t + \delta \boldsymbol{r}_t$ reproduces the equations given above for $\dot{\boldsymbol{R}}_t$ and $\delta \dot{\boldsymbol{r}}_t$. Thus, the variational principle is satisfied to first order in $\delta \boldsymbol{z}_t$. However, simply integrating the linearized Hamilton's equations backwards in time with boundary conditions $\delta \boldsymbol{x}_T = \delta \boldsymbol{p}_T = 0$ implies $\delta \boldsymbol{x}_t = \delta \boldsymbol{p}_t = 0$ everywhere. Thus, another approach is needed.

Recognizing that these equations make the action a function of $\delta \boldsymbol{x}_0$ alone, the backwards pass is equivalent to an application of Newton's method

$$\delta \boldsymbol{x}_0 = \boldsymbol{x}_{0|0} - \boldsymbol{x}_{0|T} = -[\nabla^2_{\delta \boldsymbol{x}_0^2} A]^{-1} \cdot \nabla_{\delta \boldsymbol{x}_0} A = \boldsymbol{R}_{0|0}^{-1} \cdot \boldsymbol{p}_{0|T}.$$

Moreover, the fact that $\boldsymbol{p}_{t|t} = \delta \boldsymbol{r}_t = 0$ implies the following relations

$$\boldsymbol{R}_{t|t} \cdot (\boldsymbol{x}_{t|T} - \boldsymbol{x}_{t|t}) = \boldsymbol{R}_{t|t} \cdot \delta \boldsymbol{x}_t = \delta \boldsymbol{p}_t = \boldsymbol{p}_{t|T} - \boldsymbol{p}_{t|t} = \boldsymbol{p}_{t|T}.$$

These expressions can be used to derive a variety of methods for determining $\boldsymbol{x}_{t|T}$. For instance, following Bryson and Frazier's original solution one could directly integrate Hamilton's equations

$$\begin{aligned} \dot{\boldsymbol{x}}_{t|T} &= \boldsymbol{f}(\boldsymbol{x}_{t|T}) + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_{t|T} \\ -\dot{\boldsymbol{p}}_{t|T} &= \nabla^\dagger \boldsymbol{f}_{t|T} \cdot \boldsymbol{p}_{t|T} + \nabla^\dagger \boldsymbol{h}_{t|T} \cdot \boldsymbol{R}_m \cdot \left(\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_{t|T})\right) \end{aligned}$$

backwards in time from $\boldsymbol{x}_{t|T} = \boldsymbol{x}_{t|t},\ \boldsymbol{p}_{t|T} = 0$. In practice however, these equations are often numerically unstable ([123]). Alternatively, eliminating $\boldsymbol{p}_{t|T} \to \boldsymbol{R}_{t|t} \cdot (\boldsymbol{x}_{t|T} - \boldsymbol{x}_{t|t})$ in the $\dot{\boldsymbol{x}}_{t|T}$ gives the state equation for the extended Rauch-Tung-Striebel smoother ([42])

$$\dot{\boldsymbol{x}}_{t|T} = \boldsymbol{f}(\boldsymbol{x}_{t|T}) + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{R}_{t|t} \cdot \left(\boldsymbol{x}_{t|T} - \boldsymbol{x}_{t|t}\right).$$

Note how this equation may be interpreted as backwards synchronization between the filtered and smoothed estimate. Likewise, the dual approach expands $\boldsymbol{h}(\boldsymbol{x}_{t|T}) \approx \boldsymbol{h}(\boldsymbol{x}_{t|t}) + \nabla \boldsymbol{h}_{t|t} \cdot \delta \boldsymbol{x}_t = \boldsymbol{h}(\boldsymbol{x}_{t|t}) + \nabla \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_{t|t}^{-1} \cdot \boldsymbol{p}_{t|T}$

to eliminate $\boldsymbol{x}_{t|T}$ from the $\dot{\boldsymbol{p}}_{t|T}$ equation. Ignoring higher order derivatives, one obtains

$$-\dot{\boldsymbol{p}}_{t|T} = \nabla \boldsymbol{f}_{t|t} \cdot \boldsymbol{p}_{t|T} + \nabla^{\dagger} \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_m \cdot \big(\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{x}_{t|t}) - \nabla \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_{t|t}^{-1} \cdot \boldsymbol{p}_{t|T}\big),$$

which is essentially a linearization of the successive sweep method ([122]). Suppressing the $\nabla \boldsymbol{h}_{t|t} \cdot \boldsymbol{R}_{t|t}^{-1} \cdot \boldsymbol{p}_{t|T}$ leads to the 4DVar approximation, which integrates Hamilton's equations for $\dot{\boldsymbol{p}}_{t|T}$ backwards in time, substituting $\boldsymbol{x}_{t|T} \to \boldsymbol{x}_{t|t}$ along the way.

The distinctions between these techniques all arise from how Hamilton's equations are linearized, with different choices giving to different algorithms. Thus, like the filtering solution, there is a great deal freedom to how the backwards pass is implemented. Unlike the filtering case however, no one solution stands out in terms of stability, particularly for nonlinear problems. But it is clear that certain choices are inherently unstable. And properly assessing the relative strengths and weaknesses between these methods requires a more comprehensive understanding of how they are connected.

### 2.3.6 The Hamilton-Jacobi solution

The optimality condition Eqn. (2.17) is now revisited from the point of view of classical Hamilton-Jacobi theory. Bellman's extension of this theory for optimal control problems is not required here since the controls are eliminated by imposing $\nabla_u \tilde{\mathcal{H}}_t = 0 \Longrightarrow \boldsymbol{u}_t = \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_t$. Hamilton-Jacobi theory takes $A_{t|T} := A(\boldsymbol{x}_{t|T}, t)$ as a function of $\boldsymbol{x}_{t|T}$ and substitutes $\boldsymbol{p}_{t|T} \to \nabla A_{t|T}$. Using the Hamilton-Jacobi equation

$$\frac{dA_t(\boldsymbol{x}_t)}{dt} = \langle \nabla_x A_t, \dot{\boldsymbol{x}}_t \rangle + \partial_t A_t(\boldsymbol{x}_t) = \langle \boldsymbol{p}_t, \dot{\boldsymbol{x}}_t \rangle - \mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{p}_t) \Longrightarrow -\partial_t A(\boldsymbol{x}_t) = \mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{p}_t),$$

one can derive the Riccati equation by taking its second derivative (ignoring higher order terms)

$$-\partial_t \big(\nabla^2 A_{t|T}\big) = \nabla^2_{xx} \mathcal{H}_{t|T} + \nabla^2_{xp} \mathcal{H}_{t|T} \cdot \nabla^2 A_{t|T} + \nabla^2 A_{t|T} \cdot \nabla_{px} \mathcal{H}_{t|T} + \nabla^2 A_{t|T} \cdot \nabla^2_{pp} \mathcal{H}_{t|T} \cdot \nabla^2 A_{t|T}$$

and identifying $\nabla^2 A_{t|T} \to \boldsymbol{R}_{t|T}$. Making these substitutions and rearranging terms, Eqn. (2.17) becomes

$$\begin{aligned}
\dot{\boldsymbol{p}}_{t|T} &= \nabla^2 A_{t|T} \cdot \dot{\boldsymbol{x}}_{t|T} - \nabla_x \mathcal{H}_{t|T} - \nabla^2 A_{t|T} \cdot \nabla_p \mathcal{H}_{t|T} \\
&= \nabla^2 A_{t|T} \cdot \dot{\boldsymbol{x}}_{t|T} - \nabla_x \mathcal{H}_t(\boldsymbol{x}_{t|T}, \nabla A(\boldsymbol{x}_{t|T})) \\
&= \nabla^2 A_{t|T} \cdot \dot{\boldsymbol{x}}_{t|T} + \partial_t \big(\nabla A_{t|T}\big) \\
&= \frac{d}{dt}\big(\nabla A_{t|T}\big).
\end{aligned}$$

Thus, in this context the optimality condition Eqn. (2.17) reduces to the slightly tautological statement

$$\dot{\boldsymbol{p}}_{t|T} - \frac{d}{dt}\big(\nabla A_{t|T}\big) = 0. \tag{2.19}$$

Furthermore, given the earlier remark that $\boldsymbol{p}_{t|t} = \dot{\boldsymbol{p}}_{t|t} = 0$, the Kalman filter can be derived directly from the relation ([131])

$$\frac{d}{dt}\big(\nabla A_{t|t}\big) = 0.$$

The result,

$$\dot{\boldsymbol{x}}_{t|t} = -[\nabla^2 A_{t|t}]^{-1} \cdot \nabla_x\big(\partial_t A_{t|t}\big) = [\nabla^2 A_{t|t}]^{-1} \cdot \nabla_x \mathcal{H}(\boldsymbol{x}_{t|t}, \nabla A_{t|t}), \tag{2.20}$$

bears strong resemblance to continuous formulation Newton's method. Moreover, identifying $\boldsymbol{p}_{t|t} \to \nabla A_{t|t}$ as the gradient of an objective function the assumptions $\boldsymbol{p}_{t|t} = \dot{\boldsymbol{p}}_{t|t} = 0$ may be interpreted geometrically as assuming optimality of the estimate $\boldsymbol{x}_{t|t}$ by enforcing $\nabla A_{t|t} = \mathbf{0}$ everywhere along the forward pass.

Intuitively, this makes sense if one considers a locally quadratic expansion of the action around the filtered estimate $\boldsymbol{x}_{t|t}$. These constraints specify that $\boldsymbol{x}_{t|t}$ should always lie at a local minimum $\nabla A_t = \mathbf{0}$, thereby enforcing the estimate to be 'optimal' and unbiased at all times. It is tantamount to treating each additional measurement as its own two-point boundary value problem that expands the current observation horizon, and is the geometric manifestation of the statistical property that $\boldsymbol{x}_{t|t}$ is the best linear unbiased estimate given $\boldsymbol{y}_{t'<t}$. It is closely tied to the concept of sequential quadratic programming, and carries a uniquely recursive feedback structure, which is even more explicit in discrete time.

## 2.4 Discrete time solutions

The analogous discrete time formulas can be derived in a number of ways. Many of these derivations however rely heavily on statistical properties of the problem, such as orthogonality of innovations and linearity of error covariances. Although insightful, these techniques do not always generalize to the nonlinear case, and can obscure structural and geometric features of the solution.

This section pursues several derivations of discrete time Kalman filtering and smoothing, from a deterministic point of view that stems from its interpretation as a least squares minimization problem. Results from the previous section will be revisited, using discrete time versions of the variational principle ([119]) and Hamilton-Jacobi theory ([145]). And connections to the Gauss-Newton method will be made explicit ([14, 77]).

### 2.4.1 The one-step action

From a deterministic perspective as a least squares estimation, perhaps the most straightforward derivation of the Kalman filter involves rewriting Eqn. (2.6) recursively, as a pair of one-step actions for the measurement

$$A_{n|n}(\boldsymbol{x}) = A_{n|n-1}(\boldsymbol{x}_{n|n-1}, \boldsymbol{x}) + \frac{1}{2}|\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x})|^2_{\boldsymbol{R}_m}$$

and time updates respectively

$$A_{n+1|n}(\boldsymbol{x}_{n+1|n}, \boldsymbol{x}) = A_{n|n}(\boldsymbol{x}) + \frac{1}{2}|\boldsymbol{x}_{n+1|n} - \boldsymbol{F}(\boldsymbol{x})|^2_{\boldsymbol{R}_f}.$$

The filtered solution may be derived by minimizing these objective functions as follows. For the measurement update, let

$$A_{n|n-1}(\boldsymbol{x}_{n|n-1}, \boldsymbol{x}) \approx |\delta\boldsymbol{x}_{n|n-1}|^2_{\boldsymbol{R}_{n|n-1}}$$

with $\delta\boldsymbol{x}_{n|n-1} := \boldsymbol{x} - \boldsymbol{x}_{n|n-1}$, and expand $\boldsymbol{h}(\boldsymbol{x}) \to \boldsymbol{h}(\boldsymbol{x}_{n|n-1}) + \nabla\boldsymbol{h} \cdot \delta\boldsymbol{x}_{n|n-1}$. Setting $\nabla A_{n|n} = \boldsymbol{0}$ and solving for $\delta\boldsymbol{x}_{n|n-1}$ gives

$$\delta\boldsymbol{x}_{n|n-1} = \boldsymbol{R}^{-1}_{n|n} \cdot \nabla^\dagger \boldsymbol{h}_{n|n-1} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_n - \boldsymbol{h}_{n|n-1}).$$

The covariance

$$\boldsymbol{R}_{n|n} = \boldsymbol{R}_{n|n-1} + \nabla^\dagger \boldsymbol{h}_{n|n-1} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_{n|n-1}$$

is obtained by separation of variables to complete the square so that

$$A_{n|n} \approx |\boldsymbol{x} - \boldsymbol{x}_{n|n}|^2_{\boldsymbol{R}_{n|n}},$$

where approximation is valid to first order, up to an additive constant.

Likewise, for the time update the gradient is

$$\nabla A_{n+1|n} = \begin{bmatrix} \boldsymbol{R}_{n|n-1} \cdot (\boldsymbol{x} - \boldsymbol{x}_{n|n}) - \nabla^\dagger \boldsymbol{F} \cdot \boldsymbol{R}_f \cdot \big(\boldsymbol{x}_{n+1|n} - \boldsymbol{F}(\boldsymbol{x})\big) \\ \boldsymbol{R}_f \cdot \big(\boldsymbol{x}_{n+1|n} - \boldsymbol{F}(\boldsymbol{x})\big). \end{bmatrix}$$

Taking $\boldsymbol{x} \to \boldsymbol{x}_{n|n}$ and $\boldsymbol{x}_{n+1|n} \to \boldsymbol{F}(\boldsymbol{x})$ gives $\nabla A_{n+1|n} = \boldsymbol{0}$. The covariance update can be obtained by finding $\boldsymbol{R}_{n+1|n}$ such that

$$A_{n+1|n}(\boldsymbol{x}_{n+1|n}, \boldsymbol{x}) \approx |\boldsymbol{x} - \boldsymbol{x}_{n+1|n}|^2_{\boldsymbol{R}_{n+1|n}}.$$

This may be accomplished by block diagonalizing the Hessian

$$\nabla^2 A_{n+1|n} = \begin{bmatrix} \boldsymbol{R}_{n|n-1} + \nabla^\dagger \boldsymbol{F}_{n|n} \cdot \boldsymbol{R}_f \cdot \nabla \boldsymbol{F}_{n|n} & -\nabla^\dagger \boldsymbol{F}_{n|n} \cdot \boldsymbol{R}_f \\ -\boldsymbol{R}_f \cdot \nabla \boldsymbol{F}_{n|n} & \boldsymbol{R}_f \end{bmatrix},$$

by multiplying on the left and the right by the Schur 'transition' matrices

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{B} \cdot \boldsymbol{A}^{-1} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}^\dagger \\ \boldsymbol{B} & \boldsymbol{C} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{B} \cdot \boldsymbol{A}^{-1} & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} - \boldsymbol{B} \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{B}^\dagger \end{bmatrix} \tag{2.21}$$

Applying the Sherman-Morrison-Woodbury matrix identity ([209]),[3] gives the desired result:

$$\begin{aligned} \text{Meas. update:} \quad & \begin{cases} \boldsymbol{x}_{n|n} = \boldsymbol{R}_{n|n}^{-1} \cdot \nabla^\dagger \boldsymbol{h}_{n|n-1} \cdot \boldsymbol{R}_m \cdot \big(\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_{n|n-1})\big) \\ \boldsymbol{R}_{n|n} = \boldsymbol{R}_{n|n-1} + \nabla^\dagger \boldsymbol{h}_{n|n-1} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_{n|n-1} \end{cases} \\ \text{Time update:} \quad & \begin{cases} \boldsymbol{x}_{n+1|n} = \boldsymbol{F}(\boldsymbol{x}_{n|n}) \\ \boldsymbol{R}_{n+1|n}^{-1} = \nabla \boldsymbol{F}_{n|n} \cdot \boldsymbol{R}_{n|n}^{-1} \cdot \nabla^\dagger \boldsymbol{F}_{n|n} + \boldsymbol{R}_f^{-1} \end{cases} \end{aligned} \tag{2.22}$$

### 2.4.2 Two-pass smoothing, Newton's method, and canonical structure

The above result has close ties to Newton's method. For instance, [14] showed the measurement update can be derived from the Gauss-Newton method. And [77] later gave a similar result for the time update. Smoothing results were also given in [13, 77, 10]. In particular, [10] show that the forward and backward passes arise from block diagonalization of the Hessian $\nabla^2 A$, whose block tridiagonal structure stems from its formulation as a discrete-time path integral.

These connections are now examined from the Hamiltonian point of view, by using a variant of Newton's method to directly minimize Eqn. (2.9) in canonical coordinates $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{p}\}$. Tt each iteration $i$ the system

$$\nabla^2 \mathcal{H}^{(i)} \cdot (\mathcal{Z}^{(i+1)} - \mathcal{Z}^{(i)}) = -\nabla \mathcal{H}^{(i)}$$

[3] $(\boldsymbol{A} + \boldsymbol{U} \cdot \boldsymbol{C} \cdot \boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} \cdot \boldsymbol{U} \cdot (\boldsymbol{C}^{-1} + \boldsymbol{V} \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{U})^{-1} \cdot \boldsymbol{V} \cdot \boldsymbol{A}^{-1}$

is solved. For $N = 2$:

$$\nabla^2 \mathcal{H}^{(i)} := \begin{bmatrix} \boldsymbol{R}_{0|0} & -\nabla^\dagger \boldsymbol{F}_0^{(i)} & & & \\ -\nabla \boldsymbol{F}_0^{(i)} & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} & & \\ & \boldsymbol{I} & \nabla^\dagger \boldsymbol{h}_1^{(i)} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_1^{(i)} & -\nabla^\dagger \boldsymbol{F}_1^{(i)} & \\ & & -\nabla \boldsymbol{F}_1^{(i)} & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} \\ & & & \boldsymbol{I} & \nabla^\dagger \boldsymbol{h}_2^{(i)} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_2^{(i)} \end{bmatrix}$$

$$\nabla \mathcal{H}^{(i)} := \begin{bmatrix} \boldsymbol{R}_0 \cdot (\boldsymbol{x}_0^{(i)} - \mu_0) - \nabla_x \mathcal{H}_0^{(i)} \\ \boldsymbol{x}_1^{(i)} - \nabla_p \mathcal{H}_0^{(i)} \\ \boldsymbol{p}_1^{(i)} - \nabla_x \mathcal{H}_1^{(i)} \\ \boldsymbol{x}_2^{(i)} - \nabla_p \mathcal{H}_1^{(i)} \\ \boldsymbol{p}_2^{(i)} - \nabla^\dagger \boldsymbol{h}^{(i)} \cdot \boldsymbol{R}_m \cdot \left( \boldsymbol{y}_n - \boldsymbol{h}_n^{(i)} \right) \end{bmatrix} \qquad \mathcal{Z}^{(i+1)} - \mathcal{Z}^{(i)} := \begin{bmatrix} \boldsymbol{x}_0^{(i+1)} - \boldsymbol{x}_0^{(i)} \\ \boldsymbol{p}_1^{(i+1)} - \boldsymbol{p}_1^{(i)} \\ \boldsymbol{x}_1^{(i+1)} - \boldsymbol{x}_1^{(i)} \\ \boldsymbol{p}_2^{(i+1)} - \boldsymbol{p}_2^{(i)} \\ \boldsymbol{x}_2^{(i+1)} - \boldsymbol{x}_2^{(i)} \end{bmatrix},$$

with $\mathcal{H}_n^{(i)} := \mathcal{H}_n(\boldsymbol{x}_n^{(i)}, \boldsymbol{p}_{n+1}^{(i)})$. Note that this is essentially Newton's method, without the second order derivative tensors $\nabla^2 \boldsymbol{h}$, $\nabla^2 \boldsymbol{F}$. Alternatively, it may be viewed as a modification of the Gauss-Newton method to handle the 'non-square' terms $\langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n) \rangle$.

Following [10], this system may be solved in two passes, by using the discrete time transition operators given above in Eqn. (2.21) to block diagonalize the Hessian $\nabla^2 \mathcal{H}$. On the forward pass, the lower diagonal blocks are eliminated in a process that reproduces the Kalman filter solution. The resulting linear system is block upper-diagonal, and may be solved using back substitution in the backwards pass.

In continuous time, the extended two-pass smoother was derived by linearizing the action around the around the EKF trajectory. However, there are problems with this approach in discrete time. Notably, the terms $\nabla^\dagger \boldsymbol{h}_n^{(i)} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_n^{(i)}$ and $\nabla \boldsymbol{F}_n^{(i)}$ must be evaluated at distinct points: $\boldsymbol{x}_{n|n-1}$ and $\boldsymbol{x}_{n|n}$ respectively. Also, the continuous time implies the initial conditions should be chosen such that $\boldsymbol{p}_n^{(i)} = \boldsymbol{0}$, and $\boldsymbol{x}_n^{(i)}$ coincides with the EKF solution. But this does not produce the correct result.

To reproduce Eqn. (2.22) in the forward pass, the algorithm needs to run *sequentially*, evaluating the Hamiltonian at different points for the time and measurement updates. Although this might appear problematic, since only $\boldsymbol{x}_0^{(0)} = \boldsymbol{\mu}_0$ is known initially, the recursive structure of the Kalman solution allows each $n > 0$ term to be determined in succession. Since it is not immediately clear how to modify Newton's method to achieve this result, the linear case is treated instead. With that restriction, these issues no longer apply.

The correct result may be obtained by choosing $\mathcal{Z}^{(i)} \to \mathbf{0}$ and associating $\boldsymbol{x}_n^{(i+1)} \to \boldsymbol{x}_{n|N}$, $\boldsymbol{p}_n^{(i+1)} \to \boldsymbol{p}_{n|N}$. With these specifications the linear system becomes,

$$\begin{bmatrix} \boldsymbol{R}_{0|0} & -\boldsymbol{F}_0^\dagger & & & & \\ -\boldsymbol{F}_0 & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} & & & \\ & \boldsymbol{I} & \boldsymbol{H}_1^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_1 & -\boldsymbol{F}_1^\dagger & & \\ & & -\boldsymbol{F}_1 & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} & \\ & & & \boldsymbol{I} & \boldsymbol{H}_2^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_2 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{0|2} \\ \boldsymbol{p}_{1|2} \\ \boldsymbol{x}_{1|2} \\ \boldsymbol{p}_{2|2} \\ \boldsymbol{x}_{2|2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{0|0} \cdot \boldsymbol{x}_{0|0} \\ \mathbf{0} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_1 \\ \mathbf{0} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_2 \end{bmatrix}.$$

Using the information identity

$$\boldsymbol{R}_{n|n} \cdot \boldsymbol{x}_{n|n} = \boldsymbol{R}_{n|n-1} \cdot \boldsymbol{x}_{n|n-1} + \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_n,$$

and the fact that $\boldsymbol{x}_{n+1|n} = \boldsymbol{F}_n \cdot \boldsymbol{x}_{n|n}$, the forward pass reproduces the Kalman filter:

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{R}_{0|0}^{-1} \cdot \boldsymbol{F}_0^\dagger & & & & \\ \mathbf{0} & -\boldsymbol{R}_{1|0}^{-1} & \boldsymbol{I} & & & \\ & \boldsymbol{I} & \boldsymbol{H}_{1|0}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_1 & -\boldsymbol{F}_1^\dagger & & \\ & & -\boldsymbol{F}_{1|1} & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} & \\ & & & \boldsymbol{I} & \boldsymbol{H}_2^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_2 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{0|2} \\ \boldsymbol{p}_{1|2} \\ \boldsymbol{x}_{1|2} \\ \boldsymbol{p}_{2|2} \\ \boldsymbol{x}_{2|2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{0|0} \\ \boldsymbol{x}_{1|0} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_1 \\ \mathbf{0} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_2 \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{R}_{0|0}^{-1} \cdot \boldsymbol{F}_0^\dagger & & & & \\ \mathbf{0} & -\boldsymbol{R}_{1|0}^{-1} & \boldsymbol{I} & & & \\ & \mathbf{0} & \boldsymbol{I} & -\boldsymbol{R}_{1|1}^{-1} \cdot \boldsymbol{F}_1^\dagger & & \\ & & -\boldsymbol{F}_1 & -\boldsymbol{R}_f^{-1} & \boldsymbol{I} & \\ & & & \boldsymbol{I} & \boldsymbol{H}_2^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_2 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{0|2} \\ \boldsymbol{p}_{1|2} \\ \boldsymbol{x}_{1|2} \\ \boldsymbol{p}_{2|2} \\ \boldsymbol{x}_{2|2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{0|0} \\ \boldsymbol{x}_{1|0} \\ \boldsymbol{x}_{1|1} \\ \mathbf{0} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_2 \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{R}_{0|0}^{-1} \cdot \boldsymbol{F}_0^\dagger & & & & \\ \mathbf{0} & -\boldsymbol{R}_{1|0}^{-1} & \boldsymbol{I} & & & \\ & \mathbf{0} & \boldsymbol{I} & -\boldsymbol{R}_{1|1}^{-1} \cdot \boldsymbol{F}_1^\dagger & & \\ & & \mathbf{0} & -\boldsymbol{R}_{2|1}^{-1} & \boldsymbol{I} & \\ & & & \boldsymbol{I} & \boldsymbol{H}_2^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{H}_2 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{0|2} \\ \boldsymbol{p}_{1|2} \\ \boldsymbol{x}_{1|2} \\ \boldsymbol{p}_{2|2} \\ \boldsymbol{x}_{2|2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{0|0} \\ \boldsymbol{x}_{1|0} \\ \boldsymbol{x}_{1|1} \\ \boldsymbol{x}_{2|1} \\ \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot \boldsymbol{y}_2 \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{R}_{0|0}^{-1}\cdot\boldsymbol{F}_0^\dagger & & & \\ \boldsymbol{0} & -\boldsymbol{R}_{1|0}^{-1} & \boldsymbol{I} & & \\ & \boldsymbol{0} & \boldsymbol{I} & -\boldsymbol{R}_{1|1}^{-1}\cdot\boldsymbol{F}_1^\dagger & \\ & & \boldsymbol{0} & -\boldsymbol{R}_{2|1}^{-1} & \boldsymbol{I} \\ & & & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x}_{0|2} \\ \boldsymbol{p}_{1|2} \\ \boldsymbol{x}_{1|2} \\ \boldsymbol{p}_{2|2} \\ \boldsymbol{x}_{2|2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{0|0} \\ \boldsymbol{x}_{1|0} \\ \boldsymbol{x}_{1|1} \\ \boldsymbol{x}_{2|1} \\ \boldsymbol{x}_{2|2} \end{bmatrix}$$

Setting $\boldsymbol{p}_{N+1|N} = 0$, the backward pass is described by the recursion relations

$$\begin{aligned} \boldsymbol{x}_{n|N} &= \boldsymbol{x}_{n|n} + \boldsymbol{R}_{n|n}^{-1}\cdot\boldsymbol{F}_n^\dagger\cdot\boldsymbol{p}_{n+1|N} \\ \boldsymbol{p}_{n|N} &= \boldsymbol{R}_{n|n-1}\cdot(\boldsymbol{x}_{n|N} - \boldsymbol{x}_{n|n-1}). \end{aligned}$$

Eliminating the filtered solutions reproduces the discrete time Hamilton equations from Eqn. (2.10)

$$\begin{aligned} \boldsymbol{x}_{n+1|N} &= \boldsymbol{F}_n\cdot\boldsymbol{x}_{n|N} + \boldsymbol{R}_f^{-1}\cdot\boldsymbol{p}_{n+1|N} \\ \boldsymbol{p}_{n|N} &= \boldsymbol{F}_n^\dagger\cdot\boldsymbol{p}_{n+1|N} + \boldsymbol{H}_n^\dagger\cdot\boldsymbol{R}_m\cdot(\boldsymbol{y}_n - \boldsymbol{H}_n\cdot\boldsymbol{x}_{n|N}). \end{aligned}$$

And eliminating $\boldsymbol{p}_{n|N}$ produces the discrete time analog of the Rauch-Tung-Striebel smoother Eqn. (2.15),

$$\boldsymbol{x}_{n|N} - \boldsymbol{x}_{n|n-1} = \boldsymbol{F}_n^{-1}\cdot(\boldsymbol{I} - \boldsymbol{R}_f^{-1}\cdot\boldsymbol{R}_{n+1|n})\cdot(\boldsymbol{x}_{n+1|N} - \boldsymbol{x}_{n+1|n}).$$

It is also interesting to note how the two-pass solution impacts the computational complexity of the problem. Whereas a typical implementation of Newton's method would require inverting a sparse $ND\times ND$ matrix, the block diagonalization procedure reduces this to $2N$ inversions of $D\times D$ matrices. As the computational complexity of inverting a $D\times D$ matrix is roughly $O(D^3)$, the two-pass algorithm reduces the computational overhead by roughly $O(N^2)$ by exploiting the block tridiagonal structure of the discrete-time path integral.

### 2.4.3 The second variation in discrete time

Starting from Eqn. (2.8) the derivation proceeds similarly to the continuous time solution given in Sec. (2.3.2). Expanding the action to second order gives

$$\begin{aligned} \delta A = \langle \delta\boldsymbol{x}_0, \boldsymbol{R}_0\cdot(\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0}) - \boldsymbol{p}_{0|T} - \nabla_x\mathcal{H}_{0|T}\rangle + \langle \delta\boldsymbol{x}_N, \boldsymbol{p}_{n|N} - \nabla_x\mathcal{H}_{n|N}\rangle + \\ + \sum_{n=0}^{N-1}\left\langle \begin{bmatrix} \delta\boldsymbol{x}_n \\ \delta\boldsymbol{p}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{p}_{n|N} \\ \boldsymbol{x}_{n+1|N} \end{bmatrix} - \begin{bmatrix} \nabla_x\mathcal{H}_{n|N} \\ \nabla_p\mathcal{H}_{n|N} \end{bmatrix} \right\rangle. \end{aligned}$$

$$\delta^2 A = \langle \delta\boldsymbol{x}_0, (\boldsymbol{R}_0 - \nabla_{xx}\mathcal{H}_{0|T}) - \nabla_{xp}\mathcal{H}_{0|T}\cdot\delta\boldsymbol{p}_1 - \delta\boldsymbol{p}_0)\rangle + \langle \delta\boldsymbol{x}_N, \delta\boldsymbol{p}_N - \nabla_{xx}\mathcal{H}_{n|N}\rangle$$
$$+ \sum_{n=0}^{N-1} \left\langle \begin{bmatrix} \delta\boldsymbol{x}_n \\ \delta\boldsymbol{p}_{n+1} \end{bmatrix}, \begin{bmatrix} \delta\boldsymbol{p}_n \\ \delta\boldsymbol{x}_{n+1} \end{bmatrix} - \begin{bmatrix} \nabla^2_{xx}\mathcal{H}_{n|N} & \nabla^2_{xp}\mathcal{H}_{n|N} \\ \nabla^2_{px}\mathcal{H}_{n|N} & \nabla^2_{pp}\mathcal{H}_{n|N} \end{bmatrix} \cdot \begin{bmatrix} \delta\boldsymbol{x}_n \\ \delta\boldsymbol{p}_{n+1} \end{bmatrix} \right\rangle.$$

These expressions have been resummed so that the time indices of the derivatives of the discrete Hamiltonian are consistent,

$$\sum_{n=0}^{N-1} \left\langle \begin{bmatrix} \delta\boldsymbol{x}_{n+1} \\ \delta\boldsymbol{p}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{p}_{n+1|N} \\ \boldsymbol{x}_{n+1|N} \end{bmatrix} - \begin{bmatrix} \nabla_x\mathcal{H}_{n+1|N} \\ \nabla_p\mathcal{H}_{n|N} \end{bmatrix} \right\rangle - \langle \delta\boldsymbol{x}_0, \nabla_x\mathcal{H}_{0|T}\rangle$$
$$\to \sum_{n=0}^{N-1} \left\langle \begin{bmatrix} \delta\boldsymbol{x}_n \\ \delta\boldsymbol{p}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{p}_{n|N} \\ \boldsymbol{x}_{n+1|N} \end{bmatrix} - \begin{bmatrix} \nabla_x\mathcal{H}_{n|N} \\ \nabla_p\mathcal{H}_{n|N} \end{bmatrix} \right\rangle - \langle \delta\boldsymbol{x}_0, \boldsymbol{p}_{0|T}\rangle - \langle \delta\boldsymbol{p}_N, \boldsymbol{x}_{n|N}\rangle.$$

This is the discrete time version of integration by parts [110]. Note that in contrast to the analogous expressions in continuous time, these cannot be written in compact notation involving $\boldsymbol{z}_n := \{\boldsymbol{x}_n, \boldsymbol{p}_n\}$ and $\boldsymbol{J}$, due to the mixed time indexing. Also, without this regrouping of terms the following derivation does not produce the correct result for the covariance update.

Evaluating the second variation $\delta^2 A$ along the filtered path, and introducing the discrete Riccati transformation $\delta\boldsymbol{p}_n = \boldsymbol{R}_{n|n}\cdot\delta\boldsymbol{x}_n + \delta\boldsymbol{r}_n$ into gives (for arbitrary $n$ inside the sum),

$$\left| \begin{matrix} \delta\boldsymbol{r}_n \\ \delta\boldsymbol{x}_n \\ \delta\boldsymbol{r}_{n+1} \\ \boldsymbol{R}_{n+1|n+1}\cdot\delta\boldsymbol{x}_{n+1} \end{matrix} \right|^2_{\nabla^2\mathcal{H}_n}$$

with

$$\nabla^2\mathcal{H}_n = \begin{bmatrix} \boldsymbol{0} & \frac{1}{2}\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \frac{1}{2}\boldsymbol{I} & \boldsymbol{R}_{n|n} - \nabla^2_{xx}\mathcal{H}_{n|n} & -\nabla^2_{xp}\mathcal{H}_{n|n} & -\nabla^2_{xp}\mathcal{H}_{n|n} \\ \boldsymbol{0} & -\nabla^2_{px}\mathcal{H}_{n|n} & -\nabla^2_{pp}\mathcal{H}_{n|n} & \frac{1}{2}\boldsymbol{R}^{-1}_{n+1|n+1} - \nabla^2_{pp}\mathcal{H}_{n|n} \\ \boldsymbol{0} & -\nabla^2_{px}\mathcal{H}_{n|n} & \frac{1}{2}\boldsymbol{R}^{-1}_{n+1|n+1} - \nabla^2_{pp}\mathcal{H}_{n|n} & \boldsymbol{R}^{-1}_{n+1|n+1} - \nabla^2_{pp}\mathcal{H}_{n|n} \end{bmatrix}$$

The factors of $1/2$ come from using the symmetric form

$$\langle \delta\boldsymbol{x}_n, \delta\boldsymbol{p}_{n+1}\rangle + \langle \delta\boldsymbol{p}_n, \delta\boldsymbol{x}_{n+1}\rangle \to \frac{1}{2}\Big(\langle \delta\boldsymbol{x}_n, \delta\boldsymbol{p}_{n+1}\rangle + \langle \delta\boldsymbol{p}_{n+1}, \delta\boldsymbol{x}_n\rangle + \langle \delta\boldsymbol{p}_n, \delta\boldsymbol{x}_{n+1}\rangle + \langle \delta\boldsymbol{x}_{n+1}, \delta\boldsymbol{p}_n\rangle\Big).$$

Restricting $\delta \boldsymbol{r}_n = \delta \boldsymbol{r}_{n+1} = 0$ simplifies the above expression

$$\left\langle \begin{bmatrix} \delta \boldsymbol{x}_n \\ \boldsymbol{R}_{n+1|n+1} \cdot \delta \boldsymbol{x}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{R}_{n|n} - \nabla^2_{xx} \mathcal{H}_{n|n} & -\nabla^2_{xp} \mathcal{H}_{n|n} \\ -\nabla^2_{px} \mathcal{H}_{n|n} & \boldsymbol{R}^{-1}_{n+1|n+1} - \nabla^2_{pp} \mathcal{H}_{n|n} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_n \\ \boldsymbol{R}_{n+1|n+1} \cdot \delta \boldsymbol{x}_{n+1} \end{bmatrix} \right\rangle.$$

Setting the derivatives of this term with respect to $\delta \boldsymbol{x}_n$, $\delta \boldsymbol{x}_{n+1}$ equal to zero gives a linear system, which can be solved for $\boldsymbol{R}_{n+1|n+1}$

$$\boldsymbol{R}^{-1}_{n+1|n+1} = \nabla^2_{px} \mathcal{H}_{n|n} \cdot \left( \boldsymbol{R}_{n|n} - \nabla_{xx} \mathcal{H}_{n|n} \right)^{-1} \cdot \nabla^2_{xp} \mathcal{H}_{n|n} + \nabla^2_{pp} \mathcal{H}_{n|n}. \tag{2.23}$$

This is the discrete time Riccati equation for the propagation of the approximate error covariance with $\boldsymbol{R}_{n+1|n+1} \to \boldsymbol{R}_{n+1|n}$

$$\begin{aligned} \boldsymbol{R}^{-1}_{n+1|n} &= \nabla \boldsymbol{F}_{n|n} \cdot \boldsymbol{R}^{-1}_{n|n} \cdot \nabla^\dagger \boldsymbol{F}_{n|n} + \boldsymbol{R}^{-1}_f \\ \boldsymbol{R}_{n|n} &= \boldsymbol{R}_{n|n-1} + \nabla^\dagger \boldsymbol{h}_{n|n} \cdot \boldsymbol{R}_m \cdot \nabla \boldsymbol{h}_{n|n}. \end{aligned}$$

Likewise, expanding $\delta A$ to first order around the filtering solution gives the following expression inside the sum

$$\left\langle \begin{bmatrix} \delta \boldsymbol{x}_n \\ \delta \boldsymbol{p}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{p}_{n|n} + \delta \boldsymbol{p}_n \\ \boldsymbol{x}_{n+1|n+1} + \delta \boldsymbol{x}_{n+1} \end{bmatrix} - \begin{bmatrix} \nabla_x \mathcal{H}_{n|n} \\ \nabla_p \mathcal{H}_{n|n} \end{bmatrix} + \begin{bmatrix} \nabla^2_{xx} \mathcal{H}_{n|n} & \nabla^2_{xp} \mathcal{H}_{n|n} \\ \nabla^2_{px} \mathcal{H}_{n|n} & \nabla^2_{pp} \mathcal{H}_{n|n} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_n \\ \delta \boldsymbol{p}_{n+1} \end{bmatrix} \right\rangle.$$

The Riccati transformation enforces Hamilton's equations on the variations $\delta \boldsymbol{x}_n,\ \delta \boldsymbol{p}_{n+1}$ simplifying this expression

$$\left\langle \begin{bmatrix} \delta \boldsymbol{x}_n \\ \boldsymbol{R}_{n+1|n+1} \cdot \delta \boldsymbol{x}_{n+1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{p}_{n|n} \\ \boldsymbol{x}_{n+1|n+1} \end{bmatrix} - \begin{bmatrix} \nabla_x \mathcal{H}_{n|n} \\ \nabla_p \mathcal{H}_{n|n} \end{bmatrix} \right\rangle.$$

Setting equal to zero its derivatives with respect to $\delta \boldsymbol{x}_n$, $\delta \boldsymbol{x}_{n+1}$ gives the discrete analog of Eqn. (2.17)

$$\boldsymbol{p}_{n|n} - \nabla_x \mathcal{H}_{n|n} = \boldsymbol{R}_{n+1|n+1} \cdot (\boldsymbol{x}_{n+1|n+1} - \nabla_p \mathcal{H}_{n|n}) = \boldsymbol{0}. \tag{2.24}$$

Substituting for $\nabla \mathcal{H}_{n|n}$ and rearranging terms gives

$$\begin{aligned} \boldsymbol{x}_{n+1|n+1} = \boldsymbol{F}(\boldsymbol{x}_{n|n}) &+ \boldsymbol{R}^{-1}_{n+1|n+1} \cdot \left( \boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_{n|n}) \right) \\ &- \boldsymbol{R}^{-1}_{n+1|n+1} \cdot \left( \boldsymbol{p}_{n|n} - (\nabla^\dagger \boldsymbol{F}_{n|n} + \boldsymbol{R}_{n+1|n+1} \cdot \boldsymbol{R}^{-1}_f) \cdot \boldsymbol{p}_{n+1|n+1} \right) \end{aligned}$$

Choosing to enforce these expressions independently gives a discrete analog of Eqn. (2.18)

$$\begin{aligned} \boldsymbol{x}_{n+1|n+1} &= \boldsymbol{F}(\boldsymbol{x}_{n|n}) + \boldsymbol{R}^{-1}_{n+1|n+1} \cdot \big(\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_{n|n})\big) \\ \boldsymbol{p}_{n|n} &= \big(\nabla^\dagger \boldsymbol{F}_{n|n} + \boldsymbol{R}_{n+1|n+1} \cdot \boldsymbol{R}_f^{-1}\big) \cdot \boldsymbol{p}_{n+1|n+1}. \end{aligned} \tag{2.25}$$

Note, as in the continuous case, homogeneity of the $\boldsymbol{p}_{n|n}$ equation together with the boundary conditions $\boldsymbol{p}_{0|0} = \boldsymbol{p}_{N+1|N} = 0$ implies that $\boldsymbol{p}_{n|n}$ equals zero for all $n$.

The filtered estimate for $\boldsymbol{x}_{n|n}$ is similar to the discrete time extended Kalman filter, with a few crucial differences. First, the forecast model operates on the result of the previous time update $\boldsymbol{x}_{n|n} = \boldsymbol{x}_{n|n-1}$, instead of on the analysis $\boldsymbol{x}_{n|n}$, which includes the measurement $\boldsymbol{y}_n$. The covariance is also evaluated at $\boldsymbol{R}_{n+1|n+1} = \boldsymbol{R}_{n+1|n}$, instead of at $\boldsymbol{R}_{n|n}$. These differences are important, as this result does not provide recursive separation between the measurement and time updates.

### 2.4.4 The discrete Hamilton-Jacobi solution

While it is not immediately apparent how to modify the previous derivation to obtain the desired solution, looking at the problem from the point of view of discrete Hamilton-Jacobi theory ([145]) provides some clues. Defining the action as a function of $n$

$$A_{n'} \equiv A_{n'}(\mathcal{X}_{0:n'}, \mathcal{P}_{0:n'}) := \sum_{n=0}^{n'-1} \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{p}_{n+1}),$$

the sum can be eliminated through the following recursion

$$A_{n+1} - A_n = \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{p}_{n+1}).$$

Assuming $A_n(\boldsymbol{x}_n)$ (*i.e.*, a function of $\boldsymbol{x}_n$ alone) and substituting $\boldsymbol{p}_{n+1} \to \nabla A_{n+1}$ gives the discrete Hamilton-Jacobi equation,

$$A_{n+1} - A_n = \langle \nabla A_{n+1}, \boldsymbol{x}_{n+1} \rangle - \mathcal{H}_n(\boldsymbol{x}_n, \boldsymbol{p}_{n+1}). \tag{2.26}$$

As in continuous time, the Riccati equation can be derived from the Hessian of Eqn. (2.26) and ignoring any terms with derivatives higher than second order,

$$\begin{aligned} \nabla^2 A_n + \nabla^2 A_{n+1} = \nabla^2_{xx} \mathcal{H}_n + \nabla^2 A_{n+1} \cdot \nabla^2_{pp} \mathcal{H}_n \cdot \nabla^2 A_{n+1} \\ + \nabla^2_{xp} \mathcal{H}_n \cdot \nabla^2 A_{n+1} + \nabla^2 A_{n+1} \cdot \nabla^2_{px} \mathcal{H}_n. \end{aligned}$$

All expressions in this equation are evaluated around the filtered solution. Unlike the continuous case however, this derivation requires enforcing Hamilton's equation

$$\boldsymbol{x}_{n+1} = \nabla_p \mathcal{H}_n\big(\boldsymbol{x}_n, \nabla A_{n+1}(\boldsymbol{x}_n)\big) =: \boldsymbol{\Phi}(\boldsymbol{x}_n).$$

Thus,

$$\begin{aligned}\nabla^2 A_n - \nabla^2_{xx}\mathcal{H}_n = &-\nabla\boldsymbol{\Phi}_n \cdot \nabla^2 A_{n+1} \cdot \big([\nabla^2 A_{n+1}]^{-1} - \nabla^2_{pp}\mathcal{H}_n\big) \cdot \nabla^2 A_{n+1} \cdot \nabla\boldsymbol{\Phi}_n \\ &+ \nabla^2_{xp}\mathcal{H}_n \cdot \nabla^2 A_{n+1} \cdot \nabla\boldsymbol{\Phi}_n + \nabla\boldsymbol{\Phi}_n \cdot \nabla^2 A_{n+1} \cdot \nabla^2_{px}\mathcal{H}_n.\end{aligned}$$

Substituting for

$$\nabla\boldsymbol{\Phi}_n = \big(\boldsymbol{I} - \nabla^2_{pp}\mathcal{H}_n \cdot \nabla^2 A_{n+1}\big)^{-1} \cdot \nabla_{px}\mathcal{H}_n$$

gives

$$\begin{aligned}\nabla^2 A_n - \nabla^2_{xx}\mathcal{H}_n &= \nabla\boldsymbol{\Phi}_n \cdot \nabla^2 A_{n+1} \cdot \big([\nabla^2 A_{n+1}]^{-1} - \nabla^2_{pp}\mathcal{H}_n\big) \cdot \nabla^2 A_{n+1} \cdot \nabla\boldsymbol{\Phi}_n \\ &= \nabla\boldsymbol{\Phi}_n \cdot \nabla^2 A_{n+1} \cdot \nabla^2_{px}\mathcal{H}_n \\ &= \nabla^2_{xp}\mathcal{H}_n \cdot \nabla^2 A_{n+1} \cdot \big(\boldsymbol{I} - \nabla^2_{pp}\mathcal{H}_n \cdot \nabla^2 A_{n+1}\big)^{-1} \cdot \nabla^2_{px}\mathcal{H}_n \\ &= \nabla^2_{xp}\mathcal{H}_n \cdot \big([\nabla^2 A_{n+1}]^{-1} - \nabla^2_{pp}\mathcal{H}_n\big)^{-1} \cdot \nabla^2_{px}\mathcal{H}_n.\end{aligned}$$

Assuming $\nabla^2_{px}\mathcal{H}_n = \nabla\boldsymbol{F}_n$ is invertible, gives the discrete Riccati equation

$$[\nabla^2 A_{n+1}]^{-1} = \nabla^2_{px}\mathcal{H}_n \cdot (\nabla^2 A_n - \nabla^2_{xx}\mathcal{H}_n)^{-1} \cdot \nabla^2_{xp}\mathcal{H}_n + \nabla^2_{pp}\mathcal{H}_n.$$

This assumption is not necessary however, considering that it was not required in Eqn. (2.23).

Attempting now to perform a similar derivation of the Kalman filter, the gradient of the discrete Hamilton-Jacobi equation is

$$\nabla\boldsymbol{\Phi}_{n|N} \cdot \nabla^2 A_{n+1|N} \cdot \big(\boldsymbol{x}_{n+1|N} - \nabla_p \mathcal{H}_{n|N}\big) + \big(\nabla A_{n|N} - \nabla_x \mathcal{H}_{n|N}\big) = 0.$$

Substituting expressions for $\nabla\mathcal{H}_n$, expanding around the filtered path, and rearranging terms gives

$$\begin{aligned}\big([\nabla\boldsymbol{\Phi}_{n|n} \cdot \nabla^2 A_{n+1|n+1}] \cdot \boldsymbol{x}_{n+1|n+1} - \boldsymbol{F}(\boldsymbol{x}_{n|n}) - \nabla^\dagger \boldsymbol{h}_{n|n} \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_{n|n}))\big) \\ + \big(\nabla A_{n|n} - (\nabla^\dagger \boldsymbol{F}_{n|n} + [\nabla\boldsymbol{\Phi}_{n|n} \cdot \nabla^2 A_{n+1|n+1}] \cdot \boldsymbol{R}_f^{-1}) \cdot A_{n+1|n+1}\big) = 0.\end{aligned}$$

Enforcing these two expressions separately, and using

$$\begin{aligned}
\nabla\boldsymbol{\Phi}_{n|n}\cdot\nabla^2 A_{n+1|n+1} &= \nabla_{xp}\mathcal{H}_{n|n}\cdot\left(\boldsymbol{I}-\nabla^2 A_{n+1}\cdot\nabla^2_{pp}\mathcal{H}_n\right)^{-1}\cdot\nabla^2 A_{n+1|n+1}\\
&= \nabla_{xp}\mathcal{H}_{n|n}\cdot\left([\nabla^2 A_{n+1|n+1}]^{-1}-\nabla^2_{pp}\mathcal{H}_n\right)^{-1}\\
&= \left([\nabla^2 A_{n|n}]^{-1}-\nabla^2_{xx}\mathcal{H}_n\right)^{-1}\cdot[\nabla^2_{px}\mathcal{H}_{n|n}]^{-1}\\
&= \boldsymbol{R}_{n|n}\cdot[\nabla\boldsymbol{F}_{n|n}]^{-1}
\end{aligned}$$

gives

$$\begin{aligned}
\boldsymbol{x}_{n+1|n+1} &= \boldsymbol{F}(\boldsymbol{x}_{n|n})+\nabla\boldsymbol{F}_{n|n}\cdot\boldsymbol{R}^{-1}_{n|n}\cdot\nabla^{\dagger}\boldsymbol{h}_{n|n}\cdot\boldsymbol{R}_m\cdot(\boldsymbol{y}_n-\boldsymbol{h}(\boldsymbol{x}_{n|n}))\\
\boldsymbol{p}_{n|n} &= \left(\nabla^{\dagger}\boldsymbol{F}_{n|n}+\boldsymbol{R}_{n|n}\cdot[\nabla\boldsymbol{F}_{n|n}]^{-1}\cdot\boldsymbol{R}_f^{-1}\right)\cdot\boldsymbol{p}_{n+1|n+1}.
\end{aligned}$$

The top equation is again similar to discrete time Kalman filter. Compared with Eqn. (2.25) above, the additional factor $\nabla\boldsymbol{\Phi}_n$ pulls the covariance back in time $\boldsymbol{R}_{n+1|n+1}\rightarrow\boldsymbol{R}_{n|n}$, making this result closer to the desired solution. However, the measurement term is still propagated using the tangent linear model $\nabla\boldsymbol{F}_{n|n}$, instead of the full nonlinear model. So when the forecast model linear, this result is equivalent to the Kalman filter. But it is worth noting however, that linearity was not required in the continuous time derivation of the extended Kalman filter.

Once again, the boundary conditions $\boldsymbol{p}_{0|0}=\boldsymbol{p}_{N+1|N}=0$ imply $\boldsymbol{p}_n=0$ for all $n$. However, the momentum equation also contains an additional factor $[\nabla\boldsymbol{F}_{n|n}]^{-1}$, requiring invertibility of $\nabla\boldsymbol{F}_{n|n}$, which is not usually necessary.

## 2.5 Canonical transformations for two-point boundary value problems

In physics, solutions to the Hamilton-Jacobi equation are viewed as a special type of coordinate mapping known as canonical transformation. Canonical transformations are change of canonical coordinates $\{\boldsymbol{x}_t,\boldsymbol{p}_t\}\rightarrow\{\tilde{\boldsymbol{x}}_{t'},\tilde{\boldsymbol{p}}_{t'}\}$ that preserves the form of Hamilton's equations, and thereby satisfies the necessary conditions for a stationary action path. The freedom to construct such a transformation stems from the fact that choice of generalized coordinates is not unique. The Lagrangian, for instance, is invariant to the addition of a total derivative. This amounts to a change of gauge, which is itself a type of canonical transformation, although this point has been disputed ([176]).

Within the set of possible canonical transformations, the Hamilton-Jacobi solution is particularly

singular. It represents a change of coordinates in which the new Hamiltonian is zero, rendering the new coordinates (of this generally time-dependent transformation) stationary. Consequently, if the original Hamiltonian say describes the trajectory of a particle, the new coordinates are expressed in the frame of reference of the particle.

This type of canonical transformation may described rather succinctly, through the use of *canonical generating functions*. [4] In physics, this longstanding approach is typically used to derive solutions to Hamilton's equations with certain desired properties. Generating functions play a fundamental role in perturbation theory of nonlinear systems, and has been used by [110] to derive the discrete time formulation of Hamiltonian mechanics.

Recently however, the idea has been introduced to the optimal control community ([151, 152, 150]) as an analytical technique to construct feedback solutions to two-point boundary value problems, such as Eqn. (2.5). The method simplifies the treatment of the separated boundary conditions, and offers additional insight into the problem.

The estimation analog is quite similar, albeit slightly more general due to non-homogeneous forcing terms in the Hamiltonian. But it has not yet been presented in the literature. Therefore, a derivation will now be given to complement the techniques discussed above.

### 2.5.1 Generating functions of canonical transformations

The idea behind the generating function approach is to construct a transformation between canonical coordinates $\{\boldsymbol{x}_t, \boldsymbol{p}_t\} \to \{\tilde{\boldsymbol{x}}_{t'}, \tilde{\boldsymbol{p}}_{t'}\}$ at times $t$ and $t'$ that preserves the structure of the original Hamiltonian $\mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{p}_t)$ by adding a scalar function $G_t(\cdot,\cdot)$ to the transformed Hamiltonian $\tilde{\mathcal{H}}_{t'}(\tilde{\boldsymbol{x}}_{t'}, \tilde{\boldsymbol{p}}_{t'})$. Specifically, stationary paths of the action are invariant to the addition of a total time derivative

$$\delta\left[\int dt\, \langle \boldsymbol{p}_t, \dot{\boldsymbol{x}}_t\rangle - \mathcal{H}_t(\boldsymbol{x}_t, \boldsymbol{p}_t)\right] \equiv \delta\left[\int dt\, \frac{dt'}{dt}\left(\langle \tilde{\boldsymbol{p}}_{t'}, \dot{\tilde{\boldsymbol{x}}}_{t'}\rangle - \tilde{\mathcal{H}}_{t'}(\tilde{\boldsymbol{x}}_{t'}, \tilde{\boldsymbol{p}}_{t'})\right) + \frac{dG(\cdot,\cdot)}{dt}\right]$$

where $t$ has been chosen as the independent variable. The function $G_t(\cdot,\cdot)$ has two arguments that determine the type of generating function. One argument must be from the original coordinate and one from the transformed coordinate, and thus there are four possibilities:

$$G_t^{(1)}(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_{t'}),\ G_t^{(2)}(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'}),\ G_t^{(3)}(\boldsymbol{p}_t, \tilde{\boldsymbol{x}}_{t'}),\ G_t^{(4)}(\boldsymbol{p}_t, \tilde{\boldsymbol{p}}_{t'}).$$

[4] The term 'canonical' is used to distinguish from 'mathematical' generating functions, which have an entirely different meaning.

The choice of argument determines the differential relations that $G(\cdot,\cdot)$ must satisfy to ensure $\delta A = 0$. Namely, enforcing

$$\langle \boldsymbol{p}_t, d\boldsymbol{x}_t\rangle - \langle \tilde{\boldsymbol{p}}_{t'}, d\tilde{\boldsymbol{x}}_{t'}\rangle - \big(\mathcal{H}_t(\boldsymbol{x}_t,\boldsymbol{p}_t) - \tilde{\mathcal{H}}_{t'}(\tilde{\boldsymbol{x}}_{t'},\tilde{\boldsymbol{p}}_{t'})\big) = dG(\cdot,\cdot)$$

gives the following relations

| | | | |
|---|---|---|---|
| Type 1: | $\boldsymbol{p}_t = \nabla_x G_t^{(1)}$ | $-\tilde{\boldsymbol{p}}_{t'} = \nabla_{\tilde{\boldsymbol{x}}} G_t^{(1)}$ | $-\partial_t G_t^{(1)} = \mathcal{H}_t(\boldsymbol{x}_t, \nabla_x G_t^{(1)}) - \tilde{\mathcal{H}}_{t'}(\tilde{\boldsymbol{x}}_{t'}, -\nabla_{\tilde{\boldsymbol{x}}} G_t^{(1)})$ |
| Type 2: | $\boldsymbol{p}_t = \nabla_x G_t^{(2)}$ | $\tilde{\boldsymbol{x}}_{t'} = \nabla_{\tilde{\boldsymbol{p}}} G_t^{(2)}$ | $-\partial_t G_t^{(2)} = \mathcal{H}_t(\boldsymbol{x}_t, \nabla_x G_t^{(2)}) - \tilde{\mathcal{H}}_{t'}(\nabla_{\tilde{\boldsymbol{p}}} G_t^{(2)}, \tilde{\boldsymbol{p}}_{t'})$ |
| Type 3: | $-\boldsymbol{x}_t = \nabla_p G_t^{(3)}$ | $-\tilde{\boldsymbol{p}}_t = \nabla_{\tilde{\boldsymbol{x}}} G_t^{(3)}$ | $-\partial_t G_t^{(3)} = \mathcal{H}_t(-\nabla_p G_t^{(3)}, \boldsymbol{p}_t) - \tilde{\mathcal{H}}_{t'}(\tilde{\boldsymbol{x}}_{t'}, -\nabla_{\tilde{\boldsymbol{x}}} G_t^{(3)})$ |
| Type 4: | $-\boldsymbol{x}_t = \nabla_p G_t^{(4)}$ | $\tilde{\boldsymbol{x}}_{t'} = \nabla_{\tilde{\boldsymbol{p}}} G_t^{(4)}$ | $-\partial_t G_t^{(4)} = \mathcal{H}_t(-\nabla_p G_t^{(4)}, \boldsymbol{p}_t) - \tilde{\mathcal{H}}_{t'}(\nabla_{\tilde{\boldsymbol{p}}} G_t^{(4)}, \tilde{\boldsymbol{p}}_{t'})$ |

using the Legendre transform to relate the differentials

$$\int \langle \boldsymbol{p}_t, d\boldsymbol{x}_t\rangle = \int \big(d\langle \boldsymbol{p}_t, \boldsymbol{x}_t\rangle - \langle \boldsymbol{x}_t, d\boldsymbol{p}_t\rangle\big) = \text{constant} - \int \langle \boldsymbol{x}_t, d\boldsymbol{p}_t\rangle.$$

The choice of generating function determines the form of the coordinate transformation, and is usually selected with some design goal in mind. A few simple cases stand out. For instance, $G_t^{(1)} = \langle \boldsymbol{x}_t, \tilde{\boldsymbol{x}}_{t'}\rangle$ exchanges the generalized coordinates for the position and momenta $\{\tilde{\boldsymbol{x}}_{t'}, \tilde{\boldsymbol{p}}_{t'}\} \to \{\boldsymbol{p}_t, -\boldsymbol{x}_t\}$, demonstrating their equivalence. Also, choosing $G_t^{(2)} = \langle \tilde{\boldsymbol{p}}_{t'}, \boldsymbol{g}_t(\boldsymbol{x}_t)\rangle$ results in a point change of coordinates $\tilde{\boldsymbol{x}}_{t'} = \boldsymbol{g}_t(\boldsymbol{x}_t)$ with $\boldsymbol{g}_t(\boldsymbol{x}_t) = \boldsymbol{x}_t$ giving the identity transformation. A similar result can be obtained using $G_t^{(3)}$, but not $G_t^{(1)}$ and $G_t^{(4)}$. These latter two cannot produce identity transformations. This fact can be seen by considering the Legendre transformation relating $G_t^{(1)} = G_t^{(2)} - \langle \tilde{\boldsymbol{p}}_{t'}, \tilde{\boldsymbol{x}}_{t'}\rangle = 0$ as $t' \to t$. In this limit, the Lagrangian is zero and the arguments of $G_t^{(1)}$ and $G_t^{(4)}$ lose their independence ([151]). Furthermore, considering a type two transformation that is sufficiently close to the identity

$$G_t^{(2)}(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'}) = \langle \tilde{\boldsymbol{p}}_{t'}, \boldsymbol{x}_t\rangle + \delta t\, g_t(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'})$$

$$\boldsymbol{p}_t = \nabla_x G_t^{(2)} = \tilde{\boldsymbol{p}}_{t'} + \delta t\, \nabla_x g_t$$
$$\tilde{\boldsymbol{x}}_{t'} = \nabla_{\tilde{\boldsymbol{p}}} G_t^{(2)} = \boldsymbol{x}_t + \delta t\, \nabla_{\tilde{\boldsymbol{p}}} g_t,$$

that taking $g_t = \mathcal{H}_t\big(\boldsymbol{x}_t, \nabla_x G_t^{(2)}(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'})\big)$ gives to first order

$$\tilde{\boldsymbol{x}}_{t'} = \boldsymbol{x}_t + \delta t\, \nabla_p \mathcal{H}_t \cdot \nabla^2_{x\tilde{\boldsymbol{p}}} G_t^{(2)} = \boldsymbol{x}_t + \delta t\, \nabla_p \mathcal{H}_t + O(\delta t^2) \approx \boldsymbol{x}(t+\delta t)$$
$$\tilde{\boldsymbol{p}}_{t'} = \boldsymbol{p}_t - \delta t\, \big(\nabla_x \mathcal{H}_t + \nabla_p \mathcal{H}_t \cdot \nabla^2_{xx} G_t^{(2)}\big) = \boldsymbol{p}_t + \delta t\, \nabla_x \mathcal{H}_t + O(\delta t^2) \approx \boldsymbol{p}(t+\delta t).$$

Thus, $G_t^{(2)}$ is the infinitesimal generator of time translations along the vector field of the original Hamiltonian. Applied recursively these infinitesimal time translations endow the solutions with a group structure that connects the solution at time $t$ to the solution at any other time $t'$. It is worth mentioning that the generating functions described above may also be modified to include time dependence by mapping to time-extended phase space $\boldsymbol{x}' := \{\boldsymbol{x}_t, t\},\ \boldsymbol{p}' := \{\boldsymbol{p}_t, \mathcal{H}_t\}$, but this generalization is not needed here. Rather, it can be shown directly (following *e.g.*, [110]) that a (time-independent) type two generating function of the form

$$G^{(2)}(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'}) = \langle \tilde{\boldsymbol{p}}_{t'}, \tilde{\boldsymbol{x}}_{t'} \rangle - \int_t^{t'} ds\, \langle \boldsymbol{p}_s, \dot{\boldsymbol{x}}_s \rangle - \mathcal{H}_s(\boldsymbol{x}_s, \boldsymbol{p}_s)$$

connects the solutions at times $t$ and $t'$. Specifically,

$$\nabla_{\tilde{\boldsymbol{p}}} G^{(2)} = \nabla_{\tilde{\boldsymbol{p}}} \left( \langle \tilde{\boldsymbol{p}}_{t'}, \tilde{\boldsymbol{x}}_{t'} \rangle - \langle \boldsymbol{p}_s, \boldsymbol{x}_s \rangle \big|_{s=t}^{t'} \right) - \int_t^{t'} ds\, \nabla_{\tilde{\boldsymbol{p}}} \boldsymbol{z}_s \cdot (\boldsymbol{J}^\dagger \cdot \dot{\boldsymbol{z}}_s - \nabla_z \mathcal{H}_s) = \tilde{\boldsymbol{x}}_{t'} - \nabla_{\tilde{\boldsymbol{p}}} \boldsymbol{x}_t \cdot \boldsymbol{p}_t = \tilde{\boldsymbol{x}}_{t'}$$

where $\nabla_{\tilde{\boldsymbol{p}}} \boldsymbol{x}_t = 0$ since $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{p}}_{t'}$ are assumed independent, and

$$\nabla_x G^{(2)} = \nabla_x \big( \langle \tilde{\boldsymbol{p}}_{t'}, \tilde{\boldsymbol{x}}_{t'} \rangle - \langle \boldsymbol{p}_s, \boldsymbol{x}_s \rangle \big|_{s=t}^{t'} \big) - \int_t^{t'} ds\, \nabla_x \boldsymbol{z}_s \cdot (\boldsymbol{J}^\dagger \cdot \dot{\boldsymbol{z}}_s - \nabla_z \mathcal{H}_s) = \boldsymbol{p}_t.$$

It is this time propagation property that makes the $G^{(2)}$ (and also $G^{(3)}$) type generating functions particularly useful for solving the two-point boundary value problems arising in optimal estimation and control.

### 2.5.2 Application to fixed interval smoothing

Consider now the continuous time estimation action Eqn. (2.4). The necessary conditions for a local minimum require solving Hamilton's equations $\dot{\boldsymbol{z}}_{t|T} = \boldsymbol{J} \cdot \nabla_z \boldsymbol{H}_{t|T}$ with separated boundary conditions $\boldsymbol{p}_{0|T} = \boldsymbol{R}_0 \cdot (\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0})$ and $\boldsymbol{p}_{t|T} = 0$. The solution is naturally described by a type four generating function, which connects solutions at the endpoints of the integral by specifying their canonical momenta $\boldsymbol{p}_0,\ \tilde{\boldsymbol{p}}_T$. The trouble with this approach however comes in trying to construct an explicit solution as it requires a near-identity transformation, which as discussed above, is singular for type one and type four generating functions. The solution, discussed by [151] for control problems, is to first construct the solution

using type two or three generating functions and then Legendre transform back to a type one or four.

Consider a second order Taylor expansion of a type two generating function $G^{(2)}(\boldsymbol{x}_{t|T}, \tilde{\boldsymbol{p}}_{t'|T})$ around the filtered path

$$G^{(2)}(\boldsymbol{x}_{t|T}, \tilde{\boldsymbol{p}}_{t'|T} \approx A_{t|t} + \left\langle \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}, \begin{bmatrix} \nabla A_{t|t} \\ \nabla_{\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \\ \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix} \right\rangle.$$

Also expanding the Hamiltonians to second order, neglecting any order derivatives $\nabla^3 A$, gives

$$\mathcal{H}_t(\boldsymbol{x}_{t|T}, \nabla_x G^{(2)}) \approx \mathcal{H}_{t|t} +$$
$$\left\langle \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}, \nabla_z \mathcal{H}_{t|t} + \frac{1}{2} \nabla^2_{zz} \mathcal{H}_{t|t} \cdot \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix} \right\rangle$$

$$\tilde{\mathcal{H}}(\nabla_{\tilde{\boldsymbol{p}}} G^{(2)}, \tilde{\boldsymbol{p}}_{t'|T}) \approx \tilde{\mathcal{H}}_{t'|t'} +$$
$$\left\langle \begin{bmatrix} \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}, \nabla_{\tilde{\boldsymbol{z}}} \tilde{\mathcal{H}}_{t'|t'} + \frac{1}{2} \nabla^2_{\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}} \tilde{\mathcal{H}}_{t'|t'} \cdot \begin{bmatrix} \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix} \right\rangle.$$

Writing the time dependence as

$$-\partial_t G^{(2)}_t(\boldsymbol{x}_{t|T}, \tilde{\boldsymbol{p}}_{t'|T}) = \begin{bmatrix} 1 & \tilde{\mathcal{H}}_{t'}(\nabla_{\tilde{\boldsymbol{p}}} G^{(2)}_t, \tilde{\boldsymbol{p}}_{t'|T}) \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & \mathcal{H}_t(\boldsymbol{x}_{t|T}, \nabla_x G^{(2)}_t) \end{bmatrix},$$

and matching terms in the expansion gives

$$-\partial_t \begin{bmatrix} \nabla_x A_{t|t} \\ \nabla_{\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \nabla^2_{xx} A_{t|t} \\ \boldsymbol{0} & \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \nabla_x \mathcal{H}_{t|t} \\ \nabla_p \mathcal{H}_{t|t} \end{bmatrix} - \begin{bmatrix} \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} & \boldsymbol{0} \\ \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \nabla_{\tilde{\boldsymbol{x}}} \tilde{\mathcal{H}}_{t'|t'} \\ \nabla_{\tilde{\boldsymbol{p}}} \tilde{\mathcal{H}}_{t'|t'} \end{bmatrix}$$

$$-\partial_t \begin{bmatrix} \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \\ \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \nabla^2_{xx} A_{t|t} \\ \boldsymbol{0} & \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \nabla^2_{xx} \mathcal{H}_{t|t} & \nabla^2_{xp} \mathcal{H}_{t|t} \\ \nabla^2_{px} \mathcal{H}_{t|t} & \nabla^2_{pp} \mathcal{H}_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix}$$
$$- \begin{bmatrix} \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} & \boldsymbol{0} \\ \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \nabla^2_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}} \tilde{\mathcal{H}}_{t'|t'} & \nabla^2_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{p}}} \tilde{\mathcal{H}}_{t'|t'} \\ \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{x}}} \tilde{\mathcal{H}}_{t'|t'} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} \tilde{\mathcal{H}}_{t'|t'} \end{bmatrix} \cdot \begin{bmatrix} \nabla^2_{p\tilde{\boldsymbol{x}}} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}$$

Note that the fact the second equation involves the Hessian of a scalar function immediately implies the symmetry of the matrices $\nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t}$ and $\nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t}$, as well as $\nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} = [\nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t}]^\dagger$.

The filtering solution is obtained by assuming $\tilde{\boldsymbol{H}}_{t'|t'}(\tilde{\boldsymbol{x}}_{t'}, \tilde{\boldsymbol{p}}_{t'})$ is constant, so $\dot{\tilde{\boldsymbol{z}}}_{t'} = 0$, which may be

interpreted as a canonical transformation $\tilde{\boldsymbol{z}}_{t'} \to \boldsymbol{z}_t$ from fixed initial conditions $\tilde{\boldsymbol{z}}_{t'}$ to moving coordinates $\boldsymbol{z}_t$, in the reference frame of the estimate. This choice also coincides with the Hamilton-Jacobi equation $-\partial_t G^{(2)}(\boldsymbol{x}_t, t) = \mathcal{H}_t(\boldsymbol{x}_t, \nabla_x G^{(2)})$. [5] Making this substitution in the second derivative equation above gives,

$$
\begin{aligned}
-\partial_t\left(\nabla^2_{xx} A_{t|t}\right) &= \nabla^2_{xx}\mathcal{H}_{t|t} + \nabla^2_{xx}A_{t|t}\cdot\nabla_{px}\mathcal{H}_{t|t} + \nabla_{xp}\mathcal{H}_{t|t}\cdot\nabla^2_{xx}A_{t|t} + \nabla^2_{xx}A_{t|t}\cdot\nabla_{pp}\mathcal{H}_{t|t}\cdot\nabla^2_{xx}A_{t|t} \\
-\partial_t\left(\nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t}\right) &= \left(\nabla^2_{xp}\mathcal{H}_{t|t} + \nabla^2_{xx}A_{t|t}\cdot\nabla^2_{pp}\mathcal{H}_{t|t}\right)\cdot\nabla^2_{x\tilde{\boldsymbol{p}}}A_{t|t} \\
-\partial_t\left(\nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t}\right) &= \nabla^2_{\tilde{\boldsymbol{p}}x}A_{t|t}\cdot\left(\nabla^2_{px}\mathcal{H}_{t|t} + \nabla^2_{pp}\mathcal{H}_{t|t}\cdot\nabla^2_{xx}A_{t|t}\right) \\
-\partial_t\left(\nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t}\right) &= \nabla^2_{\tilde{\boldsymbol{p}}x}A_{t|t}\cdot\nabla^2_{pp}\mathcal{H}_{t|t}\cdot\nabla^2_{x\tilde{\boldsymbol{p}}}A_{t|t}.
\end{aligned}
$$

The equation for $-\partial_t\nabla^2_{xx}A_{t|t}$ is the continuous time Riccati equation for the inverse covariance, given in Eqn. (2.16) above with $\boldsymbol{R}_{t|t} \to \nabla^2_{xx}A_{t|t}$. The other three equations also correspond to the other entries in the matrix in Eqn. (2.16), with the addition of 'transition matrices' $\nabla^2_{x\tilde{\boldsymbol{p}}}A_{t|t}$. The generating function framework thus generalizes the previous results, although the meaning behind these matrices is still not totally clear.

Regarding initial conditions, correspondence with the Riccati equation gives $\nabla^2_{xx}A_{0|0} = \boldsymbol{R}_0$. As this equation also does not depend on the matrices $\nabla^2_{x\tilde{\boldsymbol{p}}}A_{t|t}$ or $\nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}}A_{t|t}$, their initial values do not matter, at least for the filtering solution. It is interesting to note however, that one might expect (*e.g.*, following [151]) $G^{(2)}(\boldsymbol{x}_{t|T}, \tilde{\boldsymbol{p}}_{t|T})$ to generate an identity transformation as $t \to t'$. But this would require $\nabla^2_{x\tilde{\boldsymbol{p}}}A_{0|0} = \nabla^2_{\tilde{\boldsymbol{p}}x}A_{0|0} = \boldsymbol{I}$ and $\nabla^2_{xx}A_{0|0} = \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}}A_{0|0} = 0$, thereby contradicting $\nabla^2_{xx}A_{0|0} = \boldsymbol{R}_0$. Thus, the interpretation of a simple unitary transformation from fixed initial conditions to is not quite correct here.

The filtering solution comes from the equation

$$
-\partial_t\left(\nabla_x A_{t|t}\right) = \nabla_x\mathcal{H}_{t|t} + \nabla^2_{xx}A_{t|t}\cdot\nabla_p\mathcal{H}_{t|t}.
$$

For instance, substituting $-\nabla_x\mathcal{H}_{t|t} \to \dot{\boldsymbol{p}}_{t|t}$ and $\nabla_p\mathcal{H}_{t|t} \to \dot{\boldsymbol{x}}_t$ gives Eqn. (2.19) and thus Eqn. (2.17). The Kalman filter solution follows from the assumption $\boldsymbol{p}_{t|t} = 0$. An more direct (but less obvious) route is to substitute $-\partial_t A_{t|t} \to \nabla^2_{xx}A_{t|t}\cdot\dot{\boldsymbol{x}}_{t|t}$ from Eqn. (2.20). Replacing the derivatives of the Hamiltonian and assuming $\boldsymbol{p}_{t|t} = 0$ gives the desired result.

The other equation

$$
-\partial_t\left(\nabla_{\tilde{\boldsymbol{p}}}A_{t|t}\right) = \nabla^2_{\tilde{\boldsymbol{p}}x}A_{t|t}\cdot\nabla_p\mathcal{H}_{t|t}
$$

[5] Note that the same Hamilton-Jacobi equation can also be obtained from $G^{(1)}$, along with related versions from $G^{(3)}$ and $G^{(4)}$.

evidently has no impact on the filtering solution. It involves the term $\nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t}$ whose initial condition was unable to be identified, and has no effect on the Riccati equation. Also, the fact that

$$\partial_t \left( \nabla_{\tilde{\boldsymbol{p}}} A_{t|t} \right) = \dot{\tilde{\boldsymbol{x}}}_{t|t} = \nabla_{\tilde{\boldsymbol{p}}} \tilde{\mathcal{H}}_{t|t} = 0$$

seems to imply

$$\nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \cdot \nabla_p \mathcal{H}_{t|t} = \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \cdot \left( \boldsymbol{f}(\boldsymbol{x}_{t|t}) + \boldsymbol{R}_f^{-1} \cdot \boldsymbol{p}_{t|t} \right) = \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \cdot \boldsymbol{f}(\boldsymbol{x}_{t|t}) = \boldsymbol{0}.$$

everywhere along the filtering solution, which can be satisfied trivially by setting the initial condition $\nabla^2_{\tilde{\boldsymbol{p}}x} A_{0|0} = \boldsymbol{0}$.

It is therefore apparent that the Kalman filter solution is highly singular in the sense that it constrains several of the fundamental quantities appearing in the generating function solution to be identically zero. It is worth noting however, that this did not require any assumptions on the *values* of $\tilde{\boldsymbol{x}}_{t'},\ \tilde{\boldsymbol{p}}_{t'}$, only that $\tilde{\boldsymbol{H}}_{t'} =$ constant requires their time derivatives to be zero. One rough interpretation of this is that the Kalman filter solution, being the optimal stabilizing feedback control, provides optimal damping of any disturbances introduced at time $t'$ by the variables $\tilde{\boldsymbol{x}}_{t'}$ and $\tilde{\boldsymbol{p}}_{t'}$. In other words, the filtering solution is still optimal regardless of the values for $\tilde{\boldsymbol{x}}_{t'},\ \tilde{\boldsymbol{p}}_{t'}$. Obviously, more work is needed to show this explicitly, but the interpretation is intuitive and of all the techniques discussed here, the generating function approach seems to provides the clearest and most general description.

As for the backwards pass, there are several options. Taking $\boldsymbol{x}_{t|t},\ \nabla^2_{xx} A_{t|t}$ as initial conditions one could use the same approach with $G^{(2)}$ or $G^{(3)}$ generating functions to run an extended Kalman filter *backwards* in time. This gives the two filter solution of Mayne and Fraser ([121, 57, 56]), although the problem of blending the solutions to form the optimal smoothed trajectory still must be addressed.

Alternatively, one could use Legendre transform to $G^{(4)}(\boldsymbol{p}_{t|T}, \tilde{\boldsymbol{p}}_{t|T})$, which is more appropriate given the boundary conditions $\boldsymbol{p}_{t|T} = 0,\ \tilde{\boldsymbol{p}}_{0|T} = \boldsymbol{R}_0 \cdot (\boldsymbol{x}_{0|T} - \boldsymbol{x}_{0|0})$. Using the Legendre transform

$$G^{(4)}(\boldsymbol{p}_{t|T}, \tilde{\boldsymbol{p}}_{t'|T}) = G^{(2)}(\boldsymbol{x}_t, \tilde{\boldsymbol{p}}_{t'|T}) - \langle \boldsymbol{p}_{t|T}, \boldsymbol{x}_{t|T} \rangle,$$

with

$$\langle \boldsymbol{p}_{t|T}, \boldsymbol{x}_{t|T} \rangle = \langle \boldsymbol{p}_{t|t} + \delta\boldsymbol{p}_t, \boldsymbol{x}_{t|t} + \delta\boldsymbol{x}_t \rangle = \langle \boldsymbol{p}_{t|t}, \boldsymbol{x}_{t|t} \rangle + \left\langle \begin{bmatrix} \boldsymbol{x}_{t|t} \\ \boldsymbol{p}_{t|t} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \delta\boldsymbol{x}_t \\ \delta\boldsymbol{p}_t \end{bmatrix}, \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \cdot \begin{bmatrix} \delta\boldsymbol{x}_t \\ \delta\boldsymbol{p}_t \end{bmatrix} \right\rangle$$

$$\begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \boldsymbol{p}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}$$

$$\begin{bmatrix} \delta \boldsymbol{x}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix} = \begin{bmatrix} [\nabla^2_{xx} A_{t|t}]^{-1} & -[\nabla^2_{xx} A_{t|t}]^{-1} \cdot \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \delta \boldsymbol{p}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}$$

and using $\boldsymbol{p}_{t|t} = \nabla A_{t|t} = 0$, $\nabla_{\tilde{\boldsymbol{p}}} A_{t|t} = \tilde{\boldsymbol{x}}_{t'|t'}$, gives

$$G^{(4)}(\boldsymbol{p}_{t|T}, \tilde{\boldsymbol{p}}_{t'|T}) \approx A_{t|t} + \left\langle \begin{bmatrix} \delta \boldsymbol{p}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix}, \begin{bmatrix} -\boldsymbol{x}_{t|t} \\ +\tilde{\boldsymbol{x}}_{t'|t'} \end{bmatrix} + \frac{1}{2} \boldsymbol{\Xi}_{t|t} \cdot \begin{bmatrix} \delta \boldsymbol{p}_t \\ \delta \tilde{\boldsymbol{p}}_{t'} \end{bmatrix} \right\rangle.$$

$$\boldsymbol{\Xi}_{t|t} := \begin{bmatrix} -[\nabla^2_{xx} A_{t|t}]^{-1} & [\nabla^2_{xx} A_{t|t}]^{-1} \cdot \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \\ \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \cdot [\nabla^2_{xx} A_{t|t}]^{-1} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} - \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} \cdot [\nabla^2_{xx} A_{t|t}]^{-1} \cdot \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix}.$$

Note that the lower right block is the Schur complement of the Hessian matrix

$$\begin{bmatrix} \nabla^2_{xx} A_{t|t} & \nabla^2_{x\tilde{\boldsymbol{p}}} A_{t|t} \\ \nabla^2_{\tilde{\boldsymbol{p}}x} A_{t|t} & \nabla^2_{\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}} A_{t|t} \end{bmatrix}.$$

## 2.6 Symplectic structure and optimal stability

The preceding discussion points to the role of symplectic structure in the stability of solutions to the two-point boundary value problem described by Eqn. (2.5). As mentioned, boundary value problems do not inherit the same local uniqueness properties found in initial value problems, and may thus have many solutions. Moreover, even when a unique solution exists, its *representation* is unique only up to a canonical transformation. These representations are not at all equivalent, as many (if not most) choices produce solutions that are highly unstable, and cannot be directly integrated for an appreciable length of time. For instance, integrating Hamilton's Eqns. (2.5) (which itself is a canonical transformation) produces a solution that quickly becomes numerically unbounded. While this phenomenon is present in physics — where for instance even the simple harmonic oscillator requires symplectic integration methods to keep its total energy constant and bounded — in estimation, it appears to be much more severe. It is evident then, considering the set of all possible canonical transformations, that only a subset are stable enough to be explicitly integrated.

This stability is a result of combining symplectic structure and the Riccati transformation, as

illustrated by the Hamiltonian approach to steady-state Kalman filtering ([200, 177, 42]). These constraints are imposed not just on the canonical state $\delta \boldsymbol{z}_t$, but also on the fluctuations $\delta \boldsymbol{z}_t$ linearized about the filtered path. Recall that these fluctuations are described by the 'accessory' two-point boundary value problem $\delta \boldsymbol{z}_{t|t} = \boldsymbol{J} \cdot \nabla^2 \mathcal{H}_{t|t}$, obtained from enforcing the second variation $\delta^2 A = 0$. Imposing symplectic structure on the neighboring extremals, splits the eigenvalues of the (linearized) Hamiltonian into positive and negative real pairs. The Riccati transformation then inverts the positive poles, resulting in a feedback control law that shifts the eigenvalues of the error system inside the unit circle.

While this result has been known for some time, it is perhaps not as widely publicized as it should be. The fact that it makes explicit the role of symplectic structure in the optimal stability of the Kalman filter suggests there is more to be learned from viewing the problem in this way. Symplectic structure provides a great deal of geometric and group-theoretic insight into the problem ([147]), which may be leveraged to construct better fast approximation algorithms. For instance, one might consider relaxing the constraint $\boldsymbol{p}_t = 0$ when the filtered solution is far from the truth, and explicitly tracking $\boldsymbol{x}_t$ and $\boldsymbol{p}_t$ using a symplectic integration algorithm. For this and other related ideas, it would be especially helpful if the Hamiltonian technique could be extended to the time-dependent case, which to my knowledge has not yet been done.

Chapter 2, in part is being prepared for submission for publication of the material. Rey, Daniel. The dissertation author was the primary investigator and author of this material.

# 3 Observability and conditioning in dynamical inverse problems

Inverse problems vary in their degree of difficulty from trivial to impossible. Exactly where a particular problem falls on this spectrum depends on a variety of factors, all of which contribute to its conditioning. At its core, conditioning describes the sensitivity of the output (the solution) to small variations in the inputs (the data). Loosely speaking, for inverse problems it may be interpreted as the availability of information.

For static parameter estimation problems with linear models, this sensitivity is fully determined by the condition number of the observation matrix $\boldsymbol{H}$, which is the ratio of its largest and smallest singular values. It is an intrinsic property of the model, which in this case is just a matrix, and is independent of the choice of algorithm. This separation of the problem from its solution (*i.e.*, the algorithm) is a common approach in optimization and numerical analysis, where the goal is mainly to suppress round-off errors. It assumes however that the problem itself is well-posed, which is often not the case for inverse problems. If however the matrix $\boldsymbol{H}$ is singular then its condition number will be infinite and the problem satisfy Hadamards third requirement for well-posedness: *i.e.*, that the solution should change continuously with the data ([68]).

It therefore requires regularization, which broadly speaking is the process of introducing additional 'prior' information in the form of constraints that the solution unique. Where this additional information comes from depends on the problem, and may at its simplest just give priority to solutions that have desired properties like smoothness or smallest magnitude. The latter falls under the generalized framework of Tikhonov regularization ([189]) and by definition affects the conditioning of the problem, and thus the stability of the solution. These choices go both ways, in the sense that although the intent of regularization is to make the problem easier to solve, a poor choice of hyper-parameters can render even the most trivial

problem unsolvable in practice.

This is especially true in the dynamic case, where one has time-series observations and a model describing the time-evolution of the underlying state of the system. The observations are typically partial so that the observation operator is rank-deficient and therefore requires regularization, which may be provided by the model. That is, the model dynamics provides a tremendous source of additional information that allows one to distinguish between trajectories that are otherwise given what can be observed. Under certain circumstances, the information inherent in the system's dynamics allows the hypothetical 'true' state of the system to be reconstructed from noisy observations and imperfect dynamics without any additional prior knowledge of its initial state. In other cases however it can have the opposite affect, making an otherwise trivial problem computationally intractable.

The problem arises when the system is chaotic. That is, when the model contains *dynamical instabilities* that make its trajectories highly sensitive to perturbations in initial conditions. Chaos in the system injects dynamical noise into the estimation process that makes identifying the 'true' solution inherently more difficult ([159, 51]). While such behavior is quite common in the study of nonlinear dynamical systems, nonlinearity is not a requirement. Take for instance, the classic example of the diffusion equation, which is linear but quickly becomes numerically intractable when integrated backwards in time. The inverse problem of determining the initial condition for a final observed temperature distribution is therefore regarded as ill-posed. However, the solution will remain bounded for short enough periods of time, so the distinction between ill-posed and ill-conditioned problems is not always clear.

There is thus a need for a better understanding of degree of difficulty associated with inverse problems, as well as how this relates to the notion of its conditioning, the overall probability of success, and how this depends on its three elemental components: namely the model, the data, and the algorithm. With this in mind, this chapter aims to begin a discussion of the complicated trade-offs inherent in constructing the solution to ill-posed inverse problems. Beginning with a short overview of the mathematical theory of conditioning that focuses specifically on the solution to linear and nonlinear state and parameter estimation problems, these ideas are explored through a series of numerical experiments that seek to identify the strengths, weaknesses and connections between some of the most basic algorithms developed for solving these problems. Particular emphasis will be placed on the density of observations required for success, and the overall impact of dynamical instability in the model. The results will highlight some fundamental limitations of certain algorithms, and demonstrate that while some algorithms are better than others at dealing with poorly observable systems, they tend to be more computationally intensive. Thus, there is an

inherent trade-off between accuracy of solution and the efficiency of the calculation that must be addressed on a problem-specific basis.

## 3.1 Conditioning, regularization and numerical stability

This section begins with a brief introduction to the linear theory of conditioning and its nonlinear extensions. Special emphasis is placed on its application to inverse problems, including the issue of regularization and its effect on conditioning. It then concludes with a brief discussion of algorithmic stability, consistency and convergence and their relation to conditioning.

### 3.1.1 Conditioning

The conditioning of a problem is a measure of the sensitivity of its solution to errors in the input. Although the concept is typically introduced in numerical analysis of linear systems, where the goal is to suppress errors due to finite numerical precision, it applies equally well to the estimation of the parameters $\boldsymbol{x} \in \mathbb{R}^D$ from a linear model $\boldsymbol{H} \in \mathbb{R}^{D\times L}$, given observations $\boldsymbol{y} \in \mathbb{R}^L$ such that $\boldsymbol{H} \cdot \boldsymbol{x} = \boldsymbol{y}$. In this situation, the unknown parameters can be found by matrix inversion $\boldsymbol{x} = \boldsymbol{H}^{-1} \cdot \boldsymbol{y}$, provided the matrix $\boldsymbol{H}$ is nonsingular. The goal is to understand the sensitivity of the solution $\boldsymbol{x}$ due to errors in the observations $\boldsymbol{y}$, which are determined in practice by the experimental apparatus used to record the measurements, and is typically many orders of magnitude larger than machine precision.

In the linear case, a worst case estimate of this relative sensitivity is given by the condition number $\kappa$ of the model $\boldsymbol{H}$. From a parameter estimation perspective however, it is worthwhile to distinguish the hypothetical 'true' model $\boldsymbol{H}^*$, data $\boldsymbol{y}^*$ and solution $\boldsymbol{x}^*$ from their approximations $\boldsymbol{H} := \boldsymbol{H}^* - \boldsymbol{\delta H}$, $\boldsymbol{y} := \boldsymbol{y}^* - \boldsymbol{\delta y}$, and $\boldsymbol{x} := \boldsymbol{x}^* - \boldsymbol{\delta x}$, which are assumed to have errors $\boldsymbol{\delta H}$, $\boldsymbol{\delta y}$, $\boldsymbol{\delta x}$. Following [78], if we assume $\boldsymbol{H}^*$ is invertible with $|[\boldsymbol{H}^*]^{-1}||\boldsymbol{\delta H}| < 1$, it can be shown that the relative error in the solution $\boldsymbol{\delta x}$ is bounded by

$$\frac{|\boldsymbol{\delta x}|}{|\boldsymbol{x^*}|} \leq \frac{\kappa^*}{1 - \kappa^* |\boldsymbol{\delta H}|/|\boldsymbol{H^*}|} \left( \frac{|\boldsymbol{\delta y}|}{|\boldsymbol{y}^*|} + \frac{|\boldsymbol{\delta H}|}{|\boldsymbol{H}^*|} \right)$$

where

$$\kappa[\boldsymbol{H}^*] \equiv \kappa^* := |[\boldsymbol{H}^*]^{-1}||\boldsymbol{H}^*| = \frac{\sigma_{\max}}{\sigma_{\min}} \geq 1$$

is the condition number of the matrix $\boldsymbol{H}^*$ and $\sigma_{\max}$, $\sigma_{\min}$ are respectively its largest and smallest singular values.

Note that this is the condition number of the hypothetical 'true model, which is usually unknown. If we instead consider the limit $\boldsymbol{\delta H} \to 0$ that the model is perfect, the above expression simplifies to

$$\frac{|\boldsymbol{H}^* \cdot \boldsymbol{\delta x}|}{|\boldsymbol{\delta y}|} \leq \kappa^*,$$

which involves only known quantities, since the model $\boldsymbol{H}^*$ is assumed known the error $|\boldsymbol{\delta y}|$ in the measurement apparatus can typically be measured. Thus, when the model is perfect, the condition number gives an upper bound on the ratio of the observed error in the estimate $|\boldsymbol{H}^* \cdot \boldsymbol{\delta x}|$ to the error in the observations $\boldsymbol{\delta y}$. In other words, the solution becomes more unreliable as the condition number grows.

Although in the linear case the condition number is a global property of the model $\boldsymbol{H}$, it is not invariant to the choice of coordinates. Rescaling variables, for instance by changing measurement units, changes the condition number of the problem. This process is called preconditioning, and generally involves choosing a matrix $\boldsymbol{P}$ such that $\boldsymbol{P}^{-1} \cdot \boldsymbol{H}$ is better conditioned and solving the modified problem $\boldsymbol{P}^{-1} \cdot \boldsymbol{H} \cdot \boldsymbol{x} = \boldsymbol{y}$ ([82, 199]). Roughly speaking, $\boldsymbol{P}$ should be chosen to be close (in some sense) to $\boldsymbol{H}$, but easier to invert. Note that choosing $\boldsymbol{P} = \boldsymbol{H}$ the modified problem is just as difficult as the original. Thus, there are a number of trade-offs to consider when choosing a preconditioning matrix, so one often has to resort to heuristics, developed on an application specific basis.

Systems of nonlinear equations $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{y}$ are more difficult. The condition number is no longer a global property of the model, although suitable analogs were proposed by [168, 21]. Using upper and lower Lipschitz bounds function in a neighborhood of a solution $\boldsymbol{x}^*$, [168] showed that when this neighborhood shrinks to a point, the condition number of the Jacobian $\kappa[\nabla \boldsymbol{h}^*(\boldsymbol{x}^*)]$ coincides with the linear case. Note however that this definition assumes the solution is known *a-priori*, so this analysis cannot be applied experimentally.

### 3.1.2 Regularization

The assumption that the observation model $\boldsymbol{H}$ was non-singular and thus invertible at the solution $\boldsymbol{x}$ requires the mapping $\boldsymbol{H}$ to be square, so the number of states $D$ has to equal the number of measurements $L$. But this is typically not the case for inverse problems, where the the measurements are sparse $L \ll D$ and the model is degenerate, in the sense that an infinite number of solutions satisfy $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{y}$.

To remedy the situation, one must supply some additional prior information, to constrain the problem thereby removing the singularity in the model, and making the solution unique. This can be as

simple as selecting solutions with 'nice' qualities such as smoothness or continuity. Or, if one has some prior knowledge of where the solution should lie, such as the result of a previous analysis, or long-term averaged behavior, this information can be used to choose solutions that are closest (in some sense) to this background or prior estimate. This process of removing degeneracy in the model by selecting solutions with desired properties is known as regularization, and is generally described using the framework given by [190], which is now briefly discussed.

The linear parameter estimation problem can be reformulated as a quadratic optimization problem

$$\min_{\boldsymbol{x}} A(\boldsymbol{x}) := \frac{1}{2}|\boldsymbol{y}-\boldsymbol{H}\cdot\boldsymbol{x}|^2_{\boldsymbol{R}_m},$$

by noting that the stationary points require $\nabla \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m \cdot (\boldsymbol{y}-\boldsymbol{H}\cdot\boldsymbol{x}) = 0$, where the matrix $\boldsymbol{R}_m$ of inverse covariances has been added to weight the measurements based on their relative uncertainty and is assumed to be positive definite. This expression implies either $\boldsymbol{H}\cdot\boldsymbol{x} = \boldsymbol{y}$, or the observed error $\boldsymbol{y}-\boldsymbol{H}\cdot\boldsymbol{x}$ (also known as the *innovation*) lies in the null space of $\boldsymbol{H}^\dagger \cdot \boldsymbol{R}_m$. In other words, the solution $\boldsymbol{x}$ either solves $\boldsymbol{H}\cdot\boldsymbol{x} = \boldsymbol{y}$, or the error $\boldsymbol{y}-\boldsymbol{H}\cdot\boldsymbol{x}$ lies in a direction that is not resolved by the observation operator.

Tikhonov regularization introduces an additional term into the objective function

$$\min_{\boldsymbol{x}} \tilde{\boldsymbol{A}}(\boldsymbol{x}) := \frac{1}{2}|\boldsymbol{y}-\boldsymbol{H}\cdot\boldsymbol{x}|^2_{\boldsymbol{R}} + \frac{1}{2}|\boldsymbol{b}-\boldsymbol{x}|_{\boldsymbol{B}},$$

where $\boldsymbol{b}$ is an *a-priori* or background estimate of the state $\boldsymbol{x}$, and $\boldsymbol{B}$ is its weight or uncertainty in the initial guess. Since $\boldsymbol{b} \in \mathbb{R}^D$, when the matrix $\boldsymbol{B}$ is positive definite the problem will have a unique solution. The regularization selects for solutions that are close, in a least-squares sense, to $\boldsymbol{b}$. The choice of $\boldsymbol{B}$ determines how this selection process is weighted. For instance, when $|\boldsymbol{B}| \gg \boldsymbol{H}$ the solution $\boldsymbol{x}$ is skewed more towards the background estimate $\boldsymbol{x} \approx \boldsymbol{b}$. Likewise, when $|\boldsymbol{B}| \ll |\boldsymbol{H}|$ it is skewed more towards the data $\boldsymbol{y}$. Also, the choice $\boldsymbol{b} = 0$ and $\boldsymbol{B} = \alpha \boldsymbol{I}$ imposes a filter on the singular values of $\boldsymbol{H}$, which can be seen by writing the model in terms of its singular value decomposition $\boldsymbol{H} = \boldsymbol{U}\cdot\boldsymbol{S}\cdot\boldsymbol{V}^\dagger$. The solution is given by

$$\boldsymbol{x} = \boldsymbol{V}\cdot[\boldsymbol{S}^\dagger\cdot\boldsymbol{S} + \alpha\boldsymbol{I}]^{-1}\cdot\boldsymbol{S}\cdot\boldsymbol{U}^\dagger\cdot\boldsymbol{y},$$

which in the limit $\alpha \to 0$ this reduces to the pseudoinverse $\boldsymbol{H}^+$, selecting the solution with the minimum norm $|\boldsymbol{x}|$.

In this context, the condition number is given by $\kappa[\nabla^2 \tilde{\boldsymbol{A}}] = \sqrt{\lambda_{\max}/\lambda_{\min}}$, where $\lambda_{\max}$, $\lambda_{\min}$ are

the eigenvalues of the Hessian $\nabla^2 \tilde{\boldsymbol{A}} = \boldsymbol{H}^\dagger \cdot \boldsymbol{H} + \boldsymbol{B}^\dagger \cdot \boldsymbol{B}$. This follows from the fact that the eigenvalues $\lambda_i$ of $\boldsymbol{H}^\dagger \cdot \boldsymbol{H}$ are the the squares of the singular values $\sigma_i$ of $\boldsymbol{H}$. So without regularization (*i.e.*, when $\boldsymbol{B} = 0$) the condition number is infinite, since $\sigma_{\text{min}} = 0$ unless the model $\boldsymbol{H}$ happens to be full rank.

Furthermore, the optimization perspective provides a geometric interpretation of this problem. Since the eigenvalues of the Hessian are the principal curvatures of the objective function, when $\boldsymbol{H}$ is singular there are directions in which the objective function is flat, so the solution is degenerate in the sense that any perturbation along these directions is also a minimizer. The regularization term lifts this degeneracy, since it is assumed positive definite, to provide a single, unique global minimum. Also, along these lines, the preconditioning process can be viewed as rescaling the curvature of the objective function around the solution $\boldsymbol{x}^*$ to unity, so it is isotropic in all dimensions. For quadratic problems (*i.e.*, linear models), the optimal preconditioner is the Hessian $\nabla^2 A(\boldsymbol{x}^*)$ ([198]) and is independent of $\boldsymbol{x}^*$, though its inverse can be costly to compute for large dimensional systems.

In the nonlinear case, the corresponding optimization problem is no longer quadratic, so multiple local minima may exist. Also, the Hessian is more complicated, involving terms terms like $\nabla^2 \boldsymbol{h} \cdot (\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{x}))$ that depend on the data. Many nonlinear optimization methods, such as the Gauss-Newton method, ignore these terms due to their added complexity. But the validity of this approximation is difficult to determine *a-priori*, although they may be expected to hold near a local minimum where $\boldsymbol{h}(\boldsymbol{x}) \approx \boldsymbol{y}$.

Thus, it is evident that, like preconditioning, the choice of regularization also affects the conditioning of the problem. This suggests the question of whether this choice should be considered part of the problem statement or the algorithm, since in many cases the choice of regularization is a defining feature of the algorithm. This somewhat philosophical question will not be addressed here, although it is worth noting that the answer will likely be application dependent. The point however is that regularization plays a critical role in the determining conditioning of an inverse problem, and will be a main focus of our subsequent investigation.

### 3.1.3 Stability, consistency and convergence

It is apparent from this discussion that the notion of conditioning is closely related to the stability of the problem, as defined by the sensitivity of its solution to changes in both the parameters of the model (*e.g.*, dimension, grid-size) and its inputs (*i.e.*, observations). This definition of stability is akin to the idea of 'well-posedness', and can be studied, for instance using the framework described by [169], which applies to most techniques used for approximating solutions to differential equations as well as dynamical

inverse problems, including shooting methods, finite-difference, Galerkin, and spectral methods. These ideas are now briefly outlined, with some slight notational changes to fit the present discussion.

Most inverse problems can be formulated as a nonlinear system of equations $\boldsymbol{\Psi}^*(\boldsymbol{x}^*) = \mathbf{0}$ where $\boldsymbol{x}^* \in \Omega^*$. In estimation for instance, the variational approximation to expectation values of the conditional probability distribution involves finding stationary points of a scalar objective function. In practice however, these expressions are abstract and cannot be evaluated unless an analytical solution is available. To make them concrete, the problem and its solution must be discretized, which involves the introduction of a small parameter $\Delta$ such as the step size. The solution $\boldsymbol{x}_\Delta$ to the discretized problem $\boldsymbol{\Psi}_\Delta(\boldsymbol{x}_\Delta) = 0$, with $\boldsymbol{x}_\Delta \in \Omega_\Delta$, is then iteratively approximated by $\boldsymbol{x}_\Delta^{(i)}$, with $\boldsymbol{x}_\Delta^{(i)} \to \boldsymbol{x}_\Delta$ as $i \to \infty$.

The global error in the approximate solution $\boldsymbol{x}_\Delta^{(i)}$ may be measured by $e_\Delta^{(i)} := |\boldsymbol{x}_\Delta^* - \boldsymbol{x}_\Delta^{(i)}|$, which provides a means to define convergence. That is, the algorithm is then said to be convergent if there exists some finite critical value $\Delta_c > 0$ such that for all $\Delta < \Delta_c$ it produces a result such that $e_\Delta^{(i)} \to 0$ as $i \to \infty$ and $\Delta \to 0$

Note that the concept of global error requires two choices. The first is the selection of an 'optimal' discretization $\boldsymbol{x}_\Delta^*$ the 'hypothetical' solution $\boldsymbol{x}^*$. The optimality of this solution is a choice that depends on several factors including the problem, the discretization, and the algorithm. No rigorous definition is given, although some choices are natural, such as taking $\boldsymbol{x}_\Delta^*$ to coincide with $\boldsymbol{x}^*$ at nodal values. The second is the choice of metric, which may be constrained by practical considerations. Such is the case for inverse problems, where the full state of the system is not generally known, so predictability must be used to distinguish between solutions.

The convergence of an algorithm depends on two factors: its consistency and its stability. Consistency depends on the discretization of both the problem $\boldsymbol{\Psi}_\Delta$ and target solution $\boldsymbol{x}_\Delta^*$, in that the discretization is said to be consistent if the residual $|\boldsymbol{\Psi}_\Delta(\boldsymbol{x}_\Delta^*)| \to 0$ as $\Delta \to 0$. Also, convergence and consistency are said to be of order $p$ if $e_\Delta^{(i)}$ and $|\boldsymbol{\Psi}_\Delta(\boldsymbol{x}_\Delta^*)|$ are $O(\Delta^p)$ respectively, as $\Delta \to 0$ and $i \to \infty$. Stability, on the other hand, is related to boundedness of the approximate solutions. A number of definitions have been proposed, many of which have significant overlap. For instance, the $N$-stability condition requires the existence of a finite 'stability' constant $S$ such that

$$|\boldsymbol{x}_\Delta - \tilde{\boldsymbol{x}}_\Delta| \leq S\,|\boldsymbol{\Psi}_\Delta(\boldsymbol{x}_\Delta) - \boldsymbol{\Psi}_\Delta(\tilde{\boldsymbol{x}}_\Delta)| \tag{3.1}$$

for all $\boldsymbol{x}_\Delta,\ \tilde{\boldsymbol{x}}_\Delta \in \Omega_\Delta$ and $\Delta \leq \Delta_c$. This definition, includes as particular cases the notions of Lax-stability for linear initial value problems in partial differential equations and 0-stability in numerical ordinary

differential equations. Furthermore, when taken together, these properties establish the rather well-known fact that convergence and stability imply convergence.

However, the definition of stability in Eqn. (3.1) is somewhat limited due to the fact that it is required to hold globally — *i.e.*, for all $\boldsymbol{x}_\Delta \in \Omega_\Delta$. This has lead to the introduction of various local definitions, two of which were compared by [116], where it was concluded that the definition given by [97] should be favored. Keller's definition restricts the domain so that Eqn. (3.1) is only required to be bounded within a ball $B(\boldsymbol{x}^*_\Delta, \rho)$ of radius $\rho$ centered around the 'optimal' solution $\boldsymbol{x}^*_\Delta$. The value $\rho$ is called the *stability threshold*. It is infinite for linear problems, recovering the global condition Eqn. (3.1). But the nonlinear case generally exhibits a critical value $\rho_c$ above which stability is no longer guaranteed. The complementary limit — an upper bound above which the problem is guaranteed to be unstable — is not considered here.

Local stability measures cannot be applied to experimental data, as they require *a-priori* knowledge of the 'true' solution $\boldsymbol{x}^*_\Delta$. But this does not mean they are not useful. For instance, one can use simulated data approximate $\boldsymbol{x}^*_\Delta$ given the best model available, and then test these assumptions and constraints to arrive at an lower bound on the probability of success, as defined by the convergence to the desired solution.

While it is not necessarily evident that consistency and stability imply convergence in this case, [169] gives a proof based on the inverse mapping theorem, which shows that a unique solution exists within the ball $B(\boldsymbol{x}^*_\Delta, \rho)$. But this proof assumes that the mapping $\boldsymbol{\Psi}_\Delta$ is continuous, and between spaces of identical dimension. Further work is required to see if these conditions may be relaxed.

### 3.1.4 Dynamical regularization

The previous section was mainly limited to the estimation of parameters from a static model, in which the regularization is derived from some *a-priori* knowledge about the solution, such as its smoothness properties or where the solution is located. When no such information is available the problem typically becomes very difficult (if not impossible), since one has no way of distinguishing between various solutions that are all equivalent with respect to the observation operator. In these cases, it is often necessary to break this degeneracy by making heuristic assumptions about how the solution should behave. Furthermore, even when prior information is available, overreliance on it can be dangerous at times. Its accuracy is often quite difficult to evaluate objectively, which can produce poor predictive performance especially long-term.

Dynamic processes on the other hand, where the measurements are recorded as a time-series

$\boldsymbol{y}(t)$, contain a substantial amount of information in the system dynamics. That is, if one can describe the underlying dynamical process with a mathematical model, this model can be used to regularize the problem by selecting solutions that best interpolate between the model and the data. Under some cases, this framework can provide accurate estimates of the system's trajectory, even when little or no prior information about its state.

While this *dynamical regularization* does utilize a type of prior information, but not in the usual sense of a *prior distribution* that reflects uncertainty in the initial conditions. Rather, the regularization comes from our *a-priori* belief in how the system evolves in time. Its accuracy depends on the chosen application. For instance, a deterministic model based on physical processes (such as transport or electromagnetism) and obeying strict rules (such as conservation of mass, energy, or momentum) carries considerably more confidence than a stochastic model describing say economic policies or voting tendencies, which are inherently more unreliable. The latter, while interesting in its own right, will not be considered here. Instead, the scope of this discussion will be limited to deterministic dynamical processes, where the time evolution of the system is known with absolute certainty. That is, the dynamical model is known precisely, and is without error.

## 3.2 Methodology

This study examines how the conditioning of the problem impacts our ability to observe, estimate, and predict complex chaotic behavior by establishing approximate lower bounds on the probability of success of the combined observation-analysis-forecast system. In general, this success rate depends on its three core components: the model, the data, and the algorithm. So a primary goal is to how the likelihood of success scales with key parameters of the problem — such as the choice of algorithm, number of measurements $L$, the length of the estimation window $T = N\,dt$ — and identify critical values at which the probability of success drops sharply.

A computational framework will be introduced to estimate these values numerically, through a series of Monte-Carlo calculations. It may be viewed as an extension of earlier work by [174]. In addition to providing insight into the complex relationships underlying predictability, observability, and conditioning, the results will also demonstrate the existence of fundamental limits to these endeavors, which persist even under near-perfect conditions.

These results will be illustrated using the simple (yet instructive) dynamical model known as *Lorenz 96*, which is now described.

### 3.2.1 Lorenz 96 and extensive chaos

The Lorenz 96 model was introduced in [118] to study the predictability of the atmosphere, as described by a field of $D \geq 3$ scalars discretized on a periodic lattice, so $x_{D+1} = x_1$. In their simplest form, the model equations are given by

$$\frac{dx_i}{dt} = x_{i-1}(t)\left(x_{i+1}(t) - x_{i-2}(t)\right) - x_i(t) + \theta(t),$$

where the first term is a crude model of advection, the second term models dissipation and the parameter $\theta(t)$ describes external forcing. Despite its relative simplicity, this model displays a wide range of dynamical behavior. For small values of the forcing parameter solutions decay to a fixed point $x_i = \theta$. For intermediate values, most solutions are periodic, and for larger values of $\theta$ the system is chaotic. For all the numerical experiments performed here, the numerical value of $\theta(t)$ is constant 8.17, so the system is chaotic. All simulations are numerically integrated using an explicit fourth-order Runge Kutta scheme with uniform time step $dt = 0.01$.

This model was chosen both for its simplicity and for the fact that it exhibits extensive chaos [94]. That is, for large enough forcing the fractal dimension of the attractor grows linearly with the model dimension $D$. This is shown in Fig. (3.1b), which plots the Kaplan-Yorke dimension ([92]) as a function of $D$ computed from distinct initial conditions on the attractor. This metric roughly coincides with the number of positive global Lyapunov exponents (GLEs), and indicates that it scales linearly as $D_a \sim 0.69D$. On the other hand, the value of the maximum GLE $\lambda_{\max}$ behaves nonlinearly. The asymptotic region begins around $D \approx 20$ and approaches $\lambda_{\max} \sim 1.8$ as $D$ gets large. It is thus important to understand that for $D < 20$ the chaotic properties of the system may not yet be fully developed.

### 3.2.2 The average dimension of the unstable subspace

A related statistic that is of particular interest here is the average dimension of the unstable subspace $D_u$. It is calculated by averaging the number of locally unstable directions along a given trajectory. For instance, rewriting the system as a discrete time map $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n)$, at each time step the Jacobian of the discrete map $\nabla \boldsymbol{F}(\boldsymbol{x}_n)$ can be decomposed in terms of its singular values $\boldsymbol{\sigma}_n$. The number of unstable directions at time $t_n$ is obtained by counting the number of components of $\boldsymbol{\sigma}_n$ that are greater than unity.

The statistic $D_u$ is then the time average of this quantity along a trajectory. Namely,

$$D_u = \frac{1}{N}\sum_{n=0}^{N}\sum_{i=1}^{D}\Theta[\sigma_i(t_n)+\varepsilon]$$

where $\Theta[x] = 1$ if $x \geq 1$ and 0 otherwise is the Heaviside theta function, and $\varepsilon = 10^{-3}$ is a small tolerance to include any directions that are near stable.

The value of $D_u$ is also plotted in Fig. (3.1b). It scales linearly $D_u \sim 0.40D$ with the model dimension, although at a different rate than $D_a$. Thus, the average dimension of the unstable subspace is not equivalent to the Kaplan-Yorke dimension, which is roughly the number of positive GLEs.

Although the discrete time version is perhaps more intuitive, the same calculation can be done in continuous time. Consider the discrete map of an Euler step, $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) = \boldsymbol{x}_{n+1} + dt\,\boldsymbol{f}(\boldsymbol{x}_n)$. The singular values of $\nabla \boldsymbol{F}_n = \boldsymbol{I} + dt\,\nabla \boldsymbol{f}_n$ are the square root of the eigenvalues of the positive definite matrix

$$|\nabla \boldsymbol{F}_n|^2 = \boldsymbol{I} + dt\,(\nabla \boldsymbol{f}_n + \nabla^\dagger \boldsymbol{f}_n) + dt^2\,|\nabla \boldsymbol{f}_n|^2.$$

The symmetric matrix $\nabla \boldsymbol{f}_n + \nabla^\dagger \boldsymbol{f}_n$ describes the local expansion and contraction of phase space and the quantity $D_u$ can be computed by counting the number of its positive eigenvalues. Of these three terms, it is the only one that can have negative eigenvalues. Yet surprisingly not much is known about the spectrum of the sum of an arbitrary matrix and its transpose.

This dynamical model instability turns out to be rather important when constructing filtered estimates. It generates localized error growth, which must be stabilized to achieve accurate estimates. Controlling this instability requires information from the observations, and becomes increasingly difficult when measurements are sparse. The most effective filtering algorithms perform this task efficiently and robustly, without destabilizing the underlying model.

### 3.2.3 Twin experiments

The Lorenz 96 model offers a unique testbed for assessing how predictability, observability and the performance of various algorithms scale with the dimension of the model. A framework for this is now introduced. The approach involves a series of *twin-experiments*, or numerical experiments with simulated data. This provides the ‘true solution, or solutions, though here it is generally assumed that the

solution is unique.[1] The merits of twin-experiments have long been established in the geophysical and data assimilation community, as they provide a systematic way of testing various assumptions and constraints of the estimation problem. When successful, these tests provide confidence that the methods will also succeed with real data, which will inevitably be more difficult.

The probability of success for a given algorithm is estimated through a series of Monte Carlo runs. These runs may be viewed as a set of Bernoulli trials having exactly two outcomes 'success' or 'failure', which provide a direct estimate of the *success ratio* or the number of successful runs over the number of trials. Trials can be performed both locally, on a given data trajectory sampling various initial conditions, or globally, over various data trajectories sampled around the attractor.

How the sampling is performed depends on the chosen algorithm. Different methods have different ways of representing the solution, and therefore different search spaces. The requisite details will therefore be given in the context of each specific type of algorithm.

### 3.2.4 Observations and observability

The following experiments are designed to assess the impact of system observability on various standard algorithms for dynamic state estimation. However, the term observability is used loosely here. That is, although the system is 'observable' in the sense that its Lie derivatives are always full rank (see *e.g.*, [103]), in practice as the number of observations become sparse the distinction between an observable and unobservable system is not always clear. This situation occurs, for instance when one state is only weakly coupled to the rest of the system, so it may be analytically observable but unobservable from a practical standpoint ([136]). Such problems may be considered dynamically ill-conditioned. Their solution typically requires the inversion of a poorly conditioned matrix at some point during the estimation process.

Generally speaking, as the observations become increasingly sparse one eventually looses the ability to resolve the 'true' state of the system, even when the model is known perfectly and the observations contain very little noise. To be clear, the term 'sparse' here refers to spatially sparse. Observations are generated at every time-step $dt = 0.01$, and frequent enough in time that the model is only weakly nonlinear.[2] The observations are also assumed to be projections onto a uniformly distributed set of states. That is, $\boldsymbol{h}(\boldsymbol{x}) \rightarrow \boldsymbol{H} \cdot \boldsymbol{x} = \{x_{\ell_1}, \ldots, x_{\ell_L}\}$ where the indices $\ell_n$ are chosen to distribute the observations as

[1] Roughly speaking, this follows from the assumption the that model is perfect and deterministic and the estimation window is long enough to break any degeneracy associated with a non-injective observation operator $\boldsymbol{h}(\cdot)$.

[2] For a related study using temporally sparse observations, see [146, 73]

evenly as possible across the $D$ states with $\ell_1 = 1$. Many other interesting observation schemes are available (*e.g.*, lumped or random observations), but these will be pursued elsewhere. Furthermore, the study of each particular algorithm begins with arguably the simplest case, where the model and the measurements are perfect, without error (at least to machine precision). The results are then reassessed after adding simulated noise in the observations. The impact of errors in the model will be considered as part of a future study.

### 3.2.5 Critical values

One of the primary goals of this investigation is the identification of *critical values* of certain parameters at which the problem becomes ill-conditioned enough that the algorithm fails to identify the 'true' solution. Specifically, the two parameters of interest here are: 1) the critical 'number' of measurements $L_c$ and 2) the critical radius $\rho_c = |\boldsymbol{x}^{(0)} - \boldsymbol{x}^*|$ between the initial guess and the true solution. These parameters describe fundamental limits in our ability to estimate and predict the system's behavior. The goal is to understand how their values depend on other parameters of the problem, such as the degree of dynamical instability in the model and the choice of algorithm.

The search for these critical parameters uses a modified form of *exponential search* ([18]) in which the same estimation procedure is rerun several times, changing the parameter logarithmically until failure occurs. This then gives an upper and lower bound on its value that can be further refined by binary search. For example to determine $L_c$, the algorithm is rerun several times until failure occurs, starting with full observations $L = D$ and systematically reducing the number of measurements by half on each iteration. The critical value is then found by running binary search on the interval $[L_{c^-}, L_{c^+}]$.

Similarly, for the critical radius $\rho_c$ a perturbation direction $\hat{\boldsymbol{\rho}}$ is chosen at random from a unit hypersphere. The initial condition is set to $\boldsymbol{x}^{(0)} = \boldsymbol{x}^* + r\,\hat{\boldsymbol{\rho}}$, where the distance $r$ starts small enough that the result converges. The radius is then doubled at each iteration until failure occurs, or $r$ gets large enough to span the entire attractor, at which point the algorithm terminates. The interval $[r_{c^-}, r_{c^+}]$ can then be refined to get $\rho_c$.

This procedure relies on the assumption that the output is ordered, so there is only one transition between success and failure. This is not always true however. For instance, with some additional restrictions (*e.g.*, on the signal-to-noise ratio) one expects that with $L = D$ the algorithm will always be successful, and generally speaking the success ratio should increase with the number of measurements. Nevertheless, it is possible for a successful run to fail as more measurements are added, although this is not common, at least for the cases considered here.

The critical radius on the other hand, will likely have several transitions as the radius $r$ is increased, since the basin of attraction for an optimization algorithm is generally fractal. Consequently, the focus here is restricted to approximating only the region in the immediate vicinity of a solution for which convergence is guaranteed.[3] The estimate $\rho_c$ therefore offers a lower bound on the probability of success, provided $r$ is initialized to be sufficiently small.

### 3.2.6 Two criteria for success

It remains however to specify how 'success' is determined. Since ultimately these algorithms will be applied to real data, where the solution is unknown, two criteria are used:

1. The true RMS error at the end of the estimation window

$$\text{RMSE} := \sqrt{\frac{1}{D}|x(N_{est}) - x^*(N_{est})|^2}$$

2. The observed RMS deviations averaged across the prediction window

$$\text{RMSD} := \sqrt{\frac{1}{L\,(N_{pred}+1)} \sum_{n=N_{est}}^{N_{est}+N_{pred}} |y_n - h(x_n)|^2_{\boldsymbol{R}_m}}.$$

where $N_{est}$ and $N_{pred}$ are the respective lengths of the estimation and prediction windows, and $\boldsymbol{R}_m$ is the inverse covariance matrices of the measurement error. Note that the true RMSE is only valid in a twin experiment, where the true solution $\boldsymbol{x}^*$ is known. Whereas the predictive RMSD is applicable to real data. The determination of $\rho_c$ (as described above) also requires knowing $\boldsymbol{x}^*$ and is therefore only valid in a twin experiment.

For the purposes of identifying the critical parameters $L_c$ and $\rho_c$, the estimate is considered successful when the true RMSE drops below a prescribed threshold. This cutoff necessarily depends on the choice of algorithm — *e.g.*, whether it performs filtering or smoothing. For the prediction error however, some additional work is needed to quantify success as only a subset of the model state can be directly observed. However, a rough estimate of the accuracy of the solution can be obtained by measuring the time $\Delta$ it takes for the prediction error to saturate. The fact that the prediction error grows roughly as $\exp \lambda_{\max} t$ allows us to estimate the magnitude of the true error at the end of the estimation window as roughly $\sigma_x^2 \exp -\lambda_{\max} \Delta$, where $\sigma_x$ is variance of the measured states after saturation.

[3] The existence of this *immediate basin of attraction* will be discussed more in more detail in Sec. (3.4.5)

An example is shown in Fig. (3.2), which plots the prediction error as a function of $N_{pred}$, for $D = 20$, $L = 1$ for observations contaminated with a small amount of noise, drawn from a normal distribution with $N(0,1)$. The black line shows the growth of the true RMSE and the red dashed line shows the observed RMSD. Their growth is well-approximated by $\exp[\lambda_{\max}\Delta]$ where $\Delta$ is the time it takes for the prediction error to saturate at roughly $\sigma_x \approx 5$. In the following section, it will be shown that this empirical measure of success provides a reasonably accurate approximation to the 'true' $L_c$, computed using the RMSE.

Of course in practice the situation is much more complicated. The measurements and model have errors, dynamical variables have different scaling, the observation operator may be a complicated nonlinear function ,and the system may not even have global Lyapunov exponents. While it is difficult to see how this argument may be extended to the general case, this issue of estimating the true error of a solution using an empirically viable metric such as prediction has been remarkably absent from the literature. Furthermore, it is worth pointing out that this is in fact a *benefit* of a chaotic system. Namely, (at least in the simple case considered here) one can use $\lambda_{\max}$ as a natural measure for the growth rate of errors to estimate the magnitude of the true error based on predictability.

The use of prediction RMSD to evaluate success reflects the broader fact that twin experiments admit two paradigms of study: theoretical and empirical. A theoretical study assumes knowledge of the true solution, whereas an empirical study does not. While this assumption of course precludes any application to real data, it is useful nonetheless, as the results provide important information about the constraints of the problem — such as the required number of observations or the accuracy of the initial guess needed for convergence. This information can then be used to inform the design of real experiments.

## 3.3 The number of observations required for synchronization

The critical minimum number of observations $L_c$ is now assessed for a few simple filtering schemes. Recall that such methods estimate only the final state $\boldsymbol{x}_{N|N}$ given the observations $\mathcal{Y}_{0:N}$. In discrete time, they may be generally written in terms of an observe-analyze-forecast cycle

$$\begin{aligned} \boldsymbol{x}_{n|n} &= \boldsymbol{x}_{n|n-1} + \boldsymbol{K}_n \cdot (\boldsymbol{y}_n - \boldsymbol{H} \cdot \boldsymbol{x}_{n|n-1}) \\ \boldsymbol{x}_{n+1|n} &= \boldsymbol{F}(\boldsymbol{x}_{n|n}). \end{aligned} \tag{3.2}$$

The first equation incorporates the information from the measurement at time $t_n$ to produce the 'analysis' $\boldsymbol{x}_{n|n}$. The second equation evolves the estimate using the discrete time version of the full nonlinear model $\boldsymbol{F}(\cdot)$.

This observation error feedback form is quite universal among the various approaches devised for treating this problem. In optimal estimation it describes the Luenberger observer, the Kalman filter, and their various nonlinear extensions. In data assimilation, it is shared by optimal interpolation and 3DVar. In dynamical systems theory if forms the basis for a variety of feedback synchronization methods. These methods are mainly distinguished by their choice of the coupling gain matrix $\boldsymbol{K}_n$, which often determines both the accuracy and the computational efficiency of the algorithm.

Although the number of approaches that have this general form is truly vast, the focus here is restricted to three specific techniques: 1) the extended Kalman filter 2) 3DVar and 3) the synchronization method of [157]. In this order, each method may be viewed as an approximation of its predecessor that reduces the computational overhead associated with the matrix $\boldsymbol{K}_n$. That is, the extended Kalman filter computes $\boldsymbol{K}_n$ dynamically, by integrating as set of $D \times D$ ordinary differential equations and is therefore the most computationally intensive, but also the most accurate. By contrast, given the assumptions made here 3DVar uses a static $\boldsymbol{K}_n \to \boldsymbol{K}$, so it does not require any additional integration. Pecora-Carroll synchronization takes this approximation one step further, and may be viewed as a limit of 3DVar where $\boldsymbol{K} \to \infty$. It is thus slightly more efficient as it avoids the additional matrix multiplication needed to compute the coupling term.

### 3.3.1 Kalman filtering, 3DVar, and synchronization

Of the three methods considered here, the extended Kalman filter (ExtKF) has the most rigorous theoretical foundations. It is worth pointing out however that it is typically derived from statistical considerations, where both the model and the observations are subject to Gaussian noise processes, it is equally valid (and often works quite well) for deterministic problems. Although it took quite a while from its initial discovery in the early 1960s for its properties as an exponential observer to be proven [181]. However, the fact that we are still finding fundamental features of this solution — one of which will be demonstrated shortly — suggests parts of it are still not as well-understood as perhaps they should be, given its undeniable importance.

There are quite a few versions of the ExtKF in the literature. So many in fact, that it is often difficult to determine *a-priori* which one is most appropriate for a given situation. However, since the focus

is on poorly observable systems, a square-root factorization method will be used to improve the stability of the Riccati equation for the covariance, which can often become highly ill-conditioned as the number of observations becomes increasingly sparse. The covariance matrices are stored and manipulated using their symmetric square roots or Cholesky decompositions, defined here as $\boldsymbol{R}^{-1} = \boldsymbol{R}^{-\dagger/2} \cdot \boldsymbol{R}^{-1/2}$ where $\boldsymbol{R}^{-1/2}$ is an upper-right triangular matrix. Following [153], the respective time and measurement updates of the covariance can be computed directly using QR decomposition,

$$\boldsymbol{\Theta} \cdot \begin{bmatrix} \boldsymbol{R}_{n+1|n}^{-1/2} \\ \boldsymbol{0} \end{bmatrix} = \mathrm{QR} \begin{bmatrix} \boldsymbol{R}_{n|n}^{-1/2} \cdot \nabla^{\dagger} \boldsymbol{F}_n \\ \boldsymbol{R}_f^{-1/2} \end{bmatrix}$$

$$\boldsymbol{\Theta} \cdot \begin{bmatrix} \boldsymbol{C}_n^{-1/2} & \boldsymbol{C}_n^{-1/2} \cdot \boldsymbol{K}_n^{\dagger} \\ \boldsymbol{0} & \boldsymbol{R}_{n|n}^{-1/2} \end{bmatrix} = \mathrm{QR} \begin{bmatrix} \boldsymbol{R}_m^{-1/2} & \boldsymbol{0} \\ \boldsymbol{R}_{n|n-1}^{-1/2} \cdot \boldsymbol{H}_n^{\dagger} & \boldsymbol{R}_{n|n-1}^{-1/2} \end{bmatrix}.$$

In these equations, $\boldsymbol{\Theta}$ is a unitary matrix, and

$$\begin{aligned} \boldsymbol{C}_n^{-1} &:= \boldsymbol{R}_m^{-1} + \boldsymbol{H} \cdot \boldsymbol{R}_{n|n-1}^{-1} \cdot \boldsymbol{H}^{\dagger} \\ \boldsymbol{K}_n &:= \boldsymbol{R}_{n|n-1}^{-1} \cdot \boldsymbol{H}^{\dagger} \cdot \boldsymbol{C}_n = \boldsymbol{R}_{n|n}^{-1} \cdot \boldsymbol{H}^{\dagger} \cdot \boldsymbol{R}_m. \end{aligned}$$

Furthermore, using the fact that $\boldsymbol{\Theta}^{\dagger} \cdot \boldsymbol{\Theta} = \boldsymbol{I}$ along with the Sherman-Morrison-Woodbury matrix identity ([209]),[4] one can verify the update equations

$$\begin{aligned} \boldsymbol{R}_{n+1|n}^{-1} &= \nabla \boldsymbol{F}_n \cdot \boldsymbol{R}_{n|n}^{-1} \cdot \nabla \boldsymbol{F}_n^{\dagger} + \boldsymbol{R}_f^{-1} \\ \boldsymbol{R}_{n|n} &= \boldsymbol{R}_{n|n-1} + \boldsymbol{H}^{\dagger} \cdot \boldsymbol{R}_m \cdot \boldsymbol{H} \end{aligned}$$

Note that for the measurement update, the QR decomposition provides both the updated covariance $\boldsymbol{R}_{n|n}^{-1/2}$ as well as the coupling gain $\boldsymbol{K}_n$. It only requires the inverse of $\boldsymbol{C}_n^{-1/2}$, which can be efficiently computed using back-substitution since it is an upper-right triangular matrix.

One also has to specify the covariance matrices for the model and observation errors $\boldsymbol{R}_f^{-1}$, $\boldsymbol{R}_m^{-1}$ as well as the initial covariance $\boldsymbol{R}_0^{-1}$. Tuning these parameters is one of the most time-consuming tasks involved in implementing an ExtKF. This is especially so here, given the assumptions that the model and data are perfect, and that one has no prior knowledge about the initial state $\boldsymbol{x}_0$. The latter assumption also technically requires using an information filter, since $\boldsymbol{R}_0^{-1}$ is undefined. But this comes with its own issues as it is not possible to evolve the information state $\boldsymbol{z}_n := \boldsymbol{R}_{n|n} \cdot \boldsymbol{x}_{n|n}$ without linearizing the dynamics. In the

[4] $(\boldsymbol{A} + \boldsymbol{U} \cdot \boldsymbol{C} \cdot \boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1} \cdot \boldsymbol{U} \cdot (\boldsymbol{C}^{-1} + \boldsymbol{V} \cdot \boldsymbol{A}^{-1} \cdot \boldsymbol{U})^{-1} \cdot \boldsymbol{V} \cdot \boldsymbol{A}^{-1}$

end, the best and most consistent results were obtained using the simple choice $\boldsymbol{R}_f^{-1} = \boldsymbol{R}_m^{-1} = \boldsymbol{R}_0^{-1} = \boldsymbol{I}$. This also held true when simulated noise was added to the data, albeit with one important caveat that will be discussed in the context of those results.

By contrast, 3DVar methods approximate the solution to the ExtKF by ignoring the covariance update altogether. Instead, they use a fixed prior 'background' distribution $\boldsymbol{R}_b$ that penalizes deviations from a state $\boldsymbol{b}_n$, which typically is the result of the previous forecast $\boldsymbol{b}_n \to \boldsymbol{x}_{n|n-1}$. With these assumptions, the measurement update can be recast as an optimization problem

$$\boldsymbol{x}_{n|n}^{(\text{3DVar})} = \underset{\boldsymbol{x}}{\text{argmin}}\ \frac{1}{2}|\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}|^2_{\boldsymbol{R}_m} + \frac{1}{2}|\boldsymbol{x}_{n|n-1} - \boldsymbol{x}|^2_{\boldsymbol{R}_b},$$

which amounts to performing Tikhonov regularization on the measurement term ([82]). When the observation operator is linear, the optimization problem is convex and thus admits a closed form solution

$$\begin{aligned}\boldsymbol{x}_{n|n}^{(\text{3DVar})} &= (\boldsymbol{H}^\dagger\cdot\boldsymbol{R}_m\cdot\boldsymbol{H} + \boldsymbol{R}_b)^{-1}\cdot\boldsymbol{H}^\dagger\cdot\boldsymbol{R}_m\cdot(\boldsymbol{y}_n + \boldsymbol{R}_b\cdot\boldsymbol{x}_{n|n-1}) \\ &=: \boldsymbol{x}_{n|n-1} + \boldsymbol{K}^{(\text{3DVar})}\cdot(\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_{n|n-1}).\end{aligned}$$

The coupling gain matrix is time-independent

$$\begin{aligned}\boldsymbol{K}^{(\text{3DVar})} &= (\boldsymbol{H}^\dagger\cdot\boldsymbol{R}_m\cdot\boldsymbol{H} + \boldsymbol{R}_b)^{-1}\cdot\boldsymbol{H}^\dagger\cdot\boldsymbol{R}_m \\ &= \boldsymbol{R}_b^{-1}\cdot\boldsymbol{H}^\dagger\cdot(\boldsymbol{R}_m^{-1} + \boldsymbol{H}\cdot\boldsymbol{R}_b^{-1}\cdot\boldsymbol{H}^\dagger)^{-1},\end{aligned}$$

and the latter result again follows from the Sherman-Morrison-Woodbury matrix identity.

As with the ExtKF, the most difficult part of implementing this solution is arguable determining the proper values for $\boldsymbol{R}_b$ and $\boldsymbol{R}_m$. In practice, estimating $\boldsymbol{R}_m$ tends to be easier as one usually has some idea of the noise processes underlying the observations. The background matrix $\boldsymbol{R}_b$ on the other hand, must be determined from the statistics of the system. Although a number of methods exist for approximating its value, no one technique stands out above the rest.

Here however, two limits are of interest. Assuming $\boldsymbol{R}_m^{-1} \to \alpha\boldsymbol{I}$ and $\boldsymbol{R}_b^{-1} \to \beta\boldsymbol{I}$, the limit $\alpha \to 0$ with constant $\beta$ gives

$$\boldsymbol{x}_{n|n}^{(\text{3DVar-FB})} = \lim_{\alpha\to 0^+}\boldsymbol{H}^\dagger\cdot\left(\frac{\alpha}{\beta}\boldsymbol{I} + \boldsymbol{H}\cdot\boldsymbol{H}^\dagger\right)^{-1}\cdot(\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_{n|n-1}) = \boldsymbol{H}^\dagger\cdot(\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_{n|n-1})$$

so $\boldsymbol{x}_{n|n} = \boldsymbol{H}^\dagger\cdot(\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n)$. This results follows from that fact that the pseudoinverse of a general matrix

can be represented as $\boldsymbol{A}^{+} = \lim_{\delta \to 0^{+}} \boldsymbol{A}^{\dagger} \cdot (\delta \boldsymbol{I} + \boldsymbol{A} \cdot \boldsymbol{A}^{\dagger})^{-1}$. Furthermore since $\boldsymbol{H}$ is a unitary projection matrix, it follows that $\boldsymbol{H}^{+} \equiv \boldsymbol{H}^{\dagger}$.

With these assumptions, this limit of 3DVar essentially coincides with feedback synchronization techniques found in the dynamical systems literature ([4, 1]). These methods are often combined the analysis and forecast updates into one equation[5]

$$\boldsymbol{x}_{n+1}^{(\text{3DVar-FB})} = \boldsymbol{F}(\boldsymbol{x}_n) + \gamma \boldsymbol{H}^{\dagger} \cdot (\boldsymbol{y}_n - \boldsymbol{H} \cdot \boldsymbol{x}_n). \tag{3.3}$$

The strength of the coupling is controlled by a scalar parameter $\gamma$ and must be chosen appropriately. It it is too large, it will destabilize the system. And if it is too small, its impact will be negligible. Here a value of $\gamma dt = 0.01$ is used, which is known to be successful with Lorenz 96 when it is sufficiently observed ([101]). Also note that no matrix calculations are needed to compute the perturbation, so the method is less computationally intensive than the EKF.

Alternatively, one may also consider the limit $\beta^{-1} \to 0$ with $\alpha$ fixed, which gives

$$\boldsymbol{x}_{n|n}^{(\text{3DVar-PC})} = \lim_{\beta^{-1} \to 0^{+}} (\boldsymbol{H}^{\dagger} \cdot \boldsymbol{H} + \frac{\alpha}{\beta} \boldsymbol{I})^{-1} \cdot \boldsymbol{H}^{\dagger} \cdot (\boldsymbol{y}_n + \beta^{-1} \boldsymbol{x}_{n|n-1}) = \boldsymbol{H}^{\dagger} \cdot \boldsymbol{y}_n.$$

At each observation time, the measured components of $\boldsymbol{x}_{n|n}$ are simply replaced with the observations $\boldsymbol{y}_n$, while the unmeasured components are identically zero. If instead the unmeasured components are left unchanged, this limit would coincide with the method of Pecora and Carroll ([157]) — one of the first techniques developed to demonstrate synchronization in chaotic dynamical systems. It is one of the simplest estimation techniques available, as it does not require tuning any parameters, and even avoids having to compute the innovation: $\boldsymbol{y}_n - \boldsymbol{H} \cdot \boldsymbol{x}_{n|n-1}$.

### 3.3.2 The minimum observability threshold $L_c$

Approximate values for the critical minimum number of observations $L_c$ are now computed as described above as a function of $D$ for the three methods: the extended Kalman filter (EKF), feedback synchronization (3DVar-FB), and Pecora-Carroll synchronization (3DVar-PC). The initial conditions of the estimates are selected either: 1) locally, as a a random perturbation $N(0, 10^{-3})$ from the true state, or 2)

[5] There is a difference between Eqns. (3.2) and 3.3 regarding whether the perturbation is applied to the state itself or as a perturbation to the model equations. This may be considered as distinguishing between jump and drift processes in the underlying probability distribution, although it is unclear which approach is better, as they have never been directly compared.

globally, from an arbitrary point on the attractor.

These choices respectively reflect best and worst case scenarios, with the local case corresponding to an exceptionally lucky initial guess. It is meant to test the difference between the number of observations required to bring the system to a synchronized state, versus the number required to keep the system in a synchronized state. That is, it tests whether the algorithm is capable of controlling the localized chaotic instabilities in the model to keep the estimate close, given limited information from the observations. The estimation process is repeated several times, sampling different initial conditions until the statistics converge.

All simulations are performed using the same dataset taken from a single random location on the attractor. Two situations are considered; the data is either: 1) without noise, or 2) contains additive uncorrelated Gaussian noise $N(0,1)$. While it would be interesting compare results in a 'global' sense, where the data is also sampled randomly across the attractor, this will not be considered. Preliminary tests show the results given here are indicative of the attractor as a whole, and do not vary much based on the chosen initial conditions for the data.

Estimates are considered successful if their final 'true' RMSE is either within the noise ball or below $10^{-3}$. For predictability estimates, the same holds true except that the 'true' RMSE is estimated from predictive RMSD as described above. The length of the estimation and prediction windows are scaled by the largest Lyapunov exponent $\lambda_{\max}$ for each value of of $D$. Estimates are given ample time to converge (up to $100\lambda_{\max}$) and simulations are terminated early if convergence occurs before this time. Similarly, predictions are run until the RMSD saturates, so that the RMSE can be estimated from the time $\Delta$ it takes for this to occur.

Results shown in Fig. (3.3) depict six of eight possible combinations. Local estimates based on predictive RMSD are not given since these cannot be constructed in a true experiment. As expected, local estimates perform slightly better than global estimates. Predictability estimates based on RMSD provide a reasonable approximation of the true RMSE, but slightly underestimate the mean critical threshold $L_c$, computed by linear regression. This is especially so when observation noise is present.

In all cases, $L_c$ increases linearly with the resolution of the model. Without noise, the EKF does better than both 3DVar methods. But this efficiency requires an extra $D \times D$ coupled ODEs. Among the simpler approaches, 3DVar-FB consistently outperforms 3DVar-PC, and all methods require less than the analytical upper bound $L_c < 2/3\,D$ given by [106, 105] for a specific static observation operator.

The addition of observation noise effectively erases any benefits of the EKF, regardless initial

conditions. And the 15% decrease in the 3DVar-FB threshold is somewhat surprising, given the simplicity of the algorithm. These issues will be addressed further momentarily.

For 3DVar methods, the mean $L_c$ also falls roughly in the range $D_u < L_c < D_a$ given in Fig. (3.1b). That is, it lies between the time-averaged number of unstable dimensions in the model, and the number of positive global Lyapunov exponents. This supports the earlier statements about the role of chaos in determining $L_c$, and suggests that its linear scaling with the resolution $D$ is no accident, but a direct consequence of extensive chaos in the forecast model.

### 3.3.3 Synchronization, phase transitions, and the role of the unstable subspace

Constructing successful estimates using a dynamical filtering scheme requires the estimate and truth to synchronize in a way that the RMSE remains small and bounded. Although somewhat obvious in hindsight, this remarkable fact has been rediscovered and reinterpreted in many different domains and contexts. For instance, in control theory one often constructs 'observer' systems that seek to drive the error to zero asymptotically. While this term is typically reserved for cases where the underlying systems are deterministic, without modeling errors in the observations or the dynamics, it is nonetheless useful here for understanding the basic limits of system observability.

For this it is convenient to rewrite Eqn. (3.2) in a form analogous to Eqn. (3.3), namely

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{K}_n \cdot (\boldsymbol{y}_n - \boldsymbol{H} \cdot \boldsymbol{x}_n).$$

Also, let the true state be described by $\boldsymbol{x}^*_{n+1} = \boldsymbol{F}(\boldsymbol{x}^*_n)$ and the observations $\boldsymbol{y}_n = \boldsymbol{H} \cdot \boldsymbol{x}^*_n$. The the error $\boldsymbol{e}_n := \boldsymbol{x}^*_n - \boldsymbol{x}_n$ evolves as

$$\begin{aligned} \boldsymbol{e}_{n+1} &= \boldsymbol{F}(\boldsymbol{x}^*_n) - \boldsymbol{F}(\boldsymbol{x}_n) - \boldsymbol{K}_n \cdot \boldsymbol{H} \cdot \boldsymbol{e}_n \\ &\approx \big(\nabla \boldsymbol{F}_n(\boldsymbol{x}_n) - \boldsymbol{K}_n \cdot \boldsymbol{H}\big) \cdot \boldsymbol{e}_n. \end{aligned} \tag{3.4}$$

The above expansion is performed around the estimated state $\boldsymbol{x}^*_n \to \boldsymbol{x}_n + \boldsymbol{e}_n$ and is expected to hold when the error is small. The coupling matrix $\boldsymbol{K}_n$ affects the Lyapunov structure of the error dynamics, and should be selected to produce a contraction that drives it to zero asymptotically. When the system dynamics are linear and time-invariant this amounts to choosing $\boldsymbol{K}$ so that the eigenvalues of the resulting matrix $\boldsymbol{F} - \boldsymbol{K} \cdot \boldsymbol{H}$ are all inside the unit circle in the complex plane.[6] This is rather easy when the observation operator $\boldsymbol{H}$ is full rank, as choosing $\boldsymbol{K} \to (\boldsymbol{F} - \boldsymbol{G}) \cdot \boldsymbol{H}^{-1}$ will produce any desired error dynamics $\boldsymbol{G}$.

[6] In continuous time they must all be in the left half of the complex plane.

Of course, this is not very useful as assuming $\boldsymbol{H}$ is invertible implies the exact true solution is known immediately from the first set of measurements.

The more realistic case is not so straightforward. Restricting the rank of $\boldsymbol{H}$ constrains the directions along which the resulting closed loop feedback can be applied. To put it another way, let $\boldsymbol{K} =: \boldsymbol{A} \cdot \boldsymbol{H}^{\dagger}$ and rewrite Eqn. (3.4) in the following form

$$\boldsymbol{e}_{n+1} = \big(\boldsymbol{F} - \boldsymbol{A}\big) \cdot \boldsymbol{e}_n + \boldsymbol{A} \cdot \boldsymbol{P} \cdot \boldsymbol{e}_n.$$

Given our assumptions on $\boldsymbol{H}$, the matrix $\boldsymbol{P} := \boldsymbol{I} - \boldsymbol{H}^{\dagger} \cdot \boldsymbol{H}$ is the projection onto the unobserved subspace. The matrix $\boldsymbol{A}$ can be chosen at will, but if it contains elements in this subspace it can potentially destabilize the solution. On the other hand, if $\boldsymbol{A}$ is constrained to be orthogonal to this subspace so that $\boldsymbol{A} \cdot \boldsymbol{P} = 0$ then this term will not affect the solution. Thus, the problem of choosing $\boldsymbol{K}$ comes down to a structured inverse eigenvalue problem ([39]) otherwise known in optimal control as pole placement or pole assignment for which a variety of robust algorithms exist (*e.g.*, [96]).

The situation becomes more complicated however when the dynamics are nonlinear or time-varying. As pointed out by [180], simply choosing $\boldsymbol{K}_n$ such that the eigenvalues of $\boldsymbol{F}_n - \boldsymbol{K}_n$ are inside the unit circle at all times $t_n$ does not guarantee the overall product will go to zero. On the other hand, if one selects $\boldsymbol{K}_n$ so that the *singular values* of $\boldsymbol{F}_n - \boldsymbol{K}_n$ are all less than unity, the sequence is guaranteed to converge. Thus, it may be considered an inverse singular value problem, which as noted by [38] can always be converted to an inverse eigenvalue problem by using the fact that the eigenvalues of the symmetric matrix

$$\begin{bmatrix} \mathbf{0} & \nabla \boldsymbol{G}_n \\ \nabla^{\dagger} \boldsymbol{G}_n & \mathbf{0} \end{bmatrix} \tag{3.5}$$

are plus and minus the singular values of the matrix $\nabla \boldsymbol{G}_n := \nabla \boldsymbol{F}_n - \boldsymbol{K}_n \cdot \boldsymbol{H}$. This augmentation from $D \to 2D$ is reminiscent of the mapping to canonical coordinates discussed in Chap. (2). This idea will be revisited in a moment.

When the system is chaotic, the matrix $\boldsymbol{K}_n$ must be selected to control the unstable subspace of the dynamics so that the conditional Lyapunov exponents (*i.e.*, the GLEs of the coupled system) are all negative. This condition is necessary to synchronize two chaotic systems ([157]), and proceeds like a phase transition in $L$, in that either: (1) this condition is met and the systems can synchronize or (2) it is not and they do not. There is considerable freedom however as to how this condition may be satisfied. For instance, choosing $\boldsymbol{K}_n$ such that the singular vales of $\nabla \boldsymbol{G}_n$ are less than one at every $t_n$ will guarantee

convergence at least locally — provided the error is small enough that expansion in Eqn. (3.4) is valid. Yet this condition is a still bit strict, as it only needs to hold in the asymptotic sense. In other words, at any given time the largest singular value of $\nabla \boldsymbol{G}_n$ may be greater than one, as long as the overall product tends to zero.

### 3.3.4 Approximate bounds on $L_c$

Note for both 3DVar methods, the accuracy of the approximation $L_c \approx D_u \approx 0.4D$. The marked increase in performance shown by all three methods when $L$ is above $D_u$ indicates the latter may be useful as an approximate upper bound on $L_c$. The threshold $L_c \approx 0.4D$ also coincides with the transition of the maximum conditional Lyapunov exponent to a negative value, as shown in Fig. (3.4a). While this result may not hold in general for models without the homogeneous scaling and symmetry properties of Lorenz 96, it nonetheless highlights the need for adequate observations to control the unstable subspace.

The correspondence between $L_c$ and $D_u$ also suggests that perhaps each additional measurement stabilizes an additional component of the unstable subspace. This hypothesis is tested now averaging the number of singular values of $\nabla \boldsymbol{G}_n := \nabla \boldsymbol{F}_n - \boldsymbol{K}_n$ that satisfy $\sigma_m \leq 1 - \varepsilon$, with $\varepsilon = 1.e-3$. This adimensionalized conditional fraction of the unstable subspace $D'_u/D$ is plotted in Fig. (3.4b) and shows no remarkable change at this value however. This reiterates the important but subtle point made above. Although the observational feedback aims to control the growth of errors along these unstable directions, the necessary conditions for convergence are established by the time ordered product of local Jacobian matrices $\mathcal{T}_{n=0}^{N} \nabla \boldsymbol{G}_n$. Interestingly though, the values of $D'_u/D$ are quite consistent across dimensions.

Regarding an approximate lower bound on $L_c$, [106, 105] used an adaptive observation operator $\boldsymbol{H}_n$ that stabilizes only the first $r$ components of the unstable subspace. This requires full observations $L = D$, but the rank of the observation operator $\boldsymbol{H}_n$ is limited to $r$. A similar calculation is performed here. Results are shown in Fig. (3.5) for local initial conditions with no noise. This provided the lowest thresholds, with 3DVar-FB only needing $r_c \approx 0.14D$, and the EKF requiring even fewer $r_c \approx 0.015D$. Note also that under similar conditions but without the adaptive observation operator, the EKF requires $L_c \approx 0.04D$. This suggests that it is operating near optimal efficiency when the estimate is initialized near the truth.

For 3DVar-FB, these results are consistent with those of [105], who estimated $L_c \approx 0.15D$. But they only reported a marginal reduction $L_c \approx 0.12D$ for their adaptive EKF, whereas here it is roughly ten-fold. This large discrepancy is partially due to the choice of local initial conditions. For global initial

conditions, the adaptive EKF requires roughly $L_c \approx 0.08\,D$. While this is still lower the previous estimate, the difference is not nearly as large.

### 3.3.5 Assimilation in the unstable subspace (AUS)

Although it is not possible to construct this adaptive observation operator for systems that are only partially observed, these results suggest that targeting unstable manifold may help effectively reduce $L_c$. This idea is the basis for a recent strategy known as assimilation in the unstable subspace (or AUS, [196]) that projects the control perturbations to the unstable (and null) subspace. The justification is that by ignoring the stable directions where the error is naturally contracting, this approach avoids introducing unnecessary errors into the analysis.

The effect of this technique on $L_c$ is now examined, for both 3DVar-FB and the EKF. Both algorithms project their resulting control perturbation into the locally unstable (and null) subspace, but the EKF does so by projecting the analysis covariance. Results shown in Fig. (3.6) indicate that this approach improves the observational efficiency of 3DVar-FB by $5-10\,\%$.

With the EKF however, the results are markedly worse. The degraded performance disappears gradually as the tolerance distinguishing between the stable and unstable/null modes is decreased, thereby increasing the effective rank of the estimated error covariance. But the best results are obtained using the full rank EKF, without AUS.

This statement is somewhat at odds with that of [195], who suggest that AUS does not degrade the performance of the EKF. That study assumed sufficient observability however ($L = 0.5\,D$), and did not consider the observational efficiency of the method. The results here suggest that under ideal circumstances, the full rank EKF operates close to the peak observational efficiency given by the approximate lower bound in Fig. (3.5). Artificially projecting the analysis covariance into the unstable subspace appears to disrupt the observational benefits of using the EKF.

### 3.3.6 The collapse of the error covariance

The mechanism underlying the observational efficiency of the EKF is now further examined. While the precise origins of this phenomenon are not well understood, the answer is almost certainly related to the fact that, when $\boldsymbol{R}_f^{-1} = \mathbf{0}$, the Riccati equation for the estimated error covariance collapses around the unstable subspace. Notwithstanding over half a century of dedicated research into the Kalman filter, it was reported in the literature only quite recently ([194]) with numerical evidence from simulations

with the Lorenz 96 model. Formal proofs were then given by [66, 23] for the linear discrete-time case, although the numerical evidence implies these results generalize to the nonlinear case as well.

These calculations are repeated here, by running the EKF with $\boldsymbol{R}_f^{-1} = \mathbf{0}$ along the known solution until the rank of $\boldsymbol{R}_n^{-1}$ (computed with SVD) converged. The results shown in Fig. (3.7) indicate the asymptotic rank of the filter $r_\infty \sim 0.39\,D$ is in excellent agreement with $D_u$. Choosing $|\boldsymbol{R}_f^{-1}| > \mathbf{0}$ inflates the error covariance, putting a lower bound on the singular values of $\boldsymbol{R}_n^{-1}$ associated with the stable subspace. This improves the conditioning and stability of the filter by preventing it from becoming singular ([177]). The same idea is also used to stabilize ensemble Kalman methods ([51]).

Thus, the Riccati equation adaptively targets the unstable subspace of the dynamics, converging to it in the limit $\boldsymbol{R}_f^{-1} \to \mathbf{0}$. But this is not all. It also controls the growth of errors along these directions in a way that requires fewer observations than other methods. While this process appears to adaptively solve the extended inverse eigenvalue problem described by Eqn. (3.5), by shifting the poles of the linearized error matrix $\boldsymbol{G}_n$ to negate the maximum conditional GLE, the precise mechanism for this remains unclear.

This underscores an important limitation of current Kalman filtering theory. While its statistical properties have been exhaustively studied, its geometric aspects are not nearly as well established. Indeed, it took several decades before the extended Kalman filter was even proven to be an asymptotic observer ([181]). Nowhere is this more evident than in the fact that this rather fundamental property regarding the collapse of the error covariance to the unstable subspace has only just been discovered, and remains to be proven in the nonlinear case.

Thus, it is perhaps worth revisiting some of the original Kalman theory with the goal elucidating the inherent ties between its probabilistic and geometric interpretations. Based on the discussion in Chap. (2), it appears that symplectic structure plays a role in this observational efficiency, although the details are still unclear. Forging stronger links between the problem's statistical foundations and its optimal geometric substructure, will not only improve our overall understanding of one of the most singularly important algorithms ever devised, but will also motivate new and innovative approaches that exploit these optimal characteristics in a computationally efficient way.

### 3.3.7 Filtering observation noise through adaptive annealing of $\boldsymbol{R}_f$

Returning now to the poor performance of the EKF with observation noise, further examination showed many cases in which the true RMSE stabilizes to a value well below the average attractor distance, but slightly higher than the noise level. The results are therefore considered a failure, even though it does a

reasonable job of locating and tracking the truth. While this issue could be fixed by simply modifying the tolerance for success, it is perhaps better to look for a more systematic solution.

These results indicate that the observation noise is not being adequately filtered. This is a rather common issue with the EKF, indicating poorly balanced covariance matrices $\boldsymbol{R}_m$, $\boldsymbol{R}_f$ where too much weight is put on the observations. The obvious solution is to increase the relative value of $\boldsymbol{R}_f$, to put more emphasis on the dynamics. But doing so too early on in the estimation process can destabilize the results.

While the optimal tuning of these parameters has received widespread attention in the literature ([125, 126, 19]), a simpler approach will be pursued here. The idea is to give the observations more weight initially, and then slowly increase the influence of the dynamics as the filter converges. This improves stability of the estimate, when it is far from the truth, and filters out the observational noise as the solution converges.

This technique may be seen as a form of numerical continuation ([7]) or equivalently an 'annealing' of the model error. A similar approach has been recently used to improve the stability and convergence of variational estimation algorithms ([213, 214]), which will be discussed in Sec. (3.4). But this idea has not to my knowledge been applied to the EKF, although it is closely related to covariance inflation techniques needed to stabilize and improve the performance of ensemble Kalman filtering methods. The difference however is that, covariance inflation *decreases* $\boldsymbol{R}_f$ by introducing a small amount of artificial model error, whereas here $\boldsymbol{R}_f$ is increased *increased* as the estimate converges to reduce its reliance on the noisy observations.

While a variety of implementations were considered, one stood out both in terms of improved performance and overall simplicity. The idea is to link magnitude of the model error covariance to the observed RMSD, $|\boldsymbol{R}_f^{-1}| = \frac{\alpha}{\sqrt{L}}|\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n|_{\boldsymbol{R}_m}$, where $\alpha$ is a scale factor. This simple feedback rule passes the basic limiting tests. For instance, when the estimate is poor $|\boldsymbol{R}_f^{-1}|$ is large, say roughly on the scale of the size of the attractor. As the filter converges, the RMSD will tend towards unity, so $|\boldsymbol{R}_f^{-1}| \to \alpha$. In addition, when $|\boldsymbol{R}_m|$ is relatively large the RMSD will be small. This pushes $|\boldsymbol{R}_f^{-1}|$ lower to help rebalance the contributions between the observations and the model.

The benefits of this technique on $L_c$ in the presence of observation noise ($\sigma_m = 1$) are shown in Fig. (3.8). For these simulations, $\alpha = 1$ was sufficient. Compared with Figs. (3.3b) and (3.3d), the annealing technique reduces $L_c$ from $0.37D$ to $0.25D$. Although not quite as low as the noiseless case, it nonetheless reduces the value well below the approximate threshold of $D_u \approx 0.4D$ for non-adaptive methods.

These results were included as a brief example of how the adaptive tuning of the model error covariance $\boldsymbol{R}_f^{-1}$. The idea will be revisited in the next section, within the context of variational smoothing methods.

## 3.4 Conditioning and observability in 4DVar methods

The discussion now shifts from filtering to methods for solving the fixed-interval smoothing problem. Similar questions regarding conditioning and observability will be examined for a certain subset of techniques, known in the geophysics and numerical weather prediction literature as 4DVar. Estimates of both $L_c$ and the critical radius $\rho_c$ will be given for three distinct but closely related formulations of 4DVar. The latter provides a way to estimate of the minimum guaranteed success rate of a given algorithm, and can also be compared with an analytical lower bound on the radius of their basin of attraction.

In the smoothing context, another important parameter emerges from the fact that the observations are now assumed to be recorded over a (relatively short) finite time interval. The length of the estimation window plays an important role determining both the observability of a given problem and overall success rate of a particular algorithm. A key issue here is the regularity of the 4DVar objective function. Of the three techniques examined here, one will be shown to be particularly unstable when system is chaotic the estimation window is long, relative to the time scale of the chaos. This instability manifests itself in the form of exponentially many local minima, and can seriously impede the search, rendering even a trivial problem intractable. The other two methods avoid this instability, but are more computationally intensive. Each however has its own strengths and weaknesses in terms of accuracy, conditioning and overall probability of success.

### 3.4.1 Three formulations of 4DVar

Like 3DVar methods, 4DVar methods optimize a (generally nonlinear) least squares objective function that seeks to minimize a balanced measure of uncertainty between the observations and the model. The distinguishing feature of 4DVar however, is that it performs the optimization over a finite time estimation window.[7] Recall from Chap. (2) that applying Laplace's method to the statistical path integral for the conditional density $P(\mathcal{X}|\mathcal{Y})$ leads to the variational approximation, which involves the direct minimization of the log-likelihood or action functional $A(\mathcal{X}) \propto -\log[P(\mathcal{X}|\mathcal{Y})]$. Making the simplifying

[7] Hence the '4D', with the fourth dimension being time.

assumption that the errors in the system dynamics and observation process are assumed to be uncorrelated and Gaussian distributed with respective covariances $\boldsymbol{R}_f^{-1}$ and $\boldsymbol{R}_m^{-1}$, the action can be written in discrete time as

$$A(\mathcal{X}) \propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} |\boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_f}. \tag{3.6}$$

Note, as in 3DVar, the addition of the 'background' term $|\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b}$ penalizing deviations from the prior estimate $\boldsymbol{x}_b$. While this study is mainly concerned with the limit $\boldsymbol{R}_b \to \boldsymbol{0}$, this term is often needed to ensure positive definiteness of $\nabla^2 A$. The discrete time map may also be implicit $\boldsymbol{F}(\boldsymbol{x}_n, \boldsymbol{x}_{n+1})$, although this will not be needed here. Furthermore, in keeping with the previous section, the observation model is assumed to be a linear projection $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{H} \cdot \boldsymbol{x}$. The functional is not quadratic however, since here $\boldsymbol{F}(\boldsymbol{x}_n)$ is nonlinear, and may therefore contain multiple local minima.

The problem therefore comes down to finding the paths that minimize the action. These minimizing paths exponentially dominate the estimate, and the lowest minimizer gives the the *maximum a-posteriori* estimate, or conditional mode of the distribution. As previously discussed, there are a variety of iterative methods available for constructing its solution. Despite computational tradeoffs, many if not most of these techniques are formally equivalent for linear problems. Yet these methods often give vastly different results in general. This is a reflection of the fact that, in the nonlinear case, the conditioning of the problem depends on how these solutions are found. This section will focus on examining the differences between three common techniques for solving Eqn. (3.6). These differences are primarily due to choices regarding the constraints of the problem and how the solution is represented, and also give rise to their inherent computational tradeoffs.

The most straightforward approach is perhaps to minimize Eqn. (3.6) directly, as a function of the discretized path $\mathcal{X} := \{\boldsymbol{x}_n\}|_{n=0}^{N}$. This formulation, as an unconstrained nonlinear program, is common in control theory ([20]). Alternatively, as discussed in Chap. (2), the solution can instead be represented in terms of an initial condition $\boldsymbol{x}_0$ and a set of dynamical perturbations ([27]) or control variables $\mathcal{U} := \{\boldsymbol{u}_n\}|_{n=0}^{N-1}$, by imposing the constraint $\boldsymbol{u}_n := \boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n)$. With these assumptions, Eqn. (3.6) becomes a constrained nonlinear program, minimizing

$$\begin{gathered} A(\boldsymbol{x}_0, \mathcal{U}) \propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} |\boldsymbol{u}_n|^2_{\boldsymbol{R}_f} \\ \text{s.t. } \boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n. \end{gathered} \tag{3.7}$$

A variety of methods exist for solving such problems. It can be minimized directly, as a function

of $\{\boldsymbol{x}_0, \mathcal{U}\}$. Or, the constraints can be included as Lagrange multipliers $\mathcal{P} := \{\boldsymbol{p}_n\}|_{n=0}^{N}$, forming the augmented functional

$$A(\boldsymbol{x}_0, \mathcal{U}, \mathcal{P}) \propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \sum_{n=0}^{N-1} |\boldsymbol{u}_n|^2_{\boldsymbol{R}_f} + \sum_{n=0}^{N-1} \langle \boldsymbol{p}_n, \boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{u}_n \rangle.$$

This can be then minimized directly as a function of $\{(\boldsymbol{x}_0, \mathcal{U}, \mathcal{P})\}$ — *e.g.*, using Newton's method with Lagrange multipliers. This approach has convergence issues however, related to the fact that the Hessian is not positive definite ([144]). One can also enforce $\nabla_u A = 0$ and eliminate either $\mathcal{U}$ in terms of $\mathcal{P}$, or vice-versa, since $\boldsymbol{R}_f$ is assumed invertible. There are, in addition to these techniques, a variety of other methods have been devised to treat general optimization problems with constraints — see *e.g.*, [61, 202]. Given these seemingly limitless options, this study focuses on two simple approaches: namely, the direct minimization of Eqn. (3.6) and Eqn. (3.7) using the Gauss-Newton algorithm. The details of these methods will be given shortly.

It is worth noting however that these techniques have not been used much in very high dimensional problems, such as those encountered operational weather forecasting, where the number of state variables now frequently exceed $D > 10^9$ ([32]). This is due to its excessive memory requirements. Both techniques require storage of $O(D\,(N+1))$ state variables, which makes them unsuitable for these applications. These storage requirements may be reduced by assuming that the dynamical model is nearly perfect, which corresponds to the limit $\boldsymbol{R}_f^{-1} \to \boldsymbol{0}$, where both Eqn. (3.6) and Eqn. (3.7) reduce to the following constrained optimization problem of minimizing

$$\begin{aligned} A(\boldsymbol{x}_0) &\propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} \\ &\text{s.t } \boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n). \end{aligned} \tag{3.8}$$

Under this assumption, the trajectory may be represented in terms of the initial condition alone. This reduces the storage requirements to $O(D)$, and together with the use of adjoint methods for scalable evaluation of derivatives, makes this methods suitable for very large systems ([40]).

The limit $\boldsymbol{R}_f^{-1} \to \boldsymbol{0}$ is known in the data assimilation literature as 'strong constraint' 4DVar, due to the fact that the model constraints are assumed to hold exactly. The formulations given in Eqn. (3.6) and Eqn. (3.7) are likewise known as 'weak constraint' 4DVar. Strong constraint methods assume the model is perfect, without error, while weak constraints attempt to account for model error in a systematic way. Given that the experimental setup assumes the data is generated without model error, the strong constraint

formulation is particularly applicable here. In practice however, no model is perfect. Nonetheless, strong constraint methods often perform remarkably well, and have been used extensively in operational weather forecasting centers like the ECMWF ([161]). But as with most simplifying assumptions, it has its inherent limitations. Consequently, a key goal of this study is to identify these tradeoffs, and suggest ways in which they may be avoided or improved.

### 3.4.2 The Gauss-Newton method

The objective functionals in Eqns. (3.6), (3.7), and (3.8) all describe nonlinear least squares problems. That is, they all have the form

$$A(\boldsymbol{\xi}) = |\boldsymbol{\psi}(\boldsymbol{\xi})|^2_{\boldsymbol{\Gamma}}. \tag{3.9}$$

The minimization problem $\min_{\boldsymbol{\xi}\in\Omega} A(\boldsymbol{\xi})$ can be solved using a generalized version of Newton's method, which iteratively constructs a series of estimates $\boldsymbol{\xi}^{(i)}$ from the following rule

$$\boldsymbol{\xi}^{(i+1)} - \boldsymbol{\xi}^{(i)} = -\gamma^{(i)} \left[\boldsymbol{\Xi}(\boldsymbol{\xi}^{(i)})\right]^{-1} \cdot \nabla A(\boldsymbol{\xi}^{(i)}). \tag{3.10}$$

In this expression, $\gamma^{(i)} \in (0,1]$ is a line-search parameter that must be chosen appropriately to ensure a suitable decrease in the objective function value, *i.e.*, $A(\boldsymbol{\xi}^{(i)}) - A(\boldsymbol{\xi}^{(i+1)}) > \varepsilon$.

The choice of the matrix $\boldsymbol{\Xi}$ determines the type of method. For instance, $\boldsymbol{\Xi} \to \boldsymbol{I}$ gives gradient descent, and $\boldsymbol{\Xi} \to \nabla^2 A$ gives the standard Newton's method. The former is essentially guaranteed to converge, albeit slowly, to some local minimum of $A$. The latter has faster convergence but is more computationally intensive, largely due to the need to evaluate the second order derivative tensor $\nabla^2\boldsymbol{\psi}$. A third option, known as the Gauss-Newton method, neglects this term by using a first order approximation of the Hessian $\nabla^2 A(\boldsymbol{\xi}^{(i)}) \approx \nabla^\dagger \boldsymbol{\psi}^{(i)} \cdot \boldsymbol{\Gamma} \cdot \nabla \boldsymbol{\psi}^{(i)}$. It may thus be written as,

$$\boldsymbol{\xi}^{(i+1)} - \boldsymbol{\xi}^{(i)} = -\gamma^{(i)} \left[\nabla^\dagger \boldsymbol{\psi}^{(i)} \cdot \boldsymbol{\Gamma} \cdot \nabla \boldsymbol{\psi}^{(i)}\right]^{-1} \cdot \nabla^\dagger \boldsymbol{\psi}^{(i)} \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{\xi}^{(i)}). \tag{3.11}$$

This assumption simplifies the problem, improving computational efficiency and making the method particularly well suited for variational data assimilation problems ([64, 108]).

As such, it is the method of choice here. Explicit formulations will now be given for the methods described by Eqns. (3.6), (3.7), and (3.8), which will respectively be called WC-4DVar-X, WC-4DVar-U,

and SC-4DVar — beginning with the latter, as it is the simplest of the three.

For SC-4DVar, the objective function given in Eqn. (3.8) is minimized over initial conditions only: so $\boldsymbol{\xi} \to \boldsymbol{x}_0$ and $\Omega \to \mathbb{R}^D$. WC-4DVar-U has a similar form, but is extended to include the dynamical perturbations: so $\boldsymbol{\xi} \to \{\boldsymbol{x}_0, \mathcal{U}\}$ and $\Omega \to \mathbb{R}^{(N+1)D}$. Likewise, WC-4DVar-X has $\boldsymbol{\xi} \to \mathcal{X}$, and $\Omega = \mathbb{R}^{(N+1)D}$. With the following definitions,

$$\mathcal{Y} := \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix} \qquad \mathcal{H} := \begin{bmatrix} \boldsymbol{h}(\boldsymbol{x}_0) \\ \boldsymbol{h}(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}_N) \end{bmatrix} \qquad \mathcal{M} := \begin{bmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_b \\ \boldsymbol{x}_1 - \boldsymbol{F}(\boldsymbol{x}_0) \\ \vdots \\ \boldsymbol{x}_N - \boldsymbol{F}(\boldsymbol{x}_{N-1}) \end{bmatrix}$$

$$\mathcal{R}_f := \begin{bmatrix} \boldsymbol{R}_f & & \\ & \ddots & \\ & & \boldsymbol{R}_f \end{bmatrix} \qquad \mathcal{R}_m := \begin{bmatrix} \boldsymbol{R}_m & & \\ & \ddots & \\ & & \boldsymbol{R}_m \end{bmatrix},$$

the vectors $\psi$ can be written in block form as

$$\psi(\boldsymbol{x}_0) = \underbrace{\begin{bmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_b \\ \mathcal{Y} - \mathcal{H} \end{bmatrix}}_{\text{SC-4DVar}} \qquad \boldsymbol{\Gamma} \propto \begin{bmatrix} \boldsymbol{R}_b & \\ & \mathcal{R}_m \end{bmatrix}$$

$$\psi(\boldsymbol{x}_0, \mathcal{U}) = \underbrace{\begin{bmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_b \\ \mathcal{U} \\ \mathcal{Y} - \mathcal{H} \end{bmatrix}}_{\text{WC-4DVar-U}} \qquad \boldsymbol{\Gamma} \propto \begin{bmatrix} \boldsymbol{R}_b & & \\ & \mathcal{R}_f & \\ & & \mathcal{R}_m \end{bmatrix}$$

$$\psi(\mathcal{X}) = \underbrace{\begin{bmatrix} \mathcal{M} \\ \mathcal{Y} - \mathcal{H} \end{bmatrix}}_{\text{WC-4DVar-X}} \qquad \boldsymbol{\Gamma} \propto \begin{bmatrix} \boldsymbol{R}_b & & \\ & \mathcal{R}_f & \\ & & \mathcal{R}_m \end{bmatrix}$$

The partial derivatives of $\nabla\psi$ may also be expressed in block form. When $\boldsymbol{F}(\cdot)$ is explicit, it can be shown

that ([48])

$$\nabla\boldsymbol{\psi}(\boldsymbol{x}_0) = \underbrace{\begin{bmatrix} \boldsymbol{I} \\ \nabla_0\mathcal{H} \end{bmatrix}}_{\text{SC-4DVar}} \qquad \nabla\boldsymbol{\psi}(\boldsymbol{x}_0,\mathcal{U}) = \underbrace{\begin{bmatrix} \boldsymbol{I} \\ -\nabla_{\mathcal{X}}\mathcal{H}\cdot[\nabla_{\mathcal{X}}\mathcal{M}]^{-1} \end{bmatrix}}_{\text{WC-4DVar-U}} \qquad \nabla\boldsymbol{\psi}(\mathcal{X}) = \underbrace{\begin{bmatrix} \nabla_{\mathcal{X}}\mathcal{M} \\ -\nabla_{\mathcal{X}}\mathcal{H} \end{bmatrix}}_{\text{WC-4DVar-X}}$$

where

$$\nabla_{\mathcal{X}}\mathcal{M} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & & & \\ -\nabla\boldsymbol{F}_0 & \boldsymbol{I} & \ddots & & \\ \boldsymbol{0} & -\nabla\boldsymbol{F}_1 & \boldsymbol{I} & \ddots & \\ & \ddots & \ddots & \ddots & \boldsymbol{0} \\ & & \boldsymbol{0} & -\nabla\boldsymbol{F}_{N-1} & \boldsymbol{I} \end{pmatrix}$$

$$[\nabla_{\mathcal{X}}\mathcal{M}]^{-1} = \begin{bmatrix} \nabla\boldsymbol{\Phi}_{0,0} & \boldsymbol{0} & & \\ \nabla\boldsymbol{\Phi}_{1,0} & \nabla\boldsymbol{\Phi}_{1,1} & \boldsymbol{0} & \\ \vdots & \ddots & \ddots & \ddots \\ \nabla\boldsymbol{\Phi}_{N,0} & \nabla\boldsymbol{\Phi}_{N,1} & \cdots & \nabla\boldsymbol{\Phi}_{N,N} \end{bmatrix}$$

and $\boldsymbol{\Phi}_{n,m} \equiv \boldsymbol{\Phi}(\boldsymbol{x}_m,t_n,t_m) := \boldsymbol{F}_n \circ \boldsymbol{F}_{n-1} \circ \ldots \circ \boldsymbol{F}_m(\boldsymbol{x}_m)$ is the map from $\boldsymbol{x}_m \to \boldsymbol{x}_n$ for $n > m$. Its partial derivatives $\nabla\boldsymbol{\Phi}_{n,m} := \partial\boldsymbol{x}_n/\partial\boldsymbol{x}_m$ may be computed from the discrete time variational equation

$$\nabla\boldsymbol{\Phi}_{n+1,m} = \nabla\boldsymbol{F}_n \cdot \nabla\boldsymbol{\Phi}_{n,m} \qquad \nabla\boldsymbol{\Phi}_{n,n} = \boldsymbol{I}. \tag{3.12}$$

In other words, $\mathcal{F}$ maps $\mathcal{X} \to \{\boldsymbol{x}_0,\mathcal{U}\}$ and vice-versa. The inverse relationship between these two operators leads to solutions with quite distinct properties, which will be examined shortly. Also, both are expected to reduce to SC-4DVar in the perfect model limit ($\boldsymbol{R}_f^{-1} \to 0$), although this has not been explicitly shown.

### 3.4.3 Implementation details

The data and initial conditions are constructed as before. All three methods use explicit fourth-order Runge-Kutta. This avoids introducing artificial modeling errors, since the data is both generated and estimated using the same model. Also, since no prior knowledge of initial conditions is assumed, the background covariance is set to $\boldsymbol{R}_b = \boldsymbol{0}$, with $\boldsymbol{x}_b$ taken as the result of the previous iteration. This choice ensures all background entries in $\boldsymbol{\psi}$ are zero. If $\nabla\boldsymbol{\psi}$ becomes singular, the iteration is terminated.

The Jacobian matrices are evaluated directly using the variational Eqn. (3.12). Although not the

most computationally efficient approach, this study can be easily extended to analyze the impact of adjoint methods, whose accuracy have recently been brought into question ([170]). Also, the Jacobians here are expected to be highly ill-conditioned, their inverses need to be constructed carefully. Since SC-4DVar uses dense matrices, its inverse is computed via rank-truncated SVD. But this technique is not applicable to the sparse Jacobians used by the WC-4DVar methods, so their inverses are computed using sparse QR decomposition algorithms available in Matlab.

In addition, since smoothing methods are under consideration here, success is evaluated by comparing RMSE (or RMSD) across the entire estimation window. The RMSE is computed in relation to the optimal trajectory, is the same as the true solution when the observations are noiseless. But when the data has errors, these solutions no longer coincide. In this case, the optimal solution taken as the result of the optimization procedure, when initialized from the 'true' solution.

### 3.4.4 The collapse of the solution basin

The numerical study begins by examining the impact that the length of the estimation window has on the critical radius $\rho_c$. Recall that this is an approximation of the maximum distance between the initial guess and optimal solution at which the estimation no longer succeeds. Results shown in Fig. (3.9) plot estimated values of $\rho_c$ as a function of the length of the estimation window for the three 4DVar methods, with $D = 20$ and noiseless data that is both fully and partially observed $L = \{D, D/2, 1\}$.

Consider first the limit $T \to 0$. For the fully observed case $\rho_c$ is $O(1)$, so the optimal solution is globally accessible and may be found from roughly any initial point on the attractor. This is expected, as the problem is essentially trivial under these conditions. The 'true' solution is known immediately from the first set of observations. By contrast, with partial observations the critical radius decreases sharply, as the problem becomes highly ill-posed since the observation operator $\boldsymbol{h}$ is not full rank. Note also that the contraction is steeper for $L = 1$ than it is for $L = D/2$, indicating that the problem is more singular with less observations.

In this limit, the estimation window is not long enough to allow for the sufficient transfer of information from the dynamical model to the unobserved variables. The solution is therefore highly degenerate, in the sense that there are many trajectories that reproduce the observed data. This is shown explicitly in Fig. (3.10b), which plots a cross-section of the SC-4DVar action as a function of the initial deviation $|\boldsymbol{x}_0^* - \boldsymbol{x}_0|$ for various values of $T$. The slice is taken along an unobserved direction, so when the window is short the action surface is flat. As $T$ increases, the degeneracy is broken, and the landscape

becomes nearly convex before growing highly irregular.

On the other hand, in the long $T$ limit, $\rho_c$ contracts exponentially for both SC-4DVar and WC-4DVar-U, while for WC-4DVar-X it is asymptotically stable. The rate of collapse is roughly proportional to the largest global Lyapunov exponent, for reasons that will be described shortly. The contracting basin of support around the global minimum is shown more explicitly in Fig. (3.10a). For small $T$ its surface is nearly quadratic, but grows more rugged as $T$ is increased, becoming filled with exponentially many local minima. Eventually, the landscape becomes so sharply peaked that it exhibits fractal structure, in which one has to zoom in very close to see the global minimum. All this occurs despite the fact that the solution is fully observed, and therefore already known at $T = 0$. The result is that, in this limit, it is virtually impossible to find the global minimum without a highly accurate initial guess, which indicates that for certain algorithms a poor choice of certain parameters can render even a trivial problem intractable.

Thus, for SC-4DVar and WC-4DVar-U there are two competing factors: lack of observability in short term limit and dynamical instability in the asymptotic limit. Together these lead to an 'optimal' window length of roughly $T^* \approx 2/\lambda_{\max}$ where $\rho_c$ is as large as possible, although the precise value clearly depends on the number of observations. Moreover, even at the optimal length, the value $\rho_c$ is still relatively small and requires unreasonable accuracy in the initial guess to achieve successful estimates. By contrast, no such peak is evident for WC-4DVar-X. This suggests the likelihood of success should improve as the length of the window is increased, at least until about $T \approx 2 - 4/\lambda_{\max}$ where $\rho_c$ begins to level off.

These results highlight the use of the model dynamics to regularize the solution, by breaking the inherent degeneracy in partial observability. While intuitively it makes sense to try to incorporate as much of this dynamical information as possible, by using the longest available $T$, the choice of algorithm is crucial as $T$ gets long. Indeed, SC-4DVar and WC-4DVar-U are all but guaranteed to fail, if $T$ is chosen inappropriately. In some sense, the issue is that too much information is being passed to the estimate at once, causing the conditional density to collapse around the true trajectory. . On the other hand, WC-4DVar-X can evidently handle much longer estimation windows. While this seems beneficial — and it is in many circumstances — it nonetheless requires more computational resources, so there are additional tradeoffs to be considered.

The catastrophic collapse of the solution basin when $T$ is long relative to the timescale $\lambda_{\max}^{-1}$ of the chaos in the forecast model is a well-known limitation of SC-4DVar ([159, 51]). The workaround, in practice, is to use shorter windows and cycle the process, advancing the estimation window after each iteration. This approach has been widely used in many different contexts. In data assimilation, it is known

simply as 'cycling' ([115]). In estimation theory it is essentially a 'fixed-lag' smoother, and is related to the concept of 'moving horizon' estimation in control theory. It may be viewed as a hybrid filter/smoother that directly extends the dynamic filtering approach used in 3DVar to non-zero $T$. This approach will be discussed in more detail in Chap. (4).

An analytical measure of the region of guaranteed success around the global minimum is now introduced, with the goal of understanding both the mechanism behind the collapse of the solution basin in SC-4DVar, as well as the properties of the SC-4DVar-X that prevent this from occurring.

### 3.4.5 Convergence and uniqueness balls for the Gauss-Newton method

Optimization algorithms such as Newton's method and its generalizations/approximations may be viewed as discrete dynamical mappings $\boldsymbol{\xi}^{(i)} \rightarrow \boldsymbol{\xi}^{(i+1)}$ that, when successful, converge to a fixed point $\boldsymbol{\xi}^*$. Each method therefore has its own basin of attraction, defined as the set of initial conditions $\{\boldsymbol{\xi}^{(0)}\}$ that converges to $\boldsymbol{\xi}^*$. It is by now well-established that these basins, in general, exhibit fractal structure. Their complex topology, when visualized as a Julia set by drawing the boundaries between different basins of attraction, produces intricate patterns even for relatively simple functions — *e.g.*, polynomials in $D = 1$ of degree three or higher ([12]).

The notion of bounding the region of convergence for Newton's method for root-finding goes back at least as far as Kantorovich ([91]) and has been subsequently extended (*e.g.*, [192, 204, 179, 72]). However, most of the literature pertaining to the study of its basins of attraction as an optimization method has been focused exclusively on Newtons method in $D = 1$ ([75, 128]). An *a priori* lower bound on the size of the 'Newton basin' was given in [210], citing earlier work by [178, 58]. But these results were only given for polynomial functions in $D = 1$. Indeed, with the exception of [183], which examines higher order methods, a cursory search of the literature did not turn up any reference to bounds on basins of attraction for optimization methods in higher dimensions.

On the other hand, there exists another, more recent, body of literature that gives separate bounds on the *convergence and uniqueness balls* for Newtons method and its variants. Whereas Kantorovich's theorems assume the solution is unknown and bound the region of convergence based on global Lipschitz constants of the function, these results assume the solution is known and viewing the problem as an application of Banachs fixed-point theorem ([11]) provide local bounds based on generalized Lipschitz conditions (see *e.g.*, [205, 206]).

This analysis has been extended to the Gauss-Newton method as well ([113, 34, 35, 112]),

providing an upper bound on the region of guaranteed convergence and uniqueness for the solution. As these results are rather similar, the focus here will be on uniqueness ball, since it is the simpler of the two. In particular, the following bound was given by [113] in Corollary 5.2 (see also App. (3.5)).

Suppose $\boldsymbol{\xi}^*$ is a solution to Eqn. (3.9), and the derivative $\nabla\boldsymbol{\psi}$ satisfies the following conditions:

1. $\nabla\boldsymbol{\psi}(\boldsymbol{\xi})$ is continuous in the ball $B(\boldsymbol{\xi}^*, r)$
2. $\nabla\boldsymbol{\psi}(\boldsymbol{\xi})$ is full rank at the solution $\boldsymbol{\xi} \to \boldsymbol{\xi}^*$,
3. and satisfies the center Lipschitz condition

$$|\boldsymbol{\Gamma}^{1/2} \cdot (\nabla\boldsymbol{\psi}(\boldsymbol{\xi}) - \nabla\boldsymbol{\psi}(\boldsymbol{\xi}^*))| \leq \Lambda |\boldsymbol{\xi} - \boldsymbol{\xi}^*| \tag{3.13}$$

for all $\boldsymbol{\xi}$ in the open ball $B(\boldsymbol{\xi}^*, r)$ with $\boldsymbol{\Gamma}^{\dagger/2} \cdot \boldsymbol{\Gamma}^{1/2} := \boldsymbol{\Gamma}$. The solution of Eqn. (3.9) is unique in the ball $B(\boldsymbol{\xi}^*, r)$ with

$$\rho_c = \frac{2}{|[\boldsymbol{\Gamma}^{1/2} \cdot \nabla\boldsymbol{\psi}^*]^+|} \left( \frac{1}{\Lambda} - |[\nabla^\dagger \boldsymbol{\psi}^* \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}^*]^{-1}| \, |\boldsymbol{\Gamma}^{1/2} \cdot \boldsymbol{\psi}^*| \right) \tag{3.14}$$

where $\boldsymbol{\psi}^* \equiv \boldsymbol{\psi}(\boldsymbol{\xi}^*)$ and $\boldsymbol{M}^+ := [\boldsymbol{M}^\dagger \cdot \boldsymbol{M}]^{-1} \cdot \boldsymbol{M}^\dagger$ is the pseudoinverse.[8] Using the relation $\kappa[\boldsymbol{M}] = |\boldsymbol{M}^+||\boldsymbol{M}|$ this expression may be rewritten in terms of the condition number

$$\rho_c = \frac{2}{\kappa[\boldsymbol{\Gamma}^{1/2} \cdot \nabla\boldsymbol{\psi}^*]} \left( \frac{|\boldsymbol{\Gamma}^{1/2} \cdot \nabla\boldsymbol{\psi}^*|}{\Lambda} - |[\nabla^\dagger \boldsymbol{\psi}^* \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}^*]^{-1}| \, |\boldsymbol{\Gamma}^{1/2} \cdot \nabla\boldsymbol{\psi}^*| \, |\boldsymbol{\Gamma}^{1/2} \cdot \boldsymbol{\psi}^*| \right). \tag{3.15}$$

This version gives a more balanced result as $|\nabla\boldsymbol{\psi}^*|$ gets large, and provides a direct connection with the discussion of nonlinear conditioning given in Sec. (3.1).

Although the complexity of the expressions involved in Eqn. (3.15) makes it difficult to obtain analytic bounds, the terms involving $\boldsymbol{\psi}^*$, $\nabla\boldsymbol{\psi}^*$ may be directly computed, so long as the optimal solution is assumed known. The leading order behavior to $r$ appears to be determined by the condition number $\kappa[\boldsymbol{\Gamma}^{1/2} \cdot \nabla\boldsymbol{\psi}^*]$. The second term is proportional to $|\boldsymbol{\psi}^*|$, whose value is roughly proportional to magnitude of the errors in the dynamics and the observations. Its value is therefore expected to be small given the assumptions of this study.

It remains however to determine the Lipschitz constant $\Lambda$, in particular the ratio $|\nabla\boldsymbol{\psi}^*|/\Lambda$, which sets the overall length scale of $r$.[9] As discussed in App. (3.6), the calculation of $\Lambda$ depends on the second derivative tensor $|\nabla^2\boldsymbol{\psi}^*|$. While the tensor itself is easily computed, its norm is not. For instance, computing

[8] [113] also gives an argument that suggests this $r$ optimal in the scalar case ($D = 1$).

[9] Recall that the condition number is just a ratio.

the spectral norm is known to be NP-hard for a rank 3-tensor ([74]). Although the Frobenius norm can be used instead, because it is easily computable for arbitrary tensors an gives an upper bound on the spectral norm, but the tightness of this bound is not known in general. Overestimating this bound can be problematic, as leading to negative values of $r$ when $|\psi^*| \neq 0$ — *i.e.*, when the measurements or dynamics contain nonzero noise.

### 3.4.6 The inverse condition number of the Hessian

Given these difficulties, the additional terms in Eqn. (3.15) are neglected by making the approximation

$$\rho_c \approx \frac{1}{\kappa[\mathbf{\Gamma}^{1/2} \cdot \nabla \psi^*]}. \tag{3.16}$$

In other words, the radius of the uniqueness ball is taken to be the (square-root of the) inverse condition number of the approximate Hessian evaluated at the global minimum, $\nabla^2 A^* \approx |\mathbf{\Gamma}^{1/2} \cdot \nabla \psi^*|^2$. This value is easily computed. This estimate for $\rho_c$ is meant to be somewhat conservative, in the sense that many initial conditions outside this region that also converge to $\boldsymbol{\xi}^*$. It also only captures the *relative* scaling behavior of the solution basin, as its overall size is determined by the neglected terms. Also, the fact that when $\kappa$ is large $\rho_c$ is small, and vice-versa, underscores the intuition that well-conditioned problems have larger solution basins, and thus a higher inherent likelihood of success. It also provides geometric motivation for the process of preconditioning, in that lowering the condition number of the problem broadens the width the solution basin, thereby increasing the probability of success.

The lines in Fig. (3.9) show calculated values of $\rho_c$ taken from Eqn. (3.16). The absolute scale is determined by a least squares fit to the sampled estimates. The results follow the basic trends in the data in the long time limit, capturing both the collapse of the SC-4DVar and WC-4DVar-U objective functions, as well as the asymptotic stability of WC-4DVar-X. The singularity in the short $T$ limit (for partial observations) is also shown, although the collapse appears to be somewhat overpredicted — more so for $L = D/2$ than $L = 1$. This anomaly remains unexplained, but may be due to the higher order terms not calculated in Eqn. (3.15). It may be possible to analyze these terms to bound their significance in the limit $T \to 0$, but this is left for future work.

The stability issues exhibited by SC-4DVar can be understood by by examining the largest and smallest eigenvalues of the Hessian $\nabla^2 A^*$ (or equivalently the singular values of $\psi^*$). These are plotted in Fig. (3.11) as a function of $T$. In the short $T$ limit the smallest singular value behaves the same for all three methods. With partial observations it tends towards zero in the limit $T \to 0$, as the problem becomes

unobservable.

Conversely, in the long $T$ limit, the smallest singular value is roughly constant, while for SC-4DVar and WC-4DVar-U the largest grows asymptotically. The largest singular value thereby provides the dominant contribution to the condition number $\kappa[\nabla\psi^*]$ in the asymptotic limit. This is a reflection of the fact that $|\nabla\psi^*| = |\nabla\boldsymbol{h}_n \cdot \nabla\boldsymbol{\Phi}^*_{n,0}| \sim \exp[\lambda_{\max} T]$, in accordance with Oseledec's theorem ([148]), and occurs regardless of the number of observations.

With WC-4DVar-X on the other hand, the largest eigenvalue remains roughly constant, which suggests that this method naturally controls the growth of dynamical instabilities in the forecast model. This is similar to what was discussed in Sec. (3.3) for filtering algorithms. And moreover, in the linear case, it is known that WC-4DVar methods correspond to the optimal Kalman (smoothing) solution. It is interesting however, that despite being a weak constraint method, the 'control' variant WC-4DVar-U does not share the desirable asymptotic stability properties of WC-4DVar-X. This is due to the fact that, in the limit $\mathcal{U} = \mathbf{0}$, the formulation reduces to SC-4DVar, and thereby inherits its properties of asymptotic instability. Thus, although both WC-4DVar-X and WC-4DVar-U utilize weak constraints, they implement them in decidedly different ways, which results in dramatically distinct performance of the two methods.

### 3.4.7 Weak constraints, synchronization, and stability

The connections between weak constraints and asymptotic stability is now described from a synchronization perspective. These connections are perhaps most evident in the method proposed by [4], and discussed in a meteorological context by [185]. The idea is to introduce adds a coupling term (analogous to Eqn. (3.3)) into the dynamical constraints of Eqn. (3.8)

$$\begin{aligned} A(\boldsymbol{x}_0) &\propto \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n|^2_{\boldsymbol{R}_m} \\ \boldsymbol{x}_{n+1} &= \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{K}_n \cdot (\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n). \end{aligned} \tag{3.17}$$

As discussed in Sec. (3.3), the coupling term modifies the Jacobian of the dynamical model so that, with enough observations, it stabilizes the error dynamics in a neighborhood around the true solution by forcing all (conditional) global Lyapunov exponents become negative.

The inclusion of this coupling term has a similar effect on the objective function for SC-4DVar, as shown in Fig. (3.12) for both full and partial observations. Here the coupling gain matrix $\boldsymbol{K}$ is chosen to be identical to the 3DVar-FB method described in Sec. (3.3). Note that the control term causes the action

to remain convex, so its global minimum can be easily found regardless of the window length. Similar result holds if $\boldsymbol{K}$ is restricted to target only the unstable subspace, although these results are not shown.

However, as in Sec. (3.3) the effectiveness of this procedure depends on the system being sufficiently observable. As shown in Fig. (3.12), $L = D/4 = 5$ measurements are not enough to fully control the inherent chaos, resulting in an irregular objective function as $T$ gets long. While the observability requirement of roughly $L_c \approx 0.4D$ can be further reduced, *e.g.*, by using the extended Kalman filter to adaptively choose $\boldsymbol{K}$, this idea will not be pursued here in the context of Eqn. (3.17). Rather, another way of optimizing the coupling gain is now introduced, that leads to a more intuitive understanding of the benefits of using weak constraints.

Consider now including the coupling matrix in the optimization problem, by letting $\boldsymbol{K}_n = \gamma_n \boldsymbol{H}^\dagger$ as in Eqn. (3.3) and rewriting the optimization problem in Eqn. (3.17) as

$$A(\boldsymbol{x}_0) \propto \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n|^2_{\boldsymbol{R}_m} + \gamma_n^2$$
$$\text{s.t. } \boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \gamma_n \boldsymbol{H}^\dagger \cdot (\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n).$$

The objective function now includes the magnitude of the time-dependent coupling $\gamma_n$, which may be considered independent variables in the optimization. This term is added to keep the dynamical perturbations small and consistent, so that $\boldsymbol{K} \to 0$ as the estimate converges. As in Eqn. (3.3), the scalar coupling $\gamma_n$ can be generalized to a vector, which applies different weights to different components of the observations. Even more generally, it can be written as a matrix $\Gamma_n$, by including off-diagonal coupling components.

This method of dynamic state and parameter estimation ([4]) bears a strong resemblance to the constrained nonlinear program for weak constraint 4DVar given in Eqn. (3.7). The primary difference being that the former explicitly constrains its dynamical perturbations $\boldsymbol{u}_t$ to have the feedback form $\boldsymbol{u}_n = \boldsymbol{K}_n \cdot (\boldsymbol{y}_n - \boldsymbol{H}\cdot\boldsymbol{x}_n)$. Although the tradeoffs associated with this choice are not be investigated here, by penalizing $\boldsymbol{u}_n$ instead of $\gamma_n$, weak constraint 4DVar applies coupling to both the observed and unobserved states, without having to expand the search space to include the full matrix $\boldsymbol{\Gamma}_n$. The resulting feedback control attempts to synchronize the estimate with the data, to enhance the stability of the the basin of convergence around the global minimum by suppressing dynamical stabilities in the forecast model.

### 3.4.8 Critical observability limits

The critical observability limits of these methods are now examined, to understand how they scale with both: 1) the length of the estimation window $T$, and 2) the resolution of the forecast model $D$. Recall that the analogous results for filtering methods given in Sec. (3.3) used an excessively long estimation window to reduce its impact on the calculated $L_c$. For these fixed interval 4DVar methods however, $T$ must be much shorter due to the computational cost of inverting matrices of $O(DN)$ for weak constraint methods, as well as to avoid the more fundamental issues illustrated in Fig. (3.9) regarding dynamical instability in SC-4DVar and WC-4DVar-U.[10]

Critical thresholds $L_c$ are compared as a function of $T$ for $D = 20$. As expected, when $T = 0$ all methods require full observations. Both SC-4DVar and WC-4DVar-U begin to fail consistently above $T \approx 0.5/\lambda_{\max}$, while WC-4DVar-X remains stable regardless of $T$, reaffirming the results of Fig. (3.9).

Selecting $T = 0.5/\lambda_{\max}$, Fig. (3.13) plots observability limits of these three methods as a function of $D$. With no observation noise, and a highly accurate initial guess for all (even unobserved) state components, the observability requirements on all three methods are quite low ($L_c \approx 0.08D$). Such low thresholds are rather atypical however. For instance, the addition of observation noise increases them to roughly $L_c \approx D_u$, which further justifies its use as a baseline estimate for $L_c$.

Global initial conditions pushes these limits even higher — well above their filtering counterparts. This is not surprising however, as the estimation windows are significantly shorter. Reduced in temporal information raises the spatial observability limits, and produces significant differences between local and global initial conditions.

Increasing the length of the estimation window to $T = 1$ causes SC-4DVar and WC-4DVar-U to become more unstable, especially for global initial conditions. This instability does not impact the results from local initial conditions, which become more concentrated around $L_c \approx 0.07$ and $L_c \approx D_u$, with and without observation noise respectively. This implies that the initial conditions are still within the basin of attraction for these methods at $T = 1$.

If $T$ is increased further, the local results eventually become unstable as well. So neither WC-4DVar-U nor SC-4DVar will converge for *any* value of $L$. In comparing these two methods, it is also interesting that SC-4DVar remains stable for slightly higher values of $T$. This suggests that inflating the search space, from $D \to ND$, may increase the likelihood of the estimate converging to a local minimum,

[10] When a long estimation window is desired, a moving horizon or fixed-lag approach is typically adopted. These techniques will be examined in Chap. (3).

and the initial choice $\mathcal{U} = \mathbf{0}$ in effect constrains the search for the solution to nearby trajectories. It is well-known that control formulations like WC-4DVar-U require accurate initial guesses for both the initial condition $\boldsymbol{x}_0$ and the control law $\mathcal{U}$ ([30]). The latter are typically not easy to determined, although in this case one might instead consider initializing $\mathcal{U}$ with say the synchronizing feedback control law given in Eqn. (3.3). However, even with a good initial guess for $\mathcal{U}$, the problem of determining the optimal control trajectory can still be highly ill-conditioned ([20]), even more-so than its SC-4DVar counterpart.

Similar observations were recently made by [48] regarding the stability of WC-4DVar-X with the length of the estimation window. They also noted an inverse correlation between the number of spatial observations and the condition number. That is, for WC-4DVar-U, the condition number increases with the spatial resolution, whereas for WC-4DVar-X it decreases. This trend was not observed here however. To the contrary, all three methods showed improved conditioning of $\nabla\boldsymbol{\psi}^*$ as the resolution of the observations increased. In addition, based on comparing the number of iterations required for convergence, it was concluded that WC-4DVar-U performs more efficiently for the range of cases considered, which included simulations with Lorenz 96. But global errors in the initial estimate were not considered, and the results here point to WC-4DVar-X as being most efficient in this case. Moreover, the number of iterations is not always the best indicator of algorithmic efficiency. For instance, the Jacobian WC-4DVar-X is more sparse than that of WC-4DVar-U, so it may be inverted more efficiently.

This is not to say however that WC-4DVar-X is ideal for all situations. It is expected for instance, to be more sensitive than WC-4DVar-U to the frequency of observations, which here was taken to be sufficiently dense. A more detailed assessment of these tradeoffs will the subject of a future study.

### 3.4.9 Annealing $\boldsymbol{R}_f$

An analog to the annealing method described in Sec. (3.3) is now discussed, as a way of improving the performance of WC-4DVar-X. This idea of tuning the magnitude of $\boldsymbol{R}_f$ to gradually enforce the dynamical constraints of the model was first introduced in [213, 214]. In contrast to the filtering case, where $\boldsymbol{R}_f$ was increased gradually as the filtered estimate converged in time, here it is done over multiple iterations of the minimization algorithm, using the solution from the previous run to initialize the next.

Let $i$ denote the iteration and choose $\boldsymbol{R}_f^{(i)} = \alpha^{(i)}\,\boldsymbol{R}_f^{(0)}$, where $\alpha > 1.0$ is the rate at which the annealing is performed. Initially, the model error is prescribed to take some small value $|\boldsymbol{R}_f^{(0)}| \ll |\boldsymbol{R}_m|$, so that after the first iteration the estimated path closely matches the observations. As mentioned, with partial observations the solution is nearly degenerate in this limit. This degeneracy is lifted as $\boldsymbol{R}_f$ is gradually

increased at each iteration. This process continues, until a prescribed stopping point where $|\boldsymbol{R}_f| \gg |\boldsymbol{R}_m|$. At this point, model dynamics are penalized so heavily that the estimated path approximately describes a natural model trajectory.

As the value of $|\boldsymbol{R}_f|$ increases, the solution basin splits into several isolated local minima. Annealing $|\boldsymbol{R}_f|$ slowly, over multiple minimization runs, attempts to track the lowest minima throughout this transition. This idea of guiding a dynamical system through a phase transition or bifurcation by gradually annealing its parameters has roots in numerical continuation methods (see *e.g.*, [7]). The technique also resembles the approach given by [159], called quasi-static variational assimilation, which tries to track global minima by annealing the length of the assimilation window between subsequent iterations of SC-4DVar — a comparison between the two methods appears in [214].

A few practical considerations are worth also mentioning. First, the rate of annealing must be chosen, although this can be tuned empirically by decreasing $\alpha$ until the results converge. However, a smaller $\alpha$ reduces computational efficiency, so there is a trade-off that must be considered. The limit $\boldsymbol{R}_f \to \infty$ also cannot be realized numerically due to finite precision, so one has to decide how to choose the final value of $|\boldsymbol{R}_f|$. An 'optimal' value of $|\boldsymbol{R}_f|$ can be estimated using the L-curve ([70]) obtained from plotting the model and measurement errors on separate axes. Alternatively, one could also choose the 'optimal' stopping value based on the forecast deviations.

In addition, a decision needs to be made regarding how to set the relative values of $\boldsymbol{R}_f$ and $\boldsymbol{R}_m$. There are two obvious ways to do this: 1) anneal $\boldsymbol{R}_f$ with fixed $\boldsymbol{R}_m$, and 2) anneal $\boldsymbol{R}_f$ and $\boldsymbol{R}_m$ together. The former approach was used by [213, 214], where it $\boldsymbol{R}_m$ was assumed to be known. This choice is motivated in part by statistical filtering theory, where these covariance matrices have an absolute (as opposed to relative) value. The latter approach is discussed in [27, 174], where the annealing is performed with a homotopy continuation, by setting $|\boldsymbol{R}_m| = \gamma$ and $|\boldsymbol{R}_f| \to 1-\gamma$, and annealing $\gamma$ from $0 \to 1$. This idea makes use of the fact that the action is scale invariant, in the sense that it may be renormalized without changing the structure of any of the local minima, so only the relative values of $\boldsymbol{R}_m$ and $\boldsymbol{R}_f$ matter. Also, keeping the values of $\boldsymbol{R}_f$ and $\boldsymbol{R}_m$ relatively small should help prevent the objective function from becoming too poorly scaled.

Illustrative results from the annealing procedure are shown in Fig. (3.14), for Lorenz 96 with $D = 20$ and $L = \{15, 5\}$. The simulations are initialized randomly over an estimation window of length $T = 1/\lambda_{max}$, and include additive observation noise that is Gaussian distributed with mean zero and standard deviation $\sigma = 1$. The measurement error covariance remains fixed at $\boldsymbol{R}_m = 1$, while the model

error is annealed from $\boldsymbol{R}_f = 10^{-4} \to 10^4$ in 9 equally spaced steps, so $\alpha = 10^{1/5}$.

Final action levels are plotted as a function of $|\boldsymbol{R}_f|$ in Fig. (3.14a). They increase logarithmically with $|\boldsymbol{R}_f|$ until leveling off at roughly $\sqrt{|\boldsymbol{R}_f|/|\boldsymbol{R}_m|} = 3$. As $|\boldsymbol{R}_f|$ is further increased, the $L = 15$ action levels do not change, indicating they have found the global minimum. By contrast, the $L = 5$ action levels begin to diverge as different initial paths settle into various local minima.

The dashed line indicates the value of $|\boldsymbol{R}_f|$ where the Hessian of the action is most well-conditioned. This is shown more explicitly in Fig. (3.14b), which plots the inverse condition number as a function of $|\boldsymbol{R}_f|$. The peak occurs at roughly the same value of $|\boldsymbol{R}_f|$ for both $L = 15$ and $L = 5$.

The L-curve for the model error regularization is shown in Fig. (3.14c), which plots the model error $\sum_{n=0}^{N-1} |\boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n)|^2$ and measurement error $\sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2$ on an inverted log scale. Note that by convention ([70]), the error covariances $\boldsymbol{R}_m,\ \boldsymbol{R}_f$ are not included in these terms. The large dots indicate that the optimal values of $|\boldsymbol{R}_f|$ are concentrated near the bend of the L-curve, as well as the point where the action plot levels off. Thus, the L-curve criteria rather intuitively corresponds to values of $|\boldsymbol{R}_f|$ for which the objective function is optimally well-conditioned.

It is also worth noting however, that in contrast to the EKF annealing technique, this method was observed to slightly degrade the observational efficiency of WC-4DVar-X by roughly 5-10%. While this comes as a bit of a surprise, the results of [213, 214] rely heavily on sampling, and should therefore not be expected to improve $L_c$ in general. A precise explanation for this slight performance reduction is not known. However, starting with the initial path so close to the observations may increase the odds of falling into a local minimum. Similar observations were made by [174], although more work is needed to verify this claim. Nonetheless, this annealing procedure is still a valuable tool for examining the landscape of the action as a function of $|\boldsymbol{R}_f|$.

## 3.5 The uniqueness ball for the Gauss-Newton method

We now give a proof of Eqns. (3.14) and (3.15) for the uniqueness ball of the Gauss-Newton method. The original proofs are given under more general Lipschitz conditions in Theorem 4.1 and Corollary 5.2 in [113], but are simplified here for clarity.

Recall the Gauss-Newton method for solving nonlinear least squares minimization problem

$$\min_{\boldsymbol{x}} \frac{1}{2} |\boldsymbol{\psi}(\boldsymbol{x})|^2_{\boldsymbol{\Gamma}}$$

involves an iteration $\boldsymbol{x}^{(i)} \rightarrow \boldsymbol{x}^{(i+1)}$

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} - [\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^{(i)})]^{-1} \cdot \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^{(i)}).$$

Let $\boldsymbol{x}^*$ be a solution and assume the Jacobian $\nabla\boldsymbol{\psi}(\boldsymbol{x})$ is continuous in the ball $B(\boldsymbol{x}^*, r)$, is full rank at $\boldsymbol{x}^*$, and satisfies the center Lipschitz condition

$$|\boldsymbol{\Gamma}^{1/2} \cdot (\nabla\boldsymbol{\psi}(\boldsymbol{x}) - \nabla\boldsymbol{\psi}(\boldsymbol{x}^*))| \leq \Lambda |\boldsymbol{x} - \boldsymbol{x}^*| \tag{3.18}$$

for all $\boldsymbol{x} \in B(\boldsymbol{x}^*, r)$.

The proof is by contradiction. Suppose there exists another solution $\tilde{\boldsymbol{x}} \neq \boldsymbol{x}^*$ in $B(\boldsymbol{x}^*, r)$ and consider the iteration

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} - [\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1} \cdot \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^{(i)})$$

with $\boldsymbol{x}^{(0)} = \tilde{\boldsymbol{x}}$. Subtracting, $\boldsymbol{x}^*$ from both sides and using the fact that

$$[\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1} \cdot \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^*) = 0,$$

gives after some rearranging

$$\begin{aligned}
\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^* &= \boldsymbol{x}^{(i)} - \boldsymbol{x}^* - [\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1} \cdot \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \\
&= [\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1} \cdot \Big( \big(\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) - \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)})\big) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \\
&\qquad + \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \big(\nabla\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot (\boldsymbol{x}^{(i)} - \boldsymbol{x}^*) + \boldsymbol{\psi}(\boldsymbol{x}^*) - \boldsymbol{\psi}(\boldsymbol{x}^{(i)})\big) \Big) \\
&= [\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1} \cdot \Big( \big(\nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) - \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^{(i)})\big) \cdot \boldsymbol{\Gamma} \cdot \boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \\
&\quad + \nabla^{\dagger}\boldsymbol{\psi}(\boldsymbol{x}^*) \cdot \boldsymbol{\Gamma} \cdot \int_0^1 ds \, \big(\nabla\boldsymbol{\psi}(\boldsymbol{x}^*) - \boldsymbol{\nabla}\boldsymbol{\psi}(\boldsymbol{x}^* + s\,(\boldsymbol{x}^{(i)} - \boldsymbol{x}^*))\big) \cdot (\boldsymbol{x}^{(i)} - \boldsymbol{x}^*) \Big).
\end{aligned}$$

Taking the norm of both sides and using Eqn. (3.18) gives

$$\begin{aligned}|\boldsymbol{x}^{(i+1)}-\boldsymbol{x}^*| &\le |[\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^*)\cdot\boldsymbol{\Gamma}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1}|\,|(\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^*)-\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^{(i)}))\cdot\boldsymbol{\Gamma}^{\dagger/2}|\,|\boldsymbol{\Gamma}^{1/2}\cdot\boldsymbol{\psi}(\boldsymbol{x}^{(i)})| \\ &\quad +|[\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^+|\int_0^1 ds\,|\boldsymbol{\Gamma}^{1/2}\cdot(\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)-\boldsymbol{\nabla}\boldsymbol{\psi}(\boldsymbol{x}^*+s(\boldsymbol{x}^{(i)}-\boldsymbol{x}^*)))|\,|\boldsymbol{x}^{(i)}-\boldsymbol{x}^*| \\ &\le \Lambda|[\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^*)\cdot\boldsymbol{\Gamma}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1}|\,|\boldsymbol{\Gamma}^{1/2}\cdot\boldsymbol{\psi}(\boldsymbol{x}^{(i)})|\,|\boldsymbol{x}^{(i)}-\boldsymbol{x}^*|+\frac{\Lambda}{2}|[\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^+|\,|\boldsymbol{x}^{(i)}-\boldsymbol{x}^*|^2 \\ &=: q\,|\boldsymbol{x}^{(i)}-\boldsymbol{x}^*|.\end{aligned}$$

Setting

$$q := \Lambda|[\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^*)\cdot\boldsymbol{\Gamma}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1}|\,|\boldsymbol{\Gamma}^{1/2}\cdot\boldsymbol{\psi}(\boldsymbol{x}^{(i)})|+\frac{\Lambda}{2}|[\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^+|\,|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*|\le 1$$

gives a contraction mapping

$$|\boldsymbol{x}^{(i)}-\boldsymbol{x}^*|\le q^i\,|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*|$$

for all $i=0,1,\ldots$, which implies $\lim_{i\to\infty}\boldsymbol{x}^{(i)}=\boldsymbol{x}^*$ and therefore $\boldsymbol{x}^*=\boldsymbol{x}^{(0)}$. Solving for $r:=|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*|$ gives Eqn. (3.14)

$$r\le\frac{2}{|[\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^+|}\left(\frac{1}{\Lambda}-|[\nabla^\dagger\boldsymbol{\psi}(\boldsymbol{x}^*)\cdot\boldsymbol{\Gamma}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)]^{-1}|\,|\boldsymbol{\Gamma}^{1/2}\cdot\boldsymbol{\psi}(\boldsymbol{x}^*)|\right)$$

## 3.6 The Lipschitz constant $\Lambda$

To get an expression for $\Lambda$, one approach is to assume $\nabla\boldsymbol{\psi}(\boldsymbol{\xi})$ is differentiable in $B(\boldsymbol{\xi}^*,r)$ and rewrite Eqn. (3.13) as a line integral between the convex combination $\boldsymbol{\xi}^*+s(\boldsymbol{\xi}-\boldsymbol{\xi}^*)$. Explicitly,

$$\begin{aligned}|\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{\xi})-\boldsymbol{\Gamma}^{1/2}\cdot\nabla\boldsymbol{\psi}(\boldsymbol{\xi}^*)| &= \left|\int_0^1 ds\,\boldsymbol{\Gamma}^{1/2}\cdot\nabla^2\boldsymbol{\psi}(\boldsymbol{\xi}^*+s(\boldsymbol{\xi}-\boldsymbol{\xi}^*))\cdot(\boldsymbol{\xi}-\boldsymbol{\xi}^*)\right| \\ &\le \int_0^1 ds\,|\boldsymbol{\Gamma}^{1/2}|\,|\nabla^2\boldsymbol{\psi}(\boldsymbol{\xi}^*+s(\boldsymbol{\xi}-\boldsymbol{\xi}^*))|\,|\boldsymbol{\xi}-\boldsymbol{\xi}^*|=\Lambda|\boldsymbol{\xi}-\boldsymbol{\xi}^*|\end{aligned}$$

or

$$\Lambda=\int_0^1 ds\,|\boldsymbol{\Gamma}^{1/2}|\,|\nabla^2\boldsymbol{\psi}(\boldsymbol{\xi}^*+s(\boldsymbol{\xi}-\boldsymbol{\xi}^*))|. \tag{3.19}$$

It therefore requires computing the second derivative tensor $\nabla^2\psi$. For weak constraint 4DVar this tensor is easy to compute but cumbersome to display. For SC-4DVar, it is given by

$$\nabla^2\psi(\boldsymbol{x}_0) = -\nabla^2\boldsymbol{h}_n\cdot\nabla\boldsymbol{\Phi}_{n,0}\cdot\nabla\boldsymbol{\Phi}_{n,0} - \nabla\boldsymbol{h}_n\cdot\nabla^2\boldsymbol{\Phi}_{n,0} \to \nabla\boldsymbol{h}_n\cdot\nabla^2\boldsymbol{\Phi}_{n,0},$$

where $\nabla^2\boldsymbol{h}_n$ is ignored since the measurement operator is assumed to be a projection. The term $\nabla^2\boldsymbol{\Phi}_{n,0} := \partial^2\boldsymbol{x}_n/\partial\boldsymbol{x}_0^2$ is given by second order the variational equation

$$\nabla^2\boldsymbol{\Phi}_{n+1,0} = \nabla\boldsymbol{F}_n\cdot\nabla^2\boldsymbol{\Phi}_{n,0} + \nabla^2\boldsymbol{F}_n\cdot\nabla\boldsymbol{\Phi}_{n,0}\cdot\nabla\boldsymbol{\Phi}_{n,0} \qquad \nabla^2\boldsymbol{\Phi}_{0,0} = \mathbf{0}. \tag{3.20}$$

There is added difficulty however in bounding Eqn. (3.19) within $B(\boldsymbol{\xi}^*, r)$, as it requires an expression for $|\nabla^2\psi(\boldsymbol{\xi}^* + s(\boldsymbol{\xi}-\boldsymbol{\xi}^*))|$ as a function of $r = |\boldsymbol{\xi}-\boldsymbol{\xi}^*|$. One option is to approximate it as a constant by its value at the solution $|\nabla^2\psi^*|$.[11] Or one can make an ergodic approximation, in which case $\Lambda$ is proportional to the average rate of growth of $\nabla^2\boldsymbol{\Phi}_{n,0}$. This may be described using the second order Lyapunov exponents $\lambda^{(2)}$ introduced by [52, 53] to analyze the affect of strong nonlinearities on the growth of errors in the forecast.

As discussed in [47] however, care must be taken to ensure these calculations are stable. For strong constraints, the recursion relation in Eqn. (3.20) can be solved by introducing a discrete integration factor $\boldsymbol{p}_n$, where $\boldsymbol{p}_n = \boldsymbol{p}_{n+1}\cdot\nabla\boldsymbol{F}_n$ and $\boldsymbol{p}_{n+1} = \boldsymbol{I}$, so that $\boldsymbol{p}_m = \nabla\boldsymbol{\Phi}_{n-m,0}$ for $m \le n$

$$\begin{aligned}\boldsymbol{p}_{n+1}\cdot\nabla^2\boldsymbol{\Phi}_{n+1,0} &= \boldsymbol{p}_{n+1}\cdot\nabla\boldsymbol{F}_n\cdot\nabla^2\boldsymbol{\Phi}_{n,0} + \boldsymbol{p}_{n+1}\cdot\nabla^2\boldsymbol{F}_n\cdot\nabla\boldsymbol{\Phi}_{n,0}\cdot\nabla\boldsymbol{\Phi}_{n,0}\\ &= \boldsymbol{p}_n\cdot\nabla^2\boldsymbol{\Phi}_{n,0} + \boldsymbol{p}_{n+1}\cdot\nabla^2\boldsymbol{F}_n\cdot\nabla\boldsymbol{\Phi}_{n,0}\cdot\nabla\boldsymbol{\Phi}_{n,0}.\end{aligned}$$

Expanding recursively gives

$$\begin{aligned}\boldsymbol{p}_{n+1}\cdot\nabla^2\boldsymbol{\Phi}_{n+1,0} &= \sum_{m=0}^{n}\boldsymbol{p}_{m+1}\cdot\nabla^2\boldsymbol{F}_m\cdot\nabla\boldsymbol{\Phi}_{m,0}\cdot\nabla\boldsymbol{\Phi}_{m,0}\\ &= \sum_{m=0}^{n}\nabla^\dagger\boldsymbol{\Phi}_{n-m,0}\cdot\nabla^2\boldsymbol{F}_m\cdot\nabla\boldsymbol{\Phi}_{m,0}\cdot\nabla\boldsymbol{\Phi}_{m,0}\end{aligned}$$

[11] This is related to the idea introduced by [179] of using 'point estimates' to analyze convergence of Newton's method, which has subsequently become known as Smale's $\alpha$-theory and $\gamma$-theory, and was later generalized to the Gauss-Newton method in [175] and [113]. Roughly speaking, these results guarantee convergence based on the assumption $\psi(\boldsymbol{\xi})$ is analytic with a bound on higher derivatives in its Taylor expansion about a single point.

If one assumes global Lipschitz constants $|\nabla \boldsymbol{F}| \leq \nu$ and $|\nabla^2 \boldsymbol{F}| \leq \mu$ then $|\nabla \boldsymbol{\Phi}_{n,0}| \leq \nu^n$ and

$$|\nabla^2 \boldsymbol{\Phi}_{n+1,0}| \leq (\mu \nu^n) \sum_{m=0}^{n} \nu_m = (\mu \nu^n) \frac{1-\nu^{n+1}}{1-\nu},$$

which means the term $|\nabla \boldsymbol{\psi}^*|/\Lambda$ in Eqn. (3.15) scales as

$$\frac{|\nabla \boldsymbol{\psi}^*|}{\Lambda} \approx \frac{|\nabla \boldsymbol{\psi}^*|}{|\nabla^2 \boldsymbol{\psi}^*|} \sim \frac{1-\nu}{\mu(1-\nu^{n+1})}.$$

But these expressions are unbounded for chaotic systems, where $\nu > 1$, even though the limit

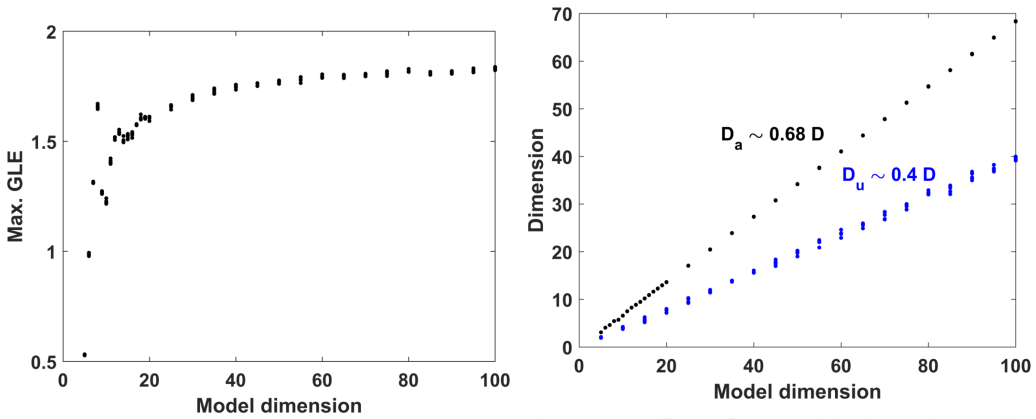
$$\lambda_{\max}^{(1)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \log[|\nabla \boldsymbol{\Phi}_{n,0}|]$$

is guaranteed to exist under the conditions of the multiplicative ergodic theorem ([148]). On the other hand, [52, 53] conjectured the relationship $\lambda_{\max}^{(2)} \approx 2\lambda_{\max}^{(1)}$ between the maximum first and second order global Lyapunov exponents ($\lambda_{\max}^{(1)}$ and $\lambda_{\max}^{(2)}$ respectively). This hypothesis was further justified by [47], who gave an analysis the case for fixed points and periodic orbits as well as a numerical calculation for chaotic trajectories. This was later proved by [188] for the scalar case. All this suggests that the ergodic average

$$\frac{|\nabla \boldsymbol{\psi}^*|}{\Lambda} \approx \exp\left[\lim_{N\to\infty} \sum_{n=0}^{N-1} \left(\log[|\nabla \boldsymbol{\Phi}_n^*|] - \log[|\nabla^2 \boldsymbol{\Phi}_n|]\right)\right] = \exp\left[(\lambda_{\max}^{(1)} - \lambda_{\max}^{(2)})N\right],$$

is expected to converge to a constant that is independent of initial conditions $\boldsymbol{x}_0$.

Chapter 3, in part is being prepared for submission for publication of the material. Rey, Daniel; Abarbanel, Henry DI. The dissertation author was the primary investigator and author of this material.




(a) Maximum GLEs $\lambda_{\max}$

(b) Fractal dimension $D_a$ (black) and unstable subspace dimension $D_u$ (blue).

**Figure 3.1**: Chaotic properties of Lorenz 96 model. The Kaplan-Yorke fractal dimension $D_a$ of the attractor, and the average dimension of the unstable subspace $D_u$, both scale linearly with the model dimension $D$. The magnitude of the maximum global Lyapunov exponents $\lambda_{\max}$ exhibits transient behavior for $D < 20$.

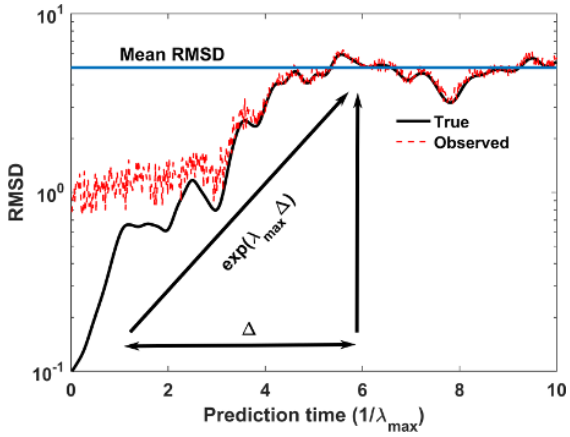
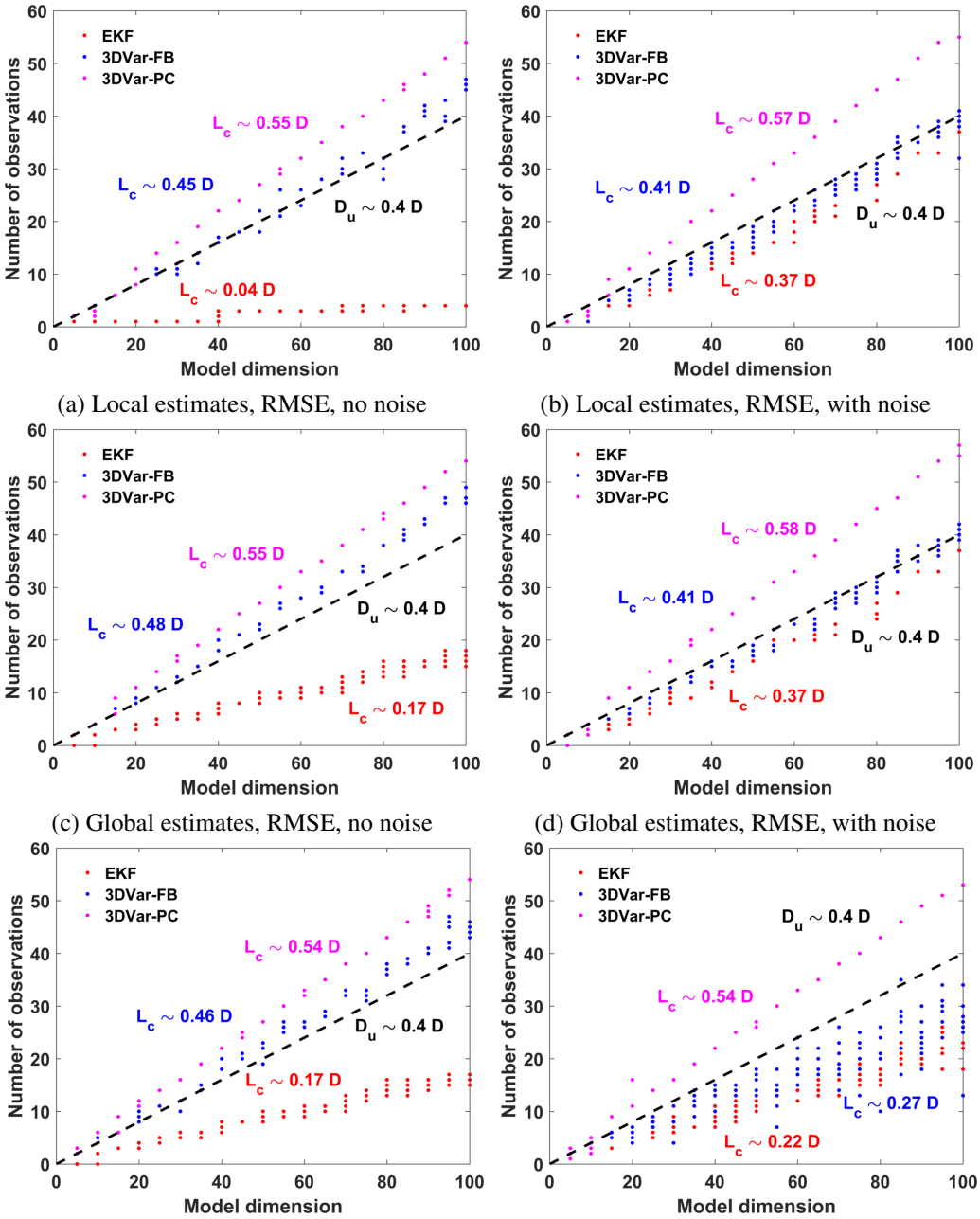



**Figure 3.2**: The magnitude of the 'true' error can be estimated using $\lambda_{\max}$ and measuring the time $\Delta$ it takes for the observed RMS deviations in the prediction to stabilize. This provides a way to quantify success with experimental data.




(a) Local estimates, RMSE, no noise

(b) Local estimates, RMSE, with noise

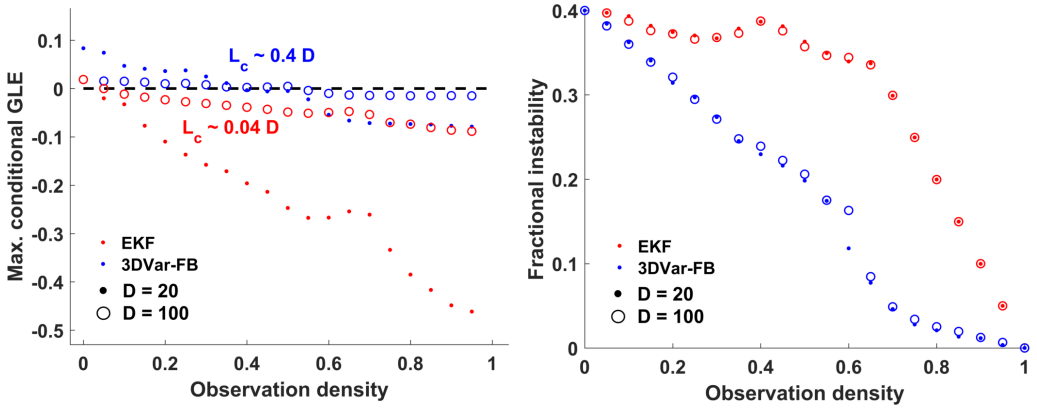
(c) Global estimates, RMSE, no noise

(d) Global estimates, RMSE, with noise

(e) Global estimates, RMSD, no noise

(f) Global estimates, RMSD, with noise

**Figure 3.3**: The critical minimum number of observations $L_c$ required to construct a successful estimate, for three related filtering methods.




(a) Max. conditional Lyapunov exponent

(b) Fractional unstable subspace $D_u/D$

**Figure 3.4**: Maximum conditional Lyapunov exponent, and fractional dimension of the unstable subspace $D_u/D$, plotted as a function of observation density $L/D$. 3DVar-FB crosses the synchronization threshold at $L_c \approx 0.4\,D$, while the EKF is more efficient at $L_c \approx 0.1\,D$. $D_u$ does not change much with the model resolution $D$.

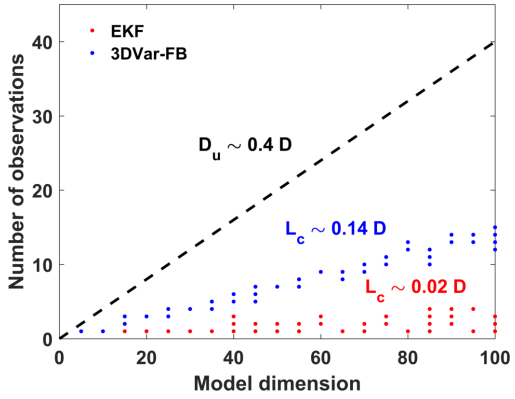
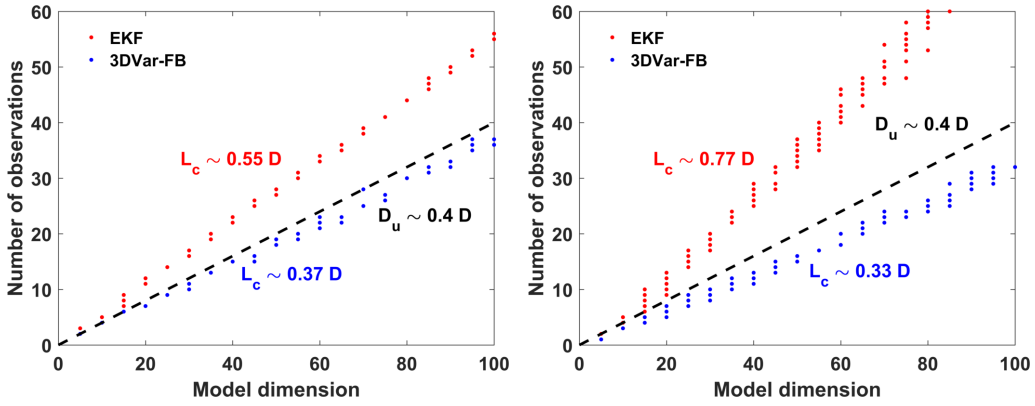
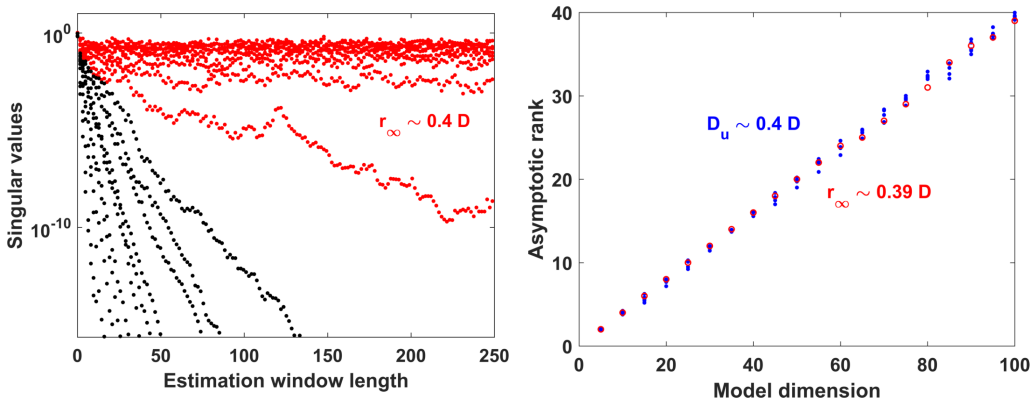



**Figure 3.5**: The critical rank of an adaptive observation operator with $L = D$ that targets and stabilizes only the largest $r$ components of the unstable subspace.



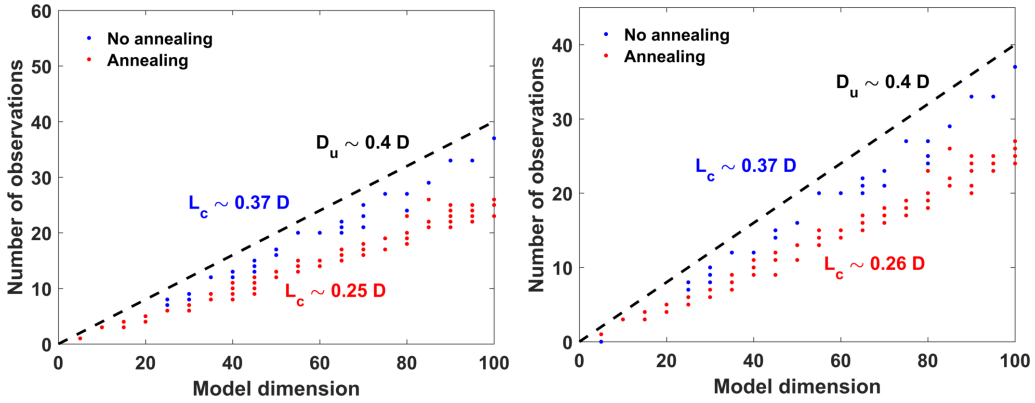

(a) Global estimates, no noise (b) Global estimates, with noise

**Figure 3.6**: The critical number of observations using assimilation in the unstable subspace (AUS), where the control perturbations are projected down onto the largest $r$ components of the unstable manifold.




(a) Singular values of $\boldsymbol{R}_n^{-1}$ for $D = 20$ (b) Asymptotic rank of $\boldsymbol{R}_n^{-1}$ as a function of $D$

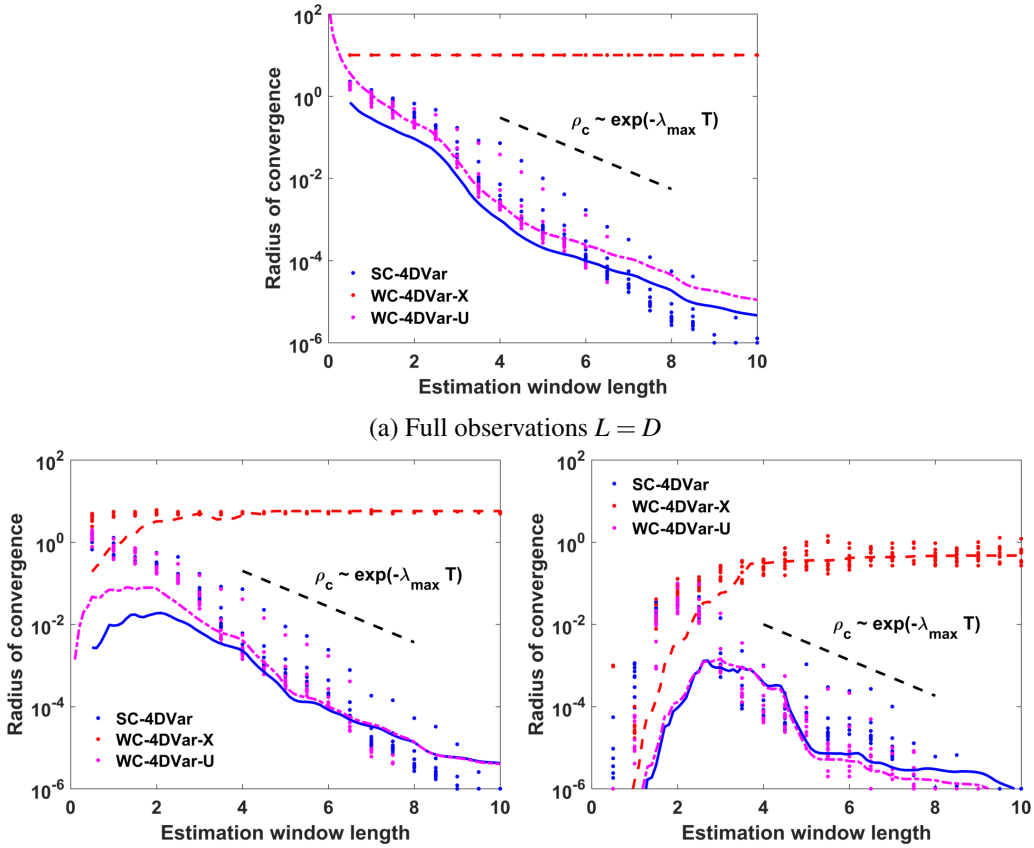
**Figure 3.7**: Collapse of the estimated error covariance for the EKF with $\boldsymbol{R}_f^{-1} = 0$. Fig. (3.7a) plots the singular values of $\boldsymbol{R}_n^{-1}$ as a function of time for $D = 20$. Fig. (3.7b) plots the asymptotic rank $r_\infty$ of $\boldsymbol{R}_n^{-1}$ as a function of $D$. The latter agrees well with the average dimension of the unstable subspace $D_u$.




(a) Local estimates, with noise

(b) Global estimates, with noise

**Figure 3.8**: Annealing the model error covariance $\boldsymbol{R}_f^{-1}$ places more weight on the model as the EKF converges. This improves its ability to filter observation noise, and reduces $L_c$ compared with Figs. (3.3b) and (3.3d)

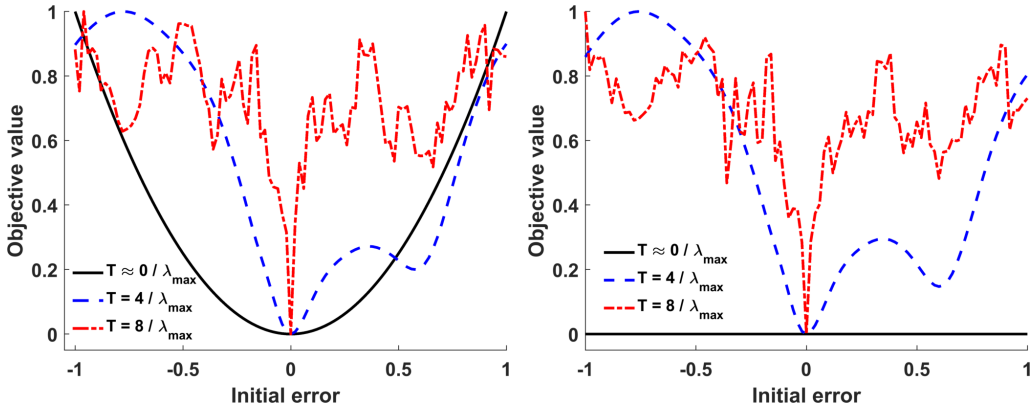



(a) Full observations $L = D$

(b) Partial observations $L = D/2$

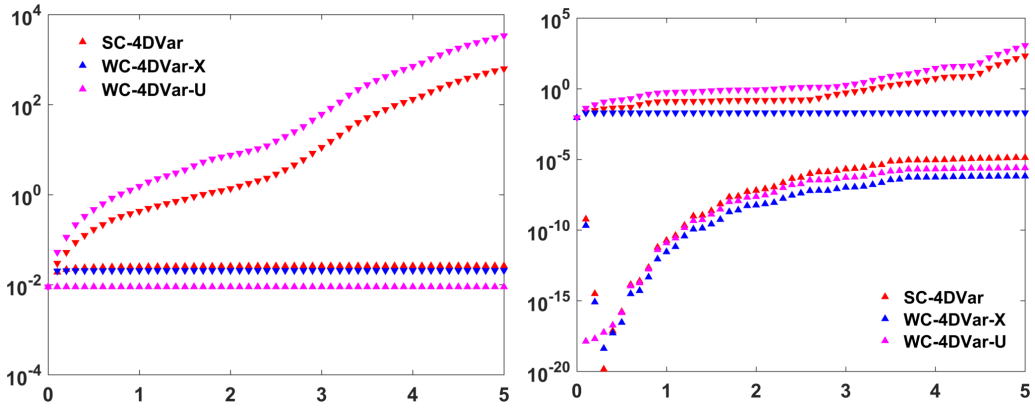
(c) Partial observations $L = 1$

**Figure 3.9**: Critical radius $\rho_c$ at which the estimation no longer succeeds, plotted as a function of the estimation window length $T$. Dots are sampled estimates, and the lines plot the inverse condition number of the Hessian $\nabla^2 A$. Note that WC-4DVar-X method is stable in the long $T$ limit, but SC-4DVar and WC-4DVar-U are not.



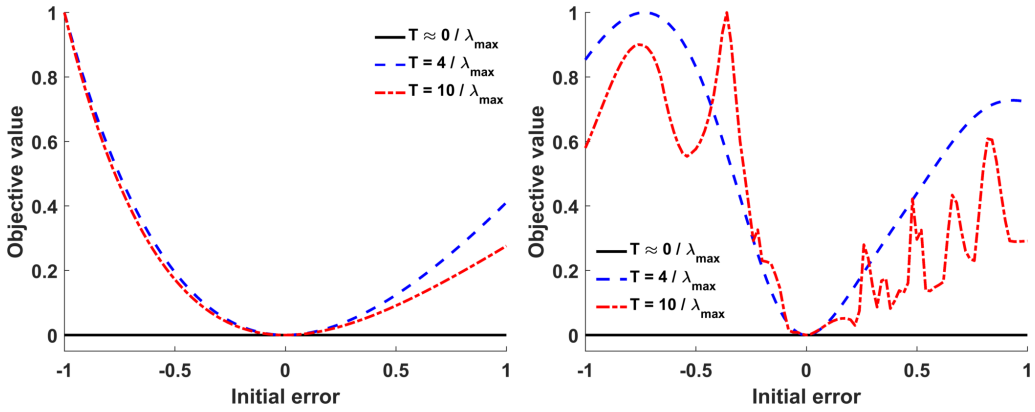

(a) Full observations $L = D$ (b) Partial observations $L = D/2$

**Figure 3.10**: Cross-sections of the SC-4DVar objective function for various values $T$ plotted as a function of the initial error between the estimate and the optimal solution $|\boldsymbol{x}_0^* - \boldsymbol{x}_0|$. As the length of the estimation window $T$ grows, the basin around the global minimum collapses, making it virtually impossible to find.



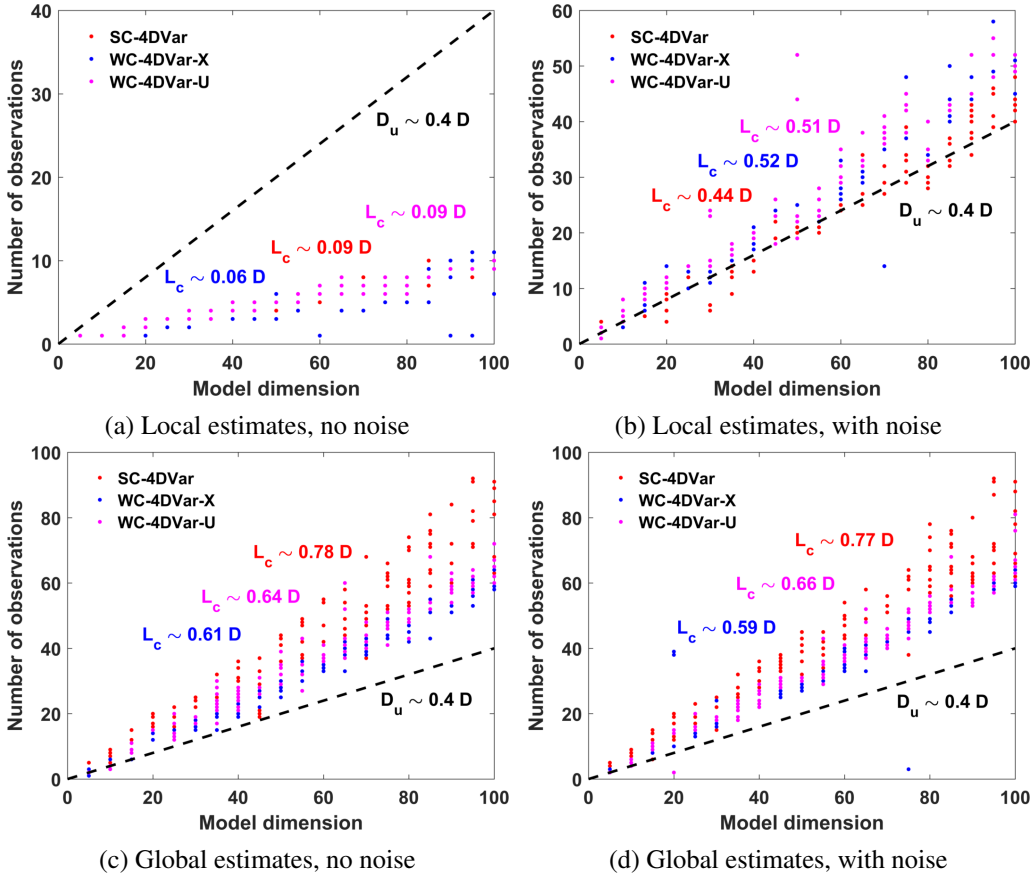

(a) Full observations $L = D$ (b) Partial observations $L = 1$

**Figure 3.11**: Largest and smallest eigenvalues of the inverse Hessian $\nabla^2 A(\boldsymbol{\xi}^*)$ as a function of estimation window length $T$. With partial observations, the smallest tends toward zero as $T \to 0$ for all three methods. In the long $T$ limit, the largest grows exponentially for SC-4DVar and WC-4DVar-U. For WC-4DVar-X, it is asymptotically stable.




(a) Partial observations $L = D/2$ (b) Partial observations $L = D/4$

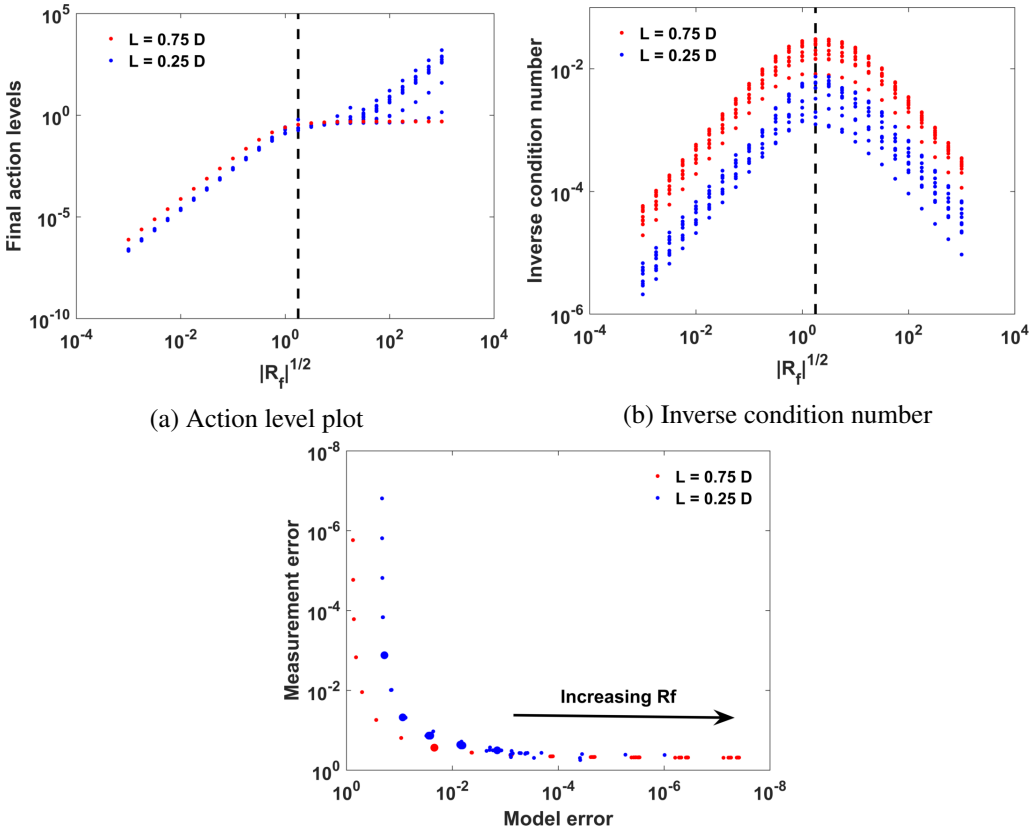
**Figure 3.12**: Cross-sections of the SC-4DVar objective function, analogous to Fig. (3.10) but including a coupling term in the dynamical constraints. When $L > L_c$, dynamical instabilities in the forecast model are suppressed, and the objective function is regular.




(a) Local estimates, no noise (b) Local estimates, with noise

(c) Global estimates, no noise (d) Global estimates, with noise

**Figure 3.13**: Calculated values of the critical minimum number of observations $L_c$ required to construct a successful estimate, for three fixed-interval 4DVar methods with $T = 0.5/\lambda_{\max}$.




(a) Action level plot

(b) Inverse condition number

(c) L-curve

**Figure 3.14**: Model error annealing for WC-4DVar-X. Dashed lines and large dots indicate values of $|\boldsymbol{R}_f|$ where the action is optimally conditioned.

# 4 Moving horizon estimation for poorly observable systems

The previous chapter developed a framework for analyzing empirical observability thresholds that arise in state and parameter estimation of nonlinear dynamical systems. It showed that for a certain chaotic model, these limits grow linearly with its resolution, at a rate that is roughly proportional to the time averaged dimension of the unstable subspace. Although the precise value of this cutoff depends in a complex way on various aspects of the problem's three core components — the forecast model, the observation system, and the data assimilation algorithm — certain estimation methods were also shown to use the available observations much more efficiently than others.

The latter point will now be explored in further detail, by examining the question: what can be done when the number of available observations are too few to reliably provide accurate estimates of the system's initial state? For instance, it is evident that Kalman-based methods are markedly more efficient in this regard, and this benefit is apparently linked to properties of the Riccati equation for the estimated error covariance. But although methods like the Ext KF are apparently capable of reducing these limits, they nonetheless contain a number of inherent tradeoffs. For one, the Ext KF simply does not scale up to very large problems, due to the memory requirements of constructing a $D \times D$ matrix. But perhaps more to the point, the method tends to become unstable as the density of observations is decreased. So that when it fails, the filtered estimate often becomes unbounded.

This chapter will focus on a technique that can both reduce the critical observability requirements and simultaneously improve the stability of existing data assimilation techniques. The idea is quite simple. Instead of using data from a single time point, use a sliding window of observations localized around the current filtered estimate. It may thus be viewed as a hybrid technique that generalizes standard smoothing and filtering methods, which are recovered in the respective long and short limits of the length of the sliding

window. This effectively enriches the set of observations using information from the forecast model, to enhance the accuracy and stability of the algorithm, albeit at some additional computational cost.

Like many good ideas, this has been rediscovered several times, in many different contexts. In optimal estimation, the linear theory of fixed-lag smoothing goes at least as far back as the late 1960s ([124]). The concept of 'moving horizon estimation' was introduced in the mid 1990s as the estimation dual to model predictive control ([8]), both of share the same motivating framework. In data assimilation, the idea of temporally extending the analysis to 'four-dimensions' was first introduced by [171], and since the 1990s has become one of the more popular techniques for generating operational numerical weather forecasts.

This idea was unknowingly rediscovered by us rather recently, although from a much different perspective. The objective was to determine whether the aforementioned observability limits of 3DVar methods may be reduced by better utilizing information in the time series of the observations. This was motivated by discrete time embedding theorems ([187, 5, 172]), developed by the nonlinear dynamics community to reconstruct the topological structure of the attractor from noisy time series observations. The resulting algorithms, which we have called time delay methods, bear strong resemblance to incremental SC-4DVar, the fixed lag Kalman smoother, and various related techniques from moving horizon estimation. Nonetheless, the theoretical connections to the discrete time embedding theorems are important, as they provide a unifying perspective that links these well-established techniques.

This chapter aims to further explicate the connections between approaches, as well as highlight some of the more subtle differences in implementation, such as the use of long overlapping estimation windows appears. The impact of these choices will be examined, to assess how it affects the overall stability of the algorithm, and its ability to reduce the number of required observations. The progression is organized much in the same way that these ideas were discovered. Starting with a discussion of the original implementation given in [166, 167], the ideas will be further generalized to connect with the EKF, 4DVar, and other methods from moving horizon estimation. Various related generalizations and extensions will then be examined, and the benefits of reusing observational data in overlapping estimation windows will be demonstrated.

## 4.1 Extracting information from the time series of observations

When the number of measurements at each time step $L$ is below the critical threshold $L_c$, one must find another means to overcome the observability deficit. While the most straightforward solution

is to simply make more observations, this may not be possible due to the constraints of the problem. An alternative approach is reduce $L_c$, by recognizing that there is additional information residing in the temporal derivatives of the observations. However, this derivative information often cannot be directly measured. And while it may instead be approximated via finite differences, the derivative operation acts as a high-pass filter, which makes the result sensitive to noise in the measurements. Alternatively, it has been known for some time in the nonlinear dynamics literature that this derivative information is also available in the *time delay* of the observations. That is, given a suitable choice of delay time $\tau = m\,dt$ where $m$ is an integer, the new information beyond $\boldsymbol{y}(t_n)$ lies in $\boldsymbol{y}(t_n \pm \tau)$. The derivative is just another (albeit less numerically robust) way of accessing this information. The process can be repeated as many times as needed to construct an $M$dimensional vector $\mathcal{Y}$.

### 4.1.1 Time delay embedding and attractor reconstruction

This idea forms the conceptual basis for the well-established technique of attractor reconstruction ([55, 2]), in which this methodology is employed as a means of identifying unambiguous orbits of a partially observable system. Mapping to a proxy space of time delayed observations inverts the projection associated with the fact that the rank of $\boldsymbol{H}$ is $L < D$. There are also discrete time extension of the Whitney embedding theorem ([187, 5, 172]) that give sufficient conditions on the number of time delays $M$ needed to ensure that the map between the two spaces is invertible. Namely, taking $M > 2D_a$ is guaranteed to be enough where $D_a$ is the fractal dimension of the attractor. No analogous theorem is available for the lower bound, and this estimate is often much larger than necessary. But there are well-known methods for computing the necessary condition, for instance by testing for false-nearest neighbors ([100]).

While these theorems suggest that any pair of uni-directionally coupled dynamical systems can be synchronized via almost all scalar observations of the driving system's state ([184]), they assume the data is noiseless. In practice, the success of this procedure depends crucially on the choice of $\tau$. If $\tau$ is too small, then not enough time has passed to allow the dynamics to adequately inform $\boldsymbol{y}(t_n \pm \tau)$. Likewise, when $\tau$ is too large, chaotic behavior in the system will eventually cause the state to decohere and inject dynamical noise into the analysis. The optimal choice lies somewhere in between these two extremes, and useful heuristics exist for selecting this value, such as the first minimum of the average mutual information between subsequent measurements ([2]). But often a suitable choice can be found simply by trial and error.

In the estimation context, the time delays are not needed to reconstruct the entire phase space. Rather, they are used to enrich the set of existing observations, to enhance control over localized dynamical

instabilities, and thereby reduce the observability threshold $L_c$. Further, under certain constraints, the number of time delays $M$ required to achieve synchronization roughly corresponds to the dimension of the unstable subspace of the attractor.

### 4.1.2 Time delay feedback control synchronization

The original time delay formulation described by [166, 167] was given in the context of feedback control synchronization, described in Sec. (3.3) of Chap. (3). The objective is to produce a filtered estimate of the system state $\boldsymbol{x}(T)$ at the end of the observation window. This may be achieved by introducing a feedback control term into the dynamical equations, which in discrete time reads

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{K}_n \cdot \big(\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)\big). \tag{4.1}$$

As in Chap. (3), to simplify the exposition, observations are made at each $dt$, and the operator $\boldsymbol{h}(\cdot) \to \boldsymbol{H}$ is taken to be a static linear projection. But the formulation is easily extended to the general case. Also, $\boldsymbol{K}_n \to \boldsymbol{H}^\dagger \cdot \boldsymbol{K}$ is taken to be static with entries only along the diagonal, which effectively restricts the coupling term to only perturb the observed model states.

The idea is to perform the estimation in the embedding space, using time delay feedback control term. Construct a set of time delayed observations $\mathcal{Y}_n$, as well as a set of time-delayed model states $\mathcal{S}_n$ and its observed outputs $\mathcal{H}_n$

$$\mathcal{Y}_n := \begin{bmatrix} \boldsymbol{y}_n \\ \boldsymbol{y}_{n+m} \\ \vdots \\ \boldsymbol{y}_{n+(M-1)m} \end{bmatrix} \qquad \mathcal{S}_n := \begin{bmatrix} \tilde{\boldsymbol{x}}_n \\ \tilde{\boldsymbol{x}}_{n+m} \\ \vdots \\ \tilde{\boldsymbol{x}}_{n+(M-1)m} \end{bmatrix} \qquad \mathcal{H}(\mathcal{S}_n) := \begin{bmatrix} \boldsymbol{h}(\tilde{\boldsymbol{x}}_n) \\ \boldsymbol{h}(\tilde{\boldsymbol{x}}_{n+m}) \\ \vdots \\ \boldsymbol{h}(\tilde{\boldsymbol{x}}_{n+(M-1)m}) \end{bmatrix}.^{1}$$

Note that these are not time 'delays' in the usual sense, but rather a time *advanced* coordinates, which for positive $\tau$ uses observations at later times. The reasoning for this stems from the need for the filter to control error growth forward in time. Both approaches are valid however, as well as a mixture of the two. These ideas are further explored in Sec. (4.3), along with a discussion of what to do at the end of the estimation window.

[1] These expressions will need to be modified to include time-dependence of $\boldsymbol{F}$, but this is straightforward.

By analogy with Eqn. (4.1), the time delay coupling can be introduced as follows

$$\mathcal{S}_{n+1} = \boldsymbol{F}(\mathcal{S}_n) + \mathcal{K}_n \cdot \big(\mathcal{Y}_n - \mathcal{H}(\mathcal{S}_n)\big).$$

It remains to specify how the estimate $\boldsymbol{x}_n$ is related to the embedded states $\tilde{\boldsymbol{x}}_n$. The simplest approach is perhaps to choose $\boldsymbol{x}_n = \tilde{\boldsymbol{x}}_n$, and set the delays $\tilde{\boldsymbol{x}}_{n+m}$ arbitrarily. In this case, one can choose a fixed coupling matrix $\tilde{\mathcal{K}}$ with nonzero elements in its off-diagonal blocks to permit the passage of information between states at different times. This idea was pursued by [156] and offers some benefits over standard 3DVar methods.

Alternatively, one can choose to impose the model constraints $\tilde{\boldsymbol{x}}_{n+1} = \boldsymbol{F}(\tilde{\boldsymbol{x}}_n)$ so that $\mathcal{S}_n$ is a function of $\boldsymbol{x}_n$. The issue however, is that once the coupling is applied, the model constraints are no longer satisfied. To repeat the procedure at the next step, requires inverting $\mathcal{H}_n \equiv \mathcal{H}_n(\mathcal{S}_n(\boldsymbol{x}_n))$. But the global inverse is in general not explicitly known. Locally however, the map can be approximated by its linearization $\nabla\mathcal{H}_n$, and inverted by solving the linear system

$$\nabla\mathcal{H}_n \cdot \delta\boldsymbol{x}_n = \mathcal{K}_n \cdot \big(\mathcal{Y}_n - \mathcal{H}_n\big), \tag{4.2}$$

where the right-hand side the desired control coupling in time delay space. In general, the linearized 'observability map' $\nabla\mathcal{H}_n$ will not a square matrix, but a stable least squares solution may be obtained using singular value decomposition (SVD) to compute its pseudoinverse $[\nabla\mathcal{H}_n]^+$. The theorems ([187, 5, 172]) guarantee this map is locally invertible, so long as $M > 2D_a$. They make no guarantees however about the sensitivity of this map, which may be highly ill-conditioned, imposing practical limits on the solution due to observation noise.

Using the localized inverse $[\nabla\mathcal{H}_n]^+$ to map the time delayed coupling perturbation back to 'state space' coordinates at $\boldsymbol{x}_n$ produces the time-delayed feedback control scheme described by [166, 167]

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) + \boldsymbol{K}_n \cdot [\nabla\mathcal{H}_n]^+ \cdot \mathcal{K}_n \cdot \big(\mathcal{Y}_n - \mathcal{H}(\boldsymbol{x}_n)\big). \tag{4.3}$$

Note there are now two coupling gain matrices $\boldsymbol{K}_n,\ \mathcal{K}_n$, which respectively operate in state space and time delay space. Rules for choosing these matrices will be given momentarily, when the connections with 4DVar and the Kalman filter are described. Also, as with the other methods in Sec. (3.3), the magnitudes of these $\boldsymbol{K}_n,\ \mathcal{K}_n$ should be rescaled with the step size $dt$.

The local Jacobian may be represented in block form

$$\nabla \mathcal{H}_n = \begin{bmatrix} \nabla \boldsymbol{h}_n \cdot \nabla \boldsymbol{\Phi}_{n,n} \\ \nabla \boldsymbol{h}_{n+m} \cdot \nabla \boldsymbol{\Phi}_{n+m,n} \\ \vdots \\ \nabla \boldsymbol{h}_{n+(M-1)m} \cdot \nabla \boldsymbol{\Phi}_{n+(M-1)m,n} \end{bmatrix}. \tag{4.4}$$

The matrix $\nabla \boldsymbol{\Phi}_{n',n} := \partial \tilde{\boldsymbol{x}}_{n'} / \partial \tilde{\boldsymbol{x}}_n$ may be directly computed from the variational equation

$$\begin{aligned} \tilde{\boldsymbol{x}}_{n+1} &= \boldsymbol{F}(\tilde{\boldsymbol{x}}_n) & \tilde{\boldsymbol{x}}_n &= \boldsymbol{x}_n \\ \nabla \boldsymbol{\Phi}_{n'+1,n} &= \nabla \boldsymbol{F}(\tilde{\boldsymbol{x}}_{n'}) \cdot \nabla \boldsymbol{\Phi}_{n',n} & \nabla \boldsymbol{\Phi}_{n,n} &= \boldsymbol{I}. \end{aligned} \tag{4.5}$$

But this requires storage and manipulation of $D \times D$ matrices, which is unsuitable for very large dimensional problems. An adjoint approach that avoids this limitation will be described in Sec. (4.3).

The estimation process proceeds recursively as follows:

1. At each time step $\boldsymbol{x}_n$, compute $\mathcal{S}(\boldsymbol{x}_n)$ and $\nabla \mathcal{S}(\boldsymbol{x}_n)$ by integrating from $t_n \to t_n + (M-1)\tau$ Eqn. (4.5)
2. Use $\mathcal{S}(\boldsymbol{x}_n)$ and $\nabla \mathcal{S}(\boldsymbol{x}_n)$ to construct $\mathcal{H}(\boldsymbol{x}_n)$ and $\nabla \mathcal{H}(\boldsymbol{x}_n)$
3. Solve Eqn. (4.2) *e.g.*, , by computing the pseudoinverse $[\nabla \mathcal{H}(\boldsymbol{x}_n)]^+$
4. Use Eqn. (4.3) to compute $\boldsymbol{x}_{n+1}$. Repeat from step 1.

To improve the robustness of the solution, the pseudoinverse in Step 3 is computed using rank truncated SVD. The rank of $[\nabla \mathcal{H}_n]^+$ needs to chosen carefully however. Useful heuristics for this are discussed in Sec. (4.3), along with related techniques for improving the conditioning of the map. Also, for large problems the SVD is to costly to compute. Other regularization techniques are available however, which will be discussed shortly.

Also, note that when $M = 1$, the time delay approach reduces to the standard feedback control given in Eqn. (4.1). Three important differences are realized however when $M > 1$.

1. It incorporates information from the time delays of the observations, which allows the coupling to act not only at given observation times, but also in between them if needed. This is useful for instance, to control nonlinear instabilities in the model that arise when the time between subsequent observations is long relative to the timescale of the dynamics.

2. All components of the model state $\boldsymbol{x}(t)$ are influenced by the control term, not just the observed components. This permits estimation of the model's fixed parameters $\boldsymbol{\theta}$, by simply treating them as additional state variables, and is accomplished without having to choose the off-diagonal components of the coupling matrix $\boldsymbol{K}_n$.

3. It extracts additional information from *existing* measurements, effectively lowering the critical observability threshold $L_c$.

### Connections to other estimation methods

This idea of using a sliding window of observations to enhance the accuracy and stability of estimation algorithms, appears frequently in the optimal estimation and data assimilation literature. However, the ties between these methods and time delay embedding are rarely mentioned. These connections will now be discussed.

### 4.1.3 Moving horizon estimation

Shortly after the introduction of the Kalman filter in 1960, it was recognized that performing 'fixed-lag' smoothing with a sliding observation window can significantly improve accuracy and stability ([99, 134, 135]). A related technique, known as the 'Gauss-Newton filter', was also proposed by Morrison around the same time as the Kalman filter, but was supposedly ignored due to its high computational demands ([138]). If so, this method may be one of the first to establish this idea, for which the linear theory dates back to [6]. These methods apparently offer a number of benefits, including the generality of the problems considered, lower demands on filter initialization accuracy, as well as better consistency with the Cramer-Rao lower bound.

Related ideas ideas were also discussed by [76], in which methods from optimal control were used to construct an observer that utilizes time delay information. However, as this approach involves computing an analytical inverse of the 'observability map $\mathcal{H}(\boldsymbol{x}_n)$, its applicability is limited by the size and complexity of the problem. Nevertheless, the motivating idea behind the method is essentially the same as the one given here.

A more general method using local inverses was proposed earlier by [65, 137, 136]. These 'Newton observers' are effectively identical the time delay filtering method given in Eqn. (4.3). And several useful results were proved regarding existence and uniqueness of the inverse, and convergence of the observer. This work also highlighted the connections between dynamical observer design and the

inversion of nonlinear maps (see *e.g.*, [141, 60]), and also helped provide the foundation for the subsequent development of 'moving horizon estimation' ([8, 163]), which is the estimation dual of 'model predictive control'. The latter was developed in the early in the 1980s, and is essentially the control formulation of the idea for using a sliding window (or a moving horizon) of observations. The theoretical development of moving horizon estimation provided a number of valuable results, including a Bayesian derivation ([71]), and extension to problems with constraints ([163]). Consistent handling of these constraints is one of the main arguments for adopting the moving horizon approach, which may be seen as a generalization of Kalman methods.

### 4.1.4 The extended Kalman Filter

The connection to the EKF was also given by [137], who suggested the observation space may be similarly enriched by substituting the embedded quantities $\boldsymbol{y}_n \to \mathcal{Y}_n$ and $\boldsymbol{h}_n \to \mathcal{H}_n$. In the limit $\boldsymbol{R}_m \to \infty$ with fixed $\boldsymbol{R}_f \propto \boldsymbol{I}$, the Kalman gain becomes the pseudoinverse $\boldsymbol{K}_n \to [\nabla\mathcal{H}(\boldsymbol{x}_n)]^+$ and reduces to the time delay filtering technique described in Eqn. (4.3). The derivation is straightforward, and starts with the error covariance update

$$\boldsymbol{R}^{-1}_{n+1|n} = \nabla\boldsymbol{F}_n \cdot \left(\boldsymbol{R}_{n|n} + \nabla^\dagger\mathcal{H}_n \cdot \boldsymbol{R}_m \cdot \nabla\mathcal{H}_n\right)^{-1} \cdot \nabla^\dagger\boldsymbol{F}_n + \boldsymbol{R}_f^{-1}$$

Let $\boldsymbol{R}_m^{-1} = \mu\boldsymbol{I}$ and $\boldsymbol{R}_f^{-1} = \varepsilon\boldsymbol{I}$ for positive scalars $\mu$ and $\varepsilon$. As long as $\boldsymbol{R}^{-1}_{n|n}$ is positive definite, $\boldsymbol{R}^{-1}_{n+1|n} = \varepsilon\boldsymbol{I}$ in the limit $\mu \to 0$. In this limit, the Kalman gain then becomes

$$\begin{aligned}\boldsymbol{K}_n &= \boldsymbol{R}^{-1}_{n|n-1} \cdot \nabla^\dagger\mathcal{H}_n \cdot \left(\nabla\mathcal{H}_n \cdot \boldsymbol{R}^{-1}_{n|n-1} \cdot \nabla^\dagger\mathcal{H}_n + \boldsymbol{R}_m^{-1}\right)^{-1} \\ &= \nabla^\dagger\mathcal{H}_n \cdot \left(\nabla\mathcal{H}_n \cdot \nabla^\dagger\mathcal{H}_n + \frac{\mu}{\varepsilon}\boldsymbol{I}\right)^{-1} \to [\nabla\mathcal{H}_n]^+.\end{aligned}$$

### 4.1.5 Strong constraint 4DVar

The connection with 4DVar is also quite direct, by noting that the SC-4DVar objective function given in Eqn. (3.8) can also be written as

$$A(\boldsymbol{x}_n) \propto |\boldsymbol{x}_n - \boldsymbol{x}_b|^2_{\boldsymbol{B}_b} + |\mathcal{Y}_n - \mathcal{H}(\boldsymbol{x}_n)|^2_{\mathcal{R}_m}, \tag{4.6}$$

where $\mathcal{R}_m$ is a time embedded version observation error matrix, with $\boldsymbol{R}_m$ on the block diagonal. Off-diagonal elements may also be included, to account for temporal correlations in the measurement errors.

Let $\boldsymbol{R}_m^{-1} \to \mu \boldsymbol{I}$ and $\boldsymbol{R}_b^{-1} \to \beta \boldsymbol{I}$. Performing Gauss-Newton minimization on this functional and taking the limit $\mu \to 0$ gives

$$\begin{aligned}\delta \boldsymbol{x}_n &= \left(\boldsymbol{R}_b + \nabla^\dagger \mathcal{H}_n \cdot \mathcal{R}_m \cdot \nabla \mathcal{H}_n\right)^{-1} \cdot \left(\boldsymbol{R}_b \cdot (\boldsymbol{x}_n - \boldsymbol{x}_b) + \nabla^\dagger \mathcal{H}_n \cdot \mathcal{R}_m \cdot (\mathcal{Y}_n - \mathcal{H}_n(\boldsymbol{x}_n))\right) \\ &= \left(\frac{\mu}{\beta}\boldsymbol{I} + \nabla^\dagger \mathcal{H}_n \cdot \nabla \mathcal{H}_n\right)^{-1} \cdot \nabla^\dagger \mathcal{H}_n \cdot \left(\mathcal{Y}_n - \mathcal{H}(\boldsymbol{x}_n)\right) \to [\nabla \mathcal{H}_n]^+ \cdot \left(\mathcal{Y}_n - \mathcal{H}(\boldsymbol{x}_n)\right).\end{aligned}$$

Strong constraint 4DVar is therefore effectively equivalent to Eqn. (4.3), except for the following differences:

1. Strong constraint 4DVar does not include the notion of a time delay or embedding dimension. But this is implicit in the choice of $\mathcal{R}_m$. So in this way, SC-4DVar may be viewed as a generalization.

2. The time delay filter uses truncated singular-value decomposition to regularize the perturbation $\delta \boldsymbol{x}_n$, while SC-4DVar performs Tikhonov regularization with a background term. This makes SC-4DVar amenable for large problems, for which the SVD is too costly to compute.

3. With the time delay method, the estimate is propagated in small increments $dt$ between analyses and observations are re-used. Strong constraint 4DVar on the other hand, typically propagates the observation window by its full length, so that observations are only used once. Despite increased computational cost and statistical arguments that caution against reusing data, propagating the window in small overlapping increments has benefits, which will be discussed in Sec. (4.2).

### 4.1.6 Weak constraint 4DVar

These connections also motivated a search for an analogous form for fixed interval smoothing with the weak constraint action (WC-4DVar-X). But while a many ideas were discussed (and several implemented) a clear reduction in $L_c$ was never observed. Moreover many of these methods required excessive computational resources, which made them unfit for even the simplest problems.

Ultimately, the connections to SC-4DVar and moving horizon estimation brought about a realization that the main benefit of these methods is that one does not have to estimate the whole window at once. But rather, by including an appropriate background term (or 'cost to-go') that in effect summarizes the past information, the use of a sliding window (or moving horizon) of observations can improve algorithmic efficiency, accuracy, and stability, to effectively reduce the critical observability limits of the system.

From this point of view, the generalization to include weak constraints is quite straightforward.

And a number of published formulations already exist in the literature on moving horizon estimation (*e.g.*, [163]).

The time delay filtering method thus shares a number of core features with the EKF and 4DVar. It brings together related ideas from a variety of fields, and adds to the base of knowledge surrounding this problem, and provides a framework for comparing new and existing algorithms.

## 4.2 Time delay methods for reducing observability limits

The effectiveness of using time delay to reduce the observability limits of the Lorenz 96 model will now be examined. This study will use a framework similar to the one described in Sec. (3.3). Two additional parameters must be selected however: the time delay $\tau = m\,dt$ and the embedding dimension $M$. Unless otherwise specified, the former is chosen to be a constant $\tau = 0.1 = 10\,dt$ (so $m = 10$). This value is roughly consistent with the average mutual information criterion, and the overall results were not sensitive to varying it a few $dt$ in either direction. The choice embedding dimension is more subtle however, as each additional time delay acts effectively as an additional measurement. Since the goal here is to characterize the minimum observational requirements, the search is performed first over $M$, for fixed $L$. Only when this fails for all values $1 \leq M \leq 2D$, is the number of 'physical' observations $L$ increased.

To clarify the exposition and provide baseline values for the reduction of $L_c$, results involving the two time delay methods discussed above (feedback synchronization and the EKF) are given first. These methods will then be compared against fixed-lag versions of 4DVar. The tradeoffs associated with long overlapping analysis windows will also be examined.

### 4.2.1 Time delay synchronization and the EKF

The minimum number of observations for time delay feedback synchronization (TDVar-FB) and the time delay EKF (TD-EKF) are plotted in Fig. (4.1) as a function of model dimension $D$. Compared with the $M = 1$ results from the previous chapter, the use of time delays reduces the observational requirements of both methods. The reduction is most pronounced for TDVar-FB, which drops from $L_c \approx 0.45\,D \rightarrow 0.06\,D$ with local initial conditions and no observation noise. With noise, it is reduced roughly by half, from $L_c \approx 0.41\,D \rightarrow 0.24\,D$.

The TD-EKF also shows some improvement, although not as much as TDVar-FB. Without observation noise, the results are roughly the same as the $M = 1$ case. When observation noise is present,

time delays reduce $L_c$ by roughly a third for both local and global initial conditions. These values are similar to those obtained from the adaptive $\boldsymbol{R}_f$ annealing technique described at the end of Sec. (3.3), suggesting that time delays provide an alternative (albeit more computationally intensive) approach to improving robustness to observation noise.

Note how the use of time delays reduces the effective observability thresholds for TDVar-FB to levels commensurate with the TD-EKF. In contrast to the $M = 1$ case, where the EKF significantly outperforms 3DVar-FB with no observation noise, this appears to hold regardless of initial conditions and the presence or absence of observation noise. That TDVar-FB also shows similar performance, is not totally surprising. Recall that it is the $|\boldsymbol{R}_m| \to \infty$ limit of SC-4DVar, which for linear problems is known to be equivalent to the EKF provided a moving horizon approach is adopted ([115]). And here, the time step $dt$ is short enough that the problem is 'almost' linear.

There is also a rough correspondence between the *total* number observations $LM$ (*i.e.*, including time delays) and the average dimension of the unstable subspace $D_u$. This is shown in Fig. (4.2) for local initial conditions, with no observation noise. Note in particular that up to $D \approx 40$, TDVar-FB follows the trend rather closely. Since $L = 1$ in this regime, the number of time delays roughly satisfies $M_c \approx D_u$. For $D > 40$, $L > 1$ observations are needed, and the pattern somewhat dissipates. The TD-EKF also exhibits lower values of $M_c$ up to $D \approx 40$, although it becomes more erratic as the dimension grows higher. On the other hand, it succeeds with $L = 1$ observation until $D \approx 80$, reaffirming that enhanced efficiency comes at the cost of stability.

With global initial conditions, or when observation noise is present, values of $L_c M_c$ are typically above $2D$, and $M_c$ appears to be uniformly distributed between 5 and 15. The analogy to SC-4DVar means that similar restrictions apply regarding the length of the horizon $T' := (M-1)\tau$ and chaos in the forecast model. As $M$ increases, so does $T'/\lambda_{max}^{-1}$, which causes the condition number of $\nabla \mathcal{H}_n$ to grow exponentially. There are many ways to improve this ill-conditioning, the majority of which fall under the framework of Tikhonov regularization. This will be discussed in more detail in Sec. (4.3).

These results indicate the extent to which the time delays act as additional measured state components, although there are limits to this interpretation. Specifically, as $D$ and $M$ grow, the problem eventually becomes ill-conditioned enough to require an increase in the number of 'physical' observations $L$. Nonetheless, the use of time delays — and in particular, the *re-use* of data in overlapping windows — provides an effective way to reduce the observational requirements of a given problem.

### 4.2.2 The choice of $\tau$

In phase space reconstruction, a suitable choice of $\tau$ is needed to distinguish between neighboring points on the attractor. When $\tau$ is too small, the dynamics have not been allowed enough time to provide adequate separation between neighboring points, and the embedding is highly susceptible to noise. Likewise, if $\tau$ is too large, neighboring points may be too far decorrelated to be of any use.

As these noise canceling properties are also be desirable in the estimation context, the choice of time delay is now examined by rerunning the previous analysis with $\tau = dt$. With this selection, there is effectively no delay. All measurements within the window are utilized at each step. And one expects $\nabla\mathcal{H}_n$ to be more ill-conditioned, as its rows are now more collinear.

Results shown in Fig. (4.3) indicate that without observation noise, the shorter delay does not make much of a difference. Somewhat surprising however, is that for global initial conditions, $L_c$ for the TD-EKF is reduced by about 25%, from $0.20\,D \rightarrow 0.15\,D$. This again points to the regularizing capability of the estimated error covariance. On the other hand, when observation noise is present, the efficiency of both algorithms are largely reduced to $M = 1$ levels. Indeed, 70% of these runs find critical values at $M = 1$.

This implies that the embedding is substantially more ill-conditioned with $\tau = dt$ than with $\tau = 10\,dt$, and thus more sensitive to observation noise. This conclusion is further supported by Fig. (4.4), which plots the singular value spectrum of $\nabla\mathcal{H}_n$ sampled across the attractor for $D = M = 20$ and $L = 1$. Also, for comparison purposes, the singular values are rescaled so that $\sigma_1 = 1$. Evidently, $\tau = 10\,dt$ is the most well conditioned, followed by $\tau = dt$ and then $\tau = 100dt$. Note also how the spectrum decreases exponentially, at roughly uniform intervals. This makes it difficult to determine where to truncate rank of the pseudoinverse, although some useful heuristics will be discussed in Sec. (4.3). But the main point here is that when observation noise is present — and it always is in practice — the success of the time delay embedding technique requires an appropriate choice of $\tau$.

### 4.2.3 Overlapping estimation windows

Operational implementations of SC-4DVar typically do not use overlapping observation windows for two primary reasons. First, reanalyzing the data in small increments $dt$ requires additional computational power, which may become prohibitively expensive for large problems. Second, there is a statistical argument made against reusing data in that, if the observations contain biased noise, those biases are introduced multiple times into the analysis.

For linear models with Gaussian errors, assimilating the data multiple times is equivalent to assimilating the same data once, provided $\boldsymbol{R}_m$ is scaled appropriately ([49]). From a fixed interval smoothing perspective, reuse of observations amounts to using a power of the conditional distribution $P(\mathcal{X}|\mathcal{Y}) \propto \exp[-A(\mathcal{X})]^{\gamma}$. This amounts to a rescaling of the objective function, and the geometry of the problem is not altered in any appreciable way.

In the nonlinear case however, the benefits of iterative methods that reanalyze the same data multiple times are quite well-established ([81, 14]. There is also more recent evidence suggesting that multiple data assimilation with long overlapping observation windows improves stability and accuracy of the analysis ([25]).

This motivates an examination of the impact of overlapping estimation windows on the minimum observability limits. The results Fig. (4.1) with $\tau = 10\,dt$ are now rerun, but after each analysis the resulting estimate is integrated forward the length of the sliding window $T' = ((M-1)\,m+1)\,dt$. This choice makes the windows disjoint, so each observation is used only once.

Results displayed in Fig. (4.5) exhibit sharply reduced performance compared with Fig. (4.1), and coincide almost exactly with the $M = 1$ results from Chap. (3). Indeed, with the exception of a few outliers, $M_c = 1$ for all cases. So the best results are obtained from running the analysis at every $dt$.

A thorough examination of critical thresholds in the *temporal* resolution of measurements will not be given here. However, preliminary results with $D = 20$ indicate that the observational benefits are lost once the time between subsequent analyses rises above $5\,dt$. And both methods fail when this period reaches roughly $15-20\,dt$, even when fully observed.

The frequency of observations is evidently just as critical to the analysis as the spatial resolution. Both exhibit critical thresholds at which estimates consistently fail. And both issues be mitigated through the use of time extended observations. Time delays effectively act as additional measurements. They improve observability by enriching the observation space with information from the forecast model. And increase the temporal frequency of the observations by allowing the analyses to be performed in between measurement times.

It is somewhat surprising these benefits seem to require the reuse of data with overlapping analysis windows, as this has no analog in the linear theory. And certainly, there are tradeoffs between accuracy, robustness and computational efficiency, which have to be addressed in the context of the particular application. But putting these issues aside, when the data is spatiotemporally sparse, time-embedded measurements appear capable of improving the effective observational limits of a given estimation algorithm.

### 4.2.4 Moving horizon estimation with 4DVar

As mentioned, the time delay feedback synchronization method is a special case of *cycled* SC-4DVar with a few differences, notably: 1) all observations are used (although time delays may be included by generalizing to a time-dependent $\mathcal{R}_m$), 2) the analysis increment $\delta\boldsymbol{x}$ is found using Tikhonov regularization instead of rank-truncated SVD, and 3) observation windows are typically disjoint. The impact of the first two on $L_c$ is now investigated — the effect of disjoint analysis windows will be examined momentarily. So for now, all methods use overlapping estimation windows that move in increments of $dt$.

The observational efficiency of the three smoothing variants from Chap. (3) is now compared using a fixed-lag approach. As before, no adjoints are used. The derivatives are all calculated directly, and minimization is performed using the full Gauss-Newton procedure. The length of the analysis window is chosen relatively short, $T = 0.25/\lambda_{\max}$, to avoid instability issues with SC-4DVar and WC-4DVar-U. The state variables are initialized with $\boldsymbol{x}_0$ as the random initial guess, with remaining path components $\mathcal{X}$ and $\mathcal{U}$ set to zero. After each measurement update, the resulting values are reused in the subsequent analysis, after shifting them by one $dt$.

As before, $\boldsymbol{R}_m$ and $\boldsymbol{R}_f$ are set to the identity. However, in this case the results were observed to benefit from the addition of a background term $\boldsymbol{R}_b = \boldsymbol{I}$. In Chap. (3), the background term was neglected under the assumption that nothing is known about the accuracy of the initial guess. Here however, this accuracy will ideally improve with each iteration. Adding the background term helps regularize the solution, by keeping the analysis from deviating too far form the previous estimate.

Calculated values of $L_c$ shown in Fig. (4.6) indicate robustness both to observation noise and the accuracy of initial conditions. Contrasting this with the fixed interval results from Fig. (3.13), where low noise and accurate initial estimates were required to reduce these limits below $L_c \approx D_u$, further highlights the benefits of the moving horizon approach. The resulting thresholds of WC-4DVar-X and WC-4DVar-U are roughly on par with the 3DVar-FB results in Fig. (3.3), with values approximately equal to $D_u$. Although SC-4DVar evidently does much better, with thresholds around $L_c \approx 0.2\,D$, this is likely related to the absence of model error from the simulated data. More realistic tests, which include model error, are needed to provide a more useful comparison.

## 4.3 Optimizations and extensions

Several implementation issues with time delay methods are now examined. These issues include: the solution to Eqn. (4.2), the choice of embedding parameters and the rank of the pseudoinverse.

### 4.3.1 Rank considerations for the pseudoinverse

There are many approaches available for solving the linear system Eqn. (4.2). Recall that since $\nabla\mathcal{H}_n$ is generally not square, a pseudoinverse must be used. This can be done simply by direct inversion

$$[\nabla\mathcal{H}_n]^+ := [\nabla^\dagger\mathcal{H}_n \cdot \nabla\mathcal{H}_n]^{-1} \cdot \nabla^\dagger\mathcal{H}_n,$$

but this technique has numerical stability problems when $\nabla\mathcal{H}_n$ is ill-conditioned. Specifically, if $\nabla\mathcal{H}_n$ has condition number $\kappa$ then $\nabla^\dagger\mathcal{H}_n \cdot \nabla\mathcal{H}_n$ has condition number $\kappa^2$. When applicable, a more robust approach is to compute the SVD of $\nabla\mathcal{H}_n = \boldsymbol{U}_n \cdot \boldsymbol{\Sigma}_n \cdot \boldsymbol{V}_n^\dagger$, and construct the pseudoinverse

$$[\nabla\mathcal{H}_n]^+ = \boldsymbol{V}_n \cdot \boldsymbol{\Sigma}_n^+ \cdot \boldsymbol{U}_n^\dagger.$$

The rectangular diagonal matrix $\boldsymbol{\Sigma}_n^+$ is defined by taking the reciprocal of each nonzero element, leaving the zeros in place. In practice, this means inverting elements whose values are above a small tolerance $\varepsilon$. Elements below this threshold are replaced with zeros. The choice of tolerance therefore determines the rank $r$ of the matrix $[\mathcal{H}_n]^+$, which as shown above plays an important role in the convergence and stability of the filter.

The default tolerance for most pseudoinverse routines in on the order of machine precision. But for this application, this choice may produce excessively large control perturbations $\delta\boldsymbol{x}_n$ that destabilize the filter. This highlights an inherent tradeoff between numerical stability and performance. While one could simply reduce the coupling terms $\boldsymbol{K}$, this was not shown to be an effective approach. Alternatively, raising this tolerance (*e.g.*, to $\varepsilon \sim 10^{-3}$) stabilize the calculations, but degrade its performance by reducing the rank $[\mathcal{H}_n]^+$ below the critical threshold $r_c \approx D_u$. These results suggest that a static tolerance may not be the best approach, prompting an exploration of other ways to choose the $r$.

One option is to choose a constant $r$ throughout the whole estimation process. As shown in the examples above, this choice provides insight into the role that $r$ plays in stabilizing the synchronization manifold. Restricting to a constant rank also makes the pseudoinverse differentiable [63]. These derivatives

may be used to calculate the conditional Lyapunov exponents and prove convergence under certain conditions. The main drawback is that this choice must be made conservatively to avoid destabilization along the entire trajectory.

A number of ideas for adaptively choosing $r$ were also explored. Of these, the best performance was obtained by choosing $r$ as large as possible so that the resulting perturbation is below a threshold set relative to the unperturbed model. That is,

$$r = \underset{1 \leq i \leq r_{max}}{\operatorname{argmax}} \left[ \frac{\delta x_i(t_n)}{F_i(\boldsymbol{x}(t_n))} \leq \varepsilon \right].$$

The advantage of choosing $r$ relative to the $\boldsymbol{F}_n$ is that it sets the scale of the perturbation, assuming the dynamics are assumed to be inherently stable. It also normalizes the effective threshold to account for different scales of the state variables, so the resulting choice is effectively invariant to say the choice of units. The main drawback of this technique is that it requires recomputing the perturbation for several choices of $r$, although the SVD only has to be computed once.

### 4.3.2 Tikhonov regularization

While SVD may be the most numerically robust way of solving Eqn. (4.2), many other viable options are available. One option to improve computational efficiency, is to rewrite Eqn. (4.2) as a least squares objective function and perform Tikhonov regularization,

$$\delta \boldsymbol{x}_n = \underset{\delta \boldsymbol{x}}{\operatorname{argmin}} |\mathcal{Y}_n - \mathcal{H}_n - \nabla \mathcal{H}_n \cdot \delta \boldsymbol{x}|^2 + |\delta \boldsymbol{x}|^2_{\boldsymbol{\Gamma}_n}. \tag{4.7}$$

The matrix $\boldsymbol{\Gamma}_n$ is chosen to give preference to solutions with desired properties. For instance, $\boldsymbol{\Gamma}_n = \varepsilon \boldsymbol{I}$ gives the pseudoinverse in the limit $\varepsilon \to 0$, although it differs from both the approaches above in that the singular values are not truncated below a certain value. Rather, the solution

$$\delta \boldsymbol{x}_n = \boldsymbol{V}_n \cdot (\varepsilon \boldsymbol{I} + \boldsymbol{\Sigma}_n^\dagger \cdot \boldsymbol{\Sigma}_n)^{-1} \cdot \boldsymbol{\Sigma}_n^\dagger \cdot \boldsymbol{U}_n^\dagger \cdot \mathcal{Y}_n$$

acts as a spectral filter that puts more weight components of $\boldsymbol{V}_n$ with singular values above $\varepsilon$, and less weight on those below.

Substituting $\boldsymbol{\Gamma}_n \to \boldsymbol{R}_b$ and introducing a measurement error $\mathcal{R}_m$ term gives incremental SC-4DVar ([40]). This was shown by [82] to perform Tikhonov regularization using the background error covariance

$\boldsymbol{R}_b$ as the filter. A main benefit of this formulation is that it allows $\delta\boldsymbol{x}_n$ to be computed efficiently, using gradient descent and adjoint methods that only require $O(D)$ memory. This makes it a viable method for high dimensional problems such as operational numerical weather forecasting, where direct computation of $\nabla\mathcal{H}_n$ is not feasible.

The EKF also performs a variant of Tikhonov regularization, except with $\boldsymbol{\Gamma}_n \to \boldsymbol{R}_n$. The estimated error covariance has some desirable properties, as exhibited by the EKF's low $L_c$. As a filter for $\nabla\mathcal{H}_n$, it targets the locally unstable subspace, reducing sensitivity to noise as well as errors in the forecast, On the other hand, it is not possible to compute $\boldsymbol{R}_n$ directly for large problems, as the Riccati equation requires storage and manipulation of $O(D^2)$ matrices. This has prompted the recent development of 'hybrid' SC-4DVar methods, which blend the static background $\boldsymbol{R}_b$ with ensemble techniques that aim to capture the forecast sensitivity ([117]). While a thorough comparison of $L_c$ among these methods is the topic of future work, there is much to be gained from studying the observability properties of these algorithms, and their relation to the full EKF.

### 4.3.3 A parallelized adjoint formulation of the Gauss-Newton method

A benefit of Tikhonov regularization is that the perturbation $\delta\boldsymbol{x}_n$ can be found by directly minimizing Eqn. (4.7). This makes this approach suitable for high dimensional problems, provided the derivatives $\nabla A$ needed by the optimization procedure are calculated using adjoint methods.

Adjoint methods may be viewed in the Hamiltonian formalism developed in Chap. (2), wherein the model constraints through the addition of a Lagrange multiplier $\boldsymbol{p}_n$, which shares the same role as the canonical momentum in Hamiltonian mechanics. Enforcing the necessary conditions for stationarity (*i.e.*, $\nabla A = 0$) gives the analog of Hamilton's equations, in the form of a two-point boundary value problem. Using these equations to integrate $\boldsymbol{x}_n$ forward from $0 \to T$, and $\boldsymbol{p}_n$ backward from $T \to 0$, produces the gradient of the objective function $\boldsymbol{p}_0 = \nabla A$. The benefit of this approach is that it uses only $O(D)$ memory. It avoids having to compute and store the $D \times D$ matrices $\boldsymbol{\Phi}_{n,0}$ required by the direct approach, using the variational Eqn. (4.5).

For instance, consider the SC-4DVar objective function given in Eqn. (3.8). Relaxing the constraints by including the constraints using a Lagrange multiplier gives

$$A(\mathcal{X},\mathcal{P}) \propto |\boldsymbol{x}_0 - \boldsymbol{x}_b|^2_{\boldsymbol{R}_b} + \sum_{n=0}^{N} |\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)|^2_{\boldsymbol{R}_m} + \langle \boldsymbol{p}_{n+1}, \boldsymbol{x}_{n+1} - \boldsymbol{F}(\boldsymbol{x}_n)\rangle.$$

Setting $\nabla A(\mathcal{X},\mathcal{P}) = 0$ yields Hamilton's equations,

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n)$$
$$\boldsymbol{p}_n = \nabla^\dagger \boldsymbol{F}_n \cdot \boldsymbol{p}_{n+1} + \nabla^\dagger \boldsymbol{h}_n \cdot \boldsymbol{R}_m \cdot \big(\boldsymbol{y}_n - \boldsymbol{h}(\boldsymbol{x}_n)\big) + \boldsymbol{R}_b \cdot (\boldsymbol{x}_b - \boldsymbol{x}_0)\,\delta(n),$$

where $\delta(n)$ is a delta function. Enforcing these equations makes $A(\mathcal{X},\mathcal{P})$ a function of $\boldsymbol{x}_0$ and $\boldsymbol{p}_{N+1}$ alone. Setting $\boldsymbol{p}_{N+1} = 0$, the gradient reduces to

$$\nabla A(\boldsymbol{x}_0) \propto \sum_{n=0}^{N} \big(\boldsymbol{\Phi}_{n+1,0} \cdot \boldsymbol{p}_{n+1} - \boldsymbol{\Phi}_{n,0} \cdot \boldsymbol{p}_n\big) = \boldsymbol{p}_0.$$

The same procedure can may be used for WC-4DVar-X and WC-4DVar-U as well ([193]). For WC-4DVar-U, the derivative with respect to the initial condition is the same as for SC-4DVar, except that now the trajectory is evaluated along the approximate model trajectory, that includes the control terms. The derivatives with respect to the control $\boldsymbol{u}_n$ is obtained from restricting the sum to times $n' \geq n$. The full gradient $\nabla A$ is still therefore computed from one forward integration of the model, and one backward integration of the adjoint. On the other hand, for WC-4DVar-X the estimated path is not an approximate model trajectory. Therefore it does not require a forward integration from $0 \to T$. The adjoint model can still be used to efficiently compute the derivatives with respect to $\boldsymbol{x}_n$. These calculations may also be parallelized, with each derivative being calculated in its own thread.

While adjoint methods enable the scalable computation of gradients, there is still the issue of how to perform the optimization. For large problems, this has led to the development of 'incremental' algorithms ([40, 193]), which split the procedure into two loops. In the 'outer loop', the full nonlinear model is run forward from the previous estimate. The problem is then linearized around the resulting trajectory, and used to initialize the 'inner loop', which minimizes

$$\delta\boldsymbol{x}_n = \underset{\delta\boldsymbol{x}}{\operatorname{argmin}} |\boldsymbol{x}_b - \boldsymbol{x}_n - \delta\boldsymbol{x}|^2_{\boldsymbol{R}_b} + |\mathcal{Y}_n - \mathcal{H}_n - \nabla\mathcal{H}_n \cdot \delta\boldsymbol{x}|^2_{\mathcal{R}_m}. \tag{4.8}$$

to find the optimal perturbation $\delta\boldsymbol{x}_n$. The inner loop is then iterated, to refine the estimate for $\delta\boldsymbol{x}_n$ until some prescribed stopping criteria is reached. Control then passes back to the outer loop, which constructs the new estimated trajectory from the resulting estimate $\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n + \delta\boldsymbol{x}_n$. The entire analysis process may then repeated using $\boldsymbol{x}_n$ as the new initial guess.

The inner loop minimization minimization can be carried in a number of ways. Most simply, gradient descent should (eventually) reach a local minimum, and can be done directly, avoiding the need

to invert any large matrices. Alternatively, faster convergence may be obtained from a Gauss-Newton approach, which involves solving a linear system,

$$\begin{aligned}(\boldsymbol{R}_b + \nabla^\dagger \mathcal{H}_n \cdot \mathcal{R}_m \cdot \nabla \mathcal{H}_n) \cdot \delta \boldsymbol{x}_n^{(i+1)} = \boldsymbol{R}_b \cdot (\boldsymbol{x}_b - \boldsymbol{x}_n - \delta \boldsymbol{x}_n^{(i)}) \\ + \nabla^\dagger \mathcal{H}_n \cdot \mathcal{R}_m \cdot (\mathcal{Y}_n - \mathcal{H}_n - \nabla \mathcal{H}_n \cdot \delta \boldsymbol{x}_n^{(i)}),\end{aligned} \tag{4.9}$$

where $i$ denotes the inner loop iteration. The solution is often obtained using a preconditioned conjugate gradient method ([139]), which exploits sparsity in the approximate Hessian. However, this still requires evaluating the product of a vector with the approximate Hessian. While $\nabla \mathcal{H}_n$ can be computed directly from Eqn. (4.5), or using a second order adjoint method ([149]), these approaches require storage and manipulation of $O(D^2)$ matrices (or higher for weak constraint methods), making them unsuitable for large problems.

How the Hessian is implemented in practice is rarely mentioned in the literature, especially in connection with the Gauss-Newton method. Evidently, the inner loop is run at a lower spatial resolution to accelerate the calculations ([107]), but whether this makes computation of $\nabla \mathcal{H}_n$ tractable remains uncertain. Also, the preconditioning transformations are constructed from the background covariance $\boldsymbol{R}_b$ ([67]), which effective ignores any dynamical information contained in $\nabla \mathcal{H}_n$. Thus, while ([108]) showed that minimizing Eqn. (4.8) is equivalent to an approximate solutionEqn. (4.9), it is unclear how the Gauss-Newton method would actually be implemented without a scalable way to evaluate $\nabla \mathcal{H}_n$.

One way to do this stems from the time embedding interpretation of SC-4DVar. The rows of $\nabla \mathcal{H}_n$ in Eqn. (4.4) correspond to the derivatives with respect to the time delayed model state. These sensitivities may be computed row by row using the adjoint model $\nabla^\dagger \boldsymbol{F}_n$. Setting $\boldsymbol{p}_{m+1}$ equal to the $\ell^{th}$ row of $\nabla \boldsymbol{h}_m$, the corresponding row $(L(m-1)+\ell)$ in $\nabla \mathcal{H}_n$ can be calculated from the recursion $\boldsymbol{p}_n = \nabla^\dagger \boldsymbol{F}_n \cdot \boldsymbol{p}_{n+1}$. This gives a way of computing $\nabla \mathcal{H}_n$ without having to compute the matrices $\boldsymbol{\Phi}_{n,0}$. The drawback of this is that it must be carried out for all $ML$ rows of the time delayed measurement operator. The upshot however, is that it be massively parallelized, by independently computing the contribution of each row on its own processor.

The benefits and feasibility of this technique for large-scale problems must still be demonstrated. It would be interesting for instance, to see if there is a discernible difference in $L_c$ among optimization techniques, such as gradient descent and Gauss-Newton, both with and without adjoint derivatives. The above results indicate the rank of the time embedded observation operator $\nabla \mathcal{H}_n$ is important to its success. For gradient descent, these results may translate to additional criteria on the number of iterations required

by the inner and outer loops. Furthermore, as the implementation of the adjoint model is (at times) difficult and cumbersome, it would also be useful to compare these results against those obtained from derivative free methods, such as the Ensemble Kalman filter. All this will be explored in future work.

### 4.3.4 Optimizing the embedding

The choice of embedding is now addressed. In the simplest case, two parameters must be selected: a uniform delay $\tau$ and the embedding dimension $M$. As mentioned above, useful heuristics exist for this choice ([55, 2]). There is however no need to restrict consideration to a fixed, uniform $\tau$. This choice is expedient and works well in practice. While some more recent efforts have focused generalizing these ideas to non-scalar signals with non-uniform delays, this is still an active area of research ([59, 158, 201]).

In estimation, the goals are different however. Attractor reconstruction is concerned with discerning the model from the data. But in estimation, the model is assumed known and may therefore be used to inform the choice of embedding. An idea for using the dynamical model to optimize the choice of embedding will now be explored.

Recall that a forward embedding was chosen for the time delay methods based on the intuition that this would enhance observation and control of the forward instabilities in the model. There is no reason however, why a backward embedding (or a mixture of the two) could not be used instead. The goal is to choose an embedding dimension $M$ and a set of delays $\{\tau_1, \tau_2, \ldots, \tau_M\}$ to optimize a local measure of observability.[2] Similar ideas have been explored in ([154, 155, 173]), as a way to guide the optimal choice of parameters $\tau$, $M$ for a static uniform embedding. This idea is now generalized to consider a non-uniform time embedding. Given the 'full' $\nabla\tilde{\mathcal{H}}_n$ matrix that contains all measurements within the estimation window, the problem is to choose a subset of its rows to minimize the condition number $\kappa_n = \sigma_n^{\max}/\sigma_n^{\min}$ of $\nabla\mathcal{H}_n$, where $\sigma_n$ are its singular values.[3]

Optimization problems involving singular values tend to be difficult, as the operation is in general discontinuous. To avoid the need for derivatives, a simple Markov Chain Monte Carlo approach will be adopted here, based on the Metropolis-Hastings algorithm ([127]). The method proceeds as follows. Choosing $M = 2D+1$ based on Takens' theorem, construct the partial $\nabla\mathcal{H}_n$ corresponding to a non-uniform set of delays $\{\tau_1, \tau_2, \ldots, \tau_M\}$ and compute its $\kappa_n$. Randomly choose one delay $\tau_m$ to increase or decrease by one $dt$. Calculate the new $\kappa_n'$. If it is lower than the previous value, accept the step. Otherwise, the step is rejected with probability $\min(1, \exp[\beta * (\kappa_n - \kappa_n')])$. The parameter $\beta$ controls the 'temperature'

[2] Also known as an 'observability index' ([111]).

[3] This problem is also known as 'column subset selection' ([197]).

of the ensemble, helps prevent it from becoming stuck in local minima. The sampling process is then repeated, until $\kappa_n$ converges to within some prescribed variance.

Tests using Lorenz 96 with $D = 20$ show evidence for clusters of measurements within the embedding space both forward and backwards in time. These clusters produce local minima of the condition number, although since the procedure is random, different results are obtained for different initial distributions. The overall patterns are reasonably consistent however, as certain times are clearly targeted and roughly spaced at $50\,dt$, five times larger than the $\tau$ used above.

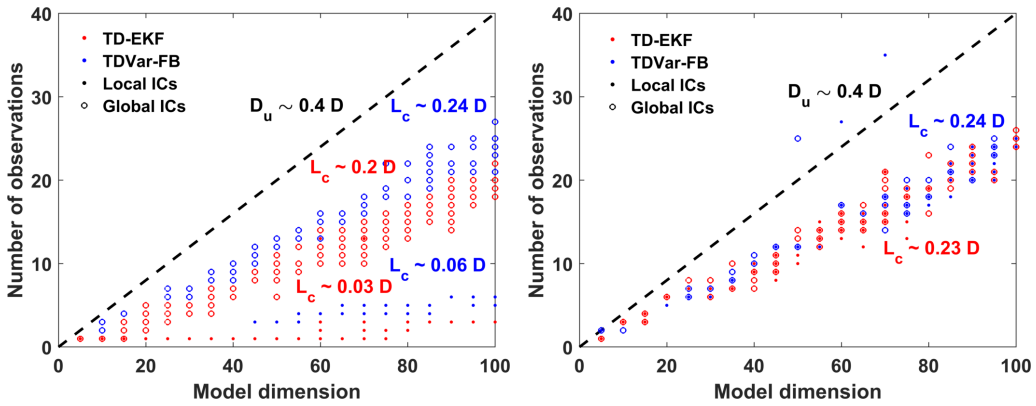
These results offer proof of concept that this 'observability sampling' scheme may be useful for selecting an optimal embedding with non-uniform delay and supports the claim that $M = 2D+1$ delays are not needed. Indeed, smaller condition numbers are typically obtained with lower $M$, although there are limits to this, as a single observation (or equivalently, taking $M$ identical delays) will produce a condition number of $\kappa_n = 1$. For this reason, the objective $\sum_i \sigma_n^i / \sigma_n^{\max}$ was observed to work somewhat better than $\kappa_n$.

The method may also be used 'on-line', by first optimizing the embedding at a fixed reference time $t_0$, then evolving forward in time by one $dt$ and repeating the sampling process. The hope is that once the initial optimization is performed, fewer iterations would be needed to update it. The conditioning of $\nabla \mathcal{H}_n$ is tied to this reference time, so the sampling procedure will update the embedding accordingly as $t_0$ is increased. This approach can also be used to help mitigate the effects at the beginning and end of the estimation window, by simply restricting $\tau_m$ to be positive and negative respectively.

Certain limitations must be overcome however for this to be a viable approach for real problems. For instance, only scalar ($L = 1$) observations are considered here. With multivariate time series, a choice has to be made about how to distribute the delays across the $L$ variables. This could be handled by the Markov chain itself, by allowing transitions from one observed component to another. But this adds an additional layer of complexity to the method, and was not considered here.

The approach also has some obvious problems with scalability, not the least of which is the computational expense of computing the singular values of the matrix $\nabla \mathcal{H}_n$ at each iteration. One option is to use a fast approximation method for condition numbers, such as the one given by [69]. Or alternatively, use a different objective function that requires less computation. These ideas will be explored in more detail elsewhere. Also, if the matrices $\nabla \mathbf{\Phi}_{n,m}$ can be stored, the trajectory does not need to be recalculated at each iteration, but rather extended as needed. Otherwise, the required rows of $\nabla \mathcal{H}_n$ can be computed in parallel using the adjoint approach described above.

Chapter 4, in part is being prepared for submission for publication of the material. Rey, Daniel; Abarbanel, Henry DI. The dissertation author was the primary investigator and author of this material.




(a) Local and global estimates, no noise (b) Local and global estimates, with noise

**Figure 4.1**: The critical minimum number of observations $L_c$ for time delay estimation methods, with $\tau = 10\,dt$.

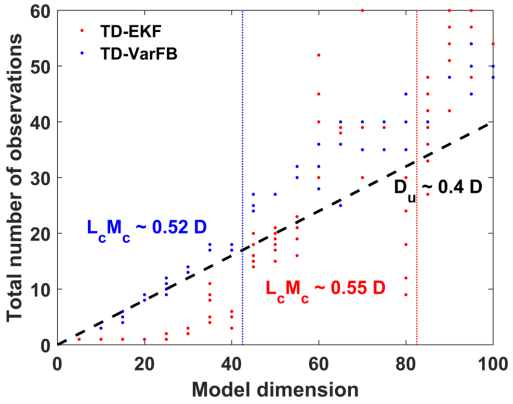
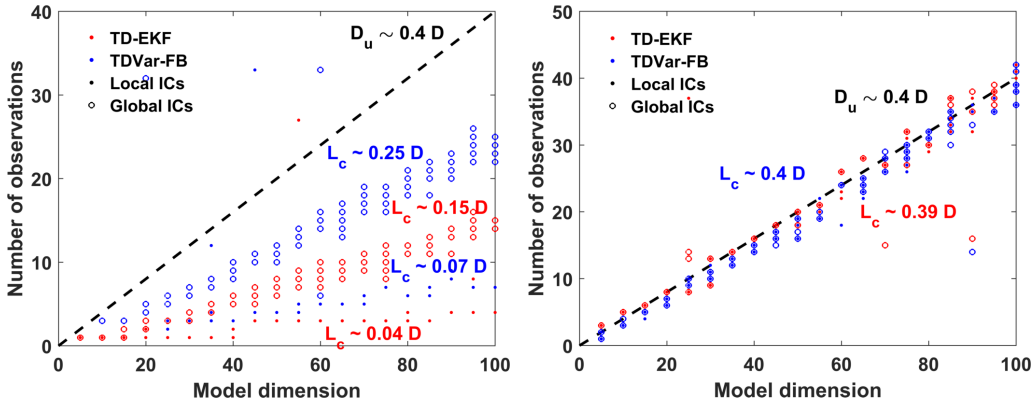



**Figure 4.2**: The critical minimum number of *total* observations $L_c M_c$, for local initial conditions with no observation noise. The trend for TDVar-FB follows $D_u$ until roughly $D \approx 40$, above which additional 'physical' measurements are required. By contrast, the TD-EKF requires only $L = 1$ until $D > 80$.




(a) Local and global estimates, no noise (b) Local and global estimates, with noise

**Figure 4.3**: The effect of the time delay is examined by repeating the experiments from Fig. (4.1) with $\tau = dt$.

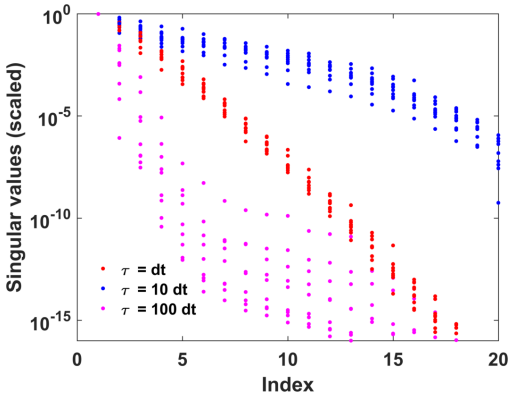
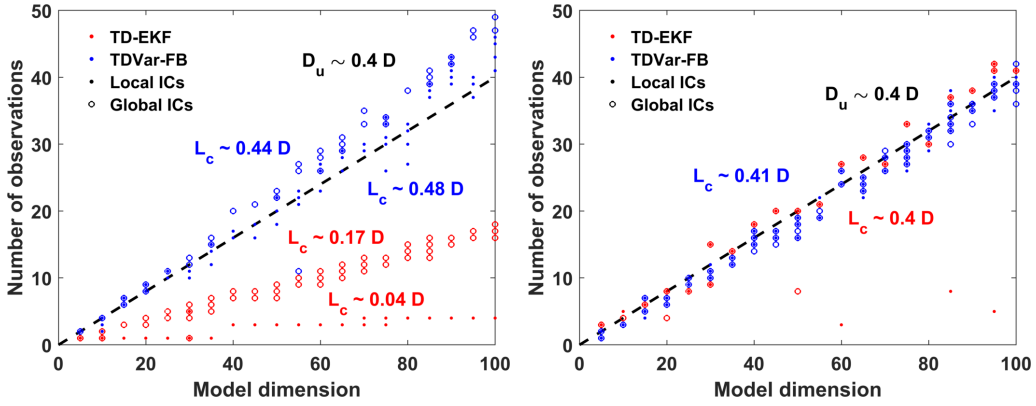
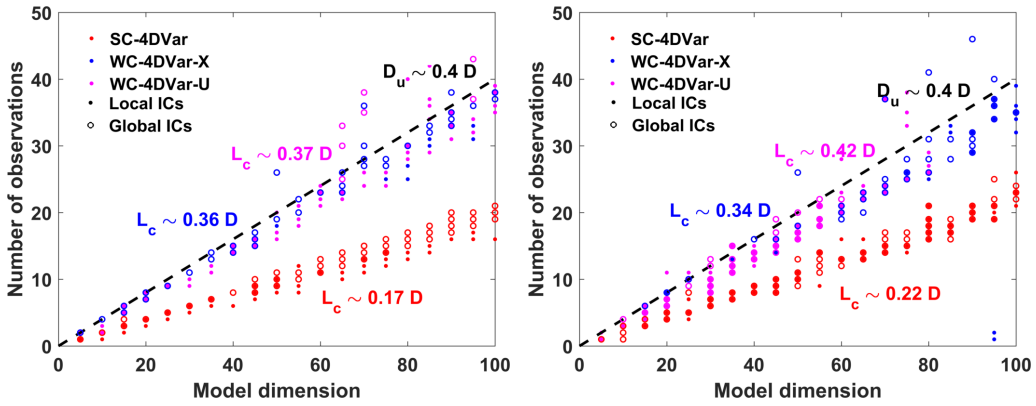



**Figure 4.4**: Singular value spectra of $\nabla\mathcal{H}$ rescaled so that $\sigma_1 = 1$, for different values of $\tau$. Points are sampled across the attractor, with $D = M = 20$, $L = 1$.




(a) Local and global estimates, no noise (b) Local and global estimates, with noise

**Figure 4.5**: Repeating Fig. (4.1) using disjoint estimation windows. The vast majority of results find $M_c = 1$ as the critical threshold, indicating that overlapping windows are needed for enhanced observational efficiency.




(a) Local and global estimates, no noise

(b) Local and global estimates, with noise

**Figure 4.6**: Critical observability limits of moving horizon estimation methods, using the three 4DVar techniques from Chap. (3).

# 5 Conclusions and future work

Estimation is one of the most ubiquitous and fundamental problems of existence. Its theoretical development goes back hundreds (if not thousands) of years, and has been shaped by many of history's greatest scientists and mathematicians. With the growing availability of both computational and data collection resources, the problem is now more important than ever.

Yet despite its undeniable importance and the overwhelming amount of attention it has received of late, there is evidently much that is still not understood about this most fundamental problem. This thesis attempts to improve this situation by examining it from a broader perspective that seeks to initiate the process of unifying shared concepts among the various fields interested in this problem. At the same time, it also proposes a number of new ideas, and tries to interpret them within this broader context.

What follows is a brief summary of the main results of each chapter. These synopses emphasize both what has been done and what is left to do. Indeed there is no shortage of remaining work. But the intention of this thesis has been to focus on the proposal and development of new ideas, leaving a more exhaustive treatment for future publications. The main ideas are brought back together in the end, to give a personal view on the future of estimation.

## 5.1 The canonical structure of optimal estimation

Chap. (2) reexamines the Kalman solution to the optimal filtering and smoothing problems from a deterministic perspective that focuses on the intersection of calculus of variation, optimal control, and classical mechanics. For fixed-interval smoothing, a number solutions have been proposed over the years. While these methods all give identical results for linear problems, this is evidently not true of their nonlinear extensions.

In particular, it appears that *sequential* (or two-pass, feasible path) smoothing algorithms have some inherent benefits. For one, they reduce computational complexity, by performing $N$ inversions

of $O(D^2)$ matrices. Compare this with direct optimization techniques *e.g.*, based on the Gauss-Newton method, which require solving linear systems of size $O(ND^2)$. Sequential methods scale better when the estimation window is long. Of course, when $D$ is extremely large inverting $O(D^2)$ matrices is also not feasible. Although there are other ways to work around this issue (*e.g.*, using adjoints or ensemble methods), these were largely left for future discussion.

More to the point however, sequential methods are unique in the sense that their recursive substructure propagates the state between observation times using the full nonlinear forecast model ([83]). By contrast, Newton's method (and its various generalizations and approximations) generates the transition using the tangent linear model. This result was derived explicitly, in both continuous and discrete time, from the estimation analog of differential dynamic programming. And it was shown how subtle differences in the various choices or assumptions in the problem statement can lead, at times, to widely different results.

This recursive substructure allows these algorithms to operate on a nearly feasible trajectory of the system. As shown in Chap. (3), this permits the forward (*i.e.*, filtering) pass to be interpreted as a dynamic form of feedback synchronization. The Kalman solution also evidently contains an inherent mechanism to target and control the unstable modes of the system, to make more efficient use of the available observations. This targeted control mechanism does not appear to be possible an approach such as WC-4DVar-X (or variants thereof), where the path is not constrained to be an approximate system trajectory.

Evidence was also presented to suggest that this recursive structure and the adaptive instability control mechanism are both a result of the role played by symplectic structure in the Kalman solution. While the Hamiltonian formulation of the optimal control problem is well-established, this representation is highly unstable. Although Hamilton's equations may be ostensibly viewed as a type of feedback synchronization, they cannot be integrated for an appreciable amount of time before the solution becomes numerically intractable. This phenomenon has been noted previously (*e.g.*, [133]), but has not garnered much attention.

Here it was shown that the Kalman solution may be derived by imposing symplectic structure on both the path $\boldsymbol{x}_t,\ \boldsymbol{p}_t$, and the first order fluctuations around it $\boldsymbol{\delta x}_t,\ \boldsymbol{\delta p}_t$. The latter in particular leads the Riccati equation for the approximate error covariance. And the Kalman filter emerges as an implicit solution to the conspicuously tautological equations $\boldsymbol{p}_t = \dot{\boldsymbol{p}}_t = 0$, upon making the Riccati transformation $\boldsymbol{p}_t \to \nabla^{\dagger} A_t(\boldsymbol{x}_t)$. From physical point of view, this transformation expresses the controls in the coordinate

system of the estimate and may be considered a type of gauge fixing. It implies that the Kalman filter solution enforces the canonical momentum (or adjoint/costate variables) to be *zero everywhere*, a result that is perhaps somewhat intuitive in hindsight. If one views $\boldsymbol{p}_t$ as the gradient of a dynamic objective function, the gradient must be zero for the estimate to be optimal with respect to the previously observed data.

This result was further elaborated from the point of view of Hamilton-Jacobi theory, both in continuous and discrete time. Although the latter approach, which is based on the discrete time action principle, clearly needs further attention. The use of generating functions was also discussed as a technique for solving two-point boundary value problems. And the previous work of ([151, 152, 150]) for feedback control systems was extended to the linear estimation problem.

Although the theoretical work described in this section is still largely incomplete, it nonetheless an alternative perspective on the problem that ties together a number of foundational ideas in estimation theory, optimization, and physics. A number of new research directions stand out in this research, most notably: 1) investigating the impact of the subtle choices that distinguish between various algorithms, 2) the use of generating functions for deriving stable feedback control/estimation algorithms, and 3) the role that symplectic structure and the canonical momentum play in the Kalman solution.

Notwithstanding the widespread use of Kalman filters/smoothers along with over half a century of dedicated theoretical efforts, there is evidently much that is still not understood about these equations — particularly, the relationship between their statistical properties and optimal geometric substructure. Unraveling these mysteries and forging connections across disciplinary boundaries will lead not only to a better understanding of what is perhaps the most elegant and important solution to the estimation problem to date, but also help fortify the theoretical foundation on which the next generation of algorithms will be built.

## 5.2 Observability and conditioning in dynamical inverse problems

The condition number of a matrix is a measure of the stability of solutions to linear systems that is independent of both the data and the algorithm. Chap. (3) discusses its extension to the nonlinear systems, with a particular emphasis on its application to dynamical inverse problems found in optimal estimation. In the estimation context, the condition number can be viewed as measuring the inherent difficulty in constructing a solution. Typically such problems are not fully observed, and are thus inherently ill-posed.

These problems therefore require regularization to break the degeneracy inherent in partial

observations, and make the solution unique. Regularization may be viewed as incorporating additional 'prior' information into the problem. This information can have many forms. For static problems, it is often simply some *a-priori* constraints or bounds on the parameters to be estimated. But for dynamical problems, knowledge of the system dynamics can also be used to constrain the solution.

Dynamical regularization can be a powerful tool, especially when the underlying processes are based on established scientific principles. Such 'high-confidence' models constrain the set of possible solutions, and thereby reduce the dependence of the solution on the accuracy of the initial estimate. In some cases, the hypothetical 'true' path can be reconstructed from totally arbitrary initial conditions. In other cases however, it can have the opposite effect. Thus, unlike for linear systems, the conditioning of dynamical inverse problems cannot be considered independently from the method for finding the solution, as it depends in a complex way on the three core components: the forecast and observation models, the data, and the estimation algorithm.

This chapter introduced techniques for analyzing these complicated dependencies. Broadly speaking, these methods may be categorized as either: 1) empirical, using only observed quantities, or 2) theoretical, using simulated data. While only the former is applicable to real experiments, the latter techniques not without benefit. They provide a way to examine the adequacy of the combined observation-analysis-modeling system under controlled conditions. Presumably, if the estimates are not successful under these rather idealized circumstances, there is little chance that they will be successful in practice.

Several estimation algorithms were compared in terms of their probability of success, as defined by when the filtered (or smoothed) RMSE drops below the noise level. A series of Bernoulli trials were repeated from random initial conditions distributed both locally (*i.e.*, near the solution) and globally across the attractor. This methodology directly compares algorithmic performance under near ideal conditions, to establish how the likelihood of constructing a successful estimate varies with certain aspects of the problem, such as: 1) the resolution of the forecast model and observations, 2) the choice of algorithm, and 3) the length of the estimation window. The results of this investigation and have led us to the following conclusions.

### 5.2.1 Filtering methods

Filtering methods necessarily require 'enough' observations to control the local dynamical instability in the model, so that the linearized error is asymptotically stable when averaged along the trajectory. This statement may be summarized in terms of the well-known necessary conditions for chaotic

synchronization given by [157], which states that the largest global Lyapunov exponent of the error system (otherwise known as the conditional LEs) must be negative.

However, the resolution of observations required to both satisfy this condition and achieve synchronization depends on the choice of algorithm. For instance, the static 3DVar filtering schemes examined here require at least roughly $L_c \approx 0.4D$ observations of the Lorenz 96 model produce synchronized estimates. This value scales linearly with the resolution of the model. It also happens to coincide with the attractor averaged dimension of the locally unstable subspace $D_u$, which turns out to provide an adequate baseline estimate for most of the methods tested here.

On the other hand, the extended Kalman filter can evidently do much better. In particular, when used as an adaptive observer under low observation noise conditions, it requires only $L_c \approx 0.04D$ for local initial conditions, and $L_c \approx 0.17D$ for global initial conditions. This value falls within the range of the approximate lower bound $r_c \approx 0.02D$ on the rank of the observation operator, with full $L = D$ observations. And it is the only method to show a significant boost in performance from starting in the vicinity of the truth.

This reduction in $L_c$ is attributed in part to the recently discovered fact that the Riccati equation for the estimated error covariance adaptively targets the unstable subspace, concentrating around it as the estimate converges. Numerical simulations show that when the EKF is run along the true solution with $\boldsymbol{R}_f \to \mathbf{0}$, the asymptotic rank of the covariance matrix is $r_\infty \approx 0.39D$, which almost exactly coincides with $D_u$.

This benefit all but disappears however when observation noise is added. Under these conditions, the critical threshold increases to $L_c \approx 0.37D$, which is roughly commensurate with $D_u$. This increase is in part due to poor tuning of the model and measurement error covariance matrices $\boldsymbol{R}_m^{-1}$, $\boldsymbol{R}_f^{-1}$. To remedy this, an annealing procedure was introduced that adaptively reduces the magnitude of $|\boldsymbol{R}_f^{-1}|$ as the estimate converges. This technique improves the observational efficiency to $L_c \approx 0.25D$.

### 5.2.2 Fixed-interval smoothing

A similar investigation was performed on three fixed-interval smoothing variants of 4DVar. One is the standard 'strong constraint' description SC-4DVar, in which the forecast model is expected to hold exactly. The others are dual versions of 'weak constraint' 4DVar, which include model errors. The WC-4DVar-X variant uses a collocated approach that performs the optimization directly in 'path space' by treating each state at each measurement time as an independent variable. By contrast, WC-4DVar-U uses

an alternative description, which involves the initial state $\boldsymbol{x}_0$ and a discretized set of additive controls $\mathcal{U}$. The minimization was performed using a basic Gauss-Newton approach with no line-search.

A approximate measure of the critical radius $\rho_c$ of the basin of convergence was introduced as a means to compare these methods. This value represents the initial distance to the global minimum above which convergence is no longer guaranteed. Both SC-4DVar and WC-4DVar-U show an exponential collapse in $\rho_c$ as the length $T$ of the estimation window is increased. The collapse proceeds at a rate roughly equal to the largest global Lyapunov exponent, suggesting chaos in the forecast model as a primary cause.

This conclusion was supported by an analysis based on the radius of the uniqueness ball for the Gauss-Newton method. In particular, the relative scaling of the basin was shown to be dominated by the condition number of the Hessian of the objective function. For both SC-4DVar and WC-4DVar-U when $\mathcal{U} \approx \mathbf{0}$, the magnitude of the Hessian depends directly on the variational matrix, whose magnitude is known to grow at a rate commensurate with $\lambda_{\max}$.

This result clearly illustrates the well-known limitation of single shooting methods like SC-4DVar — that the length of the estimation window is inherently limited by the timescale of the chaos in the forecast model. However, it also makes a more subtle point that, in the context of dynamical inverse problems, the success of the estimate depends crucially on how the model constraints are implemented. Chaotic sensitivity to initial conditions injects dynamical instability into the estimation process. This instability must be controlled by the algorithm, to prevent extreme ill-conditioning that can render trivial problems computationally intractable.

This can be accomplished by either limiting the length of the estimation window, or using a more stable but also more computationally intensive algorithm, such as WC-4DVar-X whose basin of convergence appears to be asymptotically stable with increasing $T$. This stability is ostensibly one of the main reasons for using a weak constraint method, especially in a situation like this where model error is not explicitly added. But these results make it clear that this benefit depends on the chosen formulation. Indeed, without an adequate initial guess for the 'control' variables $\mathcal{U}$, WC-4DVar-U performs markedly worse than both WC-4DVar-X and SC-4DVar.

This defect is particularly evident in the observability calculations. With local initial conditions, all three methods give similar results. But with global initial conditions, WC-4DVar-U is not able to find the global minimum for *any* value of $L$ — even without observation noise.

Global initial conditions also increases the thresholds of the other two smoothers, which require

more observations than any of the filters. However, the smoothers are working with a much smaller estimation window. And their success is measured by the average RMSE across the entire window, which is a considerably more difficult task than just identifying the final state.

### 5.2.3 Limitations and future work

The methodology described here is meant to serve as a framework to guide future investigation into the complex issues regarding observability and conditioning of dynamical inverse problems, particularly those related to optimal estimation. As with any preliminary effort however, a number of limitations must be overcome for this approach to be applied to more realistic systems.

For instance, one of the benefits of the the Lorenz 96 model is that it is nearly isotropic, in the sense that the dynamical behavior of each variable is roughly identical. This property is particularly convenient, as it allows the sampled perturbations to be chosen randomly from a unit sphere. This approach is not expected to work as well for more complex models where the variables are scaled differently. Such applications may require many more samples for the statistics to converge.

Another remaining challenge will to design more realistic observation schemes based on limited number of parameters. Again, the Lorenz 96 model is particularly convenient, in the sense that one parameter is needed to capture its spatial resolution — at least given the 'uniform' observation scheme chosen here. Among other things, this allows for the use of binary search, which is not possible with more than one parameter. If more than one parameter must be optimized, a more complex procedure is needed (*e.g.*, branch and bound).

A number of other observation schemes could also be investigated with Lorenz 96, such as random, localized, or dynamic measurements that simulate the orbits of satellites over a rotating earth. These options may be directly compared using the techniques described here, and will be the subject of a future study.

Model error was also not considered here. As such, an important open question is how best to extend this analysis to systematically include these errors. From a deterministic perspective, they could be simulated by injecting prescribed disturbances into the model. There is an extension to the Lorenz 96 model that does just this, which may serve as a good starting point. One could also examine other models of varying dimensionality, to investigate how finite cutoffs in model resolution affect model error statistics. Similar comparisons have been used in computational fluid dynamics, to inform the development of sub-grid scale models of turbulent flow. The use of data assimilation methods to validate such models

seems promising, and may also help improve estimation techniques for multi-scale coupled earth systems.

More generally, within the variational approximation to the statistical path integral, there are two fundamental questions:

1. How many solutions (*i.e.*, isolated local minima of the estimation action) are needed to achieve desired levels of uncertainty?

2. What is the likelihood of finding those solutions?

The present methodology attempts to address the latter, while the former is the subject of an entire field of uncertainty quantification. Connecting the two has obvious benefits, but determining how best to proceed appears somewhat difficult.

For example, the path integral formulation dictates that the statistical weight of any given solution is exponentially proportional to its action level. A multi-modal conditional distribution $p(\mathcal{X}|\mathcal{Y})$ will therefore exhibit localized solutions whose action levels are in some sense 'close to' the global minimum. Formalizing this notion of proximity runs into problems however. This is largely due to the fact that the distribution $p(\mathcal{X}|\mathcal{Y})$ is not normalized. As such, the geometry of the action manifold (*i.e.*, the location of the local minima) is not altered by additive and multiplicative scale factors. This scale invariance of the action is often exploited by optimization techniques, which typically try to rescale the objective function to improve the conditioning of the problem. However, a multiplicative scale factor changes the separation between local minima, and thus also the relative statistical weight of the solutions. Furthermore, since the error covariance matrices $\boldsymbol{R}_m^{-1},\ \boldsymbol{R}_f^{-1}$ are rarely known *a-priori*, it seems unrealistic to rely on them to have precise values.

Despite these concerns, it would not be difficult to include model error in the numerical studies presented here. Among other things, this would again provide an empirical characterization of both the adequacy of the observation scheme and the efficacy of the data assimilation algorithms. How useful these results are in practice relies heavily on how closely the parameterized model error reflects the truth, which is unknown. But this does not mean the results are not useful, because as mentioned, if the combined observation-analysis-forecast system is incapable of generating accurate estimates under simulated conditions, there is little hope that they will be successful in operational ones.

Regarding the second issue, the analytical results presented here are largely incomplete. Recall for instance that the additional terms in the uniqueness ball were not able to be accurately computed. These terms are likely important, as they set the overall scale for the solution basin and include the impact of errors in the model and observations. The problem of bounding the basin of guaranteed convergence around

a local minimum should also be treated from the empirical point of view, where prior knowledge of the solution is not assumed. This will likely involve revisiting some of these ideas on dynamical optimization from the point of view of Kantorovich's theorem on Newton's method ([91]), which gives conditions for the existence of a root from an arbitrary initial guess, so prior knowledge of the solution is not required.

These results on the convergence/uniqueness balls are also only applicable to static, unconstrained implementations of the Gauss-Newton method. They therefore need to be extended to both algorithms with constraints, such as augmented Lagrangian or barrier methods, as well as dynamic optimization methods such as two pass smoothers. Indeed, given the rather remarkable observable efficiency of Kalman based methods, it would be beneficial to have a definition of conditioning that is updated dynamically with the filter, which could perhaps be accomplished by using the action as a Lyapunov function.

While Kalman methods are known to be formally equivalent to the Gauss-Newton method ([14]), they are apparently capable of using the available observations more efficiently than any other technique examined here. The evidence for this stems from the fact that, under certain conditions, the observability thresholds for the EKF are considerably below $D_u$. This suggests that the estimated error covariance acts as an 'optimal' dynamic preconditioner: a claim supported by its recently discovered targeting of the locally unstable modes of the forecast model. The work of [95] also appears to support this conclusion, albeit in a control context. But this point of view does not appear to be common in the literature.

Although the precise mechanism for this observational efficiency remains unclear, consider the remarks at the end of Chap. (2) that describe the role of symplectic structure in the placement of the poles of the optimal closed-loop control feedback matrix. These remarks suggest that symplectic structure may also be responsible for its observational efficiency. Proof of this conjecture however requires a more comprehensive understanding of how symplectic structure impacts the conditioning of the resulting gain matrix. For this, it would be useful to develop a dynamical version of the Hamiltonian arguments made in Sec. (2.6).

The existence of critical observability thresholds, below which the probability of reliably generating an accurate estimate becomes effectively zero, suggests a basic limit in our ability to reconstruct the state of the system from partial observations. The fundamental nature of these limits is supported by the fact that they persist even under otherwise ideal circumstances (*i.e.*, no model or observation errors), scale linearly with the resolution of the present model, and are well-approximated by the attractor averaged dimension of its unstable subspace. While in certain cases the effective limit can admittedly be reduced below this baseline threshold, it comes at the expense of robustness, stability, and additional computational

overhead.

This perspective on observability is a considerable departure from the well-established control theoretic measures, which for nonlinear problems tend to be stated in the form of local necessary conditions ([103]). If these conditions are not satisfied, state reconstruction is guaranteed to fail — at least locally. Yet there are inherent limitations to this. For one, it does not take into account the dynamics of the estimate, which may be observable on certain parts of the attractor, an unobservable on others. More to the point however, there appears to exist a large number of problems (such as the ones considered here) that satisfy these conditions, but are nonetheless too ill-conditioned to admit accurate and reliable estimates.

Thus, the rapidly increasing size and complexity of our mathematical models has created a demand for more precise empirical and analytical techniques to validate the adequacy of the observation system. The methodology introduced here is intended to serve as a preliminary step towards realizing these goals. While it may not for all situations, it has already identified some important limitations to the current state of the art, as well as some ways in which these may be overcome.

## 5.3 Moving horizon estimation for poorly observable systems

One way to improve the observational efficiency of a given estimation algorithm involves temporally extending the observation space, to include the measurements not just at the current time, but neighboring times as well. This idea was introduced in Chap. (4), from the point of view of discrete time embedding theorems found in optimal control and nonlinear dynamics ([187, 5, 172]). These theorems were historically developed in conjunction with methods for reconstructing the topological structure of chaotic attractors from discrete scalar time-series observations.

In estimation however, these theorems are used quite differently. The differences are largely due to the assumption that, in estimation, one has knowledge of an approximate dynamical forecast model describing the time evolution of observed time series. In attractor reconstruction, the embedding procedure provides a way to distinguish between otherwise ambiguous orbits of the partially observed dynamical system. Given a low dimensional system with enough data, it can be used to build an empirical model or map to predict roughly where the next observation will occur, and may therefore be viewed as a way of inferring the system dynamics from the data. In estimation, by contrast, a known forecast model is used to enrich the set of existing observations, by exploiting latent information in their time delays. That is, it uses measurements not just time of the current estimate, but in a sliding window around the current time.

The technique itself is not new. The use of time extended observations is evidently well-known to

improve the accuracy and stability of a number of filtering algorithms, including the Kalman filter (which becomes the fixed-lag Kalman smoother), and 3DVar (which becomes 4DVar). It is also the motivating idea behind moving horizon estimation, which is the estimation dual of model predictive control. The fact that these relatively well-established ideas were not known to us prior to the development of this technique underscores a central theme of this thesis: namely, the need for better communication of shared ideas among the various fields interested in the optimal estimation problem.

The novelty of our contributions are therefore based upon the following aspects. From an analytical point of view, it unifies these existing algorithms with the discrete time embedding theorems, and the notion of feedback-based control synchronization in the proxy space of time delays. However despite their rather profound theoretical importance, these theorems are of limited use in practice since the maps they generate are often too ill-conditioned to be inverted when observation noise present — which is always.

Nevertheless, applying the methodology developed in Chap. (3) demonstrates quite clearly how the use of time extended measurements reduces the effective observational requirements on the system. Most notably, under near optimal conditions, the time delay extended Kalman filter (TD-EKF) requires observing only $0.03\,D$ states of the Lorenz 96 system, up to $D = 100$. Under similar conditions, time delay synchronization (TDVar-FB) requires only slightly more observations ($0.06\,D$). And when observation noise is present, the two methods are indistinguishable. So TDVar-FB produces threshold levels commensurate with the TD-EKF, without the additional $O(D \times D)$ overhead of the estimated error covariance. This exceptional combination or observational and computational efficiency has likely contributed to the popularity of fixed-lag SC-4DVar methods in applications such as operational weather forecasting.

The inclusion of a time delay between observations appears to be beneficial, particularly when observation noise is present. Lower thresholds are also obtained by reusing observations between subsequent analyses. Disjoint estimation windows, by contrast, did not show any improvement in observational efficiency over the non-time-embedded case. Similar observations regarding the re-use of observations were made by [25], although critical observability thresholds were not considered. Despite some inherent tradeoffs, such as increased computational overhead and somewhat reduced noise filtering capability, the reuse of data appears to improve the odds of success when the observations are spatiotemporally sparse.

This chapter connects the use of time extended observations in optimal estimation with discrete time embedding theorems and attractor reconstruction techniques from nonlinear dynamics. The results

provide evidence to support the claim that the use of time delayed observations is capable of reducing the effective observability requirements of the system. Furthermore, the broad applicability of this approach suggests that it may help guide the development of a new generation of estimation algorithms, which targets problems with sparse observations. The next steps along this line of research will be to incorporate these ideas into more modern algorithms, and determine their respective observability thresholds. For the Ensemble Kalman Filter, this appears to already have been accomplished by [24, 25]. For particle filters and related Gaussian mixture models, this has not yet been done, although the implicit framework laid out by [37, 36] appears to be amenable for this task. To summarize the conclusions of [165], the combination of these methods may permit using the power of particle methods for representing general multi-modal densities, and the power of time delays for accurately tracking the locations of the modes.

## 5.4 A personal view on the future of estimation

The estimation problem is a fundamental challenge that touches a wide range of mathematical and scientific disciplines. Yet its broad applicability poses some inherent problems in its dedicated study, as it is increasingly difficult to get a global perspective on the problem that encompasses the full spectrum of existing results. The number of new algorithms has exploded over the past few decades, but many of these ideas are proposed without adequate comparison or connection to the broader base of literature. This has made it difficult to determine *a-priori* which techniques are best for a given problem. And perhaps more concerning, the stovepiped nature of these results has produced situation where the same idea gets reinvented several times, within different contexts. As a newcomer to this field, I can attest to this from personal experience.

This thesis begins the process of bringing together some of these shared ideas. In particular, Chap. (2) revisits the classical theory of optimal estimation from a deterministic perspective, forging new ties with Hamiltonian mechanics. Indeed, the parallels between estimation and physics are quite extraordinary ([43]), with considerable overlap between many subfields, including: classical mechanics, electromagnetism, as well as statistical and quantum mechanics. Estimation has always been an important part of physics, particularly for model validation. While viewing the estimation problem as an extension of the physics has potential to unlock a number of new research directions, as as noted by Mitter ([129]): "it is best to proceed by analogy with care". Concepts such as unitarity and causality, which play such an important role in physics, may be irrelevant to estimation, as estimated state is not required to obey physical laws. But the evidence presented here suggests that there are certain benefits to satisfying these

constraints, at least approximately. And fully realizing these advantages requires a more comprehensive understanding of their similarities and differences.

Furthermore, it is also apparent that the practical limits to observation, estimation, and prediction are not well established. Moreover, the current theory, which is based largely on ideas from optimal control, does not adequately resolve these limits. This issue was explored in Chap. (3) using a computational framework that identifies these limits as a phase transition to a synchronized state. These thresholds can be reduced to an extent — *e.g.*, using time extended observations methods described in Chap. (4). But even under near perfect conditions, there still appears to be a basic limit to our ability to observe, estimate, and predict complex dynamical behavior. Understanding where this threshold lies with respect to our current technological capability, and determining how best to address any shortcomings identified from this analysis, will undoubtedly be of the utmost importance in the years to come.

# Bibliography


[1] Henry Abarbanel. *Predicting the future: completing models of observed complex systems*. Springer, 2013.

[2] Henry DI Abarbanel. *Analysis of Observed Chaotic Data*. Springer New York, 1996.

[3] Henry DI Abarbanel. Effective actions for statistical data assimilation. *Physics Letters A*, 373(44):4044–4048, 2009.

[4] Henry DI Abarbanel, Daniel R Creveling, Reza Farsian, and Mark Kostuk. Dynamical state and parameter estimation. *SIAM Journal on Applied Dynamical Systems*, 8(4):1341–1381, 2009.

[5] Dirk Aeyels. Generic observability of differentiable systems. *SIAM Journal on Control and Optimization*, 19(5):595–603, 1981.

[6] AC Aitken. On fitting polynomials to weighted data by least squares. *Proceedings of the Royal Society of Edinburgh*, 54:1–11, 1935.

[7] Eugene L Allgower and Kurt Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2012.

[8] Frank Allgöwer, Thomas A Badgwell, Joe S Qin, James B Rawlings, and Steven J Wright. Nonlinear predictive control and moving horizon estimationan introductory overview. In *Advances in control*, pages 391–449. Springer London, 1999.

[9] Ienkaran Arasaratnam and Simon Haykin. Cubature kalman smoothers. *Automatica*, 47(10):2245–2250, 2011.

[10] Aleksandr Y Aravkin, Bradley B Bell, James V Burke, and Gianluigi Pillonetto. Kalman smoothing and block tridiagonal systems: new connections and numerical stability results. *arXiv preprint arXiv:1303.5237*, 2013.

[11] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. Math*, 3(1):133–181, 1922.

[12] Krzysztof Barański, Núria Fagella, Xavier Jarque, and Bogusława Karpińska. Connectivity of julia sets of newton maps: A unified approach. *arXiv preprint arXiv:1501.05488*, 2015.

[13] Bradley M Bell. The iterated kalman smoother as a gauss–newton method. *SIAM Journal on Optimization*, 4(3):626–636, 1994.

[14] Bradley M Bell and Frederick W Cathey. The iterated kalman filter update as a gauss-newton method. *IEEE Transactions on Automatic Control*, 38(2):294–297, 1993.

[15] Richard Bellman. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences*, 42(10):767–769, 1956.

[16] Richard Bellman. Invariant imbedding and computational methods in radiative transfer. In *SIAM-AMS Proceedings*, volume 1, page 95. American Mathematical Society, 1969.

[17] VE Beneš. Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics: An International Journal of Probability and Stochastic Processes*, 5(1-2):65–92, 1981.

[18] Jon Louis Bentley and Andrew Chi-Chih Yao. An almost optimal algorithm for unbounded searching. *Information processing letters*, 5(3):82–87, 1976.

[19] Tyrus Berry and Timothy Sauer. Adaptive ensemble kalman filtering of non-linear systems. *Tellus A*, 65, 2013.

[20] J.T. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming: Second Edition*. Advances in Design and Control. SIAM, 2010.

[21] Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale. The condition number for nonlinear problems. In *Complexity and Real Computation*, pages 217–236. Springer, 1998.

[22] Hans Georg Bock and Karl-Josef Plitt. A multiple shooting algorithm for direct solution of optimal control problems. *Proceedings of the IFAC World Congress*, 1984.

[23] Marc Bocquet, Karthik S Gurumoorthy, Amit Apte, Alberto Carrassi, Colin Grudzien, and Christopher KRT Jones. Degenerate kalman filter error covariances and their convergence onto the unstable subspace. *arXiv preprint arXiv:1604.02578*, 2016.

[24] Marc Bocquet and Pavel Sakov. Joint state and parameter estimation with an iterative ensemble kalman smoother. *Nonlinear Processes in Geophysics*, 20(5):803–818, 2013.

[25] Marc Bocquet and Pavel Sakov. An iterative ensemble kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 140(682):1521–1535, 2014.

[26] Hendrik Wade Bode and Claude Elwood Shannon. A simplified derivation of linear least square smoothing and prediction theory. *Proceedings of the IRE*, 38(4):417–425, 1950.

[27] Jochen Bröcker. On variational data assimilation in continuous time. *Quarterly Journal of the Royal Meteorological Society*, 136(652):1906–1919, 2010.

[28] Roger W Brockett. Nonlinear systems and nonlinear estimation theory. In *Stochastic systems: The mathematics of filtering and identification and applications*, pages 441–477. Springer Netherlands, 1981.

[29] AE Bryson and M Frazier. Smoothing for linear and nonlinear dynamic systems. In *Proceedings of the optimum system synthesis conference*, pages 353–364, 1963.

[30] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. CRC Press, 1975.

[31] Robert H Cameron and William T Martin. Transformations of weiner integrals under translations. *Annals of Mathematics*, pages 386–396, 1944.

[32] Carla Cardinali. Data assimilation: Observation impact on the short range forecast, ecmwf lecture notes. http://www.ecmwf.int/publications/, 2013. Accessed: January 2017.

[33] KP Bharani Chandra, Da-Wei Gu, and Ian Postlethwaite. Square root cubature information filter. *IEEE Sensors Journal*, 13(2):750–758, 2013.

[34] Jinhai Chen and Weiguo Li. Convergence of gauss–newtons method and uniqueness of the solution. *Applied mathematics and computation*, 170(1):686–705, 2005.

[35] Jinhai Chen and Weiguo Li. Convergence behaviour of inexact newton methods under weak lipschitz condition. *Journal of Computational and Applied Mathematics*, 191(1):143–164, 2006.

[36] Alexandre Chorin, Matthias Morzfeld, and Xuemin Tu. Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2):221–240, 2010.

[37] Alexandre J Chorin and Xuemin Tu. Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, 106(41):17249–17254, 2009.

[38] Moody T Chu. Numerical methods for inverse singular value problems. *SIAM journal on numerical analysis*, 29(3):885–903, 1992.

[39] Moody T Chu and Gene H Golub. Structured inverse eigenvalue problems. *Acta Numerica*, 11:1, 2002.

[40] PHILIPPE Courtier, J-N Thépaut, and A Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

[41] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[42] John L Crassidis and John L Junkins. *Optimal estimation of dynamic systems*. CRC press, 2011.

[43] Fred Daum and Jim Huang. Exact particle flow for nonlinear filters: seventeen dubious solutions to a first order linear underdetermined pde. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 64–71. IEEE, 2010.

[44] Frederick Daum. Exact finite-dimensional nonlinear filters. *IEEE Transactions on Automatic Control*, 31(7):616–622, 1986.

[45] Persi Diaconis, Susan Holmes, and Richard Montgomery. Dynamical bias in the coin toss. *SIAM review*, 49(2):211–235, 2007.

[46] PAM Dirac. Lectures on quantum mechanics belfer graduate school of science monographs, no. 2. *Yeshiva University Press, New York*, 369:375–388, 1964.

[47] Ute Dressler and J Doyne Farmer. Generalized lyapunov exponents corresponding to higher derivatives. *Physica D: Nonlinear Phenomena*, 59(4):365–377, 1992.

[48] Adam El-Said. *Conditioning of the weak-constraint variational data assimilation problem for numerical weather prediction*. PhD thesis, University of Reading, 2015.

[49] Alexandre A Emerick and Albert C Reynolds. History matching time-lapse seismic data using the ensemble kalman filter with multiple data assimilations. *Computational Geosciences*, 16(3):639–659, 2012.

[50] Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

[51] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, 2009.

[52] J Doyne Farmer and John J Sidorowich. Predicting chaotic time series. *Physical review letters*, 59(8):845, 1987.

[53] J Doyne Farmer and John J Sidorowich. *Exploiting chaos to predict the future and reduce noise*, volume 279. World Scientific, Singapore, 1988.

[54] J Doyne Farmer and John J Sidorowich. Optimal shadowing and noise reduction. *Physica D: Nonlinear Phenomena*, 47(3):373–392, 1991.

[55] Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

[56] D Fraser and J Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4):387–390, 1969.

[57] Donald Charles Fraser. *New technique for optimal smoothing of data*. PhD thesis, MIT, 1968.

[58] Joel Friedman. On the convergence of newton's method. *Journal of Complexity*, 5(1):12–33, 1989.

[59] Sara P Garcia and Jonas S Almeida. Nearest neighbor embedding with different time delays. *Physical Review E*, 71(3):037204, 2005.

[60] NE Getz and Jerrold E Marsden. A dynamic inverse for nonlinear maps. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 4, pages 4218–4223. IEEE, 1995.

[61] Philip E Gill, Walter Murray, and Michael A Saunders. Snopt: An sqp algorithm for large-scale constrained optimization. *SIAM review*, 47(1):99–131, 2005.

[62] Igor Vladimirovich Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

[63] Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.

[64] Serge Gratton, Amos S Lawless, and Nancy K Nichols. Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.

[65] JW Grizzle and PE Moraal. Newton, observers and nonlinear discrete-time control. In *Decision and Control, 1990., Proceedings of the 29th IEEE Conference on*, pages 760–767. IEEE, 1990.

[66] Karthik S Gurumoorthy, Colin Grudzien, Amit Apte, Alberto Carrassi, and Christopher KRT Jones. Rank deficiency of kalman error covariance matrices in linear time-varying system with deterministic evolution. *SIAM Journal on Control and Optimization*, 55(2):741–759, 2017.

[67] Stephen A Haben. *Conditioning and preconditioning of the minimisation problem in variational data assimilation*. PhD thesis, University of Reading, 2011.

[68] MI Hadamard. On problems in partial derivatives, and their physical significance. *Princeton University Bulletin*, 13:82–88, 1902.

[69] William W Hager. Condition estimates. *SIAM Journal on Scientific and Statistical Computing*, 5(2):311–316, 1984.

[70] Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, 34(4):561–580, 1992.

[71] Eric L Haseltine and James B Rawlings. Critical evaluation of extended kalman filtering and moving-horizon estimation. *Industrial & engineering chemistry research*, 44(8):2451–2460, 2005.

[72] WM Häußler. A kantorovich-type convergence analysis for the gauss-newton-method. *Numerische Mathematik*, 48(1):119–125, 1986.

[73] Kevin Hayden, Eric Olson, and Edriss S Titi. Discrete data assimilation in the lorenz and 2d navier–stokes equations. *Physica D: Nonlinear Phenomena*, 240(18):1416–1425, 2011.

[74] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[75] RG Holt and IB Schwartz. Newton's method as a dynamical system: global convergence and predictability. *Physics Letters A*, 105(7):327–333, 1984.

[76] HJC Huijberts, T Lilge, and Hendrik Nijmeijer. Nonlinear discrete-time synchronization via extended observers. *International Journal of Bifurcation and Chaos*, 11(07):1997–2006, 2001.

[77] Jeffrey Humpherys, Preston Redd, and Jeremy West. A fresh look at the kalman filter. *SIAM review*, 54(4):801–823, 2012.

[78] Eugene Isaacson and Herbert Bishop Keller. *Analysis of numerical methods*. Courier Corporation, 1994.

[79] David H Jacobson. New second-order and first-order algorithms for determining optimal control: A differential dynamic programming approach. *Journal of Optimization Theory and Applications*, 2(6):411–440, 1968.

[80] David H Jacobson and David Q Mayne. *Differential dynamic programming*. North-Holland, 1970.

[81] Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[82] Christine Johnson, Brian J Hoskins, and Nancy K Nichols. A singular vector perspective of 4d-var: Filtering and interpolation. *Quarterly Journal of the Royal Meteorological Society*, 131(605):1–19, 2005.

[83] John Bagterp Jørgensen. *Moving horizon estimation and control*. PhD thesis, Technical University of Denmark, Department of Chemical Engineering, 2005.

[84] Simon J Julier and Jeffrey K Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[85] N Kadakia, D Rey, J Ye, and HDI Abarbanel. Symplectic structure of statistical variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):756–771, 2017.

[86] Thomas Kailath. An innovations approach to least-squares estimation–part i: Linear filtering in additive white noise. *IEEE transactions on automatic control*, 13(6):646–655, 1968.

[87] Thomas Kailath. A view of three decades of linear filtering theory. *IEEE Transactions on Information Theory*, 20(2):146–181, 1974.

[88] Thomas Kailath and Paul Frost. An innovations approach to least-squares estimation–part ii: Linear smoothing in additive white noise. *IEEE Transactions on Automatic Control*, 13(6):655–660, 1968.

[89] Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(3):95–108, 1961.

[90] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[91] Leonid Vital'evich Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

[92] James Kaplan and James Yorke. Chaotic behavior of multidimensional difference equations. *Functional Differential equations and approximation of fixed points*, pages 204–227, 1979.

[93] Hilbert J Kappen, Joaquın Marro, Pedro L Garrido, and Joaquın J Torres. An introduction to stochastic control theory, path integrals and reinforcement learning. In *AIP conference proceedings*, volume 887, pages 149–181. AIP, 2007.

[94] A Karimi and Mark R Paul. Extensive chaos in the lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):043105, 2010.

[95] E Kaszkurewicz, A Bhaya, and PRV Ramos. A control-theoretic view of diagonal preconditioners. *International journal of systems science*, 26(9):1659–1672, 1995.

[96] Jaroslav Kautsky, Nancy K Nichols, and Paul Van Dooren. Robust pole assignment in linear state feedback. *International Journal of control*, 41(5):1129–1155, 1985.

[97] HB Keller. Approximation methods for nonlinear problems with application to two-point boundary value problems. *Mathematics of Computation*, 29(130):464–474, 1975.

[98] Joseph B Keller. The probability of heads. *The American Mathematical Monthly*, 93(3):191–197, 1986.

[99] Carl N Kelly and Brian DO Anderson. On the stability of fixed-lag smoothing algorithms. *Journal of the Franklin Institute*, 291(4):271–281, 1971.

[100] Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.

[101] Mark Kostuk. *Synchronization and statistical methods for the data assimilation of HVc neuron models*. PhD thesis, University of California San Diego, Department of Physics, 2012.

[102] Jayesh H Kotecha and Petar M Djuric. Gaussian sum particle filtering. *IEEE Transactions on signal processing*, 51(10):2602–2612, 2003.

[103] Arthur J Krener and Witold Respondek. Nonlinear observers with linearizable error dynamics. *SIAM Journal on Control and Optimization*, 23(2):197–216, 1985.

[104] Harold J Kushner. On the differential equations satisfied by conditional probablitity densities of markov processes, with applications. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2(1):106–119, 1964.

[105] KJH Law, D Sanz-Alonso, A Shukla, and AM Stuart. Controlling unpredictability with observations in the partially observed lorenz'96 model. *arXiv preprint arXiv:1411.3113*, 2014.

[106] KJH Law, D Sanz-Alonso, A Shukla, and AM Stuart. Filter accuracy for chaotic dynamical systems: fixed versus adaptive observation operators. *arXiv preprint arXiv:1411.3113*, 2014.

[107] Amos S Lawless and Nancy K Nichols. Inner-loop stopping criteria for incremental four-dimensional variational data assimilation. *Monthly Weather Review*, 134(11):3425–3435, 2006.

[108] AS Lawless, S Gratton, and NK Nichols. An investigation of incremental 4d-var using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society*, 131(606):459–476, 2005.

[109] Jane B Lawrie and I David Abrahams. A brief historical perspective of the wiener–hopf technique. *Journal of Engineering Mathematics*, 59(4):351–358, 2007.

[110] Melvin Leok and Jingjing Zhang. Discrete hamiltonian variational integrators. *IMA Journal of Numerical Analysis*, page drq027, 2011.

[111] Christophe Letellier, Luis A Aguirre, and Jean Maquet. Relation between observability and differential embeddings for nonlinear dynamics. *Physical Review E*, 71(6):066213, 2005.

[112] Chong Li, Nuchun Hu, and Jinhua Wang. Convergence behavior of gauss–newtons method and extensions of the smale point estimate theory. *Journal of Complexity*, 26(3):268–295, 2010.

[113] Chong Li, Wen-Hong Zhang, and Xiao-Qing Jin. Convergence and uniqueness properties of gauss-newton's method. *Computers & Mathematics with Applications*, 47(6-7):1057–1067, 2004.

[114] Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO (1)*, pages 222–229, 2004.

[115] Zhijin Li and IM Navon. Optimality of variational data assimilation and its relationship with the kalman filter and smoother. *Quarterly Journal of the Royal Meteorological Society*, 127(572):661–683, 2001.

[116] JC Lopez-Marcos. A difference scheme for a nonlinear partial integrodifferential equation. *SIAM journal on numerical analysis*, 27(1):20–31, 1990.

[117] Andrew C Lorenc, Neill E Bowler, Adam M Clayton, Stephen R Pring, and David Fairbairn. Comparison of hybrid-4denvar and hybrid-4dvar data assimilation methods for global nwp. *Monthly Weather Review*, 143(1):212–229, 2015.

[118] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

[119] Jerrold E Marsden and Matthew West. Discrete mechanics and variational integrators. *Acta Numerica 2001*, 10:357–514, 2001.

[120] David Mayne. A second-order gradient method for determining optimal trajectories of non-linear discrete-time systems. *International Journal of Control*, 3(1):85–95, 1966.

[121] DQ Mayne. A solution of the smoothing problem for linear dynamic systems. *Automatica*, 4(2):73–92, 1966.

[122] SR McReynolds. *A Successive Sweep Method for Solving Optimal Programming Problems*. PhD thesis, Harvard University, 1965.

[123] Stephen Ralph McReynolds. Fixed interval smoothing-revisited. *Journal of Guidance, Control, and Dynamics*, 13(5):913–921, 1990.

[124] James S Meditch. On optimal linear smoothing theory. *Information and Control*, 10(6):598–615, 1967.

[125] Raman Mehra. On the identification of variances and adaptive kalman filtering. *IEEE Transactions on automatic control*, 15(2):175–184, 1970.

[126] Raman Mehra. Approaches to adaptive filtering. *IEEE Transactions on automatic control*, 17(5):693–698, 1972.

[127] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[128] JR Miller and JA Yorke. Finding all periodic orbits of maps using newton methods: sizes of basins. *Physica D: Nonlinear Phenomena*, 135(3):195–211, 2000.

[129] Sanjoy K Mitter. On the analogy between mathematical problems of non-linear filtering and quantum physics. Technical report, MIT, 1980.

[130] Sanjoy K Mitter. Existence and non-existence of finite dimensional filters. *Mathematical theory of optimization*, pages 1–10, 1981.

[131] Sanjoy K Mitter. Approximations for nonlinear filtering. In *Nonlinear Stochastic Problems*, pages 339–345. Springer Netherlands, 1983.

[132] Sanjoy K Mitter. Filtering and stochastic control: A historical perspective. *IEEE Control Systems*, 16(3):67–76, 1996.

[133] SK Mitter. Successive approximation methods for the solution of optimal control problems. *Automatica*, 3(3-4):135–149, 1966.

[134] John B Moore. Discrete-time fixed-lag smoothing algorithms. *Automatica*, 9(2):163–173, 1973.

[135] John B Moore and Peter KS Tam. Fixed-lag smoothing for nonlinear systems with discrete measurements. *Information Sciences*, 6:151–160, 1973.

[136] PE Moraal and JW Grizzle. Asymptotic observers for detectable and poorly observable systems. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 108–114. IEEE, 1995.

[137] PE Moraal and JW Grizzle. Observer design for nonlinear systems with discrete-time measurements. *IEEE Transactions on automatic control*, 40(3):395–404, 1995.

[138] Norman Morrison. *Tracking filter engineering*. The Institution of Engineering and Technology, 2012.

[139] IM Navon and David M Legler. Conjugate-gradient methods for large-scale minimization in meteorology. *Monthly Weather Review*, 115(8):1479–1502, 1987.

[140] Otto Neugebauer. *The exact sciences in antiquity*. Brown University Press, 1957.

[141] S Nicosia, A Tornambe, and P Valigi. Use of observers for the inversion of nonlinear maps. *Systems & control letters*, 16(6):447–455, 1991.

[142] Yves Nievergelt. A tutorial history of least squares with applications to astronomy and geodesy. *Journal of Computational and Applied Mathematics*, 121(1):37–72, 2000.

[143] Benjamin Noack, Daniel Lyons, Matthias Nagel, and Uwe D Hanebeck. Nonlinear information filtering for distributed multisensor data fusion. In *American Control Conference (ACC), 2011*, pages 4846–4852. IEEE, 2011.

[144] Jorge Nocedal and Stephen J Wright. Sequential quadratic programming. *Numerical optimization*, pages 529–562, 2006.

[145] Tomoki Ohsawa, Anthony M Bloch, and Melvin Leok. Discrete hamilton–jacobi theory. *SIAM Journal on Control and Optimization*, 49(4):1829–1856, 2011.

[146] Eric Olson and Edriss S Titi. Determining modes for continuous data assimilation in 2d turbulence. *Journal of statistical physics*, 113(5):799–840, 2003.

[147] SJ Orfanidis. A group-theoretical approach to optimal estimation and control. *Journal of mathematical analysis and applications*, 97(2):393–416, 1983.

[148] Valery Iustinovich Oseledec. A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc*, 19(2):197–231, 1968.

[149] DI Papadimitriou and KC Giannakoglou. Direct, adjoint and mixed approaches for the computation of hessian in airfoil design problems. *International journal for numerical methods in fluids*, 56(10):1929–1943, 2008.

[150] Chandeok Park. *The Hamilton-Jacobi theory for solving optimal feedback control problems with general boundary conditions*. PhD thesis, The University of Michigan, 2006.

[151] Chandeok Park and Daniel J Scheeres. Solutions of the optimal feedback control problem using hamiltonian dynamics and generating functions. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 2, pages 1222–1227. IEEE, 2003.

[152] Chandeok Park and Daniel J Scheeres. Formulation of a hamiltonian cauchy problem for solving optimal feedback control problems. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, pages 2793–2798. IEEE, 2005.

[153] PooGyeon Park and Thomas Kailath. New square-root algorithms for kalman filtering. *IEEE Transactions on Automatic Control*, 40(5):895–899, 1995.

[154] Ulrich Parlitz, Jan Schumann-Bischoff, and Stefan Luther. Local observability of state variables and parameters in nonlinear modeling quantified by delay reconstruction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(2):024411, 2014.

[155] Ulrich Parlitz, Jan Schumann-Bischoff, and Stefan Luther. Quantifying uncertainty in state and parameter estimation. *Physical Review E*, 89(5):050902, 2014.

[156] Diego Pazó, A Carrassi, and Juan M López. Data assimilation by delay-coordinate nudging. *Quarterly Journal of the Royal Meteorological Society*, 2016.

[157] Louis M Pecora and Thomas L Carroll. Synchronization in chaotic systems. *Physical review letters*, 64(8):821, 1990.

[158] Louis M Pecora, Linda Moniz, Jonathan Nichols, and Thomas L Carroll. A unified approach to attractor reconstruction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(1):013110, 2007.

[159] Carlos Pires, Robert Vautard, and Olivier Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48(1):96–121, 1996.

[160] Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge Univ Press, 1936.

[161] Florence Rabier, H Järvinen, E Klinker, J-F Mahfouf, and A Simmons. The ecmwf operational implementation of four-dimensional variational assimilation. i: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126(564):1143–1170, 2000.

[162] C Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.

[163] Christopher V Rao, James B Rawlings, and David Q Mayne. Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations. *IEEE transactions on automatic control*, 48(2):246–258, 2003.

[164] Herbert E Rauch, CT Striebel, and F Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[165] James B Rawlings and Bhavik R Bakshi. Particle filtering and moving horizon estimation. *Computers & chemical engineering*, 30(10):1529–1541, 2006.

[166] Daniel Rey, Michael Eldridge, Mark Kostuk, Henry DI Abarbanel, Jan Schumann-Bischoff, and Ulrich Parlitz. Accurate state and parameter estimation in nonlinear systems with sparse observations. *Physics Letters A*, 378(11):869–873, 2014.

[167] Daniel Rey, Michael Eldridge, Uriel Morone, Henry DI Abarbanel, Ulrich Parlitz, and Jan Schumann-Bischoff. Using waveform information in nonlinear data assimilation. *Physical Review E*, 90(6):062916, 2014.

[168] Werner C Rheinboldt. On measures of ill-conditioning for nonlinear equations. *Mathematics of Computation*, 30(133):104–111, 1976.

[169] JM Sanz-Serna. Two topics in nonlinear stability. *Advances in numerical analysis*, 1:147–174, 1991.

[170] JM Sanz-Serna. Symplectic runge–kutta schemes for adjoint equations, automatic differentiation, optimal control, and more. *SIAM Review*, 58(1):3–33, 2016.

[171] Yoshikazu Sasaki. Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12), 1970.

[172] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3):579–616, 1991.

[173] Jan Schumann-Bischoff, Stefan Luther, and Ulrich Parlitz. Estimability and dependency analysis of model parameters based on delay coordinates. *Physical Review E*, 94(3):032221, 2016.

[174] Jan Schumann-Bischoff, Ulrich Parlitz, Henry DI Abarbanel, Mark Kostuk, Daniel Rey, Michael Eldridge, and Stefan Luther. Basin structure of optimization based state and parameter estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(5):053108, 2015.

[175] Michael Shub and Steve Smale. Complexity of bézouts theorem iv: probability of success; extensions. *SIAM Journal on Numerical Analysis*, 33(1):128–148, 1996.

[176] ZK Silagadze. Gauge transformations are canonical transformations, redux. *arXiv preprint arXiv:1409.0692*, 2014.

[177] Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[178] Steve Smale. The fundamental theorem of algebra and complexity theory. *Bulletin of the American Mathematical Society*, 4(1):1–36, 1981.

[179] Steve Smale. *Newtons method estimates from data at one point*. Springer, 1986.

[180] Paul So, Edward Ott, and WP Dayawansa. Observing chaos: Deducing and tracking the state of a chaotic system from limited observation. *Physical Review E*, 49(4):2650, 1994.

[181] Yongkyu Song and Jessy W Grizzle. The extended kalman filter as a local asymptotic observer for nonlinear discrete-time systems. In *American Control Conference, 1992*, pages 3365–3369. IEEE, 1992.

[182] Eduardo D Sontag. A concept of local observability. *Systems & Control Letters*, 5(1):41–47, 1984.

[183] Bart D Stewart. Attractor basins of various root-finding methods. Technical report, DTIC Document, 2001.

[184] Toni Stojanovski, Ulrich Parlitz, Ljupčo Kocarev, and Richard Harris. Exploiting delay reconstruction for chaos synchronisation. *Physics Letters A*, 233(4):355–360, 1997.

[185] Nozomi Sugiura, Shuhei Masuda, Yosuke Fujii, Masafumi Kamachi, Yoichi Ishikawa, and Toshiyuki Awaji. A framework for interpreting regularized state estimation. *Monthly Weather Review*, 142(1):386–400, 2014.

[186] Hector J Sussmann and Jan C Willems. 300 years of optimal control: from the brachystochrone to the maximum principle. *IEEE Control Systems Magazine*, 17(3):32–44, 1997.

[187] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[188] Thomas J Taylor. On the existence of higher order lyapunov exponents. *Nonlinearity*, 6(3):369, 1993.

[189] Andrei N Tikhonov and Vasilii Arsenin. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

[190] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39(5):195–198, 1943.

[191] Emanuel Todorov and Yuval Tassa. Iterative local dynamic programming. In *Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL'09. IEEE Symposium on*, pages 90–95. IEEE, 2009.

[192] Joseph Frederick Traub and H Woźniakowski. Convergence and complexity of newton iteration for operator equations. *Journal of the ACM (JACM)*, 26(2):250–258, 1979.

[193] Yannick Tr'emolet. Accounting for an imperfect model in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2483–2504, 2006.

[194] A Trevisan and L Palatella. On the kalman filter error covariance collapse into the unstable subspace. *Nonlinear Processes in Geophysics*, 18(2):243–250, 2011.

[195] Anna Trevisan and Luigi Palatella. The extended kalman filter and its reduction to the unstable subspace. *Poster contribution presented at Ecodyc2010, Dresden (Germany). http://www. mpipks-dresden. mpg. de/ecodyc10/Contributions/Trevisan. pdf*, 2010.

[196] Anna Trevisan and Francesco Uboldi. Assimilation of standard and targeted observations within the unstable subspace of the observation–analysis–forecast cycle system. *Journal of the atmospheric sciences*, 61(1):103–113, 2004.

[197] Joel A Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

[198] J Tshimanga, S Gratton, AT Weaver, and A Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 134(632):751–769, 2008.

[199] Peter Jan Van Leeuwen, Yuan Cheng, and Sebastian Reich. *Nonlinear data assimilation*. Springer, 2015.

[200] D Vaughan. A nonrecursive algebraic solution for the discrete riccati equation. *IEEE Transactions on Automatic Control*, 15(5):597–599, 1970.

[201] I Vlachos and D Kugiumtzis. State space reconstruction from multiple time series. In *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference*, page 378, 2009.

[202] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

[203] Shiyuan Wang, Jiuchao Feng, and K Tse Chi. A class of stable square-root nonlinear information filters. *IEEE Transactions on Automatic Control*, 59(7):1893–1898, 2014.

[204] XH Wang. The convergence ball on newtons method. *Chinese Science Bulletin, A Special Issue of Mathematics, Physics, Chemistry*, 25:36–37, 1980.

[205] Xinghua Wang. Convergence of newtons method and inverse function theorem in banach space. *Mathematics of Computation of the American Mathematical Society*, 68(225):169–186, 1999.

[206] Xinghua Wang. Convergence of newton's method and uniqueness of the solution of equations in banach space. *IMA Journal of Numerical Analysis*, 20(1):123–134, 2000.

[207] MC Whatley. The role of kinematics in hamiltonian dynamics: Momentum as a lagrange multiplier. *American Journal of Physics*, 58(10):1006–1011, 1990.

[208] Wing Shing Wong. Theorems on the structure of finite dimensional estimation algebras. *Systems & control letters*, 9(2):117–124, 1987.

[209] Max A Woodbury. Inverting modified matrices. *Memorandum report*, 42:106, 1950.

[210] Chee-Keng Yap. *Fundamental problems of algorithmic algebra*, volume 49. Oxford University Press Oxford, 2000.

[211] Shing-Tung Yau and Stephen S-T Yau. Real time solution of nonlinear filtering problem without memory i. *Mathematical Research Letters*, 7(5/6):671–694, 2000.

[212] Stephen S-T Yau. Finite-dimensional filters with nonlinear drift. i: A class of filters including both kalman-bucy and benes filters. *J. Math. Systems, Estimation, and Control*, 4:181–203, 1994.

[213] J Ye, N Kadakia, PJ Rozdeba, HDI Abarbanel, and JC Quinn. Precision variational approximations in statistical data assimilation. *Nonlinear Processes in Geophysics Discussions*, 1:1603–1620, 2014.

[214] Jingxin Ye, Daniel Rey, Nirag Kadakia, Michael Eldridge, Uriel I Morone, Paul Rozdeba, Henry DI Abarbanel, and John C Quinn. Systematic variational method for statistical nonlinear state and parameter estimation. *Physical Review E*, 92(5):052901, 2015.

[215] Moshe Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(3):230–243, 1969.