Moving Horizon Estimation (MHE) in a Distributional Framework

Vishaal Krishnan   Sonia Martínez

University of California, San Diego

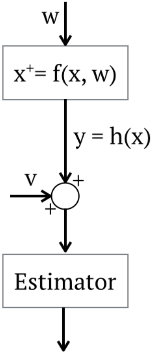
American Control Conference 2019

The Setting

System

$$x_{k+1} = f(x_k, w_k)$$
$$y_k = h(x_k) + v_k$$

- Lipschitz:
  f, h are Lipschitz-continuous
- Noise characteristics:
  $w_k,\ v_k$ (bounded, i.i.d.)



Optimization-based estimation:
Given data $\mathbf{y}_{0:T} = (y_0, \ldots, y_T)$, get estimates $\widehat{x}_k$ of x_k by an optimization procedure

Optimization-based Estimation

Full Information Estimation (FIE):

- Use all the data within the window $0 : T$
- Static optimization problem to obtain estimate:

$$\hat{x}_0 = \arg \min_{z_0, \mathbf{w}, \mathbf{v}} J(\mathbf{y}_{0:T}, \hat{\mathbf{y}}_{0:T})$$

Moving Horizon Estimation (MHE):

- Use a moving window of length $N << T$
- At time instant $k + N$, use data from $k + 1 : k + N$
- Recursive (online) optimization to obtain estimates:

$$\hat{x}_{k+1} \in \arg \min_{z_{k+1}, \mathbf{w}, \mathbf{v}} \gamma(f(\hat{x}_k, w_k), z_{k+1}) + J(\mathbf{y}_{k+1:k+N}, \hat{\mathbf{y}}_{k+1:k+N})$$

- Prediction + Correction

MHE vs Probabilistic Estimation

- MHE is a point estimator
- Estimation is fundamentally about dealing with uncertainty (Initial conditions, Measurement/Process noise)
- Estimates are essentially probability distributions, tracking the distribution is important:
  Kalman filter – Mean + Covariance define the distribution in Linear + Gaussian setting
  Bayesian estimation – Updating the prior distribution with measurements + likelihood to obtain posterior
- How about in the moving-horizon setting?
  Characterizing the evolution in nonlinear, non-Gaussian setting is difficult

Objectives of this work

1. **Point estimation vs Probabilistic estimation:**
   A unifying framework for MHE and Bayesian estimation
2. **Consistency/Asymptotic stability of MHE:**
   Connection to classical observability notions
   Existing results built on Input-Output-to-State Stability
3. **Computational bottleneck in MHE:**
   A theory of Fast-MHE (consistency, robustness)

Plan

1. Setup the Fast-MHE formulation
   (as a point estimator)
2. Lift the point estimator to the space of distributions
   (obtain the probabilistic version)
3. Realize the unification principle
   (MHE vs Bayes/Particle filter)
4. Establish **Observability** $\Rightarrow$ **Consistency**
5. Characterize robustness

Computational Bottleneck

The recursive formulation:

$$\widehat{x}_{k+1} \in \arg \min_{z_{k+1}, \mathbf{w}, \mathbf{v}} \gamma(f(\widehat{x}_k, w_k), z_{k+1}) + J(\mathbf{y}_{k+1:k+N}, \widehat{\mathbf{y}}_{k+1:k+N})$$

- Is computationally intensive
- Decision variables in optimization – $z_{k+1}, \mathbf{w}, \mathbf{v}$
  Scales with size of horizon – Need for fast MHE
- **Fast MHE:** Drop $\mathbf{w}, \mathbf{v}$ from the optimization[1]
  Propagate by noiseless system: $x_{k+1} = f(x_k)$; $y_k = h(x_k)$

[1] Alessandri, Gaggero, *TAC 2017*

Proximal GD formulation of Fast-MHE

Online proximal gradient descent:

$$\widehat{x}_{k+1} \in \arg\min_z \ \frac{1}{2}|z - f(\widehat{x}_k)|^2 + \eta G_k(z)$$

$G_k(z) = J(\mathbf{y}_{k+1:k+N},\ h \circ f^{1:N}(z))$
(assume G_k has l-Lipschitz gradient)

Objective function **strongly convex** for $\eta < \frac{1}{l}$:

$$\begin{aligned}\widehat{x}_{k+1} &= \text{prox}_{\eta G_k} \circ f(\widehat{x}_k) \\ &= f(\widehat{x}_k) - \eta \nabla G_k(\widehat{x}_{k+1})\end{aligned}$$

Alternative: Regular gradient descent

$$\widehat{x}_{k+1} = f(\widehat{x}_k) - \eta \nabla G_k(\widehat{x}_k)$$

Fast-MHE in a distributional framework

Lift $\widehat{x}_{k+1} = \arg\min_z \frac{1}{2}|z - f(\widehat{x}_k)|^2 + \eta G(z)$ to $\mathcal{P}(\mathbb{X})$:
(in the space of distributions over state space $\mathbb{X}$)

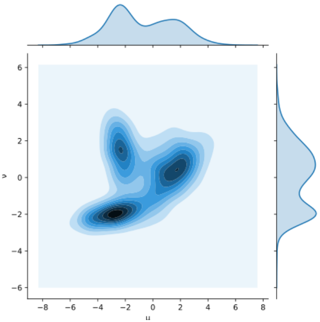
$$\mu_{k+1} = \arg\min_{\nu} \frac{1}{2} W_2^2(f_{\#}\mu_k, \nu) + \eta \mathbb{E}_{\nu}[G]$$

W_2 is the L^2-Wasserstein distance

L^2-Wasserstein distance

$$W_2^2(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{X}\times\mathbb{X}} |x-y|^2 d\pi(x,y)$$

- Optimal transport cost from μ to ν (w.r.t. Euclidean distance)
- Defines a metric on $\mathcal{P}(\mathbb{X})$



Unification of MHE and Bayes

Distributional MHE:

$$\mu_{k+1} = \arg\min_{\nu} \; \frac{1}{2} W_2^2(f_{\#}\mu_k, \nu) + \eta \mathbb{E}_{\nu}[G]$$

Replace Wasserstein $\frac{1}{2}W_2^2$ by a general D:

$$\mu_{k+1} = \arg\min_{\nu} \; D(f_{\#}\mu_k, \nu) + \eta \mathbb{E}_{\nu}[G]$$

Observation:
If $D = D_{KL}$ (KL-divergence) – Bayesian estimation (Implement by Sequential Monte Carlo/Particle filter)

Plan

1. Setup the Fast-MHE formulation
   (as a point estimator) ✓
2. Lift the point estimator to the space of distributions
   (obtain the probabilistic version) ✓
3. Realize the unification principle
   (MHE vs Bayes/Particle filter) ✓
4. Establish **Observability** $\Rightarrow$ **Consistency**
5. Characterize robustness

State Estimation as an Inverse problem

1. State Estimation is fundamentally an inverse problem
   Output measurements $\rightarrow$ Underlying state
2. Well-posedness of the inverse problem
   Observability $\Rightarrow$ (locally) unique solutions

Linear vs. Nonlinear estimation:

	Linear	Nonlinear
Observable	Unique solution	Isolated solutions
Unobservable	Linear subspace of solutions	Submanifold of solutions

Observability

Noiseless system

$$x_{k+1} = f(x_k)$$
$$y_k = h(x_k)$$

Construct output maps

$$\Sigma_T(x) = (h(x), h \circ f(x), \ldots, h \circ f^T(x))$$
$$\Sigma(x) = (h(x), h \circ f(x), \ldots)$$

Observability notions

1. Strong (local) observability: There is a finite T s.t. ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$ is a set of isolated points
2. Weak (local) observability: $\Sigma^{-1}(\mathbf{y}_{0:\infty})$ is a set of isolated points

Result: Observability characterized by a rank condition[2] on Σ_T, Σ

[2] H. Nijmeijer, "Observability of autonomous discrete time non-linear systems: a geometric approach", *IJC 1982*

Consistency of MHE

Consistency notion:
Asymptotic stability of estimator in the absence of noise
$\lim_{k\to\infty} |x_k - \widehat{x}_k| = 0$ when $\mathbf{w} = 0, \ \mathbf{v} = 0$

Does observability $\Rightarrow$ consistency of MHE?
Open questions from literature:

- Results in literature[3] establish consistency from Input-Output-to-State Stability (IOSS)
- IOSS not connected to well-posedness (observability)
- Results on consistency of Fast-MHE assume convexity of cost (very restrictive for nonlinear systems)
- Robustness guarantees for Fast-MHE

[3] Rao, Rawlings, Mayne, "Constrained state estimation for nonlinear discrete-time systems", *TAC 2000*

Consistency from Observability

Basin of attraction: $\mathcal{C}_k$ is the basin of attraction of $\Sigma_T{}^{-1}(\mathbf{y}_{k:k+T})$ for gradient descent on G_k

Assumption (Positive invariance)

There exists $\alpha > (1 - \sqrt{1 - 2lL})l^{-1}$ *s.t. for all* $\eta \in (0, \alpha)$, *we have* $\mathrm{prox}_{\eta G_k^N} \circ f(\mathcal{C}_{k-1}) \subseteq \mathcal{C}_k$.

Theorem (Asymptotic stability of W_2-MHE)

For a strongly observable system, the MHE with $\eta \in \left(\frac{1-\sqrt{1-2lL}}{l}, \min\left\{\alpha, \frac{1}{l}\right\}\right)$ *is consistent.*

Robustness

Robustness of MHE to:

1. Uncertainty in initial conditions – $W_2(\mu_0, \delta_{x_0})$
2. Noise – $\mathbf{w}$, $\mathbf{v}$

Theorem (Robustness of W_2-MHE)

The W_2-MHE satisfies:

$$W_2(\mu_k, \delta_{x_k}) \leq \beta(W_2(\mu_0, \delta_{x_0}), k) + \gamma(\mathbf{w}_{0:k}, \mathbf{v}_{0:k})$$

for $\beta \in \mathcal{KL}$, $\gamma \in \mathcal{K}$.

Future Work

1. Compare different estimators within the unified framework (performance w.r.t. estimation error, rate of convergence, etc)
2. Connection between well-posedness and stability notions (observability/detectability vs IOSS)
3. Effect of horizon length on estimator performance
4. Coupling b/w point estimates to improve MHE performance

Thank You