2019 18th European Control Conference (ECC)
Napoli, Italy, June 25-28, 2019

Causally Stable Approximation of Optimal Maps in Maximal Value Constrained Least-Squares Optimization*

Omer Tanovic[1] and Alexandre Megretski[1]

Abstract— **In this paper, we consider a problem of designing discrete-time systems which are optimal in frequency-weighted least squares sense subject to a maximal output amplitude constraint. In such problems, the optimality conditions do not provide an explicit way of generating the optimal output as a real-time implementable transformation of the input, due to causal instability of the resulting dynamical equations and sequential nature in which criterion function is revealed over time. On the other hand, under some mild conditions, the optimal system has exponentially fading memory which suggests existence of arbitrarily good finite-latency approximations. In this paper, we extend the method of balanced truncation for linear systems to the class of nonlinear models with weakly contractive operators. We then propose a causally stable finite-latency nonlinear system which returns high-quality approximations to the optimal map. The proposed system is obtained by a careful truncation of an infinite dimensional state space representation of the optimal system, as suggested by the derived generalization of the balanced truncation algorithm.**

I. INTRODUCTION

Convex quadratic programs (QP) with box constraints (or inequality constraints in general) are ubiquitous in science and engineering problems, and some examples are: modeling of the ocean circulation [1], support vector machines [2], constrained linear quadratic optimal control [3], etc. Various methods have been proposed for solving box constrained QP in a finite-dimensional setting: active set methods (see [4] and references therein), gradient projection and conjugate gradients [5], Newton iteration [6], primal-dual methods [7], etc. Such methods commonly rely on computer-aided optimization solvers and require non-negligible computation power, which makes them unfavorable in applications that have strict power budget.

The infinite-dimensional bound-constraint quadratic programs are even more computationally demanding. An important instance is the infinite-horizon linear quadratic regulation problem (LQR) with bounded control. This problem is mostly addressed approximately, where model predictive control (MPC) has probably been the most popular method for approximately solving infinite-horizon constrained LQR. Such MPC schemes rely on replacing infinite-horizon with a receding (i.e. finite) one, where, in general, an easier finite-dimensional optimization problem is resolved at every time instance [8]-[11].

*This work was supported by the National Science Foundation under award number 1743938.

[1]Authors are with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA {otanovic, ameg}@mit.edu

Another instance of the infinite-dimensional setup emerges when one wants to design discrete-time systems which are optimal in the sense of some frequency-weighted least squares criterion subject to maximal output amplitude constraints. In particular, such optimization problems serve to represent a number of peak-to-average-power ratio (PAPR) reduction objectives which are of significant importance in modern communication systems [12]-[14]. It is known for such optimization problems that, in general, the optimality conditions do not provide an explicit way of generating the optimal output as a real-time implementable transformation of the input. This is due to instability of the resulting dynamical equations as well as sequential nature in which criterion function is revealed over time Therefore, at each time instance, the knowledge of the whole history of the input signal should be known ahead of time in order to calculate the current sample of the optimal output signal. Due to difficulties in obtaining an explicit optimal solution, receding horizon optimization, i.e. model predictive control, appears to be a natural way of addressing these problems. Unfortunately, the high cost associated with MPC computations at every time step makes it unfavorable in power and time-sensitive applications such as those in signal processing for communication systems.

In [15], it was shown that, under some mild assumptions, the optimal system has exponentially fading memory. A causal and stable nonlinear system was proposed which, under an L1 dominance assumption about the equation coefficients, returns high-quality approximations to the optimal solution. The L1 dominance is a very strong condition and potentially diminishes the practical usefulness of the result.

In this paper, we propose a real-time realizable algorithm which returns high-quality approximations to the optimal map. The algorithm exploits the optimality conditions and is realized as a causally stable finite-latency nonlinear discrete-time system, and is allowed to look ahead at the input signal over a finite horizon (and is, therefore, of finite latency). A bound on the approximation error is derived by extending the method of balanced truncation for linear systems to a class of nonlinear models which include the optimal system under consideration. The algorithm does not rely on any special assumptions about the least squares criterion, except for convexity, and, therefore, provides a much stronger result than the one derived in [15]. Fading memory of the optimal system justifies the finite horizon assumption and suggests that such approach can serve as a cheaper alternative to standard MPC-based algorithms, since it does not rely on resolving an optimization problem at every time instant.

II. NOTATION AND TERMINOLOGY

$\mathbb{R}, \mathbb{Z}, \mathbb{N}$ are the usual sets of real, integer, and positive integer numbers. For an element w of a (real) Hilbert space H, $|w|$ denotes the norm. $\ell(X)$ is the real vector space of all functions $x : \mathbb{Z} \to X$, interpreted as *discrete-time (DT) signals*, with $x(t)$ used for the value of x at $t \in \mathbb{Z}$. For $x \in \ell(X)$, the *L2 norm* $|x| \in [0, \infty]$ is defined by $|x| = \left(\sum_t |x(t)|^2\right)^{\frac{1}{2}}$, where $|x(t)|$ is the norm in X. $\ell^2(X)$ is the subset of finite energy signals from $\ell(X)$, treated as a Hilbert space, with the norm $|x|$ defined above. $E = \{e_i\}_{i=-\infty}^{\infty} \subset \ell^2(\mathbb{R})$ such that $e_i(t) = 1$ for $t = i$ and $e_i(t) = 0$ otherwise, is the standard orthonormal basis in $\ell^2(\mathbb{R})$. We use shorthand notation $W = W(\omega)$ to denote the Fourier transform of signal $w \in \ell^2(\mathbb{R})$, instead of the standard notation $W = W(e^{j\omega})$. Systems are viewed as operators $\ell^2(X) \to \ell^2(Y)$. $\mathbf{G}w$ denotes the response of system $\mathbf{G}$ to signal w (even when $\mathbf{G}$ is not linear), and the *series composition* $\mathbf{K} = \mathbf{QG}$ of systems $\mathbf{Q}$ and $\mathbf{G}$ is the system mapping w to $\mathbf{Q}(\mathbf{G}w)$.

For a bounded linear operator $\mathbf{A} : \ell^2(X) \to \ell^2(Y)$, $\mathbf{A}' : \ell^2(Y) \to \ell^2(X)$ denotes the adjoint operator of $\mathbf{A}$. The matrix of $\mathbf{A}$, in the standard bases $\{e_i\}_{i=-\infty}^{\infty}$ and $\{\tilde{e}_i\}_{i=-\infty}^{\infty}$ of $\ell^2(X)$ and $\ell^2(Y)$, respectively, is denoted as $A = (A_{ij})_{i,j=-\infty}^{\infty}$. In this paper, $\mathbf{A}$ and A will be used interchangeably to denote the same operator. For any bounded (not necessarily linear) operator $\mathbf{T} : \ell^2(X) \to \ell^2(Y)$, the operator norm $\|\mathbf{T}\|$ of $\mathbf{T}$ is defined as $\|\mathbf{T}\| = \sup_{x \in \ell^2(X), x \neq 0} |\mathbf{T}x|/|x|$.

For a positive real number r, function $\mathrm{sat}_r : \mathbb{R} \to [-r, r]$ is defined by

$$\mathrm{sat}_r(\xi) = \begin{cases} \xi, & |\xi| < r \\ r\xi/|\xi|, & |\xi| \geq r \end{cases}.$$

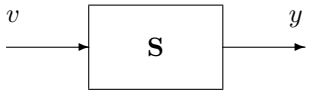
Similarly, operator $\mathrm{Sat}_r : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ is defined by

$$y = \mathrm{Sat}_r(x) \Leftrightarrow y(t) = \mathrm{sat}_r(x(t)), \forall t \in \mathbb{Z}.$$

III. PROBLEM FORMULATION

The problem setup presented in this section is (with some minor modifications) taken from [15].

We aim to optimize and implement efficiently discrete-time signal processing systems with scalar input v and scalar output y:



where the output $y = \mathbf{S}v$ is expected to be optimal, in the sense of minimizing a certain objective defined in terms of input v. For a fixed $r > 0$, let $\Omega_r = \{w \in \ell^2(\mathbb{R}) : |w(t)| \leq r, \ \forall t \in \mathbb{Z}\}$. Let $\alpha : \mathbb{R} \to \mathbb{C}$ and $\beta : \mathbb{R} \to \mathbb{C}$ be trigonometric polynomials mapping $\omega \in \mathbb{R}$ to $\alpha(\omega) \geq \epsilon > 0$ and $\beta(\omega)$, respectively. For every discrete-time signal $v \in \ell^2(\mathbb{R})$, the scalar signal $y = \mathbf{S}v \in \ell^2(\mathbb{R})$ should have samples $|y(t)| \leq r$, and minimize the functional

$$J_{\alpha,\beta}(v, y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \alpha(\omega)|Y(\omega)|^2 d\omega - \frac{1}{\pi} \int_{-\pi}^{\pi} \mathrm{Re}\{Y(\omega)'\beta(\omega)V(\omega)\} d\omega \quad (1)$$

where $V = V(\omega)$ and $Y = Y(\omega)$ are the Fourier transforms of v and y, respectively. Therefore, we are trying to solve the time-domain-value-constrained frequency-weighted least squares optimization problem

$$\min_y J_{\alpha,\beta}(v, y), \quad \text{subject to } y \in \Omega_r. \quad (2)$$

Let us denote this optimization problem as $\mathbb{P} = \mathbb{P}(\alpha, \beta, \Omega_r, v)$. It is clear that $\mathbb{P}$ is a convex infinite-dimensional quadratic problem with box constraints, which is feasible (see [15]) and has a unique solution due to strict positivity of α. Let $\mathbf{T}_\alpha$ and $\mathbf{T}_\beta$ be the finite unit sample response LTI systems with frequency responses $\alpha(\omega)$ and $\beta(\omega)$, respectively. The necessary and sufficient conditions of optimality of $\mathbb{P}$ can be written as (see [15] or [16]):

$$y = \mathrm{Sat}_r(y - \mathbf{T}_\alpha y + \mathbf{T}_\beta v). \quad (3)$$

Moreover, with $w = \mathbf{T}_\beta v$ and $\mathbf{H} = \mathbf{I} - \mathbf{T}_\alpha$, the above optimality condition can be written as

$$y = \mathrm{Sat}_r(\mathbf{H}y + w). \quad (4)$$

Let $H = H(\omega)$ and $h = h(t)$ be the frequency response and unit sample response of $\mathbf{H}$, respectively, and let T be the order of the trigonometric polynomial $\alpha = \alpha(\omega)$. It follows that $h(-t) = h(t)$ for all $t \in \mathbb{Z}$, and $h(t) = 0$ for all $|t| > T$. The optimal condition (4) can now be written sample-wise as

$$y(t) = \mathrm{sat}_r\left(\sum_{\tau=-T}^{T} h(\tau)y(t+\tau) + w(t)\right). \quad (5)$$

Due to the strict convexity of $\mathbb{P}$, and hence uniqueness of the optimal solution, equation (5) defines a system which maps input signal w into output signal y. We denote this system as $\mathbf{S}^*$ and refer to it as "the optimal system" or "the optimal map", in the rest of this paper. It can be seen from (5) that, in general, the optimal system $\mathbf{S}^*$ is nonlinear and noncausal. Moreover, the optimality condition (5) is not attractive as a description of a real-time implementable system $\mathbf{S}$ mapping w to the optimal y. Intuitively, it is clear that a necessary condition for the existence of a finite-latency system, which is a good approximation (e.g., in ℓ_∞ or H_∞ sense) of the optimal system $\mathbf{S}^*$, is that $\mathbf{S}^*$ possesses some type of 'near-finite' memory. That is, one hopes that system $\mathbf{S}^*$ for any two input signals that are close in the recent past and future, but not necessarily close in the remote past and future, yields present outputs which are close. Indeed, it has been shown in [15] that if $\mathbf{H}$ has strictly positive frequency response then system $\mathbf{S}^*$ has exponentially fading memory (for the exact definitions, statements and proofs see [15]). In the following sections, we show that, with careful truncation and adequate non-linear stability analysis, the optimal system $\mathbf{S}^*$ can be approximated arbitrarily well by a finite-latency system.

IV. MAIN RESULTS

In this section, we first state and prove an extension of the classical method of balanced truncation to a class of nonlinear models with weakly contractive operators. We use this result to show that a certain nonlinear model yields high-quality approximations to the optimal map of problem (2).

A. Preliminaries

We first give some preliminary definitions, results and assumptions which will be used in the following sections.

Definition 1: Let $\mathbf{S} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a bounded linear operator with matrix (in the standard basis) $S = (s_{ij})_{i,j=-\infty}^{\infty}$. We say that $\mathbf{S}$ is a banded operator if there exists a positive integer N such that $s_{ij} = 0$ for all $|i-j| > N$. Minimal integer N for which this is true is called the bandwidth of $\mathbf{S}$, in which case we say that $\mathbf{S}$ is an N-banded operator.

Definition 2: Let $\mathbf{S} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a bounded linear operator with matrix (in the standard basis) $S = (s_{ij})_{i,j=-\infty}^{\infty}$. We say that $\mathbf{S}$ is a Laurent operator if there exists $f \in \ell^2(\mathbb{R})$ such that $s_{ij} = f(i-j)$ for all $i, j \in \mathbb{Z}$. Such f is called the symbol of S.

Lemma 3: Let $\mathbf{S} : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be an N-banded operator such that $\|\mathbf{S}\| < 1$. For $\gamma \in (0,1]$ let $\mathbf{D}_\gamma \in \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a bounded linear operator such that $(\mathbf{D}_\gamma w)(t) = \gamma^{|t|} w(t)$ for all $w \in \ell^2(\mathbb{R})$, and let $\mathbf{S}_\gamma : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be uniquely defined by $\mathbf{S}_\gamma \mathbf{D}_\gamma = \mathbf{D}_\gamma \mathbf{S}$. There exists $\gamma_0 \in (0,1)$ such that $\|\mathbf{S}_\gamma\| < 1$ for all $\gamma \in (\gamma_0, 1]$.

Proof: Operator $\mathbf{S}_\gamma$ is bounded. This immediately follows from Hölder's inequality:

$$\|\mathbf{S}_\gamma\|^2 \le \|\mathbf{S}_\gamma\|_1 \|\mathbf{S}_\gamma\|_\infty \le \left(\gamma^{-N} \sum_{t=-N}^{N} |f(t)|\right)^2 < \infty.$$

Let $g : (0,1] \to (0,\infty)$ be defined by $\gamma \mapsto g(\gamma) = \|\mathbf{S}_\gamma\|$. Clearly, $g(1) < 1$. Let $F = F(\omega)$ be the Fourier transform of f. Since $\|\mathbf{S}\| < 1$ then $|F(\omega)| < 1$ for all $\omega \in [0, 2\pi)$, and hence $|f(t)| < 1$ for all t. From the definition of the operator norm, we have that

$$\|\mathbf{S}_\gamma\| = \sup_{|u|=|v|=1} |(u, \mathbf{S}_\gamma v)| \ge \sup_{u,v \in E} |(u, \mathbf{S}_\gamma v)| \ge \sup_{i,j \in \mathbb{Z}} |s_\gamma^{i,j}|,$$

where $S_\gamma = (s_\gamma^{i,j})$ is the matrix of $\mathbf{S}_\gamma$ in the standard basis of $\ell^2(\mathbb{R})$. By definition,

$$s_\gamma^{i,j} = \begin{cases} \gamma^{|i|-|j|} f(i-j), & |i-j| \le N, \\ 0, & \text{otherwise} \end{cases}.$$

Hence, for all $|t| \le N$, there exist, large enough, $|i|$ and $|j|$ such that $s_\gamma^{i,j} = \gamma^{-|t|} f(t)$. Therefore,

$$\|\mathbf{S}_\gamma\| \ge \max_{|t| \le N} \gamma^{-|t|} |f(t)|.$$

Let $c = \min\{|f(0)|, \min_{0<|t|\le N} |f(t)|^{\frac{1}{|t|}}\}$. Then $c \le \epsilon$ and $g(c) > 1$, so, by the continuity of g, there exists $\gamma_0 \in (c,1)$ such that $g(\gamma_0) = 1$ and $g(\gamma) < 1$ for all $\gamma \in (\gamma_0, 1)$. This concludes the proof. ∎

In the rest of this paper, and without loss of generality, we assume that operator $\mathbf{H}$ is a contraction, that is, $||\mathbf{H}|| < 1$ or, equivalently, $|H(\omega)| < 1$ for all $\omega \in [0, 2\pi)$ (even more, we can assume that $0 < H(\omega) < 1$). Indeed, let $\alpha_0 > 0$ such that $\alpha(\omega) < \alpha_0$ for all $\omega \in [0, 2\pi)$ (such α_0 exists since α is a continuous function of ω). Let $\tilde{J}_{\alpha,\beta} = \frac{1}{\alpha_0} J_{\alpha,\beta}$. Optimization problem $\mathbb{P}$ is now equivalent to the one of minimizing $\tilde{J}_{\alpha,\beta}$ subject to $\|y\|_\infty \le r$. We denote this problem as $\tilde{\mathbb{P}}$. The necessary and sufficient condition of optimality of $\tilde{\mathbb{P}}$ is now given as

$$y = \mathrm{Sat}_r(\tilde{\mathbf{H}} y + \tilde{w}),$$

where $\tilde{\mathbf{H}} = \mathbf{I} - \mathbf{T}_{\alpha/\alpha_0}$ and $\tilde{w} = \mathbf{T}_{\beta/\alpha_0} v$. This implies that $\tilde{H}(\omega) = 1 - \frac{1}{\alpha_0}\alpha(\omega) \in (0,1)$, and, therefore, optimal problem $\mathbb{P}$ is equivalent to the one for which operator $\mathbf{H}$ is a contraction.

B. Generalized Balanced Truncation Theorem

We now state and prove a result on upper bounds on error of approximating a certain class of nonlinear systems by appropriately chosen reduced order models, similar to those of the classical balanced truncation algorithm for linear systems.

Let X, V and Y be Hilbert spaces. In general, X can be infinite dimensional. Consider now systems $\mathbf{G} : \ell^2(V) \to \ell^2(Y)$ and $\tilde{\mathbf{G}} : \ell^2(V) \to \ell^2(Y)$ described by the following state space models

$$\mathbf{G} : x(t+1) = \varphi(Ax(t) + Bv(t)), \quad y(t) = Cx(t), \quad (6)$$

$$\tilde{\mathbf{G}} : \tilde{x}(t+1) = \Theta\varphi(A\tilde{x}(t) + Bv(t)), \quad \tilde{y}(t) = C\tilde{x}(t), \quad (7)$$

where $A : X \to X$, $B : V \to X$, $C : X \to V$ and $\Theta : X \to X$ are bounded linear operators, Θ is a projection, i.e., $\Theta^2 = \Theta$, and $\varphi : X \to X$ is a diagonal operator in the standard basis in X. The following theorem gives an upper bound on error of approximating $\mathbf{G}$ with $\tilde{\mathbf{G}}$.

Theorem 4: Let $\sigma_1, \sigma_2 > 0$ be positive real numbers and $P = P' > 0$, $Q = Q' > 0$ be positive definite self-adjoint operators satisfying the following Lyapunov inequalities

$$P - APA' \ge \frac{1}{\sigma_1^2} BB', \qquad Q - A'QA \ge \frac{1}{\sigma_2^2} C'C. \quad (8)$$

Let Θ and φ satisfy the following conditions

$$(P^{-1} - Q)(I - \Theta) = 0, \qquad P^{-1} + Q \ge \Theta(P^{-1} + Q)\Theta, \quad (9)$$

$$(\varphi(u) + \varphi(v))' P^{-1} (\varphi(u) + \varphi(v)) \le (u+v)' P^{-1} (u+v), \quad (10)$$

$$(\varphi(u) - \varphi(v))' Q (\varphi(u) - \varphi(v)) \le (u-v)' Q (u-v), \forall u, v \in X, \quad (11)$$

Then

$$||\mathbf{G} - \tilde{\mathbf{G}}|| \le 2\sigma_1\sigma_2.$$

Proof:

Step 1: Show that P and Q satisfy the following dissipation inequalities:

$$\sigma_1^2|v|^2 \geq (Ax+Bw)'P^{-1}(Ax+Bw) - x'P^{-1}x, \quad (12)$$

$$-\frac{1}{\sigma_2^2}|Cx|^2 \geq (Ax)Q(Ax) - x'Qx, \quad \forall v \in V, \forall x \in X. \quad (13)$$

Indeed, dissipation inequality (13) immediately follows from the second inequality in (8). The first inequality in (8), after a congruence transformation by $P^{-1/2}$ and some algebraic manipulation, is equivalent to

$$\left\| \begin{bmatrix} P^{\frac{1}{2}}A'P^{-\frac{1}{2}} \\ \frac{1}{\sigma_1}B'P^{-\frac{1}{2}} \end{bmatrix} \right\| \leq 1. \quad (14)$$

Therefore, inequality

$$\left\| \begin{bmatrix} P^{-\frac{1}{2}}AP^{\frac{1}{2}} & \frac{1}{\sigma_1}P^{-\frac{1}{2}}B \end{bmatrix} \right\| \leq 1, \quad (15)$$

holds as well. After some straightforward algebraic manipulation, (15) is equivalent to

$$\begin{bmatrix} P^{\frac{1}{2}}A'P^{-1}AP^{\frac{1}{2}} & \frac{1}{\sigma_1}P^{\frac{1}{2}}A'P^{-1}B \\ \frac{1}{\sigma_1}B'P^{-1}AP^{\frac{1}{2}} & \frac{1}{\sigma_1^2}B'P^{-1}B \end{bmatrix} \leq \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \quad (16)$$

Identity matrices on the right-hand side of (16) are assumed to be of appropriate (possibly different) dimensions. Inequality (16), after a congruence transformation by the block diagonal matrix $D = \text{diag}(P^{-\frac{1}{2}}, \sigma_1 I)$, is equivalent to the dissipation inequality (12).

Step 2: Show that the following dissipation inequality holds:

$$\sigma(v, Cx - C\tilde{x}) \geq V(x^+, \tilde{x}^+) - V(x, \tilde{x}), \quad (17)$$

where

$$\sigma(v,e) = 4\sigma_1^2|v|^2 - \sigma_2^{-2}|e|^2$$
$$V(x,\tilde{x}) = (x+\tilde{x})'P^{-1}(x+\tilde{x}) + (x-\tilde{x})'Q(x-\tilde{x})$$

To show this, consider the following state space model of the error system $\mathbf{G} - \tilde{\mathbf{G}}$ mapping w to $e = y - \tilde{y}$:

$$\begin{aligned} x^+ &= \varphi(Ax+Bw), \\ \tilde{x}^+ &= \Theta\varphi(A\tilde{x}+Bw), \\ e &= Cx - C\tilde{x}. \end{aligned} \quad (18)$$

The positive definite quadratic form $V = V(x,\tilde{x})$ can be re-written as follows:

$$V(x,\tilde{x}) = x'(P^{-1}+Q)x + 2x'(P^{-1}-Q)\tilde{x} + \tilde{x}'(P^{-1}+Q)\tilde{x}. \quad (19)$$

By expanding $V(x^+, \tilde{x}^+)$, we have the sequence of inequalities as shown in (20), where shorthand notation $z = Ax+Bw$ and $\tilde{z} = A\tilde{x} + Bw$ was used. The first inequality used (4), the second inequality used (10)-(11), and the last inequality used (12)-(13). This implies that (17) holds and, furthermore, that $\|\mathbf{G} - \tilde{\mathbf{G}}\| \leq 2\sigma_1\sigma_2$. This concludes the proof. ■

Theorem 4 is a generalization of the well known result on upper bounds of H-infinity error for the exact implementation of the balanced truncation algorithm for linear systems [17]. Indeed, let $\varphi = I$ and let (A, B, C) be the balanced realization of system $\mathbf{G}$, where the controllability and observability gramians W_c and W_o, respectively, satisfy $W_c = W_o = \Sigma > 0$ for a block diagonal balanced gramian Σ. Let σ be the smallest Hankel singular value of $\mathbf{G}$, and let $\Sigma = \text{diag}(\Sigma_0, \sigma I)$ with block diagonal Σ_0. In the classical balanced truncation method, one aims at truncating states of $\mathbf{G}$ that correspond to the lower-right block σI of Σ. Therefore, the projection matrix Θ is defined as $\Theta = \text{diag}(I, 0)$, where the dimension of the zero matrix 0 corresponds to that of the σI submatrix. It now follows that the dissipation inequalities (8) are satisfied (with equality) for $\sigma_1 = \sigma_2 = \sqrt{\sigma}$ and $P = Q = \frac{1}{\sigma}\Sigma$. Expressions in (9)-(11) hold by the definitions of P, Q and Θ. Therefore, the balanced truncation error bound follows from Theorem 4, i.e., the upper bound on H-infinity error of approximating $\mathbf{G}$ with $\tilde{\mathbf{G}}$ is 2σ.

C. Main Theorem

In this section we first propose a finite-latency nonlinear discrete-time system that approximates the optimal solution of (2) with arbitrary precision. We then give an upper bound on the approximation error.

As before, we assume that the optimal system $\mathbf{S}^*$ maps signal $w \in \ell^2(\mathbb{R})$ to $y \in \ell^2(\mathbb{R})$, as defined by

$$y(t) = \text{sat}_r\left(\sum_{\tau=-T}^{T} h(\tau)y(t+\tau) + w(t) \right). \quad (21)$$

For a given integer $m > T$, let system $\mathbf{M}_m : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R}^{2m+1})$, mapping w to $\hat{v}$, be defined by

$$\hat{v}(t) = [w(t-m+1) \quad w(t-m+2) \quad \dots \quad w(t+m+1)]^T. \quad (22)$$

Let system $\hat{\mathbf{T}}_m : \ell^2(\mathbb{R}^{2m+1}) \to \ell^2(\mathbb{R})$ be defined by the following state space model

$$\hat{x}(t+1) = \text{sat}_r(\hat{A}\hat{x}(t) + \hat{v}(t)), \quad \hat{y}(t) = \hat{C}\hat{x}(t), \quad (23)$$

where $\hat{x}(t), \hat{v}(t) \in \mathbb{R}^{2m+1}$, and matrices $\hat{A} = (\hat{A}_{ij})_{i,j=1}^{2m+1} \in \mathbb{R}^{(2m+1)\times(2m+1)}$ and $\hat{C} = (\hat{C}_j)_{j=1}^{2m+1} \in \mathbb{R}^{1\times(2m+1)}$ are

$$\begin{aligned} V(x^+,\tilde{x}^+) &= \varphi(z)'(P^{-1}+Q)\varphi(z) + 2\varphi(z)'(P^{-1}-Q)\Theta\varphi(\tilde{z}) + \varphi(\tilde{z})'\Theta(P^{-1}+Q)\Theta\varphi(\tilde{z}) \\ &\leq \varphi(z)'(P^{-1}+Q)\varphi(z) + 2\varphi(z)'(P^{-1}-Q)\varphi(\tilde{z}) + \varphi(\tilde{z})'(P^{-1}+Q)\varphi(\tilde{z}) \\ &= (\varphi(z)+\varphi(\tilde{z}))'P^{-1}(\varphi(z)+\varphi(\tilde{z})) + (\varphi(z)-\varphi(\tilde{z}))'Q(\varphi(z)-\varphi(\tilde{z})) \\ &\leq (z+\tilde{z})'P^{-1}(z+\tilde{z}) + (z-\tilde{z})'Q(z-\tilde{z}) \quad \leq \quad 4\sigma_1^2|w| - \sigma_2^{-1}|Cx - C\tilde{x}| - V(x,\tilde{x}) \end{aligned} \quad (20)$$

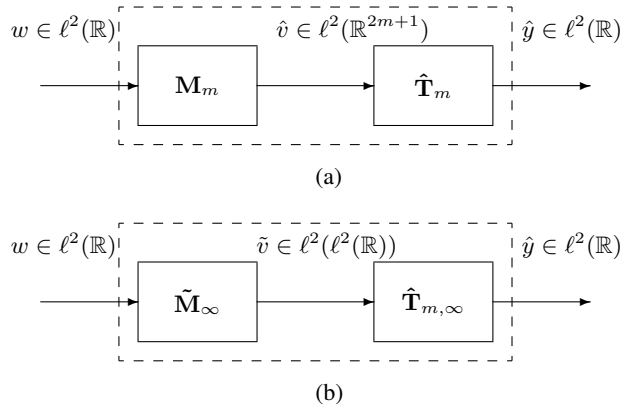


Fig. 1. Equivalent representations of the approximate system $\hat{\mathbf{S}}_m$: a) $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_m\mathbf{M}_m$ (state-space model $\hat{\mathbf{T}}_m$ is finite dimensional) and b) $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_{m,\infty}\tilde{\mathbf{M}}_\infty$ (state-space model $\hat{\mathbf{T}}_{m,\infty}$ is infinite dimensional).

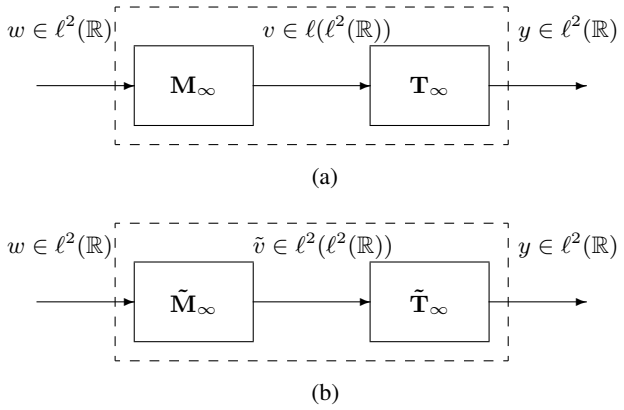


Fig. 2. Equivalent representations of the optimal system $\mathbf{S}^*$: a) $\mathbf{S}^* = \mathbf{T}_\infty\mathbf{M}_\infty$ (subsystem $\mathbf{M}_\infty$ is unbounded) and b) $\mathbf{S}^* = \tilde{\mathbf{T}}_\infty\tilde{\mathbf{M}}_\infty$ (subsystem $\tilde{\mathbf{M}}_\infty$ is bounded)

defined by $\hat{A}_{ij} = h(i-j+1), \forall i,j \in \{1,\dots,2m+1\}$, $\hat{C}_k = 1$ for $k = m$ and $\hat{C}_k = 0$ otherwise.

Systems $\hat{\mathbf{T}}_m$ and $\mathbf{M}_m$ are clearly time-invariant systems, where the former is nonlinear and causally stable while the latter is linear and non-causal but of finite latency (equal to $m+1$). Let system $\hat{\mathbf{S}}_m : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ mapping w to $\hat{y} = \hat{\mathbf{S}}_m w$ be defined as a series interconnection $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_m\mathbf{M}_m$ of $\mathbf{M}_m$ and $\hat{\mathbf{T}}_m$, see Fig. 1(a).

The following theorem establishes that the L2-induced gain of the error of approximating the optimal system $\mathbf{S}^*$ with the finite-latency system $\hat{\mathbf{S}}_m$ is bounded by $\epsilon = c\rho^m$ for some $c > 0$ and $\rho \in (0,1)$.

Theorem 5: There exist $\rho \in (0,1)$ and $c > 0$ such that $\|\mathbf{S}^* - \hat{\mathbf{S}}_m\| \le c\rho^m$ for all $m \in \mathbb{Z}, m > T$.

Proof:

Step 1: Represent system $\mathbf{S}^*$ as a series interconnection of two stable systems: a finite-latency system and a system represented by an infinite-dimensional state space model.

To show this, let $S : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a forward-shift operator defined by $(Sw)(t) = w(t+1)$ for all $w \in \ell^2(\mathbb{R})$. Let $\mathbf{M}_\infty : \ell^2(\mathbb{R}) \to \ell(\ell^2(\mathbb{R}))$ be an unbounded operator mapping w to $v = \mathbf{M}_\infty w$ such that $v(t) = S^t w$. Let operators $A : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$, $B : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ and $C : \ell^2(\mathbb{R}) \to \mathbb{R}$ be defined by

$$(A\xi)(t) = \sum_{\tau=-T}^{T} h(\tau)\xi(t-\tau+1), \quad B\xi = \xi, \quad C\xi = \xi(0),$$

for all $\xi \in \ell^2(\mathbb{R})$. Clearly, operator A is a $(T+1)$-banded Laurent operator whose symbol $f = f(t)$ is defined by $f(t) = h(t-1)$, for all $t \in \mathbb{Z}$. This implies that A is a contraction, due to $|H(\omega)| < 1$ for all $\omega \in [0, 2\pi)$ [18]. Let now system $\mathbf{T}_\infty : \ell(\ell^2(\mathbb{R})) \to \ell^2(\mathbb{R})$, mapping v to $y = \mathbf{T}_\infty v$, be defined by the following infinite-dimensional state space model

$$\mathbf{T}_\infty : x(t+1) = \mathrm{Sat}_r(Ax(t)+Bv(t)), \;\; y(t) = Cx(t). \quad (24)$$

It immediately follows, from the definition (5) of the optimal map $\mathbf{S}^*$ and the above construction of $\mathbf{M}_\infty$ and $\mathbf{T}_\infty$, that $\mathbf{S}^* = \mathbf{T}_\infty\mathbf{M}_\infty$, see Fig. 2(a).

A necessary assumption for the generalized balanced truncation algorithm from theorem 4 is that the system to be approximated is driven by square summable signals. Due to unboundedness of $\mathbf{M}_\infty$, the input of $\mathbf{T}_\infty$ is not square summable, and theorem 4 cannot be directly used to establish useful bounds on approximation error. In order to mitigate this, we introduce a suitable coordinate re-scaling as follows.

Let $D : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a bounded linear operator such that $(Dw)(t) = \gamma_0^{|t|} w(t)$, $\forall w \in \ell^2(\mathbb{R})$, where $\gamma_0 \in (0,1)$ and $\|DAD^{-1}\| < 1$. The existence of such γ_0 is guaranteed by Lemma 3. If we apply coordinate transformation $\tilde{x} = Dx$ to (24), we get

$$\mathbf{T}_\infty : \tilde{x}(t+1) = \varphi(\tilde{A}\tilde{x}(t)+\tilde{B}Dv(t)), \;\; y(t) = \tilde{C}\tilde{x}(t), \quad (25)$$

where $\varphi = D\,\mathrm{Sat}_r\,D^{-1}$, $\tilde{A} = DAD^{-1}$, $\tilde{B} = DBD^{-1} = I$ and $\tilde{C} = CD^{-1} = C$. Since operators D and Sat_r are both diagonal operators, and Sat_r has Lipschitz constant equal to 1, it follows that the Lipschitz constant of the operator φ is also equal to 1. In the rest of this proof, we assume that $\delta \in (0,1)$ is such that $\|\tilde{A}\|^2 \le 1 - \delta < 1$.

Let $\tilde{\mathbf{M}}_\infty : \ell^2(\mathbb{R}) \to \ell^2(\ell^2(\mathbb{R}))$, mapping w to $\tilde{v}$, be such that $\tilde{v}(t) = DS^t w$. Consider a system $\tilde{\mathbf{T}}_\infty$ described by the following state space model

$$\tilde{\mathbf{T}}_\infty : \tilde{x}(t+1) = \varphi(\tilde{A}\tilde{x}(t) + \tilde{B}\tilde{v}(t)), \;\; y(t) = \tilde{C}\tilde{x}(t), \quad (26)$$

It now clearly follows that system $\mathbf{S}^*$ can be represented as a series interconnection $\mathbf{S}^* = \tilde{\mathbf{T}}_\infty\tilde{\mathbf{M}}_\infty$ of $\tilde{\mathbf{M}}_\infty$ and $\tilde{\mathbf{S}}_\infty$, see Fig. 2(b). It is not hard to show that $\tilde{\mathbf{M}}_\infty$ is bounded and $||\tilde{\mathbf{M}}_\infty|| = \left(\frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}}$ Indeed, for $\tilde{v} = \tilde{\mathbf{M}}_\infty w$, we have that

$$|\tilde{v}|^2 = \sum_{t=-\infty}^{\infty}\sum_{\tau=-\infty}^{\infty} \gamma_0^{2|t-\tau|}|w(\tau)|^2 = \frac{1+\gamma_0^2}{1-\gamma_0^2}|w|^2.$$

Step 2: Represent system $\hat{\mathbf{S}}_m$ as a series interconnection of two stable systems: a finite-latency system and a system represented by an infinite-dimensional state space model.

Let $\Theta_m : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be a projection operator such that $(\Theta_m w)(t) = w(t)$ for $-m+1 \leq t \leq m+1$, and $(\Theta_m w)(t) = 0$ otherwise. Let $\hat{X} = \{\Theta_m u : u \in \ell^2(\mathbb{R})\}$. Consider a system $\hat{\mathbf{T}}_{m,\infty} : \ell^2(\ell^2(\mathbb{R})) \to \ell^2(\mathbb{R})$, mapping $\tilde{v}$ to $\hat{y}$, defined by the following infinite-dimensional state space model

$$\hat{\mathbf{T}}_{m,\infty} : \hat{x}(t+1) = \Theta_m \varphi(\tilde{A}\hat{x}(t) + \tilde{B}\tilde{v}(t)),\ \hat{y}(t) = \tilde{C}\hat{x}(t). \tag{27}$$

where $\hat{x}(t) \in \hat{X}$ for all $t \in \mathbb{Z}$.
It is not hard to see that $\hat{\mathbf{S}}_m = \hat{\mathbf{T}}_{m,\infty}\tilde{\mathbf{M}}_\infty$, see Fig. 1(b). Indeed, the state space model of $\hat{\mathbf{T}}_{m,\infty}$ is formally infinite-dimensional but in fact only $2m+1$ state components are nonzero, and those exactly correspond to the state variables of the subsystem $\hat{\mathbf{T}}_m$ of $\hat{\mathbf{S}}_m$.

Step 3: Find $\sigma_1 > 0$, $\sigma_2 > 0$, $P : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$, and $Q : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ that satisfy the assumptions of Theorem 4 for $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, Θ_m, and φ as given above.

To find this, let us first assume that $\delta \in (0,1)$, as chosen in Step 1, is such that $\|\tilde{A}\|^2 \leq 1-\delta$. It immediately follows that inequality $\tilde{A}\tilde{A}' \leq I - \delta I$ holds. Moreover, since $\tilde{B} = I$, the inequality $I - \tilde{A}I\tilde{A}' \geq \delta \tilde{B}\tilde{B}'$ holds as well. It now follows that $\sigma_1 = \frac{1}{\sqrt{\delta}}$ and $P = I$ satisfy the first inequality in (8).

For an arbitrary, but fixed, $\rho_0 \in (0,1)$, let $Q : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ be defined as $(Qw)(t) = \rho_0^{|t|-m} w(t)$ for $|t| \leq m-1$, and $(Qw)(t) = w(t)$ otherwise, for all $w \in \ell^2(\mathbb{R})$. Similar to the proof of Lemma 3, it can be shown that there exist $\delta_0 \in (0,1)$ and $\rho_0 \in (0,1)$ (ρ_0 does not depend on m) such that $\|Q^{1/2}\tilde{A}Q^{-1/2}\|^2 \leq 1 - \delta_0 < 1$. This implies that

$$I - Q^{-1/2}\tilde{A}'Q\tilde{A}Q^{-1/2} \geq \delta_0 I,$$

and, moreover,

$$Q - \tilde{A}'Q\tilde{A} \geq \delta_0 Q \geq \frac{\delta_0}{\rho_0^m}\tilde{C}'\tilde{C},$$

Therefore, the above defined Q and $\sigma_2 = \sqrt{\frac{\rho_0^m}{\delta_0}}$ satisfy the second inequality in (8).

From the definition of P, Q, and Θ_m it immediately follows that (9) is true, while (10) and (11) follow from the fact that φ is diagonal and has Lipschitz constant equal to 1, and P and Q are positive definite diagonal operators.

Step 4: The following series of inequalities hold

$$\begin{aligned}\|\mathbf{S}^* - \hat{\mathbf{S}}_m\| &= \|(\tilde{\mathbf{T}}_\infty - \hat{\mathbf{T}}_{m,\infty})\tilde{\mathbf{M}}_\infty\| \\ &\leq \|\tilde{\mathbf{T}}_\infty - \hat{\mathbf{T}}_{m,\infty}\|\|\tilde{\mathbf{M}}_\infty\| \\ &\leq \left(\frac{4}{\delta\delta_0}\cdot\frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}} \rho_0^{m/2}.\end{aligned}$$

Therefore, $\rho = \sqrt{\rho_0}$ and $c = \left(\frac{4}{\delta\delta_0}\cdot\frac{1+\gamma_0^2}{1-\gamma_0^2}\right)^{\frac{1}{2}}$ satisfy the condition of Theorem 5. This concludes the proof. ■

V. CONCLUSIONS

In this paper, a problem of designing discrete-time systems which are optimal in frequency-weighted least squares sense subject to a maximal output amplitude constraint was considered. An extension to the method of balanced truncation for linear systems to a certain class of nonlinear models was derived. A causally stable finite-latency nonlinear system which returns high-quality approximations to the optimal map was proposed. The approximate system was obtained by a careful truncation of an infinite dimensional state space representation of the optimal system, as suggested by the derived generalization of the balanced truncation method.

REFERENCES

[1] U. Oreborn, "A direct method for sparse nonnegative least squares problems," Technical Report, Department of Mathematics, Thesis, Linköping University, Linköping, Sweden, 1986.

[2] E. Osuna, R. Freund, and F. Girosi, "Support Vector Machines: Training and Applications," Technical Report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1997.

[3] G. Goodwin, M. M. Seron, and J.A. de Doná, Constrained Control and Estimation: An Optimisation Approach, 1st edition, Springer, 2010.

[4] P. Hungerländer and F. Rendl, "A Feasible Active Set Method for Strictly Convex Quadratic Problems with Simple Bounds," *SIAM Journal on Optimization*, vol. 25, no. 3, pp. 1633-1659, 2015.

[5] D. P. Bertsekas, "On the Goldstein-Levitin-Polyak gradient projection method," *in IEEE Transactions on Automatic Control*, vol. 21, no. 2, pp. 174-184, April 1976.

[6] W. Li and J. Swetits, "A Newton Method for Convex Regression, Data Smoothing, and Quadratic Programming with Bounded Constraints," *SIAM Journal on Optimization*, vol. 3, no. 3, pp. 466-488, 1993.

[7] P.M. Pardalos,Y. Ye, and C.-G. Han, An interior-point algorithm for large-scale quadratic problems with box constraints, *Analysis and Optimization of Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 413–422, 1990.

[8] M. Sznaier and M. J. Damborg, "Suboptimal control of linear systems with state and control inequality constraints," *26th IEEE Conference on Decision and Control*, Los Angeles, California, USA, 1987, pp. 761-762.

[9] D. J. Chmielewski and V. Manousiouthakis, "On constrained infinite-time linear quadratic optimal control,"*35th IEEE Conference on Decision and Control*, Kobe, Japan, 1996, pp. 1319-1324 vol.2.

[10] P. Grieder, F. Borrelli, F. Torrisi, and M. Morari, "Computation of the constrained infinite time linear quadratic regulator," *Automatica*, vol. 40, no. 4, 2004, pp. 701-708.

[11] G. Stathopoulos, M. Korda and C. N. Jones, "Solving the Infinite-Horizon Constrained LQR Problem Using Accelerated Dual Proximal Methods," *in IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1752-1767, April 2017.

[12] J. G. Proakis and M. Salehi, *Digital Communications*. McGraw-Hill, 2007.

[13] R. Reine and Z. Zang, "A quadratic programming approach in pulse shaping filter design to reducing PAPR in OFDM systems," *2013 19th Asia-Pacific Conference on Communications (APCC)*, Denpasar, 2013, pp. 572-576.

[14] J. J. Sochacki, "A Preoptimized Peak to Average Power Ratio Pulse Shaping Filter and Its Effect on System Specifications," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 7, pp. 2137-2145, July 2016.

[15] O. Tanovic and A. Megretski, "Real-Time Realization of a Family of Optimal Infinite-Memory Non-Causal Systems", *6th IFAC Conference on Nonlinear Model Predictive Control (NMPC)*, Madison, WI, 2018, pp. 168-173.

[16] A. A. Goldstein, "Convex Programming in Hilbert Space", *Bull. Amer. Math. Soc.*, vol. 70, no. 5, pp. 709-710, 1964.

[17] K. Zhou, J.C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, 1996.

[18] A. E. Frazho and W. Bhosri, *An Operator Perspective on Signals and Systems*, Chapter 2 "Toeplitz and Laurent Operators", Birkhäuser Basel, pp. 23-40, 2010.