Regularized Approximation

Fangling Du

Summer 2024

1 Introduction

Regularized approximation techniques are crucial in the field of convex optimization, particularly when dealing with ill-posed problems or problems where the solution needs to meet specific practical requirements. Here are some details about the background.

Background

- **Ill-Posed Problems:** Many real-world problems do not have a unique or stable solution. For example, when solving $Ax = b$ where A is an ill-conditioned matrix or nearly singular, leading to unstable solutions. Regularization helps stabilize these solutions by adding additional constraints or modifying the objective function.

- **Overfitting:** In machine learning and statistical modeling, overfitting occurs when a model captures the noise in the data rather than the underlying trend. Regularization techniques such as Ridge Regression (Tikhonov Regularization) and Lasso (l_1-Norm Regularization) add penalties to the model complexity, thus reducing overfitting.

- **Noise:** Real-world data is often noisy and contains measurement errors. Regularization methods help in obtaining solutions that are robust to these inaccuracies by smoothing or filtering out the noise.

Applications: signal processing, statistical estimation, and optimal design.

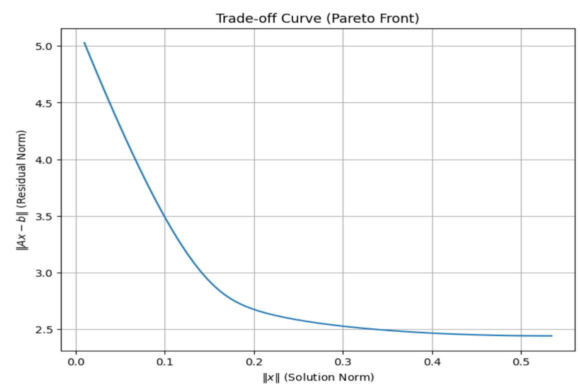
2 Bi-Criterion formulation

The goal is to balance two objectives: the residual norm$\|Ax - b\|$ and the solution norm$\|x\|$.

Formulation

$$minimize(w.r.t.R^2_+) \quad (\|Ax - b\|, \|x\|)$$

2.1 Trade-off curve



2.2 Pareto optimal points

Pareto optimality is a key concept in multi-objective optimization, where we aim to optimize multiple conflicting objectives simultaneously. A solution is considered Pareto optimal if there is no other solution that improves one objective without worsening another.

3 Regularization Techniques

This is a scalarization method to solve the bi-criterion problem.

3.1 Scalarization methods

Weighted sum

$$min\|Ax - b\| + \gamma\|x\|$$

Weighted sum of squares

$$min\|Ax - b\|^2 + \delta\|x\|^2$$

3.2 Tihkonov regularization

It is also known as Ridge Regression and the method minimizes the sum of squared residuals and a penalty term. The main idea is to limit the size of model parameters by adding a penalty term, thereby improving the generalization ability of the model.

Quadratic optimization problem

$$minimize\|Ax - b\|_2^2 + \delta\|x\|_2^2 = x^T(A^TA + \delta I)x - 2b^TAx + b^Tb$$

The Tihkonov regularization problem has the analytical solution

$$x = (A^TA + \delta I)^{-1}A^Tb$$

3.3 Smoothing regularization

Here we add a regularization term of the form $\|Dx\|$ in place of $\|x\|$, where the matrix D represents an approximate differentiation or second-order differentiation operator, so $\|Dx\|$ represents a measure of the variation or smoothness of x.

Tihkonov regularized problem

$$minimize\|Ax - b\|_2^2 + \delta\|\triangle x\|_2^2$$

The parameter δ is used to control the amount of regularization required, or to plot the optimal trade-off curve of fit versus smoothness.

Further

$$minimize\|Ax - b\|_2^2 + \delta\|\triangle x\|_2^2 + \eta\|x\|_2^2$$

We can add many regularization terms where δ is used to control the smoothness of the approximate solution and η is used to control its size.

3.4 l_1-norm regularization

Finding sparse solution

$$minimize\|Ax - b\|_2 + \gamma\|x\|_1$$

3.5 Examples

3.5.1 Example: Optimal input design

Consider a dynamical system

$$y(t) = \sum_{\tau=0}^{t} h(\tau)u(t-\tau), t = 0, 1, ..., N$$

Goal:
1) Output tracking

$$J_{track} = \frac{1}{N+1}\sum_{t=0}^{N}(y(t) - y_{des}(t))^2$$

2) Small input

$$J_{mag} = \frac{1}{N+1}\sum_{t=0}^{N}u(t)^2$$

3) Small input variations

$$J_{der} = \frac{1}{N}\sum_{t=0}^{N-1}(u(t+1) - u(t))^2$$

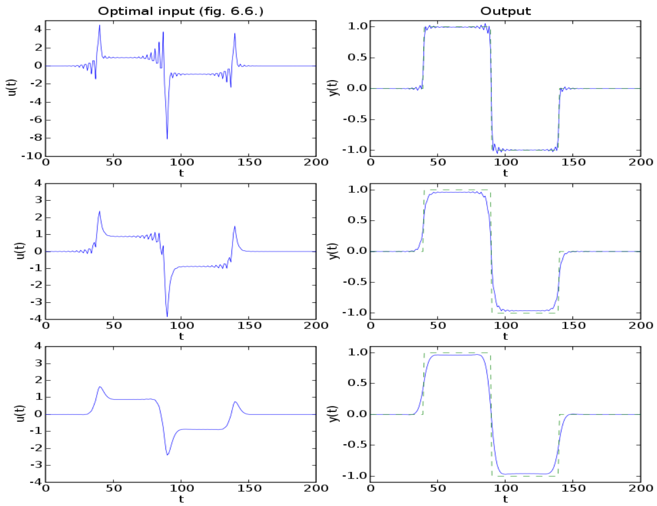
This can be traded off by minimizing the weighted sum

$$J_{track} + \delta J_{der} + \eta J_{mag}$$

Here is a specific example. With N=200, and impulse response

$$h(t) = \frac{1}{9}(0.9)^t(1 - 0.4cos(2t))$$

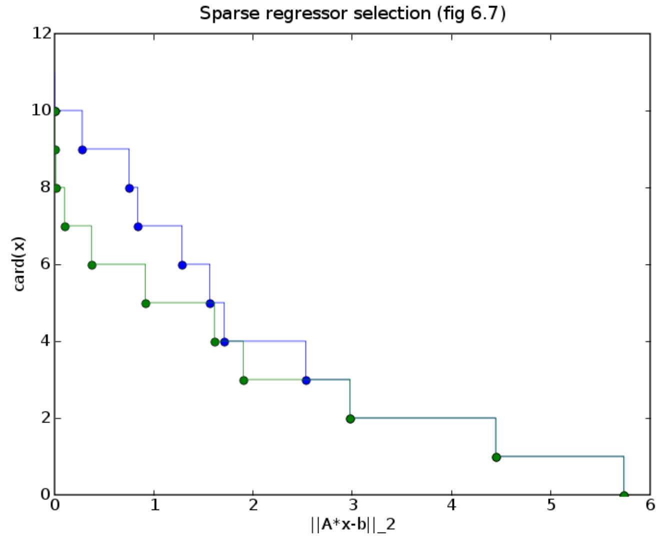
The optimal input and corresponding output for three values of the regularization parameters δ and η are as below. (See fig 6.6)



3.5.2 Example: Regressor selection problem

By varying the parameter γ, we can sweep out the optimal trade-off curve. Here is a specific example. (See fig 6.7) The problem is to choose the subset of k regressors to be used, and the associated coefficients. The problem is

$$minimize\|Ax - b\|_2 \text{ subject to } \textbf{card}(x) \leq k$$



4 Reconstruction, smoothing, and de-nosing

In reconstruction problems, we start with a signal represented by a vector. The coefficients correspond to the value of some function of time, evaluated (or sampled, in the language of signal processing) at evenly spaced points. Usually, we have $x_i \approx x_{i+1}$. The signal is corrupted by an additive noise:

$$x_{cor} = x + v$$

where the noise is unknown, small, and rapidly varying. The goal is to form an estimate $\hat{x}$ of the original signal x, given the corrupted signal x_{cor}. This process is called signal reconstruction or de-nosing. Most reconstruction methods end up performing some sort of smoothing operation on x_{cor} to produce $\hat{x}$, so the process is also called smoothing. The reconstruction problem in this case can be expressed as

bi-criterion problem

$$minimize(w.r.t.\mathbf{R}^2_+) \quad (\|\hat{x} - x_{cor}\|_2, \phi(\hat{x}))$$

Our goal is to find a signal that is as smooth as possible under the l_2-norm as close to the contaminated signal and make a trade-off between closeness and smoothness to effectively reconstruct the original signal.

4.1 Quadratic smoothing

Smoothing function

$$\Phi_{quad}(x) = \sum_{i=1}^{n-1}(x_{i+1} - x_i)^2 = \|Dx\|_2^2$$

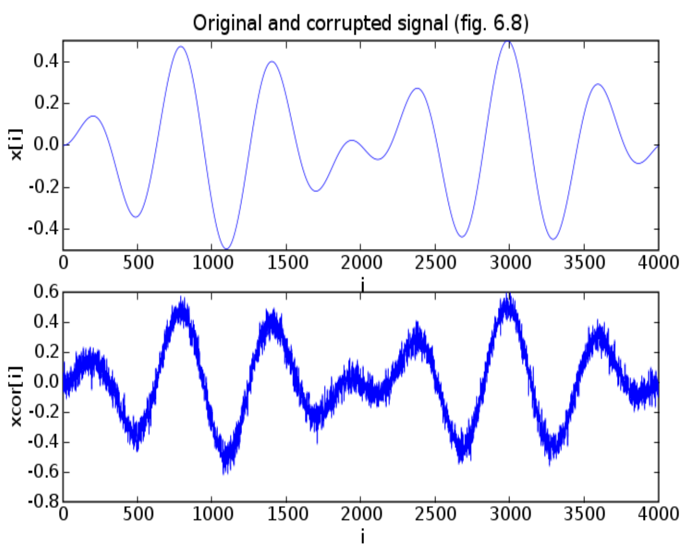
In this function, $D \in R^{(n-1)\times n}$ is the bidiagonal matrix. (See fig 6.8-fig 6.10)

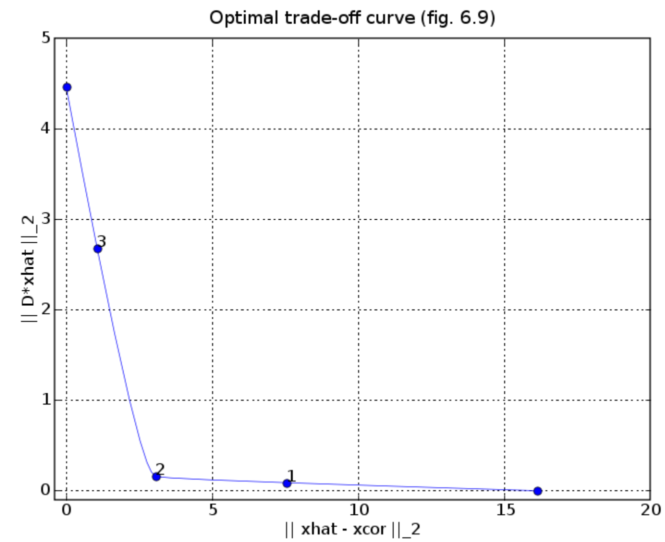
4.2 Total variation reconstruction

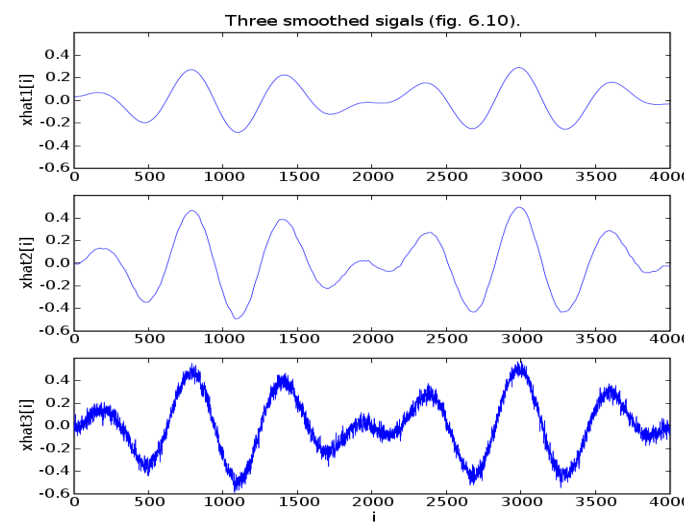
Smoothing function

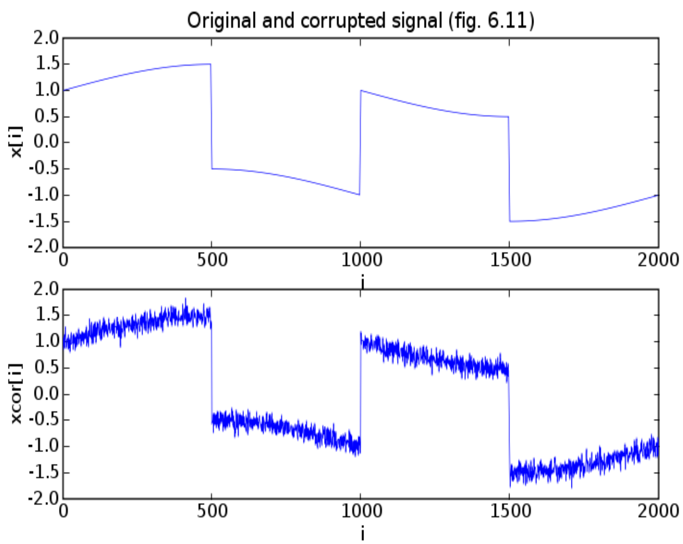
$$\Phi_{tv}(\hat{x}) = \sum_{i=1}^{n-1}|\hat{x_{i+1}} - \hat{x_i}| = \|D\hat{x}\|_1$$

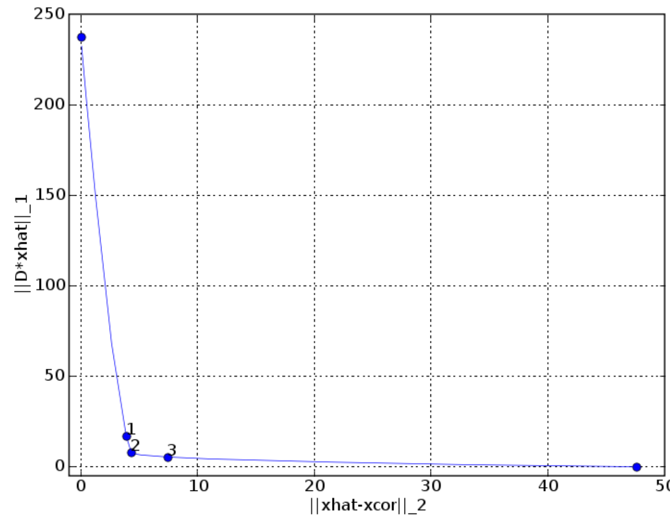
(See fig 6.11-fig 6.14)

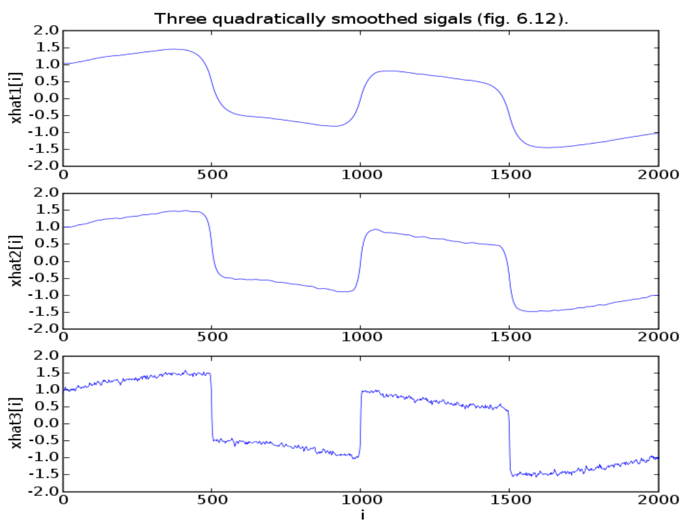


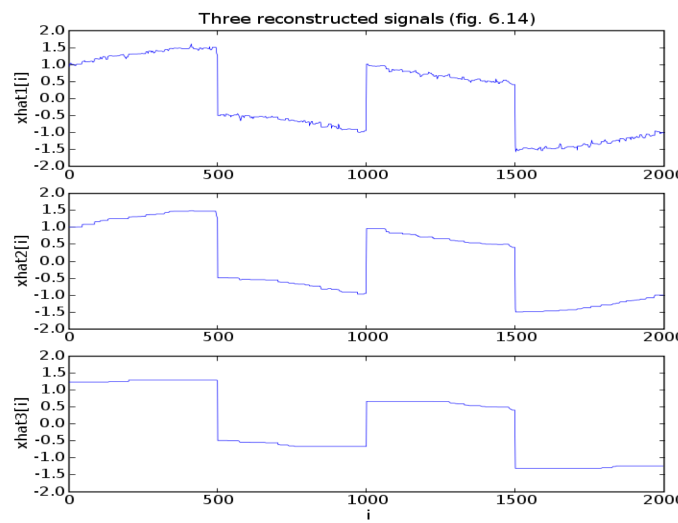





Original and corrupted signal (fig. 6.11)
x[i]
xcor[i]
i


||D*xhat||_1
||xhat-xcor||_2
1
2
3


Three quadratically smoothed sigals (fig. 6.12).
xhat1[i]
xhat2[i]
xhat3[i]
i


Three reconstructed signals (fig. 6.14)
xhat1[i]
xhat2[i]
xhat3[i]
i

5 Summary

5.1 l_1-norm regularization and l_2-norm regularization

- When we perform feature selection, we can use the l_1-norm regularization $\|Dx\|_1 = \sum_i |Dx_i|$. It tends to produce sparse solutions, and it is a convex optimization problem that can be solved by standard convex optimization algorithms.
  Applications: Lasso regression, Sparse coding

- When we want to prevent overfitting and do not care about sparsity, we can use the l_2-norm regularization $\|Dx\|_2 = \sqrt{\sum_i (Dx_i)^2}$. It tends to produce smaller but non-zero weights. It will not produce sparse solutions like the l_1-norm regularization but will smoothly reduce all weights, which helps prevent model overfitting. In addition, it limits the size of weights and makes the model less sensitive to noise in the training data.
  Applications: Ridge regression, Neural network

5.2 Quadratic smoothing and Total variation reconstruction

- Quadratic smoothing works well when the original signal is very smooth, and the noise is rapidly varying. However, this method will attenuate or remove the fast changes in the original signal because it imposes a large penalty on fast changes.

- Total variation reconstruction also assigns a large value to rapidly changing signal $\hat{x}$, but this method imposes a relatively small penalty on $|x_{i+1} - x_i|$, meaning that it is more tolerant of rapid changes in the signal and does not strongly weaken these features. So it is more suitable for signals or images that contain significant edges or discontinuities.