MATH3027: Optimization 2022

Week 3: Least squares

Prof. Richard Wilkinson
Please send any comments or mistakes to r.d.wilkinson@nottingham.ac.uk

This week we explore a very important class of problems with different applications in maths, science, and engineering. Least Squares problems frequently arise when we want to fit a model to match observations/data, in part because assuming Gaussian errors leads to consideration of the sum of squares. After a brief historical presentation, we will formulate the classical linear least squares problem and its solution through the normal equations. We will explore connections to data fitting, and introduce a regularization term to cast an optimization problem that is central in signal denoising applications. Finally, we will discuss the formulation of nonlinear least squares problems.

A Bit of History

In January 1, 1801, the Italian monk Giuseppe Piazzi, discovered a faint, nomadic object through his telescope in Palermo, correctly believing it to reside in the orbital region between Mars and Jupiter. Piazzi watched the object for 41 days but then fell ill, and shortly thereafter the wandering star strayed into the halo of the Sun and was lost to observation. The newly-discovered planet had been lost, and astronomers had a mere 41 days of observation covering a tiny arc of the night from which to attempt to compute an orbit and find the planet again.

Figure 1: Giusepe Piazzi (1746-1826), Italian priest, mathematician, and astronomer. His observations led to the discovery of the dwarf planet Ceres.

The dean of the French astrophysical establishment, Pierre-Simon Laplace, declared that the orbit recovery simply could not be done. In Germany, the 24 years old German mathematician Car Friedrich Gauss had considered that this type of problem, to determine a planet's orbit from a limited handful of observations- "commended itself to mathematicians by its difficulty and elegance." Gauss discovered a method for computing the planet's orbit using only three of the original observations and successfully predicted where Ceres might be found (now considered to be a dwarf planet). The prediction catapulted him to worldwide acclaim.

More than 200 years later, in 2019 American computer scientist Katie Bouman used similar mathematical methods, which heavily rely on optimization and large-scale astronomical observation datasets, to recover the first image of a black hole.

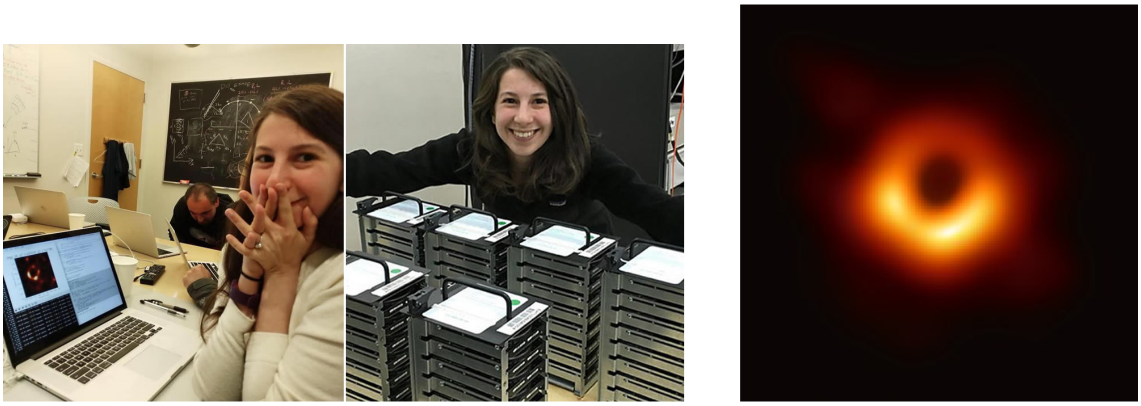

Figure 2: 200 years later, Katie Bouman (1990-) used similar optimization techniques as Gauss (and a bit of supercomputing power) to recover the first image of a black hole.

Formulating the Linear Least Squares Problem

Consider the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \qquad \left(\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^{m}\right)$$

If $m > n$, i.e., we have more equations than unknowns, we say the system is *over-determined*; when $m < n$ it is *under-determined.* Usually, when a system is over-determined, the equations will be inconsistent and no solution will exist. In this situation, a common approach for finding an approximate solution[1] is to choose the solution of the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2\,, \tag{LLS}$$

which is the same as (check!)

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + \|\mathbf{b}\|^2 \right\}.$$

Differentiating gives

$$\nabla f = 2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b} \quad \text{and } \nabla^2 f = 2\mathbf{A}^\top\mathbf{A},$$

and therefore, the optimal solution $\mathbf{x}_{LS}$, is a solution of $\nabla f(\mathbf{x}) = 0$, namely,

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{x}_{\text{LS}} = \mathbf{A}^T\mathbf{b} \leftarrow \text{ the normal equations.}$$

The *normal equations* are also a linear system, but now it is a square system as $\mathbf{A}^\top\mathbf{A}$ is $n \times n$. If $\mathbf{A}^\top\mathbf{A}$ is invertible, which is the case if and only if $\mathbf{A}^\top\mathbf{A} \succ 0$, then we have the unique global (why?) minimum solution

$$\mathbf{x}_{\text{LS}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}\,.$$

A Numerical Example. Consider the inconsistent linear system

$$\begin{aligned} x_1 + 2x_2 &= 0\\ 2x_1 + x_2 &= 1\\ 3x_1 + 2x_2 &= 1 \end{aligned}$$

To find the least squares solution, we will solve the normal equations:

$$\begin{pmatrix} 1 & 2\\ 2 & 1\\ 3 & 2 \end{pmatrix}^T \begin{pmatrix} 1 & 2\\ 2 & 1\\ 3 & 2 \end{pmatrix} \begin{pmatrix} x_1\\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2\\ 2 & 1\\ 3 & 2 \end{pmatrix}^T \begin{pmatrix} 0\\ 1\\ 1 \end{pmatrix}$$

[1] We can always formulate the problem in this way. If the equations are consistent, then the unique solution to $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is the solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$.

which is the same as

$$\begin{pmatrix} 14 & 10 \\ 10 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \Rightarrow \mathbf{x}_{LS} = \begin{pmatrix} 15/26 \\ -8/26 \end{pmatrix} .$$

Note that $\mathbf{A}\mathbf{x}_{LS} = (-0.038; 0.846; 1.115)$, so that the errors are

$$\mathbf{b} - \mathbf{A}\mathbf{x}_{LS} = \begin{pmatrix} 0.038 \\ 0.154 \\ -0.115 \end{pmatrix} \Rightarrow \text{ sq. error } = 0.038^2 + 0.154^2 + (-0.115)^2 = 0.038 \,.$$

What if $\mathbf{A}^\top\mathbf{A}$ is not invertible, i.e., $rank(\mathbf{A}^\top\mathbf{A}) = r < n$? In this case, $\mathbf{A}^\top\mathbf{A}$ has r strictly positive eigenvalues, with the remaining $n - r$ eigenvalues equal to 0. Recall that as $\mathbf{A}^\top\mathbf{A}$ is symmetric, we can diagonalize it as

$$\mathbf{A}^\top\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{u}_i^\top$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix with $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$ (the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$, which form an orthonormal basis for $\mathbb{R}^n$), and $\mathbf{D} = \text{diag}(d_1, \ldots, d_r, 0, 0, \ldots)$ an $n \times n$ diagonal matrix.

We can use the *pseudo-inverse*[2] of $\mathbf{A}^\top\mathbf{A}$, which is defined to be

$$\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top = \sum_{i=1}^{r} \frac{\mathbf{u}_i \mathbf{u}_i^\top}{d_i}$$

where $\mathbf{D}^{-1} := \text{diag}(\frac{1}{d_1}, \ldots, \frac{1}{d_r}, 0, 0, \ldots)$, to solve the normal equations. This gives

$$\mathbf{x}_{LS} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{A}^\top\mathbf{b}.$$

But note that $\sum_{i=r+1}^{n} \alpha_i \mathbf{u}_i$ is in the null space of $\mathbf{A}^\top\mathbf{A}$, and so we have an $n - r$ dimensional space of solutions

$$\mathbf{x} = \mathbf{x}_{LS} + \sum_{i=r+1}^{n} \alpha_i \mathbf{u}_i$$

all of which minimize $||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$.

Theorem. *If $rank(\mathbf{A}^\top\mathbf{A}) = r < n$, then for any $\alpha_{r+1}, \ldots, \alpha_n \in \mathbb{R}$,*

$$\mathbf{x} = \mathbf{x}_{LS} + \sum_{i=r+1}^{n} \alpha_i \mathbf{u}_i$$

is a minimizer of $||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$. The solution with minimum norm is $\mathbf{x}_{LS}$.

Proof. See the exercises. □

[2] $\mathbf{B}^+$ is a pseudo-inverse of $\mathbf{B}$ if $\mathbf{B}^+\mathbf{B}\mathbf{B}^+ = \mathbf{B}^+$ and $\mathbf{B}\mathbf{B}^+\mathbf{B} = \mathbf{B}$. If we in addition have $\mathbf{B}^+\mathbf{B} = \mathbf{B}\mathbf{B}^+ = I$ then $\mathbf{B}^+ = \mathbf{B}^{-1}$.

Data Fitting

We revisit, from an optimization viewpoint, a fundamental problem in statistics, namely, linear regression. Assume we have a dataset $(\mathbf{s}_i, t_i), i = 1, 2, \ldots, m$, where $\mathbf{s}_i \in \mathbb{R}^n$ and $t_i \in \mathbb{R}$. Assume that an approximate linear relation holds:

$$t_i \approx \mathbf{s}_i^T \mathbf{x}, \quad i = 1, 2, \ldots, m\,.$$

The corresponding least squares/regression problem reads

$$\min_{\mathbf{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left(\mathbf{s}_i^T \mathbf{x} - t_i \right)^2 ,$$

or equivalently

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{S}\mathbf{x} - \mathbf{t}\|_2^2\,.$$

where

$$S = \begin{pmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \vdots \\ \mathbf{s}_m^T \end{pmatrix} \in \mathbb{R}^{m \times n}, \ \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{pmatrix} \in \mathbb{R}^{m \times 1}.$$

That is, we want to pick the best $\mathbf{x}$ (in terms of minimizing the sum of squares) to predict t from $\mathbf{s}$[3].

Polynomial Fitting

In the previous section, we assumed the relationship between input ($\mathbf{s}$) and response (t) was linear. But non-linear relationships can also be cast as linear least-squares problems. For example, consider fitting a polynomial relationship given a set of points in $\mathbb{R}^2$: $(u_i, y_i), i = 1, 2, \ldots, m$, for which the following approximate relation holds for some $a_0, \ldots, a_d$:

$$\sum_{j=0}^{d} a_j u_i^j \approx y_i, \quad i = 1, \ldots, m$$

which in vector notation is

$$\mathbf{u}_i^\top \mathbf{a} \approx y_i$$

where $\mathbf{u}_i^\top = (1, u_i, u_i^2, \ldots, u_i^d)$, and $\mathbf{a} = (a_0, \ldots, a_d)$.

[3] Note we usually use very different notation in statistics:

$$\min_{\beta \in \mathbb{R}^n} \|X\beta - y\|^2$$

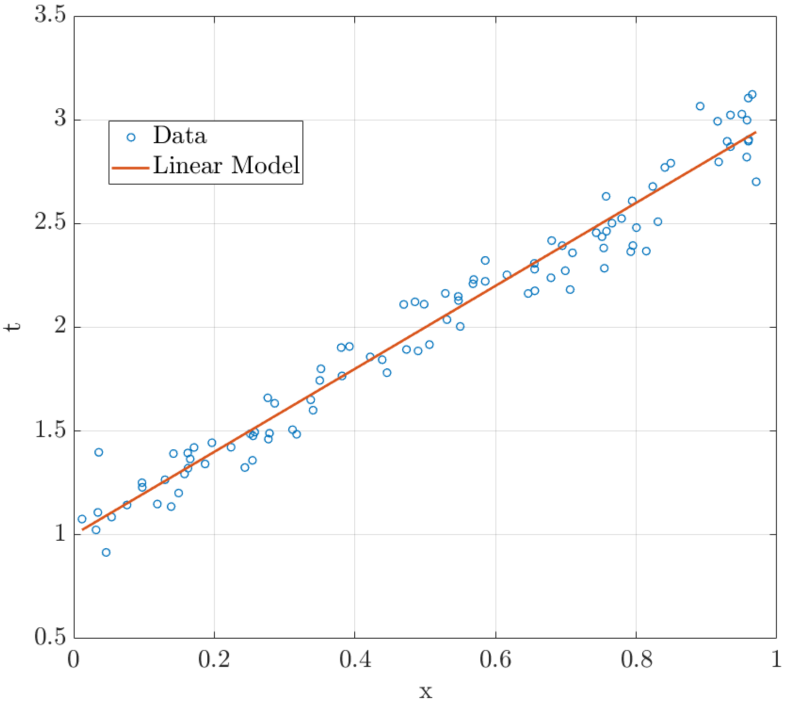

Figure 3: The typical situation in linear regression. A set of scattered data (s_i, t_i) (in blue) suggests a linear relation between s and t. Among all the possible linear models, there is an optimal choice (in red) which minimizes the sum of squared errors between the model and the measurements.

The associated linear system is:

$$\overset{\mathbf{U}}{\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix}} \overset{\mathbf{a}}{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix}} \overset{\approx}{\approx} \overset{\mathbf{y}}{\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}}$$

The least squares solution is well defined if the $m \times (d+1)$ matrix $\mathbf{U}$ is of full column rank. This is true when all the u_i 's are different from each other.

Regularized Least Squares

There are several situations in which the least squares solution does not give rise to a good estimate of the "true" vector $\mathbf{x}$. In these cases, a regularized problem (called regularized least squares (RLS)) is often solved:

$$(\mathbf{RLS}) \qquad \min_{\mathbf{x}} \|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda R(\mathbf{x})\,.$$

Here, λ is the regularization parameter and $R(\cdot)$ is the regularization function (also called a penalty function). A common choice is a quadratic regularization function:

$$\min \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{Dx}\|_2^2 \,.$$

The optimal solution of the above problem is (why?)

$$\mathbf{x}_{\text{RLS}} = \left(\mathbf{A}^T\mathbf{A} + \lambda \mathbf{D}^T\mathbf{D}\right)^{-1} \mathbf{A}^T\mathbf{b} \,.$$

What kind of assumptions are needed to assure that $\mathbf{A}^T\mathbf{A} + \lambda\mathbf{D}^T\mathbf{D}$ is invertible (when $\lambda > 0$)? (answer: $\text{Null}(\mathbf{A}) \cap \text{Null}(\mathbf{D}) = \{\mathbf{0}\}$)[4].

Note that since the Hessian $\mathbf{A}^T\mathbf{A} + \lambda\mathbf{D}^T\mathbf{D} \succeq 0$, any stationary point is a global minimum. The case where $\mathbf{D} = \mathbf{I}$ is called *ridge regression* in statistics, or *Tikhonov regularization* in applied mathematics.

Denoising

A very important application of linear least squares and regularization techniques is the denoising of signals (acoustic, images). Suppose that a noisy measurement of a signal $\mathbf{x} \in \mathbb{R}^n$ is given by:

$$\mathbf{b} = \mathbf{x} + \mathbf{w} \,,$$

where $\mathbf{x}$ is the “true” unknown signal, $\mathbf{w}$ is the unknown noise and $\mathbf{b}$ is the (known) observation vector. Note that the least squares problem:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|_2^2 \,,$$

is meaningless, as we would trivially recover $\mathbf{x} = \mathbf{b}$, without any denoising. We need to enrich the optimization problem by adding a suitable regularization term, exploiting some *a priori* information about the signal. For example, if we know that the signal is “smooth" in some sense, then $R(\cdot)$ can be chosen as a penalization of “sudden variations” in the signal

$$R(\mathbf{x}) = \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 \ ,$$

where the regularization $R(\cdot)$ can also be written as $R(\mathbf{x}) = \|\mathbf{Lx}\|^2$ where $\mathbf{L} \in \mathbb{R}^{(n-1)\times n}$ is given by

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} .$$

[4] here $\text{Null}(\mathbf{A})$ refers to the nullspace of the matrix or kernel, $\text{Ker}(\mathbf{A}) := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{0}\}$.

The resulting regularized least squares problem is

$$\min_{\mathbf{x}} \underbrace{\|x-b\|^2}_{\text{fitting}} + \underbrace{\lambda\|Lx\|^2}_{\text{denoisig}} .$$

The direct solution of this problem leads to (why?)

$$\mathbf{x}_{\text{RLS}}(\lambda) = \left(\mathbf{I} + \lambda\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{b} .$$

See Figures 4 and 5 for an example, and to see how the choice of λ affects the reconstructed curve.

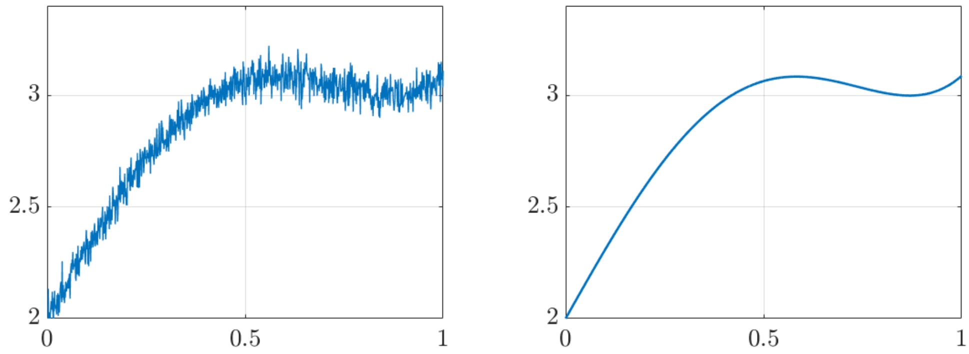

Figure 4: Signal denoising. We only have access to the noisy signal (left), and we would like to recover a “clear” signal (right) by solving a regularized least squares problem.

Nonlinear Least Squares

The least squares problem $\min \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is often called linear least squares. In some applications we are given a set of nonlinear equations:

$$f_i(\mathbf{x}) \approx b_i, \quad i = 1, 2, \ldots, m .$$

The nonlinear least squares (NLS) problem is the one of finding an $\mathbf{x}$ solving the problem

$$\min_{\mathbf{x}} \sum_{i=1}^{m} (f_i(\mathbf{x}) - b_i)^2 \tag{NLS}$$

possibly with the addition of a regularization term as above. In contrast to linear least squares, there is no easy way to solve NLS problems. However, there are some dedicated algorithms for this problem, which we will explore later in the module.

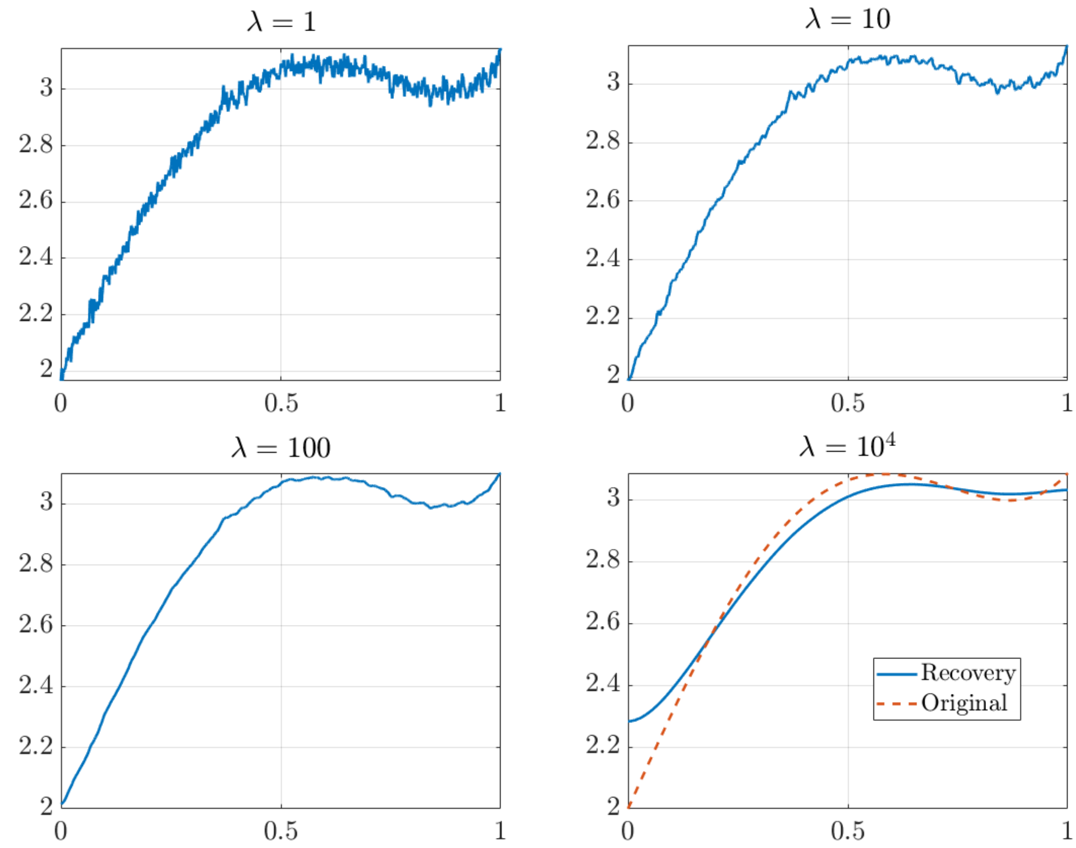

Figure 5: Signal denoising for different regularization parameters. Increasing the regularization parameters leads to better denoising, however, as the parameter becomes too large, the fit with the original signal is lost. What is the limit as λ grows?

Checklist

The idea of this checklist is to help you to self-evaluate your progress and understanding of the subject, and to give you some guidance on where to focus. If you can tick all the boxes it means you're doing alright, otherwise you need to study a bit more, grab a book, watch the videos, or seek help from classmates, the lecturers, or the demonstrators. Try to fill as many gaps as quickly as possible.

And remember to do the E.1's!

Learning Outcome	Check
I understand how an overdetermined linear system can be solved as an optimization problem.	
I understand derivation of optimality conditions for LLS.	
I can solve LLS by hand for small systems.	
I understand the relation between LLS and linear regression.	
I understand the role of a regularization parameter.	
I can redo the optimality conditions for Regularized Least Squares.	
I understand the application of RLS to signal denoising.	
I understand that LLS can be extended to nonlinear models	

Exercises

1. Prove that $\mathbf{A}^\top\mathbf{A} \succeq 0$. When is $\mathbf{A}^T\mathbf{A}$ positive definite?
2. What is the core numerical task for solving linear least squares problem? And when might it be computationally difficult to perform this task?
3. Show that $\sum_{i=1}^{r} \frac{\mathbf{u}_i\mathbf{u}_i^\top}{d_i}$ is a pseudo-inverse of $\mathbf{A}^\top\mathbf{A}$, where $d_i, \mathbf{u}_i$ are eigenpairs of $\mathbf{A}^\top\mathbf{A}$.
4. Let $\lambda_i, \mathbf{u}_i$ for $i = 1, \dots, n$ be the eigenpairs of $\mathbf{A}^\top\mathbf{A}$ with $\mathbf{A}^\top\mathbf{A}\mathbf{u}_i = \lambda_i\mathbf{u}_i$ and $\|\mathbf{u}_i\| = 1$. Prove that if $rank(\mathbf{A}^\top\mathbf{A}) = r < n$, then for any $\alpha_{r+1}, \dots, \alpha_n \in \mathbb{R}$,

   $$\mathbf{x} = \mathbf{x}_{LS} + \sum_{i=r+1}^{n} \alpha_i\mathbf{u}_i$$

   is a minimizer of $||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$.

   Show that the solution with minimum norm is $\mathbf{x}_{LS}$.
5. Consider the inconsistent linear system

   $$\begin{aligned} x_1 + x_2 &= 0 \\ 2x_1 + 2x_2 &= 1 \end{aligned}$$

   Characterize all the ‘solutions’ of this system and find the minimum norm solution.
6. Suppose we have a noisy set of scales, that we know give unbiased measurements, but which have a zero-mean error that is distributed as a Gaussian random variable with variance σ^2. We have three objects we wish to weigh, which have true (unknown) weights w_1, w_2, w_3. We use the scales four times to weight firstly just object 1, then object 1 and 2, then object 2 and 3, and finally all three objects. The four recorded weights are

   $$y_1 = 1.04, \quad y_2 = 2.73 \quad y_3 = 6.06 \quad y_4 = 5.15.$$

   What is the maximum likelihood estimate of the 3 unknown weights? Recall that the pdf of a Gaussian random variable is

   $$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(y - \mu)^2\right).$$
7. A commonly used regularizer is to use the L_1 norm, i.e., to minimise

   $$\min ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \lambda||\mathbf{D}\mathbf{x}||_1^2 .$$

   Why might it be difficult to solve this problem?

8. Consider the sum of squares

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{x})^2.$$

Show that

$$\nabla f(\mathbf{x}) = J(\mathbf{x})^\top \mathbf{r}(\mathbf{x}) \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = J(\mathbf{x})^\top J(\mathbf{x}) + \sum_{j=1}^{n} r_j(\mathbf{x}) \nabla^2 r_j(\mathbf{x})$$

where

$$J(\mathbf{x}) = \begin{pmatrix} \nabla r_1(\mathbf{x})^\top \\ \vdots \\ \nabla r_n(\mathbf{x})^\top \end{pmatrix} \text{ and } \mathbf{r}(\mathbf{x}) = \begin{pmatrix} r_1(\mathbf{x}) \\ \vdots \\ r_n(\mathbf{x}) \end{pmatrix}$$

Check this simplifies to the linear case when $r_i(\mathbf{x}) = \mathbf{a}_i^\top \mathbf{x} - y_i$.