Estimation of Heating Parameters using Likelihood Maximization

Daniel Snyder, Yale University
Optimization Techniques, Fall 2017
Professor Sekhar Tatikonda

December 14, 2017

Contents

0.1 Abstract

The driving question for this project is: given an observed time series of potentials that vary according to a decaying exponential, $T(t) = a + be^{-ct}$, can the parameters (a, b, c) be estimated accurately? How long of a time series and how many samples would be required to obtain smallest standard error in estimating a particular parameter? Are these estimates unbiased? To answer these questions, we carry out a maximization of the likelihood of the parameters given an observed time series under some assumptions about noise. Newton's method is used for this optimization since the likelihood function is known and all its derivatives are calculable. This paper will go into the steps taken to make Newton's method converge even though the objective may violate the convexity requirement. The paper concludes with a review of how optimization could be used to continue this investigation.

0.2 Problem Formulation

Ill-posed problem This problem can be seen as an inverse heat conduction problem (IHCP) if the potential is thought of as temperature. Given the temperatures at certain times, can we estimate the rate of heat input at each time? This problem belongs to a class of ill-posed problems. These problems are called ill-posed because they are inherently unstable. The solutions can be extremely sensitive to measurement error making the estimates unreliable [1]. This notion of problems being well-posed was first mentioned by Hadamard in 1923. A problem is referred to as well-posed if certain conditions are met regarding the solution's existence, uniqueness, and stability [2].

In this paper we restrict ourselves to investigate cases where our parametric model of the time dependence of potential is correct. This will be a parameter estimation. We will also carry out sensitivity analysis of the estimates using Monte Carlo simulation to relate sample parameters to estimate standard error.

Time dependence model A noise corrupted version of the time series is observed. The true, uncorrupted time dependence of the potential is assumed to obey a first order differential equation.

$$c\frac{dT}{dt} = a - f(t) \tag{1}$$

In words, the rate of change of potential must be proportional to potential difference between the object and its surroundings. This proportionality constant is parameter c of the model. This is a separable, linear differential equation and can be solved.

$$
\begin{aligned}
\frac{dT}{a - f(t)} &= c \cdot dt \\
-\log(a - f(t)) &= ct + b \\
f(t) &= a - be^{-ct}, c \geq 0
\end{aligned}
\tag{2}
$$

Probability Model We will take measurements of the object's potential, T_i, at specified times, t_i, that we can choose. We will assume that the observed time series is a Gaussian noise corrupted version of a time series that obeys the governing differential equation. The true parameters (a, b, c) that generated the time series are unknown and will be estimated. The noise is assumed to be i.i.d. Gaussian.

$$T_i = a - be^{-ct_i} + \epsilon_i \tag{3}$$

Where we have:

$$\mathbb{E}(\epsilon_1, ..., \epsilon_n) = 0 \tag{4}$$

And

$$Var(\epsilon_1, ..., \epsilon_n) = \sigma^2 I \tag{5}$$

Since the noise term, ϵ_i, are independent random variables, the likelihood function will be written as the product of marginal likelihoods. This likelihood will be a function of three variables: a, b, c.

0.3 Notation

Use the following notation
Gaussian density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \tag{6}$$

Let T be the observed potentials

$$T = (T_1, ..., T_n) \tag{7}$$

Let t be the time values for observations

$$t = (t_1, ..., t_n) \tag{8}$$

Let Θ be the parameter vector (a, b, c)

$$\Theta = (a, b, c) \tag{9}$$

Let $f(t|\Theta)$ be the time dependent function giving the potential at time t using parameters (a, b, c).

$$f(t|\Theta) = a - be^{-ct} \tag{10}$$

Difference between the observed potential vector and the theoretical potential vector using parameters (a,b,c)

$$g_i = f(t_i|\Theta) - T_i \tag{11}$$

Define maximization objective as the joint likelihood function of parameters (a,b,c) given time series (T,t). This assumes i.i.d. Gaussian noise.

$$L(\Theta|T,t) = (2\pi\sigma^2)^{\frac{-n}{2}} \prod_{i=1}^{n} \exp\left[\frac{g_i^2}{2\sigma^2}\right] \tag{12}$$

Take minus natural log since this function has the same minimizer as the objective.

$$-\log L(\Theta|T,t) = \frac{n}{2}\log[2\pi\sigma^2] + \sum_{i=1}^{n} \frac{g_i^2}{2\sigma^2} \tag{13}$$

Taking minus log of the objective makes differentiation easier and converts the problem to a minimization. The negative log likelihood function is the objective to be minimized. The solution to this minimization problem will be the maximally likely parameters given the observed time series. The optimal value will be the estimate of the parameters.

0.4 Objective Statement

We would like to know the parameters $\Theta = (a,b,c)$ that minimize the negative log likelihood function given the observed data T, t. The minimizer of this function is also the minimizer of the sum of squared errors at each time step.

$$\begin{aligned} \Theta^* = \underset{\Theta}{\operatorname{argmin}} \frac{n}{2}\log[2\pi\sigma^2] + \sum_{i=1}^{n} \frac{g_i^2}{2\sigma^2} \\ c \geq 0 \end{aligned} \tag{14}$$

Constraints The only constraint is the rate constant c must be positive so the exponential term decreases in magnitude as time increases. In other words, the potential of the object must always get closer to the equilibrium potential as time passes. Later in the paper, usage of a barrier for c is discussed, but it proved to be unnecessary.

The constant b is interpretable as the initial potential difference between object and surroundings at time zero. A good initial estimate of this parameter should be the current estimate of a minus the first potential reading.

0.5 Methods

Newton's Method Newton's method can be used to minimize convex functions. In Newton's method, we approximate the function at each iteration as a quadratic function - and then minimize that function to find the next iterate. We write the objective as a Taylor series and take only the first and second order terms. As we get closer to the optimal point, the quadratic approximation gets better and better.

0.5.1 Theory

Consider the general quadratic function $x \in \mathbb{R}^n$:

$$f(x) = x^\intercal P x + q^\intercal x + r \tag{15}$$

The minimizer of this quadratic function is found in closed form as:

$$x^+ = x - \frac{\nabla f(x)}{[\nabla^2 f(x)]^{-1}} = x - \frac{Px - q}{P^{-1}} \tag{16}$$

In Newton's method, we perform a fixed point iteration on x, taking the next x to be the minimizer of the quadratic approximation of f at x.
So to perform this fixed point iteration, we'll need the first and second derivatives of the objective function. Here are the first and second derivatives of the negative log likelihood function described above. Again, the g_i represent the error at each time in the predicted time series using parameters (a, b, c).

$$\frac{\partial \ell}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} g_i \frac{\partial g_i}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} g_i \tag{17}$$

$$\frac{\partial \ell}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^{n} g_i \frac{\partial g_i}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^{n} g_i e^{-ct_i} \tag{18}$$

$$\frac{\partial \ell}{\partial c} = \frac{1}{\sigma^2} \sum_{i=1}^{n} g_i \frac{\partial g_i}{\partial c} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (-bt_i g_i) e^{-ct_i} \tag{19}$$

$$\nabla \ell \equiv \left(\frac{\partial \ell}{\partial a} \frac{\partial \ell}{\partial b} \frac{\partial \ell}{\partial c} \right)^\intercal \tag{20}$$

And the second partial derivatives.

$$\frac{\partial^2 \ell}{\partial a^2} = \frac{n}{\sigma^2} \tag{21}$$

$$\frac{\partial^2 \ell}{\partial b^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} e^{-2ct_i} \tag{22}$$

$$\frac{\partial^2 \ell}{\partial c^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} g_i t_i^2 b e^{-ct_i} - t_i b e^{-ct_i} \tag{23}$$

$$\frac{\partial^2 \ell}{\partial a \partial b} = \frac{1}{\sigma^2}\sum_{i=1}^{n} e^{-ct_i} \tag{24}$$

$$\frac{\partial^2 \ell}{\partial a \partial c} = \frac{1}{\sigma^2}\sum_{i=1}^{n} -t_i b e^{-ct_i} \tag{25}$$

$$\frac{\partial^2 \ell}{\partial b \partial c} = \frac{1}{\sigma^2}\sum_{i=1}^{n} -t_i b e^{-2ct_i} - g_i t_i e^{-ct_i} \tag{26}$$

0.5.2 Convexity

For theoretical results about the convergence of Newton's method to hold, the objective function must be convex. If this condition is not met, the algorithm could converge to a local optimum. The nature of the objective function does depend on a random noise vector and so any statements about the convexity, smoothness, or strongness of this objective may need to be made with probability. There could be some observed patterns of noise that cause the function to become more or less strongly convex.

0.5.3 Sensitivity to initial point

Not all initial points yield convergence Using simulated data with known parameters (a, b, c, σ^2) algorithm converged, in some cases. Poor choice of initial point caused the algorithm to fail to find the optimal point. With certain choices of initial point, the procedure diverges. This may be due to objective function being not strictly convex. There may be some regions of the function that are very flat (very small gradient) and can lead the algorithm to converge to sub-optimal points.

Initial point selection In practice, the initial point must be chosen without knowledge of the true parameters. Through exploration, I have found that making an educated guess at the parameters for the optimization starting point can improve the chances that the algorithm converges. Since the potentials are assumed to asymptotically approach the parameter a, a fair initial guess at the parameter a is the final potential measurement, T_n. Similarly, a fair initial guess at the parameter b is the difference $T_1 - T_n$. To make an initial guess at the

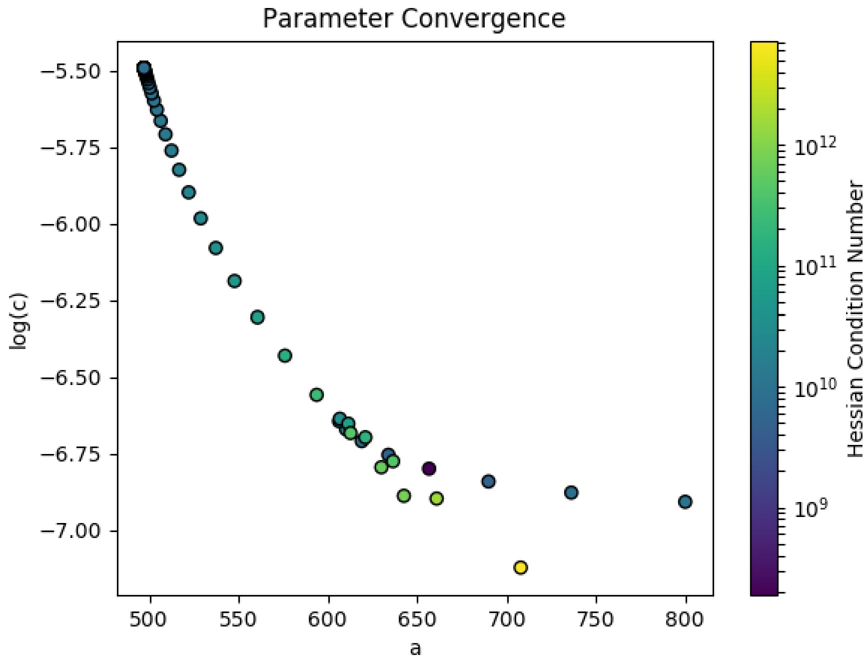

Figure 1: Parameter convergence

rate constant c, we assume that our initial guesses of a and b are correct and compute the approximate half-life of the potential. The initial guess at rate constant would then be:

$$c_{init} = \frac{log(2)}{t_{half}} \tag{27}$$

0.5.4 Converged Case

The following parameters resulted in convergence of the algorithm. Note that the optimal point need not take the same value as the true parameters that underlaid the simulation. The optimal value depends on the noise corrupted time series that was observed.

-	initial	optimal	true
a	800	497.05	500
b	-200	-399.43	-400
c	.001	0.004119	.004

Figure 1 shows the iterates in in the $a - log(c)$ plane. The iterates start on the right, move to the left, make a reversal, then ultimately converge to the optimal solution. One issue with Newton's method is that the Hessian matrix

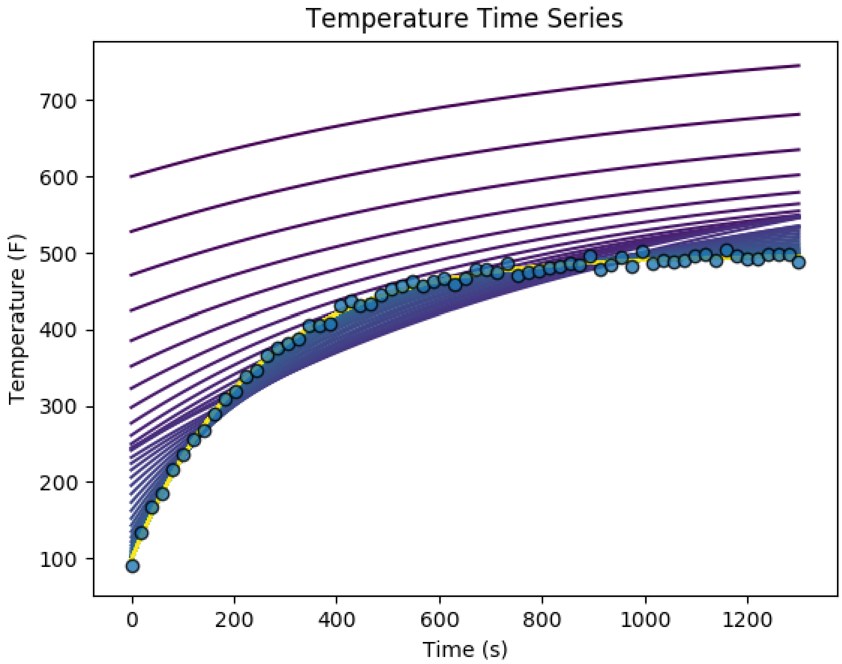

Figure 2: Time series model convergence

(second partial derivative matrix) must be invertible. Potential numerical issues could arise if the Hessian matrix becomes poorly conditioned (ratio of largest to smallest eigenvalue is large). To quantify this, the condition number of the Hessian matrix is mapped to colors of the points (on a log scale).

Time series interpretation The iterations can also be mapped to the time series space. Here is the theoretical time series plotted against the observed time series. Convergence occurs when the model time series is closest to the observed time series in a least-squares sense. Figure 2 shows that the model starts out from a point where there is large error and converges to a point of smallest squared error.

0.5.5 Diverged Case

In other cases, using different initial points caused the parameters to diverge and the optimal point was not found. Figure 3 shows the iteration end but the optimal point was not found. Figure 4 shows the model time series diverging to a poor prediction of the actual observed time series. In this case, the parameter c became very large and was never recovered.

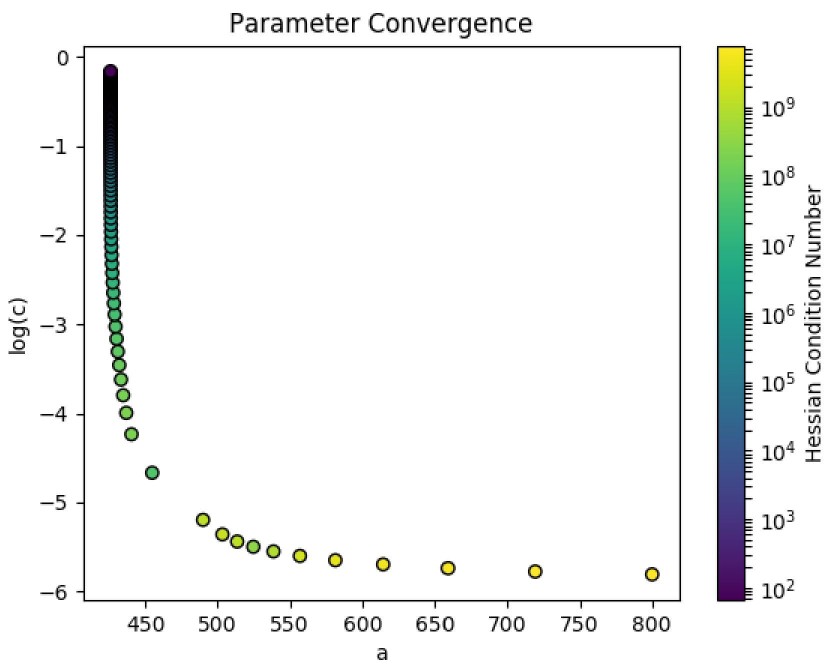

Figure 3: Parameter divergence

-	initial	optimal	true
a	800	426.00364	500
b	-200	-336.50224	-400
c	.003	0.8492	.004

Changing the initial choice for the parameter c caused the algorithm to diverge. This could be due to the function not being convex (or lacking strongness or smoothness). The exit criteria for the optimization was a threshold on the norm of the gradient. This means that the gradient got small enough in magnitude that the exit criteria was satisfied, yet the likelihood was not maximized. This is likely due to numerical issues where the function is too flat and finding a descent direction is not stable.

0.6 Optimization of sampling parameters

To answer the question of choosing the parameters of the time series acquisition, we must quantify the effect of these parameters on the metric of error of estimating the parameters. Random simulation is performed to relate sampling and noise parameters to estimate uncertainty. Here we use Monte Carlo simu-

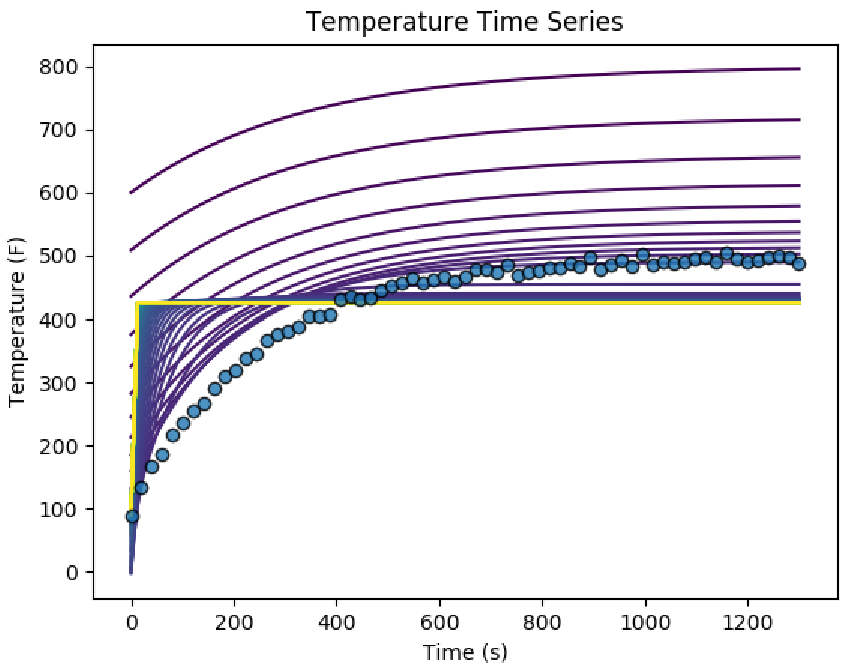

Figure 4: Time series model divergence

lation to estimate the bias and standard error of estimation of the parameters using maximum likelihood. This estimation procedure is repeated for randomly generated time series data with fixed parameters. The bias and standard error can be estimated from the errors observed in the Monte Carlo. Let $\hat{A}_i$ be the i^{th} estimate of A from the i^{th} out of N randomization of the observed time series.

$$\begin{aligned} Bias(\hat{A}) &= \text{median}(\hat{A}_i) \\ SE_{\hat{A}}^2 &= \frac{1}{N}\sum_{i=1}^{n}(\hat{A}_i - A)^2 \end{aligned} \tag{28}$$

0.6.1 Bias of Estimation

The bias in the estimate of each parameter is calculated as the difference between the median estimate and the true parameter value. The median was used instead of the average since in each Monte Carlo analysis, some of the estimates diverged and produced incorrect optimal points. Each median calculation operated with 100 random estimates. The randomness of the estimates was achieved by randomizing the noise applied to the uncorrupted time series. Figure 5 shows that there is positive bias in the estimates of A for short total measurement time and higher measurement count. Conversely, there is negative bias in the estimates for lower sample counts and lower total measurement time. Overall, it appears that most estimates were unbiased when the sample count was around 60.

0.6.2 Standard Error of Estimation

The standard error proved somewhat difficult to calculate due to the fact that not all cases in each Monte Carlo simulation converged. Careful selection of parameters can produces cases where a large enough proportion of cases converge and reasonable estimates of standard error can be produced. Figure 8 through Figure 10 show the dependence of standard error of estimation on the sampling parameters. For the asymptotic value parameter a, smallest standard error of estimation was fond when using short total time of observation and highest sample count. This trend was similar for the other two parameters. Minimum standard error was achieved with shorter, higher frequency sampling. On the other hand, this regime of sampling was shown to have higher bias.

0.7 Future Work

In this paper, we have covered a methodology of randomly generating observed time series, producing estimates of parameters, and computing bias and standard error for those estimates. A practical question is: how do I choose between low bias and low standard error? What is the optimal sampling regime for a fixed set of heating parameters? What is the optimal sampling regime for an

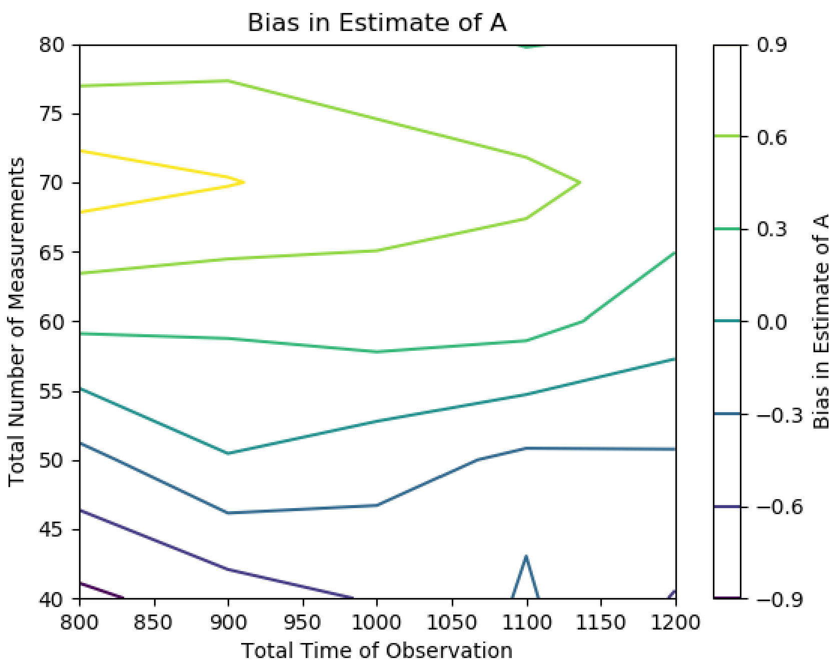

Figure 5: Bias in the estimate of A

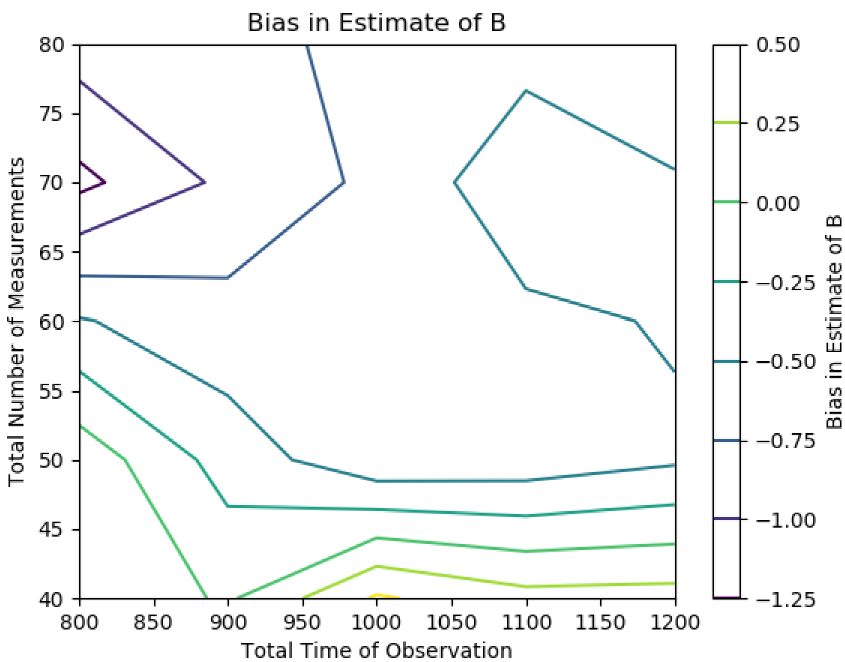

Figure 6: Bias in the estimate of B

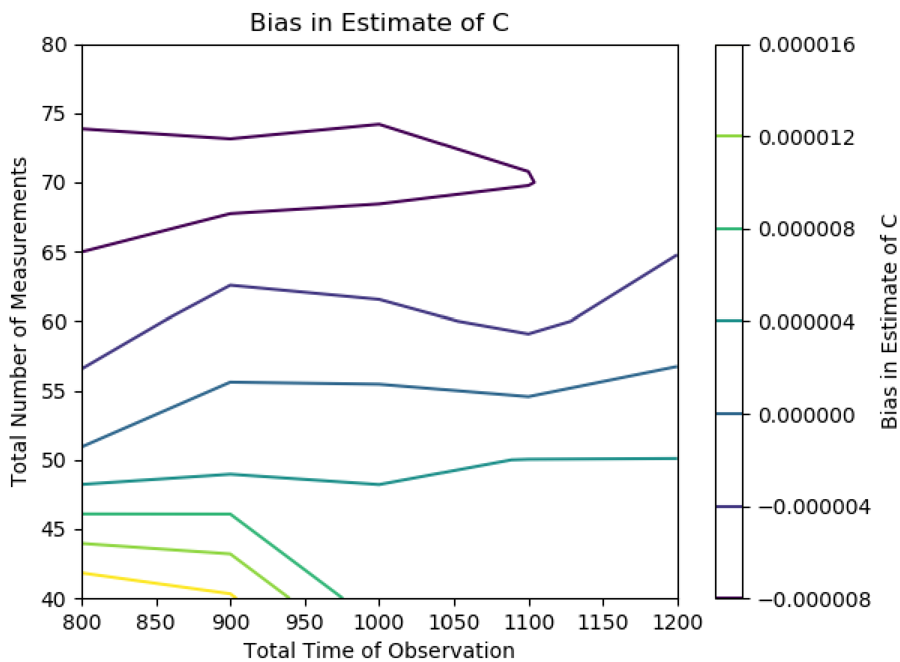

Figure 7: Bias in the estimate of C

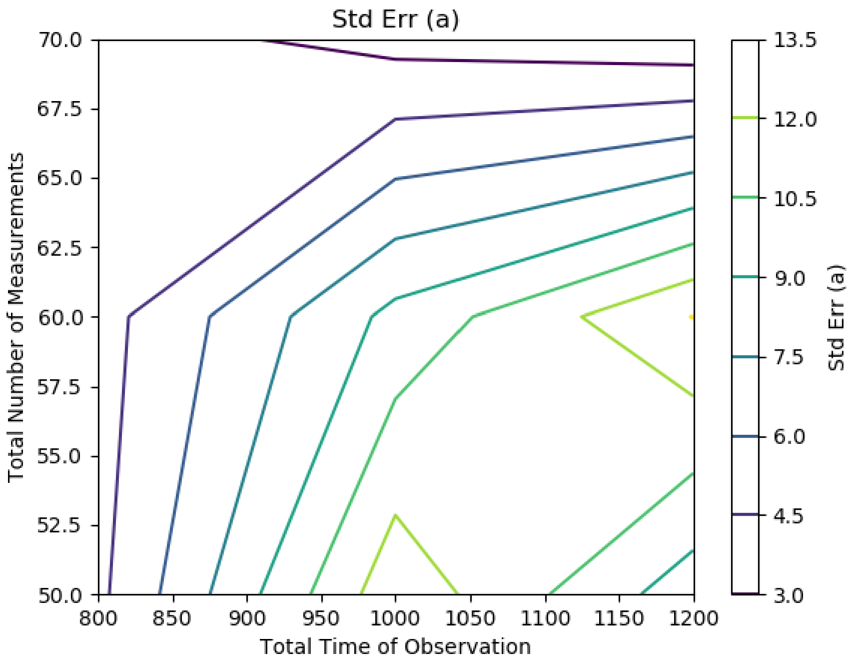

Figure 8: Standard Error in the Estimating A

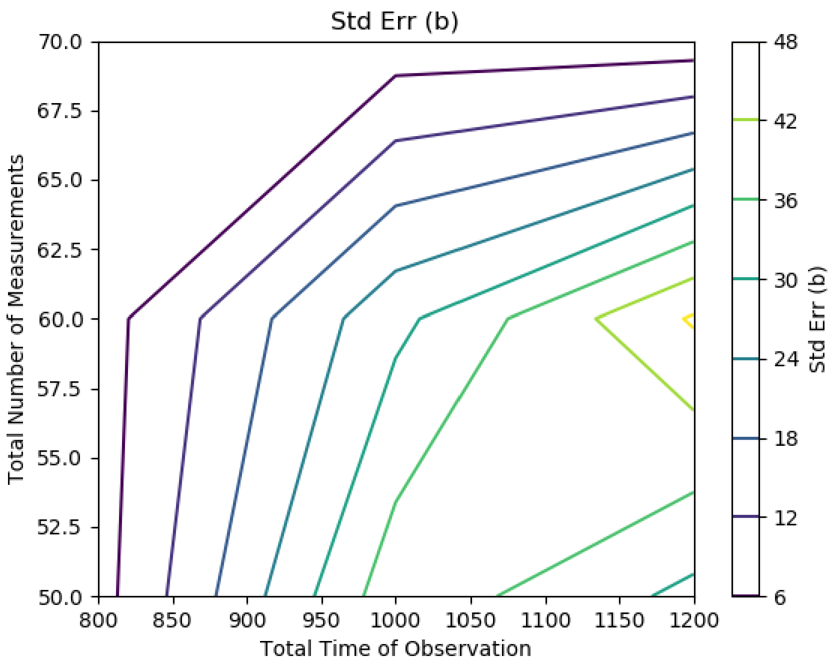

Figure 9: Standard Error in the Estimating B

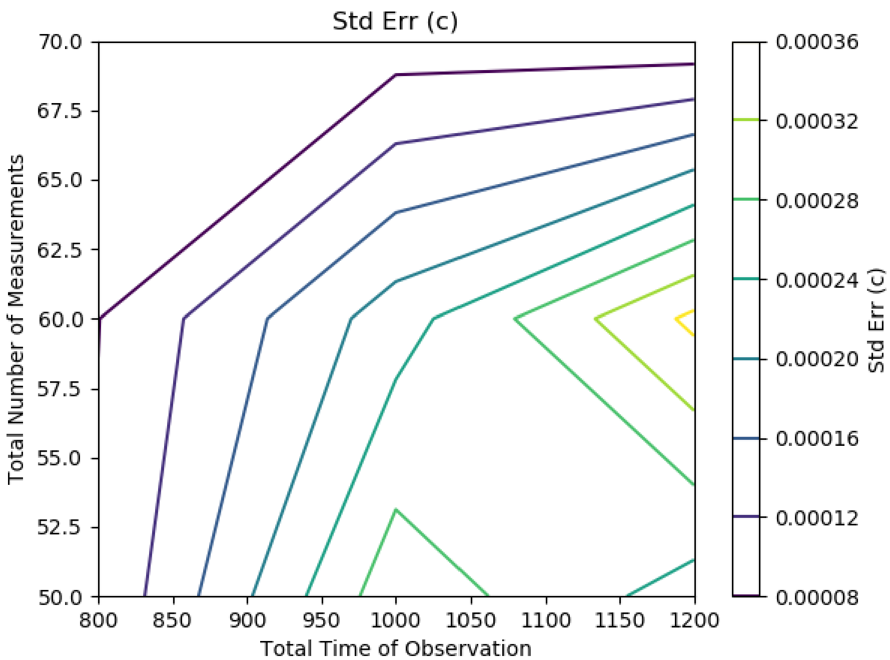

Figure 10: Standard Error in the Estimating C

unknown set of heating parameters? Another objective function could be constructed to quantify the negative utility of a given sampling regime as a function of bias and standard error of estimation. This negative utility minimization is a two parameter optimization would yield a choice of optimal sampling parameters for the task of estimating heating parameters. Since this function I have described is not known and neither are its derivatives, a first order method of optimization would need to be employed. For this task, we could choose gradient descent with backtracking line search since it only requires being able to evaluate the function's value and does not require knowledge of the function's derivative at a point.

0.8 Code

Python was chosen as the language for this project because of its wide array of libraries and availability of excellent debugging tools. The object oriented style of Python also helped with quick prototyping of optimization constructs.

0.8.1 Challenges

Initially, there were problems getting the algorithm to converge. Considering that the parameter c must be positive, I tried applying a logarithmic barrier function to the objective to keep c positive. This has marginal impact on the convergence of the procedure. Using logic to obtain a decent estimate for the initial point had a larger impact on the probability of convergence.

0.8.2 GitHub Page

The calculations and plots were generated in Python 3.5 using NumPy and Matplotlib 2.1. The code I have written to perform the calculations and produce plots for this project is available on GitHub.
GitHub Page Link

Bibliography

[1] Beck, James V. Ben Blackwell, Charles R. St Clair Jr.
Inverse Heat Conduction Problems Ill-posed problems.
Wiley Interscience Publication. (1985)
ISBN 0-471-08319-4

[2] Stolz, G., Jr. Numerical Solutions to an Inverse Problem of Heat Conduction for Simple Shapes *J. Heat Transfer* **82**, 20-26 (1960)