Topics in Large-Scale
Sparse Estimation and Control

by

Tarek Sami Rabbani

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Mechanical Engineering
and the Designated Emphasis
in
Computational Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Laurent El Ghaoui, Chair
Professor Andrew Packard, Co-Chair
Professor Francesco Borrelli
Professor Alexandre Bayen

Spring 2013

Topics in Large-Scale
Sparse Estimation and Control

Abstract

Topics in Large-Scale
Sparse Estimation and Control

by

Tarek Sami Rabbani

Doctor of Philosophy in Engineering-Mechanical Engineering
and the Designated Emphasis
in
Computational Science and Engineering

University of California, Berkeley

Professor Laurent El Ghaoui, Chair

In this thesis, we study two topics related to large-scale sparse estimation and control. In the first topic, we describe a method to eliminate features (variables) in ℓ_1-regularized convex optimization problems. The elimination of features leads to a potentially substantial reduction in computational effort needed to solve such problems, especially for large values of the penalty parameter. Our method is not heuristic: it only eliminates features that are guaranteed to be absent after solving the optimization problem. The feature elimination step is easy to parallelize and can test each feature for elimination independently. Moreover, the computational effort of our method is negligible compared to that of solving the convex problem.

We study the case of ℓ_1-regularized least-squares problem (a.k.a. LASSO) extensively and derive a closed-form sufficient condition for eliminating features. The sufficient condition can be evaluated by few vector-matrix multiplications. For comparison purposes, we present a LASSO solver that integrates SAFE with the Coordinate Descent method. We call our method CD-SAFE, and we report the number of computations needed for solving a LASSO problem using CD-SAFE and using the plain Coordinate Descent method. We observe at least a 100 fold reduction in computational complexity for dense and sparse data-sets consisting of millions of variables and millions of observations. Some of these data-sets can cause memory problems when loaded, or need specialized solvers. However, with SAFE, we can extend LASSO solvers capabilities to treat large-scale problems, previously out of their reach. This is possible, because SAFE eliminates variables and thus portions of our data at the outset, before loading it into our memory.

We also show how our method can be extended to general ℓ_1-regularized convex problems. We present preliminary results for the Sparse Support Vector Machine and Logistic Regression problems.

In the second topic of the thesis, we derive a method for open-loop control of open channel flow, based on the Hayami model, a parabolic partial differential equation resulting from a simplification of the Saint-Venant equations. The open-loop control is represented as infinite series using differential flatness, for which convergence is assessed. Numerical simulations show the effectiveness of the approach by applying the open-loop controller to irrigation canals modeled by the full Saint-Venant equations.

We experiment with our controller on the Gignac Canal, located northwest of Montpellier, in southern France. The experiments show that it is possible to achieve a desired water flow at the downstream of a canal using the Hayami model as an approximation of the real-system. However, our observations of the measured water flow at the upstream controlled gate made us realize some actuator limitations. For example, deadband in the gate opening and unmodeled disturbances such as friction in the gate-opening mechanism, only allow us to deliver piece-wise constant control inputs. This fact made us investigate a way to compute a controller that respects the actuator limitations. We use the CD-SAFE algorithm, to compute such open-loop control for the upstream water flow. We compare the computational effort needed to obtain an open-loop control with certain dynamics using the CD-SAFE algorithm and the plain Coordinate Descent algoirthm. We show that with CD-SAFE we are able to obain an open-loop control signal with cheaper computations.

To my family...

Contents

List of Figures

List of Tables

Acknowledgments

I am forever indebted to Professor Laurent El Ghaoui, who guided me through out my academic years and made sure to set me on the right career path. I also like to thank Professor Alexandre Bayen for advising me while I was an undergraduate intern in his lab. Alex support was a cornerstone in my academic achievements and a key part of my Masters and PhD theses. I must also thank Professor Andy Packard who advised me during all my year in Berkeley on various academic and non-academic matters. I am grateful to have had the opportunity to work with Professor Francesco Borrelli in teaching the Control Systems course (ME 132). Finally, I would like to thank Francesco, Alex, Laurent and Andy for serving on my PhD thesis committee.

Part I

Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems

Chapter 1

Introduction

In the first part of the thesis, we present a method that can eliminate variables in an ℓ_1-regularized convex optimization problem, that arises in statistics [54], signal processing [8], machine learning, engineering and other fields. The variables or features selected for elimination is done in a cheap pre-processing step a-priori to solving the problem. Our method is novel because it is not a heuristic, any feature eliminated is guaranteed to be absent after solving the problem with all its original features. Thus we give our method the name Safe Feature Elimination (SAFE). Although we concentrate in this work on the ℓ_1-regularized least-squares problem or the LASSO, the main idea behind our method can be generalized to other convex problems involving ℓ_1 regularization.

Performing SAFE depends only on the features of interest for elimination and thus our method can be run independently of other features in the model and in parallel. For extremely large datasets, where there are millions of observations, high dimensions or both, the bottle-neck in processing the data can be just in loading the data into memory. SAFE provides a way to resolve this issue by reducing dimensions or eliminating features. SAFE is a preprocessing method that can complement the specific solvers of a particular ℓ_1-regularized convex problem and possibly introduce huge savings in computational complexity by eliminating features. For example, the interior-point method for the LASSO in [27] has a complexity of order $O\left(\min(n,m)^2 \max(n,m)\right)$ flops, where n is the number of variables (features) and m the number of data points. Hence it is of interest to be able to efficiently eliminate features in a pre-processing step and reduce the memory requirements and computational cost for solving the problem.

An interesting fact is that SAFE can be very aggressive at removing features at some particular values of the regularization parameter. The specific application we have in mind involves large data sets of text documents, and sparse matrices based on occurrence, or other score, of words or terms in these documents. We seek extremely sparse optimal solutions, even if this means operating at values of the penalty parameter that are substantially larger than those dictated by a pure concern for predictive accuracy. This fact opens the hope that, at least for the application considered, the number of features eliminated by using our

SAFE method is high enough to allow a dramatic reduction in computing time and memory requirements.

1.1 Related Work

Feature selection methods are often used to accomplish dimensionality reduction, and are of utmost relevance for data-sets of massive dimension, see for example [14]. These methods, when used as a pre-processing step, have been referred to in the literature as *screening* procedures [14, 15]. They typically rely on univariate models to score features, independently of each other, and are usually computationally fast. Classical procedures are based on correlation coefficients, two-sample t-statistics or chi-square statistics [14]; see also [18] and the references therein for an overview in the specific case of text classification. Most screening methods might remove features that could otherwise have been selected by the regression or classification algorithm. However, some of them were recently shown to enjoy the so-called "sure screening" property [15]: under some technical conditions, no relevant feature is removed, with probability tending to one.

Screening procedures typically ignore the specific regression or classification task to be solved after feature elimination. In this work, we propose to remove features based on the supervised learning problem considered, that is on both the structure of the loss function and the problem data. Unlike screening procedures, the features in SAFE are eliminated according to a sufficient, in general conservative condition.

1.2 Contributions

The first contribution of this part of the dissertation is the formulation of the Safe Feature Elimination problem (chapters 3 and 5), or the SAFE test problem. The SAFE test problem is a convex optimization problem, whose solution can be used to construct a sufficient condition for eliminating features in an ℓ_1-regularized convex optimization problem. The formulation depends on the structure of the loss function and (only) on the feature to be tested for elimination. The formulation allows us to construct sufficient conditions for eliminating multiple features at the same time, or in parallel, because each SAFE test problem for some feature is independent of the others.

The second contribution is the closed-form solution of the SAFE test problem when the loss function of the convex problem considered is the square-loss function (chapters 3 and 4), i.e. the ℓ_1-regularized least-squares problem. The explicit solution allows us to construct the sufficient condition for eliminating features very efficiently, with a negligible total cost of computations.

1.3 Organization

The first part of the dissertation is organized as follows. In chapter 2, we introduce the LASSO problem or the ℓ_1-regularized least-squares problem. The chapter is self-contained, we explain all the background needed to derive our Safe Feature Elimination method, including the concepts of a Dual Problem, weak and strong duality. We also describe the solution of the LASSO problem as a function of the ℓ_1 regularization used. Although in this chapter we derive the dual problem and strong duality theory for the LASSO, the procedures and concepts used in the derivation are the same for other convex problems.

In chapter 3, we derive the Safe Feature Elimination method for the LASSO. We explain how to use SAFE to reduce memory requirements for solving a LASSO problem and how to improve the computational cost of LASSO solvers. We verify the advantages that SAFE provides using numerical experiments with datasets obtained from text classification problems.

In chapter 4, we derive a SAFE method more aggressive at removing features than the one introduced in chapter 3. We express the sufficient condition for eliminating features in closed-form, thus allowing us to efficiently implement the method. We also integrate our SAFE test in a Coordinate-Descent algorithm for solving the LASSO, we call our algorithm CD-SAFE. We experiment with our algorithm using large-scale datasets, some are large enough to cause memory problems when just loading them[1]. We show that it is possible to treat datasets with fewer computational operations than using the plain Coordinate-Descent algorithm. This improvement in computational complexity allows us to extend the reach of the Coordinate-Descent algorithm to large-scale problems.

In chapter 5, we adapt our Safe Feature Elimination method for the LASSO to a more general class of l_1- regularized convex problems. We show some preliminary results for deriving SAFE methods for the ℓ_1-regularized hinge loss function (also known as ℓ_1-regularized support vector machine) and ℓ_1-regularized logistic regression.

1.4 Notations

For notations, we represent scalars by lower-case font and vectors by lower-case bold font like $\boldsymbol{v} = [v_1, ..., v_n]^T$. For any vector $\boldsymbol{v}$, we consider the following representation of sub-vectors:

$$\boldsymbol{v}_{\mathcal{A}} = \left[v_{a_1}, ..., v_{a_{|\mathcal{A}|}}\right],$$

where $\mathcal{A} = \left\{a_1, ..., a_{|\mathcal{A}|}\right\}$ is an index set and $|\mathcal{A}|$ is its cardinality. Sometimes we refer to $\mathcal{A}_k$ as the k^{th} element in the set $\mathcal{A}$. Depending on context, we use small Greek letters to refer to vectors like $\boldsymbol{\theta} = [\theta_1, .., \theta_m]^T$ or scalars like ν. We refer to matrices by uppercase bold font

[1]We assume a machine is used with 8 GB of RAM .

like $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, and similarly we represent sub-matrices by

$$\boldsymbol{X}_{:,\mathcal{A}} = \left[\mathbf{x}_{a_1}, ..., \mathbf{x}_{a_{|\mathcal{A}|}}\right].$$

We also refer to the i^{th} entry of a vector $\boldsymbol{x}_j$ as $\boldsymbol{x}_j(i)$.

Chapter 2

All about LASSO

2.1 Introduction

Least Absolute Shrinkage and Selection Operator (LASSO) [54] is the ℓ_1-regularized least-squares problem,

$$\mathcal{P}(\lambda) \ : \ \phi(\lambda) := \min_{\boldsymbol{w}} \frac{1}{2} \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + \lambda \left\| \boldsymbol{w} \right\|_1 , \tag{2.1}$$

where $\boldsymbol{X} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)^T \in \mathbb{R}^{m \times n}$ is the feature matrix of observations, $\boldsymbol{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ is a given set of m observations, $\boldsymbol{y} \in \mathbb{R}^m$ is the response vector,$\lambda > 0$ is a regularization parameter and $\boldsymbol{w} \in \mathbb{R}^n$ is the optimization variable.

The ℓ_1 norm regularization introduces some attractive properties to the solution of the LASSO problem. One of these important properties is that a solution $\boldsymbol{w}^\star$ of $\mathcal{P}(\lambda)$ is sparse or has few non-zero elements. More specifically, there exist a sequence of increasing values of λ: $0 = \lambda_k < \cdots < \lambda_0 = \lambda_{\max}$ where the solution $\boldsymbol{w}^\star$ is piece-wise linear as a function of λ with breakpoints at $\lambda = \lambda_i$, $i = 0, ..., k-1$ and $\boldsymbol{w}^\star = \boldsymbol{0}$ for all $\lambda \geq \lambda_{\max}$ (see section 2.3 for more details).

To illustrate the sparsity of the LASSO solution, we show the regularization path for the LASSO problem, the solution as a function of λ, solved on the diabetes dataset [43] in Figure 2.1(a). We also show the number of non-zero elements in the solution in Figure 2.1(b). The diabetes dataset has 10 features, at $\lambda/\lambda_0 = 1$, all elements in the solution are zero as shown in Figure 2.1(b). For $\lambda/\lambda_0 < 1$, breakpoints happen when a zero entry of $\boldsymbol{w}^\star$ becomes a non-zero, or vice-versa. Generally, the number of non-zeros in the solution increases for lower values of the regularization parameter.

The concepts and derivations presented in this chapter are crucial for the understanding of our Safe Feature Elimination method for the LASSO. We start by deriving a dual problem of the LASSO in section 2.2. We derive the strong duality result of the LASSO and present the optimality conditions. We then derive the regularization path result in section 2.3.

2.2 A Dual Problem of the LASSO

The Safe Feature Elimination method (SAFE) method relies on duality and optimality conditions. We review and derive the appropriate facts for the LASSO.

Generally, a dual problem is a transformation on the primal problem, and has properties that are related to the primal problem. For instance, it provides a lower bound on the value of the objective function of the primal problem. For the LASSO problem, strong duality holds and the value of the objective function of the primal problem at optimum can be recovered by solving its dual. Moreover, knowing the optimal solution of the dual problem, or the dual optimal point, allows us to identify the zeros and non-zeros of the optimal solution of the primal problem.

A dual to the LASSO problem (2.1) is

$$\mathcal{D}(\lambda) \ : \ \phi'(\lambda) := \max_{\boldsymbol{\theta}} \ G(\boldsymbol{\theta}) \ : \ \left|\boldsymbol{\theta}^T \boldsymbol{x}_k\right| \leq \lambda, \ k = 1, \ldots, n, \tag{2.2}$$

with $\boldsymbol{x}_k \in \mathbb{R}^m$, $k = 1, \ldots, n$, the k-th column of $\boldsymbol{X}$ and $G(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{y}\|_2^2 - \frac{1}{2} \|\boldsymbol{\theta} + \boldsymbol{y}\|_2^2$. In this context, we call $\mathcal{P}(\lambda)$ the primal problem, $\boldsymbol{w}$ the primal variable, and $\boldsymbol{w}^\star$ a primal optimal point. The dual problem $\mathcal{D}(\lambda)$ is a convex optimization problem with dual variable $\boldsymbol{\theta} \in \mathbb{R}^m$. We call $\boldsymbol{\theta}$ dual feasible when it satisfies the constraints in $\mathcal{D}(\lambda)$. Figure 2.2 shows the geometry of the feasibility set in the dual space.

Weak duality implies that the quantity $G(\boldsymbol{\theta})$ gives a lower bound on the optimal value $\phi(\lambda)$ for any dual feasible point $\boldsymbol{\theta}$, i.e. $G(\boldsymbol{\theta}) \leq \phi'(\lambda) \leq \phi(\lambda)$, $\left|\boldsymbol{\theta}^T \boldsymbol{x}_k\right| \leq \lambda$, $k = 1, \ldots, n$. Since strong duality holds for the LASSO, the optimal value of $\phi'(\lambda)$ achieves $\phi(\lambda)$, $\phi'(\lambda) = \phi(\lambda)$, at $\boldsymbol{\theta}^\star$ the dual optimal point. Furthermore, we can construct a dual optimal point from a primal optimal point using the relation $\boldsymbol{\theta}^\star = \boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}$. In addition, knowledge of $\boldsymbol{\theta}^\star$ allows us to identify the zeros in $\boldsymbol{w}^\star$ by checking the optimality condition

$$\left|\boldsymbol{\theta}^{\star T} \boldsymbol{x}_k\right| < \lambda \Rightarrow \boldsymbol{w}^\star(k) = 0. \tag{2.3}$$

In this section, we derive the aforementioned dual problem in (2.2), in addition to the weak and strong duality results.

2.2.1 Dual Problem and Weak Duality

We derive (2.2) by introducing an equivalent problem to $\mathcal{P}(\lambda)$,

$$\mathcal{P}(\lambda) \ : \ \phi(\lambda) := \min_{\boldsymbol{w}, \boldsymbol{z}} \frac{1}{2} \|\boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 \ : \ \boldsymbol{z} = \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}, \tag{2.4}$$

where we have defined a new slack variable $\boldsymbol{z} \in \mathbb{R}^m$, and have set this variable equal to $\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$, i.e. $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$. We form the Lagrangian function by associating the dual variable, $\boldsymbol{\theta} \in \mathbb{R}^m$, to the equality constraint,

$$L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 + \boldsymbol{\theta}^T \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} - \boldsymbol{z}\right).$$

The partial maximization of $L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ has the same value as the objective function of (2.4) subject to the equality constraints. This fact,

$$\max_{\boldsymbol{\theta}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) = \begin{cases} \frac{1}{2} \|\boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 : & \boldsymbol{z} = \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}, \\ +\infty & \text{otherwise}, \end{cases}$$

can be recognized by substituting any infeasible point $(\boldsymbol{z}, \boldsymbol{w})$ with $\boldsymbol{z} \neq \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$ into both functions. By convention, the objective function of $\mathcal{P}(\lambda)$, takes the value $+\infty$ for infeasible points. Taking $\boldsymbol{\theta}(i) = t\mathbf{sign}\left(\boldsymbol{a}_i^T \boldsymbol{w} - y_i - z_i\right),\ i = 1, ..., m$, with $t \to +\infty$, the supremum of $L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ also takes the value $+\infty$ and thus the two formulations are equivalent.

We rewrite the primal problem in terms of the Lagrangian function,

$$\mathcal{P}(\lambda) \; : \; \phi(\lambda) := \min_{\boldsymbol{w}, \boldsymbol{z}} \max_{\boldsymbol{\theta}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}), \tag{2.5}$$

and use the min-max inequality,

$$\max_{\boldsymbol{\theta}} \min_{\boldsymbol{w}, \boldsymbol{z}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) \leq \min_{\boldsymbol{w}, \boldsymbol{z}} \max_{\boldsymbol{\theta}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}). \tag{2.6}$$

We define the dual function

$$g(\boldsymbol{\theta}) = \min_{\boldsymbol{w}, \boldsymbol{z}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}),$$

and call the problem,

$$\mathcal{D}(\lambda) \; : \; \phi'(\lambda) := \max_{\boldsymbol{\theta}} \; g(\boldsymbol{\theta}), \tag{2.7}$$

a dual problem of the LASSO. We note that $g(\boldsymbol{\theta})$ provides a lower bound on $\phi(\lambda)$ of $\mathcal{P}(\lambda)$ for any feasible dual variable $\boldsymbol{\theta}$, i.e. $g(\boldsymbol{\theta}) \leq \phi'(\lambda) \leq \phi(\lambda)$. This result is noted as weak duality.

Expression of $g(\boldsymbol{\theta})$. We write $g(\boldsymbol{\theta})$ as a minimization problem over each variable $\boldsymbol{w}$ and $\boldsymbol{z}$,

$$\begin{aligned} g(\boldsymbol{\theta}) &= \min_{\boldsymbol{w}, \boldsymbol{z}} \; L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}), \\ &= \min_{\boldsymbol{w}, \boldsymbol{z}} \; \frac{1}{2} \|\boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 + \boldsymbol{\theta}^T \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} - \boldsymbol{z}\right), \\ &= \min_{\boldsymbol{z}} \; \left(\frac{1}{2} \|\boldsymbol{z}\|_2^2 - \boldsymbol{\theta}^T \boldsymbol{z}\right) + \min_{\boldsymbol{w}} \; \left(\lambda \|\boldsymbol{w}\|_1 + \boldsymbol{\theta}^T \boldsymbol{X}\boldsymbol{w}\right) - \boldsymbol{\theta}^T \boldsymbol{y}, \end{aligned}$$

then we decompose the variables into summations,

$$\begin{aligned} g(\boldsymbol{\theta}) &= \min_{\boldsymbol{z}} \sum_{i=1}^{m} \left(\frac{1}{2} z_i^2 - \theta_i z_i\right) + \min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(\lambda |w_i| + \left(\boldsymbol{x}_i^T \boldsymbol{\theta}\right) w_i\right) - \boldsymbol{\theta}^T \boldsymbol{y}, \\ &= \sum_{i=1}^{m} \min_{z_i} \left(\frac{1}{2} z_i^2 - \theta_i z_i\right) + \lambda \sum_{i=1}^{n} \left(\min_{w_i} |w_i| + \frac{1}{\lambda} \left(\boldsymbol{x}_i^T \boldsymbol{\theta}\right) w_i\right) - \boldsymbol{\theta}^T \boldsymbol{y}. \end{aligned} \tag{2.8}$$

We define the following two optimization problems that appear in (2.8),

$$T_1(\alpha) := \min_z \ \frac{1}{2}z^2 - \alpha z, \tag{2.9}$$

and

$$T_2(\alpha) := \min_w \ |w| - \alpha w. \tag{2.10}$$

The optimal solution $z^\star$ of $T_1(\alpha)$ is $z^\star = \alpha$, and $T_1(\alpha)$ takes the value

$$T_1(\alpha) = -\frac{1}{2}\alpha^2.$$

The optimal solution $w^\star$ of $T_2(\alpha)$, satisfies the conditions,

$$w^\star = \left\{ w \;\middle|\; \begin{cases} w \geq 0 & \alpha = 1, \\ w \leq 0 & \alpha = -1, \\ w = 0 & |\alpha| < 1, \end{cases} \right\}, \tag{2.11}$$

and $T_2(\alpha)$ takes the value

$$T_2(\alpha) = \begin{cases} 0 & |w| \leq 1, \\ -\infty & \text{otherwise.} \end{cases}$$

Using the results above, we obtain

$$g(\boldsymbol{\theta}) = \begin{cases} G(\boldsymbol{\theta}) & \left|\boldsymbol{x}_i^T\boldsymbol{\theta}\right| \leq \lambda \ i = 1, ..., n, \\ -\infty & \text{otherwise,} \end{cases} \tag{2.12}$$

with $G(\boldsymbol{\theta}) = -\frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2 - \boldsymbol{\theta}^T\boldsymbol{y}$.

A Weak Dual Problem. We substitute the expression of $g(\boldsymbol{\theta})$ in (2.7) and we obtain

$$\mathcal{D}(\lambda) \ : \ \phi'(\lambda) := \max_{\boldsymbol{\theta}} \ G(\boldsymbol{\theta}) \ : \ \left|\boldsymbol{x}_i^T\boldsymbol{\theta}\right| \leq \lambda, \ i = 1, ..., n, \tag{2.13}$$

with $G(\boldsymbol{\theta}) = \frac{1}{2}\left\|\boldsymbol{y}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\theta} + \boldsymbol{y}\right\|_2^2 := -\frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2 - \boldsymbol{\theta}^T\boldsymbol{y}$. For any point $\boldsymbol{\theta} \in \mathbb{R}^m$, $g(\boldsymbol{\theta})$ gives a lower bound on the objective value of the LASSO problem $\phi(\lambda)$. For non-feasible $\boldsymbol{\theta}$, $\boldsymbol{\theta} \notin \left\{\boldsymbol{\theta} \,\middle|\, \left|\boldsymbol{x}_i^T\boldsymbol{\theta}\right| \leq \lambda, \ i = 1, ..., n\right\}$, the lower bound is trivial ($-\infty$). More interesting lower bounds are obtained when $\boldsymbol{\theta}$ is feasible. The best lower bound is $\phi'(\lambda)$ and is obtained at $\boldsymbol{\theta}^\star$ the dual optimal point of $\mathcal{D}(\lambda)$, i.e. $\phi'(\lambda) = G(\boldsymbol{\theta}^\star)$. We call the gap between $\phi(\lambda)$ and $\phi'(\lambda)$, $g(\lambda) = \phi(\lambda) - \phi'(\lambda)$, the duality gap and it is always non-negative, i.e. $g(\lambda) \geq 0$.

2.2.2 Strong Duality

For the LASSO problem, strong duality is obtained, which is to say equality holds in (2.6) and the duality gap is zero, i.e. $g(\lambda) = 0$. As a consequence of strong duality, the solutions, $\boldsymbol{\theta}$, $\boldsymbol{w}$, $\boldsymbol{z}$ are the same in both formulations of (2.6). This allows us to make two conclusions on the relation between the primal solution $\boldsymbol{w}^\star$ and the solution of the dual problem $\boldsymbol{\theta}^\star$. From the optimal solution of (2.9), we have $\boldsymbol{\theta}^\star = \boldsymbol{z}^\star := \boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}$, and from the optimal solution of (2.10), we have

$$\boldsymbol{w}^\star = \left\{ \boldsymbol{w}^\star \;\middle|\; \begin{cases} \boldsymbol{w}^\star(i) \geq 0 & \boldsymbol{x}_i^T\boldsymbol{\theta}^\star = -\lambda, \\ \boldsymbol{w}^\star(i) \leq 0 & \boldsymbol{x}_i^T\boldsymbol{\theta}^\star = \lambda, \quad , i = 1, ..., n \\ \boldsymbol{w}^\star(i) = 0 & \left|\boldsymbol{x}_i^T\boldsymbol{\theta}^\star\right| < \lambda, \end{cases} \right\}. \tag{2.14}$$

In this section, we prove the Strong Duality theorem for the LASSO.

Theorem 2.2.1 (Strong Duality of the LASSO) *Consider the Dual Problem $\mathcal{D}(\lambda)$ in (2.13), and assume that $\phi'(\lambda)$ is finite and $\boldsymbol{\theta}^\star$ is an optimal solution. Let $\mathcal{I}$ be the set of all indices, $\mathcal{I} = \{1, ..., n\}$, and $\mathcal{A}^\pm$ be the sets of active constraints at $\boldsymbol{\theta}^\star$:*

$$\boldsymbol{x}_i^T\boldsymbol{\theta}^\star = \lambda : \; i \in \mathcal{A}^-, \quad \boldsymbol{x}_i^T\boldsymbol{\theta}^\star = -\lambda : \; i \in \mathcal{A}^+, \quad \left|\boldsymbol{x}_i^T\boldsymbol{\theta}^\star\right| < \lambda : \; i \notin \mathcal{A} = \mathcal{A}^- \cup \mathcal{A}^+,$$

Then the following statements hold true:

1. *There exist a $\boldsymbol{w}^\star = (w_1, ..., w_n)^T \in \mathbb{R}^n$ that satisfies*

$$w_i \geq 0 : \; i \in \mathcal{A}^+, \quad w_i \leq 0 : \; i \in \mathcal{A}^-, \quad w_i = 0 : \; i \notin \mathcal{A}, \qquad \sum_{i \in \mathcal{A}} \boldsymbol{x}_i w_i = \boldsymbol{\theta}^\star + \boldsymbol{y}.$$

2. *The point $\boldsymbol{w}^\star$ is the optimal solution of the LASSO problem given in (2.1).*

3. *The value $\phi'(\lambda)$ achieves $\phi(\lambda)$ at the optimal solution $\boldsymbol{\theta}^\star$.*

Proof: Let $\bar{\boldsymbol{X}}^\pm$ be the matrices defined by the indices $\mathcal{A}^\pm$,

$$\bar{\boldsymbol{X}}^+ = \boldsymbol{X}_{:,\mathcal{A}^+} \in \mathbb{R}^{m \times \left|\mathcal{A}^+\right|}, \quad \bar{\boldsymbol{X}}^- = \boldsymbol{X}_{:,\mathcal{A}^-} \in \mathbb{R}^{m \times \left|\mathcal{A}^-\right|},$$

respectively. We assume that there is no $\bar{\boldsymbol{w}}^+ \succeq \boldsymbol{0}$ with $\bar{\boldsymbol{w}}^+ \in \mathbb{R}^{\left|\mathcal{A}^+\right|}$, and no $\bar{\boldsymbol{w}}^- \preceq \boldsymbol{0}$ with $\bar{\boldsymbol{w}}^- \in \mathbb{R}^{\left|\mathcal{A}^-\right|}$, such that $\bar{\boldsymbol{X}}^+\bar{\boldsymbol{w}}^+ + \bar{\boldsymbol{X}}^-\bar{\boldsymbol{w}}^- = \boldsymbol{\theta}^\star + \boldsymbol{y}$, i.e. We need to show that there exist $\bar{\boldsymbol{w}}^+ \in \mathbb{R}^{\left|\mathcal{A}^+\right|}$ and $\bar{\boldsymbol{w}}^- \in \mathbb{R}^{\left|\mathcal{A}^-\right|}$ that satisfy

$$\boldsymbol{\theta}^\star + \boldsymbol{y} \notin S = \left\{ \bar{\boldsymbol{X}}^+\bar{\boldsymbol{w}}^+ + \bar{\boldsymbol{X}}^-\bar{\boldsymbol{w}}^- \,\middle|\, \bar{\boldsymbol{w}}^+ \succeq \boldsymbol{0}, \; \bar{\boldsymbol{w}}^- \preceq \boldsymbol{0} \right\}.$$

By the strict separating hyperplane theory, applied to $\boldsymbol{\theta}^\star + \boldsymbol{y}$ and $\mathcal{S}$, there exist a $\boldsymbol{u}$ such that

$$\boldsymbol{u}^T \left(\boldsymbol{\theta}^\star + \boldsymbol{y}\right) > \boldsymbol{u}^T \bar{\boldsymbol{X}}^+ \bar{\boldsymbol{w}}^+ + \boldsymbol{u}^T \bar{\boldsymbol{X}}^- \bar{\boldsymbol{w}}^-,$$

for all $\bar{\boldsymbol{w}}^+ \succeq \mathbf{0}$ and $\bar{\boldsymbol{w}}^- \preceq \mathbf{0}$.

Evaluating the right-hand side at $\bar{\boldsymbol{w}}^+ = \mathbf{0}$ and $\bar{\boldsymbol{w}}^- = \mathbf{0}$, we obtain

$$\boldsymbol{u}^T \left(\boldsymbol{\theta}^\star + \boldsymbol{y}\right) > 0 \geq \boldsymbol{u}^T \bar{\boldsymbol{X}}^+ \bar{\boldsymbol{w}}^+ + \boldsymbol{u}^T \bar{\boldsymbol{X}}^- \bar{\boldsymbol{w}}^-.$$

Taking the right-hand side of the above inequality,

$$0 \geq \boldsymbol{u}^T \bar{\boldsymbol{X}}^+ \bar{\boldsymbol{w}}^+ + \boldsymbol{u}^T \bar{\boldsymbol{X}}^- \bar{\boldsymbol{w}}^-, \tag{2.15}$$

and evaluating it at $\bar{\boldsymbol{w}}^- = \mathbf{0}$, we obtain

$$0 \geq \boldsymbol{u}^T \bar{\boldsymbol{X}}^+ \bar{\boldsymbol{w}}^+.$$

Since $\bar{\boldsymbol{w}}^+ \succeq \mathbf{0}$, we have $\left(\bar{\boldsymbol{X}}^+\right)^T \boldsymbol{u} \preceq 0$. Similarly we evaluate (2.15) at $\bar{\boldsymbol{w}}^- = \mathbf{0}$ and we obtain $\left(\bar{\boldsymbol{X}}^-\right)^T \boldsymbol{u} \succeq 0$.

We consider $\boldsymbol{\theta} = \boldsymbol{\theta}^\star + t\boldsymbol{u}$. We have the following fact, $\boldsymbol{\theta}$ is feasible, i.e. $\left|\boldsymbol{x}_i^T \boldsymbol{\theta}\right| \leq \lambda,\ i \in \mathcal{I}$, for some sufficiently small negative values of t . Consider any index $i \in \mathcal{A}^-$, the inequality

$$\boldsymbol{x}_i^{-T} \boldsymbol{\theta} = \boldsymbol{x}_i^{-T} \boldsymbol{\theta}^\star + t\boldsymbol{x}_i^{-T} \boldsymbol{u} = \lambda + t\boldsymbol{x}_i^{-T} \boldsymbol{u} \leq \lambda,$$

holds true for $t \leq 0$ since $\boldsymbol{x}_i^{-T} \boldsymbol{u} \geq 0$. And, for any index $i \in \mathcal{A}^+$, the inequality

$$\boldsymbol{x}_i^{+T} \boldsymbol{\theta} = \boldsymbol{x}_i^{+T} \boldsymbol{\theta}^\star + t\boldsymbol{x}_i^{+T} \boldsymbol{u} = -\lambda + t\boldsymbol{x}_i^{+T} \boldsymbol{u} \geq -\lambda,$$

holds true for $t \leq 0$ since $\boldsymbol{x}_i^{+T} \boldsymbol{u} \leq 0$. When the index $i \notin \mathcal{A} = \mathcal{A}^- \cup \mathcal{A}^+$, the inequality

$$\left|\boldsymbol{x}_i^T \boldsymbol{\theta}\right| = \left|\boldsymbol{x}_i^T \boldsymbol{\theta}^\star + t\boldsymbol{x}_i^T \boldsymbol{u}\right| \leq \lambda,$$

holds true for sufficiently small negative t. More specifically, it holds true for any $t \in [\bar{t}, 0]$ with

$$\bar{t} = \inf \left\{ \frac{1}{\left|\boldsymbol{x}_i^T \boldsymbol{u}\right|} \left(\lambda - \left|\boldsymbol{x}_i^T \boldsymbol{\theta}^\star\right|\right) \mid \quad i \notin \mathcal{A} \right\}.$$

Finally, we evaluate the objective function of the Dual Problem $\mathcal{D}(\lambda)$ at $\boldsymbol{\theta}$,

$$\begin{aligned} G(\boldsymbol{\theta}) &= \frac{1}{2} \left\|\boldsymbol{y}\right\|_2^2 - \frac{1}{2} \left\|\boldsymbol{\theta}^\star + t\boldsymbol{u} + \boldsymbol{y}\right\|_2^2, \\ &= \frac{1}{2} \left\|\boldsymbol{y}\right\|_2^2 - \frac{1}{2} \left\|\boldsymbol{\theta}^\star + \boldsymbol{y}\right\|_2^2 - \frac{1}{2} \left\|t\boldsymbol{u}\right\|_2^2 - t\boldsymbol{u}^T \left(\boldsymbol{\theta}^\star + y\right). \end{aligned}$$

We have the following inequality,

$$G(\boldsymbol{\theta}) \;\; \geq \;\; \frac{1}{2}\left\|\boldsymbol{y}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\theta}^\star + \boldsymbol{y}\right\|_2^2,$$

when the condition

$$-\frac{1}{2}\left\|t\boldsymbol{u}\right\|_2^2 - t\boldsymbol{u}^T\left(\boldsymbol{\theta}^\star + \boldsymbol{y}\right) \leq 0, \tag{2.16}$$

holds true. Assuming $t \leq 0$, we have

$$\frac{1}{2}t^2 + t\tilde{\boldsymbol{u}}^T\left(\boldsymbol{\theta}^\star + y\right) \leq 0,$$

with $\tilde{\boldsymbol{u}} = \frac{\boldsymbol{u}}{\left\|\boldsymbol{u}\right\|_2^2}$. The inequality above is equivalent to

$$t \geq -2\tilde{\boldsymbol{u}}^T\left(\boldsymbol{\theta}^\star + y\right),$$

or

$$t \in \left[-2\tilde{\boldsymbol{u}}^T\left(\boldsymbol{\theta}^\star + y\right), 0\right].$$

This is a contradiction, because we have constructed a feasible point $\boldsymbol{\theta}$ with an objective value greater than $G(\boldsymbol{\theta}^\star)$.

We conclude that there exist $\bar{\boldsymbol{w}}^+ \succeq \mathbf{0}$ and $\bar{\boldsymbol{w}}^- \preceq \mathbf{0}$, such that

$$\boldsymbol{\theta} + \boldsymbol{y} = \bar{\boldsymbol{X}}^+\bar{\boldsymbol{w}}^+ + \bar{\boldsymbol{X}}^-\bar{\boldsymbol{w}}^-.$$

Therefore, for $\boldsymbol{w}^\star \in \mathbb{R}^n$, such that

$$\boldsymbol{w}^\star = \left\{ w_i \;\middle|\; \begin{cases} w_i = 0 & i \notin \mathcal{A} \\ w_i = \bar{\boldsymbol{w}}^+(k_i^+) & i \in \mathcal{A}^+ \,,\; i \in \mathcal{I} \\ w_i = \bar{\boldsymbol{w}}^-(k_i^-) & i \in \mathcal{A}^- \end{cases} \right\}, \tag{2.17}$$

with $k_i^\pm = \{k \,|\mathcal{A}^\pm{}_k = i\}$, we have the following statement,

$$w_i \geq 0 :\; i \in \mathcal{A}^+, \quad w_i \leq 0 :\; i \in \mathcal{A}^-, \quad w_i = 0 :\; i \notin \mathcal{A}, \qquad \sum_{i \in \mathcal{A}} \boldsymbol{x}_i w_i = \boldsymbol{\theta} + \boldsymbol{y},$$

holds true and the first part of our theorem is proved.

We prove that $\boldsymbol{w}^\star$ is the optimal solution of the LASSO problem given in (2.1), by checking the sub-gradient condition

$$\boldsymbol{c}(\boldsymbol{w}) = \boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right) \in -\lambda\partial\left\|\boldsymbol{w}\right\|_1, \tag{2.18}$$

with

$$\partial\left\|\boldsymbol{w}\right\|_1 = \left\{ \boldsymbol{g}(i) \;\middle|\; \begin{cases} \boldsymbol{g}(i) = 1 & \boldsymbol{w}(i) \geq 0 \\ \boldsymbol{g}(i) = -1 & \boldsymbol{w}(i) \leq 0 \,,\; i \in \mathcal{I} \\ |\boldsymbol{g}(i)| < 1 & \boldsymbol{w}(i) = 0 \end{cases} \right\},$$

at $\boldsymbol{w} = \boldsymbol{w}^\star$. From the first part of our theorem, there exists a $\boldsymbol{w}^\star$ such that $\boldsymbol{\theta}^\star = \boldsymbol{X}\boldsymbol{w}^\star - y$, with properties defined in (2.17), if $\boldsymbol{\theta}^\star$ is the optimal solution of $\mathcal{D}(\lambda)$ and

$$\boldsymbol{x}_i^T\boldsymbol{\theta}^\star = \lambda :\ i \in \mathcal{A}^-, \quad \boldsymbol{x}_i^T\boldsymbol{\theta}^\star = -\lambda :\ i \in \mathcal{A}^+, \quad \left|\boldsymbol{x}_i^T\boldsymbol{\theta}^\star\right| < \lambda :\ i \notin \mathcal{A} = \mathcal{A}^- \cup \mathcal{A}^+.$$

We conclude that the sub-gradient condition $\boldsymbol{c}(\boldsymbol{w}^\star) \in -\lambda\partial\boldsymbol{g}$ in (2.18) holds true and $\boldsymbol{w}^\star$ is the optimal solution of the LASSO problem.

Finally, we prove that $\phi'(\lambda)$ achieves $\phi(\lambda)$ by substituting $\boldsymbol{\theta}^\star = \boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}$ in $G(\boldsymbol{\theta})$. We have

$$\begin{aligned}
\phi'(\lambda) &= -\frac{1}{2}\left\|\boldsymbol{\theta}^\star\right\|_2^2 - \boldsymbol{\theta}^{\star T}\boldsymbol{y}, \\
&= \frac{1}{2}\left\|\boldsymbol{\theta}^\star\right\|_2^2 - \boldsymbol{\theta}^{\star T}\left(\boldsymbol{\theta} + \boldsymbol{y}\right), \\
&= \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}\right\|_2^2 - \boldsymbol{w}^{\star T}\boldsymbol{X}^T\boldsymbol{\theta}^\star, \\
&= \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}\right\|_2^2 - \sum_{i\in\mathcal{A}^+}\left(-\lambda\bar{\boldsymbol{w}}^+(i)\right) \\
&\quad - \sum_{i\in\mathcal{A}^-}\left(\lambda\bar{\boldsymbol{w}}^-(i)\right).
\end{aligned}$$

We recognize that

$$\left\|\boldsymbol{w}^\star\right\|_1 = \sum_{i\in\mathcal{A}^+}\bar{\boldsymbol{w}}^+(i) + \sum_{i\in\mathcal{A}^-}\left(-\bar{\boldsymbol{w}}^-(i)\right),$$

and $\phi'(\lambda)$ reduces to

$$\begin{aligned}
\phi'(\lambda) &= \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}\right\|_2^2 + \lambda\left\|\boldsymbol{w}^\star\right\|_1, \\
&= \phi(\lambda).
\end{aligned}$$

■

2.2.3 Optimality Conditions and Geometric Interpretation

The optimality conditions introduced by the strong duality theorem has geometric interpretations that can help in understanding our Safe Feature Elimination method presented in Chapter 3.

Let $\boldsymbol{c} = (c_1, ..., c_n) \in \mathbb{R}^n$ be the feature matrix and optimal dual-point correlation vector, i.e. $\boldsymbol{c}(\lambda) = \boldsymbol{X}^T\boldsymbol{\theta}^\star(\lambda)$ with $\boldsymbol{\theta}^\star(\lambda)$ the optimal dual point of (2.13) at λ. By the optimality conditions of Theorem 2.2.1, we have

$$\begin{aligned}
c_i(\lambda) = \lambda &\Rightarrow w_i \leq 0, \\
c_i(\lambda) = -\lambda &\Rightarrow w_i \geq 0, \\
|c_i(\lambda)| < \lambda &\Rightarrow w_i = 0,
\end{aligned}$$

for $i = 1, ..., n$. To illustrate these optimality conditions, we solve the LASSO with feature matrix $\boldsymbol{X}$ and response $\boldsymbol{y}$ obtained from the the diabetes dataset [43]. In Figure 2.3, we show a plot of the quantity $c_i(\lambda)$ and the corresponding LASSO solution $\boldsymbol{w}_i$ for two features out of the 10 features in the model .

In Figure 2.3(a), we notice that both features have values of $c(\lambda) \in]-\lambda, \lambda[$, where $|c(\lambda)| = \lambda$ is shown in the red dotted-line and the vertical dotted-lines represent the breakpoints of the regularization path. For lower values of the regularization parameter λ, $c_i(\lambda)$ associated with one (blue) feature of the two features takes the value λ and its corresponding weight w_i takes a non-positive value in Figure 2.3(b). Similarly, the $c_i(\lambda)$ for the other (green) feature takes the value $-\lambda$ and its corresponding weight w_i takes a nonnegative value.

Another geometric interpretation of the inequality $|c_i(\lambda)| < \lambda$ can be seen in the dual space $\boldsymbol{\theta}$. When the point $\boldsymbol{\theta}^\star$ is inside a slab $\mathcal{S} = \left\{\boldsymbol{\theta} \,\middle|\, \left|\boldsymbol{x}_k^T\theta\right| \leq \lambda\right\}$ defined by the feature $\boldsymbol{x}_k$, i.e. $\boldsymbol{\theta}^\star \in \mathcal{S}$, then strict inequality holds as shown in Figure 2.4.

Some algorithms, like the interior point method of [27], do not return exact zeros in the solution of the LASSO problem. The optimality conditions are used as a proxy to determine the zero entries of the primal solution by forming (or using) an approximate of the optimal dual point $\boldsymbol{\theta}^\star$ and then checking the inequality $|c_i(\lambda)| < \lambda$. When $\theta^\star$ is not a good estimate, the optimality conditions might result in setting some non-zero entries of $\boldsymbol{w}^\star$ to zero. In Appendix A, we provide a method for thresholding the solution based on controlling the perturbation of the objective function that is induced by thresholding.

The optimality conditions obtained from strong duality allow us to know the zero entries of the LASSO solution without actually solving the LASSO primal problem. In Figure 2.5, we recover all the zero entries for the diabetes dataset by only solving the dual problem. This fact is useful in deriving our Safe Feature Elimination method, which is essentially a cheap way for finding the zero elements of the LASSO solution at optimum without solving the LASSO problem.

2.3 LASSO Regularization Path

The LASSO solution for all parameters $\lambda \geq 0$ is refereed to as the regularization path and reads:

$$\boldsymbol{w}^\star(\lambda) = \frac{\lambda_k - \lambda}{\lambda_k - \lambda_{k+1}}\boldsymbol{w}^{(k+1)} + \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}}\boldsymbol{w}^{(k)}, \qquad \lambda^{(k+1)} \leq \lambda \leq \lambda^{(k)},\ k = 0, ..., k_{\max} - 1,$$

where $\boldsymbol{w}^{(k)}$ is the solution of the LASSO with $\lambda = \lambda^{(k)}$, $\lambda_0 = \left\|\boldsymbol{X}^T\boldsymbol{y}\right\|_\infty$ and $\lambda^{(k)},\ k = 0, ..., k_{\max}$ is an increasing sequence, i.e. λ: $0 = \lambda^{(k_{\max})} < \cdots < \lambda^{(0)}$. The solution for $\lambda > \lambda^{(0)}$ is $\boldsymbol{w}^\star = 0$. In this section, we derive this result and provide an algorithm for building the regularization path for a particular LASSO problem with feature matrix $\boldsymbol{X}$ and

response $\boldsymbol{y}$. We start by looking at the LASSO dual problem and then we deduce the primal solution as a function of λ.

2.3.1 The Dual Solution Path

We consider the dual problem $\mathcal{D}(\lambda)$ and construct the dual optimal solution $\boldsymbol{\theta}^\star$ for all $\lambda \geq 0$. We notice that $G(\theta)$ is strongly convex and admits $\boldsymbol{\theta}^\star = -\boldsymbol{y}$ as its global optimal solution when the point $-\boldsymbol{y}$ is feasible, i.e. $\lambda > \lambda^{(0)} := \left\|\boldsymbol{X}^T\boldsymbol{y}\right\|_\infty$. When $\lambda = \lambda^{(0)}$, one of the inequality constraints is active, we have $\left|\boldsymbol{x}_j^T\boldsymbol{\theta}\right| = \lambda$ for some index j, and the constraint remains active until some value $\lambda = \lambda^{(1)}$.

We provide a template problem in the following proposition that will help us derive the dual solution for all $\lambda > 0$.

Proposition 2.3.1 *Consider the optimization problem*

$$\begin{aligned}\mathcal{P}_\lambda(\tilde{\boldsymbol{y}}, \boldsymbol{l}, \boldsymbol{u}, \lambda_u) \ : \ \boldsymbol{\theta}^\star(\lambda) = \arg\min_{\boldsymbol{\theta}} \ \|\boldsymbol{\theta} + \tilde{\boldsymbol{y}}\|_2^2 \ : \\ 0 \leq \lambda \leq \lambda_u \\ \lambda l_i \leq \boldsymbol{x}_i^T\boldsymbol{\theta} \leq \lambda u_i, \qquad i = 1, ..., n\end{aligned}$$

with $\boldsymbol{l} = (l_1, ..., l_n) \in \mathbb{R}^n$, $u = (u_1, ..., u_n) \in \mathbb{R}^n$, $\boldsymbol{x}_k \in \mathbb{R}^m$ the k-th column of feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, response vector $\tilde{\boldsymbol{y}} \in \mathbb{R}^m$, and optimization variable $\boldsymbol{\theta} \in \mathbb{R}^m$. Assuming $\tilde{\boldsymbol{y}}$ is feasible for $\lambda \leq \lambda_u$ then $\mathcal{P}_\lambda$ admits the solution $\boldsymbol{\theta}^\star(\lambda) = -\tilde{\boldsymbol{y}}$ for $\lambda \in [\ \lambda_l, \lambda_u]$ with

$$\lambda_l = \min_\lambda \left\{\lambda \ : \ \lambda l_i \leq -\boldsymbol{x}_i^T\tilde{\boldsymbol{y}} \leq \lambda u_i, \ i = 1, ..., n\right\}.$$

Proof: Consider $\underline{\boldsymbol{\lambda}}(i) = -\boldsymbol{x}_i^T\tilde{\boldsymbol{y}}/l_i$ and $\bar{\boldsymbol{\lambda}}(i) = -\boldsymbol{x}_i^T\tilde{\boldsymbol{y}}/u_i$. We find indices j_1 and j_2, such that $j_1 = \{i \,|\underline{\boldsymbol{\lambda}}(i) = \sup_i \ \underline{\boldsymbol{\lambda}}(i)\}$ and $j_2 = \left\{i \,\middle|\bar{\boldsymbol{\lambda}}(i) = \sup_i \ \bar{\boldsymbol{\lambda}}(i)\right\}$. By construction, we have $\lambda_l = \max(\lambda_{j_1}, \lambda_{j_2})$. The function $\|\boldsymbol{\theta} + \tilde{\boldsymbol{y}}\|_2^2$ is strictly convex and admits $\boldsymbol{\theta}^\star = -\tilde{\boldsymbol{y}}$ as a global minimum when it is feasible, i.e. $\boldsymbol{\theta}^\star = -\tilde{\boldsymbol{y}}$ is the solution for all $\lambda \in [\lambda_l, \lambda_u]$. ■

A recursive method. We recognize that $\mathcal{D}(\lambda)$ has the same solution $\boldsymbol{\theta}^\star$ as $\mathcal{P}_\lambda(\boldsymbol{y}, \ \boldsymbol{l}, \ \boldsymbol{u}, \ \lambda_u)$ with $\boldsymbol{l} = -\boldsymbol{1}$, $\boldsymbol{u} = 1$ and $\lambda_u \to \infty$. Following Proposition 2.3.1 and defining $\lambda^{(0)} := \lambda_l$ with λ_l computed in the proposition, the dual solution for $\lambda \geq \lambda^{(0)}$ is $\boldsymbol{\theta}^\star = -\boldsymbol{y}$.

We represent the active constraint j at $\lambda = \lambda^{(0)}$ by $\boldsymbol{x}_j^T\boldsymbol{y} = -\alpha_j\lambda^{(0)}$ with $\alpha_j = u_j$ if $-\boldsymbol{x}_j^T\boldsymbol{y}$ attains its upper bound at $\lambda^{(0)}$ and $\alpha_j = l_j$ if the lower bound is attained.

We then investigate the solution $\boldsymbol{\theta}^\star(\lambda)$ for $\lambda \leq \lambda^{(0)}$ by using Proposition 2.3.1 with some new parameters $\tilde{\boldsymbol{y}}, \ \boldsymbol{l}, \ \boldsymbol{u}, \ \lambda_u$. We start by expressing the vectors $\boldsymbol{\theta}$ and $\boldsymbol{y}$ in terms of the

normal vector of the active constraint $\boldsymbol{x}_j$,

$$\begin{aligned}\boldsymbol{\theta} &= \left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{\theta}',\ \boldsymbol{x}_j^T\boldsymbol{\theta}'=0,\\ \boldsymbol{y} &= \left(\boldsymbol{x}_j^T\boldsymbol{y}\right)\frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{y}',\ \boldsymbol{x}_j^T\boldsymbol{y}'=0.\end{aligned}$$

We add those two constraints and obtain the following equivalent problem,

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\theta}'}\quad & \|\boldsymbol{\theta}+\boldsymbol{y}\|_2^2\ :\\ & \lambda l_i \le \boldsymbol{x}_i^T\boldsymbol{\theta}\le\lambda u_i,\qquad i=1,...,n,\\ & \boldsymbol{\theta}=\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{\theta}',\\ & \boldsymbol{y}=\left(\boldsymbol{x}_j^T\boldsymbol{y}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{y}',\\ & \boldsymbol{x}_j^T\boldsymbol{\theta}'=0,\ \boldsymbol{x}_j^T\boldsymbol{y}'=0.\end{aligned}$$

In the problem above, we removed the term $\frac{1}{2}\|\boldsymbol{y}\|_2^2$ from $\mathcal{D}(\lambda)$ and interchanged the maximization with a minimization for convenience. We substitute $\boldsymbol{x}_j^T\boldsymbol{y}=-\alpha_j\lambda^{(0)}$, and express the equality constraints as implicit constraints,

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\theta}'}\quad & \left\|\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{\theta}'+\left(-\alpha_j\lambda^{(0)}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{y}'\right\|_2^2\ :\\ & \lambda l_i-\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_i^T\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}\le\boldsymbol{x}_i^T\boldsymbol{\theta}'\le\lambda u_i-\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_i^T\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2},\quad i=1,...,n,\\ & \boldsymbol{\theta}=\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{\theta}',\\ & \boldsymbol{x}_j^T\boldsymbol{\theta}'=0,\ \boldsymbol{x}_j^T\boldsymbol{y}'=0.\end{aligned}$$

The objective function of the above optimization problem can be decomposed into

$$\|\boldsymbol{\theta}'+\boldsymbol{y}'\|_2^2+\frac{1}{\|\boldsymbol{x}_j\|_2^2}\left\|\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha_j\lambda^{(0)}\right\|_2^2,$$

where we have used $\boldsymbol{x}_j^T\boldsymbol{\theta}'=0$, and $\boldsymbol{x}_j^T\boldsymbol{y}'=0$. We obtain the optimization problem

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\theta}'}\quad & \|\boldsymbol{\theta}'+\boldsymbol{y}'\|_2^2+\tfrac{1}{\|\boldsymbol{x}_j\|_2^2}\left\|\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha_j\lambda^{(0)}\right\|_2^2\ :\\ & \lambda l_i-\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_i^T\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}\le\boldsymbol{x}_i^T\boldsymbol{\theta}'\le\lambda u_i-\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_i^T\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2},\quad i=1,...,n,\\ & \boldsymbol{\theta}=\left(\boldsymbol{x}_j^T\boldsymbol{\theta}\right)\tfrac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}+\boldsymbol{\theta}',\\ & \boldsymbol{x}_j^T\boldsymbol{\theta}'=0,\ \boldsymbol{x}_j^T\boldsymbol{y}'=0.\end{aligned}$$

We assume that, at optimum, the feature that got active at $\lambda = \lambda^{(0)}$, will remain active when we decrease λ below $\lambda^{(0)}$, i.e. for $\lambda < \lambda^{(0)}$ we have $\boldsymbol{x}_j^T \boldsymbol{\theta} = \alpha_j \lambda$. Thus, we add the constraint $\boldsymbol{x}_j^T \boldsymbol{\theta} = \alpha_j \lambda$ and the above optimization problem reduces to

$$\begin{aligned} \min_{\boldsymbol{\theta}'} \quad & \|\boldsymbol{\theta}' + \boldsymbol{y}'\|_2^2 + \frac{\alpha_j^2 \left(\lambda - \lambda^{(0)}\right)^2}{\|\boldsymbol{x}_j\|_2^2} \ : \\ & \lambda l_i' \leq \boldsymbol{x}_i^T \boldsymbol{\theta}' \leq \lambda u_i', \qquad i = 1, ..., n \end{aligned}$$

with $l_i' = \left(l_i - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2} \alpha_j\right)$, $u_i' = \left(u_i - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2} \alpha_j\right)$ for $i \neq j$ and $l_i' = u_i' = 0$ for $i = j$. The problem above has the same solution $\boldsymbol{\theta}^\star$ as $\mathcal{P}_\lambda(\boldsymbol{y}', \boldsymbol{l}', \boldsymbol{u}', \lambda^{(0)})$, for $\lambda \in [\lambda^{(1)}, \lambda^{(0)}]$ with $\lambda^{(1)} = \lambda_l$ and λ_l the regularization parameter provided by Proposition 2.3.1. We apply the method recursively to find the solution $\boldsymbol{\theta}^\star$ for all intervals $[\lambda^{(k+1)}, \lambda^{(k)}]$, $k = 0, ..., k_{\max} - 1$ with $k_{\max}$ the index of $\lambda^{(k)} = 0$.

We still need to check that $\boldsymbol{\theta}^\star(\lambda)$ computed for $\lambda \in [\lambda^{(1)}, \lambda^{(0)}]$ is optimal since we assumed that the feature that got active at $\lambda = \lambda^{(0)}$ remains active for some $\lambda < \lambda^{(0)}$. From the optimality condition of constrained convex maximization problems, a dual point $\boldsymbol{\theta}^\star$ is optimal for $\mathcal{D}(\lambda)$ if the inequality

$$-\nabla G(\boldsymbol{\theta}^\star) \left(\boldsymbol{\theta}^\star - \boldsymbol{\theta}\right) \geq 0$$

holds true for all $\boldsymbol{\theta}$ such that $\lambda l_i \leq \boldsymbol{x}_i^T \boldsymbol{\theta} \leq \lambda u_i$, $i = 1, ..., n$.

The optimal point we found is $\boldsymbol{\theta}^\star = (\lambda \alpha_j) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2} - \boldsymbol{y}'$ with $\boldsymbol{y}' = \boldsymbol{y} + \left(\lambda^{(0)} \alpha_j\right) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}$, and the optimality condition evaluates to

$$\left(\boldsymbol{\theta}^\star + \boldsymbol{y}\right) \left(\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right) \geq 0,$$

$$\left((\lambda \alpha_j) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2} - \left(\lambda^{(0)} \alpha_j\right) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}\right)^T \left(\boldsymbol{\theta} - (\lambda \alpha_j) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2} + \boldsymbol{y}'\right) \geq 0,$$

$$\frac{\left(\lambda - \lambda^{(0)}\right)}{\|\boldsymbol{x}_j\|_2^2} \alpha_j \boldsymbol{x}_j^T \left(\boldsymbol{\theta} - (\lambda \alpha_j) \frac{\boldsymbol{x}_j}{\|\boldsymbol{x}_j\|_2^2}\right) \geq 0,$$

$$\frac{\left(\lambda - \lambda^{(0)}\right)}{\|\boldsymbol{x}_j\|_2^2} \alpha_j \left(\boldsymbol{x}_j^T \boldsymbol{\theta} - \lambda \alpha_j\right) \geq 0.$$

Assuming that the upper bound is active, i.e. $\alpha_j = 1$, then $\frac{\left(\lambda - \lambda^{(0)}\right)}{\|\boldsymbol{x}_j\|_2^2} \alpha_j < 0$, and we have $\boldsymbol{x}_j^T \boldsymbol{\theta} \leq \lambda$. If the lower bound is active, i.e. $\alpha_j = -1$ and $\frac{\left(\lambda - \lambda^{(0)}\right)}{\|\boldsymbol{x}_j\|_2^2} \alpha_j > 0$, then we have $-\lambda \leq \boldsymbol{x}_j^T \boldsymbol{\theta}$. In both cases, the optimality condition is satisfied and $\boldsymbol{\theta}^\star$ is optimal for $\lambda \in [\lambda^{(1)}, \lambda^{(0)}]$.

Algorithm 1 summarizes the procedure for building the regularization path for the dual optimal point $\boldsymbol{\theta}^\star$.

2.3.2 The Primal Solution Path

We start with the dual solution from Algorithm 1,

$$\boldsymbol{\theta}^\star(\lambda) = \left(\lambda - \lambda^{(k)}\right) \frac{\alpha_{j(k)} \boldsymbol{x}_{j(k)}}{\left\|\boldsymbol{x}_{j(k)}\right\|_2^2} + \boldsymbol{\theta}^{(k)},\ \lambda \in [\lambda^{(k+1)}, \lambda^{(k)}], \tag{2.19}$$

with $\boldsymbol{\theta}^{(0)} = -\boldsymbol{y}$, $k = 0, ..., k_{\max} - 1$. We also recall the optimality condition of Theorem 2.2.1: there exist a $\boldsymbol{w}^\star$ such that

$$\boldsymbol{\theta}^\star = \boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y},$$

and $\boldsymbol{w}^\star$ is the optimal solution of the LASSO.

Using induction on (2.19), we can write

$$\boldsymbol{\theta}^{(k)} = \sum_{i=1}^{k} \left(\lambda^{(i)} - \lambda^{(i-1)}\right) \frac{\alpha_{j(i-1)} \boldsymbol{x}_{j(i-1)}}{\left\|\boldsymbol{x}_{j(i-1)}\right\|_2^2} - \boldsymbol{y}.$$

By inspection, we have

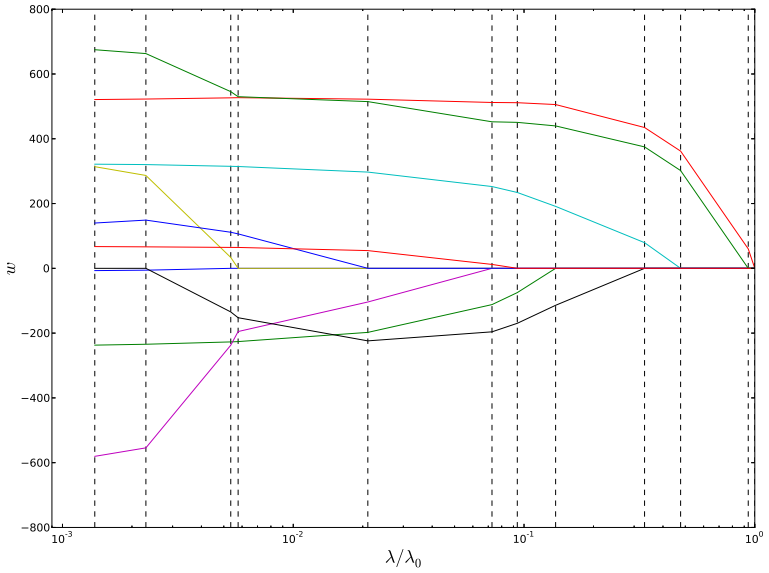
$$\boldsymbol{w}^{(k)} = \sum_{i=1}^{k} \left(\lambda^{(i)} - \lambda^{(i-1)}\right) \frac{\alpha_{j(i-1)} \boldsymbol{e}_{j(i-1)}}{\left\|\boldsymbol{x}_{j(i-1)}\right\|_2^2},$$

with $\boldsymbol{e}_{j(k)}$ the unit vector with all entries zeros except at $j(k)$. Finally, we express $\boldsymbol{w}^\star(\lambda)$ for $\lambda \in \left[\lambda^{(k+1)}, \lambda^{(k)}\right]$ in terms of $\boldsymbol{w}^{(k)}$ and $\boldsymbol{w}^{(k+1)}$,

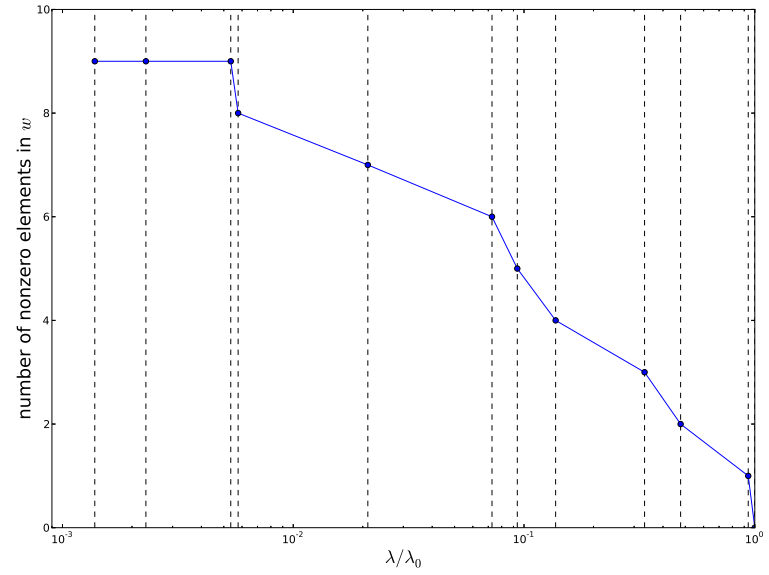
$$\boldsymbol{w}^\star(\lambda) = \frac{\lambda_k - \lambda}{\lambda_k - \lambda_{k+1}} \boldsymbol{w}^{(k+1)} + \frac{\lambda - \lambda_{k+1}}{\lambda_k - \lambda_{k+1}} \boldsymbol{w}^{(k)}, \qquad \lambda^{(k+1)} \leq \lambda \leq \lambda^{(k)},\ k = 0, ..., k_{\max} - 1.$$

2.4 Conclusion

In this chapter, we introduced the LASSO problem and explained the background needed to derive our Safe Feature Elimination method. We have derived the dual problem of the LASSO and stated the optimality conditions. The conclusions we can make on the values of the primal solution of the LASSO from the solution of its dual problem, constitute the basics for deriving SAFE as will be seen in the next chapter.



(a)



(b)

Figure 2.1: Sparsity of the LASSO solution. (a) The regularization path for the LASSO problem. The LASSO solution $\boldsymbol{w}^\star$ is solved for each value of λ, on the diabetes dataset [43]. Each color represents one element of the vector $\boldsymbol{w}^\star$. The diabetes dataset has 10 features, at $\lambda/\lambda_0 = 1$, all elements in the solution are zero. For lower values of the regularization parameter, i.e. $\lambda/\lambda_0 < 1$, breakpoints happen when a zero entry of $\boldsymbol{w}^\star$ becomes a non-zero, or vice-versa. These breakpoints are represented by the vertical dashed-lines. (b) The number of non-zeros in the solution for different values of λ.

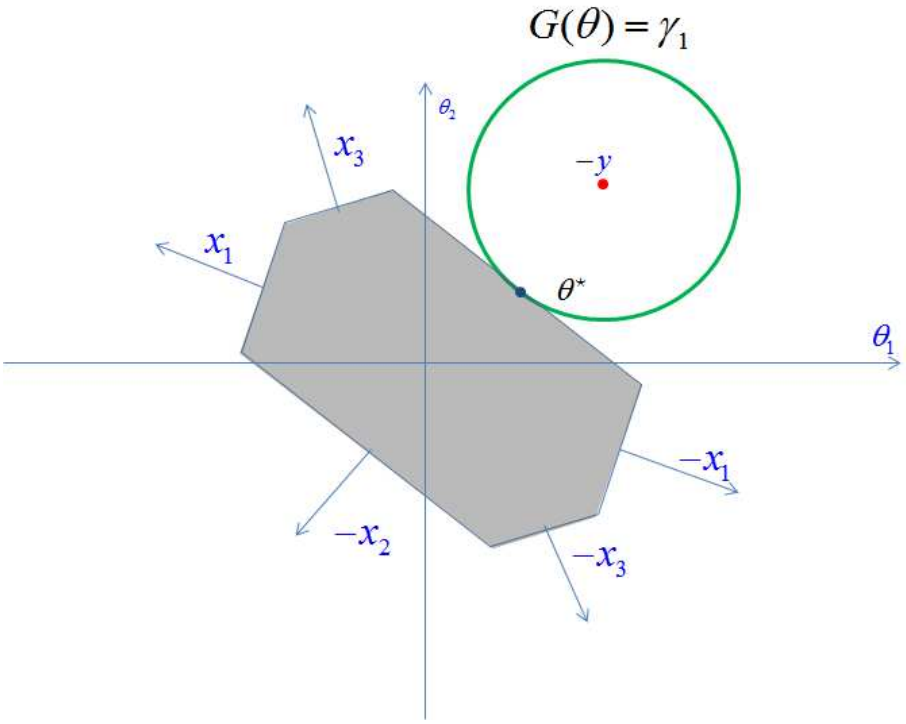
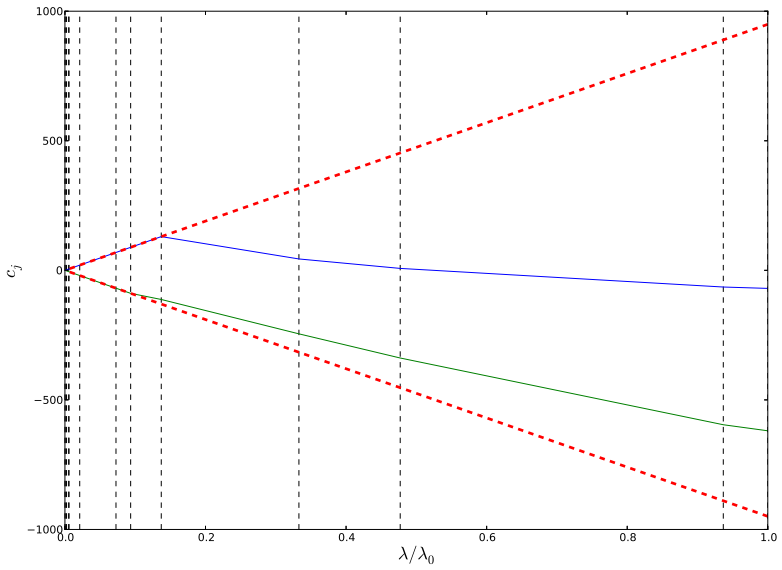
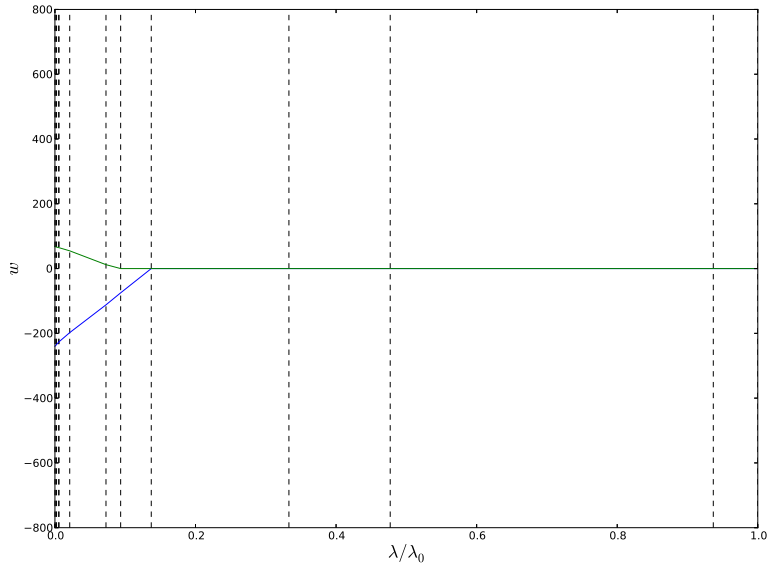


Figure 2.2: Geometry of the dual problem $\mathcal{D}(\lambda)$. The Grey shaded polytope shows the feasibility set of $\mathcal{D}(\lambda)$. The feasibility set is the intersection of n slabs in the dual space corresponding to the n features $\boldsymbol{x}_k$, $k = 1, ..., n$, i.e. the intersection of $\left|\boldsymbol{x}_k^T \boldsymbol{\theta}\right| \leq \lambda$, $k = 1, ..., n$. The level set $\{\boldsymbol{\theta} \,|G(\theta) = \gamma_1,\ \gamma_1 = G(\boldsymbol{\theta}^\star)\}$, corresponds to the optimal value of the dual function and is tangent to the feasibility set at the dual optimal point $\boldsymbol{\theta}^\star$.



(a)



(b)

Figure 2.3: Illustration of the LASSO optimality condition. The feature matrix $\boldsymbol{X}$ and response $\boldsymbol{y}$ used to generate these figures are obtained from the diabetes dataset [43]. (a) A plot of $c_i(\lambda) = \boldsymbol{x}_i^T \boldsymbol{\theta}^\star(\lambda)$ for two features. The red dashed-lines represent the curve $|c(\lambda)| = \lambda$ and the vertical dotted-lines represent the breakpoints of the regularization path. For lower values of the regularization parameter λ, $c_i(\lambda)$ associated with the blue feature takes the value λ and its corresponding weight w_i in (b) takes a non-positive value. Similarly, the $c_i(\lambda)$ for the green feature takes the value $-\lambda$ in (a) and its corresponding weight w_i takes a nonnegative value in (b).

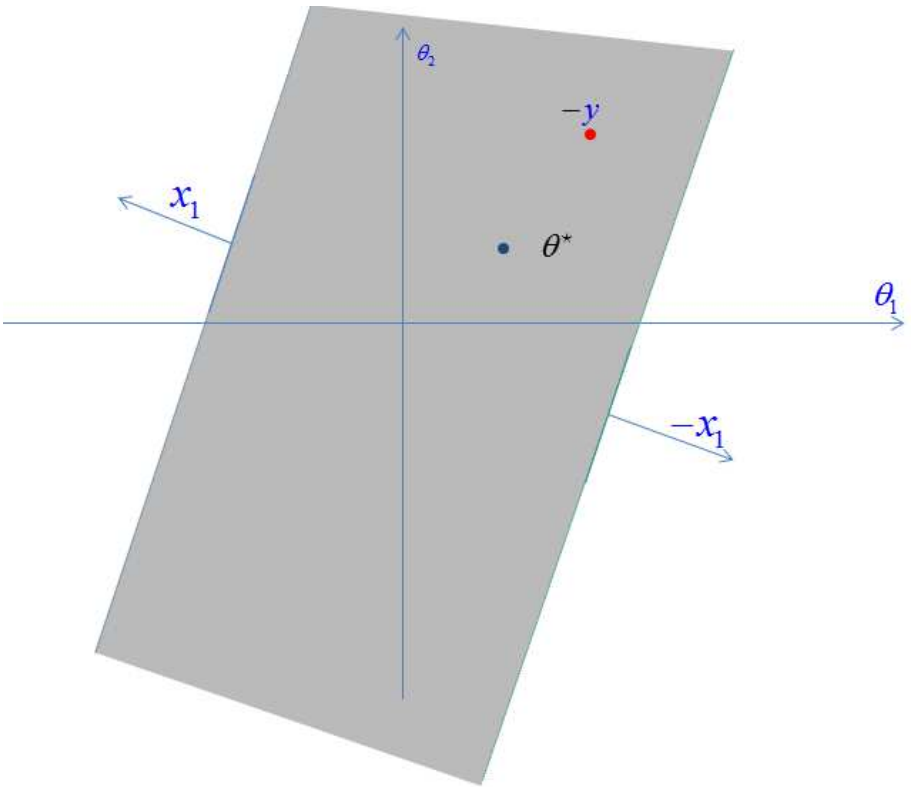
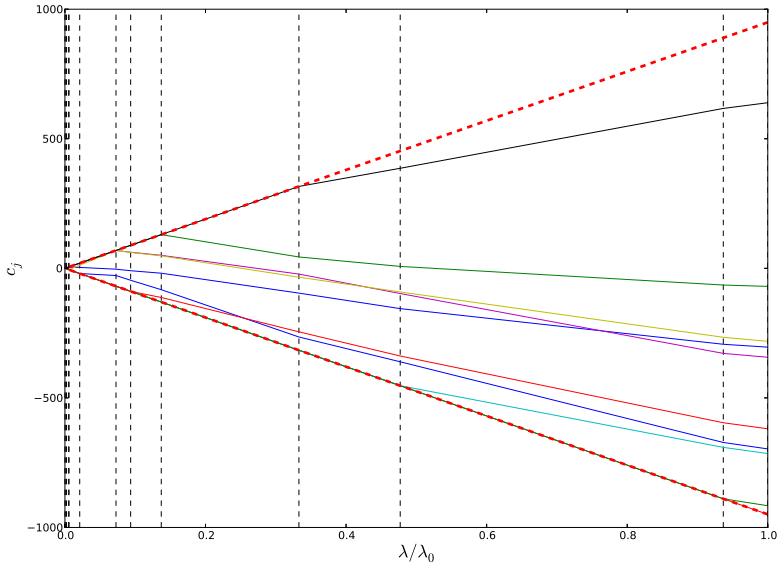
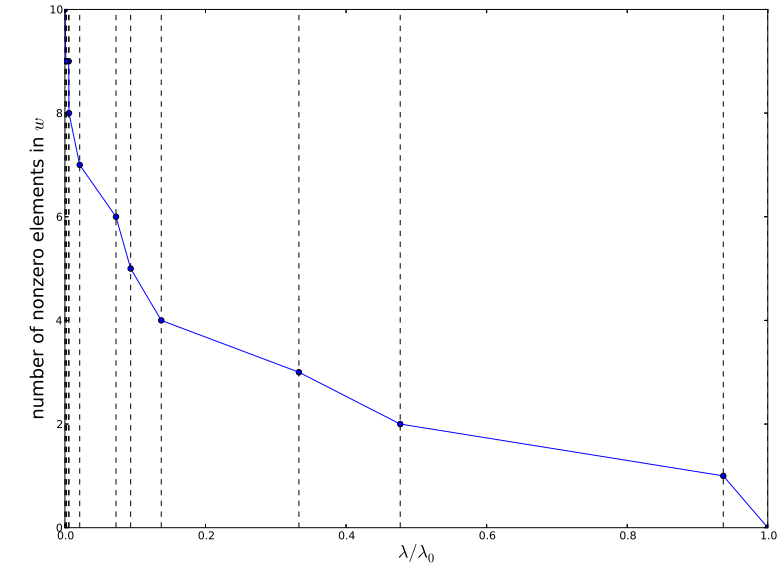


Figure 2.4: Geometry of the inequality $|c_i(\lambda)| < \lambda$, with $c_i = \boldsymbol{x}_i^T \boldsymbol{\theta}^\star$ in dual variable space. The Grey shaded region is the slab corresponding to feature $\boldsymbol{x}_i$, i.e. $\left\{\boldsymbol{\theta} \mid \left|\boldsymbol{\theta}^T \boldsymbol{x}_i\right| \leq \lambda\right\}$. The test $\left|\boldsymbol{\theta}^{\star T} \boldsymbol{x}_i\right| < \lambda$ is a strict inequality when the point $\boldsymbol{\theta}^\star$ is in the interior of the slab defined by the feature $\boldsymbol{x}_k$ (in this case $k = 1$). When the dual optimal point is inside a slab defined by feature $\boldsymbol{x}_k$, the strong duality optimality condition implies that the k-th entry of the primal optimal solution $\boldsymbol{w}^\star$ is zero or $\boldsymbol{w}^\star(k) = 0$.



(a)



(b)

Figure 2.5: Recovering the non-zero elements of a LASSO problem using the Dual problem and optimality conditions. The feature matrix $\boldsymbol{X}$ and response $\boldsymbol{y}$ used to generate these figures are obtained from the diabetes dataset [43]. (a) A plot of $c_i(\lambda) = \boldsymbol{x}_i^T \boldsymbol{\theta}^\star(\lambda)$ for all features in the model. The red dashed-lines represent the curve $|c(\lambda)| = \lambda$ and the vertical dotted-lines represent the breakpoints of the regularization path. (b) The number of non-zero elements in the solution of the LASSO is computed by checking the number of features that satisfy the inequality $|c_i(\lambda)| < \lambda$. Graphically, this corresponds to the colored lines that are encapsulated with the red dashed-line envelop in (a).

Algorithm 1 Recursive computation of the dual solution $\boldsymbol{\theta}^\star(\lambda)$ over the intervals $[\lambda^{(k+1)}, \lambda^{(k)}]$ defined by the regularization path.

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$.
initialize

1. Set $\boldsymbol{l}^{(0)} = -\mathbf{1} \in \mathbb{R}^m$, $\boldsymbol{u}^{(0)} = \mathbf{1} \in \mathbb{R}^m$, $\boldsymbol{y}^{(0)} = \boldsymbol{y}$.
2. Solve for $\mathcal{P}_\lambda(\boldsymbol{y}^{(0)}, \boldsymbol{l}^{(0)}, \boldsymbol{u}^{(0)}, \infty)$ in Proposition 2.3.1. Obtain $\lambda^{(0)} = \lambda_l$ and index j.
3. Set $j(0) = j$, $\alpha_{j(0)} = -\frac{\left(\boldsymbol{x}_{j(0)}^T \boldsymbol{y}^{(0)}\right)}{\lambda^{(0)}}$, $\boldsymbol{\theta}^{(0)} := \boldsymbol{\theta}^\star(\lambda^{(0)}) = -\boldsymbol{y}$ and $k = 0$.

repeat

1. Define $\boldsymbol{l}^{(k+1)}$ and $\boldsymbol{u}^{(k+1)}$ such that

$$l_i^{(k+1)} = \left(l_i^{(k)} - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_{j(k)}}{\left\|\boldsymbol{x}_{j(k)}\right\|_2^2}\alpha_{j(k)}\right), u_i^{(k+1)} = \left(u_i^{(k)} - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_{j(k)}}{\left\|\boldsymbol{x}_{j(k)}\right\|_2^2}\alpha_{j(k)}\right), \text{ for } i \neq j,$$

   and $l_i^{(k+1)} = u_i^{(k+1)} = 0$, for $i = j$.

2. Define $\boldsymbol{y}^{(k+1)}$ such that

$$\boldsymbol{y}^{(k+1)} = \boldsymbol{y}^{(k)} + \left(\lambda^{(k)}\alpha_{j(k)}\right)\frac{\boldsymbol{x}_{j(k)}}{\left\|\boldsymbol{x}_{j(k)}\right\|_2^2}.$$

3. Solve for $\mathcal{P}_\lambda(\boldsymbol{y}^{(k+1)}, \boldsymbol{l}^{(k+1)}, \boldsymbol{u}^{(k+1)}, \lambda^{(k)})$ in Proposition 2.3.1. Set index $j(k+1) = j$, $\alpha_{j(k+1)}$, and $\lambda^{(k+1)}$.
4. Set the optimal dual point to $\boldsymbol{\theta}^\star(\lambda) = \left(\lambda - \lambda^{(k)}\right)\frac{\alpha_{j(k)}\boldsymbol{x}_{j(k)}}{\left\|\boldsymbol{x}_{j(k)}\right\|_2^2} + \boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}^{(k+1)} := \boldsymbol{\theta}^\star(\lambda^{(k+1)})$.
5. Increment k. $k = k + 1$.

until $\lambda^{(k)} = 0$
Set $k_{\max} = k$
terminate

Chapter 3

Safe Feature Elimination for the LASSO

3.1 Introduction

Safe Feature Elimination (SAFE) is a method that can cheaply identify some of the zero entries in a LASSO solution, **a-priori** to solving the LASSO problem. Recovering the sparsity pattern of a LASSO solution allows us to reduce memory requirements and computational costs when solving the problem. We present the following proposition.

Proposition 3.1.1 *Consider the LASSO problem*

$$\mathcal{P}(\lambda) \ : \ \phi(\lambda) := \min_{\boldsymbol{w}} \frac{1}{2} \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + \lambda \left\| \boldsymbol{w} \right\|_1 ,$$

with $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ *the feature matrix,* $\boldsymbol{y} \in \mathbb{R}^m$ *the response vector,* $\lambda > 0$ *the regularization parameter,* $\boldsymbol{w} \in \mathbb{R}^n$ *the optimization variable, and* $\boldsymbol{w}^\star$ *the optimal solution.*

Let $\mathcal{E}$ *be a set of indicies, with* $|\mathcal{E}| = e$ *such that* $\boldsymbol{w}^\star_{\mathcal{E}} = \boldsymbol{0}_e \in \mathbb{R}^e$*. Without loss of generality, assume* $\mathcal{E} = \{1, .., e\}$ *and* $\boldsymbol{X} = \left(\boldsymbol{X}_{\mathcal{E}}, \bar{\boldsymbol{X}}\right)$*, then* $\boldsymbol{w}^\star = (\boldsymbol{0}_e, \bar{\boldsymbol{w}}^\star)$*, where* $\bar{\boldsymbol{w}}^\star$ *is the solution of*

$$\min_{\bar{\boldsymbol{w}}} \frac{1}{2} \left\| \bar{\boldsymbol{X}}\bar{\boldsymbol{w}} - \boldsymbol{y} \right\|_2^2 + \lambda \left\| \bar{\boldsymbol{w}} \right\|_1 .$$

■

Thus, we can eliminate the features corresponding to any identified zero entries of $\boldsymbol{w}^\star$, and we can construct a LASSO solution of the original problem using a reduced feature matrix. The reduction in the feature-matrix size, allows LASSO algorithms presented in [4, 13, 27, 42, 11, 21, 20] and references therein, to possibly solve the LASSO with less memory requirements and fewer computational cost.

In Section 2.2.3, we recovered the sparsity of the LASSO solution without solving the LASSO problem. Solving the LASSO dual problem and using its optimality conditions to eliminate features has the drawback of expensive computations. In fact, solving the dual problem, is as expensive as solving the primal problem. SAFE is a method that provides a trade-off between the number of features eliminated and the amount of computations performed. Generally, SAFE is conservative in eliminating features but computationally very cheap, it has the cost of few (one or two) vector-matrix multiplications, yet it eliminates enough features especially at large values of the regularization parameter.

We describe the basic idea of the Safe Feature Elimination method and derive a theorem for eliminating features in section 3.2. We then derive a more aggressive method for eliminating features in section 3.3. In section 3.4, we describe how to use SAFE for reducing memory limit problems and reducing running time when solving the LASSO. Finally, in section 3.5, we explore the benefits of SAFE by running numerical experiments with data derived from text classification problems, as well as randomly generated data.

3.2 The SAFE method for the LASSO

Recall the dual problem of the LASSO is

$$\mathcal{D}(\lambda) \ : \ \phi'(\lambda) := \max_{\boldsymbol{\theta}} \ G(\boldsymbol{\theta}) \ : \ \left|\boldsymbol{x}_i^T\boldsymbol{\theta}\right| \leq \lambda, \ i = 1, ..., n,$$

with $G(\boldsymbol{\theta}) = \frac{1}{2}\left\|\boldsymbol{y}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\theta} + \boldsymbol{y}\right\|_2^2$, and by defining $\boldsymbol{c}(\lambda) = \boldsymbol{X}^T\boldsymbol{\theta}^\star(\lambda)$, the optimality condition is

$$\lambda > |c_k(\lambda)| \implies w_k^\star = 0, \tag{3.1}$$

with c_k the k-th entry of $\boldsymbol{c}$. In this section, we describe the basic idea behind SAFE and derive a SAFE-LASSO theorem for eliminating features.

3.2.1 Basic idea

Since computing $\boldsymbol{\theta}^\star(\lambda)$, and thus $\boldsymbol{c}(\lambda)$, is not an option, the basic idea is to use a sufficient condition for (3.1) instead. Consider the following sufficient condition: If $c_i(\lambda) \in \left[c_k^l(\lambda), c_k^u(\lambda)\right]$ and $\lambda > c$ for all $c \in \left[c_k^l(\lambda), c_k^u(\lambda)\right]$, then $\lambda > |c_k(\lambda)|$ and $w_k^\star = 0$. When using such sufficient condition, the number of features eliminated is conservative, depending on the interval, $\left[c_k^l(\lambda), c_k^u(\lambda)\right]$, used. Figure 3.1 and Figure 3.3 illustrate the sufficient condition using two methods for computing the bounds $\left[c_k^l(\lambda), c_k^u(\lambda)\right]$.

To derive such bounds for $c_i(\lambda)$, we start with the basic optimality condition

$$\lambda > |c_k(\lambda)| := \max(-\boldsymbol{x}_k^T\boldsymbol{\theta}^\star, \boldsymbol{x}_k^T\boldsymbol{\theta}^\star)$$

An equivalent formulation of the above condition is

$$\lambda > \max(P(\boldsymbol{x}_k), P(-\boldsymbol{x}_k)),$$

where $P(\boldsymbol{x})$ is the optimal value of the convex optimization problem

$$P(\boldsymbol{x}) := \max_{\theta} \boldsymbol{x}^T\boldsymbol{\theta} \ : \ \boldsymbol{\theta} = \boldsymbol{\theta}^\star. \tag{3.2}$$

We then relax the constraint $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ and replace it with $\boldsymbol{\theta} \in \Theta_1$ where Θ_1 is a set that contains $\boldsymbol{\theta}^\star$, i.e. $\boldsymbol{\theta}^\star \in \Theta_1$. We have

$$P'(\boldsymbol{x}, \Theta) := \max_{\boldsymbol{\theta}} \boldsymbol{x}^T\boldsymbol{\theta} \ : \ \boldsymbol{\theta} \in \Theta,$$

and the value of $P'(\boldsymbol{x}, \Theta)$ is always at least equal to $P(\boldsymbol{x})$, i.e. $P'(\boldsymbol{x}, \Theta) \geq P(\boldsymbol{x})$. Thus, if $\lambda \geq P'(\boldsymbol{x}, \Theta)$, then $\lambda \geq P(\boldsymbol{x})$ and we conclude the sufficient condition: if

$$\lambda > \max(P'(\boldsymbol{x}_k, \Theta_1), P'(-\boldsymbol{x}_k, \Theta_1)), \tag{3.3}$$

then $\lambda > \max(P(\boldsymbol{x}_k), P(-\boldsymbol{x}_k))$ and $w_k^\star = 0$. We call $P'(\boldsymbol{x}, \Theta)$ the SAFE test problem, and it gives the lower and upper bounds on $c_k(\lambda)$, $c_k^l(\lambda) = -P'(-\boldsymbol{x}_k, \Theta)$ and $c_k^u(\lambda) = P'(\boldsymbol{x}_k, \Theta)$, respectively. Note that if the inequality in (3.3) holds true, then this is equivalent to saying that $\lambda > c$ for all $c \in \left[c_k^l(\lambda), c_k^u(\lambda)\right]$. Figure 3.1 shows a geometric interpretation of the inequality $\lambda > \max(P(\boldsymbol{x}_k), P(-\boldsymbol{x}_k))$ for one of the features in the diabetes dataset [43].

3.2.2 Obtaining Θ_1 by dual scaling

The point $\boldsymbol{\theta}^\star = -\boldsymbol{y}$ is optimal for the LASSO at $\lambda = \lambda_0 := \left\|\boldsymbol{X}^T\boldsymbol{y}\right\|_\infty$ (see Section 2.3). We construct a feasible point $\boldsymbol{\theta}_s$ for $\mathcal{D}(\lambda)$ by dual scaling, $\boldsymbol{\theta}_s = -\boldsymbol{y}\frac{\lambda}{\lambda_0}$ and propose the following set, Θ_1, that contains $\boldsymbol{\theta}^\star(\lambda)$.

Proposition 3.2.1 *Consider the LASSO dual problem*

$$\mathcal{D}(\lambda) \ : \ \phi'(\lambda) := \max_{\boldsymbol{\theta}} \ G(\boldsymbol{\theta}) \ : \ \left|\boldsymbol{x}_i^T\boldsymbol{\theta}\right| \leq \lambda, \ i = 1, ..., n,$$

with $\boldsymbol{\theta}^\star(\lambda)$ the optimal solution at λ. Then the set

$$\Theta_1 = \left\{\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta} = -\boldsymbol{y} + \|\boldsymbol{y}\|_2 \left(1 - \frac{\lambda}{\lambda_0}\right)\boldsymbol{v}, \ \|\boldsymbol{v}\|_2 \leq 1\right\}.$$

contains $\boldsymbol{\theta}^\star(\lambda)$.

Proof: We start by the definition of optimality, $\boldsymbol{\theta}^\star(\lambda)$ is optimal for $\mathcal{D}(\lambda)$ if $G(\boldsymbol{\theta}^\star(\lambda)) \geq G(\boldsymbol{\theta})$ for all feasible $\boldsymbol{\theta}$ at λ. Since $\boldsymbol{\theta}_s = -\boldsymbol{y}\frac{\lambda}{\lambda_0}$ is feasible, we have

$$\frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta}^\star + \boldsymbol{y}\|_2^2 \geq \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta}_s + \boldsymbol{y}\|_2^2,$$

$$-\frac{1}{2}\left\|\boldsymbol{\theta}^{\star}+\boldsymbol{y}\right\|_2^2 \geq -\frac{1}{2}\left\|\boldsymbol{y}\left(1-\frac{\lambda}{\lambda_0}\right)\right\|_2^2,$$

$$\left\|\boldsymbol{\theta}^{\star}+\boldsymbol{y}\right\|_2^2 \leq \left\|\boldsymbol{y}\right\|_2^2\left(1-\frac{\lambda}{\lambda_0}\right)^2.$$

Thus $\boldsymbol{\theta}^{\star}$ belongs to the set

$$\Theta_1 = \left\{\boldsymbol{\theta}\,\middle|\,\left\|\boldsymbol{\theta}+\boldsymbol{y}\right\|_2^2 \leq \left\|\boldsymbol{y}\right\|_2^2\left(1-\frac{\lambda}{\lambda_0}\right)^2\right\}.$$

The inequality

$$\left\|\boldsymbol{\theta}+\boldsymbol{y}\right\|_2^2 \leq \left\|\boldsymbol{y}\right\|_2^2\left(1-\frac{\lambda}{\lambda_0}\right)^2,$$

is equivalent to

$$\boldsymbol{\theta} = -\boldsymbol{y}+\left\|\boldsymbol{y}\right\|_2\left(1-\frac{\lambda}{\lambda_0}\right)\boldsymbol{v},\ \left\|\boldsymbol{v}\right\|_2 \leq 1.$$

Thus, the set Θ_1 takes the form

$$\Theta_1 = \left\{\boldsymbol{\theta}\,\middle|\,-\boldsymbol{y}+\left\|\boldsymbol{y}\right\|_2\left(1-\frac{\lambda}{\lambda_0}\right)\boldsymbol{v},\ \left\|\boldsymbol{v}\right\|_2 \leq 1\right\}.$$

■

3.2.3 Solving the SAFE test problem

Using the proposed set Θ_1, the safe test problem reads

$$\begin{aligned} P'(\boldsymbol{x}) := \quad & \max_{\boldsymbol{\theta},\boldsymbol{v}}\ \boldsymbol{x}^T\boldsymbol{\theta}: \\ & \boldsymbol{\theta} = -\boldsymbol{y}+\left\|\boldsymbol{y}\right\|_2\left(1-\tfrac{\lambda}{\lambda_0}\right)\boldsymbol{v}, \\ & \left\|\boldsymbol{v}\right\|_2 \leq 1. \end{aligned}$$

We eliminate $\boldsymbol{\theta}$ from the constraints of the problem and obtain

$$\begin{aligned} P'(\boldsymbol{x}) := \quad & \max_{\boldsymbol{v}}\ -\boldsymbol{y}^T\boldsymbol{x}++\left\|\boldsymbol{y}\right\|_2\left(1-\tfrac{\lambda}{\lambda_0}\right)\boldsymbol{x}^T\boldsymbol{v}: \\ & \left\|\boldsymbol{v}\right\|_2 \leq 1. \end{aligned}$$

The problem admits an optimal solution $\boldsymbol{v}^{\star} = \frac{\boldsymbol{x}}{\left\|\boldsymbol{x}\right\|_2}$ and

$$P'(\boldsymbol{x}) = -\boldsymbol{y}^T\boldsymbol{x}+\left\|\boldsymbol{x}\right\|_2\left\|\boldsymbol{y}\right\|_2\left(1-\frac{\lambda}{\lambda_0}\right). \tag{3.4}$$

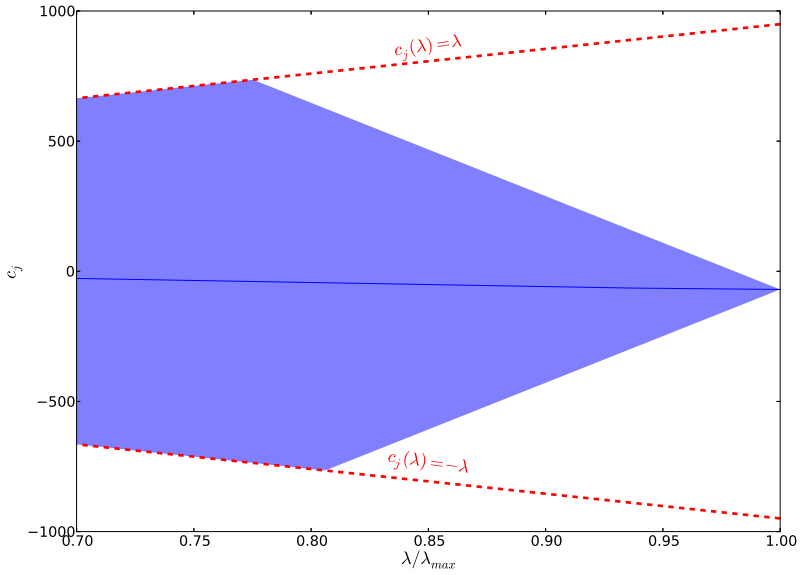


Figure 3.1: SAFE test problem bounds on $c_j(\lambda) = \boldsymbol{x}_j^T \boldsymbol{\theta}^\star(\lambda)$. A plot of $c_j(\lambda) = \boldsymbol{x}_k^T \boldsymbol{\theta}^\star(\lambda)$ for one of the features in the diabetes dataset [43] is shown in blue. The shaded region is $\{c \,|c \in \left[c_k^l(\lambda), c_k^u(\lambda)\right]\}$ with $c_k^l(\lambda) = -P'(-\boldsymbol{x}_k)$ and $c_k^u(\lambda) = P'(\boldsymbol{x}_k)$. The red dashed-lines represent the curve $|c(\lambda)| = \lambda$. The test $\lambda > \max(P'(\boldsymbol{x}_k), P'(-\boldsymbol{x}_k))$ is true when, graphically, the shaded region is inside the dashed-line red envelope. Since, by construction, $c_j(\lambda)$ is inside the shaded region, then $|c_j(\lambda)| < \lambda$ and $w_j = 0$.

3.2.4 Basic SAFE LASSO theorem

To eliminate feature k from the feature matrix $\boldsymbol{X}$, we need the inequality

$$\lambda > \max(P'(\boldsymbol{x}_k, \Theta_1), P'(-\boldsymbol{x}_k, \Theta_1)), \tag{3.5}$$

to hold true. The inequality above, using (3.4), reduces to

$$\lambda > \left|\boldsymbol{y}^T \boldsymbol{x}_k\right| + \|\boldsymbol{x}_k\|_2 \|\boldsymbol{y}\|_2 \left(1 - \frac{\lambda}{\lambda_0}\right).$$

The sufficient condition can be reduced further to read

$$\lambda > \rho_k \lambda_0,$$

with

$$\rho_k = \frac{\left|\boldsymbol{y}^T \boldsymbol{x}_k\right| + \|\boldsymbol{x}_k\|_2 \|\boldsymbol{y}\|_2}{\lambda_0 + \|\boldsymbol{x}_k\|_2 \|\boldsymbol{y}\|_2},$$

and $\lambda_0 = \left\|\boldsymbol{X}^T \boldsymbol{y}\right\|_\infty$. We summarize the result in the following theorem:

Theorem 3.2.1 (Basic SAFE-LASSO) *For the LASSO problem $\mathcal{P}(\lambda)$, and denoting by $\boldsymbol{x}_k$ the k-th feature of the feature matrix $\boldsymbol{X}$, the condition*

$$\lambda > \rho_k \lambda_0, \text{ with } \rho_k = \frac{\left|\boldsymbol{y}^T \boldsymbol{x}\right| + \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}{\lambda_0 + \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}, \ \lambda_0 = \max_{1 \leq j \leq n} \left|\boldsymbol{y}^T \boldsymbol{x}_j\right|,$$

allows to safely remove the k-th feature from feature matrix $\boldsymbol{X}$. ■

The computational complexity of running this test through all the features is $O(mn)$, with a better count if the data is sparse. The main computational burden in the test is actually independent of λ, and can be done once and for all: it suffices to rank features according to the values of ρ_k, $k = 1, ..., n$. Note that this test accurately predicts the sparsity of $\boldsymbol{w}^\star$ at $\lambda = \lambda_0$ for which all the features can be safely removed, that is, $\boldsymbol{w}^\star = \boldsymbol{0}$ is optimum for $\mathcal{P}(\lambda_0)$.

In the case of scaled data sets, for which $\|\boldsymbol{y}\|_2 = 1$ and $\|\boldsymbol{x}_k\|_2 = 1$ for every k, ρ_k has a convenient geometrical interpretation:

$$\rho_k = \frac{1 + |\cos \alpha_k|}{1 + \max\limits_{1 \leq j \leq n} |\cos \alpha_j|},$$

where α_k is the angle between the k-th feature $\boldsymbol{x}_k$ and the response vector $\boldsymbol{y}$. Our test then consists in eliminating features based on how closely they are aligned with the response, *relative* to the most closely aligned feature. For scaled data sets, our test is very similar to standard correlation-based feature selection [15]; in fact, for scaled data sets, the ranking of features it produces is then exactly the same. The big difference here is that our test is not heuristic, as it only eliminates features that are *guaranteed* to be absent when solving the full-fledged sparse supervised learning problem.

3.3 SAFE with tighter bounds on $\boldsymbol{\theta}^\star$

In this section, we assume that we are interested in solving the LASSO at some λ and we have knowledge of a LASSO solution $\boldsymbol{w}_0^\star$ at a regularization parameter λ_0, with $\lambda \leq \lambda_0$. This is a reasonable assumption as many algorithms, like coordinate descent, interior point method, LARS like algorithms (see Section 2.3), provide such points.

In this case, we can construct a tighter bound on $\boldsymbol{\theta}^\star(\lambda)$, i.e. construct a smaller set Θ such that $\boldsymbol{\theta}^\star \in \Theta$, than what we presented in Section 3.2.2.

3.3.1 Constructing Θ

We express Θ as the intersection of two sets Θ_1 and Θ_2, where each set contains $\boldsymbol{\theta}^\star$, but corresponds to different optimality conditions.

We construct Θ_1 similar to Section 3.2.2, using the optimality condition of $\mathcal{D}(\lambda)$: $\boldsymbol{\theta}^\star$ is a dual optimal point if $G(\boldsymbol{\theta}^\star) \geq G(\boldsymbol{\theta})$ for all dual feasible points $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_s$ be a dual feasible point to $\mathcal{D}(\lambda)$, and $\gamma := G(\boldsymbol{\theta}_s)$. Obviously $G(\boldsymbol{\theta}^\star) \geq \gamma$ and the set $\Theta_1 := \{\boldsymbol{\theta} \mid G(\boldsymbol{\theta}) \geq \gamma\}$ contains $\boldsymbol{\theta}^\star$, i.e. $\boldsymbol{\theta}^\star \in \Theta_1$.

One way to obtain a lower bound γ is by dual scaling. We set $\boldsymbol{\theta}_s$ to be a scaled feasible dual point in terms of $\boldsymbol{\theta}_0^\star$, $\boldsymbol{\theta}_s := s\boldsymbol{\theta}_0^\star$ with $s \in \mathbb{R}$ constrained so that $\boldsymbol{\theta}_s$ is a dual feasible point for $\mathcal{D}(\lambda)$, that is, $\|\boldsymbol{X}^T\boldsymbol{\theta}_s\|_\infty \leq \lambda$ or $|s| \leq \lambda/\lambda_0$. We then set γ according to the convex optimization problem:

$$\gamma = \max_s \left\{ G(s\boldsymbol{\theta}_0^\star) \ : \ |s| \leq \frac{\lambda}{\lambda_0} \right\} = \max_s \left\{ \beta_0 s - \frac{1}{2}s^2\alpha_0 \ : \ |s| \leq \frac{\lambda}{\lambda_0} \right\},$$

with $\alpha_0 := \boldsymbol{\theta}_0^{\star T}\boldsymbol{\theta}_0^\star > 0$, $\beta_0 := |\boldsymbol{y}^T\boldsymbol{\theta}_0^\star|$. We obtain

$$\gamma = \frac{\beta_0^2}{2\alpha_0}\left(1 - \left(1 - \frac{\alpha_0}{\beta_0}\frac{\lambda}{\lambda_0}\right)_+^2\right). \tag{3.6}$$

We construct Θ_2 by applying a first order optimality condition on $\mathcal{D}(\lambda_0)$: $\boldsymbol{\theta}_0^\star$ is a dual optimal point if $\boldsymbol{g}^T(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\star) \leq 0$ for every dual point $\boldsymbol{\theta}_0$ that is feasible for $\mathcal{D}(\lambda_0)$, where $\boldsymbol{g} := -\nabla G(\boldsymbol{\theta}_0^\star) = \boldsymbol{\theta}_0^\star + \boldsymbol{y}$. For $\lambda \leq \lambda_0$, any dual point $\boldsymbol{\theta}$ feasible for $\mathcal{D}(\lambda)$ is also dual feasible for $\mathcal{D}(\lambda_0)$ ($|\boldsymbol{\theta}^T\boldsymbol{x}_k| \leq \lambda \leq \lambda_0 \ \ k = 1, \ldots, n$). Since $\boldsymbol{\theta}^\star$ is dual feasible for $\mathcal{D}(\lambda_0)$, we conclude $\boldsymbol{\theta}^\star \in \Theta_2 := \left\{\boldsymbol{\theta} \mid g^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0^\star) \geq 0\right\}$.

Figure 3.2(a) shows the geometry of Θ_1, Θ_2 and Θ in the dual space; Figure 3.2(b) shows the geometric interpretation of the inequality test when it is applied to the set Θ.

3.3.2 SAFE-LASSO theorem

Our criterion to identify the k-th zero in $\boldsymbol{w}^\star$ and thus remove the k-th feature (column) from the feature matrix X in problem $\mathcal{P}(\lambda)$ becomes

$$\lambda > \max(P'(\boldsymbol{x}_k, \Theta), P'(-\boldsymbol{x}_k, \Theta)).$$

We parameterize Θ in terms of $\boldsymbol{\theta}_0^\star$ and γ, and $P'(\boldsymbol{x}_k, \Theta)$ is represented as $P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma)$, where $P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma)$ is the optimal value of the convex optimization problem:

$$P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) := \max_\theta \boldsymbol{x}^T\boldsymbol{\theta} \ : \ G(\boldsymbol{\theta}) \geq \gamma, \ \boldsymbol{g}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0^\star) \geq 0. \tag{3.7}$$

We can express problem (B.1) in dual form as a convex optimization problem with two scalar variables, μ_1 and μ_2:

$$\begin{aligned}
P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) &= \min_{\mu_1,\mu_2 \geq 0} \max_\theta \boldsymbol{x^T\theta} + \mu_1(G(\boldsymbol{\theta}) - \gamma) + \mu_2\boldsymbol{g}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0^\star) \\
&= \min_{\mu_1,\mu_2 \geq 0} -\mu_1\gamma - \mu_2\boldsymbol{g}^T\boldsymbol{\theta}_0^\star + \max_{\boldsymbol{\theta}} \boldsymbol{x}^T\boldsymbol{\theta} + \mu_1 G(\boldsymbol{\theta}) + \mu_2\boldsymbol{g}^T\boldsymbol{\theta} \\
&= \min_{\mu_1,\mu_2 \geq 0} -\mu_1\gamma - \mu_2\boldsymbol{g}^T\boldsymbol{\theta}_0^\star + \mu_1 \max_{\boldsymbol{\theta}} \left(\frac{\boldsymbol{x}^T - \mu_1\boldsymbol{y}^T + \mu_2\boldsymbol{g}^T}{\mu_1}\theta - \frac{1}{2}\|\boldsymbol{\theta}\|_2^2\right)
\end{aligned}$$

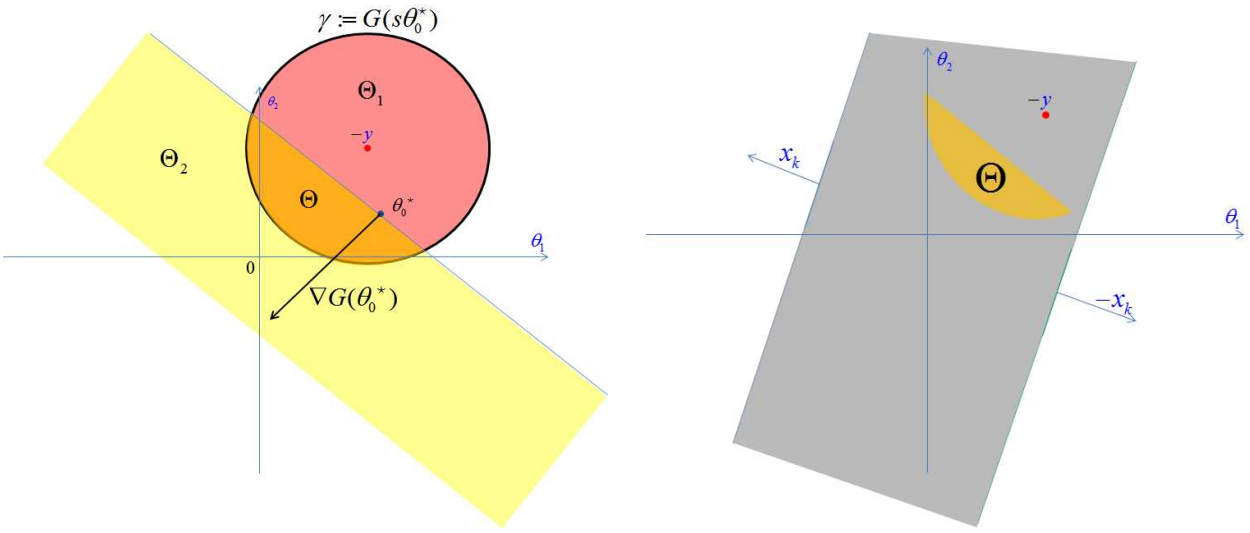


Figure 3.2: (a) Sets containing $\boldsymbol{\theta}^\star$ in the dual space. The set $\Theta_1 := \{\boldsymbol{\theta} \mid G(\boldsymbol{\theta}) \geq \gamma\}$ shown in red corresponds to a ball in the dual space with center $-\boldsymbol{y}$. The set $\Theta_2 := \left\{\boldsymbol{\theta} \mid \boldsymbol{g}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0^\star) \leq 0\right\}$ with $\boldsymbol{g} := \nabla G(\boldsymbol{\theta}_0^\star)$ shown in yellow corresponds to a half space with supporting hyperplane passing through $\boldsymbol{\theta}_0^\star$ and normal to $\nabla G(\boldsymbol{\theta}_0^\star)$. The set $\Theta = \Theta_1 \cap \Theta_2$ shown in orange contains the dual optimal point $\boldsymbol{\theta}^\star$. (b) Geometry of the inequality test $\lambda > \left|\boldsymbol{\theta}^T \boldsymbol{x}_k\right|$, $\forall \boldsymbol{\theta} \in \Theta$. The Grey shaded region is the slab corresponding to feature $\boldsymbol{x}_k$, i.e. $\left\{\boldsymbol{\theta} \mid \boldsymbol{\theta}^T \boldsymbol{x}_k \leq \lambda\right\}$. The test $\lambda > \left|\boldsymbol{\theta}^T \boldsymbol{x}_k\right|$, $\forall \boldsymbol{\theta} \in \Theta$ is a strict inequality when the entire set Θ (shown in orange) is inside the slab defined by the feature $\boldsymbol{x}_k$. In such case, the dual optimal point $\boldsymbol{\theta}^\star \in \Theta$ is also inside the slab and by (3.1), we conclude $\boldsymbol{w}^\star(k) = 0$.

We obtain:

$$P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) \quad = \quad \min_{\mu_1, \mu_2 \geq 0} L(\mu_1, \mu_2) \tag{3.8}$$

with

$$L(\mu_1, \mu_2) = -\boldsymbol{x}^T \boldsymbol{y} + \frac{\mu_1}{2} D^2 + \frac{1}{2\mu_1} \|\boldsymbol{x}\|_2^2 + \frac{\mu_2^2}{2\mu_1} \|\boldsymbol{g}\|_2^2 + \frac{\mu_2}{\mu_1} \boldsymbol{x}^T \boldsymbol{g} - \mu_2 \|\boldsymbol{g}\|_2^2 , \tag{3.9}$$

and $D := \left(\|\boldsymbol{y}\|_2^2 - 2\gamma\right)^{1/2}$.

To solve (3.8), we take the derivative of (3.9) w.r.t μ_2 and set it to zero

$$\mu_2 \|\boldsymbol{g}\|_2^2 + \boldsymbol{x^T g} - \mu_1 \|\boldsymbol{g}\|_2^2 = 0.$$

This implies that $\mu_2 = \max(0, \mu_1 - \frac{\boldsymbol{x}^T \boldsymbol{g}}{\|\boldsymbol{g}\|_2^2})$. When $\mu_1 \leq \frac{\boldsymbol{x}^T \boldsymbol{g}}{\|\boldsymbol{g}\|_2^2}$, we have $\mu_2 = 0$, $\mu_1 = \frac{\|\boldsymbol{x}\|_2}{D}$ and $P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma)$ takes the value:

$$P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) = -\boldsymbol{y}^T \boldsymbol{x} + \|\boldsymbol{x}\|_2 D.$$

On the other hand, when $\mu_1 \geq \frac{\boldsymbol{x}^T\boldsymbol{g}}{\|\boldsymbol{g}\|_2^2}$, we take the derivative of (3.9) w.r.t μ_1 and set it to zero:

$$\tilde{D}^2\mu_1^2 = \Psi^2,$$

with $\Psi = \left(\|\boldsymbol{x}\|_2^2 - \frac{\left(\boldsymbol{x}^T\boldsymbol{g}\right)^2}{\|\boldsymbol{g}\|_2^2}\right)^{1/2}$ and $\tilde{D} = \left(D^2 - \|\boldsymbol{g}\|_2^2\right)^{1/2}$. Substituting μ_1 and μ_2 in (3.8), $P(\gamma, \boldsymbol{x})$ takes the value:

$$P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) = \boldsymbol{\theta}_0^{\star T}\boldsymbol{x} + \Psi\tilde{D}.$$

Figure 3.3 shows a comparison between the bound computed using $P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma)$ and that of (3.4).

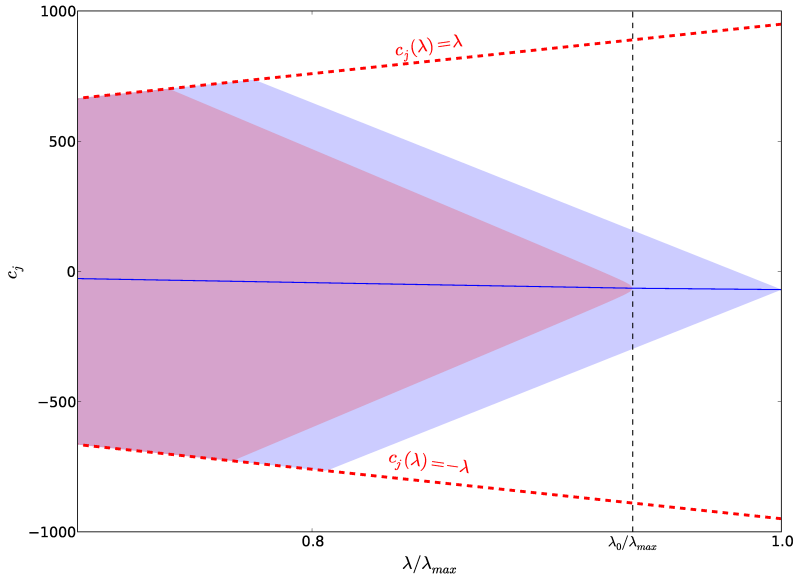


Figure 3.3: Comparison of two SAFE test bounds on $c_j(\lambda)$. Knowledge of a LASSO solution at some regularization parameter λ_0 leads to better bounds on the quantity $c_j(\lambda)$. Figure 3.1 is superposed with the red shaded region, which is computed using (3.11). The bound $c_k^l(\lambda) = -P(-\boldsymbol{x}_k, \boldsymbol{\theta_0}^\star, \gamma))$ and $c_k^u(\lambda) = P(\boldsymbol{x}_k, \boldsymbol{\theta_0}^\star, \gamma))$, accurately predict the value of $c_k(\lambda)$ at $\lambda = \lambda_0$. For $\lambda \leq \lambda_0$, the the bound estimates are tighter than those provided in (3.5), graphically the red shaded region is always inside the blue one.

Theorem 3.3.1 (SAFE-LASSO) *Consider the LASSO problem $\mathcal{P}(\lambda)$. Let $\lambda_0 \geq \lambda$ be a value for which an optimal solution $\boldsymbol{w}_0^\star \in \mathbb{R}^n$ is known. Denote by x_k the k-th feature (column) of the matrix X. Define*

$$\mathcal{E} = \{k \mid \lambda > \max(P(\boldsymbol{x}_k, \boldsymbol{\theta_0}^\star, \gamma), P(-\boldsymbol{x}_k, \boldsymbol{\theta_0}^\star, \gamma)\}, \tag{3.10}$$

where

$$P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma) = \begin{cases} \boldsymbol{\theta}_0^{\star T}\boldsymbol{x}_k + \Psi_k \tilde{D}(\gamma) & \|\boldsymbol{g}\|_2^2 \|\boldsymbol{x}_k\|_2 \geq D(\gamma)\boldsymbol{x}_k^T\boldsymbol{g}, \\ -\boldsymbol{y}^T\boldsymbol{x}_k + \|\boldsymbol{x}_k\|_2 D(\gamma) & \|\boldsymbol{g}\|_2^2 \|\boldsymbol{x}_k\|_2 \leq D(\gamma)\boldsymbol{x}_k^T\boldsymbol{g}, \end{cases} \tag{3.11}$$

with

$$\boldsymbol{\theta}_0^\star = \boldsymbol{X}\boldsymbol{w}_0^\star - \boldsymbol{y}, \;\; \boldsymbol{g} := \boldsymbol{\theta}_0^\star + \boldsymbol{y}, \;\; \alpha_0 := \boldsymbol{\theta}_0^{\star T}\boldsymbol{\theta}_0^\star, \;\; \beta_0 := |\boldsymbol{y}^T\boldsymbol{\theta}_0^\star|, \;\; \gamma := \tfrac{\beta_0^2}{2\alpha_0}\left(1 - \left(1 - \tfrac{\alpha_0}{\beta_0}\tfrac{\lambda}{\lambda_0}\right)_+^2\right),$$
$$D(\gamma) = \left(\|\boldsymbol{y}\|_2^2 - 2\gamma\right)^{1/2}, \;\; \tilde{D}(\gamma) = \left(D(\gamma)^2 - \|\boldsymbol{g}\|_2^2\right)^{1/2}, \;\; \Psi_k := \left(\|\boldsymbol{x}_k\|_2^2 - \tfrac{\left(\boldsymbol{x}_k^T\boldsymbol{g}\right)^2}{\|\boldsymbol{g}\|_2^2}\right)^{1/2}.$$

Then, for every index $e \in \mathcal{E}$, the e-th entry of $\boldsymbol{w}^\star$ is zero, i.e. $\boldsymbol{w}^\star(e) = 0$, and feature $\boldsymbol{x}_e$ can be safely eliminated from $\boldsymbol{X}$ a priori to solving the LASSO problem $\mathcal{P}(\lambda)$ ■

When we don't have access to a solution $\boldsymbol{w}_0^\star$ of $\mathcal{P}(\lambda_0)$, we can set $\boldsymbol{w}_0^\star = \boldsymbol{0}$ and $\lambda_0 = \lambda_{\max} := \|\boldsymbol{X}^T\boldsymbol{y}\|_\infty$. In this case, the inequality test $\lambda > \max(P(\boldsymbol{x}_k, \boldsymbol{\theta_0}^\star, \gamma), P(\boldsymbol{x}, \boldsymbol{\theta_0}^\star, \gamma)$ reduces to the form in Theorem 3.2.1, i.e. $\lambda > \rho_k\lambda_0$, with

$$\rho_k = \frac{\|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2 + |\boldsymbol{y}^T\boldsymbol{x}_k|}{\|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2 + \lambda_0}.$$

3.3.3 SAFE for LASSO with intercept problem

The SAFE-LASSO theorem can be applied to the LASSO with intercept problem

$$\mathcal{P}_{\text{int}}(\lambda) \;:\; \phi(\lambda) := \min_{\boldsymbol{w},\nu} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} + \boldsymbol{1}\nu - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1,$$

with $\nu \in \mathbb{R}^m$ the intercept term, by using a simple transformation. We solve for the optimal ν by taking the gradient of the objective function with respect to ν and set it to zero,

$$\left(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{1}\nu - \boldsymbol{y}\right)^T \boldsymbol{1} = 0.$$

We obtain $\nu = \bar{y} - \bar{\boldsymbol{X}}\boldsymbol{w}$ with $\bar{y} = (1/m)\boldsymbol{1}^T\boldsymbol{y}$, $\bar{\boldsymbol{X}} = (1/m)\boldsymbol{1}^T\boldsymbol{X}$ and $\boldsymbol{1} \in \mathbb{R}^m$ the vector of ones . Using the expression of ν, $\mathcal{P}_{\text{int}}(\lambda)$ can be expressed as

$$\mathcal{P}_{\text{int}}(\lambda) \;:\; \phi(\lambda) := \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{X}_{\text{cent}}\boldsymbol{w} - \boldsymbol{y}_{\text{cent}}\|_2^2 + \lambda\|\boldsymbol{w}\|_1,$$

with $\boldsymbol{X}_{\text{cent}} := \boldsymbol{X} - \boldsymbol{1}\bar{\boldsymbol{X}}$ and $\boldsymbol{y}_{\text{cent}} = \boldsymbol{y} - \boldsymbol{1}\bar{y}$. Thus the SAFE-LASSO theorem can be applied to $\mathcal{P}_{\text{int}}$ and eliminate features (columns) from $\boldsymbol{X}_{\text{cent}}$.

3.3.4 SAFE for elastic net

The elastic net problem

$$\mathcal{P}_{\text{elastic}}(\lambda) \ : \ \phi(\lambda) := \min_{w} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 + \frac{1}{2}\epsilon \|\boldsymbol{w}\|_2^2 ,$$

can be expressed in the form of $\mathcal{P}(\lambda)$ by replacing $\boldsymbol{X}$ and $\boldsymbol{y}$ with $X_{\text{elastic}} = \left(\boldsymbol{X}^T, \sqrt{\epsilon} I\right)^T$ and $\boldsymbol{y}_{\text{elastic}} = \left(\boldsymbol{y}^T, \boldsymbol{0}^T\right)^T$. This transformation allows us to apply the SAFE-LASSO theorem on $\mathcal{P}_{\text{elastic}}(\lambda)$ and eliminate features from $\boldsymbol{X}_{\text{elastic}}$.

3.4 Using SAFE

In this section, we illustrate the use of SAFE and detail the relevant algorithms.

3.4.1 SAFE for reducing memory limit problems

SAFE can extend the reach of LASSO solvers to larger size problems than what they could originally handle. In this section, we are interested in solving for $\boldsymbol{w}_d^\star$ the solution of $\mathcal{P}(\lambda_d)$ under a memory constraint of loading only M features. We can compute $\boldsymbol{w}_d^\star$ by solving a sequence of problems, where each problem has a number of features less than our memory limit M. We start by finding an appropriate λ where our SAFE method can eliminate at least $n-M$ features, we then solve a reduced size problem with $L_F \leq M$ features, where $L_F = |\mathcal{E}^c|$ is the number of features left after SAFE and $\mathcal{E}^c = \{1, \ldots, n\} \setminus \mathcal{E}$ is the complement of the set $\mathcal{E}$ in the SAFE-LASSO theorem. We proceed to the next stage as outlined in Algorithm 2.

We use a bisection method to find an appropriate value of λ for which SAFE leaves $L_F \in [M - \epsilon_F, M]$ features, where ϵ_F is a number of feature tolerance. The bisection method on λ is outlined in Algorithm 3.

3.4.2 SAFE for LASSO run-time reduction

In some applications like [22], it is of interest to solve a sequence of problems $\mathcal{P}(\lambda_1), \ldots \mathcal{P}(\lambda_s)$ for decreasing values of the penalty parameters, i.e. $\lambda_1 \geq \ldots \geq \lambda_s$. The computational complexities of LASSO solvers depend on the number of features and using SAFE might result in run-time improvements. For each problem in the sequence, we can use SAFE to reduce the number of features a priori to using our LASSO solver as shown in Algorithm 4.

Algorithm 2 SAFE for reducing memory limit problems

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$, penalty parameter λ_d , memory limit M and LASSO solver: `LASSO`, i.e. $\boldsymbol{w}^\star = \texttt{LASSO}(\boldsymbol{X}, \boldsymbol{y}, \lambda)$.
initialize $\lambda_0 = \|\boldsymbol{X}^T \boldsymbol{y}\|_\infty$, $\boldsymbol{w}_0^\star = \boldsymbol{0} \in \mathbb{R}^n$,
repeat

1. *Use SAFE to search for a λ with $LF \leq M$.* Obtain λ and $\mathcal{E}$. % L_F is the number of features left after SAFE and $\mathcal{E}$ is the set defined in the SAFE-LASSO theorem.
2. **if** $\lambda < \lambda_d$ **then** $\lambda = \lambda_d$, apply SAFE to obtain $\mathcal{E}$ **end if.**
3. Compute the solution $\boldsymbol{w}^\star$. $\boldsymbol{w}^\star(\mathcal{E}^c) = \texttt{LASSO}(\boldsymbol{X}_{:,\mathcal{E}^c}, \boldsymbol{y}, \lambda)$, $\boldsymbol{w}^\star(\mathcal{E}) = 0$; % $\boldsymbol{w}^\star(\mathcal{E}^c)$ and $\boldsymbol{X}_{:,\mathcal{E}^c}$ are the elements and columns of $\boldsymbol{w}^\star$ and $\boldsymbol{X}$ defined by the set $\mathcal{E}^c$, respectively. $\mathcal{E}^c = \{1, \ldots, n\} \setminus \mathcal{E}$ is the complement of the set $\mathcal{E}$.
4. $\lambda_0 := \lambda$, $\boldsymbol{w}_0^\star = \boldsymbol{w}^\star$.

until $\lambda_0 = \lambda_d$

3.5 Numerical results

In this section, we explore the benefits of SAFE by running numerical experiments[1] with different LASSO solvers. We present two kinds of experiments to highlight the two main benefits of SAFE. One kind, in our opinion the most important, shows how memory limitations can be reduced, by allowing to treat larger data sets. The other focuses on measuring computational time reduction when using SAFE a priori to the LASSO solver.

We have used a variety of available algorithms for solving the LASSO problem. We use acronyms to refer to the following methods: IPM stands for the Interior-Point Method for LASSO described in [27]; GLMNET corresponds to the Generalized Linear Model algorithm described in [20]; TFOCS corresponds to Templates for First-Order Conic Solvers described in [4]; FISTA and Homotopy stand for the Fast Iterative Shrinkage-Thresholding Algorithm and homotopy algorithm, described and implemented in [56], respectively. Some methods (like IPM, TFOCS) do not return exact zeros in the final solution of the LASSO problem and the issue arises in evaluating the cardinality. In appendix A, we discuss some issue related to the thresholding of the LASSO solution.

In our experiments, we use data sets derived from text classification sources in [19]. We use medical journal abstracts from PubMed represented in a bag-of-words format, where stop words have been eliminated and capitalization removed. The dimensions of the feature matrix X we use from PubMed is $m = 1,000,000$ abstracts and $n = 127,025$ features

[1] In our experiments, we have used an Apple Mac Pro 64-bit workstation, with two 2.26 GHz Quad-Core Intel Xeon processors, 8 MB on-chip shared L3 cache per processor, with 6 GB SDRAM, operating at 1066 MHz.

Algorithm 3 Bisection method on λ.

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$, penalty parameter λ_0 with LASSO solution $\boldsymbol{w}_0^\star$, tolerance $\epsilon_F > 0$, memory limit M, and maximum iterations $k_{\max}$.
initialize $l = 0$, $u = \lambda_0$, $k = 0$.
repeat

1. Set $\lambda := (l + u)/2$.
2. Use the SAFE-LASSO theorem to obtain $\mathcal{E}$.
3. Set $L_F = |\mathcal{E}^c|$.
4. **if** $L_F > M$ **then** set $l := \lambda$ **else** set $u := \lambda$ **end if**
5. $k = k + 1$.

until $(M - L_F \leq \epsilon_F$ and $L_F \leq M\ )$ or $k > k_{\max}$.

(words). There is a total of $82,209,586$ non-zeros in the feature matrix, with an average of about 645 non-zeros per feature (word). We also use data-sets derived from the headlines of *The New York Times,* (NYT) spanning a period of about 20 years (from 1985 to 2007). The number of headlines in the entire NYT data-set is $m = 3,241,260$ and the number of features (words) is $n = 159,943$. There is a total of $14,083,676$ non-zeros in the feature matrix, with an average of about 90 non-zeros per feature.

In some applications such as [22], the goal is to learn a short list of words that are predictive of the appearance of a given query term (say, "lung" or "china") in the abstracts of medical journals or NYT news. The LASSO problem can be used to produce a summarization of the query term across the many abstracts or headlines considered. To be manageable by a human reader, the list of predictive terms should be very short (say at most 100 terms) with respect to the size of the dictionary n. To produce such a short list, we solve the LASSO problem $\mathcal{P}(\lambda)$ with different penalty parameters λ, and choose the appropriate penalty λ that would generate enough non-zeros in the LASSO solution (around 100 non-zeros in our case).

3.5.1 SAFE for reducing memory limit problems

We experiment with PubMed data-set which is too large to be loaded into memory, and thus not amenable to current LASSO solvers. As described before, we are interested in solving the LASSO problem for a regularization parameter that would result in about 100 non-zeros in the solution. We implement Algorithm 2 with a memory limit $M = 1,000$ features, where we have observed that for the PubMed data loading more than $1,000$ features causes memory problems in the machine and platform we are using. The memory limit is approximately

Algorithm 4 Recursive SAFE for the Lasso

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$, a sequence of penalty parameters $\lambda_s \leq \ldots \leq \lambda_1 \leq \|\boldsymbol{X}^T \boldsymbol{y}\|_\infty$, and LASSO solver: `LASSO`.
initialize $\lambda_0 = \|\boldsymbol{X}^T \boldsymbol{y}\|_\infty$, $\boldsymbol{w}_0^\star = \boldsymbol{0} \in \mathbb{R}^n$.
for $i = 1$ **until** $i = s$ **do**

1. Set $\lambda_0 = \lambda_{i-1}$, and $\lambda = \lambda_i$.
2. Use the SAFE-LASSO theorem to obtain $\mathcal{E}$.
3. Compute the solution $\boldsymbol{w}^\star$. $\boldsymbol{w}^\star(\mathcal{E}^c) = \texttt{LASSO}(X(\mathcal{E}^c, :), y, \lambda)$, $\boldsymbol{w}^\star(\mathcal{E}) = 0$. % $\boldsymbol{w}^\star(\mathcal{E}^c)$ and $\boldsymbol{X}_{:,\mathcal{E}^C}$ are the elements and columns of $\boldsymbol{w}^\star$ and $\boldsymbol{X}$ defined by the set $\mathcal{E}^c$, respectively. $\mathcal{E}^c = \{1, \ldots, n\} \setminus \mathcal{E}$ is the complement of the set $\mathcal{E}$.
4. Set $\boldsymbol{w}_0^\star = \boldsymbol{w}^*$.

end for

two orders of magnitudes less than the original number of features n, i.e. $M \approx 0.01n$. Using Algorithm 2, we were able to solved the LASSO problem for $\lambda = 0.04\lambda_{max}$ using a sequence of 25 LASSO problem with each problem having a number of features less than $M = 1,000$. Figure 3.4 shows the simulation result for the PubMed data-set.

3.5.2 SAFE for LASSO run-time reduction

We have used a portion of the NYT data-set corresponding to all headlines in year 1985, the corresponding feature matrix has dimensions $n = 38,377$ features and $m = 192,182$ headlines, with an average of 21 non-zero per feature. We solved the plain LASSO problem and the LASSO problem with SAFE as outlined in Algorithm 4 for a sequence of λ logarithmically distributed between $0.03\lambda_{max}$ and λ_{max}. We have used four LASSO solvers, IPM, TFOCS, FISTA and Homotopy to solve the LASSO problem. Figure 3.5(a)shows the computational time saving when using SAFE. Figure 3.5(b) shows the number of features we used to solve the LASSO problem when using SAFE, and the number of non-zeros in the solution. We realize that when using Algorithm 4, we solve problems with a number of features at most $10,000$ instead of $n = 38,377$ features, this reduction has a direct impact on the solving time of the LASSO problem as demonstrated in Figure 3.5(a).

3.5.3 SAFE for LASSO with intercept problem

We return to the LASSO with intercept problem discussed in Section 3.3.3. We generate a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ with $m = 500$, $n = 10^6$. The entries of $\boldsymbol{X}$ has a $\mathcal{N}(0, 1)$ normal distributed and sparsity density $d = 0.1$. We also generate a vector of coefficients $\omega \in \mathbb{R}^n$

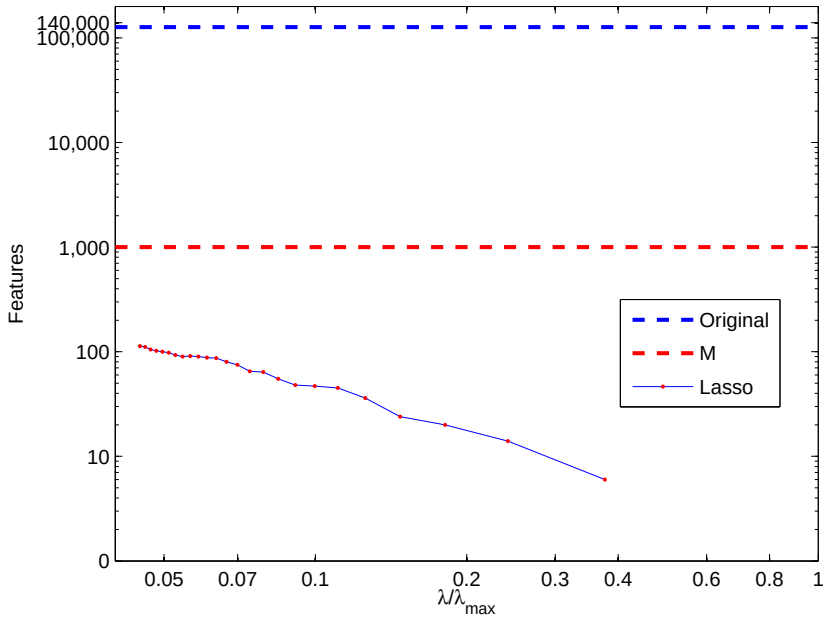


Figure 3.4: A LASSO problem solved for the PubMed data-set and $\lambda = 0.04\lambda_{max}$ using a sequence of 25 smaller size problems. Each LASSO problem in the sequence has a number of features L_F that satisfies the memory limit $M = 1,000$, i.e $L_F \leq 1,000$.

with 50 non-zero entries. The response $\boldsymbol{y}$ is generated by setting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\omega} + 0.01\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a vector in $\mathbb{R}^m$ with $\mathcal{N}(0,1)$ distribution. We use GLMNET implemented in `R` to solve the LASSO problem with intercept. The generated data, $\boldsymbol{X}$ and $\boldsymbol{y}$ can be loaded into `R` , yet memory problems occur when we try to solve the LASSO problem. We use Algorithm 2 with memory limit $M = 10,000$ features and $\lambda = 0.33\lambda_{max}$. Figure 3.6 shows the number of non-zeros in the solution of the 352 sequence of problems used to obtain the solution at $\lambda = 0.33\lambda_{\max}$.

3.6 Conclusion

In this chapter, we presented the basic idea behind the SAFE method. We derived two simple theorems to discard features safely and cheaply. The numerical results showed that the SAFE method is aggressive at eliminating features for large values of the penalty parameter and can reduce the computational cost of obtaining LASSO solutions for large-scale sparse problems. This property of SAFE can also extend the reach of LASSO solvers to problems previously out of there reach.

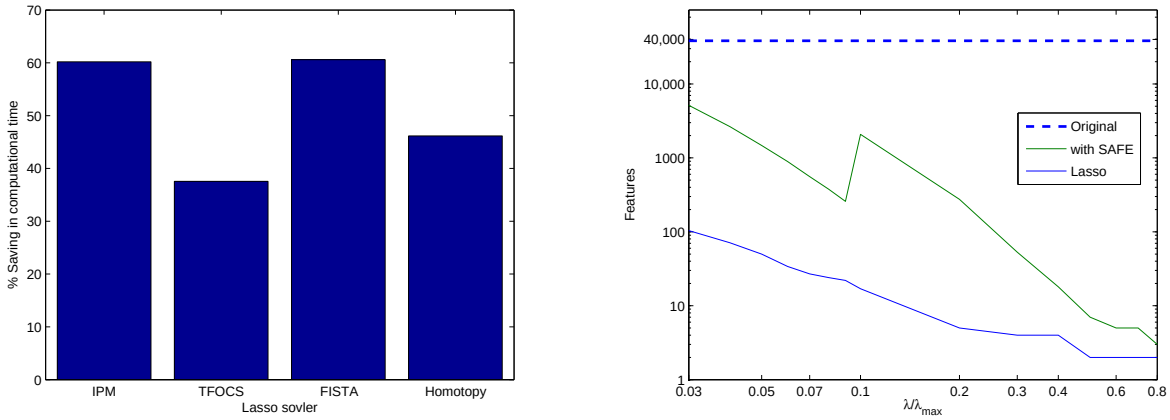


Figure 3.5: (a) Computational time savings. (b) Lasso solution for the sequence of problem between $0.03\lambda_{max}$ and λ_{max}. The green line shows the number of features we used to solve the LASSO problem after using Algorithm 4.

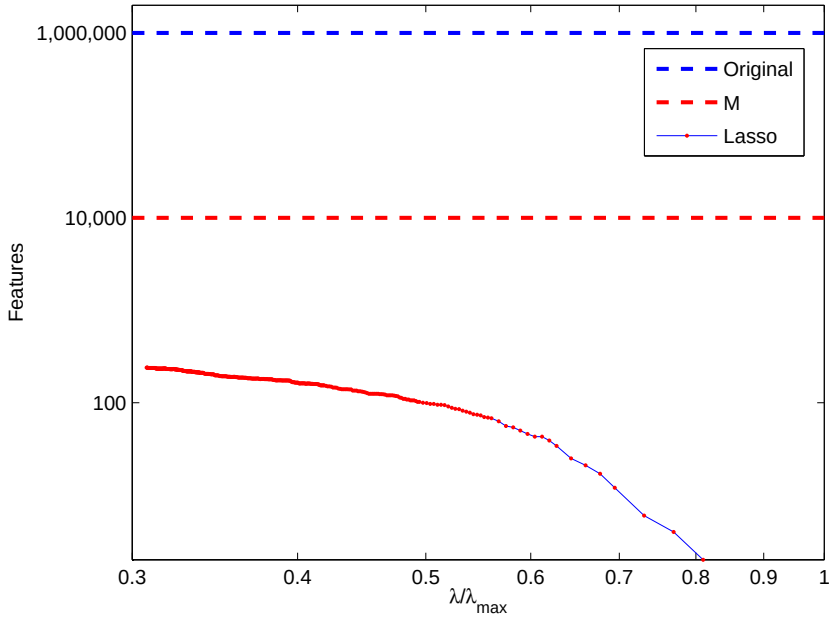


Figure 3.6: A LASSO problem with intercept solved for randomly generated data-set and $\lambda = 0.33\lambda_{max}$ using a sequence of 352 smaller size problems. Each LASSO problem in the sequence has a number of features L_F that satisfies the memory limit $M = 10,000$, i.e $L_F \leq 1000$.

Chapter 4

SAFE in the LOOP

4.1 Introduction

Recall, it is possible to eliminate a feature $\boldsymbol{x}_k$ from a feature matrix $\boldsymbol{X}$ in the LASSO problem

$$\mathcal{P}(\lambda) \ : \ \phi(\lambda) := \min_{\boldsymbol{w}} \frac{1}{2} \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + \lambda \left\| \boldsymbol{w} \right\|_1, \tag{4.1}$$

by checking the SAFE test

$$\lambda > \max(P(\boldsymbol{x}_k, \Theta), P(-\boldsymbol{x}_k, \Theta)),$$

where $P(\boldsymbol{x}_k, \Theta)$ is a convex optimization problem of the form

$$P(\boldsymbol{x}, \Theta) := \max_{\boldsymbol{\theta}} \boldsymbol{x}^T \boldsymbol{\theta} \ : \ \boldsymbol{\theta} \in \Theta,$$

and Θ is a set that contains the dual optimal point of the LASSO dual problem

$$\mathcal{D}(\lambda) \ : \ \phi'(\lambda) := \max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) \ : \ \left| \boldsymbol{\theta}^T \boldsymbol{x}_k \right| \leq \lambda, \ k = 1, \ldots, n, \tag{4.2}$$

with $G(\boldsymbol{\theta}) = \frac{1}{2} \left\| \boldsymbol{y} \right\|_2^2 - \frac{1}{2} \left\| \boldsymbol{\theta} + \boldsymbol{y} \right\|_2^2$. In this chapter, we solve the SAFE test problem for a non-empty set of the form

$$\Theta\left(\boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma\right) = \left\{ \boldsymbol{\theta} \left| G(\boldsymbol{\theta}) \geq \gamma, \ \boldsymbol{\eta}^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}_s \right) \geq 0 \right. \right\},$$

where $\boldsymbol{\eta}$, $\boldsymbol{\theta}_s$, and γ are general, i.e. not constrained as in Section 3.3.1. We then define the parameters of $\Theta\left(\boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma\right)$ so that it contains the dual optimal point. Finally, we derive a SAFE theorem for the LASSO, more aggressive at eliminating features than the preceeding SAFE theorems (Theorem 3.2.1 and Theorem 3.3.1).

The SAFE theorem we derive in this chapter has a similar closed-form inequality test to Theorem 3.2.1, i.e. if $\lambda > \rho_k \lambda_0$ then eliminate feautre $\boldsymbol{x}_k$. The closed-form solution allows

us to integrate SAFE into LASSO solvers in a more effecient way than Chapter 3, without the need of implementing iterative algorithms like Algorithm 3. In addition, mathematical terms that ρ_k depends on are usually precomputed by many LASSO algorithms.

In this chapter, we derive a SAFE-LASSO theorem with a closed-form sufficient condition in section 4.2. We then implement a Coordinate-Descent (CD) method and integrate our SAFE method in it in section 4.3. Finally, in section 4.4, we show using numerical experiments how SAFE can improve computational complexity and extend the reach of the CD method to larger size problems.

4.2 A better SAFE method for the LASSO

In this section, we present the steps necessary to derive the SAFE-LASSO Theorem.

4.2.1 Solving the SAFE test problem

The SAFE test problem

$$P(\boldsymbol{x}, \Theta) := \max_{\boldsymbol{\theta}} \boldsymbol{x}^T \boldsymbol{\theta} \ : \ \boldsymbol{\theta} \in \Theta,$$

with

$$\Theta\left(\boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma\right) = \left\{ \boldsymbol{\theta} \,\middle|\, G(\boldsymbol{\theta}) \geq \gamma,\ \boldsymbol{\eta}^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}_s\right) \geq 0 \right\},$$

admits a closed form solution as shown in the following propostion.

Proposition 4.2.1 (SAFE-LASSO test Problem) *Consider the problem*

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) := \max_{\boldsymbol{\theta}} \boldsymbol{x}^T \boldsymbol{\theta} \ : \ G(\boldsymbol{\theta}) \geq \gamma,\ \boldsymbol{\eta}^T \left(\boldsymbol{\theta} - \boldsymbol{\theta}_s\right) \geq 0, \tag{4.3}$$

with $G(\boldsymbol{\theta}) = -\frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2 - \boldsymbol{y}^T\boldsymbol{\theta}$. *Assume that strong duality holds and a solution of the problem is attained. Let* $\boldsymbol{g}_s = \boldsymbol{\theta}_s + \boldsymbol{y}$ *, then* $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ *takes the value*

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{y}^T\boldsymbol{x} + \left\|\boldsymbol{x}\right\|_2 D & \left\|\boldsymbol{x}\right\|_2 \left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D\left(\boldsymbol{\eta}^T\boldsymbol{x}\right), \\ \frac{1}{\left\|\boldsymbol{\eta}\right\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y} + \psi\tilde{D} & \textit{otherwise}, \end{cases} \tag{4.4}$$

with

$$D = \left(\left\|\boldsymbol{y}\right\|_2^2 - 2\gamma\right)^{1/2},$$

$$\tilde{D} = \left(-2\gamma - \frac{\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right)^2}{\left\|\boldsymbol{\eta}\right\|_2^2} + \left\|\boldsymbol{y}\right\|_2^2\right)^{1/2},$$

and

$$\psi = \left(\left\|\boldsymbol{x}\right\|_2^2 - \frac{1}{\left\|\boldsymbol{\eta}\right\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2\right)^{1/2}.$$

Proof: See Appendix B for proof. ■

4.2.2 Definning Θ

Similar to Section 3.3, we assume that we are interested in solving the LASSO at some λ and we have knoweldge of a LASSO solution $\boldsymbol{w}_0$ at a regularization parameter λ_0. We find γ in the inequality $G(\boldsymbol{\theta}) \geq \gamma$ by dual scaling. Assuming $\boldsymbol{\theta}_s$ is a feasible point for the LASSO dual problem at λ, then the dual optimal point $\boldsymbol{\theta}^\star$ is in the set

$$\{\boldsymbol{\theta} \,|G(\boldsymbol{\theta}) \geq \gamma := G(\boldsymbol{\theta}_s)\}.$$

We set $\boldsymbol{\eta}$ and $\boldsymbol{\theta}_s$ for the halfspace inequality $\boldsymbol{\eta}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0$ using the following proposition.

Proposition 4.2.2 *Consider the LASSO problem with regularization parameter λ and assume we know a LASSO solution $\boldsymbol{w}_0$ at λ_0. Define $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, and $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$. Then, any feasible point $\boldsymbol{\theta}$ for the dual problem $\mathcal{D}(\lambda)$ is in the half-space*

$$\boldsymbol{\eta}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0.$$

Proof: See Appendix B for proof. ■

Since the half-space contains all feasible points of the dual problem at λ and $G(\boldsymbol{\theta}) \geq \gamma$ contains the dual optimal point, we conclude that the set

$$\Theta(\boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \left\{\boldsymbol{\theta} \,\middle|\, G(\boldsymbol{\theta}) \geq \gamma,\ \boldsymbol{\eta}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0\right\},$$

with $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$, $\gamma = G(\boldsymbol{\theta}_s)$, and $G(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta} + \boldsymbol{y}\|_2^2$ contains the dual optimal point $\boldsymbol{\theta}^\star$.

4.2.3 Evaluating the SAFE test

For the $\boldsymbol{\eta}$ and $\boldsymbol{\theta}_s$, we used to define our set Θ, the SAFE test problem can be further simplified.

Proposition 4.2.3 *Consider the SAFE test problem $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ in (4.3), and assume we know a LASSO solution $\boldsymbol{w}_0$ at a regularization parameter λ_0. Defining $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$, $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, and $\gamma = G(\boldsymbol{\theta}_s)$, the SAFE test problem reduces to*

$$P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{\delta}_0(k) + \frac{1}{\lambda_0}\|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2 |\lambda - \lambda_0| & \boldsymbol{\sigma}_1(k)\lambda \leq \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{\alpha_0 - \lambda\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) - \boldsymbol{\delta}_0(k) + \boldsymbol{\psi}(k)M|\lambda - \lambda_0| & \textit{otherwise}, \end{cases}$$

with $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, $\beta_0 = \|\boldsymbol{w}_0\|_1$, $\alpha_0 := \boldsymbol{w}_0^T\boldsymbol{\delta}_0 = \boldsymbol{y}^T\boldsymbol{\tau}$,

$$\boldsymbol{\sigma}_1(k) = \|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \beta_0,$$

$$\boldsymbol{\sigma}_2(k) = \|y\|_2 - \alpha_0 \|\boldsymbol{x}_k\|_2 ,$$
$$M = \frac{-\beta_0\lambda_0 + \|\boldsymbol{y}\|_2^2 - \alpha_0}{\lambda_0^2} - \frac{\beta_0^2}{\alpha_0 - \beta_0\lambda_0},$$

and

$$\boldsymbol{\psi}(k) = \left(\|\boldsymbol{x}_k\|_2^2 - \frac{1}{\alpha_0 - \beta_0\lambda_0}\boldsymbol{\delta}_1^2(k) \right)^{1/2}, \; k = 1, ...n.$$

Proof: See Appendix B for proof. ■

Thus, in order to evaluate the solution of the SAFE test problem, we need our feature matrix $\boldsymbol{X}$ in computing only three terms, $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, and $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$. The term $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$ can be evaluated offline, the term $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$ is usually computed by all LASSO algorithms as it is necessary to evaluate the cost function, or the gradient of the cost function when optimizing (2.1), and the term $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$ is evaluated most of the time by LASSO algorithms. Coordinate-Descent, and interior point methods are examples of such algorithms. Thus, the worst-case computational complexity is $O(mn)$ operations, which is the cost of a few vector-matrix multiplications in case we were to evaluate all the terms, $\boldsymbol{\tau}$, $\boldsymbol{\delta}_0$, and $\boldsymbol{\delta}_1$. Sparse matrices have a better count (less) of operations.

We note that the second case of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ is active when λ is close enough to λ_0, independent of the feature $\boldsymbol{x}_k$. More specifically when

$$(\lambda_0 - \lambda) < \lambda_0 \frac{\|\boldsymbol{\tau}\|_2}{\|y\|_2}.$$

We derive this result by evaluating $\boldsymbol{\sigma}_1(k)\lambda > \boldsymbol{\sigma}_2(k)\lambda_0$ for some feature $\boldsymbol{x}_k$, and then simplifying it to

$$\lambda_0 \|\boldsymbol{x}_k\|_2 (-\lambda\boldsymbol{\beta}_0 + \alpha_0) > \|y\|_2 (\lambda_0 - \lambda) \boldsymbol{\delta}_1(k). \tag{4.5}$$

A sufficient condition for (4.5) to hold true is

$$\lambda_0 \|\boldsymbol{x}_k\|_2 (-\lambda_0\boldsymbol{\beta}_0 + \alpha_0) > \|y\|_2 (\lambda_0 - \lambda) \boldsymbol{\delta}_1(k),$$

since $\lambda_0 \geq \lambda$. We recall that $\|\boldsymbol{\tau}\|_2^2 = -\lambda_0\boldsymbol{\beta}_0 + \alpha_0$, and the inequality holds true if $\boldsymbol{\delta}_1(k) < 0$. When $\boldsymbol{\delta}_1(k)$ is positive, we have

$$(\lambda_0 - \lambda) < \|\boldsymbol{x}_k\|_2 \frac{\lambda_0}{\boldsymbol{\delta}_1(k)} \frac{\|\boldsymbol{\tau}\|_2^2}{\|y\|_2}.$$

A sufficient condition for the above inequality to hold true is

$$(\lambda_0 - \lambda) < \lambda_0 \frac{\|\boldsymbol{\tau}\|_2}{\|y\|_2},$$

since $\frac{1}{\|\boldsymbol{x}_k\|_2\|\boldsymbol{\tau}\|_2^2} \leq \frac{1}{\boldsymbol{\delta}_1(k)}$.

Using the expression of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$, we evaluate the inequality test

$$\lambda < \max(P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma), P(-\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)),$$

using the following proposition.

Proposition 4.2.4 *Consider the LASSO (2.1) with feature matrix $\boldsymbol{X}$ and response $\boldsymbol{y}$. Also assume we know an optimal solution $\boldsymbol{w}_0$ at some regularization parameter λ_0, then for $\lambda \leq \lambda_0$ the test*

$$\lambda < max(P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma), P(-\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)),$$

with $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ the safe test problem in (4.3), $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$, $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, $\gamma = G(\boldsymbol{\theta}_s)$ and $G(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta} + \boldsymbol{y}\|_2^2$ is equivalent to

$$\lambda > \rho_k \lambda_0,$$

with $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, $\beta_0 = \|\boldsymbol{w}_0\|_1$, $\alpha_0 := \boldsymbol{w}_0^T\boldsymbol{\delta}_0 = \boldsymbol{y}^T\boldsymbol{\tau}$,

$$\boldsymbol{\sigma}_1^+(k) = \|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \boldsymbol{\beta}_0,$$

$$\boldsymbol{\sigma}_1^-(k) = -\|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \boldsymbol{\beta}_0,$$

$$\boldsymbol{\sigma}_2(k) = \|y\|_2 - \alpha_0 \|\boldsymbol{x}_k\|_2,$$

$$M = \frac{-\beta_0\lambda_0 + \|\boldsymbol{y}\|_2^2 - \alpha_0}{\lambda_0^2} - \frac{\beta_0^2}{\alpha_0 - \beta_0\lambda_0},$$

$$\boldsymbol{\psi}(k) = \left(\|\boldsymbol{x}_k\|_2^2 - \frac{1}{\alpha_0 - \beta_0\lambda_0}\boldsymbol{\delta}_1^2(k)\right)^{1/2}, \ k = 1, ...n.,$$

$$\rho_k = \max(\rho_k^-, \rho_k^+),$$

$$\rho_k^+ = \begin{cases} \frac{-\boldsymbol{\delta}_0(k) + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2}{\lambda_0 + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2} & \boldsymbol{\sigma}_1^+(k)\lambda \leq \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{\left(-\boldsymbol{\delta}_0(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \boldsymbol{\psi}(k)M\lambda_0}{\left(-\boldsymbol{\delta}_1(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \lambda_0 + \boldsymbol{\psi}(k)M\lambda_0} & otherwise, \end{cases}$$

and

$$\rho_k^- = \begin{cases} \frac{+\boldsymbol{\delta}_0(k) + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2}{\lambda_0 + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2} & \boldsymbol{\sigma}_1^-(k)\lambda \leq \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{-\left(-\boldsymbol{\delta}_0(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \boldsymbol{\psi}(k)M\lambda_0}{-\left(-\boldsymbol{\delta}_1(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \lambda_0 + \boldsymbol{\psi}(k)M\lambda_0} & otherwise. \end{cases}$$

Proof: See Appendix B for proof. ■

4.2.4 SAFE-LASSO theorem

We summarize the SAFE result in the following theorem:

Theorem 4.2.1 (SAFE-LASSO) *Consider the LASSO problem*

$$\min_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 ,$$

with $\boldsymbol{X}$ the feature matrix, $\boldsymbol{y}$ the response, $\boldsymbol{w}$ the optimization variable, $\lambda > 0$ the regularization parameter and let $\lambda_0 \geq \lambda$ be a value for which an optimal solution $\boldsymbol{w}_0 \in \mathbb{R}^n$ is known. Denoting by $\boldsymbol{x}_k$ the k-th feature (column) of the feature matrix $\boldsymbol{X}$, the condition

$$\lambda > \rho_k \lambda_0,$$

with ρ_k defined in 4.2.4, allows to safely remove the k-th feature from feature matrix $\boldsymbol{X}$. ■

4.3 SAFE in a Coordinate-Descent (CD) algorithm

In this section, we derive the Coordinate-Descent (CD) algorithm for solving the LASSO Problem. In addition, we present an algorithm that integrates SAFE with Coordinate-Descent.

4.3.1 Coordinate-Descent for the LASSO

The Coordinate-Descent (CD) method loops over each entry w_j of $\boldsymbol{w}$, and updates its value based on the optimization problem

$$w_j = \arg\min_{w_j} \quad \frac{1}{2} \left\| \boldsymbol{X}_{:,\mathcal{I}\backslash j} \boldsymbol{w}_{\mathcal{I}\backslash j} + \boldsymbol{x}_j w_j - \boldsymbol{y} \right\|_2^2 + \lambda \left\| \boldsymbol{w}_{\mathcal{I}\backslash j} + w_j \right\|_1 , \tag{4.6}$$

with $\mathcal{I} = \{1, ...n\}$ the set of all indices, and $\mathcal{I}\backslash j$ the set of all indicies excluding the index j. For notational convinience, we drop the $\mathcal{I}$ set notation and write $\boldsymbol{X}_{\backslash j}$ instead of $\boldsymbol{X}_{:,\mathcal{I}\backslash j}$, to donate a matrix composed by all the features of $\boldsymbol{X}$ except for the j-th feature (column). Similarly, we refer to $\boldsymbol{w}_{\backslash j}$ as the vector of all entries of $\boldsymbol{w}$ except for entry j. We find w_j in (4.6) by writing the subgradient optimality condition,

$$\boldsymbol{x}_j^T \left(\boldsymbol{X}_{\backslash j} \boldsymbol{w}_{\backslash j} + \boldsymbol{x}_j w_j - \boldsymbol{y} \right) + \lambda \partial w_j = 0,$$

or

$$w_j = \frac{1}{\|\boldsymbol{x}_j\|_2^2} \left(\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\backslash j} \boldsymbol{w}_{\backslash j} \right) - \lambda \partial w_j \right) , \tag{4.7}$$

with $\partial w_j = \text{sign}(w_j)$ and $\text{sign}(0) \in [-1, 1]$. We recognize that w_j can take values based on three cases.

First Case: If $\lambda \geq \left|\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right)\right|$, then $w_j = 0$.

Second Case: Assuming $\lambda < \left|\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right)\right|$, if $\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) > 0$, then $\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) \pm \lambda > 0$ and the right-hand side of (4.7) is non-negative. Thus, $w_j = \frac{1}{\|\boldsymbol{x}_j\|_2^2}\left(\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) - \lambda\right)$.

Third Case: Assuming $\lambda < \left|\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right)\right|$, if $\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) < 0$, then $\boldsymbol{x}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) \pm \lambda < 0$ and the right-hand side of (4.7) is non-positive. Thus, $w_j = \frac{1}{\|\boldsymbol{x}_j\|_2^2}\left(x_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{\setminus j}\boldsymbol{w}_{\setminus j}\right) + \lambda\right)$.

We summarize the Coordinate-Descent method in Algorithm 5.

Algorithm 5 Coordinate Descent method for the LASSO.

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$, a penalty parameter λ, an initial guess $\boldsymbol{w}$, number of maximum iterations $i_{\max}$, and tolerance ϵ.
initialize $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, $\boldsymbol{\alpha}(j) = \|\boldsymbol{x}_j\|_2$, $j = 1, ...n$, and set $\mathcal{S} = \{1, ..., n\}$.
for $i = 1$ **until** $i = i_{\max}$ **do**
Set $\boldsymbol{w}_p = \boldsymbol{w}$.
for j **in** $\mathcal{S}$ **do**

1. Set $w^- = w_j$, $\tilde{\delta}_1 = \boldsymbol{\delta}_1(j) - \alpha_j^2 w_j$.
2. **if** $\lambda \geq \left|\boldsymbol{\delta}_0(j) - \tilde{\delta}_1\right|$, **then** update $w_j := 0$.
3. **if** $\lambda < \left|\boldsymbol{\delta}_0(j) - \tilde{\delta}_1\right|$ **and** $\boldsymbol{\delta}_0(j) - \tilde{\delta}_1 > 0$, **then** update $w_j := \frac{1}{\alpha_j^2}\left(\boldsymbol{\delta}_0(j) - \tilde{\delta}_1 - \lambda\right)$.
4. **else** update $w_j := \frac{1}{\alpha_j^2}\left(\boldsymbol{\delta}_0(j) - \tilde{\delta}_1 + \lambda\right)$.
5. Update $\boldsymbol{\tau} := \boldsymbol{\tau} + \boldsymbol{x}_j\left(w_j - w^-\right)$ and $\boldsymbol{\delta}_1 := \boldsymbol{\delta}_1 + \left(w_j - w^-\right)\boldsymbol{X}^T\boldsymbol{x}_j$.

end for
if $\|\boldsymbol{w}_p - \boldsymbol{w}\|_2 < \epsilon\|\boldsymbol{w}\|_2$, **then** terminate.
end for

4.3.2 SAFE in the Coordinate-Descent loop

We integrate SAFE in the CD method by using an algorithm similar to Algorithm 2. Recall that Algorithm 2 defines a sequence of decreasing regularization parameters $\lambda_{\max} > \lambda_1 > ... > \lambda_d$, based on the features that SAFE can eliminate at each stage. It starts with $\lambda_{\max}$ and then uses SAFE to find a $\lambda_1 < \lambda_{\max}$ for which at most M feature are not eliminated, where $M \in \mathbb{N}^+$ is some memory limit. The step is repeated until we reach our desired

regularization parameter λ_d. Since the CD method provides us with the terms, $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, then we can apply SAFE with less computational effort. More specifically, we have only to sort the values of $\rho_k,\ k = 1, ..., n$, provided by Theorem 4.2.1 and use this ordering to find λ_1 instead of using a bisection method like Algorithm 3. The following algorithm presents the steps necessary to integrate SAFE into the Coordinate-Descent algorithm for solving the LASSO.

Algorithm 6 CD-SAFE. A Coordinate-Descent algorithm with SAFE integrated in the loop of its iterations.

given a feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, response $\boldsymbol{y} \in \mathbb{R}^m$, penalty parameter λ_d , memory limit M.
initialize $\boldsymbol{w}_0 = \boldsymbol{0} \in \mathbb{R}^n$, $\boldsymbol{\tau} = \boldsymbol{0} \in \mathbb{R}^m$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{0} \in \mathbb{R}^n$ and $\lambda_0 = \|\boldsymbol{\delta}_0\|_\infty$.
repeat

1. *Use SAFE-LASSO (4.2.1) with* $\boldsymbol{w}_0$, λ_0, $\boldsymbol{\tau}$, $\boldsymbol{\delta}_0$, and $\boldsymbol{\delta}_1$. Obtain $\boldsymbol{\rho}$.
2. Sort $\boldsymbol{\rho}$, $\rho_{i_1} < ... < \rho_{i_n}$, with i_j the index of the j-th largest element in $\boldsymbol{\rho}$.
3. Set $l = \inf\left\{j \,\middle|\, \rho_{i_j} = \boldsymbol{\rho}(i_M)\right\} - 1$ and $\lambda = \rho_{i_l}\lambda_0$.
4. **if** $\lambda < \lambda_d$ **then** set $\lambda = \lambda_d$.
5. Construct the set $\mathcal{S} = \{i_l, ..., i_1\}$.
6. *Use 5 to solve the LASSO at* λ. Initialize with $\mathcal{S}$, $\boldsymbol{w}_0$, $\boldsymbol{\tau}$, $\boldsymbol{\delta}_0$, and $\boldsymbol{\delta}_1$.
7. Obtain from 5 the updated $\boldsymbol{w}_0$, $\boldsymbol{\tau}$, and $\boldsymbol{\delta}_1$.
8. Set $\lambda_0 := \lambda$.

until $\lambda_0 = \lambda_d$

4.4 Numerical results

In this section, we explore the benefits of integrating SAFE with the Coordinate-Descent method by running numerical experiments[1] on different datasets. We present two kinds of experiments, to illustrate the two main benifites of SAFE. In the first experiment, we run CD-SAFE (Algorithm 6) and CD (Algorithm 5) on synthetic data of different sizes and in the second experiment we run CD-SAFE on three large-scale problems.

[1]In our experiments, we have used an Apple Mac Pro 64-bit workstation, with two 2.26 GHz Quad-Core Intel Xeon processors, 8 MB on-chip shared L3 cache per processor, with 6 GB SDRAM, operating at 1066 MHz.

The synthetic data sets are generated using a dense feature matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with $m = 100$, and $n = 10^3,\ 5\times10^3,\ 10\times10^3$. The entries of $\boldsymbol{X}$ has a $\mathcal{N}(0,1)$ normal distribution. We also generate a vector of coefficients $\omega \in \mathbb{R}^n$ with 100 non-zero entries. The response $\boldsymbol{y}$ is generated by setting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\omega} + 0.01\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a vector in $\mathbb{R}^m$ with $\mathcal{N}(0,1)$ distribution. We also use publicly available[2] datasets presented in [26]. The `TFIDF-2006`, and `LOG1P-2006` datasets are compiled from financial data, and `KDD2010b` is a dataset compiled from an online tutoring system, where features are built based on the answers of students to some questions. The statistics of these datatsets is summarized in Table 4.1.

Data set	m	n	nnz	range of $\boldsymbol{y}$
TFIDF-2006	$16,087$	$150,360$	$19,971,015$	$[-7.90, -0.52]$
LOG1P-2006	$16,087$	$4,272,227$	$96,731,839$	$[-7.90, -0.52]$
KDD2010b	$19,264,097$	$29,890,095$	$566,345,888$	$\{0,1\}$

Table 4.1: Feature matrix $\boldsymbol{X}$ statistics for different datasets. The number of observations is m , the number of features or variables is n, and the number of non-zero entries in the feature matrix $\boldsymbol{X}$ is nnz.

4.4.1 CD-SAFE and computational complexity

In our first experiment, the synthetic data we use is small enough for us to solve the LASSO problem using both algorithms, CD and CD-SAFE. We measure the number of iterations needed in order for each algorithm to reach a tolerance $\epsilon = 10^{-2}$. In each iteration we solve the problem (4.6) for some index $j,\ j = 1, ..., n$ of $\boldsymbol{w}$.

The LASSO problem is solved for a range of regularizatoin parameters $[\lambda_{\min}, \lambda_{\max}]$, where at $\lambda_{\min}$ we have at least 50 non-zeros in the solution. In Figure 4.1, we show the simulation results when the memory limit M in Algorithm 6 is set to 100, i.e. $M = 100$. The results show for the different dimensions of feature matrix $\boldsymbol{X}$, that the number of iterations is improved by 10 to 100 folds. With CD-SAFE it takes less iterations to reach the same tolerance for a given problem.

4.4.2 CD-SAFE for reducing memory limit problems

We have used the largest datasets publicly available we could find to carry out the numerical experiements. Loading the datasets (Table 4.1) alone causes memory problems for the machine we are using. While the CD method can be run in parallel and their in no need to load the entire dataset at once, it is still required to scan all the features. In the case of the `KDD2010b` dataset, for example, there are about 30×10^6 features and we know apriori that almost all of these features are going to be inactive at the optimal solution. With CD-SAFE

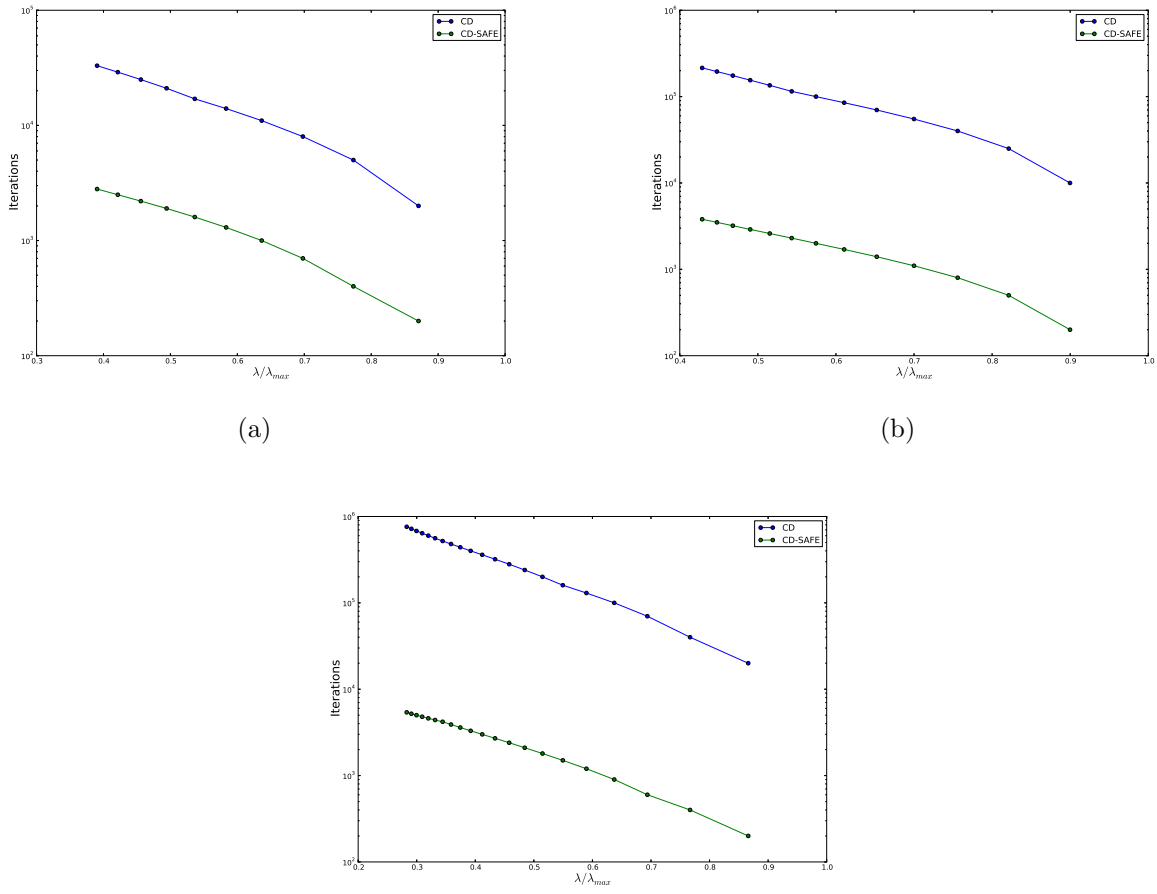
[2]Data sets can be found at http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

most of these inactive feature are discarded and thus we can handle large datasets and find solutions of interest very cheaply.

In Figure 4.2, we solve the LASSO problem using the datasets in Table 4.1, for a range of regularization parameters $[\lambda_{\min}, \lambda_{\max}]$, where at $\lambda_{\min}$ we have at least 50 non-zeros in the solution. We set the memory limit M in Algorithm 6 to 100 and the tolerance to $\epsilon = 10^{-2}$. In all simulations, it took less than $30,000$ iterations to reach the tolerance ϵ when solving the LASSO at regularization parameter $\lambda_{\min}$. This is a considerable improvement in computational complexity. Although we didn't solve the LASSO problems using the CD algorithm, but assuming that we need to scan the features only once to acheive the tolerance ϵ, then CD-SAFE still introduces at least 10^3 orders of magnitudes less compuations, i.e instead of 30×10^6 iterations, we need only 30×10^3 iterations to reach a tolerance ϵ. In addition to the improvements in computation complexity, CD-SAFE also resolved memoy problems since we are solving instances of the LASSO problem with a small feature matrix $\boldsymbol{X}_{\text{reduced}} \in \mathbb{R}^{m \times 100}$.

4.5 Conclusion

We have adapted the SAFE test problem from chapter 3 and derived an aggressive test for removing features for the LASSO problem. The closed-form solution of the sufficient condition for removing features allowed us to integrate SAFE efficiently with LASSO solvers. In this chapter, we integrated SAFE with the Coordinate-Descent algorithm and we called our algorithm CD-SAFE. SAFE allowed us to extend the reach of the CD algorithm to larger-size problems and allowed to reduce the compuational complexity by allowing us to reach a specific tollerance with fewer iterations.

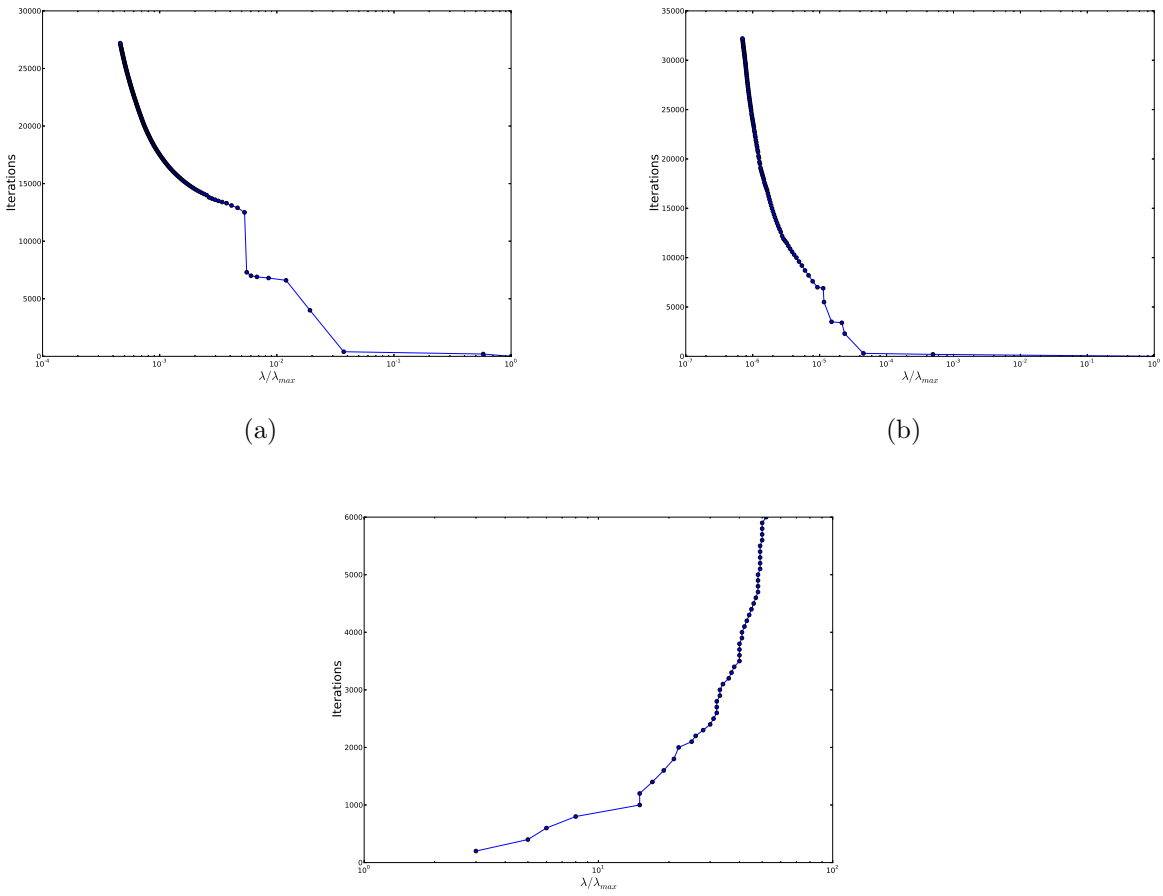


(a)

(b)

(c)

Figure 4.1: Number of iterations needed to reach a stationary tolerance of $\epsilon = 10^{-2}$ for the CD and CD-SAFE algorithms solved using synthetic data. The simulation results show that CD-SAFE provides at least 10 or 100 folds of less iterations to reach the same tolerance as the CD algorithm. The feature matrix used has the dimenstion $m = 1000$ observations and (a) $n = 1000$ features, (b) $n = 5000$ features and (c) $n = 10,000$.



(a) (b)

(c)

Figure 4.2: The LASSO (4.1) solved over a range of regularization parameters $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, using the CD-SAFE Algorithm (Algorithm 6). The plot shows the iterations needed to solve the LASSO problem at a particular λ. Each iteration is an instant of the problem (4.6) solved for some index of the solution w_i. (a) LOG1P-2006 dataset. (b) TFIDF-2006 dataset. (c) KDD2010b dataset.

Chapter 5

SAFE Applied to General ℓ_1-Regularized Convex Problems

5.1 Introduction

The SAFE-LASSO result presented in chapter 3 for the LASSO problem (2.1) can be adapted to a more general class of l_1- regularized convex problems. We consider the family of problems

$$\mathcal{P}(\lambda) \ : \ \phi(\lambda) := \min_{\boldsymbol{w},\nu} \sum_{i=1}^{m} f(\boldsymbol{a}_i^T \boldsymbol{w} + b_i v + c_i) + \lambda \left\|\boldsymbol{w}\right\|_1 , \tag{5.1}$$

where f is a closed convex function, and non-negative everywhere, $\boldsymbol{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, $\boldsymbol{b}$, $\boldsymbol{c} \in \mathbb{R}^m$ are given. The LASSO problem is a special case of (5.1) with $f(\zeta) = (1/2)\zeta^2$, $\boldsymbol{a}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ the observations, $\boldsymbol{c} = -\boldsymbol{y}$ is the (negative) response vector, and $\boldsymbol{b} = \boldsymbol{0}$. Hereafter, we refer to the LASSO problem as $\mathcal{P}_{\text{LASSO}}(\lambda)$ and to the general class of l_1-regularized problems as $\mathcal{P}(\lambda)$.

In this chapter, we outline the steps necessary to derive a SAFE method for the general problem $\mathcal{P}(\lambda)$ in section 5.2. We show some preliminary results for deriving SAFE methods when $f(\zeta)$ is the hing loss function, $f_{\text{hi}}(\zeta) = (1-\zeta)_+$ in section 5.3, and the logistic loss function $f_{\log}(\xi) = \log(1 + e^{-\xi})$ in section 5.4.

5.2 General SAFE

In this section, we derive a Safe Feature Elimination method for eliminating features in an l_1- regularized convex problem.

5.2.1 Dual Problem

The first step is to devise the dual of problem (5.1), which is

$$\mathcal{D}(\lambda) \;:\; \phi(\lambda) = \max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) \;:\; \boldsymbol{\theta}^T \boldsymbol{b} = 0, \;\; |\boldsymbol{\theta}^T \boldsymbol{x}_k| \leq \lambda, \;\; k = 1, \ldots, n, \tag{5.2}$$

where

$$G(\boldsymbol{\theta}) := \boldsymbol{c}^T \boldsymbol{\theta} - \sum_{i=1}^{m} f^*(\theta_i) \tag{5.3}$$

with $f^*(\vartheta) = \max_\xi \xi\vartheta - f(\xi)$ the conjugate of the loss function $f(\zeta)$, and $\boldsymbol{x}_k$ the k-th column or feature of the feature matrix $\boldsymbol{X} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)^T \in \mathbb{R}^{m \times n}$. $G(\boldsymbol{\theta})$ is the dual function, which is, by construction, concave. We assume that strong duality holds and primal and dual optimal points are attained. Due to the optimality conditions for the problem (see [5]), constraints for which $|\boldsymbol{\theta}^T \boldsymbol{x}_k| < \lambda$ at optimum correspond to a zero element in the primal variable: $\boldsymbol{w}^\star(k) = 0$, i.e.

$$\left|\boldsymbol{\theta}^{\star T} \boldsymbol{x}_k\right| < \lambda \Rightarrow \boldsymbol{w}^\star(k) = 0. \tag{5.4}$$

5.2.2 Optimality set Θ

For simplicity, we consider only the set $\Theta := \{\boldsymbol{\theta} \,|\, G(\boldsymbol{\theta}) \geq \gamma\}$ which contains $\boldsymbol{\theta}^\star$ the dual optimal point of $\mathcal{D}(\lambda)$. One way to get a lower bound γ is to find a dual point $\boldsymbol{\theta}_s$ that is feasible for the dual problem $\mathcal{D}(\lambda)$, and then set $\gamma = G(\boldsymbol{\theta}_s)$.

To obtain a dual feasible point, we can solve the problem for a higher value $\lambda_0 \geq \lambda$ of the penalty parameter. (In the specific case examined below, we will see how to set λ_0 so that the vector $\boldsymbol{w}_0^\star = \boldsymbol{0}$ at optimum.) This provides a dual point $\boldsymbol{\theta}_0^\star$ that is feasible for $\mathcal{D}(\lambda_0)$, which satisfies $\lambda_0 = \|\boldsymbol{X}\boldsymbol{\theta}_0\|_\infty$. In turn, $\boldsymbol{\theta}_0^\star$ can be scaled so as to become feasible for $\mathcal{D}(\lambda)$. Precisely, we set $\boldsymbol{\theta}_s = s\boldsymbol{\theta}_0$, with $\|\boldsymbol{X}\boldsymbol{\theta}_s\|_\infty \leq \lambda$ equivalent to $|s| \leq \lambda/\lambda_0$. In order to find the best possible scaling factor s, we solve the one-dimensional, convex problem

$$\gamma(\lambda) := \max_s G(s\boldsymbol{\theta}_0) \;\; : |s| \leq \frac{\lambda}{\lambda_0}. \tag{5.5}$$

Under mild conditions on the loss function f, the above problem can be solved by bisection in $O(m)$ time. By construction, $\gamma(\lambda)$ is a lower bound on $\phi(\lambda)$. We can generate an initial point $\boldsymbol{\theta}_0^\star$ by solving $\mathcal{P}(\lambda_0)$ with $\boldsymbol{w}_0 = \boldsymbol{0}$. We get

$$\min_{v_0} \sum_{i=1}^{m} f(b_i v_0 + c_i) = \min_{v_0} \max_{\theta_0} \boldsymbol{\theta}_0^T (\boldsymbol{b} v_0 + \boldsymbol{c}) - \sum_{i=1}^{m} f^* \left(\boldsymbol{\theta}_0(i)\right) = \max_{\boldsymbol{\theta}_0 \,:\, \boldsymbol{b}^T \boldsymbol{\theta}_0 = 0} G(\boldsymbol{\theta}_0).$$

Solving the one-dimensional problem above can be often done in closed-form, or by bisection, in $O(m)$. Choosing $\boldsymbol{\theta}_0^\star$ to be any optimal for the corresponding dual problem (the one on

the right-hand side) generates a point that is dual feasible for it, that is, $G(\boldsymbol{\theta}_0^\star)$ is finite, and $\boldsymbol{b}^T\boldsymbol{\theta}_0 = 0$.

The point $\boldsymbol{\theta}_0^\star$ satisfies all the constraints of problem $\mathcal{D}(\lambda)$, except perhaps for the constraint $\|\boldsymbol{X}\boldsymbol{\theta}\|_\infty \leq \lambda$, i.e. $\|\boldsymbol{X}\boldsymbol{\theta}_0^\star\|_\infty > \lambda$. Hence, if $\lambda \geq \lambda_0 := \|\boldsymbol{X}\boldsymbol{\theta}_0^\star\|_\infty$, then $\boldsymbol{\theta}_0^\star$ is dual optimal for $\mathcal{D}(\lambda)$ and by the optimality condition (5.4) we have $\boldsymbol{w}^\star = 0$. Note that, since $\boldsymbol{\theta}_0^\star$ may not be uniquely defined, λ_0 may not necessarily be the smallest value for which $\boldsymbol{w}^\star = \mathbf{0}$ is optimal for the primal problem.

5.2.3 SAFE method

Assume that a lower bound γ on the optimal value of the learning problem $\phi(\lambda)$ is known: $\gamma \leq \phi(\lambda)$. (Without loss of generality, we can assume that $0 \leq \gamma \leq \sum_{i=1}^m f(c_i)$). The test

$$\lambda > \max(P(\gamma, \boldsymbol{x}_k), P(\gamma, -\boldsymbol{x}_k)),$$

allows to eliminate the k-th feature from the feature matrix $\boldsymbol{X}$, where $P(\gamma, \boldsymbol{x})$ is the optimal value of a convex optimization problem with two constraints:

$$P(\gamma, \boldsymbol{x}) := \max_{\boldsymbol{\theta}} \boldsymbol{\theta}^T\boldsymbol{x} \ : \ G(\boldsymbol{\theta}) \geq \gamma, \ \ \boldsymbol{\theta}^T\boldsymbol{b} = \mathbf{0}. \tag{5.6}$$

Since $P(\boldsymbol{\gamma x})$ decreases when γ increases, the closer $\phi(\lambda)$ is to its lower bound γ, the more aggressive (accurate) our test is.

By construction, the dual function G is decomposable as a sum of functions of one variable only. This particular structure allows to solve problem (5.6) very efficiently, using for example interior-point methods, for a large class of loss functions f. Alternatively, we can express the problem in dual form as a convex optimization problem with two scalar variables:

$$P(\gamma, \boldsymbol{x}) = \min_{\mu>0,\, \nu} -\gamma\mu + \mu \sum_{i=1}^m f\left(\frac{\boldsymbol{x}(i) + \mu c_i + \nu b_i}{\mu}\right). \tag{5.7}$$

Note that the expression above involves the perspective of the function f, which is convex (see [5]). For many loss functions f, the above problem can be efficiently solved using a variety of methods for convex optimization, in (close to) $O(m)$ time. We can also set the variable $\nu = 0$, leading to a simple bisection problem over μ. This amounts to ignore the constraint $\boldsymbol{\theta}^T\boldsymbol{b} = \mathbf{0}$ in the definition of $P(\gamma, \boldsymbol{x})$, resulting in a more conservative test. More generally, any pair (μ, ν) with $\mu > 0$ generates an upper bound on $P(\gamma, \boldsymbol{x})$, which in turn corresponds to a valid, perhaps conservative, test.

5.3 SAFE for Sparse Support Vector Machine

We turn to the sparse support vector machine classification problem:

$$\mathcal{P}_{\rm hi}(\lambda) \;:\; \phi(\lambda) := \min_{\boldsymbol{w},v} \sum_{i=1}^{m}(1 - y_i(\boldsymbol{z}_i^T\boldsymbol{w} + v))_+ + \lambda\|\boldsymbol{w}\|_1, \tag{5.8}$$

where $\boldsymbol{z}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ are the data points, and $\boldsymbol{y} \in \{-1, 1\}^m$ is the label vector. The above is a special case of the generic problem (5.1), where $f(\zeta) := (1 - \xi)_+$ is the hinge loss, $\boldsymbol{b} = \boldsymbol{y}$, $\boldsymbol{c} = \boldsymbol{0}$, and the feature matrix $\boldsymbol{X}$ is given by $\boldsymbol{X} = [y_1\boldsymbol{z}_1, \ldots, y_m\boldsymbol{z}_m]^T$, so that $\boldsymbol{x}_k = [y_1\boldsymbol{z}_1(k), \ldots, y_m\boldsymbol{z}_m(k)]^T$.

We denote by $\mathcal{I}_+, \mathcal{I}_-$ the set of indicies corresponding to the positive and negative classes, respectively, and denote by $m_\pm = |\mathcal{I}_\pm|$ the associated cardinalities. We define $\underline{m} := \min(m_+, m_-)$. Finally, for a generic data vector $\boldsymbol{x}$, we set $\boldsymbol{x}^\pm = \boldsymbol{x}_{\mathcal{I}_\pm} \in \mathbb{R}^{m_\pm}$, $k = 1, \ldots, n$, the vectors corresponding to each one of the classes.

The dual problem takes the form

$$\mathcal{D}_{hi}(\lambda) \;:\; \phi(\lambda) := \max_{\boldsymbol{\theta}} G_{\rm hi}(\boldsymbol{\theta}) \;:\; -\boldsymbol{1} \le \boldsymbol{\theta} \le 0,\;\; \boldsymbol{\theta}^T\boldsymbol{y} = 0,\;\; |\boldsymbol{\theta}^T\boldsymbol{x}_k| \le \lambda,\;\; k = 1, \ldots, n. \tag{5.9}$$

with $G_{\rm hi}(\boldsymbol{\theta}) = \boldsymbol{1}^T\boldsymbol{\theta}$.

5.3.1 Test, γ given

Let γ be a lower bound on $\phi(\lambda)$. The optimal value obtained upon setting $w = 0$ in (5.8) is given by

$$\min_v \sum_{i=1}^{m}(1 - y_i v)_+ = 2\min(m_+, m_-) := \gamma_{\max}. \tag{5.10}$$

Hence, without loss of generality, we may assume $0 \le \gamma \le \gamma_{\max}$.

The feature elimination test hinges on the quantity

$$\begin{array}{rcl} P_{\rm hi}(\gamma, \boldsymbol{x}) & = & \max_{\boldsymbol{\theta}}\; \boldsymbol{\theta}^T x \;:\; \boldsymbol{1}^T\boldsymbol{\theta} \ge \gamma,\;\; \boldsymbol{\theta}^T\boldsymbol{y} = 0,\;\; -\boldsymbol{1} \le \boldsymbol{\theta} \le 0 \\ & = & \min_{\mu>0,\,\nu}\; -\gamma\mu + \mu\sum_{i=1}^{m} f_{\rm hi}\left(\frac{x_i - \nu y_i}{\mu}\right) \\ & = & \min_{\mu>0,\,\nu}\; -\gamma\mu + \sum_{i=1}^{m}(\mu + \nu y_i - x_i)_+. \end{array} \tag{5.11}$$

In appendix D.1, we show that for any $\boldsymbol{x}$, the quantity $P(\gamma, \boldsymbol{x})$ is finite if and only if $0 \le \gamma \le \gamma_{\max}$, and can be computed in $O(m \log m)$ computations, or less with sparse data, via a closed-form expression. That expression is simpler to state for $P_{\rm hi}(\gamma, -\boldsymbol{x})$:

$$\begin{array}{rcl} P_{\rm hi}(\gamma, -\boldsymbol{x}) & = & \sum_{j=1}^{\lfloor\gamma/2\rfloor} \bar{x}_j - (\frac{\gamma}{2} - \lfloor\frac{\gamma}{2}\rfloor)(\bar{x}_{\lfloor\gamma/2\rfloor+1})_+ \\ & & + \sum_{j=\lfloor\gamma/2\rfloor+1}^{\underline{m}}(\bar{x}_j)_+,\;\; 0 \le \gamma \le \gamma_{\max} = 2\underline{m}, \\ & & \bar{x}_j := \boldsymbol{x}_{[j]}^+ + \boldsymbol{x}_{[j]}^-,\;\; j = 1, \ldots, \underline{m}, \end{array}$$

with $\boldsymbol{x}_{[j]}$ the j-th largest element in a vector $\boldsymbol{x}$, and with the convention that a sum over an empty index set is zero. Note that in particular, since $\gamma_{\max} = 2\underline{m}$:

$$P_{\text{hi}}(\gamma_{\max}, -\boldsymbol{x}) = \sum_{i=1}^{\underline{m}} (\boldsymbol{x}^+_{[j]} + \boldsymbol{x}^-_{[j]}).$$

5.3.2 SAFE-SVM theorem

Following the construction proposed in section 5.2.2 for the generic case, we select $\gamma = G_{\text{hi}}(\boldsymbol{\theta})$, where the point $\boldsymbol{\theta}$ is feasible for (5.9), and can found by the scaling method outlined in section 5.2.2, as follows. The method starts with the assumption that there is a value $\lambda_0 \geq \lambda$ for which we know the optimal value γ_0 of $\mathcal{P}_{\text{hi}}(\lambda_0)$.

Specific choices for λ_0, γ_0. Let us first detail how we can find such values λ_0, γ_0.

We can set a value λ_0 such that $\lambda > \lambda_0$ ensures that $\boldsymbol{w} = 0$ is optimal for the primal problem (5.8). The value that results in the least conservative test is $\lambda_0 = \lambda_{\max}$, where $\lambda_{\max}$ is the smallest value of λ above which $\boldsymbol{w} = \boldsymbol{0}$ is optimal:

$$\lambda_{\max} := \min_{\boldsymbol{\theta}} \ \|\boldsymbol{X}\boldsymbol{\theta}\|_\infty \ : \ -\boldsymbol{\theta}^T\boldsymbol{1} \geq \gamma_{\max}, \ \ \boldsymbol{\theta}^T\boldsymbol{y} = 0, \ \ -\boldsymbol{1} \leq \boldsymbol{\theta} \leq \boldsymbol{0}. \tag{5.12}$$

Since $\lambda_{\max}$ may be relatively expensive to compute, we can settle for an upper bound $\overline{\lambda}_{\max}$ on $\lambda_{\max}$. One choice for $\overline{\lambda}_{\max}$ is based on the test derived in the previous section: we ask that it passes for all the features when $\lambda = \overline{\lambda}_{\max}$ and $\gamma = \gamma_{\max}$. That is, we set

$$\begin{array}{rcl} \overline{\lambda}_{\max} & = & \max_{1\leq k\leq n} \ \max\left(P_{\text{hi}}(\gamma_{\max}, \boldsymbol{x}_k), P_{\text{hi}}(\gamma_{\max}, -\boldsymbol{x}_k)\right) \\ & = & \max_{1\leq k\leq n} \ \max\left(\sum_{i=1}^{\underline{m}} (\boldsymbol{x}_k^+)_{[j]} + (\boldsymbol{x}_k^-)_{[j]}, \sum_{i=1}^{\underline{m}} (-\boldsymbol{x}_k^+)_{[j]} + (-\boldsymbol{x}_k^-)_{[j]}\right). \end{array} \tag{5.13}$$

By construction, we have $\overline{\lambda}_{\max} \geq \lambda_{\max}$, in fact:

$$\begin{array}{rcl} \overline{\lambda}_{\max} & = & \max_{1\leq k\leq n} \ \max_\theta \ |\boldsymbol{x}_k^T\boldsymbol{\theta}| \ : \ -\boldsymbol{1}^T\boldsymbol{\theta} \geq \gamma_{\max}, \ \ \boldsymbol{\theta}^T\boldsymbol{y} = 0, \ \ -\boldsymbol{1} \leq \theta \leq 0 \\ & = & \max_{\boldsymbol{\theta}} \ \|\boldsymbol{X}\boldsymbol{\theta}\|_\infty \ : \ -\boldsymbol{1}^T\boldsymbol{\theta} \geq \gamma_{\max}, \ \ \boldsymbol{\theta}^T\boldsymbol{y} = 0, \ \ -\boldsymbol{1} \leq \boldsymbol{\theta} \leq 0, \end{array}$$

The two values $\lambda_{\max}, \overline{\lambda}_{\max}$ coincide if the feasible set is a singleton, that is, when $m_+ = m_-$. On the whole interval $\lambda_0 \in [\lambda_{\max}, \overline{\lambda}_{\max}]$, the optimal value of problem $\mathcal{P}_{\text{hi}}(\lambda_0)$ is $\gamma_{\max}$.

Dual scaling. The remainder of our analysis applies to any value λ_0 for which we know the optimal value $\gamma_0 \in [0, \gamma_{\max}]$ of the problem $\mathcal{P}_{\text{hi}}(\lambda_0)$.

Let $\boldsymbol{\theta}_0$ be a corresponding optimal dual point (as seen shortly, the value of $\boldsymbol{\theta}_0$ is irrelevant, as we will only need to know $\gamma_0 = \boldsymbol{1}^T\boldsymbol{\theta}_0$). We now scale the point $\boldsymbol{\theta}_0$ to make it feasible for $\mathcal{P}_{\text{hi}}(\lambda)$, where λ ($0 \leq \lambda \leq \lambda_0$) is given. The scaled dual point is obtained as $\boldsymbol{\theta} = s\boldsymbol{\theta}_0$, with

s solution to (5.5). We obtain the optimal scaling $s = \lambda/\lambda_0$, and since $\gamma_0 = -\mathbf{1}^T\boldsymbol{\theta}_0$, the corresponding bound is

$$\gamma(\lambda) = \mathbf{1}^T(s\boldsymbol{\theta}_0) = s\gamma_0 = \gamma_0\frac{\lambda}{\lambda_0}.$$

Our test takes the form

$$\lambda > \max\left(P_{\text{hi}}(\gamma(\lambda), \boldsymbol{x}), P_{\text{hi}}(\gamma(\lambda), -\boldsymbol{x})\right).$$

Let us look at the condition $\lambda > P_{\text{hi}}(\gamma(\lambda), -\boldsymbol{x})$:

$$\exists\, \mu \geq 0,\ \nu\ :\ \lambda > -\gamma(\lambda)\mu + \sum_{i=1}^{m}(\mu + \nu y_i + x_i)_+,$$

which is equivalent to:

$$\lambda > \min_{\mu\geq 0,\nu} \frac{\sum_{i=1}^{m}(\mu + \nu y_i + x_i)_+}{1 + (\gamma_0/\lambda_0)\mu}.$$

The problem of minimizing the above objective function over variable ν has a closed-form solution. In appendix D.2, we show that for any vectors $\boldsymbol{x}^{\pm} \in \mathbb{R}^{m_\pm}$, we have

$$\Phi(\boldsymbol{x}^+, \boldsymbol{x}^-) := \min_{\nu} \sum_{i=1}^{m_+}(x_i^+ + \nu)_+ + \sum_{i=1}^{m_-}(x_i^- - \nu)_+ = \sum_{i=1}^{\underline{m}}(x_{[i]}^+ + x_{[i]}^-)_+,$$

with $\boldsymbol{x}_{[j]}$ the j-th largest element in a vector $\boldsymbol{x}$. Thus, the test becomes

$$\lambda > \min_{\mu\geq 0} \frac{\sum_{i=1}^{\underline{m}}(2\mu + \boldsymbol{x}_{[i]}^+ + \boldsymbol{x}_{[i]}^-)_+}{1 + (\gamma_0/\lambda_0)\mu}.$$

Setting $\kappa = \lambda_0/(\lambda_0 + \gamma_0\mu)$, we obtain the following formulation for our test:

$$\lambda > \min_{0\leq\kappa\leq 1} \sum_{i=1}^{\underline{m}}((1-\kappa)\frac{2\lambda_0}{\gamma_0} + \kappa(\boldsymbol{x}_{[i]}^+ + \boldsymbol{x}_{[i]}^-))_+ = \frac{2\lambda_0}{\gamma_0}G(\frac{\gamma_0}{2\lambda_0}\overline{\boldsymbol{x}}), \tag{5.14}$$

where $\overline{\boldsymbol{x}}_i := \boldsymbol{x}_{[i]}^+ + \boldsymbol{x}_{[i]}^-$, $i = 1, \ldots, \underline{m}$, and for $\boldsymbol{z} \in \mathbb{R}^m$, we define

$$G(z) := \min_{0\leq\kappa\leq 1} \sum_{i=1}^{m}(1 - \kappa + \kappa z_i)_+.$$

We show in appendix D.3 that $G(\boldsymbol{z})$ admits a closed-form expression, which can be computed in $O(d\log d)$, where d is the number of non-zero elements in vector $\boldsymbol{z}$. By construction, the test removes all the features if we set $\lambda_0 = \lambda_{\max}$, $\gamma_0 = \gamma_{\max}$, and when $\lambda > \lambda_{\max}$.

Theorem 5.3.1 (SAFE-SVM) *Consider the SVM problem $\mathcal{P}_{\rm hi}(\lambda)$ in (5.8). Denote by $\boldsymbol{x}_k$ the k-th row of the matrix $[y_1 z_1, \ldots, y_m z_m]$, and let $\mathcal{I}_\pm := \{i \ : \ y_i = \pm 1\}$, $m_\pm := |\mathcal{I}_\pm|$, $\underline{m} := \min(m_+, m_-)$, and $\gamma_{\max} := 2\underline{m}$. Let $\lambda_0 \geq \lambda$ be a value for which the optimal value $\gamma_0 \in [0, \gamma_{\max}]$ of $\mathcal{P}_{\rm sq}(\lambda_0)$ is known. The following condition allows to remove the k-th feature vector $\boldsymbol{x}_k$:*

$$\lambda > \frac{2\lambda_0}{\gamma_0} \max\left(G(\frac{\gamma_0}{2\lambda_0}\overline{x}_k), G(\frac{\gamma_0}{2\lambda_0}\underline{x}_k) \right), \tag{5.15}$$

where $(\overline{\boldsymbol{x}}_k)_i := (\boldsymbol{x}_k)^+_{[i]} + (\boldsymbol{x}_k)^-_{[i]}$, $(\underline{\boldsymbol{x}}_k)_i := (-\boldsymbol{x}_k)^+_{[i]} + (-\boldsymbol{x}_k)^-_{[i]}$, $i = 1, \ldots, \underline{m}$, and for $z \in \mathbb{R}^m$:

$$G(z) = \min_z \ \frac{1}{1-z} \sum_{i=1}^{p} (z_i - z)_+ \ \ : \ \ z \in \{-\infty, 0, (z_j)_{j \,:\, z_j < 0}\}$$

A specific choice for λ_0 is $\overline{\lambda}_{\max}$ given by (5.13), with corresponding optimal value $\gamma_0 = \gamma_{\max}$. ■

5.4 SAFE for Sparse Logistic Regression

We now consider the sparse logistic regression problem:

$$\mathcal{P}_{\rm lo}(\lambda) \ \ : \ \ \phi(\lambda) := \min_{\boldsymbol{w}, v} \ \sum_{i=1}^{m} \log\left(1 + \exp(-y_i(\boldsymbol{z}_i^T \boldsymbol{w} + v))\right) + \lambda \|\boldsymbol{w}\|_1, \tag{5.16}$$

with the same notation as in section 5.3. The dual problem takes the form

$$\begin{aligned} \mathcal{D}_{\rm lo}(\lambda) \ \ : \ \ \phi(\lambda) : \max_{\boldsymbol{\theta}} \ & \sum_{i=1}^{m} \left(\theta_i \log(-\theta_i) - (1+\theta_i)^T \log(1+\theta_i)\right) \ \ : \\ & -\mathbf{1} \leq \boldsymbol{\theta} \leq 0, \ \ \boldsymbol{\theta}^T \boldsymbol{y} = 0, \\ & |\boldsymbol{\theta}^T \boldsymbol{x}_k| \leq \lambda, \ \ k = 1, \ldots, n. \end{aligned} \tag{5.17}$$

5.4.1 Test, γ given

Assume that we know a lower bound on the problem, $\gamma \leq \phi(\lambda)$. Since $0 \leq \phi(\lambda) \leq m \log 2$, we may assume that $\gamma \in [0, m \log 2]$ without loss of generality. We proceed to formulate problem (5.7). For given $\boldsymbol{x} \in \mathbb{R}^m$, and $\gamma \in \mathbb{R}$, we have

$$P_{\log}(\gamma, \boldsymbol{x}) \ \ = \ \ \min_{\mu > 0, \, \nu} \ -\gamma\mu + \mu \textstyle\sum_{i=1}^{m} f_{\log}\left(\frac{x_i + y_i \nu}{\mu}\right), \tag{5.18}$$

which can be computed in $O(m)$ by two-dimensional search, or by the dual interior-point method described in appendix. (As mentioned before, an alternative, resulting in a more

conservative test, is to fix ν, for example $\nu = 0$.) Our test to eliminate the k-th feature takes the form

$$\lambda > T_{\log}(\gamma, \boldsymbol{x}_k) := \max(P_{\log}(\gamma, \boldsymbol{x}_k), P_{\log}(\gamma, -\boldsymbol{x}_k)).$$

If γ is known, the complexity of running this test through all the features is $O(nm)$. (In fact, the terms in the objective function that correspond to zero elements of x are of two types, involving $f_{\log}(\pm\nu/\mu)$. This means that the effective dimension of problem (5.18) is the cardinality d of vector x, which in many applications is much smaller than m.)

5.4.2 Obtaining a dual feasible point

We can construct dual feasible points based on scaling one obtained by choice of a primal point (classifier weight) $\boldsymbol{w}_0$. This in turn leads to other possible choices for the bound γ.

For $\boldsymbol{w}_0 \in \mathbb{R}^n$ given, we solve the one-dimensional, convex problem

$$v_0 := \arg\min_{\boldsymbol{b}} \sum_{i=1}^{m} f_{\log}(y_i \boldsymbol{x}_i^T \boldsymbol{w}_0 + y_i \boldsymbol{b}).$$

This problem can be solved by bisection in $O(m)$ time [27]. At optimum, the derivative of the objective is zero, hence $\boldsymbol{y}^T \boldsymbol{\theta}_0 = 0$, where

$$\boldsymbol{\theta}_0(i) := -\frac{1}{1 + \exp(y_i \boldsymbol{x}_i^T \boldsymbol{w}_0 + y_i v_0)}, \quad i = 1, \ldots, m.$$

Now apply the scaling method seen before, and set γ by solving problem (5.5).

5.4.3 A specific example of a dual point

A convenient, specific choice in the above construction is to set $\boldsymbol{w}_0 = \boldsymbol{0}$. Then, the intercept v_0 can be explicitly computed, as $v_0 = \log(m_+/m_-)$, where $m_\pm = |\{i \ : \ y_i = \pm 1\}|$ are the class cardinalities. The corresponding dual point $\boldsymbol{\theta}_0$ is

$$\boldsymbol{\theta}_0(i) = \begin{cases} -\frac{m_-}{m} & (y_i = +1) \\ -\frac{m_+}{m} & (y_i = -1), \end{cases} \qquad i = 1, \ldots, m. \tag{5.19}$$

The corresponding value of λ_0 is (see [27]):

$$\lambda_0 := \|\boldsymbol{X}^T \boldsymbol{\theta}_0\|_\infty = \max_{1 \le k \le n} |\boldsymbol{\theta}_0^T \boldsymbol{x}_k|.$$

We now compute $\gamma(\lambda)$ by solving problem (5.5), which expresses as

$$\gamma(\lambda) = \max_{|s| \le \lambda/\lambda_0} G_{\log}(s\boldsymbol{\theta}_0) = \max_{|s| \le \lambda/\lambda_0} \ -m_+ f_{\log}^*(-s\frac{m_-}{m}) - m_- f_{\log}^*(-s\frac{m_+}{m}). \tag{5.20}$$

The above can be solved analytically: it can be shown that $s = \lambda/\lambda_0$ is optimal.

5.4.4 Solving the bisection problem

In this section, we are given $\boldsymbol{c} \in \mathbb{R}^m$, $\gamma \in (0, m \log 2)$, and we consider the problem

$$F^* := \min_{\mu>0} \; F(\mu) \;\; := \;\; -\gamma\mu + \mu \sum_{i=1}^m f_{\log}(\boldsymbol{c}(i)/\mu). \tag{5.21}$$

Problem (5.21) corresponds to the problem (5.18), with ν set to a fixed value, and $\boldsymbol{c}(i) = y_i x_i$, $i = 1, \ldots, m$. We assume that $\boldsymbol{c}(i) \neq 0$ for every i, and that $\kappa := m \log 2 - \gamma > 0$. Observe that $F^* \leq F_0 := \lim_{\mu \to 0^+} \; F(\mu) = \mathbf{1}^T \boldsymbol{c}_+$, where $\boldsymbol{c}_+$ is the positive part of vector $\boldsymbol{c}$.

To solve this problem via bisection, we initialize the interval of confidence to be $[0, \mu_u]$, with μ_u set as follows. Using the inequality $\log(1 + e^{-x}) \geq \log 2 - (1/2)x_+$, which is valid for every x, we obtain that for every $\mu > 0$:

$$F(\mu) \geq -\gamma\mu + \mu \sum_{i=1}^m \left(\log 2 - \frac{(\boldsymbol{c}(i))_+}{2\mu} \right) = \kappa\mu - \frac{1}{2} \mathbf{1}^T \boldsymbol{c}_+.$$

We can now identify a value μ_u such that for every $\mu \geq \mu_u$, we have $F(\mu) \geq F_0$: it suffices to ensure $\kappa\mu - (1/2)\mathbf{1}^T \boldsymbol{c}_+ \geq F_0$, that is,

$$\mu \geq \mu_u := \frac{(1/2)\mathbf{1}^T \boldsymbol{c}_+ + F_0}{\kappa} = \frac{3}{2} \frac{\mathbf{1}^T \boldsymbol{c}_+}{m \log 2 - \gamma}.$$

5.4.5 Algorithm summary

An algorithm to check if a given feature can be removed from a sparse logistic regression problem works as follows.

Given: λ, k $(1 \leq k \leq n)$, $f_{\log}(x) = \log(1+e^{-x})$, $f^*_{\log}(\vartheta) = (-\vartheta)\log(-\vartheta) + (\vartheta+1)\log(\vartheta+1)$.

1. Set $\lambda_0 = \max_{1 \leq k \leq n} |\boldsymbol{\theta}_0^T \boldsymbol{x}_k|$, where $\boldsymbol{\theta}_0(i) = -m_-/m$ $(y_i = +1)$, $\boldsymbol{\theta}_0(i) = -m_+/m$ $(y_i = -1)$, $i = 1, \ldots, m$.

2. Set
$$\gamma(\lambda) := -m_+ f^*_{\log}(-\frac{\lambda}{\lambda_0}\frac{m_-}{m}) - m_- f^*_{\log}(-\frac{\lambda}{\lambda_0}\frac{m_+}{m}).$$

3. Solve via bisection a pair of one-dimensional convex optimization problems
$$P_\epsilon = \min_{\mu>0} \; -\gamma(\lambda)\mu + \mu \sum_{i=1}^m f_{\log}(\epsilon y_i (\boldsymbol{x}_k)_i/\mu) \;\; (\epsilon = \pm 1),$$
each with initial interval $[0, \mu_u]$, with
$$\mu_u = \frac{3}{2} \frac{\sum_{i=1}^m (\epsilon y_i (\boldsymbol{x}_k)_i)_+}{m \log 2 - \gamma}.$$

4. If $\lambda > \max(P_+, P_-)$, the k-th feature can be safely removed.

5.5 Conclusion

In this chapter, we have generalized the Safe Feature Elimination method to a class of ℓ_1-Regularized Convex Problems. The steps and concepts used in deriving SAFE-LASSO are adapted in deriving SAFE for the general convex loss function. We have studied the specific case of the hing loss function, and the logistic loss function by using a bound on the optimal solution of the dual problem of the form $G(\theta) \geq \gamma$. We also presented a closed-form solution for the SAFE test in the case of hing loss function and a numerical algorithm for the logistic regression loss function case.

Part II

Application in the Control of Large-Scale Open-Channel Flow Systems

Chapter 6

Control of an Irrigation Canal

6.1 Introduction

With a population of more than six billion people, food production from agriculture must be raised to meet increasing demand. While irrigated agriculture provides 40% of the total food production, it represents 80% of the freshwater consumption worldwide. In summer and drought conditions, efficient management of scarce water resources becomes crucial. The majority of irrigation canals are managed manually, however, with large water losses leading to low water efficiency.

Irrigation canals can be viewed and modeled as delay systems since it takes time for the water released at the upstream end to reach the user located downstream. We thus present an open-loop controller that can deliver water at a given location at a specified time. The development of this controller requires a method for inverting the equations that describe the dynamics of the canal in order to parameterize the controlled input as a function of the desired output. The Saint-Venant equations [51] are widely used to describe water discharge in a canal. Since these equations are not easy to invert, we use a simplified model, called the Hayami model. We use differential flatness to invert the dynamics of the system and to design an open-loop controller.

We experiment with our controller on the Gignac Canal, located northwest of Montpellier, in southern France. Our comprehensive simulations, and real experiments show that it is possible to achieve a desired water flow at the downstream of a canal using the Hayami model as an approximation of the real-system. However, our observations of the measured water flow at the upstream controlled gate made us realize some actuator limitations. For example, deadband in the gate opening and unmodeled disturbances such as friction in the gate-opening mechanism, only allow us to deliver piece-wise constant control inputs. This fact made us investigate a way to compute a controller that respects the actuator limitations. We use the CD-SAFE algorithm presented in chapter 4, to compute such open-loop control for the upstream water flow.

In this chapter, we model the open channel flow system in section 6.2. We present the Saint-Venant equation, a non-linear model for capturing the dynamics of an open channel flow, and the Hayami model, a simplified linear model. In section 6.3, we invert the hayami model to obtain an open loop controller for controlling the downstream water flow by manipulating the upstream water flow. In section 6.4, we present some simulation results using the the software *simulation of irrigation canals* (SIC), which implements a semi-implicit Preissmann scheme to solve the nonlinear Saint-Venant equations for open-channel one-dimensional flow. In section 6.5, we describe the implementation of our open-loop controller for real-time irrigation operations using a *supervision, control, and data acquisition* (SCADA) system with automatic centralized controller. In section 6.6, we model the open channel flow of water by a first order differential equation with delay model, we derive a controller by solving a LASSO problem with the CD-SAFE algorithm.

6.2 Modeling Open Channel Flow

6.2.1 Saint-Venant Equations

The Saint-Venant equations for water discharge in a canal are named after Adhmar Jean-Claude Barr de Saint-Venant, who derived these equations in 1871 in a note to the *Comptes-Rendus de l'Acadmie des Sciences de Paris* [51]. This model assumes one-dimensional flow, with uniform velocity over the cross section of the canal. The effect of boundary friction is accounted for through an empirical law such as the Manning-Strickler friction law [53]. The average canal bed slope is assumed to be small, and the pressure is assumed to be hydrostatic. Under these assumptions, the Saint-Venant equations are given by

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0, \tag{6.1}$$

$$\frac{\partial Q}{\partial t} + \frac{\partial \left(Q^2/A\right)}{\partial x} + gA\frac{\partial H}{\partial x} = gA(S_b - S_f), \tag{6.2}$$

where $A(x,t)$ is the wetted cross-sectional area, $Q(x,t)$ is the water discharge (m^3/s) through the cross section $A(x,t)$, $H(x,t)$ is the water depth, $S_f(x,t) = \frac{Q^2 n^2}{A^2 R^{4/3}}$ is the friction slope, $R(x,t) = \frac{A}{P}$ is the hydraulic radius, $P(x,t)$ is the wetted perimeter, n is the Manning coefficient ($\mathrm{s\text{-}m}^{-1/3}$), S_b is the bed slope, and g is the gravitational acceleration. Equation (6.1) expresses conservation of mass, while (6.2) expresses conservation of momentum.

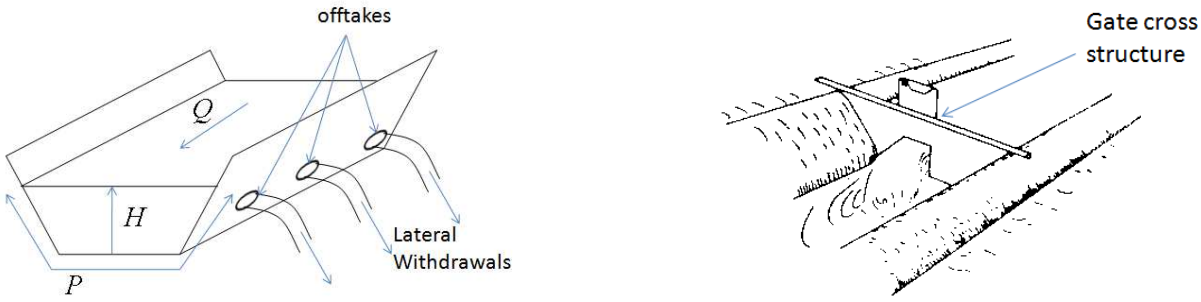
Equations (6.1), (6.2) are completed by boundary conditions at cross structures, such as gates or weirs, where the Saint-Venant equations are not valid. Figure 6.1 illustrates some of the Saint-Venant equations parameters and shows a gate cross structure. The cross structure at the downstream end of the canal can be modeled by a static relation between the water discharge $Q(L,t)$ and the water depth $H(L,t)$ at $x = L$ given by

$$Q(L,t) = W(H(L,t)),, \tag{6.3}$$

where $W(\cdot)$ is derived from hydrostatic laws. For a weir overflow structure, this relation is given by

$$Q(L,t) = C_w\sqrt{2g}L_w\left(H(L,t) - H_w\right)^{3/2},$$

where g is the gravitational acceleration, L_w is the weir length, H_w is the weir elevation, and C_w is the weir discharge coefficient.



(a) (b)

Figure 6.1: Irrigation canal. (a) shows the flow Q, water depth H, and wetted perimeter P. Lateral withdrawals are taken from offtakes. We assume that offtakes are located at the downstream of the canal, and no variables associated with lateral withdrawals are shown in the Saint-Venant equations (6.1) and (6.2). (b) shows a gate cross structure, which can be used to control the water discharge in the canal.

6.2.2 A Simplified Linear Model

A simplified version of the Saint-Venant equations is obtained by neglecting the inertia terms $\frac{\partial Q}{\partial t} + \frac{\partial\left(Q^2/A\right)}{\partial x}$ in the momentum equation (6.2), which leads to the diffusive wave equation [48]. Linearizing the Saint-Venant equations about a nominal water discharge Q_0 and water depth H_0 yields the Hayami equations

$$D_0\frac{\partial^2 q}{\partial x^2} - C_0\frac{\partial q}{\partial x} = \frac{\partial q}{\partial t}, \tag{6.4}$$

$$B_0\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} = 0, \tag{6.5}$$

where $C_0 = C_0(Q_0)$, $D_0 = D_0(Q_0)$ are the nominal wave celerity and diffusivity, which depend on Q_0, and B_0 is the average bed width. The quantities $q(x,t)$ and $h(x,t)$ are the deviations from the nominal water discharge and water depth, respectively. Figure 6.2 illustrates the relevant notation.

The linearized boundary condition at the downstream end $x = L$ is given by

$$q(L,t) = bh(L,t), \tag{6.6}$$

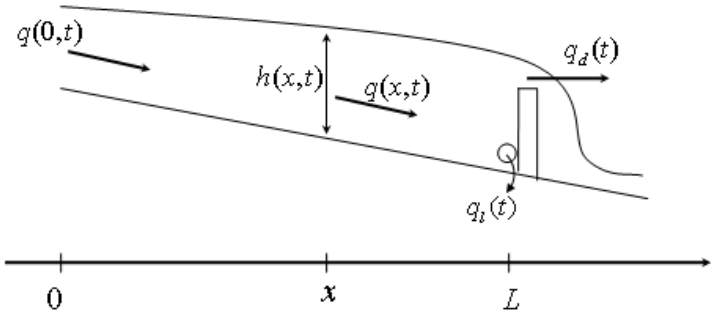


Figure 6.2: Longitudinal schematic profile of a hydraulic canal. A canal is a structure that directs water flow from an upstream location to a downstream location. Water offtakes are assumed to be located at the downstream of the canal. The variables $q(x,t)$, $h(x,t)$, $q_{\mathrm{d}}(t)$, and $q_{\mathrm{l}}(t)$ are the deviations from the nominal values of water discharge, water depth, desired downstream water discharge, and lateral withdrawal, respectively.

where b is the linearization constant equal to $\frac{\partial W}{\partial H}(H_0)$. The value of b depends on the hydraulic structure geometry, including its length, height, and discharge coefficient of the weir. The initial conditions are defined by the deviations from their nominal values, which are assumed to be zero initially, that is,

$$q(x,0) = 0, \tag{6.7}$$
$$h(x,0) = 0. \tag{6.8}$$

6.3 Flatness-based Open-loop Control

6.3.1 Open-loop Control of a Canal Pool

We develop a feedforward controller for water discharge in an open-channel hydraulic system. The system of interest is a hydraulic canal with a cross structure at the downstream end as shown in Figure 6.2. We assume that the desired downstream water discharge $q_{\mathrm{d}}(t)$ is specified in advance, based on scheduled user demands. The control problem consists of determining the upstream water discharge $q(0,t)$ that has to be delivered in order to meet the desired downstream water discharge $q_{\mathrm{d}}(t)$. This inverse problem is an open-loop control problem. Note that by linearization, computing $q(0,t)$ as a function of $q_d(t)$ is equivalent to determining $Q(0,t)$ as a function of $Q_{\mathrm{d}}(t) = Q_0 + q_{\mathrm{d}}(t)$.

The upstream water discharge $q(0,t)$ is the solution of the open-loop control problem defined by the Hayami model equations (6.4), (6.5), initial conditions (6.7), (6.8), and boundary condition (6.6). Differential flatness, as described in appendix F "What is Differential Flatness?", provides a way to solve this open loop control problem in the form of a pa-

rameterization of the input $u(t) = q(0,t)$ as a function of the desired output $y(t) = q_{\mathrm{d}}(t)$. Specifically the controller can be expressed in closed form

$$u(t) = e^{\left(-\frac{\alpha^2}{\beta^2}t - \alpha L)\right)} \left(T_1(t) - \kappa T_2(t) + \frac{B_0}{b} T_3(t) \right), \tag{6.9}$$

where the algebraic equations of T_1, T_2, and T_3 are

$$T_1(t) \triangleq \sum_{i=0}^{\infty} \frac{d^i(e^{\frac{\alpha^2}{\beta^2}t} y(t))}{dt^i} \frac{\beta^{2i} L^{2i}}{(2i)!}, \tag{6.10}$$

$$T_2(t) \triangleq \sum_{i=0}^{\infty} \frac{d^i(e^{\frac{\alpha^2}{\beta^2}t} y(t))}{dt^i} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!}, \tag{6.11}$$

$$T_3(t) \triangleq \sum_{i=0}^{\infty} \frac{d^{i+1}(e^{\frac{\alpha^2}{\beta^2}t} y(t))}{dt^{i+1}} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!}, \tag{6.12}$$

$\alpha \triangleq \frac{C_0}{2D_0}$, $\beta \triangleq \frac{1}{\sqrt{D_0}}$, and $\kappa \triangleq \frac{B_0}{b}\frac{\alpha^2}{\beta^2} - \alpha$.

The convergence of the infinite series (6.10)-(6.12) can be guaranteed when the desired output function $y(t)$ and its derivatives are bounded in a specific sense. More specifically, the sum of the infinite series (6.9) converges when the desired output $y(t) = q_{\mathrm{d}}(t)$ is a Gevrey function of order r lower than 2. A Gevrey function $y(t)$ is defined by the following property. For all non negative n, the n^{th} derivative $y^{(n)}(t)$ of a Gevrey function $y(t)$ of order r has bounded derivatives which satisfy the inequality

$$\sup_{t \in [0,T]} \left| y^{(n)}(t) \right| < m \frac{(n!)^r}{l^n},$$

where m and l are constant positive scalars. See appendix H for more details on deriving the open-loop control expression, proof of convergence, and the numerical assessment of the feed-forward controller.

6.4 Assessment of the Performance of the Method in Simulation

Before field implementation, it is necessary to test the method in simulation. We simulate the controller defined by (6.9), called hereafter the Hayami controller, on the nonlinear Saint-Venant model.

6.4.1 Simulation of Irrigation Canals

The simulations are carried out using the software *simulation of irrigation canals* (SIC) [1], which implements a semi-implicit Preissmann scheme to solve the nonlinear Saint-Venant equations (6.1), (6.2) for open-channel one-dimensional flow [1, 40]. Instead of defining a fictitious canal, we use a realistic geometry corresponding to a stretch of the Gignac canal (see section 6.5 for more details on the Gignac canal) to evaluate the open-loop control in simulation. The considered stretch is 4940 m long, with an average bed slope $S_b = 3.8 \times 10^{-4}$ m/m, an average bed width $B_0 = 2$ m, and Manning coefficient $n = 0.024$ s-m$^{-1/3}$.

6.4.2 Parameter Identification

The simulations are performed on a realistic canal geometry, which is neither prismatic nor uniform. Consequently, it is not possible to express C_0, D_0, and b analytically in terms of the physical parameters such as the canal geometry and water discharge. For this reason, it is necessary to empirically estimate the parameters C_0, D_0, and b of the Hayami model that would best approximate the water discharge governed by the Saint-Venant equations (6.1), (6.2). The identification is done with an upstream water discharge in the form of a step input. The water discharges are monitored at the upstream and downstream positions. The identification is performed by finding the parameter values that minimize the least-squares error between the downstream water discharge computed by the Hayami model and the downstream water discharge simulated by SIC. The identification is performed using data generated by simulating the Saint-Venant equations around a nominal water discharge $Q_0 = 0.400$ m^3/s. The identification leads to the parameters $C_0 = 0.84$ m/s, $D_0 = 634$ m^2/s, and $b = 0.61$ m^2/s.

6.4.3 Desired Water Demand

The water demand curve is approximated from predicted consumption or by information from farmers about their consumption intentions. User consumption requirements at offtakes are usually modeled by a demand curve in the form of a step function. However, depending on the canal model used, this demand may require high values of upstream water discharge. We define the demand curve to be a linear transformation of a Gevrey function of the form $y(t) = q_1 \phi_\sigma(t/T)$, where q_1 and T are constants, and $\phi_\sigma(t)$ is a Gevrey function of order $1 + 1/\sigma$ called the dimensionless bump function. The chosen Gevrey function allows a transition from zero water discharge for $t \leq 0$ to a water discharge equal to q_1 for $t \geq T$. The function $\phi_\sigma(t)$ is illustrated in Figure 6.3 for various values of σ.

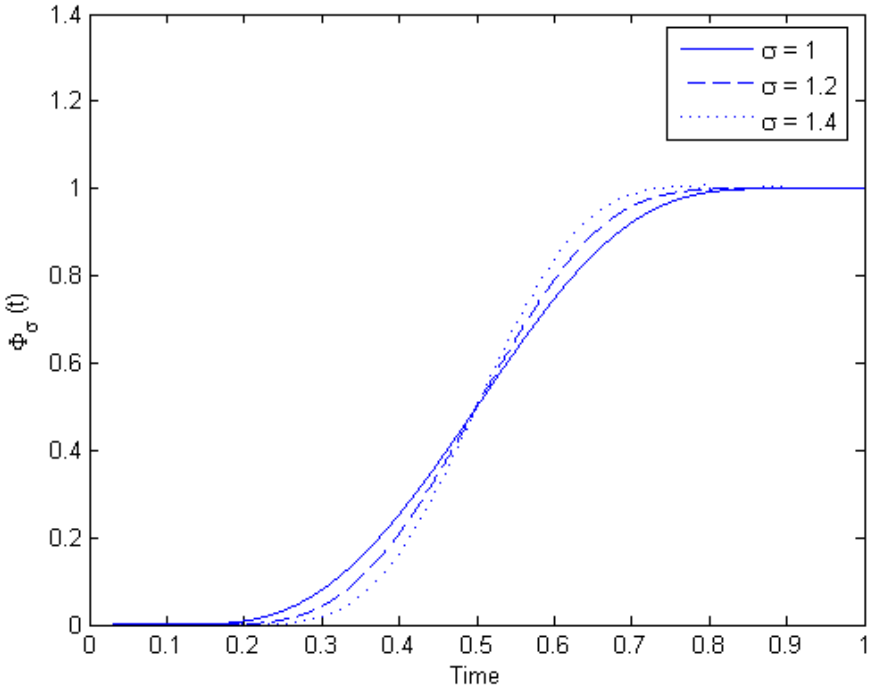


Figure 6.3: Dimensionless bump function. The bump function $\phi_\sigma(t)$ is a Gevrey function of order $1 + 1/\sigma$.

6.4.4 Simulation Results

The Hayami control (6.9) is computed using the estimated parameters C_0, D_0, and b. The downstream water discharge is defined by $y(t) = q_1\phi_\sigma(t/T)$, where $q_1 = 0.1$ m^3/s, $\sigma = 1.4$, and $T = 3$ h. Figure 6.4 shows the control $u(t)$ and the desired output $y(t)$.

The upstream water discharge (6.9) is simulated with SIC to compute the corresponding downstream water discharge. Figure 6.5 shows the downstream water discharge and the desired downstream water discharge.

Although the open-loop control is based on the linear Hayami model, the relative error between the downstream water discharge and the desired downstream water discharge, defined by $e_{\text{rel}}(t) = \left|\frac{q(L,t)-y(t)}{Q_0}\right|$, is less than 0.3%.

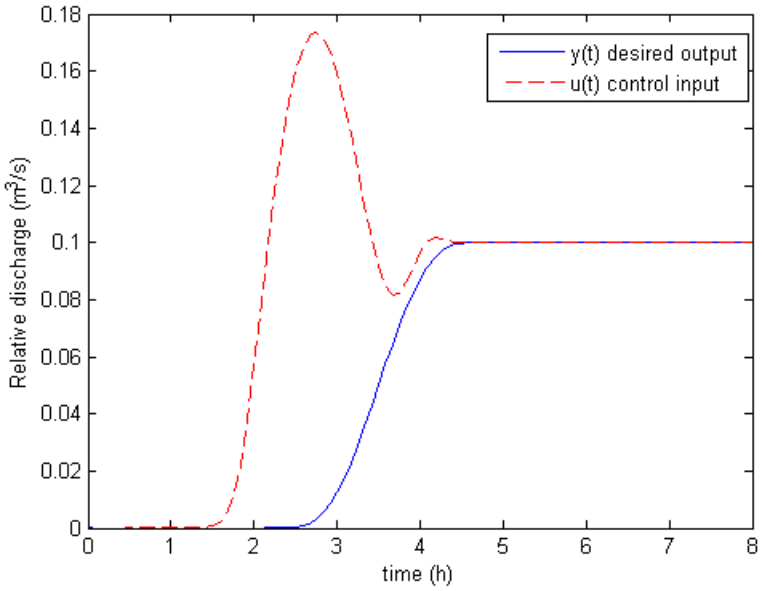


Figure 6.4: Hayami control input signal. The control input $u(t) = q(0, t)$ is computed using the differential flatness method applied to the Hayami model and a desired downstream water discharge $y(t)$.

6.5 Implementation on the Gignac Canal in Southern France

Experiments are performed on the Gignac Canal, located northwest of Montpellier, in southern France. The main canal is 50 km long, with a feeder canal, 8 km long, and two branches on the left and right banks of Hrault river, 27 km and 15 km long, respectively. Figure 6.6 shows a map of the feeder canal with its left and right branches.

As shown in Figure 6.7(a), the canal separates at Partiteur station into two branches, namely, the right branch and the left branch. The canal is equipped at each branch with an automatic regulation gate with position sensors as shown in Figure 6.7(b). Piezo resistive sensors are used to measure the water level by measuring the resistance in the sensor wires. An ultrasonic velocity sensor measures the average water velocity, see Figure 6.7(c). The velocity measurement, water-level measurement, and the geometric properties of the canal at the gate determine the water discharge.

We are interested in controlling the water discharge into the right branch of the canal. The cross section of the right branch is trapezoidal with average bed slope of $S_b = 0.00035$ m/m. The Gignac canal is equipped with a SCADA system, which enables the implementation of controllers. Data from sensors and actuators of the four gates at Partiteur are collected by a

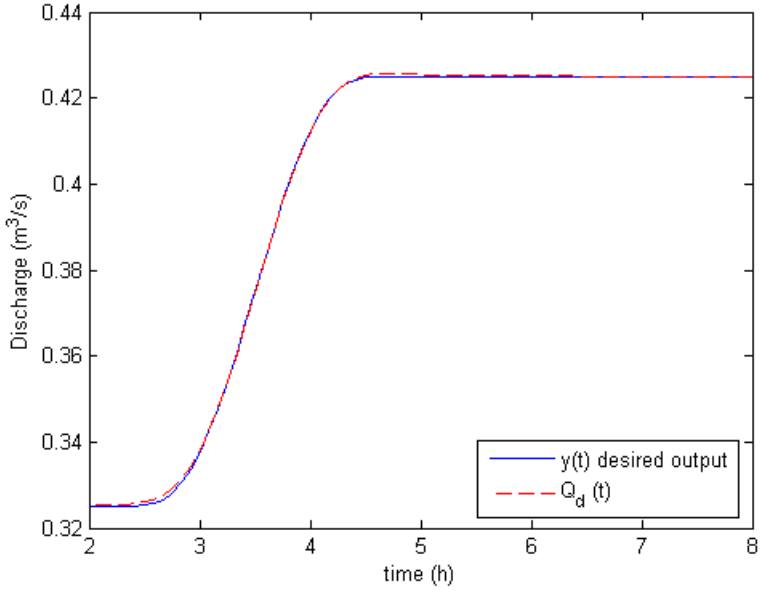


Figure 6.5: Hayami model based control applied to the Saint-Venant model. The downstream water discharge is computed using SIC software. The downstream water discharge $Q_{\mathrm{d}}(t)$ is the output obtained by applying the Hayami control on the full nonlinear model (Saint-Venant model). Although the open-loop control is based on the Hayami model, the relative error between the downstream water discharge and the desired downstream water discharge is less than 0.3%.

control station at the left branch as shown in Figure 6.8. The information is communicated by radio frequency signals every five minutes to a receiving antenna, located in the main control center, a few kilometers away. The data are displayed and saved in a database, while commands to the actuators are sent back to the local controllers at the gates. We use the SCADA system to perform open-loop control in real time. In this experiment, we are interested in controlling the gate at the right branch of the Partiteur station to achieve a desired water discharge five kilometers downstream at Avencq station. The gate opening at Partiteur is computed to deliver the upstream water discharge; for details, see appendix G, "How to Impose a Discharge at a Gate?".

6.5.1 Results Obtained Assuming Constant Lateral Withdrawals

We now estimate the canal parameters for the canal between Partiteur and Avencq. The nominal water discharge is $Q_0 = 0.640$ m^3/s. The identification is done using real sensor data, and leads to the estimates $C_0 = 1.35$ m/s, $D_0 = 893$ m^2/s, and $b = 0.17$ m^2/s. We define a downstream water discharge by $y(t) = q_1 \phi_\sigma(t/T)$, where $q_1 = -0.1$ m^3/s, $\sigma = 1.4$,

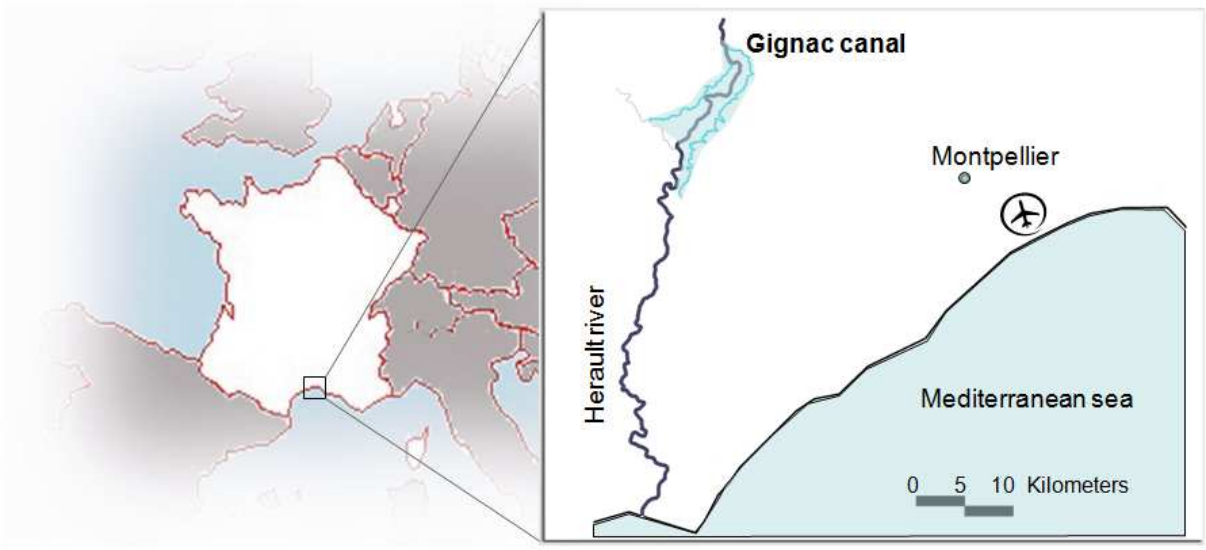


Figure 6.6: Location of Gignac canal in southern France. The canal takes water from the Hrault river, to feed two branches that irrigate a total area of 3000 hectare, where vineyards are located.

and $T = 3.2$ h. The upstream water discharge is computed using (6.9). Figure 6.9 shows the desired downstream water discharge and the upstream water discharge, to be applied at the upstream with the measured discharges at each location, respectively.

The actuator limitations include a deadband in the gate opening of 2.5 cm and unmodeled disturbances such as friction in the gate-opening mechanism. Although the downstream water discharge is tracked well until $t \approx 3.4$ h, a steady-state error of 0.03 m^3/s is evident. This error does not seem to be due to the actuator limitations, but rather to simplifications in the model assumptions, not necessarily satisfied in practice. In particular, we assume constant lateral withdrawals, whereas in reality the lateral withdrawals are driven by gravity. Such gravitational lateral withdrawals vary with the water level, as opposed to lateral withdrawals by pumps, which can be assumed constant.

6.5.2 Modeling the Effects of Gravitational Lateral Withdrawals

The gravitational lateral withdrawals in an offtake is a function of the water level in the canal just upstream of the offtake. Typically, the flow through an underflow offtake is proportional to the square root of the upstream water level. As a first approximation, we linearize this relation, and assume that the offtakes are located at the downstream end of the canal. Then, instead of being constant, the lateral flow is proportional to the downstream water level. The downstream gravitational lateral withdrawals can be seen as a local feedback between the

level and the water discharge. The dynamical model of the canal is then modified as

$$q_{\text{lateral}}(t) = b_1 h(L,t), \tag{6.13}$$

where b_1 is the linearization constant of gravitational lateral withdrawals. We combine the output equation $y(t) = q_{\text{d}}(t) = bh(L,t)$ with the conservation of water discharge at $x = L$, $q(L,T) = q_{\text{lateral}}(t) + q_{\text{d}}(t) = (b + b_1)h(L,t)$, to obtain

$$y(t) = Gq(L,t),$$

where $G = \frac{b}{b+b_1}$. The effect of gravitational lateral withdrawals is thus expressed by a gain factor G, which is less than 1. This gain factor G explains why the released upstream water discharge must be larger than the desired downstream water discharge to account for the gravitational lateral withdrawals. The control (6.9) does not account for the gain factor G, which leads to a steady-state error in the downstream water discharge. Feedback control can provide a solution for this steady-state error by including an integral control component. However, since we are using open-loop control, we need to include the gain-factor effect in this controller to reduce the steady-state error.

The open-loop control is deduced by replacing b with $b_{\text{eq}} = b + b_1$ in (6.9) and the expression of κ , and replacing $y(t)$ by $q(L,t) = G^{-1}y(t)$. The open-loop control for the gravitational lateral withdrawals case is

$$u_{\text{gravitational}}(t) = \frac{1}{G} e^{\left(-\frac{\alpha^2}{\beta^2}t - \alpha L\right)} \left(T_1(t) - \kappa T_2(t) + \frac{B_0}{b_{\text{eq}}} T_3(t) \right). \tag{6.14}$$

In the case of gravitational lateral withdrawals, the open-loop control depends on the parameters G, C_0, D_0, and b_{eq}. These parameters need to be estimated using the same method outlined for the constant lateral withdrawals.

6.5.3 Results Obtained Accounting for Gravitational Lateral Withdrawals

The Saint-Venant equations with the open-loop control input are simulated using SIC software, in order to evaluate the impact of gravitational lateral withdrawals on the output.

Simulation Results

The simulations are carried out on a test canal of length $L = 4940$ m, average bed slope $S_b = 3.8 \times 10^{-4}$, average bed width $B_0 = 2$ m, Manning coefficient $n = 0.024$ s-m$^{-1/3}$, and gravitational lateral withdrawals distributed along its length. Identification is performed about a nominal water discharge $Q_0 = 0.400$ m^3/s. The identification leads to the parameter estimates $G = 0.90$, $C_0 = 0.87$ m/s, $D_0 = 692.34$ m^2/s, and $b_{\text{eq}} = 0.62$ m^2/s for the

gravitational lateral withdrawals, and to $C_0 = 0.84$ m/s, $D_0 = 1100.72$ m^2/s, and $b = 0.75$ m^2/s for the constant lateral withdrawals. The downstream water discharge is defined by $y(t) = q_1\phi_\sigma(t/T)$, where $q_1 = 0.1$ m^3/s, $\sigma = 1.4$, and $T = 8$ h. Figure 6.10 shows the upstream water discharge $u(t)$ and $u_{\text{gravitational}}(t)$ for constant and gravitational lateral withdrawals, respectively.

We notice that the open-loop control that accounts for gravitational lateral withdrawals has a steady-state above the desired output to compensate for the variable withdrawal of water. The upstream water discharge $u(t)$ is simulated with SIC to compute the corresponding downstream water discharge. Figure 6.11 shows the SIC simulation results.

Experimental Results

Estimation of the canal parameters between Partiteur and Avencq is performed as described above for the Hayami model that accounts for gravitational lateral withdrawals. The nominal water discharge is $Q_0 = 0.480$ m^3/s. The identified parameters of the Hayami model are $G = 0.70$, $C_0 = 1.08$ m/s, $D_0 = 444$ m^2/s, and $b = 0.27$ m^2/s. The downstream water discharge is defined by $y(t) = q_1\phi(t/T)$, where $q_1 = 0.1$ m^3/s, $\sigma = 1.4$, and $T = 5$ h. The upstream water discharge $u_{\text{gravitational}}(t)$ is computed using (6.14). Figure 6.12 shows the desired downstream water discharge, the numerical control computed by (6.14), the experimental control achieved by the physical system, and the measured downstream water discharge. The relative error between the measured downstream water discharge and the desired downstream water discharge is less than 9%, despite the fact that the delivered upstream water discharge is perturbed due to actuator limitations.

6.6 Deriving a More Realistic Controller using LASSO

Our experimental results performed on the Gignac canal and shown in Figure 6.12, suggest that the system can be modeled using a simpler model, like a first-order differential equation with delay. Deadband in the gate opening and unmodeled disturbances such as friction in the gate-opening mechanism, limit our control of the water flow at the control gate. Thus, it is of interest to compute a control input $u(t)$, which has few changes in its values. We first examine the performance of a model consisting of a first-order differential equation with delay,

$$K\frac{\partial q(L,t)}{\partial t} + q(L,t) = u(t-\tau), \tag{6.15}$$

$$q(0,t) = 0, \tag{6.16}$$

where $u(t) = q(0,t)$ is the open-loop control representing the upstream discharge at the controller gate and K is a constant. We then experiment using different inputs derived from the model in (6.15).

6.6.1 Capturing the system dynamics

We estimate the parameters K and τ of (6.15) to best approximate the water discharge monitored at the upstream and downstream positions of the canal between Partiteur and Avencq. The identification is performed by finding the parameter values that minimize the least-squares error between the downstream water discharge computed by (6.15) and the downstream water discharge measured at Avencq. We also estimate the parameters C_0, D_0, and b of the Hayami model as described in section 6.4.2. In Figure 6.13, we show the upstream and downstream discharge and the simulated downstream discharge for each model. The identified parameters we used in the simulation are $K = 3138.15$ s, and $\tau = 1198.75$ s for (6.15), and $C_0 = 1.10$ m/s, $D_0 = 1539.09$ m^2/s, and $b = 6.18$ m^2/s for the Hayami model. We compute the fitting error

$$e_f = \frac{\|Q(L,t) - Q_m(L,t)\|_2}{\|Q_m(L,t)\|_2},$$

where $Q_m(L,t)$ is the water discharge measured at the downstream of the canal, and $Q(L,t)$ is the simulated downstream discharge, for each model. We obtain a fitting error $e_f = 1.87\%$ for the Hayami model and a fitting error $e_f = 1.88\%$ for the first-order with delay model. We conclude that the first-order with delay model can approximate the real-system with the same accuracy as the Hayami model. This conclusion agrees with our observation of the behavior of the canal system in the experiment shown in Figure 6.12.

6.6.2 Controller Design

We identify the parameters τ, and K for a realistic canal geometry, which is neither prismatic nor uniform. The identification is done with an upstream water discharge in the form of a step input. The identification is performed using data generated by simulating the Saint-Venant equations around a nominal water discharge $Q_0 = 0.677$ m^3/s. The identification leads to the parameters $\tau = 1203$ s, and $K = 3133.79$ s.

We assume a sampling time $T_s = 30$ s and we design our control input based on the solution of the following optimization problem:

$$\boldsymbol{u} = \arg\min_{\boldsymbol{u}} \|\boldsymbol{u} - \tilde{\boldsymbol{y}}\|_2^2 + \lambda \|\boldsymbol{D}\boldsymbol{u}\|_1, \tag{6.17}$$

with $\boldsymbol{u} \in \mathbb{R}^n$ the control vector, $\tilde{\boldsymbol{y}}(i) = \frac{\partial q(L,t)}{\partial t}\Big|_{t=t(i)+\tau} + q(L, t(i)+\tau) \in \mathbb{R}^n$ and $\boldsymbol{D} \in \mathbb{R}^{(n-1)\times n}$ the first-order difference matrix

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & -1 & \\ & & & & 1 & -1 \end{bmatrix}.$$

When the regularization parameter λ is zero, i.e. $\lambda = 0$, the control input $\boldsymbol{u}$ is simply the solution of the inverse problem of (6.15). The inverse problem computes a discharge $q(0,t)$ such that the downstream discharge is exactly equal to $q(L,t)$. When λ is non-negative, the entries of $(\boldsymbol{D}\boldsymbol{u})_i$ are biased toward the value zero, i.e. $u_i = u_{i+1}$ and thus the control input will have fewer changes in its value as desired.

The downstream water discharge is defined by $y(t) = q_1\phi_\sigma(t/T)$, where $q_1 = -0.1$ m^3/s, $\sigma = 1.4$, and $T = 2.5$ h. The upstream water discharge (6.17) is simulated with different values of the regularization parameter λ. Figure 6.14 shows the control input or the upstream water discharge, the downstream water discharge and the desired downstream water discharge for $\lambda = 0.001\lambda_{\max},\ 0.01\lambda_{\max},\ 0.03\lambda_{\max},\ \lambda_{\max}$ with $\lambda_{\max} = 6.3 \times 10^4$. We notice that there is a trade-off between large regularization parameters and the error between the desired and simulated downstream discharge. Large values of the regularization parameters bias the control input $u(t)$ to remain constant. Thus, it is possible to choose a more realistic control input $u(t)$ from (6.17) with a value of λ that satisfies some design parameters. In our case, it is important to achieve the desired-steady state and the controller shown in Figure 6.14 (c) for $\lambda = 0.03\lambda_{\max}$ would be a strong candidate.

6.6.3 Computing the Control Input

Problem (6.17) can be transformed to the LASSO problem,

$$\min_{\boldsymbol{\theta}} \ \|\boldsymbol{A}\boldsymbol{\theta} - \tilde{\boldsymbol{y}}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 , \tag{6.18}$$

with

$$A = \begin{bmatrix} 1 & & & \\ 1 & -1 & & \\ \vdots & & \ddots & \\ 1 & -1 & \cdots & -1 \end{bmatrix}.$$

The control input $\boldsymbol{u}$ can be recovered from $\boldsymbol{\theta}$ using the relation $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{\theta}$. In our case, we have used a sampling time $T_s = 30$ s, and a total simulation time $T_t = 10$ h. The problem has $1,201$ features, more than 80% of which are discarded at the optimal solution, i.e. $\theta_i = 0$ for more than 80% of its $1,201$ entries. The CD-SAFE algorithm we developed in chapter 4, is suited for such kind of problems. Were we to use a sampling time $T_s = 3$ s instead, we would have a LASSO problem with $12,010$ features and memory problems will prevail if we don't save the matrix $\boldsymbol{A}$ in the appropriate format, or if we don't use a special solver. However, with Safe Feature Elimination, many of these unimportant features can be discarded at the outset, before solving or forming the matrix $\boldsymbol{A}$. Thus SAFE allows us to obtain a solution of (6.18) without the need of any special treatment for the particular problem we are solving. We use CD-SAFE and the plain CD algorithm to compare the number of iterations needed to obtain a solution with 130 non-zeros. Figure 6.15 shows that with CD-SAFE we need

100 folds of iteration less than the CD algorithm to solve the LASSO problem in (6.18). We also report in Figure 6.16 the number of iterations needed for each algorithm to compute a control input $u(t)$ as a function of the changes in $u(t)$.

6.7 Conclusion

This chapter applied a flatness-based controller for an open channel hydraulic canal. The controller was tested by computer simulation using Saint-Venant equations and real experimentation on the Gignac canal, in southern France. The initial model that assumes constant lateral withdrawals is improved to take into account gravitational lateral withdrawals, which vary with the water level. Accounting for gravitational lateral withdrawals decreased the steady state error from 6.2% (constant lateral withdrawals assumption) to 1% (gravitational lateral withdrawals assumption). The flatness based open-loop controller is thus able to compute the upstream water discharge corresponding to a desired downstream water discharge, taking into account the gravitational withdrawals along the canal reach.

The actuator limitations were addressed by deriving a simpler model, a first order differential equation with delay. The simpler model was used to compute a more realistic open-loop controller. The CD-SAFE algorithm was compared to the CD algorithm in computing such control signal , and it was shown that SAFE enables the computations of the control input with less memory requirement and reduced computational effort.

(a)

(b)

(c)

Figure 6.7: Gignac canal. The main canal is 50 km long, with a feeder canal of 8 km, and two branches on both the left and right banks of Hrault river. The left branch, which is 27 km long, and the right branch, which is 15 km long, originate at the Partiteur station. (a) shows the left and right branches of Partiteur station. (b) shows an automatic regulation gate at the right branch used to control the water discharge. (c) shows the ultrasonic velocity sensor that measures the average water velocity.

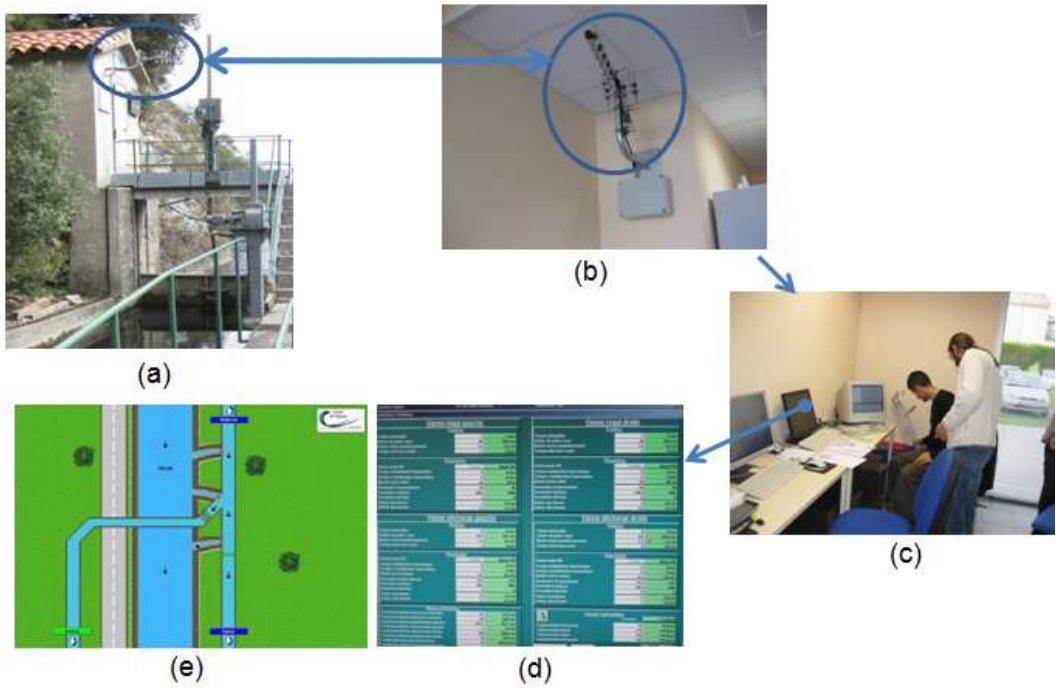


Figure 6.8: SCADA (supervision, control, and data acquisition) system. The SCADA system manages the canal by enabling the monitoring of the water discharge and by controlling the actuators at the gates. Data from sensors and actuators on the four gates at Partiteur are collected by a control station equipped with an antenna (a). The information is communicated by radio frequency signals every five minutes to a receiving antenna (b), located in the main control center, a few kilometers away (c). The data are displayed and saved in a database, while commands to the actuators are sent back to the local controllers at the gates (d)-(e). The SCADA performs open-loop control in real time.

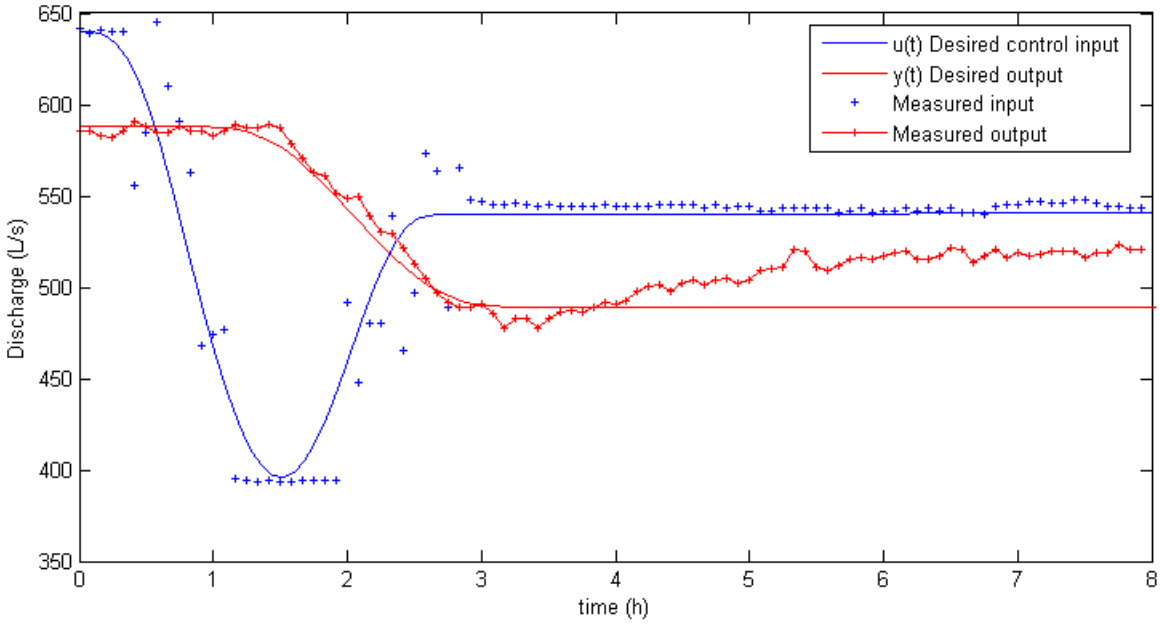


Figure 6.9: Implementation results of the Hayami controller on the Gignac canal. The Hayami open-loop control $u(t)$ is applied to right branch of Partiteur using the SCADA system. The measured output (downstream water discharge) follows the desired curve, except at the end of the experiment. This discrepancy cannot be explained solely by the actuator limitations, but rather is due to simplifications in the model assumptions.

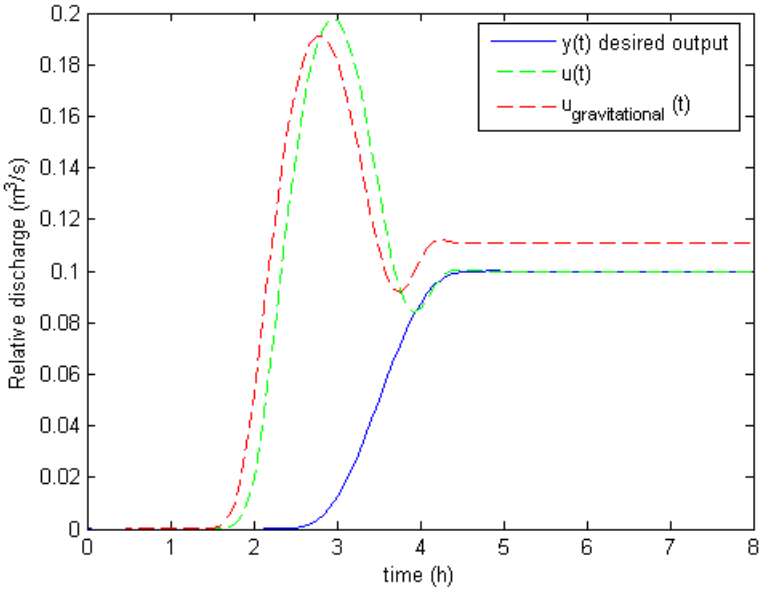


Figure 6.10: Hayami control taking into account the effect of gravitational lateral withdrawals. The control input is computed with the Hayami model (with constant and gravitational lateral withdrawals). As expected, to account for gravitational lateral withdrawals, the open-loop control $u_{\text{gravitational}}(t)$ needs to release more water than is required at the downstream end.

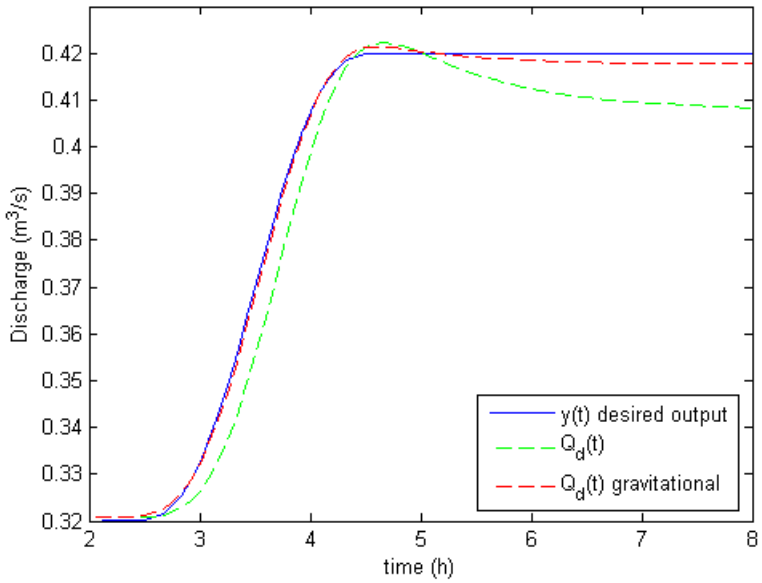


Figure 6.11: Comparison of the desired and simulated downstream water discharges. The downstream water discharge, $Q_d(t)$ and $Q_d(t)$ gravitational, is computed by solving the Saint-Venant equations with upstream water discharges $u(t)$ and $u_{\text{gravitational}}(t)$, respectively. Accounting for gravitational lateral withdrawals enables the controller to follow the desired output. This result is obtained on a realistic model of SIC, which is different from the simplified Hayami model used for control design.

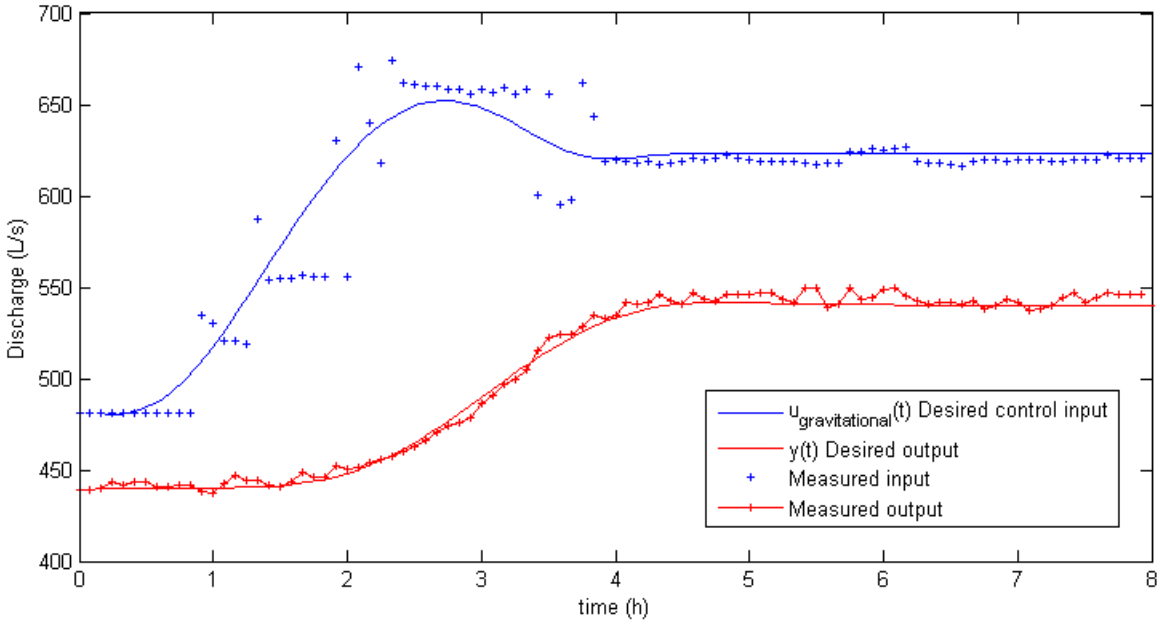
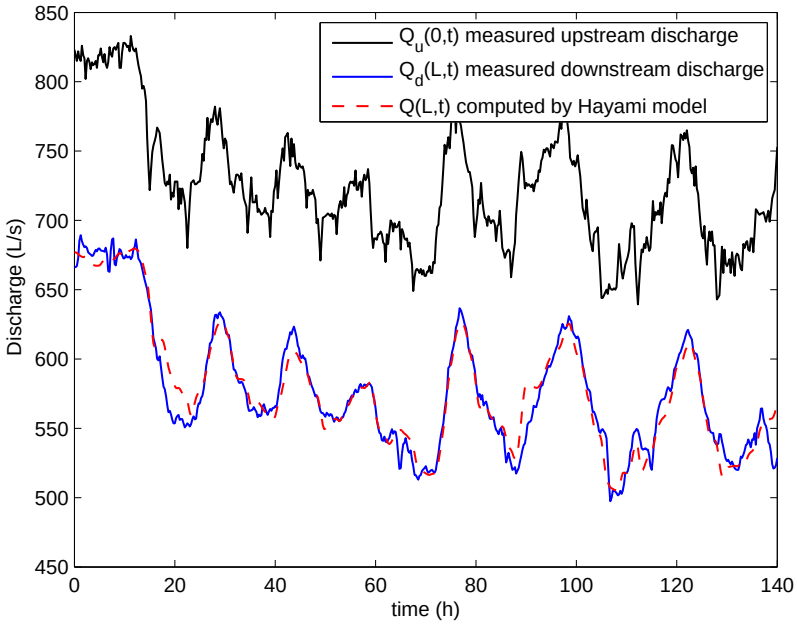
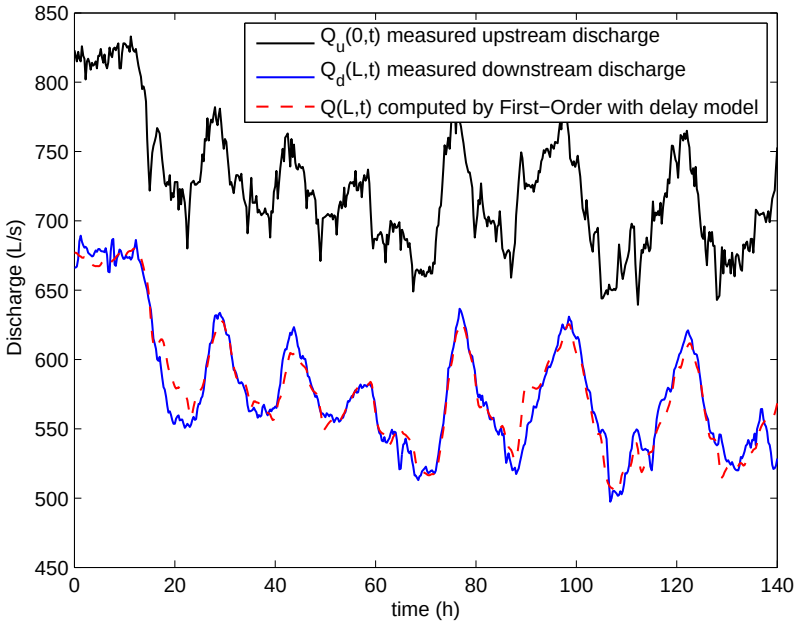


Figure 6.12: Implementation results of the Hayami controller on the Gignac canal. The Hayami controller assumes gravitational lateral withdrawals. The relative error between the measured downstream water discharge and the desired downstream water discharge is less than 9%, despite the fact that the delivered upstream water discharge is perturbed due to actuator limitations.



(a)



(b)

Figure 6.13: System identification using (a) Hayami model and (b) First order with delay model.

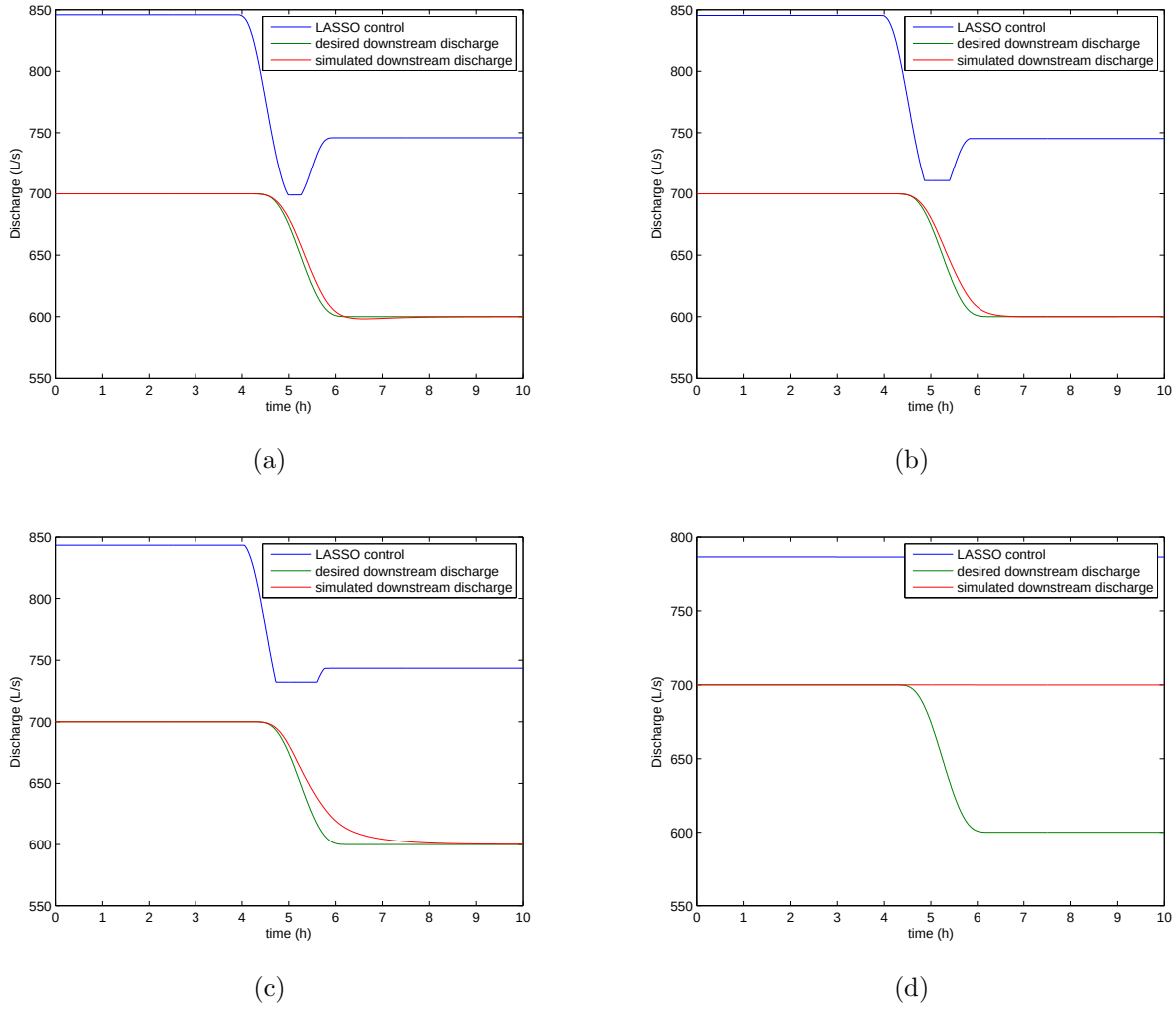


Figure 6.14: First order with delay model based control applied to the Saint-Venant model. The downstream water discharge $Q_{\mathrm{d}}(t)$ is the output obtained by applying the control input $u(t)$ of (6.17). We present four cases of the control input $u(t)$ corresponding to the four regularization parameters, (a) $\lambda = 0.001\lambda_{\max}$, (b) $\lambda = 0.01\lambda_{\max}$, (c) $\lambda = 0.03\lambda_{\max}$, and (d) $\lambda = \lambda_{\max}$, with $\lambda_{\max} = 6.3 \times 10^4$. We notice that there is a trade-off between large regularization parameters and the error between the desired and simulated downstream discharge. Large values of the regularization parameters bias the control input $u(t)$ to be constant.

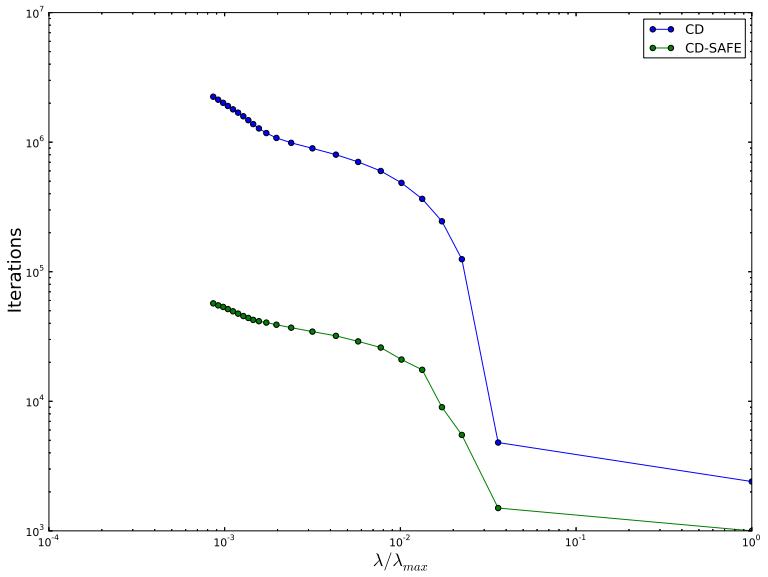


Figure 6.15: Number of iterations needed to reach a stationary tolerance of $\epsilon = 10^{-2}$ for the CD and CD-SAFE algorithms solved using feature matrix $\boldsymbol{A}$ and response $\tilde{\boldsymbol{y}}$. The simulation results show that CD-SAFE provides at least 10 or 100 folds of less iterations to reach the same tolerance as the CD algorithm.

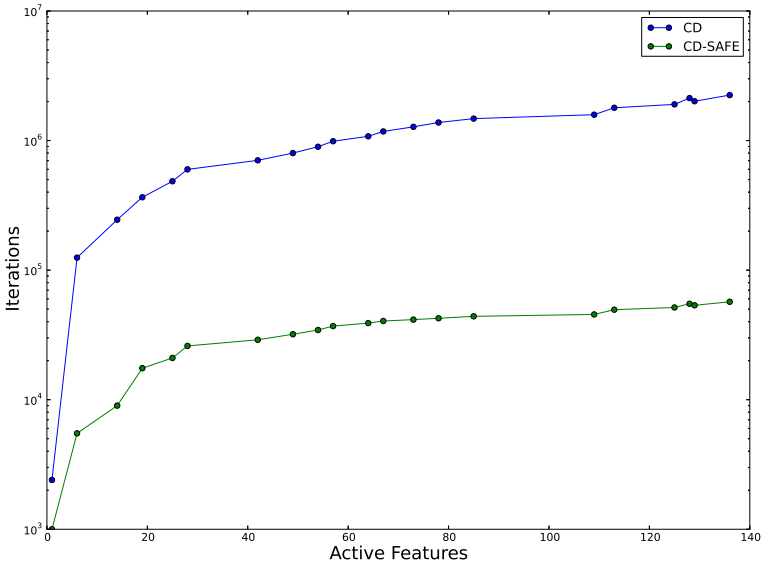


Figure 6.16: Number of iterations needed for the CD and CD-SAFE algorithms as a function of the number of changes in $u(t)$.

Part III

Appendix

Appendix A

On Thresholding Methods for the LASSO

A.1 Introduction

ℓ_1- regularized convex optimization problems or sparse classification algorithms may return an optimal solution vector with many small, but not exactly zero, elements. This implies that we need to choose a thresholding rule to decide which elements can be set to zero. In this appendix, we discuss an issue related to the thresholding rule originally proposed for the interior point method for Logistic Regression algorithm in [28], and propose a new thresholding rule.

A.2 The KKT thresholding rule

Recall that the primal problem for the LASSO is

$$\phi(\lambda) = \min_{w} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1, \tag{A.1}$$

with $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\lambda > 0$ and that the strong duality optimality conditions imply that, at optimum, $\boldsymbol{x}_k^T\left(\boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}\right) = \lambda \text{sign}(w_k^\star)$, with $\boldsymbol{x}_k$ the $k-th$ column of $\boldsymbol{X}$, $w_k^\star$ the $k-th$ entry of $\boldsymbol{w}^\star$ and $\text{sign}(0) \in [-1, 1]$. The ideas of [28] suggests that the following thresholding rule can be proposed: at optimum, set component w_k to 0 whenever

$$|\boldsymbol{x}_k^T\left(\boldsymbol{X}\boldsymbol{w}^\star - \boldsymbol{y}\right)| \leq 0.9999\lambda. \tag{A.2}$$

We refer to this rule as the "KKT" rule.

The interior point algorithm or IPM-LASSO algorithm in [27] takes as input a "duality gap" parameter ϵ, which controls the relative accuracy on duality gap. When comparing

the IPM code results with other algorithms such as Generalized Linear Model algorithm (GLMNET) described in [20], we observed chaotic behaviors when applying the KKT rule, especially when the duality gap parameter ϵ is not small enough. More surprisingly, some components w_k with absolute values not close to 0 can be thresholded. This suggests that the KKT rule should only be used for problems solved with a small enough duality gap ϵ. However, setting the duality gap to a small value can dramatically slow down computations. In our experiments, changing the duality gap from $\epsilon = 10^{-4}$ to 10^{-6} (resp. 10^{-8}) increased the computational time by 30% to 40% (resp. 50 to 100%).

A.3 An alternative method

We propose an alternative thresholding rule, which is based on controlling the perturbation of the objective function that is induced by thresholding.

Assume that we have solved the LASSO problem above, with a given duality gap parameter ϵ. If we denote by $\boldsymbol{w}^\star$ the optimal solution obtained by the IPM algorithm, $\boldsymbol{w}^\star$ is ϵ-sub-optimal, that is, achieves a value

$$\phi^* = \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\right\|_2^2 + \lambda\left\|\boldsymbol{w}\right\|_1,$$

with $0 \leq \phi^* - \phi(\lambda) \leq \epsilon\phi(\lambda)$.

For a given threshold $\tau > 0$, consider the thresholded vector $\tilde{\boldsymbol{w}}(\tau)$ defined as

$$\tilde{w}_k(\tau) \;=\; \begin{cases} 0 & \text{if } |w_k^*| \leq \tau, \\ w_k^* & \text{otherwise,} \end{cases} \quad k = 1, \ldots, n.$$

We have $\tilde{\boldsymbol{w}}(\tau) = \boldsymbol{w}^* + \boldsymbol{\delta}(\tau)$ where the vector of perturbation $\boldsymbol{\delta}(\tau)$ is such that

$$\delta_k(\tau) \;=\; \begin{cases} -w_k^* & \text{if } |w_k^*| \leq \tau, \\ 0 & \text{otherwise,} \end{cases} \quad k = 1, \ldots, n.$$

Note that, by construction, we have $\|\boldsymbol{w}^*\|_1 = \|\boldsymbol{w}^* + \boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_1$. Also note that if $\boldsymbol{w}^*$ is sparse, so is $\boldsymbol{\delta}$.

Let us now denote by ϕ_τ the LASSO objective that we obtain upon replacing the optimum solution $\boldsymbol{w}^*$ with its thresholded version $\tilde{\boldsymbol{w}}(\tau) = \boldsymbol{w}^* + \boldsymbol{\delta}(\tau)$:

$$\phi_\tau \;:=\; \frac{1}{2}\|\boldsymbol{X}(\boldsymbol{w}^* + \boldsymbol{\delta}(\tau)) - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}^* + \boldsymbol{\delta}(\tau)\|_1.$$

Since $\boldsymbol{w}(\tau)$ is (trivially) feasible for the primal problem, we have $\phi_\tau \geq \phi(\lambda)$. On the other hand,

$$\begin{aligned} \phi_\tau &= \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}^* + \boldsymbol{\delta}(\tau)\|_1 + \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\delta}(\tau)\|_2^2 + \boldsymbol{\delta}(\tau)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}) \\ &\leq \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}^*\|_1 + \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\delta}(\tau)\|_2^2 + \boldsymbol{\delta}(\tau)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}). \end{aligned}$$

For a given $\alpha > 1$, the condition

$$\mathcal{C}(\tau) := \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\delta}(\tau)\|_2 + \boldsymbol{\delta}(\tau)^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}) \leq \kappa\phi^*, \;\; \kappa := \frac{1+\alpha\epsilon}{1+\epsilon} - 1 \geq 0, \tag{A.3}$$

allows to write

$$\phi(\lambda) \leq \phi_\tau \leq \; (1+\alpha\epsilon)\phi(\lambda).$$

The condition (A.3) then implies that the thresholded solution is sub-optimal, with relative accuracy $\alpha\epsilon$.

Our proposed thresholding rule is based on the condition (A.3). Precisely, we choose the parameter $\alpha > 0$, then we set the threshold level τ by solving, via line search, the largest threshold τ allowed by condition (A.3):

$$\tau_\alpha = \arg\max_{\tau \geq 0} \left\{ \tau \;:\; \|\boldsymbol{X}\boldsymbol{\delta}(\tau)\|_2 \leq \left(\sqrt{\frac{1+\alpha\epsilon}{1+\epsilon}} - 1\right) \|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|_2 \right\}.$$

The larger α is, the more elements the rule allows to set to zero; at the same time, the more degradation in the objective we observe: precisely, the new relative accuracy is bounded by $\alpha\epsilon$. The rule also depends on the duality gap parameter ϵ. We refer to the thresholding rule as TR(α) in the sequel. In practice, we observe that the value $\alpha = 2$ works well, in a sense made more precise below.

The complexity of the rule is $O(mn)$. We have the optimal dual variable $\boldsymbol{\theta}^* = \boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}$ is returned by IPM-LASSO and the matrix $\boldsymbol{X}^T\boldsymbol{\theta}^*$ is computed once for all in $O(mn)$. We then sort the optimal vector $\boldsymbol{w}^*$ so that $|w^*_{(1)}| \leq \ldots \leq |w^*_{(n)}|$, and set $\tau = \tau_0 = |w^*_{(n)}|$, so that $\delta_k(\tau_0) = -w^*_k$ and $\tilde{w}_k(\tau_0) = 0$ for all $k = 1, \ldots, n$. The product $\boldsymbol{X}\boldsymbol{\delta}(\tau_0)$ is computed in $O(mn)$, while the product $\boldsymbol{\delta}(\tau_0)^T(\boldsymbol{X}^T\boldsymbol{\theta}^*)$ is computed in $O(n)$. If the quantity $\mathcal{C}(\tau_0) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\delta}(\tau_0)\|_2 + \boldsymbol{\delta}(\tau_0)^T(\boldsymbol{X}^T\boldsymbol{\theta}^*)$ is greater than $\kappa\phi^\star$, then we set $\tau = \tau_1 = |w^*_{(n-1)}|$. We have $\delta_k(\tau_1) = \delta_k(\tau_0)$ for any $k \neq (n)$ and $\delta_{(n)}(\tau_1) = 0$. Therefore, $\mathcal{C}(\tau_1)$ can be deduced from $\mathcal{C}(\tau_0)$ in $O(n)$. We proceed by successively setting $\tau_k = |w^*_{(n-k)}|$ until we reach a threshold τ_k such that $\mathcal{C}(\tau_k) \leq \kappa\phi^*$.

A.4 Simulation study.

We conducted a simple simulation study to evaluate our proposed method and compared it to the KKT thresholding rule. Both methods were further compared to the results returned by the `glmnet R` package. GLMNET algorithm returns exact zeros in the optimal solution, and we have chosen the corresponding sparsity pattern as the "ground truth", which the IPM should recover.

We first experimented with synthetic data. We generated samples of the pair $(\boldsymbol{X}, \boldsymbol{y})$ for various values of (m, n). We present the results for $(m, n) = (5000, 2500)$ and $(m, n) =$

$(100, 500)$. The number s of relevant features was set to $\min(m, n/2)$. Features were drawn from independent $\mathcal{N}(0, 1)$ distributions and $\boldsymbol{y}$ was computed as $\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{w} + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(0, 0.2)$ and $\boldsymbol{w}$ is a vector of $\mathbb{R}^n$ with first s components equal to $0.1 + 1/s$ and remaining $n - s$ components set to 0. Because `glmnet` includes an unpenalized intercept while IPM method does not, both $\boldsymbol{y}$ and $\boldsymbol{X}$ were centered before applying either methods to make their results comparable.

Results are presented on Figure A.1. First, the KKT thresholding rule was observed to be very chaotic when the duality gap was set to $\epsilon = 10^{-4}$ (we recall here that the default value for the duality gap in IPM `MATLAB` implementation is $\epsilon = 10^{-3}$), while it was way better when duality gap was set to $\epsilon = 10^{-8}$ (somehow justifying our choice of considering the sparsity pattern returned by `glmnet` as the ground truth). Therefore, for applications where computational time is not critical, running IPM method and applying KKT thresholding rule should yield appropriate results. However, when computational time matters, passing the duality gap from, say, 10^{-4} to 10^{-8}, is not a viable option. Next, regarding our proposal, we observed that it was significantly better than KKT thresholding rule when the duality gap was set to 10^{-4} and equivalent to KKT thresholding rule for a duality gap of 10^{-8}. Interestingly, setting $\alpha = 1.5$ in (A.3) generally enabled to achieved very good results for low values of λ, but lead to irregular results for higher values of λ (in the case $m = 100$, results were unstable for the whole range of λ values we considered). Overall, the choices $\alpha = 2$, 3 and 4 lead to acceptable results. A little irregularity remained with $\alpha = 2$ for high values of λ, but this choice of α performed the best for lower values of λ. As for choices $\alpha = 3$ and $\alpha = 4$, it is noteworthy that the results were all the better as the dimension n was low.

A.4.1 Real data examples

We also applied our proposed method and compared it to KKT rule (A.2) on real data sets arising in text classification. More precisely, we used the New York Times headlines data set presented in Section 3.5. We successively ran IPM-LASSO method with duality gap set to 10^{-4} and 10^{-8} and compared the number of active features returned after applying KKT thresholding rule (A.2) and TR (1.5), TR (2), TR (3) and TR (4). Results are presented on Figure A.2. Because we could not applied `glmnet` on this data set, the ground truth was considered as the result of KKT rule, when applied to the model returned by IPM-LASSO when ran with duality gap set to 10^{-10}. Applying KKT rule on the model built with a duality gap of 10^{-4} lead to very misleading results again, especially for low values of λ. In this very high-dimensional setting ($n = 38377$ here), our rule generally resulted in a slight "underestimation" of the true number of active features for the lowest values of λ when the duality gap was set to 10^{-4}. This suggests that the "optimal" α for our rule might depend on both n and λ when the duality gap is not small enough. However, we still observe that our proposed method introduces significant improvments over KKT rule when the duality gap is set to 10^{-4}.

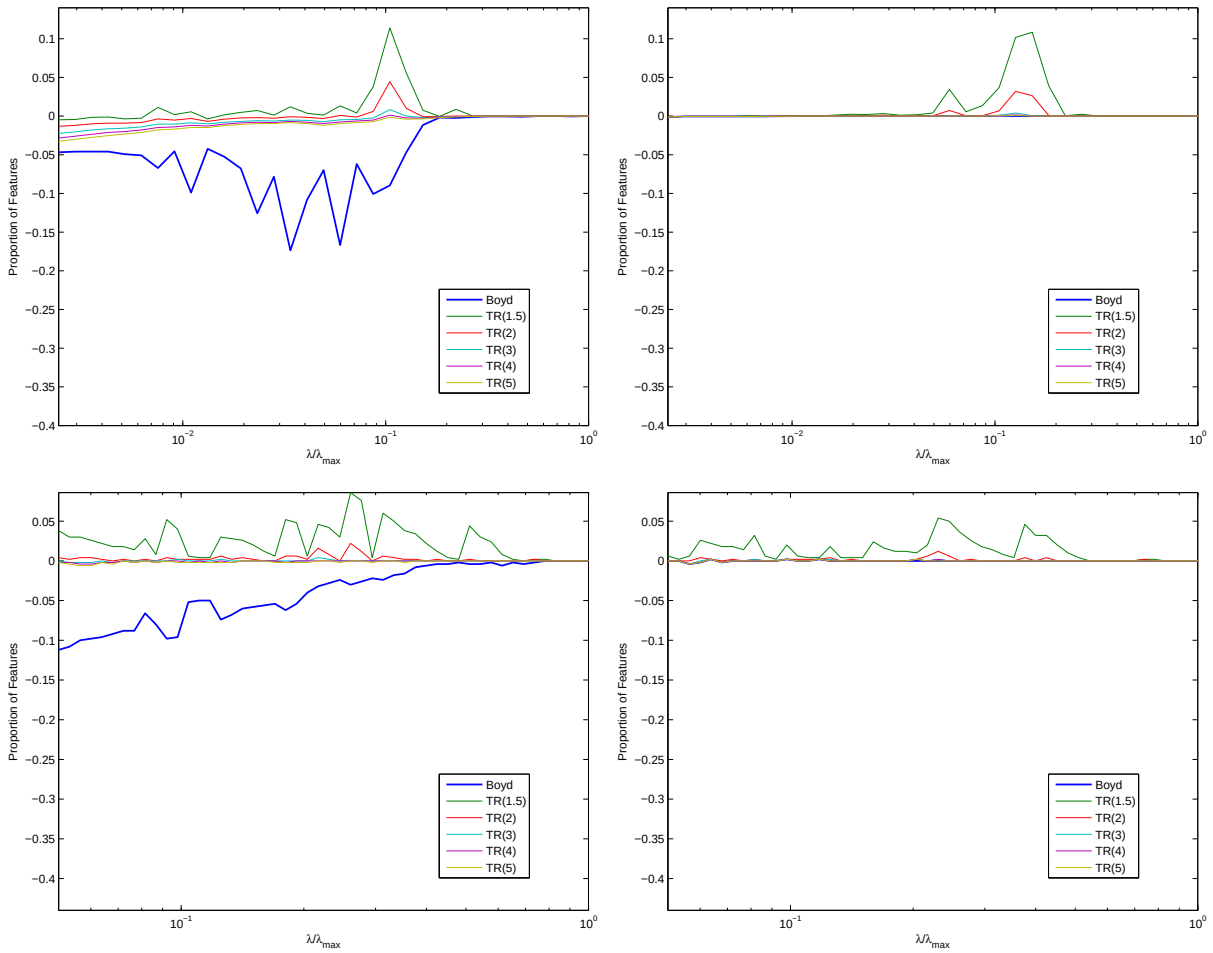


Figure A.1: Comparison of several thresholding rules on synthetic data: the case $m = 5000$, $n = 100$ (*top panel*) and $m = 100$, $n = 500$ (*bottom panel*) with duality gap in IPM method set to (i) 10^{-4} (*left panel*) and (iii) 10^{-8} (*right panel*). The curves represent the differences between the number of active features returned after each thresholding method and the one returned by `glmnet` (this difference is further divided by the total number of features n). The graphs present the results attached to six thresholding rules: the one proposed by [28] and five versions of our proposal, corresponding to setting α in (A.3) to 1.5, 2, 3, 4 and 5 respectively. Overall, these results suggest that by setting $\alpha \in (2, 5)$, our rule is less sensitive to the value of the duality gap parameter in IPM-LASSO than is the rule proposed by [28].

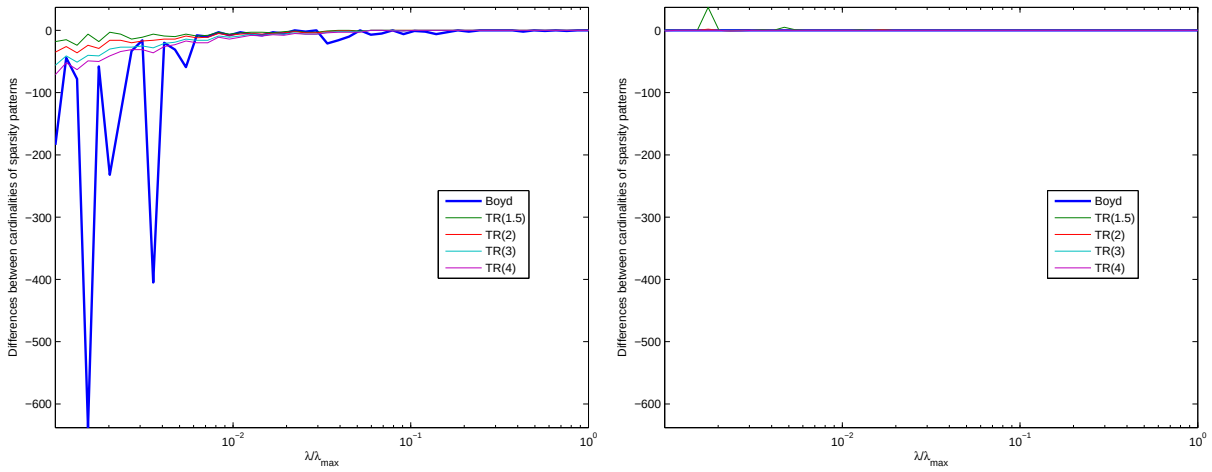


Figure A.2: Comparison of several thresholding rules on the NYT headlines data set for the topic "China" and year 1985. Duality gap in IPM-LASSO was successively set to 10^{-4} (*left panel*) and 10^{-8} (*right panel*). The curves represent the differences between the number of active features returned after each thresholding method and the one returned by the KKT rule when duality gap was set to 10^{-10}. The graphs present the results attached to five thresholding rules: the KKT rule and four versions of our rule, corresponding to setting α in (A.3) to 1.5, 2, 3 and 4 respectively. Results obtained following our proposal appear to be less sensitive to the value of the duality gap used in IPM-LASSO. For instance, for the value $\lambda = \lambda_{\max}/1000$, the KKT rule returns 1758 active feature when the duality gap is set to 10^{-4} while it returns 2357 features for a duality gap of 10^{-8}.

Appendix B

SAFE Derivations

In this appendix, we present the proof of propositions related to the SAFE test problem, which arise in Section 4.2.

Proposition B.0.1 (SAFE-LASSO test Problem) *Consider the problem*

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) := \max_{\boldsymbol{\theta}} \boldsymbol{x}^T \boldsymbol{\theta} \ : \ G(\boldsymbol{\theta}) \geq \gamma, \ \boldsymbol{\eta}^T (\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0. \tag{B.1}$$

with $G(\boldsymbol{\theta}) = -\frac{1}{2} \|\boldsymbol{\theta}\|_2^2 - \boldsymbol{y}^T \boldsymbol{\theta}$. *Assume that strong duality holds and a solution of the problem is attained. Let* $\boldsymbol{g}_s = \boldsymbol{\theta}_s + \boldsymbol{y}$ *, then* $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ *takes the value*

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{y}^T \boldsymbol{x} + \|\boldsymbol{x}\|_2 D & \|\boldsymbol{x}\|_2 \left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right) \leq D \left(\boldsymbol{\eta}^T \boldsymbol{x}\right), \\ \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T \boldsymbol{x}\right) \boldsymbol{\eta}^T \boldsymbol{g}_s - \boldsymbol{x}^T \boldsymbol{y} + \psi \tilde{D} & \textit{otherwise}, \end{cases} \tag{B.2}$$

with

$$D = \left(\|\boldsymbol{y}\|_2^2 - 2\gamma\right)^{1/2},$$

$$\tilde{D} = \left(-2\gamma - \frac{\left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2}{\|\boldsymbol{\eta}\|_2^2} + \|\boldsymbol{y}\|_2^2\right)^{1/2},$$

and

$$\psi = \left(\|\boldsymbol{x}\|_2^2 - \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2\right)^{1/2}.$$

Proof: We express problem (B.1) in dual form as a convex optimization problem with two scalar variables, μ_1 and μ_2,

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{\eta},\boldsymbol{\theta}_s,\gamma) &= \min_{\mu_1,\mu_2\geq 0}\max_{\boldsymbol{\theta}} \boldsymbol{x}^T\boldsymbol{\theta}+\mu_1\left(G(\boldsymbol{\theta})-\gamma\right)+\mu_2\boldsymbol{\eta}^T\left(\boldsymbol{\theta}-\boldsymbol{\theta}_s\right), \\ &= \min_{\mu_1,\mu_2\geq 0} -\mu_1\gamma-\mu_2\boldsymbol{\eta}^T\boldsymbol{\theta}_s+\max_{\theta}\boldsymbol{x}^T\boldsymbol{\theta}+\mu_1 G(\boldsymbol{\theta})+\mu_2\boldsymbol{\eta}^T\boldsymbol{\theta}, \\ &= \min_{\mu_1,\mu_2\geq 0} -\mu_1\gamma-\mu_2\boldsymbol{\eta}^T\boldsymbol{\theta}_s \\ &\quad +\mu_1\max_{\theta}\left(\frac{\boldsymbol{x}^T-\mu_1\boldsymbol{y}^T+\mu_2\boldsymbol{\eta}^T}{\mu_1}\boldsymbol{\theta}-\frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right). \end{aligned} \tag{B.3}$$

The maximization problem

$$\max_{\theta}\left(\frac{\boldsymbol{x}^T-\mu_1\boldsymbol{y}^T+\mu_2\boldsymbol{\eta}^T}{\mu_1}\boldsymbol{\theta}-\frac{1}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right),$$

in (B.3) admits a solution at $\boldsymbol{\theta}=\frac{1}{\mu_1}\left(\boldsymbol{x}+\mu_2\boldsymbol{\eta}\right)-\boldsymbol{y}$. We substitute the value of $\boldsymbol{\theta}$ in (B.3) and obtain

$$P(\boldsymbol{x},\boldsymbol{\eta},\boldsymbol{\theta}_s,\gamma) = \min_{\mu_1,\mu_2\geq 0} L(\mu_1,\mu_2), \tag{B.4}$$

with

$$L(\mu_1,\mu_2)=-\mu_1\gamma-\mu_2\boldsymbol{\eta}^T\boldsymbol{\theta}_s+\frac{1}{2\mu_1}\left\|\boldsymbol{x}-\mu_1\boldsymbol{y}+\mu_2\boldsymbol{\eta}\right\|_2^2. \tag{B.5}$$

Solving the dual form. We take the partial derivative of $L(\mu_1,\mu_2)$ with respect to μ_2, and set it to 0,

$$\begin{aligned} -\eta^T\boldsymbol{\theta}_s+\frac{2}{2\mu_1}\boldsymbol{\eta}^T(\boldsymbol{x}-\mu_1\boldsymbol{y}+\mu_2\boldsymbol{\eta}) &= 0, \\ \eta^T\boldsymbol{x}+\mu_2\left\|\boldsymbol{\eta}\right\|_2^2-\mu_1\boldsymbol{\eta}^T\boldsymbol{\theta}_s-\mu_1\boldsymbol{\eta}^T\boldsymbol{y} &= 0, \end{aligned}$$

or

$$\mu_2=\frac{1}{\left\|\boldsymbol{\eta}\right\|_2^2}\left(-\boldsymbol{\eta}^T\boldsymbol{x}+\mu_1\boldsymbol{\eta}^T\boldsymbol{g}_s\right).$$

Since μ_2 is contrained to be non-negative, we write $\mu_2=\max\left(0,\alpha\mu_1-\beta\right)$, with $\alpha=\frac{\boldsymbol{\eta}^T\boldsymbol{g}_s}{\left\|\boldsymbol{\eta}\right\|_2^2}$, and $\beta=\frac{\boldsymbol{\eta}^T\boldsymbol{x}}{\left\|\boldsymbol{\eta}\right\|_2^2}$. We recognize two cases: $\alpha\mu_1\leq\beta$ and $\alpha\mu_1>\beta$.

First case: When $\alpha\mu_1 \leq \beta$, we have $\mu_2 = 0$. We find μ_1 by setting $\mu_2 = 0$ in (B.5) and by taking its partial derivative with respect to μ_1,

$$L(\mu_1, 0) = -\mu_1\gamma + \frac{1}{2\mu_1}\left\|\boldsymbol{x}\right\|_2^2 + \frac{1}{2}\mu_1\left\|\boldsymbol{y}\right\|_2^2 - \boldsymbol{x}^T\boldsymbol{y},$$

$$\frac{\partial L(\mu_1, 0)}{\partial \mu_1} = -\gamma - \frac{1}{2\mu_1^2}\left\|\boldsymbol{x}\right\|_2^2 + \frac{1}{2}\left\|\boldsymbol{y}\right\|_2^2 = 0.$$

We obtain

$$\left\|\boldsymbol{y}\right\|_2^2 - 2\gamma = \frac{\left\|\boldsymbol{x}\right\|_2^2}{\mu_1^2},$$

or $\mu_1 = \frac{\left\|\boldsymbol{x}\right\|_2}{D}$ with $D = \left(\left\|\boldsymbol{y}\right\|_2^2 - 2\gamma\right)^{1/2}$. The condition $\mu_1 \leq \beta$ is equivalent to $\left\|\boldsymbol{x}\right\|_2 \left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)$.

The corresponding dual variable $\boldsymbol{\theta}$ for this (μ_1, μ_2) is $\boldsymbol{\theta} = \frac{\boldsymbol{x}}{\mu_1} - \boldsymbol{y}$ and $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \boldsymbol{x}^T\boldsymbol{\theta}$ takes the value,

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = -\boldsymbol{y}^T\boldsymbol{x} + \left\|\boldsymbol{x}\right\|_2 D \;:\; \left\|\boldsymbol{x}\right\|_2 \left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D\left(\boldsymbol{\eta}^T\boldsymbol{x}\right).$$

Second case: When $\alpha\mu_1 \geq \beta$, we have $\mu_2 = \frac{1}{\left\|\boldsymbol{\eta}\right\|_2^2}\left(-\boldsymbol{\eta}^T\boldsymbol{x} + \mu_1\boldsymbol{\eta}^T\boldsymbol{g}_s\right)$. We take the partial derivative of

$$L(\mu_1, \mu_2) = -\mu_1\gamma - \mu_2\boldsymbol{\eta}^T\boldsymbol{\theta}_s + \frac{1}{2\mu_1}\left\|\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right\|_2^2,$$

with respect to μ_1, using the chain-rule, and set it to zero. We obtain,

$$\begin{aligned} -\gamma - \left(\frac{\boldsymbol{\eta}^T\boldsymbol{g}_s}{\left\|\boldsymbol{\eta}\right\|_2^2}\right)\boldsymbol{\eta}^T\boldsymbol{\theta}_s - \frac{1}{2\mu_1^2}\left\|\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right\|_2^2 \\ + \frac{1}{\mu_1}\left(-\boldsymbol{y} + \frac{\boldsymbol{\eta}^T\boldsymbol{g}_s}{\left\|\boldsymbol{\eta}\right\|_2^2}\eta\right)^T\left(\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right) &= 0. \end{aligned} \tag{B.6}$$

We simplify the expression by calling $\sigma_1 = \left(\frac{\boldsymbol{\eta}^T\boldsymbol{g}_s}{\left\|\boldsymbol{\eta}\right\|_2^2}\right)$, $\sigma_2 = \boldsymbol{\eta}^T\boldsymbol{\theta}_s$, $\sigma_3 = -\gamma - \sigma_1\sigma_2$. We have,

$$\sigma_3 - \frac{1}{2\mu_1^2}\left\|\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right\|_2^2 + \frac{1}{2\mu_1}\left(-\boldsymbol{y} + \sigma_1\boldsymbol{\eta}\right)^T\left(\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right) = 0.$$

We expand the two terms, $\left\|\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right\|_2^2$ and $\left(-\boldsymbol{y} + \sigma_1\boldsymbol{\eta}\right)^T\left(\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right)$ inside the above expression. We have

$$\begin{aligned} \left\|\boldsymbol{x} - \mu_1\boldsymbol{y} + \mu_2\boldsymbol{\eta}\right\|_2^2 &= \left\|\boldsymbol{x}\right\|_2^2 + \mu_1^2\left\|\boldsymbol{y}\right\|_2^2 + \mu_2^2\left\|\boldsymbol{\eta}\right\|_2^2 \\ &\quad -2\mu_1\boldsymbol{x}^T\boldsymbol{y} + 2\mu_2\boldsymbol{x}^T\boldsymbol{\eta} - 2\mu_1\mu_2\boldsymbol{y}^T\boldsymbol{\eta}, \end{aligned}$$

and

$$\begin{aligned}(-\boldsymbol{y}+\sigma_1\boldsymbol{\eta})^T(\boldsymbol{x}-\mu_1\boldsymbol{y}+\mu_2\boldsymbol{\eta}) &= -\boldsymbol{y}^T\boldsymbol{x}+\mu_1\|\boldsymbol{y}\|_2^2-\mu_2\boldsymbol{y}^T\boldsymbol{\eta}\\ &\quad +\sigma_1\boldsymbol{\eta}^T\boldsymbol{x}-\mu_1\sigma_1\boldsymbol{\eta}^T\boldsymbol{y}+\mu_2\sigma_1\|\boldsymbol{\eta}\|_2^2.\end{aligned}$$

We substitute the above expressions in (B.6) and obtain

$$\begin{aligned}&\sigma_3-\frac{\|\boldsymbol{x}\|_2^2}{2\mu_1^2}-\frac{1}{2}\|\boldsymbol{y}\|_2^2-\frac{\mu_2^2}{2\mu_1^2}\|\boldsymbol{\eta}\|_2^2+\frac{1}{\mu_1}\boldsymbol{x}^T\boldsymbol{y}-\frac{\mu_2}{\mu_1^2}\boldsymbol{x}^T\boldsymbol{\eta}+\frac{\mu_2}{\mu_1}\boldsymbol{y}^T\boldsymbol{\eta}\\ &-\frac{1}{\mu_1}\boldsymbol{y}^T\boldsymbol{x}+\|\boldsymbol{y}\|_2^2-\frac{\mu_2}{\mu_1}\boldsymbol{y}^T\boldsymbol{\eta}+\frac{1}{\mu_1}\sigma_1\boldsymbol{\eta}^T\boldsymbol{x}-\sigma_1\boldsymbol{\eta}^T\boldsymbol{y}+\frac{\mu_2}{\mu_1}\sigma_1\|\boldsymbol{\eta}\|_2^2 = 0.\end{aligned}$$

The equation above can be simplified furthermore to read,

$$\begin{aligned}&\sigma_3-\frac{\|\boldsymbol{x}\|_2^2}{2\mu_1^2}+\frac{1}{2}\|\boldsymbol{y}\|_2^2-\frac{\mu_2^2}{2\mu_1^2}\|\boldsymbol{\eta}\|_2^2+\frac{1}{\mu_1}\boldsymbol{x}^T\boldsymbol{y}-\frac{\mu_2}{\mu_1^2}\boldsymbol{x}^T\boldsymbol{\eta}\\ &-\frac{1}{\mu_1}\boldsymbol{y}^T\boldsymbol{x}+\frac{1}{\mu_1}\sigma_1\boldsymbol{\eta}^T\boldsymbol{x}-\sigma_1\boldsymbol{\eta}^T\boldsymbol{y}+\frac{\mu_2}{\mu_1}\sigma_1\|\boldsymbol{\eta}\|_2^2 = 0,\end{aligned}$$

or

$$\begin{aligned}&\left(2\sigma_3+\|\boldsymbol{y}\|_2^2-2\sigma_1\boldsymbol{\eta}^T\boldsymbol{y}\right)\mu_1^2-\|\boldsymbol{x}\|_2^2-\mu_2^2\|\boldsymbol{\eta}\|_2^2-2\mu_2\boldsymbol{x}^T\boldsymbol{\eta}\\ &\qquad +2\mu_1\sigma_1\boldsymbol{\eta}^T\boldsymbol{x}+2\mu_1\mu_2\sigma_1\|\boldsymbol{\eta}\|_2^2 = 0.\end{aligned}$$

We simplify the terms $-\mu_2^2\|\boldsymbol{\eta}\|_2^2$, $-2\mu_2\boldsymbol{x}^T\boldsymbol{\eta}$ and $2\mu_1\mu_2\sigma_1\|\boldsymbol{\eta}\|_2^2$ in the above equation. We have

$$\mu_2=\left(-\frac{1}{\|\boldsymbol{\eta}\|_2^2}\boldsymbol{\eta}^T\boldsymbol{x}+\mu_1\sigma_1\right),$$

$$\mu_2^2=\frac{1}{\|\boldsymbol{\eta}\|_2^4}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2+\mu_1^2\sigma_1^2-2\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\sigma_1\mu_1,$$

$$-\mu_2^2\|\boldsymbol{\eta}\|_2^2=-\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2-\mu_1^2\sigma_1^2\|\boldsymbol{\eta}\|_2^2+2\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\sigma_1\mu_1,$$

$$-2\mu_2\boldsymbol{x}^T\boldsymbol{\eta}=+2\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2-2\mu_1\sigma_1\left(\boldsymbol{\eta}^T\boldsymbol{x}\right),$$

and

$$2\mu_2\mu_1\sigma_1\|\boldsymbol{\eta}\|_2^2=-2\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)(\mu_1\sigma_1)+2\mu_1^2\sigma_1^2\|\boldsymbol{\eta}\|_2^2.$$

After simplification, the partial derivative of $L(\mu_1,\mu_2)$, with respect to μ_1 reads

$$\begin{aligned}&\left(2\sigma_3+\|\boldsymbol{y}\|_2^2-2\sigma_1\boldsymbol{\eta}^T\boldsymbol{y}+\sigma_1^2\|\boldsymbol{\eta}\|_2^2\right)\mu_1^2\\ &\qquad -\|\boldsymbol{x}\|_2^2+\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2 = 0,\end{aligned}$$

or

$$\tilde{D}^2\mu_1^2 = \psi^2,$$

with

$$\tilde{D}^2 = -2\gamma - \frac{\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right)^2}{\|\boldsymbol{\eta}\|_2^2} + \|\boldsymbol{y}\|_2^2,$$

and

$$\psi^2 = \|\boldsymbol{x}\|_2^2 - \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2.$$

The corresponding dual variable $\boldsymbol{\theta}$ for this (μ_1, μ_2) is

$$\boldsymbol{\theta} = \frac{\boldsymbol{x}}{\mu_1} + \frac{\mu_2}{\mu_1}\boldsymbol{\eta} - \boldsymbol{y}.$$

We simplify $\frac{\mu_2}{\mu_1}$ to

$$\frac{\mu_2}{\mu_1} = \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(-\frac{\boldsymbol{\eta}^T\boldsymbol{x}}{\mu_1} + \boldsymbol{\eta}^T\boldsymbol{g}_s\right),$$

and we obtain

$$\boldsymbol{\theta} = \frac{\boldsymbol{x}}{\mu_1} - \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2^2}\frac{\boldsymbol{\eta}^T\boldsymbol{x}}{\mu_1} + \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|_2^2}\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{y}.$$

We then compute the value of $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$,

$$\begin{aligned}
P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) &= \frac{\|\boldsymbol{x}\|_2^2}{\frac{\psi}{\tilde{D}}} - \frac{(\boldsymbol{\eta}^T\boldsymbol{x})^2}{\|\boldsymbol{\eta}\|_2^2\frac{\psi}{\tilde{D}}} + \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y}, \\
&= \frac{\tilde{D}}{\psi}\left(\|\boldsymbol{x}\|_2^2 - \frac{(\boldsymbol{\eta}^T\boldsymbol{x})^2}{\|\boldsymbol{\eta}\|_2^2}\right) + \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y}, \\
&= \frac{\tilde{D}}{\psi}\psi^2 + \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y}, \\
&= \frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right)\boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y} + \psi\tilde{D}.
\end{aligned}$$

Recall that the value of $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ above is computed under the assumption that $\alpha\mu_1 \geq \beta$, or

$$\psi\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \geq \tilde{D}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right).$$

Finally, to obtain the result of (B.2), we need to prove that

$$\psi\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \geq \tilde{D}\left(\boldsymbol{\eta}^T\boldsymbol{x}\right), \tag{B.7}$$

and

$$\|\boldsymbol{x}\|_2\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \geq D\left(\boldsymbol{\eta}^T\boldsymbol{x}\right),$$

with

$$\tilde{D} = \left(-2\gamma - \frac{\left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2}{\|\boldsymbol{\eta}\|_2^2} + \|\boldsymbol{y}\|_2^2 \right)^{1/2},$$

$$\psi = \left(\|\boldsymbol{x}\|_2^2 - \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2 \right)^{1/2},$$

and $D = \left(\|\boldsymbol{y}\|_2^2 - 2\gamma\right)^{1/2}$, are equivalent.

We assume $\boldsymbol{\eta}^T \boldsymbol{g}_s \geq 0$, and we take the square of both sides of the inequality in (B.7), we have

$$\begin{aligned} \psi^2 \left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2 &\geq \tilde{D}^2 \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2 \\ \left(\|\boldsymbol{x}\|_2^2 - \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2 \right) \left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2 &\geq \left(D^2 - \frac{\left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2}{\|\boldsymbol{\eta}\|_2^2} \right) \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2 \\ \|\boldsymbol{x}\|_2^2 \left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right)^2 &\geq D^2 \left(\boldsymbol{\eta}^T \boldsymbol{x}\right)^2 \end{aligned}$$

Taking the square-root of both sides of the inequality, we obtain

$$\|\boldsymbol{x}\|_2 \left(\boldsymbol{\eta}^T \boldsymbol{g}_s\right) \geq D \left(\boldsymbol{\eta}^T \boldsymbol{x}\right).$$

Similarly, we can prove the same result for $\boldsymbol{\eta}^T \boldsymbol{g}_s \leq 0$. ■

Proposition B.0.2 *Consider the LASSO problem with regularization parameter λ and assume we know a LASSO solution $\boldsymbol{w}_0^\star$ at λ_0. Define $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, and $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$. Then any feasible point $\boldsymbol{\theta}$ for the dual problem $\mathcal{D}(\lambda)$ is in the half-sapce*

$$\boldsymbol{\eta}^T (\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0.$$

Proof: A dual point $\boldsymbol{\theta}$ is feasible for $\mathcal{D}(\lambda)$, if $\boldsymbol{\theta}$ satisfies the following inequalities

$$-\lambda \leq \boldsymbol{x}_i^T \boldsymbol{\theta} \leq \lambda, \ i = 1, ..., n. \tag{B.8}$$

We start by computing the term $\boldsymbol{\eta}^T \boldsymbol{\theta}$,

$$\begin{aligned} \boldsymbol{\eta}^T \boldsymbol{\theta} &= \frac{1}{\|\boldsymbol{w}_0\|_1} \boldsymbol{w}_0^T \boldsymbol{X}^T \boldsymbol{\theta}, \\ &= \frac{1}{\|\boldsymbol{w}_0\|_1} \sum_{i \in \mathcal{A}} \boldsymbol{w}_0(i) \boldsymbol{x}_i^T \boldsymbol{\theta}, \\ &= \frac{1}{\|\boldsymbol{w}_0\|_1} \sum_{i \in \mathcal{A}} |\boldsymbol{w}_0(i)| \operatorname{sign}(\boldsymbol{w}_0(i)) \boldsymbol{x}_i^T \boldsymbol{\theta}. \end{aligned}$$

Using (B.8), we have

$$\text{sign}(\boldsymbol{w}_0(i))\boldsymbol{x}_i^T\boldsymbol{\theta} \geq -\lambda,$$

and

$$\begin{aligned}\boldsymbol{\eta}^T\boldsymbol{\theta} &\geq \frac{1}{\|\boldsymbol{w}_0\|_1}\sum_{i\in\mathcal{A}}|\boldsymbol{w}_0(i)|\,(-\lambda),\\ &= -\lambda.\end{aligned}$$

Finally, we prove that $\boldsymbol{\eta}^T\boldsymbol{\theta}_s = -\lambda$, we recall the optimality conditions from Theorem 2.2.1,

$$\boldsymbol{x}_i^T\boldsymbol{\theta}_0 = \lambda_0,\ i \in \mathcal{A}^- \implies \boldsymbol{w}_0(i) \leq 0,$$

and

$$\boldsymbol{x}_i^T\boldsymbol{\theta}_0 = -\lambda_0,\ i \in \mathcal{A}^+ \implies \boldsymbol{w}_0(i) \geq 0.$$

We combine both conditions into one compact form,

$$\boldsymbol{x}_i^T\boldsymbol{\theta}_0 = -\text{sign}(\boldsymbol{w}_0(i))\lambda_0,\ i \in \mathcal{A} = \mathcal{A}^- \cup \mathcal{A}^+,$$

and write

$$\begin{aligned}\boldsymbol{\eta}^T\boldsymbol{\theta}_s &= \frac{1}{\|\boldsymbol{w}_0\|_1}\boldsymbol{w}_0^T\boldsymbol{X}^T\boldsymbol{\theta}_s,\\ &\quad \frac{1}{\|\boldsymbol{w}_0\|_1}\boldsymbol{w}_0^T\boldsymbol{X}^T\left(\frac{\lambda}{\lambda_0}\boldsymbol{\theta}_0\right),\\ &= \frac{\lambda}{\lambda_0\|\boldsymbol{w}_0\|_1}\sum_{i\in\mathcal{A}}\boldsymbol{w}_0(i)\boldsymbol{x}_i^T\boldsymbol{\theta}_0,\\ &= \frac{-\lambda}{\lambda_0\|\boldsymbol{w}_0\|_1}\sum_{i\in\mathcal{A}}\boldsymbol{w}_0(i)\text{sign}(\boldsymbol{w}_0(i))\lambda_0,\\ &= \frac{-\lambda}{\|\boldsymbol{w}_0\|_1}\sum_{i\in\mathcal{A}}|\boldsymbol{w}_0(i)|\,,\\ &= -\lambda\ .\end{aligned}$$

Thus, if $\boldsymbol{\theta}$ is feasible for $\mathcal{D}(\lambda)$, then $\boldsymbol{\theta}$ is in the half-space defined by

$$\boldsymbol{\eta}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_s) \geq 0.$$

■

Proposition B.0.3 *Consider the SAFE test problem $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ in (B.1), and assume we know a LASSO solution $\boldsymbol{w}_0$ at a regularization parameter λ_0. Defining $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0 / \|\boldsymbol{w}_0\|_1$, $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0 \frac{\lambda}{\lambda_0}$, and $\gamma = G(\boldsymbol{\theta}_s)$, the SAFE test problem reduces to*

$$P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{\delta}_0(k) + \frac{1}{\lambda_0} \|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2 |\lambda - \lambda_0| & \boldsymbol{\sigma}_1(k)\lambda \leq \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{\alpha_0 - \lambda\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) - \boldsymbol{\delta}_0(k) + \boldsymbol{\psi}(k) M |\lambda - \lambda_0| & \textit{otherwise.} \end{cases}$$

with $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, $\beta_0 = \|\boldsymbol{w}_0\|_1$, $\alpha_0 := \boldsymbol{w}_0^T\boldsymbol{\delta}_0 = \boldsymbol{y}^T\boldsymbol{\tau}$,

$$\boldsymbol{\sigma}_1(k) = \|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \boldsymbol{\beta}_0,$$

$$\boldsymbol{\sigma}_2(k) = \|y\|_2 - \alpha_0 \|\boldsymbol{x}_k\|_2,$$

$$M = \frac{-\beta_0\lambda_0 + \|\boldsymbol{y}\|_2^2 - \alpha_0}{\lambda_0^2} - \frac{\beta_0^2}{\alpha_0 - \beta_0\lambda_0},$$

and

$$\boldsymbol{\psi}(k) = \left(\|\boldsymbol{x}_k\|_2^2 - \frac{1}{\alpha_0 - \beta_0\lambda_0}\boldsymbol{\delta}_1^2(k) \right)^{1/2}, \; k = 1, ...n.$$

Proof: From Proposition B.0.1, the SAFE test problem $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$, takes the closed form solution,

$$P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{y}^T\boldsymbol{x} + \|\boldsymbol{x}\|_2 D & \|\boldsymbol{x}\|_2 \left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D \left(\boldsymbol{\eta}^T\boldsymbol{x}\right), \\ \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T\boldsymbol{x}\right) \boldsymbol{\eta}^T\boldsymbol{g}_s - \boldsymbol{x}^T\boldsymbol{y} + \psi\tilde{D} & \text{otherwise,} \end{cases}$$

with $\boldsymbol{g}_s = \boldsymbol{\theta}_s + \boldsymbol{y}$,

$$D = \left(\|\boldsymbol{y}\|_2^2 - 2\gamma\right)^{1/2},$$

$$\tilde{D} = \left(-2\gamma - \frac{\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right)^2}{\|\boldsymbol{\eta}\|_2^2} + \|\boldsymbol{y}\|_2^2 \right)^{1/2},$$

and

$$\psi = \left(\|\boldsymbol{x}\|_2^2 - \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T\boldsymbol{x}\right)^2 \right)^{1/2}.$$

Simplifying $P(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$. We express the different variables appearing in the closed form solution in terms of $\boldsymbol{\tau}$, $\boldsymbol{\delta}_0$, $\boldsymbol{\delta}_1$, β_0, and α_0. We start by evaluating the second case of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$, and we call

$$\boldsymbol{P}_2(k) = \frac{1}{\|\boldsymbol{\eta}\|_2^2} \left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \boldsymbol{x}_k^T\boldsymbol{\eta} - \boldsymbol{x}_k^T\boldsymbol{y} + \psi\tilde{D}.$$

We simplify each term appearing in the above expression,

$$\boldsymbol{X}^T\boldsymbol{\eta} = \frac{1}{\beta_0}\boldsymbol{\delta}_1,$$

$$\begin{aligned}\boldsymbol{\eta}^T\boldsymbol{g}_s &= \boldsymbol{\eta}^T\left(\boldsymbol{\theta}_0\frac{\boldsymbol{\lambda}}{\boldsymbol{\lambda}_0}+\boldsymbol{y}\right),\\ &= -\lambda+\frac{\boldsymbol{\alpha}_0}{\boldsymbol{\beta}_0},\end{aligned}$$

$$\begin{aligned}\|\boldsymbol{\tau}\|_2^2 &= \boldsymbol{\tau}^T\boldsymbol{\tau},\\ &= \boldsymbol{\tau}^T\left(\boldsymbol{\theta}_0+\boldsymbol{y}\right),\\ &= -\beta_0\lambda_0+\boldsymbol{\tau}^T\boldsymbol{y},\\ &= -\beta_0\lambda_0+\alpha_0,\end{aligned}$$

and we obtain

$$\boldsymbol{P}_2(k) = \frac{\alpha_0-\lambda\beta_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)-\boldsymbol{\delta}_0(k)+\boldsymbol{\psi}\tilde{D}.$$

We express the values of $\boldsymbol{\psi}=(\psi_1,...,\psi_n)$ as,

$$\boldsymbol{\psi}(k) = \left(\|\boldsymbol{x}_k\|_2^2-\frac{1}{\alpha_0-\beta_0\lambda_0}\boldsymbol{\delta}_1^2(k)\right)^{1/2},\ k=1,...n.$$

We simplify $\tilde{D}$,

$$\begin{aligned}\tilde{D} &= -2\gamma-\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right)^2+\|\boldsymbol{y}\|_2^2,\\ &= \frac{\|\boldsymbol{\theta}_0\|_2^2}{\lambda_0^2}\lambda^2+2\frac{\boldsymbol{y}^T\boldsymbol{\theta}_0}{\lambda_0}\lambda-\frac{1}{\|\boldsymbol{\eta}\|_2^2}\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right)^2+\|\boldsymbol{y}\|_2^2,\\ &= \left(\frac{\|\boldsymbol{\theta}_0\|_2^2}{\lambda_0^2}-\frac{1}{\|\boldsymbol{\eta}\|_2^2}\right)\lambda^2+2\left(\frac{\boldsymbol{y}^T\boldsymbol{\theta}_0}{\lambda_0}+\frac{\left(\boldsymbol{\eta}^T\boldsymbol{y}\right)}{\|\boldsymbol{\eta}\|_2^2}\right)\lambda-\frac{\left(\boldsymbol{\eta}^T\boldsymbol{y}\right)^2}{\|\boldsymbol{\eta}\|_2^2}+\|\boldsymbol{y}\|_2^2,\\ &= \left(\frac{-\beta_0\lambda_0+\|\boldsymbol{y}\|_2^2-\alpha_0}{\lambda_0^2}-\frac{\beta_0^2}{\|\boldsymbol{\tau}\|_2^2}\right)\lambda^2+2\left(\frac{\alpha_0}{\lambda_0}+\frac{\beta_0}{\|\boldsymbol{\tau}\|_2^2}\alpha_0-\frac{\|\boldsymbol{y}\|_2^2}{\lambda_0}\right)\lambda\\ &\quad -\frac{\left(\alpha_0\right)^2}{\|\boldsymbol{\tau}\|_2^2}+\|\boldsymbol{y}\|_2^2,\end{aligned}$$

and reduce it to

$$\tilde{D} = M\left(\lambda-\lambda_0\right)^2,$$

with

$$M = \frac{-\beta_0\lambda_0 + \|\boldsymbol{y}\|_2^2 - \alpha_0}{\lambda_0^2} - \frac{\beta_0^2}{\alpha_0 - \beta_0\lambda_0}.$$

The value of $\boldsymbol{P}_2$ is

$$\boldsymbol{P}_2 = \frac{\alpha_0 - \lambda\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1 - \boldsymbol{\delta}_0 + \boldsymbol{\psi}M\left|\lambda - \lambda_0\right|.$$

Similarly we define

$$\boldsymbol{P}_1(k) = -\boldsymbol{y}^T\boldsymbol{x}_k + \|\boldsymbol{x}_k\|_2 D,\ k = 1, ..., n,$$

and express it as

$$\boldsymbol{P}_1(k) = -\boldsymbol{\delta}_0(k) + \frac{1}{\lambda_0}\|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2\left|\lambda - \lambda_0\right|,\ k = 1, ..., n.$$

Simplifying the condition $\|\boldsymbol{x}_k\|_2\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D\left(\boldsymbol{\eta}^T\boldsymbol{x}_k\right)$. The condition $\|\boldsymbol{x}_k\|_2\left(\boldsymbol{\eta}^T\boldsymbol{g}_s\right) \leq D\left(\boldsymbol{\eta}^T\boldsymbol{x}_k\right)$, translates to

$$\begin{aligned}
\|\boldsymbol{x}_k\|_2\left(-\lambda + \frac{\boldsymbol{\alpha}_0}{\boldsymbol{\beta}_0}\right) &\leq \|y\|_2\left(1 - \frac{\lambda}{\lambda_0}\right)\frac{1}{\beta_0}\boldsymbol{\delta}_1(k),\\
\lambda_0\|\boldsymbol{x}_k\|_2\left(-\lambda\boldsymbol{\beta}_0 + \alpha_0\right) &\leq \|y\|_2\left(\lambda_0 - \lambda\right)\boldsymbol{\delta}_1(k),\\
\lambda\left(\|y\|_2\boldsymbol{\delta}_1(k) - \lambda_0\|\boldsymbol{x}_k\|_2\boldsymbol{\beta}_0\right) &\leq \|y\|_2\lambda_0 - \alpha_0\lambda_0\|\boldsymbol{x}_k\|_2,\\
\lambda\left(\|y\|_2\boldsymbol{\delta}_1(k) - \lambda_0\|\boldsymbol{x}_k\|_2\boldsymbol{\beta}_0\right) &\leq \lambda_0\left(\|y\|_2 - \alpha_0\|\boldsymbol{x}_k\|_2\right).
\end{aligned}$$

We call

$$\begin{aligned}
\boldsymbol{\sigma}_1(k) &= \|y\|_2\boldsymbol{\delta}_1(k) - \lambda_0\|\boldsymbol{x}_k\|_2\boldsymbol{\beta}_0,\\
\boldsymbol{\sigma}_2(k) &= \|y\|_2 - \alpha_0\|\boldsymbol{x}_k\|_2,
\end{aligned}$$

and the final expression of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$, reduces to

$$P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma) = \begin{cases} -\boldsymbol{\delta}_0(k) + \frac{1}{\lambda_0}\|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2\left|\lambda - \lambda_0\right| & \boldsymbol{\sigma}_1(k)\lambda \leq \boldsymbol{\sigma}_2(k)\lambda_0,\\ \frac{\alpha_0 - \lambda\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) - \boldsymbol{\delta}_0(k) + \boldsymbol{\psi}(k)M\left|\lambda - \lambda_0\right| & \text{otherwise.}\end{cases}$$

■

Proposition B.0.4 *Consider the LASSO (2.1) with feature matrix $\boldsymbol{X}$ and response $\boldsymbol{y}$. Also assume we know an optimal solution $\boldsymbol{w}_0$ at some regularization parameter λ_0, then for $\lambda \leq \lambda_0$ the test*

$$\lambda < max(P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma), P(-\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)),$$

with $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ the safe test problem in (B.1),$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{w}_0/\|\boldsymbol{w}_0\|_1$, $\boldsymbol{\theta}_0 = \boldsymbol{X}\boldsymbol{w}_0 - \boldsymbol{y}$, $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0\frac{\lambda}{\lambda_0}$, $\gamma = G(\boldsymbol{\theta}_s)$ and $G(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\theta} + \boldsymbol{y}\|_2^2$ is equivalent to

$$\lambda > \rho_k\lambda_0,$$

with $\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{w}_0$, $\boldsymbol{\delta}_0 = \boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\delta}_1 = \boldsymbol{X}^T\boldsymbol{\tau}$, $\beta_0 = \|\boldsymbol{w}_0\|_1$, $\alpha_0 := \boldsymbol{w}_0^T\boldsymbol{\delta}_0 = \boldsymbol{y}^T\boldsymbol{\tau}$,

$$\boldsymbol{\sigma}_1^+(k) = \|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \boldsymbol{\beta}_0,$$

$$\boldsymbol{\sigma}_1^-(k) = -\|y\|_2 \boldsymbol{\delta}_1(k) - \lambda_0 \|\boldsymbol{x}_k\|_2 \boldsymbol{\beta}_0,$$

$$\boldsymbol{\sigma}_2(k) = \|y\|_2 - \alpha_0 \|\boldsymbol{x}_k\|_2,$$

$$M = \frac{-\beta_0\lambda_0 + \|\boldsymbol{y}\|_2^2 - \alpha_0}{\lambda_0^2} - \frac{\beta_0^2}{\alpha_0 - \beta_0\lambda_0},$$

$$\boldsymbol{\psi}(k) = \left(\|\boldsymbol{x}_k\|_2^2 - \frac{1}{\alpha_0 - \beta_0\lambda_0}\boldsymbol{\delta}_1^2(k)\right)^{1/2}, \ k = 1, ...n.,$$

$$\rho_k = \max(\rho_k^-, \rho_k^+),$$

$$\rho_k^+ = \begin{cases} \frac{-\boldsymbol{\delta}_0(k) + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2}{\lambda_0 + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2} & \boldsymbol{\sigma}_1^+(k)\lambda \le \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{\left(-\boldsymbol{\delta}_0(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \boldsymbol{\psi}(k)M\lambda_0}{\left(-\boldsymbol{\delta}_1(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \lambda_0 + \boldsymbol{\psi}(k)M\lambda_0} & \textit{otherwise}, \end{cases}$$

and

$$\rho_k^- = \begin{cases} \frac{+\boldsymbol{\delta}_0(k) + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2}{\lambda_0 + \|\boldsymbol{y}\|_2\|\boldsymbol{x}_k\|_2} & \boldsymbol{\sigma}_1^-(k)\lambda \le \boldsymbol{\sigma}_2(k)\lambda_0, \\ \frac{-\left(-\boldsymbol{\delta}_0(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \boldsymbol{\psi}(k)M\lambda_0}{-\left(-\boldsymbol{\delta}_1(k) + \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right) + \lambda_0 + \boldsymbol{\psi}(k)M\lambda_0} & \textit{otherwise}. \end{cases}$$

Proof: We start by evaluating

$$\lambda > -\boldsymbol{\delta}_0(k) + \frac{1}{\lambda_0}\|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2 (\lambda_0 - \lambda),$$

the first case of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$. The above expression can be reduced to

$$\frac{\lambda}{\lambda_0} > \frac{-\boldsymbol{\delta}_0(k) + \|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2}{\lambda_0 + \|\boldsymbol{y}\|_2 \|\boldsymbol{x}_k\|_2}.$$

The other case of $P(\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ is

$$\lambda > \frac{\alpha_0 - \lambda\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) - \boldsymbol{\delta}_0(k) + \boldsymbol{\psi}(k)M(\lambda_0 - \lambda).$$

We reduce the above expression,

$$\lambda\left(1 + \frac{\beta_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) + \boldsymbol{\psi}(k)M\right) > \frac{\alpha_0}{\alpha_0 - \lambda_0\beta_0}\boldsymbol{\delta}_1(k) - \boldsymbol{\delta}_0(k) + \boldsymbol{\psi}(k)M\lambda_0,$$

$$\lambda\left(\lambda_0+\frac{\beta_0\lambda_0-\alpha_0+\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)+\boldsymbol{\psi}(k)M\lambda_0\right)>\frac{\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)-\boldsymbol{\delta}_0(k)+\boldsymbol{\psi}(k)M\lambda_0,$$

$$\lambda\left(\lambda_0+\frac{\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)-\boldsymbol{\delta}_1(k)+\boldsymbol{\psi}(k)M\lambda_0\right)>\frac{\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)-\boldsymbol{\delta}_0(k)+\boldsymbol{\psi}(k)M\lambda_0,$$

and we obtain

$$\frac{\lambda}{\lambda_0}>\frac{\left(-\boldsymbol{\delta}_0(k)+\frac{\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right)+\boldsymbol{\psi}(k)M\lambda_0}{\left(-\boldsymbol{\delta}_1(k)+\frac{\alpha_0}{\alpha_0-\lambda_0\beta_0}\boldsymbol{\delta}_1(k)\right)+\lambda_0+\boldsymbol{\psi}(k)M\lambda_0}.$$

Similarly, we express $\lambda < P(-\boldsymbol{x}_k, \boldsymbol{\eta}, \boldsymbol{\theta}_s, \gamma)$ by flipping the signs of $\boldsymbol{\delta}_0(k)$ and $\boldsymbol{\delta}_1(k)$. ■

Appendix C

Expression of $P(\gamma, x)$, general case

We show that the quantity $P(\gamma, x)$ defined in (5.6) can be expressed in dual form (5.7). This is a simple consequence of duality:

$$\begin{aligned}
P(\gamma, x) &= \max_\theta\ \theta^T x\ :\ G(\theta) \geq \gamma,\ \ \theta^T b = 0 \\
&= \max_\theta\ \min_{\mu>0,\,\nu}\ \theta^T x + \mu(G(\theta) - \gamma) - \nu\theta^T b \\
&= \min_{\mu>0,\,\nu}\ \max_\theta\ \theta^T x + \mu(-y^T\theta - \textstyle\sum_{i=1}^m f^*(\theta(i)) - \gamma) - \nu\theta^T b \\
&= \min_{\mu>0,\,\nu}\ -\gamma\mu + \max_\theta\ \theta^T(x - \mu y - \nu z) - \mu\textstyle\sum_{i=1}^m f^*(\theta(i)) \\
&= \min_{\mu>0,\,\nu}\ -\gamma\mu + \mu\left(\max_\theta\ \tfrac{1}{\mu}\theta^T(x - \mu y - \nu z) - \textstyle\sum_{i=1}^m f^*(\theta(i))\right) \\
&= \min_{\mu>0,\,\nu}\ -\gamma\mu + \mu\textstyle\sum_{i=1}^m f\left(\frac{x_i - \mu y(i) - \nu b_i}{\mu}\right).
\end{aligned}$$

Appendix D

SAFE test for SVM

In this appendix, we examine various optimization problems involving polyhedral functions in one or two variables, which arise in section 5.3.1 for the computation of $P_{\rm hi}(\gamma, x)$ as well as in the SAFE-SVM theorem of section 5.3.2.

D.1 Computing $P_{\rm hi}(\gamma, x)$

We first focus on the specific problem of computing the quantity defined in (5.11). To simplify notation, we will consider the problem of computing $P_{\rm hi}(\gamma, -x)$, that is:

$$P_{\rm hi}(\gamma, -x) = \min_{\mu\geq 0,\, \nu} \; -\gamma\mu + \sum_{i=1}^{m} (\mu + \nu y_i + x_i)_+, \tag{D.1}$$

where $y \in \{-1, 1\}^m$, $x \in \mathbb{R}^m$ and γ are given, with $0 \leq \gamma \leq \gamma_0 := 2\min(m_+, m_-)$. Here, $\mathcal{I}_\pm := \{i \,:\, y_i = \pm 1\}$, and $x^+ = (x_i)_{i\in\mathcal{I}_+}$, $x^- = (x_i)_{i\in\mathcal{I}_-}$, $m_\pm = |\mathcal{I}_\pm|$, and $\underline{m} = \min(m_+, m_-)$. Without loss of generality, we assume that both x^+, x^- are both sorted in descending order: $x_1^\pm \geq \ldots \geq x_{m_\pm}^\pm$.

Using $\alpha = \mu + \nu$, $\beta = \mu - \nu$, we have

$$\begin{array}{rcl} P_{\rm hi}(\gamma, -x) & = & \min_{\alpha+\beta\geq 0} \; -\frac{\gamma}{2}(\alpha+\beta) + \sum_{i=1}^{m_+}(x_i^+ + \alpha)_+ \\ & & + \sum_{i=1}^{m_-}(x_i^- + \beta)_+, \\ & = & \min_{\alpha,\,\beta} \; \max_{t\geq 0} \; -\frac{\gamma}{2}(\alpha+\beta) + \sum_{i=1}^{m_+}(x_i^+ + \alpha)_+ \\ & & + \sum_{i=1}^{m_-}(x_i^- + \beta)_+ - t(\alpha+\beta), \\ & = & \max_{t\geq 0} \; \min_{\alpha,\,\beta} \; -(\frac{\gamma}{2}+t)(\alpha+\beta) \\ & & + \sum_{i=1}^{m_+}(x_i^+ + \alpha)_+ + \sum_{i=1}^{m_-}(x_i^- + \beta)_+, \\ & = & \max_{t\geq 0} \; F(\frac{\gamma}{2}+t, x^+) + F(\frac{\gamma}{2}+t, x^-), \end{array} \tag{D.2}$$

where, for $h \in \mathbb{R}$ and $x \in \mathbb{R}^p$, $x_1 \geq \ldots \geq x_p$, we set

$$F(h, x) := \min_z \; -hz + \sum_{i=1}^{p}(z + x_i)_+, \tag{D.3}$$

Expression of the function F. If $h > p$, then with $z \to +\infty$ we obtain $F(h, x) = -\infty$. Similarly, if $h < 0$, then $z \to -\infty$ yields $F(h, x) = -\infty$. When $0 \le h \le p$, we proceed by expressing F in dual form:

$$F(h, x) = \max_u \; u^T x \;:\; 0 \le u \le \mathbf{1}, \;\; u^T \mathbf{1} = h.$$

If $h = p$, then the only feasible point is $u = \mathbf{1}$, so that $F(p, x) = \mathbf{1}^T x$. If $0 \le h < 1$, choosing $u_1 = h$, $u_2 = \ldots = u_p = 0$, we obtain the lower bound $F(h, x) \ge h x_1$, which is attained with $z = -x_1$.

Assume now that $1 \le h < p$. Let $h = q + r$, with $q = \lfloor h \rfloor$ the integer part of h, and $0 \le r < 1$. Choosing $u_1 = \ldots = u_q = 1$, $u_{q+1} = r$, we obtain the lower bound

$$F(h, x) \ge \sum_{j=1}^{q} x_j + r x_{q+1},$$

which is attained by choosing $z = -x_{q+1}$ in the expression (D.3).

To summarize:

$$F(h, x) = \begin{cases} h x_1 & \text{if } 0 \le h < 1, \\ \sum_{j=1}^{\lfloor h \rfloor} x_j + (h - \lfloor h \rfloor) x_{\lfloor h \rfloor + 1} & \text{if } 1 \le h < p, \\ \sum_{j=1}^{p} x_j & \text{if } h = p, \\ -\infty & \text{otherwise.} \end{cases} \tag{D.4}$$

A more compact expression, valid for $0 \le h \le p$ if we set $x_{p+1} = x_p$ and assume that a sum over an empty index sets is zero, is

$$F(h, x) = \sum_{j=1}^{\lfloor h \rfloor} x_j + (h - \lfloor h \rfloor) x_{\lfloor h \rfloor + 1}, \;\; 0 \le h \le p.$$

Note that $F(\cdot, x)$ is the piece-wise linear function that interpolates the sum of the h largest elements of x at the integer break points $h = 0, \ldots, p$.

Expression of $P_{\text{hi}}(\gamma, -x)$. We start with the expression found in (D.2):

$$P_{\text{hi}}(\gamma, -x) = \max_{t \ge 0} \; F(\frac{\gamma}{2} + t, x^+) + F(\frac{\gamma}{2} + t, x^-).$$

Since the domain of $F(\cdot, x^+) + F(\cdot, x^-)$ is $[0, \underline{m}]$, and with $0 \le \gamma/2 \le \gamma_0/2 = \underline{m}$, we get

$$P_{\text{hi}}(\gamma, -x) = \max_{\gamma/2 \le h \le \underline{m}} G(h, x^+, x^-) := F(h, x^+) + F(h, x^-).$$

Since $F(\cdot, x)$ with $x \in \mathbb{R}^p$ is a piece-wise linear function with break points at $0, \ldots, p$, a maximizer of $G(\cdot, x^+, x^-)$ over $[\gamma/2, \underline{m}]$ lies in $\{\gamma/2, \lfloor \gamma/2 \rfloor + 1, \ldots, \underline{m}\}$. Thus,

$$P_{\rm hi}(\gamma, -x) = \max\left(G(\frac{\gamma}{2}, x^+, x^-), \max_{h \in \{\lfloor \gamma/2 \rfloor + 1, \ldots, \underline{m}\}} G(h, x^+, x^-) \right).$$

Let us examine the second term, and introduce the notation $\bar{x}_j := x_j^+ + x_j^-$, $j = 1, \ldots, \underline{m}$:

$$\begin{aligned} \max_{h \in \{\lfloor \gamma/2 \rfloor + 1, \ldots, \underline{m}\}} G(h, x^+, x^-) &= \max_{h \in \{\lfloor \gamma/2 \rfloor + 1, \ldots, \underline{m}\}} \sum_{j=1}^{h} (x_j^+ + x_j^-) \\ &= \sum_{j=1}^{\lfloor \gamma/2 \rfloor + 1} \bar{x}_j + \sum_{j=\lfloor \gamma/2 \rfloor + 2}^{\underline{m}} (\bar{x}_j)_+, \end{aligned}$$

with the convention that sums over empty index sets are zero. Since

$$G(\frac{\gamma}{2}, x^+, x^-) = \sum_{j=1}^{\lfloor \gamma/2 \rfloor} \bar{x}_j + (\frac{\gamma}{2} - \lfloor \frac{\gamma}{2} \rfloor) \bar{x}_{\lfloor \gamma/2 \rfloor + 1},$$

we obtain

$$P_{\rm hi}(\gamma, -x) = \sum_{j=1}^{\lfloor \gamma/2 \rfloor} \bar{x}_j + \max\left((\frac{\gamma}{2} - \lfloor \frac{\gamma}{2} \rfloor) \bar{x}_{\lfloor \gamma/2 \rfloor + 1}, \bar{x}_{\lfloor \gamma/2 \rfloor + 1} + \sum_{j=\lfloor \gamma/2 \rfloor + 2}^{\underline{m}} (\bar{x}_j)_+ \right).$$

An equivalent expression is:

$$\begin{aligned} P_{\rm hi}(\gamma, -x) &= \textstyle\sum_{j=1}^{\lfloor \gamma/2 \rfloor} \bar{x}_j - (\frac{\gamma}{2} - \lfloor \frac{\gamma}{2} \rfloor)(-\bar{x}_{\lfloor \gamma/2 \rfloor + 1})_+ + \sum_{j=\lfloor \gamma/2 \rfloor + 1}^{\underline{m}} (\bar{x}_j)_+, \quad 0 \le \gamma \le 2\underline{m}, \\ &\quad \bar{x}_j := x_j^+ + x_j^-, \quad j = 1, \ldots, \underline{m}. \end{aligned}$$

The function $P_{\rm hi}(\cdot, -x)$ linearly interpolates the values obtained for $\gamma = 2q$ with q integer in $\{0, \ldots, \underline{m}\}$:

$$P_{\rm hi}(2q, -x) = \sum_{j=1}^{q} \bar{x}_j + \sum_{j=q+1}^{\underline{m}} (\bar{x}_j)_+.$$

D.2 Computing $\Phi(x^+, x^-)$

Let us consider the problem of computing

$$\Phi(x^+, x^-) := \min_{\nu} \sum_{i=1}^{m_+} (x_i^+ + \nu)_+ + \sum_{i=1}^{m_-} (x_i^- - \nu)_+,$$

with $x^{\pm} \in \mathbb{R}^{m_\pm}$, $x_1^{\pm} \geq \ldots \geq x_{m_\pm}^{\pm}$, given. We can express $\Phi(x^+, x^-)$ in terms of the function F defined in (D.3):

$$\begin{aligned}
\Phi(x^+, x^-) &= \min_{\nu_+,\nu_-} \textstyle\sum_{i\in\mathcal{I}_+}(x_i^+ + \nu^+)_+ \\
&\quad + \textstyle\sum_{i\in\mathcal{I}_-}(x_i^- - \nu^-)_+ \ : \ \nu^+ = \nu^-, \\
&= \max_h \ \min_{\nu^+,\nu^-} \ -h(\nu^+ - \nu^-) \\
&\quad + \textstyle\sum_{i\in\mathcal{I}_+}(x_i^+ + \nu^+)_+ + \sum_{i\in\mathcal{I}_-}(x_i^- - \nu^-)_+, \\
&= \max_h \ \min_{\nu^+,\nu^-} \ -h\nu^+ + \textstyle\sum_{i\in\mathcal{I}_+}(x_i^+ + \nu^+)_+ \\
&\quad + h\nu^- + \textstyle\sum_{i\in\mathcal{I}_-}(x_i^- - \nu^-)_+, \\
&= \max_h \ \Big(\min_\nu \ -h\nu + \textstyle\sum_{i\in\mathcal{I}_+}(x_i^+ + \nu)_+\Big) \\
&\quad + \Big(\min_\nu \ -h\nu + \textstyle\sum_{i\in\mathcal{I}_-}(x_i^- + \nu)_+\Big) \quad (\nu_+ = -\nu_- = \nu), \\
&= \max_h \ F(h, x^+) + F(h, x^-), \\
&= \max_{0\leq h\leq \underline{m}} \ F(h, x^+) + F(h, x^-), \\
&= \max(A, B, C),
\end{aligned}$$

where F is defined in (D.3), and

$$A = \max_{0\leq h<1} F(h, x^+) + F(h, x^-), \quad B := \max_{1\leq h<\underline{m}} F(h, x^+) + F(h, x^-)), \quad C = F(\underline{m}, x^+) + F(\underline{m}, x^-).$$

We have

$$A := \max_{0\leq h<1} \ F(h, x^+) + F(h, x^-) = \max_{0\leq h<1} \ h(x_1^+ + x_1^-) = (x_1^+ + x_1^-)_+.$$

Next:

$$\begin{aligned}
B &= \max_{1\leq h<\underline{m}} \ F(h, x^+) + F(h, x^-) \\
&= \max_{q\in\{1,\ldots,\underline{m}-1\}, r\in[0,1[} \ \sum_{i=1}^{q}(x_i^+ + x_i^-) + r(x_{q+1}^+ + x_{q+1}^-) \\
&= \max_{q\in\{1,\ldots,\underline{m}-1\}} \sum_{i=1}^{q}(x_i^+ + x_i^-) + (x_{q+1}^+ + x_{q+1}^-)_+ \\
&= (x_1^+ + x_1^-) + \sum_{i=2}^{\underline{m}}(x_i^+ + x_i^-)_+.
\end{aligned}$$

Observe that

$$B \geq C = \sum_{i=1}^{\underline{m}}(x_i^+ + x_i^-).$$

Moreover, if $(x_1^+ + x_1^-) \geq 0$, then $B = \sum_{i=1}^{\underline{m}}(x_i^+ + x_i^-)_+ \geq A$. On the other hand, if $x_1^+ + x_1^- \leq 0$, then $x_i^+ + x_i^- \leq 0$ for $2 \leq j \leq \underline{m}$, and $A = \sum_{i=1}^{\underline{m}}(x_i^+ + x_i^-)_+ \geq x_1^+ + x_1^- = B$.

In all cases,

$$\Phi(x^+, x^-) = \max(A, B, C) = \sum_{i=1}^{\underline{m}} (x_i^+ + x_i^-)_+.$$

D.3 SAFE-SVM test

Now we consider the problem that arises in the SAFE-SVM test (5.14):

$$G(z) := \min_{0 \le \kappa \le 1} \sum_{i=1}^{p} (1 - \kappa + \kappa z_i)_+,$$

where $z \in \mathbb{R}^p$ is given. (The SAFE-SVM condition (5.14) involves $z_i = \gamma_0/(2\lambda_0)(x_{[i]}^+ + x_{[i]}^-)$, $i = 1, \ldots, p := \underline{m}$.) We develop an algorithm to compute the quantity $G(z)$, the complexity of which grows as $O(d \log d)$, where d is (less than) the number of non-zero elements in z.

Define $\mathcal{I}_\pm = \{i \,:\, \pm z_i > 0\}$, $k := |\mathcal{I}_+|$, $h := |\mathcal{I}_-|$, $l = \mathcal{I}_0$, $l := |\mathcal{I}_0|$.

If $k = 0$, $\mathcal{I}_+$ is empty, and $\kappa = 1$ achieves the lower bound of 0 for $G(z)$. If $k > 0$ and $h = 0$, that is, $k + l = p$, then $\mathcal{I}_-$ is empty, and an optimal κ is attained in $\{0, 1\}$. In both cases ($\mathcal{I}_+$ or $\mathcal{I}_-$ empty), we can write

$$G(z) = \min_{\kappa \in \{0,1\}} \sum_{i=1}^{p} (1 - \kappa + \kappa z_i)_+ = \min(p, S_+), \quad S_+ := \sum_{i \in \mathcal{I}_+} z_i,$$

with the convention that a sum over an empty index set is zero.

Next we proceed with the assumption that $k \neq 0$ and $h \neq 0$. Let us re-order the elements of $\mathcal{I}_-$ in decreasing fashion, so that $z_i > 0 = z_{k+1} = \ldots = z_{k+l} > z_{k+l+1} \ge \ldots \ge z_p$, for every $i \in \mathcal{I}_+$. (The case when $\mathcal{I}_0$ is empty is handled simply by setting $l = 0$ in our formula.) We have

$$G(z) = k + l + \min_{0 \le \kappa \le 1} \left\{ \kappa\alpha + \sum_{i=k+l+1}^{p} (1 - \kappa + \kappa z_i)_+ \right\},$$

where, $\alpha := S_+ - k - l$. The minimum in the above is attained at $\kappa = 0, 1$ or one of the break points $1/(1 - z_j) \in (0, 1)$, where $j \in \{k + l + 1, \ldots, p\}$. At $\kappa = 0, 1$, the objective function of the original problem takes the values S_+, p, respectively. The value of the same objective

function at the break point $\kappa = 1/(1-z_j)$, $j = k+l+1, \ldots, p$, is $k+l+G_j(z)$, where

$$\begin{aligned} G_j(z) &:= \frac{\alpha}{1-z_j} + \sum_{i=k+l+1}^{p} \left(\frac{z_i - z_j}{1-z_j}\right)_+ \\ &= \frac{\alpha}{1-z_j} + \frac{1}{1-z_j} \sum_{i=k+l+1}^{j-1} (z_i - z_j) \\ &= \frac{1}{1-z_j} \left(\alpha - (j-k-l-1)z_j + \sum_{i=k+l+1}^{j-1} z_i\right) \\ &= \frac{1}{1-z_j} \left(S_+ - (j-1)z_j - (k+l)(1-z_j) + \sum_{i=k+l+1}^{j-1} z_i\right) \\ &= -(k+l) + \frac{1}{1-z_j} \left(\sum_{i=1}^{j-1} z_i - (j-1)z_j\right). \end{aligned}$$

This allows us to write

$$G(z) = \min\left(p, \sum_{i=1}^{k} z_i, \min_{j \in \{k+l+1,\ldots,p\}} \frac{1}{1-z_j}\left(\sum_{i=1}^{j-1} z_i - (j-1)z_j\right)\right).$$

The expression is valid when $k+l=p$ ($h=0$, $\mathcal{I}_-$ is empty), $l=0$ ($\mathcal{I}_0$ is empty), or $k=0$ ($\mathcal{I}_+$ is empty) with the convention that the sum (resp. minimum) over an empty index set is 0 (resp. $+\infty$).

We can summarize the result with the compact formula:

$$G(z) = \min_z \frac{1}{1-z} \sum_{i=1}^{p} (z_i - z)_+ \ : \ z \in \{-\infty, 0, (z_j)_{j\,:\,z_j<0}\}.$$

Let us detail an algorithm for computing $G(z)$. Assume $h > 0$. The quantity

$$\underline{G}(z) := \min_{k+l+1 \le j \le p} (G_j(z))$$

can be evaluated in less than $O(h)$, via the following recursion:

$$\begin{array}{rcl} G_{j+1}(z) & = & \frac{1-z_j}{1-z_{j+1}} G_j(z) - j\frac{z_{j+1}-z_j}{1-z_{j+1}} \\ \underline{G}_{j+1}(z) & = & \min(\underline{G}_j(z), G_{j+1}(z)) \end{array} \quad , \quad j = k+l+1, \ldots, p, \tag{D.5}$$

with initial values

$$G_{k+l+1}(z) = \underline{G}_{k+l+1}(z) = \frac{1}{1-z_{k+l+1}} \left(\sum_{i=1}^{k+l} z_i - (k+l)z_{k+l+1}\right).$$

On exit, $\underline{G}(z) = \underline{G}_p$.

Our algorithm is as follows.

Algorithm for the evaluation of $G(z)$.

1. Find the index sets $\mathcal{I}_+$, $\mathcal{I}_-$, $\mathcal{I}_0$, and their respective cardinalities k, h, l.
2. If $k = 0$, set $G(z) = 0$ and exit.
3. Set $S_+ = \sum_{i=1}^{k} z_i$.
4. If $h = 0$, set $G(z) = \min(p, S_+)$, and exit.
5. If $h > 0$, order the negative elements of z, and evaluate $\underline{G}(z)$ by the recursion (D.5). Set $G(z) = \min(p, S_+, \underline{G}(z))$ and exit.

The complexity of evaluating $G(z)$ thus grows in $O(k + h \log h)$, which is less than $O(d \log d)$, where $d = k + h$ is the number of non-zero elements in z.

Appendix E

Computing $P_{\log}(\gamma, x)$ via an interior-point method

We consider the problem (5.18) which arises with the logistic loss. We can use a generic interior-point method [5], and exploit the decomposable structure of the dual function $G_{\log}$. The algorithm is based on solving, via a variant of Newton's method, a sequence of linearly constrained problems of the form

$$\min_{\theta} \ \tau x^T\theta + \log(G_{\log}(\theta) - \gamma) + \sum_{i=1}^{m} \log(-\theta - \theta^2) \ : \ z^T\theta = 0,$$

where $\tau > 0$ is a parameter that is increased as the algorithm progresses, and the last terms correspond to domain constraints $\theta \in [-1, 0]^m$. As an initial point, we can take the point θ generated by scaling, as explained in section 5.2.2. Each iteration of the algorithm involves solving a linear system in variable δ, of the form $H\delta = h$, with H is a rank-two modification to the Hessian of the objective function in the problem above. It is easily verified that the matrix H has a "diagonal plus rank-two" structure, that is, it can be written as $H = D - gg^T - vv^T$, where the $m \times m$ matrix D is diagonal and $g, v \in \mathbb{R}^m$ are computed in $O(m)$. The matrix H can be formed, as the associated linear system solved, in $O(m)$ time. Since the number of iterations for this problem with two constraints grows as $\log(1/\epsilon)O(1)$, the total complexity of the algorithm is $\log(1/\epsilon)O(m)$ (ϵ is the absolute accuracy at which the interior-point method computes the objective). We note that memory requirements for this method also grow as $O(m)$.

Appendix F

What is Differential Flatness?

The theory of differential flatness, consists of a parameterization of the trajectories of a system by one of its outputs, called the *flat output* and its derivatives [17]. Let us consider a system $\dot{x} = f(x, u)$, where the state x is in $\mathbb{R}^n$, and the control input u is in $\mathbb{R}^m$. The system is said to be *flat*, and admits z, where $\dim(z) = \dim(u)$, for flat output, and the state x can be parameterized by z and its derivatives. More specifically, the state x can be written as $x = h(z, \dot{z}, ..., z^{(n)})$, and the equivalent dynamics can be written as $u = g(z, \dot{z}, ..., z^{(n+1)})$.

In the context of partial differential equations, the vector x can be thought of as infinite dimensional. The notion of differential flatness extends to this case, and for a differentially flat system of this type, the evolution of x can be parameterized using an input u, which often is the value of x at a given point. A system with a flat output can then be parameterized as a function of this output. This parameterization enables the solution of open loop control problems, if this flat output is the one that needs to be controlled. The open-loop control input can then directly be expressed as a function of the flat output. This parameterization also enables the solution of motion planning problems, where a system is steered from one state to another. Differential flatness is used to investigate the related problem of motion planning for heavy chain systems [45], as well as the Burgers equation [46], the telegraph equation [16], the Stefan equation [12], and the heat equation [32].

Parameterization can be achieved in different ways depending on the type of the problem. Laplace transform is widely used [45], [46], [16] to invert the system. The equations can be transformed back from the Laplace domain to the time domain, thus resulting in the flatness parameterization. Alternative methods can be used to compute the parameterization in the time domain directly. For example, the Cauchy-Kovalevskaya form [32, 31] consists in parameterizing the solution of a partial differential equation in $X(\zeta, t)$, where $\zeta \in [0, 1]$ and $t \in \mathbb{R}^+$, as a power series in space multiplied by time varying coefficients, that is $X(\zeta, t) = \sum_{i=0}^{+\infty} a_i(t)\frac{\zeta^i}{i!}$. Here, $X(\zeta, t)$ is the state of the system and $a_i(t)$ is a time function. The usual approach consists in substituting the Cauchy-Kovalevskaya form in the governing partial differential equation and boundary conditions; a relation between $a_i(t)$ and the flat

output $y(t)$ or its derivatives can then be found, for example, $a_i(t) = y^{(i)}(t)$, where $y^{(i)}(t)$ is the i^{th} derivative of $y(t)$, which leads to the final parameterization, in which $a_i(t)$ is written in terms of the desired output $y(t)$.

Appendix G

How to Impose a Discharge at a Gate?

Once a desired open-loop water discharge is computed, it needs to be imposed at the upstream end of the canal. In open channel flow, it is not easy to impose a water discharge at a gate. Indeed, once a gate is opened or closed, the upstream and downstream water levels at the gate change quickly and modify the water discharge, which is a function of the water levels on both sides of the gate. One possibility would be to use a local slave controller that operates the gate in order to deliver a given water discharge. But due to operational constraints, it is usually not possible to operate the gate at a high sampling rate. As an example, some large gates can not be operated more than few times an hour because of motor constraints, which directly limits the operation of the local controller.

Several methods were developed by hydraulic engineers to perform this control input based on the gate equation G.1, which provides a good model for the flow through the gate [38]. The problem can be described as depicted in Figure G.1. Two pools are interconnected with a hydraulic structure, a submerged orifice (also applicable for more complex structures). The gate opening is to be controlled to deliver a required flow from pool 1 to pool 2.

The hydraulic cross-structure is assumed to be modeled by a static relation between the water discharge through the gate Q, the water levels upstream and downstream of the gate Y_1, Y_2, respectively, and the gate opening W

$$Q = C_d \sqrt{2g} L_g W \sqrt{Y_1 - Y_2}, \tag{G.1}$$

where C_d is a discharge coefficient, L_g is the gate width, and g is the gravitational acceleration. This nonlinear model can be linearized for small deviations q, y_1, y_2, w from the reference water discharge value Q, water levels Y_1, Y_2, and gate opening W, respectively. This linearization leads to the equation

$$q = k_u (y_1 - y_2) + k_w w,$$

where the coefficients k_u and k_w are obtained by differentiating G.1 with respect to Y_1, Y_2, and W, respectively.

Various inversion methods can be applied either to the nonlinear or to the linear model to obtain a gate opening W necessary to deliver a desired water discharge through the gate, usually during a sampling period T_s. The static approximation method assumes constant water levels Y_1 and Y_2 during the gate operation period T_s. This approximation leads to an explicit solution of the gate opening W in the linear model assumption. The characteristic approximation method uses the characteristics for zero slope rectangular frictionless channel to approximate the water levels. The linear version of the model also leads to an explicit expression for the gate opening. The dynamic approximation method uses the linearized Saint-Venant equations to predict the water levels. This method can be thought of a global method, because it considers the global dynamics of the canal to predict the gate opening necessary to deliver the desired flow. In [38], the three methods are compared by simulation and tested by experimentation on the Gignac canal. The dynamic approximation methods has shown to better predict the gate opening necessary to obtain a desired average water discharge [38].

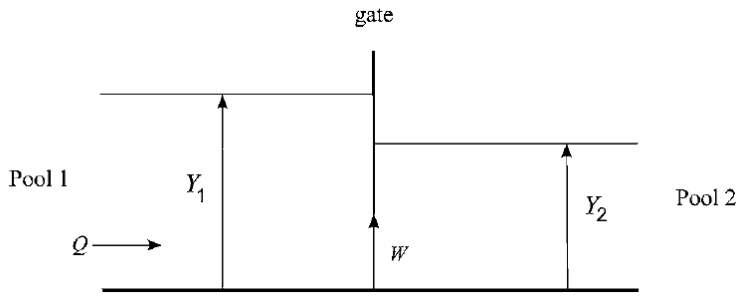


Figure G.1: Gate separating two pools. The gate opening W controls the water flow from Pool 1 to Pool 2. The water discharge can be computed from the water levels Y_1, Y_2, and the gate opening W [38].

Appendix H

Feed-Forward Control of Open Channel Flow Using Differential Flatness

This appendix derives a method for open-loop control of open channel flow, based on the Hayami model, a parabolic partial differential equation resulting from a simplification of the Saint-Venant equations. The open-loop control is represented as infinite series using differential flatness, for which convergence is assessed. A comparison is made with a similar problem available in the literature for thermal systems. Numerical simulations show the effectiveness of the approach by applying the open-loop controller to irrigation canals modeled by the full Saint-Venant equations.

H.1 Introduction

The water resources is a motivation for research on automation of management of water distribution systems. Large amounts of fresh water are lost due to poor management of open-channel systems. This appendix focuses on the management of canals used to convey water from the resource (generally a dam located upstream) to a specific downstream location. Due to the fluctuations of water needs, water demand changes with time. This change in demand calls for the efficient operations of open-channel systems to avoid overflows and to supply desired flow rates at pre-specified time instants.

Automation techniques based on optimization and control have the potential to provide more efficient management strategies than manual techniques. They rely on flow models, in particular the Saint-Venant equations [51] or simplified versions of these equations to describe one-dimensional hydraulic systems. Water level regulation and control of the water flow are among the methods used to improve the efficiency of irrigation systems. These techniques allow engineers to regulate the flow in hydraulic canals and therefore to irrigate

large areas according to user specified demands.

In this appendix, the specific problem of controlling the downstream flow in a one-dimensional hydraulic canal by the upstream discharge is investigated. Several approaches to this problem have already been described in the literature. The majority of these approaches use linear controllers to control the (nonlinear) dynamics of the canal system. Such methods include transfer function analysis for Saint-Venant equations [34] which in turn allows the use of classical control techniques for feedback control [7, 35]. Alternatively, Riemann invariants for hyperbolic conservation laws as in [9, 24] can be used to construct Lyapunov functions, used for stabilization purposes. Adjoint methods [52] have been used for estimation and control, via sensitivity analysis. More closely related to the present study, open-loop control methods have been developed either by computing the solutions of the flow equations backwards using discretization and finite difference methods [3], [2], or using a finite dimensional approximation in the frequency domain [36], [44]. Our approach is to design an open-loop controller which parametrizes the upstream discharge explicitly as a function of the desired downstream discharge at a given location using differential flatness (based on Cauchy-Kovalevskaya series). It can be shown using Lyapunov stability method that the open-loop system is stable [30, 29], which provides another justification for the usefulness of open-loop control of the considered system.

In the context of partial differential equations, differential flatness was used to investigate the related problem of heavy chains motion planning [45], as well as Burgers equation in [46] or the telegraph equation in [16]. The theory of differential flatness, which was first developed in [17], consists in a parametrization of the trajectories of a system by one of its outputs, called the "flat output".

Starting from the classical Saint-Venant equations, widely used, to model unsteady flows in rivers [51], we present a model simplification and a linearization which lead to the Hayami partial differential equation as shown in [41]. The practicality of using the Hayami equation lies in the fact that only two numerical parameters are needed to characterize flow conditions: celerity and diffusivity. The original Saint-Venant equations require the knowledge of the full geometry of the canal and of the roughness coefficient, which make it impractical for long rivers where these parameters are more difficult to infer [33].

The problem of controlling the Hayami equation was already investigated [37] with transfer function analysis, and in [33] for parameter estimation. The Hayami equation [25] is closely linked to the diffusive wave equation with quadratic source terms, which has been studied in [12] and [39]. The difference between our problem and the aforementioned problem is the nature of the boundary conditions: indeed, unlike for heat transfer problems, one cannot impose a value for the downstream discharge (respectively heat flux). In river flow, there are hydraulic structures such as weirs or gates which impose a static relation between water elevation and the flow. In fact, we show that the solution of our problem is a composite of the solution in [39] and an additional new term which captures the boundary condition set by the hydraulic structure, therefore required to solve the specific problem of interest in this study.

The appendix is organized as follows: a description of the physical problem and the system of equations to be solved is first introduced (section H.2). Then, in section H.3, a solution of these equations is derived using differential flatness. The convergence of the infinite series controller is studied and an upper bound on the truncation error is computed as a function of the approximating terms. Moreover, a numerical assessment of the open-loop controller is finally presented and discussed in section H.4. In particular, the difference with controllers synthesized in the context of heat transfer is illustrated through numerical simulation. Applications of the controller on the fully nonlinear Saint-Venant model are presented to show the usefulness of the proposed method for a full nonlinear system.

H.2 Physical Problem

The system of interest is a hydraulic canal of length L. For simplicity, the canal is assumed to have a uniform rectangular cross-section but more complex geometries can easily be taken into account. In this section we present the equations that govern the system, the Saint-Venant equations. We then derive the Hayami model which is a simplification of these equations.

H.2.1 Saint-Venant Equations

The Saint-Venant equations [51] are generally used to describe unsteady flows in rivers or canals [41]. These equations assume one-dimensional flow, with uniform velocity over the cross-section. The effect of boundary friction and turbulence is accounted for through resistance laws such as the Manning-Strickler formula [53], the average channel bed slope is assumed to be small, and the pressure is hydrostatic. Under these assumptions, these equations are written as follows:

$$A_t + Q_x = 0 \tag{H.1}$$

$$Q_t + \left(\frac{Q^2}{A}\right)_x + gA(Y_d)_x = gA(S_b - S_f) \tag{H.2}$$

with $A(x,t)$ the wetted cross-sectional area (m^2), $Q(x,t)$ the discharge (m^3/s) across section $A(x,t)$, $Y_d(x,t)$ the water depth (m), $S_f(x,t)$ the friction slope (m/m), S_b the bed slope (m/m), and g the gravitational acceleration (m^2/s). For rectangular cross sectional geometries, these variables are linked by the following relations: $A(x,t) = Y_d(x,t)B_0$, $Z(x,t) = Y_d(x,t) + S_b(L-x)$ and $Q(x,t) = V(x,t)A(x,t)$ where $Z(x,t)$ is the absolute water elevation (m), $V(x,t)$ is the mean water velocity (m/s) across section $A(x,t)$, and B_0 is the bed width (m). Equation (H.1) is referred to as the *mass conservation equation*, and equation (H.2) is called the *momentum conservation equation*. We assume that there is a cross-structure at the downstream end of the canal, which can be modeled by a static

relation between Q and Z at $x = L$, i.e:

$$Q(L,t) = W(Z(L,t)) \tag{H.3}$$

where $W(\cdot)$ is an analytical function. For a weir structure, this relation can be assumed to be $Q(L,t) = C_w\sqrt{2g}L_w\left(Z(L,t) - Z_w\right)^{3/2}$ where g is the gravitational acceleration, L_w is the weir length, Z_w is the weir elevation, and C_w is the weir discharge coefficient . A similar static relation holds in the case of a gate structure.

H.2.2 Hayami Model

Depending on the characteristics of the river, some terms in the momentum equation (H.2) can be neglected, which allows us to simplify the two equations and to assemble them into a single partial differential equation. As shown in [37], assuming that the inertia terms $Q_t + \left(\frac{Q^2}{A}\right)_x$ can be neglected with respect to $gA(Y_d)_x$ will lead to the diffusive wave model:

$$B_0(Y_d)_t + Q_x = 0 \tag{H.4}$$

$$Z_x = -S_f \tag{H.5}$$

The two equations can be combined and will lead to the diffusive wave equation [33]:

$$Q_t + CQ_x - DQ_{xx} = 0 \tag{H.6}$$

where $Q(x,t)$ is the flow (m^3/s), C and D usually known as the celerity and the diffusivity are non linear functions of the flow. Linearizing equation (H.4) around a reference discharge Q_0 (i.e. $Q(x,t) = Q_0 + q(x,t)$) leads to the Hayami equation:

$$q_t + C_0 q_x - D_0 q_{xx} = 0$$

where $q(x,t)$ is the deviation from the nominal flow Q_0, $C_0(Q_0)$ and $D_0(Q_0)$ are the nominal celerity and diffusivity which depend on Q_0. We call Z_0 the reference elevation, and assume that $Z(x,t) = Z_0 + z(x,t)$, therefore equation (H.4) can be linearized as follows:

$$B_0 z_t + q_x = 0$$

where we have substituted $(Y_d)_t$ by $\left(Z - S_b(L-x)\right)_t = Z_t$ before linearizing. The right boundary condition (H.3) is also linearized and becomes:

$$q(L,t) = bz(L,t)$$

where b is the linearization constant (m^2/s). The value of this constant depends on the weir geometry: length, height, and discharge coefficient.

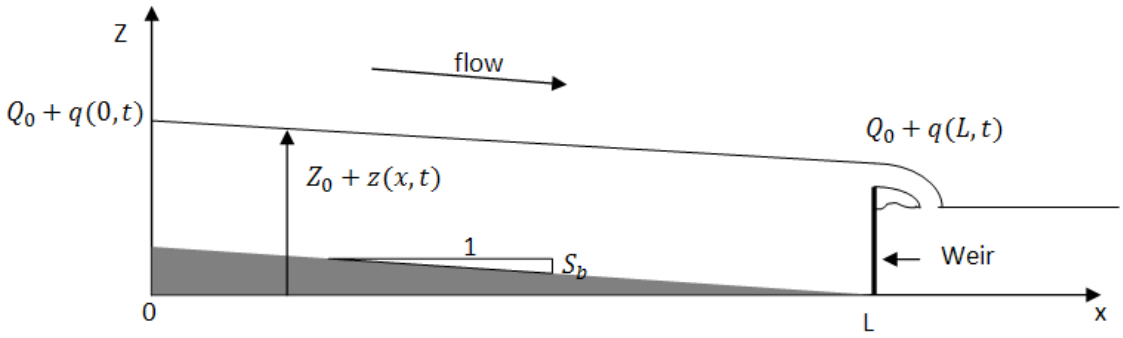


Figure H.1: Schematic representation of the canal with weir structure.

H.2.3 Open-Loop Control Problem

The control problem illustrated in Figure H.1, consists in determining the control $u(t) = q(0,t)$, i.e. the flow of the upstream discharge that yields the desired downstream discharge $y(t) = q(L,t)$, where $y(t)$ is a user-defined flow profile over time at the end of the canal.

We therefore have to solve a feed-forward control problem for a system with boundary control (in the present case upstream discharge). The dynamics are modeled by the following partial differential equations:

$$\forall x \in]0, L[\ \forall t \in]0, T] \qquad D_0 q_{xx} - C_0 q_x = q_t, \tag{H.7}$$

$$\forall x \in]0, L[\ \forall t \in]0, T] \qquad B_0 z_t + q_x = 0. \tag{H.8}$$

A boundary condition is imposed at $x = L$ by equation (H.9):

$$\forall t \in]0, T]\ q(L,t) = bz(L,t) = y(t), \tag{H.9}$$

where $y(t)$ is the desired output, and initial conditions defined by the deviations from the nominal values:

$$\begin{aligned} \forall x \in]0, L[\quad q(x,0) &= 0, \\ \forall x \in]0, L[\quad z(x,0) &= 0. \end{aligned}$$

We are looking for the appropriate control $u(\cdot)$ that will generate the $y(t)$ defined by (H.9), where $u(\cdot)$ is defined by:

$$\forall t \in]0, T]\ u(t) = q(0,t). \tag{H.10}$$

H.3 Computation of the Open Loop Control Input for the Hayami Model

In this section we solve the control problem given by equations (H.7-H.9) and try to parametrize the flow $q(x,t)$ in terms of the discharge $q(L,t)$ or $y(t)$. We will produce a solution to this

problem using differential flatness based on Cauchy-Kovalevskaya decomposition, and study the convergence of the obtained infinite series.

H.3.1 Cauchy-Kovalevskaya Decomposition

Following [47], equation (H.7) can be transformed into the heat equation. Let us consider the following transformation:

$$q(x,t) = h(x,t)p(x,t) \tag{H.11}$$

where $h(x,t) = e^{\left(-\frac{\alpha^2}{\beta^2}t+\alpha(x-L)\right)}$, $\alpha = \frac{C_0}{2D_0}$, and $\beta = \frac{1}{\sqrt{D_0}}$. We have:

$$\begin{aligned} p_t h(x,t) &= q_t + \frac{\alpha^2}{\beta^2}q, \\ p_x h(x,t) &= q_x - \alpha q, \\ p_{xx} h(x,t) &= q_{xx} - 2\alpha q_x + \alpha^2 q. \end{aligned}$$

Substituting in equation (H.7), $p(x,t)$ satisfies:

$$p_t = \frac{1}{\beta^2}p_{xx}. \tag{H.12}$$

The problem (H.7) - (H.9) can now be reformulated as follows

$$\forall x \in]0, L[\ \forall t \in]0, T] \qquad p_t = \frac{1}{\beta^2}p_{xx}, \tag{H.13}$$

$$\forall x \in]0, L[\ \forall t \in]0, T] \qquad B_0 z_t = -h(x,t)\left(p_x + \alpha p\right), \tag{H.14}$$

$$\forall t \in]0, T] \qquad p(L,t) = f(t)y(t), \tag{H.15}$$

where $f(t) = e^{\frac{\alpha^2}{\beta^2}t}$. The system of equations (H.13)-(H.15) can be used for a Cauchy-Kovalevskaya decomposition [6, 31] and the solution of the PDE, $p(x,t)$ (resp. $z(x,t)$), can be expressed in the Cauchy-Kovalevskaya power series decomposition, in the present case as a function of $p(L,t)$ (resp. $z(L,t)$) and all its derivatives. The Cauchy-Kovalevskaya decomposition is a standard way of parametrizing the input as a function of the output for parabolic and linear PDEs [31, 39, 12]. In the present case, it can be shown to be equivalent to Laplace decomposition [10] which produces the same parametrization, using spectral analysis. We assume the following form for p and z:

$$p(x,t) = \sum_{i=0}^{+\infty} p_i(t)\frac{(x-L)^i}{i!}, \tag{H.16}$$

$$z(x,t) = \sum_{i=0}^{+\infty} z_i(t)\frac{(x-L)^i}{i!}. \tag{H.17}$$

where $p_i(t)$ and $z_i(t)$ are C^∞ functions. We have: $p_t = \sum_{i=0}^{+\infty} \dot{p}_i \frac{(x-L)^i}{i!}$, $p_{xx} = \sum_{i=0}^{+\infty} p_{i+2} \frac{(x-L)^i}{i!}$ where $\dot{p}_i$ denotes the time derivative of $p_i(t)$. After substitution in equation (H.13), we obtain:

$$\sum_{i=0}^{+\infty} \dot{p}_i \frac{(x-L)^i}{i!} = \frac{1}{\beta^2} \sum_{i=0}^{+\infty} p_{i+2} \frac{(x-L)^i}{i!}.$$

Equating the coefficients of $\frac{(x-L)^i}{i!}$ gives for all $i \in \mathbb{N}$:

$$p_{i+2}(t) = \beta^2 \dot{p}_i(t). \tag{H.18}$$

Additionally, it follows from equation (H.16) and equation (H.17) that $p_0 = p(L,t)$ and $z_0 = z(L,t)$. We still need a condition on p_1 to be able to express every p_i as a function of p_0. We combine equation (H.14) and equation (H.15) to obtain a boundary condition on p at $x = L$. We have:

$$z_t = \sum_{i=0}^{+\infty} \dot{z}_i \frac{(x-L)^i}{i!}.$$

So that $\dot{z_0} = z_t(L,t)$, and equation (H.14), with $x = L$ gives:

$$B_0 \dot{z}_0 + e^{-\frac{\alpha^2}{\beta^2}t} (p_1 + \alpha p_0) = 0. \tag{H.19}$$

In addition, equation (H.15) gives: $p_0 = b z_0 e^{\frac{\alpha^2}{\beta^2}t}$. Differentiating this equation with respect to time, we get:

$$\dot{z}_0 = \frac{1}{b} \left(\dot{p}_0 - \frac{\alpha^2}{\beta^2} p_0 \right) e^{-\frac{\alpha^2}{\beta^2}t},$$

and eventually; plugging back into equation (H.19), we obtain: $p_1 = -\frac{B_0}{b}\dot{p}_0 + \kappa p_0$, where $\kappa = \frac{B_0}{b}\frac{\alpha^2}{\beta^2} - \alpha$. Using the induction relation (H.18) and the expression of p_0 and p_1, we can compute separately the odd and even terms:

$$\begin{aligned} p_{2i} &= \beta^{2i} p_0^{(i)}, \\ p_{2i+1} &= \kappa \beta^{2i} p_0^{(i)} - \frac{B_0}{b} \beta^{2i} p_0^{(i+1)}, \end{aligned}$$

where $p_0^{(i)}$ stands for the i^{th} time derivative of $p_0(t)$. Therefore, we can formally write $p(x,t)$ as follows:

$$\begin{aligned} p(x,t) &= \sum_{i=0}^{+\infty} \beta^{2i} p_0^{(i)} \frac{(x-L)^{2i}}{(2i)!}, \\ &\quad + \sum_{i=0}^{+\infty} \beta^{2i} \left(\kappa p_0^{(i)} - \frac{B_0}{b} p_0^{(i+1)} \right) \frac{(x-L)^{2i+1}}{(2i+1)!}. \end{aligned}$$

From equation (H.15), we deduce that $p_0(t) = f(t)y(t)$. The final parametrization of the flow $q(x,t)$ will have the form:

$$q(x,t) = h(x,t)\left(T_1(x,t) + \kappa T_2(x,t) - \frac{B_0}{b}T_3(x,t)\right), \tag{H.20}$$

where

$$T_1(x,t) = \sum_{i=0}^{+\infty}(fy)^{(i)}\frac{\beta^{2i}(x-L)^{2i}}{(2i)!}, \tag{H.21}$$

$$T_2(x,t) = \sum_{i=0}^{+\infty}(fy)^{(i)}\frac{\beta^{2i}(x-L)^{2i+1}}{(2i+1)!}, \tag{H.22}$$

$$T_3(x,t) = \sum_{i=0}^{+\infty}(fy)^{(i+1)}\frac{\beta^{2i}(x-L)^{2i+1}}{(2i+1)!}. \tag{H.23}$$

Equation (H.20) relates the discharge variation $q(x,t)$ as a function of the desired flat output $y(t)$ which corresponds to the discharge $q(L,t)$ at the downstream end of the canal. The output $y(t)$ is sometimes referred to as "flat", which in the present context means that it is possible to express the input of the system $u(t)$ explicitly as a function of the desired output $y(t)$ and its derivatives (a formal definition of differential flatness is available in [17], for general systems). This also defines the parametrization of the state $q(x,t)$ as a function of the same derivatives. The present decomposition, chosen for this study, is the Cauchy-Kovalevskaya form, which is appropriate for parabolic equations such as the one presented in this appendix. This solution is formal, until the convergence of the infinite series is assessed. An alternate derivation of equation (H.20) was produced using Laplace techniques, and provides the same algebraic result [10].

H.3.2 Convergence of the Infinite Series

We now give the formal proof of convergence of the series in equation (H.20). We assume that the flat output $y(t)$ is a Gevrey function [49] of order $\gamma > 0$, i.e.:

$$\exists\, m,\, l > 0 \quad \forall n \in \mathbb{N} \quad \sup_{t\in\mathbb{R}} \left|y^{(n)}(t)\right| < m\frac{(n!)^\gamma}{l^n}. \tag{H.24}$$

$f(t) = e^{\frac{\alpha^2}{\beta^2}t}$ is Gevrey of order 0, and therefore is Gevrey of order γ. The product of two Gevrey functions of same order is a Gevrey function of the same order, as a consequence, $f(t)y(t)$ is Gevrey of order $\gamma > 0$. We will use the Cauchy-Hadamard theorem [23] which states that the radius of convergence of the Taylor series $\sum_{i=0}^{+\infty} a_n x^n$ is $\frac{1}{\limsup_{n\to+\infty}|a_n|^{1/n}}$. The radius

of convergence for $T_3(x,t)$ is given by:

$$\frac{1}{\rho} = \limsup_{i\to+\infty} \left(\frac{\beta^{2i} \left|(fy)^{(i+1)}(t)\right|}{(2i+1)!} \right)^{\frac{1}{2i+1}},$$

where ρ is the radius of convergence around L. We can find an upper bound to $\frac{1}{\rho}$ by inducing the property of bounds on a Gevrey function of order $\gamma > 0$ from equation (H.24),

$$\begin{aligned}
\frac{1}{\rho} &\leq \limsup_{i\to+\infty} \left(\frac{\beta^{2i} m^{\frac{((i+1)!)^\gamma}{l^{i+1}}}}{(2i+1)!} \right)^{\frac{1}{2i+1}}, \\
&\leq \limsup_{i\to+\infty} \frac{\beta^{\frac{2i}{2i+1}} m^{\frac{1}{2i+1}}}{l^{\frac{i+1}{2i+1}}} \left(\frac{((i+1)!)^\gamma}{(2i+1)!} \right)^{\frac{1}{2i+1}}, \\
&\sim \limsup_{i\to+\infty} \frac{\beta}{\sqrt{l}} \frac{i+1}{2i+1} \left(\frac{i+1}{e} \right)^{\frac{(\gamma-2)i+(\gamma-1)}{2i+1}}, \qquad \text{(H.25)} \\
&\sim \begin{cases} +\infty & \gamma > 2, \\ \frac{\beta}{2\sqrt{l}} & \gamma = 2, \\ 0 & \gamma < 2, \end{cases}
\end{aligned}$$

where in equation (H.25) we have used the fact that $((i+1)!)^{\frac{1}{i+1}} \sim \frac{i+1}{e}$, and $((2i+1)!)^{\frac{1}{2i+1}} \sim \frac{2i+1}{e}$ as an immediate consequence of the Stirling formula. Also we have used $m^{\frac{1}{2i+1}} \sim 1$, $\beta^{\frac{2i}{2i+1}} \sim \beta$ and $l^{\frac{i+1}{2i+1}} \sim \sqrt{l}$. This will ensure an infinite radius of convergence for $\gamma < 2$. Similar calculations can be held for $T_1(x,t)$ and $T_2(x,t)$ leading to the following conclusions:

- Equation (H.20) converges with an infinite radius of convergence for the choice of a Gevrey function $y(t)$ of order $\gamma < 2$.
- For $\gamma = 2$, the radius of convergence is greater than $\frac{2\sqrt{l}}{\beta}$, which provides convergence of the series for $x \in [L - \frac{2\sqrt{l}}{\beta}, L]$, given the definition of $x \in [0, L]$.
- We can draw no conclusions on the convergence of the series when $\gamma > 2$.

H.4 Numerical Assessment of the Performance of the Feed-Forward Controller

In this section, we compute the control command $u(t)$ by evaluating equation (H.20) at $x = 0$. We subsequently simulate the controller numerically on the Hayami model equations (H.7)-(H.9) in order to evaluate their behavior before testing them on the Saint-Venant equations. This section successively investigates numerical simulations for the Hayami and the Saint-Venant models and the performance of the controller on both models.

H.4.1 Hayami Model Simulation

From section H.3.2, the infinite series convergence is ensured by choosing $y(t)$ to be a Gevrey function of order $\gamma < 2$. To meet this convergence condition following [31], we introduce the bump function $\phi_\sigma(t) : \mathbb{R} \to \mathbb{R}$ defined as

$$\phi_\sigma(t) \;=\; \begin{cases} 0 & t < 0, \\ \dfrac{\int_0^{t/T} \exp(-1/((\tau(1-\tau))^\sigma)d\tau}{\int_0^1 \exp(-1/((\tau(1-\tau))^\sigma)d\tau} & 0 \le t \le T\,, \\ 1 & t > T, \end{cases} \tag{H.26}$$

where $\sigma > 1$, $T > 0$. The Gevrey order of the bump function is $1 + 1/\sigma$. The function $\phi_\sigma(t)$ is used in [12, 17, 31, 39, 50], it is strictly increasing from 0 at $t = 0$ to 1 at $t = T$ with zero derivatives at $t = 0$ and $t = T$. The larger the σ parameter is, the faster is the slope of transition. Figure H.2 shows a plot of the bump function for different values of σ and $T = 1$. Setting $y(t) = q_1\phi_\sigma(t)$ will allow us to have a transition from zero discharge flow for $t \le 0$ to a discharge flow equal to q_1 for $t \ge T$, where q_1 is a constant. Note that the bump function was chosen because of its Gevrey properties, we guarantee an infinite radius of convergence for $\sigma > 1$ ($\gamma < 2$ as described in section H.3.2). As can be inferred from the previous proof, the proposed method only applies to functions with proper radius of convergence, by equation (H.25). This is due to the fact that in general, the reachable set (i.e. the set of attainable $y(\cdot)$ functions) from input functions $u(\cdot)$ is not equal to the whole state space of output functions. In other words, not all functions $y(\cdot)$ can be synthesized by a function $u(\cdot)$.

The upstream discharge or the control input $u(t)$ can be computed by substituting $x = 0$ in equation (H.20). We obtain:

$$u(t) = h(0,t)\left(T_1(0,t) + \kappa T_2(0,t) - \frac{B_0}{b}T_3(0,t)\right). \tag{H.27}$$

Evaluation of the Truncation Error

For practical implementation purposes, one needs to know how many terms should be included in the numerical computation. This can be done by computing an upper bound on the truncation error. When the infinite series, $T_1(0,t)$, $T_2(0,t)$, and $T_3(0,t)$, in equation (H.27) are truncated, this generates an approximation error which needs to be evaluated. We use the Gevrey assumption in equation (H.24) and write:

$$|T_1(0,t)| \le \sum_{i=0}^{\infty} b_i, \quad b_i = m\frac{(i!)^\gamma}{l^i}\frac{[\beta L]^{2i}}{(2i)!},$$

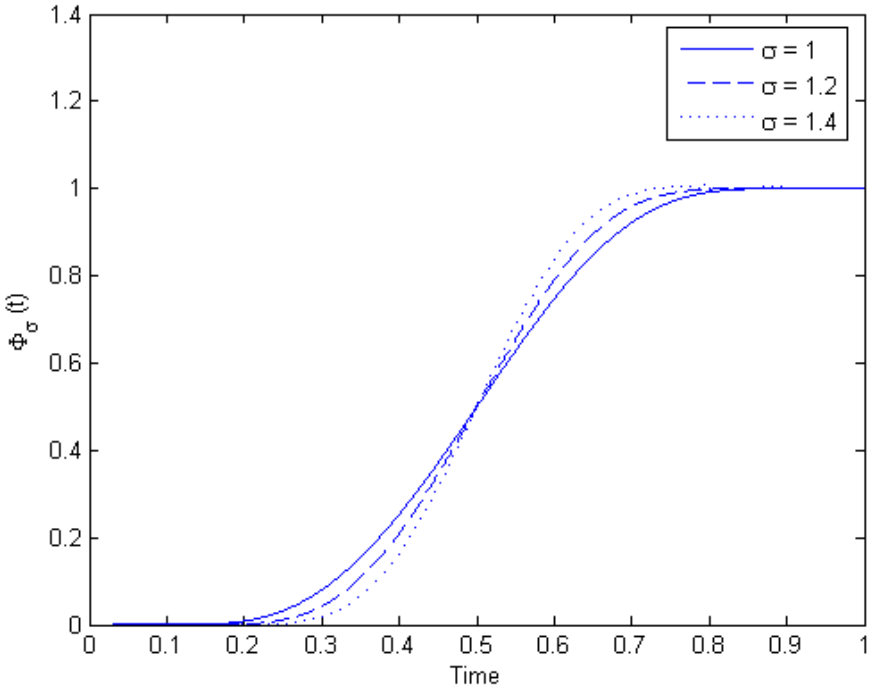


Figure H.2: Bump function described by equation (H.26) plotted for different values of σ and $T = 1$.

$$|T_2(0,t)| \leq \sum_{i=0}^{\infty} c_i, \quad c_i = m\frac{(i!)^\gamma}{l^i}\frac{\beta^{2i}L^{2i+1}}{(2i+1)!},$$

$$|T_3(0,t)| \leq \sum_{i=0}^{\infty} d_i, \quad d_i = m\frac{((i+1)!)^\gamma}{l^{i+1}}\frac{\beta^{2i}L^{2i+1}}{(2i+1)!}.$$

To evaluate the approximation error of $T_1(0,t)$ when truncated, we study the series b_i. The series b_i satisfies the relation $b_{i+1} = \mathcal{E}_1(i)b_i$ where $\mathcal{E}_1(i) = \frac{(i+1)^\gamma}{(2i+2)(2i+1)}\frac{\beta^2L^2}{l}$. The function $\mathcal{E}_1(i)$ is decreasing towards zero:

$$\frac{d(\mathcal{E}_1(i))}{di} = \frac{(1+i)^\gamma(\gamma - 3 + 2i(\gamma-2))}{2(1+3i+2i^2)^2}\frac{\beta^2L^2}{l} < 0 \quad \forall\, \gamma < 2,$$

and $\mathcal{E}_1(i) \sim \frac{\beta^2L^2}{4l}i^{\gamma-2}$ for large values of i. Thus, for $\gamma < 2$, this implies that, for any small constant $\epsilon < 1$, there exists a unique integer i_1 such that $\mathcal{E}_1(i_1) \leq \epsilon$ and $\mathcal{E}_1(i_1 - 1) > \epsilon$. Since $\mathcal{E}_1(i)$ is strictly decreasing, we have $\mathcal{E}_1(j) \leq \mathcal{E}_1(i_1) \leq \epsilon$ for any $j \geq i_1$. Thus $b_{j+1} \leq b_j\epsilon$ and $b_{j+k} \leq b_j\epsilon^k$ $\forall j \geq i_1$, $\forall k \geq 0$. $T_2(0,t)$, and $T_3(0,t)$ satisfy similar properties, which can be summarized by: for any $\epsilon < 1$, there exist $j \geq 0$, such that:

$$b_{j+k} \leq b_j\epsilon^k, c_{j+k} \leq c_j\epsilon^k, d_{j+k} \leq d_j\epsilon^k \quad \forall k \geq 0. \tag{H.28}$$

This result provides us with an upper bound on the truncation error, which is quantified by writing equation (H.27) as a sum of the truncated series and the truncation error:

$$u(t) = u_j(t) + e_j(t),$$

where

$$\begin{aligned} u_j(t) &= \left(\sum_{i=0}^{j-1} (fy)^{(i)} \frac{\beta^{2i} L^{2i}}{(2i)!} - \kappa \sum_{i=0}^{j-1} (fy)^{(i)} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!} \right. \\ &\quad \left. + \frac{B_0}{b} \sum_{i=0}^{j-1} (fy)^{(i+1)} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!} \right), \end{aligned}$$

$$\begin{aligned} e_j(t) &= h(0,t) \left(\sum_{i=j}^{+\infty} (fy)^{(i)} \frac{\beta^{2i} L^{2i}}{(2i)!} - \kappa \sum_{i=j}^{+\infty} (fy)^{(i)} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!} \right. \\ &\quad \left. + \frac{B_0}{b} \sum_{i=j}^{+\infty} (fy)^{(i+1)} \frac{\beta^{2i} L^{2i+1}}{(2i+1)!} \right). \end{aligned} \tag{H.29}$$

We now use the geometric series upper bound given by equation (H.28) to compute an upper bound of the truncation error, for a large enough j:

$$\begin{aligned} |u(t) - u_j(t)| &= |e_j(t)| \\ &\leq h(0,t) \left(b_j \sum_{k=0}^{\infty} \epsilon^k + |\kappa| \, c_j \sum_{k=0}^{\infty} \epsilon^k \right. \\ &\quad \left. + \frac{B_0}{b} d_j \sum_{k=0}^{\infty} \epsilon^k \right), \\ &\leq \frac{h(0,t)}{1-\epsilon} \left(b_j + |\kappa| \, c_j + \frac{B_0}{b} d_j \right). \end{aligned} \tag{H.30}$$

Therefore, an upper bound on the truncation error of approximating $u(t)$ using j terms of the infinite series can be found, and it is linear in the coefficients b_j, c_j, and d_j.

Numerical Simulation

For the numerical simulation, we consider incrementing the flow by $1\,m^3/s$ from its nominal flow $Q_0 = 2.5\,m^3/s$ in 1 hour ($T = 3600$ seconds). We take $\sigma = 2$ which implies $y(t)$ to be a Gevrey-function of order 1.5 thus satisfying the convergence condition in section H.3.2. The model parameters are $L = 1000\,m$, $C_0 = 20\,m/s$, $D_0 = 1800\,m^2/s$, $B_0 = 7\,m$, and $b = 1\,m^2/s$.

The infinite series of the control input $u(t)$ is approximated using j terms. The value of j is determined by evaluating the L_2 norm of the truncation error as a function of j which is given by:

$$\|e_j\| = \left(\int_0^{T_{\text{sim}}} |e_j(t)|^2 \, d\tau \right)^{\frac{1}{2}}, \tag{H.31}$$

where T_{sim} is the simulation time. We compute the L_2 norm of the upper bound error:

$$\begin{aligned} \|e_j\| \quad \leq \quad & \frac{\sqrt{2}}{2(1-\epsilon)} \frac{\beta}{\alpha} e^{\left(-\frac{\alpha^2}{\beta^2} T_{\text{sim}} - \alpha L\right)} \sqrt{e^{2\frac{\alpha^2}{\beta^2} T_{\text{sim}}} - 1} \\ & \left(b_j + |\kappa| \, c_j + \frac{B_0}{b} d_j \right). \end{aligned} \tag{H.32}$$

Figure H.3 shows a comparison between the L_2 norms of the upper bound computed by equation (H.32) and the real error computed by equation (H.29) until numerical convergence (the residual goes to machine accuracy for 76 terms). We notice that our upper bound is conservative, (the real error may be two orders of magnitude smaller). Nonetheless, it gives a sufficient condition useful for computational purposes. Figure H.4 shows the effect of adding more terms on the relative error $e_{\text{rel}}(t) = \left| \frac{u(t) - u_j(t)}{Q_0 + u(t)} \right|$. We choose $j = 10$ which yields an error of $\|e_j\| \sim 10^{-3}$, and solve equations (H.7), (H.8), (H.9), and (H.10) using the Crank-Nicholson scheme. The numerical solution at $x = L$ or $q(L,t)$ is compared to $y(t)$, the desired downstream discharge flow. The results of this simulation are shown in figure H.5.

The discharge at the downstream follows the desired discharge accurately which validates our control input. We can now compare our result to other problems from the literature.

Comparison with the Heat Equation

In the context of thermal systems [31], an explicit open loop controller was derived for the heat equation with zero gradient boundary conditions. With some simple transformations in time and space we can relate the results [31] to our problem. The transformed version of the equations of [31] has the following form:

$$\forall x \in]0, L[\; \forall t \in]0, T] \qquad D_0 q_{xx} - C_0 q_x = q_t, \tag{H.33}$$

$$\forall t \in]0, T] \qquad q_x(L,t) = 0, \tag{H.34}$$

$$\forall x \in]0, L[\qquad q(x,0) = 0,$$

$$\forall t \in]0, T] \qquad y(t) = q(L,t),$$

$$\forall t \in]0, T] \qquad u(t) = q(0,t).$$

The solution of the control input for this particular problem is:

$$u_{\text{heat}}(t) = h(0,t) \left(T_1(0,t) - \alpha T_2(0,t) \right). \tag{H.35}$$

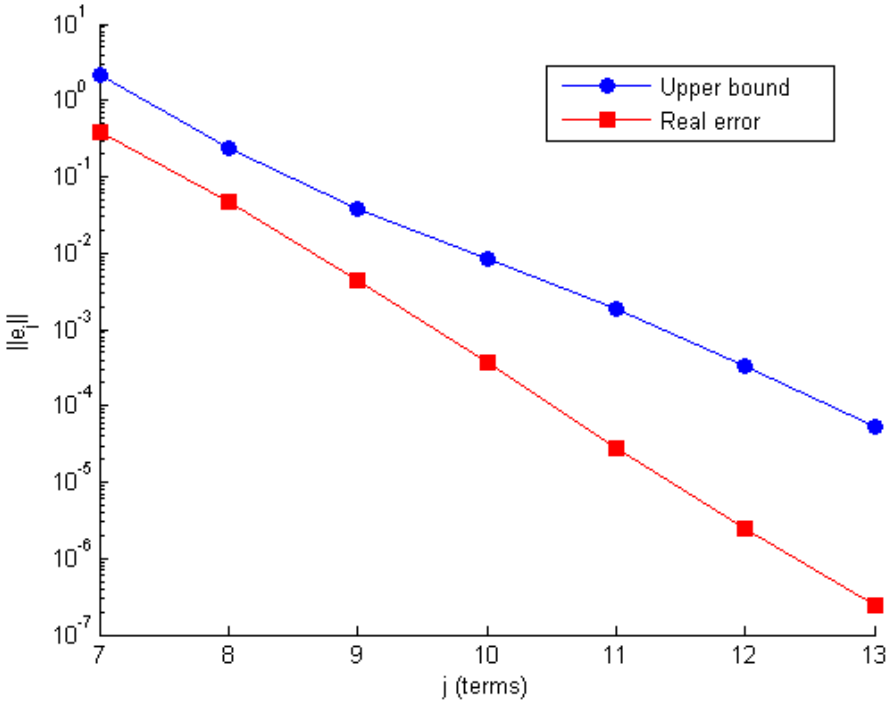


Figure H.3: L_2 norm of the error $e_j(t)$ defined by equation (H.31) as a function of the terms used j. The upper bound is computed using equation (H.32) and the real error is computed until numerical convergence.

We can vary the value of the variable b in equation (H.27), and observe its effect on $u(t)$. This physically corresponds to changing the height or the width of the weir located at the downstream end of the canal. Figure H.6 shows the effect of varying b on the control input u.

We can see that by increasing the value of b, the function of $u(t)$ numerically converges to $u_{\text{heat}}(t)$ described by equation (H.35). This can be seen directly by inspection of the limit of equation (H.27) as b tends to $+\infty$ which would result in equation (H.35). Substituting $\kappa = \frac{B_0}{b}\frac{\alpha^2}{\beta^2} - \alpha$ into equation (H.27), we obtain:

$$u(t) = u_{\text{heat}}(t) + u_{\text{b}}(t),$$

where

$$u_{\text{b}}(t) = h(0,t)\frac{B_0}{b}\left(\frac{\alpha^2}{\beta^2}T_2(0,t) - T_3(0,t)\right).$$

As b tends to $+\infty$, the boundary effect becomes negligible, and equation (H.27) converges in the limit to equation (H.35), i.e. in the limit u and u_{heat} are identical. If we were to use the controller in equation (H.35) to control our problem with $b = 1\, m^2/s$, we would

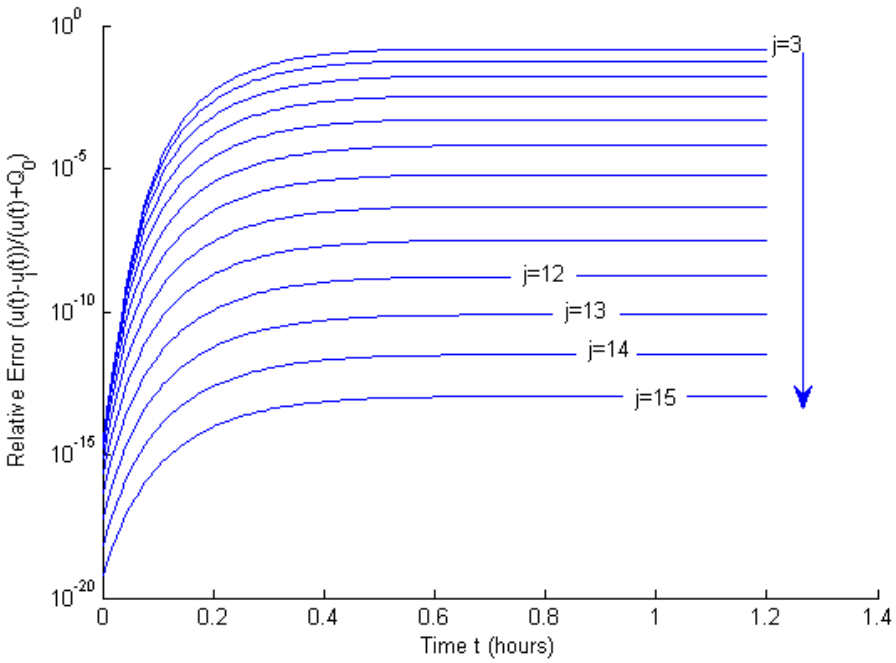


Figure H.4: Effect of adding more terms on the relative error $e_{\text{rel}}(t) = \left|\frac{u(t)-u_j(t)}{Q_0+u(t)}\right|$ for consecutive values of j starting from $j = 3$ to $j = 15$.

obtain the results shown in figure H.7. The effect can be seen in the transition which takes approximately 1.6 hours instead of one hour. This shows the considerable importance of boundary conditions on the dynamics of the flow transfer. It is therefore very important to take into account the appropriate physical boundary conditions in the open-loop control design to ensure a scheduled water distribution.

H.4.2 Saint-Venant Model Simulation

In numerous cases, controlling the Saint-Venant equations directly is impractical because of the required knowledge for the geometry of the canal and the Saint-Venant parameters defined in section H.2.1. For this reason we have used a simplification of the model to arrive to the Hayami equation which requires only two parameters, C_0 and D_0. The coefficient b, which represents the downstream boundary condition, can easily be inferred from the weir equation. In this section we show numerically that a calibrated Hayami model would provide us with an open-loop control law that steers the Saint-Venant equation solution at $x = L$ or the flow discharge at the weir to the desired discharge accurately. For the purpose of the simulation we use SIC, a computer program developed by Cemagref [40, 1] to simulate the upstream discharge and the measurement discharge at the downstream. SIC solves the

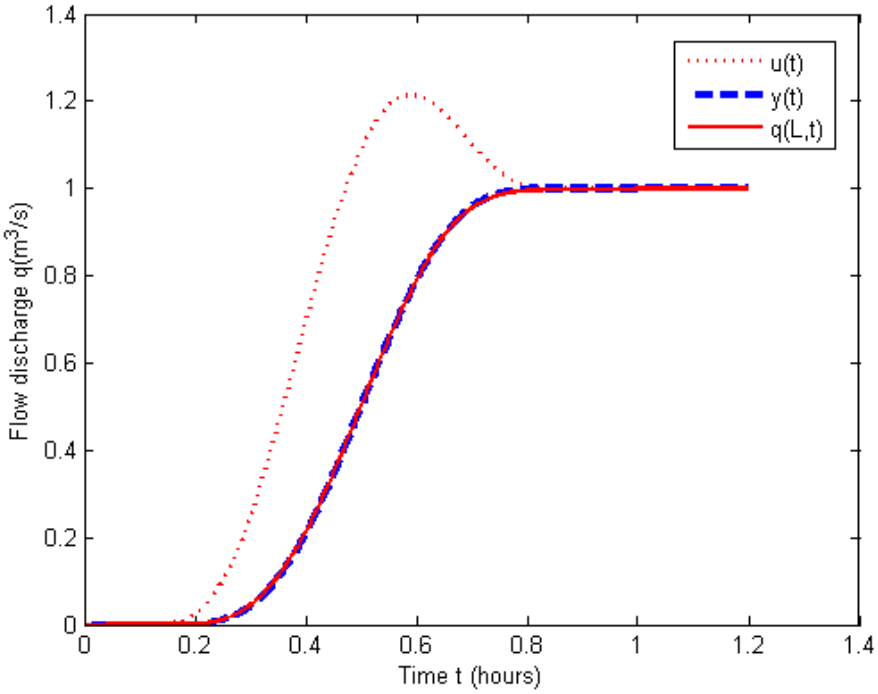


Figure H.5: Results of the numerical simulation of feed-forward control of the Hayami equation. The desired downstream discharge is $y(t)$, the upstream discharge is $u(t)$, and the downstream discharge computed by solving the Hayami model with $b = 1\, m^2/s$ is $q(L, t)$.

full nonlinear Saint-Venant equations using a finite difference scheme standard in hydraulics (Preissmann scheme).We also study the effect of uncertainties of the Saint-Venant equation parameters on the open-loop control system performance.

Hayami Model Identification

The purpose of model identification is to identify the parameters C_0, D_0 and b corresponding to the Hayami model and its boundary condition parameter that would best approximate the real flow governed by the Saint-Venant equations. This is done with an upstream discharge in a form of a step input, the flow discharges are monitored at the upstream and downstream positions. The hydraulic identification is done classically by finding the values of C_0, D_0 and b that minimize the error between the computed downstream discharge by the solution of the Crank-Nicholson scheme [55] and the measured one. We therefore have to solve the following

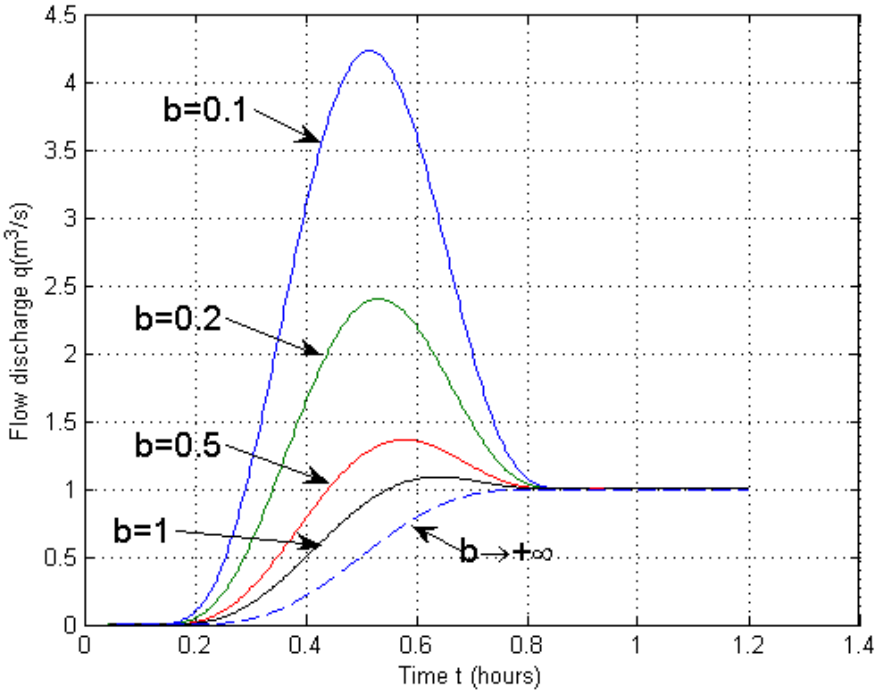


Figure H.6: Effect of varying b (m^2/s) on the upstream discharge or control input $u(t)$.

optimization problem:

$$\min_{C_0,D_0,b>0} \int_0^{T_{\text{sim}}} |q_{\text{SIC}}(\tau) - q_{\text{CN}}(C_0, D_0, b, \tau)|^2 \, d\tau,$$

where q_{SIC} is the downstream flow generated by SIC, and q_{CN} is the downstream flow generated by the Crank-Nicholson scheme, T_{sim} is the simulation time usually larger than the period needed to reach steady state. The nonlinear optimization problem was solved by the MATLAB nonlinear least-square curve fitting function (lsqnonlin). The identification was done using Saint Venant equations generated data. In our case, the identification was performed around a steady flow regime of $0.4\ m^3/s$, canal of length $L = 4887\ m$, and bed width $B_0 = 2\ m$. The average bottom slope is 3.8×10^{-4}, the Manning coefficient is $0.0213\ m^{-1/3}s$, and the weir discharge coefficient is 0.35. This leads to the following parameters: $C_0 = 0.88\ m/s$, $D_0 = 660.19\ m^2/s$, and $b = 0.16\ m^2/s$. Identification of coefficients of the Hayami equation is standard in hydraulics, and has shown to work well in practice [33].

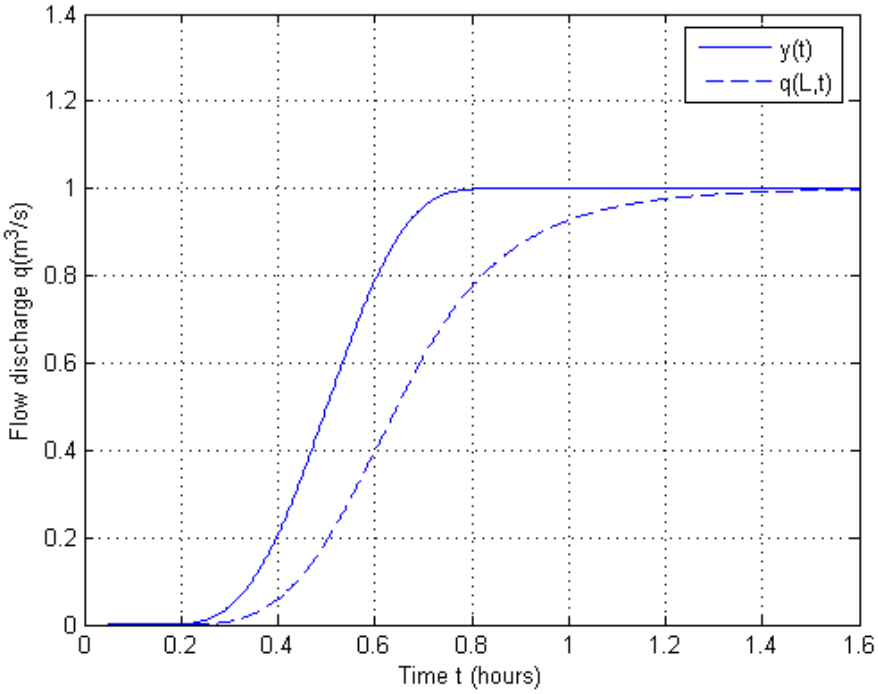


Figure H.7: Consequence of neglecting the boundary conditions in calculating the upstream discharge. The desired downstream discharge is $y(t)$, and the downstream discharge calculated by solving the Hayami model with $b = 1\,m^2/s$ and control input of equation (H.35) is $q(L, t)$.

Saint-Venant Control

The experimental canal we would like to simulate has the same properties as the one we have used for identification in the previous section. We are interested in raising the flow at the downstream from $0.4\,m^3/s$ to $0.5\,m^3/s$ in 5 hours. Setting the variables in section H.4.1 to $q_1 = 0.1\,m^3/s$, $T = 5$ hours, and $\sigma = 1.1$ will define the downstream profile $y(t)$. The control input or the discharge at the upstream can be calculated and the results are shown in figure H.8.

We notice that the open-loop control designed with the Hayami model performs very well on the full nonlinear Saint-Venant equations. As can be seen in Figure H.8, the reference output and the actual output achieved by the Hayami controller on the full Saint-Venant equations are visually almost identical, which confirms the practicality of the method for implementation on canals. This shows that the Hayami model is practical for the design of open-loop control when the corresponding parameters are identified. We extended our results by evaluating the uncertainties on the system parameters, and studying their effect on the performance of the open-loop control system.

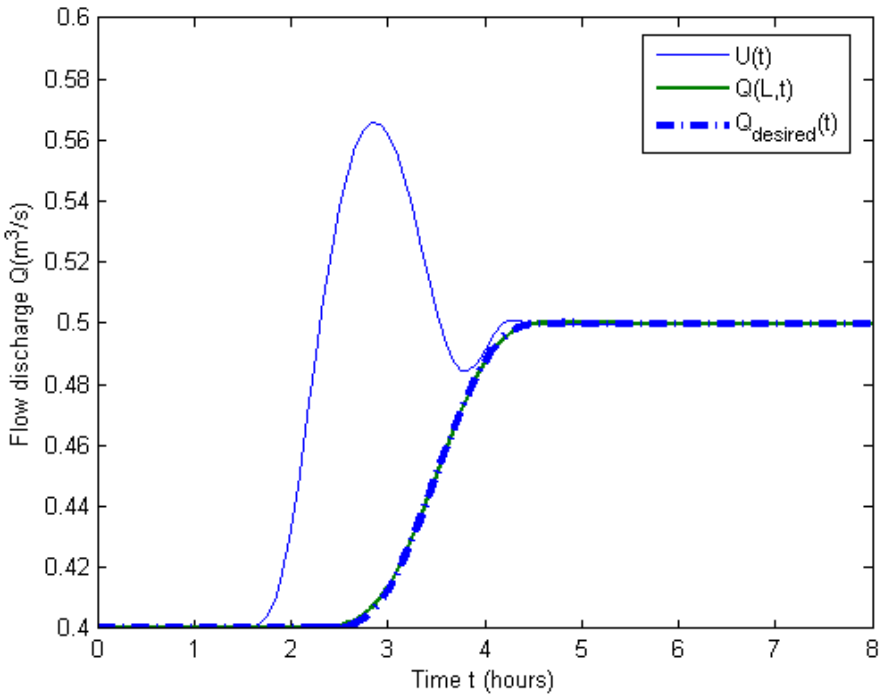


Figure H.8: Results of the implementation of our controller on the full nonlinear Saint-Venant equations. The desired downstream discharge is $Q_{\text{desidred}}(t) = Q_0 + y(t)$, the downstream discharge calculated by solving the Saint-Venant equations in SIC is $Q(L,t) = Q_0 + q(L,t)$, and the control input of the canal is $U(t) = Q_0 + u(t)$ where $u(t)$ is calculated using the Hayami model open-loop controller. The nominal flow in the canal is $Q_0 = 0.4\, m^3/s$.

Sensitivity Analysis

We study the effect of parameter uncertainties in the full nonlinear Saint-Venant model on the downstream discharge. We compute the control input using nominal values of Saint-Venant equations parameters and simulate it with models which incorporate some uncertainties. We specifically study the effect of the Manning and discharge coefficients uncertainties. We experiment with $+/-$ 20% variations on the nominal values and compare the downstream discharges of each scenario.

We use the experimental canal described in section H.4.2 with nominal Manning coefficient $n = 0.0213\ m^{-1/3}s$, and weir discharge coefficient $C_w = 0.35$. The control input of section H.4.2 is simulated under four different scenarios which define the $+/-$ 20% variations on the nominal values: Scenario 1: $n = 0.0256\ m^{-1/3}s$, $C_w = 0.42$, Scenario 2: $n = 0.0170\ m^{-1/3}s$, $C_w = 0.28$, Scenario 3: $n = 0.0256\ m^{-1/3}s$, $C_w = 0.28$, and Scenario 4: $n = 0.0170\ m^{-1/3}s$, $C_w = 0.42$. Figure H.9 shows the result of the sensitivity analysis.

We observe that the dominant effect is due to uncertainties in the Manning coefficient.

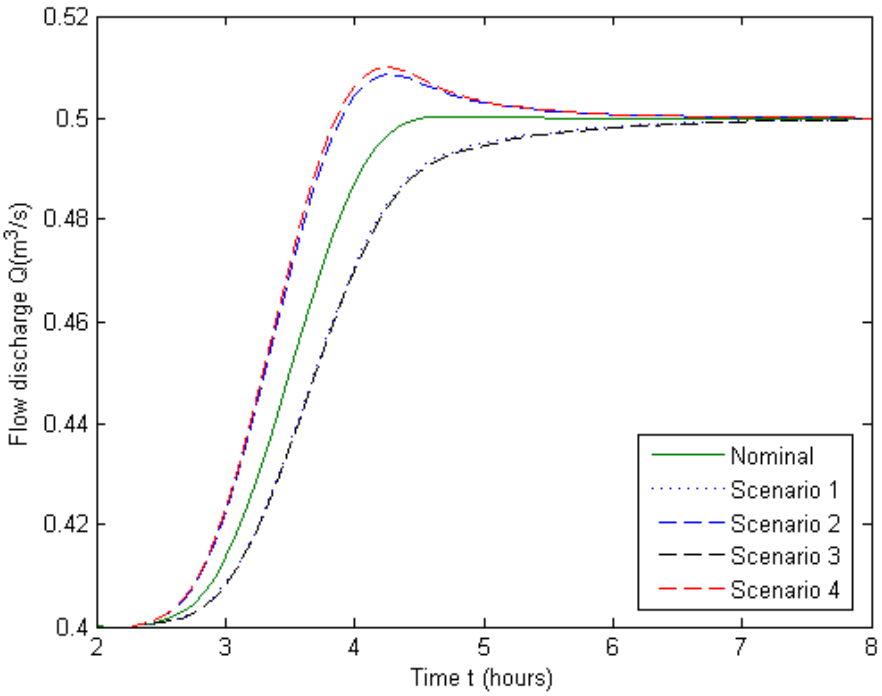


Figure H.9: Simulation results when the Manning and weir discharge coefficients are perturbed around their nominal values ($n = 0.0213$, $C_w = 0.35$). The downstream discharge is computed with four different scenarios. The scenarios correspond to $+/-$ 20% uncertainties on the nominal Manning and weir discharge coefficients. Scenario 1: $n = 0.0256\ m^{-1/3}s$, $C_w = 0.42$, Scenario 2: $n = 0.0170\ m^{-1/3}s$, $C_w = 0.28$, Scenario 3: $n = 0.0256\ m^{-1/3}s$, $C_w = 0.28$, Scenario 4: $n = 0.0170\ m^{-1/3}s$, $C_w = 0.42$.

Underestimating the Manning coefficient as in scenarios 1 and 3, leads to a delay in the downstream discharge delivery (approximately two hours delay to reach the desired downstream discharge). A larger Manning coefficient means more friction and this slows down the upstream discharges to reach the downstream location. In scenarios 2 and 4 (overestimating the Manning coefficient), the downstream discharge reaches its desired value one hour earlier. The peak in the upstream discharge is not fully filtered by the dynamics of the canal and leads to an overshoot in the discharge. The overshoot stabilizes at the desired downstream discharge ($0.5\ m^3/s$) after two hours.

Overestimating (scenarios 2 and 3), or underestimating (scenarios 1 and 4) the weir discharge coefficient has a very minor, yet opposite effect to uncertainties in the Manning coefficient. Downstream discharges in scenarios 1 and 4 are above the ones in scenarios 3 and 2, respectively. In all cases, the downstream discharge reaches a steady state equal to the desired one with a delay of two hours.

H.5 Conclusion

This appendix introduces a method to design an open-loop control based on the Hayami model for open channel flow control using differential flatness. The controller is obtained as an infinite series (Cauchy-Kovalevskaya decomposition) in terms of the desired downstream discharge flow. We have given sufficient conditions on the downstream profiles to ensure convergence. The effect of the boundary condition is also investigated and compared to previous studies realized for thermal systems. The simulations show satisfactory results for controlling the full Saint-Venant equations.

Bibliography

[1] J.-P. Baume et al. "SIC: a 1D Hydrodynamic Model for River and Irrigation Canal Modeling and Regulation". In: *Mtodos Numricos em Recursos Hidricos* 7 (2005). Ed. by Editor Rui Carlos Vieira da Silva Coppetec Fundacao, pp. 1–81.

[2] E. Bautista and A.J. Clemmens. "Response of ASCE task committee test cases to open-loop control measures". In: *Journal of Irrigation and Drainage Engineering* 125.4 (1999), pp. 179–188.

[3] E. Bautista, A.J. Clemmens, and T. Strelkoff. "Comparison of numerical procedures for gate stroking". In: *Journal of Irrigation and Drainage Engineering* 123.2 (1997), pp. 129–136.

[4] S.R. Becker, E.J. Candes, and M. Grant. "Templates for convex cone problems with applications to sparse signal recovery". In: *Stanford University Technical Report* (2010).

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004. ISBN: 0521833787.

[6] A. Bressan. *Hyperbolic systems of conservation laws: the one-dimensional Cauchy problem*. Oxford, UK: Oxford University Press, 2000.

[7] M. Cantoni et al. "Control of Large-Scale Irrigation Networks". In: *Proceedings of the IEEE* 95.1 (2007), pp. 75–91. ISSN: 0018-9219.

[8] S.S. Chen, D.L. Donoho, and M.A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM review* 43 (2001), p. 129.

[9] J.M. Coron, B. D'Andrea-Novel, and G. Bastin. "A Lyapunov approach to control irrigation canals modeled by Saint-Venant equations". In: *Proceedings of European Control Conference, Karlsruhe, Germany* (1999).

[10] F. Di Meglio et al. "Feed-forward river flow control using differential flatness". In: *Proceedings of the 47th IEEE Conference on Decision and Control, Cancun, Mexico* 1 (December 2008), pp. 3903–3910.

[11] David L. Donoho and Yaakov Tsaig. "Fast solution of l_1-norm minimization problems when the solution may be sparse". In: *IEEE Transactions on Information Theory* 54.11 (2008), pp. 4789–4812.

[12] W. Dunbar et al. "Motion planning for a nonlinear Stefan problem". In: *ESAIM: Control, Optimisation and Calculus of Variations* 9 (2003), pp. 275–296.

[13] Bradley Efron et al. "Least angle regression (with discussion)". In: *Annals of Statistics* 32 (2004), pp. 407–499.

[14] Jianqing Fan and Jinchi Lv. "A selective overview of variable selection in high dimensional feature space". In: *Statistica Sinica* 20 (2010), pp. 101–148.

[15] Jianqing Fan and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5 (2008), pp. 849–911.

[16] M. Fliess et al. "Active signal restoration for the telegraph equation". In: *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix* 2 (1999), pp. 1107–1111.

[17] M. Fliess et al. "Flatness and defect of non-linear systems: introductory theory and examples". In: *International Journal of Control* 61(6) (1995), pp. 1327–1361.

[18] George Forman. "An extensive empirical study of feature selection metrics for text classification". In: *Journal of Machine Learning Research* 3 (2003), pp. 1289–1305.

[19] A. Frank and A. Asuncion. *UCI Machine Learning Repository*. 2010. URL: `http://archive.ics.uci.edu/ml`.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: `http://www.jstatsoft.org/v33/i01`.

[21] Jerome Friedman et al. "Pathwise coordinate optimization". In: *The Annals of Applied Statistics* 1.2 (2007), pp. 302–332.

[22] Brian Gawalt et al. "Discovering word associations in news media via feature selection and sparse classification". In: *MIR '10: Proceedings of the international conference on Multimedia information retrieval*. Philadelphia, Pennsylvania, USA, 2010, pp. 211–220.

[23] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. New York, NY: Courier Dover Publications, 2003.

[24] J. de Halleux et al. "Boundary feedback control in networks of open-channels". In: *Automatica* 39 (2003), pp. 1365–1376.

[25] S. Hayami. "On the propagation of flood waves". In: *Bulletin of the Disaster Prevention Institute* 1 (1951), pp. 1–16.

[26] Chia-Hua Ho and Chih-Jen Lin. "Large-scale Linear Support Vector Regression". In: *Journal of Machine Learning Research* 13 (2013), pp. 3323–3348.

[27] Seung-Jean Kim et al. "An interior-point method for large-scale l_1-regularized least squares". In: *IEEE Journal on Selected Topics in Signal Processing* 1.4 (2007), pp. 606–617.

[28] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. "An interior-point method for large-scale l_1-regularized logistic regression". In: *Journal of Machine Learning Research* 8 (2007), pp. 1519–1555.

[29] M. Kristic and A. Smyshlyaev. *Boundary control of PDEs: A course on backstepping designs*. Philadelphia, PA: SIAM, 2008.

[30] M. Krstic. "Personal communication. UC Berkeley, October 2007". In: ().

[31] B. Laroche, P. Martin, and P. Rouchon. "Motion planning for a class of partial differential equations with boundary control". In: *Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL* 3 (1998), pp. 3494–3497.

[32] B. Laroche, P. Martin, and P. Rouchon. "Motion planning for the heat equation". In: *International Journal of Robust and Nonlinear Control* 10.8 (2000), pp. 629–643.

[33] X. Litrico. "Nonlinear diffusive wave modeling and identification for open-channels". In: *J. Hydraul. Eng.* 127.4 (2001), pp. 313–320.

[34] X. Litrico and V. Fromion. "Frequency modeling of open channel flow". In: *J. Hydraul. Eng.* 130.8 (2004), pp. 806–815.

[35] X. Litrico and V. Fromion. "H_∞ control of an irrigation canal pool with a mixed control politics". In: *IEEE Trans. Control Systems Technology* 14.1 (2006), pp. 99–111.

[36] X. Litrico, V. Fromion, and G. Scorletti. "Robust feedforward boundary control of hyperbolic conservation laws". In: *Networks and Heterogeneous Media* 2.4 (2007), pp. 715–729.

[37] X. Litrico and D. Georges. "Robust continuous-time and discrete-time flow control of a dam-river system: (I) Modelling". In: *Applied mathematical modelling* 23.11 (1999), pp. 809–827.

[38] X. Litrico et al. "Conversion from discharge to gate opening for the control of irrigation canals". In: *Journal of Irrigation and Drainage Engineering* 134.3 (2008), pp. 305–314.

[39] A.F. Lynch and J. Rudolph. "Flatness-based boundary control of a nonlinear parabolic equation modelling a tubular reactor". In: *Nonlinear control in the year 2000, London* Lecture Notes in Control and Information Sciences 259 (2000), pp. 45–54.

[40] P.-O. Malaterre. *SIC 4.20, Simulation of Irrigation Canals*. http://www.cemagref.net/sic/sicgb.htm. 2006.

[41] R. Moussa and C. Bocquillon. "Criteria for the choice of flood-routing methods in natural channels". In: *Journal of Hydrology* 186 (1996), pp. 1–30.

[42] Mee Young Park and Trevor Hastie. "ℓ_1-regularization path algorithm for generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4 (2007), pp. 659–677.

[43] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[44] N. Petit and P. Rouchon. "Dynamics and solutions to some control problems for water-tank systems". In: *IEEE Transactions on Automatic Control* 47.4 (2002), pp. 594–609.

[45] N. Petit and P. Rouchon. "Flatness of heavy chain systems". In: *SIAM Journal on Control and Optimization* 40 (2) (2001), pp. 475–495.

[46] N. Petit et al. "Motion planning for two classes of nonlinear systems with delays depending on the control". In: *Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL* 1 (December 1998), pp. 1007–1011.

[47] A.D. Polyanin. *Handbook of linear partial differential equations for engineers and scientists*. London, UK: Chapman & Hall/CRC, 2002.

[48] T. Rabbani et al. "Feed-forward control of open channel flow using differential flatness". In: *IEEE Transactions on Control Systems Technology* (to appear 2009).

[49] L. Rodino. *Linear partial differential operators in Gevrey spaces*. River Edge, NJ: World Scientific, 1993.

[50] J. Rudolph. "Planning trajectories for a class of linear partial differential equations: an introduction". In: *Sciences et Technologies de l'Automatique. Electronic Journal: http://www. esta. see. asso. fr* (2004).

[51] A. J. C. Barré de Saint-Venant. "Théorie du mouvement non-permanent des eaux avec application aux crues des rivières à l'introduction des marées dans leur lit". In: *Comptes rendus de l'Académie des Sciences* 73 (1871), pp. 148–154, 237–240.

[52] B.F. Sanders and N.D. Katopodes. "Adjoint sensitivity analysis for shallow-water wave control". In: *Journal of Engineering Mechanics* 126.9 (2000), pp. 909–919.

[53] T. Sturm. *Open channel hydraulics*. New York, NY: McGraw-Hill Science Engineering, 2001.

[54] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 0035-9246.

[55] A. Tveito and R. Winther. *Introduction to partial differential equations: a computational approach*. Springer, Berlin, 1998.

[56] A. Yang et al. "Fast ℓ_1-minimization algorithms and an application in robust face recognition: a review". In: *University of California at Berkeley Technical report UCB/EECS-2010-13* (2010).