Homework: Regularization

November 23, 2023

This is the third assignment for the course 'Numerical Linear Algebra'. The goal of this assignment is to understand regularization as a method of solving linear systems. It will touch on ill-posed problems, standard regularization methods and methods of selecting the optimal regularization parameters. **Please note that the questions are deliberately open-ended. Unlike in the exercise sessions, you are encouraged to make your own analysis, provide your own results and construct your own report in your own way.** The assignment is individual. Don't be discouraged if not everything works in the end, but show your work and report what you tried and whether you know why something did not work. Your report should be a standalone text, should read pleasantly and should not refer to this assignment. **You are allowed to copy-paste from this assignment. Take special care of structure and visual presentation.**

1 Theory

The concept of ill-posedness was first studied by Jacques Hadamard in the beginning of the 20th century. Hadamard essentially defined a problem to be well-posed ("bien posé") if the solution is unique and if it is a continuous function of the data. Contrary, an ill-posed problem is one that is not uniquely solvable or not continuous as a function of the data i.e. if small perturbation of the data can cause large perturbations in the solution. The typical example of an ill-posed problem is a Fredholm integral equation of the first kind with a square integrable kernel

$$g(s) = \int_a^b K(s,t) f(t)\, dt, \quad c \le s \le d$$

with given K and g, where f is an unknown solution.

There are certain finite-dimensional discrete problems that have properties very similar to those of ill-posed problems, such as being highly sensitive to high frequency perturbations. Examples include the discretizations of the above integral equations. We can be more precise and say that a system of the form

$$A\mathbf{x} = \mathbf{b}, A \in \mathbb{R}^{m\times n}$$

is a discrete ill-posed problem if both the following conditions are satisfied:

1. the singular values of A decay gradually to zero
2. the ratio between largest and smallest singular values is large

Criterion 2 implies that the matrix A is ill-conditioned, while criterion 1 implies that there is no “nearby” problem with a well-conditioned coefficient matrix and with well-determined numerical rank.

Being ill-posed is not a fundamental barrier to systems being solvable. Rather, the ill-conditioning means that standard solution procedures are not useful. Indeed, the many small singular values of a discrete ill-posed problem essentially make the problem (numerically) under determined. One has to resort to more sophisticated methods, incorporating additional information about the solution, in order to compute a useful solution. This is the central idea behind regularization methods. When such ‘side constraints’ are introduced, one must give up the requirement $Ax = \mathbf{b}$ exactly in the linear system and instead seek a solution that provides a fair balance between minimizing some cost function that encodes the side constraints and minimizing the residual norm $\|A\mathbf{x} - \mathbf{b}\|_2$.

Most (but not all) of the theory and code needed for this assignment can at least in part be found in the `regtools` package and its accompanying manual, available at `https://www.mathworks.com/matlabcentral/fileexchange/52-regtools`.

1.1 Tikhonov regularization

Undoubtedly, the most common and well-known form of regularization is Tikhonov regularization. Here, the idea is to define the regularized solution $\mathbf{x}_\lambda$ (as a function of λ) as the minimizer of the following weighted combination of the residual norm and a side constraint

$$\mathbf{x}_\lambda = \arg\min_{\mathbf{x}}\{\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda^2\|L(\mathbf{x} - \mathbf{x}^*)\|_2^2\}, \tag{1}$$

where the regularization parameter λ controls the weight given to minimization of the side constraint relative to minimization of the residual norm. The side constraint is captured by the matrix L. One typical example is $L = I_n$. In this case, a large λ (strong regularization) favors a small solution norm at the cost of large residual norm, while a small λ has the opposite effect. Other choices for L are also possible. For example, L can be a discrete version of the second derivative operator. We can even arrange it so that

$$\mathbf{x}_\lambda = \arg\min_{\mathbf{x}}\{\lambda_0^2\|A\mathbf{x} - \mathbf{b}\|_2^2 + \sum_i \lambda_i^2\|L_i(\mathbf{x} - \mathbf{x}^*)\|_2^2\}. \tag{2}$$

In this homework $\mathbf{x}^* = 0$ is assumed. If the different L_i's are (discrete) derivatives, the rightmost term in 2 is called a Sobolev norm. **Give some examples of discrete derivatives of order 1 and 2. Take into account that discrete derivatives can be forward or backwards!**
These are not the only possible such L_i's. Another important class is seminorm matrices, which include norm constraints only on a subdomain of the solution and boundary condition restraints. **What would this look like? Give a simple example. Show that the expression 2 is equivalent to equation 1 with L the Cholesky factor (in Matlab convention) of $\sum_i \lambda_i^2 L_i^T L_i$. Can this be done? Be precise.** With λ one also controls the sensitivity of the regularized solution $\mathbf{x}_\lambda$ to perturbations in A and $\mathbf{b}$, and the perturbation bound is proportional to λ^{-1}. Numerical methods for actually computing an optimal λ will be discussed later. You have seen that in case of $L = I_n$, the solution is given by

$$\begin{aligned}\mathbf{x}_\lambda &= \sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2}\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i \\ &= \sum_{i=1}^n f_i\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i\end{aligned}$$

Use theorem 1 and the course screencast to show that in the case that L is not the identity we have

$$\mathbf{x}_\lambda = \sum_{i=1}^p f_i\frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{x}_i + \sum_{i=p+1}^n \mathbf{u}_i^T\mathbf{b}\,\mathbf{x}_i$$

with filter factors

$$f_i = \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2}$$

Theorem 1. *Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$ with $m \geq n \geq p$, there exists orthogonal $U \in \mathbb{R}^{m\times n}$ and $V \in \mathbb{R}^{p\times p}$ together with an invertible $X \in \mathbb{R}^{n\times n}$ such that*

$$A = U \begin{pmatrix} \Sigma & \\ & I_{n-p} \end{pmatrix} X^{-1}$$

and

$$B = V[M \;\; 0]X^{-1}$$

with both $\Sigma = diag(\sigma_1, \ldots, \sigma_p)$ and $M = diag(\mu_1, \ldots, \mu_p)$ diagonal matrices with nonnegative entries such that $0 \leq \sigma_1 \ldots \leq \sigma_p \leq 1 \geq \mu_1 \geq \ldots \geq \mu_p > 0$. The values $\gamma_i := \sigma_i/\mu_i$ are referred to as the generalized singular values. In addition we have $\mu_i^2 + \sigma_i^2 = 1$.

Mind the difference in ordering of the σ's to the usual ordering! Similarly to the usual discrete Picard condition, the more general case for L also leads to a Picard condition, but now in terms of the generalized singular values. That is, the discrete Picard condition in case of $L \neq I_n$ is satisfied whenever the $|u_i^T b|$ (with U from the GSVD) decay at least as fast as the generalized singular values.

1.2 TSVD/TGSVD and DSVD/DGSVD regularization

A fundamental observation regarding the Tikhonov method is that it circumvents the ill-conditioning of A by introducing a new problem with a well-conditioned coefficient matrix with full rank (**which matrix?**). A different way to treat the ill-conditioning of A is to derive a new problem with a well-conditioned rank deficient coefficient matrix (well-conditioned meaning that the nonzero singular values span a limited range). A fundamental result about rank deficient matrices is the Eckart-Young theorem. It states that the closest rank-k approximation A_k to A (measured in 2-norm) is obtained by truncating the SVD expansion at k, i.e., A_k is given by

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

The truncated SVD (TSVD) regularization method is based on this observation in that one solves the problem

$$\min \|\mathbf{x}\|_2 \quad s.t. \quad \min \|A_k\mathbf{x} - \mathbf{b}\|.$$

The solution to this problem is given by

$$\mathbf{x}_k = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i$$

Let us note that the TSVD solution $\mathbf{x}_k$ is the only solution that has no component in the numerical null-space of A, spanned by columns of V with numbers from $k+1$ to n. Instead of using filter factors 0 and 1 as in TSVD, one can introduce a smoother cut-off by means of filter factors f_i defined as

$$f_i = \frac{\sigma_i}{\sigma_i + \lambda}$$

thus getting the damped SVD (DSVD). The new filter factors decay slower than Tikhonov filter factors and thus introduce less filtering. We can extend these to the generalized context as well. In this case we have for the truncated GSVD (TGSVD) that

$$\mathbf{x}_k = X \begin{pmatrix} \hat{\Sigma}_k & \\ & I_{n-p} \end{pmatrix} U^T \mathbf{b}$$

in which $\hat{\Sigma}_k = \operatorname{diag}(0, \ldots, 0, \sigma_{p-k+1}^{-1}, \ldots, \sigma_p^{-1})$. **Show that if $L = I$ the TSVD solution and TGSVD solution coincide. Hint: test your argument on the provided matrix `Test.mat`, using `cgsvd` from `regtools` and `svd` from Matlab. Verify that all your claims hold.** Again, we can dampen this by, rather than using filter factors 0 and 1, using the filter factors

$$f_i = \frac{\gamma_i}{\gamma_i + \lambda}.$$

1.3 Conjugate gradient regularization

The conjugate gradient method is an iterative Krylov method for the solution of linear systems. It is only defined for symmetric positive semi-definite systems, but it can in our case be applied to the normal equations

$$A^T A \mathbf{x} = A^T \mathbf{b}$$

the solution of which will be called $\mathbf{x}_*$. **You can assume that $A^T A$ is positive definite.** This is given in algorithm 1. Typically, $\mathbf{x}_0 = 0$ is assumed. **If you find it simpler, you can restrict your derivations to this case. Analyze the algorithm (briefly). In particular, show that $\mathbf{r}_i = \mathbf{b} - A\mathbf{x}_i$. You might need this later...** Some notation:

Algorithm 1: Conjugate Gradient Algorithm for the normal equations (CGLS)

```
input : Starting vector $\mathbf{x}_0$, number of iterations $k$
output: CG solution $\mathbf{x}_k$
init compute initial residual $\mathbf{r}_0 := \mathbf{b} - A\mathbf{x}_0$ and initial auxiliary vector $\mathbf{d}_0 := A^T\mathbf{r}_0$
for $i = 1 \dots k$ do
    $a_i = \|A^T\mathbf{r}_{i-1}\|_2^2 / \|\mathbf{d}_{i-1}\|_A^2$
    $\mathbf{x}_i = \mathbf{x}_{i-1} + a_i\mathbf{d}_{i-1}$
    $\mathbf{r}_i = \mathbf{r}_{i-1} - a_iA\mathbf{d}_{i-1}$
    $\beta_i = \|A^T\mathbf{r}_i\|_2^2 / \|A^T\mathbf{r}_{i-1}\|_2^2$
    $\mathbf{d}_i = A^T\mathbf{r}_i + \beta_i\mathbf{d}_{i-1}$
end
```

- $\mathbf{r}_i^{LS} = A^T\mathbf{r}_i$
- $\mathcal{K}_i(A^TA, \mathbf{r}_0^{LS}) = \operatorname{span}\{\mathbf{r}_0^{LS}, A^TA\mathbf{r}_0^{LS}, \dots, (A^TA)^{i-1}\mathbf{r}_0^{LS}\}$ is the Krylov subspace associated to the CGLS iteration i.
- $\mathbf{e}_i := \mathbf{x}_* - \mathbf{x}_i$
- $\langle \mathbf{v}, \mathbf{w} \rangle_A := \langle A\mathbf{v}, A\mathbf{w} \rangle$, and the norm $\|\mathbf{v}\|_A$ and orthogonality $\mathbf{v} \perp_A \mathbf{w}$ are defined correspondingly.

Now the conjugate gradient iteration can be described completely as the system of recurrences that generates the (unique) sequence of iterates satisfying $\{\mathbf{x}_i \in \mathbf{x}_0 + \mathcal{K}_i(A^TA, \mathbf{r}_0^{LS})\}_i$ such that at step i, $\|\mathbf{e}_i\|_A$ is minimized. **You do not have to prove this, but you can certainly try. Do show that $\|\mathbf{e}_i\|_A$ being minimized in this case is equivalent to $\mathbf{e}_i \perp_A \mathcal{K}_i(A^TA, \mathbf{r}_0^{LS})$. Show that $\mathbf{r}_i^{LS} = A^*A\mathbf{e}_i$. Use this to show that $\mathbf{e}_i \perp_A \mathcal{K}_i(A^TA, \mathbf{r}_0^{LS})$ is also equivalent to $\mathbf{r}_i^{LS} \perp \mathcal{K}_i(A^TA, \mathbf{r}_0^{LS})$. Also use it to show the following two simple lemmas:**

Lemma 1. *Let $\mathbb{P}_{i,1}$ denote the space of polynomials of degree (at most) i with constant coefficient 1. For the CGLS iteration i it then holds that*

$$\mathbf{e}_i = q(A^TA)\mathbf{e}_0$$

with $q \in \mathbb{P}_{i,1}$.

Lemma 2. *For any polynomial p and at any iterate i we have*

$$\mathbf{e}_i = p(A^TA)\mathbf{e}_0 \iff \mathbf{r}_i^{LS} = p(A^TA)\mathbf{r}_0^{LS}$$

Believe it or not, with these two lemmas you are now ready to prove the following important theorem:

Theorem 2. *At CGLS iteration k (at the end of the CGLS procedure) it holds that*

$$\mathbf{x}_* - \mathbf{x}_k = R_k(A^TA)(\mathbf{x}_* - \mathbf{x}_0)$$

with the Ritz polynomial R_k *given by*

$$R_k(t) := \prod_{j=1}^{k} \frac{\theta_j^{(k)} - t}{\theta_j^{(k)}},$$

in which $\theta_j^{(k)}$ denotes the jth Ritz value of A^TA i.e. the eigenvalues of $Q_k^* A^T A Q_k$, with Q_k the orthogonal basis of $\mathcal{K}_k(A^TA, \mathbf{r}_0^{LS})$.

Prove this. Use that if θ is a zero of the polynomial $p(t)$, then $p(t) = (\theta - t)\tilde{p}(t)$, with $\tilde{p}$ one degree lower. In matrix polynomial form this reads as $p(t) = (\theta I_k - A^TA)\tilde{p}(A^TA)$. Hint 1: Can you find a vector in your equations that can be written as $Q_k z$ for some z? Hint 2: When are two polynomials that share the same zeros equal?
Conclude from theorem 2 that the filter factors for $\mathbf{x}_k$ obtained by the CG method (looking at the normal equations!) are given by

$$f_i = 1 - R_k(\sigma_i^2)$$

with σ_i the ith singular value of A. Use this to explain the regularizing effect of the CG method. What is the regularization parameter? Note that the CG method does not include an L matrix! You should also be extremely careful with CG since the convergence can be delayed due to round-off errors in finite precision arithmetic.

2 Methods of choosing the regularization parameter

In this section we outline two methods for selecting the regularization parameter. Many of these are implemented in the package `regtools`, so you don't need to implement them. **Only for the CG method parameter selection methods are not implemented in `regtools`!**

2.1 The L-curve

The L-curve criterion is already known to you from the theory. One is always interested in finding the 'corner' of the L-curve. Explain why in your report. For a continuous regularization parameter λ one may compute the curvature of the curve $(\log \|A\mathbf{x}_\lambda - \mathbf{b}\|_2, \log \|L\mathbf{x}_\lambda\|_2)$ (with λ the regularization parameter) and seek the point with maximum curvature, which is defined as L-curve corner. **You could do this automatically, which is what `regtools` does, but you don't have to do this. Simply write code that generates the L-curve and find the optimal parameter by studying the resulting graph. Note that even though we have used the notation λ for the regularization parameter, which suggests a continuum, an L-curve is equally well-defined for discrete regularization parameters!**

2.2 Generalized cross-validation

GCV is based on the philosophy that if an arbitrary element $\mathbf{b}_i$ of the right-hand side $\mathbf{b}$ is left out, then the corresponding regularized solution should predict this observation well, and the choice of regularization parameter should be independent of an orthogonal transformation of $\mathbf{b}$. This (through some complex theory that you do not have to worry about) leads to choosing the regularization parameter, which minimizes the GCV function

$$G = \frac{\|A\mathbf{x}_{\text{REG}} - \mathbf{b}\|_2^2}{(\text{trace}(I_m - AA^I))^2}$$

where A^I is a matrix which produces the regularized solution $\mathbf{x}_{REG}$ when multiplied with $\mathbf{b}$, i.e. $\mathbf{x}_{\text{REG}} = A^I\mathbf{b}$. For instance, in the case of Tikhonov regularization (with $L = I_m$) we have that $A^I = VF\Sigma^{-1}U^T$ (you can easily check this) when $A = U\Sigma V^T$ is the singular value decomposition of A and F is the corresponding diagonal matrix of filter factors at the parameter λ. Note that G is defined for both continuous and discrete regularization parameters. **Show the following, more general statement:**

Lemma 3. *We have that trace*$(I_m - AA^I) = m - (n - p) - \sum_{i=1}^{p} f_i$

What are n and p for CG? Assume $x_0 = 0$ and A^TA positive definite.

3 Application: the watchmaker and the bowl of soup

In this section we introduce a slightly silly application of the regularization methods described.

3.1 Context

Imagine you are a watchmaker, specializing in wristwatches. As such you have very sensitive hands and fingertips. In addition, you cannot allow them to be burned. Today, your partner has made you a bowl of soup, heated in the microwave and set it down at your workbench. You notice that the bowl is ceramic, which means it is probably very hot. Contrary, you deduce from the lack of steam that the soup is room temperature. Unfortunately, there is no cloth around that is thick enough to protect your hands from the scorching ceramic over the long trip back to the microwave. Fortunately though, you are skilled in integral equations and you can tell temperatures using your fingers with an accuracy of about $1e-3$ degrees. And so you put your finger in the soup, close to the bowl, but not touching, and measure the temperature as a function of time. You use this to determine the temperature of the bowl by solving *the inverse heat equation.*

3.2 The inverse heat equation

We model this as the inverse of the following heat equation:

$$\begin{aligned}\frac{\partial}{\partial t}T &= \frac{\partial^2}{\partial^2 x}T \\ T(x,0) &= 0 \\ T(0,t) &= f(t)\end{aligned}$$

for $0 < x < \infty$ and $0 \leq t < \infty$. Here, the temperature $T = 0$ actually corresponds to room temperature, that is 20 degrees Celsius. The origin ($x = 0$) corresponds to the bowl. The variable t in our example is measured in minutes. By inverse we mean: while the above calculates T given the boundary condition f , we wish to compute f from the observed $T(y,t) := g(t)$ at a distance y from the origin. It is known that this is given by

$$Kf = g$$

with

$$(Kf)(t) = \frac{y}{2\sqrt{\pi}} \int_0^t \frac{f(\xi)}{(t-\xi)^{3/2}} \exp(\frac{-y^2}{4(t-\xi)})\, d\xi.$$

This is a Volterra integral equation, but can be transformed into a Fredholm equation by extension. It is thus immediately suspect: the discretization of this problem is probably ill-posed.

3.3 Provided data

You have been provided a matrix K, obtained by the mid-point rule discretization of the operator K. In addition you are given an exact solution f and a perturbed (random normally distributed noise, $\sigma = 1e-3$) right-hand-side $\mathbf{g}$. Both are given over a time range of 10 minutes.

3.4 Extra information about the solution

Since regularization is all about using information about the solution, I will now provide some information that you can use. You can verify this by looking at the exact solution given. **Not every piece of information needs to be used, or might even be helpful!**

(1) The solution is not oscillatory. This can be achieved by requiring that its second derivative is small

(2) The solution is eventually zero. In fact, if $t > 5$, f is negligible

(3) The temperature is always positive i.e. $f > 0$

(4) Initially, the temperature is stationary, that is, $g'(0) = 0$

3.5 What you have to do [1]

Download from the course Toledo the mat-file (`Temp.mat`) with a pregenerated matrix K, perturbed right-hand side $\mathbf{g}$ and solution $\mathbf{f}$. You are not given the exact right-hand side. Load your file in Matlab. Inspect the problem. Is it Ill-posed? What is its condition number? As a reference for comparison later, solve the system using Matlab's `mldivide`. Connect your analysis to a pleasing graphical representation. Apply the regularization methods:

[1] This section is to be interpreted as entirely in bold font

(1) Tikhonov: classical ($L = I_m$) and advanced $L \neq I_m$. Try to incorporate information about the solution into L. If you use Sobolev and/or semi-norm constraints, compare some choices and interpret/report on the best one.

(2) SVD based: TSVD, DSVD, TGSVD, DGSVD. Again, for the generalized version, incorporate additional information about the solution.

(3) Conjugate gradient based. Here you do not have to incorporate additional information about the solution.

Compare L-curve parameter selection and GCV parameter selection for all methods. Hint: it is normal that some combinations of regularization-parameter selection do not give good results. If you encounter one or more combinations that do not work, try to explain why.

4 Quotation and questions

Points are awarded based on the correctness of your results and the quality of your code (40%), the insight demonstrated in your report (40%) and the organization and presentation of your report (20%). Don't hesitate to email me at simon.dirckx@kuleuven.be if anything is unclear or if you suspect something is wrong. If you are stuck on something you can always ask for a hint. This will be taken into account in the evaluation but will not severely impact your grade.