On Observability and Stability of Moving-Horizon Estimation in a Distributional Framework

Vishaal Krishnan and Sonia Martínez

Abstract— **In this work, we propose a unifying framework in the space of probability measures for gradient-based and sampling-based moving-horizon estimation methods. We begin with an investigation of the classical notion of strong local observability of nonlinear systems and its relationship to optimization-based state estimation. We then present a general moving-horizon estimation framework for strongly locally observable systems, as an iterative minimization scheme in the space of probability measures. This framework allows for the minimization of the estimation cost with respect to different metrics and divergences. In particular, we consider two variants, which we name W_2-MHE and KL-MHE, where the minimization scheme uses the 2-Wasserstein distance and the KL-divergence respectively. The W_2-MHE yields a gradient-based estimator whereas the KL-MHE yields a particle filter, for which we investigate asymptotic stability and robustness properties. Stability results for these moving-horizon estimators are derived in the distributional setting, against the backdrop of the classical notion of strong local observability which, to the best of our knowledge, differentiates it from other previous works. We also present results from numerical simulations to demonstrate the performance of these estimators.**

I. Introduction

Moving-horizon estimation (MHE) is an optimization-based state estimation method that uses the most recent measurements within a moving time horizon to recursively update state estimates. In principle, its optimization-based formulation enables it to handle nonlinearities and state constraints much more effectively than other known methods. This, coupled with the availability of increasingly powerful, inexpensive computing platforms, has brought new impetus to the adoption of moving-horizon methods in novel estimation applications.

The origins of MHE can be traced back to limited-memory optimal filters, see [1] for an early work. Further theoretical investigations have broadly been directed at establishing their asymptotic stability [2]–[4] and robustness [5]–[7] properties. These results have primarily been built upon underlying assumptions of input/output-to-state (IOSS) stability [8], which is adopted as the notion of detectability. However, alternative foundations for stability results relying on other classical notions of observability, such as strong observability [9], have remained unexplored. The connection between nonlinear observability theory and estimation problems runs deep, see [10] and, more recently [11], and it is worthwhile to study this unexamined connection in the context of optimization-based estimation methods.

On the other hand, the problem of state estimation is fundamentally about dealing with uncertainty, manifested as uncertainty in the initial conditions and/or in the evolution of the system in the presence of unknown disturbances. This can be appropriately formulated in the space of probability measures over the state space of the system. Recent advances in gradient flows in the space of probability measures [12], and the corresponding discrete-time movement-minimizing schemes [13] present powerful theoretical tools that can be leveraged for recursive optimization-based estimation methods such as MHE, and can serve as a unifying framework for their design and analysis.

Another important consideration for the applicability of MHE is the cost of computation, which increases with the length of the horizon. The most widely adopted formulation of MHE requires solving an optimization problem at every time instant, with both state estimate and disturbances taken as decision variables. This approach, in general, tends to be computationally intensive, which poses a hurdle for implementation in real-time. This has motivated the search for fast MHE that just implement one or more iterations of an optimization algorithm at every time instant. Recently, in [14], [15], the authors develop such method for noiseless systems and provided theoretical guarantees for convergence. However, these works assume the cost function is convex, which is restrictive for general nonlinear systems, and not well connected to notions of observability.

Statement of contributions: In this work, we begin with the well-studied notion of strong local observability of nonlinear, discrete-time systems and investigate its relationship to the optimization-based state estimation problem. To handle uncertain initial conditions and the possible nonuniqueness of solutions to the estimation problem, we generalize the basic setting as an optimization problem in the space of probability measures over the state space. In particular, we formulate the MHE as a proximal-gradient optimization in this space, with a nonconvex, time-varying cost function. This distributional formulation serves as a unifying framework for moving-horizon estimation and allows us to develop different classes of estimators by varying the metric used in the proximal operator. In this way, we consider the Wasserstein metric and the KL-divergence, which yield the more familiar MHE and a particle filter, respectively, after a Monte-Carlo sampling procedure. Following this, we present an analysis of the convergence and robustness properties of these filters in the distributional setting, under assumptions of strong local

The authors gratefully acknowledge support from grants FA9550-18-1-0158 and DARPA (Lagrange) N660011824027.

The authors are with the Department of Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla CA 92093 USA (email: v6krishn@ucsd.edu; soniamd@ucsd.edu).

observability.

II. Notation and preliminaries

In this section, we introduce the notation and mathematical preliminaries relevant to this paper.

We denote by $\nabla = \left(\frac{\partial}{\partial x_1}, \dots \frac{\partial}{\partial x_n}\right)$ the gradient operator in $\mathbb{R}^d$. For any $x \in \mathcal{X} \subset \mathbb{R}^d$, we let $\mu \in \mathcal{P}(\mathcal{X})$ be an absolutely continuous probability measure on $\mathcal{X} \subset \mathbb{R}^d$. We denote by ρ the corresponding density function, where $\mathrm{d}\,\mu = \rho\,\mathrm{dvol}$, with vol being the Lebesgue measure. Let M be a subset of a metric space $(\mathcal{X}, d)$. The distance $d(x, M)$ of a point $x \in \mathcal{X}$ to the set M is given by $d(x, M) = \inf_{y \in M} d(x, y)$. We denote by $\langle p, q\rangle$ the inner product of functions $p, q : \mathcal{X} \to \mathbb{R}$ with respect to the Lebesgue measure vol, given by $\langle p, q\rangle = \int_{\mathcal{X}} pq\,\mathrm{dvol}$. Let $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be a smooth real-valued function on the space of probability measures on $\mathcal{X} \subset \mathbb{R}^d$. We denote by $\frac{\delta F}{\delta \mu}(x)$ the derivative of F with respect to μ, see [16], such that a perturbation $\delta\mu$ of the measure results in a perturbation $\delta F = \int_{\mathcal{X}} \frac{\delta F}{\delta \mu} d(\delta\mu)$. Given a map $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu \in \mathcal{P}(\mathcal{X})$, in the space of probability measures $\mathcal{P}(\mathcal{X})$, we let $\nu = \mathcal{T}_{\#}\mu$ denote the pushforward measure of μ by $\mathcal{T}$, where for a measurable set $\mathcal{B} \subset \mathcal{T}(\mathcal{X})$, we have $\nu(\mathcal{B}) = \mathcal{T}_{\#}\mu(\mathcal{B}) = \mu(\mathcal{T}^{-1}(\mathcal{B}))$.

We now introduce the notion of l-smoothness that underlies the results on convergence of gradient descent methods.

Definition 1: ***(l-Smoothness).*** A function $p : \mathcal{X} \to \mathbb{R}$ is called *l-smooth* if for any $x, y \in \mathcal{X}$, we have $|\nabla p(y) - \nabla p(x)| \le l\|y - x\|$.

The following lemma can be easily verified for l-smooth functions:

Lemma 1: ***(l-Smooth functions).*** For an l-smooth function $p : \mathcal{X} \to \mathbb{R}$ and any $x, y \in \mathcal{X}$, we have $|p(y) - p(x) - \langle \nabla p(x), y - x\rangle| \le \frac{l}{2}\|y - x\|^2$. •

We now define the proximal operator on a metric (or pseudo metric) space $(\mathcal{Y}, d)$ endowed with a metric (or pseudo metric) d, with respect to a function $F : \mathcal{Y} \to \mathbb{R}$, as follows:

$$\mathrm{prox}_F^d(y) = \arg\min_{\tilde{y} \in \mathcal{Y}} \frac{1}{2} d^2(\tilde{y}, y) + F(\tilde{y}).$$

We now introduce the notion of lower semicontinuity of set-valued maps, which underlies some of the results on optimization-based state estimation in this paper.

Definition 2: ***(Lower semicontinuity of set-valued maps).*** A point-to-set mapping $L : \mathcal{Z} \subset \mathbb{R} \rightrightarrows \mathbb{R}^d$ is lower semicontinuous at a point $\alpha \in \mathcal{Z}$ if for any $x \in L(\alpha)$ and sequences $\{\alpha_i\} \subseteq \mathcal{Z}$, $\{x_i\} \subseteq \mathbb{R}^d$ with $\{\alpha_i\} \to \alpha$, $\{x_i\} \to x$ such that $x_i \in L(\alpha_i)$ for all i, it holds that $x \in L(\alpha)$. If L is lower semicontinuous at every $\alpha \in \mathcal{Z}$, then L is said to be lower semicontinuous on $\mathcal{Z}$.

III. Observability notions

In this paper, we consider systems of the form:

$$\Omega : \begin{cases} x_{k+1} = f(x_k, w_k), \\ y_k = h(x_k) + v_k, \end{cases} \tag{1}$$

where $f : \mathbb{X} \times \mathbb{W} \to \mathbb{X}$ and $h : \mathbb{X} \to \mathbb{Y}$, $w_k \in \mathbb{W}$ is the process noise, $v_k \in \mathbb{V}$ is the measurement noise at time instant k, and sets of appropriate dimension $\mathbb{X} \subset \mathbb{R}^{d_X}$, $\mathbb{Y} \subset \mathbb{R}^{d_Y}$, $\mathbb{W} \subset \mathbb{R}^{d_W}$ and $\mathbb{V} \subset \mathbb{R}^{d_V}$.

Assumption 1: ***(Lipschitz continuity).*** The functions f and h are Lipschitz continuous, such that $\|f(x_1, w_1) - f(x_2, w_2)\| \le c_f^{(1)}\|x_1 - x_2\| + c_f^{(2)}\|w_1 - w_2\|$ and $\|h(x_1) - h(x_2)\| \le c_h\|x_1 - x_2\|$.

Assumption 2: ***(Noise characteristics).*** The noise sequences $\{w_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$ are i.i.d samples from distributions ω and ν (with supports in $\mathbb{W}$ and $\mathbb{V}$). The sets $\mathbb{W}$ and $\mathbb{V}$ are bounded, with $|w_k| \le W$ and $|v_k| \le V$. Moreover, we assume that $\mathbb{E}[w_k] = 0$ and $\mathbb{E}[v_k] = 0$.

We also introduce the following autonomous system corresponding to (1):

$$\Sigma : \begin{cases} x_{k+1} = f(x_k, 0) = f_0(x_k), \\ y_k = h(x_k). \end{cases} \tag{2}$$

With a slight abuse of notation, for any $x \in \mathbb{X}$, we let $\Sigma_T(x) = \big(h(x), h \circ f_0(x), \dots, h \circ f_0^T(x)\big)$ be the sequence of outputs over a horizon of length $T + 1$ for the system (2) from the state $x \in \mathbb{X}$. Similarly, for the system (1), we let $\Omega_{\mathbf{w}_{i:j}}(x) = (h(x), h \circ f(x, w_i), \dots, h \circ f(\dots f(f(x, w_i), w_{i+1}), \dots, w_j)$, for some sequence of process noise samples $\{w_k\}$, where $\mathbf{w}_{i:j} = (w_i, \dots, w_j)$.

We now introduce the notion of strong local observability used in this paper:

Definition 3: ***(Strong local observability).*** The system Σ defined in (2) is called *strongly locally observable* if there exists a $T_0 \in \mathbb{N}$ such that for any given $\mathbf{y}_T = \Sigma_T(x) \in \mathbb{Y}^{T+1}$ and $T \ge T_0$, we have that ${\Sigma_T}^{-1}(\mathbf{y}_T)$ is a set of isolated points, and, in addition, ${\Sigma_{T_1}}^{-1}(\mathbf{y}_1) = {\Sigma_{T_2}}^{-1}(\mathbf{y}_2)$, for all $\mathbf{y}_1 = \Sigma_{T_1}(x)$ and $\mathbf{y}_2 = \Sigma_{T_2}(x)$, and $T_1, T_2 \ge T_0$. We call T_0 the *minimum horizon length* of Σ.

The above definition is equivalent to the definitions contained in [9] and [18]. We have restated it in a manner that is suitable for the optimization-based estimation framework that is considered in this paper. For systems with process noise, of the form Ω in (1), we introduce the notion of almost sure strong local observability.

Definition 4: ***(Almost sure strong local observability).*** The system Ω defined in (1) is called *almost surely strongly locally observable* if there exists a $T^w \in \mathbb{N}$ such that, given a process noise sequence $\mathbf{w}_{0:T-1} \in \mathbb{W}^T$, for $T \ge T^w$, and any $\mathbf{y} = \Omega_{\mathbf{w}_{0:T-1}}(x) \in \mathbb{Y}^{T+1}$ and $T \ge T^w$, we have that $\Omega^{-1}_{\mathbf{w}_{0:T-1}}(\mathbf{y})$ is a set of isolated points almost surely. More precisely, the set of noise sequences $\mathbf{w}_{0:T-1}$ for which $\Omega^{-1}_{\mathbf{w}_{0:T-1}}(\mathbf{y})$ is not a set of isolated points, is of measure zero. Moreover, we call T^w the *minimum horizon length* of Ω.

We now present a fundamental result that characterizes strong local observability via a rank condition.

Lemma 2: ***(Observability rank condition [9]).*** The system Σ is locally strongly observable with minimum horizon length T_0 if and only if $\mathrm{Rank}(\nabla\Sigma_T(x)) = \dim(\mathbb{X})$ for all $T \ge T_0$ and $x \in \mathbb{X}$. The system Ω is almost surely locally strongly observable with minimum horizon length T^w if and only if $\mathrm{Rank}(\nabla\Omega_{\mathbf{w}_{0:T-1}}(x)) = \dim(\mathbb{X})$ almost surely for all $T \geq T^w$. •

We make the following assumption in the rest of the paper:

Assumption 3: ***(Strong local observability).***

1) The system Σ^0 in (2) is strongly locally observable with minimum horizon length T_0.
2) The system Σ in (1) is almost surely strongly locally observable with minimum horizon length T^w.

IV. Optimization-based state estimation

We now begin by addressing the state estimation problem for the autonomous system Σ, and develop a recursive moving horizon estimator for it.

A. Full Information Estimation (FIE)

Let $\{y_k\}_{k\in\{0\}\cup\mathbb{N}}$ be a sequence of measurements generated by the system Σ. Let $\{0,\dots,T\}$ be a time horizon such that $T \geq T_0$, the minimum horizon length of the system Σ, and denote $\mathbf{y}_{0:T} = (y_0,\dots,y_T)$. The problem of estimation essentially aims at characterizing ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$, which is an inverse problem, and the problem of optimal estimation aims to solve it through an optimization. Assumptions 1, and 3, on Lipschitz continuity and strong local observability, respectively, ensure that the inverse problem is locally well-posed, as defined in [19].

To formulate the inverse problem as an optimization, consider a convex function $J_T(\mathbf{y}_{0:T},\cdot) : \mathbb{Y}^{T+1} \to \mathbb{R}_{\geq 0}$ such that $J_T(\mathbf{y}_{0:T},\xi) = 0$ if and only if $\xi = \mathbf{y}_{0:T}$. Now, the problem of interest becomes:

$$x_0 \in \arg\min_{x\in\mathbb{X}} J_T(\mathbf{y}_{0:T}, \Sigma_T(x)). \tag{3}$$

In the above, $\mathbf{y}_{0:T}$ is the data in the estimation problem, which is given. Since the objective is to solve the original inverse problem, and we would like to use gradient descent-based methods, we would like that every local minimizer of $J_T(\mathbf{y}_{0:T},\Sigma_T(x))$ belongs to the set ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$, or, in other words, that every local minimizer is also global. We therefore make the following additional assumption on the system Σ and the choice of J_T. For a conciseness of notation, in the following assumption and lemma, we let $J_T(\cdot) = J_T(\mathbf{y}_{0:T},\cdot)$, suppressing the data $\mathbf{y}_{0:T}$ in the notation where useful, and is understood from context.

Assumption 4: ***(Lower semicontinuity of sublevel sets).*** We assume that, for all $T \geq T_0$, the convex function $J_T : \mathbb{Y}^{T+1} \to \mathbb{R}$ is such that the set-valued map $\mathcal{S}_{\mathbb{X}}(\alpha) = {\Sigma_T}^{-1}\left(\mathcal{S}^{J_T}_{\mathbb{Y}^{T+1}}(\alpha) \cap \Sigma_T(\mathbb{X})\right)$ is lower semicontinuous, where $\mathcal{S}^{J_T}_{\mathbb{Y}^{T+1}}(\alpha) = \{\xi \in \mathbb{Y}^{T+1} | J_T(\xi) \leq \alpha\}$.

The above assumption ensures that $J_T\left(\mathbf{y}_{0:T},\Sigma_T(\cdot)\right)$ satisfies the condition for the local minimizers to be global (Theorem 1 from [20]). The following lemma provides a sufficient condition for it to hold.

Lemma 3: ***(Second-order sufficient condition for lower semicontinuity).*** Assumption 4 holds if for any $x \in \mathbb{X}$ such that $\nabla\left(J_T(\mathbf{y}_{0:T},\Sigma_T(x))\right) = 0$ we have $J_T(\mathbf{y}_{0:T},\Sigma_T(x)) = 0$, or the following condition holds when $J_T(\mathbf{y}_{0:T},\Sigma_T(x)) \neq 0$ for any $v \in \mathbb{R}^{d_X}$, $v \neq 0$:

$$\frac{\langle \nabla^2\Sigma_T[v,v](x), \nabla J_T(\Sigma_T(x))\rangle}{\|\nabla\Sigma_T[v]\|^2} \leq -\lambda_{\max}\left(\mathrm{Hess}\, J_T\Big|_{\Sigma_T(x)}\right)$$

•

The final inequality in Lemma 3 merely states that those critical points at which the cost function does not reach the global minimum value are local maximizers.

We are now ready to present the following theorem that establishes the equivalence between the inverse problem of characterizing the set ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$ and the optimization (3).

Theorem 1: ***(Inverse as minimizer).*** Under Assumptions 3 and 4, for any $T \geq T_0$, it holds that $z \in {\Sigma_T}^{-1}(\mathbf{y}_{0:T})$ if and only if z is a local minimizer of $J_T(\mathbf{y}_{0:T},\Sigma_T(\cdot))$.

Theorem 1 suggests that the state estimates for the system Σ can be obtained by minimizing $J_T(\mathbf{y}_{0:T},\Sigma_T(\cdot))$ over a horizon of length $T \geq T_0$. This is also called the full information estimation (FIE) problem in the optimal state estimation literature [2], [7], as it works with the entire sequence of output measurements over the horizon $\{0,\dots,T\}$.

We let $\mathcal{C}_k$ be the basin of attraction of ${\Sigma_T}^{-1}(\mathbf{y}_{k:k+T})$. We now lift the FIE problem (3) to the space of probability measures over $\mathbb{X}$ as a minimization in expectation of the estimation objective function:

$$\mu_0 \in \arg\min_{\mu\in\mathcal{P}(\mathbb{X})} \mathbb{E}_\mu\left[J_T(\mathbf{y}_{0:T},\Sigma_T(\cdot))\right]. \tag{4}$$

The above formulation allows us to capture information about the (probably many) optimal estimates through a probability measure μ_0, and help encode other distributional constraints, which will be considered in future work.

In the following, we develop recursive moving-horizon estimators that generate sequences $\{\mu_k\}_{k\in\mathbb{N}}$ of probability measures in $\mathcal{P}(\mathbb{X})$ as estimates. We then obtain practically implementable estimators using Monte Carlo methods to sample from the measures μ_k.

B. Moving Horizon Estimation (MHE)

In the previous section, we presented a full information estimation (FIE) problem for the autonomous system Σ, which uses the entire measurement sequence over a horizon of length $T \geq T_0$. However, the minimum horizon length T_0 may be large, which would make the estimation computationally intensive. We therefore adopt a moving-horizon estimation method which, at any time instant $k+N$, uses the output measurements from the horizon $\{k+1,\dots,k+N\}$ (of length $N < T_0$), and the state estimate at the time instant $k-1$, to obtain the state estimate at instant k, recursively.

We let $G^N_k(z) = J_{N-1}\left(\mathbf{y}_{k+1:k+N},\Sigma_N(z)\right)$ be the objective function over the horizon $\{k+1,\dots,k+N\}$, at the time instant $k+N$, where $\mathbf{y}_{k+1:k+N} = (y_{k+1},\dots,y_{k+N})$.

Assumption 5: ***(Moving horizon cost).*** We make the following assumptions on the cost function G^N_k:

1) the cost G^N_k is l-smooth,
2) it holds that $|G^N_{k+1}(f_0(z)) - G^N_k(z)| \leq L\|\nabla G^N_k(z)\|^2$,

3) the previous constants are such that $lL \leq \frac{1}{2}$,
4) for any two δ-adjacent measurements $\mathbf{y}, \tilde{\mathbf{y}} \in \mathbb{Y}^{T+1}$, such that $\|\mathbf{y} - \tilde{\mathbf{y}}\| \leq \delta$ and with corresponding costs G_k^N and $\widetilde{G}_k^N$, for $k \in \{0, \ldots, T\}$ and $N \leq T-k$, we have $\|\nabla(G_k^N(x) - \widetilde{G}_k^N(x))\| \leq l\delta$ for all $x \in \mathbb{X}$.

We now formulate the general moving horizon estimation method as follows:

$$\begin{aligned} &\mu_k \in \arg\min_{\mu \in \mathcal{P}(\mathbb{X})} D(\mu, f_{0\#}\mu_{k-1}) + \eta \mathbb{E}_\mu \left[G_k^N\right], \\ &\text{given} \;\; \mu_0 \in \mathcal{P}(\mathbb{X}), \end{aligned} \tag{5}$$

where D is a (pseudo) metric in $\mathcal{P}(\mathbb{X})$. We obtain implementable observers from the above formulation by sampling from the measures by Monte Carlo methods. As discussed in the ensuing sections, using the 2-Wasserstein distance W_2 yields the more familiar MHE formulation, whereas with the KL-divergence we obtain a moving-horizon particle filter. Hence, this formulation is proposed as a distributional unifying framework for moving-horizon estimation, where different estimators are generated by different choices of D.

We now introduce the following asymptotic stability notion for estimators that will be used in investigating the properties of the estimators we design.

Definition 5: ***(Asymptotic stability of state estimator).*** We call an estimator of the form (5) an *asymptotically stable observer* for the system Σ if the sequence of estimates $\{\mu_k\}_{k\in\mathbb{N}}$ is such that $\lim_{k\to\infty} \mu_k({\Sigma_T}^{-1}(\mathbf{y}_{k:k+T})) = 1$ for $T \geq T_0$.

V. A W_2-Moving-Horizon Estimator

In this section, we derive a moving-horizon estimator, which we refer to as the W_2-MHE, to generate a sequence of probability distributions $\{\mu_k\}_{k\in\mathbb{N}}$. This is based on the one-step minimization scheme of [12] in $\mathcal{P}(\mathbb{X})$ w.r.t. the Wasserstein metric W_2, which we extend to the moving-horizon setting. For every $k > 0$, consider:

$$\begin{aligned} &\mu_k \in \arg\min_{\mu \in \mathcal{P}(\mathbb{X})} \frac{1}{2} W_2^2(\mu, f_{0\#}\mu_{k-1}) + \eta \mathbb{E}_\mu \left[G_k^N\right], \\ &\text{given} \;\; \mu_0 \in \mathcal{P}(\mathbb{X}). \end{aligned} \tag{6}$$

We let $\mathcal{K}_k$ be the support of μ_k, with $\mathcal{K}_0 \subseteq \mathcal{C}_0$, where $\mathcal{C}_0$ is as defined earlier in Section IV-A.

We represent the above in a compact form using the proximal operator on $\mathcal{P}(\mathbb{X})$ associated with $\mathbb{E}_\mu \left[G_k^N\right]$ and w.r.t. the Wasserstein metric W_2, which we denote as follows:

$$\mu_k \in \operatorname{prox}_{\eta G_k^N}^{W_2} (f_{0\#}\mu_{k-1}), \quad k > 0. \tag{7}$$

The objective functional in (6) is not necessarily convex, which implies that the image of the proximal mapping (7) is not necessarily a singleton.

The sample-update scheme and implementable filter for the W_2-MHE formulation is given by:

$$\begin{aligned} &z_k \in \arg\min_z \frac{1}{2} \left|z - f_0(z_{k-1})\right|^2 + \eta G_k^N(z), \quad k > 0, \\ &z_0 \sim \mu_0 \in \mathcal{P}(\mathbb{X}). \end{aligned} \tag{8}$$

Lemma 4: ***(Strong convexity).*** For $\eta < l^{-1}$, the objective function in (8) is strongly convex, and therefore $\operatorname{prox}_{\eta G_k^N}(f_0(x))$ is a singleton for any $x \in \mathbb{X}$.

A. Asymptotic stability of W_2-MHE

We present the asymptotic stability result for W_2-MHE in this section, before which we introduce the following assumption on positive invariance of the discrete-time dynamics defined by the map $\operatorname{prox}_{\eta G_k^N} \circ f$.

Assumption 6: ***(Positive invariance).*** We assume that there exists $\alpha > (1 - \sqrt{1-2lL})l^{-1}$ such that for all $\eta \in (0, \alpha)$, we have $\operatorname{prox}_{\eta G_k^N} \circ f(\mathcal{C}_{k-1}) \subseteq \mathcal{C}_k$.

The above assumption ensures that under the discrete-time dynamics defined by the map $\operatorname{prox}_{\eta G_k^N} \circ f$, any sequence starting in the basin of attraction $\mathcal{C}_0$ of ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$ remains within the basins of attraction $\mathcal{C}_k$ of ${\Sigma_T}^{-1}(\mathbf{y}_{k:k+T})$ at the subsequent instants of time $k \in \mathbb{N}$.

We are now ready to present the asymptotic stability result for W_2-MHE:

Theorem 2: ***(Asymptotic stability of W_2-MHE).*** The estimator (6), under Assumptions 3 to 6, with a constant step size $\eta \in \left(\frac{1-\sqrt{1-2lL}}{l}, \min\left\{\alpha, \frac{1}{l}\right\}\right)$, is an asymptotically stable observer for the system Σ.

B. Robustness of W_2-MHE

We now characterize the performance of the estimator (6) on the system Ω in (1). Since the true process and measurement noise sequences remain unknown, we are interested in the robustness properties of the estimator (9), in the form of an upper bound by the norms of the disturbance sequences on the estimation error.

We begin by constructing a reference estimator that recursively generates the estimate sequence, given the true disturbance sequences $\{w_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$, as follows:

$$\begin{aligned} &\bar{\mu}_k \in \arg\min_{\mu \in \mathcal{P}(\mathbb{X})} \frac{1}{2} W_2^2(\mu, f_{0\#}\bar{\mu}_{k-1}) + \eta \mathbb{E}_\mu \left[\bar{G}_k^N\right], \\ &\text{given} \;\; \bar{\mu}_0 \in \mathcal{P}(\mathbb{X}). \end{aligned} \tag{9}$$

where, we employ for conciseness $\mathbf{w} \equiv \mathbf{w}_{k:k+N-1} = (w_k, \ldots, w_{k+N-1})$ and $\mathbf{v} \equiv \mathbf{v}_{k+1:k+N} = (v_{k+1}, \ldots, v_{k+N})$, so that $\bar{G}_k^N(z) \equiv \bar{G}_k^N(z, \mathbf{w}, \mathbf{v}) = J_{N-1}\left(\mathbf{y}_{k+1:k+N}, \Omega_{\mathbf{w}_{k:k+N-1}}(z) + \mathbf{v}_{k+1:k+N}\right)$. Note that $G_k^N = \bar{G}_k^N\big|_{\mathbf{w}=0,\mathbf{v}=0}$. We let $\bar{\mathcal{K}}_k$ be the support of $\bar{\mu}_k$, with $\bar{\mathcal{K}}_0 \subseteq \bar{\mathcal{C}}_0$, where the definition of $\bar{\mathcal{C}}_k$ is similar to that of $\mathcal{C}_k$ but taking the noise $\{w_k\}$ and $\{v_k\}$ into account.

Assumption 7: ***(l-Smoothness w.r.t. disturbances).*** We assume that $\|\nabla G_k^N(z) - \nabla \bar{G}_k^N(z)\| \leq l_w \|(\mathbf{w}_{k:k+N-1}, \mathbf{v}_{k+1:k+N})\|$ for all $z \in \mathbb{X}$.

Following Theorem 2, under the same set of underlying assumptions, we infer that the reference estimator (9) is almost surely an asymptotically stable observer for the system Ω, given a particular realization of the disturbances $\{w_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$.

We now present the following theorem on the robustness of the estimator (6), characterized by a bound on the error in the estimates generated by (6) with respect to the estimates generated by the reference estimator (9):

Theorem 3: ***(Robustness of W_2-MHE).*** Under Assumptions 1, 3, 5, and 7, given the estimate sequences $\{\mu_k\}_{k\in\mathbb{N}}$

generated by (6) and $\{\bar{\mu}_k\}_{k\in\mathbb{N}}$ generated by the reference estimator (9), with $\mu_0 = \bar{\mu}_0$, we have $W_2(\mu_k, \bar{\mu}_k) \leq \frac{c_f^{(2)}}{c_f^{(1)}} W C_k + \frac{\eta l_w \sqrt{N}}{c_f^{(1)}}(W+V)C_k$, for all $k \in \mathbb{N}$, where $C_k = \sum_{\ell=1}^{k} (\frac{c_f^{(1)}}{1-\eta l})^{\ell}$.

VI. A KL-Moving-Horizon Estimator

In this section, we derive a moving-horizon estimator, which we refer to as KL-MHE, to generate a sequence of probability distributions $\{\mu_k\}_{k\in\mathbb{N}}$. Using the KL-divergence D_{KL} as the choice of pseudo-distance in the moving-horizon formulation (5). to obtain:

$$\begin{aligned} &\mu_k \in \arg \min_{\mu \in \mathcal{P}(\mathbb{X})} D_{\text{KL}}(\mu \| f_{0\#}\mu_{k-1}) + \eta \mathbb{E}_{\mu}\left[G_k^N\right], \\ &\text{given} \quad \mu_0 \in \mathcal{P}(\mathbb{X}). \end{aligned} \tag{10}$$

We represent the above in a compact form using the proximal operator on $\mathcal{P}(\mathbb{X})$ associated with $\mathbb{E}_{\mu}\left[G_k^N\right]$ and w.r.t. the KL-divergence, which we denote as follows:

$$\mu_k \in \text{prox}_{\eta G_k^N}^{D_{\text{KL}}}\left(f_{0\#}\mu_{k-1}\right), \quad k > 0. \tag{11}$$

We note that any local minimizer μ_k of (10) is a critical point of the objective functional, and, therefore, it satisfies:

$$c = \frac{\delta}{\delta\mu}\left[D_{\text{KL}}(\mu\|f_{0\#}\mu_{k-1}) + \eta\mathbb{E}_{\mu}\left[G_k^N\right]\right]\Bigg|_{\mu=\mu_k},$$

where c is a constant (from the constraint $\int_{\mathbb{X}} d\mu(x) = 1$, for $\mu \in \mathcal{P}(\mathbb{X})$, due to which the first variation is defined up to an additive constant). From the above, we get:

$$c = \log\left(\frac{\rho_k(x)}{f_{0\#}\rho_{k-1}(x)}\right) + \eta G_k^N(x),$$

where for any $\ell \in \{0, 1, \ldots\}$, ρ_ℓ is the density function corresponding to the measure μ_ℓ. Therefore, the corresponding recursive update scheme for the density function is given by:

$$\rho_k(x) = c_k\left(f_{0\#}\rho_{k-1}(x)\right)\exp\left(-\eta G_k^N(x)\right), \tag{12}$$

where c_k is the normalization constant. We note that the above is a particle filter formulation, with the horizon cost G_k^N defining the weighting function. Implementable filters are obtained by a Sequential Monte Carlo method, see [21]. We now present the asymptotic stability result for KL-MHE:

Theorem 4: ***(Asymptotic stability of KL-MHE).*** The estimator (10), under Assumptions 1 to 4, is an asymptotically stable observer for the system Σ.

VII. Simulation Results

In this section, we present results from numerical simulations of the estimators studied in this paper. The simulations were performed in MATLAB (version R2017a) on a 2.5 GHz Intel Core i5 processor.

We considered the following nonlinear discrete-time system:

$$\begin{aligned} x_1(k+1) &= x_1(k) + \tau x_2(k), \\ x_2(k+1) &= x_2(k) - \tau\frac{x_1(k)}{1 + |x_1(k)|^2 + |x_2(k)|^2} + w_k, \\ y(k) &= x_1(k) + v_k, \end{aligned}$$

with $\tau = 0.1$, w_k and v_k are i.i.d disturbances, sampled uniformly from the intervals $[-0.1, 0.1]$ and $[-0.15, 0.15]$ respectively.

We first present the simulation results for W_2-MHE. We ran 30 trials of the estimator (8) on the same measurement sequence, with randomly generated initial conditions and over a time horizon of length $T = 100$. The length of the moving horizon was chosen to be $N = 10$. Figure 1 contains the plots of the mean of the estimates along with the true states. The root mean squared error (RMSE) for the mean state estimate sequences were found to be ${z_1}^{\text{RMSE}} = 0.0856$ and ${z_2}^{\text{RMSE}} = 0.0846$ for the estimates of x_1 and x_2, respectively. The average time for computing the state estimate through the minimization (8) using the $fminunc$ function in MATLAB was observed to be $t_{\text{comp}} = 0.012 \pm 0.02s$.

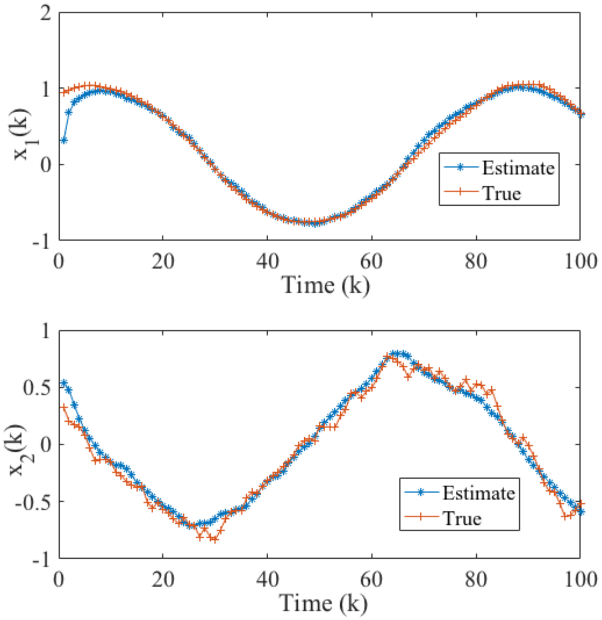

Fig. 1. Mean state estimates from 30 trials of W_2-MHE.

We then implemented the estimator (10) with 30 samples, over a time horizon of length $T = 100$. The length of the moving horizon was chosen to be $N = 10$. Figure 2 contains the plots of the mean of the estimates along with the true states. The root mean squared error (RMSE) for the mean state estimate sequences were found to be ${z_1}^{\text{RMSE}} = 0.1073$ and ${z_2}^{\text{RMSE}} = 0.1144$ for the estimates of x_1 and x_2, respectively. The average run-time for the minimization (10) by a resampling method was observed to be $t_{\text{comp}} = (4.8 \pm 0.4) \times 10^{-4}s$.

In simulation, with 30 samples, we find that the W_2-MHE performs better with respect to the root mean squared error, while the KL-MHE is much faster. The performance of the KL-MHE is determined by the richness of the sample set and effectiveness of the resampling procedure, choices that depend on context and experience. In this manuscript, we did not attempt to investigate improvements in performance with respect to these choices. The performance of W_2-MHE does not necessarily improve with the richness of the sample set, but for systems for which ${\Sigma_T}^{-1}(\mathbf{y}_{0:T})$ is not a singleton, a richer sample set allows for a more complete characterization of the set of feasible estimates.

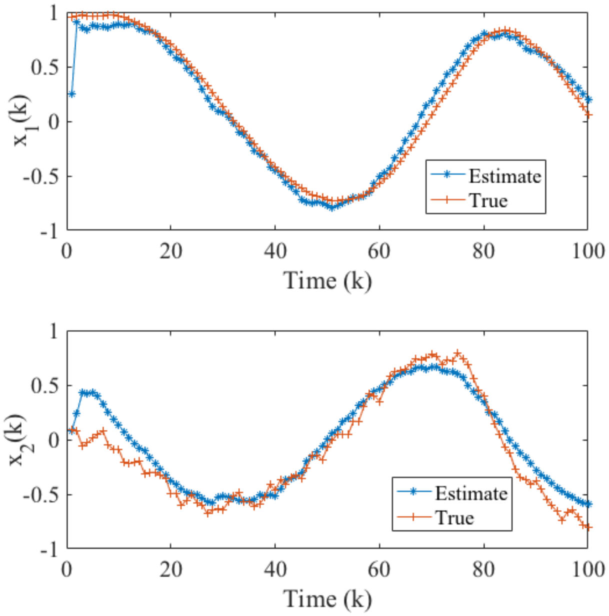

Fig. 2. Mean state estimates from KL-MHE with 30 samples.

VIII. Conclusions and Future Work

In this work, we laid out a unifying distributional framework for moving-horizon estimation. We have clearly established the connection between the classical notion of strong local observability and the stability of moving horizon estimation, for nonlinear discrete-time systems. As an extension to this work, we intend to include distributional constraints in the moving horizon estimation framework. Future work will be devoted to more extensive simulations to more closely characterize the performance of the proposed estimators in practice. Another important consideration in the estimation problem is the rate of convergence of the observer, and it is of interest to obtain convergence rate bounds for the moving-horizon estimators proposed in this paper. A comparison of the rates of convergence for various choices of the (pseudo) metric in the unifying MHE formulation will be also undertaken in our future work.

References

[1] A. Jazwinski, “Limited memory optimal filtering,” *IEEE Transactions on Automatic Control*, vol. 13, no. 5, pp. 558–563, 1968.

[2] C. Rao, J. Rawlings, and D. Mayne, “Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations,” *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

[3] A. Alessandri, M. Baglietto, and G. Battistelli, “Moving-horizon state estimation for nonlinear discrete-time systems: New stability results and approximation schemes,” *Automatica*, vol. 44, no. 7, pp. 1753–1765, 2008.

[4] A. Wynn, M. Vukov, and M. Diehl, “Convergence guarantees for moving horizon estimation based on the real-time iteration scheme,” *IEEE Transactions on Automatic Control*, vol. 59, no. 8, pp. 2215–2221, 2014.

[5] L. Ji, J. Rawlings, W. Hu, A. Wynn, and M. Diehl, “Robust stability of moving horizon estimation under bounded disturbances,” *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3509–3514, 2016.

[6] M. Müller, “Nonlinear moving horizon estimation in the presence of bounded disturbances,” *Automatica*, vol. 79, pp. 306–314, 2017.

[7] W. Hu, “Robust stability of optimization-based state estimation,” *arXiv preprint arXiv:1702.01903*, 2017.

[8] J. Hespanha, D. Liberzon, D. Angeli, and E. Sontag, “Nonlinear norm-observability notions and stability of switched systems,” *IEEE Transactions on Automatic Control*, vol. 50, no. 2, pp. 154–168, 2005.

[9] H. Nijmeijer, “Observability of autonomous discrete time non-linear systems: a geometric approach,” *International Journal of Control*, vol. 36, no. 5, pp. 867–874, 1982.

[10] T. S. Lee, K. P. Dunn, and C. B. Chang, “On observability and unbiased estimation of nonlinear systems,” in *System Modeling and Optimization*, pp. 258–266, Springer, 1982.

[11] J. Tsinias and C. Kitsos, “Observability and state estimation for a class of nonlinear systems,” *arXiv preprint arXiv:1803.08386*, 2018.

[12] F. Santambrogio, “{Euclidean, metric, and Wasserstein} gradient flows: an overview,” *Bulletin of Mathematical Sciences*, vol. 7, no. 1, pp. 87–154, 2017.

[13] G. Peyré, “Entropic approximation of wasserstein gradient flows,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2323–2351, 2015.

[14] A. Alessandri and M. Gaggero, “Moving-horizon estimation for discrete-time linear and nonlinear systems using the gradient and newton methods,” in *IEEE Int. Conf. on Decision and Control*, pp. 2906–2911, IEEE, 2016.

[15] A. Alessandri and M. Gaggero, “Fast moving horizon state estimation for discrete-time systems using single and multi iteration descent methods,” *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4499–4511, 2017.

[16] L. Evans, *Partial differential equations*. Graduate studies in mathematics, Providence (R.I.): American Mathematical Society, 1998.

[17] A. L. Gibbs and F. E. Su, “On choosing and bounding probability metrics,” *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[18] F. Albertini and D. D’Alessandro, “Observability and forward–backward observability of discrete-time nonlinear systems,” *Mathematics of Control, Signals and Systems*, vol. 15, no. 4, pp. 275–290, 2002.

[19] A. Kirsch, *An introduction to the mathematical theory of inverse problems*, vol. 120. Springer Science & Business Media, 2011.

[20] I. Zang and M. Avriel, “On functions whose local minima are global,” *Journal of Optimization Theory and Applications*, vol. 16, no. 3-4, pp. 183–190, 1975.

[21] A. Doucet, N. D. Freitas, and N. Gordon, “Sequential Monte Carlo methods in practice,” Springer, 2001.