A GENERALIZED PERSIDSKII THEOREM AND ITS APPLICATIONS TO NONSMOOTH GRADIENT DYNAMICAL SYSTEMS

Eugenius Kaszkurewicz *,1 **Amit Bhaya** *

* *Dept. of Electrical Engineering, Federal University of Rio de Janeiro, PEE/COPPE/UFRJ, P.O. Box 68504, Rio de Janeiro, RJ 21945-970, Brazil*

Abstract: A generalized Persidskii-like theorem is derived and shown to be applicable to the stability analysis of a class of gradient dynamical systems with discontinuous right hand sides. These dynamical systems arise from the steepest descent technique applied to a variety of problems suitably formulated as constrained minimization problems. The problems susceptible to this approach include linear programming problems and specifically the k-winners-take-all problem, the problem of solving underdetermined linear systems arising in least squares support vector machines, and quadratic programming problems associated to the support vector machine approach to classification. The advantage of the proposed analysis is the derivation of simple convergence conditions.

Keywords: differential equation with discontinuous right hand side, Filippov solution, Liapunov function, gradient dynamical system, linear programming, support vector machine, k-winners-take-all problem.

1. INTRODUCTION

Gradient dynamical systems to solve linear programming problems were first proposed by Pyne and subsequently analyzed in several papers and, notably, in the book Utkin (1992). An excellent account of this work can be found in Chong *et al.* (1999), which contains the first complete and rigorous analysis of a particular class of gradient dynamical systems with discontinuous right hand sides. In fact, we will use the latter, but will develop an alternative analysis that is applicable to a larger class of problems, starting from a Liapunov function first proposed by Persidskii (1969) and further developed in Hsu *et al.* (2000) as well as Kaszkurewicz and Bhaya (2000). This so called Persidskii type diagonal Liapunov function leads to conditions (for the linear programming problem) that are more tractable than those obtained by Chong *et al.* (1999). The class of gradient systems with discontinuous right hand sides considered in this paper can be regarded as neural networks with discontinuous activation functions (Cichocki and Unbehauen, 1993).

The notation is standard. We denote column vectors by boldface lowercase letters, like $\mathbf{x}, \mathbf{c}$. Scalars are represented by lowercase italic letters, like x or Greek letters, like γ, while sets are represented by capital Greek letters such as Δ. Matrices are denoted by uppercase boldface letters, like $\mathbf{P}$. Vector functions are denoted by $\mathbf{f}(\mathbf{x})$ which, unless otherwise specified, are diagonal type functions, i.e.,

$$\mathbf{f}(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^T,$$

where x_j are the components of vector $\mathbf{x}$, for $j = 1, \ldots, n$. Finally, the step-like set-valued functions $\mathrm{hsgn}, \mathrm{uhsgn}, \mathrm{sgn}$ are defined for use below. Note that

[1] This research was partially financed by Project Nos. 140811/2002-8, 551863/2002-1, 471262/03-0 of CNPq, and also by the agencies CAPES & FAPERJ. E-mail: {eugenius, amit}@nacad.ufrj.br

all are discontinuous at the origin and these functions will usually occur as components of diagonal functions, which are denoted using the same abbreviations.

$$\mathrm{hsgn}(a) := \begin{cases} 0, \text{ if } a > 0 \\ \in [-1, 0], \text{ if } a = 0 \\ -1, \text{ if } a < 0 \end{cases} \quad \mathrm{uhsgn}(a) := \begin{cases} 1, \text{ if } a > 0 \\ \in [0, 1], \text{ if } a = 0 \\ 0, \text{ if } a < 0 \end{cases}$$

$$\mathrm{sgn}(x) = \begin{cases} -1, \text{ if } x < 0 \\ \in [-1, 1], \text{ if } a = 0 \\ 1, \text{ if } x > 0. \end{cases}$$

2. PERSIDSKII TYPE RESULTS FOR NONSMOOTH DYNAMICAL SYSTEMS

Persidskii systems and the corresponding diagonal type functions were first presented in (Persidskii, 1969), where absolute stability was proved for such systems by means of diagonal type Lyapunov functions. Kaszkurewicz and Bhaya (2000) studied such systems in further detail, while Hsu *et al.* (2000) analyzed a class of Persidskii systems with discontinuous right hand side. The latter result is generalized in this section. Consider the following Persidskii-type system:

$$\dot{\mathbf{x}}(t) = -\mathbf{A}\mathbf{f}(\mathbf{x}(t)), \tag{1}$$

where $\mathbf{x} = (\bar{\mathbf{x}}^T, \hat{\mathbf{x}}^T)^T$, $\bar{\mathbf{x}} \in \mathbb{R}^p$, $\hat{\mathbf{x}} \in \mathbb{R}^q$, $\mathbf{f}(\mathbf{x}) = (\bar{\mathbf{x}}^T, \mathbf{g}^T(\hat{\mathbf{x}}))^T$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $n = p + q$ and the vector function $\mathbf{g} : \mathbb{R}^q \to \mathbb{R}^q$ satisfies the following assumptions:

a) $\mathbf{g}(\hat{\mathbf{x}})$ is a piecewise continuous diagonal type function,

$$\mathbf{g}(\hat{\mathbf{x}}) = (g_1(\hat{x}_1), \ldots, g_q(\hat{x}_q))^T, \tag{2}$$

b) $\hat{x}_i g_i(\hat{x}_i) \geq 0$, $i = 1, \ldots, q$;

c) $\mathbf{g}$ is continuous almost everywhere (i.e., the points at which it is discontinuous form a set M of Lebesgue measure zero).

Furthermore, when the g_is are chosen as hsgn, uhsgn, sgn functions, the set M is described by the intersection of surfaces $\bigcap_i \{\mathbf{x} : g_i(\hat{x}_i) = 0\}$, which is referred to as a *surface of discontinuity*. Since system (1) has discontinuous righthand side, its solutions must be considered in the sense of Filippov (1988). According to Filippov's theory, when trajectories are not confined to the surface of discontinuity, the solutions are considered in the usual sense, otherwise the solutions of (1) are the solutions of the following differential inclusion:

$$\dot{\mathbf{x}}(t) \in G(\mathbf{x}(t)), \tag{3}$$

where $\mathbf{x}$ is absolutely continuous, defined almost everywhere within an interval, and the set G is described as the convex hull containing all the limiting values of $g(\mathbf{x})$, when $\mathbf{x} \to \mathbf{x}' \in M$. We define the set G by means of the equivalent control method (Utkin, 1992).

The following is the main result of this paper.

Theorem 1. Consider the Persidskii-type system (1). If there exist a symmetric positive semi-definite matrix $\mathbf{S}$ and a positive definite block diagonal matrix $\mathbf{D}$ such that

$$\mathbf{D} = \left(\begin{array}{c|c} \mathbf{D}_{11} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{D}_{22} \end{array} \right),$$

where the block $\mathbf{D}_{11}$ is symmetric positive-definite and $\mathbf{D}_{22}$ is positive diagonal, such that $\mathbf{A} = \mathbf{SD}$, then the trajectories of (1) converge to the invariant set $\mathcal{A} := \{\mathbf{x} : \mathbf{f}(\mathbf{x}) \in \mathcal{N}(\mathbf{A})\}$.

Proof. (sketch only) Let the following nonsmooth candidate Liapunov function of Lure-Persidskii type, associated to the Persidskii-type system (1) be:

$$V(\mathbf{x}) = \frac{1}{2}\bar{\mathbf{x}}^T \mathbf{D}_{11} \bar{\mathbf{x}} + \sum_{i=p+1}^{p+q} d_{ii}^{(2)} \int_0^{\hat{x}_i} g_i(\tau) d\tau - \left(\frac{1}{2}\bar{\mathbf{x}}^{*T} \mathbf{D}_{11} \bar{\mathbf{x}}^* + \sum_{i=p+1}^{p+q} d_{ii}^{(2)} \int_0^{\hat{x}_i^*} g_i(\tau) d\tau \right), \tag{4}$$

where $\mathbf{x}^{*T}$ belongs to the invariant set $\mathcal{A} := \{\mathbf{x} : \mathbf{f}(\mathbf{x}) \in \mathcal{N}(\mathbf{A})\}$, $d_{ii}^{(2)}$ are elements of a positive diagonal matrix $\mathbf{D}_{22}$ and matrices $\mathbf{D}_{11}$ and $\mathbf{D}_{22}$ are the diagonal elements of a block diagonal matrix $\mathbf{D}$. If $\mathbf{A} = \mathbf{SD}$, the time derivative of (4) along the trajectories of (1),

$$\dot{V} = \nabla V^T \dot{\mathbf{x}} = -\mathbf{f}^T(\mathbf{x})\mathbf{DSDf}(\mathbf{x}). \tag{5}$$

Notice that if $\mathbf{S}$ is positive semi-definite, then $\dot{V}$ is negative semi-definite. Consider the Lure-Persidskii function (4) and its time derivative (5). Notice that $\dot{V} = 0$ iff $\mathbf{Af}(\mathbf{x}) = \mathbf{0}$ and, consequently, $\dot{\mathbf{x}} = \mathbf{0}$ for $\mathbf{f}(\mathbf{x}) \in \mathcal{N}(\mathbf{A})$. Thus, $\mathcal{A} = \{\mathbf{x} : \mathbf{f}(\mathbf{x}) \in \mathcal{N}(\mathbf{A})\}$ is an invariant set. Since, by assumption, $\mathbf{A} = \mathbf{SD}$, (5) can be written as

$$\dot{V} = -\mathbf{f}^T(\mathbf{x})\mathbf{DSDf}(\mathbf{x}). \tag{6}$$

Given the fact that $\mathbf{f}(\mathbf{x})$ is discontinuous in a set M, we analyze $\dot{V}$ under three assumptions:

i) $\mathbf{x}(t) \notin M$, i.e., the trajectories of system (1) are not confined to the surface of discontinuity. In this case, the solutions of (1) exist in the usual sense, thus since $\mathbf{S}$ is positive semi-definite, it is immediate that $\dot{V}(\mathbf{x}) \leq 0$;

ii) $\mathbf{x}(t) \in M$, for $t \in [t_0, t_f]$, i.e., the trajectories of system (1) are confined to the surface of discontinuity during a certain time interval. In this case, the vectors $\mathbf{SDf}(\mathbf{x})$ are described by some vector $\mathbf{e}$ such that $\dot{\mathbf{x}} = -\mathbf{e} \in G(\mathbf{x})$ and we have $\dot{V} = -\mathbf{e}^T\mathbf{e} \leq 0$.

iii) Components of $\hat{\mathbf{x}}_i(t)$, for some i are at points of discontinuity of the corresponding g_i, while the remaining $\hat{\mathbf{x}}_j(t)$, $j \neq i$ are not at points of discontinuity of the corresponding g_j: this case can be treated in a way similar to items i) and ii) above.

Consequently, from items i), ii), iii) and using the nonsmooth version of LaSalle's theorem (Shevitz and Paden, 1994), the trajectories of (1) converge to the invariant set $\mathcal{A}$. Theorem 1 is an extension of the result in Hsu *et al.* (2000) and it provides a general convergence result for Persidskii systems with discontinuous righthand sides. An extension of theorem 1 is also needed in applications to linear programming for the following modification of the dynamical system (1)

$$\dot{\mathbf{x}}(t) = -\mathbf{A}\mathbf{f}(\mathbf{x}(t)) - \mathbf{c}, \qquad (7)$$

where $\mathbf{c}$ is a constant vector, and the other symbols are as defined above. The extension of theorem 1 appears in section 5.0.1.

3. SOLVING LINEAR SYSTEMS

This section considers the problem of finding a solution to an underdetermined system of algebraic linear equations of the form:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad (8)$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ has full row rank, $m \leq n$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$. The least absolute deviation or L_1 approach is to solve the following unconstrained optimization problem:

$$\text{Minimize } E(\mathbf{x}) = \|\mathbf{r}(\mathbf{x})\|_1, \ \mathbf{x} \in \mathbb{R}^n, \qquad (9)$$

where $\mathbf{r}(\mathbf{x}) := \mathbf{A}\mathbf{x} - \mathbf{b}$, and $\|\cdot\|_1$ denotes the 1-norm. A solution of the optimization problem (9) is regarded as a solution of the system of linear equations (8) in the L_1 sense, and is well known to have good robustness properties in regard to outliers and bad data (Portnoy and Koenker, 1997), as opposed to the popular least squares (L_2) solution.

Let $r_i : \mathbb{R}^n \to \mathbb{R}$ be the components of vector $\mathbf{r}$. The set $\Delta := \{\mathbf{x} : \mathbf{r}(\mathbf{x}) = \mathbf{0}\}$ is defined as:

$$\Delta = \bigcap_{i=1}^{m} \Delta_i; \ \Delta_i := \{\mathbf{x} : r_i(\mathbf{x}) = 0\}, \qquad (10)$$

The minimum of the energy function E in (9) is $\mathbf{r} = \mathbf{0}$, consequently, a solution of problem (9) is a vector $\mathbf{x}^* \in \mathbb{R}^n$ such that $\mathbf{x}^* \in \Delta$. Notice that E is convex in $\mathbf{r}$, thus its unique minimizer is the zero vector $\mathbf{r}^* = \mathbf{0}$, and it is nondifferentiable at $\mathbf{r} = \mathbf{0}$. The optimization problem (9) is solved in the present paper by mapping it into a gradient system of the form: $\dot{\mathbf{x}} = -\mathbf{M}\nabla E(\mathbf{x})$, where $\mathbf{M} = \mathrm{diag}\ (\mu)$, is a positive diagonal matrix with $\mu \in \mathbb{R}^n$. In the context of neural networks, the objective function E is referred to as a *computational energy function*, while $\mathbf{M}$ is referred to as a *learning matrix*, and used to improve convergence speed (Cichocki and Unbehauen, 1993). Consider the following gradient system associated to problem (9):

$$\dot{\mathbf{x}} = -\mathbf{M}\mathbf{A}^T\mathrm{sgn}(\mathbf{r}). \qquad (11)$$

Notice that the function $E(\mathbf{x})$ in (9) is nondifferentiable at $\Delta_i = 0$, leading to the discontinuous righthand side of (11).The solutions of (11) are considered in the sense of Filippov (Filippov, 1988), and the set Δ is referred to as a *surface of discontinuity*. If the trajectories of (11) are confined to Δ, this motion is said to be a *sliding motion* or, equivalently, the system is said to be in *sliding mode*. This is equivalent to saying that the motion occurs in the hyperplane tangent to the surface of discontinuity. Further details about sliding modes can be found in (Utkin, 1992; Edwards and Spurgeon, 1998).

3.1 Convergence analysis

Convergence analysis is performed using a Persidskii form of the gradient system (11) in conjunction with the corresponding candidate diagonal type Lyapunov function. The Persidskii form of (11) is obtained by premultiplying (11) by the matrix $\mathbf{A}$. Observe that since $\dot{\mathbf{r}} = \mathbf{A}\dot{\mathbf{x}}$, from (11) we get:

$$\dot{\mathbf{r}} = -\mathbf{A}\ \mathbf{M}\ \mathbf{A}^T\mathrm{sgn}(\mathbf{r}). \qquad (12)$$

The following proposition holds.

Proposition 1. The Persidskii system (12) is equivalent to the original gradient system (11), in the sense that $\dot{\mathbf{r}} \equiv \mathbf{0}$ iff $\dot{\mathbf{x}} \equiv \mathbf{0}$.

Proposition 1, which has a simple proof omitted here, is necessary since it ensures that the convergence results derived for the Persidskii system (12) also hold for the original gradient system (11). Since system (12) has a discontinuous righthand side, we choose following nonsmooth candidate diagonal type Lyapunov function (Hsu *et al.*, 2000):

$$V(\mathbf{r}) = \sum_{i=1}^{m} \int_0^{r_i} \mathrm{sgn}(\tau)\, d\tau. \qquad (13)$$

Observe that i) $V(\mathbf{r}) > 0$ for $\mathbf{r} \neq \mathbf{0}$; ii) $V(\mathbf{r}) = 0$ if and only if $\mathbf{r} = \mathbf{0}$. The time derivative of V along the trajectories of (12) is given by $\dot{V} = \nabla V^T \dot{\mathbf{r}}$, i.e.,

$$\dot{V}(\mathbf{r}) = -\mathrm{sgn}^T(\mathbf{r})\ \mathbf{A}\ \mathbf{M}\ \mathbf{A}^T\mathrm{sgn}(\mathbf{r}). \qquad (14)$$

Notice that since $\mathbf{A}$ has full row rank and $\mathbf{M}$ is positive definite, then $\mathbf{A}\ \mathbf{M}\ \mathbf{A}^T$ is also positive definite. Consequently, $\dot{V} \equiv 0$ if and only if $\mathrm{sgn}(\mathbf{r}) \equiv \mathbf{0}$ implying $\dot{\mathbf{r}} \equiv \mathbf{0}$ and, from Proposition 1, $\dot{\mathbf{x}} \equiv \mathbf{0}$ and we have:

Theorem 2. The trajectories of system (11) converge, from any initial conditions, to the solution set of the system of linear equations (8) in finite time and remain in this set thereafter. Moreover, the convergence time t_f satisfy the bound $t_f \leq V(\mathbf{r}_0)/\lambda_{min}(\mathbf{A}\mathbf{M}\mathbf{A}^T)$, where $\mathbf{r}_0 := \mathbf{r}(\mathbf{x}_0)$.

4. SUPPORT VECTOR MACHINES (SVM)

Given two classes A and B, the classical pattern recognition problem of finding the best surface that separates the elements of two given classes can be described as follows. Consider the following training pairs:

$$(y_1, \mathbf{z}_1), \ldots, (y_m, \mathbf{z}_m), \quad y_i \in \{-1, 1\}, \tag{15}$$

where the vectors $\mathbf{z}_i$ belong to the input space and the scalars y_i define the position of the vectors $\mathbf{z}_i$ in relation to the surface that separates the classes, i.e., if $y_i = +1$ the vector $\mathbf{z}_i$ is located above the separating surface and if $y_i = -1$, this vector is located below the separating surface. If given a set of pairs as in (15), a single hyperplane can be chosen such that $\forall\, i,\; y_i = \pm 1$, then the set of points $\{\mathbf{z}_i\}_{i=1}^m$ is said to be *linearly separable*.

Consider two classes A and B, not necessarily linearly separable, identified as $y_A = +1$ and $y_B = -1$, respectively. The problem of finding the best hyperplane $\Pi := \{\mathbf{u} : \mathbf{u}^T\mathbf{z} + c = 0\}$ that separates the elements of classes A and B is modeled by the following quadratic optimization problem (Cristianini and Shawe-Taylor, 2000):

$$\min \xi(\mathbf{u}, \mathbf{e}, c) = \frac{1}{2}\mathbf{u}^T\mathbf{u} + b\sum_{k=1}^{N} e_k^p \tag{16}$$
$$\text{s.t. } y_i(\mathbf{u}^T\mathbf{z}_i + c) \geq 1 - e_i;\; e_i \geq 0, \quad i = 1, \ldots, m.$$

where p is a positive integer, $\mathbf{u}$, $\mathbf{z}_i \in \mathbb{R}^n$ and $e_i, \in \mathbb{R}$. The quantity $y_i(\mathbf{u}^T\mathbf{z} + c)$ is defined as the margin of the input $\mathbf{z}$ with respect to the hyperplane Π. The hyperplane Π that solves problem (16) gives the soft margin hyperplane, in the sense that the number of training errors is minimal (Schölkopf and Smola, 2002; Cristianini and Shawe-Taylor, 2000).The slack variables e_i are introduced in order to provide tolerance to misclassifications.

For nonlinear classification, a feature function ϕ, that maps the input space into a higher dimensional space is introduced. In this case, the constraints of problem (16) become $y_i(\mathbf{u}^T\phi(\mathbf{z}_i) + c) \geq 1 - e_i, \quad i = 1, \ldots, m$. The traditional approach is to solve the dual of (16), since in this case, instead of the function ϕ, another class of functions, known as kernel functions and defined as $K(\mathbf{z}, \mathbf{z}_i) = \phi^T(\mathbf{z})\phi(\mathbf{z}_i)$ is used, with the advantage that it is not necessary to know the feature function ϕ. The feature function ϕ is defined implicitly by the kernel which is assumed to satisfy the Mercer conditions (Schölkopf and Smola, 2002; Cristianini and Shawe-Taylor, 2000).

4.1 LS-SVM

The LS-SVM model is a modification of the original SVM model (16), in which the inequality constraints are replaced by equality constraints. The LS-SVM is modeled by the following constrained optimization problem (Suykens *et al.*, 2002):

$$\min \xi_{ls}(\mathbf{u}, \mathbf{e}, c) = \frac{1}{2}\mathbf{u}^T\mathbf{u} + \frac{b}{2}\sum_{i=1}^{n} e_i \tag{17}$$
$$\text{s.t. } y_i(\mathbf{u}^T\phi(\mathbf{z}_i) + c) = 1 - e_i,\; i = 1, \ldots, m$$

The dual problem of (17) is given by the following system of linear equations, also known as a Karush-Kuhn-Tucker (KKT) linear system (Suykens *et al.*, 2002):

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \mathbf{Q} + b^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} c \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right], \tag{18}$$

where $\mathbf{Q}$ is a symmetric matrix given by $q_{ij} = y_i y_j K(\mathbf{z}_i, \mathbf{z}_j)$ and K is defined by the kernel $K(\mathbf{z}, \mathbf{z}_j) = \phi^T(\mathbf{z})\,\phi(\mathbf{z}_j)$. In the LS-SVM model, the problem of determining the best separating surface for classes A and B is reduced to solving the system of linear equations (18), which has a full rank coefficient matrix if $b^{-1} \neq -\lambda_i(\mathbf{Q})$, $\forall i$. Thus by Theorem 2, the trajectories of the gradient system (11), with $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{b}$ defined as in (18), converge in finite time to the solution of (17).

Numerical examples and further details can be found in Ferreira *et al.* (2005).

4.2 ν-SVC for nonlinear separation

There are several formulations for the nonlinear separation problem (Cristianini and Shawe-Taylor, 2000). The ν-SVM formulation (Schölkopf *et al.*, 2000) fits into the general formulation of Theorem 1 and is modeled by the following constrained optimization problem:

$$\min \tau(\mathbf{w}, \xi, \eta) = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\eta + \frac{1}{m}\sum_{i=1}^{m} e_i \tag{19}$$
$$\text{s.t. } y_i(\mathbf{w}^T\phi(\mathbf{z}_i) + c) \geq \eta - \xi_i, \text{ and } \eta \geq 0,\, e_i \geq 0,$$

where ϕ is a feature map function, which provides the classifier with the ability of performing nonlinear discrimination of patterns. The additional parameter ν controls the number of margin errors and support vectors (Schölkopf and Smola, 2002). The dual of the constrained optimization problem (19) is as follows:

$$\min \xi_\nu(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \tag{20}$$
$$\text{s.t. } \mathbf{y}^T\boldsymbol{\alpha} = 0, \tag{21}$$
$$\boldsymbol{\alpha} - m^{-1}\mathbf{1}_m \leq \mathbf{0}_m \text{ and} \tag{22}$$
$$\mathbf{B}\boldsymbol{\alpha} - \nu\mathbf{h} \geq \mathbf{0}_{m+1}. \tag{23}$$

where the column vectors $\mathbf{0}_m, \mathbf{1}_m \in \mathbb{R}^m$, $\mathbf{0}_{m+1} \in \mathbb{R}^{m+1}$, $\boldsymbol{\alpha} \in \mathbb{R}^m$, $q_{ij} = y_i y_j \phi(\mathbf{z}_i)^T\phi(\mathbf{z}_j)$ and the matrix $\mathbf{B}^T := (\mathbf{1}_m \;\; \mathbf{I}_m)$ and the column vector $\mathbf{h}^T := (1 \;\; \mathbf{0}_m)$. Notice that the objective function ξ_ν is homogeneously quadratic in $\boldsymbol{\alpha}$ and the constraints of the quadratic programming problem are linear. Let $\mathbf{r} := \mathbf{B}\boldsymbol{\alpha} - \nu\mathbf{h}$, $x = \mathbf{y}^T\boldsymbol{\alpha}$ and $\mathbf{v} := \boldsymbol{\alpha} - m^{-1}\mathbf{1}$.

Applying the exact penalty function method to (20) we obtain:

$$\min E(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} + \rho|x| - \gamma\sum_{i=1}^{n}\min(0, r_i) + \beta\sum_{i=1}^{n}\max(0, v_i). \tag{24}$$

The gradient system $\dot{\boldsymbol{\alpha}} = -\nabla E(\boldsymbol{\alpha})$ associated to the unconstrained optimization problem (24) is given by:

$$\dot{\boldsymbol{\alpha}} = -\mathbf{Q}\boldsymbol{\alpha} - \rho\,\mathbf{y}\,\mathrm{sgn}(x) - \gamma\mathbf{B}^T\mathrm{hsgn}(\mathbf{r}) - \beta\mathrm{uhsgn}(\mathbf{v}), \tag{25}$$

4.2.1. Convergence analysis: Define the function $\mathbf{f}$ and the vector $\boldsymbol{\theta}$ as:

$$\mathbf{f}(\boldsymbol{\theta}) := (\boldsymbol{\alpha}^T, \mathrm{sgn}(x), \mathrm{hsgn}^T(\mathbf{r}), \mathrm{uhsgn}^T(\mathbf{v}))^T \tag{26}$$

$$\boldsymbol{\theta} := (\boldsymbol{\alpha}, x, \mathbf{r}, \mathbf{v}). \tag{27}$$

An augmented dynamical system in the vector $\boldsymbol{\theta}$ can be written as the Persidskii system:

$$\dot{\boldsymbol{\theta}} = -\mathbf{A}\mathbf{f}(\boldsymbol{\theta}), \tag{28}$$

where the matrix $\mathbf{A}$ is:

$$\mathbf{A} = \begin{pmatrix} \mathbf{Q} & \rho\mathbf{y} & \gamma\mathbf{B}^T & \beta\mathbf{I} \\ \mathbf{y}^T\mathbf{Q} & \rho\mathbf{y}^T\mathbf{y} & \gamma\mathbf{y}^T\mathbf{B}^T & \beta\mathbf{y}^T \\ \mathbf{B}\mathbf{Q} & \rho\mathbf{B}\mathbf{y} & \gamma\mathbf{B}\mathbf{B}^T & \beta\mathbf{B} \\ \mathbf{Q} & \rho\mathbf{y} & \gamma\mathbf{B}^T & \beta\mathbf{I} \end{pmatrix}.$$

Matrix $\mathbf{A}$ can be factored into the form $\mathbf{A} = \mathbf{S}\mathbf{D}$, where:

$$\mathbf{S} := \left(\begin{array}{c|ccc} \mathbf{I} & \mathbf{y} & \mathbf{B}^T & \mathbf{I} \\ \hline \mathbf{y}^T & \mathbf{y}^T\mathbf{y} & \mathbf{y}^T\mathbf{B}^T & \mathbf{y}^T \\ \mathbf{B} & \mathbf{B}\mathbf{y} & \mathbf{B}\mathbf{B}^T & \mathbf{B} \\ \mathbf{I} & \mathbf{y} & \mathbf{B}^T & \mathbf{I} \end{array}\right), \mathbf{D} := \left(\begin{array}{c|ccc} \mathbf{Q} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \rho & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \gamma\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \beta\mathbf{I} \end{array}\right). \tag{29}$$

Theorem 3. If $\mathbf{Q}$ is positive definite then, for any positive constants ρ, γ and β, the trajectories of the system (25) converge to the solution of the dual quadratic programming problem (20).

The proof of this theorem follows directly from Theorem 1 observing that the positive definiteness of matrix $\mathbf{Q}$, assumed in theorem 3, is achieved by choosing positive-definite kernels in the implementations. Observe also that the convergence of the gradient system does not depend on the parameter ν.

5. LINEAR PROGRAMMING SOLUTION OF THE KWTA PROBLEM

Urahama and Nagao (1995) formulated the KWTA problem as the following integer programming problem:

$$\begin{aligned} &\max_{\mathbf{x}} \mathbf{c}^T\mathbf{x} \\ &\text{s.t. } \mathbf{1}^T\mathbf{x} = k,\ \mathbf{x} \in \{0,1\}^n, \end{aligned} \tag{30}$$

converted it into a nonlinear programming problem, and solved it by minimizing an associated Lagrangian function. In fact, the integer programming problem above can be relaxed to the following LP problem with bounded variables:

$$\begin{aligned} &\max_{\mathbf{x}} \mathbf{c}^T\mathbf{x} \\ &\text{s.t. } \mathbf{1}^T\mathbf{x} = k,\ \mathbf{x} \in [0,1]^n \end{aligned} \tag{31}$$

where $\mathbf{c} = [c_1, \ldots, c_n]^T$, $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^{n\times 1}$, $k \le n \in \mathbb{N}$ is a nonnegative integer and $\mathbf{x} \in \mathbb{R}^{n\times 1}$. The following proposition states that the integer programming problem (30) and its relaxed version (31) have the same solution $\mathbf{x}^*$.

Proposition 2. Consider the LP problem (31), and let the components of vector $\mathbf{c}$ be distinct. Then, the solution of the LP problem (31) is unique and presents k components equal to one, which, correspondingly, multiply the k largest components of vector $\mathbf{c}$ in the objective function z, while the $n-k$ remaining components are equal to zero.

Applying the penalty function method, we have:

$$\min E(\mathbf{x}, \gamma, \rho) = -\mathbf{c}^T\mathbf{x} - \gamma\left(\sum_{j=1}^{n}\min(0, x_j) - \sum_{i=1}^{n} x_j^+\right) + \rho\,|\mathbf{1}^T\mathbf{x} - k| \tag{32}$$

where, for each j

$$x_j^+ = \begin{cases} x_j - 1 & \text{if } x_j > 1 \\ 0 & \text{if } x_j \le 1. \end{cases}$$

Consider the gradient system $\dot{\mathbf{x}} = -\nabla E(\mathbf{x})$, that minimizes E, which is given by:

$$\dot{\mathbf{x}} = \mathbf{c} - \gamma[\mathrm{hsgn}(\mathbf{x}) + \mathrm{uhsgn}(\mathbf{x})] - \rho\mathbf{1}\mathrm{sgn}(\mathbf{1}^T\mathbf{x} - k) \tag{33}$$

5.0.1. Convergence results Let $\mathcal{X}$ be the solution set of the LP problem (31). Convergence to the set $\mathcal{X}$ is defined, after (Utkin, 1992, pg. 229), as follows.

$$\lim_{t\to\infty} \min_{\mathcal{X}} \|\mathbf{x}(t) - \mathbf{x}^*\| = 0,\ \mathbf{x}^* \in \mathcal{X},$$

Convergence of the system described by equation (33) occurs in two steps, commonly known as the *reaching phase* and the *sliding mode*, in the literature on variable structure systems (Utkin, 1992). Our analysis is, accordingly, also in two steps. First, we derive sufficient conditions for the reaching phase, i.e., to ensure convergence to the feasible set of problem (31), which is given by the intersection

$$\Omega := \Pi \cap \Gamma, \tag{34}$$

where $\Pi := \{\mathbf{x} : \mathbf{1}^T\mathbf{x} - k = 0\}$ and $\Gamma := \{\mathbf{x} : x_j \in [0,1],\ \text{for each } j\}$. In common with methods that use discontinuous switching functions (hsgn, uhsgn, sgn), the dynamical system (33) has the pleasant property of a finite time reaching phase (31), which means that there exists $\bar{t} < \infty$ such that $\min_{\mathcal{X}} \|\mathbf{x}(t) - \mathbf{x}^*\| \to 0$, $\mathbf{x}^* \in \mathcal{X}$, as $t \to \bar{t}$ (Chong *et al.*, 1999).

Premultiplying (33) by the row vector $\mathbf{1}^T$ and noticing that $\dot{r} = \mathbf{1}^T\dot{\mathbf{x}}$ we get,

$$\dot{r} = \mathbf{1}^T\mathbf{c} - \gamma\mathbf{1}^T\mathbf{h}(\mathbf{x}) - \rho\mathbf{1}^T\mathbf{1}\mathrm{sgn}(r). \tag{35}$$

Writing equations (33) and (35) in vector notation, we get

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{r} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{1}^T\mathbf{c} \end{bmatrix} - \begin{bmatrix} \mathbf{I}_n & \mathbf{1} \\ \mathbf{1}^T & \mathbf{1}^T\mathbf{1} \end{bmatrix} \begin{bmatrix} \gamma\,\mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & \rho \end{bmatrix} \begin{bmatrix} \mathbf{h}(\mathbf{x}) \\ \mathrm{sgn}(r) \end{bmatrix}. \tag{36}$$

Defining

$$\mathbf{D} = \begin{bmatrix} \gamma\,\mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & \rho \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{I}_n & \mathbf{1} \\ \mathbf{1}^T & \mathbf{1}^T\mathbf{1} \end{bmatrix}, \tag{37}$$

using the standard choice of the following Persidskii diagonal type “sum of integral-of-nonlinearities” Lyapunov function:

$$V(\mathbf{x}, r) = \gamma \sum_{j=1}^{n} \int_0^{x_j} h(\tau)d\tau + \rho \int_0^{r} \mathrm{sgn}(\tau)d\tau, \tag{38}$$

and recalling that in the reaching phase, $\mathbf{x} \notin \Omega$, i.e., $\mathbf{h}(\mathbf{x}) \neq \mathbf{0}$ or $\mathbf{1}^T\mathbf{x} - k \neq 0$, it is possible to derive the following lemma.

Lemma 1. Consider the system of ordinary differential equations (33). Provided that γ and ρ satisfy the following inequality

$$\min(\gamma^2, n\rho^2, n(\gamma - \rho)^2) \geq \|\mathbf{c}\|_1(\gamma + \rho) \tag{39}$$

then, for any initial condition, the trajectories reach the set Ω, defined in (34), in finite time and remain in this set thereafter.

Lemma 1, the proof of which is omitted but available on request, results in tractable convergence conditions because the inequality (39) leads to bounds that are easy to calculate, and depend only on the 1-norm of vector $\mathbf{c}$, as the reader can easily check.

It remains to show that the LP problem (31) and the unconstrained problem (32) have the same solution $\mathbf{x}^*$ and that, in sliding mode, the trajectories of the dynamical system (33) converge to the solution of the LP problem. These straightforward proofs are omitted for lack of space (see Ferreira *et al.* (2003)). We now have all the elements needed to ensure that the network modeled by the gradient system (33) is a “*k*-winners-take-all” network, stated in the theorem below.

Theorem 4. Consider the network described by the system (33),and assume that the network gains γ and ρ satisfy the inequality (39). Then, given a vector $\mathbf{c} \in \mathbb{R}^n$, with distinct components, and a positive integer $k \leq n$, the network described by (33) is a KWTA network, i.e., trajectories from all initial conditions converge to the solution. □

6. CONCLUDING REMARKS

Several examples from different areas show that the generalized Persidskii theorem and its corollaries proposed in this paper lead to simply computable conditions for convergence of dynamical systems with discontinuous right hand sides. This is, in the case of the KWTA and SVM problems, a considerable simplification of general LP results proposed earlier in the literature (Chong *et al.*, 1999). Another area of potential applications is in the congestion control of communication networks: a gradient search algorithm for a non-differentiable objective function is mentioned in Wu *et al.* (2001, p.1276), and a sliding mode approach is used by Lagoa *et al.* (2004) in this problem.

REFERENCES

Chong, E. K. P., S. Hui and S. H. Żak (1999). An analysis of a class of neural networks for solving linear programming problems. *IEEE Trans. Automatic Control* **44**(11), 1995–2006.

Cichocki, A. and R. Unbehauen (1993). *Neural Networks for Optimization and Signal Processing*. John Wiley and Sons.

Cristianini, N. and J. Shawe-Taylor (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Edwards, C. and S. K. Spurgeon (1998). *Sliding mode control: Theory and Applications*. Taylor & Francis.

Ferreira, L. V., E. Kaszkurewicz and A. Bhaya (2003). Synthesis of a k-winners-take-all neural network using linear programming with bounded variables. In: *Procedings of the International Joint Conference on Neural Networks 2003, Portland, OR, USA*. Vol. 3. pp. 2360–2365.

Ferreira, L. V., E. Kaszkurewicz and A. Bhaya (2005). Solving systems of linear equations via gradient systems with discontinuous righthand sides: application to LS-SVM. *IEEE Trans. Neural Networks* **16**(2), to appear.

Filippov, A. F. (1988). *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, Dordrecht.

Hsu, L., E. Kaszkurewicz and A. Bhaya (2000). Matrix-theoretic conditions for the realizability of sliding manifolds. *Systems & Control Letters* **40**, 145–152.

Kaszkurewicz, E. and A. Bhaya (2000). *Matrix Diagonal Stability in Systems and Computation*. Birkhäuser, Boston.

Lagoa, C. M., H. Che and B. A. Movischoff (2004). Adaptive control algorithms for decentralized optimal traffic engineering in the Internet. *IEEE-ACM Transactions on Networking* **12**(3), 415–428.

Persidskii, S. K. (1969). Problem of absolute stability. *Automation and Remote Control* **12**, 1889–1895.

Portnoy, S. and R. Koenker (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science* **12**(4), 279–300.

Schölkopf, B., A. J. Smola, R. C. Williamson and P. L. Barlett (2000). New support vector algorithms. *Neural Computation* **12**(5), 1207–1245.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. The MIT Press.

Shevitz, D. and B. Paden (1994). Lyapunov stability of nonsmooth systems. *IEEE Transactions on Automatic Control*.

Suykens, J. A., T. V. Gestel, J. D. Brabanter, B. D. Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific. Singapore.

Urahama, K. and T. Nagao (1995). K-winners-take-all circuit with o(n) complexity. *IEEE Transactions on Neural Networks* **6**(3), 776–778.

Utkin, V. I. (1992). *Sliding Modes in Control and Optimization*. Springer-Verlag, Berlin.

Wu, G., E. K. P. Chong and R. Givan (2001). Congestion control via online sampling. In: *Proc. IEEE Infocom*. Anchorage, AK. pp. 1271–1280.