Constrained least squares as the limit solution of a regularized least squares problem

Saurabh V. Pendse*[1]

[1] Department of Computer Science and Engineering
The Maharaja Sayajirao University of Baroda, Gujarat, India

May 23, 2010

Abstract

One of the most flexible techniques for function approximation in statistics is the basis function based linear regression. By choosing a sufficiently flexible set of basis functions, one can model complicated functions accurately. Smoothness or complexity of the resulting functions can be controlled by introducing a regularization term in the objective function. In this technical note, we describe how one can solve the function estimation problem in the presence of simple linear constraints. In particular, we show that the constrained least squares solution can be obtained as the limiting solution of a particular regularized least squares problem.

1 Notation

Vectors will be denoted using lower case letters in bold face, for example $\boldsymbol{x_i}, \boldsymbol{y}$. Matrices will be denoted using upper case letters in bold face, for example

*To whom correspo dence should be addressed. e-mail: sau2pen@gmail.com

$\boldsymbol{X}, \boldsymbol{A}$. The jth element of a vector $\boldsymbol{x_i}$ will be denoted by x_{ij} whereas the jth element of a vector $\boldsymbol{y}$ will be denoted by y_j. The ijth element of a matrix $\boldsymbol{X}$ will be denoted by X_{ij}. The transpose of a matrix $\boldsymbol{A}$ will be denoted by $\boldsymbol{A}^T$ and its inverse will be denoted by $\boldsymbol{A}^{-1}$. We will use $\mathbf{0}$ to denote a vector or matrix of all zeros whose size should be clear from context.

2 Introduction

Suppose we are given n input vectors, $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ each of size $p \times 1$. Corresponding to each $\boldsymbol{x_i}$ we observe an output variable y_i to get a set of n observed outputs $y_1, y_2, \ldots, y_n$. A linear model explaining the observed y_i in terms of $\boldsymbol{x_i}$ is given by:

$$y_i = \boldsymbol{x_i}^T \boldsymbol{\psi} + \varepsilon_i \tag{2.1}$$

where $\boldsymbol{\psi}$ is a $p \times 1$ coefficient vector and ε_i is a scalar random variable representing the unobserved noise corruption with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2$. We also assume that ε_i is independent across i. A linear equation such as 2.1 has limited flexibility and might not be appropriate for modeling real world data. This problem is easily solved by first transforming the $p \times 1$ input vectors $\boldsymbol{x_i}$ into $m \times 1$ vectors $\boldsymbol{\phi}(\boldsymbol{x_i})$ where $\boldsymbol{\phi} : \mathbf{R}^p \to \mathbf{R}^m$ is a non-linear function which takes p dimensional vectors to m dimensional vectors. Each component ϕ_i of $\boldsymbol{\phi}$ represents one basis function. Next, we construct a linear model in the transformed space:

$$y_i = \boldsymbol{\phi}(\boldsymbol{x_i})^T \boldsymbol{\beta} + \varepsilon_i \tag{2.2}$$

Here $\boldsymbol{\beta}$ is a $m \times 1$ coefficient vector corresponding to the m basis functions. To simplify notation define:

$$\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T \text{ is a } n \times 1 \text{ vector} \tag{2.3}$$

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}(\boldsymbol{x_1}), \boldsymbol{\phi}(\boldsymbol{x_2}), \ldots, \boldsymbol{\phi}(\boldsymbol{x_n})]^T \text{ is a } n \times m \text{ matrix} \tag{2.4}$$

$$\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T \text{ is a } n \times 1 \text{ vector} \tag{2.5}$$

In this work, we will assume that $n > m$ and that $\text{rank}(\boldsymbol{\Phi}) = m$. Then we can re-write 2.2 as:

$$\boldsymbol{y} = \boldsymbol{\Phi}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.6}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{I}_n$. Given $\boldsymbol{y}$ and $\boldsymbol{\Phi}$ the problem is to estimate the unknown coefficient vector $\boldsymbol{\beta}$. In the least squares solution

approach, we estimate $\boldsymbol{\beta}$ by solving the optimization problem:

$$\min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) \tag{2.7}$$

Taking the gradient of 2.7 and equating it to $\mathbf{0}$ at the least squares solution point $\boldsymbol{\beta}_{LS}$ we get:

$$\frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{\beta})|_{\boldsymbol{\beta}_{LS}} = 2\boldsymbol{\Phi}^T (\boldsymbol{\Phi}\boldsymbol{\beta}_{LS} - \boldsymbol{y}) = \mathbf{0} \tag{2.8}$$

Since $\boldsymbol{\Phi}$ has full column rank, upon simplification, we get:

$$\boldsymbol{\beta}_{LS} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} \tag{2.9}$$

A desirable property of $\boldsymbol{\beta}_{LS}$ is that it is an unbiased estimator of $\boldsymbol{\beta}$ in the presence of noise $\boldsymbol{\varepsilon}$. Assuming that the true $\boldsymbol{\beta}$ is given, it is easy to see from 2.9 and 2.6 that $E(\boldsymbol{\beta}_{LS}) = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T E(\boldsymbol{y}) = \boldsymbol{\beta}$. Since we are given only a finite number of points n, the error function $f(\boldsymbol{\beta})$ can be made arbitrarily small by selecting a sufficiently flexible set of basis functions $\boldsymbol{\phi}(\boldsymbol{x})$. For example, any function $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\beta}$ that passes through the given points $\boldsymbol{y}$ exactly will make the objective $f(\boldsymbol{\beta}) = 0$. How can we impose conditions on the estimated function $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\beta}$ to avoid a non-sensical noisy solution? One way is to add a penalty term to the objective function $f(\boldsymbol{\beta})$ that will encourage smoothness in the estimated function $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\beta}$. Let $g(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\beta} = \sum_{i=1}^m \phi_i(\boldsymbol{x})\,\beta_i$ then

$$\frac{\partial^2 g(\boldsymbol{x})}{\partial x_k \partial x_l} = \sum_{i=1}^m \frac{\partial^2 \phi_i(\boldsymbol{x})}{\partial x_k \partial x_l}\,\beta_i \tag{2.10}$$

Using the Cauchy-Schwarz inequality we can write:

$$\left[\frac{\partial^2 g(\boldsymbol{x})}{\partial x_k \partial x_l}\right]^2 \leq \sum_{i=1}^m \left[\frac{\partial^2 \phi_i(\boldsymbol{x})}{\partial x_k \partial x_l}\right]^2 \sum_{i=1}^m \beta_i^2 \tag{2.11}$$

For a fixed set of basis functions $\phi_i(\boldsymbol{x})$ we can impose smoothness by forcing $\left[\frac{\partial^2 g(\boldsymbol{x})}{\partial x_k \partial x_l}\right]^2$ to have small values for all k and l. It is clear from 2.11 that this can be attained by making $\sum_{i=1}^m \beta_i^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}$ small. This motivates the addition of a penalty term proportional to $\boldsymbol{\beta}^T\boldsymbol{\beta}$ into $f(\boldsymbol{\beta})$ in 2.7 to get a regularized least squares problem also known as ridge regression (see [1]):

$$\min_{\boldsymbol{\beta}} \; f_{reg}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) + \lambda\, \boldsymbol{\beta}^T\boldsymbol{\beta} \tag{2.12}$$

Here the parameter λ controls the degree of regularization or smoothness. Higher values of λ force a smoother function and lower values of λ encourage a rougher function. In addition to imposing regularization it might be desirable for $\boldsymbol{\beta}$ to satisfy certain linear constraints. Suppose $\boldsymbol{G}$ is a $q \times m$ matrix with $q < m$ and $\text{rank}(\boldsymbol{G}) = q$ and let $\boldsymbol{c}$ be a $q \times 1$ vector. We would like $\boldsymbol{\beta}$ to satisfy the constraints:

$$\boldsymbol{G}\,\boldsymbol{\beta} = \boldsymbol{c} \tag{2.13}$$

We propose to introduce the constraints in 2.13 as "soft" constraints into the objective function in 2.12. We do not require the exact satisfaction of these soft constraints but only encourage their satisfaction. Define the modified optimization problem which encourages satisfaction of 2.13 as:

$$\min_{\boldsymbol{\beta}} \; f_{con}(\boldsymbol{\beta}) = \underbrace{(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})}_{\text{model fit}} + \underbrace{\lambda\,\boldsymbol{\beta}^T\boldsymbol{\beta}}_{\text{smoothness}} + \underbrace{\theta\,(\boldsymbol{G}\,\boldsymbol{\beta} - \boldsymbol{c})^T(\boldsymbol{G}\,\boldsymbol{\beta} - \boldsymbol{c})}_{\text{soft constraints}} \tag{2.14}$$

First we obtain a solution $\boldsymbol{\beta}(\lambda, \theta)$ of the problem 2.14 and then prove that $\lim_{\theta \to \infty} \boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{\beta}^*(\lambda)$ where $\boldsymbol{\beta}^*(\lambda)$ solves the "hard" constraint optimization problem:

$$\min_{\boldsymbol{\beta}} \; f_{hcon}(\boldsymbol{\beta}) = \underbrace{(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta})}_{\text{model fit}} + \underbrace{\lambda\,\boldsymbol{\beta}^T\boldsymbol{\beta}}_{\text{smoothness}} \tag{2.15}$$

$$\text{subject to :} \quad \underbrace{\boldsymbol{G}\,\boldsymbol{\beta} = \boldsymbol{c}}_{\text{hard constraint}}$$

3 The soft constrained problem

Suppose $\boldsymbol{\beta}(\lambda, \theta)$ is a solution of the soft constrained problem 2.14. The necessary conditions for optimality are:

$$\frac{\partial f_{con}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}(\lambda,\theta)} = \left[-2\boldsymbol{\Phi}^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\beta}) + 2\lambda\,\boldsymbol{\beta} + 2\theta\,\boldsymbol{G}^T(\boldsymbol{G}\boldsymbol{\beta} - \boldsymbol{c})\right]|_{\boldsymbol{\beta}(\lambda,\theta)} = \boldsymbol{0} \tag{3.1}$$

Rearranging we get:

$$[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m + \theta\,\boldsymbol{G}^T\boldsymbol{G}]\,\boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{\Phi}^T\boldsymbol{y} + \theta\,\boldsymbol{G}^T\boldsymbol{c} \tag{3.2}$$

Note that $[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m + \theta\,\boldsymbol{G}^T\boldsymbol{G}]$ is invertible for all λ and θ since $\boldsymbol{\Phi}$ is $n \times m$ with rank m and $n > m$. Thus we can write the unique solution $\boldsymbol{\beta}(\lambda, \theta)$ for 2.14 as:

$$\boldsymbol{\beta}(\lambda, \theta) = [\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m + \theta\,\boldsymbol{G}^T\boldsymbol{G}]^{-1}\{\boldsymbol{\Phi}^T\boldsymbol{y} + \theta\,\boldsymbol{G}^T\boldsymbol{c}\} \tag{3.3}$$

To improve computational stability, we can use the Sherman-Morrison-Woodbury formula to compute the matrix inverse in 3.3:

$$[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m + \theta\,\boldsymbol{G}^T\boldsymbol{G}]^{-1} = \quad \boldsymbol{R} - \boldsymbol{R}\,\boldsymbol{G}^T \left[\frac{1}{\theta}\boldsymbol{I}_q + \boldsymbol{G}\boldsymbol{R}\boldsymbol{G}^T\right]^{-1} \boldsymbol{G}\,\boldsymbol{R} \tag{3.4}$$

where

$$\boldsymbol{R} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1} \tag{3.5}$$

Note also that:

$$\frac{\partial^2 f_{con}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T} = 2\,[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m + \theta\,\boldsymbol{G}^T\boldsymbol{G}] \tag{3.6}$$

Clearly $\frac{\partial^2 f_{con}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}$ is positive definite and hence $\boldsymbol{\beta}(\lambda, \theta)$ is the unique minimizer of 2.14.

4 The limit solution as $\theta \to \infty$

A simple way of controlling the "softness" of the constraint 2.13 is to vary the constraint controlling parameter θ in the objective function 2.14. In this section we study the properties of the soft constrained solution as $\theta \to \infty$.

Proposition 4.1. *Suppose $\boldsymbol{\beta}(\lambda, \theta)$ is a solution to the soft constrained problem 2.14 then we have:*

1. $$\lim_{\theta\to\infty} \boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{G} \lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{c} \tag{4.1}$$

2. $$\lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\left\{\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}^T\boldsymbol{\delta}\right\} \tag{4.2}$$

   where

$$\boldsymbol{\delta} = \left[\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1}\left\{\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{c}\right\} \tag{4.3}$$

Proof. Since $\boldsymbol{\beta}(\lambda, \theta)$ is a solution to 2.14 it must satisfy the necessary conditions for optimality in 3.2. Pre-multiplying both sides of equation 3.2 by $\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}$ we get:

$$\left[\boldsymbol{I}_q + \theta\, \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]\, \boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\, \boldsymbol{I}_m)^{-1}\left[\boldsymbol{\Phi}^T\boldsymbol{y} + \theta\, \boldsymbol{G}^T\boldsymbol{c}\right] \tag{4.4}$$

Since $(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}$ is symmetric and positive definite, the $q \times q$ matrices $\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T$ and $\left[\boldsymbol{I}_q + \theta\, \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]$ are both invertible. Therefore we get:

$$\boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \left[\boldsymbol{I}_q + \theta\, \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1} \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\, \boldsymbol{I}_m)^{-1}\left[\boldsymbol{\Phi}^T\boldsymbol{y} + \theta\, \boldsymbol{G}^T\boldsymbol{c}\right] \tag{4.5}$$

Re-arranging the right hand side we can write:

$$\boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \left[\frac{\boldsymbol{I}_q}{\theta} + \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1} \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\, \boldsymbol{I}_m)^{-1}\left[\frac{\boldsymbol{\Phi}^T\boldsymbol{y}}{\theta} + \boldsymbol{G}^T\boldsymbol{c}\right] \tag{4.6}$$

Taking the limit as $\theta \to \infty$ on both sides and noting that $\frac{1}{\theta} \to 0$ as $\theta \to \infty$ we get:

$$\lim_{\theta\to\infty} \boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{G} \lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = \left[\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1} \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\lambda\, \boldsymbol{I}_m)^{-1}\left[\boldsymbol{G}^T\boldsymbol{c}\right] \tag{4.7}$$

Simplifying equation 4.7 we get:

$$\lim_{\theta\to\infty} \boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{G} \lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{c} \tag{4.8}$$

This proves part 1 of the proposition. Now re-arranging equation 3.2 we can write:

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m)\, \boldsymbol{\beta}(\lambda, \theta) + \boldsymbol{G}^T\left[\underbrace{\theta}_{\text{tends to } \infty}\ \underbrace{\{\boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) - \boldsymbol{c}\}}_{\text{tends to } 0}\right] = \boldsymbol{\Phi}^T\boldsymbol{y} \tag{4.9}$$

In the above equation $\theta \to \infty$ and from 4.8 $\lim_{\theta\to\infty} \{\boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) - c\} \to \boldsymbol{0}$. Let

$$\lim_{\theta\to\infty} \theta\ \{\boldsymbol{G}\boldsymbol{\beta}(\lambda, \theta) - c\} = \boldsymbol{\delta} \tag{4.10}$$

where δ is a $q \times 1$ vector. Taking $\lim_{\theta\to\infty}$ on both sides of 4.9 and using 4.10 we get:

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\, \boldsymbol{I}_m) \lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) + \boldsymbol{G}^T\boldsymbol{\delta} = \boldsymbol{\Phi}^T\boldsymbol{y} \tag{4.11}$$

Pre-multiplying both sides of 4.11 by $\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}$ we get:

$$\boldsymbol{G}\lim_{\theta\to\infty}\boldsymbol{\beta}(\lambda,\theta) + \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\boldsymbol{\delta} = \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \quad (4.12)$$

Using 4.8 and noting the non-singularity of $\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T$ we can solve for $\boldsymbol{\delta}$ to get:

$$\boldsymbol{\delta} = \left[\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1}\left\{\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{c}\right\} \quad (4.13)$$

With this value of $\boldsymbol{\delta}$ we can use 4.11 to get:

$$\lim_{\theta\to\infty}\boldsymbol{\beta}(\lambda,\theta) = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\left\{\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}^T\boldsymbol{\delta}\right\} \quad (4.14)$$

This proves part 2 of the proposition. □ □

5 Equivalence to solution of KKT system

In this section, we will show that the solution to the hard constrained problem 2.15 as obtained by a direct solution of Karush-Kuhn-Tucker (KKT) optimality conditions coincides with the solution to the soft constrained problem as the constraint enforcing parameter $\theta \to \infty$.

Proposition 5.1. *Suppose $\boldsymbol{\beta}(\lambda,\theta)$ is a solution to the soft constrained problem 2.14 and let $\boldsymbol{\beta}^*(\lambda)$ be a solution to the hard constrained problem 2.15 then*

$$\lim_{\theta\to\infty}\boldsymbol{\beta}(\lambda,\theta) = \boldsymbol{\beta}^*(\lambda) \quad (5.1)$$

Proof. First consider the hard constrained problem 2.15. The Lagrangian for this constrained problem is:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\alpha}) = f_{hcon}(\boldsymbol{\beta}) - \boldsymbol{\alpha}^T\left\{\boldsymbol{G}\,\boldsymbol{\beta} - \boldsymbol{c}\right\} \quad (5.2)$$

where $\boldsymbol{\alpha}$ is a $q\times 1$ vector of Lagrange multipliers for the q constraints in 2.13. If $\boldsymbol{\beta}^*(\lambda)$ is a solution to 2.15 and $\boldsymbol{\alpha}^*$ the corresponding Lagrange multiplier then the KKT necessary conditions for optimality are:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\alpha})|_{\boldsymbol{\beta}^*(\lambda),\boldsymbol{\alpha}^*} = \left[\frac{\partial}{\partial\boldsymbol{\beta}}f_{hcon}(\boldsymbol{\beta}) - \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\alpha}^T\left\{\boldsymbol{G}\,\boldsymbol{\beta} - \boldsymbol{c}\right\}\right]_{\boldsymbol{\beta}^*(\lambda),\boldsymbol{\alpha}^*} = \boldsymbol{0} \quad (5.3)$$

$$\boldsymbol{G}\,\boldsymbol{\beta}^*(\lambda) - \boldsymbol{c} = \boldsymbol{0} \quad (5.4)$$

Now from 2.15 we have:

$$\frac{\partial}{\partial \boldsymbol{\beta}} f_{hcon}(\boldsymbol{\beta})|_{\boldsymbol{\beta}^*(\lambda),\boldsymbol{\alpha}^*} = 2\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{I}_m\right)\boldsymbol{\beta}^*(\lambda) - 2\boldsymbol{\Phi}^T\boldsymbol{y} \tag{5.5}$$

Also

$$\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\alpha}^T\left\{\boldsymbol{G}\,\boldsymbol{\beta} - \boldsymbol{c}\right\}|_{\boldsymbol{\beta}^*(\lambda),\boldsymbol{\alpha}^*} = \boldsymbol{G}^T\boldsymbol{\alpha}|_{\boldsymbol{\beta}^*(\lambda),\boldsymbol{\alpha}^*} = \boldsymbol{G}^T\boldsymbol{\alpha}^* \tag{5.6}$$

Substituting 5.5 and 5.6 we can re-write the KKT optimality conditions as:

$$\begin{gathered} 2\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{I}_m\right)\boldsymbol{\beta}^*(\lambda) - 2\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}^T\boldsymbol{\alpha}^* = \boldsymbol{0} \\ \boldsymbol{G}\,\boldsymbol{\beta}^*(\lambda) - \boldsymbol{c} = \boldsymbol{0} \end{gathered} \tag{5.7}$$

Pre-multiplying both sides of the first equation in 5.7 by $\frac{1}{2}\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}$ we get:

$$\boldsymbol{G}\,\boldsymbol{\beta}^*(\lambda) - \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\left(\frac{\boldsymbol{\alpha}^*}{2}\right) = \boldsymbol{0} \tag{5.8}$$

Now $\boldsymbol{G}\,\boldsymbol{\beta}^*(\lambda) = \boldsymbol{c}$ from the second equation in 5.7. Noting the non-singularity of $\left(\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right)$ we can solve for $\left(\frac{\boldsymbol{\alpha}^*}{2}\right)$ to get:

$$\left(\frac{\boldsymbol{\alpha}^*}{2}\right) = \left[\boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{G}^T\right]^{-1}\left\{\boldsymbol{c} - \boldsymbol{G}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}\right\} \tag{5.9}$$

Comparing 5.9 and 4.3 from Proposition 4.1 it is clear that:

$$\left(\frac{\boldsymbol{\alpha}^*}{2}\right) = -\boldsymbol{\delta} \tag{5.10}$$

Solving for $\boldsymbol{\beta}^*(\lambda)$ from the first equation in 5.7 we get:

$$\boldsymbol{\beta}^*(\lambda) = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I}_m\right)^{-1}\left\{\boldsymbol{\Phi}^T\boldsymbol{y} + \boldsymbol{G}^T\left(\frac{\boldsymbol{\alpha}^*}{2}\right)\right\} \tag{5.11}$$

From 5.10 and 5.11 we can write:

$$\boldsymbol{\beta}^*(\lambda) = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I}_m\right)^{-1}\left\{\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}^T\boldsymbol{\delta}\right\} \tag{5.12}$$

Comparing 5.12 and 4.14 we get:

$$\lim_{\theta\to\infty}\boldsymbol{\beta}(\lambda,\theta) = \boldsymbol{\beta}^*(\lambda) = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\,\boldsymbol{I}_m)^{-1}\left\{\boldsymbol{\Phi}^T\boldsymbol{y} - \boldsymbol{G}^T\boldsymbol{\delta}\right\} \tag{5.13}$$

which proves the proposition. □ □

6 Illustrative Experiment

To illustrate the results in Propositions 4.1 and 5.1 we performed a simple numerical experiment. We generated example data for $\boldsymbol{y}$ and $\boldsymbol{\Phi}$ as follows:

1. In MATLAB notation suppose $\boldsymbol{e}$ =`ones(100,1)` and $\boldsymbol{u}$ =`linspace(0,1,100)`^{T}. Suppose $\boldsymbol{U}$ be a diagonal matrix with the elements of $\boldsymbol{u}$ on the diagonal, in other words $U(i,i) = u_i$

2. Let $\boldsymbol{\Phi} = [\boldsymbol{e}, \boldsymbol{u}, \boldsymbol{U}\,\boldsymbol{u}, \boldsymbol{U}^2\,\boldsymbol{u}, \boldsymbol{U}^3\,\boldsymbol{u}]$. Here $\boldsymbol{u}$ is a linear basis function, $\boldsymbol{U}\,\boldsymbol{u}$ is a quadratic basis function, $\boldsymbol{U}^2\,\boldsymbol{u}$ is a cubic basis function and so on

3. We generated a random coefficient vector $\boldsymbol{\beta}$ =`unifrnd(-1,1,5,1)`$\times 10$

4. Next we generate $\boldsymbol{y} = \boldsymbol{\Phi}\,\boldsymbol{\beta}$ and set $\lambda = 1$

5. We chose $\boldsymbol{G}$ =`unifrnd(0,1,2,5)` and $\boldsymbol{c}$ =`unifrnd(0,1,2,1)` to create a set of 2 linear constraints on $\boldsymbol{\beta}$

Note that in this dummy example $n = 100$, $m = 5$ and $q = 2$. Next we chose a 170×1 vector of potential values for θ in 2.14 as:

$$\boldsymbol{\theta}_{vec} = \texttt{[linspace(0,30,100),linspace(30,100,50),linspace(100,1000,20)]}^T$$

For each i from 1 to 170 we set $\theta = \theta_{vec}(i)$ and solve the soft constrained problem 2.14 using 3.3 to get $\boldsymbol{\beta}(\lambda, \theta)$. For the same θ we also solve the hard constrained problem 2.15 using 5.13 and 4.3 to get $\boldsymbol{\beta}^*(\lambda)$. By Proposition 5.1 we have $\lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{\beta}^*(\lambda)$. To verify this numerically we simply plot the components of $\boldsymbol{\beta}(\lambda, \theta)$ on the x-axis and the corresponding components of $\boldsymbol{\beta}^*(\lambda)$ on the y-axis. As $\boldsymbol{\beta}(\lambda, \theta)$ gets closer and closer to $\boldsymbol{\beta}^*(\lambda)$ with increasing θ, the plotted points should get closer to the $y = x$ line and eventually lie on the $y = x$ line. Figure 1 shows a plot of $\boldsymbol{\beta}(\lambda, \theta)$ versus $\boldsymbol{\beta}^*(\lambda)$ for 4 values $\theta = 0, 4, 20, 1000$. It can be seen that as θ increases the components of $\boldsymbol{\beta}(\lambda, \theta)$ converge to the components of $\boldsymbol{\beta}^*(\lambda)$.

7 Discussion

In this work we considered least squares problems with smoothness regularization and enforcement of constraints in the soft form by setting up the optimization problem 2.14. First, we solve 2.14 to get a solution $\boldsymbol{\beta}(\lambda, \theta)$

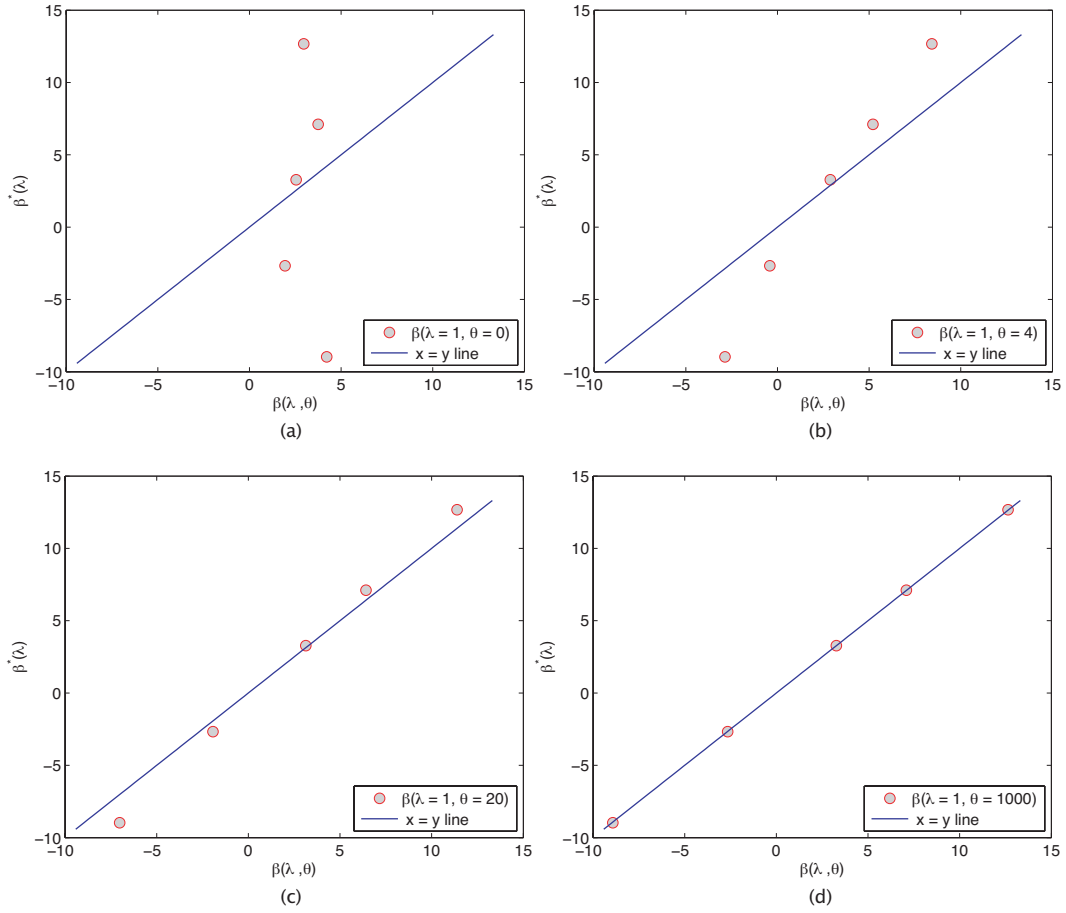

Figure 1: Figure shows the behavior of regularized and soft constrained solutions $\boldsymbol{\beta}(\lambda, \theta)$ for various values of θ for a fixed $\lambda = 1$. The x-axis shows components of $\boldsymbol{\beta}(\lambda, \theta)$. The corresponding components of the hard constrained solution $\boldsymbol{\beta}^*(\lambda)$ are shown on the y-axis. It can be seen that as $\theta \to \infty$ the agreement between $\boldsymbol{\beta}(\lambda, \theta)$ and $\boldsymbol{\beta}^*(\lambda)$ increases.

that is parameterized by λ and θ. Next, we study the behavior of this solution as $\theta \to \infty$ in Proposition 4.1 and show in Proposition 5.1 that in fact $\lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta) = \boldsymbol{\beta}^*(\lambda)$ where $\boldsymbol{\beta}^*(\lambda)$ is the solution to the hard constrained optimization problem 2.15. We also illustrate this convergence phenomenon via a simple illustrative experiment in Section 6.

Enforcement of constraints in a soft manner as in 2.14 is similar to the quadratic penalty method in constrained optimization [2]. Results related to limiting behavior such as 5.1 have also been proven under more general conditions in [2] given the assumption of convexity of the intermediate quadratic penalty problems. It is worth noting that the limiting process $\lim_{\theta\to\infty} \boldsymbol{\beta}(\lambda, \theta)$ could be difficult to carry out numerically because of the ill conditioning inherent in the problem as $\theta \to \infty$ [2]. However, we have circumvented this difficulty in the present case by using the Sherman-Morrison-Woodbury analytical matrix inversion in 3.4.

8 Conclusion

Solution to problems such as 2.14 could find many practical applications since they not only impose smoothness via the parameter λ on the inferred function but also enable the modeler to impose soft constraints via the parameter θ. By varying λ and θ, a family of solutions with various properties is obtained. It is re-assuring to know that increasing θ will eventually lead to the fully constrained solution corresponding to the problem 2.15.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[2] J. Nocedal and S.J. Wright. *Numerical Optimization*. New York:Springer, 1999.