Closest Feasible Points Invariance: A System Property to Characterize Systems with Actuator Saturation

Masoud Soroush, *Senior Member*, *IEEE; Fellow, AIChE*

***Abstract*— This article introduces a new system property called the closest feasible points (CFP) invariance to characterize systems with actuator saturation. Systems that possess this invariance property include diagonal matrices, completely decentralized (completely decoupled) linear dynamical systems, and dynamical systems with a nonsingular input-independent characteristic (decoupling) matrix that can be made diagonal with row or column rearrangements. However, a single-input single-output system may not possess this property. This system property has implications and applications in control, where actuator saturation is common. For example, when an actuator saturates, the closed-loop performance of a CFP non-invariant plant under a controller that is not a solution to a constrained optimal control problem, may degrade considerably. The definition of this property guides the derivation of optimal CFP non-invariance compensators that decrease the control performance degradation gracefully in CFP non-invariant plants. This work characterizes the plants for which clipping and direction preservation of controller outputs are optimal.**

I. Introduction

An invariant is a property of a class of mathematical objects that does not change under certain transformations applied to the objects. It usually reflects an intrinsic property of the objects. Examples are as follows. The observability, detectability, controllability, and stabilizability of linear time-invariant dynamical systems are invariant under invertible linear coordinate transformations [1]. The notion of invariance has also been defined for sets. A set is said to be positively invariant with respect to a dynamical system, if every solution of the dynamical system originating inside the set is globally defined and stays within the set at every time instant [2]. Controllability, observability, and stabilizability of linear systems are invariant with respect to expansion-contraction processes under certain conditions [3–6].

When a control signal (controller output), c, is sent to actuators (Fig. 1), the actuators implement the control signal as it is, only if the control signal is within the lower and upper limits of the actuators. Otherwise, at least an actuator saturates; that is, the actuator clips the control signal component before applying the plant input (u) corresponding to the signal, to the plant. In this case, the control system performance may degrade significantly due to two phenomena: (a) integral windup when the controller is dynamical; and (b) the implemented plant input corresponding to the clipped control signal is not 'optimal'. The former phenomenon is caused by the state variables of the dynamical controller not being properly informed of the actual controller action (plant input) applied to the plant under control [7, 8]. To decrease the former control performance degradation, anti-windup compensators have been proposed to properly inform the states of a controller, of the actual controller action that the plant under control is subjected to [7, 8]. The latter phenomenon is due to the plant output response to the plant input corresponding to $\mathrm{sat}(c)$ not being 'closet' to the plant output response to the control signal, c. To address the latter control performance degradation, compensators have been proposed [9, 10]. Also, relevant to this work are the efforts that have been made to characterize the class of dynamical systems that can benefit most from constrained optimal control (e.g., model predictive control) [11, 12].

This paper introduces a new system property called the closest feasible points (CFP) invariance that allows for the characterization of the systems that suffer from the latter control performance degradation. Indeed, this work describes a projection operation on a closed hyperrectangle, which is a convex compact subset of the Euclidian space. The projection operation is in general non-invariant under a transformation S (since the norm is not invariant under transformation), but this work identifies those transformations with respect to which the projection operation is invariant. Several implications and applications of this system property are considered. For continuous-time dynamical systems that do not possess this invariance property, an optimal CFP non-invariance compensator is proposed. The ability of the compensators to gracefully decrease control quality degradation in the presence of actuator saturation is shown via numerical simulations of an example.

Section II describes the property. Section III applies special cases of the definition to three classes of systems to determine the subclasses under which CFPs are invariant. Section IV presents a CFP non-invariance (CFPN) compensator for continuous-time dynamical systems, and it compares the performances of the CFPN compensator with clipping and direction preservation, via numerical simulations.

II. Closest Feasible Points Invariance

Let $\Omega = \{u |\, u_{i,min} \le u_i \le u_{i,max},\ i = 1, \cdots, m\} \subset \mathbb{R}^m$ — where $u_{i,min}$ and $u_{i,max}$ $(u_{i,min} < u_{i,max})$, $i = 1, \cdots, m$, are finite scalars — be the set of all feasible values that the plant input u can take (Ω is a convex compact subset of the Euclidean space), $c \in \mathbb{R}^m$ be the control signal, and $u^I \in \Omega$ represent the feasible plant input that is 'closest' to the control signal c, in the input hyperspace.

Fig. 2 graphically explains the CFP invariance. Let $S: \mathbb{R}^m \to \mathbb{R}^q$ represent a system, where $m \le q$. If for every control signal $c \in \mathbb{R}^m$, the response of the system S to the u^I; i.e., $S * u^I$, is 'closest' (in terms of a norm) to the response of

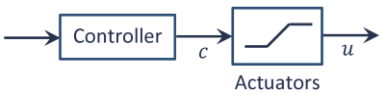

Fig.1. Each actuator is assumed to be linear (when unsaturated) and static.

* This work was supported by the U.S. National Science Foundation under Grant No. CBET-1704915.

M. Soroush is with the Department of Chemical and Biological Engineering, Drexel University, Philadelphia, PA 19104, USA (e-mail: ms1@drexel.edu).

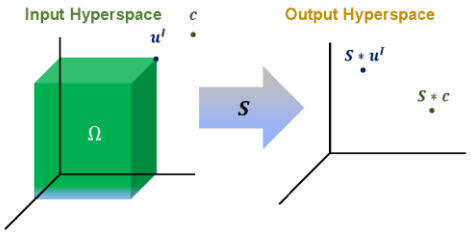

Fig. 2. Graphical description of the CFP invariance. u^l is the closest feasible plant input to the control signal $c \in \mathbb{R}^m$. The response of the system to u^l, $S*u^l$, is closest to the response of the system to the control signal c, $S*c$, for every $c \in \mathbb{R}^m$. The hyperrectangle represents the convex compact set of all feasible inputs, Ω.

the system to the control signal c; i.e., $S*c$, then the closest feasible points are said to be invariant under the system S. Here, "$*$" represents an operator. The next definition describes this property mathematically.

Definition 1: Let $\|c\|_p$ denote a p-norm of a vector c and $S: \mathbb{R}^m \to \mathbb{R}^q$ represent a system, where $m \le q$. If

$$\arg\left\{\min_{u\in\Omega}\|u-c\|_p\right\} = \arg\left\{\min_{u\in\Omega}\|S*u - S*c\|_p\right\} \quad (1)$$

for every $c \in \mathbb{R}^m$, then the p-norm closest feasible points are said to be invariant under the system S. Here,

$$\arg\left\{\min_{u\in\Omega}\|u-c\|_p\right\}$$

represents the feasible plant input that is p-norm closest to the control signal c, in the input hyperspace.

III. Application to Three Classes of Systems

Definition 1 with different types of norms can be applied to different classes of systems to identify the subclasses that possess the corresponding CFP invariance property. In this section, three classes of systems are considered, and to each class the definition with a specific type of the norm is applied.

A. Linear Static Square Systems

Theorem 1: Let the system $S: \mathbb{R}^m \to \mathbb{R}^q$ represent a full rank, $q \times m$ ($m \le q$) matrix and the p norm in (1) be the L^2 norm (Euclidean norm). For every $c \in \mathbb{R}^m$,

$$\arg\left\{\min_{u\in\Omega}\|Su-Sc\|_2\right\} = \mathrm{sat}(c) = \arg\left\{\min_{u\in\Omega}\|u-c\|_2\right\} \quad (2)$$

where

$$\mathrm{sat}(c) = \begin{bmatrix} \mathrm{sat}_1(c_1) \\ \vdots \\ \mathrm{sat}_m(c_m) \end{bmatrix}, \quad (3)$$

$$\mathrm{sat}_i(c_i) \stackrel{\text{def}}{=} \begin{cases} u_{i,min}, & c_i \le u_{i,min} \\ c_i, & u_{i,min} < c_i < u_{i,max} \\ u_{i,max}, & u_{i,max} \le c_i \end{cases}, \ i = 1,\cdots,m$$

if the positive definite matrix S^TS is diagonal or can be made diagonal with row or column rearrangements.

Proof: Let:

$$u^* = \arg\left\{\min_{u\in\Omega}\|Su-Sc\|_2\right\} = \arg\left\{\min_{u\in\Omega}\|Su-Sc\|_2^2\right\}.$$

In the case that the matrix S^TS is diagonal, the Lagrange function is:

$$L = \sum_{i=1}^{m}[S^TS]_{ii}(u_i-c_i)^2 + \sum_{i=1}^{m}\mu_i(u_i-u_{i,max}) + \sum_{i=1}^{m}\tilde{\mu}_i(-u_i+u_{i,max}) \quad (4)$$

where $[S^TS]_{ij}$ is the ith-row jth-column element of the matrix

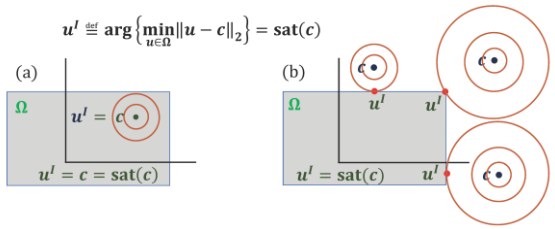

Fig. 3. $\mathrm{sat}(c)$ is the closest feasible point to c for every $c \in \mathbb{R}^m$. (a) $c \in \Omega$, (b) $c \notin \Omega$.

S^TS, and μ_i, $\tilde{\mu}_i \ge 0$, $i = 1,\cdots,m$ are the Lagrange multipliers. The necessary conditions of optimality (Karush–Kuhn–Tucker conditions) [13] applied to this constrained minimization are:

$$2[S^TS]_{ii}(u_i^* - c_i) + \mu_i - \tilde{\mu}_i = 0, \quad i = 1,\cdots,m \quad (5)$$

$$\mu_i, \tilde{\mu}_i \ge 0, \quad i = 1,\cdots,m \quad (6)$$

$$\mu_i(u_i^* - u_{i,max}) = 0, \quad i = 1,\cdots,m \quad (7)$$

$$\tilde{\mu}_i(-u_i^* + u_{i,min}) = 0, \quad i = 1,\cdots,m \quad (8)$$

$$u_{i,min} \le u_i^* \le u_{i,max}, \quad i = 1,\cdots,m \quad (9)$$

As the Hessian matrix of the Lagrange function is positive definite ($[S^TS]_{ii} > 0$, $i = 1,\cdots,m$), the conditions of (5)–(9) are the necessary and sufficient conditions of optimality. Conditions of (7) and (8) indicate that for every i, μ_i and $\tilde{\mu}_i$ cannot be nonzero simultaneously.

- If $u_{i,min} < c_i < u_{i,max}$, according to (5), (6), (7) and (8) $\mu_i = \tilde{\mu}_i = 0$ and $u_i^* = c_i$.
- If $u_{i,max} \le c_i$, according to (5), (6), (7) and (8) $\mu_i \ne 0$, $\tilde{\mu}_i = 0$, and $u_i^* = u_{i,max}$.
- If $c_i \le u_{i,min}$, according to (5), (6), (7) and (8) $\mu_i = 0$, $\tilde{\mu}_i \ne 0$, and $u_i^* = u_{i,min}$.

Therefore, $u_i^* = \mathrm{sat}_i(c_i), i = 1,\cdots,m$. In other words:

$$\arg\left\{\min_{u\in\Omega}\|Su-Sc\|_2\right\} = u^* = \mathrm{sat}(c) \quad (10)$$

when S^TS is diagonal. (10) implies that:

$$\arg\left\{\min_{u\in\Omega}\|u-c\|_2\right\} = \arg\left\{\min_{u\in\Omega}\|Iu-Ic\|_2\right\} = \mathrm{sat}(c)$$

where I is the identity matrix. Q.E.D.

Given a control signal c, the locus of the plant inputs u that yield a specific value of $\|u-c\|_2$ is a hypersphere (Fig.3). However, the locus of plant inputs u that yield a specific value of $\|Su-Sc\|_2$ is in general a m-dimensional ellipsoid (hyperellipsoid) (Fig.4). The semiaxes of the hyperellipsoid are given by $s_i = \sigma_i^{-0.5}\vartheta_i$, $i = 1,\cdots,m$ (Fig.4) [14], where $\vartheta_1,\cdots,\vartheta_m$ are the eigenvectors of S^TS; and $\sigma_1,\cdots,\sigma_m$ are the singular values of S. In other words, the eigenvectors determine directions of the semiaxes and the eigenvalues determine lengths of the semiaxes. These imply that the identity of (2) holds if the eigenvectors of S^TS are parallel to the standard basis vectors [14]. Recall that the eigenvectors of S^TS are orthogonal, as S^TS is a positive definite, symmetric matrix. As S and I are both full rank, Ω is a Chebychev set with respect to both norms in (2).

B. Control-Affine Nonlinear Continuous-Time Dynamical Square Systems

Let S be a control-affine nonlinear continuous-time dynamical system in the form:

$$S: \begin{cases} \dfrac{dx(t)}{dt} = f(x(t)) + g(x(t))u(t), & u \in \Omega \subset \mathbb{R}^m \\ y(t) = h(x(t)) \end{cases} \quad (11)$$

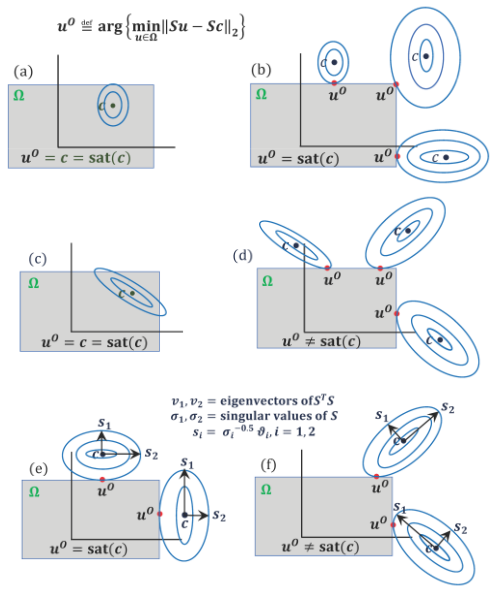

Fig. 4. The response of the system S to $u^I = \text{sat}(c)$ is closest to the response of the system to c for every for every $c \in \mathbb{R}^m$, if the eigenvectors of S^TS are parallel to the standard basis vectors. In other words, $u^O = u^I$, if the eigenvectors of S^TS are parallel to the standard basis vectors.

where $y \in \mathbb{R}^q$, $x \in \mathbb{R}^n$, and $m \le q$. The following assumptions are made: the vector fields $g_1(x), \cdots, g_m(x)$, $h(x)$, and $f(x)$ are smooth, where $g_j(x)$ represents the jth column of the matrix $g(x)$; every system output y_j, $j = 1, \cdots, q$, has a finite relative order (degree), r_j, which is the smallest integer for which the row vector $L_g L_f^{r_j-1} h_j(x) \neq 0$; and the characteristic (decoupling) matrix of the system:

$$\boldsymbol{C}(x) = \begin{bmatrix} L_g L_f^{r_1-1} h_1(x) \\ \vdots \\ L_g L_f^{r_q-1} h_q(x) \end{bmatrix} \tag{12}$$

is full rank. Here, L_f and L_{g_j} are Lie derivatives (in the directions of the vectors f and g_j, respectively).

For a control-affine nonlinear continuous-time dynamical system in the form of (11), consider the p norm:

$$\|c(t)\|_p \stackrel{\text{def}}{=} \sqrt{\sum_{\ell=1}^m \int_t^{t+T} |c_\ell(\tau)|^2 d\tau} \tag{13}$$

where the time horizon T is sufficiently small.

Theorem 2: For a control-affine nonlinear continuous-time dynamical system in the form of (11),

$$\arg\left\{\min_{u(\tau)\in\Omega} \|S * u(t) - S * c(t)\|_p\right\} = \arg\left\{\min_{u(\tau)\in\Omega} \|u(t) - c(t)\|_p\right\}$$

where $\tau \in [t, t+T]$, if $\boldsymbol{C}^T\boldsymbol{C}$ can be made diagonal with column or row rearrangements; that is, if the eigenvectors of $\boldsymbol{C}^T\boldsymbol{C}$ are parallel to the standard basis vectors.

Proof: For the system of (11), as T is sufficiently small, given the value of the vector of state variables at the present time instant t, $x(t)$, the system output responses to $c(t)$ and $u(t)$, denoted by $\bar{y}(\tau)$ and each $y(\tau)$, respectively, are obtained by using truncated Taylor series expansions of the system output responses around the current time, t:

$$\bar{y}_i(\tau) = \sum_{l=0}^{r_i} L_f^l\, h_i(x(t)) \frac{(\tau-t)^l}{l!} + \frac{(\tau-t)^{r_i}}{r_i!} L_g L_f^{r_i-1} h_i(x(t)) c(t) + \text{h. o. t.}, \quad i = 1, \cdots, q \tag{14}$$

$$y_i(\tau) = \sum_{l=0}^{r_i} L_f^l\, h_i(x(t)) \frac{(\tau-t)^l}{l!} + \frac{(\tau-t)^{r_i}}{r_i!} L_g L_f^{r_i-1} h_i(x(t)) u(t) + \text{h. o. t.}, \quad i = 1, \cdots, q \tag{15}$$

where $\tau \in [t, t+T]$. Using (14) and (15),

$$\min_{u(\tau)\in\Omega} \|S * u(t) - S * c(t)\|_p = \min_{u(\tau)\in\Omega} \|y(t) - \bar{y}(t)\|_p = \min_{u(\tau)\in\Omega} \left\{ \sqrt{\sum_{i=1}^q \left[L_g L_f^{r_i-1} h_i(x(t)) [u(t) - c(t)] \sigma_i \right]^2} \right\} \tag{16}$$

$$= \min_{u(\tau)\in\Omega} \left\| \boldsymbol{Q}\boldsymbol{C}(x(t)) u(t) - \boldsymbol{Q}\boldsymbol{C}(x(t)) c(t) \right\|_2$$

where $\boldsymbol{Q} = \text{diag}\{\sigma_i\}$,

$$\sigma_i = \sqrt{\int_t^{t+T} \left(\frac{(\tau-t)^{r_i}}{r_i!} \right)^2 d\tau} = \frac{1}{r_i!} \frac{T^{r_i+0.5}}{\sqrt{2r_i+1}}, \quad i = 1, \cdots, q$$

When $\boldsymbol{C}^T\boldsymbol{C}$ can be made diagonal with column or row rearrangements; that is, the eigenvectors of $\boldsymbol{C}^T\boldsymbol{C}$ are parallel to the standard basis vectors, according to Theorem 1:

$$\arg\left\{ \min_{u(\tau)\in\Omega} \left\| \boldsymbol{Q}\boldsymbol{C}(x(t)) u(t) - \boldsymbol{Q}\boldsymbol{C}(x(t)) c(t) \right\|_2 \right\} = \text{sat}(c(t)) \tag{17}$$

as $\boldsymbol{Q}$ is diagonal. Also,

$$\|u(t) - c(t)\|_p = \sqrt{\sum_{\ell=1}^m \int_t^{t+T} |u_\ell(\tau) - c_\ell(\tau)|^2 d\tau} = \sqrt{\sum_{\ell=1}^m T |u_\ell(t) - c_\ell(t)|^2} = \sqrt{T} \|u(t) - c(t)\|_2$$

Thus,

$$\min_{u(\tau)\in\Omega} \|u(t) - c(t)\|_p = \min_{u(\tau)\in\Omega} \sqrt{T} \|u(t) - c(t)\|_2 = \min_{u(\tau)\in\Omega} \|u(t) - c(t)\|_2 = \min_{u(\tau)\in\Omega} \|Iu(t) - Ic(t)\|_2$$

And according to Theorem 1:

$$\arg\left\{ \min_{u(\tau)\in\Omega} \|Iu(t) - Ic(t)\|_2 \right\} = \text{sat}(c(t)).$$

Thus,

$$\arg\left\{ \min_{u(\tau)\in\Omega} \|u(t) - c(t)\|_p \right\} = \text{sat}(c(t)).$$

Q.E.D.

Remark 1: Theorem 2 states that the p-norm closest feasible points are invariant under a control-affine nonlinear continuous-time dynamical system in the form of (11), if its $\boldsymbol{C}^T\boldsymbol{C}$ can be made diagonal with column or row rearrangements. Requiring $\boldsymbol{C}^T\boldsymbol{C}$ to become diagonal with column or row rearrangements does not require $\boldsymbol{C}^T\boldsymbol{C}$ to be independent of the state variables, x. In other words, the eigenvectors of $\boldsymbol{C}^T\boldsymbol{C}$ need to be parallel to the standard basis vectors at every time instant t, while the eigenvalues of $\boldsymbol{C}^T\boldsymbol{C}$ may depend on x.

Consider the following conintuous-time system examples:

$$\begin{cases} \dot{x}_1 = -2x_1 + x_2 + u_1 - 10u_2 \\ \dot{x}_2 = 6x_1 - 3x_2 + 0.1u_1 + 2u_2 \\ y_1 = x_1 + 5x_2 \\ y_2 = -0.1x_1 + x_2 \end{cases} \tag{18}$$

$$\begin{cases} \dot{x}_1 = -2x_1 + x_2 + u_1 - 10u_2 \\ \dot{x}_2 = 6x_1 - 3x_2 + 0.1u_1 + 2u_2 \\ y_1 = x_1 \\ y_2 = x_2 \end{cases} \tag{19}$$

The system of (18) has the CFP invariace property, but that of (19) does not, as their characteristic (decoupling) matrices, respectively, are:

$$\begin{bmatrix}1.5 & 0\\ 0 & 3\end{bmatrix},\quad \begin{bmatrix}1 & -10\\ 0.1 & 2\end{bmatrix}.$$

C. Nonlinear Discrete-Time Dynamical Square Systems

Let S be a delay-free nonlinear discrete-time dynamical system of the form:

$$S:\ \begin{cases}x(k+1)=\Phi\big(x(k),u(k)\big), & u\in\Omega\subset\mathbb{R}^m\\ y(k)=h\big(x(k)\big)\end{cases} \tag{20}$$

where $y\in\mathbb{R}^q$, $x\in\mathbb{R}^n$, and $m\le q$. The following assumptions are made: the vector fields $\Phi(x,u)$ and $h(x)$ are smooth; every system output y_j has a finite relative order (degree), R_j, which is the smallest integer for which $y_j\big(k+R_j\big)$ explicitly depends on $u(k)$; and the characteristic (decoupling) matrix of the system:

$$\overline{\boldsymbol{C}}\big(x(k),u(k)\big)=\frac{\partial}{\partial u}\begin{bmatrix}h_1^{R_1}\big(x(k),u(k)\big)\\ \vdots\\ h_q^{R_q}\big(x(k),u(k)\big)\end{bmatrix} \tag{21}$$

is full rank, where:

$$h_i^0\big(x(k)\big)\stackrel{\text{def}}{=}h_i\big(x(k)\big)=y_i(k),\quad i=1,\cdots,q \tag{22}$$

$$h_i^1\big(x(k)\big)\stackrel{\text{def}}{=}h_i^0\big(x(k+1)\big)=h_i^0\Big(\Phi\big(x(k),u(k)\big)\Big)=y_i(k+1),\quad i=1,\cdots,q \tag{23}$$

$$h_i^{R_i-1}\big(x(k)\big)\stackrel{\text{def}}{=}h_i^{R_i-2}\big(x(k+1)\big)=h_i^{R_i-2}\Big(\Phi\big(x(k),u(k)\big)\Big)=y_i(k+R_i-1),\ i=1,\cdots,q \tag{24}$$

$$h_i^{R_i}\big(x(k),u(k)\big)\stackrel{\text{def}}{=}h_i^{R_i-1}\big(x(k+1)\big)=h_i^{R_i-1}\Big(\Phi\big(x(k),u(k)\big)\Big)=y_i(k+R_i),\quad i=1,\cdots,q \tag{25}$$

As the system of (20) is delay-free, $R_i=1, i=1,\cdots,q$. For a nonlinear discrete-time dynamical system in the form of (20), we consider the norm:

$$\|c(k)\|_p\stackrel{\text{def}}{=}\sqrt{\textstyle\sum_{\ell=1}^{m}\sum_{j=0}^{1}[c_\ell(k+j)]^2} \tag{26}$$

Theorem 3: For a nonlinear discrete-time dynamical system in the form of (20),

$$\arg\Big\{\min_{u(k),u(k+1)\in\Omega}\|S*u(k)-S*c(k)\|_p\Big\}=\arg\Big\{\min_{u(k),u(k+1)\in\Omega}\|u(k)-c(k)\|_p\Big\}$$

if the characteristic (decoupling) matrix, $\overline{\boldsymbol{C}}$, is independent of u and $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ can be made diagonal with column and row rearrangements.

Proof: Using (22) to (25),

$$\|S*u(k)-S*c(k)\|_p=\|y(k)-\bar{y}(k)\|_p=$$

$$=\Big\{\sum_{\ell=1}^{q}\big[h_\ell^0\big(x(k)\big)-h_\ell^0\big(x(k)\big)\big]^2+\left\{\sum_{\ell=1}^{q}\sum_{j=0}^{1}[y_\ell(k+j)-\bar{y}_\ell(k+j)]^2\right\}^{0.5}\sum_{\ell=1}^{q}\big[h_\ell^1\big(x(k),u(k)\big)-h_\ell^1\big(x(k),c(k)\big)\big]^2\Big\}^{0.5}=$$

$$\sqrt{\sum_{\ell=1}^{q}\big[h_\ell^1\big(x(k),u(k)\big)-h_\ell^1\big(x(k),c(k)\big)\big]^2}=$$

$$\left\|\begin{bmatrix}h_1^1\big(x(k),u(k)\big)\\ \vdots\\ h_q^1\big(x(k),u(k)\big)\end{bmatrix}-\begin{bmatrix}h_1^1\big(x(k),c(k)\big)\\ \vdots\\ h_q^1\big(x(k),c(k)\big)\end{bmatrix}\right\|_2.$$

As the characteristic (decoupling) matrix is independent of u,

$$\left\|\begin{bmatrix}h_1^1\big(x(k),u(k)\big)\\ \vdots\\ h_q^1\big(x(k),u(k)\big)\end{bmatrix}-\begin{bmatrix}h_1^1\big(x(k),c(k)\big)\\ \vdots\\ h_q^1\big(x(k),c(k)\big)\end{bmatrix}\right\|_2=\big\|\overline{\boldsymbol{C}}\big(x(k)\big)u(k)-\overline{\boldsymbol{C}}\big(x(k)\big)c(k)\big\|_2$$

Thus,

$$\min_{u(k),u(k+1)\in\Omega}\|S*u(k)-S*c(k)\|_p=\min_{u(k),u(k+1)\in\Omega}\big\|\overline{\boldsymbol{C}}\big(x(k)\big)[u(k)-c(k)]\big\|_2 \tag{27}$$

According to Theorem 1, as $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ can be made diagonal with row and column rearrangements,

$$\arg\Big\{\min_{u(k)\in\Omega}\big\|\overline{\boldsymbol{C}}\big(x(k)\big)[u(k)-c(k)]\big\|_2\Big\}=\mathrm{sat}(c(k)).$$

As the performance index in (27) is independent of $u(k+1)$, $u(k)=\mathrm{sat}(c(k))$ and every $u(k+1)\in\Omega$ is a solution of the minimization problem of (27). Therefore, $u(k)=\mathrm{sat}(c(k))$ and $u(k+1)=\mathrm{sat}(c(k+1))$ is a solution of the minimization problem. In other words,

$$\min_{u(k),u(k+1)\in\Omega}\big\|\overline{\boldsymbol{C}}\big(x(k)\big)[u(k)-c(k)]\big\|_2=[\mathrm{sat}(c(k))\quad\mathrm{sat}(c(k))]^T$$

As

$$\min_{u(k),u(k+1)\in\Omega}\|u(k)-c(k)\|_p=\min_{u(k),u(k+1)\in\Omega}\sqrt{\sum_{\ell=1}^{m}\sum_{j=0}^{1}[u_\ell(k+j)-c_\ell(k+j)]^2}=\min_{u(k),u(k+1)\in\Omega}\sqrt{\sum_{\ell=1}^{m}[U_\ell(k)-C_\ell(k)]^2}=\min_{u(k),u(k+1)\in\Omega}\|U(k)-C(k)\|_2,$$

where $U(k)=[u(k)\quad u(k+1)]^T$ and $C(k)=[c(k)\quad c(k+1)]^T$.

$$\arg\Big\{\min_{u(k),u(k+1)\in\Omega}\|U(k)-C(k)\|_p\Big\}=\arg\Big\{\min_{u(k),u(k+1)\in\Omega}\|U(k)-C(k)\|_2\Big\}=[\mathrm{sat}(c(k))\quad\mathrm{sat}(c(k))]^T.$$

Q.E.D.

Remark 2: Theorem 3 states that the p-norm closest feasible points are invariant under a nonlinear discrete-time dynamical system in the form of (20), if the characteristic (decoupling) matrix of the system is independent of u and $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ can be made diagonal with column or row rearrangements. Requiring the characteristic (decoupling) matrix of a system to be independent of u and $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ to become diagonal with column or row rearrangements does not require the matrix to be independent of the state variables, x. In other words, the eigenvectors of $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ need to be parallel to the standard basis vectors at every time instant t, while the eigenvalues of $\overline{\boldsymbol{C}}^T\overline{\boldsymbol{C}}$ may depend on x.

Consider the following three dynamical system examples:

$$\begin{cases}x(k+1)=0.995x(k)+0.1u(k)+0.1\cos(3u(k))\\ y(k)=x(k)\end{cases} \tag{28}$$

$$\begin{cases} x_1(k+1) = -3x_1(k) + {x_2}^3(k) - 3u_1(k) + 30u_2(k) \\ x_2(k+1) = {x_1}^2(k) - x_2(k) + 9u_1(k) + u_2(k) \\ y_1(k) = x_1(k) \\ y_2(k) = x_2(k) \end{cases} \quad (29)$$

$$\begin{cases} x_1(k+1) = -3x_1(k) + {x_2}^3(k) - 3u_1(k) + 30u_2(k) \\ x_2(k+1) = {x_1}^2(k) - x_2(k) + 9u_1(k) + u_2(k) \\ y_1(k) = x_1(k) - 30\,x_2(k) \\ y_2(k) = 3x_1(k) + x_2(k) \end{cases} \quad (30)$$

The systems of (28) and (29) do not have this CFP invariace property but that of (30) has, because their characteristic (decoupling) matrices, respectively, are:

$$0.1 - 0.3\sin(3u), \ \begin{bmatrix} -3 & 30 \\ 9 & 1 \end{bmatrix}, \ \begin{bmatrix} -273 & 0 \\ 0 & 91 \end{bmatrix}.$$

Note that the system of (28) is single-input single-output, but it lacks the CFP invariance property.

IV. Closest-feasible-points Non-Invariance Compensator

As pointed out in the Introduction, in the presence of actuator saturation, the performance of a control system may degrade significantly due to: (a) integral windup when the controller is dynamical; and (b) CFP non-invariance (CFPN), that is, the closest feasible points being non-invariant under the plant that is subjected to control.

The definition of the CFP invariance guides how to derive a CFPN compensator (CFPNC) that optimally mitigates the control performance degradation due to the CFPN of a plant. Given a control signal, c, such a compensator calculates the optimal feasible plant input, u^O, that yields a plant output response closest to the plant output response to c. In the case that the controller is dynamical, the states of the controller must be informed of the calculated optimal feasible plant input properly (Fig.5), as it is common in every anti-integral-windup scheme.

Following the approaches used in [10], for continuous-time systems in the form of (11), given a control signal and a measurement of state variables a time instant t, $c(t)$ and $x(t)$, an optimal feasible plant input that yields a plant output response closest to the plant output response to $c(t)$ is proposed to be calculated by solving the following constrained minimization problem at each time instant t:

$$\min_{u(t)\in\Omega} \|S*u(t) - S*c(t)\|_p = \min_{u(t)\in\Omega} \|y(t) - \bar{y}(t)\|_p \quad (31)$$

where

$$\|y(t)\|_p \stackrel{\text{def}}{=} \sqrt{\sum_{\ell=1}^{q} w_\ell \int_t^{t+T} |y_\ell(\tau)|^2 d\tau} \quad (32)$$

T is a sufficiently small time-horizon, and $w_1, \cdots, w_q$ are positive scalar constants, which allow one to adjust the effects of input constraints on controlled variables; the higher is the value of a weight, the higher is the importance of the controlled variable tied to the weight, and the less will be the effects of the input constraints on the controlled variable.

Corollary 1: For continuous-time systems in the form of (11):

$$\min_{u(t)\in\Omega} \|S*u(t) - S*c(t)\|_p = \min_{u(t)\in\Omega} \|y(t) - \bar{y}(t)\|_p =$$

$$= \min_{u(t)\in\Omega} \left\|\tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))u(t) - \tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))c(t)\right\|_2 \quad (33)$$

where $\tilde{\boldsymbol{Q}} = \text{diag}\{\sigma_i\sqrt{w_i}\}$.

Proof: Similarly to the Proof of Theorem 2, when the time horizon, T, is sufficiently small, using (14) and (15), one can write:

Fig. 5. Calculation of an optimal feasible plant input based on an unconstrained controller output (control signal) using a CFPN compensator. The feedback is needed to prevent integral windup, if the controller has dynamics.

$$\|y(t) - \bar{y}(t)\|_p = \left\{\sqrt{\sum_{i=1}^{q} w_i \left[L_g L_f^{r_i-1} h_i(x(t))[u(t) - c(t)]\sigma_i\right]^2}\right\} = \left\|\tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))u(t) - \tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))c(t)\right\|_2$$

Q.E.D.

Thus, given a control signal and a measurement of state variables a time instant t, $c(t)$ and $x(t)$, the constrained optimization of (33) can be solved to obtain the optimal feasible plant input corresponding to the control signal $c(t)$. This constrained optimization problem can be solved easily using the computationally efficient, globally-converging, simple method [15]:

$$u^{\ell+1} = \text{sat}\left(d^{-1}P(u^\ell - c) + u^\ell\right), \quad \ell = 0,1,\cdots \quad u^0 = c$$

where $P = [p_{ij}] = \boldsymbol{C}^T\boldsymbol{C}$ and $d = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} {p_{ij}}^2}$.

In the case that the p-norm closest feasible points are invariant under a dynamical system in the form of (11); that is, the eigenvectors of C^TC are parallel to the standard basis vectors at every time instant t,

$$\arg\left\{\min_{u(t)\in\Omega} \left\|\tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))u(t) - \tilde{\boldsymbol{Q}}\boldsymbol{C}(x(t))c(t)\right\|_2\right\} = \text{sat}(c(t))$$

indicating that in this special case, clipping is optimal; that is, $\text{sat}(c)$ is optimal in the sense of (31).

Example. Consider the plant of (19) with the input constraints: $u_{1,min} = -1$, $u_{1,max} = +1$, $u_{2,min} = -2$, *and* $u_{2,max} = +2$. This plant, which lacks the CFP invariance property, is controlled using the static I-O linearizing state feedback:

$$\begin{cases} c_1 = (2A + 10B)/3 \\ c_2 = (-0.1A + B)/3 \end{cases} \quad (34)$$

where

$$A = \frac{y_{sp,1} - x_1}{\beta_1} + 2x_1 - x_2, \ B = \frac{y_{sp,2} - x_2}{\beta_2} - 6x_1 + 3x_2$$

which induces the closed-loop plant output responses: $\beta_1\dot{y}_1 + y_1 = y_{sp,1}$ and $\beta_2\dot{y}_2 + y_2 = y_{sp,2}$ in the absence of constraints. As the state feedback of (34) has no dynamics, the control quality does not degrade due to integral windup when an actuator saturates.

Other existing methods of calculating a feasible u based on a control signal (controller output), c, are clipping [7, 8]:

$$u = \text{sat}(c), \quad (35)$$

and direction preservation [16]:

$$u_j = u_{j,ss} + (c_j - u_{j,ss})\min\{\rho_1, \cdots, \rho_m\}, \ j = 1, \cdots, m \quad (36)$$

where,

$$\rho_j = \left[\text{sat}_j(c_j) - u_{j,ss}\right]/\left[c_j - u_{j,ss}\right], \quad j = 1, \cdots, m$$

and $u_{j,ss}$ is the steady-state (equilibrium) value of u_j.

Fig.6 compares three cases for which feasible plant inputs are calculated by the CFPN compensator, clipping, and direction preservation, given a controller output c. Direction preservation yields an optimal plant input in the sense of (33)

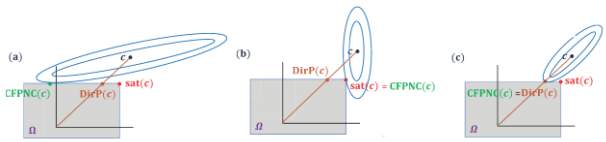

Fig. 6. A comparison of feasible plant inputs calculated by the CFPNC, clipping, and direction preservation, given a controller output c. (a) $\mathrm{sat}(c) \neq \mathrm{DirP}(c) \neq \mathrm{CFPNC}(c)$; (b) $\mathrm{sat}(c) = \mathrm{CFPNC}(c) \neq \mathrm{DirP}(c)$; (c) $\mathrm{DirP}(c) = \mathrm{CFPNC}(c) \neq \mathrm{sat}(c)$.

when the controller output vector and an eigenvector of $\boldsymbol{C}^T\boldsymbol{C}$ are coincidant (Fig.6), while clipping yields an optimal plant input in the sense of (33) when the eigenvectors of $\boldsymbol{C}^T\boldsymbol{C}$ are parallel to the standard basis vectors (Fig.6).

Fig. 7 depicts the input and output responses of the plant under the state feedback of (34) in four cases: (a) when there are no constraints; and (b), (c) and (d) when the input constraints are present and the CFPNC of (32), clipping, (35), and direction preservation, (36), are implemented separately. As can be seen in Fig. 7, for this plant that lacks the CFP invariance property, the control performances are very different in the three cases (b), (c) and (d). As expected, the CFPN compensator provides the constrained plant output response that is closest to the unconstrainted one. Under direction preservation it takes a long time for the controlled variables to reach their setpoint values, and under clipping the initial conditions are not in the domain of attraction of the closed-loop control system; in this case, under the CFPN compensator and direction preservation, the domain of attraction of the closed-loop system is larger.

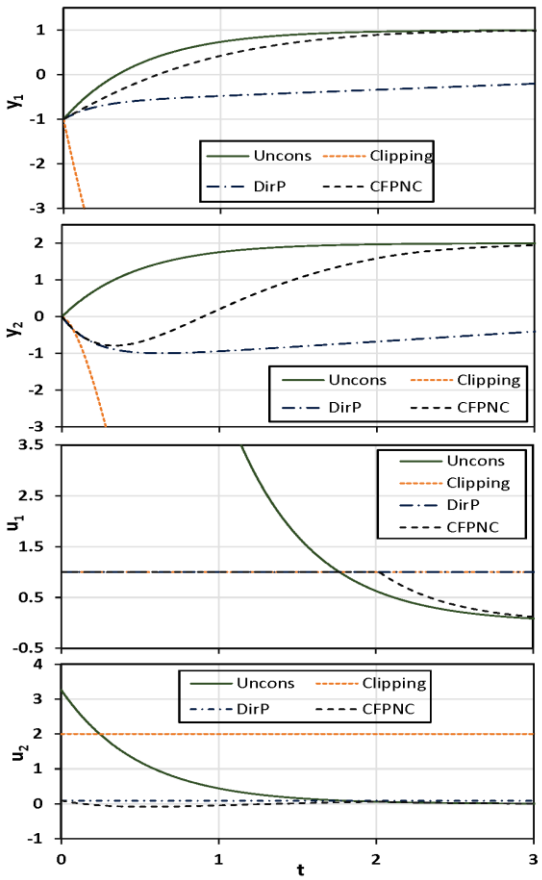

Fig. 7. Input and output profiles of the plant of (19) under the state feedback of (34). Uncons = no input constraints. DirP = direction preservation. $\beta_1 = \beta_2 = 0.5$, $y_{sp,2} = 1$, $y_{sp,2} = 2$, $u_{1,ss} = u_{2,ss} = 0$, $x_1(0) = -1$, and $x_2(0) = 0$.

V. Conclusion

This paper introduced a new system property called the closest feasible points invariance to characterize systems with actuator limits. A few system classes and norm types were considered, and in two of these cases implications and applications of this property in control were explored and discussed. The presence or absence of this invariance property in a system has no relation with the dimension of the system. The definition of this property guided the derivation of a CFP non-invariance compensator that gracefully decreases the control quality degradation in continuous-time plants that lack the CFP invariance property. This work also characterized the plants for which clipping and direction preservation of controller outputs are optimal.

Acknowledgment

The author would like to thank Tamer Basar, Richard Braatz, Prodromos Daoutidis, Nikolaos Kazantzis, Costas Kravaris, and Ali Mesbah for their invaluable feedback on this work.

References

[1] C.T. Chen, *Linear System Theory and Design*. 3rd Ed. Oxford University Press, Inc., pp.144–157, 1999.

[2] F. Blanchini, F. and S. Miani, *Set-theoretic Methods in Control*, Boston: Birkhäuser, pp.99–100, 2008.

[3] D. Chu and D. D. Siljak, "A canonical form for the inclusion principle of dynamic systems," *SIAM Journal on Control and Optimization*, vol. 44, no.3, pp.969–990, 2005.

[4] G. J. Pappas, G. Lafferriere, and S. Sastry, "Hierarchically consistent control systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 6, pp. 1144–1160, Jun 2000.

[5] L. Bakule, J. Rodellar, J. M. Rossell, and P. Rubió. "Preservation of controllability-observability in expanded systems," *IEEE Transactions on Automatic Control*, vol. 46, no. 7, 1155–1162, Jul 2001.

[6] S. S. Stankovic and D. D. Siljak, "Model abstraction and inclusion principle: A comparison," *IEEE Transactions on Automatic Control*, vol. 47, no. 3, pp. 529–532, Aug 2002.

[7] S. Sajjadi-Kia and F. Jabbari, "Multi-stage anti-windup compensation for open-loop stable plants," *IEEE Transactions on Automatic Control,* vol. 56, no. 9, pp. 2166–2172, Sept 2011.

[8] N. Kapoor, A. R. Teel, and. Daoutidis, "An anti-windup design for linear systems with input saturation," *Automatica*, vol. 34, no. 5 pp. 559–574, May 1998.

[9] M. Soroush and N. Mehranbod, "Optimal compensation for directionality in processes with a saturating actuator," *Computers & Chemical Engineering*, vol. 26, no. 11, pp. 1633–1641, Nov 2002.

[10] M. Soroush and S. Valluri, "Optimal directionality compensation in processes with input saturation nonlinearities," *International Journal of Control*, vol. 72, no. 17, pp.1555–1564, Jan 1999.

[11] D. L. Ma, J. G. VanAntwerp, M. Hovd, and R.D. Braatz, "Quantifying the potential benefits of constrained control for a large-scale system." *IEE Proceedings-Control Theory and Applications*, 149(5), 423-432, Sep 2002.

[12] M. Soroush, F. S. Rantow, & Y. Dimitratos, "Control quality loss in analytical control of input-constrained processes," *Industrial & Engineering Chemistry Research*, vol. 45, no. 25, pp. 8528–8538, Dec 2006.

[13] D. Bertsekas. *Nonlinear Programming, 3rd ed*. Athena Scientific, Belmont, Massachusetts, 2016, pp. 307–367.

[14] E. Rimon and S.P. Boyd. "Obstacle collision detection using best ellipsoid fit," *Journal of Intelligent and Robotic Systems*, vol. 18, no. 2, pp.105–126, Feb. 1997.

[15] R. Barnard, "Continuous-time implementation of optimal-aim controls," *IEEE Transactions on Automatic Control*, vol. 21, no.3, pp. 432–434, Jun 1976.

[16] P. Campo and M. Morari, "Robust control of processes subject to saturation nonlinearities," *Computers & Chemical Engineering*, vol. 14, no. 4-5, pp. 343–358, May 1990.