Bayesian system identification
Accounting for model error for improved robustness to sparse and noisy data

Machine Learning in Science & Engineering (Virtual)
December 13-14, 2020

Alex Gorodetsky and Nicholas Galioto,
Aerospace Engineering, UM
Motivation 1: autonomous dynamical systems
Noisy data and fast dynamics

- Goal: control and plan in systems with unknown system dynamics

- Problem: system identification must happen from limited onboard sensor measurements in real-time
  - Data is noisy
  - Data is sparse
  - Dynamics/sensors might change
Motivation 2: indirectly observed PDE systems

Time-varying Reaction-diffusion PDE

\[
\frac{\partial C_1}{\partial t} = \theta_1 \frac{\partial^2 C_1}{\partial x^2} + 0.1 - C_1 + \theta_3 C_1^2 C_2 \\
\frac{\partial C_2}{\partial t} = \theta_2 \frac{\partial^2 C_2}{\partial x^2} C_2 + 0.9 - C_1^2 C_2
\]
Characterization of the problem

1. Time series $\mathcal{Y}_n$: noisy, irregularly timed, partially informative data
2. Learn model parameters $\theta$
3. Make probabilistic predictions about the future or for different conditions

$$p(x(t) | \mathcal{Y}_n) = \int p(x(t) | \theta)p(\theta | \mathcal{Y}_n)d\theta$$
1. Provide the probabilistic framework and relation to existing approaches
2. Describe the algorithm
3. Demonstrate significant robustness compared to state-of-the-art
4. Extend to Hamiltonian Systems
Probabilistic modeling framework
Joint parameter-state inference problem

Parameterized Hidden Markov Model
Generic formulation: linear, physical model, dictionary of bases, NN, etc.

\[ X_{k+1} = \Psi(X_k; \theta) + \eta_k \]
\[ Y_k = h(X_k) + \nu_k, \]
\[ \eta_k \sim \mathcal{N}(0, \Sigma) \]
\[ \nu_k \sim \mathcal{N}(0, \Gamma) \]

1. Accounts for model parameter uncertainty through \( \theta \)
2. Accounts for model form uncertainty through \( \eta_k \)
3. Accounts for measurement noise through \( \nu_k \)
Bayesian inference
Updating models with data

Prior \( p(\theta, X_n | I) \)

Bayes’ Rule
\[
p(\theta, X_n | Y_n, I) = \frac{p(\theta, X_n | I) p(Y_n | \theta, X_n, I)}{p(Y_n | I)}
\]

Posterior \( p(\theta, X_n | Y_n, I) \)

For our Markovian system
\[
p(\theta, X_n | Y_n, I) = \frac{1}{Z} \left[ \prod_{i=1}^{n} p(y_i | X_i, \theta) \right] \left[ \prod_{i=1}^{n} p(X_i | X_{i-1}, \theta) \right] p(\theta | I)
\]

\( Z \)
Log posterior is a generalized objective function
Many existing approaches recovered under simplifying assumptions

\[
\log L(\theta, \mathcal{X}; \mathcal{Y}) \propto -\frac{1}{2} \sum_{k=1}^{n} \| y_k - h(x_k, \theta_h) \|^2 + \frac{1}{2} \sum_{k=1}^{n} \| x_k - \Psi(x_{k-1}, \theta_\Psi) \|^2
\]

Identity observations: \( y_k = x_k \)

Deterministic dynamics: \( x_k = \Psi(x_{k-1}) \)

Least Squares for Propagators:
\[
\log L(\theta; \mathcal{Y}) \propto -\frac{1}{2} \sum_{k=2}^{n} \| y_k - A(\theta_\Psi) y_{k-1} \|^2
\]

\[
\Psi(y_{k-1}, \theta_\Psi) = A(\theta_\Psi) y_{k-1}
\]

Least Squares for Vector Fields:
\[
\log L(\theta; \mathcal{Y}) \propto -\frac{1}{2} \sum_{k=2}^{n} \| \frac{y_k - y_{k-1}}{\Delta t} - \Xi(y_{k-1}) \theta_\Psi \|^2
\]

\[
\Xi(y_{k-1}) \theta_\Psi = x_k
\]

\[
\Psi(y_{k-1}, \theta_\Psi) = \Xi(y_{k-1}) \theta_\Psi \Delta t + y_{k-1}
\]
Log posterior is a generalized objective function
Many existing approaches recovered under simplifying assumptions

1. Identity observations: Hills et. al. 2015, Raissi, 2018, Qin et.al 2019
3. Many of these works look at parameterizing models (e.g., via NN architectures)
4. We are looking at the construction of the objective
Log posterior is a generalized objective function
Many existing approaches recovered under simplifying assumptions

1. Identity observations: Hills et. al. 2015, Raissi, 2018, Qin et.al 2019
3. Many of these works look at parameterizing models (e.g., via NN architectures)
4. We are looking at the construction of the objective

Special cases: linearity and noiseless, invertible measurements → DMD and SINDy

Full Negative Log Posterior

\[-L_p(\theta, \mathcal{X}_n \mid \mathcal{Y}_n) \propto \sum_{i=1}^{n} \| y_i - h(X_i) \|^2 + \sum_{i=2}^{N} \| X_i - \Psi(X_{i-1}, \theta) \|^2 - \log p(\theta)\]

**MLE ↔ DMD estimator**
- linear model \( \Psi(X_k, \theta) = \theta X_k \)
- identity observation \( h = I \)
- noiseless measurements \( \Gamma = 0 \)
- identity process noise \( \Sigma = I \)

**MAP ↔ SINDy**
- linear-subspace model \( \phi(x) \theta \)
- identity observation \( h = I \)
- noiseless measurements \( \Gamma = 0 \)
- identity process noise \( \Sigma = I \)
- a time-stepping scheme
Dynamic mode decomposition DMD
Linear mappings in observable space (Koopman)

1. Shift observations

\[ Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_{m-1} \end{bmatrix} \]
\[ Y' = \begin{bmatrix} y_2 & y_3 & \cdots & y_m \end{bmatrix}. \]

2. Seek \( Y' = AY \)

3. Solve the optimization problem

\[
A = \arg \min_{\tilde{A}} \sum_{k=2}^{N} \| y_k - \tilde{A}y_{k-1} \|^2.
\]

Approximate the time derivative via finite differences

$$\theta = \arg\min_{\tilde{\theta}} \sum_{i=2}^{N} \left\| \frac{y_i - y_{i-1}}{dt} - \phi(y_{i-1})\tilde{\theta} \right\|^2 + \lambda \|\tilde{\theta}\|_1$$

Accounting for all uncertainties
Inference with the marginal likelihood

Goals: learn and predict

▶ Learn the model: \( p(\theta \mid Y_n, I) \)
▶ Make (future) predictions: \( p(X_k \mid Y_n, I) = \int p(X_k \mid \theta) p(\theta \mid Y_n, I) d\theta \)

Marginal posterior — required by Optimization/Markov Chain Monte Carlo

\[
p(\theta \mid Y_n, I) = \int p(\theta, X_n \mid Y_n, I) dX_n = \frac{1}{Z} \underbrace{p(Y_n \mid \theta)}_{\text{marginal likelihood}} p(\theta)
\]

Evaluating the marginal posterior

Recursive evaluation $p(\theta \mid \mathcal{Y}_n)$

1: for $k = 1$ to $n$ do
2: Predict $p(X_k \mid \theta, \mathcal{Y}_{k-1}) = \int p(X_k \mid \theta, X_{k-1})p(X_{k-1} \mid \theta, \mathcal{Y}_{k-1})dX_{k-1}$
3: Compute the evidence $p(y_k \mid \theta, \mathcal{Y}_{k-1}) = \int p(y_k \mid \theta, X_k)p(X_k \mid \theta, \mathcal{Y}_{k-1})dX_k$
4: Update filter $p(X_k \mid \theta, \mathcal{Y}_k) = \frac{p(y_k \mid \theta, X_k)p(X_k \mid \theta, \mathcal{Y}_{k-1})}{p(y_k \mid \mathcal{Y}_{k-1})}$
5: Update posterior $p(\theta \mid \mathcal{Y}_k) = \frac{p(y_k \mid \theta, \mathcal{Y}_{k-1})p(\theta \mid \mathcal{Y}_{k-1})}{p(y_k \mid \mathcal{Y}_{k-1})}$
6: end for
State estimation and Bayesian Filtering

Gaussian filters propagate the mean and variance

\[ \frac{dm}{dt} \quad \frac{dC}{dt} \quad m \leftarrow m + K \Delta y_k \quad C \leftarrow C - KSK^T \quad \frac{dm}{dt} \quad \frac{dC}{dt} \quad m \leftarrow m + K \Delta y_k \quad C \leftarrow C - KSK^T \]

- **Exact filters**
  - Kalman filter - only optimal for linear problems
  - Fokker-Planck + Bayes rule

- **Approximate filters**
  - Extended Kalman filter - accuracy suffers for nonlinear problems
  - Nonlinear Gaussian filters (ensemble, unscented, cubature)
  - Particle filters

- **Embedded for approximate likelihood computation** Khalil et. al. 2015, Noh 2019, Drovandi et. al. 2019
Linear pendulum: DMD model is “correct”
Correct DMD model still does not recover system

Reconstruction

(a) $x_1, \sigma = 10^{-2}, n = 8$
(b) $x_2, \sigma = 10^{-2}, n = 8$
(c) $x_1, \sigma = 10^{-1}, n = 40$
(d) $x_2, \sigma = 10^{-1}, n = 40$
Linear pendulum: DMD model is “correct”
Correct DMD model still does not recover system

Reconstruction

Prediction

(a) $x_1, \sigma = 10^{-2}, n = 8$
(b) $x_2, \sigma = 10^{-2}, n = 8$
(c) $x_1, \sigma = 10^{-1}, n = 40$
(d) $x_2, \sigma = 10^{-1}, n = 40$
Linear pendulum: DMD model is “correct”
Correct DMD model still does not recover system
Vanderpole system (2D nonlinear oscillator)

Even with correct model, SINDy cannot robustly recover the system

Reconstruction

(a) Reconstruction of $x_1$

(b) Reconstruction of $x_2$

(c) Reconstruction of $x_1$

(d) Reconstruction of $x_2$
Vanderpol system (2D nonlinear oscillator)
Even with correct model, SINDy cannot robustly recover the system

Reconstruction

(a) Reconstruction of $x_1$
(b) Reconstruction of $x_2$

Phase Diagram

(a) $\sigma = 10^{-3}, n = 2000$
(b) $\sigma = 2.5 \times 10^{-1}, n = 200$
Lorenz 63
With only 300 noisy data points, we recover the attractor

\[
\begin{align*}
\dot{x} &= \sigma (y - x) \\
\dot{y} &= x (\rho - z) - y \\
\dot{z} &= xy - \beta z
\end{align*}
\]
Lorenz 63
With only 300 noisy data points, we recover the attractor

\[
\begin{align*}
\dot{x} &= \sigma (y - x) \\
\dot{y} &= x (\rho - z) - y \\
\dot{z} &= xy - \beta z
\end{align*}
\]

Posterior Lyapunov exp.
Why does our objective lead to better reconstruction? Geometry of the posterior

1. Attempt to learn a linear pendulum
   - Top row (20 data points) middle (40), bottom (80)
   - Left column: deterministic dynamics
   - Middle column: noiseless observations
   - Right column: our objective

2. Neglecting process noise leads to many local min.

3. Neglecting measurement noise smooths objective too much

4. Our approach maintains geometry and global min.
Specializing to Hamiltonian systems

- Hamiltonian systems are reversible and preserve certain invariants (energy)
  \[ \dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q} \]

- Dynamics $\Phi$ become a mixture of a leap frog and Hamiltonian parameterization:
  \[ H = \frac{1}{2} p^T M^{-1}(q)p + U(q, p) \]

\[
\Psi(q_k, p_k; \theta_\Psi) = \left[ \begin{array}{l}
q_k + \Delta t p_k - \frac{\Delta t^2}{2} \frac{\partial U(q, p, \theta_\Psi)}{\partial q} \\
q_k
\end{array} \right] \bigg|_{q_k} + \left[ \begin{array}{l}
p_k - \frac{\Delta t}{2} \left( \frac{\partial U(q, p, \theta_\Psi)}{\partial q} \right)_{q_k} + \frac{\partial U(q, p, \theta_\Psi)}{\partial q} \right]_{q_{k+1}}
\]

- We parameterize the potential energy $U$
- We both learn the Hamiltonian and assume the data is from a symplectic process
Hénon Heiles Potential

\[ U(q_1, q_2) = \frac{1}{2} q_1^2 + \frac{1}{2} q_2^2 + q_1^2 q_2 - \frac{1}{3} q_2^3 \]

Phase plots

- (a) Leapfrog
- (b) Runge-Kutta
- (c) True
- (d) Position Data

Hamiltonian over time

- (a) Leapfrog Integrator
- (b) Runge-Kutta Integrator

The method equipped with RK must learn a smaller Hamiltonian to compensate for being non-conservative

Relative mean error:
- Leapfrog: 0.7%
- Runge-Kutta: 1.3%
Hénon Heiles Trajectory

**Posterior estimates of $q_1$ trajectory**

- **Data**
- **LF Posterior**
- **Leapfrog**
- **RK Posterior**
- **Runge-Kutta**
- **Truth**

**Process noise marginal posteriors**

Symplectic approach learns a model with an order of magnitude greater certainty
FPU Chain

\[ U(q_1, q_2, q_3) = \sum_{i=1}^{2} \frac{(q_{i+1} - q_i)^2}{2} + \frac{(q_{i+1} - q_i)^4}{40} \]

**Relative mean error:**
Leapfrog: 4.3%
Runge-Kutta: 6.7%

**Data Generation:**
- \( n = 26 \)
- \( \Delta t = 1.5 \)
- \( \sigma_q = 0.01 \)
- \( \sigma_p = 0.02 \)
FPU Chain Trajectory

Process noise marginal posteriors

Leapfrog

Runge-Kutta
Conclusions and Acknowledgements

1 Conclusions
   - Probabilistic models provide flexible and capable frameworks for system ID
     - Parameter uncertainty
     - Model uncertainty
     - Measurement uncertainty
   - MCMC approaches are feasible for moderately large problems
   - Significant gains can be achieved under non-optimal conditions over DMD and SINDy
   - How do we know the size of the latent space?

2 Papers on my website
   www.alexgorodetsky.com/publications.html

3 Funding: DARPA PAI and AIRA, AFOSR Computational Mathematics
Computational science for autonomy

Compression enabled control and estimation

Real-time autonomy

Source code: github.com/goroda, papers: alexgorodetsky.com

Thanks!