MEDICAL IMAGING

Lecture Skriptum

Benjamin Bergmann

2026

Contents

Inverse Problems

What is an Inverse Problem?

There exit a "Forward Problem" which estimate the effect from the cause and then there is inverse Problem which estimates the cause from the effect. In the medical context that would be finding the cause illness given from a certain syntom/effect. Typically, the forward problem is "easy" and well described. The challenge here is: We need to solve the inverse problem given only the observed effect of the forward problem.

As an Example from the real world: forward problem: The street becomes wet when it rains. backward problem would be: We observe that the street is wet. Why?

There are multiple different causes:
- Rain
- Fog
- Cleaning

And this can be already problematic as we have multiple different options for what the cause could be.

Example 1 — Computer Tomography .

Forward Problem X-ray emitter and detector rotating around the body. Detectors measure the number of photons passing through the body and hitting the detector

Inverse Problem Reconstruct the interior of the body from the measured detector signals.

Note that a CT Scan can be very large in file size. A scan from shoulder to belt line is already 18GB of data for just a single scan. So we bascially have y and we want to get to x

Example 2 — Deconvolution . **Forward Problem** Observe a blurred image

$$f = k * u$$

on a domain $\Omega \subset \mathbb{R}^2$.

Inverse Problem Estimate the sharp image $u : \Omega \to \mathbb{R}$ given the blur kernel $k : \Omega \times \Omega \to \mathbb{R}_+$ On of the oldest clasical methods to do that is the Wiener Filter. Deconvolution is linked to Fourier F:

$$f = k * u$$

$$F(f) = F(k) \odot F(u)$$

If we wanna do the inverse:

$$F^{-1}\{F(f)\} = F^{-1}\{F(k) \odot F(u)\} = f$$

where $\odot$ is a pointwise multiplication. So a estimate $\hat{u}$ would be

$$\hat{u} = F^{-1}\left(\frac{F(f)}{F(k)}\right)$$

The only problem here is when we have 0 frequencies in the kernel. The Wiener Filtering introduces

$$\hat{u} = F^{-1}\left(\frac{F(f)}{I\sigma^2 F(k)}\right)$$

What is an Inverse Problem? (formal)

Definition 1 (Inverse Problem) .

Given a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $x \in \mathbb{R}^n$ the forward problem is $y = Ax \in \mathbb{R}^m$. The inverse problem is: Given A and y, estimate x.

Vector Space

Definition 2 (Vector Space) . A non-empty set V is a vector space over a field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ if there are operations of vector addition: $+ : V \times V \to V$ and scalar multiplication: $\cdot : \mathbb{F} \times V \to V$ satisfying the following axioms:

Vector addition

1. $u + v \in V \quad \forall u, v \in V$
2. $u + v = v + u$
3. $(u + v) + w = u + (v + w) \quad \forall u, v, w \in V$
4. $\exists 0 \in V : u + 0 = u \quad \forall u \in V$
5. $\forall u \in V : \exists -w : u + (-w) = 0$

Scalar multiplication

1. $av \in V \quad \forall a \in \mathbb{F}, \forall v \in V$
2. $(ab)v = a(bv) \quad \forall a, b \in \mathbb{F}, v \in V$
3. $a(u + v) = au + av \quad \forall a \in \mathbb{F}, \forall u, v \in V$
4. $(a + b)v = av + bv \quad \forall a, b \in \mathbb{F}, \forall v \in V$
5. $\exists 1 \in \mathbb{F} : 1 * u = u \quad \forall u \in V$

Example 3 — Vector Space .

- $\mathbb{R}^n = \left\{(x_1, ..., x_n)^T : x_1, ..., x_n \in \mathbb{R}\right\}$
- $\mathcal{C}(\mathbb{R}^n, \mathbb{R})$ set of function $f : \mathbb{R}^n \to \mathbb{R}$ that are continuous
- $\mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ set of function $f : \mathbb{R}^n \to \mathbb{R}$ that are continuous and once continuously differentiable
- $L^2(\mathbb{R}^n, \mathbb{R}) = \left\{f : \mathbb{R}^n \to \mathbb{R} : \int_{\mathbb{R}^n} |f(x)|^2 \, dx < \infty\right\}$ Lebesgue space
- $H^1(\mathbb{R}^n, \mathbb{R}) = \left\{f \in L^2(\mathbb{R}^n, \mathbb{R}) : \int_{\mathbb{R}^n} |f'(x)|^2 \, dx < \infty\right\}$ Sobolev space $(p = 2)$, Hilbert space

Inverse Problem

Definition 3 (Inverse Problem) . Let X, Y be vector spaces and $A : X \to Y$. The forward problem is defined as $y = Ax$ for any $x \in X$. The inverse problem is to find $x \in X$ such that $Ax = y$ for any $y \in Y$.

So we want to get $A^{-1}(y) = \hat{x}$

Well-Posedness (Hadamard)

We can now start to categorize inverse problems:

Definition 4 (Well-Posedness) . The inverse problem $Ax = y$ is well-posed if:

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

If one condition fails, the problem is ill-posed.

Example 4 — Well-Posedness . Is this example well posed?

Let $X, Y \in \mathbb{R}$ and $A : \mathbb{R} \to \mathbb{R}, x \to x^2$

Answer:

- Existence: for $y = -1$ no solution exists (if we would map to $\mathbb{R}^+$ it would be okay)
- Uniqueness: for $y = 1$, $x = \pm 1$ which is not unique
- Stability: yes, since A is continuous

Example 5 — Well-Posedness . Let $X, Y \in \mathbb{R}^2$ and $A = \begin{pmatrix} 2 & 3 \\ 1 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

Is the inverse problem $Ax = y$ for $y \in Y$ well-posed?

- Existence: $\exists A^{-1}$? Since $\det(A) = 4 - 3 = 1 \neq 0$, the matrix is invertible.
- Uniqueness: Yes, because $\det(A) \neq 0$.
- Stability: Yes, as A^{-1} is continuous.

Inner Product

Definition 5 (Inner Product) . An inner product on a vector space Y over a $\mathbb{F}$ is a map

$$\langle ., . \rangle : Y \times Y \to \mathbb{F}$$

with the following properties:

1. Symmetry: $\langle x, y \rangle = \overline{\langle y, x \rangle} \quad x, y \in Y$
2. Additivity: $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle \quad x, y, z \in Y$

3. Homogeneity: $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle \quad x, y \in Y \quad \lambda \in \mathbb{R}$
4. Positivity: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$

Vector Norm

Definition 6 (Inner Product) . A vector norm is a vector space Y over a field F is a map $\|.\| : Y \to \mathbb{R}$ with:

1. **NON-NEGATIVITY** $\|x\| \geq 0 \quad \forall x \in V, \|x\| = 0 \Leftrightarrow x = 0$
2. **POSITIVE HOMOGENEITY** $\|\lambda x\| = |\lambda| \ \|x\| \ \forall x \in Y, \lambda \in \mathbb{F}$
3. **TRIANGLE INEQUALITY** $\|x + y\| \leq \|x\| + \|y\| \ x, y \in V$

Example 6 — vector norm .

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \quad x \in X \subset \mathbb{R}^n$$

Matrix Norm

Definition 7 (Inner Product) . Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be vector norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Given a matrix $A \in \mathbb{R}^{m \times n}$, the **induced matrix norm** $\|A\|_{a,b}$ is defined as:

$$\|A\|_{a,b} = \max_{x \in \mathbb{R}^n : \|x\|_a \leq 1} \|Ax\|_b = \sup_{\{x \in \mathbb{R}^n \setminus \{0\}\}} \frac{\|Ax\|_b}{\|x\|_a}$$

$$\|Ax\|_b \leq \|A\|_{ab} \|x\|_a$$

Example 7 — Matrix norm .

- If $a, b = 2$: $\|A\|_{2,2} = \|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^T A)}$
- If $a, b = 1$: $\|A\|_{1,1} = \|A\|_1 = \max_j \sum_i |A_{ij}|$
- If $a, b = \infty$: $\|A\|_\infty = \max_i \sum_j |A_{ij}|$

Injection, Surjection, Bijection

Definition 8 (Injection, Surjection, Bijection) . These properties of mappings $A : X \to Y$ are defined as

- **Injection:** $A : X \to Y$ is injective if $Ax_1 = Ax_2 \Rightarrow x_1 = x_2$.
- **Surjection:** $A : X \to Y$ is surjective if $\forall y \in Y, \exists x \in X : Ax = y$.

- **Bijection:** $A : X \to Y$ is bijective if it is both injective and surjective. $\forall y \in Y, \exists! x \in X : Ax = y \Leftrightarrow \exists A^{-1} : x = A^{-1}y$.

Null Space and Range Space

Definition 9 (Null Spance and Range Space) . Let $A : X \to Y$ where X, Y are vector spaces.
- **Nullspace of A:** $N(A) = \{x \in X : Ax = 0\}$
- **Range space of A:** $R(A) = \{Ax \in Y : x \in X\}$

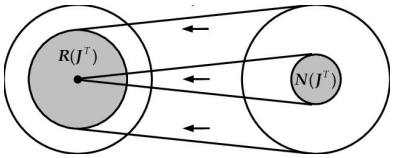

Figure 1 : Null space and range space taken from here

Connection to Hadamard's Definition

- **Existence** $\Leftrightarrow$ Surjection $\Leftrightarrow$ $R(A) = Y$
- **Uniqueness** $\Leftrightarrow$ Injection $\Leftrightarrow$ $N(A) = \{0\}$
- **Existence & Uniqueness** $\Leftrightarrow$ Bijection

Definition of the Linear Inverse Problem

Definition 10 (Linear Inverse Problem) . Given $A : X \to Y$ and observation $y \in Y$ the inverse problem is called linear if A is linear which means that $A(\alpha x_1 + \beta x_2) = \alpha A(x_1) + \beta A(x_2)$

Example 8 — Linear Inverse Problem . A... is the Radon transform

$$(Ax)_i = y_i = \int_{\Gamma_i} x(s)ds$$

$$A(\hat{x}) = A(\lambda_1 \cdot x_1 + \lambda_2 x_2) = \hat{y}_i = \int_{\Gamma_i} \hat{x}(s)ds = \int_{\Gamma_i} \lambda_1 x_1(s) + \lambda_2 \cdot x_2(s)ds$$

$$= \lambda_1 \underbrace{\int_{\Gamma_i} x_1(s)ds}_{y_i^1} + \lambda_2 \underbrace{\int_{\Gamma_i} x_2(s)ds}_{y_i^2} = \lambda_1 y_i^1 + \lambda_2 y_i^2 = \lambda_1 A(x_1)_i + \lambda_2 A(x_2)_i$$

Nullspace of linear A $\Rightarrow \{0\} \in \mathcal{N}(A)$

Decomposition of Square Matrices

Definition 11 (Decomposition of the Square Matrix) . Let $A \in \mathbb{R}^{n\times n}$, recall Eigenvalues λ_i and Eigenvectors v_i:

$$Av_i = \lambda_i v_i \quad \text{for } i = 1, ..., n$$

$$\det(A - \lambda I) = 0$$

If v_i are linearly independent: $Av_i = \lambda_i v_i \Rightarrow AQ = Q\Lambda \Rightarrow A = Q\Lambda Q^{-1}$ Where $Q = (v_1, ..., v_n)$.

Remark. If A is hermitian $\Leftrightarrow A^* = A$, we have that all λ_i are real & v_i are orthonormal.

$$v_i^T v_j = 0 \quad \text{for } i \neq j$$

$$A = Q\Lambda Q^T$$

Singular Value Decomposition

Definition 12 (Singular Value Decomposition) . Let $X \in \mathbb{R}^n, Y \in \mathbb{R}^m$ be an inverse problem $Ax = y$ with a $A \in \mathbb{R}^{m\times n}$. The Goal:

$$A = U\Lambda V^T$$

- $U \in \mathbb{R}^{m\times p}$, $\Lambda \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{p\times n}$
- p is the number of non-zero singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p > 0$.

Link between SVD and Eigendecompostion

$$A \in \mathbb{R}^{m\times n}$$

$$\begin{cases} (1) & Ax = y \\ (2) & A^T\hat{x} = \hat{y} \end{cases} \Leftrightarrow \underbrace{\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}}_{B\in\mathbb{R}^{(m+n)\times(m+n)}} \cdot \begin{pmatrix} \hat{x} \\ x \end{pmatrix} = \begin{pmatrix} y \\ \hat{y} \end{pmatrix}$$

$B = B^T$: $\quad Bw_i = \lambda_i w_i$

$$\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \lambda_i \begin{pmatrix} u_i \\ v_i \end{pmatrix} \Leftrightarrow \begin{cases} \text{1st:} & Av_i = \lambda_i u_i \\ \text{2nd:} & A^T u_i = \lambda_i v_i \end{cases}$$

1st:

$$\lambda_i Av_i = \lambda_i^2 u_i$$

$$A(\lambda_i v_i) = \lambda_i^2 u_i$$

$$AA^T u_i = \lambda_i^2 u_i$$

$$U = (u_1 \mid ... \mid u_m)$$

2nd:

$$A^T(\lambda_i u_i) = \lambda_i^2 v_i$$

$$A^T A v_i = \lambda_i^2 v_i$$

$$V = (v_1 \mid ... \mid v_n)$$

Solving Inverse Problems

We have 3 possible cases:
- Eigenvalue Problem ($p = n = m$)
- Least Squares ($p = m > n$)
- Minimum Length ($p = n > m$)

Eigenvalues ($p = n = m$)

Let $A \in \mathbb{R}^{n \times n}$. Eigendecomposition:

$$Av_i = \lambda_i v_i \quad i = 1, ..., n$$

Assuming A is invertible (assume $\exists A^{-1}$):

$$A^{-1}Av_i = \lambda_i A^{-1} v_i$$

Since $A^{-1}A = I_d$:

$$v_i = \lambda_i A^{-1} v_i$$

Rearranging for A^{-1} (where $\lambda_i \neq 0$):

$$\frac{1}{\lambda_i} v_i = A^{-1} v_i$$

Conclusions:
- Eigenvectors of A and A^{-1} are the same.
- Eigenvalues of $\lambda(A^{-1}) = \frac{1}{\lambda(A)}$.
- This implies $\lambda_i \neq 0$, which is equivalent to $\det(A) \neq 0$.

Least Squares ($m > n$)

We have $Ax = y$ and $A \in \mathbb{R}^{m \times n}$ and $m > n$ which leads us to a overdetermined system

$$e_i = a_i^T x - y_i$$

Idea: minimize the squared error

$$\hat{x} = \arg\min E(x) := \frac{1}{2} \sum_{i=1}^{m} \left(a_i^T x - y_i\right)^2 = \frac{1}{2} \left\|Ax - y\right\|_2^2 = \frac{1}{2} \left\|e\right\|_2^2$$

Here we define $e = Ax - y$. How do we solve this optimization problem?

$\nabla E(x) = 0$ where $\nabla E(x) \in \mathbb{R}^n$

$$\frac{\partial e}{\partial x} \frac{\partial E}{\partial e} = \frac{\partial e}{\partial x} \frac{1}{2} 2e = A^T e = A^T (Ax - y) = 0$$

$$(A^T A)x = A^T y$$

$$x = \left(A^T A\right)^{-1} A^T y$$

Example 9 — 2x2 CT Reconstruction .

$$x \in \mathbb{R}^4 \quad y \in \mathbb{R}^5$$

x_1	x_2
x_3	x_4

Table 1 : Grid representation of variables x_i

We send rays through the matrix in 3 directions (top to bottowm, diagonal and left to right):

$y_1 = x_1 + x_3$ (Column 1 sum)

$y_2 = x_2 + x_4$ (Column 2 sum)

$y_3 = x_1 + x_2$ (Row 1 sum)

$y_4 = x_3 + x_4$ (Row 2 sum)

$y_5 = x_1 + x_4$ (Diagonal sum)

We can bring this now in the form $Ax = y$ which leads us to equation that looks like this:

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

Minimum Length $(p = n > m)$

Let $Ax = y$ with $A \in \mathbb{R}^{m \times n}$. Since $n > m$, $(A^T A)^{-1}$ does not exist. This is an **underdetermined system**. If we had multiple solutions we would exactly solve $Ax = y$. We pick one using a priori knowledge:

$$\min_x \frac{1}{2} \|x\|_2^2 \quad \text{s.t.} \quad Ax = y$$

Recap: Lagrange Multipliers We want to solve $\min_x E(x)$ subject to $C(x) = 0$. For that we define Lagrangian:

$$\mathcal{L}(x, \tau) = E(x) + \langle C(x), \tau \rangle$$

where $\tau \in \mathbb{R}^m$ is the Lagrange multiplier. Then we can find a solution by setting $\nabla \mathcal{L}(x, \tau) = 0$

$$\frac{\partial}{\partial x} \mathcal{L} = \frac{\partial E}{\partial x} + \frac{\partial C}{\partial x} \tau = 0$$

$$\frac{\partial}{\partial \tau} \mathcal{L} = C(x) = 0$$

From the initial setting where we want to find the minimum of x we can define now E and C:

$$E(x) = \frac{1}{2} \|x\|_2^2$$

$$C(x) = y - Ax = 0 \quad \Leftrightarrow \quad h(x, \tau) = \frac{1}{2} \|x\|_2^2 + \langle y - Ax, \tau \rangle$$

Then we can write out the Lagrane Terms and solve them for x:

$$\frac{\partial}{\partial x}\mathcal{L} = x - A^T\tau = 0$$

$$x = A^T\tau$$

$$\frac{\partial}{\partial \tau}\mathcal{L} = y - Ax = 0$$

$$y = Ax = A(A^T\tau) = (AA^T)\tau$$

$$\tau = (AA^T)^{-1}y$$

$$x = A^T(AA^T)^{-1}y$$

Generalized Inverse

Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and the inverse problem $Ax = y$ with $A \in \mathbb{R}^{m\times n}$.

We define the generalized inverse as:

$$A_g^{-1} = \left(U_p\Lambda_p V_p^T\right)^{-1} = \left(V_p^T\right)^{-1}\Lambda_p^{-1}U_p^{-1} = V_p\Lambda_p^{-1}U_p^T$$

Check if Generalized Inverse computes the exact Least Squares and Minimum Length solutions:

I. $p = m = n$***:***

$$A_g^{-1} = V_p\Lambda_p^{-1}U_p^T \quad |\cdot A = U_p\Lambda_p V_p^T$$

$$A_g^{-1}A = V_p\Lambda_p^{-1}\underbrace{U_p^TU_p}_{I}\Lambda_p V_p^T = V_p\underbrace{\Lambda_p^{-1}\Lambda_p}_{I}V_p^T = I$$

II. $p = m > n$***:***

$$\begin{aligned}
x &= (A^TA)^{-1}A^Ty \\
&= \left(\left(U_p\Lambda_pV_p^T\right)^T\left(U_p\Lambda_pV_p^T\right)\right)^{-1}\left(U_p\Lambda_pV_p^T\right)^Ty \\
&= \left(V_p\Lambda_p\underbrace{U_p^TU_p}_{\text{Id}}\Lambda_pV_p^T\right)^{-1}\left(V_p\Lambda_pU_p^T\right)y \\
&= \left(V_p\Lambda_p^2V_p^T\right)^{-1}V_p\Lambda_pU_p^Ty \\
&= V_p\Lambda_p^{-2}\underbrace{V_p^TV_p}_{\text{Id}}\Lambda_pU_p^Ty \\
&= V_p\Lambda_p^{-1}U_p^Ty = A_g^{-1}y
\end{aligned}$$

III. $p = m < n$***:***

$$\begin{aligned} x &= A^T(AA^T)^{-1}y \\ &= (V_p\Lambda_p U_p^T)(U_p\Lambda_p V_p^T V_p\Lambda_p U_p^T)^{-1}y \\ &= (V_p\Lambda_p U_p^T)(U_p\Lambda_p^2 U_p^T)^{-1}y \\ &= V_p\Lambda_p \underbrace{U_p^T U_p}\Lambda_p^{-2}U_p^T y \\ &= V_p\Lambda_p^{-1}U_p^T y = A_g^{-1}y \end{aligned}$$

IV. $0 < p < \min(m,n)$***:***

However, A_g^{-1} still exists. It computes a solution that interpolates between Least Squares & Minimum Length solutions.

Regularization

Consider polynomial regression

$$p(a) = \sum_{i=1}^{n} x_i a^{i-1} = x_1 \cdot 1 + x_2 \cdot a + ... + x_n a^{n-1}$$

Where x represents the coefficients of the polynomial.

$$p(a) \Leftrightarrow Ax = \begin{pmatrix} 1 & a_1 & a_1^2 & ... & a_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_m & a_m^2 & ... & a_m^{n-1} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

How to choose n?
- manually
- very large + regularization

We can incorporate Prior Knowledge: Least squares problem + regularization:

$$\hat{x} = \arg\min_x \frac{1}{2}\left\|Ax - y\right\|_2^2 + R(x)$$

Example: $R(x) = \frac{\lambda}{2}\left\|x\right\|_2^2$ (Tikhonov regularization, aka weight decay)

$$\frac{1}{2}\left\|Ax - y\right\|^2 + \frac{\lambda}{2}\left\|x\right\|_2^2 \to \min$$

We derive it an set it to 0:

$$\frac{1}{2} \cdot 2 \cdot A^T(Ax - y) + \frac{\lambda}{2} \cdot 2x = 0$$

$$A^T Ax + \lambda x = A^T y$$

$$x = (A^T A + \lambda I)^{-1} A^T y$$

Regularization Types

Name	$R(x)$	Intuition
Tikhonov	$\lambda \|Gx\|_2^2$ $(\|Gx - \hat{x}\|_2^2)$	Existence of Inverse
L^2	$\lambda \|x\|_2^2$ $(G = I)$	Minimum length/norm
H^1	$\lambda \|\nabla x\|_2^2$ $(G = \nabla)$	Smooth gradients
L^1	$\lambda \|x\|_1$	Sparse solutions
Total variation (TV)	$\lambda \|\nabla x\|_1$	Sparse gradients (piece-wise constant solutions)

The Proximal Mapping

1. Projection onto a set S

$$\text{proj}_S(x) = \arg\min_{y \in S} \frac{1}{2} \|x - y\|_2^2$$

2. Proximal mapping of a function $g(x)$

$$\text{prox}_g(x) = \arg\min_y \frac{1}{2} \|x - y\|_2^2 + g(y)$$

where we can for example put in the indicator function of the set S

$$g(y) = \begin{cases} 0 & \text{if } y \in S \\ \infty & \text{else} \end{cases}$$

Example 10 — Soft Thresholding . Let's set: $g(x) = |x|$

To find the proximal mapping for the absolute value (L1 norm), we solve:

$$\text{prox}_{|\cdot|}(x) = \arg\min_y \frac{1}{2} |x - y|^2 + |y|$$

The derivative of $|y|$ is:

$$\frac{d}{dy} |y| = \begin{cases} 1 & y > 0 \\ [-1, 1] & y = 0 \\ -1 & y < 0 \end{cases}$$

We simplify and solve

$$\frac{d}{dy}\left(\frac{1}{2} |x - y|^2\right) + \partial g(y) = 0$$

- **Case $y > 0$:** $-(x - y) + 1 = 0 \quad \Rightarrow \quad y = x - 1 > 0 \quad \Rightarrow \quad x > 1$
- **Case $y < 0$:** $-(x - y) - 1 = 0 \quad \Rightarrow \quad y = x + 1 < 0 \quad \Rightarrow \quad x < -1$

- **Case $y = 0$:** $-(x-0) + [-1,1] = 0 \quad \Rightarrow \quad x \in [-1,1]$

Thus, the Soft Thresholding operator is:

$$\text{prox}_{|\cdot|}(x) = \begin{cases} x-1 & \text{if } x > 1 \\ x+1 & \text{if } x < -1 \\ 0 & \text{else} \end{cases}$$

This would now be for example the sparse solution that we have seen in the table above. The proximal map kills all of the small graidents and shirks the rest of the gradients. It is essentially a way of first performing graident descent and then you look with the proximal map where to go and at this point you look then how suitable it is or how much penalty the regularizer gives you. In our case the regularizer shrinks the gradient and sets all the values in ± 1 to 0.

A Probabilistic Perspective of Regularization

Assume observed measurements $y_i \in \mathbb{R}^m$ follow a Gaussian distribution:

$$y \sim \mathcal{N}(Ax, \Sigma) \iff p(y|x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|Ax - y\|^2_{\Sigma^{-1}}\right)$$

Moreover, we know the gradients of the solution follows a Gaussian prior:

$$\nabla x \sim \mathcal{N}(0, \eta I) \iff p(x) = |2\pi\eta I|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\eta \|x\|^2\right)$$

Using Bayes' Rule to find the posterior distribution:

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}$$

$$\log p(x|y) = \log p(y|x) + \log p(x) - \log p(y)$$

$$\log p(x|y) = -\frac{1}{2} \|Ax - y\|^2_{\Sigma^{-1}} - \log Z_1 - \frac{1}{2\eta} \|x\|^2 - \log Z_2 - \log p(y)$$

Since Z_1, Z_2, and $p(y)$ are constants that do not depend on x:

$$\max_x \log p(x|y) = \max_x -\frac{1}{2} \|Ax - y\|^2_{\Sigma^{-1}} - \frac{1}{2\eta} \|x\|^2$$

$$\min_x -\log p(x|y) = \min_x \underbrace{\frac{1}{2} \|Ax - y\|^2_{\Sigma^{-1}}}_{\substack{D(x,y) \text{ (Data Fidelity)} \\ \alpha \propto \log(p(y|x))}} + \underbrace{\frac{1}{2\eta} \|x\|^2}_{\substack{R(x) \text{ (Regularizer)} \\ \alpha \propto \log(p(x))}}$$

Conclusion: The variational formulation of inverse problems corresponds to the Maximum A Posteriori (MAP) estimation.

X-rays and Computed Tomography

Discovery of X-rays

In 1895, Wilhelm Röntgen noticed "rays of mysterious origin", which he called X-rays. Within a month (22.12.1895), the first radiograph of the hand of Röntgen's wife was made in Würzburg. This immediate application to imagine the human body marks the birth of medical imaging.

Nature and Properties of X-rays

X-rays are electromagnetic waves. They are a form of ionizing radiation—radiation with enough energy to eject electrons from an atom.

What needs to hold: Bound energy < Unbound energy + Electron Energy.

The binding energy is 13.6 eV which is the binding energy of hydrogen. For a Medical CT you need around 100keV, for Mammography you need around 20keV.

Forms of Ionizing Radiation

1. **Particulate Radiation**: Any subatomic particle (proton, neutron, electron) with enough kinetic energy to be ionizing.
2. **Electromagnetic Radiation**: Can act as a wave or a particle (photon). If energy > 13.6 eV (binding energy of hydrogen electron), it is considered ionizing.

Remark.

$$E = h\nu$$

and

$$\lambda = \frac{c}{\nu}$$

Where:
- h: Planck's constant
- ν: frequency
- λ: wavelength
- c: speed of light

Interaction of Energetic Electrons with Matter

- **Collision transfer** (99% $\rightarrow$ heat): Collision with other electrons until kinetic energy is exhausted. If they bump into each other, then energy can be transfered to the other electorn which then will emit infrared photons, which is heat.
- **Radiative transfer** (1% $\rightarrow$ X-ray):
  - Eject inner shell electron, generating **characteristic X-ray radiation**.
  - Electron flies close to the atom nucleus and is braked by nucleus, generating **Bremsstrahlung X-ray**.

Interaction of Electromagnetic Radiation with Matter

- **Photoelectrical Effect**: Photon ejects an inner shell electron. The energy is $h\nu - E_B$. Filling the hole yields characteristic X-rays or Auger electrons.
- **Compton Scattering**: Photon interacts with outer-shell electrons, yielding a Compton electron and a scattered photon with less energy.

Remark. **Handwritten formula for Compton energy:**

$$E_c = h\nu - h\nu' = h(\nu - \nu')$$

Attenuation of Electromagnetic Radiation

Consider a narrow beam geometry with an X-ray source and a detector.

Definition 13 (Beer-Lambert Law Derivation) . Let N be the number of photons leaving the source and N' be the photons hitting the detector. Suppose n photons are lost in a thickness Δx:

$$n = N\mu\Delta x$$

The change in photons is:

$$\Delta N = N' - N = -n = -\mu N\Delta x$$

In the limit $\Delta x \rightarrow 0$:

$$dN = -\mu N dx \rightarrow \frac{dN}{N} = -\mu dx$$

Integrating both sides:

$$\int \frac{dN}{N} = -\int \mu dx \rightarrow \log(N) = -\int \mu dx + C$$

For $x = 0$, $N(0) = N_0$, thus $C = \log(N_0)$.

$$N(x) = N_0 \exp\left(-\int \mu dx\right)$$

Intensity I is proportional to photon count, so $I = I_0 \exp(-\int \mu(s)ds)$.

Narrow Beam vs. Broad Beam

- **Broad beam**: Scattering (Compton effect) causes photons to hit the detector from multiple angles, and the monoenergetic assumption often fails.
- **Rescue**: Use detector collimation to ensure only primary (non-scattered) rays are measured.

Projection Radiographic System

Definition 14 (Basic Imaging Equation) .

$$I(x) = \int S_0(E) \exp\left(-\int \mu(x,E)ds\right) dE$$

Simplification: Assuming monoenergetic X-rays with effective energy E:

$$y = -\log\left(\frac{I}{I_0}\right) = \int \mu(s)ds$$

Blurring and Noise

- **Blurring sources**: Penumbra (due to focal spot size), Compton scattering, and detector resolution.
- **Noise**: Photon counting follows a Poisson distribution $N \sim \text{Pois}(|(N))$, so the variance is $\sigma^2 = |(N)$.
- **Signal-to-Noise Ratio (SNR)**: To increase SNR, one can increase the photon count or use contrast agents.

Computed Tomography (CT)

Tomography (from Greek **tomos** "slice" and **grapho** "to write") involves imaging by sectioning a volume using projected radiographs from different orientations.

Radon Transform

For a 2D object $f(x,y)$, the projection $g(\theta,\rho)$ at angle θ and distance ρ is given by the line integral:

$$g(\theta,\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - \rho)dxdy$$

A collection of these projections is called a **sinogram**.

Reconstruction Methods

1. **Backprojection**: Project the measured values back onto the image plane.

$$b(x,y) = \int_0^{\pi} g(\theta, x\cos\theta + y\sin\theta)d\theta$$

Problem: Results in a blurry image (1/r blurring).

2. **Filtered Backprojection (FBP)**: Apply a high-pass filter (Ramp filter $|q|$) to the projections in the frequency domain before backprojecting.

> **Theorem 1** (Fourier-Slice Theorem) . The 1D Fourier Transform of a projection at angle θ is equal to a slice of the 2D Fourier Transform of the original image at that same angle.

Artifacts and Hounsfield Units

- **Aliasing**: Streak artifacts due to insufficient number of projections.
- **Beam Hardening**: Caused by energy-selective attenuation; low-energy photons are absorbed more easily, shifting the spectrum toward "harder" (higher energy) X-rays.

> **Definition 15** (Hounsfield Units (HU)) . Standardized scale to compare CT scans:
>
> $$h = 1000 \cdot \frac{\mu - \mu_{\text{Water}}}{\mu_{\text{Water}} - \mu_{\text{Air}}}$$

| Substance | HU | | :— | :— | | Air | -1000 | | Fat | -120 to -90 | | Water | 0 | | Muscle | $+35$ to $+55$ | | Bone | $+300$ to $+1900$ |

> *Remark.* **Historical Note**: The development of CT was funded in part by EMI (the Beatles' record label), leading to Hounsfield's Nobel Prize.

Learned Reconstruction Methods

Recall: Inverse Problems

Let $X = \mathbb{R}^n$ be the image space and $Y = \mathbb{R}^m$ be the measurement space. The inverse problem is defined as:

$$Ax = y$$

where $A \in \mathbb{R}^{m \times n}$ is the forward operator.

Instances in Medical Imaging:
- Computed Tomography (CT):
  - y is the sinogram data.
  - A is the Radon transform.
- Reconstruction variants:
  - **Full-view CT**: Dense sampling of projections.
  - **Sparse-view CT**: Reduced number of projections (ill-posed problem).

Remark. **Handwritten Flowchart:** CT Acquisition $\rightarrow$ X-ray Projection Data (y) $\rightarrow$ Filtered Backprojection (A^{-1}) $\rightarrow$ Reconstructed Image (x).

Deep Learning Approaches

There are three main paradigms for integrating Deep Learning into the reconstruction pipeline:

1. Post-processing: Applying a Neural Network (NN) to an initial reconstruction (e.g., FBP) to remove artifacts.
$$y \rightarrow \text{FBP} \rightarrow x_{\text{initial}} \rightarrow \mathbb{N} \rightarrow x_{\text{final}}$$
2. Pre-processing: Applying a NN to the raw data (sinogram/k-space) before reconstruction.
$$y \rightarrow \mathbb{N} \rightarrow y_{\text{full}} \rightarrow \text{FBP} \rightarrow x$$
3. Learned Inverse / Model-based Reconstruction: Replacing or augmenting the reconstruction operator itself.

Post-processing Approach: FBPConvNet

The FBPConvNet uses a U-Net architecture to refine sparse-view FBP reconstructions.
- Architecture: U-Net with skip connections and concatenation.
- Spatial Dimension: 512×512.
- Operations: 3×3 convolutions, Batch Normalization (BN), ReLU, and 2×2 max pooling.

Example 11 — Performance Comparison . Results for sparse-view CT reconstruction:
- FBP: SNR 24.06
- Total Variation (TV): SNR 29.64
- FBPConvNet: SNR 35.38

Reference: Jin et al. (2017), "Deep convolutional neural network for inverse problems in imaging".

Pre-processing Approach: RAKI

RAKI (Scan-specific Robust Artificial-neural-networks for K-space Interpolation) is a database-free method for fast MRI imaging.
- It learns to interpolate missing k-space data from the auto-calibration signal (ACS) of the specific scan.
- Outperforms classical GRAPPA, especially at high acceleration rates (Rate 4 to 6).

Reference: Akçakaya et al. (2019), "Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction".

Model-based Reconstruction

In model-based approaches, we estimate the solution via a reconstruction operator $B(y)$ that approximates the inverse A^{-1}.

Definition 16 (Variational Formulation) .

$$B(y) = \arg\min_x \frac{1}{2}\|Ax - y\|_2^2 + R(x)$$

Where:
- $\frac{1}{2}\|Ax - y\|_2^2$ is the data fidelity term.
- $R(x)$ is the regularization term (prior knowledge).

Learned Inversion: AUTOMAP

AUTOMAP learns the entire mapping from sensor domain to image domain using a deep network.
Reference: Zhu et al. (2018), "Image reconstruction by domain-transform manifold learning," Nature.

Learned Model-based Reconstruction

Modern methods focus on learning the regularization functional $R(x)$ or the optimization steps.

Key Learning Principles:
1. Bilevel Optimization: Learning parameters by solving an optimization problem within another.
2. Contrastive Learning: Learning representations by comparing positive and negative pairs.
3. Distribution Matching: Ensuring the reconstructed distribution matches the ground truth distribution.

4. Plug & Play (PnP): Using a pre-trained deep denoiser as a proximal operator in iterative algorithms.

> *Remark.* The evaluation of learned regularization is a critical area of current research (e.g., Hertrich et al., 2025).

Plug & Play Optimization

PnP replaces the traditional proximal operator with a deep denoiser D_σ:

$$x^{k+1} = D_\sigma\big(x^k - \eta \nabla f(x^k)\big)$$

This allows the use of state-of-the-art denoisers without explicitly defining $R(x)$.

References: Venkatakrishnan et al. (2013); Zhang et al. (2021).

Magnetic Resonance Imaging

From Spin to Magnetic Resonance Imaging

The study of MRI often begins from a classical physics viewpoint, where we accept the existence of nuclear spin without diving into the full quantum mechanics motivation. Reference: Prince, J. L., & Links, J. M. (2014). **Medical imaging signals and systems**. Pearson.

Nuclear Spin, Magnetic Dipole Moment, and Torque

A rotating object with mass m leads to angular momentum:

$$\vec{L} = \vec{r} \times (m\vec{v})$$

> *Remark.* **Handwritten Note**: The spin of a proton $(1H)$ leads to magnetic angular momentum $\vec{I}$. It is modeled as a magnetic dipole with a moment $\vec{\mu} = \gamma\vec{I}$, where γ is the gyromagnetic ratio.

Gyromagnetic Ratios

| Element | Gyromagnetic ratio $\frac{\gamma}{2\pi}$ (MHz/T) | | :— | :— | | $1H$ | 42.58 | | ^{3}He | 32.43 | | ^{23}Na | 11.26 | | $31P$ | 17.24 |

Interaction with external Magnetic Field B_0

Exposure to an external magnetic field $\vec{B}_0$ leads to a torque $\vec{\tau}$ that attempts to align the magnetic moment $\vec{\mu}$:

$$\vec{\tau} = \vec{\mu} \times \vec{B}_0$$

- Thermal Motion: In the absence of a field, random orientation means no net magnetization (humans are not inherently magnetic).
- Magnetization: In the presence of B_0, thermal motion is still present, but the magnetic moments align enough to create a small bulk magnetization $\overrightarrow{M} = \sum_i \vec{\mu}_i$.

The Two Effects of B_0

1. Magnetization: The magnitude M is given by:

$$M = \frac{\rho\gamma^2\hbar B_0}{4kT}$$

2. Precession: The magnetic momentum precesses around the external field, similar to a top in a gravitational field.

Definition 17 (Larmor Frequency) . The frequency of precession is the Larmor frequency:

$$\omega_0 = \gamma B_0$$

For a proton ($1H$), $\frac{\gamma}{2\pi} \approx 42.6$ MHz/T. This is a key equation for MR imaging.

Interaction with Radiofrequency field B_1

When an RF field B_1 is applied at the Larmor frequency, it tips the magnetization away from the longitudinal axis.

- Bloch Equation (simplified): $\frac{dM}{dt} = \gamma(M \times B_1)$.
- Flip Angle: $\alpha = \gamma B_1 t$.

The resulting magnetization has two components:

- Longitudinal component: Parallel to B_0.
- Transversal component: Perpendicular to B_0, which induces a current in the receiver coil (signal reception).

Relaxation and Contrast

Definition 18 (Longitudinal Relaxation (T1)) . Recovery of the M_z component after an RF pulse:

$$M_z = M_0\left(1 - e^{-\frac{t}{T_1}}\right)$$

Definition 19 (Transversal Relaxation (T2)) . Decay of the M_{xy} component:

$$M_{xy} = M_0 e^{-\frac{t}{T_2}}$$

Contrast Information

By tailoring the Repetition Time (TR) and Echo Time (TE), we can choose the most suitable contrast to differentiate structures:

- T1-weighted: Short TR, short TE.
- T2-weighted: Long TR, long TE ($M_{xy} = M_0 e^{-\frac{t}{T_2}}$).
- Proton Density (PD) weighted: Long TR, short TE ($M_z = M_0\left(1 - e^{-\frac{t}{T_1}}\right)$).

Image Encoding (Gradients)

To get an image, spatial information must be encoded using gradient fields $\vec{G}$. The local Larmor frequency becomes position-dependent:

- X-gradient: $\omega(x) = \omega_0 + \gamma G_x x$
- Y-gradient: $\omega(y) = \omega_0 + \gamma G_y y$
- Z-gradient: $\omega(z) = \omega_0 + \gamma G_z z$

This allows for slice selection (z-axis) and frequency/phase encoding (x and y axes) to fill the k-space, which is then transformed into an image via a 2D Fourier Transform.

Image Registration

What is Image Registration?

Image registration is the process of transforming different sets of data into one coordinate system.

Definition 20 (Fundamental Components) .
- **Fixed image** $f(x)$: The reference image that remains stationary.
- **Moving image** $m(x)$: The image that is deformed to match the fixed image.
- **Transformation** T: A mapping $T : x \to T(x)$ that defines how the moving image is warped.
- **Warped image**: The result of applying the transformation to the moving image, denoted as $(m \circ T)(x) = m(T(x))$.

Variational Approach to Registration

Registration is typically formulated as an optimization problem where we seek the optimal transformation parameters θ:

Theorem 2 (Variational Formulation) .

$$\min_{\theta} S(f, m \circ T_\theta) + R(T_\theta)$$

Where:
- $S(f, m \circ T_\theta)$ is the Similarity Metric (measures how well the images match).
- $R(T_\theta)$ is the Regularization term (ensures the transformation is physically plausible or smooth).

Transformation Models

1. Global Linear Transformation Models:
   - **Rigid**: Rotation and translation (6 degrees of freedom in 3D).
   - **Affine**: Includes scaling and shearing.
2. Non-linear Transformation Models:
   - Allows for local deformations (e.g., organ movement, breathing).
   - Often parameterized by B-Splines or displacement fields.

Remark. **Handwritten Note on Interpolation**: When warping an image, we often need to calculate values at non-integer coordinates. B-Splines (1D and 3D cases) are commonly used for smooth interpolation.

Similarity Metrics

The choice of similarity metric depends on whether the images are from the same modality (intra-modal) or different modalities (inter-modal).

- Sum of Squared Differences (SSD): Best for intra-modal images with linear intensity relationships.

$$S_{\text{SSD}} = \int (f(x) - m(T(x)))^2 dx$$

- Normalized Cross Correlation (NCC): Robust to linear intensity changes.
- Normalized Gradient Field (NGF): Matches the edges/gradients of the images.
- Mutual Information (MI): The standard for multi-modal registration (e.g., MR to CT). It measures the statistical dependence between image intensities.

Regularization

Regularization prevents "unrealistic" warping, such as folding the image onto itself.

- Implicit regularization: Built into the model architecture or transformation model (e.g., low-resolution B-spline grid).
- Explicit regularization: A penalty term added to the loss function (e.g., Diffusion, Elastic, or Total Variation regularizers).

Optimization and Deep Learning

Optimization Tricks

- Coarse-to-fine strategy: Start by registering downsampled (low-res) versions of the images and gradually increase resolution to avoid local minima.
- Sequential complexity: Start with rigid/affine transforms before moving to non-linear deformations.

Deep Learning Approaches

Deep learning has shifted registration from iterative optimization to "one-shot" prediction.

- VoxelMorph: A CNN-based framework that learns to predict the displacement field between two images in a single forward pass.
- Implicit Neural Representations (INR): Representing the transformation as a continuous function $T(x)$ parameterized by a neural network (e.g., using periodic activation functions like SIREN).

Example 12 — Key References .

- Balakrishnan et al. (2019), **VoxelMorph: a learning framework for deformable medical image registration**.
- Wolterink et al. (2022), **Implicit neural representations for deformable image registration**.

Image Segmentation

What is Image Segmentation?

Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects).

- Goal: Assign a semantic label to every pixel (2D) or voxel (3D) in an image.
- Output: A dense label map.

Example 13 — Medical Applications .
- Tumor delineation: Identifying the boundaries of a tumor in MRI or CT scans.
- Organ segmentation: Separating liver, lungs, or heart from surrounding tissue.
- Lesion burden estimation: Quantifying the total area or volume affected by a disease.

Segmentation vs. Other Tasks

It is important to distinguish segmentation from other computer vision tasks:
- Image Classification: Assigning a single label to the entire image (e.g., "glioma" vs "no tumor").
- Object Detection: Identifying objects and drawing bounding boxes around them.
- Segmentation: Identifying the exact shape of the object at the pixel level.

Clinical Significance

- Quantitative Analysis: Allows for precise measurement of volume, shape, and thickness.
- Treatment Planning: Essential for radiotherapy (delineating the target volume) and surgery.
- Longitudinal Follow-up: Comparing scans over time to check for progression or stability using criteria like RANO.
- Sensitivity: In medicine, small boundary errors can have a massive clinical impact.

Mathematical Formulation

Definition 21 (Segmentation as a Labeling Function) . Let $\Omega \subset \mathbb{R}^d$ be the image domain ($d = 2, 3$). An image x is a function $x : \Omega \to \mathcal{C}$ (color space).

The segmentation problem is defined as finding a mapping:

$$S : \Omega \to \mathcal{L}$$

Where $\mathcal{L} = \{0, 1, ..., K\}$ is the set of labels (0 is usually the background).

Remark. **Handwritten Note on Partitioning:** The labeling induces a partitioning of the domain into $\Omega_k = \{i \in \Omega \texttt{ mid } S(i) = k\}$. These partitions must be:
- Non-overlapping: $\Omega_k \cap \Omega_l = \emptyset$ for $k \neq l$.
- Complete: $\bigcup_{\{l \in \mathcal{L}\}} \Omega_l = \Omega$.

Types of Segmentation

- Binary: Separating a single foreground structure from the background ($K = 1$).
- Multi-class: Segmenting multiple anatomical structures ($K > 1$).
- Semantic: All pixels of the same class (e.g., all cells) share one label.
- Instance: Separating individual objects (e.g., each individual cell gets a unique ID).
- Panoptic: Combines semantic and instance segmentation.

Classical Segmentation Methods

1. Thresholding:
   - **Global**: One value for the entire image (e.g., Otsu's method).
   - **Local**: Adaptive thresholds based on local neighborhoods.
2. Region-based:
   - **Region Growing**: Starts with seed points and expands to similar neighbors.
   - **Watershed**: Interprets the image as a topographic map and "floods" it from local minima.
3. Graph Cuts:
   - Represents the image as a graph where pixels are nodes.
   - Minimizes an energy function $E(x)$ consisting of unary (likelihood) and pairwise (smoothness) costs.
   - Solved using min-cut/max-flow algorithms.

Remark. **Handwritten relation**: The pairwise term in Graph Cuts is related to anisotropic Total Variation (TV).

Deep Learning for Segmentation

U-Net

The U-Net is the gold standard for medical image segmentation.
- Architecture: Symmetric encoder (contracting path) and decoder (expansive path).
- Skip Connections: Concatenate high-resolution features from the encoder to the decoder to preserve spatial detail.
- Training Objective: Usually Weighted Cross Entropy (CE).

V-Net

Designed for volumetric (3D) medical images.
- Objective: Uses the Dice Loss to handle class imbalance (e.g., when the tumor is much smaller than the background).

Theorem 3 (Dice Loss (Binary)) .

$$D(\theta) = 1 - \frac{2\sum_{\{i=1\}}^{I} \widehat{\ }\{p\}_i s_i}{\sum_{\{i=1\}}^{I} \widehat{\ }\{p\}_i + \sum_{\{i=1\}}^{I} s_i}$$

Advanced Architectures

- nnU-Net: A "self-configuring" method that automatically adapts the U-Net architecture and hyperparameters to a specific dataset.
- UNETR: Uses Transformers as the encoder to capture long-range dependencies, paired with a U-shaped decoder.
- Segment Anything Model (SAM): A promptable foundation model for segmentation, recently adapted for medical images (SAM-Med).

Segmentation Loss Odyssey

A combination of different loss terms is often used in practice.

- Distribution-based:
  - **Weighted Cross Entropy**: Penalizes errors in rare classes more heavily.
  - **Focal Loss**: Focuses on hard-to-classify pixels by down-weighting easy ones.
- Region-based:
  - **Dice Loss** and **IoU (Jaccard)**: Measure the overlap between prediction and ground truth.
  - **Tversky Loss**: Generalization of Dice that allows controlling the trade-off between False Positives and False Negatives.
- Boundary-based:
  - **Hausdorff Distance (HD)**: Penalizes the distance between the boundaries of the predicted and ground truth masks.

Evaluation

Validation should follow the Metrics Reloaded recommendations to ensure results are clinically meaningful and statistically sound.

Federated Learning

Data Protection in Healthcare

Data protection is critical in healthcare due to the high sensitivity of patient records (medical history, genetics, diagnoses). It is regulated by laws such as:

- **GDPR (EU)**: General Data Protection Regulation, which regulates how personal data of EU residents is collected, stored, and processed.
- **HIPAA (USA)**: Health Insurance Portability and Accountability Act, which sets national standards for protecting sensitive patient health information.

Breaches of these regulations can lead to identity theft, loss of trust, and severe legal penalties.

Personal Data and Re-identification

Definition 22 (Personal Data (GDPR)) . Personal data is any information relating to an identified or identifiable living individual. Data that has been de-identified or pseudonymized but can still be used to re-identify a person remains personal data.

Anonymization: To be truly anonymized, the process must be irreversible.

Remark. **Re-identification Risk**: A famous study by Sweeney (2000) showed that 87% of US citizens can be uniquely identified using only their ZIP code, birth date, and sex.

From Centralized to Federated Learning

- Centralized ML: Training data from all sources is moved to a central server.
- Distributed On-Site Learning: Models are trained locally at each site with no information exchange.
- Federated Learning (FL): A collaborative learning approach where data remains at the source, and only model updates (weights) are shared with a central server.

Comparison: Centralized vs. Federated Learning

| Feature | Centralized | Federated | | :— | :— | :— | | Data location | Cloud / Central server | Distributed nodes (edge) | | Training | Primarily in the cloud | Primarily at the edge | | Communication | Nodes share local data | Nodes share model weights | | Privacy | Low user data privacy | High user data privacy | | Heterogeneity | Cannot handle easily | Can operate on heterogeneous data |

Centralized FL - Mathematical Formulation

Let $D = (x_i, y_i)_{i=1}^n$ be a dataset distributed to K clients C_k where $k \in \{1, ..., k\}$. We denote by $P = \{1, ...n\}$ and each client has a subset P_k such that $P = \bigcup_{k=1}^n P_k$. The goal is to solve:

$$\min_w f(w) = \min \frac{1}{n} \sum_{i=1}^{N} f_{i(w)}$$

where $f_{i(w)} = l(x_i, y_i, w)$ is a loss function. Then we have that the total loss function

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_{i(w)} = \sum_{i=1}^{n} \frac{1}{n} f_{i(w)} = \sum_{k=1}^{K} \frac{1}{n} n_f F_{k(w)}$$

with $F_{k(w)} = \frac{1}{n_k} \sum_{i \in P_k} f_{i(w)}$ which is the loss at the distributed clients. So the loss function of the sample is the same, but now we combined the indices into the clients and then we write it by n_k. It is still the same thing, we just shifted the indices. At the client we do the same thing as globally.

Remark. **Iterative Learning Concept**:
1. Central server chooses a model and transmits it to nodes.
2. Nodes train the model locally with their own data.
3. Nodes upload local updates to the server.
4. Server pools results and generates a new global model.

Algorithms: FedSGD and FedAVG

- FedSGD: A simple version where each client performs one step of gradient descent per round.
- FedAVG: Substantially reduces communication by allowing clients to perform multiple local epochs before aggregating.

Theorem 4 (FedAVG Update Rule) . The server aggregates weights from a subset of sampled clients S_t:

$$w_{t+1} \to \sum_{k \in S_t} \frac{n_k}{n} w_{t+1}^k$$

Non-IID Data Challenges

In FL, data is typically not independent and identically distributed (Non-IID).
1. Feature distribution skew: Different demographics or devices ($P_{k(x)}$ varies).
2. Label distribution skew: Different distribution of labels ($P_{k(y)}$ varies).
3. Concept shift: Same feature, different labels (e.g., inter-reader variability).

Solution: SCAFFOLD uses control variables to correct for "client drift" caused by non-IID data.

Personalization Techniques

To improve performance on heterogeneous data, models can be personalized:
- Personalization Layers: Splitting the model into global layers (shared) and local layers (private to each client).

- FedBN: Keeping Batch Normalization parameters local to account for feature shifts.
- Hypernetworks: Using a central network to predict personalized model parameters for each client based on their data distribution.

Privacy and Security in FL

Despite data staying local, FL is vulnerable to several attacks:
1. Inference Attacks: Inferring class representatives, membership, or even training samples from gradients (Deep Leakage from Gradients).
2. Malicious Server: A server using a GAN to reconstruct client data.
3. Poisoning Attacks: Backdoor or replacement attacks to manipulate the global model.

Federated Learning with Differential Privacy (DP)

Definition 23 (Differential Privacy) . A mechanism M satisfies (ε, δ)-DP if for any two adjacent datasets D, D' differing by one individual:

$$P[M(D) \in S] \leq e^{\varepsilon} P[M(D') \in S] + \delta$$

Remark. **Handwritten Sensitivity derivation for FedAvg**: The sensitivity S_k of the update is given by the maximum change in weights. To ensure DP, updates must be:
1. Clipped: $|(w) = \frac{w}{\max\left(1, \frac{\|w\|_2}{C}\right)}$.
2. Noised: Adding Gaussian noise $n \sim N(0, \sigma^2 I)$ proportional to the sensitivity.

FedAVG with DP (*Pseudocode*)

- For each round t:
  - Sample clients k.
  - Clients update local weights w^k.
  - Clip local updates.
  - Add noise to the clipped updates.
  - Server aggregates noised weights and broadcasts the new global model.

Microscopy

Why Microscopy matters in Medicine?

Microscopy reveals structure and function at the cellular and tissue level, which is critical for diagnosis, research, and therapy decisions.

Key medical applications:
- Histopathology: For example, cancer diagnosis through tissue examination.
- Hematology: Analysis of blood smears.
- Infectious disease identification: Detecting pathogens.
- Cell biology & drug discovery: Understanding cellular mechanisms.

Remark. **Handwritten Workflow:**
1. Endoscopic Biopsy → 2. Gross Examination → 3. Tissue Fixation/Embedding → 4. Microtomy/Staining → 5. Microscopic Evaluation.

Why Machine Learning?

Traditional manual microscopy analysis is:
- Time-intensive: Pathologists must manually scan large slides.
- Subjective: High variability between different practitioners.
- Hard to scale: Difficult to handle large datasets of high-resolution slides.

Machine Learning (ML) Advantages:
- Automates repetitive tasks.
- Delivers quantitative measures (e.g., cell counts, morphology).
- Enables pattern discovery beyond human perception.

Microscopy Modalities Overview

Modality	Contrast Mechanism	Advantages	Limitations
Brightfield	Absorption by stains (H&E)	Cheap, clinical standard	Requires staining
Phase Contrast	Phase shifts (refractive index)	Live cell imaging (no stain)	Low molecular specificity
Fluorescence	Fluorophore emission	High specificity, multi-channel	Photobleaching, blur
Confocal	Pinhole rejection	3D optical sectioning	Slower, phototoxicity
Electron (TEM)	Electron scattering	Extremely high res (< 1 nm)	Expensive, destructive

Table 2 : Comparison of common microscopy modalities.

Brightfield Microscopy

White light passes through the sample, and the image is based on absorption by stains. This is the most used method in standard histology.

Definition 24 (Staining) . Biological tissues are largely transparent. Stains (like Hematoxylin & Eosin / H&E) bind selectively to cellular components (e.g., nuclei vs. cytoplasm) to convert biochemical differences into visible intensity differences.

Other Modalities

- Fluorescence Microscopy: Uses fluorophores that absorb excitation light and emit light at a longer wavelength.
- Confocal Microscopy A laser scanning technique using a pinhole to reject out-of-focus light, allowing for 3D "optical sectioning".
- Electron Microscopy: Uses electrons instead of photons for resolution up to 1,000,000x. Includes TEM (internal structure) and SEM (surface topology).

Key Challenges in Medical Imaging

1. Data: Expert annotations are expensive and time-consuming (pathologists spend hours per slide).
2. Whole Slide Images (WSI): Images can be massive (e.g., $100,000 \times 100,000$ pixels, 10GB per image).
3. Class Imbalance: Tasks often involve "rare events" like mitoses.
4. Domain Shifts: Variations in scanner types, staining protocols, and patient populations.

Multiple Instance Learning (MIL)

Due to the size of WSIs and the lack of pixel-level labels, we often use Weakly Supervised Learning through MIL.

Definition 25 (Multiple Instance Learning (MIL)) . Instead of individual labeled samples, we have bags of instances $X_j = \{x_{\{j1\}}, x_{\{j2\}}, ..., x_{\{jK\}}\}$.
- A bag is labeled $Y = 0$ if all instances are negative.
- A bag is labeled $Y = 1$ if at least one instance is positive.

Theorem 5 (Permutation Invariance) . A MIL scoring function $S(X)$ must be symmetric (invariant to the order of instances). It can be decomposed as:

$$S(X) = g\left(\sum_{\{xinX\}} f(x)\right)$$

Deep MIL Approaches

1. Instance-level approach: f is an instance classifier; scores are aggregated.
2. Embedding-level approach: f maps instances to low-dimensional embeddings, which are then pooled to create a bag representation for the classifier g.

Attention-based MIL Pooling

The bag representation z is computed as a weighted sum of instance embeddings h_k:

$$z = \sum_{\{k=1\}}^{K} a_k h_k$$

Where the attention weights a_k are:

$$a_k = \frac{\exp\left(w^T \tanh(Vh_k)\right)}{\sum_{\{j=1\}}^{K} \exp\left(w^T \tanh\left(Vh_j\right)\right)}$$

Gated variant: $a_k \propto \exp(w^T(\tanh(Vh_k) \odot \sigma(Uh_k)))$.