Inverse Problems

Dirk Lorenz

Summer term 2022

Contents

Introductory remarks 2

1 Introduction and motivation 4

2 Examples and basic notions 10

3 Hilbert spaces 14

4 The singular value decomposition and the pseudo-inverse 18

5 Regularization 23

6 Tikhonov regularization 30

7 Spectral regularization 37

8 Parameter choice and error estimates 42

9 Convergence rates and smoothness spaces 47

10 Convergence rates for spectral regularization 52

11 Iterative regulariztion 58

12 A Bayesian perspective on regulariztion 65

13 Discretization by projection 70

Introductory remarks

These are the lecture notes for the lecture "Inverse Problems" I held in the summer term 2023 at TU Braunschweig. The lecture is aimed at students from mathematics, financial mathematics, computational science and engineering as well as data science. Solid knowledge in analysis and linear algebra is needed. Helpful would be a background in functional analysis, especially the notions of Hilbert space and linear operators, but we will also provide a little background on these topics in the course.

Inverse problems are problems where one want to find some cause which can only measured indirectly, i.e. one can only observe the effects, but not the cause itself. Hence, it is quite applied as a math topic, but still one can do serious mathematics and the charm of the field lies in the tight connection of theoretical results and real world applications.

The books on inverse problems one can find also vary from quite theoretical to very applied. Here is a short commented list of books:

1. The books Engl u. a. [1996]; Rieder [2013] are on the theoretical side of inverse problems. While they also present applications, they focus on the underlying theory. The latter one (Rieder [2013]) is in German and written in the style of a textbook. The more recent lecture notes Clason [2020] are also quite theoretical.

2. The older Groetsch [1993] is also written as a text book, focuses on theory, but targets readers with less background in mathematics.

3. The book Moura Neto und da Silva Neto [2012] uses less mathematics and also targets students with less background in math. It also does not focus that much on theory.

Braunschweig, June 15, 2023

Dirk Lorenz
d.lorenz@tu-braunschweig.de

[Clason 2020] Clason, Christian: *Regularization of inverse problems.* arXiv preprint arXiv:2001.00617. 2020

[Engl u. a. 1996] Engl, Heinz W. ; Hanke, Martin ; Neubauer, Andreas: *Regularization of inverse problems.* Bd. 375. Springer Science & Business Media, 1996

[Groetsch 1993] Groetsch, Charles: *Inverse problems in the mathematical sciences.* Bd. 52. Springer, 1993

[Moura Neto und da Silva Neto 2012] Moura Neto, Francisco D. ; Silva Neto, Antônio José da: *An introduction to inverse problems with applications.* Springer Science & Business Media, 2012

[Rieder 2013] Rieder, Andreas: *Keine Probleme mit inversen Problemen: eine Einführung in ihre stabile Lösung.* Springer-Verlag, 2013

1 Introduction and motivation

The central topic of this lecture are inverse and ill-posed problems. Both the terms "inverse" and "ill-posed" are not clearly defined up to now (and will be hard to pin down exactly). Instead of defining them right away, we start with a motivating example.

Example 1.1 (Differentiation). The problem of finding the derivative g' of a given function g is quite straightforward as long as symbolic computations are considered. However, finding the slope of a function that is not given as analytic expression, but can only be evaluated through a black box, is more involved. We will show, that it is even inverse and ill-posed in some sense.

Mathematically, we would like to invert the operator A which takes a function f (for simplicity defined on $[0,1]$) to its integral, i.e. $Af = g$ with

$$Af(x) := g(x) := \int_0^x f(t)\mathrm{d}t.$$

Hence, the task is: Given some g, find f such that $Af = g$.

We would like to measure errors in the data g and also in the reconstruction f and hence, we introduce norms for these quantities. We use the following norms:

$$\begin{aligned}\|f\|_C &:= \|f\|_\infty := \max\{|f(x)| \mid x \in I\}\\ \|f\|_{C^1} &:= \|f\|_\infty + \|f'\|_\infty\\ \|f\|_{C^k} &:= \sum_{l=0}^{k} \|f^{(l)}\|_\infty.\end{aligned}$$

In fact, these norms turn the appropriate vector spaces into normed spaces: For an interval I let

$$\begin{aligned}C(I) &:= \{f : I \to \mathbb{R} \mid f \text{ continuous}\},\\ C^1(I) &:= \{f : I \to \mathbb{R} \mid f \text{ continuously differentiable}\},\\ C^k(I) &:= \{f : I \to \mathbb{R} \mid f \text{ } k\text{-times continuously differentiable}\}.\end{aligned}$$

We may omit the argument I and just write C and C^k if I is clear from the context or does not play a role. We also denote $C^0 = C$ which is consistent with the notation for C^k. These norms induce a notion of convergence: We say that $f_n \to f$ in C^k if $\|f_n - f\|_{C^k} \to 0$. Convergence in C is exactly uniform convergence and convergence in C^k means that the functions as well as their first k derivatives converge uniformly.

We can model the operator A as a map between various spaces, e.g. we can write $A : C([0,1]) \to C([0,1])$. Differentiation is a left inverse of A: We have that $DAf(x) = D(\int_0^x f(t)\mathrm{d}t) = f(x)$. In this sense, the problem of calculating the derivative is the inverse problem to calculating the integral.

It's not a right inverse, since $ADf(x) = \int_0^x f'(x)\mathrm{d}x = f(x) - f(0)$.

Now let us argue that the inverse problem of differentiation is ill-posed while the *direct* problem of integration is well posed: The map $A : C([0,1]) \to C([0,1])$ is linear and bounded. Linearity follows from the known rules for integrals and boundedness is

seen as follows: For a function $f \in C$ we have

$$\begin{aligned}\|Af\|_C &= \max\left\{ |\int_0^x f(t)\mathrm{d}t| \mid 0 \le x \le 1 \right\} \\ &\le \max\left\{ \int_0^x |f(t)|\mathrm{d}t \mid 0 \le x \le 1 \right\} \le \max|f(x)| = \|f\|_C.\end{aligned}$$

This shows that A is bounded and even that the operator norm of A fulfills $\|A\| \le 1$.

How about continuity of the inverse operation? Consider g with $g(0) = 0$ and $g' = f$ (i.e, $Af = g$) and let us perturb g slightly to

$$g^\delta(x) = g(x) + \delta \sin(nx)$$

for some $\delta > 0$ and $n \in \mathbb{N}$. Then we have

$$(g^\delta)'(x) = g'(x) + \delta n \cos(nx) =: f^\delta(x).$$

Hence, we have $Af = g$ and $Af^\delta = g^\delta$. Moreover, we easily see that

$$\begin{aligned}\|g - g^\delta\|_C &= \delta \\ \|f - f^\delta\|_C &= \max\{\delta n \cos(nx) \mid 0 \le x \le 1\} = n\delta.\end{aligned}$$

If we couple $\delta = 1/\sqrt{n}$ we get

$$\|g - g^\delta\| = \delta = \tfrac{1}{\sqrt{n}} \overset{n\to\infty}{\longrightarrow} 0, \quad \text{but} \quad \|f - f^\delta\|_C = n\delta = \sqrt{n} \overset{n\to\infty}{\longrightarrow} \infty.$$

This shows that small perturbations in g may lead to large perturbations in the derivative $f = g'$, and hence, taking the derivative is unstable, and thus ill-posed.

At first, this may seem like a hopeless situation when is comes to numerical differentiation. We always have some round-off error, so does that mean that numerical differentiation can not work?

Assume that $g : [0,1] \to \mathbb{R}$ is a differentiable function, but we don't have a formula for g but for every x we get the value $g(x)$. How can we find $g'(x)$? The simplest idea may be to use a central difference quotient

$$\frac{g(x+h) - g(x-h)}{2h} =: D_h g(x).$$

This operator D_h is again a linear operator, and, in some sense, an approximation of the derivative D. Using Taylor expansion, we get for $g \in C^2$ that

$$g(x \pm h) = g(x) \pm hg'(x) + \tfrac{g''(\xi_\pm)}{2} h^2$$

with some $\xi_\pm$ between x and $x+h$ and $x-h$, respectively. This gives us an estimate

$$\|g' - D_h g\|_\infty = \max\left\{ |g'(x) - \left(g'(x) + \tfrac{h}{2} \tfrac{g''(\xi_+) - g''(\xi_i)}{2} \right)| \right\} \le \tfrac{h}{2} \|g''\|_\infty$$

This is an estimate for the error we make when we use an approximation of the derivative, and hence, we call this error the *approximation error*.

Now we assume that we have an error in our available data, i.e. we do not get the exact values $g(x)$, but slightly perturbed data

$$g^\delta(x) = g(x) + w(x) \quad \text{with} \quad \|w\|_\infty \leq \delta.$$

We are interested in the *total error*, i.e. for $f = g'$ we would like to compute or estimate

$$\|f - D_h g^\delta\|_\infty,$$

which is the error between the unknown exact derivative f and the quantity $D_h g^\delta$ we can actually compute. This error has a natural decomposition as

$$\|f - D_h g^\delta\|_\infty \leq \|f - D_h g\|_\infty + \|D_h g - D_h g^\delta\|_\infty$$

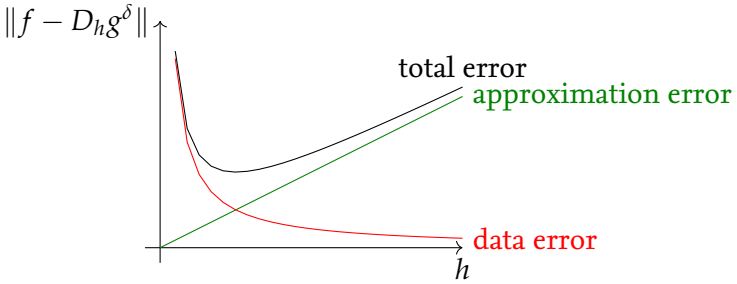
where we inserted the term $\pm D_h g$ and used the triangle inequality. The first term on the right is the approximation error which we just estimated already. The second term is called *data error* and is also simple to estimate:

$$\|D_h g - D_h g^\delta\|_\infty = \|D_h w\|_\infty = \max\left\{ \left| \tfrac{w(x+h)-w(x-h)}{2h} \right| \right\} \leq \tfrac{\delta}{h}.$$

So we get for the total error

$$\|f - D_h g^\delta\|_\infty \leq h\tfrac{\|g''\|_\infty}{2} + \tfrac{\delta}{h}.$$

Overall, we see a very typical behavior:



For a fixed *noise level* δ, there is a tradeoff between large and small parameters h: For small h the data error gets big, while for large h the approximation error gets big. Somewhere in the middle there is an optimum and a little calculus shows that the parameter h which minimizes our upper bound for the total error is

$$h^* = \sqrt{\tfrac{2\delta}{\|g''\|_\infty}}.$$

With this value we get

$$\|f - D_{h^*} g^\delta\|_\infty \leq \sqrt{2\|g''\|_\infty \delta}.$$

We see that even when the operation of differentiation is ill-posed, we can still get a stable approximation of the derivative from noisy data. However, the error is not as small as one could have hoped: We obtain an error in the order of $\sqrt{\delta}$ for noise of size δ. △

It is quite instructive, to see that higher smoothness can lead to a better total error: Assume that g''' exists, derive a better estimate for the approximation error (by using more terms of the Taylor expansion, calculate the optimal h and deduce that the total error can be of order $\delta^{2/3}$ in this case.

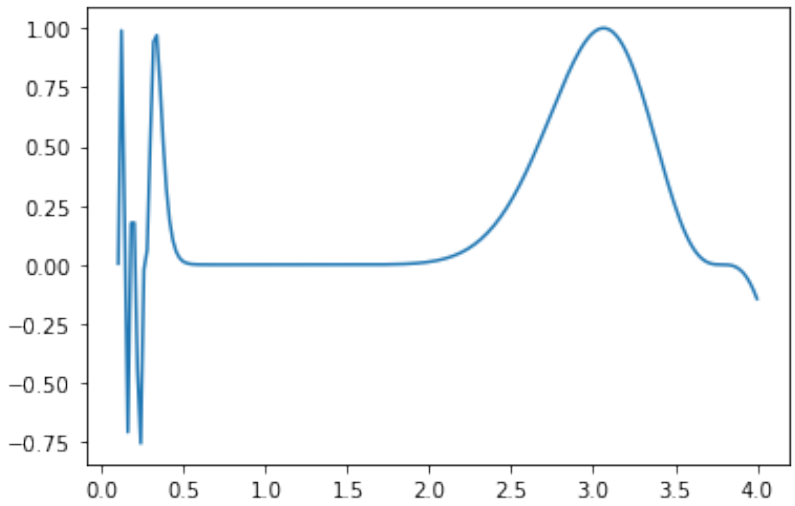
Here are a few important takeaways from the above example:

- The total error in the solution of an inverse problem decomposes into an approximation error and a data error: Good approximation leads to a amplification of the error in the data, and keeping the data error small needs a large approximation error.
- For a fixed noise level δ there is a tradeoff between approximation error and data error.
- A helpful estimate of the approximation error needs a smoothness assumption on the unknow data (in our case we needed that $g'' = f'$ exists and is bounded).
- The total error is, even in the best case, not of the order of the error in the data, but worse.

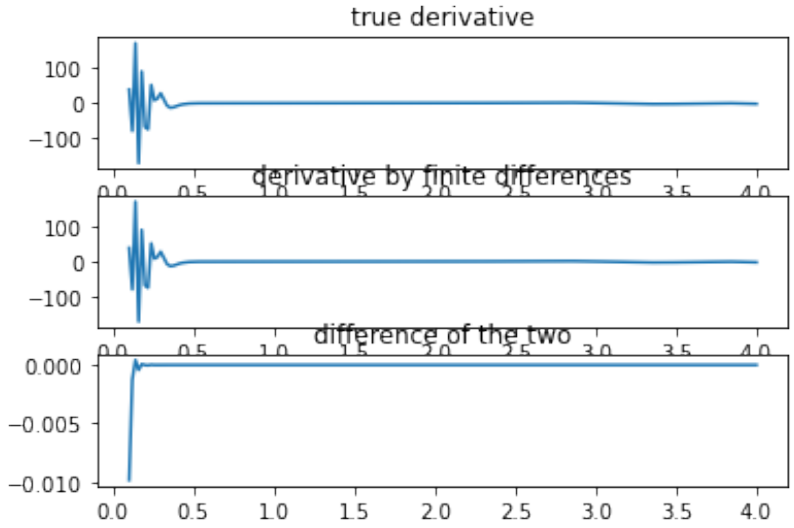
Remark 1.2. Our results is actually useful in practice: If you want to evaluate derivatives of functions numerically by a finite difference approximation, one usually uses $h = \sqrt{\text{eps}}$ where eps is the machine precision. For double precision numbers eps $\approx 10^{-16}$, so $h = 10^{-8}$ is recommended.

Here is an example (written in Python):

```
# import libraries
import numpy as np
import matplotlib.pyplot as plt

# define functions
def f(x):
    return np.sin(np.log(x)**4)**3


def fprime(x):
    return 3*np.sin(np.log(x)**4)**2*np.cos(np.log(x)**4)*4*np.log(x)**3/x


def Df(x,h):
    return (f(x+h)-f(x-h))/2/h

# some x values
x = np.linspace(0.1,4,200,dtype='float64')
# plot function
plt.plot(x,f(x))
plt.show()
```

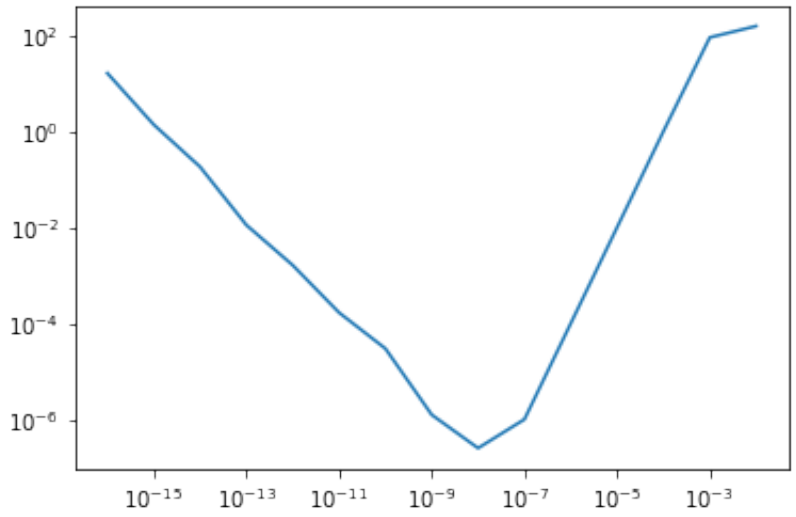



```
# Choose some stepsize
h = 1e-5
# plot derivative and approximation by finite differences
fig, axs = plt.subplots(3)
axs[0].plot(x, fprime(x))
axs[0].set_title('true derivative')
axs[1].plot(x, Df(x,h))
axs[1].set_title('derivative by finite differences')
axs[2].plot(x, fprime(x)-Df(x,h))
axs[2].set_title('difference of the two')
plt.show()
```




```
# compare derivative and approximation for different values of h
h = np.logspace(-2,-16,15)
e = np.zeros(15)
for k in range(0,15):
```

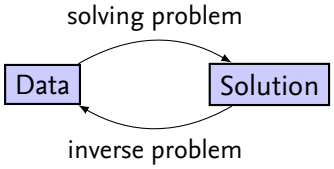
```
    e[k] = np.max(np.abs(fprime(x)-Df(x,h[k])))

plt.loglog(h,e)
plt.show()

# square root of machine precision is close to optimal h:
print('sqrt(eps) = ', np.sqrt(np.finfo(x.dtype).eps))
```




```
sqrt(eps) =  1.4901161193847656e-08
```

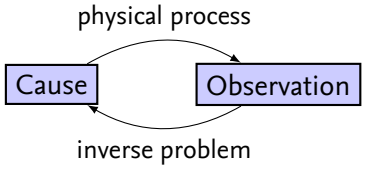
2 Examples and basic notions

The notion of "inverse problems" is vague and hard to pin down. What is a "problem" anyway? Well, a problem always has "data" and "solution", i.e. something that is given, and something that is wanted. Solving the problem means, taking the data, do some computations and arrive at a solution. For every problem, there is an "inverse problem", namely: Having some solution, what is the corresponding data that gave rise to this solution?



In a more physical context one could frame inverse problems as follows:

> An inverse problems asks for some *cause* that is behind a given *observation*.



Example 2.1 (Parameter identification in PDEs). Assume that you can observe the heat distribution u of some matter in a domain Ω at a given time $t = T > 0$. What was the heat distribution at time $t = 0$? This is an inverse problem. To formulate it mathematically, we use the heat equation. The distribution of heat follows the partial differential equation

$$\begin{aligned} u_t(t,x) &= \Delta u(t,x) \quad \text{in } [0,T] \times \Omega \\ u(0,x) &= u_0(x) \quad \text{in } \Omega \\ \partial_n u(t,x) &= 0 \quad \text{in } [0,T] \times \partial\Omega. \end{aligned}$$

The forward problem would be: Given the initial data u_0, calculate $u(T,x)$. The corresponding inverse problem is: Given a measurement of $u(T,x)$, find initial data u_0 that explains the measurement.

In the context of partial differential equations one can formulate numerous inverse problems. Consider the following problem:

$$\begin{aligned} u_t - Lu &= f \quad \text{in } [0,T] \times \Omega \\ u(0,x) &= u_0(x) \quad \text{in } \Omega \\ \partial_n u &= g \quad \text{on } [0,T] \times \partial\Omega \end{aligned}$$

with some differential operator L, initial data u_0, source term f, and boundary data g. The forward problem would be to compute

u from knowledge of u_0, f and g, but there are various inverse problems and here are just two:

- Given $u(T,\cdot)$, f and g, find u_0.
- Given u, and u_0, find f and g.

You can find more inverse problems easily. △

Example 2.2 (Computerized tomography). The basic concept of CT is to measure the intensity of X-ray beams from a source with known intensity after passing through the body at a fixed plane. It is assumed that these beams travel on a straight line and their intensity is attenuated proportially to some material constant one is interested in reconstructing, which is usually the density. Denoting this material constant by u and $x = x(L,t)$ the point in which the X-ray beam associated with the line L passes at time t, this can be modeled by the ordinary differential equation

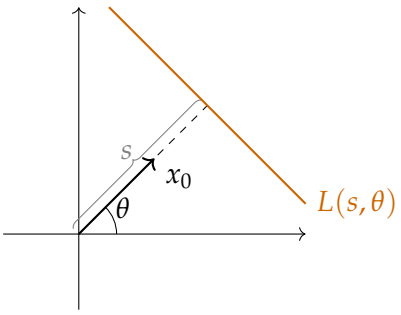
$$I_L'(t) = -u(x(L,t))I_L(t),$$

so if $I_L(T)$ is measured for some time $T > 0$ where the beam passed the object of interest,

$$-\log\Big(\frac{I_L(T)}{I_L(0)}\Big) = \int_0^T u(x(L,t))\mathrm{d}t.$$

The left-hand side is known while the right-hand side constitutes the integral of u associated with the line L up to some factor. The principle of computed tomography is obtain these integrals by emitting and measuring X-ray beams along all possible lines. This is typically done by placing an X-ray point source on one side of the object, installing a detector array on the other and side rotating source and detector simultaneously, giving the intensities for all lines passing through a region of interest.

Mathematically, the reconstruction problem in CT is now to compute u given all of its line integrals. Placing the origin in the center of the object of interest of radius $R > 0$, a line L passing through the domain can uniquely be associated an angle $\theta \in [-\pi/2, \pi/2[$ and offset $s \in \mathbb{R}$ at which the line crosses the axis spanned by $x_0(\theta) = (\cos(\theta), \sin(\theta))$.



We denote the corresponding line by $L = L(s,\theta)$. Given $u^0 :]-R,R[\times [-\pi/2,\pi/2[\to \mathbb{R}$, the CT reconstruction problem is to find $u^\dagger : \mathbb{R}^2 \to \mathbb{R}$ such that

$$(Ru)(s,\theta) := \int_{L(s,\theta)} u^\dagger \mathrm{d}x = u^0(s,\theta) \qquad \text{for all} \quad (s,\theta)$$

where the mapping R is called the *Radon transform*. The inverse problem related to the Radon transform is the main problem in tomographic reconstruction. △

Here is the classical definition of "ill posedness" which is due to Hadamard and dates back to the 1920s.

Definition 2.3 (Well- and ill-posed problems)**.** Let X and Y be topological spaces and $A : X \to Y$. We say that the problem "solve $Ax = y$" is *well posed* if

(a) The equation $Ax = y$ has a solution for every $y \in Y$, and

(b) this solution is unique, and

(c) the solution depends continuously on the data.

In short: the problem is well posed if the inverse $A^{-1} : Y \to X$ exists and is continuous.

If one of these conditions is not fulfilled, we call the problem *ill-posed*.

In this lecture we will treat *linear* inverse problems. We will always assume *Hilbert spaces* X and Y and a linear, bounded operator $A : X \to Y$.The space X is the *solution space* and Y is the *data space*. The forward problem is "given x, evaluate Ax". Since inverse problems always assume measurement data with error, the inverse problem is:

We will denote the set of linear and bounded operators from X to Y by $L(X,Y)$.

> Given measured data $g^\delta \in Y$ which fulfills
>
> $$\|Af^\dagger - g^\delta\| \leq \delta$$
>
> for some known *noise level* δ and an *unknown true solution* $f^\dagger$, find a good approximation to $f^\dagger$.

For linear operators one can characterize the ill-posedness of the problem "Solve $Ax = y$" quite explicitly:

(a) $Ax = y$ has a solution for every y exactly if A is surjective (also called onto), i.e. if $\mathrm{rg}(A) = Y$.

(b) Solutions of $Ax = y$ are unique exactly if A is injective (also called one-to-one), i.e. if $\ker(A) = \{0\}$

(c) Solutions of $Ax = y$ depend continuously on y exactly if A^{-1} is bounded.

A more quantitative way to describe ill-posedness of a problem is the notion of *condition* or conditioning of a problem. We say that a problem is

well condtioned if small changes in the data lead to small changes in the solution

ill conditioned (or badly conditioned) if small changes in the data lead to large errors in the solution.

Example 2.4 (Function evaluation). Here the problem is simply "given x evaluate $y = f(x)$. We consider a perturbation $x + \Delta x$ of x. The solution changes to $y + \Delta y = f(x + \Delta x)$ and by linearization we get

$$|\Delta y| \approx |f'(x)||\Delta x|.$$

So we say that the problem is ill-conditioned (with respect to absolute errors) if $|f'(x)|$ is large.

If we consider relative errors, we get

$$\frac{|\Delta y|}{|y|} \approx \frac{|f'(x)||\Delta x|}{|f(x)|} = \frac{|f'(x)||x|}{|f(x)|}\frac{|\Delta x|}{|x|}$$

so we say that the problem is ill-conditioned with respect to relative error if $|f'(x)||x|/|f(x)|$ is large. △

Example 2.5 (Solving linear equations). Consider an invertible square matrix A and the problem: given b, find the solution of $Ax = b$. Changing the data to $b + \Delta b$, the new solution fulfills

$$A(x + \Delta x) = b + \Delta b.$$

We see that the change in the solution is $\Delta x = A^{-1}\Delta b$. Using the operator norm we get that $\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|$ and we see that the problem is ill-conditioned (with respect to absolute errors) if $\|A^{-1}\|$ is large. If we consider relative errors, we get (using $\|b\| = \|Ax\| \leq \|A\|\|x\|$

$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}\Delta b\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta b\|}{\|x\|}\frac{\|b\|}{\|b\|} \leq \|A^{-1}\|\|A\|\frac{\|\Delta b\|}{\|b\|}.$$

△

Definition 2.6. The condition number of a square matrix A is

$$\operatorname{cond}(A) = \begin{cases} \|A\|\|A^{-1}\| & : \text{ if } A \text{ is invertible} \\ \infty & : \text{ else.} \end{cases}$$

Strictly speaking, the problem "Solve $Ax = b$" is only ill-posed if A is not invertible. Practically, a large condition number of A will still lead to a large increase of the error, so we consider the problem still ill conditioned or badly conditioned if the condition number is large.

To understand all these things better, we introduce the notion of Hilbert spaces, linear bounded operators between these spaces and the singular value decomposition of these operators.

3 Hilbert spaces

Now we introduce abstract notions of vector spaces that we will use throughout the lecture. The main notions are *Hilbert spaces*. In a nutshell, they are spaces which behave the sae as the euclidean space $\mathbb{R}^n$ when it comes to geometric and analytical structures like length, and orthogonality.

First we define inner products:

Definition 3.1. A real *inner product space* X is a real vector space that is equipped with an *inner product* $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ which fulfills

$$\begin{aligned} \langle x, y \rangle &= \langle y, x \rangle \\ \langle \alpha x + y, z \rangle &= \alpha \langle x, z \rangle + \langle y, z \rangle \\ \langle x, x \rangle &> 0 \quad \text{if } x \neq 0. \end{aligned}$$

An inner product always induces a norm $\|x\| = \sqrt{\langle x, x \rangle}$ (which can be shown using the Cauchy-Schwarz inequality $\langle x, y \rangle \leq \|x\| \|y\|$).Since any norm induces a notion of convergence by saying that $x_n \overset{n\to\infty}{\longrightarrow} x$ if $\|x_n - x\| \to 0$ we can also talk about *completeness* and this is the property that turns inner product spaces into Hilbert spaces.

On a Hilbert space X we will denote the norm by $\|x\|_X$, but sometimes the subscript may be dropped, when the norm is clear from the context.

Definition 3.2. A real *Hilbert space* is a complete real inner product space, i.e. a real inner product space with the property that every Cauchy sequence in the space converges, i.e. for every sequence x_n in X which fulfills

$$\forall \epsilon > 0 \, \exists N \, \forall m, n \geq N : \|x_n - x_m\| \leq \epsilon$$

there exists some limit x^*, i.e. $x_n \overset{n\to\infty}{\longrightarrow} x^*$ (i.e. we have $\|x_n - x^*\| \overset{n\to\infty}{\longrightarrow} 0$).

A little bit sloppy one could say that a Hilbert space is an inner product space that "you can't leave with 'convergent sequences'"

Example 3.3.

1. The space $\mathbb{R}^n$ with standard inner product

$$\langle x, y \rangle := x \cdot y := x^T y = \sum_{i=1}^{n} x_i y_i$$

is a Hilbert space of dimension n. The corresponding norm $\|x\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$ is the well known euclidean norm. The standard inner product on $\mathbb{R}^d$ is also called *dot product* and we will also use the notation $x \cdot y$ for it. The euclidean norm of a vector x will also be denoten by $|x|$.

2. An example of an infinite dimensional Hilbert space is

$$\ell^2 := \left\{ (x_n)_{n=1,2,\ldots} \;\middle|\; \sum_{i=1}^{\infty} x_n^2 < \infty \right\}$$

of square summable sequences. When equipped with the inner product $\langle x, y\rangle := \sum_{i=1}^{\infty} x_i y_i$ it is a Hilbert space. The corresponding norm is denoted by

$$\|x\|_2 := \Big(\sum_{i=1}^{\infty} x_i^2\Big)^{1/2}.$$

3. A different example of an infinite dimensional Hilbert space is the *Lebesgue space* L^2 of square integrable functions. For some domain $\Omega \subset \mathbb{R}^d$ (i.e. a non-empty, connected and open subset of $\mathbb{R}^d$) we can define

$$L^2(\Omega) := \left\{ f : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega f(x)^2 \mathrm{d}x < \infty \right\}$$

and equipped with the inner product

$$\langle f, g\rangle_{L^2} := \int_\Omega f(x) g(x) \mathrm{d}x$$

this is a Hilbert space as well. The corresponding norm is denoted by

$$\|f\|_{L^2} := \Big(\int_\Omega f(x)^2 \mathrm{d}x\Big)^{1/2}.$$

The are Sobolev spaces of higher order, i.e. for every $k \in \mathbb{N}$ there is a Sobolev space $H^k(\Omega)$ which incorporates partial derivatives up to k-th order, but we will not define them here.

You may have noted that there is a problem here: This is not a norm, since there are functions $f \neq 0$ with $\|f\|_{L^2(\Omega)} = 0$. This problem can be solved: These functions have the property of being zeros "almost everywhere" and one can "quotient out" all these functions from $L^2(\Omega)$. This means that one identifies two functions f and g if $\int (f-g)^2 = 0$. The full theory behind this can be found in books on real anaylsis or measure theory.

4. Slightly more complicated are the *Sobolev spaces* which generalize the Lebesgue space L^2 by also incorporating derivatives. The space $H^1(\Omega)$ is

$$H^1(\Omega) := \left\{ f : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega f(x)^2 \mathrm{d}x + \int_\Omega |\nabla f(x)|^2 \mathrm{d}x < \infty \right\}$$

and it is equipped with the inner product

$$\langle f, g\rangle_{H^1} := \int_\Omega f(x) g(x) \mathrm{d}x + \int_\Omega \nabla f(x) \cdot \nabla g(x) \mathrm{d}x$$

where the second integral is over the dot product of the gradients. The corresponding H^1-norm is

$$\|f\|_{H^1} := \left(\int_\Omega f(x)^2 \mathrm{d}x + \int_\Omega |\nabla f(x)|^2 \mathrm{d}x \right)^{1/2}$$

△

As important as the spaces are linear operators between these spaces. We will call a linear map A from one Hilbert space X to another Y an *operator* and say that an operator A is *bounded* if there exists a constant C such that $\|Ax\|_Y \leq C\|x\|_X$ for all x.The infimum over all such constants is the *operator norm* of A, denoted by $\|A\|$. Other ways to define the operator norm are

Bounded operators are exactly the continuous operators.

$$\|A\| = \sup_{\|x\|_X=1} \|Ax\|_Y = \sup_{x\neq 0} \frac{\|Ax\|_Y}{\|x\|_X}.$$

The set of all bounded linear operators from a Hilbert space X to another Hilbert space Y is denoted by $L(X,Y)$. For every $A \in L(X,Y)$ we have the so-called *adjoint* operator A^* defined by

$$\forall x \in X, y \in Y : \langle x, A^*y\rangle = \langle Ax, y\rangle .$$

Note that the adjoint maps $A : Y \to X$ and it is also bounded with $\|A^*\| = \|A\|$. In $\mathbb{R}^n$, operators are just matrices and the adjoint with respect to the standard inner product is just the transpose of the matrix.

Example 3.4.

1. Linear operators $A \in L(\mathbb{R}^n, \mathbb{R}^m)$, i.e. from one euclidean space to another can be identified with matrices $A \in \mathbb{R}^{m\times n}$, i.e. the application of the operator A to a vector x is given by matrix-vector multiplication Ax. In this case these is a simple expression for the operator norm of a matrix: It holds that

   $$\|A\| = \sqrt{\lambda_{\max}(A^TA)},$$

   i.e. the operator norm is the square root of the largest eigenvalue of the matrix A^TA. Note that A^TA is symmetric and positive definite and thus, only has real and eigenvalues greater or equal zero. The adjoint of matrix is given by the transposed matrix as it holds that

   $$(Ax)\cdot y = (Ax)^T \cdot y = x^TA^Ty = x^T(A^Ty) = x\cdot(A^Ty).$$

2. A class of linear operators between L^2 spaces $A \in L(L^2(\Omega_1), L^2(\Omega_2))$ is given by so called *integral operators*. These are given by

   $$Af(y) = \int_{\Omega_1} k(x,y)f(x)\mathrm{d}x$$

   for some function $k : \Omega_1 \times \Omega_2 \to \mathbb{R}$. For these operators to be well defined and bounded one needs that k is square integrable, i.e. $k \in L^2(\Omega_1 \times \Omega_2)$. This can be seen by an application of the Cauchy-Schwarz inequality for the L^2 inner product as follows:

   $$\|Af\|_{L^2}^2 = \int_{\Omega_2} \left(\int_{\Omega_1} k(x,y)f(x)\mathrm{d}x \right)^2 \mathrm{d}y.$$

The inner integral is the L^2 inner product between $k(x,\cdot)$ and f and hence, we have

$$\begin{aligned}\|Af\|_{L^2}^2 &\leq \int_{\Omega_2} \left(\int_{\Omega_1} k(x,y) f(x) \mathrm{d}x \right)^2 \mathrm{d}y \\ &\leq \int_{\Omega_2} \int_{\Omega_1} k(x,y)^2 \mathrm{d}x \int_{\Omega_1} f(x)^2 \mathrm{d}x \mathrm{d}y \\ &= \|k\|_{L^2(\Omega_1 \times \Omega_2)}^2 \|f\|_{L^2(\Omega_1)}^2.\end{aligned}$$

This also gives the upper bound $\|A\| \leq \|k\|_{L^2(\Omega_1 \times \Omega_2)}$ (which is usually not strict).

We compute the adjoint of an integral operator using Fubinis theorem to interchange the order of integrals

$$\begin{aligned}\langle Af, g \rangle_{L^2(\Omega_2)} &= \int_{\Omega_2} Af(y) g(y) \mathrm{d}y \\ &= \int_{\Omega_2} \int_{\Omega_1} k(x,y) f(x) \mathrm{d}x g(y) \mathrm{d}y \\ &= \int_{\Omega_1} f(x) \underbrace{\int_{\Omega_2} k(x,y) g(y) \mathrm{d}y}_{= A^* g(x)} \mathrm{d}x = \langle f, A^* g \rangle .\end{aligned}$$

△

There is another result that we will use frequently:

Theorem 3.5 (Dominated convergence for series)**.** *Let $a_{m,n} \in \mathbb{R}$ with $a_{m,n} \overset{m \to \infty}{\longrightarrow} a_n^*$. If there exists a sequence b_n with $|a_{m,n}| \leq b_n$ for all m, n and $\sum_n b_n < \infty$, then it holds that*

$$\sum_n a_{m,n} \overset{m \to \infty}{\longrightarrow} \sum_n a_n^*.$$

Informally: If we have a sequence of series, where the summands converge, we can pull the limit under the series if there is a dominating sequence which is summable.

4 The singular value decomposition and the pseudo-inverse

We will build the section on the spectral theorem for compact operators. Recall the notion of compact operator:

Definition 4.1. Some $A \in L(X, Y)$ is *compact*, if it holds that

$$(x_n) \text{ bounded in } X \implies (Ax_n) \text{ has convergent subsequence in } Y.$$

We will denote the set of compact operators from X to Y by $K(X, Y)$. Directly from the definition we get that $K(X, Y) \subset L(X, Y)$.

Here are some equivalent descriptions of compact operators for those who know what weak convergence in Hilbert spaces mean:

1. A is compact if it maps bounded sets in X to precompact sets in Y (i.e. their closure is compact). (For bounded operators we only have that they map bounded sets to bounded sets.)
2. A is compact if it maps weakly convergent subsequences to strongly convergent ones, i.e. Ax_n is (strongly) convergent in Y whenever x_n is weakly convergent in X. (For bounded operators we only have that they map strongly convergent sequences to strongly convergent sequences and weakly convergent sequences to weakly convergent sequences.)

Example 4.2. For $A \in L(X, Y)$ we can say:

1. If the range of A is finite dimensional and A is bounded, then A is compact, as bounded sequences in finite dimensional spaces always have convergent subsequences.
2. The identity $\mathrm{id} : X \to X$ is always bounded, but only compact if X is finite dimensional.
3. If A is compact and B is bounded than AB is compact (if defined). Similarly, BA is compact (if defined).
4. Also the adjoint operator A^* is compact if A is compact.
5. Finally, if K_n is a sequence of compact operators and we have that $\|K_n - K\| \to 0$, then K is compact as well.

△

A class of non-trivial compact operators are *integral operators*:

Example 4.3. Let $X = L^2(\Omega_1)$ and $Y = L^2(\Omega_2)$ and $k \in L^2(\Omega_1 \times \Omega_2)$ and define the operator

$$Kx(t) = \int_{\Omega_1} k(s, t)x(s)\mathrm{d}s.$$

We have seen in Example 3.4 that K is bounded with operator norm $\|K\| \leq \|k\|_{L^2(\Omega_1 \times \Omega_2)}$.

However, K is also compact! To see this, note that we can approximate the function k by "simple functions", i.e. there is sequence k_n of functions of the form

$$k_n(s, t) = \sum_{i,j} \alpha_{ij} \mathbb{1}_{E_i}(s) \mathbb{1}_{F_j}(t)$$

for some disjoint sets $E_i \subset \Omega_1$ and $F_j \subset \Omega_2$ such that k_n approximated k, more precisely $\|k - k_n\|_{L^2(\Omega_1 \times \Omega_2)} \to 0$. The respective integral operators K_n all have finite dimensional range and hence, are compact. Moreover it holds that

Here we use $\mathbb{1}_E$ for the so-called *characteristic function* of the set E, i.e. the function which is 1 on E and zero elsewhere.

$$\|K - K_n\| \leq \|k - k_n\|_{L^2(\Omega_1 \times \Omega_2)} \to 0$$

and thus, K is compact as well. △

One central theorem for compact operators is the following:

Theorem 4.4 (Spectral theorem for compact, selfadjoint operators)**.** *Let X be a Hilbert space and $K \in K(X, X)$ be selfadjoint. Then there exists and orthonormal basis (u_n) of $\operatorname{cl}\operatorname{rg}(K)$ and $\lambda_n \in \mathbb{R} \setminus \{0\}$ such that*

$$Kx = \sum_n \lambda_n \langle x, u_n \rangle u_n.$$

If the dimension of $\operatorname{cl}\operatorname{rg}(K)$ is infinite, we also have $\lambda_n \to 0$.

The proof can be found in H.W. Alt's book "Linear functional analysis" where this theorem is Theorem 12.12.

Remark 4.5. Plugging in $x = u_m$, we get that $Ku_m = \sum_n \lambda_n \langle u_m, u_n \rangle u_n = \lambda_m u_m$ and we see that the u_n are actually eigenvectors of K for eigenvalues λ_n. By convention, one sorts the eigenvalues by decreasing magnitude, i.e. $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$.

From the spectral theorem we can deduce the existence of the singular value decomposition (SVD):

Theorem 4.6 (Singular value decomposition)**.** *For every $K \in K(X, Y)$ there exist*

(i) an orthonormal basis (u_n) of $\operatorname{cl}\operatorname{rg}(K) \subset Y$,

(ii) an orthonormal basis (v_n) of $\operatorname{cl}\operatorname{rg}(K^*) \subset X$,

(iii) numbers $\sigma_1 \geq \sigma_2 \geq \cdots > 0$

such that for all n

$$Kv_n = \sigma_n u_n, \quad \text{and} \quad K^* u_n = \sigma_n v_n$$

and for all $x \in X$

$$Kx = \sum_n \sigma_n \langle x, v_n \rangle u_n.$$

Proof. Since K^*K is selfadjoint and compact, we get from the spectral theorem the existence of λ_n and v_n such that

$$K^*Kx = \sum_n \lambda_n \langle x, v_n \rangle v_n.$$

Since $\lambda_n \|v_n\|_X^2 = \langle \lambda_n v_n, v_n \rangle = \langle K^*Kv_n, v_n \rangle = \langle Kv_n, Kv_n \rangle = \|Kv_n\|_X^2 > 0$ we get that $\lambda_n > 0$. Now we define

$$\sigma_n = \sqrt{\lambda_n}, \quad \text{and} \quad u_n = \tfrac{1}{\sigma_n} Kv_n.$$

Checking that the claimed equalities hold as well as checking orthonormality of the u_n is a routine calculation. □

Remark 4.7. (a) We call (σ_n, u_n, v_n) the *singular system* of K.

(b) We also get the singular value decomposition of K^*, namely

$$K^*y = \sum_n \sigma_n \langle y, u_n \rangle v_n.$$

(c) The σ_n are called *singular values*, the u_n are *left singular vectors* and the v_n are *right singular vectors*.

(d) The singular vectors can be used to project onto the closures of the ranges of K and K^*, namely

$$P_{\operatorname{cl}\operatorname{rg}(K)} y = \sum_n \langle y, u_n \rangle u_n, \quad P_{\operatorname{cl}\operatorname{rg}(K^*)} x = \sum_n \langle x, u_n \rangle u_n.$$

(e) We have $\sum_n |\langle x, v_n \rangle|^2 = \|P_{\operatorname{cl}(\operatorname{rg}(K))}(x)\|_X \leq \|x\|_X$.

The singular value decomposition also allows to describe the boundary of the range:

Theorem 4.8 (Picard condition)**.** *Let $K \in K(X, Y)$ with singular system (σ_n, u_n, v_n) and let $y \in \operatorname{cl}(\operatorname{rg}(K))$. Then $y \in \operatorname{rg}(K)$ exactly if*

$$\sum_n \frac{|\langle y, u_n \rangle|^2}{\sigma_n^2} < \infty. \tag{P}$$

Proof. Let $y \in \operatorname{rg}(K)$, then there is x with $y = Kx$ and we have

$$\langle y, u_n \rangle = \langle Kx, u_n \rangle = \langle x, K^* u_n \rangle = \sigma_n \langle x, v_n \rangle .$$

We get

$$\sum_n \frac{|\langle y, u_n \rangle|^2}{\sigma_n^2} = \sum_n |\langle x, v_n \rangle|^2 \leq \|x\|_X^2 < \infty.$$

Conversely, the $y \in \operatorname{cl}(\operatorname{rg}(K))$ fulfill (P). We define $x_N = \sum_{n=1}^N \frac{1}{\sigma_n} \langle y, u_n \rangle v_n$ and from (P) it follows that x_N is a Cauchy sequence, and thus,

$$x_N \to \sum_n \tfrac{1}{\sigma_n} \langle y, u_n \rangle v_n =: x.$$

Finally, we get

$$\begin{aligned} Kx &= K\left(\sum_n \tfrac{1}{\sigma_n} \langle y, u_n \rangle v_n \right) = \sum_n \tfrac{1}{\sigma_n} \langle y, u_n \rangle Kv_n = \sum_n \langle y, u_n \rangle u_n \\ &= P_{\operatorname{cl}(\operatorname{rg}(K))} y = y \end{aligned}$$

which shows that $y \in \operatorname{rg}(K)$. □

With the singular value decomposition, we can define the so-called *Moore-Penrose pseudo-inverse* (often just called pseudo-inverse).

Definition 4.9 (Pseudo-inverse)**.** Let $K \in K(X, Y)$ with singular system (σ_n, u_n, v_n). Then the *pseudo-inverse* of K, is $K^\dagger : \operatorname{rg}(K) \oplus \operatorname{rg}(K)^\perp \to X$ defined by

$$K^\dagger y = \sum_n \tfrac{1}{\sigma_n} \langle y, u_n \rangle v_n.$$

We denote by $D(K^\dagger) := \operatorname{rg}(K) \oplus \operatorname{rg}(K)^\perp \subset Y$ the *domain* of the pseudo-inverse.

Remark 4.10.

1. Note that

$$\begin{aligned} K^\dagger K x &= \sum_n \tfrac{1}{\sigma_n} \langle Kx, u_n \rangle v_n = \sum_n \tfrac{1}{\sigma_n} \langle x, K^* u_n \rangle v_n \\ &= \sum_n \langle x, v_n \rangle v_n = P_{\operatorname{cl}\operatorname{rg}(K^*)} x = P_{\ker(K)^\perp} x, \end{aligned}$$

   i.e. $K^\dagger K = P_{\ker(K)^\perp}$.

2. Similarly, we have

$$\begin{aligned} KK^\dagger y &= \sum_n \sigma_n \left\langle K^\dagger y, v_n \right\rangle u_n = \sum_n \sigma_n \left\langle \sum_m \tfrac{1}{\sigma_m} \langle y, u_m \rangle v_m, v_n \right\rangle u_n \\ &= \sum_{m,n} \sigma_n \tfrac{1}{\sigma_m} \langle y, u_m \rangle \langle v_m, v_n \rangle u_n = \sum_n \langle y, u_n \rangle u_n = P_{\operatorname{cl}(\operatorname{rg}(K))} y \\ &= P_{\ker(K^*)^\perp}, \end{aligned}$$

   i.e. $KK^\dagger = P_{\operatorname{cl}(\operatorname{rg}(K))} = P_{\ker(K^*)^\perp}$.

3. We have $K^\dagger y = 0$ if $y \in \operatorname{rg}(K)^\perp$, i.e. $\ker(K^\dagger) = \operatorname{rg}(K)^\perp$.

4. Since (v_n) is a basis of $\operatorname{cl}\operatorname{rg}(K^*) = \ker(K)^\perp$, we have that $\operatorname{rg}(K^\dagger) = \operatorname{cl}\operatorname{rg}(K^*) = \ker(K)^\perp$. Note that $\operatorname{rg}(K)$ is in general not closed (for compact operators it is only closed if it is finite dimensional), i.e. it is not a Hilbert space.

By the above remark, the pseudo-inverse is actually a kind of an inverse, namely of $K|_{\ker(K)^\perp} : \ker(K)^\perp \to \operatorname{rg}(K)$. There is a little more to say:

Theorem 4.11. *For every $y \in D(K^\dagger)$ it holds that the equation $Kx = y$ has a unique* minimum norm solution *which is $x^\dagger = K^\dagger y$, i.e. $x^\dagger$ is a least squares solution of minimal norm, i.e. it holds that*

$$\|Kx^\dagger - y\|_Y = \min\{\|Kx - y\|_Y \mid x \in X\} \quad \textit{and}$$
$$\|x^\dagger\|_X = \min\{\|z\|_X \mid z \textit{ is a least squares solution of } Kx = y\}.$$

The first equality defines "least squares solutions".

Moreover, the set of all least squares solutions is $x^\dagger + \ker(K)$.

Proof. That $x^\dagger = K^\dagger y$ is a least squares solution follows from Remark 4.10, 2.:

$$\|Kx^\dagger - y\|_Y = \|KK^\dagger y - y\|_Y = \|P_{\operatorname{cl}(\operatorname{rg}(K))} y - y\|_Y.$$

Now recall that the orthogonal projection $P_{\operatorname{cl}(\operatorname{rg}(K))} y$ is the closest point to y within the closure of range of K.

If x' is any least squares solution x' we can write $x' = x^\dagger + v$ with $v \in \ker(K)$, but since $x^\dagger \in \ker(K)^\perp$ we have (by the Pythagorean theorem)

$$\|x'\|_X^2 = \|x^\dagger\|_X^2 + \|v\|_X^2 \geq \|x^\dagger\|_X^2$$

which shows that $x^\dagger$ has minimal norm among all least squares solutions. □

Remark 4.12. The pseudo inverse can also be defined for general bounded linear operators (not necessarily compact ones) $A \in L(X, Y)$. There one defines the $A^\dagger y$ as the unique minimum norm least squares solution (and has to show that this is a meaningful definition). All properties of the pseudo inverse we have shown are still fulfilled in this case.

We will use the pseudo-inverse also for merely bounded operators in the following.

5 Regularization

We have seen in the previous section that the pseudo-inverse solves two of the problems with ill-posed linear problems: Existence and uniqueness. A little bit more explicit: The problem of existence is (roughly) solved by moving to least squares solutions (i.e. minimizing the residual $\|Kx - y\|_Y$ rather than solving $Kx = y$) and the problem of uniqueness is solved by considering minimum norm solutions, i.e. among all (least squares) solution we pick the one with minimal norm. What about the remaining problem of instability?

Before we answer that, we note the following fact:

Lemma 5.1. *If $K \in K(X, Y)$ has the singular system $(\sigma_n, u_n v_n)$, then we have $\|K\| = \sigma_1$.*

The proof is a good exercise

Unfortunately, this shows that the pseudo-inverse is, in general, not bounded: If $\operatorname{rg}(K)$ is infinite dimensional, we have from Theorem 4.6 that $\sigma_n \to 0$. But this implies

$$\|K^\dagger u_n\| = \|\sum_m \tfrac{1}{\sigma_m} \langle u_n, u_m \rangle v_m\| = \|\tfrac{1}{\sigma_n} v_n\| = \tfrac{1}{\sigma_n} \overset{n\to\infty}{\longrightarrow} \infty$$

and thus, $K^\dagger$ can not be bounded. The pseudo-inverse even helps to make the instability quite quantifiable: Consider the case that $y^\dagger = Kx^\dagger$ for $x \in \ker(K)^\perp$. Then $x^\dagger$ is the minimum norm least squares solution of $Kx = y^\dagger$. Let's assume that we have measurement data y^δ instead of $y^\dagger$ and let us assume moreover, that we know that we have a small measurement error, i.e. $\|y^\dagger - y^\delta\|_Y \leq \delta$ for some known $\delta > 0$. Then the "noise" ns the data is

$$\eta = y^\delta - y^\dagger \in Y.$$

Let us blindly apply the pseudo inverse to y^δ:

$$K^\dagger y^\delta = K^\dagger(y^\dagger + \eta) = x^\dagger + K^\dagger \eta = x^\dagger + \sum_n \tfrac{1}{\sigma_n} \langle \eta, u_n \rangle v_n.$$

We see that the contribution of the noise is amplified unboundedly, i.e. the component $\langle \eta, u_n \rangle$ of the noise in the n-th singular vector u_n is amplified by a factor of $1/\sigma_n$ and these factors grow beyond all bounds. Hence: If the noise contains contributions from singular vectors that correspond to small singular values, they get amplified a lot. Unfortunately, this is the standard situation: Singular vectors for small singular values tend to be oscillatory (i.e. be of high frequency) and hence, noise always tends to be amplified.

Example 5.2 (Discretized inverse problems). One can check this observation numerically. After discretization, an inverse problem still reads as $\mathbf{Kx} = \mathbf{y}^\delta$ with $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{K} \in \mathbb{R}^{m\times n}$. The singular value decomposition exists as well and if we write the

It's worth to consider the finite dimensional case here: If $K = U\Sigma V^T$ is the singular value decomposition, then $\|K\| = \sigma_1$ and $\|K^\dagger\| = 1/\sigma_k$ where k is the smallest singular value. The *condition number* of K is defined as $\kappa(K) = \|K\|\|K^\dagger\|$ and hence, equal the ratio of the largest and smallest singular value of K. For inverse problems in infinite dimensions, the condition number can the infinite as there may be arbitrarily small singular values.

singular vectors $\mathbf{u}_i$ and $\mathbf{v}_j$ as colums in matrices $\mathbf{U}$ and $\mathbf{V}$ and the singular values σ_i on the diagonal of a matrix $\mathbf{\Sigma}$, we get

$$\mathbf{K}\mathbf{x} = \sum_i \sigma_i \langle \mathbf{x}, \mathbf{v}_i \rangle \mathbf{u}_i = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}.$$

The pseudo inverse is

$$\mathbf{K}^\dagger\mathbf{y} = \sum_i \tfrac{1}{\sigma_i} \langle \mathbf{y}, \mathbf{u}_i \rangle \mathbf{v}_i = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^T\mathbf{y}. \qquad (*)$$

where $\mathbf{\Sigma}^\dagger$ has the values $1/\sigma_i$ on the diagonal.

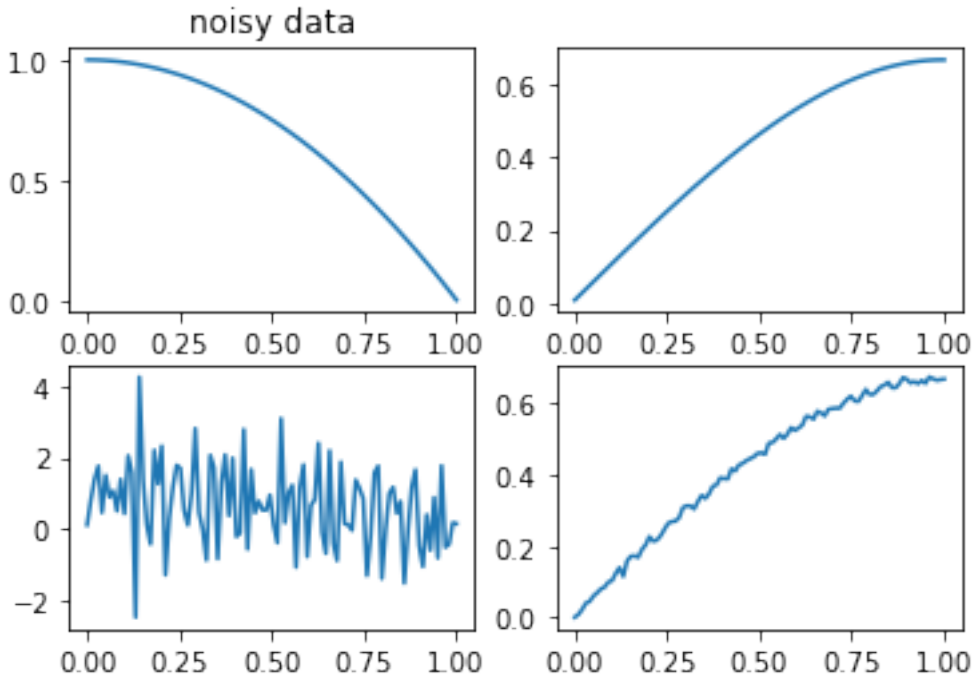
Let us consider a (quite simple) discrete approximation of the inverse problem of differentiation, i.e. the inversion of A given by $Af(x) = \int_0^x f(t)\mathrm{d}t$. This operator can be (roughly) discretized by the matrix

$$\mathbf{A} = \frac{1}{n}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Here is an example of the naive reconstruction (we can use a direct solve here, since the matrix is actually invertible (it is square and it smallest singular value is positive, but quite small). We could, in principle, also use `pinv` to calculate the pseudo-inverse or use the formula (*)).

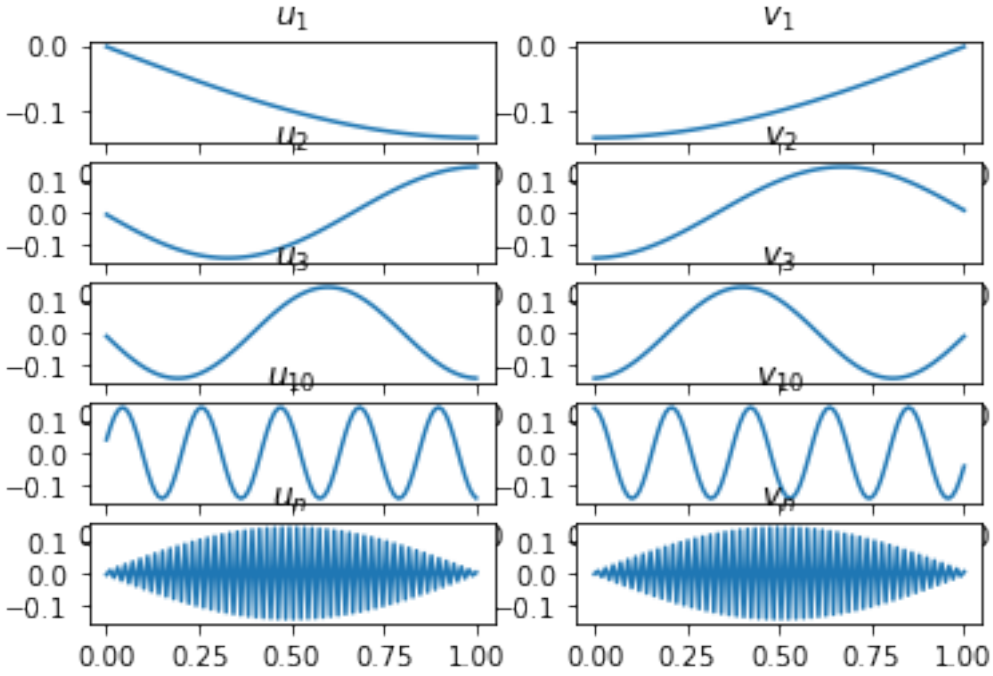
```
import numpy as np
import matplotlib.pyplot as plt

# problem size and matrix
n = 100
A = np.tril(np.ones((n,n)))/n

# discretized interval
t = np.linspace(0,1,n)

# true solution
xdag = 1-t**2
# noise free data
ydag = A@xdag

# noisy data
eta = np.random.randn(n);
eta /= np.sum(eta)

# noise level
delta = 0.05
ydelta = ydag  + delta*eta
```

```
# naive reconstruction
x = np.linalg.solve(A,ydelta)


fig, axs = plt.subplots(2,2)
axs[0,0].plot(t,xdag)
axs[0, 0].set_title('true solution')
axs[0,1].plot(t,ydag)
axs[0, 0].set_title('true data')
axs[1,0].plot(t,x)
axs[0, 0].set_title('naive reconstruction')
axs[1,1].plot(t,ydelta)
axs[0, 0].set_title('noisy data')
plt.show()
```

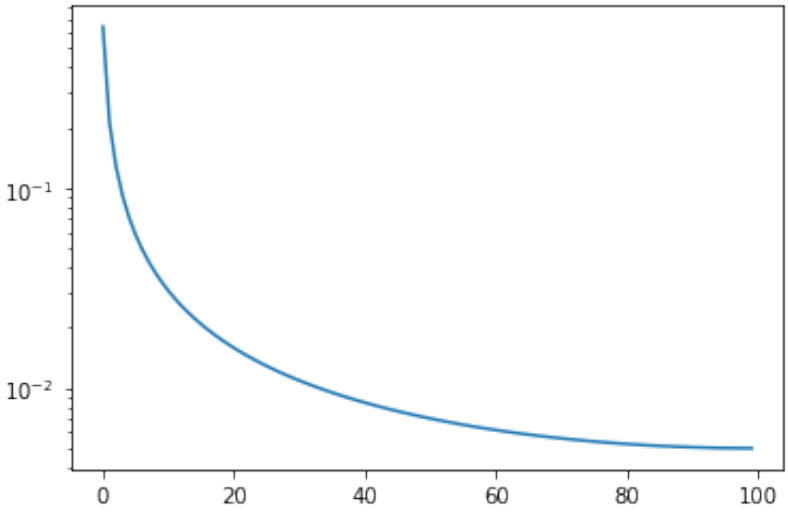



```
# compute svd
U,S,VT = np.linalg.svd(A)

# show some singular vectors
fig, axs = plt.subplots(5,2)
axs[0,0].plot(t,U[:,0])
axs[0,0].set_title('$u_1$')
axs[0,1].plot(t,VT[0,:])
axs[0,1].set_title('$v_1$')
axs[1,0].plot(t,U[:,1])
axs[1,0].set_title('$u_2$')
axs[1,1].plot(t,VT[1,:])
axs[1,1].set_title('$v_2$')
axs[2,0].plot(t,U[:,2])
```

```
axs[2,0].set_title('$u_3$')
axs[2,1].plot(t,VT[2,:])
axs[2,1].set_title('$v_3$')
axs[3,0].plot(t,U[:,9])
axs[3,0].set_title('$u_{10}$')
axs[3,1].plot(t,VT[9,:])
axs[3,1].set_title('$v_{10}$')
axs[4,0].plot(t,U[:,-1])
axs[4,0].set_title('$u_n$')
axs[4,1].plot(t,VT[-1,:])
axs[4,1].set_title('$v_n$')
plt.show()
```




```
# plot singular vectors in semilog plot
plt.semilogy(S)
plt.show()
```



△

Now, let us fix our aim and let us define what a "regularization" shall be:

Definition 5.3 (Regularization)**.** Let $A \in L(X,Y)$. A *regularization* of $A^\dagger$ is a family of continuous maps $R_\alpha : Y \to X, \alpha > 0$ such that for all $y \in D(A^\dagger)$ it holds that

$$R_\alpha y \xrightarrow{\alpha \to 0} A^\dagger y.$$

If all R_α are linear, we speak of a *linear regularization*. The parameter α is called *regularization parameter*.

As a matter of fact, any linear regularization can not be uniformly bounded (as they approximate an unbounded operator):

Theorem 5.4. *Let $A \in L(X,Y)$ and R_α be a linear regularization of $A^\dagger$. If $A^\dagger$ is unbounded, then it holds that $\|R_\alpha\|_X \xrightarrow{\alpha \to 0} \infty$.*

This follows from the so-called uniform boundedness prinicple (also known as Banach-Steinhaus Theorem) and we do not discuss the proof here.

Now let us discuss the various error we defined in the example in Section 2: The *totel error* $\|R_\alpha y^\delta - x^\dagger\|$ (also called regularization error) can be decomposed (using $x^\dagger = A^\dagger y^\dagger$ and the triangle inequality) in the case of linear regularization into *data error* and *approximation error*

$$\begin{aligned} \|R_\alpha y^\delta - x^\dagger\|_X &\leq \|R_\alpha y^\delta - R_\alpha y^\dagger\|_X + \|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X \\ &\leq \|R_\alpha\|\delta + \|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X. \end{aligned} \tag{*}$$

Again, we see that one needs to choose the regularization parameter α carefully: At least we need to be able to balance the term $\|R_\alpha\|\delta$ as it blows up for small α. On the other hand, the term $\|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X$ tends to be small for small α and large for large α (exactly as we have seen in Section 2).

We will often denote the regularized reconstruction by $x_\alpha^\delta := R_\alpha y^\delta$. We will also use the notation $x_\alpha := R_\alpha y^\dagger$ for the (in general

unknown) regularized reconstruction from idealized noiseless data. With this notation, our error decomposition is

$$\underbrace{\|x_\alpha^\delta - x^\dagger\|_X}_{\text{total error}} \leq \underbrace{\|x_\alpha^\delta - x_a\|_X}_{\text{data error}} + \underbrace{\|x_\alpha - x^\dagger\|_X}_{\text{approximation error}} .$$

Definition 5.5 (Parameter choice)**.** A function

$$\alpha :]0,\infty[\times Y \to]0,\infty[, \quad (\delta, y^\delta) \to \alpha(\delta, y^\delta)$$

is called a *parameter choice rule*. We distinguish further: α is

(i) an *a priori choice rule* if α does not depend on y^δ,

(ii) an *a posteriori choice rule* if α depends on δ and y^δ, and

(iii) a *heuristic rule* is α does not depend on δ.

An *a priori* rule can be devised without having seen the actual data (it only needs knowledge of the noise level), hence one can, in principle, construct the operator $R_{\alpha(\delta)}$ a priorily, before the data has arrived; hence, the name.

Definition 5.6 (Convergent regularization)**.** If R_α is a regularization of $A^\dagger$ and α is a parameter choice rule we say that (R_α, α) is a *convergent regularization method* if for all $y^\dagger \in D(A^\dagger)$ it holds that

$$\sup\left\{ \|R_{\alpha(\delta,y^\delta)}y^\delta - A^\dagger y^\dagger\|_X \;\middle|\; \|y^\delta - y^\dagger\|_Y \leq \delta \right\} \xrightarrow{\delta\to 0} 0.$$

In other words: We want that the *worst case reconstruction error* goes to zero, i.e. even if the noisy data y^δ is as bad as possible, given the noise level δ.

Informally: We demand that in the regime of vanishing noise, we shall be able to approximate the true solution $x^\dagger = A^\dagger y^\dagger$ as good as possible. On the one hand, this sounds like a meaningless demand, since usually the noise level stays fixed. On the other hand, it sounds like something that should be the bare minimum: If we can not even guarantee this, what is the point of regularization at all? Finally, it sounds quite ambitious, given that we already know that we try to approximate unbounded (i.e. discontinuous) operators with continuous ones.

Example 5.7 (Truncated SVD). Here is a simple idea for a regularization method for: For $K \in K(X,Y)$ with singular system (σ_n, u_n, v_n) and $\alpha > 0$ we define

$$R_\alpha y = \sum_{\sigma_n > \alpha} \tfrac{1}{\sigma_n} \langle y, u_n\rangle v_n$$

i.e. we cut off the small singular values which lead to unboundedness of the pseudo-inverse. These R_α are indeed bounded operators:

$$\|R_\alpha y\|_X^2 = \sum_{\sigma_n > \alpha} \tfrac{1}{\sigma_n^2} |\langle y, u_n\rangle|^2 \leq \sup\left\{ \tfrac{1}{\sigma_n^2} \mid \sigma_n \geq \alpha \right\} \|y\|_Y^2 \leq \tfrac{1}{\alpha^2}\|y\|_Y^2$$

i.e. $\|R_\alpha\| \leq \frac{1}{\alpha}$.

Let us investigate if the truncated SVD is a convergent regularization method, i.e. if we can find a suitable parameter choice: To that end, we use our standard error decomposition (*)

$$\begin{aligned}\|R_\alpha y^\delta - K^\dagger y^\dagger\|_X &\leq \|R_\alpha\|\delta + \|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X \\ &\leq \tfrac{\delta}{\alpha} + \|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X.\end{aligned}$$

We estimate the approximation error further

$$R_\alpha y^\dagger - A^\dagger y^\dagger = \sum_{\sigma_n \geq \alpha} \tfrac{1}{\sigma_n} \left\langle y^\dagger, u_n \right\rangle v_n - \sum_n \tfrac{1}{\sigma_n} \left\langle y^\dagger, u_n \right\rangle v_n = -\sum_{\sigma_n < \alpha} \tfrac{1}{\sigma_n} \left\langle y^\dagger, u_n \right\rangle v_n.$$

This gives us

$$\|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X^2 \leq \sum_{\sigma_n < \alpha} \tfrac{1}{\sigma_n^2} |\left\langle y^\dagger, u_n \right\rangle|^2.$$

Together we have

$$\begin{aligned} \|R_\alpha y^\delta - K^\dagger y^\dagger\|_X &\leq \|R_\alpha\|\delta + \|R_\alpha y^\dagger - A^\dagger y^\dagger\|_X \\ &\leq \tfrac{\delta}{\alpha} + \sqrt{\sum_{\sigma_n < \alpha} \tfrac{1}{\sigma_n^2} |\langle y^\dagger, u_n \rangle|^2}. \end{aligned}$$

Now we see: The second summand is the "rest of a convergent series" (recall the Picard condition, Theorem 4.8) and the smaller α, the later the rest of the series starts. Hence, we have

$$\sqrt{\sum_{\sigma_n < \alpha} \tfrac{1}{\sigma_n^2} |\langle y^\dagger, u_n \rangle|^2} \to 0 \quad \text{for} \quad \alpha \to 0.$$

For the first term we need that $\alpha(\delta) \to 0$ slower than δ. In conclusion: Any $\alpha(\delta)$ with

$$\alpha(\delta) \xrightarrow{\delta \to 0} 0, \quad \frac{\delta}{\alpha(\delta)} \xrightarrow{\delta \to 0} 0$$

is a valid (a priori) parameter choice rule and we can claim that the truncated SVD together with this rule is a convergent regularization method. △

One could take, for example, $\alpha(\delta) = \sqrt{\delta}$ (or $= \delta^\kappa$ for $0 < \kappa < 1$, for that matter).

6 Tikhonov regularization

The problem of instability of the solution of $Ax = y^\delta$ comes from the small singular values which are the eigenvalues of the self-adjoint operator A^*A. Another way to understand this, is via the normal equation: Some x is a minimizer of $\|Ax - y^\delta\|_X^2$ exactly if it solves the *normal equation*

$$A^*Ax = A^*y^\delta.$$

However, in general minimizers of $\|Ax - y^\delta\|_X^2$ do not exist (i.e. the normal equation does not have solutions) and this is (in the case of compact A) due to the eigenvalues of A^*A converging to zero. To avoid this problem, we can simply shift them to be strictly positive: If σ_i^2 are the eigenvalues of A^*A, then the eigenvalues of $A^*A + \alpha \operatorname{id}$ are $\sigma_n^2 + \alpha \geq \alpha > 0$. Hence, instead of the normal equation, we consider for $\alpha > 0$ regularized normal equations

$$(A^*A + \alpha \operatorname{id})x = A^*y^\delta.$$

Since the operator $A^*A + \alpha \operatorname{id}$ is always invertible, we can write this as

$$x_\alpha^\delta = (A^*A + \alpha \operatorname{id})^{-1}A^*y^\delta$$

and this method is called *Tikhonov regularization*. The shift of the singular values is one motivation for Tikhonov regularization. In fact, Tikhonov regularization also corresponds to a regularized least squares problem.

Theorem 6.1. *Let $A \in L(X, Y)$. The regularized normal equation $(A^*A + \alpha \operatorname{id})x = A^*y^\delta$ has a unique solution x_α^δ which is exactly the unique minimum of the Tikhonov functional*

$$T_\alpha(x; y^\delta) := \tfrac{1}{2}\|Ax - y^\delta\|_Y^2 + \tfrac{\alpha}{2}\|x\|_X^2.$$

Proof. A minimizer x of the Tikhonov function is characterized by the condition that $T_\alpha(x + th; y^\delta) \geq T_\alpha(x; y^\delta)$ for all $t \in \mathbb{R}$ and $h \in X$. Starting from the left hand side we get

$$\begin{aligned}
T_\alpha(x + th; y^\delta) &= \tfrac{1}{2}\|Ax + tAh - y^\delta\|_Y^2 + \tfrac{\alpha}{2}\|x + th\|_X^2 \\
&= \tfrac{1}{2}\|Ax - y^\delta\|_Y^2 + \left\langle Ax - y^\delta, tAh \right\rangle + \tfrac{1}{2}\|tAh\|_Y^2 \\
&\quad + \tfrac{\alpha}{2}\|x\|_X^2 + \alpha \langle x, th \rangle + \tfrac{\alpha}{2}\|th\|_X^2 \\
&= T_\alpha(x; y^\delta) + t\left\langle A^*(Ax - y^\delta) + \alpha x, h \right\rangle + t^2(\tfrac{1}{2}\|Ah\|_Y^2 + \tfrac{\alpha}{2}\|h\|_X^2).
\end{aligned}$$

We see that $T_\alpha(x + th; y^\delta) \geq T_\alpha(x; y^\delta)$ holds for all t and h exactly if

$$\left\langle A^*(Ax - y^\delta) + \alpha x, h \right\rangle = 0$$

for all $h \in X$ and this is exactly the case when $A^*(Ax - y^\delta) + \alpha x = 0$ which is just the regularized normal equation. Uniqueness of the minimizer follows since the Tikhonov functional is strictly convex. □

The description of Tikhonov regularization as a minimization framework allows for another interpretation: The regularization is a compromise of two things, namely finding a reconstruction x_α^δ that has a good *data fit*, i.e. it produces a small value for the residual $\|Ax - y^\delta\|_Y$, but, at the same time, also does not blow up, i.e. it has a small norm $\|x\|_X$. These two demands are weighted by the regularization parameter α. Regularization methods that build upon the idea of minimizing a functional that balances the demands of data fit and "reasonable reconstruction" are also called "variational regularization methods" (as the theory that deals with minimization problems in infinite dimensional spaces is called "calculus of variations"). Aiming at a reconstruction with a bounded norm seems like a valid idea, but one may know a little more about the unknown solution. If we assume that we have a rough idea of the unknown $x^\dagger$, i.e. we know that x^0 is a good guess, we can of course minimize

$$T_\alpha(x; y^\delta, x^0) := \tfrac{1}{2}\|Ax - y^\delta\|_Y^2 + \tfrac{\alpha}{2}\|x - x^0\|_X^2.$$

Similar to the proof of Theorem 6.1 one shows that the unique minimizer here is given as a solution of

$$(A^*A + \alpha \,\mathrm{id})x = A^*y^\delta + \alpha x^0.$$

Remark 6.2 (Numerical realization of Tikhonov regularization). Tikhonov regularization is popular, because its implementation is pretty straight forward. Let us consider the discrete case where $\mathbf{A} \in \mathbb{K}^{m\times n}$ and $\mathbf{y}^\delta \in \mathbb{R}^m$. Then the regularized normal equation (for $\mathbf{x}^0 = 0$

$$(\mathbf{A}^T\mathbf{A} + \alpha I_n)\mathbf{x} = \mathbf{A}^T\mathbf{y}^\delta$$

is a square linear system in n dimensions and the matrix $(\mathbf{A}^T\mathbf{A} + \alpha I_n)$ is symmetric positive definite. Hence, there are many methods available to solve the problem numerically (one method is the method of conjugate gradients).

Both the overdetermined case $m > n$ (where non-existence of solutions is a problem, due to measurement error) and the underdetermined case $m < n$ (where non-uniqueness is a problem, due to not enough data) of can be considered here.

Is Tikhonov regularization indeed a convergence regularization method? To answer this question, we should find a parameter choice rule. We will analyze this question with the help of the singular value decomposition.

Theorem 6.3. *Let $K \in K(X, Y)$ have the singular system (σ_n, u_n, v_n). Then solution x_α^δ of $(A^*A + \alpha \,\mathrm{id})x = A^*y^\delta$ is given by*

$$x_\alpha^\delta = \sum_n \tfrac{\sigma_n}{\sigma_n^2+\alpha} \left\langle y^\delta, u_n \right\rangle v_n.$$

Proof. It holds that $x_\alpha^\delta = P_{\ker(A)}x_\alpha^\delta + P_{\ker(A)^\perp}x_\alpha^\delta = P_{\ker(A)}x_\alpha^\delta + \sum_n \left\langle x_\alpha^\delta, v_n \right\rangle v_n$. Since (σ_n^2, v_n, v_n) is the spectral decomposition of

A^*A we get $A^*Ax_\alpha^\delta = \sum_n \sigma_n^2 \langle x_\alpha^\delta, v_n\rangle v_n$. Also $A^*y^\delta = \sum_n \sigma_n \langle y^\delta, u_n\rangle v_n$. Thus, the regularized normal equation is

$$\sum_n \sigma_n^2 \left\langle x_\alpha^\delta, v_n \right\rangle v_n + \alpha \left(P_{\ker(A)}(x_\alpha^\delta) + \sum_n \left\langle x_\alpha^\delta, v_n \right\rangle v_n \right) = A^*y^\delta$$
$$= \sum_n \sigma_n \left\langle y^\delta, u_n \right\rangle v_n.$$

We see that necessarily $P_{\ker(A)}(x_\alpha^\delta) = 0$ and that

$$\sum_n (\sigma_n^2 + \alpha) \left\langle x_\alpha^\delta, v_n \right\rangle v_n = \sum_n \sigma_n \left\langle y^\delta, u_n \right\rangle v_n.$$

Comparing coefficients shows that $\langle x_\alpha^\delta, v_n\rangle = \frac{\sigma_n}{\sigma_n^2+\alpha} \langle y^\delta, u_n\rangle$ which shows the claim. □

The representation of x_α^δ from Theorem 6.3 via the singular value decomposition is called *spectral representation*. We use it to prove the following result on regularization:

Theorem 6.4 (Tikhonov with a-priori parameter choice)**.** *For an a-priori parameter choice $\alpha(\delta)$ that fulfills*

$$\alpha(\delta) \to 0 \quad \frac{\delta^2}{\alpha(\delta)} \to 0 \quad \textit{for} \quad \delta \to 0$$

it holds that Tikhonov regularization is a convergent regularization method, i.e. it holds that $x_\alpha^\delta := (A^*A + \alpha \operatorname{id})^{-1}A^*y^\delta \to x^\dagger := A^\dagger y^\dagger$ whenever $\|y^\delta - y^\dagger\| \le \delta$ and $\delta \to 0$.

Proof. We set $x_\alpha = (A^*A + \alpha \operatorname{id})^{-1}A^*y^\dagger$ and decompose

$$x_\alpha^\delta - x^\dagger = x_\alpha^\delta - x_\alpha + x_\alpha - x^\dagger$$
$$= \underbrace{(A^*A + \alpha \operatorname{id})^{-1}A^*(y^\delta - y^\dagger)}_{\text{data error}} + \underbrace{(A^*A + \alpha \operatorname{id})^{-1}A^*y^\dagger - A^\dagger y^\dagger}_{\text{approx. error}}.$$

The data error fulfills

$$(A^*A + \alpha \operatorname{id})^{-1}A^*(y^\delta - y^\dagger) = \sum_n \tfrac{\sigma_n}{\sigma_n^2+\alpha} \left\langle y^\delta - y^\dagger, u_n \right\rangle v_n,$$

and hence, its norm is

$$\|(A^*A + \alpha \operatorname{id})^{-1}A^*(y^\delta - y^\dagger)\|_Y^2 = \sum_n \left(\tfrac{\sigma_n}{\sigma_n^2+\alpha}\right)^2 |\left\langle y^\delta - y^\dagger, u_n \right\rangle|^2$$
$$\le \left(\sup_{0\le\sigma\le\|A\|} \frac{\sigma}{\sigma^2+\alpha} \right)^2 \|y^\delta - y^\dagger\|_Y^2$$

For the approximation error and we use that $x^\dagger = A^\dagger y^\dagger$ implies $\langle x^\dagger, v_n\rangle = \frac{1}{\sigma_n} \langle y^\delta, u_n\rangle$ to get

$$(A^*A + \alpha \operatorname{id})^{-1}A^*y^\dagger - A^\dagger y^\dagger = \sum_n \tfrac{\sigma_n}{\sigma_n^2+\alpha} \left\langle y^\dagger, u_n \right\rangle v_n - \sum_n \tfrac{1}{\sigma_n} \left\langle y^\dagger, u_n \right\rangle v_n$$
$$= \sum_n \left(\tfrac{\sigma_n^2}{\sigma_n^2+\alpha} - 1\right) \left\langle x^\dagger, v_n \right\rangle v_n.$$

Together we arrive at the error estimate

$$\|x_\alpha^\delta - x^\dagger\|_X \leq \left(\sup_{0\leq\sigma\leq\|A\|} \tfrac{\sigma}{\sigma^2+\alpha} \right) \delta + \sqrt{\sum_n \left(\tfrac{\sigma_n^2}{\sigma_n^2+\alpha} - 1\right)^2 |\langle x^\dagger, v_n\rangle|^2}. \tag{*}$$

We estimate the supremum by

$$\sup_{0\leq\sigma\leq\|A\|} \tfrac{\sigma}{\sigma^2+\alpha} \leq \tfrac{1}{2\sqrt{\alpha}}.$$

We maximize over all $\sigma > 0$: The derivative of $\sigma/(\sigma^2+\alpha)$ is $((\sigma^2+\alpha) - 2\sigma^2)/(\sigma^2+\alpha)^2$ and hence, vanishes exactly at $\sigma = \sqrt{\alpha}$. Plugging this in gives the result.

By assumption $\delta/\sqrt{\alpha} \to 0$ for $\delta \to 0$, and thus, the first term on the right hand side of (*) goes to zero for $\delta \to 0$.

Now we consider the square of second term in (*), which we write as $\sum_n a_n(\alpha)$ with $a_n(\alpha) = \left(\tfrac{\sigma_n^2}{\sigma_n^2+\alpha} - 1\right)^2 |\langle x^\dagger, v_n\rangle|^2$. It holds that $a_n(\alpha) \to |\langle x^\dagger, v_n\rangle|^2$ for $\alpha \to 0$. We have the (very coarse) estimate $\left(\tfrac{\sigma_n^2}{\sigma_n^2+\alpha} - 1\right)^2 \leq 4$ and hence $\sum_n a_n(\alpha) \leq 4\|x\|_X^2$,and by the dominated convergence theorem (Theorem 3.5), we get that the full sum $\sum_n a_n(\alpha) \to 0$ for $\alpha \to 0$. This proves the theorem. □

The previous theorem shows that Tikhonov is indeed a convergent regularization method. However, we did not get an explicit error estimate for the total error $\|x_\alpha^\delta - x^\dagger\|_X$. While we could bound the data error by

$$\|x_\alpha^\delta - x_\alpha\|_X \leq \tfrac{\delta}{\sqrt{\alpha}},$$

we did not get an effective bound on the approximation error $\|x_\alpha - x^\dagger\|_X$. This is a general fact:

Theorem 6.5 (No general worst case error bound for ill-posed problems)**.** *Let $(R_\alpha, \alpha(\delta, y^\delta))$ be a convergent regularization method for $A^\dagger$. If there exists a function $\psi :]0,\infty[\to]0,\infty[$ with $\psi(\delta) \xrightarrow{\delta\to 0} 0$ such that for all $y^\dagger \in D(A^\dagger)$*

$$\sup\left\{ \|R_{\alpha(\delta,y^\delta)} y^\delta - A^\dagger y^\dagger\|_X \mid y^\dagger \in D(A^\dagger),\ y^\delta \in Y,\ \text{with } \|y^\dagger - y^\delta\|_Y \leq \delta \right\} \leq \psi(\delta)$$

then $A^\dagger$ is bounded.

The significance of this theorem is as follows: If $A^\dagger$ is bounded, we can get a nice error bound $\|A^\dagger y^\delta - x^\dagger\|_X \leq \|A^\dagger\|\delta$ and hence, the problem is not ill-posed.

Proof. Let $y^\dagger, y_n \in D(A^\dagger)$ with $\|y^\dagger - y_n\|_Y = \delta_n \xrightarrow{n\to\infty} 0$. Then we have

$$\|A^\dagger y_n - A^\dagger y^\dagger\|_Y \leq \|A^\dagger y_n - R_{\alpha(\delta,y_n)} y_n\| + \|R_{\alpha(\delta,y_n)} y_n - A^\dagger y\|.$$

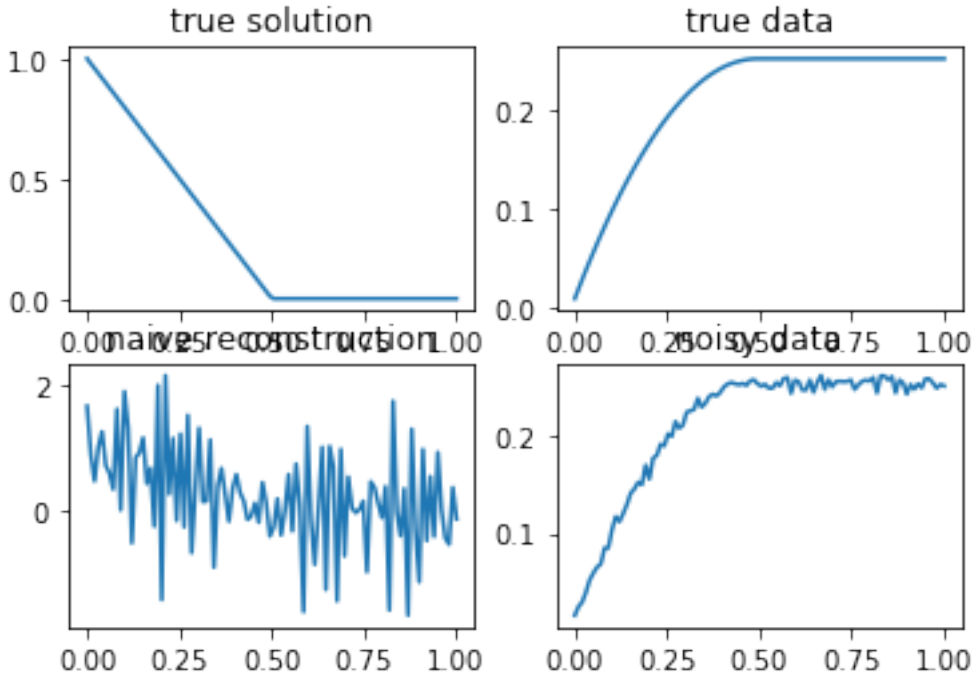
By our assumption, we have that both terms on the right are bounded by $\psi(\delta_n)$, i.e.

$$\|A^\dagger y_n - A^\dagger y^\dagger\|_Y \leq 2\psi(\delta_n) \xrightarrow{n\to\infty} 0.$$

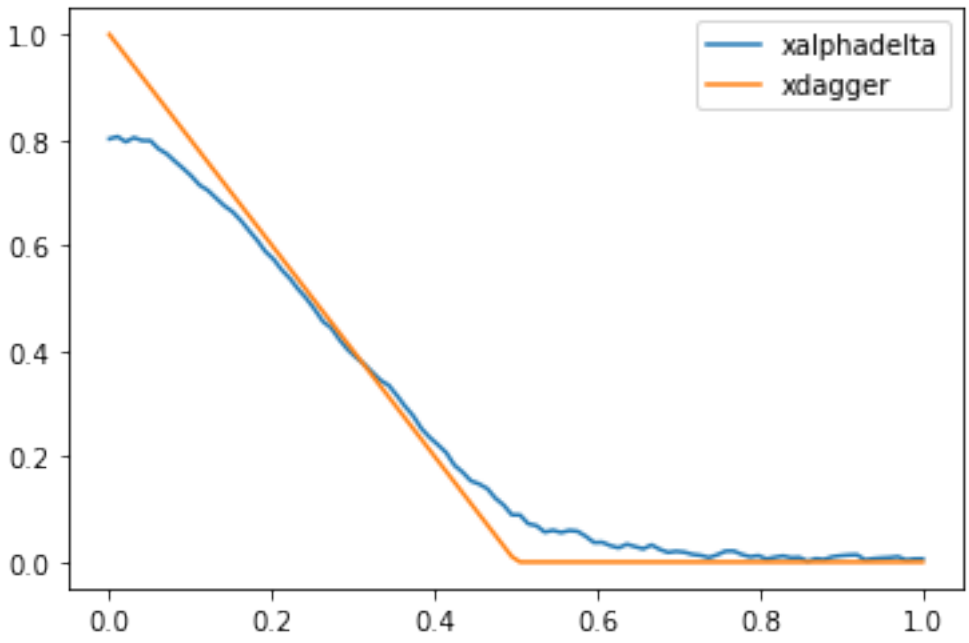
But this means that $A^\dagger$ is continuous at $y^\dagger$ and since $A^\dagger$ is linear, this shown continuity everywhere. □

Here is an example of Tikhonov regularization in practice:

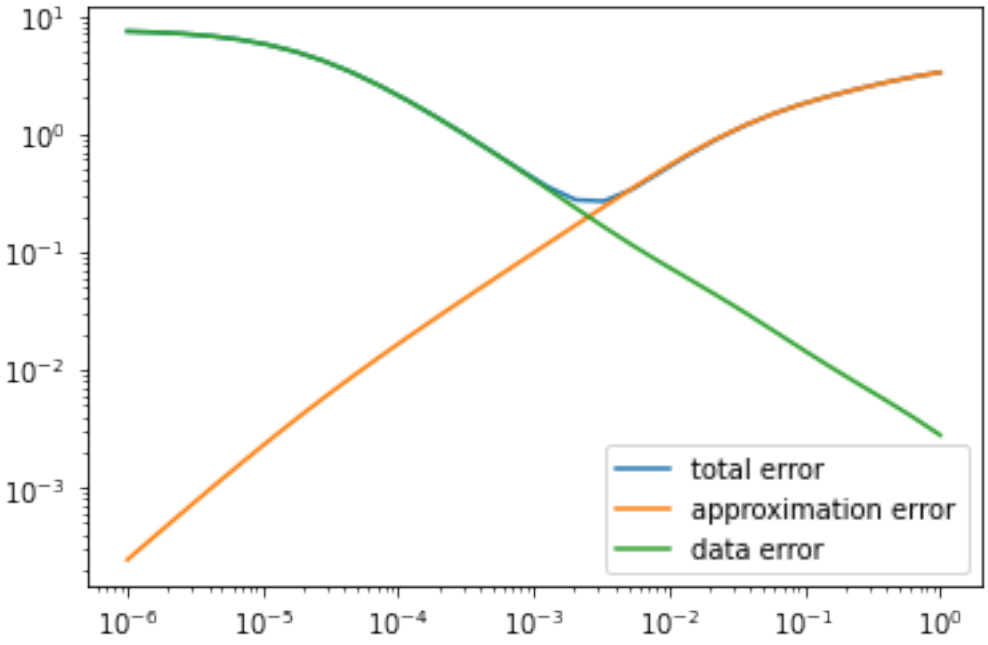
```
import numpy as np
import matplotlib.pyplot as plt

# problem size and matrix
n = 100
A = np.tril(np.ones((n,n)))/n

# discretized interval
t = np.linspace(0,1,n)

# true solution
xdag = np.maximum(1-2*t,0)
# noise free data
ydag = A@xdag

# noisy data
eta = np.random.randn(n);
eta /= np.linalg.norm(eta)

# noise level
delta = 0.05
ydelta = ydag  + delta*eta


# naive reconstruction
x = np.linalg.solve(A,ydelta)

fig, axs = plt.subplots(2,2)
axs[0,0].plot(t,xdag)
axs[0,0].set_title('true solution')
axs[0,1].plot(t,ydag)
axs[0,1].set_title('true data')
axs[1,0].plot(t,x)
axs[1,0].set_title('naive reconstruction')
axs[1,1].plot(t,ydelta)
axs[1,1].set_title('noisy data')
plt.show()
```




```
# reconstruct with Tikhonov
# regularization parameter
alpha = 0.01
# compute reconstruction
xalphadelta = np.linalg.solve(A.T@A + alpha*np.identity(n),A.T@ydelta)
plt.plot(t,xalphadelta,label='xalphadelta')
plt.plot(t,xdag,label='xdagger')
plt.legend()
plt.show()
```



```
# Plot for N different errors to illustrate error decomposition
N = 30
alphas = np.logspace(0,-6,N)
totalError = np.zeros(N)
approximationError = np.zeros(N)
dataError = np.zeros(N)
# reconstruct and compute errors
for k in range(N):
    alpha = alphas[k]
    xalphadelta = np.linalg.solve(A.T@A + alpha*np.identity(n),A.T@ydelta)
    xalpha = np.linalg.solve(A.T@A + alpha*np.identity(n),A.T@ydag)
    totalError[k] = np.linalg.norm(xalphadelta-xdag)
    approximationError[k] = np.linalg.norm(xalpha-xdag)
    dataError[k] = np.linalg.norm(xalphadelta-xalpha)

# Show errors is loglog-plot
plt.loglog(alphas,totalError,label='total error')
plt.loglog(alphas,approximationError,label='approximation error')
plt.loglog(alphas,dataError,label='data error')
plt.legend()
plt.show()
```



7 Spectral regularization

We have analyzed the regularization properties of the truncated singular value decomposition and Tikhonov regularization in Example 5.7 and Theorem 6.4. If you inspect the arguments, you'll note that they are actually almost the same in both cases. In this section we will start to derive a general theory for linear regularization. This theory will contain the truncated SVD as well as Tikhonov regularization as special cases.

Recall that the truncated SVD is

$$R_\alpha y = \sum_{\sigma_n > \alpha} \tfrac{1}{\sigma_n} \langle y, u_n \rangle v_n$$

while we could express Tikhonov regularization as

$$R_\alpha y = \sum_n \tfrac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_n \rangle v_n.$$

Both methods can be written in the following form:

$$R_\alpha y = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n \langle y, u_n \rangle v_n \tag{R}$$

with some function φ_α:

For $\varphi_\alpha(\lambda) = \frac{1}{\lambda+\alpha}$ we get

$$R_\alpha y = \sum_n \tfrac{1}{\sigma_n^2 + \alpha}\sigma_n \langle y, u_n \rangle v_n,$$

i.e. exactly Tikhonov regularization. If we set

$$\varphi_\alpha(\lambda) = \begin{cases} \frac{1}{\lambda} & : \quad \lambda \geq \alpha \\ 0 & : \quad \text{else,} \end{cases}$$

we get

$$R_\alpha y = \sum_{\sigma_n^2 \geq \alpha} \tfrac{1}{\sigma_n^2}\sigma_n \langle y, u_n \rangle v_n = \sum_{\sigma_n > \sqrt{\alpha}} \tfrac{1}{\sigma_n} \langle y, u_n \rangle v_n,$$

which is (up to a different scaling of the regularization parameter) the truncated SVD from Example 5.7.

Remark 7.1. Note that we could have written $R_\alpha y = \sum_n f_\alpha(\sigma_n) \langle y, u_n \rangle v_n$ with some function f_α as well. However,there is a reason why we did chose this slightly complicated form: Regularization methods approximate the minimum norm solution of the normal equation $K^*Kx = K^*y$, i.e. $x = (K^*K)^\dagger K^*y$. If we express everything with the SVD of K we first get that $(K^*K)^\dagger z = \sum_n \sigma_n^{-2} \langle z, v_n \rangle v_n$ and hence, for $z = K^*y$

$$\begin{aligned} x &= \sum_n \sigma_n^{-2} \langle K^*y, v_n \rangle v_n = \sum_n \sigma_n^{-2} \langle y, Kv_n \rangle v_n \\ &= \sum_n \sigma_n^{-2} \langle y, \sigma_n u_n \rangle v_n = \sum_n \sigma_n^{-2}\sigma_n \langle y, u_n \rangle v_n. \end{aligned}$$

To mimic this formula, we express regularization methods as

$$R_\alpha y = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n \langle y, u_n\rangle v_n.$$

and need that $\varphi_\alpha(\lambda) \approx 1/\lambda$ for R_α being close to $A^\dagger$.

We will use the following *functional calculus* for compact operators: If K is compact with singular system (σ_n, u_n, v_n) and $f : [0, \|K\|^2] \to \mathbb{R}$ is piecewise continuous and bounded, we define another operator $f(K^*K) : X \to Y$ by

$$f(K^*K)x := \sum_n f(\sigma_n^2) \langle x, v_n\rangle v_n + f(0)P_{\ker(K)}x.$$

The additional term $P_{\ker(K)}x$ takes into account that $f(0) \neq 0$ and makes the identity $\mathrm{id} = f(K^*K)$ for $f \equiv 1$ correct.

We observe that the series always converges (since f is only evaluated on the bounded interval $[0, \|K\|^2]$) and we also get that

$$\|f(K^*K)\| \leq \sup_n |f(\sigma_n^2)| + f(0) \leq 2 \sup_{\lambda\in[0,\|K\|^2]} |f(\lambda)| < \infty$$

which shows that $f(K^*K) \in L(X, X)$.

With functional calculus we can write

$$R_\alpha y = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n \langle y, u_n\rangle v_n = \varphi_\alpha(K^*K)K^*y.$$

Example 7.2 (Absolute value of a compact operator). For $f(t) = t$ we get that $f(K^*K) = K^*K$ and for $f(t) = \sqrt{t}$ we define $|K| := f(K^*K)$. It holds that

$$|K|x = \sum_n \sigma_n \langle x, v_n\rangle v_n.$$

△

We state some properties of the absolute value of an operator as we will use it later when we derive convergence rates in abstract smoothness spaces in Sections 9 and 10.

Lemma 7.3 (Properties of functional calculus). *Let $K \in K(X, Y)$.*

(i) For $s, r > 0$ it holds that $|K|^{r+s} = |K|^r|K|^s$.

(ii) For all $r > 0$ the operator $|K|^r$ is self-adjoint.

(iii) For all $x \in X$ it holds that $\||K|x\|_Y = \|Kx\|_y$.

(iv) It holds that $\mathrm{rg}(|K|) = \mathrm{rg}(K^*)$.

Proof. (i) This is a direct computation (using that v_n is orthonormal)

$$\begin{aligned}|K|^{r+s}x &= \sum_n \sigma_n^{r+s} \langle x, v_n\rangle v_n = \sum_n \sigma_n^r \left\langle \sum_m \sigma_m^s \langle x, v_m\rangle v_m, v_n \right\rangle v_n \\ &= \sum_n \sigma_n^r \langle |K|^s x, v_n\rangle v_n.\end{aligned}$$

(ii) Again a direct computation

$$\langle |K|^r x, z\rangle = \sum_n \sigma_n^r \langle x, v_n\rangle \langle v_n, z\rangle = \langle x, |K|^r z\rangle .$$

(iii) Using the first two points we get

$$\||K|x\|_X^2 = \langle |K|x, |K|x\rangle = \langle |K|^2 x, x\rangle = \langle K^*Kx, x\rangle = \langle Kx, Kx\rangle = \|Kx\|_Y^2.$$

(iv) If (σ_n, u_n, v_n) is the singular system of K, then K^* has the singular system (σ_n, v_n, u_n) and $|K|$ has the singular system (σ_n, v_n, v_n). Now note that $x \in \operatorname{rg}(K^*)$ exactly if $Kx \in \operatorname{rg}(KK^*)$ and $x \perp \ker(K)$. The Picard condition (Theorem 4.8) for $Kx \in \operatorname{rg}(KK^*)$ is

$$\infty > \sum_n \sigma_n^{-4} |\langle Kx, u_n\rangle|^2 = \sum_n \sigma_n^{-4} |\langle x, K^* u_n\rangle|^2 = \sum_n \sigma_n^{-2} |\langle x, v_n\rangle|^2$$

which is exactly the Picard condition for $x \in \operatorname{rg}(|K|)$ (for which $x \perp \ker(K)$ is necessary anyway).

□

We will investigate regularization methods of the form (R). The following definition well be useful, as we will see:

Definition 7.4 (Regularizing filter)**.** Let $K \in K(X, Y)$ with $\kappa = \|K\|^2$ and SVD (σ_n, u_n, v_n). A family $\varphi_\alpha : [0, \kappa] \to \mathbb{R}$ of piecewise continuous and bounded functions is called *regularizing filter* if it fulfills

(i) For all $\lambda \in]0, \kappa]$ it holds that

$$\varphi_\alpha(\lambda) \xrightarrow{\alpha \to 0} \frac{1}{\lambda}.$$

(ii) There exists $C_\varphi > 0$ such that for all $\lambda \in]0, \kappa]$ and $\alpha > 0$ it holds that

$$\lambda |\varphi_\alpha(\lambda)| \leq C_\varphi.$$

Now we aim to prove that regularizing filters give rise to convergent regularization methods. First we collect three useful facts in a lemma:

Lemma 7.5 (Fundamental lemma of regularization theory)**.** *If φ_α is a regularizing filter and $R_\alpha = \varphi_\alpha(K^*K)K^*$ we have*

(1) $\|KR_\alpha\| \leq C_\varphi$

(2) $\|R_\alpha\| \leq \sqrt{C_\varphi} \sup_{\lambda \in]0, \|K\|^2]} \sqrt{|\varphi_\alpha(\lambda)|}$

(3) $K^\dagger y - R_\alpha y = \sum_n (1 - \sigma_n^2 \varphi_\alpha(\sigma_n)^2) \left\langle x^\dagger, v_n \right\rangle v_n$ *for $y \in D(K^\dagger)$ and $x^\dagger = K^\dagger y$.*

Proof. We first compute

$$KR_\alpha y = K\varphi_\alpha(K^*K)K^*y = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n \langle y, u_n\rangle Kv_n = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n^2 \langle y, u_n\rangle u_n$$

and then get that

$$\begin{aligned}\|KR_\alpha y\|_Y^2 &= \sum_n |\varphi_\alpha(\sigma_n^2)\sigma_n^2 \langle y, u_n\rangle|^2 \\ &\leq \sup_n |\varphi_\alpha(\sigma_n^2)\sigma_n^2|^2 \|y\|_Y^2\end{aligned}$$

which, by definition of the constant C_φ, implies the claim (1). For the claim (2) compute

$$\begin{aligned}\|R_\alpha y\|_X^2 &= \langle R_\alpha y, R_\alpha y\rangle = \sum_n \varphi_\alpha(\sigma_n^2)\sigma_n \langle y, u_n\rangle \langle R_\alpha y, v_n\rangle \\ &= \sum_n \varphi_\alpha(\sigma_n^2) \langle y, u_n\rangle \langle R_\alpha y, K^* u_n\rangle \\ &= \sum_n \varphi_\alpha(\sigma_n^2) \langle y, u_n\rangle \langle KR_\alpha y, u_n\rangle \\ &\leq \sup_n |\varphi_\alpha(\sigma_n^2)| \sum_n \langle y, u_n\rangle \langle KR_\alpha y, u_n\rangle \\ &\leq \sup_n |\varphi_\alpha(\sigma_n^2)| \left(\sum_n \langle y, u_n\rangle^2\right)^{1/2} \left(\sum_n \langle KR_\alpha y, u_n\rangle\right)^{1/2} \\ &\qquad \text{(by Cauchy-Schwarz)} \\ &\leq \sup_n |\varphi_\alpha(\sigma_n^2)| \|y\| \|KR_\alpha y\| \\ &\leq \sup_n |\varphi_\alpha(\sigma_n^2)| C_\varphi \|y\|_Y^2 \qquad \text{(by claim (1))}\end{aligned}$$

which proves the claim. Finally, for claim (3) we note that if $x^\dagger = K^\dagger y$, then $K^*Kx^\dagger = K^*y$ and thus

$$R_\alpha y = \varphi_\alpha(K^*K)K^*y = \varphi_\alpha(K^*K)K^*Kx$$

and we get

$$K^\dagger y - R_\alpha y = (\mathrm{id} - \varphi_\alpha(K^*K)K^*K)x^\dagger = \sum_n (1 - \sigma_n^2\varphi_\alpha(\sigma_n^2)) \left\langle x^\dagger, v_n\right\rangle v_n.$$

□

Theorem 7.6 (Regularization with regularizing filters)**.** *Let φ_α be a regularizing filter and $R_\alpha = \varphi_\alpha(K^*K)K^*$. Then it holds for all $y \in D(A^\dagger)$ that*

$$R_\alpha y \xrightarrow{\alpha\to 0} K^\dagger y.$$

Proof. By Lemma 7.5 (3) we have

$$\|K^\dagger y - R_\alpha y\|_X^2 = \sum_n (1 - \sigma_n^2\varphi_\alpha(\sigma_n^2))^2 |\left\langle x^\dagger, v_n\right\rangle|^2.$$

Since $\varphi_\alpha(\lambda) \to 1/\lambda$ for $\alpha \to 0$ we get $(1 - \sigma_n^2\varphi_\alpha(\sigma_n^2)) \to 0$ for $\alpha \to 0$. Moreover, $|1 - \sigma_n^2\varphi_\alpha(\sigma_n^2)| \leq 1 + C_\varphi$ and hence, the convergence $R_\alpha y \to K^\dagger y$ follows from the dominated convergence theorem (Theorem 3.5). □

Next we will show that there is a general strategy to construct a parameter choice rule that turns regularizing filters into convergent regularization methods.

8 Parameter choice and error estimates

We will now investigate the problem of a-priori parameter choice for general spectral regularization methods of the type $R_\alpha = \varphi_\alpha(K^*K)K^*$ for a regularizing filter φ_α.

Theorem 8.1. *Let $K^\dagger$ be non-continuous and R_α be a regularization of $K^\dagger$. Then it holds: An a priori parameter choice $\alpha(\delta)$ fulfills that $\|R_{\alpha(\delta)}y^\delta - x^\dagger\|_X \to 0$ for $\delta \to 0$ exactly if*

(i) $\alpha(\delta) \to 0$ *for* $\delta \to 0$, *and*

(ii) $\delta \sup_{0<\lambda\leq\|K\|^2} \left\{ \sqrt{|\varphi_\alpha(\lambda)|} \right\} \to 0$ *for* $\delta \to 0$.

Proof. We start with our standard error decomposition

$$\|R_\alpha y^\delta - x^\dagger\|_X \leq \|R_\alpha\|\delta + \|R_\alpha y^\dagger - K^\dagger y^\dagger\|_X.$$

By Theorem 7.6 and $\alpha(\delta) \to 0$ we get that the second term on right hand side goes to zero. By Lemma 7.5 (2) we have that $\|R_\alpha\| \leq \sup_{\lambda\in]0,\|K\|^2]} \left\{ \sqrt{|\varphi_\alpha(\lambda)|} \right\}$ and hence, the first term goes to zero as well.

Conversely, assume that either (i) or (ii) does not hold. Let's start with the case where (i) does not hold. Then $R_{\alpha(\delta)}$ does not converge to $K^\dagger$ pointwise. Hence, we can even set $y^\delta = y \in D(K^\dagger)$, $x^\dagger = K^\dagger y$ and get that $\|R_{\alpha(\delta)}y^\delta - x^\dagger\|_X = \|R_{\alpha(\delta)}y - K^\dagger y\| \not\to 0$.

If (i) is fulfilled, but (ii) not, there exists δ_n with $\delta_n \overset{n\to\infty}{\longrightarrow} 0$ such that $\delta_n\|R_{\alpha(\delta_n)}\| > \epsilon$ for some ϵ. Hence, there exists a sequence $z_n \in Y$ with $\|z_n\|_Y = 1$ and $\delta_n\|R_{\alpha(\delta_n)}z_n\|_X > \epsilon$. Now let $y \in D(K^\dagger)$ and set $y_n := y + \delta_n z_n$. Then $\|y - y_n\| = \delta$, but

$$R_{\alpha(\delta_n)}y_n - K^\dagger y = (R_{\alpha(\delta_n)}y - K^\dagger y) + \delta_n R_{\alpha(\delta_n)}z_n \not\to 0.$$

□

Now we aim for more sophisticated error estimates. What is needed, is a better estimate of the approximation error. As we have seen in Theorem 6.5, this is not possible without additional assumptions.

By Lemma 7.5 (3) we have that=

$$K^\dagger y^\dagger - R_\alpha y^\dagger = \sum_n (1 - \sigma_n^2\varphi_\alpha(\sigma_n^2)) \left\langle x^\dagger, v_n \right\rangle v_n \qquad (*)$$

if $x^\dagger = K^\dagger y^\dagger$. However, bounding this error just in terms of α is not possible, since the decay of the terms $\langle x^\dagger, v_n\rangle$ is not known (the sequence has to be square summable, but that's basically all we know). Here is a simple way to get a useful error bound:

We assume that our true solution $x^\dagger$ is in $\mathrm{rg}(K^*)$.

How does that help? Well, in this case we have some $w^\dagger$ with $x^\dagger = K^* w^\dagger$, we get from (*)

$$\begin{aligned}\|K^\dagger y^\dagger - R_\alpha y^\dagger\|_X^2 &= \sum_n (1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2 |\left\langle x^\dagger, v_n\right\rangle|^2 \\ &= \sum_n (1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2 |\left\langle K^* w^\dagger, v_n\right\rangle|^2 \\ &= \sum_n (1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2 |\left\langle w^\dagger, K v_n\right\rangle|^2 \\ &= \sum_n (1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2 \sigma_n^2 |\left\langle w^\dagger, v_n\right\rangle|^2. \qquad (**)\end{aligned}$$

Now we may get an error bound, if we can control the coefficients $(1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2\sigma_n^2$. Expressed in the variable $\lambda = \sigma^2$ this says that we have to control the function $\lambda \mapsto (1-\lambda\varphi_\alpha(\lambda))\sqrt{\lambda}$. Let us investigate the situation for the truncated SVD and Tikhonov regularization:

Example 8.2.

1. For the truncated SVD we have $\varphi_\alpha(\lambda) = 1/\lambda$ for $\lambda \geq \alpha$ and $= 0$ else. So we get

   $$(1-\lambda\varphi_\alpha(\lambda))\sqrt{\lambda} = \begin{cases} 0 & : \quad \lambda \geq \alpha \\ \sqrt{\lambda} & : \quad \lambda < \alpha \end{cases}$$

   and thus, $(1-\lambda\varphi_\alpha(\lambda))\sqrt{\lambda} \leq \sqrt{\alpha}$ or, equivalently

   $$(1-\sigma^2\varphi_\alpha(\sigma^2))^2\sigma^2 \leq \alpha.$$

   Using this in (**), we obtain for the approximation error

   $$\begin{aligned}\|K^\dagger y^\dagger - R_\alpha y^\dagger\|_X^2 &= \sum_n (1-\sigma_n^2\varphi_\alpha(\sigma_n^2))^2\sigma_n^2 |\left\langle w^\dagger, v_n\right\rangle|^2 \\ &\leq \alpha \sum_n |\left\langle w^\dagger, v_n\right\rangle|^2 \leq \alpha \|w\|_Y^2\end{aligned}$$

   and thus

   $$\|K^\dagger y^\dagger - R_\alpha y^\dagger\|_X \leq \sqrt{\alpha}\|w^\dagger\|_Y.$$

   Note the we change the threshold in comparison to Example 5.7: There we took $1/\sigma_n$ if $\sigma_n < \alpha$ and here we use $1/\sigma_n = \sigma_n/\sigma_n^2$ if $\sigma_n^2 > \alpha$, i.e. for $\sigma_n > \sqrt{\alpha}$.

2. For Tikhonov regulrization we have $\varphi_\alpha(\lambda) = 1/(\lambda+\alpha)$ and we get

   $$(1-\lambda\varphi_\alpha(\lambda))\sqrt{\lambda} = (1-\tfrac{\lambda}{\lambda+\alpha})\sqrt{\lambda} = \tfrac{\alpha\sqrt{\lambda}}{\lambda+\alpha}.$$

   We want to maximize the right hand side over $\lambda \geq 0$. To this end we define $f(\lambda) = \frac{\sqrt{\lambda}}{\lambda+\alpha}$, calculate $f'(\lambda) = \frac{1}{2}\frac{\alpha\lambda^{-1/2}-\lambda^{1/2}}{(\lambda+\alpha)^2}$ and see that $f'(\lambda) = 0$ for $\lambda = \alpha$. Hence, we get that $\frac{\alpha\sqrt{\lambda}}{\lambda+\alpha} \leq \frac{\alpha\sqrt{\alpha}}{\alpha+\alpha} = \frac{1}{2}\sqrt{\alpha}$.

   Again, using this in (**) gives $\|K^\dagger y^\dagger - R_\alpha y^\dagger\|_X^2 \leq \frac{\alpha}{4}\|w^\dagger\|_Y^2$ and thus

   $$\|K^\dagger y^\dagger - R_\alpha y^\dagger\|_X \leq \tfrac{\sqrt{\alpha}}{2}\|w^\dagger\|_Y.$$

We conclude: If the unknown solution fulfills $x^\dagger \in \mathrm{rg}(K^*)$, then the total error of both the truncated SVD and Tikhonov regularization can be estimated by

$$\|x_\alpha^\delta - x^\dagger\|_X \leq \delta \|R_\alpha\| + \sqrt{\alpha} C$$

for some constant C (independent of α and δ). This our first quantitative error bound, i.e. an upper bound for the total error that is explicit in δ and α. We also know that $\|R_\alpha\| \leq \sup_\lambda \sqrt{|\varphi_\alpha(\lambda)|}$ and for both the truncated SVD and Tikhonov regularization we conclude by a simple calculation that $\|R_\alpha\| \leq 1/\sqrt{\alpha}$. This gives the even more explicit error estimate

$$\|x_\alpha^\delta - x^\dagger\|_X \leq \tfrac{\delta}{\sqrt{\alpha}} + \sqrt{\alpha} C.$$

Now we can even choose the regularization parameter α in an optimal way: We can minimize the right hand side over α and see that the minimum is attained for $\alpha(\delta) = \delta/C$ and this gives the *error estimate*

$$\|x_\alpha^\delta - x^\dagger\|_X \leq 2\sqrt{C}\sqrt{\delta}.$$

Even if we do not know the constant C, we could still set α proportional to δ, i.e. $\alpha(\delta) = c\delta$ for some constant c, and obtain

$$\|x_\alpha^\delta - x^\dagger\|_X \leq \mathcal{O}(\delta^{1/2}).$$

Results of this form are called *convergence rates* of regularization methods.

Assumption on the unknown solution $x^\dagger$ such as $x^\dagger \in \mathrm{rg}(K^*)$ are called *source conditions*. △

We will come back to error estimates and convergence rates later.

Now we briefly discuss the most popular a posteriori parameter choice rule: Recall that our standing assumption is that the true data $y^\dagger$ and our measured data y^δ always fulfill $\|y^\dagger - y^\delta\|_Y \leq \delta$. The main idea now is to look at the residuum for a reconstruction $R_\alpha y^\delta$, i.e. to consider

$$\|KR_\alpha y^\delta - y^\delta\|_Y.$$

The residuum for the minimum norm solution $x^\dagger$ fulfills $\|Kx^\dagger - y^\delta\|_Y = \|y^\dagger - y^\delta\|_Y \leq \delta$, thus it seems reasonable to not aim for a smaller residuum for any other reconstruction. This is the idea of the following:

> **Morozov's discrepancy principle:** For some $\delta > 0$ and y^δ with $\|y^\dagger - y^\delta\|_Y \leq \delta$ choose $\alpha = \alpha(\delta, y^\delta)$ (as large as possible) such that
>
> $$\|KR_\alpha y^\delta - y^\delta\|_Y \leq \tau\delta$$
>
> for some $\tau > 1$.

We want to choose α as large as possible, to have the most stable reconstruction.

Remark 8.3. This principle does not work without assumptions: For $y^\dagger \in \operatorname{rg}(K)^\perp \setminus \{0\}$ and exact data $y^\delta = y^\dagger$ (i.e. $\delta = 0$) even the minimum norm solution $x^\dagger$ fulfills

$$\|Kx^\dagger - y^\delta\|_Y = \|KK^\dagger y^\dagger - y^\dagger\|_Y = \|P_{\overline{\operatorname{rg} K}} y^\dagger - y^\dagger\| = \|y^\dagger\|_Y > 0 = \tau\delta.$$

Therefore one usually assumes that the range $\operatorname{rg} K$ is dense in Y (since then $\operatorname{rg} K^\perp = \{0\}$).

For a practical realization of Morozov's discrepancy principle one usually defines a decreasing sequence $\alpha_n \to 0$, computes $R_{\alpha_n} y^\delta$ for $n = 1, 2, \dots$ and stops when $\|R_{\alpha_n} y^\delta - y^\delta\|_Y \leq \tau\delta$ for the first time. This always works if the range of K is dense:

Theorem 8.4. *Let $R_\alpha = \varphi_\alpha(K^*K)K^*$ be a regularization of $K^\dagger$, $\operatorname{rg} K$ dense in Y, α_n be a strictly decreasing null sequence and $\tau > 1$. Then it holds: For all $y^\dagger \in D(K^\dagger)$, all $\delta > 0$ and y^δ with $\|y^\dagger - y^\delta\|_Y \leq \delta$ there exists an n^* such that for all $n < n^*$*

$$\|KR_{\alpha_{n^*}} y^\delta - y^\delta\|_Y \leq \tau\delta < \|KR_{\alpha_n} y^\delta - y^\delta\|_Y.$$

Proof. We study $\|KR_\alpha y^\delta - y^\delta\|$ in dependence on α. Using the filter we get

$$\begin{aligned}\|KR_\alpha y^\delta - y^\delta\|^2 &= \sum_n (1 - \sigma_n^2 \varphi_\alpha(\sigma_n^2))^2 |\langle y^\delta, u_n \rangle|^2 + \|P_{\operatorname{rg}(K)^\perp}(y^\delta)\|^2 \\ &= \sum_n (1 - \sigma_n^2 \varphi_\alpha(\sigma_n^2))^2 |\langle y^\delta, u_n \rangle|^2\end{aligned}$$

since $\operatorname{rg}(K)^\perp = \{0\}$. As we have seen in Theorem 7.6, the right hand side goes to zero which proves the claim. □

If we do not assume that $\operatorname{rg}(K)$ is dense, we only get that $\|KR_\alpha y^\delta - \delta\| \to \|P_{\operatorname{rg}(K)^\perp} y^\delta\| \leq \|y^\delta\|$. Hence, we need that $\|y^\delta\| \leq \delta$ for Morozov's discrepancy principle to work. This seems reasonable: There should be "more signal than noise".

We will show later that Morozov's principle does indeed give a convergent regularization and now make a few remarks on *heuristic* rules: First and foremost, there a negative result, named *Bakushinkii veto*.

Theorem 8.5 (Bakushinkii veto)**.** *Let R_α be a regularization for $K^\dagger$. If there exists a heuristic choice rule $\alpha = \alpha(y^\delta)$ such that (R_α, α) is a convergent regulariztion, then $K^\dagger$ is continuous.*

Proof. Let $\alpha : Y \to]0, \infty[$ by such a parameter choice. Now let $y \in D(K^\dagger)$ and consider $y_n \in D(K^\dagger)$ with $y_n \to y$. But then (trivially) $\|y_n - y_n\|_Y \leq \delta$ for every $\delta > 0$ and by definition of convergent regularization we get $\|K^\dagger y_n - R_{\alpha(y_n)} y_n\|_Y = 0$, i.e. $R_{\alpha(y_n)} y_n = K^\dagger y_n$. Moreover, we have for $\delta_n = \|y - y_n\|_Y$ (again by definition of convergent regulariztion) that

$$K^\dagger y_n = R_{\alpha(y_n)} y_n \to K^\dagger y$$

which proves continuity of $K^\dagger$. □

This theorem shows that heuristic rules only exist for inverse problems that aren't ill-posed. Despite this negative result there are several heuristic rules that work remarkably well in practice. This phenomenon is still not fully understood. One explanation could be that one usually faces a perturbation $y^\delta = y^\dagger + \eta$ by noise η in practice, but our theory uses general $\eta \in Y$, i.e. also perturbations which do not look like noise at all are considered. Some rules that work well in practice are the quasi-optimality principle, the Hanke-Raus rule, the L-curve method, and generalized cross validation.

9 Convergence rates and smoothness spaces

In this section we will focus on the question on how to establish convergence rates, i.e. under what circumstances we can find a function $\psi :]0,\infty[\to]0,\infty[$ with $\psi(\delta) \to 0$ for $\delta \to 0$ such that

$$\|R_\alpha y^\delta - K^\dagger y^\dagger\|_X \leq \psi(\delta)$$

for a-priori or a-posteriori parameter choice rules. Recall that by Theorem 6.5 this can not hold without any further assumptions, i.e. it can't be true if we consider $y^\dagger$ arbitrary in the range of K (or, equivalently, $x^\dagger$ arbitrary in X).

However, we have seen in Example 8.2 that such a result can be achieved for the truncated SVD and Tikhonov regularization (for the a-priori choice rule $\alpha(\delta) = C\sqrt{\delta}$ and $\psi(\delta) = C\sqrt{\delta}$) if we assume $x^\dagger \in \operatorname{rg}(K^*) \subsetneq X$. This assumption is some kind of "abstract smoothness assumption". The notion of smoothness that we will need will be formulated in terms of the operator K and may look confusing at first:

Definition 9.1. Let X, Y be Hilbert spaces and $K \in K(X,Y)$. For $\nu \geq 0$ we define the subspaces $X^\nu \subset X$ as

$$X^\nu := \operatorname{rg}(|K|^\nu) = \left\{ |K|^\nu z \;\middle|\; z \in \ker(K)^\perp \right\}.$$

Some first observations:

- For $\nu = 2k$ we have that
$$X^{2k} = \operatorname{rg}(|K|^{2k}) = \operatorname{rg}(\sqrt{K^*K}^{2k}) = \operatorname{rg}((K^*K)^k).$$
- If $\nu > \mu$, then $|K|^\nu = |K|^\mu |K|^{\nu-\mu}$, i.e. $\operatorname{rg}(|K|^\nu) \subset \operatorname{rg}(|K|^\mu)$ and thus $X^\nu \subset X^\mu$, i.e. the spaces get smaller, the larger the ν. The boundary case is $X^0 = \ker(K)^\perp$.
- The spaces X^ν are characterized by summability assumptions of the coefficients in the singular basis: If there is z such that $x = |K|^\nu z$ we have by definition of $|K|^\nu$ that $x = \sum_n \sigma_n^\nu \langle z, v_n\rangle u_n$ and hence
$$\sum_n \sigma_n^{-2\nu} |\langle x, v_n\rangle|^2 = \sum_n \sigma_n^{-2\nu}\sigma_n^{2\nu} |\langle z, v_n\rangle|^2 = \|z\|_X^2 < \infty.$$
In other words: For $x \in X^\nu$ we need that the sequence $\langle x, v_n\rangle$ decays fast enough such that the decay of $|\langle x, v_n\rangle|^2$ compensates the growth of $\sigma_n^{-2\nu}$.

The last observation motivates the following definition:

Definition 9.2. On X^ν we define the norm

$$\|x\|_\nu^2 := \|z\|_X^2 = \sum_n \sigma_n^{-2\nu} |\langle x, v_n\rangle|^2, \quad X^\nu = \{x \mid \|x\|_\nu < \infty\}$$

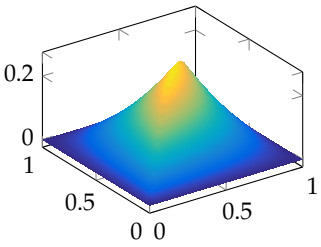
We call these norms *ν-norms*.

Note that the notation X^ν and $\|\cdot\|_\nu$ do not include their dependency on K. Since these spaces are only considered for one K at a time, this usually does not lead to confusion.

In a certain sense, these norms measure smoothness, i.e. some ν-norm of x is finite, only if x is somehow smooth and the larger we can take ν while $\|x\|_\nu$ stays finite, the smoother x is. A bit more precise: The X^ν spaces demand a certain decay of the expansion coefficients with respect to the singular basis; the faster the coefficients $\langle x, v_n\rangle$ decay, the "smoother" the x is. As we have already observed: The singular vectors v_n with large n are usually highly oscillating and thus, they can not contribute much to a function that is in some X^ν with large ν. In the following example we can make this a bit more precise:

Example 9.3. We consider the following integral operator:

$$Kf(t) = \int_0^1 k(t,s)f(s)\mathrm{d}s, \quad k(t,s) = \begin{cases} t(1-s) & : \quad t \le s \\ s(1-t) & : \quad s \le t \end{cases}.$$



This operator maps $L^2([0,1])$ into itself and is compact (cf. Example 4.3). One can show that

$$Kf = g \iff \begin{cases} -g'' = f, \text{ and} \\ g(0) = g(1) = 0 \end{cases}.$$

Not a full proof but the first calculations: The equation $g = Kf$ is

$$\begin{aligned} g(t) &= \int_0^1 k(s,t)f(s)\mathrm{d}s \\ &= \int_0^t s(1-t)f(s)\mathrm{d}s + \int_t^1 t(1-s)f(s)\mathrm{d}s \\ &= (1-t)\int_0^t sf(s)\mathrm{d}s + t\int_t^1 (1-s)f(s)\mathrm{d}s. \end{aligned}$$

We directly see that $g(0) = g(1) = 0$ follows. We take the derivative on both sides (using the known rules) gives

$$\begin{aligned} g'(t) &= -\int_0^t sf(s)\mathrm{d}s + (1-t)tf(t) + \int_t^1 (1-s)f(s)\mathrm{d}s - t(1-t)f(t) \\ &= -\int_0^t sf(s)\mathrm{d}s + \int_t^1 (1-s)f(s)\mathrm{d}s. \end{aligned}$$

The second derivative is

$$g''(t) = -tf(t) - (1-t)f(t) = -f(t).$$

In principle one can argue similarly in the opposite direction as well.

Moreover K is selfadjoint, since $k(s,t) = k(t,s)$. One can also show that the SVD of K is given by

$$\sigma_n = \tfrac{1}{(\pi n)^2}, \quad u_n(t) = v_n(t) = \sqrt{2}\sin(\pi n t).$$

This allows us to characterize the spaces X^ν a quite explicitly: By Definition 9.2 we have

$$f \in X^\nu \iff \|f\|_\nu^2 = 2\pi^{4\nu} \sum_n n^{4\nu} |\langle f, \sin(\pi n \cdot)\rangle|^2 < \infty.$$

Now one can show the following: For $f \in C^{2\nu}([0,1])$ with $f^{(2k)}(0) = f^{(2k)}(1) = 0$ for $k = 0, \dots, \nu - 1$ it holds that

$$\|f\|_\nu = \|f^{(2\nu)}\|_{L^2}.$$

We start from the right and use that $v_n(t) = \sqrt{2}\sin(\pi n t)$ is an orthonormal basis of $L^2([0,1])$ to get

$$\|f^{(2\nu)}\|_{L^2}^2 = \sum_n |\left\langle f^{(2\nu)}, v_n \right\rangle|^2.$$

For the inner products we use integration by parts two times, the boundary conditions of f and that the sine functions vanish at the boundary to get

$$\begin{aligned}
\left\langle f^{(2\nu)}, v_n \right\rangle &= \int_0^1 f^{(2\nu)}(t)\sqrt{2}\sin(\pi n t)\mathrm{d}t \\
&= \underbrace{f^{(2\nu-1)}(t)\sqrt{2}\sin(\pi n t)\Big|_0^1}_{=0} - \int_0^1 f^{(2\nu-1)}(t)\sqrt{2}(\pi n)\cos(\pi n t)\mathrm{d}t \\
&= \underbrace{f^{(2\nu-2)}(t)\sqrt{2}(\pi n)\cos(\pi n t)\Big|_0^1}_{=0} + \int_0^1 f^{(2\nu-2)}\sqrt{2}(\pi n)^2\sin(\pi n t)\mathrm{d}t \\
&= (\pi n)^2 \left\langle f^{(2\nu-2)}, v_n \right\rangle.
\end{aligned}$$

Recursively, this gives $\left\langle f^{(2\nu)}, v_n \right\rangle = (\pi n)^{2\nu} \langle f, v_n \rangle$ and hence

$$\|f^{(2\nu)}\|_{L^2}^2 = 2\pi^{4\nu} \sum_n n^{4\nu} |\langle f, \sin(\pi n \cdot)\rangle|^2 = \|f\|_\nu^2$$

as claimed.

This can be used to rigorously prove that it holds

$$X^\nu = \overline{\left\{ f \in C^{(2\nu)}([0,1]) \;\middle|\; f^{(2k)}(0) = f^{(2k)}(1) = 0 \right\}}^{\|\cdot\|_\nu},$$

i.e. the space X^ν is the closure of the space of 2ν-times continuous differentiable functions with respective boundary conditions with respect to the ν-norm. △

Our aim is to construct methods (R_α, α) such that the total error $\|R_\alpha y^\delta - x^\dagger\|_X$ is small. First we will establish a baseline and analyze the question of "how good can a reconstruction method $R : Y \to X$ be in general". We introduce a little more notation:

Definition 9.4 (Worst case error in ν-spaces)**.** Let $K \in K(X,Y)$ and $R : Y \to X$ continuous with $R0 = 0$. We define

$$E_\nu(\delta,\rho,R) = \sup\left\{ \|Ry^\delta - x^\dagger\|_X \;\middle|\; x^\dagger \in X^\nu, \|Kx^\dagger - y^\delta\|_Y \le \delta, \|x^\dagger\|_\nu \le \rho \right\}.$$

This quantity is the *worst case total error of the method R over solutions in an X^ν-ball of radius ρ and noise level δ*. Furthermore we define

$$E_\nu(\delta,\rho) = \inf\{E_\nu(\delta,\rho,R) \mid R : Y \to X \text{ continous with } R0 = 0\}.$$

This is the *best possible worst case error of any method over solutions in an X^ν-ball of radius ρ and noise level δ*.

This "best possible worst case error" may look a bit weird, but is actually not hard to quantify:

Theorem 9.5. *For $K \in K(X,Y)$ it holds that*

$$E_\nu(\delta,\rho) \ge \sup\{\|x\|_X \mid \|Kx\|_Y \le \delta, \|x\|_\nu \le \rho\} =: e_\nu(\delta,\rho).$$

The quantity on the right hand side has a simple explanation: It is a so-called *modulus of continuity* of $K^\dagger$ restricted to some set defined by δ, ν and ρ. These three parameters have the following meaning:
δ: Noise level
ν: Degree of smoothness
ρ: "Largeness" in the smoothness class.
The noise level is often available (or can be estimated), the smoothness may be guessed, but the largeness in the smoothness class is basically never known.

Proof. Let $x \in X^\nu$ with $\|Kx\|_Y \le \delta$ and $\|x\|_\nu \le \rho$ and $R : Y \to X$ be arbitrary (given the conditions). We set $y = Kx$ and $y^\delta = 0$. Then this x is in the feasible set for the supremum in the definition of $E_\nu(\delta,\rho,R)$ and thus $E_\nu(\delta,\rho,R) \ge \|x\|_X$. Taking the supremum over all these x we get the inequality $e_\nu(\delta,\rho) \le E_\nu(\delta,\rho,R)$. The claim follows by taking the infimum over all R. □

The following theorem shows, how large the lower bound $e_\nu(\delta,\rho)$ of the best worst case error can be:

Theorem 9.6. *Let $K \in K(X,Y)$. Then it holds for all $\nu,\rho,\delta > 0$ that*

$$e_\nu(\delta,\rho) \le \delta^{\frac{\nu}{\nu+1}} \rho^{\frac{1}{\nu+1}}.$$

Moreover, there exists a sequence δ_n with $\delta_n \overset{n\to\infty}{\longrightarrow} 0$ such that there is equality along that sequence.

Proof. Let $x \in X^\nu$ with $\|x\|_\nu \le \rho$ and $\|Kx\|_Y \le \delta$. Then there is $z \in X$ such that $x = |K|^\nu z$. We would like to estimate $\|x\|_X = \||K|^\nu z\|_X$ in terms of $\|z\|_X = \|x\|_\nu$ and $\|Kx\|_Y$. To that end we use the following result (also known as *interpolation inequality*): For $r > s \ge 0$ and all x it holds that $\||K|^s x\|_X \le \||K|^r x\|_X^{\frac{s}{r}} \|x\|_X^{1-\frac{s}{r}}$.

The definition of $|K|^s$ gives

$$\||K|^s x\|_X = \sum_n \sigma_n^{2s} |\langle x, v_n\rangle|^2.$$

Now we define sequences $a_n = \sigma_n^{2s}|\langle x, v_n\rangle|^{2\frac{s}{r}}$ and $b_n = |\langle x, v_n\rangle|^{2-2\frac{r}{s}}$ and numbers $p = r/s$ and $q = r/(r-s)$. We use the Hölder inequality to get

$$\begin{aligned}\||K|^s x\|_X^2 &= \sum_n \sigma_n^{2s}|\langle x, v_n\rangle|^2 \\ &= \sum_n a_n b_n \le \left(\sum_n a_n^p\right)^{1/p} \cdot \left(\sum_n b_n^q\right)^{1/q} \\ &= \left(\sum_n \sigma_n^{2r}|\langle x, v_n\rangle|^2\right)^{s/r} \cdot \left(\sum_n |\langle x, v_n\rangle|^2\right)^{(r-s)/r} \\ &= \||K|^r x\|_X^{2s/r}\|x\|_X^{2(r-s)/r}\end{aligned}$$

which proves the claim.

We use this claim with $s = \nu$ and $r = \nu + 1$ and

$$\begin{aligned}\|x\|_X = \||K|^\nu z\|_X &\le \||K|^{\nu+1} z\|_X^{\frac{\nu}{\nu+1}} \|z\|_X^{\frac{1}{\nu+1}} \\ &= \|K\underbrace{|K|^\nu z}_{=x}\|_X^{\frac{\nu}{\nu+1}} \|z\|_X^{\frac{1}{\nu+1}} \le \delta^{\frac{\nu}{\nu+1}} \rho^{\frac{1}{\nu+1}}.\end{aligned}$$

For the equality we set $\delta_n = \rho\sigma_n^{\nu+1}$ and $x_n = \rho|K|^\nu v_n$. One can show that this gives indeed equality.

Note that by definition of $|K|^\nu$ we have $x = \rho\sigma_n^\nu v_n$ and by definition of the ν-norm we have that $\|x\|_\nu = \|\rho v_n\|_X = \rho$ and $\|Kx\|_Y = \|\rho\sigma_n^\nu K v_n\|_Y = \rho\sigma_n^\nu\|\sigma_n u_n\| = \rho\sigma_n^{\nu+1} = \delta_n$ as needed. From $\delta_n = \rho\sigma_n^{\nu+1}$ we get that $\sigma_n = (\delta_n/\rho)^{1/(\nu+1)}$ and thus

$$\|x\|_X = \rho\sigma_n^\nu = \rho\left(\tfrac{\delta_n}{\rho}\right)^{\frac{\nu}{\nu+1}} = \delta_n^{\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}}$$

as desired.

□

10 Convergence rates for spectral regularization

The above Theorems 9.5 and 9.6 give a benchmark with which we can compare regularization methods: They can not do better than the right hand side $\delta^{\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}}$ for data $x^\dagger$ from X^ν with $\|x^\dagger\|_\nu \leq \rho$.

We fix this in the following definition:

Definition 10.1. A regularization method (R_α, α) is called *optimal* for parameters ρ and ν, if for all $x^\dagger$ with $\|x^\dagger\|_\nu \leq \rho$ and y^δ with $\|Kx^\dagger - y^\delta\|_Y \leq \delta$ it holds that

$$\|R_\alpha y^\delta - x^\dagger\|_X = \delta^{\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}}.$$

We call the method *order optimal* for parameters ρ and ν if there exists some C such that for all $x^\dagger$ as above it holds that

$$\|R_\alpha y^\delta - x^\dagger\|_X = C\delta^{\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}}.$$

Finally, we call a method *order optimal* for ν, if for all $x^\dagger \in X^\nu$ and y^δ with $\|Kx^\dagger - y^\delta\| \leq \delta$ there exists C such that

$$\|R_\alpha y^\delta - x^\dagger\| \leq C\delta^{\frac{\nu}{\nu+1}}.$$

The assumptions $\|x^\dagger\|_\nu \leq \rho$ or $x^\dagger \in X^\nu$ are called *source conditions* and the element z with $|K|^\nu z = x^\dagger$ is called *source element.*

In the case of Example 9.3, the source condition $x^\dagger \in X^\nu$ means that $x^\dagger$ is 2ν-times (weakly) differentiable (with additional boundary conditions) with 2ν-th weak derivative z which is an L^2-function.

Recall the Definition 7.4 of a regularizing filter: A family of functions φ_α (piecewise continuous and bounded) on the interval $[0, \kappa]$ ($\kappa = \|K\|^2$) is a regularizing filter, if for $\lambda > 0$ it holds that

$$\varphi_\alpha(\lambda) \xrightarrow{\alpha\to 0} \tfrac{1}{\lambda}, \quad \lambda|\varphi_\alpha(\lambda)| \leq C_\varphi$$

for some $C_\varphi > 0$. Theorem 7.6 showed that regularizing filters indeed lead to convergent regularizations. Now we want to answer the question when a regularizing filter is optimal or order optimal.

The key to such results are estimates of the approximation error under the assumption that $x^\dagger$ lies in a ρ-ball in X^ν, i.e. under the assumption that $x^\dagger$ is "ν-smooth" and "ρ-large".

We can express such estimates for a given filter φ_α with the functions

$$\begin{aligned}\omega_\nu(\alpha) &:= \sup_{0<\lambda\leq\kappa} \lambda^{\nu/2}|r_\alpha(\lambda)| \quad (\text{recall } r_\alpha(\lambda) = 1 - \lambda\varphi_\alpha(\lambda)) \\ &= \sup_{0<\lambda\leq\kappa} \lambda^{\nu/2}|1 - \lambda\varphi_\alpha(\lambda)|.\end{aligned}$$

These functions can be used to get bounds on the approximation error, the depend on α.

Lemma 10.2. *Let $y^\dagger \in D(K^\dagger)$ and $x^\dagger \in X^\nu$ with $\|x^\dagger\|_\nu \leq \rho$. Further define $x_\alpha = R_\alpha y^\dagger$. Then it holds for all $\alpha > 0$ that*

$$\begin{aligned}\|x_\alpha - x^\dagger\|_X &\leq \omega_\nu(\alpha)\rho, \\ \|Kx_\alpha - Kx^\dagger\|_Y &\leq \omega_{\nu+1}(\alpha)\rho.\end{aligned}$$

Proof. If $\|x^\dagger\|_\nu \leq \rho$ we have that $x^\dagger = |K|^\nu w = (K^*K)^{\nu/2} w$ with $\|w\|_X \leq \rho$. Then from Lemma 7.5 (3) and using $\langle x^\dagger, v_n \rangle = \langle (K^*K)^{\nu/2} w, v_n \rangle = \sigma_n^\nu \langle w, v_n \rangle$ we get

$$\begin{aligned} x^\dagger - x_\alpha &= \sum_n (1 - \sigma_n^2 \varphi_\alpha(\sigma_n^2)) \left\langle x^\dagger, v_n \right\rangle v_n \\ &= \sum_n r_\alpha(\sigma_n^2) \sigma_n^\nu \langle w, v_n \rangle v_n. \end{aligned}$$

Taking the squared norm gives

$$\begin{aligned} \|x^\dagger - x_\alpha\|_X^2 &= \sum_n (r_\alpha(\sigma_n^2)\sigma_n^\nu)^2 |\langle w, v_n \rangle|^2 \\ &\leq \omega_\nu(\alpha)^2 \sum_n |\langle w, v_n \rangle|^2 \leq \omega_\nu(\alpha)^2 \|w\|_X^2 \end{aligned}$$

as claimed. For the second claim recall from Lemma 7.3 that $\||K|x\|_X = \|Kx\|_Y$ and thus

$$\|Kx_\alpha - Kx^\dagger\|_Y = \||K|(x_\alpha - x^\dagger)\|.$$

Moreover,

$$|K|(x_\alpha - x^\dagger) = (K^*K)^{1/2} r_\alpha(K^*K)(K^*K)^{\nu/2} w = \sum_n \sigma_n r_\alpha(\sigma_n^2) \sigma_n^\nu \langle w, v_n \rangle v_n.$$

Since $|\sigma_n r_\alpha(\sigma_n^2)\sigma_n^\nu|^2 = |r_\alpha(\sigma_n^2)\sigma_n^{\nu+1}|^2 \leq \omega_{\nu+1}(\alpha)^2$ the second claim follows similar to the first one. □

Now we can show how to achieve order optimal regularization:

Theorem 10.3 (Order optimal a-priori parameter choice)**.** *Let $K \in K(X, Y)$. If φ_α is a regularizing filter for which fulfills*

$$\begin{aligned} \sup_{0<\lambda\leq\|K\|^2} |\varphi_\alpha(\lambda)| &\leq C_\varphi \alpha^{-1} \\ \omega_\nu(\alpha) &\leq C_\nu \alpha^{\nu/2}. \end{aligned}$$

If the a-priori choice α fulfills

$$c \left(\frac{\delta}{\rho}\right)^{\frac{2}{\nu+1}} \leq \alpha(\delta) \leq C \left(\frac{\delta}{\rho}\right)^{\frac{2}{\nu+1}}$$

for some $0 < c < C$, then (R_α, α) is an order optimal regularization method in the sense of Definition 10.1.

Proof. As always we start with our error decomposition

$$\|x_{\alpha(\delta)}^\delta - x^\dagger\|_X \leq \delta \|R_{\alpha(\delta)}\| + \|x_{\alpha(\delta)} - x^\dagger\|_X.$$

From Lemma 7.5 (2) and our first assumption we know that

$$\|R_{\alpha(\delta)}\| \leq \sqrt{C_\varphi} \sqrt{\sup_{0<\lambda\leq\|K\|^2} |\varphi_{\alpha(\delta)}(\lambda)|} \leq C_\varphi \alpha(\delta)^{-1/2}.$$

From Lemma 10.2 and our second assumption we get

$$\|x_{\alpha(\delta)} - x^\dagger\|_X \leq \omega_\nu(\alpha(\delta))\rho \leq C_\nu \alpha(\delta)^{\nu/2}\rho.$$

We use these estimates in the error decomposition and use the upper and lower bound of the parameter choice to get

$$\begin{aligned}\|x^\delta_{\alpha(\delta)} - x^\dagger\|_X &\leq C_\varphi \alpha(\delta)^{-1/2}\delta + C_\nu \alpha(\delta)^{\nu/2}\rho \\ &\leq C_\varphi c^{-1/2}\delta^{-\frac{1}{\nu+1}}\rho^{\frac{1}{\nu+1}}\delta + C_\nu C^{\nu/2}\delta^{\frac{\nu}{\nu+1}}\rho^{-\frac{\nu}{\nu+1}}\rho \\ &= (C_\varphi c^{-1/2} + C_\nu C^{\nu/2})\delta^{\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}}.\end{aligned}$$

□

Let us investigate the few filters we know if they can lead to order optimal methods To that end, let us collect the inequalities that we need:

Example 10.4 (Order optimality of the truncated SVD). The filter is $\varphi_\alpha(\lambda) = 1/\lambda$ for $\lambda \geq \alpha$ and $= 0$ else. Thus $r_\alpha(\lambda) = 0$ for $\lambda \geq \alpha$ and $= 1$ else. We get

$$\sup_\lambda |\varphi_\alpha(\lambda)| = \tfrac{1}{\alpha} \text{ (attained at } \lambda = \alpha) \qquad \implies C_\varphi = 1,$$

and

$$\omega_\nu(\alpha) = \sup_\lambda \lambda^{\nu/2}|r_\alpha(\lambda)| = \alpha^{\nu/2} \text{ (attained at } \lambda = \alpha) \quad \implies C_\nu = 1.$$

We see that the truncated SVD is indeed an order optimal regularization method (for any $\nu > 0$)! As such, the method can make use of any smoothness in the (unknown) solution. The smoother $x^\dagger$, the better the convergence rate will be since the exponent $\nu/(\nu+1)$ of δ in Theorem 10.3 will be larger for larger ν. △

Example 10.5 (Order optimality and saturation for Tikhonov regularization). The filter is $\varphi_\alpha(\lambda) = (\lambda + \alpha)^{-1}$ and thus

$$r_\alpha(\lambda) = 1 - \lambda\varphi_\alpha(\lambda) = 1 - \tfrac{\lambda}{\lambda+\alpha} = \tfrac{\alpha}{\lambda+\alpha}.$$

We get

$$\sup_\lambda |\varphi_\alpha(\lambda)| = \tfrac{1}{\alpha} \text{ (attained at } \lambda = 0) \qquad \implies C_\varphi = 1.$$

For the other condition we need to investigate the supremum of $\lambda^{\nu/2}|r_\alpha(\lambda)| = \alpha\frac{\lambda^{\nu/2}}{\lambda+\alpha}$. We define $f(\lambda) = \frac{\lambda^{\nu/2}}{\lambda+\alpha}$ and note:

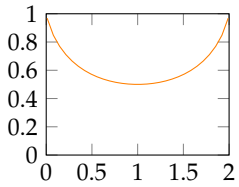
- $\nu > 2$: The function f is unbounded on $]0, \infty[$ (since $\nu/2 > 1$) and hence, the supremum is infinite.
- $\nu = 2$. Here $f(\lambda) < 1$ and even $f(\lambda) \to 1$ for $\lambda \to \infty$ and thus

$$\sup_\lambda \lambda^{\nu/2}|r_\alpha(\lambda)| = \alpha = \alpha^{\nu/2}, \implies C_1 = 1.$$

- $0 < \nu < 2$: Here we have $f(0) = 0$ and $f(\lambda) \to 0$ for $\lambda \to 0$ and hence a finite maximum exists. The condition $f'(\lambda) = 0$ is $0 = (\frac{\nu}{2}\lambda^{\frac{\nu}{2}-1}(\lambda+\alpha) - \lambda^{\frac{\nu}{2}})/(\lambda+\alpha)^2$ which holds exactly if $\lambda^{\frac{\nu}{2}-1}(\frac{\nu}{2}(\lambda+\alpha) - \lambda) = 0$ Since $0 < \lambda < 2$ the only solution is $\lambda = \frac{\nu}{2-\nu}\alpha$ and corresponds to a maximum. We plug this is and get

$$\begin{aligned}\sup_\lambda \lambda^{\nu/2}|r_\alpha(\lambda)| &= \alpha\left(\tfrac{\nu}{2-\nu}\alpha\right)^{\nu/2}\left(\tfrac{\nu}{2-\nu}\alpha+\alpha\right)^{-1}\\ &= \alpha\left(\tfrac{\nu}{2-\nu}\right)^{\nu/2}\alpha^{\frac{\nu}{2}}\left(\tfrac{2}{2-\nu}\right)^{-1}\alpha^{-1}\\ &= \left(\tfrac{\nu}{2-\nu}\right)^{\frac{\nu}{2}}\tfrac{2-\nu}{2}\alpha^{\nu/2}\\ &= \tfrac{\nu^{\nu/2}(2-\nu)^{(2-\nu)/2}}{2}\alpha^{\nu/2} \implies C_\nu = \tfrac{\sqrt{\nu^\nu(2-\nu)^{2-\nu}}}{2}.\end{aligned}$$

The constant C_ν is not as bad as it may look. Here is C_ν is dependence on ν:



In conclusion: Tikhonov regularization is an order optimal regularization method for $0 \le \nu \le 2$ but not for $\nu > 2$. As such, it can take advantage of smoothness up to the space $X^2 = \operatorname{rg}|K|^2 = \operatorname{rg}(K^*K)$.

△

While the a-priori rule from Theorem 10.3 indeed leads to order optimal methods, one drawback is that they need knowledge about both ν and ρ. Without knowledge of ρ one could still choose $\alpha(\delta) \propto \delta^{2/(\nu+1)}$ and obtain on order optimal method (just go through the estimates in the proof and see that you get the right exponent for δ in the end).

The symbol $\propto$ indicates that the left hand side is proportional to the right hand side, i.e. that $\alpha(\delta) = C\delta^{2/(\nu+1)}$ for some C. We could also consider $c\delta^{2/(\nu+1)} \le \alpha(\delta) \le C\delta^{2/(\nu+1)}$

Morozov's discrepancy principle (the only a-posteriori method we know) does not need any knowledge about ν or ρ. Remarkably, this parameter choice also turns out to be order optimal (but in slightly less cases). Before we can formulate this, we make one more definition:

Definition 10.6. Let φ_α be a regularizing filter with $\sup_{0<\lambda\le\|K\|^2}|\varphi_\alpha(\lambda)| \le C_\varphi\alpha^{-1}$. We say that this filter has *qualification* ν_0 if $\omega_\nu(\alpha) \le C_\nu\alpha^{\nu/2}$ is fulfilled for all $0 \le \nu \le \nu_0$. (It it holds for all $\nu \ge 0$ we say that the qualification is $\nu_0 = \infty$).

Theorem 10.7 (Optimality of Morozov's discrepancy principal)**.** *The φ_α be a regularizing filter with qualification $\nu_0 > 0$ and let*

$$\tau > \sup_{\alpha>0,\ 0<\lambda\le\kappa}|r_\alpha(\lambda)| =: C_r.$$

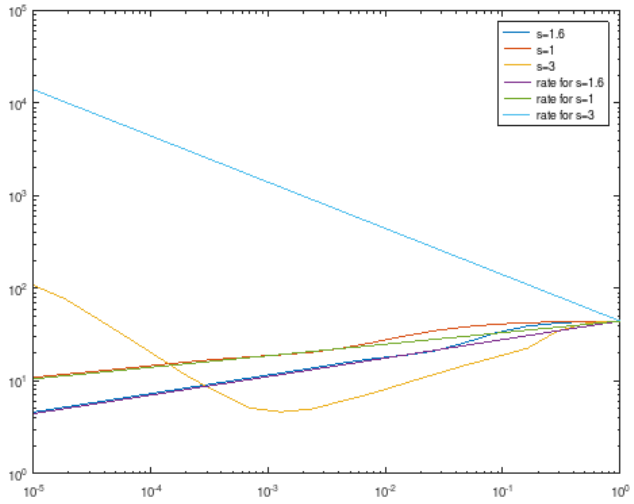
Then the parameter choice defined by Morozov's discrepancy principle with this τ (cf. Section 8) is an order optimal regularization method for $0 \le \nu \le \nu_0 - 1$.

Unfortunately, the proof does not fit into the lecture. It can be found in the lecture notes "Regularization of inverse problems" by Christian Clason, Theorem 5.11, available at https://arxiv.org/abs/2001.00617.

Here is an example that shows that the results of Theorem 10.3 is indeed quite close to what one can observe in practice:

```
% Dimension of the problem
n = 2000;
% deriv2 from the regu-toolbox (from MATLAB's file exchange) implements the
% "inverse of the negative second derivative" from Example 8.3
[A,~,~] = deriv2(n);

% The constant 1 function should be in X^nu for nu<1/4.
% We take the edge case anyway.
x = ones(n,1);
nu = 1/4;


% now do three choices for alpha of the form alpha = delta^s
% one optimal, the other too large and small, resp.
% According to Theorem 9.3 the optimal s is s=2/(nu+1) = 2/(5/4) = 1.6
% expected rate for alpha = delta^s is C*delta^r with r= min(1-s/2,s*nu/2)
% See the proof of Theorem 9.3
s1 = 1.6;
r1 = min(1-s1/2,s1*nu/2)
s2 = 1;
r2 = min(1-s2/2,s2*nu/2)
s3 = 3;
r3 = min(1-s3/2,s3*nu/2)


% Some noise levels going to zero
deltas = logspace(0,-5,20);
% Some noise levels going to zero
% Precompute A'*A
ATA = A'*A;
y = A*x;
for k = 1:length(deltas)
  % for each delta construct data with that noise level
  delta = deltas(k);
  noise = randn(n,1);noise = noise/norm(noise);
  ydelta = y + delta*noise;
  C = 1; % C could be anything. It's choice does not affect the rate,
         % but it does affect the values or the errors.
  % these are our alphas
  alpha1 = C*delta^s1;
  alpha2 = C*delta^s2;
  alpha3 = C*delta^s3;

  % Precompute A'*ydelta
  ATydelta = A'*ydelta;
  % reconstruct by Tikhonov
```

```
  x1 =(ATA+alpha1*eye(n))\(ATydelta);
  x2 =(ATA+alpha2*eye(n))\(ATydelta);
  x3 =(ATA+alpha3*eye(n))\(ATydelta);

  % measure the errors
  error1(k) = norm(x1 - x);
  error2(k) = norm(x2 - x);
  error3(k) = norm(x3 - x);
end

offset = error1(1); %used to adjust the plots
% plor total errors:
loglog(deltas,error1,deltas,error2,deltas,error3)
hold on
% plot the expected rates
loglog(deltas,offset*deltas.^r1,...
   deltas,offset*deltas.^r2,...
   deltas,offset*deltas.^r3)
legend('s=1.6', 's=1', 's=3','rate for s=1.6', 'rate for s=1', 'rate for s=3')

r1 = 0.2000
r2 = 0.1250
r3 = -0.5000
```



One should add that the plot changes quite a bit if we move to even smaller noise levels. There we will see that the plot for $s = 3$ (much too small regularization parameter) will go down again, contrary to what has been predicted by theory. This may due to effects of the finite precision of floating point arithmetic.

11 Iterative regulariztion

The idea of iterative regularization is to use iterative methods that can either solve $Kx = y$ or minimize $\|Kx - y\|_Y^2$ exactly in the case where y is in the range of K (or the domain of the pseudo inverse in the latter case), apply them to the case with some y^δ instead of y, even though y^δ is not in the range of K. We expect that the method will not converge in this case and hence, we stop them at some point. The simplest iterative regularization method is the *Landweber method* and we motivate the method two times:

Example 11.1 (Landweber as a fixed point iteration). We start from the normal equations $K^*Kx = K^*y$ and rewrite them as

$$x = x - \omega(K^*Kx - K^*y) = x - \omega K^*(Kx - y)$$

for some $\omega \in \mathbb{R}$ We turn this into a fixed point iteration

$$x_{n+1} = x_n - \omega K^*(Kx_n - y).$$

For the sake of simplicity we start with $x_0 = 0$. We analyze the convergence of the iteration by Banach's fixed point theorem. The map under consideration is $x \mapsto x - \omega K^*(Kx - y)$, so we analyze

$$\|x - \omega K^*(Kx - y) - (x' - \omega K^*(Kx' - y)\|_X = \|(I - \omega K^*K)(x - x')\|_X.$$

We see that the iteration map is a contraction if $\|\mathrm{id} - \omega K^*K\| < 1$. If this holds, we can see inductively that from $x_0 = 0 \in \mathrm{rg}(K^*)$ it follows that $x_n \in \mathrm{rg}(K^*)$ as well, and hence we get convergence $x_n \to x^\dagger = K^\dagger y$. For $y^\delta \notin \mathrm{rg}(K)$ we can't expect convergence, and need to stop early. Here the stopping index m act as regularization parameter, but to be consistent with our convention "smaller regularization parameter is less regularization" it's more like $\alpha = 1/m$ is the regularization parameter. △

Lemma 11.2. *If* $0 < \omega < 2/\|K\|^2$, *then* $\|\mathrm{id} - \omega K^*K\| \leq 1$.

Proof. For the singular values σ_j of K we know that $0 < \sigma_j \leq \|K\|$. The operator $\mathrm{id} - \omega K^*K$ is self adjoint and has eigenvalues $1 - \omega\sigma_j^2 \in]1 - \omega\|K\|^2, 1[$ and since $0 < \omega < 2/\|K\|^2$ we have that the eigenvalues of $\mathrm{id} - \omega K^*K$ lie in $]-1, 1]$ as needed. □

As a consequence, convergence of the Landweber method does not follow from Banach fixed point theorem (it does so, if K is injective and the smallest singular value exists and is positive, but then the problem is well posed).

Here is another view on the Landweber method.

Example 11.3 (Landweber as gradient descent on the least squares functional). We can also start with the least squares functional

$$f(x) = \tfrac{1}{2}\|Kx - y\|_Y^2.$$

To calculate the derivative of f we do

$$\begin{aligned} f(x+h) &= \tfrac{1}{2}\|K(x+h)-y\|_Y^2 = \tfrac{1}{2}\|Kx-y+Kh\|_Y^2 \\ &\tfrac{1}{2}\|Kx-y\|_y^2 + \langle Kx-y, Kh\rangle + \tfrac{1}{2}\|Kh\|_Y^2 \\ &= f(x) + \langle K^*(Kx-y), h\rangle + \varphi(h) \end{aligned}$$

with $\varphi(h) = \frac{1}{2}\|Kh\|_Y^2$. Since $\varphi(h)/\|h\|_X \leq \|K\|^2\|h\| \to 0$ for $h \to 0$ we get that the gradient of f is

$$\nabla f(x) = K^*(Kx-y).$$

Hence, gradient descent for f with constant stepsize ω is

$$x_{n+1} = x_n - \omega\nabla f(x_n) = x_n - \omega K^*(Kx-y)$$

which is exactly the Landweber iteration we in the previous example. △

The next lemma shows how the n-th iterate can be written explicitly:

Lemma 11.4. *If $x_0 = 0$, then the m-th iterate of the Landweber method with stepsize ω is given by*

$$x_n = \omega \sum_{n=0}^{m-1} (\mathrm{id} - \omega K^*K)^n K^* y.$$

Proof. We prove this by induction: For $m = 1$ we have

$$x_1 = \omega K^* y = \omega(\mathrm{id} - K^*K)^0 K^* y.$$

For the induction step we start with

$$\begin{aligned} x_{m+1} &= x_m - \omega K^*(Kx_m - y) = (\mathrm{id} - \omega K^*K)x_m + \omega K^* y \\ &= (\mathrm{id} - \omega K^*K)\left(\omega \sum_{n=0}^{m-1} (\mathrm{id} - \omega K^*K)^n K^* y\right) + \omega K^* y \\ &= \omega \sum_{n=0}^{m-1} (\mathrm{id} - \omega K^*K)^{n+1} K^* y + \omega(\mathrm{id} - \omega K^*K)^0 K^* y \\ &= \omega \sum_{m=0}^{m} (\mathrm{id} - \omega K^*K)^n K^* y. \end{aligned}$$

□

Hence, m steps of the Landweber method are the same as

$$x_m = \varphi_m(K^*K)K^*y$$

with the filter function

$$\varphi_m(\lambda) = \omega \sum_{n=0}^{m-1} (1-\omega\lambda)^n.$$

Using the geometric sum $\sum_{k=0}^{m} q^k = \frac{1-q^{m+1}}{1-q}$ this gives

$$\varphi_m(\lambda) = \omega \sum_{n=0}^{m-1} (1-\omega\lambda)^n = \omega\frac{1-(1-\omega\lambda)^m}{1-(1-\omega\lambda)} = \frac{1-(1-\omega\lambda)^m}{\lambda}. \quad (1)$$

Theorem 11.5. *Let φ_m be defines by* (1)*. Then it holds that $R_m = \varphi_m(K^*K)K^*$ defined a regularization if $0 < \omega < 2/\|K\|^2$.*

Proof. By Theorem 7.6 we only need to show that φ_m is a regularizing filter, i.e. that $\varphi_m(\lambda) \to 1/\lambda$ for $m \to \infty$ and $0 < \lambda < \|K\|^2$ (recall that $\alpha = 1/m$ act as regularization parameter) and that $\lambda\varphi_m(\lambda)$ is uniformly bounded for all m.

Since we have $0 < \omega < 2/\|K\|^2$ we have for all λ with $0 < \lambda \leq \|K\|^2$ that

$$-1 < 1 - \omega\lambda < 1,$$

and hence $(1-\omega\lambda)^m \to 0$ for $m \to \infty$. Moreover we have for $0 \leq \lambda \leq \|K\|^2$

$$\lambda|\varphi_m(\lambda)| = |1 - (1-\omega\lambda)^m| \leq 2 =: C_\varphi.$$

□

We can even show that the Landweber method is an order optimal method.

Recall from Section 10 that this holds for a-priori parameter choices of the form $\alpha(\delta) \propto \delta^{2/(\nu+1)}$ so here we should stop after about $m^* \approx \delta^{-2/(\nu+1)}$ iterations.

Theorem 11.6 (Landweber is order optimal)**.** *Let $0 < \omega < 2/\|K\|^2$. Then the Landweber method with a-priori rule $m^* = m(\delta) \propto \delta^{-2/(\nu+1)}$ is an order optimal regularization method for any $\nu > 0$.*

Proof. We use Theorem 10.3 and need to show that

$$\sup_{0<\lambda\leq\|K\|^2} |\varphi_m(\lambda)| \leq C_\varphi m$$
$$\omega_\nu(m) = C_\nu m^{-\nu/2}$$

(recall that $\alpha = 1/m$ and don't confuse ω_ν with the stepsize ω).

For the first estimate consider recall that $-1 < 1 - \omega\lambda < 1$ and hence, by Bernoulli's inequality

$$|\varphi_m(\lambda)| = \tfrac{|1-(1-\omega\lambda)^m|}{\lambda} = \tfrac{1-(1-\omega\lambda)^m}{\lambda} \leq \tfrac{1-(1-m\lambda\omega)}{\lambda} = \omega m.$$

From above we already had $C_\varphi = 2$, so now we should set $C_\varphi = \max(2, \omega)$. To estimate $\omega_\nu(\lambda)$ we use consider

$$\lambda^{\nu/2}(1 - \lambda\varphi_m(\lambda)) = \lambda^{\nu/2}(1-\omega\lambda)^m$$

and substitute $t = m\lambda$. Then this expression becomes

$$h(t) = \left(\tfrac{t}{m}\right)^{\nu/2}(1 - \tfrac{\omega t}{m})^m.$$

We use the elementary inequality $(1 - \frac{x}{m})^m \leq e^{-x}$ and get

$$h(t)m^{-\nu/2}t^{\nu/2}e^{-\omega t}.$$

The derivative is

$$\begin{aligned} h'(\lambda) =& m^{-\nu/2}\left(\tfrac{\nu}{2}t^{\nu/2-1}e^{-\omega t} + t^{\nu/2}(-\omega)e^{-\omega t}\right) \\ =& m^{-\nu/2}t^{\nu/2-1}e^{-\omega t}\left(\tfrac{\nu}{2} - t\omega\right) \end{aligned}$$

and thus, h has a global maximum at $t = \nu/(2\omega)$. This shows that

$$h(t) \leq m^{-\nu/2} \left(\tfrac{\nu}{2\omega}\right)^{\nu/2} e^{-\nu/2},$$

and thus

$$\omega_\nu(\lambda) \leq C_\nu m^{-\nu/2},$$

i.e. $C_\nu = \left(\tfrac{\nu}{2\omega}\right)^{\frac{\nu}{2}} e^{-\nu/2}$. □

Note that iterative methods are well suited for Morozov's discrepancy principle. One just monitors $\|Kx_m - y^\delta\|_Y$ during the iteration and stops at the first m^* such that $\|Kx_{m^*} - y^\delta\|_Y \leq \tau\delta$.

One can actually show a little bit more here: Let us denote by x_m^δ the m-th iterate of the Landweber method with y^δ instead of y.

Theorem 11.7. *If $Kx_m^\delta - y^\delta \neq 0$, then it holds for stepsizes $0 < \omega < 2/\|K\|^2$ that*

$$\|Kx_{m+1}^\delta - y^\delta\|_Y \leq \|Kx_m^\delta - y^\delta\|_Y.$$

Moreover, if $\|Kx_m^\delta - y^\delta\|_Y > 2\delta$ and $0 < \omega < 1/\|K\|^2$ we even have

$$\|x_{m+1}^\delta - x^\dagger\|_X < \|x_m^\delta - x^\dagger\|_X.$$

Proof. We compute from the iteration

$$\begin{aligned} Kx_{m+1}^\delta - y^\delta &= K((\mathrm{id} - \omega K^*K)x_m^\delta + \omega K^* y^\delta - y^\delta \\ &= (\mathrm{id} - \omega K^*K)(Kx_m^\delta - y^\delta). \end{aligned}$$

For stepsize $\omega \in]0, 2/\|K\|^2[$ we get that $\|\mathrm{id} - \omega K^*K\| \leq 1$ and thus shows the first claim.

For the second claim we write $z_m^\delta := y^\delta - Kx_m^\delta$ and $y = Kx^\dagger$ and get

$$\begin{aligned} \|x_{m+1}^\delta - x^\dagger\|_X^2 &= \|x_m^\delta - x^\dagger - \omega K^*(Kx_m^\delta - y^\delta)\|_X^2 \\ &= \|x_m^\delta - x^\dagger\|_X^2 + 2\omega \left\langle x_m^\delta - x^\dagger, K^* z_m^\delta \right\rangle + \omega^2 \|K^* z_m^\delta\|_X^2 \\ &= \|x_m^\delta - x^\dagger\|_X^2 + 2\omega \left\langle Kx_m^\delta - Kx^\dagger, z_m^\delta \right\rangle + \omega^2 \|K^* z_m^\delta\|_X^2 \\ &= \|x_m^\delta - x^\dagger\|_X^2 + \omega \left\langle z_m^\delta + 2Kx_m^\delta - 2y, z_m^\delta \right\rangle + \omega(\omega \|K^* z_m^\delta\|_X^2 - \|z_m^\delta\|_Y^2. \end{aligned}$$

We aim to show that the last two terms are negative. For the first term we compute

$$\begin{aligned} \left\langle z_m^\delta + 2Kx_m^\delta - 2y, z_m^\delta \right\rangle &= \left\langle y^\delta - Kx_m^\delta + 2Kx_m^\delta - 2y, z_m^\delta \right\rangle \\ &= \left\langle y^\delta + Kx_m^\delta - 2y, z_m^\delta \right\rangle \\ &= 2 \left\langle y^\delta - y, z_m^\delta \right\rangle - \|z_m^\delta\|_Y^2 \\ &\leq 2\delta \|z_m^\delta\|_Y^2 - \|z_m^\delta\|_Y^2 \\ &= (2\delta - \|Kx_m^\delta - y^\delta\|_Y)\|z_m^\delta\|_Y < 0. \end{aligned}$$

For the second term we use $\omega < 1/\|K\|^2$ to get

$$\omega \|K^* z_m^\delta\|_X^2 \leq \omega \|K\|^2 \|z_m^\delta\|_Y^2 < \|z_m^\delta\|_Y^2$$

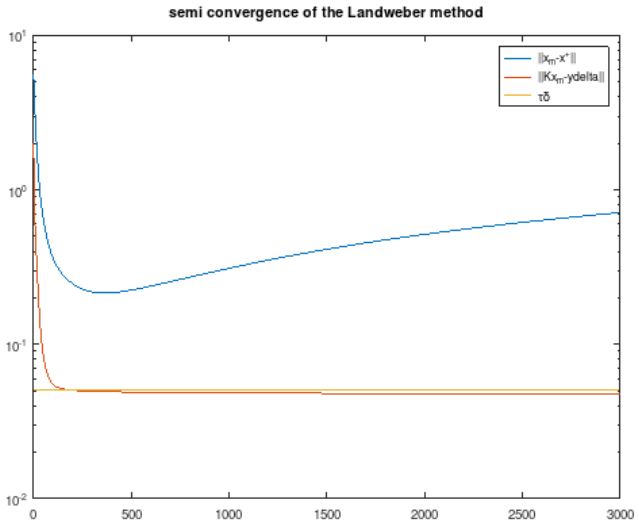
which shows that the last term is negative as well. □

The theorem shows two important things: First, the Landweber method always decreases the residual but, more importantly, even the distance to the *true* solution decreases if the residual is larger than 2δ, so it is always beneficial to use $\tau < 2$ in Morozov's discrepancy principle. Note that the Landweber method can always be applied when one is able to apply the operator K and its adjoint K^*. The is no need for singular value decomposition (as for the TSVD) and also we do not need to solve linear systems (like for Tikhonov regularization). One downside of the Landweber method is that it usually needs a lot of iterations until a good reconstruction is achieved (or Morozov's discrepancy principle kicks is). In practice one can use other iterative methods that solve the normal equations, e.g. the method of conjugate gradients (CG) which converges much faster. One can show (with very different tool than here) that, combined with the discrepancy principle, is indeed a regularization method.

Here is an example with the Landweber iteration:

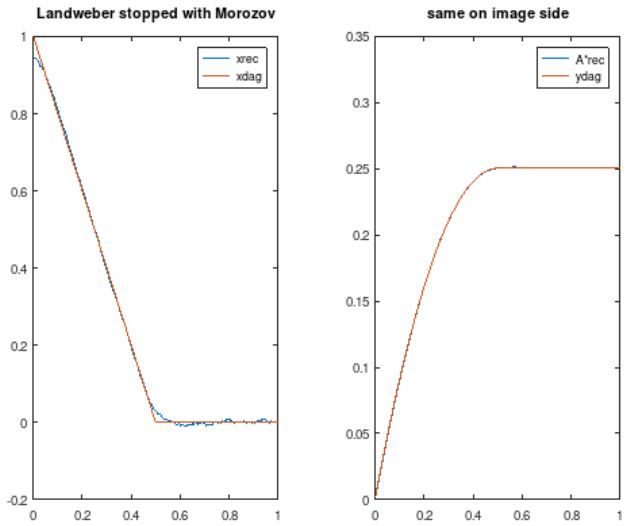
```
% problem size and matrix
n = 500;
A = tril(ones(n))/n;

% discretized interval
t = linspace(0,1,n)';

% some true solutions. Uncomment the one you want to use
%xdag = 1-t.^2;
xdag = max(1-2*t,0);
%xdag = (t<0.5);
% true data
ydag = A*xdag;

% noisy data
eta = randn(n,1); eta = eta/norm(eta); % normalized noise
delta = 0.05;                          % noise level
ydelta = ydag + delta*eta;

% stepsize for the Landweber iteration
normA = norm(A);
omega = 1/normA;
% constant for Morozov's discrepancy principle
tau = 1.01;

% number of iterations
```

```
m = 3000;
% initialization
x = zeros(n,1);

residual = zeros(m,1);
error = zeros(m,1);
stopped = false;
for k=1:m
  x = x - omega*A'*(A*x-ydelta);
  error(k) = norm(x-xdag);
  residual(k) = norm(A*x-ydelta);
  % if the residual falls below the noise level for the first time
  % record the index and the reconstruction at that time
  if stopped==false && residual(k)<tau*delta
    mstar = k;
    xrec = x;
    stopped = true;
  end
end

semilogy(1:m,error,1:m,residual,1:m,tau*delta*ones(m,1))
title('semi convergence of the Landweber method')
legend('||x_m-x^+||', '||Kx_m-ydelta||', '\tau\delta')
```




```
fprintf('stopping index: m* = %d\n',mstar)

subplot(1,2,1)
plot(t,xrec,t,xdag)
title('Landweber stopped with Morozov')
legend('xrec','xdag')
subplot(1,2,2)
```

```
plot(t,A*xrec,t,ydag)
title('same on image side')
legend('A*rec','ydag')

stopping index: m* = 221
```



12 A Bayesian perspective on regulariztion

In this section we would like to draw connections between regularization (especially Tikhonov regularization) and probabilistic approaches to inverse problems. This will be much simpler if we consider everything to be finite dimensional, i.e. $K \in \mathbb{R}^{m\times n}$ is a matrix, $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. We start by modeling noise stochastically. Our data is $y^\delta = Kx^\dagger + \eta$ where $\eta \in \mathbb{R}^m$ is a random vector, i.e. a realization of a random variable H. A vector valued random variable is a map $H : \Omega \to \mathbb{R}^m$ on some probability space Ω (which will actually not play a role). On Ω there is probability measure P and the random variable H generates a probability distribution on $\mathbb{R}^m$ by

$$\mu(B) = P(H^{-1}(B))$$

for all Borel sets $B \subset \mathbb{R}^m$. However, P is never needed in practice and we only need to know the probability distribution π_{noise} in $\mathbb{R}^m$ since then

$$\mu(B) = P(H \in B) = \int_B \pi_{\text{noise}}(\eta)\mathrm{d}\eta.$$

The mean value (or expectation) of H is

$$\mathbb{E}(H) = \int_{\mathbb{R}^m} \eta \, \mathrm{d}\pi_{\text{noise}}(\eta) = \int \eta\, \pi_{\text{noise}}(\eta)\mathrm{d}\eta$$

and the covariance is

$$\operatorname{cov}(H) = \mathbb{E}((H - \mathbb{E}(H))(H - \mathbb{E}(H))^T) = \int (\eta - E(H))(\eta - E(H))^T \pi_{\text{noise}}(\eta)\mathrm{d}\eta.$$

Example 12.1 (Gaussian noise). The most simple (and also most widely used) example of additive noise is Gaussian noise. Usually we assume that the noise has zero mean and for simplicity we assume that the components of the random vector are independent and identically distributed (i.i.d), each with variance σ^2. Then the probability distribution of one entry of H is

$$\tfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}.$$

Hence, the full probability distribution of H is

$$\begin{aligned}\pi_{\text{noise}}(\eta) &= \prod_{i=1}^m \tfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\eta_i^2}{2\sigma^2}} \\ &= \tfrac{1}{\sigma^m\sqrt{2\pi}^m} e^{\frac{-\|\eta\|^2}{2\sigma^2}}.\end{aligned}$$

The covariance of H is $\operatorname{cov}(H) = \sigma^2 I_m$ (where I_m is the $m \times m$ identity matrix). △

The goal of probabilistic approaches is to gain as much information as possible about the *posterior distribution* $\pi_{\text{posterior}}(x \mid y)$, i.e. the distribution of the solution x, given that the data y has been observed. The posterior distribution is, by Bayes theorem,

$$\pi_{\text{posterior}}(x \mid y) = \frac{\pi(y \mid x)\pi_{\text{prior}}(x)}{\pi(y)}.$$

where $\pi_{\text{prior}}(x)$ is the so-called *prior distribution*, $\pi(y \mid x)$ is the probability of measuring the data y given that x is the solution and $\pi(y)$ is the probability of the data y.

Example 12.2 (Maximum a-posteriori estimation). One crucial information that one can get from the posterior distribution is the mode of the distribution which is nothing else that the maximum $x^* \in \operatorname{argmax}_x \pi_{\text{posterior}}(x \mid y)$. This x^* is the most likely x given the measured data y and the assumptions we made. This gives us a *point estimate* for our solution x and this one is called the *maximum a-posteriori estimator* or MAP estimator. The computation of the MAP estimator amounts the solution of a maximization problem in n dimensions. △

The MAP estimator may be the most likely x^*, but is it is not necessarily a typical one. One example of this phenomenon already occurs for very simple distributions: A standard Gaussian distribution has its mode at zero while a sample that you draw has expected norm $\sqrt{n}$ in n dimensions. One can even show that it holds that the probability that the norm $\|x\|_2$ of a vector with independent standard Gaussian entries deviates from $\sqrt{n}$ is very small, more precisely

$$P\left(|\|x\|_2 - \sqrt{n}| \geq t\right) \leq 2\exp(-ct^2)$$

for some c.

Example 12.3 (Conditional mean estimator). Another point estimate of a distribution is its mean/expected value. Hence, we could also consider the so-called *conditional mean* of the posterior which is

$$x^* = \mathbb{E}(x \mid y) = \int_{\mathbb{R}^n} x\, \pi(x \mid y) \mathrm{d}x$$

(provided that the integral exists). The computation of the conditional mean amount to the computation of n integrals (recall that $x \in \mathbb{R}^n$) over the full $\mathbb{R}^n$. Since n is the dimension of the solution, this is by no means an easy task and standard approximation techniques for integrals (such as the trapezoidal rule) can not be applied. △

In the following we will only consider the MAP estimate further. For the computation of the MAP estimator we maximize $\pi_{\text{posterior}}(x|y)$ over x. Using Bayes theorem we note that we do not need to know anything about $\pi(y)$, but only $\pi(y \mid x)$ and $\pi_{\text{prior}}(x)$ are needed.

Example 12.4 (Additive Gaussian noise again). If we assume that our solution X and the noise H are independent, the probability density of H does not change, when we condition it on the realization $X = x$. Since $y = Kx + \eta$ we also see that the Y conditioned on $X = x$ is distributed like H but translated by Kx, i.e.

$$\pi(y \mid x) = \pi_{\text{noise}}(y - Kx). \tag{*}$$

In the case a Gaussian noise as above we get

$$\pi(y \mid x) = \frac{1}{\sigma^m \sqrt{2\pi}^m} e^{\frac{-\|y-Kx\|^2}{2\sigma^2}}.$$

△

Collection what we have so far we see that we still need to specify the prior distribution for x. This is up to us; we can design a prior distribution in any way we like. More precisely, we should design the prior such that it reflects all the prior information that we have about the solution. Once we have the prior, we can start thinking about how to compute the MAP estimator. If we assume a Gaussian prior, we actually end up with Tikhonov regularization:

Theorem 12.5 (MAP for Gaussian prior and Gaussian noise gives Tikhonov regularization)**.** *Assume that $K \in \mathbb{R}^{m\times n}$, $y^\delta \in \mathbb{R}^m$ be the data, x_0 be an initial guess and $\sigma, \tau > 0$. Further let the distribution of the noise and the prior be*

$$\pi_{\text{noise}}(\eta) = \frac{1}{\sigma^m \sqrt{2\pi}^m} e^{\frac{-\|\eta\|^2}{2\sigma^2}}$$
$$\pi_{\text{prior}}(x) = \frac{1}{\tau^n \sqrt{2\pi}^n} e^{\frac{-\|x-x_0\|^2}{2\tau^2}}.$$

Then the MAP estimator for x from y^δ is

$$x^* = (K^*K + \tfrac{\sigma^2}{\tau^2}\,\mathrm{id})^{-1}(K^*y^\delta + x^0)$$

Proof. The posterior distribution is

$$\pi(x \mid y^\delta) \propto \pi(y^\delta \mid x)\pi_{\text{prior}}(x)$$

Using (*) and the definition of π_{noise} and π_{prior} we get

$$\begin{aligned}\pi(x \mid y^\delta) &\propto \pi_{\text{noise}}(y^\delta - Kx)\pi_{\text{prior}}(x)\\ &= \frac{1}{\sigma^m \sqrt{2\pi}^m} e^{\frac{-\|y^\delta-Kx\|^2}{2\sigma^2}} \frac{1}{\tau^n \sqrt{2\pi}^n} e^{\frac{-\|x-x_0\|^2}{2\tau^2}}.\end{aligned}$$

To maximize this we equivalently maximize the logarithm of $\pi(x \mid y^\delta)$ (since everything is positive and the logarithm is monotone). This gives us the maximization problem.

$$\begin{aligned}&\operatorname*{argmax}_x \log\left(\frac{1}{\sigma^m \sqrt{2\pi}^m} e^{\frac{-\|y^\delta-Kx\|^2}{2\sigma^2}} \frac{1}{\tau^n \sqrt{2\pi}^n} e^{\frac{-\|x-x_0\|^2}{2\tau^2}}\right)\\ = &\operatorname*{argmax}_x \left[-m\log(\sigma\sqrt{2\pi}) - \frac{\|Kx-y^\delta\|^2}{2\sigma^2} - n\log(\tau\sqrt{2\pi}) - \frac{\|x-x_0\|^2}{2\tau^2}\right].\end{aligned}$$

Since we only maximize with respect to x we can neglect the additive terms that do not depend on x and also scale by positive numbers o get

$$\begin{aligned} x^* &\in \operatorname*{argmax}_x -\tfrac{\|Kx-y^\delta\|^2}{2\sigma^2} - \tfrac{\|x-x_0\|^2}{2\tau^2} \\ &= \operatorname*{argmin}_x \tfrac{\|Kx-y^\delta\|^2}{2\sigma^2} + \tfrac{\|x-x_0\|^2}{2\tau^2} \\ &= \operatorname*{argmin}_x \tfrac{1}{2}\|Kx-y^\delta\|^2 + \tfrac{\sigma^2}{2\tau^2}\|x-x_0\|^2. \end{aligned}$$

We recognize the Tikhonov functional with regularization parameter $\frac{\sigma}{\tau}$. The minimizer x^* is given by the solution of

$$K^*(Kx^* - y^\delta) + \tfrac{\sigma^2}{\tau^2}(x^* - x^0) = 0$$

which proves the claim. □

The choice of the prior distribution is basically an art. Many suitable priors exists for various types of data. If we consider additive Gaussian noise as above and a prior of the form $\pi_{\text{prior}}(x) = e^{-\alpha\Phi(x)}$ one get, similarly to the above theorem, that the MAP estimate is gives as a solution of the minimization problem

$$\min_x \tfrac{1}{2}\|Kx-y^\delta\|^2 + \tfrac{\sigma^2\alpha}{2}\Phi(x).$$

The noise distribution, however, is dictated by the noise model and if one does not have additive Gaussian noise, one can still formulate a regularization method:

Example 12.6 (Poisson noise). The Poisson distribution models rare events. One example comes from photography-like applications with very low light as it occurs, for example, in electron microscopy. There each pixel collects incoming photos over a short time span. The number of incoming photons in some pixel p, when measured and averaged over a long time span, gives the true intensity $y(p)$. Over a short time span, one collects a finite number of photons and the stochastic model for this number is that it is distributed according to the Poisson distribution with parameter $\lambda = y(p)$ (the parameter λ is also the expected value of the distribution), this means that the probability to collect $y^\delta(p) = k$ photons in pixel p is

$$P(y^\delta(p) = k) = \tfrac{y(p)^k e^{-y(p)}}{k!}.$$

Hence, the conditional probability that $y^\delta(p)$ is measured if $y(p)$ is the true value is

$$\pi(y^\delta(p) \mid y(p)) = \tfrac{y(p)^{y^\delta(p)} e^{-y(p)}}{(y^\delta(p)!}.$$

We still assume that the noise is the pixels is independent, i.e we have

$$\pi(y^\delta \mid y) = \prod_p \frac{y(p)^{y^\delta(p)} e^{-y(p)}}{(y^\delta(p)!}.$$

If we assume that the image prior π_{prior} is of the form $\pi_{\text{prior}}(x) \propto e^{-\alpha\Phi(x)}$ for some function Φ, the MAP estimate for x with $y = Kx$ from y^δ (where y^δ is a version of y that is corrupted by Poisson noise) is

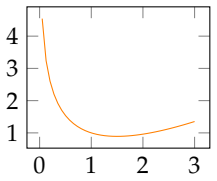
$$\operatorname*{argmax}_x \pi(y^\delta \mid Kx)\pi_{\text{prior}}(x) = \operatorname*{argmax}_x \prod_p \frac{(Kx(p))^{y^\delta(p)} e^{-Kx(p)}}{(y^\delta(p)!} \cdot e^{-\alpha\Phi(x)}.$$

Equivalently, we minimize the negative logarithm of the objective which is

$$\begin{aligned} &\operatorname*{argmin}_x \Big[-\log(\pi(y^\delta \mid Kx)) - \log(\pi_{\text{prior}}(x)) \Big] \\ = &\operatorname*{argmin}_x \sum_p \Big[-y^\delta(p)\log(Kx(p)) + Kx(p) \Big] + \alpha\Phi(x). \end{aligned}$$

△

The negative log of the noise prior may look strange at first. However, note that the function $f_a(t) = -a\log(t) + t$ has a unique minimum at $t = a$ (here for $a = 1.5$).



13 Discretization by projection

In our last section we finally treat some discretization methods.

The backbone of discretization are projection operators. A bounded linear operator $P : X \to X$ on a normed space is a *projection onto a subspace* U if $Px \in U$ for all $x \in X$ and $Px = x$ for $x \in U$. Moreover it holds that $P^2 = P$ and $\|P\| \geq 1$. If X is a Hilbert space and P is self-adjoint, then P is an orthonormal projection and it holds for all $u \in U$ that

$$\|Px - x\| \leq \|u - x\|,$$

i.e. Px is the best approximation from U to x.

Definition 13.1. Let X, Y be Banach spaces, $K : X \to Y$ be bounded and linear and $X_n \subset X$, $Y_m \subset Y$ be n- and m-dimensional subspaces, respectively. Further, let $Q_m : Y \to Y_m$ be a projection onto Y_m. The *projection method* for solving $Kx = y$ is to solve the problem

$$Q_m K x_n = Q_m y, \text{ for } x_n \in X_n.$$

If we choose bases $\{\hat{x}_1, \dots, \hat{x}_n\}$ and $\{\hat{y}_1, \dots, \hat{y}_m\}$ of X_n and Y_m, respectively, we can write

$$Q_n y = \sum_{i=1}^{n} \beta_i \hat{y}_i, \quad \text{and} \quad Q_n K \hat{x}_j = \sum_{i=i}^{n} A_{ij} \hat{y}_i.$$

The solution x_n can written as $\sum_{j=1}^{n} \alpha_j \hat{x}_j$ and thus, the coefficients can be determined by the linear system

$$\sum_{j=1}^{n} A_{ij} \alpha_j = \beta_i.$$

Plugging the expansion for x_n into $Q_m K x_n = Q_m y$ gives

$$\sum_{j=1}^{n} \alpha_j Q_m K \hat{x}_j = \sum_{i=1}^{m} \beta_i \hat{y}_i$$

and using $Q_n K \hat{x}_j = \sum_{i=i}^{n} A_{ij} \hat{y}_i$ gives the result.

Example 13.2 (Galerkin method). Let X and Y be Hilbert spaces, X_n, Y_m as above and Q_m an orthogonal projection onto Y_m. The equation $Q_m K x_n = Q_m y$ is then equivalently expressed as the so-called Galerkin equations

$$\langle K x_n, z_m \rangle = \langle y, z_m \rangle \quad \text{for all } z_m \in Y_m. \tag{*}$$

Choosing bases as above gives us the system

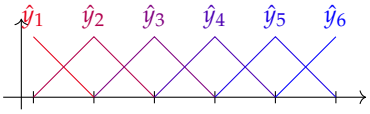
$$\sum_{j=1}^{n} \alpha_j \underbrace{\langle K \hat{x}_j, \hat{y}_i \rangle}_{=:A_{ij}} = \underbrace{\langle y, \hat{y}_i \rangle}_{=:\beta_i}, \quad i = 1, \dots, m. \tag{**}$$

△

Example 13.3 (Collocation method). Here we have any Banach space X but fix $Y = C([a, b])$. We choose so-called *collocation points* $a = t_1 < \cdots < t_m = b$ and consider the subspace Y_m as the

space of functions that are continuous and piecewise linear on the intervals $[t_i, t_{i+1}]$ (also known as the space of linear splines). As operator Q_m we take the "linear interpolation operator", i.e.

$$Q_m y = \sum_{i=1}^{m} y(t_i)\hat{y}_i$$



where the $\hat{y}_i$ are the linear splines which are 1 at t_i and zero at t_j with $j \neq i$. The projected equation $Q_m K x_n = Q_m y$ is then equivalent to

$$(Kx_n)(t_i) = y(t_i), \quad i = 1, \dots, n.$$

We choose an n-dimensional subspace X_n of X and a basis $\{\hat{x}_j \mid j = 1, \dots, n\}$ of X_n. Then we can express $x_n \in X_n$ by $x_n = \sum_{j=1}^{n} \alpha_j \hat{x}_j$. The collocations equations for $x_n \in X_n$ then become

$$K \sum_{j=1}^{n} \alpha_j \hat{x}_j(t_i) = y(t_i).$$

The left hand side is $\sum_{j=1}^{n} K\hat{x}_j(t_i)\alpha_j$ and we see that the collocation equations are equivalent to $A\alpha = \beta$ with

$$\beta_i = y(t_i), \quad A_{ij} = K\hat{x}_j(t_i).$$

△

Example 13.4 (Galerkin and collocation for integral equations). Now consider the special case $K : L^2([a,b]) \to L^2([c,d])$,

$$Kx(t) = \int_a^b k(t,s)x(s)\mathrm{d}s = y(t), \quad t \in [c,d].$$

The Galerkin method uses the values (cf. (**))

$$A_{ij} = \int_c^d \int_a^b k(t,s)\hat{x}_j(s)\hat{y}_i(t)\mathrm{d}s\,\mathrm{d}t, \quad \beta_i = \int_c^d y(t)\hat{y}_i(t)\mathrm{d}t$$

while the collocation method uses

$$A_{ij} = \int_a^b k(t_i,s)\hat{x}_j(s)\mathrm{d}s, \quad \beta_i = y(t_i).$$

Note that the entries for the Galerkin method are harder to compute (more integrals...). △

In the following we will assume $m = n$ and that the following conditions are fulfilled: K is injective, the union $\bigcup_n X_n$ is dense in X and $Q_n K|_{X_n} : X_n \to Y_n$ is invertible.

Then a solution of $Q_n K x_n = Q_n y$ exists and is given by

$$x_n = R_n y_n, \quad \text{with} \quad R_n := (Q_n K|_{X_n})^{-1} Q_n : Y \to X_n.$$

We say that the projection method is *convergent* if for every $x \in X$ it holds that

$$R_n K x \overset{n\to\infty}{\longrightarrow} x.$$

Note that what we are doing here is to consider projection methods as regularizations! We can consider $\alpha = 1/n$ as regularization parameter.

Not every projection method converges, but there is a simple condition that ensures convergence:

Theorem 13.5. *Under our standing assumptions it holds that $x_n = R_n y$ converges to x for every $y = Kx$ exactly if there exists $c > 0$ such that*

$$\|R_n K\| \leq c. \tag{+}$$

Moreover, if this is fulfilled, then (with the same c)

$$\|x_n - x\|_X \leq (1 + c) \min_{z_n \in X_n} \|z_n - x\|_X.$$

Proof. First assume that the method converges, i.e. that $R_n K x \overset{n\to\infty}{\longrightarrow} x$ for every x. Then the assertion follows from the uniform boundedness principle.

For us, the other direction is more interesting: Let $\|R_n K\|$ be bounded. For $z_n \in X_n$ we have that

$$R_n K z_n = (Q_n K|_{X_n})^{-1} Q_n K z_n = (Q_n K|_{X_n})^{-1} Q_n K|_{X_n} z_n = z_n$$

and thus, $R_n K$ is a projection. We conclude that

$$x_n - x = (R_n K - \mathrm{id})x = (R_n K - \mathrm{id})(x - z_n).$$

We obtain $\|x_n - x\|_X \leq (c+1)\|x - z_n\|_X$ and taking the minimum over all z_n shows the inequality. The convergence $x_n \overset{n\to\infty}{\longrightarrow} x$ follows since $\bigcup_n X_n$ is dense in X. □

Here is an error estimate for the Galerkin method. To express it, we define the *synthesis operator* $S_n^X : \mathbb{R}^n \to X$ in X by $S_n^X \alpha = \sum_j \alpha_j \hat{x}_j$ and similarly for S_n^Y. We define the quantities

If we choose orthonormal bases, then we get $a_n = \|S_n^X\| = 1$ and also $b_n = 1$.

$$a_n = \|S_n^X\| = \max\left\{ \|S_n^X \alpha\|_X \;\middle|\; \|\alpha\|_2 = 1 \right\}$$
$$b_n = \max\left\{ \|\beta\|_2 \;\middle|\; \|S_n^Y \beta\|_Y = 1 \right\}$$

Theorem 13.6. *Assume that the Galerkin equations (*) from Example 13.2 are uniquely solvable.*

(a) Let $y^\delta \in Y$ with $\|y - y^\delta\|_Y \leq \delta$ and x_n^δ be the solution of

$$\left\langle K x_n^\delta, z_n \right\rangle = \left\langle y^\delta, z_n \right\rangle \quad \text{for all } z_n \in Y_n.$$

Then it holds that

$$\|x_n^\delta - x\|_X \leq \|R_n\|\delta + \|R_n K x - x\|_X.$$

(b) Let A and β be given by (**) from Example 13.2 and let $\|\beta - \beta^\delta\| \leq \delta$ hold and let λ_n be the smallest singular value of A. Let α^δ be the solution of $A\alpha^\delta = \beta^\delta$ and define $x_n^\delta = \sum_{j=1}^n \alpha_j \hat{x}_j$. Then it holds

$$\|x_n^\delta - x\|_X \leq \tfrac{a_n}{\lambda_n}\delta + \|R_n K x - x\|_X$$

Proof. For part (a) we simply use the standard error decomposition

$$\|x_n^\delta - x\|_X \leq \|x_n^\delta - R_n y\| + \|R_n y - x\| \leq \|R_n\|\|y^\delta - y\|_Y + \|R_n K x - x\|_X$$

from which the estimate follows.

For part (b) we just need to estimate the data error in the above decomposition. We write $R_n x = \sum_{j=1}^n \alpha_j \hat{x}_j$. Since $x_n^\delta - R_n y = \sum_{j=1}^n (\alpha_j^\delta - \alpha_j)\hat{x}_j = S_n^X(\alpha_n^\delta - \alpha)$ we get

$$\|x_n^\delta - R_n y\|_X \leq a_n \|\alpha_n^\delta - \alpha\|_2 = a_n \|A^{-1}(\beta^\delta - \beta)\|_2 \leq \tfrac{a_n}{\lambda_n}\delta$$

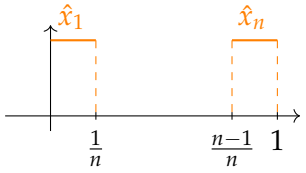
as desired. □

Example 13.7 (Collocation method for the inverse integration problem). We consider the simple problem $Kx(t) = \int_0^t x(s)\mathrm{d}s$ with $K : C([0,1]) \to C([0,1])$. We have to choose the collocation points t_i and the basis $\hat{x}_j$ of X_n. Once we have done this, the linear equation $A\alpha = \beta$ is given by $\beta_i = y(t_i)$ and

$$A_{ij} = K\hat{x}_j(t_i) = \int_0^{t_i} \hat{x}_j(s)\mathrm{d}s.$$

First we choose the basis $\hat{x}_j$. Let us choose the characteristic functions on the intervals $I_j = [\frac{j-1}{n}, \frac{j}{n}[$:

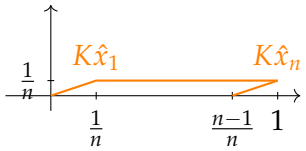
$$\hat{x}_j(t) = \chi_{I_j}(t).$$



,

The functions $K\hat{x}_j$ are

$$K\hat{x}_j(t) = \begin{cases} 0 & : \ t \leq \frac{j-1}{n} \\ t - \frac{j-1}{n} & : \ \frac{j-1}{n} \leq t \leq \frac{j}{n} \\ \frac{1}{n} & : \ t \geq \frac{j}{n} \end{cases}$$



Depending on the collocation points we get different linear systems:

1. We choose $t_i = \frac{i-1}{n}$, $i = 1, \dots, n+1$, i.e. we have $m = n+1$. This gives us

$$A_{ij} = K\hat{x}_j(x_i) = \begin{cases} 0 & : \quad i \leq j \\ \frac{1}{n} & : \quad \text{else.} \end{cases}, \quad A = \frac{1}{n} \begin{pmatrix} 0 & & \\ 1 & \ddots & \\ \vdots & \ddots & 0 \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n+1 \times n}.$$

2. As a variant of the first choice we could only take the left ends, i.e. $t_i = \frac{i-1}{n}$, $i = 1 \dots, n$, or the right ends $t_i = \frac{i}{n}$, $i = 1, \dots, n$ and get

$$A = \frac{1}{n} \begin{pmatrix} 0 & & & \\ 1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ 1 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad A = \frac{1}{n} \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n},$$

   respectively.

3. We choose the middle points of the intervals $t_i = \frac{i-\frac{1}{2}}{n}$. This way we get

$$A_{ij} = \frac{1}{n} \begin{pmatrix} \frac{1}{2} & & & \\ 1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ 1 & \cdots & 1 & \frac{1}{2} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

△