

M. S. Z. Tienstra
msz.tienstra@gmail.com
Master Thesis

Regularization of Ill-posed Statistical Inverse Problems

A data driven method for choosing the regularization
parameters in Tikhonov regularization.

Thesis Supervisors: Dr. Tristan van Leeuwen,
Dr. Evgeny Verbitsky



Mathematisch Instituut, Universiteit Leiden
Date: 28-02-2022

Summary

In this thesis we discuss a data driven way to chose the regularization parameter in Tikhonov regularization of ill-posed inverse problems. We then implement and compare three different kinds of iterative methods for solving the minimization problem. Common parameter choice methods are a-prior rules, discrepancy principle, L-curve method, and cross-validation, but we would like to somehow choose the regularization parameter based on the observed data. We show in this thesis that we can choose λ , the regularization parameter, from the data by posing the inverse problem in the the Bayesian framework. We gain insight into how the regularization parameter depends on the variance of model. Namely that λ consists of two parts, 1) the variance from the observational noise, and 2) the variance of the underlying ground truth parameter. In the particular case of normal likelihood and additive normal noise we get a posterior distribution where solving for the MAP estimate is equivalent to solving the Tikhonov regularization problem with $\lambda = \beta/\alpha$. If we bound the variance parameters of the observational noise and x , the resulting minimization problem is well-posed. In the special case above we have closed form solutions x, α, β and see that the minimization problem is bi-convex so coordinate decent method seems a natural choice. We alternatively implement gradient based methods in the case that we don't have closed form solutions. Using the partial solutions we can solve the resulting minimization problem numerically. This extends to the case when we have non-normal priors on x .

Contents

1 Inverse Problems	3
1.1 Introduction	3
1.2 Basic Formulation	3
1.3 Goals	5
1.4 Outline	6
2 Regularization	6
2.1 Ill-posedness	7
2.2 Stabilization	10
2.3 Tikhonov Regularization Revisited	11
3 Statistical and Probability	13
3.1 Probability Theory	14
3.2 Some statistics definitions	15
4 Statistical Inverse Problems	17
4.1 Bayes Formula	17
4.2 Connection to Tikhonov Regularization	20
5 Bayesian Regularization	21
5.1 Empirical Bayesian Method	21
5.2 Well-posedness	21
5.3 Hierarchical Bayesian Method	22
5.4 Well-posedness	23
6 Numerical Methods	24
6.1 Method 1	24
6.2 Method 2	26
6.3 Method 3	28
7 Implementation	30
7.1 Example	30
7.2 Ill-posedness	31
7.3 Regularization	31
7.4 Convergence and Consistency	39
7.5 Sensitivity	39
8 Conclusions	40
9 Appendix	41
9.1 Sensitivity Plots	41

1 Inverse Problems

In this section we give a brief introduction to functional analytic inverse problems, and why they are usefully and interesting to study. We then give a brief overview of the research done and summarize the goals of the thesis. We then discuss the outline.

1.1 Introduction

What are inverse problems? To understand what is inverse about inverse problem, we must first define the direct problem. The direct problem simulates the effects from known causal parameters. However, in inverse problems we observe only the effects and want to infer the causes. For example the direct problem can be a mathematical model such as those found in physics, where the direct problem comes from natural laws written as systems of differential equations. One such example is seismic inversion where we wish to infer the underlying causes from observed measurements on some subsurface. Other cause and effect examples are in image reconstruction, magnetic resonance imaging, tomography, and heat diffusion. Inverse problems also arise in non-physical situations such as in root finding, matrix inversion, and differentiation. Below we give an example where the inverse problems is to recover an image.

Example 1.1 (Image Processing). *Suppose we observe an 2-D digital image. We can obtain this by convolving the original image with some filter and adding noise. The resulting observation is a distorted and noisy version of the original image. We can mathematically model this by the discrete model*

$$y_i = \sum_{j \in \mathbb{Z}^2} a_{i,j} x_j$$

where y_i is the measured image, x_j are the pixel values our image. $i = (i_1, i_2)$ and $j = (j_1, j_2)$ and $a : \mathbb{Z}^2 \times \mathbb{Z}^2 \rightarrow \mathbb{R}$ the filter. The direct model is

$$a(t) = \int_{\Omega} a(t - y) x(y) dy$$

where $a(t)$ is the image represented as function, $\Omega \subset \mathbb{R}^2$ is the image domain, and $k : \Omega \times \Omega \rightarrow \mathbb{R}$ is the convolution kernel. The inverse problem is then deconvolution.

1.2 Basic Formulation

In inverse problems, the goal is to recover the unknown parameter x from observation y . We can write this as

$$y = Ax \tag{1}$$

where We want to solve the following equation for x given $y \in \mathcal{Y}$ is the observed/measured data, $x \in \mathcal{X}$ is the unknown parameter, and $A : \mathcal{X} \rightarrow \mathcal{Y}$ is

the forward linear operator that relates how x relates to y . We assume there exists some \bar{x} such (1) holds, and that the forward problem linking the x to y is well-defined. We typically observe only noisy measurements of the x , and express this as

$$y = A(x, e) \quad (2)$$

It is assumed throughout this paper that e is some additive noise so we can write

$$y = Ax + e \quad (3)$$

Examples of inverse problems abound in many areas in science and mathematics. In mathematical modeling we know the parameters of the model, and their relation and therefore can predict the outcome. In inverse problems we are interested in the opposite where we see only the outcome but not the parameters that cause these observations or how they are related. Below we return to the example of convolution as an inverse problem that arises in imaging and signal processing. This demonstrates that finding a solution to (3) is not trivial.

Example 1.2. (*DeConvolution [1]*)

Let $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{R})$ be the space of square integrable functions. Let $A : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ such that

$$(Af)(x) = g \circ f = \int_{\mathbb{R}} g(x-y) f(y) dy$$

Let $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The Fourier transform of (Af) is

$$\mathcal{F}(Af)(\xi) = \int_{\mathbb{R}} e^{-i\xi x} Af(x) dx = \hat{g}\hat{f}(\xi)$$

If $Af = 0 \implies \hat{f} = 0 \implies f = 0$ so A is injective. So the solution is unique and exists. The solution to $Af = h$ is given by

$$f(x) = \mathcal{F}^{-1}(\hat{g}^{-1}\hat{h})(x)$$

The solution is not well defined for an arbitrary $h \in L^2(\mathbb{R})$. Suppose we observe small errors in h . As \hat{g}^{-1} grows exponentially, h may no longer be in the range of A . So the integral does not converge, and no solution exists.

Inverse problems can be categorized by the forward operator. They are either linear or non-linear. Most inverse problems arising from physical systems are non-linear, but in this thesis we will study the simpler case of discrete finite linear inverse problems. A variety of classical examples of inverse problems can be found in [8] and [6]. Inverse problems are often inherently ill-posed. Regularization of inverse problems is a major area of research that tries to reconstruct a good estimate of the causal parameters given the data. Regularization methods have been studied in [6], [3], and more recently [9]. We can also use Tikhonov

regularization of inverse problems, with ℓ_2 regularization as analogous to ridge regression in linear regression. [14]. Tikhonov regularization gives us the best point estimator given some regularization parameter. The choice of regularization parameter is somewhat arbitrary or assumes we have access to unknown information such as the noise level. We can reformulate inverse problems using Bayesian statistics. The major benefit of this, is that we can mathematical incorporate the uncertain of the model parameters and by considering them as random variables defined by (conditional) distributions. An overview of statistical inverse problems can be found in [2], which introduces the finite setting, and [11] which is focused on the infinite setting. The regularized solution is a good estimate conditional on the regularization parameter, that balances interpolating the data points versus some other desired property such as smoothness. This balance is the bias-variance tradeoff. The Bayesian solution is the posterior (conditional) distribution computed from a likelihood and a prior using Bayes formula. These two methods seems to be at opposed ends. However, if we suppose normal noise, normal priors and normal likelihood, resulting in a normal posterior distribution computing the MAP estimate is equivalent to computing the Tikhonov regularized solution with regularization parameter depending on the precision parameters of the prior and the noise. Coupling these two methods we can formally write down our assumptions about the model [11] and infer the regularization parameters from the data alone [7],[10]. In this nice setting computing the MAP estimate of the posterior distribution is equivalent to solving the minimization problem in the Tikhonov regularization where we minimize over solution as well as the regularization parameters. In [7] study this case in detail proving convergence and consistency of this estimator as well as proposing and implementing an alternating algorithm to numerical solve the minimization problem. This alternating algorithm relies on the closed form solution to gradient in all directions of the unknown parameters. We build on this by first examining and implementing their method, and then proposing and implementing numerical methods that can solve the minimisation problem when we do not have the closed form is not the case.

1.3 Goals

We have two main goals. They are as follows

- Present a data driven way to choose the regularization parameter in Tikhonov regularization. We will show that this is a well-posed problem. We will also show that resulting estimate for ground truth converges to the ground truth as the noise level goes to zero, as showing that the residuals are a consistent estimator for variance of the model.
- Derive numerical methods to solve the resulting minimization problem from regularization the inverse problem. The first method implemented was proposed by [7]. We extend their work by implement additional algorithms in to handle more a more general case. We will show that these

methods converge to a minimum. We then implement our methods¹ and test them on simple example. Through this example we will also explore the effect of changing the varying the hyper priors on the convergence of the methods.

1.4 Outline

The outline to the thesis is as follows. In Chapter 2 we explain one of the key questions in inverse problems. Is the formulated problem of solving for x well-posed. We define what ill-posedness is in a specific setting, and then discuss numerous situation in which we encounter ill-posedness. This then naturally leads us to the resolution of ill-posed problems where we explain how we can stabilize the solution through regularization. So far everything until then has been in the non statistical finite dimensional vector space setting, and we transition to the statistical setting in Chapters 3 and 4. Chapter 3 is a brief over view the statistical and probability notation and definitions used to define statistical inverse problems. Then chapter 4 briefly introduces statistical inverse problems, and overviews some examples, and how they are related to the functional analytical setting. Chapter 5 is really the first main purpose of this thesis where we explain how from the Bayesian setting of inverse problems we can have a data driven method to infer the regularization parameters from the observations. We prove that the purely empirical Bayesian method is ill-posed in certain cases, and that a hierarchical model resolves this. Chapters 6 and 7 contain the second major half of this thesis. In this chapter we design three different numerical algorithms to numerical solve the resulting minimization problem derived in chapter 5. We show that these converge to a critical point of regularization functional. Then in chapter 7 we implement the methods in python, and explain the results. We also look at the models sensitively to the choice of hyper priors and the convergence and consistency. In chapter 8, we conclude with a discussion of where the thesis could go next and how we can improve on the results seen in chapter 7.

2 Regularization

We have seen that solving for x is more than just inverting the forward operator. In fact in the above example was an ill-posed problem. The direct inverse problem is often ill-posed, so we need a method to solve for x such that resulting problem is well-posed. Even when a unique solution exists or when we can define a unique solution, the solution is not guaranteed to depend on the the data continuously. To resolve these problems will introduce regularization into the direct inverse problem. The resulting problem will be well-posed and the resulting solution will be regularized.

¹<https://github.com/Tienstra/BayesianRegularization>

2.1 Ill-posedness

Hadamard defined a well-posed problem as one that meets all of the following conditions [6]:

Definition 2.1. *A problem is well posed if the following three condition hold*

1. *Existence: There exists a solution*
2. *Uniqueness: The solution is unique.*
3. *Stability: The solution depends continuously on the observed data.*

Now let $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}^n$, and suppose that we wish to solve the following for x

$$y = Ax$$

where now $y \in \mathbb{R}^n$, $x \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times m}$.

Remark 2.1. *We then consider the following cases:*

1. *If A is a square matrix, and A has full rank, then A is invertible. We then have that $x = A^{-1}y$ is the solution to the above.*
2. *if $n > m$, and $\text{rank}(A) = m$, then the system of equations is undetermined and inconsistent. So no solution can exist.*
3. *if $n < m$, and $\text{rank}(A) = n$, then the system of equations is overdetermined, and a solution exist but may not be unique.*
4. *If A is not full rank then by the rank-nullity theorem, the dimension of the null space can be greater than 0. In this case the solution may not exist and/or may not be unique.*

So a unique solution exists if and only if $\ker(A) = 0$. Then we have that

$$x = A^{-1}y$$

To check if this solution is stable we recall the following two definitions. Since we are in the finite dimensional case, A can be represented by a singular system.

Definition 2.2. *Let $A \in \mathbb{R}^{n \times m}$, then the singular value decomposition of A is a factorization of A ,*

$$A = U\Sigma V^*$$

where $U \in \mathbb{R}^{n \times n}$ orthogonal matrix, $\Sigma \in \mathbb{R}^{n \times m}$ rectangular diagonal matrix with non-negative entries, and $V \in \mathbb{R}^{m \times m}$ orthogonal matrix. The diagonal entries of Σ , denoted by σ_i are the singular values of A , and are listed in descending order. That is $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$. We sometimes write $U = [u_1, \dots, u_n]$ and $V = [v_1, \dots, v_m]$ where u_i and v_i are orthogonal basis for \mathbb{R}^n and \mathbb{R}^m respectively. The singular system of A is then $(u_i, v_i, \sigma_i)_{1 \leq i \leq \min(n,m)}$.

and

Definition 2.3. Let A be an invertible matrix. Let σ_{\min} and σ_{\max} be the minimum and maximum eigenvalues of A respectively. Then the condition number of $\kappa(A)$ is

$$\begin{aligned}\kappa(A) &= \|A^{-1}\| \|A\| \\ &= \frac{\sigma_{\max}}{\sigma_{\min}}\end{aligned}$$

Example 2.1. Suppose we would like to solve the following equation for x ,

$$y = Ax$$

Let δy be the error in y . Assume that A is inevitable. Then the $A(x + \delta x) = y + \delta y$, so $(x + \delta x) = A^{-1}(y + \delta y) = A^{-1}y + A^{-1}\delta y$. The error in the solution is then $A^{-1}\delta y$. We can compute the ratio of the relative error in the solution compared to the relative error in y as

$$\frac{\|A^{-1}\delta y\|}{\|\delta y\|} \frac{\|y\|}{\|A^{-1}y\|}$$

Then we see have from the above definition that

$$\begin{aligned}\kappa(A) &= \|A^{-1}\| \|A\| \\ &= \frac{\sigma_{\max}}{\sigma_{\min}} \\ &= \max_{\delta y, y \neq 0} \left\{ \frac{\|A^{-1}\delta y\|}{\|\delta y\|} \frac{\|y\|}{\|A^{-1}y\|} \right\} \\ &= \max_{\delta y \neq 0} \left\{ \frac{\|A^{-1}\delta y\|}{\|\delta y\|} \right\} \max_{y \neq 0} \left\{ \frac{\|Ay\|}{\|y\|} \right\}\end{aligned}$$

where $\|y\|$ is the euclidean norm, $\|A\|$ is the induced matrix norm, and $\sigma_{\max}, \sigma_{\min}$ are the maximum and minimum singular values of A respectively. So then

$$\frac{\|x - x_\delta\|}{\|x\|} \leq \kappa(A) \frac{\|y - y_\delta\|}{\|y\|}.$$

Example 2.2 (Matrix Inversion [3]). Let $y \in \mathbb{C}^n, x \in \mathbb{C}^n, A \in \mathbb{C}^{n \times n}$. Assume that A is symmetric positive definite. From the above, we can write

$$A = \sum_{i=1}^n \sigma_i a_i a_i^T$$

where σ_i are the eigenvalues of A ordered such that $\sigma_1 \geq \sigma_2 \geq \dots > 0$, and eigenvectors $a_i \in \mathbb{R}^n$ where $a_i \perp a_j$ for $i \neq j$. Assume we observe y_δ where $y_\delta = Ax^\delta$. Then we have that

$$x - x^\delta = \sum_{i=1}^n \sigma_i^{-1} a_i a_i^T (y - y^\delta).$$

The error between x and the estimate x^δ is

$$\begin{aligned} \|x - x^\delta\|_2^2 &= \sum_{i=1}^n \sigma_i^{-2} \|a_i\|^2 |a_i^T(y - y^\delta)|^2 \\ &\leq \sigma_n^{-2} \|y - y^\delta\|_2^2 \\ &\leq \sigma_n^{-1} \|y - y^\delta\|_2^2 \\ &\leq \kappa(A) \delta \end{aligned}$$

Let $y = Ax$ and $y^\delta = Ax + e$, such that $\|y - y^\delta\|_2^2 \leq \delta \kappa$. Suppose that $\kappa(A) \ll \infty$, then the solution depends continuously on the data, as the relative error in x is bounded by the relative error in y times a small constant. On the other hand if $\kappa(A)$ is very large, in which case A is ill-conditioned, then the solution is does not depend continuously on the data as a small change in y results in a large change in x . Since $\kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$ we see that a large $\kappa(A)$ occurs if σ_{\min} is very small. If $\sigma_i \rightarrow 0$ very quickly then $\kappa(A) \rightarrow \infty$. So stability is determined by the decay of the singular values of A . To guarantee a stable solution, we need to bound the singular values of A away from zero.

Suppose now $n \geq m$ and $y \notin \text{Ran}(A)$. Suppose also that A has full rank. The SVD of A is

$$A = U_m \Sigma_m V_m^* \quad (4)$$

Then

$$Ax = U_m U_m^* y$$

since U_m is an orthogonal matrix $U_m U_m^*$ projects y onto the range of A . The solution is then given by

$$\hat{x} = V_m \Sigma_m^{-1} U_m^* y = A^\dagger y$$

where we define A^\dagger below.

Definition 2.4. Let $A \in \mathbb{R}^{n \times n}$ with $\text{rank}(A) = r \leq \min\{n, m\}$. Then using SVD of A the Moore-Penrose pseudo is defined as

$$A^\dagger = V_r \Sigma_r U_r^* \quad (5)$$

where U_r, V_r, Σ_r are defined in definition 2.2.

Claim 2.1. The least squares solution to $y = Ax$ is given by

$$x_{LS} := \min_x \|Ax - y\|_2^2 \equiv V_m \Sigma_m U_m^* y \equiv A^\dagger y \quad (6)$$

where $A = U_m \Sigma_m V_m^*$ is the singular decomposition of the matrix operator A , $U_m = (u_1, \dots, u_m), V_m = (v_1, \dots, v_m)$, are the left and right m singular vectors and Σ_m is the diagonal matrix with the first m singular values. A^\dagger is the Moore-Penrose pseudo inverse of A .

Suppose now that $n < m$, and A has full rank. The solution exists but is not unique. The solution we would like then is the minimum norm solution. In this case the solution is given by

$$x = x' + \sum_{i=1}^n \frac{\langle y, u_i \rangle}{\sigma_i}$$

where $x' \in \ker(A)$. Since $n < m$, the $\ker(A) = \text{span}(v_n + 1, \dots, v + m)$. So $x' = Vc$ with $V = [v_n + 1, \dots, v + m]$. so the minimum norm solution is when $x' = 0$, and the solution does not contribute to the $\ker(A)$. So the minimum norm solution is

$$\hat{x} = V_n \Sigma_n^{-1} U^* y = A^\dagger y$$

Is the least squares or minimum norm solution stable? Using the SVD of the pseudo inverse of A we get that

$$\hat{x} = \sum_{i=1}^r \frac{\langle x_i, y \rangle}{\sigma_i} v_i$$

where $r = \min(m, n)$. We see that the continuity of \hat{x} is depending on the singular values σ_i . If σ_i is small then $\langle u_i, y \rangle$ can be large amplifying the $v_i^t h$ component of y . So it is possible that $v_i^t h$ with small singular values exaggerate the noise of y . So the solution is not continuous. So when is the solution stable?

Definition 2.5. For $y = Ax$, y satisfies the Picard condition if the Fourier coefficients $\langle u_i, y \rangle$ as derived above decay faster than σ_i , the singular values defined above. That is

$$\sum_{i=1}^r \left| \frac{\langle u_i, y \rangle}{\sigma_i} \right|^2 < \infty$$

2.2 Stabilization

Recall that we can decompose the mean square error of an estimator into its bias and variance. For x_{LS} , the bias is zero but the variance can be so high that the solution is still ill-posed as we have seen. To lower the variance we can introduce a biased estimator. Ideally we would end up with a lower MSE over all. One way to avoid dividing by large singular values to regularize the σ_i 's by with some regularization function \mathcal{R}_α , with regularization parameter α . The regularization solution is then

Definition 2.6.

$$x_\alpha = V_k \mathcal{R}_\alpha(\sigma_k) U_k^* y \tag{7}$$

where \mathcal{R}_α is the regularizing function depending on regularization parameter α .

Possible regularization functions are thresholding functions, such as TVSD, or shifting functions such as Tikhonov regularization.

Definition 2.7. *The Tikhonov regularization solution is*

$$x_\alpha = \sum_{i=1}^r \frac{\sigma_i \langle x_i, y \rangle}{\sigma_i^2 + \alpha} v_i \quad (8)$$

where $\mathcal{R}_\alpha(\sigma) = \sigma / (\sigma^2 + \alpha)$, and α is the regularization parameter.

In the above we modify pseudo inverse by adding some weight to the singular values. The denominator is then bounded by α even if $\sigma \rightarrow 0$. When $\sigma_i \gg \alpha$ the ratio $\frac{\sigma_i \langle x_i, y \rangle}{\sigma_i^2 + \alpha}$ is relatively unchanged. In the case where $\sigma_i \ll \alpha$, the ratio is decreased, thus overall decreasing the variance \hat{x} . The consequence of this is that resulting estimator x_α will be a biased. We can compare the difference between the non-regularized solution \hat{x} and the regularized solutions x_α . This difference displays the bias-variance trade.

Definition 2.8. *Let $\hat{x} = A^\dagger y$ the non-regularized solution with A^\dagger pseudo inverse. Let $\hat{x}_\alpha = A_\alpha^\dagger y^\delta$ the regularized solution with A_α^\dagger regularized pseudo inverse.*

$$\|\hat{x} - x_\alpha\| \leq \|(A^\dagger - A_\alpha^\dagger)y\| + \|A^\dagger(y - y^\delta)\|$$

The bias is measured as $\|(A^\dagger - A_\alpha^\dagger)y\|$ and variance is measured as $\|A^\dagger(y - y^\delta)\|$.

When $\alpha \rightarrow 0$, $A^\dagger = A_\alpha^\dagger \implies \|(A^\dagger - A_\alpha^\dagger)\| = 0$. Now that we have a good estimator we would like know how good this estimator is. To do this we can compute the mean squared error as

Definition 2.9. *Let $x = Ay$ be the true parameter. Let $x_\alpha = A_\alpha^\dagger y^\delta$ be the regularized solution. The measure of how close this estimator is to the truth is given by*

$$\|x - x_\alpha\| \leq \|x - A_\alpha^\dagger y\| + \|A_\alpha^\dagger(y - y^\delta)\|$$

2.3 Tikhonov Regularization Revisited

Above we defined everything in terms of SVD. But there is a variational formulation of Tikhonov regularization that turns solving for x into an optimization problem.

Definition 2.10. *Let $\lambda > 0$ be a fixed constant. The Tikhonov regularized solution x_λ to (3) $x_\lambda \in \mathcal{X}$ is the minimum of the functional*

$$\mathcal{R}_\lambda(x) = \|Ax - y\|^2 + \lambda\|x\|^2 \quad (9)$$

assuming that such a minimizer exists. $\mathcal{R}_\lambda(x) : \mathcal{X} \rightarrow \mathcal{Y}$ and λ is called the regularization parameter.

To find the estimate for \bar{x} is now an optimization problem, where we want to minimize $\mathcal{R}_\lambda(x)$ for some fixed λ . We get the following scheme

1. Minimize: $\min_{x \in \mathbb{R}^m} (\|Ax - y\|_2^2 + \lambda \|Lx\|)$. This can also be written as some constrained optimization problem.
2. The solution is $x_e^\lambda = (A^*A + \lambda L^*L)^{-1} A^*y$. Note that if there is no noise in the model, so we need no regularization, then we get back the least-squares solution.

We will denote the functional $\min_{x \in \mathbb{R}^m} (\|Ax - y\|_2^2 + \lambda \|Lx\|)$ by $\mathcal{J}(x)$, which consist of two portions, the data fidelity term $\|Ax - y\|_2^2$, and the regularization term $\|x\|_2^2$. We can check that the problem of minimizing $\mathcal{R}_\lambda(x)$ for a given λ is well-posed problem, by checking that

1. For fixed λ , $\mathcal{R}_\lambda(x)$ is well defined
2. For fixed λ , $\mathcal{R}_\lambda(x)$ is continuous in \mathcal{Y} .
3. We can select λ such that if $y \rightarrow A(\bar{x})$, then $\mathcal{R}_\lambda(x) \rightarrow \bar{x}$.

In this setting, we can check well-posedness by looking at the SVD of the regularized solution. We can find the solution to the minimization problem by writing down the normal equation. We get that

$$x_\alpha = (A^*A + \alpha L)^{-1} A^*y = V(\Sigma_r^2 + \alpha L)^{-1} \Sigma_U^* y \quad (10)$$

If the regularization guarantees stability, and $\ker A \cap \ker L = \{0\}$, then (12) is a well-posed problem. The estimate x_e^λ , depends on fixed λ , so we must need some method to choose λ such that the solution to the optimization problem is continuous (condition 3 in the above). Common methods to choose λ are via

1. a-prior rules knowing the noise level
2. Discrepancy principle
3. L-curve
4. Cross validation ²

Definition 2.11. Assume we know the noise level, and denote the noise level by e . We can then a-prior choose $\alpha(e)$, the regularization parameter now depending on e . This is called an a-prior rule. This is called convergent if and only if

$$\begin{aligned} \lim_{e \rightarrow 0} \alpha(e) &= 0 \\ \lim_{e \rightarrow 0} e \|A_{\alpha(e)}^\dagger\| &= 0 \end{aligned}$$

²An example of choosing the regularization parameter in ridge regression can be found in [14]

Claim 2.2. *If the $\alpha(e)$ as defined above is convergent then the total error $\|A^\dagger y_e A^\dagger y\|_2^2 \rightarrow 0$ as $e \rightarrow 0$. So we have consistency.*

Definition 2.12. *The discrepancy principle chooses α a-posterior depending on both y_e and e , such that*

$$\|AA^\dagger y_e - y_e\|_2^2 \leq \eta e$$

for $\eta > 1$ fixed. If $y_e \in \ker(A^\dagger)$ then no such α can exist.

Definition 2.13. *The L-curve method chooses α heuristically via a minimization problem*

$$\min_{\alpha > 0} \|A_\alpha^\dagger\|_2^2 \|AA_\alpha^\dagger y_e - y_e\|_2^2$$

the optimal α should lie at the corner of the curve $\|A_\alpha^\dagger\|_2^2 \|AA_\alpha^\dagger y_e - y_e\|_2^2$. We do not have consistency with this choice of α .

So far we have seen a a-prior rule, a posterior rule, and a heuristic rule, but we now introduce a more data driven rule, that being choosing α via cross validation. Later on we will see that there is a connection between this rule, and the Bayesian method we describe in this thesis. We would like to somehow choose λ from the data alone. We also want that as $\sigma \rightarrow 0 \rightarrow 0$ so \hat{x} converges to x_{LS} should it exist. To do this we will use Bayesian statistics, and frame (3) in the Bayesian framework. With certain choices, we will see that the Bayesian interpretation of (3), is closely related to the Tikhonov regularization we described above.

Finally we remark that

$$\mathcal{R}_\lambda(x) = \|Ax - y\|^2 + \lambda\|x\|^2 \quad (11)$$

for $\lambda \in (0, \infty)$ can be written as [14]

$$\mathcal{R}_\lambda(x) = (1 - \lambda)\|Ax - y\|^2 + \lambda\|x\|^2 \quad (12)$$

for $\lambda \in (0, 1)$ where here we see that that regularization parameter can be seen as balancing the fit versus the smoothing. At the end points when $\lambda = 0$ we get interpolation of the data points, and when $\lambda = 1$ we get over-smoothing. In the rest of this thesis we explicitly define how to choose λ from the data, and why indeed when λ is minimized we have also balanced the fit versus the smoothing.

3 Statistical and Probability

Below we review some measure theoretic probability facts and definitions that are necessary for defining Bayes formula for inverse problems. In the second section we recap the definitions of certain distributions as characterized by their probability density functions. We also review some algebra rules for the multi-variate normally distributed vectors.

3.1 Probability Theory

Recall that a probability space consists of a sample space Ω , a σ -algebra \mathcal{F} , and a probability measure \mathbb{P} . In this thesis we consider only σ -finite measures, that is measures that are countable unions of finite measure measurable sets. Particularly we use the σ -finite Lebesgue measure on \mathbb{R}^n . We denote a measurable space as $(X, \mathcal{B}(X))$ where X is some (metric) space and $\mathcal{B}(X)$ is the borel σ -algebra. In the case that $X = \mathbb{R}$ we know that $\mathcal{B}(X)$ is generated by the open intervals $(a, b]$ for $a, b \in \mathbb{R}$. We can extend this to \mathbb{R}^n . Also recall that random variable, f , is a measurable map $f : \Omega \rightarrow X$ and induces a probability measure X .

Definition 3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X, \mathcal{B}(X))$ a measurable space then the measure μ induced by the random variable $f : \Omega \rightarrow X$ is defined as

$$\mu(A) = \mathbb{P}(f^{-1}(A)) = \{\omega \in \Omega \mid f(\omega) \in A\}, A \in \mathcal{B}(X)$$

where μ is the distribution f . We denote this as $f \sim \mu$.

Definition 3.2. Let μ and ν be measures on a measure space (x, Σ) . Then we have the following

1. If $\nu(A) = 0 \implies \mu(A) = 0$ for all $A \in \Sigma$, then μ is dominated by ν and we say that μ is absolutely continuous with respect to ν . We denote this as $\mu \ll \nu$
2. If $\mu \ll \nu$ and $\nu \ll \mu$ then μ and ν are equivalent.
3. Let A and $B \in \mathcal{B}(X)$ be disjoint sets such that $A \cup B = \mathcal{B}(X)$ and $\mu(A) = 0$ while $\nu(B) = 0$, then μ and ν are mutually singular. We denote this as $\mu \perp \nu$.

We will now also state the Radon–Nikodym theorem which we can use to relate random variables to their probability density functions, as well as proving that the conditional posterior distribution is a solution to an Bayesian inverse problem if Bayes rule holds.

Theorem 3.1. Let (X, Σ) be a measurable space. Then let μ and ν be σ -finite measures defined on this space. Suppose also that $\nu \ll \mu$. Then there exists a unique up to a μ -null set Σ measurable function $f : X \rightarrow [0, \infty)$ such that for all $A \subset X$,

$$\nu(A) = \int_A f d\mu$$

We call f the Radon–Nikodym derivative and denote it as $\frac{d\nu}{d\mu}$.

In the case that the measure space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ for some finite n , and $X \sim \nu$ and $\mu = leb(\cdot)$, then by the Radon–Nikodym theorem, $f \in L^1(\mathbb{R}^n)$ is the unique probability density function for $X \sim \nu$.

What follows is a few definitions to define conditional distributions. with these we can precisely write down the posterior distribution as a conditional distribution.

Definition 3.3. Let $\mathcal{G} \subset \mathcal{F}$ be a sub- σ -algebra. Let y be a \mathcal{G} measurable function. We call $y : \Omega \rightarrow X$ a conditional expectation of a random variable $f : \Omega \rightarrow X$ with respect to \mathcal{G} if

$$\int_{\mathcal{G}} f d\mathbb{P} = \int_{\mathcal{G}} y d\mathbb{P}$$

Definition 3.4. Let $\mathcal{G} \subset \mathcal{F}$ be a sub- σ -algebra. The condition probability of $B \in \mathcal{B}(X)$, given \mathcal{G} is

$$\mathbb{P}(B | \mathcal{G}) = \mathbb{E}(1_B | \mathcal{G})$$

Definition 3.5. Let $(\mu(\cdot, \omega))_{\omega \in \Omega}$ be a family of probability distributions on $(X, \mathcal{B}(X))$. Then $(\mu(\cdot, \omega))_{\omega \in \Omega}$ is a regular conditional distribution of f given $\mathcal{G} \subset \mathcal{F}$ if

$$\mu(B, \cdot) = \mathbb{E}(1_B(f) | \mathcal{G}) \text{ a.s.}$$

for all $B \in \mathcal{B}(X)$. If f is as defined about then such a a regular conditional distribution exists.

Remark 3.1. These theorems are use full when we let $\mathcal{G} = \sigma(y)$ be the sub- σ -algebra generated by the observations $y = AX + e$ in the Bayesian setting. If we let π_{post} denote the posterior distribution and π_{prior} denote the prior distribution on x , then we have that

$$\pi_{post}(B, y(\omega)) = \mathbb{E}(1_B(x) | \sigma(y))(\omega)$$

for $B \in \mathcal{B}(X)$. So then

$$\pi_{post}(B, y) = \pi_{prior}(B | y).$$

3.2 Some statistics definitions

Denote random variables by capital letters. For example X, Y, E . Denote realization of these random variables by the corresponding lower case letter for $Y = y$, y is one realization of Y .

Definition 3.6. Multivariate normal distribution A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ if its probability density function is give by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \quad (13)$$

for $\mu \in \mathbb{R}$ and $\sigma > 0$.

Now let $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ a non-negative symmetric matrix. Then we can define $X \sim \mathcal{N}(\mu, \Sigma)$ for X a random n -dimensional random vector.

Definition 3.7. A random vector $X \sim \mathcal{N}(\mu, \Sigma)$ if and only if its probability density function is given by

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x - \mu)^T \Sigma^T (x - \mu)} \quad (14)$$

for parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ symmetric positive definite matrix.

Proof. See Aard van der Vaart chapter 2 lemma 2.3 dictaat. [12] \square

Theorem 3.2. If $c \in \mathbb{R}^n$ and X is an n -dimensional random vector such that $X \sim \mathcal{N}(\mu, \Sigma)$. Then $c + X \sim \mathcal{N}(c + \mu, \sigma)$.

Theorem 3.3. If X is an n -dimensional random vector such that $X \sim \mathcal{N}(\mu, \Sigma)$, and $A \in \mathbb{R}^{m \times m}$ a fixed matrix with rank $m \leq n$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$ is an m dimensional normally distributed random vector with mean $A\mu$ and covariance $A\Sigma A^T$.

Remark 3.2. If $Z \sim \mathcal{N}(0, I)$ and $X \sim \mathcal{N}(\mu, \Sigma)$ then we can write X as and $X = \mu + \Sigma^{1/2}Z$.

Remark 3.3. If $X \sim \mathcal{N}(\mu, \Sigma)$ is an n -dimensional random vector such that each X_i is independent from X_j for $i \neq j$, then Σ is a diagonal matrix with the variance of each X_i on the diagonal.

Definition 3.8. Let $\alpha > 0$, the the Gamma function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$. A random variable X is Gamma distributed with parameters $\alpha, \beta > 0$ if the pdf is characterized by

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0.$$

We denote this as $X \sim \text{Gamma}(\alpha, \beta)$. [13]

We can extend this definition to a random matrix symmetric X of dimension $p \times p$. In which we have the Wishart distribution.

Definition 3.9. Let M be a $p \times p$ positive definite matrix. Then the multivariate Gamma function Γ_p for a given α is define as

$$\Gamma_p(M) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$$

here $|M|$ is the determinate of M and $\text{tr}(M)$ is the trace. For X a random variable then, X is Gamma distributed if the pdf is characterized by

$$f_X(x) = \frac{|x|^{(n-p-1)/2} e^{-\text{tr}(M^{-1}x)/2}}{2^{\frac{np}{2}} |M|^{n/2} \Gamma_p(\frac{n}{2})}$$

where $n > p - 1$ is the degrees of freedom.

4 Statistical Inverse Problems

We use a Bayesian interpretation of our model in order to choose the regularization parameter λ from the data. To do this we will consider y, x, e to be random variables and A as some fixed carefully chosen operator. The model (2) can then be written as

$$Y = A(X, E) \quad (15)$$

where X, Y, E are random vectors defined on the probability space $\Omega = \Omega_1 \times \Omega_2$ such that $X : \Omega_1 \rightarrow \mathbb{R}^m$, $Y : \Omega_2 \rightarrow \mathbb{R}^n$, and $E : \Omega \rightarrow \mathbb{R}^n$. We want to learn the relationship between X, Y, E that determine their conditional probability distributions. We can relate X after making observations of Y using Bayes formula. First note again some notation. As in the non-random case Y is the observed/measured data, with $Y = y$ the realization of Y , X is the unknown parameter again with $X = x$ the realization of X , and E is still the noise. We will now consider this noise to be additive and Y, X, E to be random vectors in $\mathbb{R}^n, \mathbb{R}^m$ and \mathbb{R}^n respectively. We then rewrite (15) as

$$Y = AX + E \quad (16)$$

The solution to the above problem is a conditional posterior distribution for X given $Y = y$. Two things are gained from posing the inverse problem in the Bayesian setting. 1) we can obtain point estimates by computing the most likely value for X which we will describe later, and compute the uncertainty of this estimate by calculating the spread of the posterior distribution. 2) If the distribution is Gaussian, then the MAP estimate connects the non-Bayesian Tikhonov regularization setting to this Bayesian setting where we can then have a better understanding of the regularization parameter and the model's variance. This insight allows us to choose λ in a data driven way.

4.1 Bayes Formula

One key aspect of Bayesian statistics is that we assume we have prior knowledge of certain parameters in our model. In this setting what we can observe are realizations of Y , and we assume that we have prior knowledge of X , mainly which values of X are occurring and at what frequency. To express this prior knowledge we place a prior distribution on X . We denote this by F_X with density π_X . We also assume that $E \sim F_E$ with density π_E and that E is independent of X . Note that we will see later that independence is key here to get the desired likelihood and resulting Tikhonov regularization. With these assumptions we can find the likelihood $Y | X$ for $X = x$.

Claim 4.1. *The likelihood $L(Y = y | X = x) = \pi_E(y - Ax)$.*

Proof. [8] Since we assume that $X \perp E$, the distribution of E conditioned on $X = x$ is unaffected. That is

$$\mu_E(B | x) = \mathbb{P}(E \in B) = \int_B \pi_E(e) de$$

where $B \in \mathcal{B}(\mathbb{R}^n)$. If we condition Y on $X = x$, then $Y = A(X) + E$ is distributed like E , with shift $A(x)$. \square

Lemma 4.1. $(X, Y) \in \mathbb{R}^m \times \mathbb{R}^n$ is a random variable with Lebesgue density $\pi(x, y) = \pi_E(y - Ax)\pi_X(x)$

We now formulate Bayes theorem which will tell us how X is depending on Y .

Theorem 4.1. Assume that the $m(y) = \int_{\mathbb{R}} \pi_E(y - Ax)\pi_X(x)dx > 0$. This is called the normalizing constant. Then $Y = y | X = x$ is a random variable with Lebesgue density $\pi(x, y) \stackrel{(rem3.1)}{=} \pi_X(x | y) = \frac{1}{m(y)}\pi_E(y - Ax)\pi_X(x)$.

Definition 4.1. Recall that $\pi_E(y - Ax)$ is the likelihood of $Y = y$ given $X = x$. Let $\phi(y; x) = -\log(\pi_E(y - Ax))$. Then $\phi(y; x)$ is called the potential function. Note that this is also called the negative log-likelihood.

Remark 4.1. Let Π and Π_x be measures on \mathbb{R}^m with densities π and π_x respectively. Then from the above theorem we have

$$\begin{aligned} \frac{d\Pi^x}{d\Pi}(x) &= \frac{1}{m(y)} \exp(-\phi(x; y)) \\ m(y) &= \int_{\mathbb{R}} \exp(-\phi(x; y)) d\Pi(x) \end{aligned}$$

so we can reformulate Bayes theorem as

$$\frac{1}{\pi_X}(x)\pi_E(y - Ax)(x | y) = \frac{1}{m(y)}\pi_X(y | x)$$

So the posterior distribution is absolutely continuous with respect to the prior and the Radon-Nikodym derivative is proportional to the likelihood.

We now have a formula to find the conditional probability for $X = x$ given our measurements $Y = y$, where we see that the conditional posterior distribution is a product of the likelihood and the prior on X .

Remark 4.2. So far we have formulated the Radon-Nikodym theorem with respect the finite dimensional case which is the setting of this paper. But using the formulation in remark (4.1), we can generalize Bayes theorem to the infinite dimensional case using Gaussian measure [11]/[2].

Before we move further, we go through two classical examples with can be found in [11].

Example 4.1. Let $x \in \mathbb{R}$, $y \in \mathbb{R}^n$ for $n \geq 1$, and let $A \in \mathbb{R}^{n \times 1} - \{0\}$. Define the observations as

$$y = Ax + e$$

where $e \sim \mathcal{N}(0, \delta^2 I)$. By Bayes theorem the conditional posterior distributing is then

$$\pi(x | y) \propto \exp\left(-\frac{1}{\delta^2} \|Ax - y\|_2^2 - \frac{1}{2}|x|^2\right)$$

The posterior is Normal and is completely characterized by its mean and covariance. The inner equation of exponential is quadratic so we can complete the square and compute the mean and variance, μ and Σ^2 as

$$\mu = \frac{\langle A, y \rangle}{\delta^2 + \|A\|_2^2} \quad \text{and} \quad \Sigma^2 = \frac{\delta^2}{\delta^2 + \|A\|_2^2}$$

We propose that as $\delta \rightarrow 0$, μ will converge to $\frac{\langle A, y \rangle}{\|A\|_2^2}$ (consistency) and that the covariance will converge to 0 (convergence). Indeed by the definition of μ and Σ we can easily see that as $\delta \rightarrow 0$, we have convergence and consistency.

Example 4.2. Let $x \in R^n$ with $n \geq 2$, and let $y \in \mathbb{R}$. Let $A \in R^{n-1} - \{0\}$. Again the observation are

$$y = Ax + e$$

where $e \sim \mathcal{N}(0, \delta^2 I)$. Assume that $x \sim \mathcal{N}(0, \Delta^2 I)$ By Bayes theorem the conditional posterior distributing is then

$$\pi(x | y) \propto \exp\left(-\frac{1}{\delta^2} |\langle A, y \rangle - y|^2 - \frac{1}{2}\langle x, \Delta^{-1}x \rangle\right)$$

Again by completing the square we have that

$$\mu = \frac{x\Delta A}{\delta^2 + \langle A, \Delta A \rangle} \quad \text{and} \quad \Sigma^2 = \Delta - \frac{(\Delta A)(\Delta A)^*}{\delta^2 + \langle A, \Delta A \rangle}$$

Then again we check what happens when the noise level goes to zero.

$$\lim_{\delta \rightarrow 0} \mu = \frac{x\Delta A}{\langle A, \Delta A \rangle} \quad \text{and} \quad \lim_{\delta \rightarrow 0} \Sigma^2 = \Delta - \frac{(\Delta A)(\Delta A)^*}{\langle A, \Delta A \rangle}$$

The $\langle \mu, A \rangle = \bar{x}$, the ground truth, and $\Sigma^2 A = 0$. So as $\delta \rightarrow 0$, uncertainty about \hat{x} goes to zero in the direction of A . There is uncertain in the directions not aligned with A . So in the undetermined case the prior plays a role even as noise the goes to zero. In the over determined case, the prior plays no role as the noise goes to zero.

Remark 4.3.

It is now natural to ask what well-posedness is in this case as the solution is not a point estimator but rather an entire distribution. We can still check if the solution exists, if it is unique and if it is stable. While this is interesting area

of research, it is beyond the scope of this thesis, but has been studied in papers [11][1].

We would now somehow like to understand and explore the posterior distribution. Already if $n > 2$ we cannot graph the posterior distributions. Common methods to explore the higher dimensional posterior distribution are either computing point estimators such as those defined below, or by MCMC sampling methods.

Definition 4.2. *The maximum a posterior estimator of x is found by maximizing the posterior distribution if the maximum exists. That is*

$$x_{MAP} = \max_{\mathbb{R}^m} \pi(x | y) \quad (17)$$

To compute this point estimator is an optimization problem. Another point estimator is the conditional mean estimator defined as

Definition 4.3. *The conditional mean estimator of x given y is*

$$x_{CM} = \mathbb{E}(x | y) = \int_{\mathbb{R}^m} x \pi_X(x | y) dx$$

To compute this point estimator is an integration problem, which can be very difficult in high dimensional settings.

We can also compute spread estimators by computing Bayesian Credible sets. These are defined as follows:

Definition 4.4. *Let $\alpha \in (0, 1)$, then a $1 - \alpha$ level credible set C_α is given by*

$$\Pi(C_\alpha | y) = \int_{C_\alpha} \pi_X(x | y) dx = 1 - \alpha$$

To compute this is a root finding problem.

4.2 Connection to Tikhonov Regularization

Now that we have seen some examples. We will explain under which circumstances we can return to the Tikhonov regularization. We will consider the independent Gaussian noise model with a normal prior. Suppose that the noise E is additive Gaussian noise with each e_i i.i.d. Suppose also that we assume X is Gaussian, and that X is smooth. We formally write this as

$$E \sim \mathcal{N}(0, \alpha I) \quad (18)$$

$$X \sim \mathcal{N}(0, \beta \Sigma) \quad (19)$$

for fixed α, β, Σ . Note that here α, β are precision parameters. Using Bayes formula we find that posterior distribution of $X = x | Y = y$ to be proportional to

$$\pi(x, \alpha, \beta) \propto \alpha/2 \|Ax - y\|^2 - \beta/2 \|Lx\|^2 + m/2\log(\alpha) + n/2\log(\beta) \quad (20)$$

and the potential is

$$\mathcal{J}(x, \alpha, \beta) = \alpha/2\|Ax - y\|^2 + \beta/2\|Lx\|^2 - m/2\log(\alpha) - n/2\log(\beta) \quad (21)$$

The resulting posterior distribution is Gaussian with mean μ and variance Σ . A Gaussian distribution is completely characterized by its mean and variance, so we compute μ and Σ . Do to this we need compute the MAP estimate of $\pi(x | \alpha, \beta)$, which is equivalent to minimizing the potential.

$$x_{MAP} = \min_x \mathcal{J}(x, \alpha, \beta) \quad (22)$$

$$= \min_x \alpha/2\|Ax - y\|^2 + \beta/2\|Lx\|^2 - m/2\log(\alpha) - n/2\log(\beta) \quad (23)$$

We can now easily see that this minimization problem is equivalent minimization problem as the Tikhonov regularization with $\lambda = \alpha/\beta$. We find that

$$\mu = (\alpha A^* A + \beta L^* L)^{-1} \alpha A^* y \quad (24)$$

$$\Sigma = (\alpha A^* A + \beta L^* L)^{-1} \quad (25)$$

5 Bayesian Regularization

Recall that we wanted a data dependant way to choose λ in this setting. We already have some indication for the choice of λ as it is the ratio of the data variance and the observed variance from the previous section. To chose requires choosing α and β such that the resulting minimization problem is well-posed.

5.1 Empirical Bayesian Method

Our first guess is to use empirical Bayesian method that would allow us to determine α and β from the data only. To do this we minimize the objective function over all three parameters.

$$(x_{MAP}, \alpha_{MAP}, \beta_{MAP}) = \min_{x, \alpha, \beta} \mathcal{J}(x, \alpha, \beta) \quad (26)$$

$$= \min_{x, \alpha, \beta} \alpha/2\|Ax - y\|^2 + \beta/2\|Lx\|^2 - m/2\log(\alpha) - n/2\log(\beta) \quad (27)$$

5.2 Well-posedness

We now check that this minimization problem is well-posed.

Claim 5.1. *The empirical Bayesian method is ill-posed.*

Proof. We wish to

$$\min_{x, \alpha, \beta} \alpha/2\|Ax - y\|^2 + \beta/2\|Lx\|^2 - m/2\log(\alpha) - n/2\log(\beta)$$

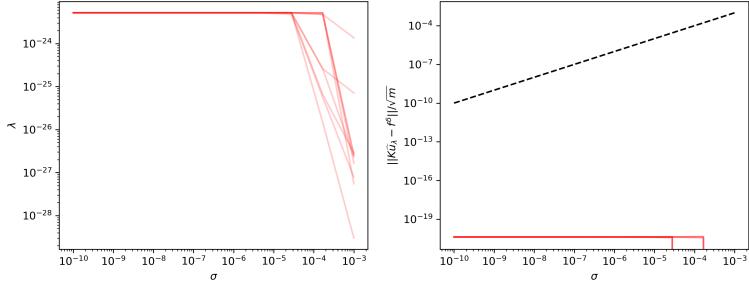


Figure 1: Convergence and consistency of Empirical Bayesian method for $\sigma \in [1e-10, 1e-3]$

Suppose that A is invertible. Recall that $y = Ax + e$. Then $\|Ax - y\| = 0$, and minimizing \mathcal{J} occurs when $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$. If $\alpha \rightarrow \infty$, and $\beta \rightarrow 0$, no regularization occurs. The bias goes to zero, but the variance is high and $\beta \rightarrow 0$ then \hat{x} goes to the zero vector. So the method is not well-posed. \square

Example 5.1. Let $A = I_n$ the identity matrix. This is invertible. Let

$$y = A\sin(x) + e$$

for $x \in [-4\pi, 4\pi]$, $e \sim \mathcal{N}(0, \sigma^2)$. We want that

$$\begin{aligned} \lim_{\sigma \rightarrow 0} \beta/\alpha &= 0 \\ \mathbb{E}\|Ax - y\|_2^2 &= n\sigma^2 \end{aligned}$$

We expect however that $\beta/\alpha = \lambda \rightarrow 0$ regardless of the σ .

We see that this is indeed the case in 1. The problem is then that minimization occurs at the bound extreme values for α, β , which we have seen results in ill-posedness.

5.3 Hierarchical Bayesian Method

We have seen that the empirical Bayesian method of regularization is not a well-posed problem. Failure occurred exactly when A is invertible leading to a solution $A^{-1}y = \hat{x}$, so that $\alpha \rightarrow \infty$ and $\beta \rightarrow 0$. Additionally if did not hold in general that if the noise level went to zero, the regularization went to zero, and the estimate for residuals was consistent. What is needed then is bound α and β . In [7] they prove that to turn the empirical Bayesian method into a well-posed problem we need to place priors on the precision parameters α and β . Since we assume a Gaussian prior on X and E the natural(conjugate) hyper priors are Gamma distributions. In this setting they prove that J a minimum to exists.

Suppose we now add the following hyper priors on the precision parameters

$$\alpha \sim \text{Gamma}(a_0, b_0) \quad (28)$$

$$\beta \sim \text{Gamma}(a_1, b_1) \quad (29)$$

Then the posterior becomes

$$p(x, \alpha, \beta | y) \propto \rho(Ax - y | \alpha) \times \pi(\alpha) \pi(x | \beta) \times \pi(\beta) \quad (30)$$

$$\propto \alpha^{n/2} e^{-\alpha/2||Ax-y||^2} \alpha^{a_0-1} e^{-b_0\alpha} \beta^{n/2} e^{-\beta/2||Lx||^2} \beta^{a_1-1} e^{-b_1\beta} \quad (31)$$

The resulting minimization problem is

$$\begin{aligned} \min_{x, \alpha, \beta} \mathcal{J}(x, \alpha, \beta) = & \alpha/2||Ax - y||^2 - (n/2 + a_0 - 1)\log(\alpha) + b_0\alpha + \\ & \beta/2||Lx||^2 - (n/2 + a_1 - 1)\log(\beta) + b_1\beta \end{aligned}$$

Note that if we let $a_0, a_1 = 1$ and $b_0, b_1 = 0$ the we recover the objective function given no hyper priors.

5.4 Well-posedness

To check if the solution is now stable we again use the optimal solutions. The partial derivatives are

$$\partial/\partial x(J(x, \alpha, \beta)) = (A^* A + \beta/\alpha L^* L)x - A^* y \quad (32)$$

$$\partial/\partial \alpha J(x, \alpha, \beta) = 1/2||Ax - y||^2 - (n/2 + a_0 - 1)/\alpha + b_0 \quad (33)$$

$$\partial/\partial \beta J(x, \alpha, \beta) = 1/2||Lx||^2 - (n/2 + a_1 - 1)/\beta + b_1 \quad (34)$$

so then the optimal solutions are

$$x = (A^* A + \beta/\alpha L^* L)^{-1} A^* y \quad (35)$$

$$\alpha = \frac{(n/2 + a_0 - 1)}{1/2||Ax - y||^2 + b_0} \quad (36)$$

$$\beta = \frac{(n/2 + a_1 - 1)}{1/2||Lx||^2 + b_1} \quad (37)$$

The resulting theorem is proven by [7], and relies on many intermediate lemmas. We restart the main theorem here and briefly prove the main theorem in the case $L = I$.

Theorem 5.1. *Let σ_0^2 denote the variance. Assume that the random variable η_i is such that $|\eta_i| \leq c_\omega \sigma_0$ (is c_ω the decay rate of the eigenvalues of which matrix?) for $i = 1, \dots, n$. Fix b_1 and let $\frac{n}{2} + a_1 - 1 \sim \sigma_0^d$ for $0 < d < 2$. Then*

$$\lim_{\sigma_0 \rightarrow 0} \|\hat{x}_\lambda - x_{LS}\|_2^2 = 0$$

that is as the variance goes to zero the regularization should also go to zero.

6 Numerical Methods

In this section we propose three different iterative methods to numerically solve $\min_{x,\alpha,\beta} \mathcal{J}(x, \alpha, \beta)$. Suppose that we have some function $f(x, y, z)$ and we want to find the minimum of this function. All critical points are then found when $\frac{\partial f}{\partial x} = 0$, $\frac{\partial f}{\partial y} = 0$, and $\frac{\partial f}{\partial z} = 0$. Suppose f is strictly convex then any local minimum on a convex set is a global minimum and in fact this minimum is unique. A common method to numerically solve for minimum of f is coordinate descent, where we used the fact that in the parameter of interested we are guaranteed to move downward by strict convexity. We can then do coordinate descent to find the minimum.

Example 6.1. Suppose we would like to minimize

$$f(x, y) = (x - y)^2$$

The function is minimized along the line $x = y$. We could have also solved this numerically by first noting that $f(x) = x^2 - 2y'x$ and $f(y) = y^2 - 2yx'$ are convex functions with first derivatives equal to $x - y'$ and $y - x'$. Suppose we start at the point $(1, 0)$. And suppose that first we want to minimize along x . Then

$$y = \operatorname{argmin}_x f(x, 0) = \operatorname{argmin}_x x^2 = 0$$

Then

$$x = \operatorname{argmin}_y f(0, y) = y^2 = 0$$

So the minimum is at $(0, 0)$ which is on the line $x = y$.

Now recall that we have computed the partial derivatives to $\mathcal{J}(x, \alpha, \beta)$ in (34) and computed the critical points as (37). However our function is only bi-(strictly) convex. But we can still implement a similar algorithm to coordinate descent by splitting \mathcal{J} into its two strictly convex parts, and do alternating minimization. The following method was proposed by [7] and we use it as a benchmark to compared to the other two algorithms.

6.1 Method 1

In this method the estimates are found by simultaneous minimize over all three parameters. That is we

$$\min_{x,\alpha,\beta} \mathcal{J}(x, \alpha, \beta) \tag{38}$$

This is done by alternating method, where at each iteration we define either a normal equation for x or a normal equation for α, β . The next best guess for

x , respectively α, β is then the root of the normal equation. Recall that these roots are the optimal solutions (37) to the partial derivatives.

$$\begin{aligned} x(\beta/\alpha) &= (A^*A + \beta/\alpha L^*L)^{-1} A^*y \\ \alpha(x) &= \frac{(n/2 + a_0 - 1)}{1/2||Ax - y||^2 + b_0} \\ \beta(x) &= \frac{(n/2 + a_1 - 1)}{1/2||Lx||^2 + b_1} \end{aligned}$$

For fixed a_0, b_0, a_1, b_1 , these defined closed form solutions for x, α, β . Suppose, then, that we start with some initial values α_0, β_0 . Then using the optimal solution for x we can define an estimate

$$\begin{aligned} x_0 &= x(\beta_0/\alpha_0) \\ &= (A^*A + \beta_0/\alpha_0 L^*L)^{-1} A^*y \end{aligned}$$

Notice that this is the normal Tikhonov regularized solution with regularization parameter β_0/α_0 . To find the next estimates for α, β we compute

$$\begin{aligned} \alpha_1 &= \alpha(x_0) \\ &= \frac{(n/2 + a_0 - 1)}{1/2||Ax_0 - y||^2 + b_0} \\ \beta_1 &= \beta(x_0) \\ &= \frac{(n/2 + a_1 - 1)}{1/2||Lx_0||^2 + b_1} \end{aligned}$$

We can repeatedly alternate between minimizing over x versus minimizing over α, β until some stopping criterion is met. The resulting algorithm is given in Algorithm (1).

In [7] the following two theorems are proven for algorithm 1.

Theorem 6.1. (*Theorem 3.1 in [7]*) Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 1. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i \in I}$ converges monotonically.

Theorem 6.2. (*Theorem 3.2 in [7]*) Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 1, then this sequence converges to a critical point of $\mathcal{J}(x, \alpha, \beta)$.

This method works when we can compute the partial derivatives of x, α, β , and have closed form solutions. This is indeed the case when we assume normal prior on x , and Gamma priors on α, β , where by we could recover the Tikhonov regularization with L^2 penalty on x . We now propose two additional methods where we assume we do not have closed form solutions. This way we can still numerically solve $\min_{x, \alpha, \beta} \mathcal{J}(x, \alpha, \beta)$ in case where different prior assumptions lead to different penalties.

Algorithm 1 Alternating Algorithm

Require: $I, x_0, \alpha_0, \beta_0, \epsilon \geq 0$

Require: $a_0, b_0, a_1, b_1 > 0$

$i \leftarrow 1$

$g \leftarrow \|\nabla \mathcal{J}(x_0, \alpha_0, \beta_0)\|_2^2$

while $i \leq I \ \& \ g < \epsilon$ **do**

$x_i \leftarrow (A^* A + \beta_{i-1} / \alpha_{i-1} L^* L)^{-1} A^* y$

$\alpha_i \leftarrow \frac{(n/2+a_0-1)}{1/2\|Ax_i-y\|^2+b_0}$

$\beta_i \leftarrow \frac{(n/2+a_1-1)}{1/2\|Lx_i\|^2+b_1}.$

$g \leftarrow \|\nabla \mathcal{J}(x_i, \alpha_i, \beta_i)\|_2^2$

$i \leftarrow i + 1$

end while

6.2 Method 2

Suppose now that we do not have closed form solutions for α, β . Define

$$\begin{aligned} \mathcal{J}_1(\alpha, \beta) = \mathcal{J}(\alpha, \beta; x) &= \alpha/2\|A\hat{x}(\alpha, \beta) - y\|^2 - (n/2 + a_0 - 1)\log(\alpha) + b_0\alpha + \\ &\quad \beta/2\|L\hat{x}(\alpha, \beta)\|^2 - (n/2 + a_1 - 1)\log(\beta) + b_1\beta \end{aligned}$$

for fixed $\hat{x}(\alpha, \beta)$. First supposing that α, β are fixed we can then minimize \mathcal{J} with

$$x \stackrel{\text{set}}{=} (A^* A + \beta/\alpha L^* L)^{-1} A^* y$$

So the inner minimization fixes x at the Tikhonov estimate for some given initial values of α, β . So want we really want to do is

$$\min_{\alpha, \beta} \left[\min_x \mathcal{J}(x, \alpha, \beta) \right]$$

The inner minimization can be done using the closed form solution of x , but to minimize over α, β we must use gradient descent that is

$$\begin{aligned} \alpha_{i+1} &= \alpha_i - \mu \partial_\alpha \mathcal{J}_1(\alpha_i, \beta_i) \\ \beta_{i+t} &= \beta_i - \mu \partial_\beta \mathcal{J}_1(\alpha_i, \beta_i) \end{aligned}$$

Thus to solve the joint minimization problem start with an initial α, β , compute $x(\beta/\alpha)$, then solve for $\hat{\alpha}, \hat{\beta}$ by taking one step along the gradient in the direction α, β . The resulting algorithm is given in Algorithm (2). We now prove the same convergence theorems for algorithm 2. First recall the following theorems

Definition 6.1. Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be metric space with distance metrics $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ respectively. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$. Then f is Lipschitz continuous if there exists a $K \in \mathbb{R}$ with $K \geq 0$ such that for all $x_1, x_2 \in \mathcal{X}$

$$\|f(x_1) - f(x_2)\|_{\mathcal{Y}} \leq L\|x_1 - x_2\|_{\mathcal{X}}$$

Algorithm 2 Gradient Descent in α, β

Require: $x_0, \alpha_0, \beta_0, \epsilon > 0$

Require: $a_0, b_0, a_1, b_1 > 0$

Require: $I > \mathcal{O}(1/\epsilon)$

$i \leftarrow 1$

$g \leftarrow \|\nabla \mathcal{J}(x_0, \alpha_0, \beta_0)\|_2^2$

while $i \leq I$ & $g < \epsilon$ **do**

$x_i \leftarrow (A^* A + \beta_{i-1} L^* L)^{-1} A^* y$

$L_\alpha \leftarrow \frac{1}{2} \|Ax_i - y\|_2 - (n/2 + a_0 - 1) + b_0$

$L_\beta \leftarrow \frac{1}{2} \|Lx_i\|_2 - (n/2 + a_1 - 1) + b_1$

$\mu_\alpha \leftarrow \epsilon / L_\alpha^2$

$\mu_\beta \leftarrow \epsilon / L_\beta^2$

$\alpha_i \leftarrow \alpha_{i-1} - \mu_\alpha \partial_\alpha \mathcal{J}_1(\alpha_{i-1}, \beta_{i-1})$

$\beta_i \leftarrow \beta_{i-1} - \mu_\beta \partial_\beta \mathcal{J}_1(\alpha_{i-1}, \beta_{i-1})$

$g \leftarrow \|\nabla \mathcal{J}(x_i, \alpha_i, \beta_i)\|_2^2$

$i \leftarrow i + 1$

end while

L is then called the Lipschitz constant of f .

Remark 6.1. Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be metric space with distance metrics $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ respectively. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$. Suppose that f is L -Lipschitz continuous. Then for all $x_1, x_2 \in \mathcal{X}$

$$\frac{\|f(x_1) - f(x_2)\|_{\mathcal{Y}}}{\|x_1 - x_2\|_{\mathcal{X}}} \leq L$$

Remark 6.2. Let $\mathcal{X}, \mathcal{Y}, f$ be defined as in the previous remark. Suppose that \mathcal{X} is closed and that f is differentiable \mathcal{X}° . Then by the mean value theorem

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2} \leq f'(z)$$

for all $x_1 < z < x_2 \in \mathcal{X}^\circ$. This implies that we can find L such that $\|f'(z)\|_{\mathcal{Y}} < L$.

Theorem 6.3. Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 2. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i \in I}$ converges monotonically.

Proof. For fixed a_0, b_0, a_1, b_1 , recall that we defined

$$\mathcal{J}_1(\alpha, \beta) = \mathcal{J}(\hat{x}, \alpha, \beta) = \min_x \mathcal{J}(x, \alpha, \beta)$$

So then

$$x_{i+1} = \underset{x}{\operatorname{argmin}} \mathcal{J}(x_i, \alpha, \beta)$$

$$\begin{pmatrix} \alpha_{i+1} \\ \beta_{i+1} \end{pmatrix} = \begin{pmatrix} \alpha_i - \mu \partial_\alpha \mathcal{J}_1(\alpha_i, \beta_i; x_{i+1}) \\ \beta_i - \mu \partial_\beta \mathcal{J}_1(\alpha_i, \beta_i; x_{i+1}) \end{pmatrix}$$

Fix a_0, b_0, a_1, b_1 such that $\mathcal{J}_1(\alpha, \beta)$ and $\mathcal{J}_2(x)$ are convex. Let L be such that $\|\nabla \mathcal{J}_1(\alpha, \beta)\|_2 < L$ and $\epsilon > 0$. Set $\mu = \epsilon/L^2$. Then

$$\mathcal{J}_1(\alpha_{i+1}, \beta_{i+1}; x_{i+1}) \leq \mathcal{J}_1(\alpha_i, \beta_i; x_{i+1})$$

This implies that

$$\mathcal{J}(x_{i+1}, \alpha_{i+1}, \beta_{i+1}) \leq \mathcal{J}(x_{i+1}, \alpha_i, \beta_i) \leq \mathcal{J}(x_i, \alpha_i, \beta_i)$$

So \mathcal{J} is monotonically decreasing and bounded below by theorem (??) so we have the \mathcal{J} converges. \square

Theorem 6.4. *Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 2, then this sequence converges to a critical point $\{x_*, \alpha_*, \beta_*\}$ of $\mathcal{J}(x, \alpha, \beta)$.*

Proof. ??? Fix a_0, b_0, a_1, b_1 such that $\mathcal{J}_1(\alpha, \beta)$ and $\mathcal{J}_2(x)$ are convex. Let L be such that $\|\nabla \mathcal{J}_1(\alpha, \beta)\|_2 < L$ and $\epsilon \geq 0$. Set $\mu = \epsilon/L^2$.

$$J_2(\alpha_i, \beta_i) \rightarrow J_2(\alpha_*, \beta_*)$$

Recall that $\hat{x}_i(\alpha_{i-1}, \beta_{i-1}) = (A^* A + \beta_{i-1}/\alpha_{i-1} L^* L)^{-1} A^* y$. Then

$$\mathcal{J}_2(\alpha_i, \beta_i; \hat{x}(\alpha_i, \beta_i)) \rightarrow \mathcal{J}_2(\alpha_*, \beta_*; \hat{x}(\alpha_*, \beta_*)) = \mathcal{J}(x_*, \alpha_*, \beta_*)$$

\square

6.3 Method 3

Suppose now that we do not have a closed form solution for x . Then define

$$\mathcal{J}_2(x) = \mathcal{J}(x; \hat{\alpha}(x), \hat{\beta}(x)) = \hat{\alpha}(x) \|Ax - y\|_2^2 + \hat{\beta}(x) \|Lx\|_2^2$$

for fixed $\hat{\alpha}(x), \hat{\beta}(x)$ and constants a_0, b_0, a_1, b_1 such that $a_0, a_1 \neq 1$ and $b_0, b_1 \neq 0$. Similarly as in method two we compute estimates for x by taking one step along the gradient of \mathcal{J} in the direction of x . Our minimization problem is then

$$\min_x \left[\min_{\alpha, \beta} \mathcal{J}(x, \alpha, \beta) \right]$$

The inner minimization can be solved by using the closed form solution for α, β given some fixed x . For the outer minimization we need to use a gradient method. To do this we compute

$$\nabla_x \mathcal{J}_2(x) = (A^* A + \beta(x)/\alpha(x) L^* L)y - A^*.$$

. And take one step along the gradient there by finding the next best guess for x given α, β

$$x_{i+1} = x_i - \mu \nabla_x \mathcal{J}_2(x_i)$$

where i denotes where we are in the iteration. So to solve the joint minimization problem start with an initial x , then minimize over α, β by computing their optimal solutions, then iterative solve for \hat{x} and update α, β . The resulting algorithm is given in Algorithm (3). We now prove the same convergence theorems

Algorithm 3 Gradient Descent in x

Require: $x_0, \alpha_0, \beta_0, \epsilon > 0$
Require: $a_0, b_0, a_1, b_1 > 0$
Require: $I > \mathcal{O}(1/\epsilon)$

```

 $i \leftarrow 1$ 
 $g \leftarrow \|\nabla \mathcal{J}(x_0, \alpha_0, \beta_0)\|_2^2$ 
while  $i \leq I \ \& \ g < \epsilon$  do
     $\alpha_i \leftarrow \frac{(n/2+a_0-1)}{1/2\|Ax_{i-1}-y\|^2+b_0}$ 
     $\beta_i \leftarrow \frac{(n/2+a_1-1)}{1/2\|Lx_{i-1}\|^2+b_1}.$ 
     $L \leftarrow \|A^*A + \beta_{i-1}/\alpha_{i-1}L^*L\|_2$ 
     $\mu \leftarrow \epsilon/L^2$ 
     $x_i \leftarrow x_{i-1} - \mu \nabla \mathcal{J}_2(x_{i-1})$ 
     $g \leftarrow \|\nabla \mathcal{J}(x_i, \alpha_i, \beta_i)\|_2^2$ 
     $i \leftarrow i + 1$ 
end while

```

for algorithm 3.

Theorem 6.5. Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 3. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i \in I}$ converges monotonically.

Proof. For fixed a_0, b_0, a_1, b_1 , recall that we defined

$$\mathcal{J}_2(x) = \mathcal{J}(x, \hat{\alpha}, \hat{\beta}) = \min_{\alpha, \beta} \mathcal{J}(x, \alpha, \beta)$$

So then

$$\begin{aligned}
(\alpha_{i+1}, \beta_{i+1}) &= \operatorname{argmin}_{\alpha, \beta} \mathcal{J}(x_i, \alpha, \beta) \\
x_{i+1} &= x_i - \mu \nabla \mathcal{J}_2(x_i; \alpha_{i+1}, \beta_{i+1})
\end{aligned}$$

Fix a_0, b_0, a_1, b_1 such that $\mathcal{J}_1(\alpha, \beta)$ and $\mathcal{J}_2(x)$ are convex. Let L be such that $\|\partial \mathcal{J}_2(x)\|_2 < L$ and $\epsilon > 0$. Set $\mu = \epsilon/L^2$. Then

$$\mathcal{J}_2(x_{i+1}; \alpha_{i+1}, \beta_{i+1}) \leq \mathcal{J}_2(x_i; \alpha_{i+1}, \beta_{i+1})$$

This implies that

$$\mathcal{J}(x_{i+1}, \alpha_{i+1}, \beta_{i+1}) \leq \mathcal{J}(x_i, \alpha_{i+1}, \beta_{i+1}) \leq \mathcal{J}(x_i, \alpha_i, \beta_i)$$

So \mathcal{J} is monotonically decreasing and bounded below by theorem (??) so we have the \mathcal{J} converges. \square

Theorem 6.6. *Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 3, then this sequence converges to a critical point of $\mathcal{J}(x, \alpha, \beta)$.*

Proof. ??? Fix a_0, b_0, a_1, b_1 such that $\mathcal{J}_1(\alpha, \beta)$ and $\mathcal{J}_2(x)$ are convex. Let L be such that $\|\partial\mathcal{J}_1(\alpha, \beta)\|_2 < L$ and $\epsilon \geq 0$. Set $\mu = \epsilon/L^2$.

$$J_1(x_i) \rightarrow J_1(x_*)$$

Recall that

$$\begin{aligned}\hat{\alpha}_i(x_{i-1}) &= \frac{(n/2 + a_0 - 1)}{1/2\|Ax_{i-1} - y\|^2 + b_0} \\ \hat{\beta}_i(x_{i-1}) &= \frac{(n/2 + a_1 - 1)}{1/2\|Lx_{i-1}\|^2 + b_1}\end{aligned}$$

Then

$$\mathcal{J}_1(x_i; \hat{\alpha}_i(x_{i-1}), \hat{\beta}_i(x_{i-1})) \rightarrow \mathcal{J}_1(x_*; \hat{\alpha}_i(x_*), \hat{\beta}_i(x_*)) = \mathcal{J}(x_*, \alpha_*, \beta_*)$$

\square

7 Implementation

7.1 Example

In this section we show the results of a particular example we used to test our implementation in python. The inverse problem we are interested in is recovering $\sin(x)$ with $x \in [-4\pi, 4\pi]$. We observe

$$y = A(\sin(x)) + \epsilon \tag{39}$$

where

$$Af(t) = \int_0^1 \frac{f(y)}{((1 + (t - y)^2)^3/2)} dy \tag{40}$$

for our implementation we discretize the the interval $[-4\pi, 4\pi]$ into n evenly spaced points, and discretize A by letting the step size be equal to m . The resulting forward problem $A(\sin(x))$ is an m system of equations with n variables which is well defined as the number of columns of A equals the number of rows of $\sin(x)$. We observe y with random Gaussian noise center at 0, and variance σ . In our experiments we let $n = m = 200$, and the noise be distributed $\mathcal{N}(0, 0.1)$.

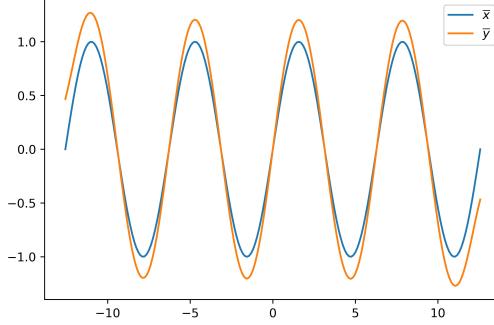


Figure 2: Ground Truth \bar{x} , where $\bar{y} = A\bar{x}$

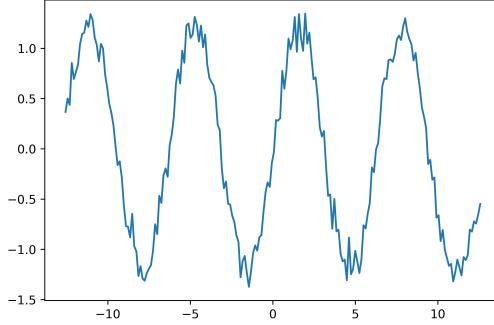


Figure 3: Noisy observation with noise level $\mathcal{N}(0, 0.1)$

7.2 Ill-posedness

In this section we looked at the ill-posedness of finding the least squares estimate. The conditioning number of A is $4887979232 \approx 10^{9.689} \gg 1$ which is much larger than 1. A is ill-conditioned so the direct inverse problem of inverting the matrix A is ill-posed. We further plot the decay of the singular values of A . We can see that the eigenvalues quickly decay to zero. Now we examine the Picard condition. Where we see that the Picard condition is not met (5), so in fact the problem is ill-posed. If we were to ignore the ill-posedness and directly invert A , which we can do since A is full rank and $\det(A) \neq 0$, the resulting solution is the least squares solution which we plot below (6). We see that this solution has high variance and does not accurately recover \bar{x} . We note that a majority of the variance is around the peaks.

7.3 Regularization

Since the problem of recovering \bar{x} by simply inverting A is an ill-posed we instead solve the well-posed problem using regularization. The resulting well-

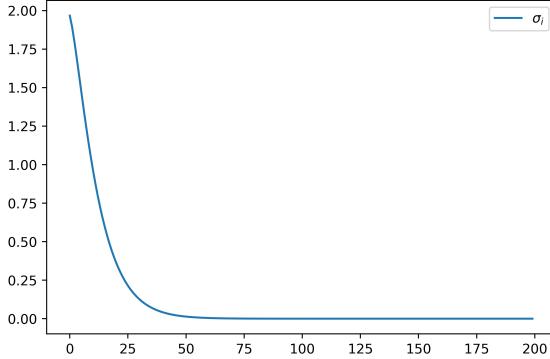


Figure 4: Decay of Singular Values of A

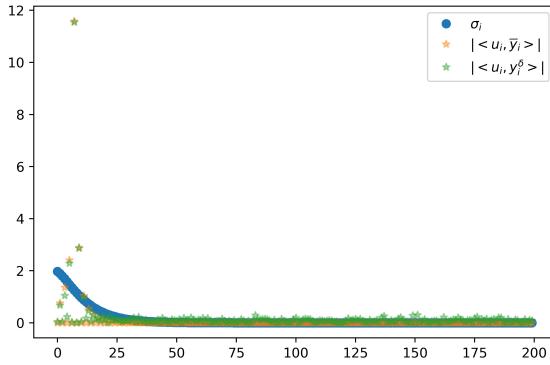


Figure 5: Picard Condition

posed problem is

$$\min_{x,\alpha,\beta} \mathcal{J}(x,\alpha,\beta) = \ell(x,\alpha,\beta \mid y) = \alpha/2\|Ax-y\|^2 - (n/2+a_0-1)\log(\alpha) + b_0\alpha + \beta/2\|Lx\|^2 - (n/2+a_1-1)\log(\beta) + b_1\beta$$

We numerically solve this by the three proposed methods in section 6.1. The stopping condition is the first order condition

$$\|\partial_x \mathcal{J}\|_2^2 + \|\partial_\alpha \mathcal{J}\|_2^2 + \|\partial_\beta \mathcal{J}\|_2^2 < tol$$

If the stopping condition is met, then we say that the algorithm has converged. The parameters were set to Recall that $\mathcal{J}(x,\alpha,\beta)$ is convex in α,β . Below we fix x and plot the contour plots. We see that the contour lines are slightly non-circular. The gradient is steeper in β than it is in α . The red dot approximates the minimum of $\mathcal{J}(\hat{x}(\alpha,\beta),\alpha,\beta)$. Finally, taking advantage of the Bayesian

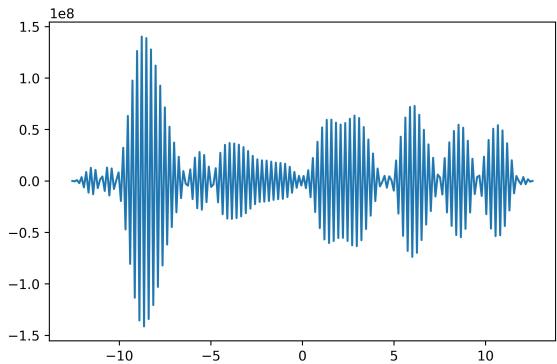


Figure 6: Least Squares solution

tol	$1e - 5$
max iter	100,000
n	200
$\alpha_{initial}$	10
$\beta_{initial}$	1
$x_{initial}$	$\kappa = \frac{ A^*y^\delta _2^2}{ AA^*y^\delta _2^2}, x = \kappa A^*y^\delta$
$a_0 = a_1$	$1 + 1e - 6$
$b_0 = b_1$	$1e - 6$

Table 1: Initial parameters

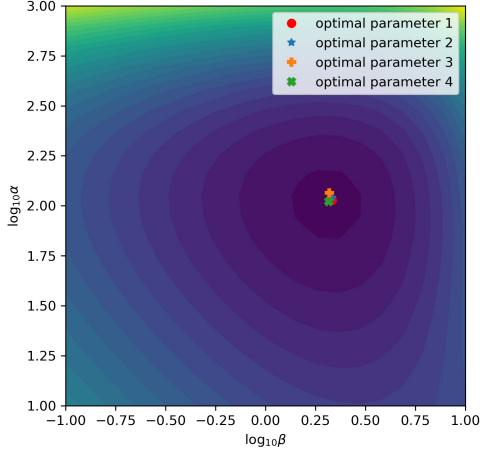


Figure 7: Contour plots of $J(\hat{x}(\alpha, \beta), \alpha, \beta) = z$. Optimal parameter found by all four algorithms.

setting, we can see these contour plots as highest posterior density confidence intervals for α, β given fixed x . Below we plot the results of running all 3 algorithms. We ran a fourth algorithm that was a modification of method 1, where we replaced the closed form solution to x in method 1 with that of a gradient method implemented from python. In figure (8) we plot the path of the objective function over all iterations that the algorithms ran. We hope to see that they are monotonically decreasing. This is indeed the case in algorithms 1 and 4. However this is not the case in 2 and 3. In algorithm 1 and 4, the objective function converged in less than 10 iterations, and we can see in the left hand plot in figure (8) that after 3 iterations the graph at \mathcal{J} is flat. Algorithm 2 and 3 also converged but much slower. These both converged in under 10,000 iterations. We can also see that the slope of \mathcal{J} in algorithm 1 and 4 is very steep in comparison to that of algorithm 2 and 3. The slope of \mathcal{J} is steeper in algorithm 2, than in algorithm 3. In figure (9) we plot the optimal parameter found at each iteration. We see that in algorithm 1 and 4, we jump very quickly to the minimum. In algorithm 2 we have a pretty continuous path to the minimum, and again in algorithm 3 something strange is happening as we can see there is a small jump to the left. In [7], they prove that for the alternating algorithm (algorithm 1) that for any initial λ_0 , the sequence of $\{\lambda_i\}$ generated by the alternating algorithm is monotonic. Moreover they prove that $\{\alpha_i\}, \{\beta_i\}$ is converging monotonically to critical points α_*, β_* . In figure (10) we plot the $\hat{\alpha}, \hat{\beta}, \hat{x}$, and $\hat{\lambda}$ over all iterations for algorithms 1 and 4. We see that minus the first initial guess for β the sequence of $\{\alpha_i, \beta_i, \lambda_i\}$ is monotonic. In figure (11) we again plot the $\hat{\alpha}, \hat{\beta}, \hat{x}$, and $\hat{\lambda}$ over all iterations for algorithms 2 and 3. We see that in algorithm 2 the sequence of $\{\alpha_i, \lambda_i\}$ is monotonic,

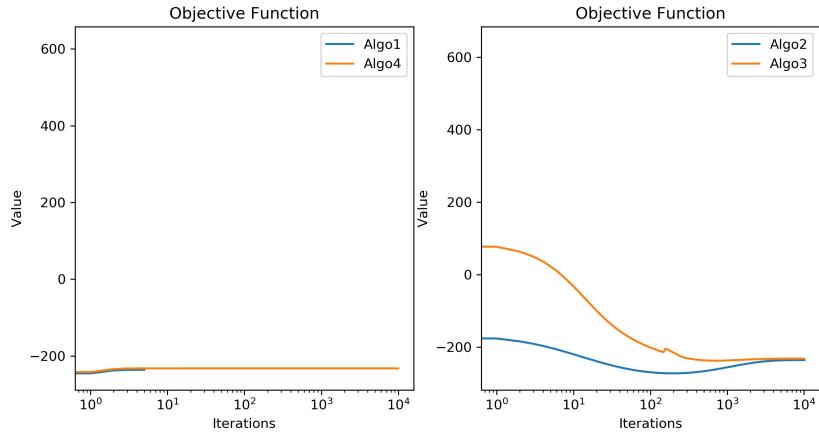


Figure 8: Plot of Objective function over all iterations. On left we plot results from Algorithm 1 and 4. On the right we plot the results of Algorithm 2 and 3.

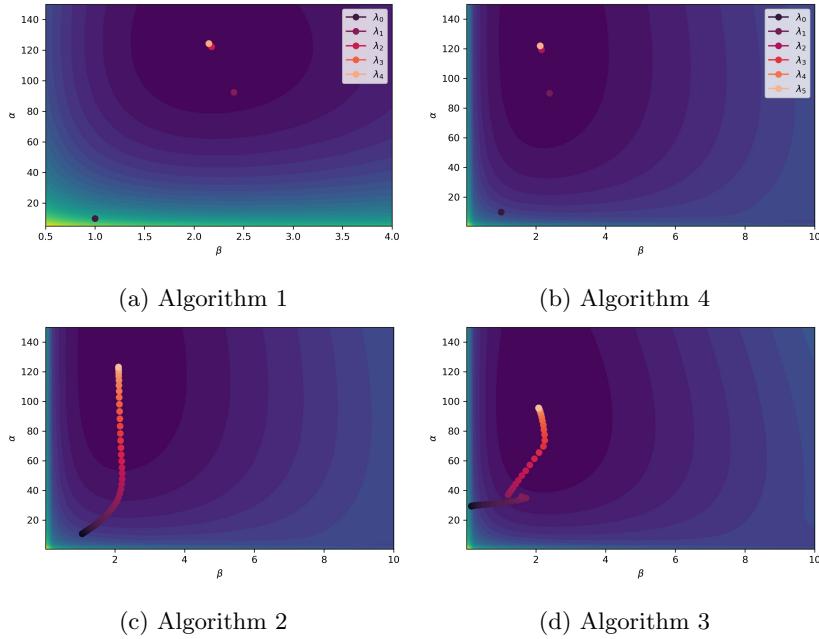
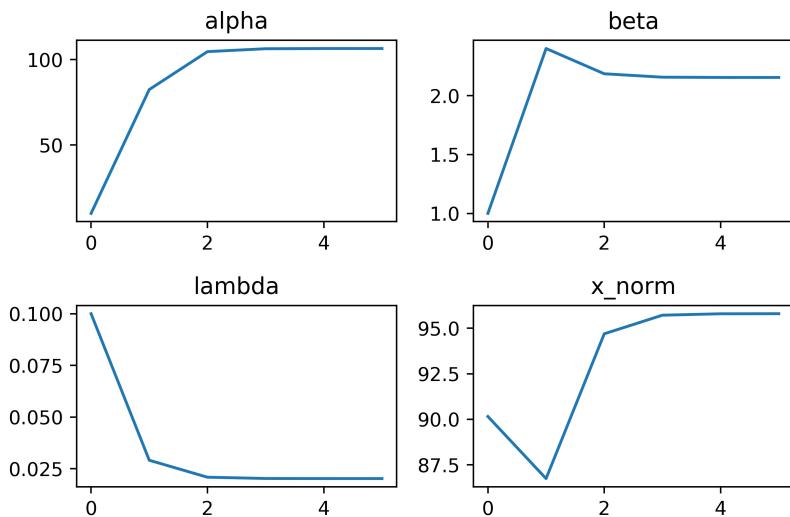
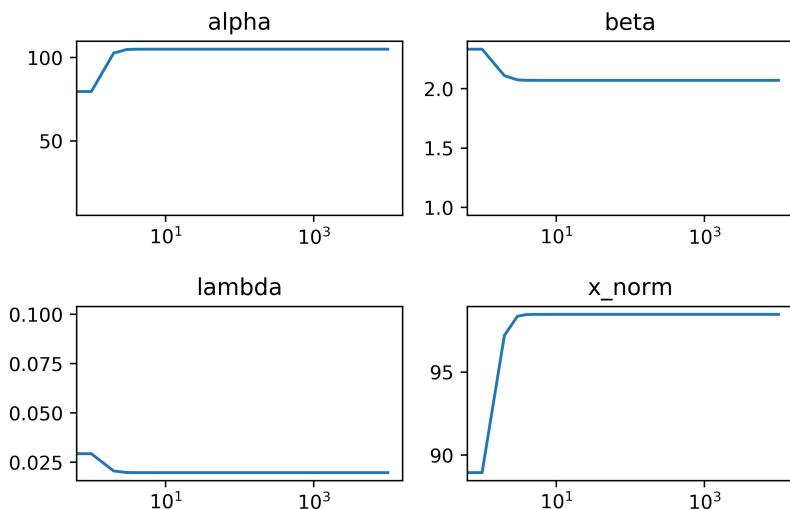


Figure 9: Plots of Optimal Parameter at Each Iteration

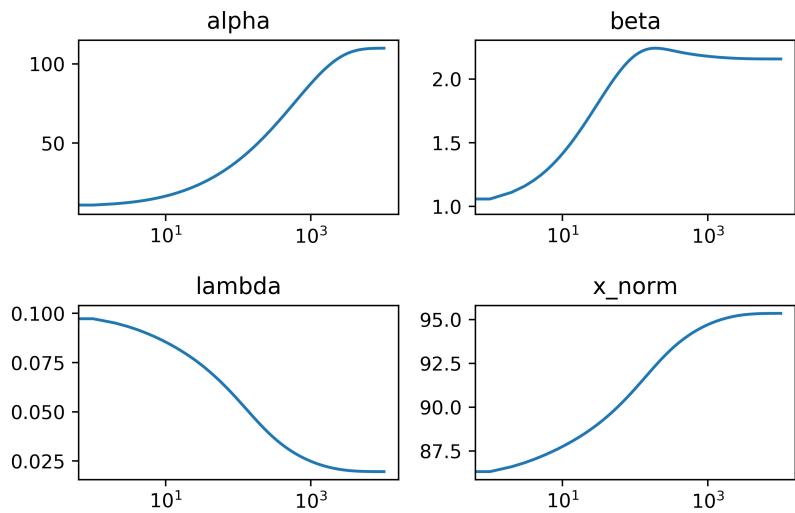


(a) Algorithm 1

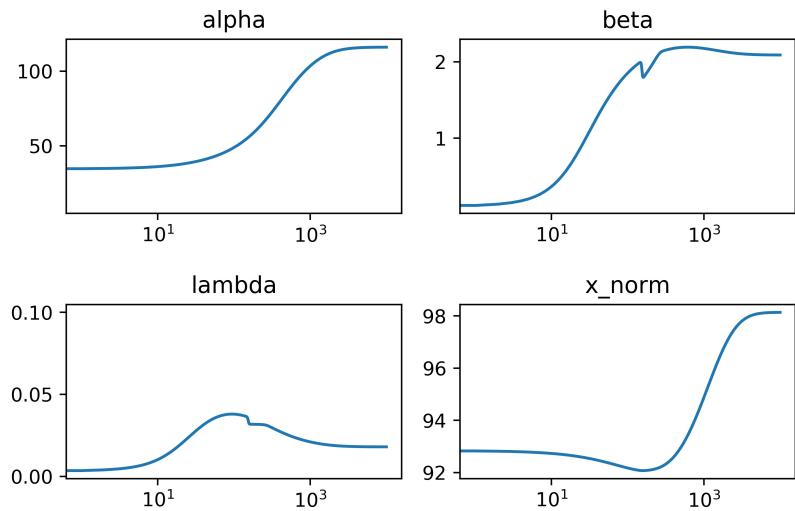


(b) Algorithm 4

Figure 10: Convergence of estimators Algorithm 1 and 4



(a) Algorithm 2



(b) Algorithm 3

Figure 11: Convergence of estimators Algorithm 2 and 3

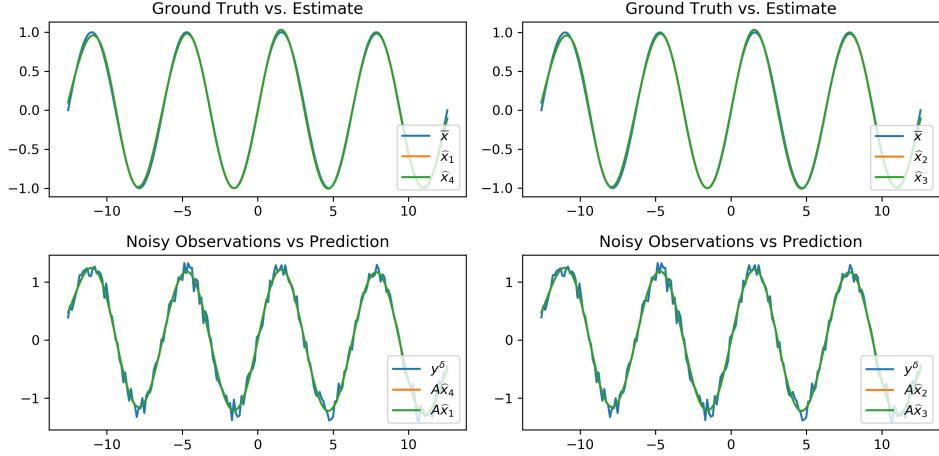


Figure 12: Top plots are of \hat{x} found by each algorithm versus the ground truth \bar{x} . Bottom plots compare the noisy observations versus $A\hat{x}$ for each Algorithm. On the left we compare Algorithm 1 and 4, and on the right we compare Algorithm 2 and 3.

	α	β	λ	$\mathcal{J}(x, \alpha, \beta)$	$\ \bar{x} - \hat{x}\ _2^2$	niter
Algo1	122.06212	2.13766	0.01751	-233.60244	0.18775	4
Algo2	117.01666	2.15817	0.01844	-235.73951	0.18871	6598
Algo3	107.86797	2.14434	0.01988	-235.74297	0.20619	6302
Algo4	88.73767	2.12964	0.02400	-234.66994	0.18082	4

Table 2: Results

However $\{\beta\}_i$ is not. In the bottom plots of figure (11) we see a strange kink that was also seen in figure (d) of (9). We are unsure why this is so.

In the upper plots of figure 12 we see that all estimates of \bar{x} are pretty close to the ground truth, and even hard to distinguish among each other. We also see that each estimate has much lower variance than that of the least squares estimate (fig. 6). In table 2 we summarize the all of our results. Overall all algorithms 1,2, and 4 found roughly the same regularization parameter. Algorithm 4 found a higher regularization parameter and resulted in a lower error.

7.4 Convergence and Consistency

In the this section we examine consistency and convergence of Algorithm 1, since this was the only one that converged and converged relatively quickly. We proposed that

$$\mathbb{E}\|A\hat{x}(\hat{\alpha}, \hat{\beta}) - y\|_2^2 = n\sigma^2 \quad \lim_{\sigma^2 \rightarrow 0} \beta/\alpha \rightarrow 0$$

We see that as the noise level of the model goes to zero, so does the regulariza-

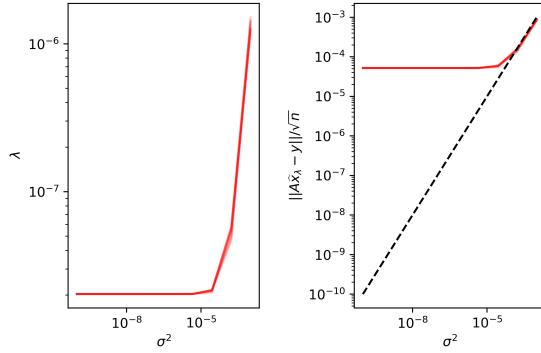


Figure 13: Regularization vs noise and Estimate of residuals vs noise

tion. We also see the effects of bounding β/α in the plot (fig 13) on the right hand side where at some level of noise Algorithm 1 slightly over estimates the noise level as it is bounded away from zero.

7.5 Sensitivity

In this section we examine the sensitive of algorithm 1, to values of the hyper priors. This was already explored in [7], but we make an additional observation that in for certain values of hyper parameters the objective function increased to a critical point rather than decreased to the minimum.

$b_0 = b_1$	α	β	λ	$\mathcal{J}(x, \alpha, \beta)$	$\ \bar{x} - \hat{x}\ _2^2$	niter
1e4	0.0100	0.0100	0.9995	47086.284332	15.344035	3
1e2	0.8967	0.8434	0.9405	156.021743	14.264391	7
1e1	45.5487	2.1735	0.0477	-259.513647	0.246932	4
1e - 2	87.9113	2.1303	0.0242	-235.477913	0.181100	4
1e - 4	88.7295	2.1296	0.0240	-234.678019	0.180828	4
1e - 6	88.7377	2.1296	0.0240	-234.669902	0.180825	4
1e - 8	88.7378	2.1296	0.0240	-234.669821	0.180825	4

Table 3: Results of varying hyper priors $b_0 = b_1$

$a_0 = a_1$	α	β	λ	$\mathcal{J}(x, \alpha, \beta)$	$\ \bar{x} - \hat{x}\ _2^2$	niter
1e4	9 206.6177	216.8017	0.0235	9.200286e+07	0.1785	4
1e2	182.3093	4.2931	0.0235	1.762785e+04	0.1785	4
1e1	92.0662	2.1680	0.0235	-1.472485e+02	0.1785	4
1e - 2	91.1637	2.1468	0.0235	-2.348479e+02	0.1785	4
1e - 4	91.1547	2.1466	0.0235	-2.357148e+02	0.1785	4
1e - 6	91.1546	2.1466	0.0235	-2.357235e+02	0.1785	4
1e - 8	91.1546	2.1466	0.0235	-2.357236e+02	0.1785	4

Table 4: Results of varying hyper priors $a_0 = a_1$

Varying b_0, b_1 the variance parameters to the hyper priors does not seem to effect convergence. We see that too large $b_0 = b_1$ leads to over regularization, but for $b_0 = b_1 < 1e - 4$ we see little change. It is strange that in the first two cases \mathcal{J} is minimized at a high positive value [figs 14, 15]. It does seem to be the case that $a_0 = a_1$ can be chosen more freely, ([7] pages 16-18), in that small or large values lead to similar regularization and convergence. However the again large values for $a_0 = a_1$ lead to a high value positive of \mathcal{J} at the critical point. See appendix for plots of the objective function and the comparison between \bar{x} vs \hat{x} and y^δ vs \widehat{Ax} .

8 Conclusions

We have shown that under the assumption that the data is coming from a normal distribution and the noise is addition normal computing the MAP estimates of the posterior distribution $p(x, \alpha, \beta | y)$ is equivalent to $\min_{x, \alpha, \beta} \mathcal{J}(x, \alpha, \beta)$. Solving the minimization problems allows us to simultaneously estimate the underlying parameter x , the regularization parameter λ , and the noise level. We have developed and implement numerical methods to solve the functional \mathcal{J} , in the case that no closed form solution exists for x , and α, β , as well as show their convergence to a minimum. We have shown that in the simple case, these algorithms work relatively well suggesting that using a gradient methods in place of the close form solution should work, as would be necessary in the case of non-smooth priors. We looked again at the effects of the hyper prior on the alternating algorithm from [7], and see that in fact they play important role. For some choice of the hyper priors the functional no longer decreases, so convergence to a minimum is not guarantee. We have seen that for some choice of hyper priors the functional actually increases. Perhaps this is because a saddle point exists or for some hyper priors \mathcal{J} is no longer convex. Additionally while developing methods 2 and 3, we saw that step size played an important role in the convergence. From the theory we know that this step size depends on the Lipschitz continuity. In our implementation, we used an ad-hoc fixed step size in method 3. The step size in method 3 is a weighted step size that accounts for the different scales of α, β , as well as the differing steepness in each direction

respectively.

Further research would be

- Analyze the role of the hyper priors and develop a method to choose them without the knowledge noise level or variance of the underlying parameter.
- Extend the parameter choice method to settings to the non-normal setting such as non smooth and spare priors and larger scale settings where $n \gg 200$.
- Adjusted algorithms 2 and 3 to be able to solve the resulting functional.
- Develop and implement an adaptive step size for faster convergence.

9 Appendix

9.1 Sensitivity Plots

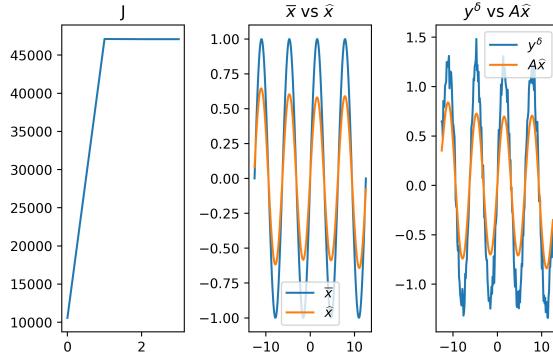


Figure 14: Results when $b_0 = b_1 = 1e4$

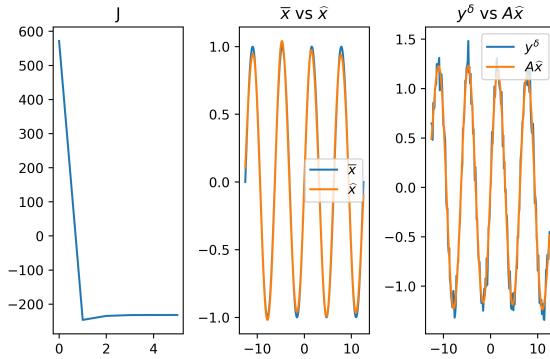


Figure 15: Results when $b_0 = b_1 = 1e-8$

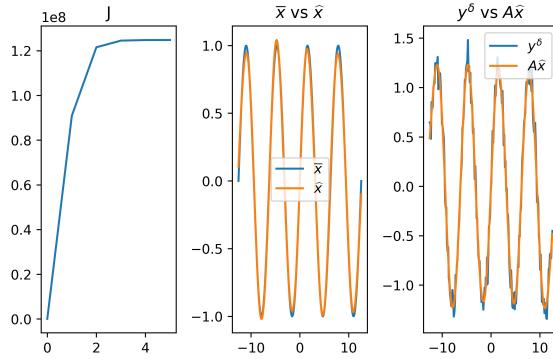


Figure 16: Results when $a_0 = a_1 = 1e4$

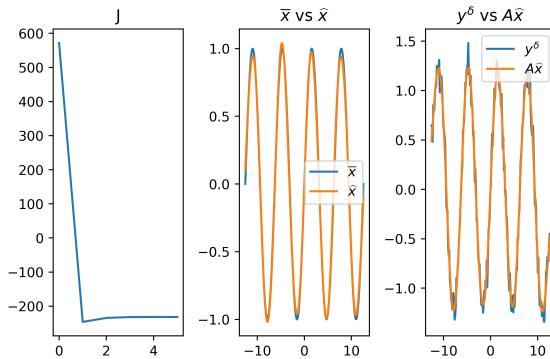


Figure 17: Results when $a_0 = a_1 = 1e-8$

References

- [1] Simon Arridge et al. “Solving Inverse Problems Using Data-driven Models”. In: *Acta Numerica* 28 (May 2019), pp. 1–174. ISSN: 14740508. doi: [10.1017/S0962492919000059](https://doi.org/10.1017/S0962492919000059).
- [2] Masoumeh Dashti and Andrew M. Stuart. *The Bayesian Approach to Inverse Problems*. June 2017. doi: [10.1007/978-3-319-12385-1_7](https://doi.org/10.1007/978-3-319-12385-1_7).
- [3] Matthias J Ehrhardt and Lukas F Lang. *Inverse Problems*. 2018.
- [4] Andrew Gelman et al. *Bayesian Data Analysis CHAPMAN HALL/CRC Texts in Statistical Science Series Series Editors Analysis of Failure and Survival Data*. 2014.
- [5] Christian Hansen. “The Discrete Picard Condition for Discrete Ill-Posed Problems”. In: *BIT* 30 (1990), pp. 658–072.
- [6] Engl Hienz, Hanke Martin, and Andreas Neubauer. *Regularization of Inverse Problems (Mathematics and Its Applications)-Springer* (1996). Vol. 1. 1996.
- [7] Bangti Jin and Jun Zou. “Augmented Tikhonov Regularization”. In: *Inverse Problems* 25 (2 2009). ISSN: 02665611. doi: [10.1088/0266-5611/25/2/025001](https://doi.org/10.1088/0266-5611/25/2/025001).
- [8] Jari Kaipio and Erkki Sommersalo. *Statistical and Computational Inverse Problems*. Vol. 160. 2004.
- [9] Felix Lucka et al. “Risk estimators for choosing regularization parameters in ill-posed problems - Properties and limitations”. In: *Inverse Problems and Imaging* 12 (5 2018), pp. 1121–1155. ISSN: 19308345. doi: [10.3934/ipy.2018047](https://doi.org/10.3934/ipi.2018047).
- [10] Ali Mohammad-Djafari. *A Full Bayesian Approach for Inverse Problems*. 2001.
- [11] A. M. Stuart. “Inverse problems: A Bayesian Perspective”. In: *Acta Numerica* 19 (May 2010), pp. 451–459. ISSN: 09624929. doi: [10.1017/S0962492910000061](https://doi.org/10.1017/S0962492910000061).
- [12] A W Van Der Vaart. *Mathematische Statistiek*. 1997.
- [13] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. 2004.
- [14] Wessel N. van Wieringen. *Lecture notes on ridge regression*. 2021. arXiv: [1509.09169 \[stat.ME\]](https://arxiv.org/abs/1509.09169).
- [15] Stephen J. Wright. “Coordinate Descent Algorithms”. In: *Mathematical Programming* 151 (1 June 2015), pp. 3–34. ISSN: 14364646. doi: [10.1007/s10107-015-0892-3](https://doi.org/10.1007/s10107-015-0892-3).