Final Project Math 56:
Unbiased Predictive Risk Estimator Method

Henry Morris

March 2023

Introduction

In class and homework we examined the problem of finding the solution to the slightly modified normal equation $(A^TA + \lambda R^TR)x = A^Tb$ which can be derived using a minimizing function, i.e. the Tikhonov regularization equation

$$\arg\min_x ||Ax - b||_2^2 + \lambda||Rx||_2^2. \tag{1}$$

Here $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, $b \in \mathbb{R}^m$, $R \in \mathbb{R}^{k\times n}$, and $\lambda > 0$. We also assume A has a full column rank.
Solving this modified normal equation, where we have a regularization matrix and a scaling parameter, by itself is not very hard or complicated. For example, we can use the numerically stable QR decomposition method to solve the equation to name just one of many methods. In fact, numpy in python has a built in function that will do this internally. However, this is not the focus of the paper.
What we are examining is how to choose the regularization parameter λ in a linear least squares problem by using the unbiased predictive risk estimator (UPRE). The goal of this method is to best choose λ such that our solution balances the fit to the data, measured of course by the norm of residuals, and the amount of regularization, measured by the norms of our solution.

UPRE Method

Before I explain how the UPRE method works and how I implemented, it is important to know this well known equation sometimes used to determine the UPRE:

$$UPRE(\lambda) = (m \times ||Ax(\lambda) - b||_2^2)/(m - ||Rx(\lambda)||_2^2)^2 + 2 \times Tr(R) \times \lambda \tag{2}$$

. This equation clearly measures the prediction error in terms of the residual error, $||Ax(\lambda) - b||_2^2$ and the regularization term, $||Rx(\lambda)||_2^2$. It follows that the first addend represents the normalized residual error of the regularized solution while the second addend just represents how much smoothing we actually are applying. I will not go into the full derivation here of this UPRE equation, which can be easily found online, but it stems from rewriting our minimization problem as a Bayesian model.
In fact, there is another equation that calculates the Tikhonov-regularized estimator unbiased predictive risk estimate differently by explicitly taking into account the effective degrees of freedom of the regularized solution. This definition is more accurate than the one I have implemented, although it costs more computationally which means it is used more often with smaller data sets while the definition above may be the better implementation for large data sets.

Now, the method of picking a λ with the UPRE method involves the calculation of UPRE in a wide range of λs, and then choosing to save the λ which corresponds to the lowest absolute value of UPRE. It is important to have the absolute value here as when λ get large,

$UPRE(\lambda)$ intuitively becomes negative.

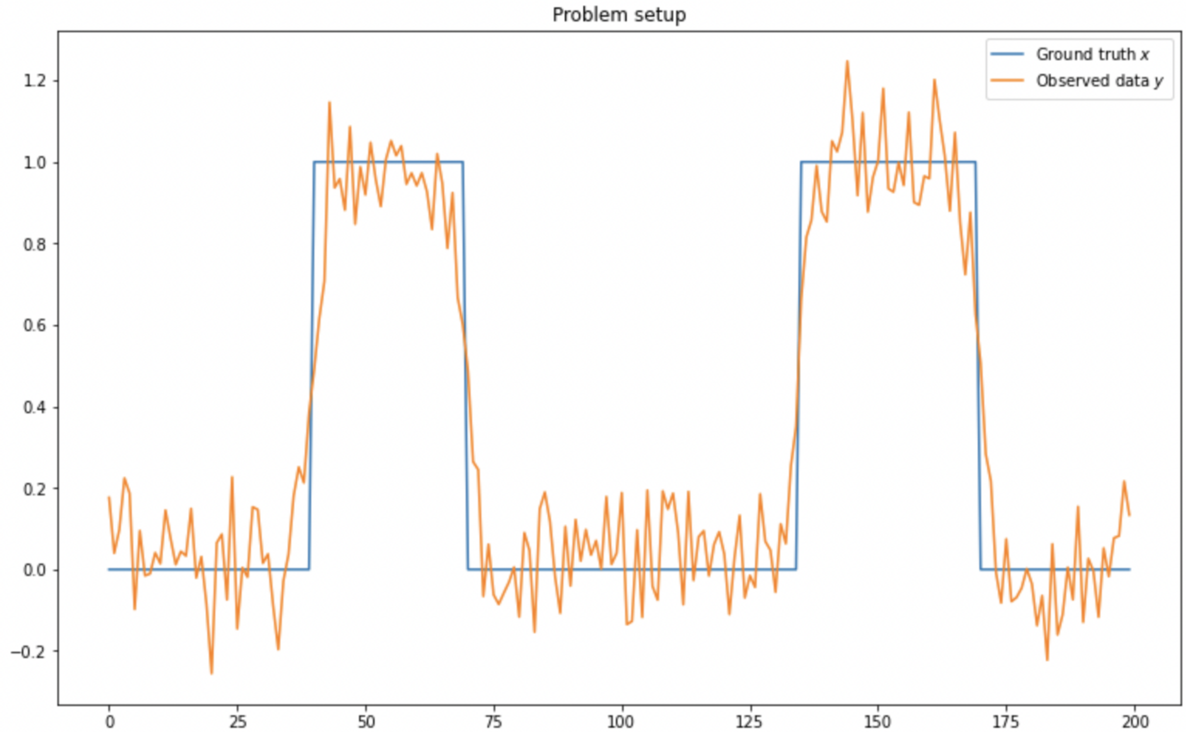
In order to determine what range of λs you want to look at, you can use numerical experimentation to get a sense of what a lower bound and upper bound would be. Within the range of λs, it also makes sense to space the guesses out logarithmically so you have greater precision at lower numbers.

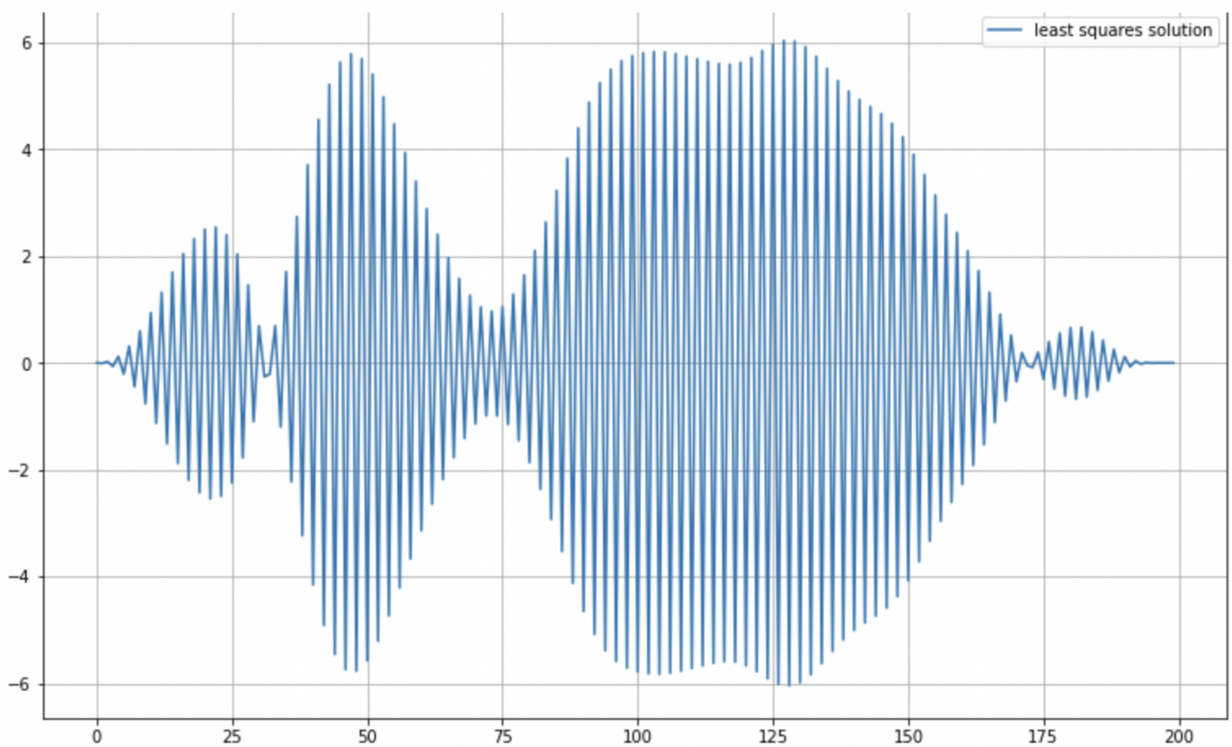
The actual coding method to determine the choice of λ can be seen via this link to my google colab https://colab.research.google.com/drive/1efCpLE1rzYjIvJ1HRnHTtRHT1ro0Nlj4?usp=sharing

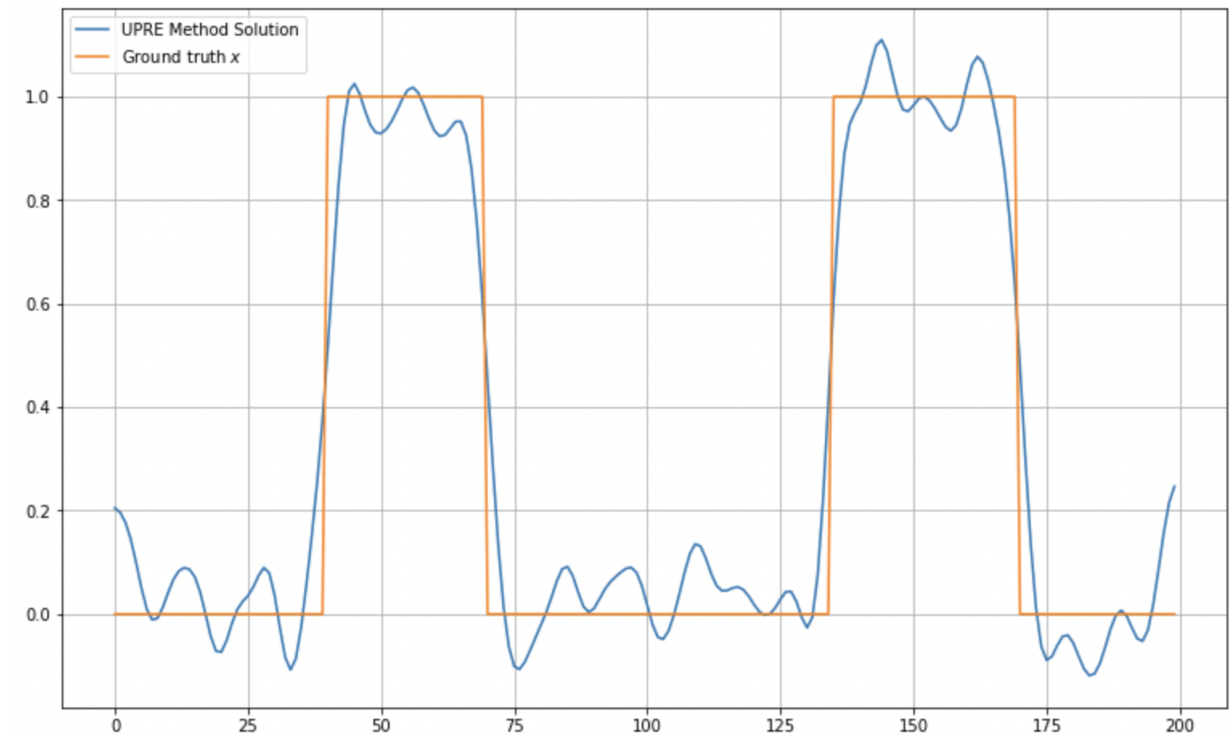
Graphical representation

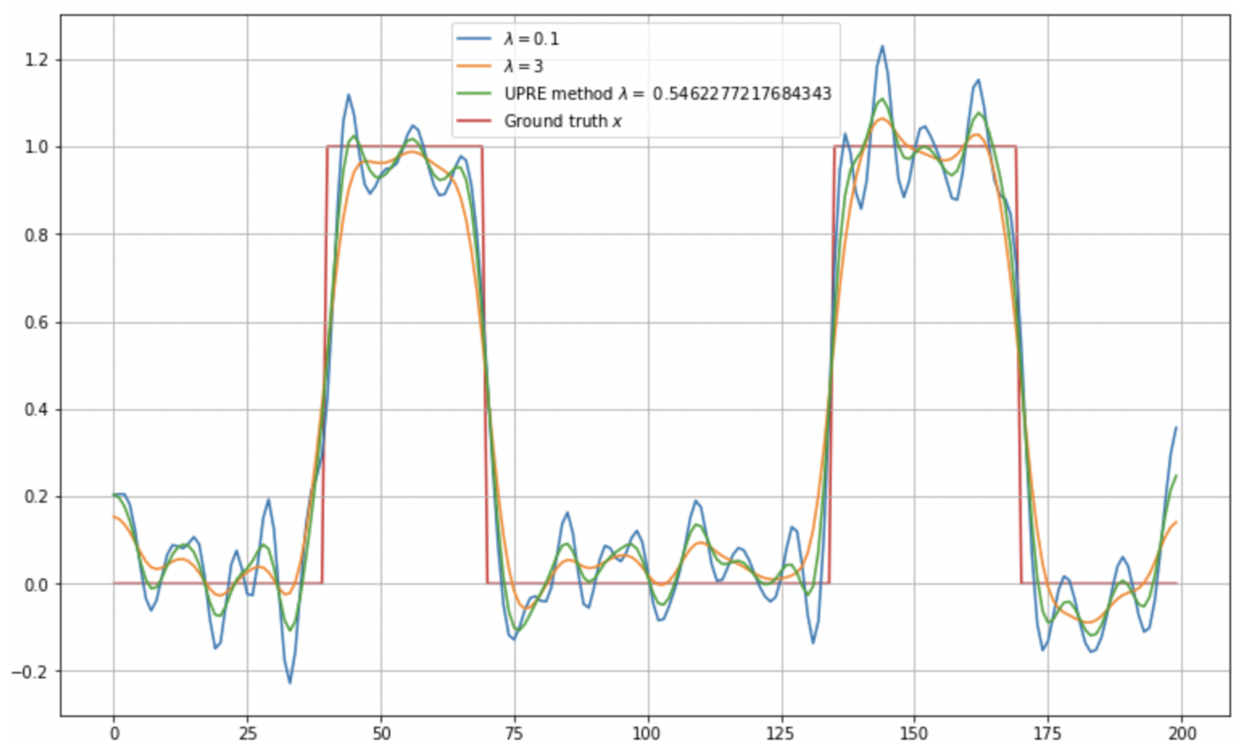
The following graphs were developed as part of HW2, with the code of Jonathon Lindbloom responsible for creating the example, a 1d blurring linear equation, we are testing our algorithms on. We have added the use of setting $\lambda = 0.5462277217684343$ in the Tikhonov regularization equation, which is what our function, using the the UPRE method described in previous section, thinks the scaling parameter should be. The following graphs and all the associated python code can be found on the the google colab as well.

1. The first graph illustrates the ground truth data, x, and the observed blurred data, y.
2. The second graph illustrates the least squares solution to the problem where we let $\lambda = 0$, i.e. there is no regularization.
3. The third graph illustrates our solution for x using the λ value we calculated with the UPRE method in comparison to the ground truth x
4. The fourth graph illustrates our solution for x using the λ value we calculated using the UPRE method, as well as the calculations for x when we set $\lambda = 0.1, 3$ in relation t0 the ground truth x. I used chose to illustrate the solutions with these values for λ so the differences between each solution were abundantly evident.


Problem setup
Ground truth x
Observed data y
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
25
50
75
100
125
150
175
200


least squares solution
6
4
2
0
−2
−4
−6
0
25
50
75
100
125
150
175
200





Graphical representation takeaways

Just looking at all these graphs visually, we can compare the UPRE solution and the solutions for different λ values to conclude that that the minimizing λ is very near our UPRE solution

λ (actually showing it is better than every other single lambda requires much more in-depth numerical analysis). The first takeaway is that λ should not be zero, meaning regularization is needed. In graph #2, we see just how terrible a solution this leads us to have, as our solution goes up and down constantly here. In graph #4, we compare the UPRE λ solution to when $\lambda = .1$ and when $\lambda = 3$, i.e. when λ is a bit higher or a bit lower than we think is the best choice for λ. When $\lambda = 0.1$, we see our solution is better as this choice for λ sways much further from the ground truth in terms of how far above and below its path travels on. We can explain this by asserting that enough regularization is present when $\lambda = 0.1$. Meanwhile, when $\lambda = 3$, we see that our solution is much smoother than when $\lambda = 0.5462277217684343$ but it also is significantly farther from the ground data during some intervals in comparison the UPRE solution. This tells us that we applied too much regularization and should make λ lower than 3.