International Journal of Control
Vol. 82, No. 8, August 2009, 1448–1468

Taylor & Francis
Taylor & Francis Group

Optimal control during feedback failure

Debraj Chakraborty[a] and Jacob Hammer[b]*

[a]*Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India;* [b]*Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA*

(*Received 4 April 2008; final version received 20 September 2008*)

The problem of controlling a perturbed open loop system so as to keep its performance errors within bounds is considered. The objective is to maximise the time during which performance errors remain below a prescribed ceiling, while the controlled system's parameters are within a specified neighbourhood of their nominal values. It is shown that there is an optimal open loop controller that achieves this objective. Conditions under which the optimal controller generates a bang-bang control input signal are characterised. In general, it is shown that the performance of the optimal controller can always be approximated by a bang-bang signal.

Keywords: optimal control; robust control; bang-bang control

1. Introduction

Oftentimes, control systems must operate temporarily without feedback. Interruptions in the feedback signal may be caused by malfunctions or disruptions in the feedback communication link, or they may be the result of efforts to reduce operating costs. In other applications, feedback channels are opened only occasionally, when system performance degrades below an acceptable level.

Consider, for example, the medical treatment of type 1 diabetes. Individuals afflicted by this condition require periodic injections of insulin in order to control the glucose concentration in their blood. Insulin is injected when glucose concentration deviates by more than a specified amount from nominal level. Insulin injection is often done by an implanted insulin infusion pump, which allows excellent control of the infusion profile. The feedback mechanism in this case consists of periodical blood analyses, which, at the present time, require the drawing of blood through finger pricks or similar irksome procedures. In order to improve patient comfort, it would be desirable to maximise the time interval between blood samplings, while maintaining blood glucose concentration within desirable bounds. Needless to say, models of the dynamics of blood glucose concentration are subject to significant errors and depend on external interferences. In this context, the objective of the present article is to develop techniques for the design of glucose infusion profiles that keep blood glucose concentrations within desirable bounds and allow the longest possible time interval between blood samplings.

Intermittent use of feedback is also of interest in other biomedical applications. Consider, for example, the treatment of cancer by chemotherapy. Here, it would be of advantage to maximise the time between observations of cancer status, observations that often require extensive testing. The methodology developed in the present article can be used to design optimal chemotherapy protocols that maximise the time between subsequent tests. Such protocols will improve patient independence and reduce costs (e.g. Panetta and Fister (2003), and others). Many additional potential applications in biomedicine are possible as well.

Another potential application can be found in networked control systems, where feedback is used only intermittently so as to reduce network traffic (e.g. Zhivogyladov and Middleton (2003), Montestruque and Antsaklis (2004), Nair, Fagnani, Zampieri, and Evans (2007), and others). Here, feedback sensors and system actuators communicate through networks that are shared by a vast number of users, with only limited network capacity available for each user. To abide by network capacity limitations, feedback can only be used intermittently. Examples of applications of networked control systems include spatially distributed resource allocation networks, highway transportation control systems, power generation and distribution networks, and others. Clearly, to minimise traffic within communications

*Corresponding author. Email: hammer@mst.ufl.edu

ISSN 0020–7179 print/ISSN 1366–5820 online

DOI: 10.1080/00207170802510265
http://www.informaworld.com

networks, it is necessary to reduce feedback and actuator use. The methodology developed in the present article can help accomplish this task by providing open loop input signals that allow operation without feedback for maximal intervals of time.

In general terms, our objective is to address the needs exhibited by such applications and others through the development of open loop controllers that maximise the duration of time during which a perturbed system can operate without feedback and not exceed acceptable error bounds. Specifically, consider a system Σ whose parameters are not accurately known. Let Σ_0 be the nominal version of Σ, and let Σ_ϵ be the system obtained when the parameters of Σ are perturbed by ϵ from their nominal values. The exact value of ϵ is not known; it is, however, assured that ϵ does not exceed a specified bound d. Now, for an input function $v(t)$, denote by $\Sigma_0 v$ the response of the nominal system and let $\Sigma_\epsilon v$ be the response of the perturbed system. The perturbation then creates a deviation of the response, given by the magnitude $|\Sigma_\epsilon v - \Sigma_0 v|$. To reduce this deviation, we employ an open loop controller that adds a 'correction signal' $u(t)$ to the input signal, so that the output signal of the system with the controller becomes $\Sigma_\epsilon(v+u)$. Comparing to the nominal output signal, we have then the deviation $|\Sigma_\epsilon(v+u) - \Sigma_0 v|$. Of course, the correction signal $u(t)$ must be independent of the perturbation ϵ, since the perturbation is not specified. Furthermore, as the feedback signal was completely disrupted at the time $t=0$, the input function $u(t)$ cannot depend on the state or the output of Σ. Let M designate the maximal deviation that is allowed, and let t_f be the duration of time during which

$$|\Sigma_\epsilon(v+u)(t) - \Sigma_0 v(t)| \leq M, \quad 0 \leq t \leq t_f. \tag{1}$$

The goal of our present discussion is to find a correction signal $u(t)$ that maximises the duration t_f, given only that the perturbation ϵ is bounded by d. To accommodate natural restrictions on the input amplitude of the system Σ, we assume that all input signals of Σ must have an amplitude not exceeding $K > 0$.

In the present article, we concentrate on some of the basic aspects of this problem. In particular, we restrict our attention to the case where the system Σ is a linear time-invariant system. To simplify the calculations, we choose simple nominal operating conditions for the nominal system Σ_0, setting the nominal initial conditions at zero and taking the zero signal as the nominal input signal (however, the initial conditions of the perturbed system Σ are not assumed to be zero). Under such nominal operating conditions, we obtain that $\Sigma_0 v = 0$; the correction signal $u(t)$ compensates for deviations caused by perturbations of the parameters and of the initial conditions. Letting x_0 be the initial condition of the perturbed system, inequality (1) reduces to

$$|\Sigma_\epsilon u(t)| \leq M \quad \text{for all } |x_0| \leq M, |\epsilon| \leq d, \text{ and } 0 \leq t \leq t_f, \tag{2}$$

where $|u(t)| \leq K$ for all t. We intend to derive the correction signal $u(t)$ that maintains the validity of (2) for the longest duration t_f. The control configuration takes the form shown in Figure 1, where the controller C generates the correction signal $u(t)$.

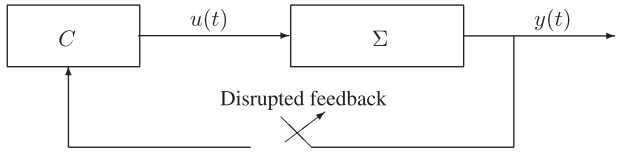

Figure 1. Feedback failure schematic diagram.

We show in § 2 that the calculation of the correction signal $u(t)$ leads to a max–min optimisation problem. In § 3, we show that this problem does have a solution, and in § 4 we characterise the conditions under which the optimal corrective signal $u(t)$ is a bang-bang signal. (Recall that a bang-bang signal is a signal whose components assume only their extreme values, switching from one extreme value to another as necessary.) Cases in which the optimal solution is not necessarily a bang-bang signal are examined in § 5, where we show that optimal performance can always be approximated by using a bang-bang correction signal $u(t)$. Thus, one can always achieve optimal, or close to optimal, performance by employing a bang-bang correction signal. Bang-bang signals are relatively easy to calculate and implement, since one only has to calculate the switching times of the signal.

Critical features of the optimal correction signal $u(t)$ are determined by a function $z(t)$ introduced in § 4, which is reminiscent of the switching function so often employed in classical time-optimal control. The optimal correction signal $u(t)$ is a bang-bang signal in intervals of time over which the function $z(t)$ is not identically zero. Like in the case of the classical switching function, components of the optimal correction signal $u(t)$ switch from one extremal value to another when the corresponding component of $z(t)$ changes sign. However, on intervals in which the function $z(t)$ is identically zero, the optimal correction signal $u(t)$ may not be a bang-bang signal; nevertheless, we show in § 5 that optimal performance can be approximated by a bang-bang corrective signal during such intervals.

The discussion in this article impinges on some of the most common practices in modern control

systems – the use of digital controllers to operate continuous-time systems. Most often, when operating a continuous-time system with a digital controller, the controller signal is kept constant between sample times (so-called 'zero order hold' method). The results of the present article show that, by using an optimal signal instead of the constant signal, the sampling interval can often be significantly increased without increasing performance errors. A simple example in this regard is provided in §4.

Our considerations in this article rely on the large body of literature available in the area max–min optimisation, including the works of Kelendzheridze (1961), Pontryagin, Boltyansky, Gamkrelidze, and Mishchenko (1962), Gamkrelidze (1965), Neustadt (1966 and 1967), Luenberger (1969), Young (1969), Warga (1972), the references cited in these works, and many others.

2. Notation and problem formulation

Let Σ be a linear time invariant continuous-time system described by the realisation

$$\Sigma : \dot{x}(t) = A'x(t) + B'u(t), \quad x(0) = x_0. \tag{3}$$

Here, $x(t)$ is the state of Σ at the time t, and $u(t)$ is the input function at the time t. We denote by n the dimension of $x(t)$ and by m the dimension of $u(t)$. Accordingly, A' and B' are constant real matrices of dimensions $n \times n$ and $n \times m$, respectively. Note that Σ is an input/state system, namely, the state $x(t)$ of Σ is available as output. We assume that the system Σ was connected in a state feedback loop until the time $t = 0$, when the feedback signal was lost. Thus, the initial state x_0 of Σ is known, being the last state value provided by the feedback.

The entries of the matrices A' and B' are not accurately known; rather, there are uncertainties about these entries. To describe these uncertainties, let $d > 0$ be a real number. Denote by Δ_A the set of all $n \times n$ matrices with entries in the interval $[-d, d]$, and let Δ_B be the set of all $n \times m$ matrices with entries in $[-d, d]$. Then,

$$A' := A + D_A \quad \text{and} \quad B' := B + D_B, \tag{4}$$

where $D_A \in \Delta_A$ and $D_B \in \Delta_B$ are unspecified matrices. Here, A and B represent the nominal values of the matrices A' and B' of (3), respectively, while $D_A \in \Delta_A$ and $D_B \in \Delta_B$ represent perturbations and uncertainties. In shorthand, denote

$$D := (D_A, D_B) \quad \text{and} \quad \Delta := \Delta_A \times \Delta_B \tag{5}$$

so that $D \in \Delta$. We refer to Δ as the *uncertainty range*. The only information available about the system Σ are the nominal matrices A and B and the uncertainty magnitude d; the entries of the matrices D_A and D_B are not specified. The performance requirement (2) can now be rewritten in the form

$$x^T(t)x(t) \le M \quad \text{for all } D \in \Delta \text{ and all } t \in [0, t_f], \tag{6}$$

where x^T is the transpose of x. The initial state x_0 satisfies $x_0^T x_0 \le M$, so that performance was within bounds when the feedback channel was disrupted. Our objective is to find an input function $u(t)$ that maximises the duration t_f.

Given two m-dimensional vector valued functions $a(t)$ and $b(t)$, we define their weighted inner product by setting

$$\langle a(t), b(t) \rangle = \int_0^\infty e^{-\alpha t} a^T(t) b(t) \mathrm{d}t, \tag{7}$$

where $\alpha > 0$ and the integral is taken in the Lebesgue sense. The weight function $e^{-\alpha t}$ allows us to include all bounded input functions in the domain over which the inner product (7) is well defined. Denote by $L_2^{\alpha,m}$ the Hilbert space of all m-dimensional Lebesgue measurable functions with the inner product (7).

Physical systems often have restrictions on the largest input signal amplitude they can tolerate. To describe these restrictions for a signal with m components, we use the pointwise norm

$$||u(t)|| = \max_{i=1,\ldots,m} |u_i(t)|,$$

where $u(t)$ is the vector $(u_1(t), u_2(t), \ldots, u_m(t))^T$ at the time t. Letting $K > 0$ be the input amplitude bound of the system Σ, it follows that the input function $u(t)$ of Σ must satisfy $\|u(t)\| \le K$ for all t. Then, all Lebesgue measurable input functions of Σ are members of the Hilbert space $L_2^{\alpha,m}$. Restricting ourselves to this set of input functions, denote by

$$U := \{u \in L_2^{\alpha,m} : ||u(t)|| \le K \quad \text{for all } t \ge 0\} \tag{8}$$

the set of all permissible input functions of Σ. In these terms, our objective is to find an input function $u \in U$ that drives the system Σ so as to preserve the state amplitude bound (6) for the longest possible time, irrespective of the perturbations that may affect Σ.

2.1 *Problem statement*

The response $x(t)$ of the system Σ depends, of course, on the perturbation matrices D_A and D_B, as well as on the input function u, so we often write $x(t, D, u)$ instead of $x(t)$, where $D = (D_A, D_B)$. Then, (6) takes the form

$$x^T(t, D, u)x(t, D, u) \le M \text{ for all } D \in \Delta \text{ and all } t \in [0, t_f]. \tag{9}$$

For a particular selection of D and u, the period of time during which the response amplitude does not exceed the bound M is characterised by the quantity

$$T(M, D, u) := \inf\{t \geq 0 : x^T(t, D, u)x(t, D, u) > M\}, \tag{10}$$

where $T(M, D, u) := \infty$ if $x^T(t)x(t) \leq M$ for all $t \geq 0$. Since $x_0^T x_0 \leq M$ by assumption, it follows that $T(M, D, u) \geq 0$. Referring to (6), we have $t_f = T(M, D, u)$ for these particular selections of D and u. As we discuss later, the form (10) turns the duration t_f into an upper semi-continuous functional of the input function, a fact that simplifies forthcoming mathematical arguments.

Among the variables of the state trajectory $x(t, D, u)$, the entries of the matrix $D = (D_A, D_B)$ are unknown and unpredictable; since no feedback is available, the input function u cannot depend on D. In order for the bound (9) to remain valid for all possible D, we must consider the 'worst case' with respect to D. This leads us to the quantity

$$T^*(M, u) := \inf_{D \in \Delta} T(M, D, u), \tag{11}$$

which describes the time duration during which inequality (9) is valid for all permissible perturbations $D \in \Delta$. Explicitly, for all $t \in [0, T^*(M, u)]$, we have $x^T(t, D, u)x(t, D, u) \leq M$ for any perturbation $D \in \Delta$.

The duration $T^*(M, u)$ still depends on the input function u, and we can choose any input function in the set U of (8). Of course, the best choice is an input function u that maximises $T^*(M, u)$, yielding the maximal duration

$$t_f^* := \sup_{u \in U} T^*(M, u). \tag{12}$$

When such an optimal input function exists, we denote it by u^*, so that $t_f^* = T^*(M, u^*)$. We can now state our objectives in formal terms.

Problem 2.1: (i) Determine whether or not there is an optimal input function $u^* \in U$ that yields the maximal duration t_f^*; and (ii) if such a function exists, describe a procedure for its computation.

As we can see from (11) and (12), the derivation of the optimal input function u^* involves the solution of a max–min optimisation problem. In the next section, we show that this optimisation problem does have a solution, so that an optimal input function u^* exists within the set U of input functions. Later, in §§ 4 and 5, we show that this optimal input function is either a bang-bang function, or it can be replaced by a bang-bang function without appreciably affecting performance. Bang-bang functions are relatively easy to compute and work with, since they are completely determined by their switching times.

3. Existence of an optimal solution

In this section, we show that there is an optimal input function $u^*(t)$ that keeps the perturbed system Σ within its error bounds for the longest possible duration. The discussion proceeds along two main steps: in the first step, we show that the set U of (8) is compact in a certain sense; and in the second step, we show that the functional $T^*(M, u)$ of (11) is continuous in an appropriate sense. Then, the existence of an optimal input function $u^*(t)$ within U follows from the fact that a continuous functional over a compact set achieves its maximum within the set. We start by reviewing a few notions from analysis (e.g. Liusternik and Sobolev (1961)).

Definition 3.1: Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

(i) A sequence $\{x^n\}$ in H *converges weakly* to an element $x \in H$ if $\lim_{n \to \infty} \langle x^n, y \rangle = \langle x, y \rangle$ for every element $y \in H$.
(ii) A subset W of H is *weakly compact* if every sequence of elements of W has a subsequence that converges weakly to an element of W.
(iii) A sequence $\{z^n\} \subset H$ is *strongly convergent* if there is an element $z \in H$ such that $\lim_{n \to \infty} \langle (z^n - z), (z^n - z) \rangle = 0$.
(iv) A set S is *strongly closed* if every strongly convergent sequence of elements of S has its limit in S.

We proceed now with the first step of our proof regarding the existence of an optimal input function $u^*(t)$.

Lemma 3.2: *The set U of* (8) *is weakly compact in the topology of the Hilbert space $L_2^{\alpha,m}$.*

Proof: By (8), the set U is bounded. Recall Alaoglu's theorem, which states that every bounded sequence in a Hilbert space contains a weakly convergent subsequence (e.g. Halmos (1982)). Hence, every sequence of elements of U has a subsequence that is weakly convergent to an element of $L_2^{\alpha,m}$. To prove weak compactness, it only remains to show that this element is a member of U. To this end, we show that U is weakly closed, namely, that every weakly convergent sequence of elements of U has its limit in U. We utilise Mazur's theorem, which states that a bounded and strongly closed convex set in Hilbert space is also weakly closed (e.g. Halmos (1982)).

To apply Mazur's theorem, note first that U is convex. Indeed, given two Lebesgue measurable

functions $v, w \in U$, we have, by the definition of U, that $\|v(t)\| \le K$ and $\|w(t)\| \le K$ for all t. Then, for a number $0 \le a \le 1$, the function $z(t) := av(t) + (1-a)w(t)$ is clearly Lebesgue measurable, and $\|z(t)\| \le a\|v(t)\| + (1-a)\|w(t)\| \le K$. Whence, $w(t) \in U$, and U is a convex set.

To show that U is strongly closed, let $u^n \in U$, $n = 1, 2, \ldots$, be a strongly convergent sequence of functions with the limit u, namely, $\lim_{n\to\infty} \langle (u^n - u), (u^n - u)\rangle = 0$. Assume, by contradiction, that $u \notin U$. Being the limit of a sequence of Lebesgue measurable functions, u is Lebesgue measurable as well. But then, in view of (8), the relation $u \notin U$ implies that there is a Lebesgue measurable subset $\delta \subset [0, \infty)$ of the time axis over which $\|u(t)\| \ge K + \epsilon$ for all $t \in \delta$, where $\epsilon > 0$ and δ has non-zero measure. As $u(t)$ is a vector of dimension m, it further follows that there is an integer $1 \le i \le m$ and a measurable subset $\delta_i \subset \delta$ of non-zero measure, such that the i-th component $u_i(t)$ of $u(t)$ satisfies

$$|u_i(t)| - K \ge \epsilon \quad \text{for all } t \in \delta_i. \tag{13}$$

Now, calculating the norm of the difference $u - u^n$, we get

$$\begin{aligned}\left\langle (u-u^n), (u-u^n)\right\rangle &= \int_0^\infty e^{-\alpha t}[u(t)-u^n(t)]^T[u(t)-u^n(t)]\mathrm{d}t \\ &\ge \int_{\delta_i} e^{-\alpha t}[u(t)-u^n(t)]^T[u(t)-u^n(t)]\mathrm{d}t \\ &\ge \int_{\delta_i} e^{-\alpha t}(u_i(t)-u_i^n(t))^2\mathrm{d}t, \end{aligned} \tag{14}$$

where $u_i^n(t)$ is the i-th component of the function $u^n(t)$. Now, since $u^n \in U$, we have that $\|u^n(t)\| \le K$ for all t, so that $|u_i^n(t)| \le K$ for all t as well. Thus, $|u_i(t) - u_i^n(t)| \ge |u_i(t)| - |u_i^n(t)| \ge |u_i(t)| - K$. Then, by (13), we have that $|u_i(t) - u_i^n(t)| \ge \epsilon$ for all $t \in \delta_i$. Substituting into (14) yields

$$\begin{aligned}\left\langle (u-u^n), (u-u^n)\right\rangle &\ge \int_{\delta_i} e^{-\alpha t}(u_i(t)-u_i^n(t))^2\mathrm{d}t \\ &\ge \int_{\delta_i} e^{-\alpha t}\epsilon^2\mathrm{d}t \end{aligned}$$

for all $n = 1, 2, \ldots$, contradicting the fact that the sequence $\{u^n\}$ is strongly convergent. Thus, $u \in U$, and the Lemma's assertion follows by Mazur's theorem. □

Our main focus in this article is on cases when the controlled system Σ is nominally unstable, namely, on cases when the nominal matrix A has at least one eigenvalue with positive real part. For such systems, controlling the error is particularly critical, as it may diverge and significantly impact system performance. The next statement indicates that, for such systems, the state trajectory $x(t)$ must escape the bound M of (9) for at least one perturbation matrix $D \in \Delta$.

Lemma 3.3: *Assume that the system Σ of (3) is nominally unstable and has a non-zero initial state. Then, for each input function $u(t) \in U$ and for every uncertainty range Δ, there is a perturbation matrix $D \in \Delta$ for which $T(M, D, u) < \infty$, where $T(M, D, u)$ is given by (10).*

The proof of Lemma 3.3 depends on an auxiliary fact, which is stated next. We denote by

$$||G|| := \max_{i=1,\ldots,q;j=1,\ldots,r} |G_{ij}|$$

the ℓ^∞-norm of a $q \times r$ matrix G with entries G_{ij}.

Lemma 3.4: *Let A^+ be an $n \times n$ block diagonal matrix*

$$A^+ = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix},$$

where A_u is an $n_u \times n_u$ matrix all of whose eigenvalues have strictly positive real parts, and A_s is an $n_s \times n_s$ matrix whose eigenvalues have non-positive real parts (possibly, $n_s = 0$). Assume that $n_u \ge 1$, and let z_0 be a non-zero vector. Then, for every real number $\epsilon > 0$, there is an $n \times n$ matrix E that satisfies the following:

(i) *The equation $\dot{z}(t) = (A^+ + E)z(t)$ has a divergent solution with $z(0) = z_0$; and*
(ii) $\|E\| \le \epsilon$.

Proof: First, partition the vector $z(t)$ into

$$z(t) =: \begin{bmatrix} p(t) \\ q(t) \end{bmatrix} \quad \text{and} \quad z_0 =: \begin{bmatrix} p_0 \\ q_0 \end{bmatrix},$$

where $p(t)$ has n_s components and $q(t)$ has n_u components. Then, for $E = 0$, we have

$$\dot{z}(t) = \begin{bmatrix} \dot{p}(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix}\begin{bmatrix} p(t) \\ q(t) \end{bmatrix},$$

so that

$$\begin{bmatrix} p(t) \\ q(t) \end{bmatrix} = \begin{bmatrix} \exp\{A_s(t)\} & 0 \\ 0 & \exp\{A_u(t)\} \end{bmatrix}\begin{bmatrix} p_0 \\ q_0 \end{bmatrix}.$$

Now, we distinguish between two cases:

Case 1: $q_0 \ne 0$. Then, $\|q(t)\| \to \infty$ as $t \to \infty$, since all eigenvalues of A_u have positive real parts. Whence $\|z(t)\| \to \infty$ as $t \to \infty$ and our claim is valid for $E = 0$. Hence, the lemma is valid in this case.

Case 2: $q_0 = 0$. Then, since $z_0 \neq 0$, we must have $p_0 \neq 0$. Let $a > 0$ be a real number, and consider the similarity transformation induced by the matrix

$$Q := \begin{bmatrix} I & 0 \\ aI & I \end{bmatrix} \quad \text{where } Q^{-1} = \begin{bmatrix} I & 0 \\ -aI & I \end{bmatrix}.$$

Define the function $y(t) := Qz(t)$, and partition

$$y(t) = \begin{bmatrix} y_s(t) \\ y_u(t) \end{bmatrix},$$

where y_s has n_s components and y_u has n_u components. Now, consider the vector $y(0) = Qz_0$. As $p_0 \neq 0$ and $q_0 = 0$, we have that $y_u(0) = ap_0 \neq 0$. Applying the similarity transformation, we get the matrix

$$A' := Q\begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix}Q^{-1} = \begin{bmatrix} A_s & 0 \\ a(A_s - A_u) & A_u \end{bmatrix}.$$

Adding to A' the perturbation matrix

$$D'_A := \begin{bmatrix} 0 & 0 \\ -a(A_s - A_u) & 0 \end{bmatrix},$$

we obtain the differential equation

$$\dot{y}(t) = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix} y(t).$$

Thus, $y_u(t)$ satisfies the equation $\dot{y}_u(t) = A_u y_u(t)$, so that $y_u(t) = \exp(A_u)y_u(0)$. In view of the fact that $y_u(0) \neq 0$ and all eigenvalues of A_u have strictly positive real parts, we obtain that $\|y_u(t)\| \to \infty$ as $t \to \infty$. Thus, $\|y(t)\| \to \infty$ as $t \to \infty$, and, considering that the matrix Q is invertible, we conclude that $\|z(t)\| \to \infty$ as well.

Returning to the original coordinate system, we add the perturbation $E := Q^{-1}D'_A Q$ to the matrix A^+ to achieve the same effect. Considering the forms of Q and Q^{-1}, it follows that $a > 0$ can be selected to satisfy

$$a\left\| Q^{-1}\begin{bmatrix} 0 & 0 \\ -(A_s - A_u) & 0 \end{bmatrix} Q \right\| \leq \epsilon;$$

then, our lemma is valid for this choice of E. This completes our proof. □

Proof (of Lemma 3.3): Let x_0 be the initial condition of the system Σ. Referring to (4), denote by $x^0(t)$ the zero input response of Σ with the perturbation matrix $D_A \in \Delta_A$. Then, we have the differential equation $\dot{x}^0(t) = (A + D_A)x^0(t)$, and the solution is $x^0(t) := e^{(A+D_A)t}x_0$. We show first that there is a perturbation matrix $D_A \in \Delta_A$ for which the norm $\|x^0(t)\|$ approaches infinity as $t \to \infty$.

Indeed, by assumption, the nominal matrix A has at least one eigenvalue with positive real part. Consequently, there is a similarity transformation $A^+ := PAP^{-1}$ that brings A into the block diagonal form

$$A^+ = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix},$$

where A_u is an $n_u \times n_u$ matrix all of whose eigenvalues have strictly positive real parts, and A_s is an $n_s \times n_s$ matrix whose eigenvalues have non-positive real parts (possibly, $n_s = 0$). Note that, by assumption, $n_u \geq 1$. Using the similarity transformation matrix, define the vector $z^0(t) := Px^0(t)$. Then, $z^0(t)$ satisfies the differential equation $\dot{z}^0(t) = A^+ z^0(t)$. Note that the initial condition $z^0(0) = Px^0(0) = Px_0$ is not zero, since the initial state x_0 is not zero by assumption and the matrix P is non-singular. In view of Lemma 3.4, there is then a perturbation matrix E for which the differential equation $\dot{z}(t) = (A^+ + E)z(t)$ has a divergent solution, when started from the initial condition $z^0(0)$.

Now, consider the effect of an input function $u(t) \in U$. Upon including the input in the differential equation and denoting the solution by $z'(t)$, we obtain

$$\dot{z}'(t) = (A^+ + E)z'(t) + B^+ u(t), \tag{15}$$

where $z'(t) = Px(t)$ and $B^+ := PB$. Then,

$$||z'(t)|| = ||Px(t)|| \leq ||P||_\infty ||x(t)||, \tag{16}$$

where $\|\cdot\|_\infty$ denotes the matrix norm induced by the l^∞-norm on $x(t)$. Since P is non-singular, (16) implies that $\|x(t)\|$ approaches infinity as $t \to \infty$, if the same is true for $z'(t)$.

Now, if $\|z'(t)\| \to \infty$ as $t \to \infty$ for the current input function $u(t)$, then the Lemma assertion is satisfied by $D := (P^{-1}EP, 0)$. Otherwise, assume that $\|z'(t)\|$ is bounded for all t, and recall that the solution of (15) has the form

$$z'(t) = z(t) + \exp[(A^+ + E)t] \int_0^t \exp[-(A^+ + E)\tau]B^+ u(\tau)\mathrm{d}\tau, \tag{17}$$

where $\lim_{t\to\infty} z(t) = \infty$ for the current D_A. Defining

$$\varphi(t) := \exp[(A^+ + E)t] \int_0^t \exp[-(A^+ + E)\tau]B^+ u(\tau)\mathrm{d}\tau,$$

we can write $z'(t) = z(t) + \varphi(t)$. As $\|z(t)\| \to \infty$ and $\|z'(t)\|$ is bounded, it follows that $\varphi(t)$ cannot be zero for all t. Choose then a real number $\delta > 0$ for which $\|\delta B^+\| \leq \epsilon$, and consider the perturbed matrix $B' := B^+ + \delta B^+ = (1 + \delta)B^+$. Using B' as the perturbed input matrix of the system Σ, it follows from the form of (17) that the solution becomes

$$\begin{aligned} z''(t) &= z(t) + (1 + \delta)\varphi(t) = z'(t) + \delta\varphi(t) \\ &= z'(t) + \delta[z'(t) - z(t)] = (1 + \delta)z'(t) - \delta z(t), \end{aligned}$$

so that $\|z''(t)\| \geq |(1+\delta)\|z'(t)\| - \delta\|z(t)\||$. Using the facts that $\lim_{t\to\infty}\|z'(t)\| < \infty$, while $\lim_{t\to\infty}\|z(t)\| = \infty$ and $\delta > 0$, we conclude that $\lim_{t\to\infty}\|z''(t)\| = \infty$. Finally, since $z''(t) = Px(t)$ and P is invertible, we obtain from (16) that $\lim_{t\to\infty}\|x(t)\| \geq \lim_{t\to\infty}\|z''t)\|/\|P\|_\infty = \infty$. Thus, the lemma is valid for the perturbation $D := (P^{-1}EP, \delta P^{-1}B^+)$, where $\epsilon > 0$ and $\delta > 0$ can be selected as small as desired. □

Lemma 3.3 shows that there is always a disturbance matrix D for which the escape time $T(M, D, u)$ is finite. Consequently, the quantity $T^*(M, u)$ of (11), which is the smallest of these escape times, must also be finite. This implies the following:

Corollary 3.5: *Assume that the system* Σ *of* (3) *is nominally unstable and is operated from a non-zero initial state, and let* $T^*(M, u)$ *be given by* (11). *Then,* $T^*(M, u) < \infty$ *for every input function* $u(t) \in U$ *and for every uncertainty range* Δ.

Our next objective is to show that there is an input function $u^*(t) \in U$ that maximises $T^*(M, u)$. We use a line of argument based on the fact that a continuous functional always attains its maximum in a compact set. Recall that we have shown in Lemma 3.2 that the set U is, in a sense, compact. We review now a weak notion of continuity that is compatible with the sense of compactness employed in Lemma 3.2 (e.g. Liusternik and Sobolev (1961)). (Throughout this article, R denotes the real numbers.)

Definition 3.6: Let S be a subset of a Hilbert space, and let $F: S \to R$ be a functional. Then, F is *weakly upper semi-continuous* at a point $z \in S$ if the following is true for every sequence $\{z_n\}_{n=1}^{\infty} \subset S$ that converges weakly to z: whenever $F(z)$ is bounded, there is, for every $\epsilon > 0$, an integer $N > 0$ such that $F(z_n) - F(z) < \epsilon$ for $n > N$. If the latter is true for all points $z \in S$, then F is *weakly upper semi-continuous*.

In order to show that the functional $T^*(M, u)$ is weakly upper semi-continuous, we need the following:

Lemma 3.7: *The functional* $T(M, D, u): U \to R$ *of* (10) *is weakly upper semi-continuous in u for any choice of M and D*.

Proof: Select a perturbation matrix $D \in \Delta$ and a bound M, and consider a sequence of input functions $u_1, u_2, \ldots \in U$ that weakly converges to the limit $u \in U$. For a time $t \geq 0$, consider the state vectors $x(t, D, u_1)$, $x(t, D, u_2), \ldots$. We claim that this sequence converges pointwise to the vector $x(t, D, u)$. Indeed, letting x_0 be the initial condition of Σ, it follows from the linear differential equation (3) that

$$x(t, D, u) = e^{A't}\left[x_0 + \int_0^t e^{-A'\tau}B'u(\tau)\mathrm{d}\tau\right]. \tag{18}$$

Defining the function

$$\rho(\tau) := \begin{cases} 1 & \text{if } \tau \leq t, \\ 0 & \text{otherwise,} \end{cases}$$

we can rewrite (18) in the form

$$x(t, D, u) = e^{A't}\left[x_0 + \int_0^\infty \rho(\tau)e^{-A'\tau}B'u(\tau)\mathrm{d}\tau\right].$$

Subtracting the contribution of the initial condition, we obtain the difference

$$\begin{aligned} v(t, D, u) &:= x(t, D, u) - e^{A't}x_0 \\ &= e^{A't}\int_0^\infty \rho(\tau)e^{-A'\tau}B'u(\tau)\mathrm{d}\tau, \end{aligned}$$

which is a linear functional of u. Recalling that weak convergence implies convergence of every linear functional of the sequence, we conclude that $\lim_{n\to\infty} v(t, D, u_n) = v(t, D, u)$ for every $t < \infty$. But then, since $x(t, D, u) = v(t, D, u) + e^{A't}x_0$, it follows that $\lim_{n\to\infty} x(t, D, u_n) = x(t, D, u)$ for every $t < \infty$.

Next, consider the following functional defined over state trajectories $x(t)$ of the system Σ:

$$\Theta(x) = \inf\{t \geq 0 : x^T(t)x(t) > M\}, \tag{19}$$

where $\Theta(x) := \infty$ if $x^T(t)x(t) \leq M$ for all $t \geq 0$. Let $x_1(t)$, $x_2(t), \ldots$ be a sequence of state trajectories that converges to the function $x(t)$ at each $t \geq 0$, and assume that $\Theta(x)$ is bounded. We show that, for any $\epsilon > 0$, there is an integer $N > 0$ that satisfies the following condition: $\Theta(x_n) - \Theta(x) < \epsilon$ for all integers $n > N$.

Clearly, if there is an integer $N > 0$ for which $\Theta(x_n) \leq \Theta(x)$ for all $n > N$, then our claim is true. So let us examine the case when there is no such N. Then, there is a divergent sequence of integers $i(1), i(2), \ldots$ such that $\Theta(x_{i(n)}) > \Theta(x)$ for all integers $n > 0$. Recall that, by our assumption, $\Theta(x) < \infty$. In view of (19), the following is true for every real number $\epsilon > 0$: there is a time $t' \in [\Theta(x), \Theta(x) + \epsilon)$ at which $x^T(t')x(t') > M$.

Further, since $x_n(t) \to x(t)$ at every $t \geq 0$, we also have that $\lim_{n\to\infty} x_n^T(t)x_n(t) = x^T(t)x(t)$ at every $t \geq 0$. Therefore, setting $t = t'$, there must be an integer $N > 0$ such that $|x_n^T(t')x_n(t') - x^T(t')x(t')| < [x^T(t')x(t') - M]/2$ for all $n \geq N$. For such n, we have $x_n^T(t')x_n(t') = x^T(t')x(t') + [x_n^T(t')x_n(t') - x^T(t')x(t')] \geq x^T(t')x(t') - [x^T(t')x(t') - M]/2 \geq x^T(t')x(t')/2 + M/2 > M$, i.e. $x_n^T(t')x_n(t') > M$. The last inequality implies that $\Theta(x_n) \leq t'$ so that, by the selection of t', we have that

$\Theta(x_n) < \Theta(x) + \epsilon$, or $\Theta(x_n) - \Theta(x) < \epsilon$, for all $n > N$. Consequently, $\Theta(x)$ is upper semi-continuous.

Finally, regarding the functional $T(M, D, u)$ of (10), note that $T(M, D, u) = \Theta(x(t, D, u))$. Now, let $\{u_n\}_{n=1}^{\infty} \subset U$ be a sequence that converges weakly to the function $u \in U$. Then, we have shown earlier in this proof that $\lim_{n\to\infty} x(t, D, u_n) = x(t, D, u)$ for every t. Combining this with the upper semi-continuity of Θ shown in the previous paragraph, it follows that $T(M, D, u)$ is weakly upper semi-continuous in u. This concludes our proof. □

We can now address the semi-continuity of the infimal time $T^*(M, u)$.

Lemma 3.8: *Assume that the system* Σ *of* (3) *is nominally unstable and has a non-zero initial state. Then, the function* $T^*(M, u)$ *of* (11) *is weakly upper semi-continuous in* u.

Proof: Our proof is based on the following mathematical fact (e.g. Willard (1970)). Let S and A be topological spaces, and let f_α be a weakly upper semi-continuous real valued function on S for each element $\alpha \in A$. If $\inf_{\alpha\in A} f_\alpha(x)$ exists at each point $x \in S$, then the function $f(x) := \inf_{\alpha\in A} f_\alpha(x)$ is weakly upper semi-continuous on S. Now, in view of Lemma 3.7, the function $T(M, D, u)$ is weakly upper semi-continuous on U for each $D \in \Delta$. Furthermore, since Σ is unstable and has a non-zero initial state, it follows by Corollary 3.5 that $\inf_{D\in\Delta} T(M, D, u) < \infty$ for every $u \in U$. Thus, by the fact quoted at the beginning of the proof, $T^*(M, u) := \inf_{D\in\Delta} T(M, D, u)$ is weakly upper semi-continuous in u. □

We are ready now to state the main result of this section: there is an input function that maximises the time during which our perturbed system's state remains within a specified error bound. This resolves part (i) of Problem 2.1.

Theorem 3.9: *Assume that the system* Σ *of* (3) *is nominally unstable and has a non-zero initial state, and let* $T^*(M, u)$ *be given by* (11). *Then, the following are valid*:

(i) *There is a maximal time* $t_f^* := \sup_{u\in U} T^*(M, u) < \infty$, *and*
(ii) *There is an input function* $u^* \in U$ *satisfying* $t_f^* = T^*(M, u^*)$.

Proof: The set U is weakly compact by Lemma 3.2 and, by Lemma 3.8, the functional $T^*(M, u)$ is weakly upper semi-continuous over U for any fixed error bound M. Consequently, we can apply the generalised Weierstrass Theorem (e.g. Zeidler (1985)), which, in our current terminology, states the following: a weakly upper semi-continuous function attains a maximum on a weakly compact set. Hence, $T^*(M, u)$ attains a maximum over the set of inputs U, and our proof concludes. □

To summarise, we have shown that, after a feedback failure occurs, there is an optimal input function $u^*(t)$ that keeps the open loop response below a specified error bound for a duration of at least t_f^*, irrespective of the perturbation matrices. While driven by the optimal input function $u^*(t)$, the actual duration of time t_f during which the system's response remains below the specified error bound depends, of course, on the entries of the perturbation matrix D. However, for all permissible perturbation matrices D, the duration t_f is never less than t_f^*, and there are values of D for which t_f gets indefinitely close to t_f^*. The optimal input function $u^*(t)$ is independent of the perturbation matrix D, as there is no feedback and no information can be deduced about D. Our next objective is to obtain a description of the optimal input function $u^*(t)$.

4. Characteristics of the optimal solution

We proceed in this section to show that the optimal input function $u^*(t)$ is often a bang-bang function, i.e. a function whose components switch between their bounds $+K$ and $-K$, lingering at no other values. Furthermore, in cases where $u^*(t)$ is not a bang-bang function, we show in § 5 that $u^*(t)$ can be replaced by a bang-bang function without significant deterioration in system performance. Thus, the solution of Problem 2.1 is closely linked to bang-bang input functions. Bang-bang input functions are desirable in engineering applications, since, being determined by their switching times, they are relatively easy to compute and implement.

4.1 *Examining the optimal solution*

To somewhat simplify the analysis of our optimisation problem 2.1, it would be convenient to reformulate it so as to make the terminal time into a constant that is not involved in the optimisation process. This can be achieved simply by introducing a scaling factor $\beta > 0$ in conjunction with the time variable, so that the actual time t is expressed as a product

$$t = \beta s,$$

where the variable s has the fixed range $0 \le s \le 1$ and the scaling factor β represents the terminal time. To obtain the maximal time duration, we then maximise the value of the scaling parameter β, rather than

maximise the length of the time interval. To this end, we introduce the variables

$$y(s) := x(\beta s) \quad \text{and} \quad v(s) := u(\beta s), \tag{20}$$

where $s \in [0, 1]$ and β is a constant parameter. In this context, we introduce the set of input functions

$$V := \{v \in L_2^{\alpha,m} : \|v(s)\| \leq K \quad \text{for all } 0 \leq s \leq 1, \text{ and } v(s) = 0 \text{ for all } s > 1\}. \tag{21}$$

Denoting $\dot{y} := \mathrm{d}y(s)/\mathrm{d}s$, we have $\dot{y} = \beta\, \mathrm{d}x/\mathrm{d}t$, and the system equation (3) takes the form

$$\Sigma : \dot{y}(s) = \beta[A'y(s) + B'v(s)], \quad 0 \leq s \leq 1, \quad y(0) = x_0. \tag{22}$$

As before, the matrices $A' = A + D_A$ and $B' = B + D_B$ are given by (4); the input function $v(s)$ is taken from the set V of (21). The new 'time variable' s is within the fixed interval $[0, 1]$, and is not subject to optimisation. To indicate the dependence of the solution $y(s)$ of (22) on the matrices $D := (D_A, D_B)$, the number β, and the input function v, we often use the expanded notation $y(s; \beta, D, v)$ instead of $y(s)$. Rephrasing (6), we are interested in values of β and in input functions $v \in V$ for which

$$y^T(s; \beta, D, v)y(s; \beta, D, v) \leq M \tag{23}$$

for all $0 \leq s \leq 1$ and for all matrices $D \in \Delta$, given that the initial condition $x_0 \neq 0$ has a magnitude $x_0^T x_0 < M$. A slight reflection shows that the maximal possible value of β is given by t_f^* of (12); we denote $\beta^* := t_f^*$. When the system Σ is nominally unstable, it follows by Theorem 3.9 that the maximal value β^* exists and that there is an input function $v^*(s)$ that forms the optimal solution, where

$$\begin{cases} v^* := u^*(\beta^* s), \ 0 \leq s \leq 1, \\ \beta^* = t_f^*; \end{cases} \tag{24}$$

here $u^*(t)$ is the optimal input function of Theorem 3.9.

Proceeding with our discussion, define the sets of matrices

$$\{A + \Delta_A\} := \{A' \in R^{n\times n} : A' = A + D_A, D_A \in \Delta_A\},$$
$$\{B + \Delta_B\} := \{B' \in R^{n\times m} : B' = B + D_B, D_B \in \Delta_B\}.$$

To further shorten the notation, we use

$$\Xi := \{A + \Delta_A\} \times \{B + \Delta_B\}. \tag{25}$$

Now, let ω be a Radon probability measure on the set

$$P := [0, 1] \times \Xi. \tag{26}$$

Given a point $(s, A', B') \in P$, let $\omega(A', B'|s)$ be the conditional probability measure induced by ω and let $\omega(s)$ be the marginal probability measure, so that

$$\omega(A', B', s) = \omega(A', B'|s)\omega(s), (s, A', B') \in P. \tag{27}$$

The following statement will shortly lead us to the conclusion that the optimal input function of Problem 2.1 is often a bang-bang function, namely, a function whose components switch between their maximal allowed values $+K$ and $-K$.

Theorem 4.1: *Assume that the conditions of Theorem* 3.9 *are valid. Let* $(v^*(s), \beta^*)$ *be a solution of Problem* 2.1 *as described by* (24), *and let* V *be the set of input functions* (21). *Then, there is a Lebesgue measurable function* $z(s) : [0, 1] \to R^m$ *such that* $z^T(s)v^*(s) \leq z^T(s)v(s)$ *for all input functions* $v \in V$ *and for almost all times* $s \in [0, 1]$.

We discuss the implications of Theorem 4.1 before providing its proof. One interesting implication of the theorem is the following. When a component of the function $z(s)$ of Theorem 4.1 is non-zero over an interval of time, then the corresponding component of the optimal input function $v^*(s)$ must equal either $+K$ or $-K$ over the same interval of time, where K is the maximal permissible input amplitude of the controlled system Σ. Indeed, assume that the j-th component $z_j(s)$ of $z(s)$ is positive over the interval $[s_1, s_2] \subset [0, 1]$, and consider the measurable input function $v(s) \in V$ whose components are given by

$$v_i(s) := \begin{cases} -K & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Then, the inequality $z^T(s)v^*(s) \leq z^T(s)v(s)$ reduces to the form $z_j(s)v_j^*(s) \leq z_j(s)(-K)$; cancelling $z_j(s) > 0$, we obtain $v_j^*(s) \leq -K$ which, due to the amplitude limitation, yields $v_j^*(s) = -K$ for all $s \in [s_1, s_2]$. Instead, if $z_j(s) < 0$ for all $s \in [s_1, s_2]$, a similar argument shows that $v_j^*(s) = K$ for all $s \in [s_1, s_2]$. We can summarise this discussion by the following statement:

Corollary 4.2: *Under the conditions and the notation of Theorem* 4.1, *assume that all components of the function* $z(s)$ *are non-zero almost everywhere in the interval* $[0, 1]$. *Then, the optimal input function* $v^*(s)$ *of Problem* 2.1 *is a bang-bang function, where*

$$v_j^*(s) := \begin{cases} -K & \text{if } z_j(s) > 0, \\ K & \text{if } z_j(s) < 0, \end{cases} \tag{28}$$

for almost all $s \in [0, 1]$ *and all* $j = 1, 2, \ldots, m$.

Bang-bang functions are preferable for design and implementation, as they are completely determined by their switching times. Example 4.11 of the next subsection demonstrates the situation considered in Corollary 4.2.

The function $z(s)$ of Theorem 4.1 is reminiscent of the classical switching function that appears in bang-bang control problems examined by Pontryagin et al. (1962). However, our current function $z(s)$ is, in some respects, different from the classical switching function. One such aspect is the fact that no conclusions can be drawn about the optimal input function $v^*(s)$ on time intervals where the corresponding components of $z(s)$ are zero. Nevertheless, we show in § 5 that, over such intervals, optimal performance can be approximated by a bang-bang input function. We turn now to some mathematical deliberations that lead to the proof of Theorem 4.1.

4.2 *Mathematical considerations*

Our arguments in this section are based to a large degree on the geometric form of the Hahn–Banach Theorem (e.g. Bourbaki (1987)), which is frequently used in the analysis of optimisation problems and can be stated as follows. Let S' and S'' be non-empty disjoint convex subsets of a topological vector space $\mathcal{B}$. Assume that the interior of S' is not empty and recall that R denotes the real numbers. Then, there is a linear functional $\ell : \mathcal{B} \to R$, not identically zero, that separates S' from S''; namely, there is a real number ρ such that $\ell(s') \le \rho \le \ell(s'')$ for all $s' \in S'$ and $s'' \in S''$. On the cross product space $R \times \mathcal{B}$, it is convenient to define the two following projections:

(i) The projection onto the reals $\Pi_r : R \times \mathcal{B} \to R : (r, b) \mapsto r$, and
(ii) The projection $\Pi^- : R \times \mathcal{B} \to \mathcal{B}$ that provides the $\mathcal{B}$ components of pairs with negative real parts, i.e. for a pair $(r, b) \in R \times \mathcal{B}$,

$$\Pi^-(r, b) := \begin{cases} b & \text{if } r < 0, \\ \phi & \text{if } r \ge 0, \end{cases}$$

where ϕ denotes the empty set. In these terms, the following consequence of the Hahn–Banach Theorem is used repeatedly in the sequel. (For a subset $C \subset \mathcal{B}$, denote by $\bar{C}$ the closure of C in $\mathcal{B}$.)

Lemma 4.3: *Let C be an open convex subset of the Banach space $\mathcal{B}$, and let S be a convex subset of $R \times \mathcal{B}$. Assume that S includes the origin $(0, 0)$, that 0 is an interior point of $\Pi_r S$, and that $0 \in \bar{C}$. Then, one of the following is true*:

(i) *There is a linear functional $\ell : \mathcal{B} \to R$, not identically zero, such that $\ell(s) \ge 0 \ge \ell(c)$ for all $s \in \Pi^- S$ and all $c \in \bar{C}$; or*
(ii) *There is an element $s \in S$ for which $\Pi_r s < 0$ and $\Pi^- s \in C$.*

Proof: Option (ii) simply states that $(\Pi^- S) \cap C \ne \phi$. Thus, it only remains to show that option (i) is valid whenever $(\Pi^- S) \cap C = \phi$. To this end, we show first that the projection $\Pi^- S$ is a convex set. Indeed, let $\Pi : R \times \mathcal{B} \to \mathcal{B} : (r, b) \mapsto b$ be the standard projection onto $\mathcal{B}$. Then, linearity implies that, for every convex set $Z \subset R \times \mathcal{B}$, the projection ΠZ is convex as well.

Let R^- be the set of all negative real numbers; then, a slight reflection shows that $R^- \times \mathcal{B}$ is a convex set. Now, since 0 is an interior point of $\Pi_r S$ by assumption, the intersection $(R^- \times \mathcal{B}) \cap S$ is not empty. Furthermore, since S is convex by assumption and the intersection of convex sets is convex, it follows that $S \cap (R^- \times \mathcal{B})$ is a convex set. As $\Pi^- S = \Pi(S \cap (R^- \times \mathcal{B}))$ and Π preserves convexity, we conclude that $\Pi^- S$ is a convex set. Applying the Hahn–Banach Theorem, we conclude that the condition $(\Pi^- S) \cap C = \phi$ implies that there is a linear functional $\ell : \mathcal{B} \to R$ (not identically zero) and a real number α such that

$$\ell(s) \ge \alpha \ge \ell(c) \quad \text{for all } s \in \Pi^- S \text{ and } c \in C.$$

Finally, since $0 \in \bar{C} \cap \overline{\Pi^- S}$ by the Lemma's assumptions, we conclude that $\alpha = 0$. This verifies option (i) of the lemma, and our proof concludes. □

Next, we review a generalised notion of the directional derivative, often referred to as the *Gateaux derivative*. Let X be a vector space over the real numbers R, let D be a subset of X, let V be a normed space, and let $T : D \to V$ be a function. Let x, $h \in D$ be two vectors, and assume that there is a real number $a(h) > 0$ such that $x + \alpha h \in D$ for all real numbers $0 \le \alpha < a(h)$. Then, the *right-sided Gateaux derivative* of T at x in the direction h is defined as the derivative of $T(x + \alpha h)$ with respect to $\alpha > 0$ evaluated at $\alpha = 0$, namely,

$$\mathcal{D}T(x; h) := \lim_{\alpha \to 0^+} \frac{1}{\alpha} \big[T(x + \alpha h) - T(x) \big]; \tag{29}$$

here, the limit is taken from the right in the sense of the norm on V. If this limit exists for all $h \in D$, then we say that T is *right-sided Gateaux differentiable* at x.

Assume now that T is right-side Gateaux differentiable at the point x. Then, the Gateaux derivative is, of course, a function of the direction h in which it is taken. The right-sided Gateaux derivative of T at x is *linear* in its direction if

$$\mathcal{D}T(x; \alpha h + \beta k) = \alpha \mathcal{D}T(x; h) + \beta \mathcal{D}T(x; k)$$

for all real numbers α, β. Linearity of the Gateaux derivative is a rather common feature, as indicated by the following statement.

Lemma 4.4: *Let X be a normed vector space over the real numbers, let D be a subset of X, let V be a normed vector space, and let $T: D \to V$ be a function. Denote by $|\cdot|$ the norm over X and the norm over V. Assume that T has a linear approximation, namely, that the following is valid for all vectors $x, \delta \in X$ for which $x+\delta \in D$: there is a linear functional $\theta(x): X \to V$ such that $T(x+\delta) = T(x) + \theta(x)(\delta) + O(\delta^2)$, where $\lim_{|\delta|\to 0} |O(\delta^2)|/|\delta| = 0$. Then, the right-sided Gateaux derivative of T at x is linear in its direction.*

Proof: A direct substitution yields that $\mathcal{D}T(x; h) := \lim_{\alpha\to 0^+} \frac{1}{\alpha}[T(x+\alpha h) - T(x)] = \lim_{\alpha\to 0^+} \frac{1}{\alpha}[T(x) + \theta(x)(\delta) + O(\delta^2) - T(x)] = \lim_{\alpha\to 0^+} \frac{1}{\alpha}[\theta(x)(\alpha h) + O((\alpha h)^2)] = \theta(x)(h) + \lim_{\alpha\to 0^+} O((\alpha h)^2)/\alpha = \theta(x)(h)$ by the lemma's assumption. Thus, $\mathcal{D}T(x; h)$ is linear in h, and our proof concludes. □

We state next a modified version of a result of Warga (1972), which will help us investigate cases in which the optimal solution of Problem 2.1 is a bang-bang function. First, some notation. For a function $T: S_1 \to S_2$, denote by T^{-1} the inverse set function of T, i.e. for a set $S \subset S_2$,

$$T^{-1}[S] := \{s \in S_1 : Ts \in S\}.$$

For a subset P of a finite dimensional metric space, denote by $C(P, R)$ the space of real valued continuous functions over P. Given a pair of real numbers ϵ, $M > 0$, let $G(-\epsilon, M)$ be the subset of $C(P, R)$ consisting of all members whose image is contained in the interval $[-\epsilon, M]$.

Lemma 4.5: *Let Q be a convex subset of a Banach space, let F be a convex set of real numbers, and let P be a compact subset of a finite dimensional metric space. Let $T_1: Q \times F \to R$ and $T_2: Q \times F \to C(P, R)$ be functions that satisfy the following requirements*:

(1) *The restriction of the function T_1 to the set $(T_2)^{-1}G(-\epsilon, M)$ attains a minimal value at the point $(q^*, f^*) \in (T_2)^{-1}G(-\epsilon, M)$.*
(2) *The functions T_1 and T_2 have continuous right-sided Gateaux derivatives that are linear in their direction at the point (q^*, f^*).*
(3) *The image $T_2[Q \times F] \subset C(P, R)$ includes only bounded functions.*

Then, there is a Radon probability measure ω over P and an ω-integrable function $\lambda: P \to R$ such that

(i) $|\lambda(p)| = 1$ *almost everywhere with respect to the measure ω;*
(ii) $\int_P \lambda(p)\mathcal{D}T_2[(q^*, f^*); (q, f) - (q^*, f^*)](p)\mathrm{d}\omega(p) \geq 0$ *for all $(q, f) \in Q \times F$;*
(iii) $\lambda(p)T_2(q^*, f^*)(p) = \max\{\lambda(p)(-\epsilon), \lambda(p)M\}$ *for ω-almost every $p \in P$.*

The proof of Lemma 4.5 depends on several auxiliary facts, which we list next.

Lemma 4.6: *Under the notation and conditions of Lemma* 4.5, *the set of pairs*

$$W(p) := \bigcup_{(q,f)\in Q\times F} \Big(\mathcal{D}T_1[(q^*, f^*); (q, f) - (q^*, f^*)], \mathcal{D}T_2[(q^*, f^*); (q, f) - (q^*, f^*)](p)\Big) \tag{30}$$

is a convex subset of R^2, for every point $p \in P$.

Proof: Let a and b be two elements of the set $W(p)$. Then, by (30), there are (q_1, f_1), $(q_2, f_2) \in Q \times F$ such that

$$\begin{aligned} a = \big(&\mathcal{D}T_1((q^*, f^*); (q_1, f_1) - (q^*, f^*)), \mathcal{D}T_2((q^*, f^*); \\ &(q_1, f_1) - (q^*, f^*))(p)\big), \\ b = \big(&\mathcal{D}T_1((q^*, f^*); (q_2, f_2) - (q^*, f^*)), \mathcal{D}T_2((q^*, f^*); \\ &(q_2, f_2) - (q^*, f^*))(p)\big). \end{aligned}$$

Now, let $\alpha, \beta \geq 0$ be two real numbers satisfying $\alpha + \beta = 1$. Then, using the linearity assumption (2) of Lemma 4.5 together with the fact that $\alpha + \beta = 1$, we can write

$$\begin{aligned} \alpha a + \beta b = &\big(\mathcal{D}T_1((q^*, f^*); \alpha[(q_1, f_1) - (q^*, f^*)]), \\ &\mathcal{D}T_2((q^*, f^*); \alpha[(q_1, f_1) - (q^*, f^*)])(p)\big) \\ &+ \big(\mathcal{D}T_1((q^*, f^*); \beta[(q_2, f_2) - (q^*, f^*)]), \\ &\mathcal{D}T_2((q^*, f^*); \beta[(q_2, f_2) - (q^*, f^*)])(p)\big) \\ = &\big(\mathcal{D}T_1((q^*, f^*); \alpha(q_1, f_1) + \beta(q_2, f_2) - (q^*, f^*)), \\ &\mathcal{D}T_2((q^*, f^*); \alpha(q_1, f_1) \\ &+ \beta(q_2, f_2) - (q^*, f^*))(p)\big). \end{aligned}$$

In view of the fact that Q and F are convex, the combination $(q_3, f_3) := \alpha(q_1, f_1) + \beta(q_2, f_2)$ belongs to $Q \times F$. Hence,

$$\begin{aligned} \alpha a + \beta b = \big(&\mathcal{D}T_1((q^*, f^*); (q_3, f_3) - (q^*, f^*)), \mathcal{D}T_2((q^*, f^*); \\ &(q_3, f_3) - (q^*, f^*))(p)\big) \in W(p), \end{aligned}$$

and $W(p)$ is a convex set. This concludes our proof. □

By leaving the point p in (30) unspecified, we obtain the set

$$S := W(\cdot) = \bigcup_{(q,f)\in Q\times F} \big(\mathcal{D}T_1((q^*, f^*); (q, f) - (q^*, f^*)), \mathcal{D}T_2((q^*, f^*); (q, f) - (q^*, f^*))(\cdot)\big), \tag{31}$$

which forms a subset of the cross product space $R \times C(P, R)$. This set has the following feature.

Lemma 4.7: *Under the notation and the assumptions of Lemma* 4.5, *the following are valid*:

(i) *The set* S *of* (31) *is a convex subset of* $R \times C(P, R)$.

(ii) *If there is a point* $(q', f') \in Q \times F$ *at which* $\mathcal{D}T_1((q^*, f^*); (q', f') - (q^*, f^*)) \neq 0$, *then* 0 *is an interior point of* $\Pi_r S$.

Proof: (i) To show that S is convex, let (r_1, c_1), $(r_2, c_2) \in S$ be two points, let $0 \leq \alpha \leq 1$ be a real number, and consider the combination $(r, c) := \alpha(r_1 c_1) + (1 - \alpha)(r_2, c_2) = (\alpha r_1 + (1 - \alpha) r_2, \alpha c_1 + (1 - \alpha) c_2)$. By the definition (31) of the set S, there are pairs (q_1, f_1), $(q_2, f_2) \in Q \times F$ such that

$$
\begin{aligned}
(r_1, c_1) = \big(&\mathcal{D}T_1((q^*, f^*); (q_1, f_1) - (q^*, f^*)), \\
&\mathcal{D}T_2((q^*, f^*); (q_1, f_1) - (q^*, f^*))(\cdot)\big), \\
(r_2, c_2) = \big(&\mathcal{D}T_1((q^*, f^*); (q_2, f_2) - (q^*, f^*)), \\
&\mathcal{D}T_2((q^*, f^*); (q_2, f_2) - (q^*, f^*))(\cdot)\big).
\end{aligned}
$$

Using assumption 2 of Lemma 4.5 regarding the linearity of the Gateaux derivatives, we can write

$$
\begin{aligned}
(r, c) = \big(&\mathcal{D}T_1((q^*, f^*); (\alpha q_1 + (1 - \alpha) q_2, \\
&c\alpha f_1 + (1 - \alpha) f_2) - (q^*, f^*)), \\
&\mathcal{D}T_2((q^*, f^*); (\alpha q_1 + (1 - \alpha) q_2, \\
&\alpha f_1 + (1 - \alpha) f_2) - (q^*, f^*))(\cdot)\big).
\end{aligned}
$$

In view of the fact that Q and F are both convex sets, it follows that $(r, c) \in S$ for all $0 \leq \alpha \leq 1$, and S is convex.

(ii) Substituting $(q, f) = (q^*, f^*)$ in (31), and using assumption (2) of Lemma 4.5 regarding the linearity of the Gateaux derivatives, it follows that $(0, 0) \in S$. Consequently, $0 \in \Pi_r S$. Further, assume that there is a point $(q', f') \in Q \times F$ at which the value $d := \mathcal{D}T_1((q^*, f^*); (q', f') - (q^*, f^*)) \neq 0$. Then, by the assumed linearity of the Gateaux derivative, it follows that, for every real number β, the number $\beta d \in \Pi_r S$. Thus, 0 is an interior point of $\Pi_r S$, and (ii) is valid. This concludes the proof. □

We introduce now an additional set that is important to our discussion. First, some notation. As usual, for a subset B of a topological space, we denote by $\mathrm{Int}(B)$ the interior of B, namely, the largest open set contained in B. Now, we shift the set of functions $\mathrm{Int}(G(-\epsilon, M))$ by subtracting the function $T_2(q^*, f^*)$ from each element, to obtain the set of continuous functions

$$
C := \mathrm{Int}(G(-\epsilon, M)) - T_2(q^*, f^*). \tag{32}
$$

Lemma 4.8: *Using the notation and the assumptions of Lemma* 4.5, *the set of continuous functions* C *of* (32) *has the following properties*:

(i) C *is an open and convex subset of* $C(P, R)$;

(ii) $0 \in \bar{C}$;

(iii) *If* $h \in C$, *then* $\gamma h \in C$ *for all* $0 < \gamma < 1$.

Proof: (i) Note that C is simply a shift of $\mathrm{Int}(G(-\epsilon, M))$ by the bounded function $T_2(q^*, f^*) \in C(P, R)$. Thus, in order to show that C is an open and convex set it is enough to show that $\mathrm{Int}(G(-\epsilon, M))$ is an open and convex set. As $\mathrm{Int}(G(-\epsilon, M))$ is an open set by the definition of the interior of a set, we only have to show that $\mathrm{Int}(G(-\epsilon, M))$ is a convex set. To this end, let $g_1, g_2 \in \mathrm{Int}(G(-\epsilon, M))$ be two functions, let $0 \leq \alpha \leq 1$ be a real number, and consider the combination $g := \alpha g_1 + (1 - \alpha) g_2$. A slight reflection shows that $g \in G(-\epsilon, M)$. To show that g is an interior point of $G(-\epsilon, M)$, note that the inclusion $g_1, g_2 \in \mathrm{Int}(G(-\epsilon, M))$ implies that there is a neighbourhood $N(g_1)$ of g_1 and a neighbourhood $N(g_2)$ of g_2 such that $N(g_1)$, $N(g_2) \subset G(-\epsilon, M)$. Assume now, by contradiction, that g is not an interior point of $G(-\epsilon, M)$. Then, for every real number $\eta > 0$, there is a function $f \in C(P, R)$ such that $|f - g| < \eta$ and $f \notin G(-\epsilon, M)$. Define the function $\phi := (f - g)$, and consider the two functions $g' := g_1 + \phi$ and $g'' := g_2 + \phi$. Then, $|g' - g_1| = |\phi| = |f - g| < \eta$ and, similarly, $|g'' - g_2| < \eta$. Consequently, for sufficiently small η, we must have $g' \in N(g_1)$ and $g'' \in N(g_2)$, so that $g', g'' \in G(-\epsilon, M)$. Furthermore, since $\alpha g' + (1 - \alpha) g'' = \alpha g_1 + (1 - \alpha) g_2 + \phi = g + \phi = f$ and $G(-\epsilon, M)$ is a convex set, it follows that $f \in G(-\epsilon, M)$, a contradiction. Thus, g is an interior point of $G(-\epsilon, M)$, and (i) is true.

(ii) By the notation of Lemma 4.5(1), we have that $(q^*, f^*) \in (T_2)^{-1} G(-\epsilon, M)$, so that $T_2(q^*, f^*) \in G(-\epsilon, M)$. Now, if $T_2(q^*, f^*) \in \mathrm{Int}(G(-\epsilon, M))$, then $0 \in C$, and it follows that $0 \in \bar{C}$ as well. Otherwise, let $0 < \delta < 1$ be a real number; a slight reflection shows that the function $(1 - \delta) T_2(q^*, f^*)$ is an interior point of $G(-\epsilon, M)$. Then, by the definition of C, the function

$$
\theta_\delta := (1 - \delta) T_2(q^*, f^*) - T_2(q^*, f^*) = -\delta T_2(q^*, f^*)
$$

is in C, and the function 0 satisfies $0 = \lim_{\delta \to 0} \theta_\delta$, so that $0 \in \bar{C}$.

(iii) Let h be a function in C. There is then a function $z \in \mathrm{Int}(G(-\epsilon, M))$ such that $z - T_2(q^*, f^*) = h$. For a real number $0 < \gamma < 1$, we can write $\gamma h = \gamma[z - T_2(q^*, f^*)]$. Now, let

$$
s := \gamma h + T_2(q^*, f^*) = \gamma z + (1 - \gamma) T_2(q^*, f^*). \tag{33}
$$

We know that $z \in \text{Int}(G(-\epsilon, M))$, so we have $-\epsilon < z(p) < M$ for all $p \in P$. Also, since $T_2(q^*, f^*) \in G(-\epsilon, M)$, we have that $-\epsilon \le T_2(q^*, f^*)(p) \le M$ for all $p \in P$. Consequently, $s(p) = \gamma z(p) + (1-\gamma) T_2(q^*, f^*)(p) < \gamma M + (1-\gamma)M = M$, and $s(p) = \gamma z(p) + (1-\gamma) T_2(q^*, f^*)(p) > \gamma(-\epsilon) + (1-\gamma)(-\epsilon) = -\epsilon$, so that $-\epsilon < s(p) < M$ for all $p \in P$. Thus, $s \in \text{Int}(G(-\epsilon, M))$, and since $\gamma h = s - T_2(q^*, f^*)$ by (33), it follows that $\gamma h \in C$, and our proof concludes. □

Lemma 4.9: *Using the notation and the assumptions of Lemma* 4.5, *assume that* $\mathcal{D}T_1((q^*, f^*); (q,f) - (q^*, f^*))$ *is not the zero function, and let C be the set of continuous functions* (32). *Then, there is a linear functional* $\ell : C(P, R) \to R$, *not identically zero, that satisfies the following inequalities*:

$$\ell\big(\mathcal{D}T_2((q^*, f^*); (q,f) - (q^*, f^*))(p)\big) \ge 0 \quad \text{for all } (q,f) \in Q \times F \text{ and all } p \in P; \text{ and} \tag{34}$$

$$\ell(c) \le 0 \text{ for all } c \in \bar{C}. \tag{35}$$

Proof: In view of Lemmas 4.7 and 4.8, the conditions of Lemma 4.3 are satisfied; whence one of the alternatives listed in Lemma 4.3 must be valid. Alternative (i) of Lemma 4.3 yields (34) and (35); we show next that alternative (ii) of Lemma 4.3 cannot be valid.

By contradiction, assume that alternative (ii) of Lemma 4.3 is valid, namely, that there is an element $s \in S$ for which $\Pi_r s < 0$ and $\Pi^- s \in C$, where C is given by (32). Recall from Lemma 4.5(1) that the point (q^*, f^*) is a minimum point of the function T_1 over the set $(T_2)^{-1}G(-\epsilon, M)$. Our proof will conclude upon showing that this fact contradicts Lemma 4.3(ii). Now, since $\mathcal{D}T_1((q^*, f^*); (q,f) - (q^*, f^*))$ is not the zero function by assumption, it follows by Lemma 4.7(ii) that 0 is an interior point of the projection $\Pi_r S$, where S is given by (31). As we have assumed that Lemma 4.3(ii) is valid, there must be a pair $(r, s) \in S$ such that $r < 0$ and $s \in C$, where C is given by (32). Explicitly, this means that there is a point $(q,f) \in (T_2)^{-1}G(-\epsilon, M)$ at which

$$\begin{cases} \mathcal{D}T_1((q^*, f^*); (q,f) - (q^*, f^*)) < 0, \text{and} \\ \mathcal{D}T_2((q^*, f^*); (q,f) - (q^*, f^*)) \in C. \end{cases} \tag{36}$$

At this pair (q,f), we define two functions $h_1 : [0,1] \to R$ and $h_2 : [0,1] \to C(P, R)$ as follows:

$$\begin{cases} h_1(\theta) := T_1(q^* + \theta(q - q^*), f^* + \theta(f - f^*)), \\ h_2(\theta) := T_2(q^* + \theta(q - q^*), f^* + \theta(f - f^*)), \end{cases} \tag{37}$$

for all $0 \le \theta \le 1$. Then, referring to (29), we obtain that the derivatives are given by

$$\begin{cases} h_1'(\theta) := \mathcal{D}T_1[(q^* + \theta(q - q^*), \\ \qquad f^* + \theta(f - f^*)); (q,f) - (q^*, f^*)], \\ h_2'(\theta) := \mathcal{D}T_2[(q^* + \theta(q - q^*), \\ \qquad f^* + \theta(f - f^*)); (q,f) - (q^*, f^*)]. \end{cases}$$

Taking $\theta = 0$, we obtain

$$\begin{cases} h_1'(0) := \mathcal{D}T_1((q^*, f^*); (q,f) - (q^*, f^*)) \\ h_2'(0) := \mathcal{D}T_2((q^*, f^*); (q,f) - (q^*, f^*)). \end{cases} \tag{38}$$

Now, we can use the Intermediate Value Theorem over the interval $[0, \delta\theta]$, $\delta\theta > 0$, to write

$$h_1(\delta\theta) = h_1(0) + h_1'(\eta)\delta\theta,$$

where $0 < \eta < \delta\theta$. Defining the quantity $r(\eta) := [h_1'(\eta) - h_1'(0)]$, we can rewrite the last equality in the form

$$h_1(\delta\theta) = h_1(0) + [h_1'(0) + r_1(\eta)]\delta\theta. \tag{39}$$

By the continuity assumption on the Gateaux derivatives listed in Lemma 4.5(2), we conclude that

$$\lim_{\delta\theta \to 0} r_1(\eta) = 0. \tag{40}$$

Now, by (36) and (38), we have that $h_1'(0) < 0$. In view of (40), there is a real number $\zeta > 0$, $\zeta < 1$, such that $|r_1(\eta)| < |h_1'(0)|/2$ for all $0 < \delta\theta < \zeta$. Then, for all $0 < \delta\theta < \zeta$, we obtain that $h_1(\delta\theta) < h_1(0)$, or using (37), we obtain that

$$T_1(q^* + \delta\theta(q - q^*), f^* + \delta\theta(f - f^*)) < T_1(q^*, f^*). \tag{41}$$

Consequently, our proof will conclude upon showing that $T_2((q^* + \delta\theta(q - q^*), f^* + \delta\theta(f - f^*)) \in G(-\epsilon, M)$, since then (41) contradicts assumption (1) of Lemma 4.5, which states that (q^*, f^*) is a minimum of T_1 on the set $(T_2)^{-1}G(-\epsilon, M)$. This would then show that alternative (ii) of Lemma 4.3 is not valid.

To show that $T_2((q^* + \delta\theta(q - q^*), f^* + \delta\theta(f - f^*)) \in G(-\epsilon, M)$, denote $c^- := \mathcal{D}T_2((q^*, f^*); (q,f) - (q^*, f^*))$; then, $c^- \in C$ by (36). Further, by the definition (32) of the set C, there is a function $c \in \text{Int}(G(-\epsilon, M))$ for which $c^- = c - T(q^*, f^*)$, or

$$\mathcal{D}T_2((q^*, f^*); (q,f) - (q^*, f^*)) = c - T(q^*, f^*). \tag{42}$$

An argument similar to the one used in the derivation of (39) and (40) leads to the analogous equations

$$h_2(\delta\theta) = h_2(0) + h_2'(0)\delta\theta + r_2(\eta)\delta\theta, \tag{43}$$

where

$$\lim_{\delta\theta\to 0} r_2(\eta) = 0. \tag{44}$$

Substituting (37) and (38) into (43), we obtain

$$\begin{aligned} &T_2(q^* + \delta\theta(q-q^*), f^* + \delta\theta(f-f^*)) \\ &\quad = T_2(q^*, f^*) + \mathcal{D}T_2((q^*, f^*); \\ &\qquad (q,f) - (q^*, f^*))\delta\theta + r_2(\eta)\delta\theta \\ &\quad = T_2(q^*, f^*) + [c - T(q^*, f^*)]\delta\theta + r_2(\eta)\delta\theta. \end{aligned}$$

This yields

$$\begin{aligned} &T_2(q^* + \delta\theta(q-q^*), f^* + \delta\theta(f-f^*)) - T_2(q^*, f^*) \\ &\quad = [c^- + r_2(\eta)]\delta\theta. \end{aligned} \tag{45}$$

Now, as $c^- \in C$ and C is an open set by Lemma 4.8(i), there is a neighbourhood $N(\chi)$ radius $\chi > 0$ around c^- that is entirely contained in C. In view of (44), there is a real number $\mu > 0$ such that $|r_2(\eta)| < \chi$ for all $\delta\theta < \mu$. Then, for all $\delta\theta < \mu$, we have that $c^- + r_2(\eta) \in C$. Applying now Lemma 4.8(iii), we conclude that $[c^- + r_2(\eta)]\delta\theta \in C$. By the definition (32) of C, there is then a function $f \in \mathrm{Int}(G(-\epsilon, M))$ such that $[c^- + r_2(\eta)]\delta\theta = f - T(q^*, f^*)$. Substituting into (45), we obtain

$$\begin{aligned} &T_2(q^* + \delta\theta(q-q^*), f^* + \delta\theta(f-f^*)) - T_2(q^*, f^*) \\ &\quad = f - T(q^*, f^*), \end{aligned}$$

so that $T_2(q^* + \delta\theta(q-q^*), f^* + \delta\theta(f-f^*)) = f \in \mathrm{Int}(G(-\epsilon, M))$. Thus, $T_2((q^* + \delta\theta(q-q^*), f^* + \delta\theta(f-f^*)) \in G(-\epsilon, M)$, and our proof concludes. □

Proof (of Lemma 4.5): The proof is based on the Riesz–Markov Representation Theorem (e.g. Evans and Gariepy (1992)), which can be stated as follows. Let P be a compact subset of R^m, and let $L : C(P, R) \to R$ be a bounded linear functional. Then, there is a positive Radon probability measure ω on P and an ω-measurable function $\lambda : P \to R$ such that $|\lambda(x)| = 1$ for ω-almost every $x \in P$ and $L(c) = \int_P \lambda c \, \mathrm{d}\omega$ for all $c \in C(P, R)$. Applying this theorem to our functional $\ell : C(P, R) \to R$ of (34) and (35), we obtain an integral representation of ℓ, namely, there is a positive Radon probability measure ω on P and an ω-integrable function $\lambda : P \to R$ such that

$$\begin{cases} |\lambda(p)| = 1 \text{ for } \omega\text{-almost all } p \in P; \\ \ell(c) = \int_P \lambda c \, \mathrm{d}\omega \text{ for all functions } c \in C(P, R); \text{ and} \\ \omega(P) > 0, \end{cases} \tag{46}$$

where the last item follows from the fact that, according to Lemma 4.9, ℓ is not identically zero. Combining (46) with (34), we obtain parts (i) and (ii) of Lemma 4.5(ii).

Next, consider the function

$$c^*(\cdot) := T_2(q^*, f^*)(\cdot) : P \to R.$$

In view of the definition (32) of the set C, every element $c \in \bar{C}$ can be written in the form $c = g - c^*$, where $g \in G(-\epsilon, M)$. Consequently, (35) takes the form $\ell(g - c^*) \le 0$ for $g \in G(-\epsilon, M)$. Using (46), we can rewrite the latter in the form

$$\begin{aligned} &\int_P \lambda(p)[c^*(p) - c(p)]\mathrm{d}\omega(p) \ge 0 \\ &\qquad \text{for all } c(p) \in G(-\epsilon, M). \end{aligned} \tag{47}$$

We show next that this inequality implies Lemma 4.5(iii). Let $C(P, [0,1])$ be the set of all continuous functions mapping P into the real interval $[0,1]$. For a pair of functions $\alpha \in C(P, [0,1])$ and $c \in G(-\epsilon, M)$, consider the continuous function

$$k(p) := (1 - \alpha(p))c^*(p) + \alpha(p)c(p). \tag{48}$$

Since the real interval $[-\epsilon, M]$ is convex and $0 \le \alpha(p) \le 1$, it follows that $k(p) \in [-\epsilon, M]$ for all $p \in P$. As $k(\cdot)$ is a continuous function, we conclude that $k \in G(-\epsilon, M)$ for all functions $\alpha \in C(P, [0,1])$. Substituting the function k into (47) and using (48), we obtain that

$$\begin{aligned} &\int_P \lambda(p) \cdot [c^*(p) - k(p)]\mathrm{d}\omega(p) \\ &\quad = \int_P \alpha(p)\lambda(p) \cdot [c^*(p) - c(p)]\mathrm{d}\omega(p) \ge 0. \end{aligned}$$

Keeping c fixed, the last expression forms a linear functional of α; as this functional is non-negative, it follows (e.g. Doob (1994)) that $\lambda(p) \cdot [c^*(p) - c(p)] \ge 0$ for ω-almost every point $p \in P$. As this argument can be repeated for every function $c \in G(-\epsilon, M)$, we conclude that

$$\begin{aligned} &\lambda(p)c^*(p) \ge \lambda(p)c(p) \quad \text{for all functions} \\ &c \in G(-\epsilon, M) \quad \text{at } \omega\text{-almost every point } p \in P. \end{aligned} \tag{49}$$

Now, according to (46), we have $|\lambda(p)| = 1$ for ω-almost all $p \in P$. Consider first a point $p \in P$ at which $\lambda(p) = 1$, and substitute into (49) the constant function $c(p) := M$ for all $p \in P$. Then, (49) yields $c^*(p) \ge M$ for ω-almost every $p \in P$. On the other hand, if $\lambda(p) = -1$, then (49) yields $-c^*(p) \ge -c(p)$, or $c^*(p) \le c(p)$; using the constant function $c(p) = -\epsilon$, the latter yields $c^*(p) \le -\epsilon$. Now, by assumption (1) of Lemma 4.5, $-\epsilon \le c^*(p) \le M$ for all $p \in P$. Consequently, the last two sentences imply that $c^*(p) = M$ when $\lambda(p) = 1$ and $c^*(p) = -\epsilon$ when $\lambda(p) = -1$, for ω-almost every $p \in P$. This proves statement (iii) of Lemma 4.5, and our proof concludes. □

Proof (of Theorem 4.1): The proof is based on Lemma 4.5. First, we identify the quantities listed in Lemma 4.5 with the quantities that appear in Theorem 4.1, by setting

$$Q := \{v(s) \in V : v(s) = 0 \quad \text{for all } s > 1\},$$

$$F := [0, t_f^* + 1] \subset R, \tag{50}$$

where t_f^* is the maximal time (12). Further, let P be the set given by (26). It is then easy to verify that Q is a convex subset of the Banach space $L_2^{\alpha,m}$; that F is a convex set of real numbers; and that P is a compact subset of the metric space $R^{(1+nn+mn)}$. Thus, Q, F and P fulfil the requirements of the corresponding quantities listed in Lemma 4.5.

Next, recalling the matrix D of (5), define the two functions

$$T_1(v(s), \beta) := -\beta, \text{and}$$

$$T_2(v(s), \beta, p) := y^T(s; \beta, D, v)y(s; \beta, D, v), \quad \text{where } p \in P. \tag{51}$$

As $y(s; \beta, D, v)$ is the solution of the linear differential equation (22), we have

$$y(s; \beta, D, v) = e^{\beta A' s}\left[x_0 + \int_0^s e^{-\beta A'\tau}\beta B' v(\tau)\mathrm{d}\mu(\tau)\right], \tag{52}$$

where $\mu(\tau)$ is the unit Lebesgue measure. Thus, the function T_2 satisfies requirement (3) of Lemma 4.5, and $T_2(v(s), \beta, \cdot) \in C(P, R)$ for all $(v(s), \beta) \in Q \times F$. Further, by (24), β^* is the maximal value of β over the set of all points $(v(s), \beta) \in Q \times F$ for which $y^T(s; \beta, D, v)y(s; \beta, D, v) \le M$. Whence, $-\beta^*$ is the minimal value of $T_1(v(s), \beta)$ over the same set, and, consequently, requirement (1) of Lemma 4.5 is valid as well. Finally, a direct examination shows that the functions T_1 and T_2 satisfy the requirements of Lemma 4.4. Thus, the Gateaux derivatives of T_1 and T_2 are linear in their direction, and whence assumption (2) of Lemma 4.5 is satisfied for T_1 and T_2.

A direct computation using (29) yields the Gateaux derivative of T_1,

$$\mathcal{D}T_1((v^*, \beta^*); (v, \beta) - (v^*, \beta^*)) = \beta^* - \beta.$$

For T_2, we calculate the Gateaux derivative in the direction $h = v - v^*$ with β at its maximal value β^*, namely, in the direction $h = (v, \beta) - (v^*, \beta^*)|_{\beta=\beta^*}$. Applying (29) to (51) while using (52), we obtain

$$\begin{aligned}\mathcal{D}T_2((v^*, \beta^*), (s, A', B'); (v - v^*))|_{\beta=\beta^*} &= y^T(s, A', B'; \beta^*, v - v^*)y(s, A', B'; \beta^*, v^*) \\ &\quad + y^T(s, A', B'; \beta^*, v^*)y(s, A', B'; \beta^*, v - v^*).\end{aligned} \tag{53}$$

As $y^T(s, A', B'; \beta^*, v - v^*)y(s, A', B'; \beta^*, v^*)$ is a scalar, we can write

$$\begin{aligned}y^T(s, A', B'; \beta^*, v - v^*)y(s, A', B'; \beta^*, v^*) &= [y^T(s, A', B'; \beta^*, v - v^*)y(s, A', B'; \beta^*, v^*)]^T \\ &= y^T(s, A', B'; \beta^*, v^*)y(s, A', B'; \beta^*, v - v^*).\end{aligned}$$

Substituting into (53) and using (52), this yields

$$\begin{aligned}\mathcal{D}T_2((v^*, \beta^*), (s, A', B'); (v - v^*))|_{\beta=\beta^*} &= 2y^T(s, A', B'; \beta^*, v^*)y(s, A', B'; \beta^*, v - v^*) \\ &= 2y^T(s, A', B'; \beta^*, v^*) \\ &\quad \int_0^s e^{\beta^* A'(s-\tau)}\beta^* B'(v(\tau) - v^*(\tau))\mathrm{d}\mu(\tau).\end{aligned} \tag{54}$$

Applying now Lemma 4.5 with $p := (\tau, A', B')$ and using the notation of (50), we conclude that there is a Radon probability measure ω over P and an ω-integrable function $\lambda : P \to R$ such that

$$|\lambda(p)| = 1 \text{ for } \omega\text{-almost all points } p \in P, \text{ and}$$
$$\int \lambda(p)\mathcal{D}T_2((v^*, \beta^*), p; (v, \beta) - (v^*, \beta^*))\mathrm{d}\omega(p) \ge 0. \tag{55}$$

Denoting $y^*(s) := y(s, A', B'; \beta^*, v^*)$, setting $\beta = \beta^*$, and substituting (54) into the last inequality, we obtain

$$\int_P \lambda(p)(y^*(s))^T\left\{\int_0^s e^{\beta^* A'(s-\tau)}\beta^* B'(v(\tau) - v^*(\tau))\mathrm{d}\mu(\tau)\right\} \mathrm{d}\omega(p) \ge 0. \tag{56}$$

It is convenient now to define the function

$$\eta(s, \tau) := \begin{cases} 1 & \text{for } 0 \le \tau \le s, \\ 0 & \text{otherwise,} \end{cases} \tag{57}$$

where $s \ge 0$. Then, (56) can be rewritten in the form

$$\int_P \lambda(p)(y^*(s))^T \left\{\int_0^1 e^{\beta^* A'(s-\tau)}\beta^* B'\eta(s, \tau)(v(\tau) - v^*(\tau))\mathrm{d}\mu(\tau)\right\} \mathrm{d}\omega(p) \ge 0.$$

Applying Fubini's Theorem (e.g. Rudin (1966)) to the last expression, we obtain

$$\int_0^1 \left\{ \int_P \lambda(p)(y^*(s))^T e^{\beta^* A'(s-\tau)} B' \eta(s,\tau) \mathrm{d}\omega(p) \right\} (v(\tau) - v^*(\tau)) \mathrm{d}\mu(\tau) \geq 0.$$

Defining the function

$$z^T(\tau) := \int_P \lambda(p)(y^*(s))^T e^{\beta^* A'(s-\tau)} B' \eta(s,\tau) \mathrm{d}\omega(p), \quad (58)$$

we can rewrite the last inequality as

$$\int_0^1 z^T(\tau)(v(\tau) - v^*(\tau)) \mathrm{d}\mu(\tau) \geq 0, \quad (59)$$

which must be valid for every function $v \in V$, where V is our set of input functions (21). Now, recall that $\omega(s, A', B')$ is a Radon probability measure over the space $[0,1] \times \Xi$, where Ξ is given by (25). Using the conditional measure $\omega(A', B'|s)$ and the corresponding marginal measure $\omega(s)$, we can rewrite (58) in the form

$$z^T(\tau) = \int_0^1 \int_\Xi \lambda(s, A', B')(y^*(s))^T e^{\beta^* A'(s-\tau)} B' \mathrm{d}\omega(A', B'|s) \eta(s,\tau) \mathrm{d}\omega(s).$$

In view of (57), this reduces to

$$z^T(\tau) = \int_\tau^1 \int_\Xi \lambda(s, A', B')(y^*(s))^T e^{\beta^* A'(s-\tau)} B' \mathrm{d}\omega(A', B'|s) \mathrm{d}\omega(s). \quad (60)$$

Finally, we show that (59) implies that the inequality

$$z^T(\tau) v(\tau) \geq z^T(\tau) v^*(\tau) \quad (61)$$

must be valid for μ-almost every $\tau \in [0,1]$ and for every function $v \in V$. To this end, assume, by contradiction, that there is an input function $v' \in V$ and a measurable set $\delta \subset [0,1]$ of non-zero μ measure such that $z^T(\tau)v'(\tau) < z^T(\tau)v^*(\tau)$ for all $\tau \in \delta$. Then we can form a new measurable input function $v'' \in V$ by setting

$$v''(\tau) := \begin{cases} v'(\tau) & \text{if } \tau \in \delta, \\ v^*(\tau) & \text{otherwise.} \end{cases}$$

Inserting this function for v into (59), we obtain

$$\int_0^1 z^T(\tau)(v''(\tau) - v^*(\tau)) \mathrm{d}\mu(\tau) = \int_\delta z^T(\tau)(v'(\tau) - v^*(\tau)) \mathrm{d}\mu(\tau) < 0,$$

contradicting (59). Thus, (61) must be valid, and the Theorem follows by changing τ into s. □

We provide now a somewhat simplified form of the function $z(s)$ of Theorem 4.1.

Lemma 4.10: *Let Ξ be given by* (25) *and let P be given by* (26). *Then, the function $z(s)$ of Theorem* 4.1 *can be expressed in the form*

$$z^T(s) = \int_s^1 \int_\Xi (y(\zeta, A', B'; \beta^*, v^*))^T e^{\beta^* A'(\zeta - s)} B' \mathrm{d}\omega(A', B'|\zeta) \mathrm{d}\omega(\zeta), \quad (62)$$

where $\omega(A', B', \zeta)$ is a Radon probability measure on P with the support

$$\Omega = \{(A', B', \zeta) \in \Xi \times [0,1] : y^T(\zeta, A', B'; \beta^*, v^*) y(\zeta, A', B'; \beta^*, v^*) = M\}. \quad (63)$$

Proof: We use the measure ω introduced in the proof of Theorem 4.1. In view of Lemma 4.5(iii), we have

$$\lambda(p) y^T(p; v^*, \beta^*) y(p; v^*, \beta^*) = \max_{a \in [-\epsilon, M]} \lambda(p) a \quad (64)$$

for ω-almost every $p \in P$, where P is given by (26). Recall from (55) that $\lambda(p) = \pm 1$ for ω-almost all $p \in P$. Now, when $\lambda(p) = 1$, then the right side of (64) is M, and, consequently, we must have $y^T(p; \beta^*, v^*) y(p; \beta^*, v^*) = M$. When $\lambda(p) = -1$, then the right side of (64) is ϵ, while the left side cannot be positive; hence, $\lambda(p) = -1$ is incompatible. Thus, we must have $\lambda(p) = 1$ for ω-almost every point $p \in P$. This implies that the measure ω has the support set $\Omega = \{p \in P : y^T(p; \beta^*, v^*) y(p; \beta^*, v^*) = M\}$, as given by (63). Finally, (62) follows directly from (60) by substituting $\lambda(p) = 1$ for ω-almost all $p \in P$ and renaming the variable τ to s. □

As we have seen in Corollary (4.2), the optimal input function that solves our optimisation problem 2.1 is a bang-bang function over time intervals where the function $z(s)$ of Theorem 4.1 is almost nowhere zero. In the next section, we show that, over intervals where $z(s)$ is identically zero, optimal performance can be approximated by using bang-bang input functions. Thus, bang-bang functions can be generally used when implementing the optimal solution. We complete this section with an example.

Example 4.11: Consider the one-dimensional system

$$\dot{x}(t) = ax(t) + u(t), \quad (65)$$

where the time constant a is subject to the uncertainty $1.2 \leq a \leq 1.4$. The system has the input bound $|u(t)| \leq 2$ for all t, and the initial condition is $x(0) = 1$. We set the bound $M := 25$, so the objective is to find an input function $u^*(t)$ that keeps the state amplitude below the bound $x^2(t) \leq 25$ (i.e. $|x(t)| \leq 5$) for the longest time, irrespective of the value of a within its uncertainty range. We show next that, in this case, $z(s) \neq 0$ for almost all $s \in [0,1]$. Thus, by Corollary 4.2, the optimal

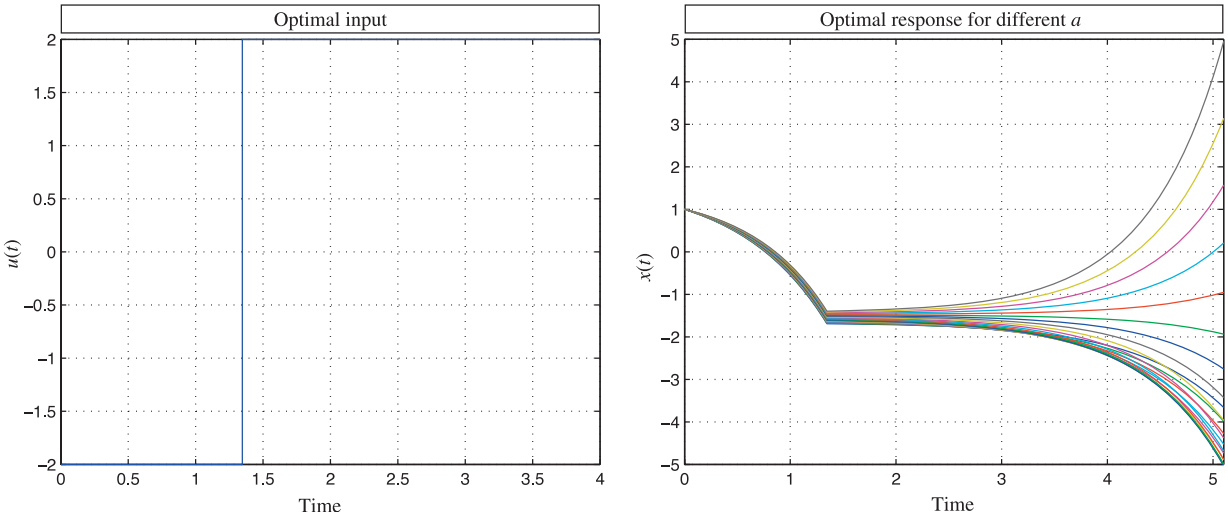

Figure 2. An example.

input function $u^*(t)$ is a bang-bang function, as depicted below. The maximal time during which all samples of the system can be kept below the prescribed error bound is $t_f^* = 5.08$ s. The trajectories for various values of a are depicted in Figure 2. To show that the optimal input function is a bang-bang function in this case, note first that the system cannot rebound to lower valued states after reaching the state $|x(t)| = 5$. Indeed, consider the error function

$$e(t) = x^2(t).$$

Using the system equation (65), we get

$$\dot{e}(t) = x(t)\dot{x}(t) = x^2(t)a + x(t)u(t) = x(t)[x(t)a + u(t)].$$

If $e(t) = 25$, we clearly need $\dot{e}(t) \leq 0$ for the error not to worsen. When $e(t) = 25$, we have either $x(t) = 5$ or $x(t) = -5$. For $x(t) = 5$, we obtain $\dot{e}(t) = x(t)[x(t)a + u(t)] = 5[5a + u(t)] > 0$ for all possible values of a and of $u(t)$. Also, for $x(t) = -5$, we have $\dot{e}(t) = -5[-5a + u(t)] > 0$ for all possible values of a and $u(t)$. Thus, once the system reaches $e(t) = 25$, it has reached the terminal time, since the error can only continue to grow. Hence, for any value of a, the process terminates when the corresponding trajectory hits the error bound M. In other words, any trajectory meets the error bound only once: at the terminal time $s = 1$. Thus, in view of Lemma 4.10, the support set of the function $z(s)$ in this case is given by the following (in this example, $B' = 1$ always).

$$\Omega = \{(a', 1, \zeta) \in [1.2, 1.4] \times \{1\} \times \{1\} : x^2(1, a', v^*) = M\}. \tag{66}$$

Note that Ω cannot be empty here, since that would imply that x^2 does not meet the bound M on the scaled time interval $[0, 1]$, contradicting what we have concluded in the previous paragraph.

Substituting the support set (66) into (62), we obtain

$$z(s) = \int_{1.2}^{1.4} x(1, a', v^*) e^{\beta^* a'(1-s)} \, \mathrm{d}\omega(a').$$

Let us now expand the exponential in the integrand into a series and integrate; this yields

$$z(s) = p_0 + p_1(1-s) + p_2(1-s)^2 + \cdots + p_m(1-s)^m + \ldots, \tag{67}$$

where

$$p_m = \int_{1.2}^{1.4} x(1, a', v^*) \frac{(\beta^* a')^m}{m!} \, \mathrm{d}\omega(a').$$

As the integrand includes the power $(a')^m$, the equality $p_m = 0$ for all $m = 0, 1, 2, \ldots$ would imply that $x(1, a, v^*) = 0$ almost everywhere with respect to the measure $\omega(a')$, contradicting the support (66). Thus, at least one of the coefficients of the power series (67) is not zero, and whence $z(s) \neq 0$ almost everywhere on the interval $(0, 1)$. By Corollary 4.2, this proves that the optimal input function is a bang-bang function in this case.

4.3 *Implications on digital control*

It is interesting to compare the optimal input function derived here with the 'zero-order hold' policy commonly employed in digital control, namely, the policy of keeping the input signal constant between feedback sampling instants. Referring to Example 4.11, we can calculate the best 'zero-order hold' for this case,

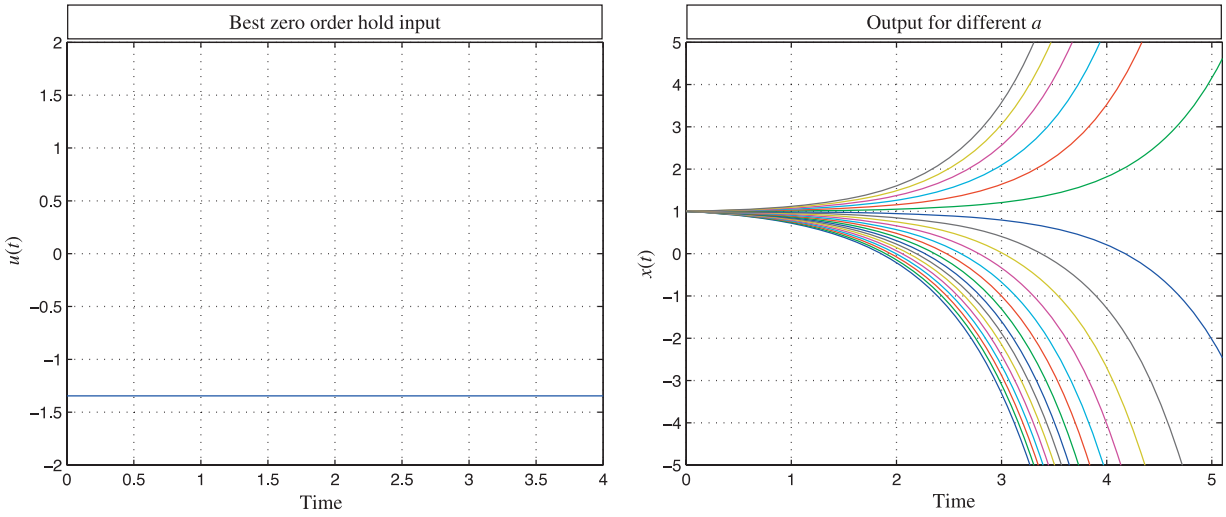

Figure 3. Comparing to a zero order hold.

namely, the constant input that will keep the output below the error bound for the longest time. We obtain that the best constant input value here is $u = -1.38$; this input yields the time $t_f = 3.26$ s during which the output error is within the prescribed bound M for all possible uncertainty values of a. The output functions for various values of a generated by this constant input are depicted in Figure 3. As we can see, t_f is substantially shorter than the optimal time $t_f^* = 5.08$ s obtained in Example 4.11.

5. Bang-bang approximation

We explore now a methodology that provides convenient means for achieving optimal, or nearly optimal, performance when controlling a system under conditions of feedback failure. Specifically, we show that the optimal time t_f^* of (12) can be achieved, or nearly achieved (depending on problem parameters), by using bang-bang input functions. As mentioned earlier, bang-bang functions are convenient for calculation and implementation, since they are determined solely by their switching times. We underscore the fact that, when the optimal input function is itself not a bang-bang function, the bang-bang functions derived below do not necessarily approximate the optimal input function; instead, they are designed to yield approximate optimal performance. Of course, the latter is the aspect most pertinent in applications.

Recall that our objective is to control the system Σ of (3) under the perturbations described by (4). The optimal input function $u^*(t)$ keeps the state trajectory of Σ below the bound M for the longest possible time t_f^* that is compatible with all perturbation matrices $D \in \Delta$. When approximating optimal performance, we allow the state trajectory of Σ to slightly exceed the bound M. Specifically, let $x(t, D, u^*)$ be the state trajectory of Σ generated by the optimal input function $u^*(t)$ for a particular uncertainty matrix $D \in \Delta$. We are seeking a bang-bang input function $u^{\pm}(t)$ which, when applied to Σ, generates a state trajectory $x(t, D, u^{\pm})$ that deviates only slightly from $x(t, D, u^*)$ for all $t \in [0, t_f^*]$ and for all $D \in \Delta$. The next statement indicates that there is such an input function.

Theorem 5.1: *Let Σ be a system that satisfies the conditions of Theorem* 3.9, *and let t_f^* be the optimal time and $u^*(t)$ the optimal input function of Theorem* 3.9. *Then, for every $\epsilon > 0$, there is a bang-bang input function $u^{\pm} \in U$ for which the following are true.*

(i) *$u^{\pm}$ has only finite number of switches, and*
(ii) *The state trajectory $x(t, D, u^{\pm})$ of Σ created by $u^{\pm}$ satisfies $\|x(t, D, u^*) - x(t, D, u^{\pm})\| < \epsilon$ for all $t \in [0, t_f^*]$ and all $D \in \Delta$.*

Proof: Fix a real number $\epsilon > 0$. Recall that all input functions $u(t)$ of Σ are bounded by K, that $t_f^* < \infty$ by Theorem 3.9, and that all perturbation matrices $D \in \Delta$ have entries of magnitude not exceeding $d > 0$. Recall that $D = (D_A, D_B)$, $\Delta = (\Delta_A, \Delta_B)$ and that $A' = A + D_A$ and $B' = B + D_B$ in (3), where $D_A \in \Delta_A$ and $D_B \in \Delta_B$. Now, let $\eta > 0$ be a real number (to be selected later). By the uniform continuity of the function $e^{A't}$, there is a real number $\delta(\eta) > 0$ such that the function

$$\mu(t', t) := e^{-A't'} - e^{-A't}$$

satisfies $\|\mu(t', t)\| \leq \eta$ for all $t', t \in [0, t_f^*]$ satisfying $|t' - t| < \delta(\eta)$. Also, let

$$\beta := \sup\{\|B + D_B\| : D_B \in \Delta_B\},$$

and let

$$N := \sup\{e^{A't} : D_A \in \Delta_A, t \in [0, t_f^*]\};$$

here, N exists due the fact that all involved quantities are bounded. Let $0 < \gamma \leq \delta(\eta)$ be any number for which $s := t_f^*/\gamma$ is an integer. We build a partition of the interval $[0, t_f^*]$ into segments of length γ, namely, the partition determined by the points $0, \gamma, 2\gamma, \ldots, (s-1)\gamma$. Recalling that the input function $u(t)$ of Σ is an m-dimensional vector with each component bounded by K, we define a bang-bang input function $u^{\pm}(t) := (u_1^{\pm}(t), u_2^{\pm}(t), \ldots, u_m^{\pm}(t))^T$ through its components as follows: for each component $u_i^{\pm}$, $i = 1, 2, \ldots, m$, select in each interval $[q\gamma, (q+1)\gamma]$ a switching time θ_{qi}, where $q = 0, 1, 2, \ldots, s-1$, $i = 1, 2, \ldots, m$, and set

$$u_i^{\pm}(t) := \begin{cases} +K & \text{for } t \in [q\gamma, \theta_{qi}), \\ -K & \text{for } t \in [\theta_{qi}, (q+1)\gamma), \end{cases}$$

where the value of θ_{qi} is selected to satisfy the equality

$$\int_{q\gamma}^{(q+1)\gamma} u_i^*(\tau)\mathrm{d}\tau = K\int_{q\gamma}^{\theta_{qi}} \mathrm{d}\tau - K\int_{\theta_{qi}}^{(q+1)\gamma} \mathrm{d}\tau$$
$$= K[2(\theta_{qi} - q\gamma) - \gamma].$$

Note that θ_{qi} exists for all $q = 0, 1, 2, \ldots, s-1$ and all $i = 1, 2, \ldots, m$ due to the fact that $|u_i^*(t)| \leq K$ for all $t \geq 0$. Then, we obtain the equality

$$\int_{q\gamma}^{(q+1)\gamma} [u_i^*(\tau) - u_i^{\pm}(\tau)]\mathrm{d}\tau = 0, \quad q = 0, 1, 2, \ldots, s-1. \tag{68}$$

Finally, let $x^{\pm}(t)$ be the state trajectory generated by the system Σ when driven by the input function $u^{\pm}(t)$, and let $x^*(t)$ be the trajectory induced by the optimal input function $u^*(t)$. Noting that the perturbation matrix D is the same in both cases (we are comparing the two input functions for the same system), one obtains (using (68))

$$||x^*(t) - x^{\pm}(t)|| = \left\| e^{A't}\left[x_0 + \int_0^t e^{-A'\tau}B'u^*(\tau)\mathrm{d}\tau\right] - e^{A't}\left[x_0 + \int_0^t e^{-A'\tau}B'u^{\pm}(\tau)\mathrm{d}\tau\right]\right\|$$
$$= \left\| e^{A't}\int_0^t e^{-A'\tau}B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau \right\|$$
$$\leq N\left\| \int_0^t e^{-A'\tau}B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau \right\|.$$

Now, let q be the largest integer for which $q\gamma \leq t$; then, continuing from the last expression, we can write

$$||x^*(t) - x^{\pm}(t)||$$
$$\leq N\left\| \sum_{r=0}^{q-1}\int_{r\gamma}^{(r+1)\gamma} e^{-A'\tau}B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau + \int_{q\gamma}^{t} e^{-A'\tau}B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau\right\|$$
$$\leq N\left\{\left\| \sum_{r=0}^{(q-1)}\left[e^{-A'r\gamma}B'\int_{r\gamma}^{(r+1)\gamma}[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau + \int_{r\gamma}^{(r+1)\gamma}\mu(\tau, r\gamma)B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau\right]\right\| + \left\| \int_{q\gamma}^{t} e^{-A'\tau}B'[u^*(\tau) - u^{\pm}(\tau)]\mathrm{d}\tau\right\|\right\}$$
$$\leq N\left\{ \sum_{r=0}^{(q-1)}\int_{r\gamma}^{(r+1)\gamma} \|\mu(\tau, r\gamma)\|\|B'\|[||u^*(\tau)|| + \|u^{\pm}(\tau)||]\mathrm{d}\tau + \int_{q\gamma}^{t}\|e^{-A'\tau}\|\|B'\|[\|u^*(\tau)\| + \|u^{\pm}(\tau)\|]\mathrm{d}\tau\right\}$$
$$\leq 2KN\beta[\eta t_f^* + N\gamma].$$

We choose now the value of η so that $2KN\beta\eta t_f^* < \epsilon/2$. Then, we choose $0 < \gamma \leq \min\{\delta(\eta), \epsilon/(4KN^2\beta)\}$ so that t_f^*/γ is an integer. For these selections, we obtain $\|x^*(t) - x^{\pm}(t)\| < \epsilon$ for all $t \in [0, t_f^*]$, and our proof concludes. □

Of course, the bang-bang input function $u^{\pm}(t)$ that replaces the optimal input function $u^*(t)$ is independent of the perturbation matrices. The cost of making ϵ smaller is an increase in the number of switches of the bang-bang function $u^{\pm}(t)$. In many practical applications, a good approximation of optimal performance can be achieved by a bang-bang input function with a relatively low number of switches, as the following example indicates:

Example 5.2: Consider the one-dimensional system $\dot{x}(t) = ax(t) + u(t)$, where the time constant a is subject to the uncertainty $1.2 \leq a \leq 1.4$. The system has the input bound $|u(t)| \leq 2$ for all t, and the initial condition $x(0) = 1$. The objective is to find an input function $u^*(t)$ that keeps the state amplitude below the bound $x^2(t) \leq 1.96$ for the longest period of time, irrespective of the value a adopts within its uncertainty range. The optimal input is shown in the left plot of Figure 4, and the corresponding state trajectories for different values of a are plotted on the right side of Figure 4, with $M = 1.96$ and $t_f = 3.7$.

As can be seen from the plot in Figure 4, the solution is bang-bang only over the time span $[0, 1.27]$.

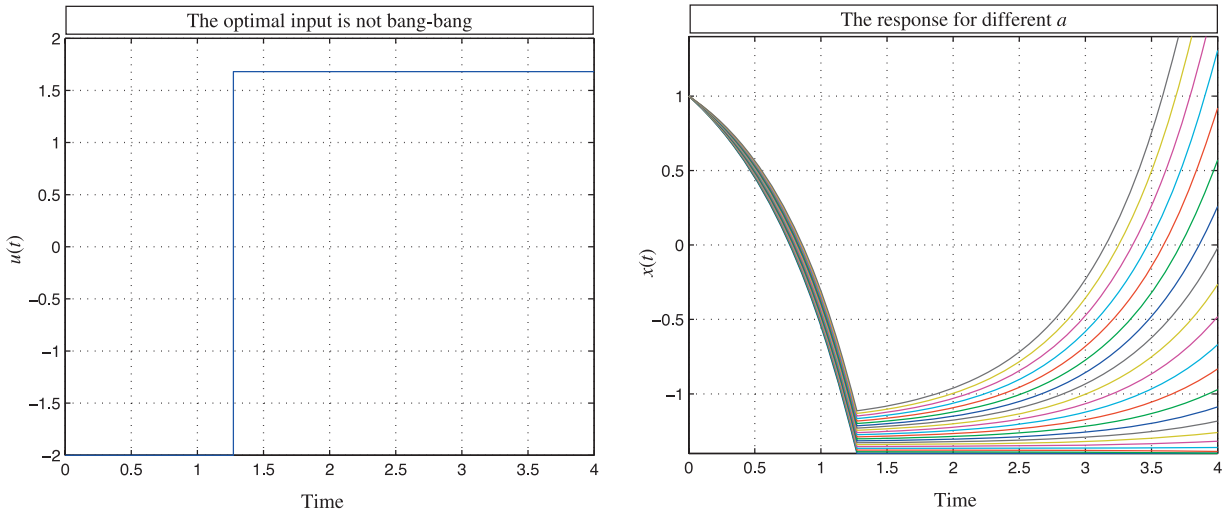

Figure 4. Non bang-bang optimal input.

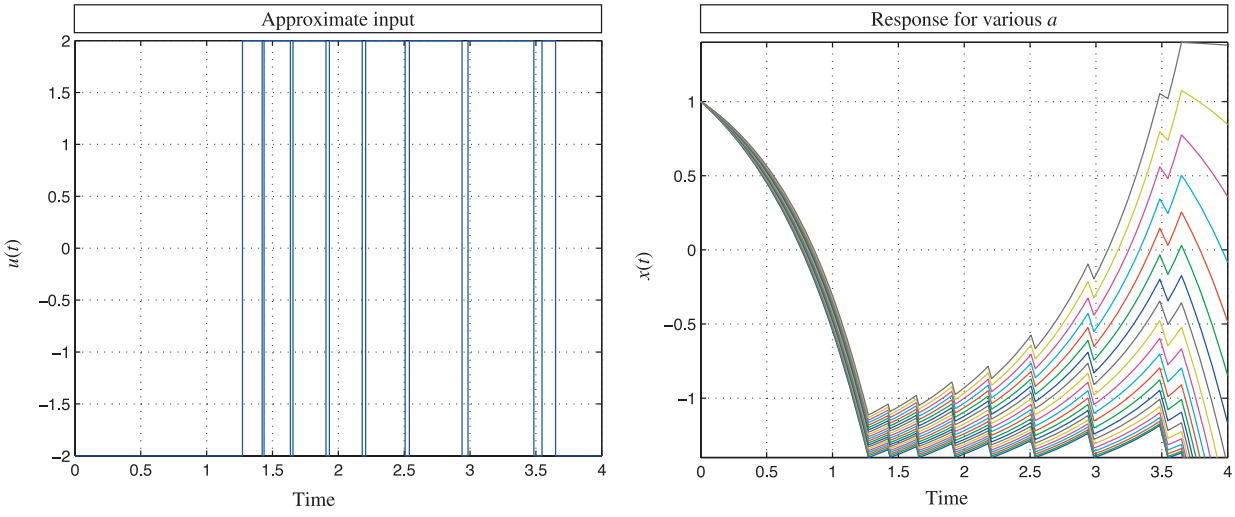

Figure 5. Bang-bang approximation with 16 switchings.

For the remaining time, the input switches to the value 1.67, not one of the values ± 2 that a bang-bang function would assume. The maximal time here is $t_f^* = 3.7\,\mathrm{s}$.

The graphs in Figure 5 demonstrate a bang-bang input function with 16 switches that approximates optimal performance. As can be seen from the accompanying state trajectories, the same maximal time of 3.7 s can be kept when the error bound is allowed to increase from $M = 1.96$ to $M = 2.01$. □

To summarise in brief terms, we have seen that bang-bang input functions can always be used to achieve optimal, or nearly optimal, performance, when the objective is to keep errors below a specified bound for the longest time without feedback.

References

Bourbaki, N. (1987), *Topological Vector Spaces*, Berlin: Springer-Verlag.

Doob, J.L. (1994), *Measure Theory*, New York: Springer-Verlag.

Evans, L.C., and Gariepy, R.F. (1992), *Measure Theory and Fine Properties of Functions*, Boca Raton, FL: CRC Press.

Gamkrelidze, R.V. (1965), 'On Some Extremal Problems in the Theory of Differential Equations with Applications to the Theory of Optimal Control', *Journal of the Society for Industrial and Applied Mathematics, Series A, on Control*, 3, 106–128.

Halmos, P.R. (1982), *A Hilbert Space Problem Book*, New York: Springer-Verlag.

Kelendzheridze, D.L. (1961), 'On the Theory of Optimal Pursuit', *Soviet Mathematics Doklady*, 2, 654–656.

Liusternik, L.A., and Sobolev, V.J. (1961), *Elements of Functional Analysis*, New York: Frederick Ungar.

Luenberger, D.G. (1969), *Optimization by Vector Space Methods*, New York: Wiley.

Montestruque, L.A., and Antsaklis, P.J. (2004), 'Stability of Model-based Networked Control Systems with Time-varying Transmission Times', *IEEE Transactions on Automatic Control*, 49(9), 1562–1572.

Nair, G.N., Fagnani, F., Zampieri, S., and Evans, R.J. (2007), 'Feedback Control Under Data Rate Constraints: An Overview', *Proceedings of the IEEE*, 108–137.

Neustadt, L.W. (1966), 'An Abstract Variational Theory with Applications to a Broad Class of Optimisation Problems I, General Theory', *SIAM Journal on Control*, 4, 505–527.

Neustadt, L.W. (1967), 'An Abstract Variational Theory with Applications to a Broad Class of Optimisation Problems II, Applications', *SIAM Journal on Control*, 5, 90–137.

Panetta, J.C., and Fister, K.R. (2003), 'Optimal Control Applied to Competing Chemotherapeutic Cell-kill Strategies', *SIAM Journal of Applied Mathematics*, 63(6), 1954–1971.

Pontryagin, L.S., Boltyansky, V.G., Gamkrelidze, R.V., and Mishchenko, E.F. (1962), *The Mathematical Theory of Optimal Processes*, New York, London: Interscience Publishers John Wiley & Sons Inc.

Rudin, W. (1966), *Real and Complex Analysis*, New York: McGraw-Hill.

Warga, J. (1972), *Optimal Control of Differential and Functional Equations*, New York: Academic Press.

Willard, S. (1970), *General Topology*, Reading, MA: Addison-Wesley.

Young, L.C. (1969), *Lectures on the Calculus of Variations and Optimal Control Theory*, Philadelphia: W.B. Saunders.

Zeidler, E. (1985), *Nonlinear Functional Analysis and its Applications III*, New York: Springer-Verlag.

Zhivogyladov, P.V., and Middleton, R.H. (2003), 'Networked Control Design for Linear Systems', *Automatica*, 39, 743–750.