Trends in Computational and Applied Mathematics, **25** (2024), e01767
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029 www.scielo.br/tcam
ORIGINAL ARTICLE
doi: 10.5540/tcam.2024.025.e01767

The NFDA-Nonsmooth Feasible Directions Algorithm applied to constructing Pareto Fronts of Ridge and Lasso Regressions

W. P. FREIRE

Received on July 4, de 2023 / Accepted on September 27, 2024

ABSTRACT. Ridge and Lasso regressions are types of linear regression, a machine learning tool for dealing with data. Based on multiobjective optimization theory, we transform Ridge and Lasso regression into bi-objective optimization problems. The Pareto fronts of the resulting problems provide a range of regression models from which the best one can be selected. We employ the NFDA-Nonsmooth Feasible Directions Algorithm devised for solving convex optimization problems to construct the Pareto fronts of Ridge and Lasso when regarded as bi-objective problems.

Keywords: Ridge regression, Lasso regression, multiobjective optimization, Pareto front.

1 INTRODUCTION

Today, massive amounts of data are collected and stored every minute. In general, these data are used for predicting future results based on previous information. Doctors, for example, would be delighted to be able to tell their patients their likelihood to develop certain diseases based on the patients' features or habits. If doctors had at their disposal reliable tools to make such predictions, patients could start treatment or even prevention before too late.

Cientists develop mathematical models that use data collected from previous experiences to produce accurate answers to important questions. One of the models used for prediction is Linear Regression [16], [22] where n outcome variables $b_1,...,b_n$ and n associated predictors or features $a_i = (a_{i1},...,a_{ip})$, $i = 1,...,n$, are observed. The goal is to build a model capable of predicting new outcomes from new features with certain reliability and also sort out the most relevant components of the features. A linear regression model is written as

$$b_i = x_0 + \sum_{j=1}^{p} a_{ij}x_j + e_i\,,\, i = 1,2,...,n$$

Department of Mathematics, Federal University of Juiz de Fora, Rua José Lourenço Kelmer s/n, 36036-900, Campus Universitário, Bairro São Pedro, Juiz de Fora, MG, Brazil – Email: wilhelm.freire@ufjf.br https://orcid.org/0000-0001-8348-0799

where $x_0, x_1, ..., x_p$ are the coefficients of the model and e_i is the error. Using matrices, the model can be rewritten as $b = Ax + e$ where $b = (b_1, ..., b_n)$, $A = (a_{ij})$ is a $n \times p$ matrix, $x = (x_1, ..., x_p)$ and $e = (e_1, ..., e_n)$.

The well known Least Squares method [13], [4] provides estimates for the coefficients by minimizing the sum of the squares of the errors or, in other words, by minimizing $\|e\|_2^2 = \|Ax - b\|_2^2$. If $A^T A$ is invertible, the solution of the least squares problem, called here ls, can be expressed through the formula $ls = (A^T A)^{-1} A^T b$. The Least Squares method produces low bias and, if n is much larger than p, low variance too [22]. However, if $p > n$, the solution of the least squares is no longer unique. Moreover, the interpretability of the model is hampered when p is large.

Shrinkage methods [16], [8], [17], [1], [11], [33] such as Ridge and Lasso reduce such drawback by restricting the size of the coefficients. Ridge imposes $\|x\|_2^2 \leq t$ as a constraint added to the least squares problem whereas the Lasso demands $\|x\|_1 \leq t$, where $t > 0$. The upside of using 1-norm is that it increases the chances of finding null coefficients because of the shape of the feasible region [16], [17]. However, the model becomes nondifferentiable. Regardless the case, a minimization problem has to be solved and the parameter t must be set previously. It is not easy to tell in advance the value of t that provides the best model. For each t fixed, the coefficients are obtained after the minimization of the sum of the squares of the errors subject to some restriction on the size of the coefficients. In practice, many values of t are set leading to as many model as values of t. The practitioner, having the models at hand, has to verify the accuracy of the models and choose the best one by performing some procedure like, for example, cross-validation [2], [25], [7], [35]. In general, the analysis of the models is a massive task and takes long time due to the large number of values of t necessary to get a good model [14]. One alternative is transforming the regression problem into a bi-objective optimization problem, constructing its Pareto front and choosing the best model from those solutions which produce the front. Each point on the front corresponds to a solution which, in its turn, corresponds to a model.

The connection between Linear Regression and Optimization is clear and, going further, Machine Learning and Optimization are strongly intertwined [3]. For example, Suttorp and Igel [36] apply multiobjective evolutionary optimization to support vector machines. Jin and Sendhoff [23] provide an overview of the research envolving Multiobjective Optimization and Machine Learning. In his Ph.D thesis [32], the author studies how some machine learning problems can be addressed by means of multiobjective optimization techniques and propose algorithms to solve such problems. In [6], the authors propose a bi-objective mixed integer linear programming to select the best model for a linear regression problem. The authors also suggest a heuristic for choosing the best point.

In this work, based on multiobjective optimization theory, we show that Ridge and Lasso regressions can be regarded as bi-objective optimization problems. However, the main contribution of the paper is the algorithm called NFDA-Nonsmooth Feasible Directions Algorithm, devised for solving convex (nonsmooth) optimization problems. We employ it to generate the Pareto fronts of the resulting optimization problems and show that NFDA can be a good tool for such task.

The paper is organized in 6 sections of which this introduction is the first. In Section 2, we state Ridge and Lasso regression problems. Section 3 is about Multiobjective Optimization. There, we present the main concepts and explain how we develop our approach. In Section 4, we present the NFDA – Nonsmooth Feasible Directions Algorithm. In Section 5, we present numerical experiments and, finally, the conclusions are presented in Section 6.

2 RIDGE AND LASSO REGRESSIONS

In this section, we state Ridge regression, the Lasso and their Lagrangian forms. The reader can find comprehensive discussion and applications in [16], [22], [17], [1], [35], [15], [24] and references therein.

Given a $n \times p$ matrix A and a n-dimensional vector b, the objective of Ridge and Lasso regression is to find a p-dimensional vector x that minimizes the squared 2-norm $\|Ax-b\|_2^2$ and satisfies $\|x\|_q^q \leq t$, where $\|x\|_q = \left(\sum_{i=1}^p |x_i|^q\right)^{\frac{1}{q}}$ and $t \geq 0$ is given. We have Ridge regression if $q = 2$ and Lasso regression if $q = 1$. The associated optimization problem is then posed as

$$\text{(P2.1)} \quad \text{minimize} \quad \|Ax-b\|_2^2 \quad \text{subject to} \quad \|x\|_q^q \leq t.$$

Problem (P2.1) is a convex constrained problem that is also smooth in Ridge case and nonsmooth in Lasso case. For $0 < t < \|ls\|_q^q$, the solution of problem (P2.1) lies on the border of the ball $\{x \in \mathbb{R}^n : ||x||_q^q = t\}$. If $t \geq \|ls\|_q^q$ the solution is exactly ls. On the other hand, if $t = 0$, the solution is obviously the null vector. As t is varied between 0 and $\|ls\|_q^q$, the solutions of (P2.1) show a trade-off between the bias and the variance of the models [35], [6]. Associated with problem (P2.1) is its Lagrangian form

$$\text{(P2.2)} \quad \text{minimize} \quad \|Ax-b\|_2^2 + \lambda \|x\|_q^q, \lambda \geq 0.$$

Lagrangian Duality [30], [20] assures a one-to-one correspondence between the original problem (P2.1) and its Lagrangian form (P2.2), i.e. for each t there is a λ that leads to the same solution and vice-versa. Therefore, it is often better solving (P2.2) instead of (P2.1). However, λ, as well as t, must be set before minimizing the Lagrangian, which is not a simple task. This previous choice can be avoided by using Multiobjective Optimization.

3 MULTIOBJECTIVE OPTIMIZATION

In this section, we explain how the scalar constrained optimization problem (P2.1) is transformed into a bi-objective optimization problem. We present the concepts, theoretical results and methods that allow us to do so. For further discussion, the reader can find excellent material in [10], [28], [21].

Given a vector function $f : X \subset \mathbb{R}^n \longrightarrow \mathbb{R}^k$, $f(x) = (f_1(x), f_2(x), ..., f_k(x))$, a multiobjective optimization problem is defined as

$$\text{(MOP)} \quad \text{minimize} \quad \{f_1(x), f_2(x), ..., f_k(x)\}, x \in X$$

where $f_1, f_2, ..., f_k : \mathbb{R}^n \longrightarrow \mathbb{R}$ are the objective functions and $X \subset \mathbb{R}^n$ is the feasible region. In general, $X = \{x \in \mathbb{R}^n | g_i(x) \leq 0\,, h_j(x) = 0\,, i = 1, ..., m\,, j = 1, ..., p\}$ with $g_i\,, h_j : \mathbb{R}^n \longrightarrow \mathbb{R}$ continuous.
Solving (MOP) means minimizing all the objective functions simultaneously. The spirit of multiobjective optimization is based on the assumption that there is conflict among the functions f_i. There being no conflict, the solution can be found by minimizing each f_i separately, in which case no method for multiobjective optimization is required. For the reader's convenience, we recall some standart definitions [28] largely used in Multiobjective Optimization.

Definition 3.1. *The feasible objective region, denoted by $Z = f(X) \subset \mathbb{R}^k$, is the image of the feasible region X. Its elements $z = (z_1, z_2, ..., z_k) \in \mathbb{R}^k$ are called objective vectors and each $z_i = f_i(x)$ is called objective value.*

Definition 3.2. *A decision vector $x^* \in X$ is Pareto optimal if there does not exist another decision vector $x \in X$ such that $f_i(x) \leq f_i(x^*)$, $i = 1, 2, ..., k$ and $f_j(x) < f_j(x^*)$ for at least one index j. An objective vector $z^* \in Z$ is Pareto optimal if the corresponding decision vector is Pareto optimal.*

Definition 3.3. *A decision vector $x^* \in X$ is weak Pareto optimal if there does not exist another decision vector $x \in X$ such that $f_i(x) < f_i(x^*)$ for all $i = 1, ..., k$. An objective vector $z^* \in Z$ is weak Pareto optimal if the corresponding decision vector is weak Pareto optimal.*

A Pareto point is a vector that is either Pareto optimal or weak Pareto optimal. The subset of the feasible objective region consisting of Pareto points is the Pareto front. The goal of Multiobjective Optimization is to find Pareto fronts. The practitioner can then sort out the best solution among those on the front. Our goal is to build effectively Pareto fronts of Ridge and Lasso by using the NFDA. The analysis of the front is not object of the paper. As mentioned before, the points on the front are obtained from the solutions of the multiobjective optimization problem equivalent to problem (P2.1).

Scalarization [9], [5], [31] is a technique widely employed for solving multiobjective optimization problems. It consists of transforming the vector function associated to the multiobjective problem (MOP) into a scalar function to be minimized.

In this paper, we focus on two scalarization methods, namely the Weighting Method and the ε-Constraint Method. We have based our approach on the relation between those methods and the equivalence between the problem (P2.1) and its Lagrangian form (P2.2). Since Ridge and Lasso can be regarded as bi-objective optimization problems, in what follows, for simplicity, we fix $k = 2$. For generalizations, the reader can see [10], [28], [21].

3.1 The Weighting Method

This method consists of associating a positive weighting coefficient to each objective function and minimizing the weighted sum of objectives. The original multiobjective problem (MOP) is then transformed into the following (scalar) problem

$$\text{(WM)} \quad \text{minimize} \quad \{w_1 f_1(x) + w_2 f_2(x)\}, x \in X, w_1, w_2 \geq 0, w_1 + w_2 = 1.$$

The solutions of (WM) are Pareto points. If $w_1, w_2 > 0$, they are Pareto optimals.

3.2 The ε-Constraint Method

In this method one of the objective functions is selected to be minimized and the remaining functions are set as constraints by imposing upper bounds to each of them. The original multiobjective optimization problem thus becomes

$$\text{(CM)} \quad \text{minimize} \quad f_1(x) \quad \text{subject to} \quad f_2(x) \leq \varepsilon, x \in X.$$

The solutions of (CM) are Pareto points.

3.3 Connections between the Weighting Method and the ε-Constraint Method

In this subsection, we lay the bases upon which our approach develops.

Theorem 3.1. *([28] , Theorem 3.2.5)*
Let $x^* \in X$ be a solution of (WM) and $w_1, w_2 \geq 0$.
(1) If $w_1 > 0$ then x^* is a solution of (CM) for f_1 as objective function and $\varepsilon = f_2(x^*)$ or
(2) If x^* is the unique solution of (WM) then x^* is a solution of (CM) with $\varepsilon = f_2(x^*)$ and f_1 as objective function.

Theorem 3.2. *([28] , Theorem 3.2.6) Let the multiobjective optimization problem be convex. If $x^* \in X$ is a solution of (CM) for f_1 as objective function and $\varepsilon = f_2(x^*)$ then there exists $w_1, w_2 \geq 0$ with $w_1 + w_2 = 1$ such that x^* is also a solution of (WM).*

Theorems 3.1 and 3.2 assert the equivalence between the Weighting Method and the ε-Constraint Method, for convex problems. We have developed our approach based on this equivalence. Let us look at Ridge and Lasso from multiobjective optimization perspective. Setting $f_1(x) = \|Ax - b\|_2^2$ and $f_2(x) = \|x\|_q^q$, we get the following bi-objective convex optimization problem

$$\text{(P3.3.1)} \quad \text{minimize} \quad \{\|Ax - b\|_2^2, \|x\|_q^q\}, x \in \mathbb{R}^n.$$

If we apply the ε-Constraint Method, the problem becomes

$$\text{(P3.3.2)} \quad \text{minimize} \quad \|Ax - b\|_2^2 \quad \text{subject to} \quad \|x\|_q^q < \varepsilon$$

which is exactly the original problem (P2.1) if we replace t with ε.

According to theorems 3.1 and 3.2, for solving (P3.3.2), we can use the Weighting Method and solve the problem

$$\text{(P3.3.3)} \quad \text{minimize} \quad w_1\|Ax-b\|_2^2 + w_2\|x\|_q^q, x \in \mathbb{R}^n$$

where w_1 and w_2 are previously fixed and must satisfy $w_1, w_2 \geq 0$, $w_1 + w_2 = 1$.
One should notice that the scalar function $w_1\|Ax-b\|_2^2 + w_2\|x\|_q^q$ is equivalent to the Lagrangian function $\|Ax-b\|_2^2 + \lambda\|x\|_q^q$ if we take $\lambda = \dfrac{w_2}{w_1}$.

(P3.3.3) is a convex, maybe nonsmooth, and unconstrained problem which can be solved by some methods available in the literature [20], [26], [27].

4 THE NFDA-NONSMOOTH FEASIBLE DIRECTIONS ALGORITHM FOR CONVEX OPTIMIZATION

In this section, we explain the NFDA-Nonsmooth Feasible Directions Algorithm, specially devised for convex unconstrained optimization problems. NFDA was firstly presented by Freire in [12] and later studied in [19].

The main aspect of NFDA is its search direction. It uses the direction employed by the FDIPA-Feasible Directions Interior Point Algorithm [18] for nonlinear but differentiable problems. Another important feature of NFDA is its clear stopping criterion.

Let us start by considering the problem

$$\text{(P4.1)} \quad \text{minimize} \quad F(x), x \in \mathbb{R}^n$$

where $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a convex, not necessarily differentiable function.

One can easily see that if we set $F(x) = w_1\|Ax-b\|_2^2 + w_2\|x\|_q^q$ we get problem (P.3.3.3).

Problem (P4.1) is equivalent to the following constrained problem

$$\text{(P4.2)} \quad \text{minimize} \quad f(x,z) = z \quad \text{subject to} \quad F(x) \leq z, (x,z) \in \mathbb{R}^n \times \mathbb{R}.$$

We recall the definition of the epigraph of a function $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$

$$\text{epi}(F) = \{(x,z) \in D \times \mathbb{R} \mid F(x) \leq z\}$$

its interior

$$(\text{epi}(F))^0 = \{(x,z) \in D \times \mathbb{R} \mid F(x) < z\}$$

and the subdifferential of F at a point $a \in D$

$$\partial F(a) = \{s \in \mathbb{R}^n | F(x) \geq F(a) + \langle s, x-a\rangle\}.$$

NFDA starts at a point $(x^1, z^1) \in (\text{epi}(F))^0$. At iteration k, having the point $(x^k, z^k) \in (\text{epi}(F))^0$, NFDA computes a supporting hyperplane h_k to the epigraph of F at $(x^k, F(x^k))$. This hyperplane

is given by $h_k(x) = F(x^k) + \langle s^k, (x - x^k)\rangle$, where a subgradient $s^k \in \partial F(x^k)$ is assumed to be available. An auxiliary linear constrained problem

$$\text{(P4.3)} \quad \text{minimize} \quad f(x,z) = z \quad \text{subject to} \quad g^k(x,z) \leq 0\,,\ (x,z) \in \mathbb{R}^n \times \mathbb{R}$$

is defined by employing the supporting hyperplanes computed so far.

The function

$$g^k = (g_1, ..., g_k) : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^k$$

is a vector function with

$$g_i : \mathbb{R}^{n+1} \longrightarrow \mathbb{R} \quad \text{given by} \quad g_i(x,z) = h_i(x) - z.$$

Let us for a moment suppose that $p^* = (x^*, z^*)$ is a regular point of (P4.3). The Karush-Kuhn-Tucker (KKT) first order necessary conditions are expressed as follows: If p^* is a local minimum of (P4.3) then there exists $\lambda^* \in \mathbb{R}^k$ such that

$$\nabla f(p^*) + \nabla g^k(p^*)\lambda^* = 0 \tag{4.1}$$

$$G^k(p^*)\lambda^* = 0 \tag{4.2}$$

$$\lambda^* \geq 0 \tag{4.3}$$

$$g^k(p^*) \leq 0 \tag{4.4}$$

where $G(p)$ is a diagonal matrix with $G_{ii}(p) \equiv g_i(p)$.

A Newton-like iteration to solve the nonlinear system of equations (4.1)-(4.2) leads to

$$\begin{pmatrix} B^k & \nabla g^k(p^k) \\ \Lambda^k \nabla g^k(p^k)^t & G^k(p^k) \end{pmatrix} \begin{pmatrix} p - p^k \\ \lambda - \lambda^k \end{pmatrix} = - \begin{pmatrix} \nabla f(p^k) + \nabla g^k(p^k)\lambda^k \\ G^k(p^k)\lambda^k \end{pmatrix} \tag{4.5}$$

where (p^k, λ^k) is the current point at the iteration k, Λ is a diagonal matrix with $\Lambda_{ii} \equiv \lambda_i$ and $B^k \equiv \nabla^2 f(p^k) + \sum_{i=1}^{k} \lambda_i^k \nabla^2 g_i(p^k)$ is the hessian of the Lagrangian function $L(p,\lambda) = f(p) + \lambda^T g(p)$ or some quasi-Newton approximation which must be symmetric and positive definite in order to ensure convergence.

Setting $d = p - p^k$, the following system can be written from (4.5)

$$B^k d + \nabla g^k(p^k)\lambda = -\nabla f(p^k) \tag{4.6}$$

$$\Lambda^k \nabla g^k(p^k)^T d + G^k(p^k)\lambda = 0. \tag{4.7}$$

The solution (d_1^k, λ_1^k) of the system (4.6)-(4.7) provides a descent direction d_1^k for f, as proved in [12] and [19]. However, d_1^k might not be a feasible direction. Indeed, if the point $p^k = (x^k, z^k) \in (\text{epi}(F))^0$ is too close to the graph of the function F, i.e. if $z^k = F(x^k)$ then $g_k(x^k, z^k) = 0$ and, therefore, from (4.7), we have $\nabla g_k(x^k, z^k)^T d_1^k = 0$.

Such trouble can be avoided by perturbing equation (4.7) by adding the matrix $-\rho_k\Lambda^k$, with $\rho_k > 0$, to its right side. A new system

$$B^k d + \nabla g^k(p^k)\bar{\lambda} = -\nabla f(p^k) \tag{4.8}$$

$$\Lambda^k \nabla g^k(p^k)^T d + G^k(p^k)\bar{\lambda} = -\rho_k \Lambda^k \tag{4.9}$$

with unknows d and $\bar{\lambda}$ is obtained. Now, equation (4.9) is equivalent to $\lambda_i^k \nabla g_i(p^k)^T d^k + g_i(p^k)\bar{\lambda}^k = -\rho_k \lambda_i^k$, $i = 1,2,...,k$. Consequently, if $g_k(p^k) = 0$ then $\nabla g_k(p^k)d^k = -\rho_k < 0$ which means that d^k is a feasible direction. The addition of the negative term $-\rho_k\Lambda^k$ produces a deflexion of d^k into the interior of the epigraph of F. The trouble now is that d^k might no longer be a descent direction for the function f. However, this property can be assured if ρ_k is properly adjusted. The system (4.8)-(4.9) can be decoupled in two systems with the same matrix

$$B^k d_1 + \nabla g^k(p^k)\lambda_1 = -\nabla f(p^k) \tag{4.10}$$

$$\Lambda^k \nabla g^k(p^k)^T d_1 + G^k(p^k)\lambda_1 = 0 \tag{4.11}$$

and

$$B^k d_2 + \nabla g^k(p^k)\lambda_2 = 0 \tag{4.12}$$

$$\Lambda^k \nabla g^k(p^k)^T d_2 + G^k(p^k)\lambda_2 = -\Lambda^k. \tag{4.13}$$

After solving systems (4.10)-(4.11) and (4.12)-(4.13), d^k is then set as

$$d^k = d_1^k + \rho_k d_2^k.$$

We wish ρ_k such that $(d^k)^T \nabla f(p^k) < 0$. The formula for d^k can be used to find an appropriate ρ_k. Indeed, we have that

$$(d^k)^T \nabla f(p^k) = (d_1^k)^T \nabla f(p^k) + \rho_k (d_2^k)^T \nabla f(p^k).$$

Thus, as $(d_1^k)^T \nabla f(p^k) < 0$, if $(d_2^k)^T \nabla f(p^k) \leq 0$ then $(d^k)^T \nabla f(p^k) < 0$, $\forall \rho_k > 0$.
If $(d_2^k)^T \nabla f(p^k) > 0$, by imposing

$$(d^k)^T \nabla f(p^k) \leq \xi (d_1^k)^T \nabla f(p^k) \text{ with } \xi \in (0,1)$$

we get

$$(d_1^k)^T \nabla f(p^k) + \rho_k (d_2^k)^T \nabla f(p^k) \leq \xi (d_1^k)^T \nabla f(p^k).$$

The latest inequality leads to

$$\rho_k \leq \frac{(\xi - 1)(d_1^k)^T \nabla f(p^k)}{(d_2^k)^T \nabla f(p^k)}.$$

Therefore, if ρ_k is chosen as just described, $d^k = d_1^k + \rho_k d_2^k$ is assured to be a feasible descent direction.

Figure 1 summarizes the previous discussion.

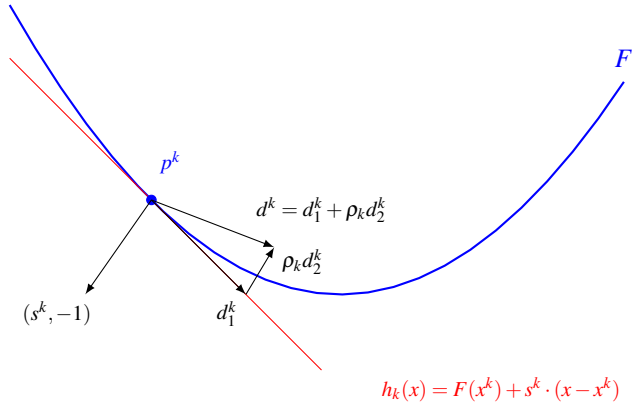


Figure 1: The NFDA search direction

A stepsize t_k is defined by

$$t_k = \min\{t_{\max}, T\}$$

where

$$t_{\max} = \max\{t : g_i^k(x^k, z^k) + t d^k \leq 0\}$$

and $T > 0$ is a predefined parameter.

An auxiliary point

$$(y^k, \omega^k) = (x^k, z^k) + \mu t_k d^k$$

is computed. Here, $\mu \in (0,1)$.

If $F(y^k) < \omega^k$ or, equivalently, if $(y^k, \omega^k) \in (\text{epi}(F))^0$ then $(x^{k+1}, z^{k+1}) = (y^k, \omega^k)$, the hyperplane

$$h_{k+1} = F(x^{k+1}) + \langle s^{k+1}, x - x^{k+1} \rangle \ , \ \ s^{k+1} \in \partial F(x^{k+1}),$$

is computed and the constraint

$$g_{k+1}(x,z) = h_{k+1}(x) - z \leq 0$$

is added to (P4.3). This procedure is called serious step.

If $F(y^k) \geq \omega^k$, then $(x^{k+1}, z^{k+1}) = (x^k, z^k)$. This is called null step. In this case, the supporting hyperplane h_{k+1} is defined by

$$h_{k+1}(x) = F(y^k) + \langle s, x - y^k \rangle \ \text{ with } \ s \in \partial F(y^k)$$

and, as before,

$$g_{k+1}(x,z) = h_{k+1} - z \leq 0$$

is added to update the auxiliary problem (P4.3).

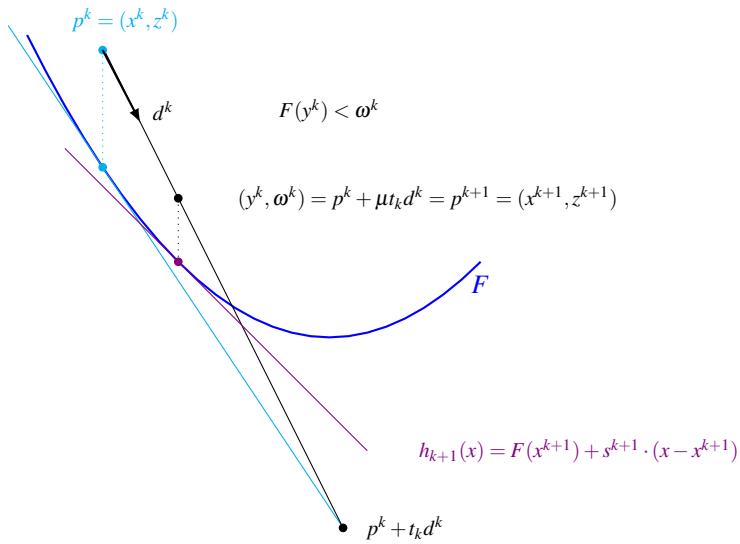


Figure 2: Serious step.

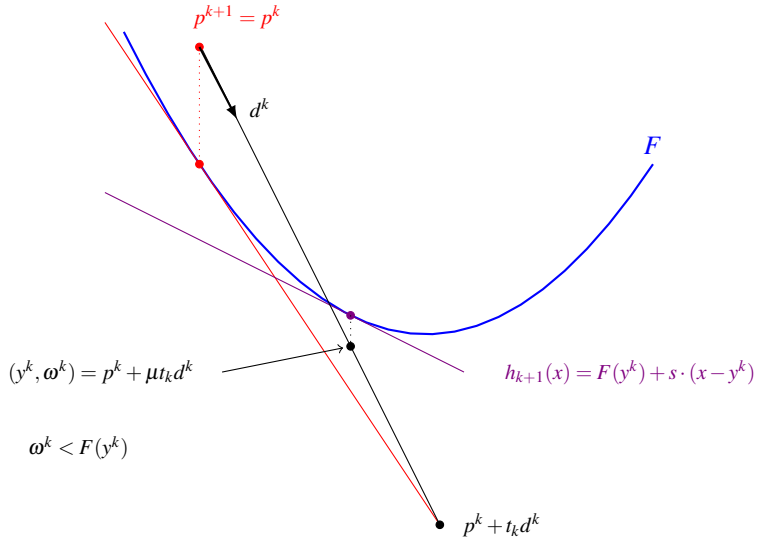


Figure 3: Null step.

It has been proved [12], [19] that the direction d^k goes to zero as k grows. This fact provides a stopping criterion for the algorithm. Moreover, the accumulation points of the limited sequence $\{(x^k, z^k)\} \in (\text{epi}(F))^0$ generated by NFDA are solutions of (P4.1). This is also proved in references [12], [19].

It is important to highlight that the problem (P4.3) is not solved. It may not even have a solution. NFDA uses its linear and, therefore differentiable, structure to get a feasible descent direction.

NFDA is very easy to be coded. It practically only requires the solution of two linear systems with the same matrix. It is robust as it does not require adjusting parameters and has shown good performance in several applications, specially in the present context. We also highlight that, in the description of NFDA given in this section, all the supporting hyperplanes are stored. There are versions of NFDA in which only part of them are kept. Readers interested in detailed discussion on NFDA, its assumptions, updating rules, convergence and other features are referred to [12], [19]. In these references, one finds a comparison between NFDA and some algorithms that use classical methods such as steepest descent and bundle methods, applied to a set of problems from the literature [27].

Algorithm 1 NFDA

Parameters

$\xi, \mu \in (0,1)$, $\varphi > 0$, $T > 0$, $\gamma > 0$.

Initial Data

$(x^1, z^1) \in (\text{epi}(F))^0$, $\lambda^1 > 0$, $B^1 \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ symmetric and positive definite.

Step k

Given $p^k = (x^k, z^k) \in (\text{epi}(F))^0$, h_k, $\lambda^k > 0$, compute $g_k(x,z) = h_k(x) - z$, $g^k = [g_1, g_2, ..., g_k]$ and $\nabla g^k = [\nabla g_1, \nabla g_2, ..., \nabla g_k]$.

Find $d_1^k, d_2^k, \lambda_1^k$ and λ_2^k by solving the systems

$$\begin{cases} B^k d_1 + \nabla g^k(p^k)\lambda_1 = -\nabla f(p^k) \\ \Lambda^k \nabla g^k(p^k)^T d_1 + G^k(p^k)\lambda_1 = 0 \end{cases}$$

and

$$\begin{cases} B^k d_2 + \nabla g^k(p^k)\lambda_2 = 0 \\ \Lambda^k \nabla g^k(p^k)^T d_2 + G^k(p^k)\lambda_2 = -\Lambda^k \end{cases}$$

where B^k is symmetric and positive definite, $G^k(p^k)$ and Λ^k are diagonal matrices with $G^k_{ii}(p^k) = g^k_i(p^k)$ and $\Lambda^k_{ii} = \max\{\lambda_i^{k-1}, \varphi\|d_1^k\|^2\}$.

If $d_2^k \nabla f(p^k) > 0$ set $\rho^k = \varphi\|d_1^k\|^2$. Else, $\rho^k = \min\left\{\varphi||d_1^k||^2, (1-\xi)\dfrac{(d_1^k)^T\nabla f(p^k)}{(d_2^k)^T\nabla f(p^k)}\right\}$.

Set $d^k = d_1^k + \rho^k d_2^k$ and $\lambda^k = \lambda_1^k + \rho^k \lambda_2^k$.

Compute $t_{\max} = \min\{t | g_i^k(x,x) + t d^k \leq 0\}$ and the stepsize $t^k = \max\{t_{\max}, T\}$.

Compute the auxiliary point $(y^k, w^k) = (x^k, z^k) + \mu t^k d^k$.

If $F(y^k) < w^k$ then $(x^{k+1}, z^{k+1}) = (y^k, w^k)$. Compute $s^{k+1} \in \partial F(x^{k+1})$ and $h_{k+1} = F(x^{k+1}) + \langle s^{k+1}, x - x^{k+1}\rangle$.

Otherwise, $(x^{k+1}, z^{k+1}) = (x^k, z^k)$, compute $s \in \partial F(y^k)$ and $h_{k+1} = F(y^k) + \langle s, x - y^k\rangle$.

Repeat Step k until $\|d^k\| \leq \gamma$.

We finalize this section by pointing out that, in order to generate the Pareto fronts of Ridge and Lasso via NFDA, one must set the function F in (P4.1) equal to $w_1\|Ax-b\|_2^2+w_2\|x\|_q^q$ and vary w_1 and w_2 as many times as the number of Pareto points desired. This procedure is equivalent to solving (P3.3.3) as many times as the number of Pareto points. For each given pair (w_1,w_2), the solution x^* of (P3.3.3) generates a point $(\|Ax^*-b\|_2^2,\|x^*\|_q^q)$ on the front. Two important points on the front are the so called end points which are $(\|b\|_q^q,0)$ and $(\|Als-b\|_2^2,\|ls\|_q^q)$, obtained by setting $t=0$ and $t=\|ls\|_q^q$ or equivalently, $w_1=0$ and $w_2=1$ or $w_1=1$ and $w_2=0$, respectively.

5 NUMERICAL EXPERIMENTS

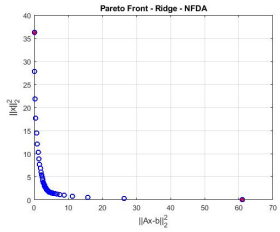
5.1 Generating and Comparing Pareto Fronts

The success of our approach relies on the resolution of problem (P3.3.3). To check the effectiveness of NFDA, we built the Pareto front of three examples using NFDA and MatLab's minimization function fminsearch. The matrices A and vectors b of examples 5.1 and 5.3 were made up whereas, in example 5.2, A and b were randomly created by Matlab. The algorithms have been implemented in MatLab (R2017b) in a microcomputer i7 of 2.60 GHz with 8.00 Gb of RAM. Figures 4, 5, 7 and 8 show the Pareto fronts of Ridge and Lasso generated by NFDA and MatLab, respectively. Subfigure (a) shows the Pareto front of example 1, subfigure (b) shows the Pareto front of example 2 and subfigure (c) shows the Pareto front of example 3. To provide a comparison, Figures 6 and 9 show the overlay of the fronts. We also used the hypervolume metric [34], [29] with the nadir point $(\|b\|_q^q,\|ls\|_q^q)$ as a reference point, to compare the algorithms (Table 1).

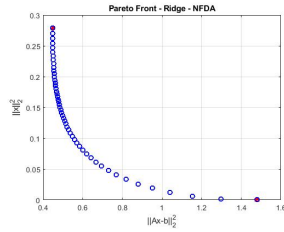
Example 5.1. $A=\begin{bmatrix}1&2\\3&4\end{bmatrix}$, $b=\begin{bmatrix}5\\6\end{bmatrix}$.

Example 5.2. $A=\begin{bmatrix}0.3008&0.8961\\0.9394&0.5975\\0.9809&0.8840\\0.2866&0.9437\\0.8008&0.5492\end{bmatrix}$, $b=\begin{bmatrix}0.7284\\0.5768\\0.0259\\0.4465\\0.6463\end{bmatrix}$.

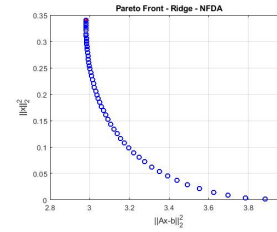
Example 5.3. $A=\begin{bmatrix}1&1&1\\0.01&0&0\\0&0.01&0\\0&0&0.01\end{bmatrix}$, $b=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$.



(a) Example 1

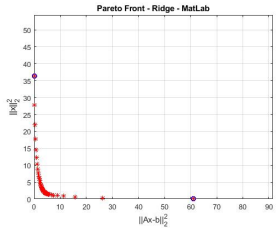


(b) Example 2

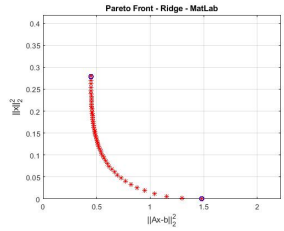


(c) Example 3

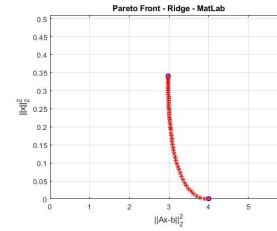
Figure 4: Ridge – NFDA.



(a) Example 1

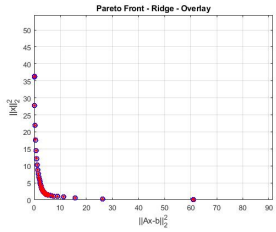


(b) Example 2

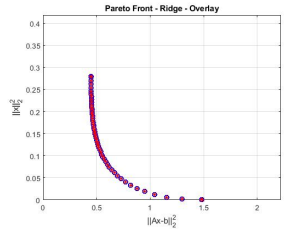


(c) Example 3

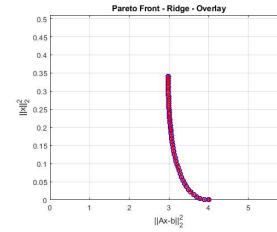
Figure 5: Ridge – MatLab.



(a) Example 1

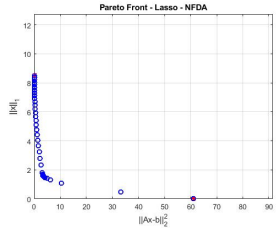


(b) Example 2

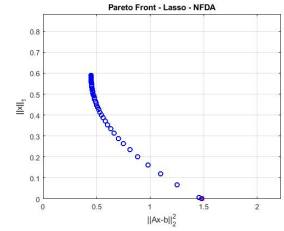


(c) Example 3

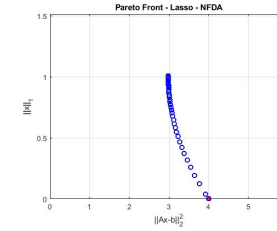
Figure 6: Ridge – Overlay.



(a) Example 1

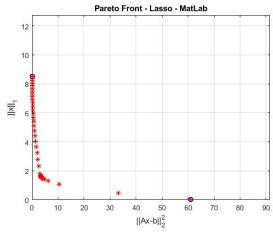


(b) Example 2

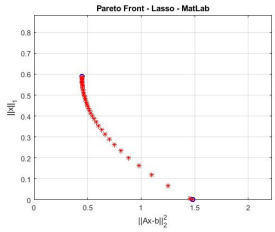


(c) Example 3

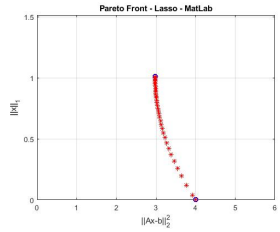
Figure 7: Lasso – NFDA.



(a) Example 1

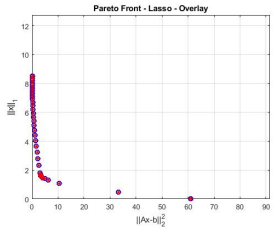


(b) Example 2

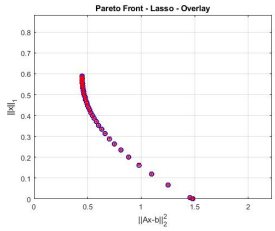


(c) Example 3

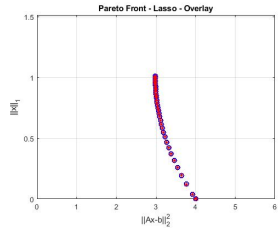
Figure 8: Lasso – MatLab



(a) Example 1



(b) Example 2



(c) Example 3

Figure 9: Lasso – Overlay.

Table 1: The Hypervolume metric.

	Ridge MatLab	Ridge NFDA	Lasso MatLab	Lasso NFDA
Example 1	2146.2	2145.9	456.9948	456.9566
Example 2	0.2481	0.2481	0.3893	0.3893
Example 3	0.2857	0.2856	0.6592	0.6590

5.2 On Time Consumed

In order to provide a comparison regarding the time, we built the Pareto front with 50 points of 7 problems with n features and p variables whose matrices and vectores were randomily created in MatLab. We recall that n is the number of rows of A or the number of features and p is the number of columns of A or the number of variables (coefficients) of the model.

Table 2 shows the average time consumed by the algorithms. Each instance was run 3 times.

Table 2: Time consumed.

size		Ridge		Lasso	
n	p	NFDA	MatLab	NFDA	MatLab
20	10	2.95 s	20.50 s	5.86 s	15.90 s
50	10	4.43 s	21.65 s	6.02 s	16.34 s
60	20	5.80 s	76.06 s	12.84 s	19.45 s
80	20	6.43 s	82.86 s	14.35 s	26.95 s
120	40	10.12 s	225.68 s	64.86 s	89.23 s
160	40	12.46 s	344.43 s	69.98 s	178.70 s
500	100	121.21 s	3884.38 s	2283.76 s	3097.74 s

6 CONCLUSION

We have shown how to transform Ridge and Lasso regressions into bi-objective optimization problems by using Multiobjective Optimization theory. We have presented NFDA, an algorithm for solving convex optimization problems, used it for generating Pareto fronts of some problems and compared the results with those obtained by the MatLab's minimization function fminsearch. The numbers suggest NFDA might be a useful tool for those who work with Ridge and Lasso regressions. Furthermore, as the Pareto front provides a range of models from which the practitioner can choose the most suitable one, it is essential to have a rule to pick the point on the front that leads to the best model. As a future work, we plan to study Pareto fronts of real regression problems to try to find some procedure that provides the best choice.

Acknowledgments

The author is grateful to Regina S. Burachik and C. Yalçin Kaya for the invitation to visit the School of Information Technology and Mathematical Sciences at University of South Australia and for the meaningful debates on the theme of this work. The author also thanks the Department of Mathematics of the Federal University of Juiz de Fora for the sabbatical leave granted in the period 2018-2019 and the University of South Australia for the full support during his academic visit.

REFERENCES

[1] V. Andriopoulos & M. Kornaros. LASSO Regression with Multiple Imputations for the Selection of Key Variables Affecting the Fatty Acid Profile of Nannochloropsis oculata. *Marine Drugs*, **21** (2023), 483. doi:10.3390/md.2109483.

[2] S. Arlot & A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4** (2010), 40–79. doi:10.1214/09-SS054.

[3] K. Bennet & E. Parrado-Hernandez. The Interplay of Optimization and Machine Learning Research. *Journal of Machine Learning Research*, **7** (2006), 1265–1281.

[4] C. Bishop. “Pattern Recognition and Machine Learning”. Springer, New York (2006).

[5] R. Burachik, C. Kaya & M. Rizvi. A New Scalarization Technique and New Algorithms to Generate Pareto Fronts. *SIAM Journal on Optimization*, **27**(2) (2017), 1010–1034. doi:10.1137/16M1083967.

[6] H. Charkhgard & A. Eshragh. A New Approach to Select the Best Subset of Predictors in Linear Regression Modelling: Bi-Objective Mixed Integer Linear Programming. *The Australian and New Zealand Industrial and Applied Mathematical Journal*, **61**(1) (2019), 64–75. doi:10.1017/S1446181118000275.

[7] V. Cherkassky & Y. Ma. Comparison of Model selection for Regression. *Neural Computation*, **15**(7) (2003), 1691–1714. doi:10.1162/089976603321891864.

[8] J. Copas. Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society, Series B, Methodological*, **45**(3) (1983), 311–354. doi:10.1111/j.2517-6161.1983.tb01258.x.

[9] J. Dutta & C. Kaya. A New Scalarization and Numerical Method for Constructing the weak Pareto Front of Multiobjective Optimization Problems. *Optimization*, **60**(8-9) (2011), 1091–1104. doi:10.1080/02331934.2011.587006.

[10] M. Ehrgott. “Multicriteria Optimization”. Springer (2005).

[11] L. Freijeiro-González, M. Febrero-Bande & W. González-Manteiga. A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *International Statistical Review*, **90**(1) (2022), 118–145. doi:10.1111/insr.12469.

[12] W. Freire. “A Feasible Directions Algorithm for Convex Nondifferentiable Optimization”. Ph.D. thesis, Federal University of Rio de Janeiro (2005). URL `http://www.optimize.ufrj.br/files/WilhelmPassarellaFreire.pdf`.

[13] G. Golub & C. Van Loan. “Matrix Computations”. John Hopkins University Press, Baltimore (1983).

[14] N. Hamada & S. Ichiki. Free Disposal Hull Condition to Verify When Efficiency Coincides with Weak Efficiency. *Journal of Optimization Theory and Applications*, **192** (2022), 248–270. doi:10.1007/s10957-021-01961-5.

[15] T. Hastie, J. Taylor, R. Tibshirani & G. Walther. Forward Stagewise Regression and the Monotone Lasso. *Eletronic Journal of Statistics*, **1** (2007), 1–29. doi:10.1214/07-EJS004.

[16] T. Hastie, R. Tibshirani & J. Friedman. “The Elements of Statistical Learning. Data Mining, Inference and Prediction”. Springer (2008).

[17] T. Hastie, R. Tibshirani & M. Wainwright. “Statistical Learning with Sparsity. The Lasso and Generalizations”. CRC Press (2016).

[18] J. Herskovits. Feasible Directions Interior Point Technique for Nonlinear Optimization. *Journal of Optimization Theory and Applications*, **99**(1) (1998), 121–146. doi:10.1023/A:1021752227797.

[19] J. Herskovits, W. Freire, M. Tanaka & A. Canelas. A Feasible Directions Method for Nonsmooth Convex Optimization. *Structural and Multidisciplinary Optimization*, **44**(3) (2011), 363–377. doi:10.1007/s00158-011-0634-y.

[20] J. Hiriart-Urruty & C. Lemarechal. “Convex Analysis and Minimization Algorithms I, II”. Springer (1993).

[21] J. Jahn. “Vector Optimization. Theory, Applications and Extensions”. Springer (2011).

[22] G. James, D. Witten, T. Hastie & R. Tibshirani. “An Introduction to Statistical Learning With Applications in R”. Springer (2013).

[23] Y. Jin & B. Sendhoff. Pareto-Based Multiobjective Machine Learning: An Overview and Cases Studies. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, **38**(3) (2008), 397–415. doi:10.1109/TSMCC.2008.919172.

[24] P. Johansson, H. Henriksen, S. Karvelsson, O. Rolfsson, M. Schønemann-Lund, M. Bestle & S. McGarrity. LASSO regression shows histidine and sphingosine 1 phosphate are linked to both sepsis mortality and endothelial damage. *European Journal of Medical Research*, **29**(1) (2024), 71. doi:10.1186/s40001-023-01612-7.

[25] Y. Jung. Multiple prediction k-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, **30**(1) (2018), 197–215. doi:10.1080/10485252.2017.1404598.

[26] K. Kiwiel. “Methods of Descent for Nondifferentiable Optimization”. Springer-Verlag (1985).

[27] M. Makela & P. Neittaanmaki. “Nonsmooth Optimization. Analysis and Algorithms with Applications to Optimal Control”. World Scientific (1992).

[28] K. Miettinen. “Nonlinear Multiobjective Optimization”. Springer Science + Business Media, LLC (1998).

[29] A. Mohammadi & A. Custódio. A trust-region approach for computing Pareto fronts in multiobjective optimization. *Computational Optimization and Applications*, **87** (2024), 149–179. doi:10.1007/s10589-023-00510-2.

[30] M. Osborne, B. Presnell & B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, **9**(2) (2000), 319–337. doi:10.2307/1390657.

[31] P. Pardalos, A. Zilinskas & J. Zilinskas. “Non-Convex Multi-Objective Optimization”. Springer (2017).

[32] M. Raimundo. “Multi-Objective Optimization in Machine Learning”. Ph.D. thesis, State University of Campinas (2018).

[33] S. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman, D. Awwad & Z. Abdalla. Optimizing Linear Regression Models with Lasso and Ridge Regression:A Study on UAE Financial Behavior during COVID-19. *Migration Letters*, **20**(6) (2023), 139–153. doi:10.59670/ml.v20i6.3468.

[34] K. Shang, H. Ishibuchi, L. He & L. Pang. A Survey on the Hypervolume Indicator in Evolutionary Multi-objective Optimization. *IEEE Transactions on Evolutionary Computation*, **25**(1) (2021), 1–20. doi:10.1109/TEVC.2020.3013290.

[35] B. Sloboda, D. Pearson & M. Etherton. An application of the LASSO and elastic net regression to assess poverty and economic freedom on ECOWAS countries. *Mathematical Biosciences and Engineering*, **20**(7) (2023), 12154–12168. doi:10.3934/mbe.2023541.

[36] T. Suttorp & C. Igel. Multi-objective Optimization of Support Vector Machines. *Studies in Computational Intelligence*, **16** (2006), 199–220. doi:10.1007/3-540-33019-49.

How to cite

W. P. Freire. The NFDA-Nonsmooth Feasible Directions Algorithm applied to constructing Pareto Fronts of Ridge and Lasso Regressions. *Trends in Computational and Applied Mathematics*, **25**(2024), e01767. doi: 10.5540/tcam.2024.025.e01767.