Automatica 44 (2008) 2745–2753

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Piecewise linear solution paths with application to direct weight optimization ☆

Jacob Roll *

Division of Automatic Control, Linköping University, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 22 January 2007
Received in revised form
12 March 2008
Accepted 25 March 2008
Available online 5 October 2008

Keywords:
Parametric programming
Piecewise quadratic programming
Function approximation
Nonlinear system identification
Nonparametric identification

ABSTRACT

Recently, pathfollowing algorithms for parametric optimization problems with piecewise linear solution paths have been developed within the field of regularized regression. This paper presents a generalization of these algorithms to a wider class of problems. It is shown that the approach can be applied to the nonparametric system identification method, Direct Weight Optimization (DWO), and be used to enhance the computational efficiency of this method. The most important design parameter in the DWO method is a parameter (λ) controlling the bias-variance trade-off, and the use of parametric optimization with piecewise linear solution paths means that the DWO estimates can be efficiently computed for all values of λ simultaneously. This allows for designing computationally attractive adaptive bandwidth selection algorithms. One such algorithm for DWO is proposed and demonstrated in two examples.

1. Introduction

In many applications, one encounters optimization problems involving a trade-off between two terms to optimize, i.e., problems of the type

$$\min_x L(x) + \lambda J(x) \tag{1}$$

where λ is a design parameter controlling the trade-off. The problem (1) is a parametric optimization problem (Guddat, Guerra Vasquez, & Jongen, 1990), or can also be viewed as a special case of multiobjective optimization (Boyd & Vandenberghe, 2004).

Examples include, e.g., many bias-variance trade-off types of problems, and can also be found in the field of regularized regression. In the paper by Efron, Hastie, Johnstone, and Tibshirani (2004), the authors present a new estimation method, least angle regression (LARS), and show that the solutions to both LARS and LASSO (Tibshirani, 1996) can be efficiently computed for all values of λ simultaneously. As pointed out in Rosset and Zhu (2004, 2007), the key to these algorithms is that the *solution paths* (i.e., the optimal solutions x to the parametric optimization problem as a function of λ) are piecewise linear as λ varies from 0 to ∞. Similar results have recently also been shown for the related nn-garrote method and grouped versions of all these methods (Yuan & Lin, 2006). In all these cases, having a single-parametric optimization problem allows for developing pathfollowing algorithms that exploit the piecewise linearity to efficiently find and represent the solution path.

This paper presents a generalization of the framework of pathfollowing algorithms for piecewise linear solution paths in Efron et al. (2004), Rosset and Zhu (2007) and Yuan and Lin (2006), and extends the problem class to a broad class of (single-) parametric piecewise quadratic programs and related problems. It is shown that the solution paths are piecewise linear, and a pathfollowing algorithm is given. For the case of quadratic plus piecewise affine cost functions, an algorithm with explicit expressions for computation of the solution path is given.

Related work can also be found in the area of model predictive control, where in recent years results in explicit model predictive control has led to a growing interest in multiparametric linear and quadratic programming. It has been shown that the solutions to different classes of problems are piecewise affine functions of the parameters (see, e.g., Bemporad, Morari, Dua, and Pistikopoulos (2002), Borrelli (2003), Pistikopoulos, Georgiadis, and Dua (2007) and Tøndel, Johansen, and Bemporad (2003)). However, it seems that piecewise quadratic problems has only very recently begun to receive attention (Mayne, Raković, & Kerrigan, 2007).

A particular example of parametric problems in the form (1) occurs in Direct Weight Optimization (Roll, 2003; Roll, Nazin, & Ljung, 2005a,b), which is a nonparametric identification/function estimation method. DWO computes pointwise function estimates, given data $\{y(t), \varphi(t)\}_{t=1}^N$ from

$$y(t) = f_0(\varphi(t)) + e(t) \tag{2}$$

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.

* Corresponding address: Autoliv Electronics AB, Teknikringen 9, SE-583 30 Linköping, Sweden. Tel.: +46 13 480 4452; fax: +46 13 367201.

E-mail address: jacob.roll@autoliv.com.

0005-1098/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2008.03.020

where f_0 is the unknown function to be estimated, $f_0 : \mathbb{R}^n \to \mathbb{R}$, and $e(t)$ are white noise terms. The idea of making pointwise function estimates has also appeared under names such as *Model on Demand*, *lazy learning* and *least commitment learning* (see, e.g., Atkeson, Moore, and Schaal (1997a,b), Bontempi, Birattari, and Bersini (1999), Stenman (1999) and references therein).

In order to estimate $f_0(\varphi^*)$ for a given point φ^*, the idea of DWO is to use a linear estimator $\hat{f}_N(\varphi^*)$

$$\hat{f}_N(\varphi^*) = \sum_{t=1}^{N} w_t y(t) \tag{3}$$

and to select the weights $w = (w_1, \ldots, w_N)$ of the estimator by convex optimization. Assuming that f_0 belongs to some function class $\mathcal{F}$, the weights can be determined by minimizing a convex upper bound on the maximum mean-squared error (MSE)[1]

$$mMSE(\varphi^*, w) = \sup_{f \in \mathcal{F}} E\left[\left(f(\varphi^*) - \hat{f}_N(\varphi^*)\right)^2 \mid \{\varphi(t)\}_{t=1}^{N}\right]. \tag{4}$$

The resulting minimization problem is convex and can be written in the following abstract form:

$$\min_{w \in \mathcal{D}} \lambda U_2(w) + V(w) \tag{5}$$

where U_2 is basically an upper bound on the squared bias, and V the variance term. The design parameter λ determines the trade-off between the flexibility of the function class and the noise variance (see Section 3 for more details).

The computed estimate will of course depend on the choice of λ controlling the bias-variance trade-off. A method for selecting λ for the case when the noise variance is known was given in Juditsky, Nazin, Roll, and Ljung (2004). It could also be chosen by using cross-validation or some other criterion (Härdle, 1990; Stenman, 1999). For all these methods, one needs to compute the DWO estimates for several different parameter values, which makes it desirable to be able to efficiently compute the entire solution path.

We will show that the developed pathfollowing algorithm can be applied to the DWO approach. This means that we can simultaneously compute the DWO values from (5) for all choices of λ, which would mean a great gain in computational efficiency. A cross-validation-based algorithm for selection of λ in DWO is also proposed.

The paper is organized as follows: Section 2 considers some specific problem classes for which the solution paths are piecewise linear, while Section 3 proposes how this property can be exploited in the DWO approach.

2. Piecewise linear solution paths

In this section, we will consider some specific classes of optimization problems of the type (1), which will be shown to have piecewise linear solution paths.

2.1. Piecewise quadratic plus piecewise affine cost function

First, we will consider a class of optimization problems in the form (1) where $J(x)$ is piecewise affine and $L(x)$ is a piecewise quadratic function. A general piecewise affine convex function can be written (Boyd & Vandenberghe, 2004)

$$J(x) = \max_k \{c_k^{\mathrm{T}} x + d_k\} \tag{6}$$

$L(x)$ is supposed to be strictly convex and in the form

$$L(x) = \frac{1}{2} x^{\mathrm{T}} Q_i x + f_i^{\mathrm{T}} x + r_i \quad \text{if } x \in \mathcal{X}_i \tag{7}$$

where $Q_i = Q_i^{\mathrm{T}}$ are positive definite, and the polyhedral regions $\mathcal{X}_i = \{x \mid \tilde{H}_i x \preccurlyeq \tilde{q}_i\}$, $i \in \mathcal{I}$ (here $\preccurlyeq$ denotes componentwise inequalities), form a partition of the x space (for simplicity, we let the regions be closed sets, which means that they will intersect at the boundaries). Furthermore, we assume that for each $\lambda \geq 0$, problem (1) has a unique, finite optimal solution.

We can now show the following lemma.

Lemma 1. *The problem*

$$\begin{aligned} \min_x \quad & \lambda \max_k \{c_k^{\mathrm{T}} x + d_k\} + L(x) \\ \text{subj. to } & Ax = b \\ & \bar{A}x \preccurlyeq \bar{b} \end{aligned} \tag{8}$$

with $L(x)$ given by (7) has a piecewise linear solution path, i.e., the optimal $x \in \mathbb{R}^n$ is a piecewise affine function of $\lambda \in [0, \infty]$.

Proof. It is easy to see that the optimum of (8), which is unique and finite for given λ according to the assumptions, changes continuously with λ.

Now, we can partition the feasible set into a number of relatively open polyhedra together with a number of points (the corners of the polyhedra), denoted P_j (i.e., either $P_j = \mathit{relint}(P_j)$ or P_j is a single point; for the definition of relative interior, see Boyd and Vandenberghe (2004)), such that on P_j, the cost function of (1) equals

$$\lambda(c_{k_j}^{\mathrm{T}} x + d_{k_j}) + \frac{1}{2} x^{\mathrm{T}} Q_{i_j} x + f_{i_j}^{\mathrm{T}} x + r_{i_j}.$$

Let the affine hull of P_j (Boyd & Vandenberghe, 2004) be described by

$$\mathit{aff}(P_j) = \{x \mid \tilde{A}_j x = \tilde{b}_j\}$$

where $\tilde{A}_j$ is chosen such that it has full row rank.

Assume that the solution to (8) for a given λ lies in P_j. Then, since this solution is either in the relative interior of P_j or the only point of P_j, it is also the solution to

$$\begin{aligned} \min_x \quad & \lambda(c_{k_j}^{\mathrm{T}} x + d_{k_j}) + \tfrac{1}{2} x^{\mathrm{T}} Q_{i_j} x + f_{i_j}^{\mathrm{T}} x + r_{i_j} \\ \text{subj. to } & \tilde{A}_j x = \tilde{b}_j \end{aligned} \tag{9}$$

But the solution to this problem can be computed as

$$\begin{aligned} x = Q_{i_j}^{-1} \Big(& \left(\tilde{A}_j^{\mathrm{T}} (\tilde{A}_j Q_{i_j}^{-1} \tilde{A}_j^{\mathrm{T}})^{-1} \tilde{A}_j Q_{i_j}^{-1} - I\right)(f_{i_j} + c_{k_j}\lambda) \\ & + \tilde{A}_j^{\mathrm{T}} (\tilde{A}_j Q_{i_j}^{-1} \tilde{A}_j^{\mathrm{T}})^{-1} \tilde{b}_j \Big) \end{aligned} \tag{10}$$

(see Roll (2007)). Here, x is linear in λ. This means that the solution to (8) must consist of a number of such linear pieces, one piece for every P_j that the solution path passes through. Hence, the solution path is piecewise linear. □

Remark 2. The strict convexity condition for $L(x)$ can be relaxed. It is sufficient that $L(x)$ is strictly convex in a neighborhood of each point on the solution path, and convex elsewhere.

[1] Note that (4) is always convex in w, regardless of how the function class $\mathcal{F}$ is chosen, so in principle we could minimize the maximum MSE directly. However, for many function classes, (4) is difficult to compute, and we have to find an upper bound instead.