2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Finite-Horizon Minimal Realizations for Model Predictive Control of Large-Scale Systems

T. J. Meijer [1,†], S. A. N. Nouwens [1], V. S. Dolk [2], B. de Jager [1], W. P. M. H. Heemels [1]

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992913

***Abstract*— In model predictive control (MPC) for large-scale applications, the computational limitations for on-line optimization often lead to the use of (relatively) short prediction horizons. In this paper, we show that as a result, the controller optimizes over only a fraction of the dynamics of the large-scale system. Based on this observation, which we will formalize, we propose a method to construct reduced-order models of minimal order, by exploiting the system-theoretic concept of finite-horizon observability, that exactly match the response of the large-scale system within a finite horizon. These so-called finite-horizon minimal realizations are used to implement equivalent MPC schemes with reduced computational effort (or the same computational effort but with a larger prediction horizon) without sacrificing accuracy/performance (as the equivalent optimization problem has the same optimizers as the original MPC problem). By computing finite-horizon minimal realizations, we can determine the dynamics as "seen" by the MPC, which can provide useful design insights, in particular, when tuning the prediction horizon. We demonstrate the strengths of our results in a numerical case study.**

***Index Terms*— Model predictive control, projection-based model order reduction, large-scale systems, observability**

I. Introduction

In model predictive control, a constrained open-loop finite-horizon optimal control problem is solved at each sampling instant. By implementing only the first optimal control action and solving the optimal control problem again at the next sampling instance using the updated (estimated) state, feedback is established–also known as receding horizon control [1]. While the on-line optimization is able to account for the presence of constraints, which is one of the main advantages of MPC, solving the optimization problem in real-time is, especially for large-scale systems, computationally expensive. In fact, these computational challenges still form a limiting factor for the broader adoption of MPC in many (fast and/or large-scale) applications [2].

Initially, MPC quickly gained popularity in the process industry, largely due to its conceptual simplicity, constraint-handling capabilities and ease of dealing with multi-input multi-output systems [3]. In recent times MPC has found its way into increasingly complex and, particularly, large-scale applications, such as, e.g., nuclear fusion, temperature-controlled medical treatments and power grid optimization and control. The computational complexity of MPC scales with the state and input dimensions as well as the length of the prediction horizon and number of inequality constraints. Several efforts towards reducing the computational complexity are undertaken, such as using tailored solvers, see, e.g., [4]–[6] for applications where the prediction horizon is the bottleneck, or reducing the number of inequality constraints, see, e.g., [7], [8]. Other approaches use Krylov methods [9], explicit MPC [10], or model reduction techniques, see, e.g., [11], [12]. This paper is concerned with the latter.

We focus specifically on MPC for large-scale systems, i.e., with large state dimension, and relatively few performance channels/outputs. As mentioned before, the on-line computational complexity of an MPC scheme scales (poorly) with the state dimension. This is particularly true for sparse formulations, in which the states are included in the decision variables and, hence, the number of required operations, when using conventional interior-point methods, scales cubically in the state dimension [6]. We can exploit specific structure in the sparse formulation to achieve complexity that grows linearly with the prediction horizon, see, e.g., [6]. As an alternative to the sparse formulation, one can also eliminate the states from the decision variables [1], which results in the dense formulation. The required computational effort for solving the dense formulation is independent of the number of states, however, we still require the computation of a particular solution, which scales quadratically with the state dimension [4]. As a result, we are typically only able to consider a short prediction horizon in large-scale applications. We will see that, as a consequence, only a small part of the dynamics is actually considered within the finite prediction horizon. In practice, it is often sufficient to consider such a short horizon since our MPC still reacts to the "unseen" dynamics via the receding-horizon (or feedback) principle, however, we do not optimize for or anticipate these dynamics in the MPC optimization. This observation, which we formalize, has useful consequences for MPC design and facilitates a reduction of the on-line computational complexity without changing the MPC feedback law. In particular, regarding the latter, we construct so-called *finite-horizon minimal realizations*, which are reduced-order models of minimal order that exactly match the output response over the finite prediction horizon for any initial condition, input sequence and disturbance sequence. In other words, we propose a reduction technique that removes the dynamics that do not play a role in the finite-horizon optimal control control problem. By formulating the MPC problem for these reduced-order models, we implement more efficient, *equivalent* MPC schemes without changing the solution to the finite-horizon optimal control problem and, thereby, the overall MPC scheme (and thus its feasibility, stability and performance properties). The finite-horizon minimal realizations can also be used to understand the dynamics actually considered by the MPC, for

[1] Tomas Meijer, Sven Nouwens, Bram de Jager and Maurice Heemels are with the Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. {t.j.meijer; s.a.n.nouwens; a.g.de.jager; m.heemels}@tue.nl

[2] Victor Dolk is with ASML, De Run 6665, 5504 DT Veldhoven, The Netherlands. victor.dolk@asml.com

† Corresponding author: T. J. Meijer.

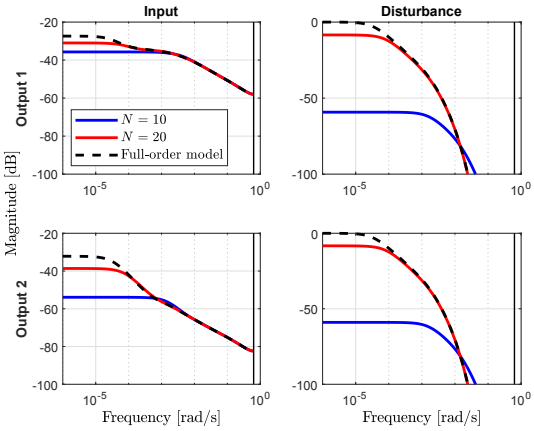


Fig. 1: Bode magnitude plot (of the N-minimal realizations computed in Section IV) visualizing the dynamics accounted for in the optimal control problem for prediction horizon N.

instance, by generating Bode plots of the "seen" dynamics, see, Fig. 1. This provides valuable insights for tuning the controller, in particular, the prediction horizon.

II. Preliminaries

Notation. The sets of real and non-negative natural numbers are denoted, respectively, $\mathbb{R} = (-\infty, \infty)$ and $\mathbb{N} = \{0, 1, 2, \ldots\}$. We denote $\mathbb{N}_{\geqslant n} = \{n, n+1, n+2, \ldots\}$ and $\mathbb{N}_{[n,m]} = \{n, n+1, n+2, \ldots, m\}$ for $n, m \in \mathbb{N}$. The symbol I is an identity matrix of appropriate dimensions while I_n denotes an n-by-n identity matrix. For a set of matrices $\{A_i\}_{i\in\mathcal{N}}$, $\mathcal{N} \subseteq \mathbb{N}$, we denote, for all $n, m \in \mathcal{N}$, $\prod_{i\in\mathbb{N}_{[n,m]}} A_i = A_n A_{n+1} \ldots A_m$ when $m \geqslant n$, and $\prod_{i\in\mathbb{N}_{[n,m]}} A_i = I$ when $m < n$. The Kronecker product of $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{r\times s}$ is denoted by $A \otimes B \in \mathbb{R}^{nr\times ms}$. For a vector $x \in \mathbb{R}^n$ (matrix $A \in \mathbb{R}^{n\times m}$), $x_{\mathcal{I}}$ ($A_{\mathcal{I},\mathcal{J}}$) denotes a vector (matrix) consisting of the rows (and columns of A) with indices in $\mathcal{I} \subset \mathbb{N}_{[1,n]}$ (and $\mathcal{J} \subset \mathbb{N}_{[1,m]}$, respectively). Finally, $y \leftarrow x$ means assign y the value of x.

A. Model predictive control

Consider the discrete-time system represented by

$$\begin{aligned} x_{k+1} &= Ax_k + B_u u_k + B_w w_k, \\ z_k &= Cx_k + D_u u_k + D_w w_k, \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$, $u_k \in \mathbb{R}^{n_u}$, $w_k \in \mathbb{R}^{n_w}$ and $z_k \in \mathbb{R}^{n_z}$ denote, respectively, the state, control input, disturbance input and (performance) output, at time instant $k \in \mathbb{N}$. We focus on an MPC formulation for (1) in which the cost function and constraints are assumed to be formulated exclusively in terms of inputs u and performance outputs z.[1] At each time instant $k \in \mathbb{N}$, the following finite-horizon optimal control problem is solved:

$$\begin{aligned} \min_{\{u_{i|k}\}_{i\in\mathcal{N}_N}} \ & \frac{1}{2}\sum_{i\in\mathcal{N}_N} \ell(z_{i|k}, u_{i|k}), \\ \text{s.t.} \ \ & x_{i+1|k} = Ax_{i|k} + B_u u_{i|k} + B_w w_{i|k}, \ \ i \in \mathcal{N}_{N-1}, \\ & z_{i|k} = Cx_{i|k} + D_u u_{i|k} + D_w w_{i|k}, \ \ i \in \mathcal{N}_N, \\ & x_{0|k} = x_k, \\ & Mz_{i|k} + Ju_{i|k} \leqslant c, \qquad\qquad i \in \mathcal{N}_N, \end{aligned} \tag{2}$$

[1] If this is not the case, extra (combinations of) states can be added to z.

where $\mathcal{N}_N := \mathbb{N}_{[0,N]}$ contains the time instances along the prediction horizon of length $N \in \mathbb{N}_{\geqslant 1}$ and $\{x_{i|k}\}_{i\in\mathcal{N}_N}$, $\{z_{i|k}\}_{i\in\mathcal{N}_N}$ and $\{u_{i|k}\}_{i\in\mathcal{N}_N}$ in (2) denote, respectively, the predicted states, outputs and future inputs. For instance, $x_{i|k}$ denotes a prediction of x_{k+i} made at time $k \in \mathbb{N}$. We assume that the (estimated) state x_k is available at time $k \in \mathbb{N}$ as well as (an estimate of) the disturbance sequence $\{w_{i|k}\}_{i\in\mathcal{N}_N}$. The function $\ell\colon \mathbb{R}^{n_z} \times \mathbb{R}^{n_u} \to \mathbb{R}_{\geqslant 0}$ is the (quadratic) stage cost, given by $\ell(z, u) = z^\top Q z + 2z^\top S u + u^\top R u$ with tuning parameters $Q \in \mathbb{S}^{n_z}_{\geqslant 0}$, $S \in \mathbb{R}^{n_z\times n_u}$ and $R \in \mathbb{S}^{n_u}_{\succ 0}$. Finally, the matrices $M \in \mathbb{R}^{n_c\times n_z}$ and $J \in \mathbb{R}^{n_c\times n_u}$ and the vector $c \in \mathbb{R}^{n_c}$ are used to capture the (polyhedral) constraints.[2] Solving (2) leads to a unique optimal input sequence denoted by $\{u^\star_{i|k}\}_{i\in\mathcal{N}_N}$ of which our MPC applies the first entry to the actual system, i.e., $u_k = u^\star_{0,k} \in \mathbb{R}^{n_u}$, and solves (2) again at time $k+1$, initialized with a new (estimated) state x_{k+1}, leading to the so-called receding horizon principle [1].

B. Projection-based model reduction

Projection-based model reduction methods function by projecting the state of (1) and its dynamics onto, respectively, lower-dimensional test and trial spaces, denoted $\mathcal{V} \subseteq \mathbb{R}^{n_x}$ and $\mathcal{W} \subseteq \mathbb{R}^{n_x}$, respectively. This results in a reduced-order model [12]

$$\hat{x}_{k+1} = \hat{A}\hat{x}_k + \hat{B}_u u_k + \hat{B}_w w_k, \tag{3a}$$

$$\hat{z}_k = \hat{C}\hat{x}_k + D_u u_k + D_w w_k, \tag{3b}$$

$$\hat{A} = W^\top A V, \ [\hat{B}_u, \hat{B}_w] = W^\top [B_u, B_w], \ \hat{C} = CV. \tag{3c}$$

The reduced state at time $k \in \mathbb{N}$ is denoted by $\hat{x}_k \in \mathbb{R}^{n_r}$, with (typically) $n_r \ll n_x$, and the output of the reduced-order system at time $k \in \mathbb{N}$ is denoted $\hat{z}_k \in \mathbb{R}^{n_z}$. The above assumes that the projection matrices $W, V \in \mathbb{R}^{n_x\times n_r}$ are bi-orthonormal, i.e., $W^\top V = I$. The projection matrices are such that $\operatorname{Im} V = \mathcal{V}$ and $\operatorname{Im} W = \mathcal{W}$. Through carefully designing $\mathcal{V}$ and $\mathcal{W}$, we ensure that the relevant/desired dynamics are captured by the reduced-order model (3). We consider Galerkin projections, where $\mathcal{V} = \mathcal{W}$ and $V = W$.

III. Problem definition

The computational complexity of (implicit) MPC is inherently large due to the required optimization that has to be carried out in real-time. In this paper, we focus on a class of large-scale systems for which the computational complexity of MPC is particularly high, due to the large state dimension.

Assumption 1. *The state dimension is significantly larger than the number of performance outputs throughout the finite prediction horizon, i.e., $n_x \gg (N+1)n_z$.*

Assumption 1 encompasses many relevant applications with (relatively) few outputs and a short prediction horizon, the latter often necessary due to the computational burden becoming excessive for longer prediction horizons.

A consequence of Assumption 1, as we will discuss in more detail later on, is that only a fraction of the dynamics is actually being considered by the controller, see Fig. 1 for

[2] Although, we adopt a quadratic stage cost and polyhedral constraints, our results do not depend on this. For any cost function and constraints, even if (2) does not admit a (set of) minimizer(s), our method will yield a reduced-order MPC scheme with the same (possibly empty set of) minimizer(s).

an illustration of this. In other words, *the computed optimal control action is independent of a large part of the dynamics*. Given this observation, the following problem is considered in this work: For the described setting and a given prediction horizon N, construct a reduced-order model (3) of minimal order such that the optimal control action $\hat{u}^\star_{0|k}$ computed by solving the associated reduced-order optimal control problem

$$\begin{aligned}\min_{\{\hat{u}_{i|k}\}_{i\in\mathcal{N}_N}} \ & \frac{1}{2}\sum_{i\in\mathcal{N}_N}\ell(\hat{z}_{i|k},\hat{u}_{i|k}), \qquad (4)\\ \text{s.t.} \ \ \hat{x}_{i+1|k} &= \hat{A}\hat{x}_{i|k}+\hat{B}_u\hat{u}_{i|k}+\hat{B}_w w_{i|k}, \ \ i\in\mathcal{N}_{N-1},\\ \hat{z}_{i|k} &= \hat{C}\hat{x}_{i|k}+D_u\hat{u}_{i|k}+D_w w_{i|k}, \ \ i\in\mathcal{N}_N,\\ \hat{x}_{0|k} &= V^\top x_k,\\ M\hat{z}_{i|k}+J\hat{u}_{i|k} &\leqslant c, \qquad\qquad i\in\mathcal{N}_N,\end{aligned}$$

exactly coincides with the optimal solution to (2), i.e., $\hat{u}^\star_{0|k} = u^\star_{0|k}$, for all $x_k\in\mathbb{R}^{n_x}$, $\{w_{i|k}\}_{i\in\mathcal{N}_N}$ with $w_{i|k}\in\mathbb{R}^{n_w}$, $k\in\mathbb{N}$.

The constructed reduced-order models can be exploited to implement an *equivalent* MPC scheme that can be solved more efficiently for the same prediction horizon N due to the lower state dimension or, as a result of the reduced computational burden, we can implement a longer prediction horizon N, which might be desired from a control performance perspective. We address the above problem using the system-theoretic notion of finite-horizon observability, which is discussed in more detail in the next section.

IV. Finite-horizon minimal realizations

The optimal control problem in (2) only depends on the solution to (1) on a finite horizon. We exploit system-theoretic properties of the system to construct reduced-order models (3) of minimal order, that are exact on this finite horizon.

Definition 1. *The reduced-order system* (3) *is said to be an N-minimal realization of* (1)*, if, for any initial condition $x_0\in\mathbb{R}^{n_x}$, input sequence $\{u_k\}_{k\in\mathcal{N}_N}$ and disturbance sequence $\{w_k\}_{k\in\mathcal{N}_N}$, with $u_k\in\mathbb{R}^{n_u}$, $w_k\in\mathbb{R}^{n_w}$, $k\in\mathcal{N}_N$, it holds, for all $k\in\mathcal{N}_N$ and with $\hat{x}_0 = V^\top x_0$, that*[3]

$$\begin{aligned}0 = \hat{C}\hat{A}^k\hat{x}_0 - CA^k x_0 + \qquad (5)\\ \sum_{i\in\mathbb{N}_{[0,k-1]}}\left(\hat{C}\hat{A}^{k-1-i}\begin{bmatrix}\hat{B}_u & \hat{B}_w\end{bmatrix} - CA^{k-1-i}\begin{bmatrix}B_u & B_w\end{bmatrix}\right)\begin{bmatrix}u_i\\ w_i\end{bmatrix},\end{aligned}$$

and there is no lower-order model for which this also holds.

Definition 1 means that, for any initial condition $x_0\in\mathbb{R}^{n_x}$, input sequence $\{u_k\}_{k\in\mathcal{N}_N}$ and disturbance sequence $\{w_k\}_{k\in\mathcal{N}_N}$, the outputs of (1) and (3), starting from the projected initial condition $\hat{x}_0 = V^\top x_0$, remain equal for (at least) the duration of the finite horizon. Unlike traditional minimality [13], (5) also accounts for initial conditions. At first glance, N-minimal realizations may seem similar to partial realizations [14], however, partial realizations do not consider initial conditions. In fact, any N-minimal realization is a partial realization, since (5) also holds for $x_0 = 0$, but the converse is not true. An important observation is that, for any N-minimal realization of (1), the optimal control problems (2) and (4) yield the same solution since the constraints and cost depend only on the output z and input u.

[3]Note that D_u and D_w are not affected by the projection performed in (3) and, hence, the direct feedthrough terms are cancelled out in (5).

A. Finite-horizon observability

The construction of our reduced-order models relies on the system-theoretic concept of finite-horizon observability, which we state in terms of the prediction horizon $N\in\mathbb{N}$.

Definition 2. *The N-step unobservable subspace $\mathcal{UO}_N\subseteq\mathbb{R}^{n_x}$ of system* (1) *is the set of all states, which can not be distinguished from the zero state in the output for zero input and disturbance within at least N steps, i.e.,*

$$\mathcal{UO}_N := \left\{x\in\mathbb{R}^{n_x}\,\middle|\, CA^i x = 0, \text{ for all } i\in\mathcal{N}_N\right\}.$$

The system (1) *is said to be N-observable, if $\mathcal{UO}_N = \{0\}$.*

It is well known that $\mathcal{UO}_N = \ker O_N$, where $O_N\in\mathbb{R}^{(N+1)n_z\times n_x}$ is the N-step observability matrix of (1), i.e.,

$$O_N^\top := \begin{bmatrix}C^\top & (CA)^\top & \dots & (CA^N)^\top\end{bmatrix}. \qquad (6)$$

Note that O_N has $(N+1)n_z$ rows, which implies that, in N steps, at least $n_x-(N+1)n_z$ states are unobservable, i.e., $\dim\mathcal{UO}_N\geqslant n_x-(N+1)n_z$. We state below that N-minimality can be verified through checking N-observability.

Proposition 1. *Suppose the system* (1) *satisfies Assumption 1. Then, any N-observable system of the form* (3) *satisfying* (5) *with $\hat{x}_0 = V^\top x_0$, for any initial condition $x_0\in\mathbb{R}^{n_x}$, input sequence $\{u_k\}_{k\in\mathbb{N}}$ and disturbance sequence $\{w_k\}_{k\in\mathbb{N}}$, $u_k\in\mathbb{R}^{n_u}$, $w_k\in\mathbb{R}^{n_w}$, $k\in\mathcal{N}_N$, is an N-minimal realization of* (1)*.*

While traditional minimality is satisfied, if the system is both (infinite-horizon) controllable and observable, our definition of N-minimality is satisfied when only N-observability holds due to the consideration of the initial condition in (5). For the sake of simplicity, we adopt the following.[4]

Assumption 2. *The system* (1) *is (infinite-horizon) observable, i.e., $\mathcal{UO}_{k-1} = \{0\}$ for all $k\in\mathbb{N}_{\geqslant n_x}$.*

B. Model reduction

Using Proposition 1, we will construct N-minimal realizations by removing the N-step unobservable states from our model. We achieve this through the so-called *N-step observability staircase form* [15]. We adopt, without loss of generality, the following assumption (we can always project z onto a lower-dimensional output space such that this holds).

Assumption 3. *The matrix C in* (1) *has full row rank.*

Any system (1) satisfying Assumptions 2-3, can be transformed using a unitary similarity transformation $\bar{x} = T^\top x$, with $T^\top T = I$, to the *N-step observability staircase form*, using the method from [15] (see Algorithm 1 below)[5], i.e.,

$$\left[\begin{array}{c}CT\\ \hline T^\top AT\end{array}\right] = \qquad (7)$$

$$\left[\begin{array}{ccccccc}C_0 & 0 & 0 & \dots & 0 & 0 & 0\\ \hline A_{0,0} & A_{0,1} & 0 & \dots & 0 & 0 & 0\\ A_{1,0} & A_{1,1} & A_{1,2} & \dots & 0 & 0 & 0\\ A_{2,0} & A_{2,1} & A_{2,2} & \dots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ A_{N-1,0} & A_{N-1,1} & A_{N-1,2} & \dots & A_{N-1,N-1} & A_{N-1,N} & 0\\ A_{N,0} & A_{N,1} & A_{N,2} & \dots & A_{N,N-1} & A_{N,N} & \star\\ \star & \star & \star & \dots & \star & \star & \star\end{array}\right],$$

[4]Our approach extends to unobservable systems by replacing (7) with the full observability staircase form in [15].

[5]The method in [15] computes the full observability staircase form. We reduce computations by stopping after N steps to obtain the form in (7).

where $C_0 \in \mathbb{R}^{n_z \times m_0}$ and $A_{i,i+1} \in \mathbb{R}^{m_i, m_{i+1}}$, $i \in \mathcal{N}_{N-1}$, have full column rank, i.e., $\operatorname{rank} C_0 = m_0$ and $\operatorname{rank} A_{i,i+1} = m_{i+1}$ with $m_{i+1} \leqslant m_i \leqslant \ldots \leqslant m_0 = n_z$. Moreover, $\star$ denotes an arbitrary block of appropriate dimensions. For the sake of completeness, we state the method from [15], slightly adapted to compute the N-step observability staircase form (7), below.

Algorithm 1 (N-step observability staircase form [15])**.** *Let* $\bar{S} = \begin{bmatrix} C^\top & A^\top \end{bmatrix}$, $T = I_{n_x}$, $r = 0$, $i = 1$, $q = n_z$ *and* $k = 0$.

1) *For* $j \in \{1, \ldots, \min\{q, n_x - r - 1\}\}$*:*
   a) *Let* $Q_j = \mathcal{H}_{r+j}(\bar{S}_{\{1,\ldots,n_x\},k+j})$.[6]
   b) $\bar{S} \leftarrow Q_j \bar{S} \operatorname{diag}(I_{n_z}, Q_j)$*;* $T \leftarrow T Q_j$.
2) *Let* $\bar{R} = \bar{S}_{\{r+1,r+2,\ldots,r+q_1\},\{k+1,k+2,\ldots,k+q\}}$*, with* $q_1 = \min\{q, n_x - r\}$*, and compute its singular value decomposition* $\bar{R} = U_i \Sigma_i V_i^\top$.
3) *Let* $\bar{\Sigma}_i$ *be* Σ_i *with all diagonal entries smaller than* $\epsilon_M \max\{\|A\|_2, \|C\|_2\}$*, where* ϵ_M *is the relative machine precision, set to zero and let* $m_i = \operatorname{rank} \bar{\Sigma}_i$.
4) *If* $m_i = 0$*, then* $i \leftarrow i - 1$ *and stop; else go to 5).*
5) *If* $r + m_i = n_x$*, then* $r = n_x$ *and stop; else, go to 6).*
6) *If* $m_i = q$*, go to 7); else,* $T \leftarrow T \operatorname{diag}(I_r, U_i, I)$.
7) $r \leftarrow r + m_i$*,* $k \leftarrow k + q$*,* $q = m_i$*,* $i \leftarrow i + 1$*. If* $i = N - 1$ *stop; else, go to 2).*

The N-step observability matrix $O_N T \in \mathbb{R}^{(N+1)n_z \times n_x}$, with O_N as in (6), evaluated for the pair $(T^\top A T, CT)$ yields

$$O_N T = \begin{bmatrix} C_0 & 0 & \ldots & 0 & 0 \ldots 0 \\ \star & C_0 A_{0,1} & \ldots & 0 & 0 \ldots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \ddots \vdots \\ \star & \star & \ldots & C_0 \prod_{i \in \mathbb{N}_{[1,N]}} A_{i-1,i} & 0 \ldots 0 \end{bmatrix}. \quad (8)$$

We prove some useful properties based on (8) below.

Lemma 1. *Given* $N \in \mathbb{N}$*, consider the system* (1) *satisfying Assumptions 1-3, the matrix* T*, which transforms* (1) *to its* N*-step observability staircase form* (7)*. Then, the last* $n_x - \bar{m}_N$ *columns of* T*, with* $\bar{m}_N \coloneqq \sum_{i \in \mathcal{N}_N} m_i \leqslant (N+1) n_z$*, form a basis for the* N*-step unobservable subspace of* (1)*, i.e.,* $\operatorname{rank} T \begin{bmatrix} 0 & I_{n_x - \bar{m}_N} \end{bmatrix}^\top = n_x - \bar{m}_N$ *and*

$$\mathcal{UO}_N = \ker O_N = \operatorname{Im} T \begin{bmatrix} 0 & I_{n_x - \bar{m}_N} \end{bmatrix}^\top. \quad (9)$$

Lemma 1 leads us to partition T according to

$$T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}, T_1 \in \mathbb{R}^{n_x \times \bar{m}_N}, T_2 \in \mathbb{R}^{n_x \times n_x - \bar{m}_N}, \quad (10)$$

such that, in view of (9), $\mathcal{UO}_N = \operatorname{Im} T_2$. We effectively truncate (8) by projecting onto $\mathcal{V} = \mathcal{W} = \operatorname{Im} T_1$. To do so, we set $V = W = T_1$ in (3), which yields

$$\left[\begin{array}{c|c} D_u & D_w \\ \hline \hat{B}_u & \hat{B}_w \end{array}\right] = \left[\begin{array}{c|cccccc} D_u & D_w & C_0 & 0 & 0 & \ldots & 0 & 0 \\ \hline & & A_{0,0} & A_{0,1} & 0 & \ldots & 0 & 0 \\ & & A_{1,0} & A_{1,1} & A_{1,2} & \ldots & 0 & 0 \\ \hat{B}_u & \hat{B}_w & A_{2,0} & A_{2,1} & A_{2,2} & \ldots & 0 & 0 \\ & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & A_{N-1,0} & A_{N-1,1} & A_{N-1,2} & \ldots & A_{N-1,N-1} & A_{N-1,N} \\ & & A_{N,0} & A_{N,1} & A_{N,2} & \ldots & A_{N,N-1} & A_{N,N} \end{array}\right]. \quad (11)$$

The obtained model (3) with (11) satisfies (5) and there is no model of lower order that satisfies also (5).

[6]Here, $\mathcal{H}_i(y)$, $y = (y_1, y_2, \ldots, y_n)$, $i \in \mathbb{N}_{[1,n]}$, denotes a Householder annihilator, which is such that $\mathcal{H}_i(y) y = (y_1, y_2, \ldots, y_{i-1}, \bar{y}_i, 0, \ldots, 0)$ with some $\bar{y}_i \in \mathbb{R}$. Let $\mathcal{I} = \{i+1, i+2, \ldots, n\}$ and $u = (0, \ldots, 0, \operatorname{sgn}(y_i)\|y_{\mathcal{I}}\|_2 + y_i, y_{\mathcal{I}})$, then, $\mathcal{H}_i(y) = I - 2uu^\top / \|u\|_2^2$.

Theorem 1. *Given* $N \in \mathbb{N}$*, consider the system* (1) *satisfying Assumptions 1-3, the matrix* T*, partitioned according to* (10)*, which transforms the system to its* N*-step observability staircase form* (7)*, and the reduced-order model* (3)*, which is constructed using* $\mathcal{V} = \mathcal{W} = \operatorname{Im} T_1$ *by setting* $W = V = T_1$*. The resulting reduced-order model* (3) *with* (11) *is an* N*-minimal realization of* (1)*.*

Theorem 1 means that, by taking $\mathcal{V} = \mathcal{W}$ with $\dim \mathcal{V} = \bar{m}_N$ and $\mathcal{V} \oplus \mathcal{UO}_N = \mathbb{R}^{n_x}$ in (3), we obtain a reduced-order model of minimal dimensions that, with initial condition $\hat{x}_0 = V^\top x_0$, yields the same response as the full-order system during the finite horizon for any initial condition x_0, input sequence $\{u\}_{k \in \mathcal{N}_N}$ and disturbance sequence $\{w_k\}_{k \in \mathcal{N}_N}$.

V. Application to Model Predictive Control

The optimal solution to (2) only depends on the N-observable part of the system, which forms only a fraction of the states of the system. Another way of looking at this is that the MPC scheme (2) implicitly performs a model reduction step. We perform this model reduction step explicitly by formulating our MPC scheme based on the finite-horizon minimal realizations obtained, in the previous section, by truncating the N-step observability staircase form (7). This has direct on-line computational benefits as the 'implicit' on-line model reduction step is carried out explicitly and off-line, thereby saving costly on-line computations.

A. Sparse and dense formulations

We formulate (4) as the following quadratic program (QP), referred to as the *sparse formulation* [4], [5]:

$$\begin{aligned} \min_{q_k} \quad & \frac{1}{2} q_k^\top H q_k, \\ \text{s.t.} \quad & F q_k = f_k, \text{ and } G q_k \leqslant g, \end{aligned} \quad (12)$$

where $q_k \in \mathbb{R}^{(N+1)(n_u+n_z)+N n_r}$ and

$$q_k = \begin{bmatrix} \hat{u}_{0|k}^\top & \hat{z}_{0|k}^\top & \hat{x}_{1|k}^\top & \hat{u}_{1|k}^\top & \hat{z}_{1|k}^\top & \ldots & \hat{x}_{N|k}^\top & \hat{u}_{N|k}^\top & \hat{z}_{N|k}^\top \end{bmatrix}^\top,$$

$$H = \begin{bmatrix} R & S^\top & & & 0 \\ S & Q & & & 0 \\ & & \multicolumn{3}{c}{\multirow{2}{*}{$I_N \otimes \begin{bmatrix} 0 & 0 & 0 \\ 0 & R & S^\top \\ 0 & S & Q \end{bmatrix}$}} \\ 0 & 0 & \multicolumn{3}{c}{} \end{bmatrix} \quad G = \begin{bmatrix} J & M & & & 0 \\ 0 & 0 & \multicolumn{3}{c}{\multirow{2}{*}{$I_N \otimes \begin{bmatrix} J & M \end{bmatrix}$}} \\ & & \multicolumn{3}{c}{} \end{bmatrix},$$

$$F = \begin{bmatrix} D_u & -I & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ \hat{B}_u & 0 & -I & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{C} & D_u & -I & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{A} & \hat{B}_u & 0 & -I & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \hat{C} & D_u & -I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & \hat{A} & \hat{B}_u & 0 & -I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \hat{C} & D_u & -I \end{bmatrix},$$

$$f_k = - \begin{bmatrix} \hat{C} V^\top x_k + D_w w_{0|k} \\ \hat{A} V^\top x_k + \hat{B}_w w_{0|k} \\ D_w w_{1|k} \\ \hat{B}_w w_{1|k} \\ \vdots \\ D_w w_{N-1|k} \\ \hat{B}_w w_{N-1|k} \\ D_w w_{N|k} \end{bmatrix}, \; g = \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix}.$$

The above formulation contains the (reduced-order) states, inputs and outputs in the decision variable q_k. We can eliminate the states and outputs from the decision variables using the null-space method [4], which decomposes q_k as

$$q_k = Z \xi_k + \hat{q}_k, \quad (13)$$

where the columns of $Z \in \mathbb{R}^{((N+1)(n_u+n_z)+Nn_r)\times(N+1)n_u}$ span the null-space of F, i.e., $FZ = 0$ and $\operatorname{rank} Z = (N+1)n_u$. The vector $\hat{q}_k \in \mathbb{R}^{(N+1)(n_u+n_z)+Nn_r}$ is any particular solution to $F\hat{q}_k = f_k$. The choice of Z and $\hat{q}_k$ is not unique. Here, we take $Z = \begin{bmatrix} Z_1 & Z_2 & \dots & Z_N \end{bmatrix}$ with $Z_i \in \mathbb{R}^{((N+1)(n_u+n_z)+Nnr)\times n_u}$ given by

$$\begin{aligned} Z_i^\top = \big[& 0_{n_u\times(i-1)(n_u+n_r+n_z)}\ I\ D_u^\top\ \hat{B}_u^\top\ 0_{n_u\times n_u}\ (\hat{C}\hat{B}_u)^\top \dots \\ & (\hat{A}^{N-i}\hat{B}_u)^\top\ 0_{n_u\times n_u}\ (\hat{C}\hat{A}^{N-i}\hat{B}_u)^\top \big], \end{aligned} \tag{14}$$

which consists of delayed impulse responses. The particular solution is given, for all $k \in \mathbb{N}$, by

$$\hat{q}_k = \begin{bmatrix} \hat{u}_{0|k}^\top\ \hat{z}_{0|k}^\top\ \hat{x}_{1|k}^\top\ \hat{u}_{1|k}^\top\ \hat{z}_{1|k}^\top \dots \hat{x}_{N|k}^\top\ \hat{u}_{N|k}^\top\ \hat{z}_{N|k}^\top \end{bmatrix}^\top, \tag{15}$$

with $\hat{x}_{i|k}$ and $\hat{z}_{i|k}$ satisfying, for $i \in \mathcal{N}_N$, $\hat{x}_{i+1|k} = \hat{A}\hat{x}_{i|k} + \hat{B}_u\hat{u}_{i|k} + \hat{B}_w w_{i|k}$ and $\hat{z}_{i|k} = \hat{C}\hat{x}_{i|k} + D_u\hat{u}_{i|k} + D_w w_{i|k}$, respectively, with $\hat{x}_{0|k} = V^\top x_k$, $u_{i|k} = 0$ and the given (estimate of the) disturbance sequence $\{w_{i|k}\}_{i\in\mathcal{N}_N}$. See [4] for further details on the computation of Z and $\hat{q}_k$. Substituting (13) into (12) yields an equivalent QP with significantly fewer decision variables, also known as the *dense formulation*,

$$\begin{aligned} \min_{\xi_k}\quad & \frac{1}{2}\xi_k^\top Z^\top H Z \xi_k + \hat{q}_k^\top H Z \xi_k, \\ \text{s.t.}\quad & GZ\xi_k \leqslant g - G\hat{q}_k, \end{aligned} \tag{16}$$

with $\xi_k \in \mathbb{R}^{(N+1)n_u}$. Although the dense QP (16) no longer contains the states as decision variables, computing Z based on the reduced-order system rather than the full-order system reduces the off-line computational cost. More importantly, the particular solution $\hat{q}_k$ for the reduced-order system is significantly cheaper to compute than for the full-order system, which yields a reduction in on-line computational complexity, as we will detail in the next section.

B. Computational complexity

We recall that the number of outputs over the horizon provides an upper bound on the reduced state dimension, i.e., $n_r = \bar{m}_N \leqslant (N+1)n_z$. Table I summarizes the computational complexity associated with solving the sparse and dense formulations (12) and (16), respectively, using conventional interior-point methods, as well as the computation of the particular solution $\hat{q}_k$ in (15). These computational complexities are listed for the MPC scheme constructed based on the full-order model (1) and the N-minimal realizations from Section IV, where the latter is obtained by replacing n_x by the upper bound on n_r and analyzing the asymptotic behaviour. For both sparse schemes, we exploit the sparsity to achieve more favourable scaling in N, see [6], for the full-order formulation this yields linear complexity in N whereas the reduced-order model, since its state dimension grows with N too, achieves quartic scaling, i.e. N^4. The complexity also depends on the number of constraints, which is not affected by our approach and, hence, is not included in Table I.

	Full-order model (1)	N-minimal (3), (11)
Sparse (12)	$O(N(n_x+n_u)^3)$	$O(N(Nn_z+n_u)^3)$
Dense (16)	$O(N^3n_u^3)$	$O(N^3n_u^3)$
$\hat{q}_k$ in (15)	$O(N(n_x^2+n_xn_w))$	$O(N^3n_z^2+N^2n_zn_w)$

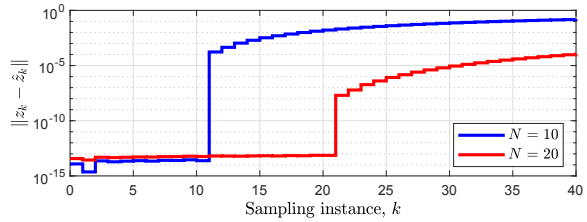
TABLE I: Comparison of computational cost.

Noting that $n_x \gg (N+1)n_z$ (see Assumption 1), a significant reduction in computational effort is achieved by using the N-minimal realizations. This is even true for the dense formulation, where solving the QP is independent of the state dimension, due to the required computation of a particular solution, which is more efficient for the reduced-order models. In fact, the N-minimal realizations result in a computational complexity that is *independent* of the underlying state dimension, while, as discussed in the previous section, the solutions to (2) and (4) (and, hence, (12) and (16)) are the same. Hence, using the finite-horizon minimal realizations computed in Section IV, we can efficiently solve (2). Using the N-minimal realizations, it is tractable to solve (2) for longer N with similar computational cost, as illustrated in a numerical example below.

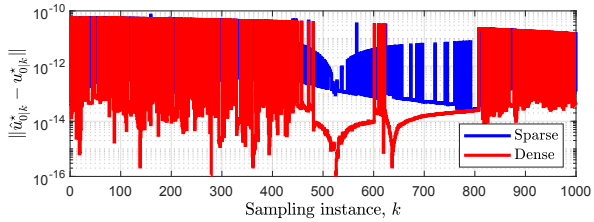
VI. Numerical case study

We demonstrate our methods on a three-dimensional thermal system (discretized in space and time), satisfying Assumptions 2-3, with $n_x = 1000$ states, $n_z = 2$ outputs, $n_u = 1$ input and $n_w = 1$ disturbance for $N = 10$ and $N = 20$, which thus also satisfies Assumption 1. The system is actuated by applying a heat load to part of the system's surface while the ambient temperature acts as a disturbance. We aim to control and impose constraints on the average temperature on the heated part of the surface and the temperature at the center of this heated surface. The N-minimal realizations, computed as in Section IV, are of order $n_r = 22$ and $n_r = 42$, respectively, which coincides with the bound in Assumption 1. Bode magnitude plots of both N-minimal realizations are shown in Fig. 1 together with the full-order model. It can be seen that, for low N, only the dynamics at higher frequencies are accurately "seen" by the controller and, as N increases, also lower-frequency dynamics become "visible". Hence, the N-minimal realizations and corresponding Bode (magnitude) plots can be very insightful for tuning, in particular, the prediction horizon N. In Fig. 2a, we simulate the response of the N-minimal realizations and plot the error between their outputs and the output of the full-order model. Theoretically, the response should be equal for at least N steps since the results of Theorem 1 apply. We observe that the error remains at the numerical noise level, which is due to finite machine precision, for N steps after which it increases, however, this is acceptable as this error does not affect the responses within the MPC's finite horizon.

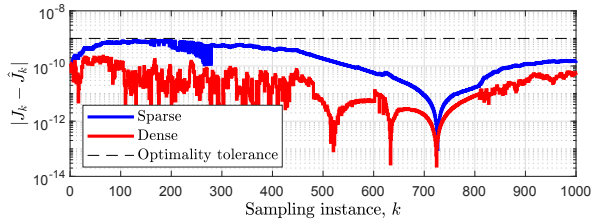
Next, we implement both the sparse QP formulation in (12) and the dense QP formulation in (16) and compare the computed control action to the one computed by solving (2) using the full-order model. The results are shown in Fig. 2b, where we see that both formulations yield a solution that is acceptably close to the full-order solution. Theoretically the computed control actions should be equal by Theorem 1, however, in practice numerical noise is introduced due to finite machine precision. Interestingly, the sparse formulation is significantly closer to $u_{0|k}^\star$ in some parts of the range $k \in [500, 800]$. This is due to numerical differences between both formulations. To see this, we look at Fig. 2c, where we show, for both schemes, the difference between the achieved optimal cost and the cost achieved when using the full-order system as a function of the sampling instance. For both the sparse and dense formulation, the achieved cost is within the



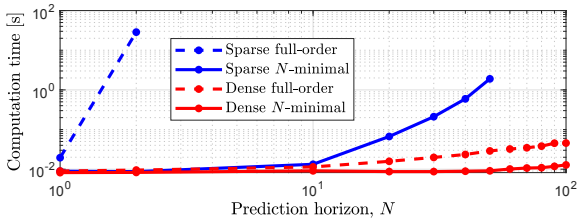
(a) Output error between the N-minimal realizations and the full system for a step input and $\hat{x}_0 = V^\top x_0$.



(b) Mismatch in computed optimal control actions $u^\star_{0|k}$, using the full-order model, and $\hat{u}^\star_{0|k}$, using sparse (12) and dense (16) formulations for the N-minimal realization ($N = 20$).



(c) Mismatch between the achieved cost J_k, using the full-order model, and $\hat{J}_k$, using sparse (12) and dense (16) formulations for the N-minimal realization ($N = 20$). For a fair comparison, we add $\frac{1}{2}\hat{q}_k^\top H \hat{q}_k$ to the cost of the dense formulation.



(d) Average computation time per time step $k \in \mathbb{N}$ to solve the sparse formulations (12) and dense formulations (16) (including computation of $\hat{q}_k$ in (15) for the full and reduced-order models.

Fig. 2: Numerical case study results.

solver's optimality tolerance centered around the cost achieved by the full-order scheme. In fact, the residual mismatch can be further reduced by tightening the tolerances.

The computational effort required for solving the different QP problems, using MATLAB's built-in quadprog function, can be seen in Fig. 2d. We directly see that the sparse formulation based on the full-order model is exceptionally costly due to the large state dimension. As a result, this formulation could only be simulated for $N \in \{1, 2\}$, however, this already shows that the complexity rapidly grows with N and that a significant reduction is achieved using the N-minimal realization (which we could tractably simulate for $N \in [1, 50]$). Both dense formulations are significantly faster than their sparse counterparts and the N-minimal realization outperforms the traditional dense formulation for higher N. The computation times of both dense formulations in Fig. 2d also include the computation of the particular solution. The computation time of the reduced-order dense formulation is (almost) constant for $N \leqslant 40$, this is due to the fact that for small horizons, solving the dense QP takes significantly more time than computing the particular solution. As N grows large, the order n_r of the N-minimal realization approaches n_x and, hence, a significant reduction in computational complexity can only be achieved when Assumption 1 holds.

VII. Conclusions

In this paper, we have presented a method to construct finite-horizon minimal realizations that can be used to formulate *equivalent* MPC schemes for large-scale applications, in particular, with few performance outputs, but with significantly reduced on-line computation time. Equivalence is meant here in the sense of both schemes having the same optimizers for the input sequence. We have demonstrated these methods in a numerical case study and shown that reduced computational complexity is indeed achieved. Similar reduction techniques based on finite-horizon controllability for applications with few actuators are the focus of future work. It is also interesting to study what consequences the observations in this paper may have on state observer design for large-scale MPC, since only a fraction of the states now needs to be estimated.

References

[1] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model predictive control: Theory, computation and design*, 2nd ed. Nob Hill Publishing, 2020.

[2] R. R. Negenborn and J. M. Maestre, "Distributed model predictive control: An overview and roadmap of future research opportunities," *IEEE Control Syst.*, vol. 34, no. 4, pp. 87–97, 2014.

[3] D. Q. Mayne, "Model predictive control: Recent developments and future promise," *Automatica*, vol. 50, pp. 2967–2986, 2014.

[4] J. Yang, T. J. Meijer, V. S. Dolk, B. de Jager, and W. P. M. H. Heemels, "A system-theoretic approach to construct a banded null basis to efficiently solve MPC-based QP problems," in *58th IEEE Conf. Decis. Control*, 2019, pp. 1410–1415.

[5] J. L. Jerez, E. C. Kerrigan, and G. A. Constantinides, "A sparse and condensed QP formulation for predictive control of LTI systems," *Automatica*, vol. 48, no. 5, pp. 999–1002, 2012.

[6] Y. Wang and S. Boyd, "Fast model predictive control using online optimization," *IEEE Trans. Autom. Control*, vol. 18, no. 2, pp. 267–278, 2010.

[7] S. A. N. Nouwens, B. de Jager, M. M. Paulides, and W. P. M. H. Heemels, "Constraint removal for MPC with performance preservation and a hyperthermia cancer treatment case study," in *60th IEEE Conf. Decis. Control*, 2021, pp. 4103–4108.

[8] M. Jost, G. Pannocchia, and M. Mönnigmann, "Online constraint removal: Accelerating MPC with a Lyapunov function," *Automatica*, vol. 57, pp. 164–169, 2015.

[9] A. Malyshev, R. Quirynen, and A. Knyazev, "Preconditioned Krylov iterations and condensing in interior point MPC method," *IFAC-PapersOnLine*, vol. 51, no. 20, pp. 388–393, 2018.

[10] A. Bemporad, F. Borrelli, and M. Morari, "Model predictive control based on linear programming–The explicit solution," *IEEE Trans. Autom. Control*, vol. 47, no. 12, pp. 1974–1985, 2002.

[11] S. Hovland, K. Willcox, and J. T. Gravdahl, "MPC for large-scale systems via model reduction and multiparametric quadratic programming," in *45th IEEE Conf. Decis. Control*, 2006, pp. 3418–3423.

[12] P. Benner, S. Gugercin, and K. Willcox, "A survey of projection-based model reduction methods for parametric dynamical systems," *SIAM Review*, vol. 57, no. 4, pp. 483–531, 2015.

[13] J. P. Hespanha, *Linear systems theory*. Princeton University Press, 2018.

[14] P. Benner and V. I. Sokolov, "Partial realizations of descriptor systems," *Syst. Control Lett.*, vol. 55, no. 11, pp. 929–938, 2006.

[15] A. Varga, "Numerically stable algorithm for standard controllability form determination," *Electron. Lett*, vol. 17, no. 2, pp. 74–75, 1981.