# Notes from lectures SF2529

Gustaf Bjurstam
bjurstam@kth.se

Lecturer: Ozan Öktem

# 1 Lecture 1 (2025-08-25)

Old exams will be very different from what we will be discussing.

We will only be dealing with linear problems.

The lectures are mathematical in nature, and there are homeworks and computer labs.

We will pick problems where the straightforward simulation is simple, but tricky to reverse engineer, this is the *Inverse Problem*.

This course is in a functional analytic setting. Ozan will try to reduce things to finite dimensions, but this is not always possible.

**Agenda**

- Examples of inverse problems
- Mathematical formalisation of what an inverse problem is
- Naïve reconstruction (methods to solve the inverse problem) wasn't time for
- *Inverse crimes* wasn't time for

## 1.1 Examples

The direct problem is generating data that replicates a system, aka a simulation. In the inverse problem, we seek the cause of existing data. We are reconstructing the input that caused the output.

Calibrating a model to data, is a common example of an inverse problem.

Signal processing is full of inverse problems, example, recover the analogue signal from the digital signal observed. Similarly, denoising a signal is an inverse problem.

Imaging is also quite full of inverse problems.

Training a neural network is also an inverse problem.

Decrypting a message is an inverse problem as well.

Since the 1950s there has been a theory built up which works for general inverse problems, originally people only used methods engineered for specific problems.

We have two problems as model cases. Deconvolution, and tomographic imaging.

### 1.1.1 Deconvolution

We are looking to invert the effect of convolution, often used to model the degradation of signals. $x * k(t) = \int_{\mathbb{R}^d} x(\tau)k(t-\tau)\,d\tau$. We are thinking of the input as a function. The signal is a function $x : \mathbb{R}^d \to \mathbb{R}$, the data $y : \mathbb{R}^d \to \mathbb{R}$. We are observing $y$ and want to recover $x$, and the kernel $k$ is known. The kernel is $k$. The kernel is the system response. Any translation invariant, linear problem, is a convolution.

For convolution, outside the sampling interval, we either use a *circular model*, meaning that $x$ is extended periodically. The other option is zero-padding.

The circular model gives $\psi_j = \sum_{i=1}^{n-1} x_j k_{(j-i)_n}$, meaning modulo $n$. Nice for using Fourier transforms/FFT.

In zero-padding, it means that $k_j = 0$ if $j \neq 0, 1, \ldots, n-1$. This is a banded matrix.

Maybe this problem is a bit too simple.

### 1.1.2 X-ray computer tomography

All the labs will be on tomography.

We want to recover an image of the interior of an object, in ultra sounds this would correspond to recording the signal that goes through the body (not what reflects).

Probe an object from different directions with a particle or wave, then you hope to recover the interior.

Radiative transport equation models the interactions between particles used for probing and the object.

too messy

In clinical settings we don't need the full model, so we will use that instead. Beer-Lambert's law is the simplest variant of radioactive transport equation and disregards many phenomena which could really happen. Assumes material is homogeneous. We call the intensity $I$, the law says $\Delta I/I_{in} = -x\Delta t$, where $\Delta t$ is the thickness and $x$ is a material constant. What happens if $t \to 0$? In continuous setting

$$I'(t) = -xI(t)$$

and thus $I = I_0 e^{-xt}$. Now consider two slabs with different material, $x$ is called the linear attenuation coefficient. Claim $I(t+\Delta t_1+\Delta t_2) = I_0 e^{-x_1\Delta t_1 - x_2\Delta t_2}$. In a general situation with continuous differences we have $x : \mathbb{R}^2 \to \mathbb{R}$, we obtain $I_{out} = I_{in}\exp(-\int x\,dt)$ (line integral). We want to find the function $x$ in the inverse problem.

More formally, we want to find $x : \mathbb{R}^d \to \mathbb{R}$ ($d = 2$ or 3), based on information measured $I(s\omega + t\omega^\perp)$ at $s\omega + t\omega^\perp$. Here $\omega$ is just a direction, i.e. a unit vector in $\mathbb{R}^d$. This is a description of coordinates on lines. The difficulty here is that what we want is in $\mathbb{R}^d$, but the data is from lines, i.e. one-dimensional.

We have $y(\omega, s) = \int_{\mathbb{R}} x(s\omega + t\omega^{\perp})\, dt$. The description of which lines we have is called acquisition geometry. Notice that the integral is a linear operator, so in the discretised version we get a matrix.

## 1.2 Mathematical formalisation

We want to recover some signal $x \in X$, from data $y \in Y$ where $y = A(x) + \varepsilon$, where $A : X \to Y$ is the forward operator which models how data is generated from the signal without noise, $\varepsilon \in Y$ models noise. In this course in particular, $A$ will be a linear operator. We will also assume that we know a useful bound for the error $\|\varepsilon\| \leq \delta$. We ignore the statistical characteristic of $\varepsilon$.

We also assume that $A$ to be a linear, bounded, operator.

# 2 Lecture 2 (2025-08-26)

**From last time**

- Naïve reconstruction (methods to solve the inverse problem)
- *Inverse crimes*

**For today**

- Hilbert spaces
- Compact operators
- Moore-Penrose inverse (pseudoinverse) didn't have time for
- Spectral theorem and SVD mentioned SVD?

## 2.1 Hadamard conditions

H1 There should be at least one solution (existence)

H2 There should be at most one solution (uniqueness)

H3 The solution should depend continuously on data (stability)

A problem where all three conditions holds, is called *well-posed*. If any fails, then it is *ill-posed*.

Unfortunately most inverse problems are ill-posed. By changing the notion of solution we can often handle conditions 1 and 2, so stability is usually the bigger problem.

## 2.2 Naïve reconstruction

Assume $A : X \to Y$ injective, with $A^{-1} : A(X) \to X$. If $y \notin A(X)$ (due to $\varepsilon$), then there is no $x \in X$ such that $y = A(x)$.

If $y \notin A(X)$, project $y \mapsto \tilde{y} \in A(x)$. We then have two naïve inversions

1. $x^* = A^{-1}\tilde{y}$
2. If $A$ is not invertible, use least-squares, solving $x^* = \operatorname{argmin}_x \|A(x) - y\|_y^2$ or $x^* = \operatorname{argmin}(\|x\|)$ such that $x$ is a solution to $A(x) = \tilde{y}$.

In order to invert the convolution we use Fourier transform $\mathscr{F}[y] = \mathscr{F}[x]\mathscr{F}[k]$. We can then just take $x = \mathscr{F}^{-1}\left[\frac{\mathscr{F}[y]}{\mathscr{F}[k]}\right]$. This clearly doesn't work if $\mathscr{F}[k]$ ever is zero, the division is the problem.

For the tomography problem in 2D, we have (harmonic analysis needed) an inversion formula. $t \in \mathbb{R}^2$

$$x(t) = A^{-1}(y)(t) = \frac{1}{2\pi}\int_0^{\pi} G(y)(\theta, \underbrace{t\cdot\omega(\theta)}_{p})\,d\theta$$

where

$$G(y)(\theta,p) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathscr{F}\left[y(\theta,\cdot)\right](\zeta)e^{i\zeta\cdot p}|\zeta|\,d\zeta = \mathscr{F}^{-1}\left[\mathscr{F}\left[y(\theta,\cdot)\right](\zeta)|\zeta|\right].$$

.

## 2.3 Inverse crimes

- Testing methods on noise-free data
- If you insist on noise-free data, you should not use the same sampling frequency in the for $m$ and $n$ (data is in $\mathbb{R}^m$ and cause in $\mathbb{R}^n$).

## 2.4 Can we work in finite dimensions?

Actual measured data is given on finitely many points. There is a sampling operator $S_y : Y \to \mathbb{R}^m$. In the *semi-discrete inverse problem* we keep $x$ continuous but use sampled data. $\bar{y} = (S_y \circ A)(x) + \bar{\varepsilon}$. There is not much to do here however, we should go fully discrete or fully continuous.

We need $E_x : \mathbb{R}^n \to X$ to extend the finite $x$ to actual functions. We want $(S_y \circ A \circ E_x) = \mathcal{A}$. If all the operators are linear, then the total transformation is also linear.

In the fully discrete problem, we want to find $\bar{x} \in \mathbb{R}^n$ such that $\bar{y} = (S_y \circ A \circ E_x)\bar{x} + \bar{\varepsilon}$. The matrix we obtain for $\mathcal{A}$ quickly gets obscenely large, and this is a huge problem in applications that we will need to handle. It is often not possible to store the entire matrix in memory.

## 2.5 Singular Value Decomposition

Any matrix $A \in \mathbb{R}^{m\times n}$ has a singular value decomposition $A = UDV^*$ where $D$ is diagonal with positive, sorted falling, values on the diagonal and $U, V$ are unitary. The condition number is given as the ratio of the largest and smallest singular value.

Problem with H3 in discrete setting is that even though the condition number is large, the discrete problem is still very much continuous. If $\operatorname{cond} A \to \infty$ as $n, m \to \infty$ that is a problem, as we do not get more accuracy with better sampling.

# 3 Lecture 3 (2025-09-01)

**For today**

- Functional analysis on real vector spaces
- Moore-Penrose inverse
- Compact operators

## 3.1 Functional analysis

**Definition 3.1.** A *normed linear space* is a vector space $X$ over $\mathbb{R}$, together with a *norm* $\|\cdot\| : X \to \mathbb{R}^{\geq 0}$, such that

(a) $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in X$,

(b) $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$, and

(c) $\|x\| = 0 \iff x = 0$.

**Definition 3.2.** A normed linear space where every Cauchy sequence converges is called a *Banach space*.

**Definition 3.3.** Let $X$ is a vector space. An *inner product* is a binary map $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ such that

(a) $\langle x_1, x_2 \rangle = \langle x_2, x_1 \rangle$,

(b) $\langle \alpha x_1, x_2 \rangle = \alpha \langle x_1, x_2 \rangle$,

(c) $\langle x_1 + x_2, x_3 \rangle = \langle x_1, x_3 \rangle + \langle x_2, x_3 \rangle$, and

(d) $\langle x, x \rangle \geq 0$, with equality only for $x = 0$.

**Definition 3.4.** A *Hilbert space* is a Banach space where the norm is defined by an inner product $\|x\| = \sqrt{\langle x, x \rangle}$.

## 3.2 Moore-Penrose Inverse

If H1 or H2 fails to hold, we change what we mean by solution. Often we go for the minimum norm solution. The Moore-Penrose inverse is a way of obtaining the minimum norm solution.

**Definition 3.5.** (a) $x^\dagger \in X$ is a least square solution to an inverse problem, if $x^\dagger \in \operatorname{argmin}_{x \in X} \|Ax - y\|_Y$

(b) $x^\dagger \in X$ is a minimum norm solution to the inverse problem if

$$x^\dagger = \begin{cases} \operatorname{argmin}_{x \in X} \|x\|_X \, , \\ x \text{ is a least squares solution.} \end{cases}$$

Assume range$(A)$ is not closed, then $X \mapsto \|Ax - y\|$ does not attain a minimum. Thus, the least squares solution does not exist. It is natural to ask for what problems a least squares solution exist.

**Theorem 3.1** (Moore-Penrose)**.** *Assume $A \in \mathcal{L}(X, Y)$, and set $\tilde{A} : \ker(A)^\perp \to A(X)$ as $\tilde{A} = A\Big|_{\ker(A)^\perp}$, then there exists a **unique** extension of $A^\dagger$ of $\tilde{A}^{-1}$ where the domain of* $\text{domain}(A^\dagger) = A(X) \oplus A(X)^\perp$ *and* $\ker A^\dagger = A(X)^\perp$.

There are three ways to define the Moore-Penrose inverse on a Banach space.

**Theorem 3.2.** *If $y \in \text{domain}(A^\dagger)$, then $x$ is a least squares solution if and only if $A^*Ax = A^*y$.*

The condition H3 fails if $A^\dagger$ does not exist or if $A^\dagger$ is not continuous.

If $A$ is a linear map, then H1-H3 are not independent.

**Theorem 3.3.** *If $X, Y$ are Banach spaces and $A \in \mathcal{L}(X, Y)$. Then H1+H2 $\implies$ H3.*

*Proof.* Any linear map between normed spaces that is bounded, is continuous. Thus, it is enough to show that $A^{-1}$ is linear and bounded. Do this perhaps □

What is the graph of an operator???

## 3.3 Compact operators

Compact operators are the infinite dimensional analogue to ill-conditioned matrices.

**Definition 3.6.** Let $X, Y$ are Banach spaces. The operator $A \in \mathcal{L}(X, Y)$ is *compact* if it maps bounded sequences in $X$ to sequences in $Y$ which has a convergent subsequence.

Think about how Theorem 3.3 and 3.4 go together.

**Theorem 3.4.** *If $A \in \mathcal{L}(X, Y)$, compact and $\dim(A(X)) = \infty$, then $A^\dagger$ is not continuous.*

**Remark.** Equivalently, the preimage of $A$ is a compact set.

Examples of compact operators are

- if $A \in \mathcal{L}(X, Y)$ has finite rank, then $A$ is compact.
- The identity mapping $I : X \to X$ is compact if and only if $X$ is finite dimensional.
- If $A \in \mathcal{L}(X, Y)$ and $B : Y \to E$ is compact, then $B \circ A$ is compact.
- If we have Hilbert spaces, $A$ is compact if and only if $A^*$ is compact.

Prove the statement about the identity mapping

**Theorem 3.5.** *Let $A \in \mathcal{L}(X, Y)$ be compact. If $A(X)$ is infinite dimensional, then $A^\dagger$ is not continuous, so H3 does not hold.*

*Proof.* $A$ is injective $\implies$ $A^\dagger = A^{-1}$. Assume $\dim A(X) = \infty$ and that $A^{-1} \in \mathcal{L}(X, Y)$. Then the identity mapping $I = A^{-1} \circ A$, is continuous and compact. Then $X$ is finite dimensional, contradicting $\dim A(X) = \infty$. □

**Theorem 3.6.** *If $A_n$ is a sequence of compact operators converging to $A$, then $A$ is compact.*

Example: $X = L^2(\Omega)$, $A : X \to X$ defined by $Ax = \int_\Omega k(s,t)x(t)\,dt$ with $k \in L^2(\Omega \times \Omega)$, then $A$ is compact.

Proof: The space $L^2(\Omega)$ has an orthonormal basis $\{\varphi_i\}$. So does $L^2(\Omega \times \Omega)$, set $\{\varphi_i(s)\varphi_j(t)\}$. We can write $k = \sum_{i,j} k_{i,j}\varphi_i\varphi_j$, where $k_{i,j}$ are the Fourier coefficients $k_{i,j} = \int_{\Omega^2} k\varphi_i\varphi_j$. Equalities are almost everywhere. Set $k_n = \sum_{i,j=0}^{n} k_{i,j}\varphi_i\varphi_j$ and $A_n(x) = \int_\Omega k_n(s,t)x(t)\,dt$. If each $A_n$ is compact, and if $A_n \to A$, then $A$ is compact. Notice that $\dim A_n(X)$ is finite, .....

Let $X_n = A_n(X)$, so $X_n$ has finite dimension.

This is theorem 3.4 in the book

$$\begin{aligned}
\|(A - A_n)x\|_{X_n}^2 &= \left\| \int_\Omega (k - k_n)(s,t)x(t)\,dt \right\|_{X_n}^2 \\
&= \int_\Omega \left| \int_\Omega (k - k_n)(s,t)x(t)\,dt \right|^2 ds \\
&\leq \int_\Omega \left( \int_\Omega |k - k_n| x\,dt \right)^2 ds \\
&\leq
\end{aligned}$$

**Definition 3.7.** On $\mathcal{L}(X, Y)$ we define the norm by $\|A\| = \sup_{\|x\|_X = 1} \|Ax\|_Y$.

# 4 Lecture 4 (2025-09-04)

**Today**

- Singular Value Decomposition of compact operator
- Formal definition of regularisation method
- Spectral regularisation
  - Truncated SVD
  - Tikhonov regularisation

## 4.1 SVD of operators

Operators might not have countably many singular values. Compact operators however, luckily, have countably many singular values.

**Theorem 4.1.** *Let $X, Y$ be Banach spaces. If $A \in \mathcal{L}(X, Y)$ is compact and $\dim(A(X)) = \infty$, then $A^\dagger : Y \to X$ is not continuous. Furthermore, $A^\dagger$ is continuous $\iff$ $A(X)$ is closed.*

**Theorem 4.2** (Spectral decomposition)**.** *Let $X$ be a Hilbert space, $N : X \to X$ is self-adjoint and compact. Then there exists an ON-basis $\{\varphi_j\} \subset X$ of $\overline{N(X)}$ and $\{\lambda_j\} \subset \mathbb{R}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ such that $N(x) = \sum_j \lambda_j \langle x, \varphi_j \rangle \varphi_j$ for all $x \in X$.*

**Theorem 4.3** (SVD of compact operator)**.** *Let $X, Y$ be Hilbert spaces, and $A \in \mathcal{L}(X, Y)$ be compact. Then there exists:*

*(a)* $\{\sigma_j\} \subset \mathbb{R}$ *s.t.* $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$,

*(b)* $\{\varphi_j\} \subset X$ *is an ON-basis of* $\ker A^\perp$, *and*

*(c)* $\{\psi_j\} \subset Y$ *is an ON-basis of* $\overline{A(X)}$

*such that*

*(i)* $A(\varphi_j) = \sigma_j \psi_j$, *and* $A * (\psi_j) = \sigma_j \varphi_j$,

*(ii)* $A(x) = \sum_j \sigma_j \langle x, \varphi_j \rangle_X \psi_j$ *for all* $x \in X$.

*(iii)* $A^*(y) = \sum_j \sigma_j \langle y, \psi_j \rangle_Y \varphi_j$ *for all* $y \in Y$.

Notice that the eigenvalues of $N = A^*A$ with eigenvectors $\{\varphi_j\}$ are also eigenvalues of $AA^*$ with eigenvectors $\{\psi_j\}$, we have $\sigma_j = \sqrt{\lambda_j}$ and $\psi_j = \frac{1}{\sigma_j} A\varphi_j$.

**Theorem 4.4** (Picard criteria)**.** *Let $A \in \mathcal{L}(X, Y)$ be compact, with SVD $\{(\sigma_j, \varphi_j, \psi_j)\}$ and $y \in \overline{A(X)}$, then*

$$y \in A(X) \iff \sum_j \frac{\left|\langle y, \psi_j \rangle_Y\right|^2}{\sigma_j^2} < \infty.$$

**Theorem 4.5** (SVD of $A^\dagger$)**.** *If $A \in \mathcal{L}(X,Y)$ compact with SVD $\{(\sigma_j, \varphi_j, \psi_j)\}$ and $y \in$ domain $A^\dagger$, then*

$$A^\dagger y = \sum_j \frac{1}{\sigma_j} \langle y, \psi_j \rangle_Y \varphi_j.$$

Notice that the right-hand side accepts any $y \in Y$, could we extend $A^\dagger$ this way?

- The SVD of $A^\dagger$ encodes the stability properties of "inverting" $A$.
- Consider $y^\delta = y + \varepsilon$, where $y$ is the ideal data and we take $\varepsilon = \delta\psi_j$. The minimum norm solutions $x^\dagger = A^\dagger y$, and $x^\delta = A^\dagger y^\delta$. The difference becomes $\frac{\delta}{\sigma_j}\varphi_j$. Which for small $\sigma_j$ might grow very large.
- We can use the singular values to put a "measure" on ill-posedness.
  - We can say that a problem is severely ill-posed if the singular values decay exponentially.
  - Moderately ill-posed if it is not severely ill-posed.

## 4.2 Regularisation method

Earlier we changed the solution to get out of the problems of H1 and H2, can we do it again to solve H3?

The first idea is to truncate the SVD, and look in a finite dimensional subspace.

The second is *filter functions* where we instead apply functions to the singular values first.

**Definition 4.1.** Let $X, Y$ be Banach spaces and $A \in \mathcal{L}(X,Y)$. A family $\{R_\alpha\}$ is a regularisation of $A^\dagger : Y \to X$ if

(a) each $R_\alpha : Y \to X$,

(b) each $R_\alpha$ is continuous

(c) $\lim_{\alpha\to 0} R_\alpha(y) = A^\dagger(y)$ for all $y \in$ domain $A^\dagger$.

If each $R_\alpha \in \mathcal{L}(X,Y)$, then the regularisation is said to be linear.

Hence, we have reformulated the inverse problem by looking for solutions to $R_\alpha$ and thus ensured H3.

**Theorem 4.6.** *Let $X, Y$ be Hilbert spaces, $A \in \mathcal{L}(X,Y)$, and $\{R_\alpha\}$ is a linear regularisation. If*

*(a) $A^\dagger$ is not continuous, then $\{R_\alpha\}$ is not uniformly bounded. Then there exists $y \in Y$ such that $\|R_\alpha y\|_X \to \infty$ as $\alpha \to 0$.*

*(b) $\|AR_\alpha\|_{\mathcal{L}(X,Y)} < \infty \implies \|R_\alpha y\|_X \to \infty$ as $\alpha \to 0$ for all $y \notin$ domain $A^\dagger$.*

In applications, we cannot expect $y \in \text{domain } A^\dagger$. Consider data $y^\delta = y + \varepsilon$ with $\left\|y^\delta - y\right\| \leq$ *delta*. And a linear regularisation $R_\alpha$, the total error $\left\|R_\alpha(y^\delta) - A^\dagger(y)\right\|_X \leq \|R_\alpha(\varepsilon)\|_X + \left\|R_\alpha y - A^\dagger y\right\|_X \leq \delta \|R_\alpha\|_{\mathcal{L}(X,Y)} + \left\|R_\alpha y - A^\dagger y\right\|_X$. First term is *data error* and second *approximation error*.

The data error does not stay bounded as $\alpha \to 0$. The approximation error goes to 0 as $\alpha \to 0$. Picking a good $\alpha$ is almost like the bias-variance trade-off.

## 4.3 Spectral regularisation

Tikhonov was the first to study regularisation that wasn't built problem specific, that it generally applicable.

**Definition 4.2** (Spectral regularisation)**.** Let $X, Y$ be Hilbert spaces and $A \in \mathcal{L}(X, Y)$, so $A^\dagger$ is not continuous, and $A$ have SVD $\{(\sigma_j, \varphi_j, \psi_j)\}$. The SVD of $A^\dagger$ is $\{(1/\sigma_j, \psi_j, \varphi_j)\}$. Define $R_\alpha : Y \to X$ as

$$R_\alpha(y) = \sum_j g_\alpha(\sigma_j) \langle y, \psi_j \rangle \varphi_j, \quad \forall y \in Y,$$

where $g_\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ is called a *spectral filter* and

(a) $g_\alpha(\sigma) \to \frac{1}{\sigma}$ as $\alpha \to 0$,

(b) $g_\alpha \leq C_\alpha$ for all $\sigma > 0$.

The first condition ensures convergence to $A^\dagger$ and the second gives continuity of $R_\alpha$.

**Definition 4.3.** *Truncated SVD* is a special case of spectral regularisation, given by choosing

$$g_\alpha(\sigma) = \begin{cases} \frac{1}{\sigma}, & \sigma \geq \alpha \\ 0, & \text{else.} \end{cases}$$

**Definition 4.4.** In *Tikhonov regularisation* we take $g_\alpha(\sigma) = \frac{\sigma}{\sigma^2 + \alpha}$.

In large scale problems it becomes impossible to compute the SVD, this causes big problems.

# 5 Lecture 5 (2025-09-08) CANCELLED

Ozan was sick :c

# 6 Lecture 6 (2025-09-11)

**Today**

- Parameter choice rules
- Convergence rates / stability estimates
- A priori parameter choice rules
- A posteriori parameter choice rules
- Spectral regularisation

## 6.1 Parameter choice rules

**Definition 6.1.** A *parameter choice rule* is a formally a mapping $\alpha : \mathbb{R}^+ \times Y \to \mathbb{R}^+$. We call $\alpha$

(i) *a priori* if it only depends on $\delta$,

(ii) *a posteriori* if it depends on both $\delta$ and $y^\delta$,

(iii) *heuristic* if it only depends on $y^\delta$.

**Definition 6.2.** A regularisation $\{R_\alpha\}$ of $A^\dagger$ is convergent with parameter choice rule $\alpha : \mathbb{R}^+ \times Y \to \mathbb{R}^+$ if

$$\lim_{\delta \to 0} \sup_{u \in Y : \|y-u\| \leq \delta} \left\| R_{\alpha(\delta,u)}(u) - A^\dagger(y) \right\|_X = 0$$

hold for any $y \in \operatorname{domain} A^\dagger$.

**Remark.** Each $u \in Y$ such that $\|u - y\|_Y \leq \delta$ can be written as $u = y + \varepsilon$ with $\|\varepsilon\|_Y \leq \delta$.

Can we determine "forms" of choice rules?

## 6.2 Convergence rates / stability estimates

**Definition 6.3.** Assume a regularisation $\{R_\alpha\}$ which is convergent with parameter choice rule $\alpha$. Assume $y^\delta = y + \varepsilon$ with $\|\varepsilon\| \leq \delta$. A *stability estimate* is a function $C : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\left\| R_{\alpha(\delta,y^\delta)}(y^\delta) - A^\dagger(y) \right\|_X \leq C(\delta)$$

where $C$ is increasing and vanishing at 0.

## 6.3 A priori choice rules

**Theorem 6.1.** *Let $\{R_\alpha\}$ be a regularisation method for $A^\dagger$. Then there exists an a priori choice rule $\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\{R_\alpha\}$ is convergent with $\alpha$.*

**Theorem 6.2.** *Let $\{R_\alpha\}$ be a linear regularisation method for $A^\dagger$. It is convergent with respect to an a priori choice rule $\alpha$ if and only if*

*(a)* $\lim_{\delta \to 0} \alpha(\delta) = 0$, *and*

*(b)* $\lim_{\delta \to 0} \delta \left\| R_\alpha \right\|_{\mathcal{L}(X,Y)} = 0$.

Often people choose $\alpha = \delta^p$ for some $p \in (0,1)$. We define the *best* $p$ as the one which achieves the fastest convergence rate. Finding the best $p$ depends on the regularisation method and requires additional information about $x^\dagger = A^\dagger y$.

## 6.4 A posteriori choice rules

We will only study one, the most widely used rule. For notation's sake, let $x_{\alpha,\delta} = R_\alpha y^\delta$, and $y = A^\dagger x^\dagger$, where $x^\dagger$ is the minimum norm least square solution. The idea is to chose $\alpha(\delta, y^\delta)$ such that

$$\left\| A(x_{\alpha(\delta,y^\delta),\delta}) - y^\delta \right\| \leq \delta.$$

In order to do this practically, we take a sequence of $\alpha_1 > \alpha_2 > \cdots > 0$ and compute $x_j = R_{\alpha_j} y^\delta$, and then just check which $\alpha$ is the largest acceptable. This is called the Morozov discrepancy principle.

## 6.5 Spectral regularisation

Let $A$ be a compact linear operator from $X$ to $Y$. Then $A$ has a singular value decomposition $\{(\sigma_j, \varphi_j, \psi_j)\}$.

**Theorem 6.3.** *If $x_\alpha = R_\alpha y$, where $\{R_\alpha\}$ is Tikhonov regularisation, then*

$$x_\alpha = \operatorname{argmin}_{x \in X} \left\| Ax - y \right\|_Y^2 + \alpha \left\| x \right\|_X^2,$$

*and $x_\alpha$ solves $(A^*A + \alpha I)x_\alpha = A^*y$.*

If we change the last term in the first equation we start talking about variational regularisation, which in some ways is much more powerful than spectral regularisation. The second term in some way embeds prior knowledge, and penalises unnatural behaviour.

# 7 Lecture 7 (2025-09-15)

**Today**

- Tikhonov regularisation as variational model
- General variational models wasn't time for
  - Interpretation
  - Existence and uniqueness
  - Convergence
  - Stability estimates
  - Computational methods
  - Examples

## 7.1 Tikhonov regularisation as a variational model

A variational model means that we define the model as the solution to an optimisation problem.

**Advantages and disadvantages of definition via SVD**

(+) Special case of spectral regularisation, that can be theoretically analysed.

(−) The SVD is not always computationally unfeasible.

(−) Requires $A$ to be linear and compact, it would be nice to be able to apply the method to a wider class of problems.

(−) Regularising property only operates by filtering singular values. Hard to encode other prior knowledge.

We address these problems through the variational formulation of Tikhonov regularisation.

First we should prove Theorem 6.3.

*Proof of Theorem 6.3.* First prove $x_\alpha = R_\alpha \implies (A^*A + \alpha I)x_\alpha = A^*y$. Notice

$$(\alpha I)x_\alpha = \sum \alpha \frac{\sigma}{\sigma^2 + \alpha} \langle y, \psi \rangle \varphi,$$

and

$$\begin{aligned}(A^*A)x_\alpha &= (A^*A) \sum \frac{\sigma}{\sigma^2 + \alpha} \langle y, \psi \rangle \varphi \\ &= \sum \frac{\sigma}{\sigma^2 + \alpha} \langle y, \psi \rangle (A^*A)\varphi \\ &= \sum \frac{\sigma}{\sigma^2 + \alpha} \langle y, \psi \rangle \sigma^2 \varphi\end{aligned}$$

Where we are allowed take $(A^*A)$ inside due to convergence and continuity. Thus,

$$(A^*A + \alpha I)x_\alpha = \sum \sigma \langle y, \psi \rangle \varphi = A^*y.$$

Now for the other direction. Assume $x \in X$ solves $(A^*A + \alpha I)x = A^*y$, show that $x = x_\alpha = R_\alpha y$ for some $\alpha > 0$. We know that $X = \ker A \oplus \ker A^\perp$, and $\ker A^\perp = \overline{\operatorname{range} A^*}$. From SVD theorem we have $\varphi_n$ is an ON-basis of $\overline{\operatorname{range} A^*} \subseteq X$, and that $A^*A\varphi = \sigma^2\varphi$. Since $x \in X$, we can write it as $x = x' + x''$, with $x' \in \overline{\operatorname{range} A^*}$, $x'' \in \ker A$, and

$$x' = \sum \langle x, \varphi \rangle \varphi$$

Which means that

$$x = \underbrace{\sum \langle x, \varphi \rangle \varphi}_{x'} + \underbrace{\pi(x)}_{x''}$$

where $\pi$ is the projection onto the $\ker A$. We know that

$$\begin{aligned}(A^*A + \alpha I)(x' + x'') &= A^*Ax' + \alpha I x' + \underbrace{A^*Ax''}_{=0} + \alpha I x'' \\ &= \sum (\sigma^2 + \alpha) \langle x, \varphi \rangle \varphi + \alpha x'' \\ &= \sum \sigma \langle y, \psi \rangle \varphi\end{aligned}$$

Since the right-hand side is in range $A^*$, we must have $x'' = 0$, so $x = x'$. Thus, since we can solve for the factor in front of each basis vector individually,

$$x' = x = \sum \frac{\sigma}{\sigma^2 + \alpha} \langle y, \psi \rangle \varphi$$

thus $x = x_\alpha$. □

**Remark.** Notice we didn't prove the optimisation statement.

The normal equations $A^*Ax = A^*y$ is equivalent to $x = \operatorname{argmin} \|Ax - y\|_Y^2$.

*Proof of optimisation part of Theorem 6.3.* Knowing that $x = R_\alpha y$ we have to show that $x_\alpha$ is the minimiser to $x' \mapsto \|Ax - y\|_Y^2 + \alpha \|x\|_X^2 = \mathcal{E}_\alpha(x')$. Equivalently, that if $x \in X$ then $\mathcal{E}_\alpha(x) > \mathcal{E}_\alpha(x_\alpha)$. Let $x \in X$, then

$$\begin{aligned}\mathcal{E}_\alpha(x) - \mathcal{E}_\alpha(x_\alpha) &= \langle Ax - y, Ax - y \rangle + \alpha \langle x, x \rangle - \langle Ax_\alpha - y, Ax_\alpha - y \rangle + \alpha \langle x_\alpha, x_\alpha \rangle \\ &= \|Ax - Ax_\alpha\|^2 + \alpha \|x - x_\alpha\|^2 + 2 \left\langle \underbrace{A^*(Ax_\alpha - y) + \alpha x_\alpha}_{=0}, x - x_\alpha \right\rangle\end{aligned}$$

Now for the other direction. Assume $\tilde{x}$ minimises $\mathcal{E}_\alpha$. We must show $R_\alpha y = \tilde{x}$. Since $\tilde{x}$ is a minimiser, so $\mathcal{E}_\alpha(x) - \mathcal{E}_\alpha(\tilde{x}) \geq 0$ for any $x \in X$. In particular, $x = \tilde{x} + tx_0$, for $t > 0$ and $x_0 \in X$. Then

$$0 \leq \mathcal{E}_\alpha(x) - \mathcal{E}_\alpha(\tilde{x}) = t^2 \|Ax_0\|^2 + t^2\alpha \|x_0\|^2 + 2t \langle A^*(A\tilde{x} - y) + \alpha\tilde{x}, x_0 \rangle$$

Divide by $t > 0$ and take limit as $t \to 0$. Then

$$\langle A^*(A\tilde{x} - y) + \alpha\tilde{x}, x_0\rangle \geq 0$$

for any $x_0 \in X$. Since $x_0$ is arbitrary we have that the left entry of the inner product is 0. (Otherwise $-x_0$ would give negative inner product which isn't allowed). Thus, $\tilde{x} = x_\alpha$. $\square$

Thus, we could define Tikhonov regularisation from the optimisation problem, but using that as a starting point makes it difficult to prove that it is formally a regularisation method.

## Variational models

Generally we write variational models as

$$R_\alpha(y) = \operatorname{argmin} \mathcal{L}_Y(Ax, y) + \alpha s(x)$$

where $\mathcal{L}_Y : Y \times Y \to \mathbb{R}^+$ measures the data fidelity and $s : X \to \overline{\mathbb{R}^+}$ is the regulariser. It is often used to encode prior knowledge.

# 8 Lecture 8 (2025-09-18)

**Today**

- Examples of variational models
- Theory of variational models

A variational model defines a mapping $R_\alpha : Y \to X$ such that

$$R_\alpha(y) \in \operatorname{argmin}_{x \in X} \mathcal{E}_{\alpha(x),y}$$

with $\mathcal{E}_\alpha(x) = \frac{1}{2}\|Ax - y\|^2 + \alpha S(x)$, where $S$ is the regulariser (in some sense, not rigorous) chosen to ensure $R_\alpha$ is a regularisation model, $S(x) = \frac{1}{2}\|x\|^2$ is classical Tikhonov regularisation.

## 8.1 Tikhonov-Philips Ellipse

Assume $\mathscr{D} : X \to Z$ linear, and $S(x) = \frac{1}{2}\|\mathscr{D}(x)\|_Z^2$.

- The Dirichlet is a special case given by $\mathscr{D}(x) = \nabla x$.
- Let $\mathscr{D} : H^1(\Omega) \to \mathcal{L}^2(\Omega, \mathbb{R}^d)$, $H^1(\Omega) = \{x \in \mathcal{L}^2 : \nabla x \in \mathcal{L}^2(\omega, \mathbb{R}^d)\}$.
- Similar, the norm on $H^1$ $S(x) = \frac{1}{2}\|x\|^2 + \|\nabla x\|^2$. Obviously you can weight this differently, introducing more parameter.

## 8.2 Total variation

The Tikhonov regularisers smooth edges, that is not always desirable. Let $X = W^{1,1}(\Omega) = \{x \in \mathcal{L}^1(\Omega) : \nabla x \in \mathcal{L}^1(\Omega, \mathbb{R}^d)\}$ The TV (total variation) functional is a mapping

$$\mathrm{TV} : X \to \mathbb{R}^+, \qquad \mathrm{TV}(x) = \int_\Omega |\nabla x|$$

In many applications $W^{1,1}$ is too restrictive, as we need jump discontinuities. So we have to extend TV to a broader class of functions.

$$\mathrm{BV}(\Omega) = \{x \in \mathcal{L}^1 : \|x\|_{\mathcal{L}^1} + \mathrm{TV} < \infty\}$$

Where

$$\mathrm{TV}(x) = \sup_{\varphi \in D(\Omega, \mathbb{R}^d)} \int_\Omega u(x)(\nabla \cdot \varphi)(t)\, dt$$

and $D(\Omega, \mathbb{R}^d) = \{\varphi \in C_c^\infty(\Omega, \mathbb{R}^d) : \sup_{t \in \Omega} |\varphi(t)|_{\mathbb{R}^d} < 1\}$. Unfortunately then, BV is no longer a Hilbert space, but only Banach.

## $\ell^1$-regularisation (Lasso)

Here $X$ is the set of sequences in $\ell^1$, and $A : \ell^1 \to \ell^2$. And $S(x) = \|x\|_{\ell^1}$. This is sparsity promoting for some reason. Only proven in 2004, compressed sensing.

## 8.3 Theory

We have still not properly defined the mapping $R_\alpha$, as we haven't shown uniqueness of the minimiser. In variational models, we define $R_\alpha$ indirectly as the solution to a minimisation problem. Thus, we have to show that the problem has a solution, and preferably that it is unique.

We also need to show that $y \mapsto R_\alpha(y)$ is a continuous mapping, and that $R_\alpha(y) \to x^\dagger$ as $\alpha \to 0$. If all of this is done, then we can actually call it a regularisation method.

The first step is quite technically demanding. So we will be trying to dumb it down. Normally this is done in Banach spaces, we try a slightly easier version in Hilbert spaces.

**Today and next time most technically demanding lectures of the course!!!!**

**Basic assumptions (Hilbert setting)**

(a) $A : X \to Y$ continuous

(b) $S : X \to \overline{\mathbb{R}}$ is *proper* and weakly *lower semi-continuous* (l.s.c.)

(c) $S$ is *coercive*

**Remark.** If $S$ is continuous and convex, then it is weakly l.s.c.

**Definition 8.1.** A function $S : X \to \overline{\mathbb{R}}$

(a) is proper if $S(x) > -\infty$ and domain $S \neq \emptyset$,

(b) is weakly l.s.c. at $x_0 \in X$ if

$$\liminf_{n\to\infty} S(x_n) \geq S(x_0)$$

whenever $x_n$ weakly converges to $x_0$,

(c) is coercive if $S(x_n) \to \infty$ whenever $\|x_n\| \to \infty$.

**Theorem 8.1.** *If the basic assumptions hold. Then $\mathcal{E}_{\alpha,y} : X \to \overline{\mathbb{R}}$ has a minimiser for any $y \in Y$ and $\alpha > 0$.*

No proof :c, but we have obtained existence. Relies on Banach-Alaoglu theorem. Called Generalised Weierstraß theorem.

**Theorem 8.2.** *If the basic assumptions hold, $S$ is strictly convex or, $S$ is convex and $A$ is injective. Then $\mathcal{E}_{\alpha,y}$ has a unique minimiser.*

**Remark.** Makes $R_\alpha$ well-defined.

We might wonder, is the data-fidelity term convex? The answer is yes, it also strictly convex if and only if it is injective.

*Proof of convexity of data-fidelity.* Need to show that for $Q(x) = \frac{1}{2}\|Ax - y\|^2$, we have

$$Q(\lambda x' + (1-\lambda)x'') < \lambda Q(x') + (1-\lambda)Q(x'')$$

A "lengthy" calculation gives:

$$Q(\lambda x' + (1-\lambda)x'') = \lambda Q(x') + (1-\lambda)Q(x'') - \frac{\lambda(1-\lambda)}{2}\|A(x'-x'')\|^2$$

the final term is clearly positive, so we have convexity. If $A$ is injective $Ax' - x'' \neq 0$, since $x' \neq x''$, so strict. If $A$ is not Something is wrong here, the inequality just changed direction. □

# 9 Lecture 9 (2025-09-22)

**Today**

- More theory for variational models
- Optimisation methods

## 9.1 Theory of Variational Models

**Theorem 9.1** (Stability)**.** *Let $R_\alpha : Y \to X$ be defined as $R_\alpha \in \operatorname{argmin} \mathcal{E}_{\alpha,y}(x)$, and that the basic assumptions hold. Also, let $y_n \to y$ in $Y$. Then the following holds*

*(a) $R_\alpha(y_n)$ converges weakly to $R_\alpha(y)$ as $n \to \infty$, and*

*(b) If $S : X \to \overline{\mathbb{R}}$ satisfies Radon-Riesz property saying that, if $x_n$ converges to $x$ weakly and $S(x_n) \to S(x)$, then $x_n$ converges to $x$ strongly. Then $R_\alpha(y_n)$ converges to $R_\alpha(y)$ strongly and $S(R_\alpha(y_n)) \to S(R_\alpha(y))$.*

The $S$-minmising least square solution is $x^\dagger = \operatorname{argmin}_{x \in L_y} S(x)$, where $L_y = \left\{ x \in X : x \text{ minimises } z \mapsto \|Az - y\|_Y^2 \right\}$.

**Definition 9.1.** We redefine *regularisation* to mean

- $R_\alpha : Y \to X$ well-defined
- $R_\alpha$ continuous
- $R_\alpha y \to x^\dagger$ when $\alpha \to 0$, and $x^\dagger$ is the $S$-minimising least square solution.

Most statements about minimum norm solutions extend to $S$-minimising solution as well. But when do we have a unique $x^\dagger$?

**Theorem 9.2.** *Assume the basic assumptions hold. And that $y \in AX$. Then we have a unique $S$-minimisig least square solution.*

**Theorem 9.3.** *Let $R_\alpha : Y \to X$ be given as in a variational model, and $S : X \to \overline{\mathbb{R}}$ satisfies the basic assumptions. Next, let $y \in AX$, and $y_n \in Y$, and $\delta_n > 0$ such that $\|y - y_n\| \leq \delta_n$ and $\delta_n \to 0$. Finally, we also have an a priori parameter choice rule $\alpha : \mathbb{R}^+ \to \mathbb{R}^+$, such that $\alpha(\delta) \to 0$, and $\frac{\delta^2}{\alpha(\delta)} \to 0$ when $\delta \to 0$. Then,*

*(a) $x_n = R_{\alpha_n} y_n$, with $\alpha_n = \alpha(\delta_n)$, then $x_n$ converges weakly to $x^\dagger$ as $n \to \infty$, and,*

*(b) if $S$ satisfies the Radon-Riesz property, then $x_n$ converges strongly to $x^\dagger$.*

## 9.2 Optimisation methods

Evaluating $R_\alpha$ for some given $y$ and $\alpha > 0$, we need to solve an optimisation problem:

$$\operatorname{argmin}_{x \in X} \frac{1}{2} \|Ax - y\|_Y^2 + \alpha S(x).$$

These problems become very high dimensional. In tomography for example, the images are $512 \times 512 \times 1000$ (1000 images), i.e. $\mathbb{R}^{512 \cdot 512 \cdot 1000}$. Variational models are quite slow, which is why they haven't been used much in medical applications.

Two ways of solving the problem

1. Optimise first, then discretise,
2. Discretise first, then optimise.

In the first version, we formulate the optimisation problem in the infinite dimensional setting, and convergence needs to be proven here. We will be taking the second route, however. It is important which inner product we get after discretisation. Just picking the Euclidean norm is not a good plan.

Henceforth, when we say convergence we just mean that $\overline{x}_n \to \overline{x}$ where

$$\overline{x} = \operatorname{argmin}_{x \in X} \|Ax - y\|_Y^2 + S(x)$$

where $X = \mathbb{R}^N$ with some inner product.

Suppose $E : \mathbb{R}^d \to \mathbb{R}$ is differentiable. We want to find $x_* = \operatorname{argmin} E$. The simplest method is just gradient descent.

### 9.2.1 Gradient Descent

Initialise by taking some $\overline{x}_0 \in \mathbb{R}^d$, and compute $\bar{x}_{n+1} = \bar{x}_n - \nu \nabla E(\bar{x}_n)$. If $\bar{x}_*$ is a fixed point, then $\nabla E = 0$.

The question is, how do we pick $\nu$?

**Definition 9.2.** We say that $E : \mathbb{R}^d \to \mathbb{R}$ has $L$-Lipschitz gradient if $E \in C^1(\mathbb{R}^d)$ and $\|\nabla E(x') - \nabla E(x'')\| \leq L \|x' - x''\|$.

**Theorem 9.4.** *If $0 < \nu < 1/L$, then $\min_{n=0,1,\ldots,N-1} \|\nabla E(x_n)\| \leq \frac{2}{\nu N}(E(x_0) - \inf_{x \in X} E(x))$, and if $E \in C^2(\mathbb{R}^d)$, then $x_n$ converges to a local minimiser for almost all $x_0$.*

# 10 Lecture 10 (2025-09-22)

**Today**

- Optimisation algorithms
- Let $E : \mathbb{R}^n \to \overline{\mathbb{R}}$.
- Iterative methods for computing a (local) minimiser to $E$.
- Methods
  - Second order methods (use the Hessian of $E$)
  - Quasi-second order methods
  - First order methods
  - Stochastic and coordinate-wise first order methods
- Order of methods listed from fast to slow convergence, and from bad to good for large scale problems.

## 10.1 Second order methods

Assume $E \in C^2(\mathbb{R}^d)$. We will use seconds order information (Hessian), which gives fast convergence. Can incorporate additional constraints via barrier functions. Most classical methods is Newton's methods $x_{k+1} = x_k - \nu(\nabla^2 E(x_k))^{-1}\nabla E(x_k)$, clearly expensive each step.

Interior point methods are a big field... All use second order information, and computational bottleneck is that each step needs to solve linear systems. Usually suitable for small and medium-sized problems.

## 10.2 Quasi-second order methods

Here we assume $E \in C^2(\mathbb{R}^d)$, but we only use first order information. Still fairly fast convergence, asymptotically the same, but in practice slightly slower. Best well know are the Quasi-Newton methods $x_{k+1} = x_k - \nu B_k \nabla(x_k)$ where $B_k$ approximated the inverse of the Hessian, but is computed from first order information. The computational bottleneck is now to form the matrix $B_k$. Can sometimes be used for large scale problems.

## 10.3 First order methods

First order methods are popular for inverse problems. Assume $E \in C^1(\mathbb{R})$, then we can use *gradient based methods*. If $E$ is not $C^1$, we have to use Proximal methods.

Iterates often cheaper to compute than in (Quasi-)second order methods. Cheaper both in compute power and memory footprint. We only use first order information, which also means slower convergence.

### 10.3.1 Gradient descent

- Suppose $E \in C^1(\mathbb{R}^d)$.
- Initialise by picking $x_0$ and $\nu$
- Updates $x_{k+1} = x_k - \nu \nabla E(x_k)$

If $x_*$ is a critical point to $E$, then $x_*$ is a fixed point to the iterations scheme. But which critical point do we get?

We would like to know how close to a critical point after a certain number of steps. Assume $E$ has an $L$-Lipschitz gradient, which means that

$$\|\nabla E(x') + \nabla E(x'')\| \leq L \|x' - x''\|.$$

There are many ways to choose $\nu$ in many ways, we would of course like to choose it to ensure fast convergence.

Example theorem from the field:

**Theorem 10.1.** *Let $E \in C^1(\mathbb{R}^d)$ be bounded from below and have $L$-Lipschitz gradient. Then*

*(a) If $\nu < 1/L$, then $\min_{k=0,\dots,n-1} \|\nabla E(x_k)\|^2 \leq \frac{2}{\nu n}(E(x_n) - E_*)$, with $E_* = \inf E$.*

*(b) If $E \in C^2$, $x_k$ converge to a local minimum of $E$ for almost all $x_0$.*

*(c) If $E$ is convex with global minimiser $x_*$, then $E(x_n) - E(x_*) \leq \frac{1}{2\nu n} \|x_0 - x_*\|^2$.*

Notice that the convergence rate is $1/n$, which is slow, and a problem in applications. Can we do better without making more assumptions about $E$? Yes, Nesterov acceleration (basis for Adam etc.).

Possible to improve the convergence rate for general $E$ by modifying the updates. We will only formulate the model.

### 10.3.2 Nesterov's accelerated gradient method

Initialise $x_0 \in \mathbb{R}^d$, $\nu > 0, \lambda_0 = \beta_0 = 0$.

Update $z_k = x_k + \beta_k(x_k - x_{k-1}), x_{k+1} = z_k - \nu \nabla E(z_k), \lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2}), \beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$. Gives convergence rate $1/n^2$, very good.

**Theorem 10.2.** *If $E \in C^1(\mathbb{R})$, convex, coercive, and has $L$-Lipschitz gradient, also $\nu < 1/L$. Then $x_n$ converges to a local minimiser of $E$ and $E(x_n) - E_* \leq \frac{2L}{(n+1)^2} \min_{z \in \Omega_E} \|z - x_0\|^2$, where $E_* = \inf E$, and $\Omega_E = \operatorname{argmin} E(x)$.*

Finding $L$ is difficult, often people use adaptive methods.

### 10.3.3 Proximal methods

Suppose $E$ is not necessarily differentiable. Basis of Proximal methods is non-smooth calculus, which we don't have time for.

The proximal operator (associated with $E$), is defined as

$$\text{prox}_E(x) = \text{argmin}_{z\in\mathbb{R}^d} E(z) + \frac{1}{2}\|z-x\|^2.$$

Why do we define it like this?

**Theorem 10.3.** *If $E$ is a closed, proper function. Then $z \mapsto E(z) + \frac{1}{2}\|z-x\|^2$ is strongly convex, which means $\text{prox}_E : \mathbb{R}^d \to \mathbb{R}^d$ is well-defined.*

In the problem, we actually want to compute $\text{prox}_{\nu E}$.

Let $\mathcal{U} \subset \mathbb{R}^d$ be closed and convex, and $E(x) = 0$ for $x \in \mathcal{U}$ and $\infty$ otherwise. Then $\text{prox}_E(x) = \text{argmin}_{z\in\mathcal{U}} \|x-z\|$.

Suppose $E \in C^1$, then $\text{prox}_{\nu E} \approx x - \nu\nabla E(x)$, for small $\nu$. If $x_*$ is a fixed point to $\text{prox}_{\nu E}$ if and only if $x_*$ is a minimiser of $E$.

Suppose $E : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^d$, then $\text{prox}_E$ maps points in $\Omega$ to $\Omega$, moving closer to a minimum of $E$, and points outside $\Omega$ are mapped to $\partial\Omega$.

Suppose

- $E = G + H$
  - $G \in C^1$, with $L$-Lipschitz gradient
  - $H$ is proper, coercive and l.s.c, but not necessarily differentiable.
- Initialise $x_0 \in \mathbb{R}^d$ and $\nu < 1/L$.
- Update $x_{k+1} = \text{prox}_{\nu H}(x_k - \nu\nabla G(x_k))$

This is sometimes called forward-backward splitting. For many examples we know analytic expressions of prox, but not for all, $\|x\|$ has known prox but not TV.

**Douglas-Ratchford**

- Let $E = G + H$
- Update $z_{k+1} = \frac{1}{2}z_k + \frac{1}{2}(\text{prox}_{\nu G}())$

google it lol

# 11 Lecture 11 (2025-09-29) (Guest lecturer: Evren Yarman)

**Today**

- Sparsity promoting regularisation

Forward problem: $F(x) = y$, have input $x$, output $y$, operator $F$. Where could we have "sparsity"? Anywhere, the inputs, outputs or in some sense the operator.

Let's consider the linear problem with $x \in \mathbb{R}^m, y \in \mathbb{R}^n$. Suppose we have a basis $e_k$ of $\mathbb{R}^n$.

Suppose $A = U\Sigma V^T$

The solution to $\operatorname{argmin}_x \|Ax - b\|^2$ is $(A^TA)^{-1}b$, the solution to $\operatorname{argmin}_B \|BAx - x\|^2$ is $B = A^T(AA^T)^{-1}$, this is the *Left inverse of* $A$, while the right inverse is $(A^TA)^{-1}$.

**Definition 11.1.** A *frame* is a set of vectors $D = \{d_k\}_{k=1}^{d \geq n}$ such that span $D = V$ and $A\|v\|^2 \leq \sum |\langle v, d_k\rangle|^2 \leq B\|v\|^2$ for all $v$ and for some real $A, B > 0$, these are called the frame bounds.

We can design the frames to be normal, i.e. $\|d_k\| = 1$. If $A = B$, we say we have simple frame bounds, and we have a tight frame.

For a tight frame, we can write $v = \frac{1}{A}\sum_{k=1}^m \langle v, d_k\rangle d_k$.

We can build dictionaries using PCA.

# 12 Lecture 12 (2025-10-02)

**Today**

- Smoothed TV
- Second order TGV
- Besov space norm
- Non-local TV

Today we are cherry-picking some examples of other functionals used in variational models.

## 12.1 Smoothed TV

Replace TV by

$$\mathrm{TV}^{\varepsilon}(x) = \int_{\Omega} \sqrt{|\nabla x(t)|^2 + \varepsilon^2}\, dt.$$

Similar to Huber functional from the project. Clearly we are introducing another parameter. Whether $\varepsilon$ is a regularisation parameter is a bit unclear, it can also be thought of as some other hyperparameter to be chosen a priori.

In a hospital setting, the optimisation problem is never actually solved, only two iterations in the solver are usually taken.

Total variation is bad at preserving texture, we call this staircasing.

## 12.2 Second order Total Generalised Variation

An attempt at recovering edges sharply, but without staircasing the texture. This is actually a whole family of regularisers, but we will only be doing second order.

- Usage of TV leads to solutions that have staircasing artefacts, $\ker \mathrm{TV} = \{\text{constant functions}\}$.
- Design a regulariser $S : X \to \mathbb{R}$ that reduces staircasing while preserving edges.

**Definition 12.1.** We have signals $x \in \mathcal{L}^1(\Omega), \Omega \subset \mathbb{R}^d$, and constants $\alpha, \beta > 0$. We define

$$\mathrm{TGV}^2_{\alpha,\beta}(x) = \sup\left\{ \int_{\omega} x(t)(\mathrm{div}^2\, \varphi)(t)\, dt : \varphi \in C_c^2(\Omega, S^{d\times d}), \|\varphi\|_\infty \le \beta, \|\mathrm{div}\, \varphi\|_\infty \le \alpha \right\}.$$

Where $C_c^2(\Omega, S^{d\times d})$ is the set of symmetric $d \times d$ matrices and each $\varphi_{i,j}$ is differentiable with compact support. Further, $(\mathrm{div}\, \varphi)_i = \sum_{j=1}^d \frac{\partial \varphi_{i,j}}{\partial t_j}$, and $\mathrm{div}^2\, \varphi = \sum_{i=1}^d \frac{\partial \varphi_{i,i}}{\partial t_i} + 2\sum_{i<j} \frac{\partial^2 \varphi_{i,j}}{\partial t_i \partial t_j}$. On the spaces $C_c^1(\Omega, S^{d\times d})$ and $C_c(\Omega, \mathbb{R}^d)$ are given as

$$\|q\|_\infty = \sup_{t\in\Omega} \left\{ \left( \sum_{i=1}^d |q_{i,i}(t)|^2 + 2\sum_{i<j} |q_{i,j}(t)|^2 \right)^{1/2} \right\}$$

and

$$\|p\|_\infty = \sup_{t\in\Omega} \left\{ \left( \sum_{i=1}^{d} |p_i(t)|^2 \right)^{1/2} \right\}.$$

### 12.2.1 Properties of $\mathrm{TGV}^2_{\alpha,\beta}$

- $\mathrm{TGV}^2_{\alpha,\beta}$ maps $L^p$ functions to the real numbers
  - Is proper
  - Convex, and
  - lower semicontinuous for $1 \leq p \leq \infty$.
- Rotation and translation invariant.
- $\mathrm{TGV}^2_{\alpha,\beta}$ can serve as a norm on $\mathrm{BGV}^2_{\alpha,\beta}{}^2_{\alpha,\beta}(\Omega) = \left\{x \in L^1(\Omega) : \mathrm{TGV}^2_{\alpha,\beta}(x) < \infty\right\}$, this is a Banach space with norm $\|x\| = \|x\|_1 + \mathrm{TGV}^2_{\alpha,\beta}(x)$.

The kernel is $\ker \mathrm{TGV}^2_{\alpha,\beta} = \left\{x : \mathrm{BGV}^2_{\alpha,\beta} : x = a \cdot t + b, a \in \mathbb{R}^d, b \in \mathbb{R}\right\}$, i.e. first order polynomials. Obviously the formulation from the definition is not very convenient for computation purposes, as such we have a reformulation which is derived from convex analysis using Fenchel duality.

$$\mathrm{TGV}^2_{\alpha,\beta}(x) = \min_{u\in\mathrm{BD}(\Omega)} \alpha \|x-u\|_{\mathcal{M}} + \beta \|\mathcal{E}(u)\|_{\mathcal{M}},$$

where BD are functions of bounded deformation, $\mathcal{E}(u) = \frac{1}{2}(\nabla u + \nabla u^T)$ and $\|\cdot\|_{\mathcal{M}}$ is a suitable matrix norm. With double optimisation problem (in the variational model), we can solve for $x$ and $u$ together.

## 12.3 Besov space norm

Another attempt at solving the issues with TV. Might skip.

## 12.4 Non-local TV

Gradient computation is very local, sometimes we want to compare things far away from each other spatially.

- Compare similarity in images or signals, from regions that are not necessarily close.
- Let $x : \Omega \to \mathbb{R}$, and $\Omega \subset \mathbb{R}^d$.
- Define $\nabla_{NL} x(s,t) = \sqrt{w(s,t)}(x(t) - s(t))$ for $s, t \in \Omega$.
- The weighting function $w : \Omega \times \Omega \to \mathbb{R}^+$ quantifies similarity between regions near $s$ and $t$.
  - Popular choice is $w(s,t) = \alpha_1 \exp\left(\frac{-(s-t)^2}{\sigma_1^2}\right) + \alpha_2 \exp\left(\frac{-(f(s)-f(t))^2}{\sigma_2^2}\right)$, where $\alpha, \sigma > 0$ and $f : \mathbb{R}^d \to \mathbb{R}^d$ is called a feature extractor.

**Definition 12.2.** We define $|\nabla_{NL} x|(t) = \sqrt{\int_\Omega \nabla_{NL} x(s,t)^2 \, ds}$ and

$$\mathrm{TV}_{NL}(x) = \int_\Omega |\nabla_{NL} x|(t) \, dt.$$

# 13 Lecture 13 (2025-10-06)

I was not able to attend this lecture, these notes are based on what was written by another student.

**Today**

- Bayesian view of inverse problems

We take here inverse problems of two flavours.

(i) Recover $x \in X$ from $Ax + \varepsilon = y$ where $y \in Y$, $\varepsilon$ is a $Y$-valued random variable, and $A \in \mathcal{L}(X, Y)$.

(ii) Consider both $\mathbb{x}$ and $\varepsilon$ to be samples from $X$-valued and $Y$-valued random variables.

Our approach to solving the first kind of problem is the same as usual. Apply a regularisation and construct the random variable $\hat{\mathbb{x}}_\alpha = R_\alpha(\mathbb{y})$, so that $\hat{x}_\alpha = R_\alpha(y)$.

## 13.1 Least squares solutions

**Theorem 13.1.** *Assume $x \in \mathbb{R}^d, y \in \mathbb{R}^n, n > d$ and $\operatorname{rank} A = d$. Furthermore, assume $\mathbb{E}[\varepsilon] = 0$ and $\operatorname{Var}[\varepsilon] = \Sigma \in \mathbb{R}^{n\times n}$. If we set $\hat{\mathbb{x}} = (A^*A)^{-1}A\mathbb{y}$, then*

*(a)* $\mathbb{E}[\hat{\mathbb{x}}] = x$, *and*

*(b)* $\operatorname{Var}[\hat{\mathbb{x}}] = (A^*A)^{-1}A^*\Sigma A(A^*A)^{-1}$.

*Furthermore, if we assume $\Sigma = \sigma^2 I$, i.e. we have uncorrelated errors with the same variance, then $\operatorname{Var}[\hat{\mathbb{x}}] = \sigma^2(A^*A)^{-1}$. Set $\hat{\mathbb{y}} = A\hat{\mathbb{x}}$, then*

*(c)* $\mathbb{E}[\hat{\mathbb{y}}] = Ax$,

*(d)* $\operatorname{Var}\hat{\mathbb{y}} = \sigma^2 A(A^*A)^{-1}A^*$,

*(e) If $\mathbb{r} = \mathbb{y} - \hat{\mathbb{y}}$, then $\mathbb{E}[\mathbb{r}] = 0$,*

*(f)* $\operatorname{Var}[\mathbb{r}] = \sigma^2(I - A(A^*A)^{-1}A^*)$, *and*

*(g)* $\mathbb{E}\left[\|\mathbb{r}\|^2\right] = (n-d)\sigma^2$, *such that $\hat{\sigma}^2 = \frac{\|\mathbb{r}\|^2}{n-d}$ becomes an unbiased estimator of $\sigma^2$.*

*Proof.* For (a), simple calculation gives

$$\begin{aligned}\mathbb{E}[\hat{\mathbb{x}}] =& \mathbb{E}\left[(A^*A)^{-1}A^*\mathbb{y}\right]\\ =& \mathbb{E}\left[(A^*A)^{-1}A^*\big(Ax+\varepsilon\big)\right]\\ =& x + (A^*A)^{-1}A^*\mathbb{E}[\varepsilon].\end{aligned}$$

Since the expectation of $\varepsilon$ is zero, this completes the proof of (a).

For (b), we see that

$$\begin{aligned}\operatorname{Var}\hat{\mathbb{x}} &= \operatorname{Cov}(\hat{\mathbb{x}}, \hat{\mathbb{x}}) \\ &= \operatorname{Cov}((A^*A)^{-1}A^*\mathbb{y}, (A^*A)^{-1}A^*\mathbb{y}) \\ &= (A^*A)^{-1}A^* \operatorname{Cov}(\mathbb{y}, \mathbb{y})A(A^*A)^{-1} \\ &= (A^*A)^{-1}A^* \operatorname{Cov}(\varepsilon, \varepsilon)A(A^*A)^{-1} \\ &= (A^*A)^{-1}A^*\Sigma A(A^*A)^{-1}\end{aligned}$$

□

**Theorem 13.2.** *Assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, then*

*(a)* $\hat{\mathbb{x}} \sim \mathcal{N}(x, \sigma^2(A^*A)^{-1})$,

*(b)* $\hat{\mathbb{x}}$ *and* $\mathbb{r}$ *are independent*

*(c)* $\hat{\mathbb{y}} \sim \mathcal{N}(Ax, \sigma^2 A(A^*A)^{-1}A^*)$,

*(d)* $\mathbb{r} \sim \mathcal{N}(Ax, \sigma^2(I - A(A^*A)^{-1}A^*))$, *and*

*(e)* $\frac{(n-d)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-d}$.

If we have $\Sigma$ symmetric positive definite, then if

$$\hat{x}_\Sigma = \operatorname{argmin}_x (y - Ax)^*\Sigma(y - Ax),$$

then we have $\hat{x}_\Sigma = (A^*\Sigma^{-1}A)^{-1}A^*\Sigma^{-1}y$, and $\mathbb{E}[\hat{x}_\Sigma] = x$, and $\operatorname{Var}\hat{x}_\Sigma = (A^*\Sigma^{-1}A)^{-1}$. This is called *weighted least squares*.

## 13.2 Confidence intervals

Suppose $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma^2$ is known, then a $100(1-\alpha)\%$ confidence ellipse for $\hat{\mathbb{x}}$ is given by

$$\left\{x \in \mathbb{R}^d : (\hat{\mathbb{x}} - x)A^*A(\hat{\mathbb{x}}) \le \sigma^2 \chi_\alpha(\alpha)\right\}$$

Does $\chi_\alpha$ actually make sense? Feels like a typo. Similarly, $100(1-\alpha)\%$ interval for $\hat{\mathbb{x}}_k$ is given by $\left[\hat{\mathbb{x}}_k - J_{\alpha/2}\sqrt{((A^*A)^{-1})_k}, \hat{\mathbb{x}}_{k,k} + J_{\alpha/2}\sqrt{((A^*A)^{-1})_{k,k}}\right]$, where $J_\alpha$ is the $\alpha$-quartile of $N(0, \sigma^2)$.

## 13.3 Moore-Penrose pseudoinverse

If $y = Ax + \varepsilon$, and $\ker A$ is non-trivial, then with $\hat{\mathbb{x}} = A^\dagger \mathbb{y}$, we have

- $\mathbb{E}[\hat{\mathbb{x}}] = A^\dagger Ax$
- $\operatorname{Var}\hat{\mathbb{x}} = \sigma^2(A^*A)^\dagger$.

## 13.4 Tikhonov regularisation

Suppose $X = \mathbb{R}^d, Y = \mathbb{R}^n$, and $y = Ax + \varepsilon$. If we set $\hat{x}_\lambda = \operatorname{argmin}_x \frac{1}{2} \|Ax - y\|^2 + +\lambda x^T S x$ for some symmetric positive semi-definite matrix $S$, then $\hat{x}_\lambda = (A^*A + \lambda S)^{-1} y$, we define $R_\lambda = (A^*A + \lambda S)^{-1}$. If $y$ is a sample from $\mathbb{y} = Ax + \varepsilon$, then $\hat{\mathbb{x}}_\lambda = R_\lambda \mathbb{y}$.

If $\mathbb{E}[\varepsilon] = 0$ and $\operatorname{Var}[\varepsilon] = \Sigma$, then $\mathbb{E}[\hat{\mathbb{x}}_\lambda] = R_\lambda A x$, and $\operatorname{Var} \hat{\mathbb{x}}_\lambda = R_\lambda \Sigma R_\lambda^T$.

## 13.5 Bayesian inverse problems

In Bayesian inverse problems we want to recover the posterior distribution of $(\mathbb{x}|\mathbb{y} = y)$, where $y$ is a single sample of $(\mathbb{y}|\mathbb{x} = x^*)$ and $\mathbb{y} = Ax + \varepsilon$. Here

- $x^*$ is the unknown, true, solution,
- $\varepsilon$ is $Y$-valued random variable, and
- $\mathbb{x}$ is $X$-valued random variable with prior $\pi$.

Even if the problem $y = Ax + \varepsilon$ is ill-posed, lifting the problem to a Bayesian statistical setting, will for many priors, make the problem well-posed.

# 14 Lecture 14 (2025-10-09)

**Today**

- (Finish from last time) Bayesian inverse problems
- Deep learning based methods

We have worked with four types of solutions

- Analytic pseudoinverse
  - Not iterative
  - Tied to specific forward problems
  - Based on having a closed form expression for $A^{-1}$
  - Have something that approximates $R_\alpha \approx A^{-1}$
  - Very fast, but weak in regularisation power
- Iterative methods with early stopping (for solving $\operatorname{argmin} \|Ax - y\|^2$).
  - Early stopping is similar to throwing out high frequency components
  - Not tied to specific problems, can even be used for non-linear problems
- Variational methods
  - Slow, actually want to solve the minimisation problem
  - The general framework can be tailored to the specific problem you are working with
  - Makes regularisation parameter choice rule computationally important
- Statistical methods
  - Richer framework
  - Able to derive confidence intervals etc.
  - Horrendously difficult to compute, and define a good prior

This was the entire landscape until recently. People wanted to combine the last two types, without paying the horrible price. Our usual priors do not give with random samples things that we would actually expect, for example in imaging of humans, a random sample of the prior will not look anything like a human. We instead want to learn the prior from data.

One idea, which was very successful, was to use a Neural denoiser $f$ such that our reconstruction is given by $f \circ A^\dagger$. Intellectually not very satisfying however, today we will speak about what has happened since then.