**Lecture notes**

# Mathematical methods of signal and image processing

Benjamin Berkels

WS 2021/22

RWTH Aachen

last change: February 28, 2022

typeset using KOMA-Script and LaTeX

# 1. Digital images and point operators

### References


[1] K. Bredies and D. Lorenz. *Mathematical Image Processing.* 1st ed. Vieweg+Teubner Verlag, 2018.

[2] R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* 3rd Edition. Prentice Hall, 2007.

[3] S. B. Damelin and W. Miller Jr. *The Mathematics of Signal Processing.* Cambridge University Press, 2012.

[4] S. Mallat. *A Wavelet Tour of Signal Processing.* 3rd ed. Elsevier, 2009.

[5] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing.* 2nd ed. Springer New York, 2006.


## 1.1. Digital images

Mathematically, an image is a mapping from a domain $\Omega \subset \mathbb{R}^d$ to a value range $V$. For instance, the mapping

$$f : (0,1)^2 \to [0,1]$$

is a typical (continuous) grayscale image. As domain, we consider a $d$-dimensional cuboid, i. e. $\Omega = (a_1, b_1) \times \cdots \times (a_d, b_d) \subset \mathbb{R}^d$. Depending on the acquisition technique, there are quite different value ranges:

- Grayscale images: $V = \mathbb{R}$ or $V = [0,1]$.
- Color images: $V = \mathbb{R}^3$ or $V = [0,1]^3$ (different color models, e. g. RGB, HSV).
- Diffusion Tensor Imaging (DTI): $V = \big\{M \in \mathbb{R}^{3\times 3} \,:\, M \text{ symmetric, positive definite}\big\}$
- Spectral Imaging: $V = \mathbb{R}^n$ or $V = [0,1]^n$ with $n \gg 1$.
- ...

**Definition 1.1** (Digital image)**.**

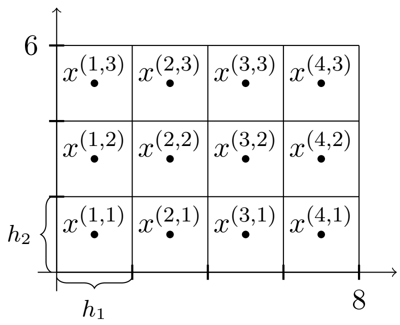
(i) $d \in \mathbb{N}$ is called *spatial dimension.*

(ii) Let $\underline{m} = (m_1, \ldots, m_d) \in \mathbb{N}^d$. The tuple $\underline{m}$ is called *image size.*

(iii) Combined, $\underline{m}$ and the cuboid $\Omega$ imply a grid: An index $\underline{j} \in \mathcal{I} \coloneqq \{1, \ldots, m_1\} \times \cdots \times \{1, \ldots, m_d\}$ corresponds to the grid node $x^{\underline{j}} = \left(x_1^{\underline{j}}, \ldots, x_d^{\underline{j}}\right) \in \Omega$, where

$$x_i^{\underline{j}} = a_i + \left(j_i - \tfrac{1}{2}\right) h_i, \; i = 1, \ldots, d.$$

Here, the *grid size* $\underline{h} \in \mathbb{R}^d$ is defined as $h_i = \frac{b_i - a_i}{m_i}$, $i = 1, \ldots, d$.




Grid for $d = 2$, $\Omega = (0, 8) \times (0, 6)$, $\underline{m} = (4, 3)$

(iv) A *grid* $\underline{X}$ is the $d$-dimensional array (tensor of order $d$, short: $d$-array) of all grid nodes, i. e.

$$\underline{X} = (x^{\underline{j}})_{\underline{j} \in \mathcal{I}}.$$

Such a grid is called *regular grid.* The special case $h_1 = \ldots = h_d$ is called *cartesian grid.*

(v) A grid implies a partition of $\Omega$ into cells

$$c_{\underline{j}} = \left\{ x \in \Omega \, : \, \left| x_i - x_i^{\underline{j}} \right| < \tfrac{h_i}{2} \text{ for all } i \in \{1, \ldots, d\} \right\}.$$

We have,

(1) $\overline{\Omega} = \bigcup_{\underline{j} \in \mathcal{I}} \overline{c_{\underline{j}}}$

(2) $c_{\underline{j}} \cap c_{\underline{k}} = \emptyset$ for $\underline{j}, \underline{k} \in \mathcal{I}$ with $\underline{j} \neq \underline{k}$

(3) $c_{\underline{j}} \neq \emptyset$ for $\underline{j} \in \mathcal{I}$

This kind of grid is called *cell-centered* and is usually used with finite differences. For multilinear finite elements, *mesh-centered* grids are more common.

(vi) If a value (e. g. a gray value) $f_{\underline{j}}$ is assigned to each cell, the values can be arranged in a $d$-array $F = (f_{\underline{j}})_{\underline{j} \in \mathcal{I}}$. This kind of $F$ is called *digital image*, *discrete image* or *pixel image.*

For actual image data one has to take into account that the value range is quantized and only contains finitely many elements. If a grayscale image is stored with a precision of $n$ bit, there are $2^n$ possible values. For grayscale images, the precision is typically 8 bit, i. e. $V = \{0, \ldots, 255\}$. Color images are usually stored with a precision of 8 bit per channel. Modern video formats like HDR10 or Dolby Vision use 10 resp. 12 bit.

(vii) The tuple $\left(c_{\underline{j}}, f_{\underline{j}}\right)$ is called *pixel* (for $d = 2$) or *voxel* (for $d = 3$). Pixel is short for "picture element" and voxel the analogue for volume elements.

(viii) A function $f : \Omega \to V$ that is constant on each cell, i. e. $f|_{c_{\underline{j}}}$ is constant for all $\underline{j} \in \mathcal{I}$, is called *pixelated.* Since

$$L \coloneqq \Omega \setminus \bigcup_{\underline{j} \in \mathcal{I}} c_{\underline{j}}$$

is a Lebesgue null set, the values can be freely chosen in the sense of the Lebesgue measure. If $x \in L$, then $x$ is between (at least) two cells, i.e. there are $\underline{j} \in \mathcal{I}$ and $i \in \{1, \dots, d\}$, such that $x \in \overline{c_{\underline{j}}}$ and $x_i = x_i^{\underline{j}} - \frac{h_i}{2}$. If multiple such indices $\underline{j}$ exist, we choose the index such that $\#\left\{i \in \{1, \dots, d\} : x_i = x_i^{\underline{j}} - \frac{h_i}{2}\right\}$ is maximal. Here, $\#A$ denotes the number of elements in the set $A$. For a pixelated function, we then assume that $f(x) = f(c_{\underline{j}})$.

*Example* 1.2. For $d = 2$, $\Omega = (0, 8) \times (0, 6)$ and $\underline{m} = (4, 3)$, we have

$$\underline{X} = \begin{pmatrix} (1,5) & (3,5) & (5,5) & (7,5) \\ (1,3) & (3,3) & (5,3) & (7,3) \\ (1,1) & (3,1) & (5,1) & (7,1) \end{pmatrix}.$$

An image on this grid is, for instance,

$$F = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 \end{pmatrix}.$$

*Remark* 1.3.

- In case $d = 1$, one more commonly talks about *signals* rather than images.
- In case $d = 2$, the cuboid $\Omega$ is usually chosen such that $h_1 = h_2 =: h$. Moreover, the cuboid is usually such that either $h = 1$ (thus $b_i - a_i = m_i$) or $h = \min \frac{1}{m_i}$ (thus $b_i - a_i = hm_i$).
- Volumetric data (e.g. MRT or CT) is mostly acquired slice-wise. Here, the grid size in a slice is usually constant ($h_1 = h_2$) but not orthogonal to the slices ($h_1 \neq h_3$). In this case, the voxels are no cubes.

*Remark* 1.4 (Scanning). Digital images can be acquired by scanning: Let $f : \Omega \to \mathbb{R}$ be a function (signal/image). If $f \in L^2(\Omega)$, a measurement of $f$ is generally described by

$$(f, \psi)_{L^2} = \int_\Omega f(y)\psi(y)\,\mathrm{d}y\,.$$

Here, $\psi \in L^2(\mathbb{R}^d)$ is a test function with $\psi \geq 0$ and $\int_{\mathbb{R}^d} \psi(x)\,\mathrm{d}x = 1$.

A simple approximation of a CCD sensor is, for example, given by the test function

$$r_h(x) = \begin{cases} \frac{1}{h^d} & \|x\|_\infty < \frac{h}{2} \\ 0 & \text{else} \end{cases} \quad \text{and } (f, r_h(x - \cdot))_{L^2} = \int_\Omega f(y) r_h(x-y)\,\mathrm{d}y =: r_x^h[f].$$

For a given grid $\underline{X}$ one gets the discrete image $(r^h_{x^{\underline{j}}}[f])_{\underline{j} \in \mathcal{I}}$.

This approach can also be used to represent a high resolution image with a smaller resolution, which is required, for instance, in so-called multilevel methods.

*Remark* 1.5 (Coordinate systems). We consider the unit square $(0, 1)^2$ as image domain. Mathematically, the origin, i.e. the point $(0, 0)$, is in the “lower left” of the domain. For digital images, when they are displayed on the screen, this is different. The origin pixel, i.e. the pixel corresponding to the index (1,1) (or (0,0) in C indexing), is “top left”. Moreover, the first coordinate index corresponds to the vertical axis, the second coordinate index to the horizontal axis. This corresponds to a rotation of the image domain by 90 degrees. This is important, for instance, when expressing angles.

## 1.2. Intensity transformations

In image processing, one distinguishes the following classes of methods, which create a new image from a given image (e. g. a denoised image from a noisy image):

- Point operators / intensity transformations
- Local operators
- Global operators

First, we address point operators, which can be used, for instance, to correct (global) illumination problems.

**Definition 1.6.** Let $f : \Omega \to \mathbb{R}$ be an image and $T : \mathbb{R} \to \mathbb{R}$. $T$ is called *intensity transformation* and $T \circ f$ is called *intensity transformed image.*

(i) For $a < b$, the mapping

$$T^{\text{clip}}_{[a,b]} : \mathbb{R} \to [a, b], s \mapsto \begin{cases} a & s \leq a \\ b & s \geq b \\ s & \text{else} \end{cases}$$

is called *intensity clipping.*

(ii) For $a < b$, the mapping

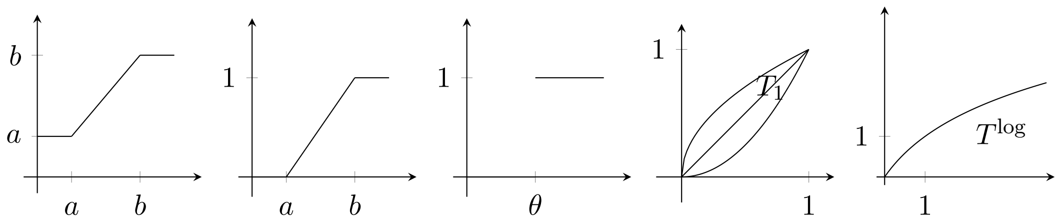
$$T^{\text{norm}}_{[a,b]} : \mathbb{R} \to [0, 1], s \mapsto T^{\text{clip}}_{[0,1]} \left( \frac{s - a}{b - a} \right)$$

is called *intensity normalization.* If $f$ is not constant and $f(\Omega)$ closed and bounded, $\hat{f} = T^{\text{norm}}[f] := T^{\text{norm}}_{[\min f, \max f]} \circ f$ has value range $[0, 1]$ and it holds that $\min \hat{f} = 0$ and $\max \hat{f} = 1$.

(iii) The mapping

$$T^{\text{threshold}}_{\theta} : \mathbb{R} \to \{0, 1\}, s \mapsto \begin{cases} 0 & s \leq \theta \\ 1 & s > \theta \end{cases}$$

is called *thresholding.* $H := T^{\text{threshold}}_{0}$ is known as *Heaviside function* and also plays a role in areas other than image processing.



(iv) For $\gamma > 0$, the mapping

$$T_\gamma : [0,1] \to [0,1], s \mapsto s^\gamma$$

is called *gamma correction.* For $\gamma < 1$, the mapping is strictly concave, for $\gamma > 1$ it is strictly convex.

(v) The mapping

$$T^{\log} : [0,\infty) \to [0,\infty), s \mapsto \log_2(1+s)$$

is called *log transformation* and is often used to visualize the *power spectrum* (absolute value of the Fourier transform). The base of the logarithm is 2, so that $[0,1]$ is mapped to $[0,1]$ bijectively.

*Remark* 1.7 (Discrete histogram)*.* If $F$ is a discrete image with value range $V$,

$$H_F : V \to \mathbb{N}_0, s \mapsto \#\left\{\underline{j} \in \mathcal{I} \,:\, f_{\underline{j}} = s\right\}$$

is called *(discrete) histogram* of $F$. It states which values occur how often in the image. Using the Kronecker delta

$$\delta_{i,j} := \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases},$$

the histogram can be expressed as a sum:

$$H_F(s) = \sum_{\underline{j} \in \mathcal{I}} \delta_{s,f_{\underline{j}}}$$

If the value range is $V := \{0,\dots,n\}$, the *(discrete) cumulative distribution function* of $F$ is defined as

$$G_F : \{0,\dots,n\} \to \mathbb{N}_0, s \mapsto \sum_{r=0}^{s} H_F(r).$$

The histogram can be easily extended to pixelated images. To this end, we first introduce some additional notation.

**Definition 1.8.**

(i) For a set $A \subset \mathbb{R}^d$,

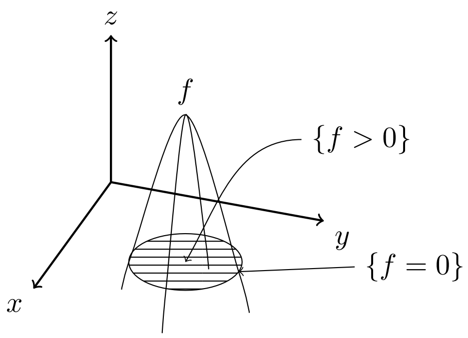
$$\chi_A : \mathbb{R}^d \to \{0,1\}, x \mapsto \begin{cases} 1 & x \in A \\ 0 & \text{else} \end{cases}$$

is called *characteristic function* of $A$.

(ii) Let $\Omega \subset \mathbb{R}^d$ be a domain and $f : \Omega \to \mathbb{R}$. For $s \in \mathbb{R}$,

$$\{f = s\} := \{x \in \Omega \,:\, f(x) = s\}$$

is called *s-level set* of $f$. Analogously, one defines $\{f \leq s\}$, the *s-sublevel set* of $f$ and $\{f \geq s\}$, the *s-superlevel set* of $f$. We have $f^{-1}(s) = \{f = s\}$, $f^{-1}((-\infty, s]) = \{f \leq s\}$ and $f^{-1}([s,\infty)) = \{f \geq s\}$. $\{f = s\}$ is also called *level line/isoline* ($d = 2$) or *level surface/isosurface* ($d = 3$).




(iii) For a bounded set $A \subset \mathbb{R}^d$,

$$|A| \coloneqq \mathrm{Vol}(A) \coloneqq \int_A \mathrm{d}x = \int_{\mathbb{R}^d} \chi_A(x)\,\mathrm{d}x$$

denotes the *volume* of $A$. In particular, $\mathrm{Vol}(A) = \lambda(A)$, where $\lambda$ is the Lebesgue measure.

*Remark* 1.9 (Histogram)*.* Let $f$ be a pixelated image and $F$ the corresponding discrete image. Then,

$$H_F(s) = \sum_{\underline{j}\in\mathcal{I}} \frac{1}{\mathrm{Vol}(c_{\underline{j}})} \int_{c_{\underline{j}}} \chi_{\{f=s\}}(x)\,\mathrm{d}x = \frac{1}{\prod h_i} \int_\Omega \chi_{\{f=s\}}(x)\,\mathrm{d}x = \frac{1}{\prod h_i} \mathrm{Vol}(\{f = s\}).$$

Except for the scaling with the cell volume, the discrete histogram is equal to the volume of the level sets. The latter is a property that can also be considered on images that are not pixelated.

This motivates to consider the following mapping as basis for a continuous histogram:

$$H_f : \mathbb{R} \to [0,\infty), s \mapsto \mathrm{Vol}(\{f = s\})$$

Since all level sets of a continuous image could be Lebesgue null sets (e. g. $f(x) = x$ on $[0,1]$), we extend $H_f$ from $\mathbb{R}$ as domain to subsets of $\mathbb{R}$ as domain, specifically to $\mathcal{B}(\mathbb{R})$, the set of Borel measurable subsets of $\mathbb{R}$:

$$H_f : \mathcal{B}(\mathbb{R}) \to [0,\infty), A \mapsto \mathrm{Vol}(\{f \in A\}).$$

Thus, $H_f$ is a (positive) measure. In particular,

$$H_f(A) = \mathrm{Vol}(\{f \in A\}) = \mathrm{Vol}(f^{-1}(A)) = \lambda(f^{-1}(A)).$$

Therefore, $H_f$ is the push-forward measure (or image measure) of the Lebesgue measure $\lambda$ under the mapping $f$, cf. Proposition A.8. For $g : \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{R}} g(s)\mathrm{d}H_f(s) = \int_\Omega (g \circ f)(x)\,\mathrm{d}\lambda(x),$$

provided $f : \Omega \to \mathbb{R}$ is $\mathcal{B}(\Omega)$-$\mathcal{B}(\mathbb{R})$ measurable and $g$ is $f(\lambda)$-integrable, cf. Proposition A.9.

The *cumulative distribution function* of $f$ is defined as

$$G_f(s) = \mathrm{Vol}(\{f \le s\}).$$

The change of this function at a position $s$ is also a measure for the histogram of $f$ at $s$. However, $G_f$ is not necessarily differentiable in the classical sense. For example, for pixelated images,

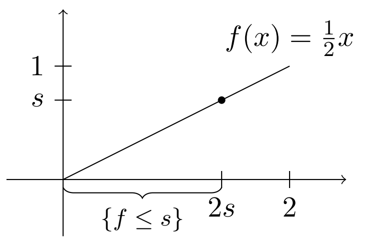
$G_f$ is a step function and thus not differentiable. One can show that $H_f$ is the distributional derivative of $G_f$ (Exercise).

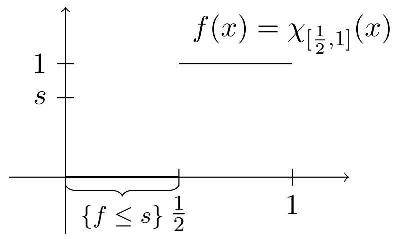
Examples:

- For $f : (0, 2) \to [0, 1], x \mapsto \frac{1}{2}x$, we have $G_f : [0, 1] \to [0, 2], s \mapsto 2s$ and $G'_f(s) = 2$.
- For $f : (0, 1) \to [0, 1], x \mapsto \chi_{[\frac{1}{2},1]}(x)$, we have

$$G_f(s) = \begin{cases} 0 & s < 0 \\ \frac{1}{2} & 0 \le s < 1 \\ 1 & s \ge 1 \end{cases} \text{ and } H_f = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

Here, $\delta_x$ denotes the Dirac delta function centered at $x$.




If $f$ is pixelated with value range $V := \{0, \dots, n\}$ and $F$ the corresponding discrete image, then

$$G_f(s) = \operatorname{Vol}\left(\bigcup_{r=0}^{s}\{f = r\}\right) = \sum_{r=0}^{s} \operatorname{Vol}(\{f = r\}) = \left(\prod h_i\right) \sum_{r=0}^{s} H_F(r) = \left(\prod h_i\right) G_F(s).$$

Except for the scaling with the cell volume, the discrete cumulative distribution function is equal to the volume of the sublevel sets.

*Remark* 1.10 (Binning). The histogram of a continuous function can be approximated discretely in a straightforward manner. In the same way, a discrete histogram can be approximated. To this end, the value range $V$ is partitioned into intervals (so-called *bins*).

Choose $m \in \mathbb{N}$ and $r_0 < r_1 < \ldots < r_m \in \mathbb{R}$. Then, the bins are defined as

$$B_k = \begin{cases} [r_{k-1}, r_k) & k \in \{1, \dots, m-1\} \\ [r_{k-1}, r_k] & k = m \end{cases}$$

and the binned discrete histogram of $f : \Omega \to \mathbb{R}$ is

$$H_k[f] = \operatorname{Vol}(\{f \in B_k\}) \text{ for all } k \in \{1, \dots, m\}.$$

If the value range of $f$ is bounded, a typical choice of the interval boundaries is

$$r_k = \begin{cases} \inf f + \frac{k}{m-1}(\sup f - \inf f) & k \in \{0, \dots, m-1\} \\ \sup f + 1 & k = m \end{cases}.$$

As example, we consider the image

$$F = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 4 & 3 & 4 & 2 \\ 7 & 1 & 4 & 1 \end{pmatrix}$$

on a cartesian grid with grid size 1 and the following bins

$$B_1 = [1,3),\ B_2 = [3,5),\ B_3 = [5,7),\ B_4 = [7,8].$$

Thus, we get

$$H_1 = 7,\ H_2 = 4,\ H_3 = 0,\ H_4 = 1.$$

*Example* 1.11 (Histogram equalization)*.* If the contrast of an image $f$ is still not sufficient after a normalization with $T^{\text{norm}}$, the contrast can potentially be further improved using the histogram:

Here, the idea is to distribute the gray values as uniformly as possible in the value range $[0,1]$. Let now $f : \Omega \to [0,1]$. We look for a bijective, increasing intensity transformation $T : [0,1] \to [0,1]$ such that

$$G_{T\circ f}(s) = s\,\mathrm{Vol}(\Omega) \text{ for all } s \in [0,1]. \tag{$*$}$$

Then, $G'_{T\circ f} \equiv \mathrm{Vol}(\Omega)$, i. e. the gray values are distributed uniformly. If $T$ fulfills $(*)$, we get

$$T(s)\,\mathrm{Vol}(\Omega) \overset{(*)}{=} G_{T\circ f}(T(s)) = \mathrm{Vol}(\{T\circ f \le T(s)\}) \overset{T \text{ bijective, increasing}}{=} \mathrm{Vol}(\{f \le s\}) = G_f(s).$$

If $f$ is such that $G_f \in C([0,1],[0,\mathrm{Vol}(\Omega)])$, $G_f$ is strictly increasing and $G_f(0) = 0$, then

$$T : [0,1] \to [0,1], s \mapsto \frac{G_f(s)}{\mathrm{Vol}\,(\Omega)}$$

has the desired properties: The continuity and strict monotonicity of $G_f$ combined with $G_f(0) = 0$ imply the invertibility of $G_f$ on $G_f([0,1]) = [0,\mathrm{Vol}(\Omega)]$. Thus, $T^{-1}(s) = G_f^{-1}(s\,\mathrm{Vol}(\Omega))$ and we get

$$G_{T\circ f}(s) = \mathrm{Vol}(\{T\circ f \le s\}) = \mathrm{Vol}(\{f \le T^{-1}(s)\}) = G_f(T^{-1}(s)) = s\,\mathrm{Vol}(\Omega).$$

If the value range is discrete, we cannot create a uniformly distributed histogram (since a gray value can only be mapped to exactly one gray value), but the idea can still be applied in this case. Let $f : \Omega \to V$ be a pixelated image with $V \coloneqq \{0,\dots,n\}$ and $F$ the corresponding discrete image. Moreover, let $\mathrm{rd} : \mathbb{R} \to \mathbb{Z}$ denote the rounding mapping, i. e. a real number is mapped to its nearest integer (rounding up if there are two nearest integers). More precisely,

$$\mathrm{rd}(x) = \begin{cases} \lfloor x \rfloor & x - \lfloor x \rfloor < \frac{1}{2} \\ \lfloor x \rfloor + 1 & \text{else} \end{cases},$$

where

$$\lfloor x \rfloor \coloneqq \max\{k \in \mathbb{Z} : k \le x\}$$

is the floor function. Then, one gets a discrete histogram equalization using the intensity transformation

$$T(s) = \mathrm{rd}\left(\frac{n}{\mathrm{Vol}(\Omega)} G_f(s)\right) = \mathrm{rd}\left(\frac{n}{\#\mathcal{I}} G_F(s)\right).$$

Using the first equality, the idea can also be applied to non-pixelated images with discrete value range.

*Remark* 1.12 (Histogram matching). If images are taken from the same scene but with different sensors, the histograms can be used to compensate sensor-dependent intensity differences.

Let $f, g : \Omega \to [0,1]$ be such that $G_f$ and $G_g$ are invertible. Moreover, let

$$T(s) = \frac{G_f(s)}{\operatorname{Vol}(\Omega)} \text{ and } S(s) = \frac{G_g(s)}{\operatorname{Vol}(\Omega)}$$

be the intensity transformations for histogram equalization. With those, the histogram of one of the images can be transferred to the other image: We have

$$\begin{aligned} G_{S^{-1}\circ T\circ f}(s) &= \operatorname{Vol}(\{S^{-1}\circ T\circ f \le s\}) = \operatorname{Vol}(\{T\circ f \le S(s)\}) = G_{T\circ f}(S(s)) \\ &= S(s)\operatorname{Vol}(\Omega) = G_g(s). \end{aligned}$$

Thus, the histograms of $S^{-1}\circ T\circ f$ and $g$ coincide.

Note that above we did not use that $G_g$ is the cumulative distribution function of an image $g$ but just the bijectivity of $G_g$ as mapping from $[0,1]$ to $[0, \operatorname{Vol}(\Omega)]$. Thus, this idea allows transferring an arbitrary, bijective, increasing distribution function to $f$.

*Example* 1.13 (Segmentation by thresholding). We consider an image $f : \Omega \to [0,1]$ of a handwritten text (e. g. scan of a note or a photo of a blackboard). For such an image, it is useful to reduce the number of gray values to two, i. e. to segment the image. This compensates differences in illumination and reduces noise. Moreover, the data size is reduced significantly. The segmentation can be achieved using $T_\theta^{\text{threshold}}$, but requires a threshold $\theta$ to be chosen.

$T_\theta^{\text{threshold}}$ partitions the image into two components, $\Sigma = \{f \le \theta\}$ and $\Omega \setminus \Sigma = \{f > \theta\}$. Now, the idea is to chose the threshold $\theta$ in such a way that it corresponds to the average of the average gray values in the two components, i. e.

$$\theta = \frac{1}{2}\left(\fint_{\{f\le\theta\}} f(x)\,\mathrm{d}x + \fint_{\{f>\theta\}} f(x)\,\mathrm{d}x\right) =: \varphi(\theta),$$

where

$$\fint_A f(x)\,\mathrm{d}x = \frac{1}{\operatorname{Vol}(A)}\int_A f(x)\,\mathrm{d}x\,.$$

In case $\operatorname{Vol}(\{f \le \theta\}) = 0$ or $\operatorname{Vol}(\{f > \theta\}) = 0$, we define

$$\fint_{\{f\le\theta\}} f(x)\,\mathrm{d}x = \theta \text{ and } \fint_{\{f>\theta\}} f(x)\,\mathrm{d}x = \theta, \text{ respectively.}$$

This equation can be solved with the fixed point iteration

$$\theta^{n+1} = \varphi(\theta^n) \text{ for } n \in \mathbb{N},$$

given an initial value $\theta^1 \in (0,1)$. A typical initial value is $\theta^1 = \frac{1}{2}(\inf f + \sup f)$. This method to determine the threshold is called *isodata algorithm*.

Since $f(\Omega) \subset [0,1]$, we have $\varphi([0,1]) \subset [0,1]$. Moreover, for $\theta_1 \leq \theta_2$,

$$\begin{aligned}
\varphi_a(\theta_2) &:= \fint_{\{f \leq \theta_2\}} f(x)\,\mathrm{d}x = \frac{1}{|\{f \leq \theta_2\}|}\left(\int_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x + \int_{\{\theta_1 < f \leq \theta_2\}} f(x)\,\mathrm{d}x\right) \\
&\geq \frac{1}{|\{f \leq \theta_2\}|}\left(\int_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x + \theta_1\,|\{\theta_1 < f \leq \theta_2\}|\right) \\
&\geq \frac{1}{|\{f \leq \theta_2\}|}\left(\int_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x + \fint_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x\,|\{\theta_1 < f \leq \theta_2\}|\right) \\
&= \frac{1}{|\{f \leq \theta_2\}|}\left(|\{f \leq \theta_1\}| + |\{\theta_1 < f \leq \theta_2\}|\right)\fint_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x \\
&= \fint_{\{f \leq \theta_1\}} f(x)\,\mathrm{d}x = \varphi_a(\theta_1).
\end{aligned}$$

This means that $\varphi_a$ is increasing. Analogously, one shows that $\varphi_b(\theta) := \fint_{\{f>\theta\}} f(x)\,\mathrm{d}x$ is increasing. Thus, as sum of two increasing functions, $\varphi$ is also increasing. Let $\theta^1 \in [0,1]$ be arbitrary but fixed. If $\theta^2 = \varphi(\theta^1) \geq \theta^1$, we get $\theta^{n+1} \geq \theta^n$, i. e. the sequence is increasing. Otherwise, we have $\theta^2 < \theta^1$ and we get $\theta^{n+1} \leq \theta^n$, i. e. the sequence is decreasing. In both cases, the sequence is monotonic and bounded, thus convergent.

Moreover, one can show: Every monotonic function $\varphi : [0,1] \to [0,1]$ has a fixed point (exercise). In case $\varphi$ is not continuous, the limit of the sequence $(\theta^n)_{n\in\mathbb{N}}$ is not necessarily a fixed point of $\varphi$ though.

The average gray values can be computed with the histogram. We have (exercise)

$$\fint_{\{f \leq \theta\}} f(x)\,\mathrm{d}x = \int_{\mathbb{R}} s\chi_{(-\infty,\theta]}(s)\mathrm{d}H_f(s) \Big/ \int_{\mathbb{R}} \chi_{(-\infty,\theta]}(s)\mathrm{d}H_f(s),$$

$$\fint_{\{f > \theta\}} f(x)\,\mathrm{d}x = \int_{\mathbb{R}} s\chi_{(\theta,\infty)}(s)\mathrm{d}H_f(s) \Big/ \int_{\mathbb{R}} \chi_{(\theta,\infty)}(s)\mathrm{d}H_f(s).$$

If $F$ is a discrete image with value range $V := \{0,\ldots,n\}$, we have

$$\varphi(\theta) = \frac{1}{2}\left(\sum_{s=0}^{\lfloor\theta\rfloor} sH_F(s) \Big/ \sum_{s=0}^{\lfloor\theta\rfloor} H_F(s) + \sum_{s=\lfloor\theta\rfloor+1}^{n} sH_F(s) \Big/ \sum_{s=\lfloor\theta\rfloor+1}^{n} H_F(s)\right)$$

and the fixed point iteration can be implemented using this expression.

# 2. Local operators

Now, we address the second class of operators in image processing. To compute a value of the new image, these operators not only use the value at the corresponding position in the input image, but also values from a local neighborhood.

*Example* 2.1 (The moving average). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a noisy image, i. e. $f = f_0 + n$. Here, $f_0 : \mathbb{R}^d \to [0,1]$ is the (sought) original image without noise and $n : \mathbb{R}^d \to \mathbb{R}$ is noise with mean 0. Since the mean of the noise is 0, a local averaging should reduce the noise. Thus, we consider

$$M_r f(x) = ⨍_{B_r(x)} f(y)\,\mathrm{d}y\,, \text{ where } B_r(x) = \left\{ y \in \mathbb{R}^d \;:\; \|y-x\|_2 < r \right\}.$$

This operation is one of the simplest methods for noise reduction and is called *moving average*. For now, we intentionally consider $\mathbb{R}^d$ here instead of $\Omega$ to avoid boundary effects.

Instead of a ball in the Euclidean norm, one typically also considers other norms. In the discrete case, for instance, the $\infty$-norm is particularly easy to compute:

$$M_r^\infty f(x) = ⨍_{B_r^\infty(x)} f(y)\,\mathrm{d}y\,, \text{ where } B_r^\infty(x) = \left\{ y \in \mathbb{R}^d \;:\; \|y-x\|_\infty < r \right\}.$$

Using $\chi_{B_r(0)}$, one can express $M_r$ in a different way:

$$M_r f(x) = \int_{\mathbb{R}^d} f(y) \left( \tfrac{1}{|B_r(0)|} \chi_{B_r(0)}(y-x) \right) \mathrm{d}y \underset{\text{Theorem B.7}}{\overset{\phi(y)=y-x}{=}} \int_{\mathbb{R}^d} f(x+y) \left( \tfrac{1}{|B_r(0)|} \chi_{B_r(0)}(y) \right) \mathrm{d}y$$

This is the typical way to express a linear filter:

**Definition 2.2.** For measurable $\psi : \Omega \to \mathbb{R}$ and $f : \mathbb{R}^d \to \mathbb{R}$,

$$(\psi \star f)(x) = \int_\Omega \psi(y) f(x+y)\,\mathrm{d}y$$

is called *cross-correlation* of $f$ and $\psi$ at position $x \in \mathbb{R}^d$. The operation

$$M_\psi : f \mapsto (x \mapsto (\psi \star f)(x))$$

is called *linear filter* with *kernel* $\psi$.

**Proposition 2.3** (Properties of the cross-correlation)**.**

*(i) Let* $1 \le p, q, r \le \infty$ *with* $\frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$ *(using the convention* $\frac{1}{\infty} = 0$*),* $f \in L^p(\Omega)$ *and* $g \in L^q(\mathbb{R}^d)$*. Then,* $(f \star g) \in L^r(\mathbb{R}^d)$ *and*

$$\|f \star g\|_{L^r(\mathbb{R}^d)} \le \|f\|_{L^p(\Omega)} \|g\|_{L^q(\mathbb{R}^d)}\,.$$

*This is the counterpart of Young's convolution inequality for the cross-correlation, cf. Proposition B.10 (i).*

*If* $\Omega = \mathbb{R}^d$*, we have* $(f \star g)(x) = (g \star f)(-x)$ *for all* $x \in \mathbb{R}^d$ *and*

$$\|f \star g\|_{L^r(\mathbb{R}^d)} = \|g \star f\|_{L^r(\mathbb{R}^d)}\,.$$

*(ii) Let $\psi \in C_c^k(\mathbb{R}^d)$ and $f \in L^p(\Omega)$ with $1 \le p \le \infty$. Then, $(f \star \psi) \in C^k(\mathbb{R}^d)$. If $\alpha$ is a multi-index with $|\alpha| \le k$, we have*

$$\frac{\partial^\alpha}{\partial x^\alpha}(f \star \psi) = f \star \frac{\partial^\alpha}{\partial x^\alpha}\psi.$$

*If $\Omega = \mathbb{R}^d$, we also have*

$$\frac{\partial^\alpha}{\partial x^\alpha}(\psi \star f) = (-1)^{|\alpha|}\frac{\partial^\alpha}{\partial x^\alpha}\psi \star f.$$

*(iii) Let $\psi \in L^1(\mathbb{R}^d)$ be such that $\psi \ge 0$ and $\int_{\mathbb{R}^d} \psi(x)\,\mathrm{d}x = 1$. Moreover, for $\epsilon > 0$, let*

$$\psi_\epsilon : \mathbb{R}^d \to \mathbb{R}, x \mapsto \tfrac{1}{\epsilon^d}\psi(\tfrac{x}{\epsilon})$$

*and let $f \in L^\infty(\mathbb{R}^d)$. If $f$ is continuous in $x \in \mathbb{R}^d$, we have*

$$\lim_{\epsilon \to 0}(\psi_\epsilon \star f)(x) = f(x).$$

*If $f$ is uniformly continuous, $\psi_\epsilon \star f$ converges to $f$ on each compact subset of $\mathbb{R}^d$ uniformly.*

**Proof.**

(i) For $x \in \mathbb{R}^d$ and $r = \infty$, we get using Hölder's inequality (cf. Lemma B.5)

$$|(f \star g)(x)| \le \int_\Omega |f(y)g(x+y)|\,\mathrm{d}y \le \|f\|_{L^p(\Omega)}\,\|g(\cdot + x)\|_{L^q(\Omega)} \le \|f\|_{L^p(\Omega)}\,\|g\|_{L^q(\mathbb{R}^d)}\,.$$

Since the right hand side is independent of $x$, taking the supremum over $x \in \mathbb{R}^d$ shows

$$\|f \star g\|_{L^\infty(\mathbb{R}^d)} \le \|f\|_{L^p(\Omega)}\,\|g\|_{L^q(\mathbb{R}^d)}\,.$$

Now consider $r = p$, which implies $q = 1$. For $p = \infty$, we have $r = \infty$, which we already covered above. For $p < \infty$, we integrate the $p$-th power of

$$|(f \star g)(x)| \le \int_\Omega |f(y)g(x+y)|\,\mathrm{d}y\,.$$

and get

$$\int_{\mathbb{R}^d} |(f \star g)(x)|^p\,\mathrm{d}x \le \int_{\mathbb{R}^d}\left(\int_\Omega |f(y)g(x+y)|\,\mathrm{d}y\right)^p \mathrm{d}x\,. \tag{$*$}$$

For $p = 1$, we get using Fubini's theorem (cf. Theorem B.6)

$$\|f \star g\|_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |(f \star g)(x)|\,\mathrm{d}x \overset{(*)}{\le} \int_\Omega |f(y)| \int_{\mathbb{R}^d} |g(x+y)|\,\mathrm{d}x\,\mathrm{d}y = \|f\|_{L^1(\Omega)}\,\|g\|_{L^1(\mathbb{R}^d)}\,.$$

For $1 < p < \infty$, we get using $p^*$ with $\frac{1}{p} + \frac{1}{p^*} = 1$ that

$$\begin{aligned}
\int_{\mathbb{R}^d} |(f \star g)(x)|^p\,\mathrm{d}x &\overset{(*)}{\le} \int_{\mathbb{R}^d}\left(\int_\Omega |f(y)|\,|g(x+y)|^{\frac{1}{p}}\,|g(x+y)|^{\frac{1}{p^*}}\,\mathrm{d}y\right)^p \mathrm{d}x \\
&\overset{\text{Hölder}}{\le} \int_{\mathbb{R}^d}\left(\int_\Omega |f(y)|^p\,|g(x+y)|\,\mathrm{d}y\right)^{\frac{p}{p}}\left(\int_\Omega |g(x+y)|\,\mathrm{d}y\right)^{\frac{p}{p^*}} \mathrm{d}x \\
&\overset{\text{Fubini}}{\le} \int_\Omega |f(y)|^p \int_{\mathbb{R}^d} |g(x+y)|\,\mathrm{d}x\,\mathrm{d}y\,\|g\|_{L^1(\mathbb{R}^d)}^{\frac{p}{p^*}} \\
&= \|f\|_{L^p(\Omega)}^p\,\|g\|_{L^1(\mathbb{R}^d)}^{\frac{p}{p^*}+1}\,.
\end{aligned}$$

Taking the $p$-th root shows

$$\|f \star g\|_{L^p(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^1(\mathbb{R}^d)} .$$

The case $r \neq p$ can be shown similarly, but is more complicated and will not be shown here. The proof for the corresponding statement for the convolution can be found in the book of Lieb and Loss[1].

In case $\Omega = \mathbb{R}^d$, using the substitution rule for multiple variables (Theorem B.7), we get for $x \in \mathbb{R}^d$ that

$$\begin{aligned}(f \star g)(x) &= \int_{\mathbb{R}^d} f(y)g(x+y)\,\mathrm{d}y \overset{\phi(y)=x+y}{=} \int_{\mathbb{R}^d} f(y-x)g(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} g(y)f(-x+y)\,\mathrm{d}y = (g \star f)(-x).\end{aligned}$$

From this and the substitution rule for multiple variables, we immediately get

$$\|f \star g\|_{L^r(\mathbb{R}^d)} = \|g \star f\|_{L^r(\mathbb{R}^d)} .$$

(ii) Let $i \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d$. We have

$$\frac{(f \star \psi)(x+\epsilon e_i) - (f \star \psi)(x)}{\epsilon} = \int_\Omega f(y) \frac{\psi(x+y+\epsilon e_i) - \psi(x+y)}{\epsilon}\,\mathrm{d}y . \tag{$*$}$$

For $g \in C_c^1(\mathbb{R})$, the difference quotient converges uniformly to $\partial_i g$ (will be shown in Lemma 2.4). Thus,

$$\frac{\psi(x+y+\epsilon e_i) - \psi(x+y)}{\epsilon} =: \psi_i^\epsilon(x+y)$$

converges uniformly to $\partial_i \psi(x+y)$ for $y \in \mathbb{R}^d$. In other words, $\|\psi_i^\epsilon - \partial_i \psi\|_{L^\infty(\mathbb{R}^d)} \to 0$ for $\epsilon \to 0$. Since $\psi$ has compact support, there is a compact set $K \subset \mathbb{R}^d$ such that $\operatorname{supp} \psi(x+\cdot) \subset K$ and $\operatorname{supp} \psi_i^\epsilon(x+\cdot) \subset K$ for all $\epsilon \in (0,1)$. With this, we get

$$\left| \int_\Omega f(y)\psi_i^\epsilon(x+y)\,\mathrm{d}y - \int_\Omega f(y)\partial_i\psi(x+y)\,\mathrm{d}y \right| \leq \underbrace{\|\psi_i^\epsilon - \partial_i\psi\|_{L^\infty(\mathbb{R}^d)}}_{\to 0 \text{ for } \epsilon \to 0} \underbrace{\|\chi_K f\|_{L^1(\Omega)}}_{\leq \|\chi_K\|_{L^{p^*}(\Omega)} \|f\|_{L^p(\Omega)}} .$$

This shows that when passing to the limit $\epsilon \to 0$ in $(*)$, integration and limit on the right hand side can be interchanged. Thus, we get

$$\partial_i(f \star \psi) = f \star \partial_i \psi.$$

Iterating this, the first claim follows. In case $\Omega = \mathbb{R}^d$, we have

$$\partial_i(\psi \star f)(x) = \partial_i\left((f \star \psi)(-x)\right) = -(f \star \partial_i\psi)(-x) = -(\partial_i\psi \star f)(x).$$

Iterating this, the second claim follows.

Note that $\partial_i(f \star \psi) = f \star \partial_i \psi$ can also be proven using Corollary B.12, which follows from the dominated convergence theorem (Theorem B.11).

[1] E. H. Lieb and M. Loss. *Analysis.* 2nd ed. Vol. 14. Graduate Studies in Mathematics. American Mathematical Society, 2001.

(iii) Exercise. [1, Theorem 3.13.3] shows the statement for uniformly continuous $f$. $\square$

**Lemma 2.4.** *Let $\Omega \subset \mathbb{R}^d$ be open and $f \in C_c^1(\Omega)$. Then, for all $i \in \{1, \dots, d\}$ and $\epsilon \in (0, \infty)$, there is a $\delta \in (0, \infty)$ with*

$$\left| \frac{f(x + he_i) - f(x)}{h} - \partial_i f(x) \right| < \epsilon \text{ for all } x \in \Omega \text{ and } h \in (-\delta, \delta) \setminus \{0\} \text{ with } x + he_i \in \Omega.$$

*In other words, the difference quotient of $f$ converges uniformly to $\partial_i f$.*

**Proof.** Let $i \in \{1, \dots, d\}$ and $\epsilon \in (0, \infty)$ be arbitrary, but fixed. Since $f \in C_c^1(\Omega)$, we get $f \in C_c^1(\mathbb{R}^d)$ by extending $f$ with zero. Thus, $\partial_i f \in C_c^0(\mathbb{R}^d)$. Since $\operatorname{supp}(\partial_i f)$ is compact, $\partial_i f$ is even uniformly continuous on its support and, since it is zero outside its support, also uniformly continuous on $\mathbb{R}^d$. In particular, there is a $\delta \in (0, \infty)$ with

$$|\partial_i f(x) - \partial_i f(y)| < \epsilon \text{ for all } x, y \in \mathbb{R}^d \text{ with } |x - y| < \delta.$$

Let $x \in \mathbb{R}^d$ and $h \in (0, \delta)$. The function $g : [0, h] \to \mathbb{R}, t \mapsto f(x + te_i)$ is continuous on $[0, h]$ and differentiable on $(0, h)$ since $f \in C^1(\mathbb{R}^d)$. By the mean value theorem, there is a $\zeta \in (0, h)$ with

$$\partial_i f(x + \zeta e_i) = g'(\zeta) \overset{\text{mean value thm.}}{=} \frac{g(h) - g(0)}{h} = \frac{f(x + he_i) - f(x)}{h}.$$

Then, $|x + \zeta e_i - x| < h < \delta$ and we get

$$\left| \frac{f(x + he_i) - f(x)}{h} - \partial_i f(x) \right| = |\partial_i f(x + \zeta e_i) - \partial_i f(x)| < \epsilon.$$

The case $0 < -h < \delta$ is shown analogously. $\square$

**Proposition 2.5.** *Let $1 < p \leq \infty$ and $q$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Moreover, let $f \in L^p(\Omega)$ and $g \in L^q(\mathbb{R}^d)$. Then, $f \star g \in C(\mathbb{R}^d)$.*

**Proof.** Let $x, h \in \mathbb{R}^d$. We have,

$$\begin{aligned} |(f \star g)(x + h) - (f \star g)(x)| &\leq \int_\Omega |f(y)g(x + h + y) - f(y)g(x + y)| \,\mathrm{d}y \\ &= \int_\Omega |f(y)| \, |g(x + h + y) - g(x + y)| \,\mathrm{d}y \\ &\overset{\text{Hölder}}{\leq} \left( \int_\Omega |f(y)|^p \,\mathrm{d}y \right)^{\frac{1}{p}} \left( \int_\Omega |g(x + h + y) - g(x + y)|^q \,\mathrm{d}y \right)^{\frac{1}{q}} \\ &\leq \|f\|_{L^p(\Omega)} \|T_h g - g\|_{L^q(\mathbb{R}^d)}. \end{aligned}$$

Here, $T_h g$ is the *translation operator* with $h$, i.e. $T_h g(x) = g(x + h)$. $\|T_h g - g\|_{L^q}$ converges to $0$ for $h \to 0$ (exercise) and with this the claim follows. $\square$

**Definition 2.6.** Let $\psi$ be as in Proposition 2.3 (iii), i.e. $\psi \in L^1(\mathbb{R}^d)$ with $\psi \geq 0$ and $\int_{\mathbb{R}^d} \psi(x) \,\mathrm{d}x = 1$. If additionally $\psi \in C_c^\infty(\mathbb{R}^d)$, then $\psi$ is called *(positive) mollifier.*

**Proposition 2.7.** *Let $\Omega \subset \mathbb{R}^d$ be open, $1 \leq p < \infty$ and $\psi$ be a mollifier. For all $f \in L^p(\Omega)$ and all $\delta > 0$, there is $\epsilon > 0$ such that*

$$\|\psi_\epsilon \star f - f\|_{L^p(\Omega)} < \delta.$$

*Moreover, $C_c^\infty(\Omega)$ is dense in $L^p(\Omega)$.*

**Proof.** Since $C_c(\Omega)$ is dense in $L^p(\Omega)$ (cf. Proposition B.8), there is a $g \in C_c(\Omega)$ with $\|f-g\|_{L^p(\Omega)} < \frac{\delta}{3}$. Then, extending $f, g$ to $\mathbb{R}^d$ with 0, using the Minkowski inequality (cf. Remark B.4) and Proposition 2.3 (i) with $r = p$ and $q = 1$, we get

$$\begin{aligned}
\|\psi_\epsilon \star f - f\|_{L^p(\Omega)} &\leq \|\psi_\epsilon \star f - \psi_\epsilon \star g\|_{L^p(\Omega)} + \|\psi_\epsilon \star g - g\|_{L^p(\Omega)} + \|g-f\|_{L^p(\Omega)} \\
&< \|(f-g) \star \psi_\epsilon\|_{L^p(\mathbb{R}^d)} + \|\psi_\epsilon \star g - g\|_{L^p(\Omega)} + \frac{\delta}{3} \\
&\leq \|f-g\|_{L^p(\mathbb{R}^d)} \|\psi_\epsilon\|_{L^1(\mathbb{R}^d)} + \|\psi_\epsilon \star g - g\|_{L^p(\Omega)} + \frac{\delta}{3} \\
&< \|\psi_\epsilon \star g - g\|_{L^p(\Omega)} + \frac{2\delta}{3}.
\end{aligned}$$

Since $g \in C_c(\Omega)$ and we extended $g$ with zero, we get $g \in C_c(\mathbb{R}^d)$. Moreover, there is a $r > 0$ such that $K := \overline{B_r(0) + \operatorname{supp}(g)} \subset \Omega$ where $A + B = \{a + b \,:\, a \in A \wedge b \in B\}$ and supp is the closed support (cf. Definition B.1). Since $\psi$ has compact support and by construction of $\psi_\epsilon$, for all $\epsilon > 0$ small enough, we have $\operatorname{supp}(\psi_\epsilon) \subset B_r(0)$ and thus

$$\operatorname{supp}(\psi_\epsilon \star g) \subset \overline{\operatorname{supp}(\psi_\epsilon) + \operatorname{supp}(g)} \subset K \subset \Omega.$$

Since $g$ is continuous and has compact support, it is also uniformly continuous and bounded. Thus, by Proposition 2.3 (iii), $\psi_\epsilon \star g$ converges uniformly to $g$ on $K$. Moreover, $\psi_\epsilon \star g$ and $g$ are zero on $\mathbb{R}^d \setminus K$. Thus, there is $\epsilon > 0$ such that $\|\psi_\epsilon \star g - g\|_{L^\infty(K)} < \frac{\delta}{3\|1\|_{L^p(K)}}$. With this, we get

$$\|\psi_\epsilon \star g - g\|_{L^p(\Omega)} = \|\psi_\epsilon \star g - g\|_{L^p(K)} \leq \|\psi_\epsilon \star g - g\|_{L^\infty(K)} \|1\|_{L^p(K)} < \frac{\delta}{3},$$

which shows the first statement. Moreover, we have

$$\|\psi_\epsilon \star g - f\|_{L^p(\Omega)} \leq \|\psi_\epsilon \star g - g\|_{L^p(\Omega)} + \|g - f\|_{L^p(\Omega)} < \frac{2\delta}{3} < \delta.$$

By Proposition 2.3 (ii), we get $\psi_\epsilon \star g \in C^\infty(\mathbb{R}^d)$. Finally, as shown above, we can choose $\epsilon > 0$ so small that additionally $\operatorname{supp}(\psi_\epsilon \star g) \subset \Omega$. Thus, we have $\psi_\epsilon \star g \in C_c^\infty(\Omega)$, which shows the claimed denseness. $\square$

*Remark* 2.8 (Discrete filter). If $F = (f_i)_{i \in \mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ is a signal and $W = (w_{-1}, w_0, w_1) \in \mathbb{R}^3$, we consider the weighted sum

$$\hat{f}_i = \sum_{j=-1}^{1} w_j f_{i+j}.$$

This corresponds to the discrete correlation of $F$ and $W$ in 1D and also holds accordingly for more general filters $W = (w_{-a}, \ldots, w_a) \in \mathbb{R}^{2a+1}$ for $a \in \mathbb{N}_0$. However, one has to take into account that discrete signals $F = (f_i)_{i \in \{1,\ldots,m\}}$ only have finitely many components, but the discrete filter needs $f_{i+j}$ for all $i \in \{1, \ldots, m\}$ and $j \in \{-1, 0, 1\}$ or $j \in \{-a, \ldots, a\}$, respectively. There are different ways to define $f_i$ for $i \in \mathbb{Z} \setminus \{1, \ldots, m\}$:

- Zero extension (or zero padding), i. e. $f_i = 0$ for $i \in \mathbb{Z} \setminus \{1, \ldots, m\}$:

$$\ldots, 0, 0, 0, f_1, f_2, \ldots, f_{m-1}, f_m, 0, 0, 0, \ldots$$

- Extension by mirroring at the boundary, i. e. $f_i = f_{2m-i}$, $i \in \{m+1, \dots, 2m-1\}$ (analogously for the left boundary):

$$\dots, f_4, f_3, f_2, f_1, f_2, \dots, f_{m-1}, f_m, f_{m-1}, f_{m-2}, f_{m-3}, \dots$$

- Periodic extension, i. e. $f_i = f_{((i-1) \bmod m)+1}$ for $i > m$ (analogously for $i < 1$):

$$\dots, f_{m-2}, f_{m-1}, f_m, f_1, f_2, \dots, f_{m-1}, f_m, f_1, f_2, f_3, \dots$$

- Constant extension, i. e. $f_i = f_1$ for $i < 1$ and $f_i = f_m$ for $i > m$:

$$\dots, f_1, f_1, f_1, f_1, f_2, \dots, f_{m-1}, f_m, f_m, f_m, f_m, \dots$$

This can be extended to discrete images in 2D in a straightforward manner:

**Definition 2.9** (Discrete cross-correlation and convolution)**.** If $F = (f_{i,j})_{(i,j)\in\mathcal{I}}$ is a discrete image and $W = (w_{k,l})_{(k,l)\in\mathcal{J}}$ with $\mathcal{J} := \big\{(k,l) \in \mathbb{Z}^2 \,:\, -a \le k \le a, -b \le l \le b\big\}$ for $a, b \in \mathbb{N}_0$, the *discrete cross-correlation* is defined as

$$(W \star F)_{i,j} = \sum_{k=-a}^{a} \sum_{l=-b}^{b} w_{k,l} f_{i+k,j+l} \text{ for } (i,j) \in \mathcal{I},$$

where $f_{i+k,j+l}$ for $(i+k, j+l) \notin \mathcal{I}$ is defined according to one of the extensions discussed above. $W$ is called *(filter) kernel* or *(filter) mask* of size $(2a+1) \times (2b+1)$. The mapping

$$(W * F)_{i,j} = \sum_{k=-a}^{a} \sum_{l=-b}^{b} w_{k,l} f_{i-k,j-l} \text{ for } (i,j) \in \mathcal{I},$$

is called *discrete convolution.*

When explicitly specifying all kernel values, we use the notation

$$W = \begin{bmatrix} w_{-a,b} & \cdots & w_{a,b} \\ \vdots & w_{0,0} & \vdots \\ w_{-a,-b} & \cdots & w_{a,-b} \end{bmatrix},$$

which is following the indexing from the figure in Definition 1.1.

*Remark* 2.10. $(W \star F)$ is linear in $F$. Thus, filtering with $W$ can be expressed as matrix. To this end, $F$ has to be interpreted as vector, e. g. using *reflected lexicographical* order (lexicographical order: $(a_1, a_2) < (b_1, b_2)$ if $a_1 < b_1$ or $a_1 = b_1 \wedge a_2 < b_2$), i. e.

$$F = \left(\begin{array}{c} f_{1,1} \\ \vdots \\ f_{m_1,1} \\ \hline f_{1,2} \\ \vdots \\ f_{m_1,2} \\ \hline \vdots \\ \hline f_{1,m_2} \\ \vdots \\ f_{m_1,m_2} \end{array}\right) = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_{m_2} \end{pmatrix}, \text{ where } F_j = \begin{pmatrix} f_{1,j} \\ \vdots \\ f_{m_1,j} \end{pmatrix}.$$

For $W = \begin{bmatrix} w_{-1,1} & w_{0,1} & w_{1,1} \\ w_{-1,0} & w_{0,0} & w_{1,0} \\ w_{-1,-1} & w_{0,-1} & w_{1,-1} \end{bmatrix}$ and zero padding, we get

$$(W \star F) = \begin{pmatrix} W_0 & W_1 & & \\ W_{-1} & W_0 & W_1 & \\ & \ddots & \ddots & \ddots \\ & & W_{-1} & W_0 \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_{m_2} \end{pmatrix}, \text{ where } W_j = \begin{pmatrix} w_{0,j} & w_{1,j} & & \\ w_{-1,j} & w_{0,j} & w_{1,j} & \\ & \ddots & \ddots & \ddots \\ & & w_{-1,j} & w_{0,j} \end{pmatrix}.$$

*Remark* 2.11 (Denoising filters).

(i) Mean value filters: The moving average over squares $M_r^\infty$ can be expressed straightforwardly as discrete filter. For an edge length of the square of 3 or 5 pixels, we get the filter kernels

$$M_3 = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ and } M_5 = \frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

of size $3 \times 3$ and $5 \times 5$, respectively. In general, for a filter size $a \times b$, the corresponding mean value filter is

$$M_{a,b} = \frac{1}{ab}\underbrace{(1, \dots, 1)}_{\in \mathbb{R}^a} \otimes \underbrace{(1, \dots, 1)}_{\in \mathbb{R}^b} \in \mathbb{R}^{a \times b}.$$

Here, $\otimes : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{m \times n}, ((x_i)_i, (y_j)_j) \mapsto (x_i y_j)_{ij}$ denotes the *tensor product.*

Let $f$ be a pixelated image (with zero padding) and $F$ the corresponding discrete image on a cartesian grid with grid size 1. Then, for $a \in \mathbb{N}$, we have

$$M^\infty_{a+\frac{1}{2}} f(x^{\underline{j}}) = (M_{2a+1,2a+1} \star F)_{\underline{j}}.$$

Further mean value filter examples:

$$\frac{1}{5}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \frac{1}{10}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Mean value filters fulfill $\sum_{k,l} w_{k,l} = 1$.

(ii) Gaussian filter: From Proposition 2.3, we know that $\psi \star f$ is at least as smooth as the kernel $\psi$. Thus, for smoothing, the *Gaussian filter* with filter width $\sigma > 0$

$$g_\sigma(x) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \exp\left(\frac{-\|x\|^2}{2\sigma^2}\right)$$

suggests itself, since $g_\sigma \in C^\infty(\mathbb{R}^d)$. The simplest way to discretize this (continuous) filter is to first evaluate $g_\sigma$ on a discrete grid and then to normalize the result, i. e.

$$(\tilde{G}_\sigma)_{k,l} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^2} \exp\left(\frac{-(k^2 + l^2)}{2\sigma^2}\right), \qquad G_\sigma = \frac{1}{\sum_{k,l}(\tilde{G}_\sigma)_{k,l}}\tilde{G}_\sigma.$$

This discretization assumes a cartesian grid with grid size 1. Otherwise, $(k^2 + l^2)$ has to be scaled corresponding to the actual grid size.

(iii) Binomial filter: We consider the following construction:

$$
\begin{aligned}
B^0 &= (1), \quad W = \frac{1}{2}(1,1) \\
B^1 &= W \star (0, B^0, 0) = \frac{1}{2}(1,1) \\
B^2 &= W \star (0, B^1, 0) = \frac{1}{4}(1,2,1) \\
B^3 &= W \star (0, B^2, 0) = \frac{1}{8}(1,3,3,1) \\
B^4 &= W \star (0, B^3, 0) = \frac{1}{16}(1,4,6,4,1) \\
&\vdots \\
B^n &= W \star (0, B^{n-1}, 0) \in \mathbb{R}^{n+1}
\end{aligned}
$$

Here, we define the correlation of $f \in \mathbb{R}^n$ with $W = (w_0, w_1)$, a filter of size 2, as

$$(W \star f)_{i=1}^{n-1} := \sum_{k=0}^{1} w_k f_{i+k}.$$

The resulting vectors match the normalized rows of Pascal's triangle. Thus, the corresponding filter is called *binomial filter*. The binomial filter of size $a \times b$ is

$$B_{a,b} = B^{a-1} \otimes B^{b-1}.$$

In particular, for size $3 \times 3$ and $5 \times 5$, we get

$$B_3 = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \text{ and } B_5 = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}.$$

For large $n$, the binomial filters are a good approximation of the Gaussian filter.

(iv) Duto blur: The idea of the *duto blur* is to overlay an image with a smoothed version of itself, i. e. one considers

$$\lambda F + (1 - \lambda)(G_\sigma \star F)$$

for $\lambda \in [0, 1]$. The result seems to preserve the sharpness of the input image, while still being blurred.

(v) Median filter: Let $a \in \mathbb{N}$. The mean value filter is the solution of a minimization problem:

$$(M_{2a+1} \star F)_{i,j} = \frac{1}{(2a+1)^2} \sum_{k=-a}^{a} \sum_{l=-a}^{a} f_{i+k,j+l} = \operatorname*{argmin}_{f \in \mathbb{R}} \sum_{k=-a}^{a} \sum_{l=-a}^{a} |f_{i+k,j+l} - f|^2$$

This follows from the necessary condition

$$0 \stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}f}\left(\sum_{k=-a}^{a}\sum_{l=-a}^{a}|f_{i+k,j+l}-f|^2\right) = 2\sum_{k=-a}^{a}\sum_{l=-a}^{a}(f-f_{i+k,j+l})$$

and the fact that the second derivative of the objective function is constant and positive. Due to the square, the minimization problem is susceptible to outliers in $F$. This can be fixed by replacing the square with the absolute value. Thus, one has to solve a minimization problem of type

$$\min_{g\in\mathbb{R}}\sum_{i=1}^{n}|g_i-g|\,,$$

where $G=(g_i)_i\in\mathbb{R}^n$ and $n$ odd. The minimizer is known to be the median of the values $g_1,\dots,g_n$ and can be computed by sorting the vector $G$: If $\tilde{G}$ contains the entries of $G$ in ascending order, i.e. $\tilde{g}_1\le\ldots\le\tilde{g}_n$, the median is the element in the middle, i.e. $\tilde{g}_{\frac{n+1}{2}}$ (in case $n$ is even, the median is not unique).

Thus, the *median filter* of an image $F$ is defined as the computation of the median over windows of size $(2a+1)\times(2a+1)$. This is a nonlinear filter that is well suited to remove outlier pixels, e.g. salt-and-pepper noise.

If, instead of the median, the minimum or maximum is considered, the corresponding filters are called *minimum filter* and *maximum filter*. These three nonlinear filters are all referred to as *rank filters*.

(vi) Bilateral filter: Using the substitution rule for multiple variables, the application of a linear filter can be expressed as follows:

$$(\psi\star f)(x)=\int_{\mathbb{R}^d}\psi(y-x)f(y)\,\mathrm{d}y$$

Thus, the weight at the position $y$ only depends on the distance of $x$ to $y$. In particular, the difference of the gray values $f(x)$ and $f(y)$ does not matter. Thus, close to an edge in image intensity, gray values on both sides of the edge are averaged leading to the blurring of edges typical for linear filters. This can be prevented if the gray value distance is taken into account:

$$B_{\psi_1,\psi_2}f(x)=\int_{\mathbb{R}^d}\psi_1(y-x)\psi_2(f(y)-f(x))f(y)\,\mathrm{d}y\Big/\int_{\mathbb{R}^d}\psi_1(y-x)\psi_2(f(y)-f(x))\,\mathrm{d}y$$

Here, $\psi_1:\mathbb{R}^d\to\mathbb{R}$ and $\psi_2:\mathbb{R}\to\mathbb{R}$ are weight functions. The denominator is necessary to normalize the adaptive weighting. $B_{\psi_1,\psi_2}$ is called *bilateral filter*. If Gaussian kernels are used as weight functions, the filter is also called *nonlinear* or *selective Gaussian filter*.

This filter can be discretized similarly to the correlation:

$$\sum_{k=-a}^{a}\sum_{l=-b}^{b}\psi_1(k,l)\psi_2(f_{i+k,j+l}-f_{i,j})f_{i+k,j+l}\Big/\sum_{k=-a}^{a}\sum_{l=-b}^{b}\psi_1(k,l)\psi_2(f_{i+k,j+l}-f_{i,j})$$

Compared to linear filters, the computation is more costly, since the normalization depends on the input image and has be computed for every pixel. Note that it is not necessary to extend the image to apply this filter. Instead, in case $(i+k,j+l)\notin\mathcal{I}$, the corresponding weight can be set to 0 due the adaptive normalization.

*Remark* 2.12 (Derivative filters). Finite difference based derivatives can be expressed as linear filters. Let $F$ be a discrete image on a cartesian grid. Then, the forward difference quotient at the position $i, j$ in direction $x_1$ is given by

$$\frac{1}{h}\left(F_{i+1,j} - F_{i,j}\right).$$

This corresponds to filtering with the kernel

$$D^+_{x_1} = \frac{1}{h}\begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Correspondingly, the backward difference quotient and the quotients in direction $x_2$ are

$$D^-_{x_1} = \frac{1}{h}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D^+_{x_2} = \frac{1}{h}\begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D^-_{x_2} = \frac{1}{h}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix}.$$

Also the central difference quotient can be expressed as linear filter:

$$D^c_{x_1} = \frac{1}{2h}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad D^c_{x_2} = \frac{1}{2h}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}.$$

The same is true for higher derivatives:

$$D^2_{x_1} = \frac{1}{h^2}\begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad D^2_{x_2} = \frac{1}{h^2}\begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

For derivative filters, we have $\sum_{k,l} w_{k,l} = 0$, whereas averaging filters fulfill $\sum_{k,l} w_{k,l} = 1$.

Note that the rows / columns containing only zeros are not required to represent the difference quotients as linear filters. They were just added to have square filter kernels. Omitting these zeros, leads to non-square filter kernels, e. g.

$$D^+_{x_1} = \frac{1}{h}\begin{bmatrix} 0 & -1 & 1 \end{bmatrix} \text{ and } D^+_{x_2} = \frac{1}{h}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

The difference quotients do not benefit from the properties of the correlation (Proposition 2.3), but have the problems typical for finite differences. For instance, consider the oscillating signal

$$\ldots, -1, 1, -1, 1, -1, 1, \ldots$$

The central difference quotient of this signal is zero everywhere, but the signal is not constant. Instead of using difference quotients, it is better to make use of the properties from Proposition 2.3.

*Remark* 2.13 (Prewitt and Sobel edge detector). Despite their limitations, difference quotients can be used for basic edge detection. To compensate for the problems of difference quotients, they are combined with averaging. More precisely, a difference quotient in one coordinate direction is combined with averaging in the other coordinate direction. If the mean value filter is used for the averaging, this leads to the *Prewitt kernels*

$$D_{x_1}^{\text{Prewitt}} = \frac{1}{2h}(-1,0,1) \otimes \frac{1}{3}(1,1,1) = \frac{1}{6h}\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix},$$

$$D_{x_2}^{\text{Prewitt}} = \frac{1}{3}(1,1,1) \otimes \frac{1}{2h}(-1,0,1) = \frac{1}{6h}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}.$$

If the binomial filter $B^2 = \frac{1}{4}(1,2,1)$ is used for the averaging instead, this leads to the *Sobel kernels*

$$D_{x_1}^{\text{Sobel}} = \frac{1}{2h}(-1,0,1) \otimes \frac{1}{4}(1,2,1) = \frac{1}{8h}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix},$$

$$D_{x_2}^{\text{Sobel}} = \frac{1}{4}(1,2,1) \otimes \frac{1}{2h}(-1,0,1) = \frac{1}{8h}\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}.$$

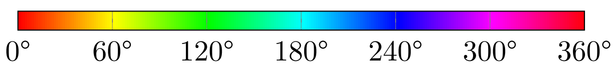
Let $F$ be a discrete image, $G_{x_1} = D_{x_1}^{\text{Prewitt}} \star F$ and $G_{x_2} = D_{x_2}^{\text{Prewitt}} \star F$. Then, $\sqrt{G_{x_1}^2 + G_{x_2}^2}$ is the gradient magnitude, i. e. a simple edge detector. Here, the squares and the square root are computed pixel-wise. Moreover, $\Theta = \operatorname{arctan2}(G_{x_1}, G_{x_2})$ are the gradient directions. Here, arctan2 is the *2-argument arctangent*, i. e.

$$\operatorname{arctan2} : \mathbb{R}^2 \setminus \{0\} \to (-\pi,\pi], (a,b) \mapsto \begin{cases} \arctan\left(\frac{b}{a}\right) & a > 0 \\ \arctan\left(\frac{b}{a}\right) + \pi & a < 0,\, b \geq 0 \\ \arctan\left(\frac{b}{a}\right) - \pi & a < 0,\, b < 0 \\ \frac{\pi}{2} & a = 0,\, b > 0 \\ -\frac{\pi}{2} & a = 0,\, b < 0. \end{cases}$$

The same is true if the Sobel kernels are used instead of the Prewitt kernels.

*Remark* 2.14. Consider an angle image $\Theta$, i. e. a discrete image with $V = [-\pi,\pi]$. While the value range technically is an interval, it intends to represent the unit circle $S^1 = \left\{x \in \mathbb{R}^2 : \|x\|_2 = 1\right\}$, i. e. an angle $\alpha$ encodes the corresponding vector $(\cos(\alpha), \sin(\alpha))$. In particular, two end points of the interval, $-\pi$ and $\pi$, are equivalent. When visualizing an angle image, this should be taken into account in order to ensure that angles are visualized similarly if they represent unit vectors that are close to each other. This would not be the case if the interval $[-\pi,\pi]$ is simply mapped to gray values similarly to how we handle values in $[0,1]$ for grayscale images. By its construction, the HSV color model offers a solution for this visualization problem.

Like RGB, HSV consist of three channels, hue (H), saturation (S) and value (V). Unlike RGB, the geometry is not cubic, but cylindrical. The hue is the shade of the color and specified as angle. The following color bar illustrates how angles are mapped to color shades:



The saturation encodes the "colorfulness" and is in $[0,1]$, where zero means colorless, i. e. white, one means full color. Finally, the value encodes the brightness and is also in $[0,1]$, where zero means completely dark, i. e. black, one means full brightness.

In order to visualize $\Theta$, we specify the visualization in HSV space, setting $H = \Theta$ (after mapping $[-\pi,\pi]$ to $[0,2\pi]$ with $\theta \mapsto (\theta+2\pi) \bmod (2\pi)$). Setting further $S = V = 1$, we get a full color, full brightness visualization of the angle image $\Theta$. Note that to display an image given in HSV space, it still needs to be converted to RGB. This is done by the mapping

$$[0,2\pi]\times[0,1]\times[0,1] \to [0,1]^3, (H,S,V) \mapsto (R,G,B) := \begin{cases} (V,t,p), & i \in \{0,6\} \\ (q,V,p), & i = 1 \\ (p,V,t), & i = 2 \\ (p,q,V), & i = 3 \\ (t,p,V), & i = 4 \\ (V,p,q), & i = 5 \end{cases},$$

where $i = \lfloor H/\frac{\pi}{3} \rfloor$, $f = \left(H/\frac{\pi}{3} - i\right)$, $p = V\cdot(1-S)$, $q = V\cdot(1-S\cdot f)$, $t := V\cdot(1-S\cdot(1-f))$.

When used to visualize gradient directions, this approach ignores the gradient magnitude. It is very easy to include this though by using the gradient as saturation, i. e. by setting $S = T^{\text{norm}}(\sqrt{G_{x_1}^2 + G_{x_2}^2})$. This way, areas with small gradient are white. If one prefers have those areas to be black instead, one can additionally has to set $V = T^{\text{norm}}(\sqrt{G_{x_1}^2 + G_{x_2}^2})$.

*Remark* 2.15. A linear filter can be used as feature extractor. As such, linear filters play an integral role in convolutional neural networks (CNNs). In a convolutional layer, multiple filters of small size, e. g. $3\times 3$ or $7\times 7$, are applied to the input data to get a multi-channel output, one channel for each filter. Only the number of filters and their size is prescribed in such a CNN, the weights (filter values) are learned by the network from training data, which creates feature extractors suitable for the desired task, e. g. classification, segmentation, etc.

*Remark* 2.16 (Canny edge detector). Edges in images or signals are positions at which the value changes significantly. First, we consider the 1D case, i. e. signals. For a smooth signal $f$, one can define the position of an edge as the position where the slope $f'$ is locally maximal. In particular, at an edge, we then also have $f'' = 0$.

In general, a signal $f$ is not smooth, but Proposition 2.3 (ii) ensures that $\psi \star f \in C^2$ and $(\psi\star f)' = -\psi'\star f$ for a suitable kernel $\psi$. Assuming some regularity of $f$ (e. g. Proposition 2.3 (iii)), with $\psi_\sigma(x) = \frac{1}{\sigma}\psi(\frac{x}{\sigma})$ we have

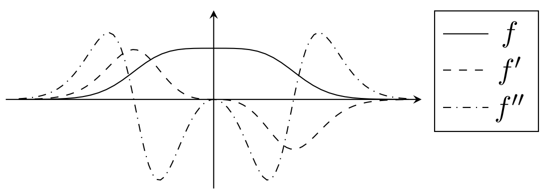
$$\lim_{\sigma\to 0}(\psi_\sigma \star f)(x) = f(x).$$

This gives us a whole family of smoothed images

$$u(x,\sigma) = (\psi_\sigma \star f)(x).$$

Edges exist on different scales, in the sense of rather "coarse" or rather "fine" edges. Thus, a kernel $\psi_\sigma$ would be desirable for which the amount of edges in $u(\cdot,\sigma)$ scales with $\sigma$. The larger $\sigma$, the fewer edges. In particular, increasing $\sigma$ should never add new edges. Moreover, the position of the edges should not move with $\sigma$. This leads to the following conditions

$$\begin{aligned} \partial_x^2 u(x,\sigma) > 0 &\Rightarrow \partial_\sigma u(x,\sigma) > 0, \\ \partial_x^2 u(x,\sigma) = 0 &\Rightarrow \partial_\sigma u(x,\sigma) = 0, \\ \partial_x^2 u(x,\sigma) < 0 &\Rightarrow \partial_\sigma u(x,\sigma) < 0. \end{aligned}$$



Solutions of the partial differential equation (PDE)

$$\partial_x^2 u(x,\sigma) = \partial_\sigma u(x,\sigma)$$

fulfill the three properties. In addition to the PDE, we obviously also need

$$u(x,0) = f(x).$$

This initial value problem also occurs in different contexts and models the distribution of heat. Thus, the PDE is called *heat equation*. The unique solution is the convolution with the Gaussian kernel, i. e.

$$u(x,\sigma) = (g_{\sqrt{2\sigma}} * f)(x).$$

The particular "speed" in $\sigma$ is not important. Thus, also

$$u(x, \tfrac{\sigma^2}{2}) = (g_\sigma * f)(x)$$

is suitable. Thus, $\psi_\sigma = g_\sigma$ is a suitable kernel for edge detection.

In case $d > 1$, $\partial_x^2 u$ is generalized with the *Laplace operator*

$$\Delta_x u(x,\sigma) = \sum_{i=1}^{d} \partial_{x_i}^2 u(x,\sigma).$$

The convolution with $g_{\sqrt{2\sigma}}$ is a solution of the heat equation for all $d \in \mathbb{N}$ (exercise). This convolution can also be expressed as correlation:

$$(g_{\sqrt{2\sigma}} * f)(x) = \int_{\mathbb{R}^d} g_{\sqrt{2\sigma}}(y) f(x-y)\,\mathrm{d}y \overset{\phi(y)=-y}{=} \int_{\mathbb{R}^d} g_{\sqrt{2\sigma}}(-y) f(x+y)\,\mathrm{d}y = (g_{\sqrt{2\sigma}} \star f)(x).$$

The *Canny edge detector* now works as follows: For a given $\sigma > 0$, size and direction of the gradient of $g_\sigma \star f$ are computed, i. e.

$$\rho(x) = \|\nabla(g_\sigma \star f)(x)\|_2 = \sqrt{(\partial_{x_1}(g_\sigma \star f)(x))^2 + (\partial_{x_2}(g_\sigma \star f)(x))^2}$$
$$\theta(x) = \operatorname{arctan2}\left(\partial_{x_1}(g_\sigma \star f)(x), \partial_{x_2}(g_\sigma \star f)(x)\right).$$

Here, the derivatives are computed using

$$\partial_{x_i}(g_\sigma \star f)(x) = (-\partial_{x_i} g_\sigma \star f)(x).$$

It is sufficient to precompute the corresponding derivative filters and then to filter the image with those.

As edges, we consider all points $x$ at which $\rho(x)$ has a strict local maximum in direction $(\sin(\theta(x)), \cos(\theta(x)))$. Discretely (assuming a cartesian grid), $\theta$ is rounded to the grid directions $0°, 45°, 90°, \ldots, 315°$ and one checks whether $\rho$ is strictly bigger at the pixel $x$ than at the two neighboring pixels in direction $\theta$ and $180° + \theta$. Finally, all edges are disregarded for which $\rho(x)$ is smaller than a specified threshold.

*Remark* 2.17 (Laplacian sharpening). We want to sharpen an unsharp image. First, we consider the 1D case. Sharpening means emphasizing edges, which are zeros of the second derivative. If one subtracts from a signal $f$ its second derivative $f''$, the edges in $f$ are amplified (cf. construction in Remark 2.16). Thus, we consider

$$f - \tau f''$$

for $\tau > 0$. The bigger $\tau$, the more the edges are amplified. In the general case, edges occur in different directions. Thus, we need a rotationally invariant classification of the second derivative. The simplest approach is the Laplace operator. Since images are usually not smooth enough, we again use the Gaussian filter and obtain

$$g_\sigma \star f - \tau\Delta(g_\sigma \star f) = (g_\sigma - \tau\Delta g_\sigma) \star f.$$

This operation is called *Laplacian sharpening*.

## 2.1. Morphological filters

*Example* 2.18 (Denoising of objects). Let $A \subset \Omega$ be a noisy object that we want to denoise. Noisy here means that there are small, isolated areas in $A$ that don't belong to the original object.

So far, we have considered image denoising. Thus, it seems to suggest itself to consider the characteristic function $\chi_A$ instead of the object itself. This way, we could apply our image denoising techniques, but would get an image whose values are not necessarily in $\{0, 1\}$ anymore. Thus, we use a different strategy.

If $x$ belongs to a noise area, we have $\chi_A(x) = 1$, but in the "proximity" of $x$, there are points $y$ with $\chi_A(y) = 0$. Therefore, we consider a so-called structuring element $B \subset \mathbb{R}^d$ that is slightly larger than the individual noise areas (and usually contains 0). Then, the function

$$f(x) = \begin{cases} 1 & \text{if } \chi_A(x+y) = 1 \text{ for all } y \in B \\ 0 & \text{else} \end{cases}$$

is free of the noisy areas. However, it describes a "thinner" object than the sought one since this operation also removes the boundary areas of the object. This is (approximately) undone by the function

$$g(x) = \begin{cases} 1 & \text{if there is a } y \in B \text{ with } f(x-y) = 1 \\ 0 & \text{else} \end{cases},$$

but noisy areas that are not removed completely in the first step are amplified again. Note that the minus-sign in $g$ is relevant, in case $B$ is not symmetric about 0. Overall, this is a simple but quite viable method for object denoising.

The two operation we have used are called erosion and dilation:

**Definition 2.19.** Let $f : \mathbb{R}^d \to \{0, 1\}$ and $B \subset \mathbb{R}^d$ nonempty. Then,

$$(f \oplus B)(x) = \begin{cases} 1 & \text{if there is } y \in B \text{ with } f(x+y) = 1 \\ 0 & \text{else} \end{cases}$$

is called *dilation* of $f$ with the *structuring element* $B$.

$$(f \ominus B)(x) = \begin{cases} 1 & \text{if } f(x+y) = 1 \text{ for all } y \in B \\ 0 & \text{else} \end{cases}$$

is called *erosion* of $f$ with the structuring element $B$.

Extending this definition to grayscale images requires a different viewpoint:

**Lemma 2.20.** *Let $f : \mathbb{R}^d \to \{0,1\}$ and $B \subset \mathbb{R}^d$ nonempty. Then,*

$$(f \oplus B)(x) = \sup_{y \in B} f(x+y) \text{ and } (f \ominus B)(x) = \inf_{y \in B} f(x+y).$$

**Proof.** Follows immediately from the definition of $\oplus$ and $\ominus$. □

This allows us to define $\oplus$ and $\ominus$ also for grayscale images:

**Definition 2.21.** Let $f : \mathbb{R}^d \to [0,1]$ be an image and $B \subset \mathbb{R}^d$ nonempty. Then,

$$(f \oplus B)(x) = \sup_{y \in B} f(x+y)$$

is called *dilation* of the image $f$ with the structuring element $B$.

$$(f \ominus B)(x) = \inf_{y \in B} f(x+y)$$

is called *erosion* of the image $f$ with the structuring element $B$.

Erosion and dilation of an image coincide with the minimum and maximum filter mentioned in Remark 2.11. There, we only introduced these filters in the discrete setting with a rectangle as structuring element.

As we have seen in Example 2.18, it may be helpful to combine $\oplus$ and $\ominus$. The two canonical combinations have special names:

**Definition 2.22.** Let $f : \mathbb{R}^d \to [0,1]$ be an image and $B \subset \mathbb{R}^d$ nonempty. Moreover, let

$$-B \coloneqq \{-y \,:\, y \in B\}.$$

Then, the operations

$$f \circ B = (f \ominus B) \oplus (-B) \text{ and } f \bullet B = (f \oplus B) \ominus (-B)$$

are called *opening* and *closing* of $f$ with $B$.

*Example* 2.23 (Segmentation with background equalization). The morphological operations can be used as pre-processing step to improve the results of the isodata algorithm from Example 1.13.

If $f$ is an image with bright writing on a dark background, $f \circ B$ is an approximation of the background (the erosion removes the thin, bright writing, the dilation corrects the shrinkage caused by the erosion). If one subtracts the estimated background from the image, i. e. considers $f - f \circ B$, brightness variations in the background and small "smudges" are compensated. In case of the blackboard image, the result is improved noticeably if the isodata algorithm is applied to $f - f \circ B$ instead of $f$. $f - f \circ B$ is called *white-top-hat transform of $f$*. The analog for closing is $f \bullet B - f$ and called *black-top-hat transform.*

In case of dark writing on a bright background, either the black-top-hat transform is used or the gray values are inverted before the white-top-hat transform is applied. This connection, i. e.

$$f \bullet B - f = -((-f) \circ B) - f = (-f) - ((-f) \circ B),$$

follows from the so-called duality of opening and closing. In addition to this, there are more useful properties.

*Remark* 2.24 (Properties of the morphological operators). Let $f, g : \mathbb{R}^d \to [0,1]$ and $B \subset \mathbb{R}^d$ nonempty. $\ominus$ and $\oplus$ have the following properties (cf. [1, Theorem 3.29]):

$$\begin{aligned} \text{duality} \quad & -(f \oplus B) = (-f) \ominus B \\ \text{translational invariance} \quad & (T_h f) \ominus B = T_h(f \ominus B) \\ & (T_h f) \oplus B = T_h(f \oplus B) \\ \text{monotonicity} \quad & f \leq g \Rightarrow f \ominus B \leq g \ominus B \wedge f \oplus B \leq g \oplus B \\ \text{distributivity} \quad & (f \wedge g) \ominus B = (f \ominus B) \wedge (g \ominus B) \\ & (f \vee g) \oplus B = (f \oplus B) \vee (g \oplus B). \end{aligned}$$

Here, $\vee$ and $\wedge$ denote the pointwise minimum and maximum, respectively, i. e.

$$(f \vee g)(x) \coloneqq \max(f(x), g(x)) \text{ and } (f \wedge g)(x) \coloneqq \min(f(x), g(x))$$

Note that the symbols $\vee$ and $\wedge$ are motivated by the following: For $f = \chi_A$, $g = \chi_B$ with $A, B \subset \mathbb{R}^d$, we have $f \vee g = \chi_{A \cup B}$ and $f \wedge g = \chi_{A \cap B}$.

Except for distributivity, $\circ$ and $\bullet$ also have these properties (cf. [1, Theorem 3.33]):

$$\begin{aligned} \text{duality} \quad & -(f \bullet B) = (-f) \circ B \\ \text{translational invariance} \quad & (T_h f) \circ B = T_h(f \circ B) \\ & (T_h f) \bullet B = T_h(f \bullet B) \\ \text{monotonicity} \quad & f \leq g \Rightarrow f \circ B \leq g \circ B \wedge f \bullet B \leq g \bullet B \end{aligned}$$

Moreover, we have

$$\begin{aligned} \text{non-increasingness} \quad & f \circ B \leq f \\ \text{non-decreasingness} \quad & f \leq f \bullet B \\ \text{idempotence} \quad & (f \circ B) \circ B = f \circ B \\ & (f \bullet B) \bullet B = f \bullet B. \end{aligned}$$

*Remark* 2.25. As Lemma 2.26 will show, the distributivity of $\ominus$ and $\oplus$ still holds when we consider an infinite number of functions. Moreover, dilation and erosion are the only translational invariant operators on binary images that fulfill this generalized distributivity. This will be formalized and shown in Proposition 2.27.

**Lemma 2.26** (Infinite distributive law for erosion and dilation)**.** *Let $(f_i)_{i \in I} \subset F(\mathbb{R}^d, [0,1])$ be a set of images indexed by a set $I$ and $B \subset \mathbb{R}^d$ nonempty. Then,*

$$\Big(\bigvee\nolimits_i f_i\Big) \oplus B = \bigvee\nolimits_i (f_i \oplus B) \textit{ and } \Big(\bigwedge\nolimits_i f_i\Big) \ominus B = \bigwedge\nolimits_i (f_i \ominus B),$$

*where*

$$\left(\bigvee_i f_i\right)(x) := \sup_i f_i(x) \text{ and } \left(\bigwedge_i f_i\right)(x) := \inf_i f_i(x)$$

*Here,* $F(A,B) := \{f : A \to B\}$ *denotes the set of all mappings from* $A$ *to* $B$.

*Proof.* Let $x \in \mathbb{R}^d$. Then,

$$\begin{aligned}\left(\left(\bigvee_i f_i\right) \oplus B\right)(x) &= \sup_{y\in B}\left(\bigvee_i f_i\right)(x+y) = \sup_{y\in B}\sup_i f_i(x+y) = \sup_i \sup_{y\in B} f_i(x+y)\\ &= \sup_i (f_i \oplus B)(x) = \left(\bigvee_i (f_i \oplus B)\right)(x).\end{aligned}$$

The second statement is shown analogously since also infimums can be interchanged. □

**Proposition 2.27.** *Let* $D$ *be a translational invariant operator on binary images, i. e.*

$$D : F(\mathbb{R}^d, \{0,1\}) \to F(\mathbb{R}^d, \{0,1\})$$

*with* $D(T_h f) = T_h D(f)$ *for all* $h \in \mathbb{R}^d$ *and* $f \in F(\mathbb{R}^d, \{0,1\})$. *If* $D(0) = 0$, $D(1) \neq 0$ *and*

$$D\left(\bigvee_i f_i\right) = \bigvee_i (D(f_i))$$

*for all sets of images* $(f_i)_{i\in I} \subset F(\mathbb{R}^d, \{0,1\})$, *then there is a nonempty set* $B \subset \mathbb{R}^d$ *such that*

$$D(f) = f \oplus B \text{ for all } f \in F(\mathbb{R}^d, \{0,1\}).$$

*If* $D(1) = 1$, $D(0) \neq 1$ *and*

$$D\left(\bigwedge_i f_i\right) = \bigwedge_i (D(f_i))$$

*for all sets of images* $(f_i)_{i\in I} \subset F(\mathbb{R}^d, \{0,1\})$, *then there is a nonempty set* $B \subset \mathbb{R}^d$ *such that*

$$D(f) = f \ominus B \text{ for all } f \in F(\mathbb{R}^d, \{0,1\}).$$

*Proof.* Let $f : \mathbb{R}^d \to \{0,1\}$ and $F = \{f = 1\}$. Then, $f = \chi_F$. Moreover, $F = \bigcup_{y\in F}\{y\}$ and thus $f = \bigvee_{y\in F} \chi_{\{y\}}$, if we additionally assume $F \neq \emptyset$[2], i. e. $f \not\equiv 0$. Combined with the assumed distributivity of $D$, we get

$$D(f) = D\left(\bigvee_{y\in F} \chi_{\{y\}}\right) = \bigvee_{y\in F}\left(D(\chi_{\{y\}})\right).$$

Noting that

$$\chi_{\{y\}}(x) = \begin{cases} 1 & x \in \{y\} \\ 0 & \text{else}\end{cases} = \begin{cases} 1 & x - y \in \{0\} \\ 0 & \text{else}\end{cases} = \chi_{\{0\}}(x-y) = (T_{-y}\chi_{\{0\}})(x)$$

for all $x \in \mathbb{R}^d$, the translational invariance of $D$ implies

$$D(\chi_{\{y\}}) = D(T_{-y}\chi_{\{0\}}) = T_{-y}D(\chi_{\{0\}}).$$

[2]This assumptions is necessary since the supremum over the empty set is $-\infty$.

For $x \in \mathbb{R}^d$, the above leads to

$$\begin{aligned} D(f)(x) &= \bigvee_{y \in F} \left(T_{-y} D(\chi_{\{0\}})\right)(x) \\ &= \begin{cases} 1 & \exists y \in \mathbb{R}^d : y \in F \wedge T_{-y} D(\chi_{\{0\}})(x) = 1 \\ 0 & \text{else} \end{cases} \\ &= \begin{cases} 1 & \exists y \in \mathbb{R}^d : f(y) = 1 \wedge D(\chi_{\{0\}})(x-y) = 1 \\ 0 & \text{else} \end{cases} \\ &\overset{(z=y-x)}{=} \begin{cases} 1 & \exists z \in \mathbb{R}^d : f(x+z) = 1 \wedge D(\chi_{\{0\}})(-z) = 1 \\ 0 & \text{else} \end{cases}. \end{aligned} \tag{$*$}$$

For $B := \{D(\chi_{\{0\}})(-\cdot) = 1\}$, we have $D(\chi_{\{0\}})(-z) = 1 \Leftrightarrow z \in B$. The assumption $D(1) \neq 0$ together with $(*)$ ensures that $B$ is nonempty. Recalling the definition of the dilation of a binary image, i. e.

$$(f \oplus B)(x) = \begin{cases} 1 & \exists y \in B : f(x+y) = 1 \\ 0 & \text{else} \end{cases}$$

we have shown $D(f) = f \oplus B$ for $f \not\equiv 0$. For $f \equiv 0$, the statement follows from the assumption $D(0) = 0$, i. e. $D(0) = 0 = 0 \oplus B$. Thus, we have shown the first claim. The second claim can be shown with the first, using the complement mapping $f \mapsto 1 - f$ (exercise). □

**Proposition 2.28** (Contrast invariance of erosion and dilation)**.** *Let $B \subset \mathbb{R}^d$ be nonempty. Then, for all continuous, increasing intensity transformations $T : \mathbb{R} \to \mathbb{R}$ and all $f : \mathbb{R}^d \to [0,1]$,*

$$T(f) \ominus B = T(f \ominus B) \text{ and } T(f) \oplus B = T(f \oplus B)$$

*Proof.* For $x \in \mathbb{R}^d$, we have

$$(T(f) \ominus B)(x) = \inf_{y \in B} T(f(x+y)) = T(\inf_{y \in B} f(x+y)) = T((f \ominus B)(x)),$$

where we have used that continuous, increasing functions can be interchanged with the infimum (exercise). The same is true for the supremum, which shows the claim for the dilation. □

*Remark* 2.29*.* Dilation and erosion of grayscale images with a structuring element can be further generalized: For this purpose, the "flat" structuring element is replaced by a function. Let $f : \mathbb{R}^d \to [0,1]$, $B \subset \mathbb{R}^d$ nonempty and $b : B \to \mathbb{R}$. Then, dilation and erosion of the image $f$ with $b$ are defined as

$$(f \oplus b)(x) := \sup_{y \in B} \left(f(x+y) + b(y)\right) \text{ and } (f \ominus b)(x) := \inf_{y \in B} \left(f(x+y) - b(y)\right).$$

With the constant function $b \equiv 0$, these are equivalent to dilatation and erosion with $B$. The function $b$ can be thought of as a nonflat structuring element that changes the graph of $f$ before the infimum or supremum is evaluated.

# 3. The frequency domain

Goal of this section is to construct a so-called *frequency domain representation* of images. There, low frequencies should describe the coarse structure of the image, while high frequencies should describe fine details like edges or textures. Such an alternative representation opens a very different way to modify an image or to determine its properties. Moreover, this approach leads to methods that are global operators.

## 3.1. The Fourier transform

**Definition 3.1.** Let $f \in L^1(\mathbb{R}^d, \mathbb{C})$ and $\omega \in \mathbb{R}^d$. Then,

$$(\mathcal{F}f)(\omega) := \hat{f}(\omega) := \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\omega} \,\mathrm{d}x$$

is called *Fourier transform* of $f$ in $\omega$. The mapping $\mathcal{F} : f \mapsto \mathcal{F}f$ is called *(continuous) Fourier transform.*

**Attention:** There are multiple definitions of the Fourier transform in the literature. Both the factor in front of the integral and the scaling of the argument of the exponential function vary.

**Proposition 3.2.** *For the Fourier transform $\mathcal{F}$, we have*

$$\mathcal{F} : L^1(\mathbb{R}^d, \mathbb{C}) \to C(\mathbb{R}^d, \mathbb{C}).$$

*Moreover, this mapping is linear and continuous.*

**Proof.** Let $(\omega_n) \subset \mathbb{R}^d$ be a sequence with $\lim\limits_{n\to\infty} \omega_n = \omega \in \mathbb{R}^d$. First we note that, for $t \in \mathbb{R}$,

$$\left|e^{it}\right| = |\cos(t) + i\sin(t)| = \sqrt{\cos^2(t) + \sin^2(t)} = 1.$$

For $g_n(x) = f(x)e^{-ix\cdot\omega_n}$, we have

$$|g_n(x)| \le |f(x)|$$

and $\lim\limits_{n\to\infty} g_n(x) = f(x)e^{-ix\cdot\omega}$. Thus, the dominated convergence theorem (Theorem B.11) implies

$$\lim_{n\to\infty} (\mathcal{F}f)(\omega_n) = \lim_{n\to\infty} \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x)e^{-ix\cdot\omega_n} \,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x)e^{-ix\cdot\omega} \,\mathrm{d}x = (\mathcal{F}f)(\omega).$$

Therefore, $\mathcal{F}f \in C(\mathbb{R}^d, \mathbb{C})$. The linearity of $\mathcal{F}$ follows from the linearity of the integral. For the continuity of $\mathcal{F}$, it suffices to show

$$\|\mathcal{F}f\|_{C(\mathbb{R}^d,\mathbb{C})} \le C \|f\|_{L^1(\mathbb{R}^d,\mathbb{C})},$$

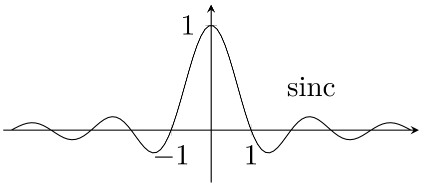
since continuity and boundedness are equivalent for linear operators (exercise). For an arbitrary $\omega \in \mathbb{R}^d$, we have

$$|(\mathcal{F}f)(\omega)| \le \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \left|f(x)e^{-ix\cdot\omega}\right| \mathrm{d}x \le \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} |f(x)| \,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{d}{2}}} \|f\|_{L^1}.$$

Thus, $\|\mathcal{F}f\|_{C(\mathbb{R}^d,\mathbb{C})} = \sup\limits_{\omega\in\mathbb{R}^d} |(\mathcal{F}f)(\omega)| \le \frac{1}{(2\pi)^{\frac{d}{2}}} \|f\|_{L^1}$, which shows the continuity of $\mathcal{F}$. □

*Example* 3.3. For $B > 0$, the Fourier transform of the characteristic function $\chi_{[-B,B]}$ can be computed analytically (exercise). We have

$$(\mathcal{F}\chi_{[-B,B]})(\omega) = \sqrt{\tfrac{2}{\pi}} B \operatorname{sinc}\left(\tfrac{B\omega}{\pi}\right).$$




Here

$$\operatorname{sinc} : \mathbb{R} \to \mathbb{R}, x \mapsto \begin{cases} \frac{\sin(\pi x)}{\pi x} & x \neq 0 \\ 1 & x = 0 \end{cases}$$

is the *(normalized) cardinal sine function.* In particular, this shows that the Fourier transform of a function with compact support does not necessarily have compact support. Moreover, sinc is not even integrable. For $n \in \mathbb{N}$, we have

$$\begin{aligned} \int_0^n |\operatorname{sinc}(x)| \,\mathrm{d}x &= \sum_{k=0}^n \int_k^{k+1} \left|\frac{\sin(\pi x)}{\pi x}\right| \mathrm{d}x \geq \sum_{k=0}^n \int_k^{k+1} \left|\frac{\sin(\pi x)}{\pi(k+1)}\right| \mathrm{d}x \\ &= \frac{1}{\pi} \sum_{k=0}^n \frac{1}{(k+1)} \int_k^{k+1} |\sin(\pi x)| \,\mathrm{d}x = \frac{1}{\pi} \sum_{k=0}^n \frac{1}{(k+1)} \frac{2}{\pi}, \end{aligned}$$

where we have used

$$\begin{aligned} \int_k^{k+1} |\sin(\pi x)| \,\mathrm{d}x &= \left| \int_k^{k+1} \sin(\pi x) \,\mathrm{d}x \right| = \left| -\frac{1}{\pi} \cos(\pi(k+1)) - \left( -\frac{1}{\pi} \cos(\pi k) \right) \right| \\ &= \left| -\frac{1}{\pi} (-1)^{k+1} - \left( -\frac{1}{\pi} (-1)^k \right) \right| = \left| \frac{2}{\pi} (-1)^{k+2} \right| = \frac{2}{\pi}. \end{aligned}$$

Since the harmonic series $\sum_{k=1}^\infty \frac{1}{k}$ diverges, this implies that also $\int_{\mathbb{R}} |\operatorname{sinc}(x)| \,\mathrm{d}x$ diverges. Thus, $\operatorname{sinc} \notin L^1(\mathbb{R})$, which shows that $\mathcal{F}$ does not map $L^1(\mathbb{R}, \mathbb{C})$ to itself.

To make the Fourier transform viable for the purposes of image processing, we need to be able to transform functions back from the frequency domain, i. e. we need the inverse transform. This requires numerous properties that are collected in the following. We will even need to introduce a new space (the Schwartz space, see Definition 3.11) to understand in which setting the Fourier is bijective.

**Lemma 3.4.** *Let $f \in L^1(\mathbb{R}^d, \mathbb{C})$, $y \in \mathbb{R}^d$ and $A \in GL(d) := \{A \in \mathbb{R}^{d \times d} : \det A \neq 0\}$. Then,*

$$\begin{aligned} \mathcal{F}(T_y f) &= M_y(\mathcal{F} f), \\ \mathcal{F}(M_y f) &= T_{-y}(\mathcal{F} f), \\ \mathcal{F}(D_A f) &= \frac{1}{|\det A|} D_{A^{-T}}(\mathcal{F} f), \\ \mathcal{F}(\overline{f}) &= \overline{D_{-\mathbb{1}}(\mathcal{F} f)}. \end{aligned}$$

*Here, $\mathbb{1}$ is the identity matrix,*

$$M_y : L^1(\mathbb{R}^d, \mathbb{C}) \to L^1(\mathbb{R}^d, \mathbb{C}), f \mapsto m_y f$$

*the* modulation *of $f$ with the function $m_y : x \mapsto e^{ix \cdot y}$ and*

$$D_A : L^1(\mathbb{R}^d, \mathbb{C}) \to L^1(\mathbb{R}^d, \mathbb{C}), f \mapsto (x \mapsto f(Ax))$$

*the* linear coordinate transformation *of $f$ with $A$.*

**Proof.** The expressions are well-defined, since $T_y$, $M_y$ and $D_A$ map both $L^1(\mathbb{R}^d, \mathbb{C})$ and $C(\mathbb{R}^d, \mathbb{C})$ to itself. For $h, y \in \mathbb{R}^d$, we have

$$\begin{aligned}(\mathcal{F} M_y T_h f)(\omega) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x+h) e^{-ix\cdot(\omega-y)} \, \mathrm{d}x \overset{z=x+h}{=} \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(z) e^{-i(z-h)\cdot(\omega-y)} \, \mathrm{d}z \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{ih\cdot(\omega-y)} \int_{\mathbb{R}^d} f(z) e^{-iz\cdot(\omega-y)} \, \mathrm{d}z = e^{i(\omega-y)\cdot h} (\mathcal{F}f)(\omega-y) \\ &= m_h(\omega-y)(\mathcal{F}f)(\omega-y) = (M_h \mathcal{F} f)(\omega-y) = (T_{-y} M_h \mathcal{F} f)(\omega).\end{aligned}$$

Since $M_0 = T_0 = \mathrm{id}$, we get the first two statements using $y = 0$ and $h = 0$, respectively. The third statement follows readily from the substitution rule for multiple variables (exercise). The last statement follows from

$$\begin{aligned} e^{-ix\cdot y} &= \cos(-x\cdot y) + i\sin(-x\cdot y) = \overline{\cos(-x\cdot y) - i\sin(-x\cdot y)} \\ &= \overline{\cos(x\cdot y) + i\sin(x\cdot y)} = \overline{e^{ix\cdot y}} = \overline{e^{-ix\cdot(-\mathbb{1}\, y)}} \end{aligned}$$

and

$$(\mathcal{F}\overline{f})(\omega) = \overline{\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) \overline{e^{-ix\cdot\omega}} \, \mathrm{d}x}. \qquad \square$$

**Definition 3.5.** Let $f : \mathbb{R}^d \to \mathbb{C}$. Then, $f$ is called *Hermitian*, if

$$\overline{f(\omega)} = f(-\omega) \text{ for all } \omega \in \mathbb{R}^d.$$

Moreover, $f$ is called *skew-Hermitian*, if

$$\overline{f(\omega)} = -f(-\omega) \text{ for all } \omega \in \mathbb{R}^d.$$

**Corollary 3.6.** *For $f \in L^1(\mathbb{R}^d, \mathbb{C})$, we have*

$$\begin{aligned} f \text{ real-valued } &\Leftrightarrow \mathcal{F}f \text{ is Hermitian} \\ f \text{ imaginary-valued } &\Leftrightarrow \mathcal{F}f \text{ is skew-Hermitian.} \end{aligned}$$

**Proof.** Both statements are a consequence of the identity $\mathcal{F}(\overline{f}) = \overline{D_{-\mathbb{1}}(\mathcal{F}f)}$. First statement:

"⇒" If $f$ is real-valued, i. e. $f = \overline{f}$, we get

$$\overline{\mathcal{F}f} = \overline{\mathcal{F}(\overline{f})} = \overline{\overline{D_{-\mathbb{1}}(\mathcal{F}f)}} = D_{-\mathbb{1}}(\mathcal{F}f).$$

"⇐" We need to show $f = \overline{f}$. Since $\mathcal{F}$ is linear and injective (the injectivity will be shown later, cf. Corollary 3.23), it is sufficient to show that $\mathcal{F}f - \mathcal{F}(\overline{f}) = 0$. This follows from

$$\mathcal{F}f - \mathcal{F}(\overline{f}) = \mathcal{F}f - \overline{D_{-\mathbb{1}}(\mathcal{F}f)} \overset{\mathcal{F}f \text{ Hermitian}}{=} \mathcal{F}f - \overline{\overline{\mathcal{F}f}} = 0.$$

The second statement is shown analogously using that $f$ imaginary-valued means $f = -\overline{f}$. $\square$

**Theorem 3.7** (Convolution theorem)**.** *For $f, g \in L^1(\mathbb{R}^d, \mathbb{C})$, we have*

$$\mathcal{F}(f * g) = (2\pi)^{\frac{d}{2}} \mathcal{F}(f)\mathcal{F}(g),$$

*i. e. the Fourier transform converts a convolution to a pointwise multiplication.*

**Proof.** The statement can be shown with Fubini's theorem and the substitution rule for multiple variables:

$$\begin{aligned}
\mathcal{F}(f * g)(\omega) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(y)g(x-y)\,\mathrm{d}y\, e^{-ix\cdot\omega}\,\mathrm{d}x \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(y)e^{-iy\cdot\omega} g(x-y)e^{-i(x-y)\cdot\omega}\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(y)e^{-iy\cdot\omega} \int_{\mathbb{R}^d} g(x-y)e^{-i(x-y)\cdot\omega}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} f(y)e^{-iy\cdot\omega}(\mathcal{F}g)(\omega)\,\mathrm{d}y = (2\pi)^{\frac{d}{2}}(\mathcal{F}f)(\omega)(\mathcal{F}g)(\omega).
\end{aligned}$$

□

**Lemma 3.8.** *For $f, g \in L^1(\mathbb{R}^d, \mathbb{C})$, we have*

$$\int_{\mathbb{R}^d} (\mathcal{F}f)(x)g(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(x)(\mathcal{F}g)(x)\,\mathrm{d}x\,.$$

**Proof.** The statement can be shown (similarly to Theorem 3.7) with Fubini's theorem (exercise). □

*Remark* 3.9. Like the convolution, the cross-correlation (Definition 2.2) can be generalized to complex valued functions. Here, the integrand is just changed slightly: For measurable $\psi : \Omega \to \mathbb{C}$ and $f : \mathbb{R}^d \to \mathbb{C}$, we define

$$(\psi \star f)(x) := \int_{\Omega} \overline{\psi(y)} f(x+y)\,\mathrm{d}y$$

as cross-correlation of $f$ and $\psi$ at position $x \in \mathbb{R}^d$. Note that this definition is consistent with the definition for real-valued functions, since complex conjugation is the identity on real numbers. The cross-correlation properties from Proposition 2.3 also hold in the complex case. Just the behavior when interchanging the arguments is slightly different for complex valued functions: For $\Omega = \mathbb{R}^d$, we get

$$\begin{aligned}
\overline{(\psi \star f)(x)} &= \int_{\mathbb{R}^d} \psi(y)\overline{f(x+y)}\,\mathrm{d}y \overset{\phi(y)=x+y}{=} \int_{\mathbb{R}^d} \psi(y-x)\overline{f(y)}\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} \overline{f(y)}\psi(-x+y)\,\mathrm{d}y = (f \star \psi)(-x).
\end{aligned}$$

**Corollary 3.10.** *For $f, g \in L^1(\mathbb{R}^d, \mathbb{C})$, we have*

$$\mathcal{F}(f \star g) = (2\pi)^{\frac{d}{2}} \overline{\mathcal{F}(f)}\mathcal{F}(g).$$

**Proof.** Using

$$(f \star g)(x) = \int_{\mathbb{R}^d} \overline{f(y)} g(x+y)\,\mathrm{d}y \overset{\phi(y)=-y}{=} \int_{\mathbb{R}^d} \overline{f(-y)} g(x-y)\,\mathrm{d}y = ((D_{-\mathbb{1}}\overline{f}) * g)(x),$$

the convolution theorem and Lemma 3.4 imply

$$\begin{aligned}\mathcal{F}(f \star g) &= \mathcal{F}((D_{-\mathbb{1}}\overline{f}) * g) = (2\pi)^{\frac{d}{2}} \mathcal{F}(D_{-\mathbb{1}}\overline{f}) \mathcal{F}(g) = (2\pi)^{\frac{d}{2}} (D_{-\mathbb{1}} \mathcal{F}(\overline{f})) \mathcal{F}(g) \\ &= (2\pi)^{\frac{d}{2}} \overline{\mathcal{F}(f)} \mathcal{F}(g).\end{aligned}$$

□

**Definition 3.11.** The *Schwartz space*, also called the *space of rapidly decreasing functions on* $\mathbb{R}^d$, is defined as

$$\mathcal{S}(\mathbb{R}^d, \mathbb{C}) := \left\{ f \in C^\infty(\mathbb{R}^d, \mathbb{C}) \,:\, C_{\alpha,\beta}(f) := \sup_{x \in \mathbb{R}^d} \left| x^\alpha \frac{\partial^\beta}{\partial x^\beta} f(x) \right| < \infty \text{ for all } \alpha, \beta \in \mathbb{N}_0^d \right\}.$$

The elements of $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$ are called *Schwartz functions*.

*Remark* 3.12. By construction, we have

$$C_c^\infty(\mathbb{R}^d, \mathbb{C}) \subset \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset C^\infty(\mathbb{R}^d, \mathbb{C}).$$

One can think of Schwartz functions as smooth functions that (including their derivatives) tend to zero faster than polynomials tend to infinity. They do not necessarily have compact support though. An example for a Schwartz function without compact support is the Gaussian kernel $g_\sigma$. To show this, one first shows that any derivative of $g_\sigma$ is the product of a polynomial and $g_\sigma$. Combined with the fact that $g_\sigma$ tends to zero faster than polynomials tend to infinity, this shows $g_\sigma \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ (exercise).

**Lemma 3.13.** *Let* $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$.

- $\frac{\partial^\gamma}{\partial x^\gamma} f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ *for all* $\gamma \in \mathbb{N}_0^d$.
- $pf \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ *for all polynomials* $p : \mathbb{R}^d \to \mathbb{C}$.
- $fg \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ *for all* $g \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$.

*Proof.* Let $\gamma \in \mathbb{N}_0^d$. We immediately get $\frac{\partial^\gamma}{\partial x^\gamma} f \in C^\infty(\mathbb{R}^d, \mathbb{C})$ since $\mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset C^\infty(\mathbb{R}^d, \mathbb{C})$. Let $\alpha, \beta \in \mathbb{N}_0^d$ be arbitrary, but fixed. Then,

$$C_{\alpha,\beta}\left( \frac{\partial^\gamma}{\partial x^\gamma} f \right) = \sup_{x \in \mathbb{R}^d} \left| x^\alpha \frac{\partial^\beta}{\partial x^\beta} \frac{\partial^\gamma}{\partial x^\gamma} f(x) \right| = C_{\alpha,\beta+\gamma}(f) < \infty,$$

which shows $\frac{\partial^\gamma}{\partial x^\gamma} f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$.

Let $p : \mathbb{R}^d \to \mathbb{C}$ be a polynomial. Then, $p$ can be expressed as weighted sum of monomials, i. e. there is a $k \in \mathbb{N}_0$ and coefficients $p_\gamma \in \mathbb{C}$ for all multi-indices with $|\gamma| \leq k$ such that $p(x) = \sum_{|\gamma| \leq k} p_\gamma x^\gamma$ for all $x \in \mathbb{R}$. Thus, for $\alpha, \beta \in \mathbb{N}_0^d$, we get

$$\begin{aligned}C_{\alpha,\beta}(pf) &= \sup_{x \in \mathbb{R}^d} \left| x^\alpha \frac{\partial^\beta}{\partial x^\beta} \left( \sum_{|\gamma| \leq k} p_\gamma x^\gamma f(x) \right) \right| = \sup_{x \in \mathbb{R}^d} \left| \sum_{|\gamma| \leq k} p_\gamma x^\alpha \frac{\partial^\beta}{\partial x^\beta} (x^\gamma f(x)) \right| \\ &\leq \sup_{x \in \mathbb{R}^d} \sum_{|\gamma| \leq k} |p_\gamma| \left| x^\alpha \frac{\partial^\beta}{\partial x^\beta} (x^\gamma f(x)) \right| \leq \sum_{|\gamma| \leq k} |p_\gamma| \sup_{x \in \mathbb{R}^d} \left| x^\alpha \frac{\partial^\beta}{\partial x^\beta} (x^\gamma f(x)) \right| \\ &= \sum_{|\gamma| \leq k} |p_\gamma| \, C_{\alpha,\beta}(q^\gamma f),\end{aligned}$$

where $q^\gamma$ is the monomial $q^\gamma(x) = x^\gamma$. This means to show $pf \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ it is sufficient to show that $qf \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ for all monomials $q$. This is follows from the triangle inequality and the general Leibniz rule for multivariable calculus, i. e.

$$\partial^\alpha(fg) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} \partial^{\alpha-\beta} f \partial^\beta g.$$

Similarly, one shows $fg \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ for $g \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ (exercise). □

**Lemma 3.14.** *For $1 \leq p \leq \infty$, we have $\mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset L^p(\mathbb{R}^d, \mathbb{C})$. More specifically, for $p < \infty$ and $q \in \mathbb{N}$ with $2qp > d$, there is $C > 0$, such that for all $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$*

$$\|f\|_{L^p} \leq C \left\| (1 + \|\cdot\|_2^{2q}) f \right\|_{L^\infty} < \infty.$$

*Proof.* Let $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$. Then, $\|f\|_{L^\infty} = C_{0,0}(f) < \infty$, which shows the claim for $p = \infty$. Now consider $p < \infty$. Let $q \in \mathbb{N}$ such that $2qp > d$. Then, one can show that $g(x) := \frac{1}{1+\|x\|_2^{2q}} \in L^p(\mathbb{R}^d)$ (exercise, the rotational symmetry of $g$ allows using Proposition B.14). Then, we get

$$\int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x = \int_{\mathbb{R}^d} \left| (1 + \|x\|_2^{2q}) f(x) \right|^p \left( \frac{1}{1 + \|x\|_2^{2q}} \right)^p \mathrm{d}x \leq \left\| (1 + \|\cdot\|_2^{2q}) f \right\|_{L^\infty}^p \|g\|_{L^p}^p < \infty,$$

where we have used that $g \in L^p(\mathbb{R}^d)$ and $(1 + \|\cdot\|_2^{2q}) f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset L^\infty(\mathbb{R}^d, \mathbb{C})$. The latter follows since $1 + \|\cdot\|_2^{2q}$ is a polynomial. □

**Lemma 3.15.** *Let $f \in C^1(\mathbb{R}^d, \mathbb{C})$ such that $\partial_i f \in L^1(\mathbb{R}^d, \mathbb{C})$ for an $i \in \{1, \ldots, d\}$ and $\sup_{x \in \mathbb{R}^d} \left| f(x) \|x\|_2^{2d} \right| < \infty$. Then,*

$$\int_{\mathbb{R}^d} \partial_i f(x) \, \mathrm{d}x = 0.$$

*Proof.* WLOG $i = 1$ (else renumber the components of $\mathbb{R}^d$). For $r > 0$, we get using Fubini's theorem (cf. Theorem B.6)

$$\begin{aligned}
\int_{[-r,r]^d} \partial_1 f(x) \, \mathrm{d}x &= \int_{[-r,r]^{d-1}} \int_{-r}^{r} \partial_1 f(x) \, \mathrm{d}x_1 \, \mathrm{d}(x_2, \ldots, x_d) \\
&= \int_{[-r,r]^{d-1}} f(r, x_2, \ldots, x_d) - f(-r, x_2, \ldots, x_d) \mathrm{d}(x_2, \ldots, x_d)
\end{aligned}$$

By our assumptions, there is $C > 0$ such that

$$\left| f(x) \|x\|_2^{2d} \right| \leq C \text{ for all } x \in \mathbb{R}^d \;\Rightarrow\; |f(x)| \leq \frac{C}{\|x\|_2^{2d}} \text{ for all } x \in \mathbb{R}^d.$$

Combined, we get

$$\begin{aligned}
\left| \int_{[-r,r]^d} \partial_1 f(x) \, \mathrm{d}x \right| &\leq \int_{[-r,r]^{d-1}} \underbrace{|f(r, x_2, \ldots, x_d) - f(-r, x_2, \ldots, x_d)|}_{\leq \frac{C}{\|(r, x_2, \ldots, x_d)\|_2^{2d}} + \frac{C}{\|(-r, x_2, \ldots, x_d)\|_2^{2d}} \leq \frac{2C}{r^{2d}}} \mathrm{d}(x_2, \ldots, x_d) \\
&\leq (2r)^{d-1} \frac{2C}{r^{2d}} = \frac{2^d C}{r^{d+1}}
\end{aligned}$$

$$\Rightarrow \lim_{r\to\infty} \int_{[-r,r]^d} \partial_1 f(x)\,\mathrm{d}x = 0.$$

On the other hand, the sequence $g_n \coloneqq \chi_{[-n,n]^d}\partial_1 f(x)$ converges pointwise to $\partial_1 f(x)$ and fulfills $|g_n(x)| \leq |\partial_1 f(x)|$ for $x \in \mathbb{R}^d$. Since $\partial_1 f \in L^1(\mathbb{R}^d, \mathbb{C})$, the dominated convergence theorem (Theorem B.11) implies

$$\lim_{n\to\infty} \int_{[-n,n]^d} \partial_1 f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \partial_1 f(x)\,\mathrm{d}x,$$

which concludes the proof, since we have already shown that this limit is zero. $\square$

*Remark* 3.16. Lemma 3.15 is the basis for integration by parts on $\mathbb{R}^d$. Let $g, h \in C^1(\mathbb{R}^d, \mathbb{C})$ such that $f = gh$ fulfills the assumptions of Lemma 3.15. Then, we get

$$0 = \int_{\mathbb{R}^d} \partial_i (gh)(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \partial_i g(x)h(x) + g(x)\partial_i h(x)\,\mathrm{d}x\,.$$

If $\partial_i gh, g\partial_i h \in L^1(\mathbb{R}^d, \mathbb{C})$, we can split the integral on the right hand side and get

$$\int_{\mathbb{R}^d} \partial_i g(x)h(x)\,\mathrm{d}x = -\int_{\mathbb{R}^d} g(x)\partial_i h(x)\,\mathrm{d}x\,.$$

With sufficient assumptions on $g$ and $h$, we can iterate this to get integration by parts for the derivative with respect to a multi-index $\alpha$, i. e.

$$\int_{\mathbb{R}^d} \frac{\partial^\alpha}{\partial x^\alpha} g(x)h(x)\,\mathrm{d}x = (-1)^{|\alpha|} \int_{\mathbb{R}^d} g(x)\frac{\partial^\alpha}{\partial x^\alpha} h(x)\,\mathrm{d}x\,.$$

**Lemma 3.17.** *Let $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, $\alpha$ a multi-index and $p^\alpha(x) = x^\alpha$. Then, $\mathcal{F}(f) \in C^\infty(\mathbb{R}^d, \mathbb{C})$,*

$$\mathcal{F}\left(\frac{\partial^\alpha}{\partial x^\alpha} f\right) = i^{|\alpha|} p^\alpha \mathcal{F}(f) \text{ and } \mathcal{F}(p^\alpha f) = i^{|\alpha|}\frac{\partial^\alpha}{\partial \omega^\alpha}\mathcal{F}(f).$$

**Proof.** We have

$$\frac{\partial^\alpha}{\partial x^\alpha}\left(e^{-ix\cdot\omega}\right) = (-i)^{|\alpha|}\omega^\alpha e^{-ix\cdot\omega} \text{ and } \frac{\partial^\alpha}{\partial \omega^\alpha}\left(e^{-ix\cdot\omega}\right) = (-i)^{|\alpha|} x^\alpha e^{-ix\cdot\omega}.$$

Since $\frac{1}{i} = -i$, we get

$$x^\alpha e^{-ix\cdot\omega} = i^{|\alpha|}\frac{\partial^\alpha}{\partial \omega^\alpha}\left(e^{-ix\cdot\omega}\right).$$

Using integration by parts (here we use that $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, $\left|e^{-ix\cdot\omega}\right| = 1$ and $\left|\frac{\partial^\beta}{\partial x^\beta}\left(e^{-ix\cdot\omega}\right)\right| = \left|\omega^\beta\right|$ for all $\beta \in \mathbb{N}_0^d$ to show that the necessary assumptions are fulfilled), we get the first claimed equality:

$$\begin{aligned}\mathcal{F}\left(\frac{\partial^\alpha}{\partial x^\alpha} f\right)(\omega) &= \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} \frac{\partial^\alpha}{\partial x^\alpha} f(x) e^{-ix\cdot\omega}\,\mathrm{d}x = (-1)^{|\alpha|}\frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} f(x)\frac{\partial^\alpha}{\partial x^\alpha}\left(e^{-ix\cdot\omega}\right)\mathrm{d}x \\ &= i^{|\alpha|}\omega^\alpha \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} f(x)e^{-ix\cdot\omega}\,\mathrm{d}x = i^{|\alpha|}p^\alpha(\omega)(\mathcal{F}f)(\omega).\end{aligned}$$

Iteratively using Corollary B.12 (here we use that $f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C}) \subset L^1(\mathbb{R}^d,\mathbb{C})$ and again that $\left|\frac{\partial^\beta}{\partial x^\beta}\left(e^{-ix\cdot\omega}\right)\right| = \left|\omega^\beta\right|$ for all $\beta \in \mathbb{N}_0^d$), one shows the second claimed equality:

$$\begin{aligned}\mathcal{F}\left(p^\alpha f\right) &= \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} x^\alpha f(x) e^{-ix\cdot\omega}\,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{d}{2}}} i^{|\alpha|}\int_{\mathbb{R}^d} f(x)\frac{\partial^\alpha}{\partial\omega^\alpha}\left(e^{-ix\cdot\omega}\right)\mathrm{d}x \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} i^{|\alpha|}\frac{\partial^\alpha}{\partial\omega^\alpha}\int_{\mathbb{R}^d} f(x) e^{-ix\cdot\omega}\,\mathrm{d}x = i^{|\alpha|}\frac{\partial^\alpha}{\partial\omega^\alpha}(\mathcal{F}f)(\omega)\end{aligned}$$

In particular, we have shown that $\frac{\partial^\alpha}{\partial\omega^\alpha}(\mathcal{F}f)$ exists. Since $\alpha$ was arbitrary, this also shows $\mathcal{F}(f) \in C^\infty(\mathbb{R}^d,\mathbb{C})$. □

**Proposition 3.18.** *For $f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$, we have $\mathcal{F}f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$. In other words, the Fourier transform maps the Schwartz space to itself. Moreover, if $(f_n)_n \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})^\mathbb{N}$ is a sequence with*

$$\lim_{n\to\infty} C_{\alpha,\beta}(f_n) = 0 \text{ for all } \alpha,\beta \in \mathbb{N}_0^d, \text{ then also } \lim_{n\to\infty} C_{\alpha,\beta}(\mathcal{F}f_n) = 0 \text{ for all } \alpha,\beta \in \mathbb{N}_0^d.$$

**Proof.** Let $f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$. From Lemma 3.17, we already know $\mathcal{F}f \in C^\infty(\mathbb{R}^d,\mathbb{C})$. Now, let $\alpha,\beta \in \mathbb{N}_0^d$ be arbitrary but fixed. For any $\omega \in \mathbb{R}^d$, we get using the notation $p^\beta(x) = x^\beta$ and applying Lemma 3.17 twice

$$\left|\omega^\alpha \tfrac{\partial^\beta}{\partial\omega^\beta}(\mathcal{F}f)(\omega)\right| = \left|\omega^\alpha \mathcal{F}(p^\beta f)(\omega)\right| = \left|\mathcal{F}\left(\tfrac{\partial^\alpha}{\partial x^\alpha}(p^\beta f)\right)(\omega)\right| \le \tfrac{1}{(2\pi)^{\frac{d}{2}}}\left\|\tfrac{\partial^\alpha}{\partial x^\alpha}(p^\beta f)\right\|_{L^1}.$$

Taking the supremum over $\omega \in \mathbb{R}^d$ shows

$$C_{\alpha,\beta}(\mathcal{F}f) \le \tfrac{1}{(2\pi)^{\frac{d}{2}}}\left\|\tfrac{\partial^\alpha}{\partial x^\alpha}(p^\beta f)\right\|_{L^1} < \infty, \tag{$*$}$$

where we have used Lemmas 3.13 and 3.14 to show $\frac{\partial^\alpha}{\partial x^\alpha}(p^\beta f) \in \mathcal{S}(\mathbb{R}^d,\mathbb{C}) \subset L^1(\Omega,\mathbb{C})$, which shows that the $L^1$-norm is finite. Since $\alpha,\beta \in \mathbb{N}_0^d$ were arbitrary, we have shown $\mathcal{F}f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$.

Now, let $(f_n)_n \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})^\mathbb{N}$ be a sequence with $C_{\alpha,\beta}(f_n) \to 0$ for $n \to \infty$ for all $\alpha,\beta \in \mathbb{N}_0^d$. Let again $\alpha,\beta \in \mathbb{N}_0^d$ be arbitrary but fixed. Inserting $f_n$ in $(*)$ and using Lemma 3.14, there are $q \in \mathbb{N}$ and $C > 0$ such that

$$C_{\alpha,\beta}(\mathcal{F}f_n) \overset{(*)}{\le} \tfrac{1}{(2\pi)^{\frac{d}{2}}}\left\|\tfrac{\partial^\alpha}{\partial x^\alpha}(p^\beta f_n)\right\|_{L^1} \overset{\text{Lemma 3.14}}{\le} \tfrac{1}{(2\pi)^{\frac{d}{2}}} C\left\|(1+\|\cdot\|_2^{2q})\tfrac{\partial^\alpha}{\partial x^\alpha}(p^\beta f_n)\right\|_{L^\infty}$$

for all $n \in \mathbb{N}$. Since $C_{\gamma,\delta}(f_n) \to 0$ for $n \to \infty$ for all $\gamma,\delta \in \mathbb{N}_0^d$, $p^\gamma \frac{\partial^\delta}{\partial x^\delta} f_n$ converges uniformly to 0 for all $\gamma,\delta \in \mathbb{N}_0^d$. Using the general Leibniz rule for multivariable calculus and the triangle inequality, this implies that also $(1+\|\cdot\|_2^{2q})\frac{\partial^\alpha}{\partial x^\alpha}(p^\beta f_n)$ converges uniformly to 0. With the inequality from above, this shows $C_{\alpha,\beta}(\mathcal{F}f_n) \to 0$ for $n \to \infty$. □

*Remark* 3.19. The second statement of Proposition 3.18 means that

$$\mathcal{F} : \mathcal{S}(\mathbb{R}^d,\mathbb{C}) \to \mathcal{S}(\mathbb{R}^d,\mathbb{C})$$

is continuous with respect to a suitable topology on the Schwartz space. This topology is not defined by a norm on this space. Instead, one defines when sequences converge in this space: A sequence $(f_n)_n \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})^\mathbb{N}$ converges to $f \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$, if

$$C_{\alpha,\beta}(f_n - f) \to 0 \text{ for } n \to \infty \text{ for all } \alpha,\beta \in \mathbb{N}_0^d.$$

Due to the linearity of $\mathcal{F}$ (cf. Proposition 3.2), it is sufficient to show the continuity in 0. Thus, let $(f_n)_n \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})^{\mathbb{N}}$ converge to 0, i. e. $C_{\alpha,\beta}(f_n) = C_{\alpha,\beta}(f_n - 0) \to 0$ for $n \to \infty$ for all $\alpha, \beta \in \mathbb{N}_0^d$. By Proposition 3.18, we get $C_{\alpha,\beta}(\mathcal{F}f_n - 0) = C_{\alpha,\beta}(\mathcal{F}f_n) \to 0$ for $n \to \infty$ for all $\alpha, \beta \in \mathbb{N}_0^d$, which shows that $\mathcal{F}f_n$ converges to $0 = \mathcal{F}0$ in the Schwartz space, proving the continuity.

**Lemma 3.20.** *The Gaussian kernel with filter width 1*

$$g(x) \coloneqq g_1(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}\|x\|^2}$$

*is an eigenfunction of the Fourier transform for the eigenvalue 1, i. e. we have $\mathcal{F}g = g$.*

**Proof.** We denote the 1D variant of the Gaussian kernel by $g^1$. $g$ is known to be separable, i. e.

$$g(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right) = \frac{1}{(2\pi)^{\frac{d}{2}}} \prod_{k=1}^{d} \exp\left(-\frac{1}{2}x_k^2\right) = \prod_{k=1}^{d} g^1(x_k).$$

From this, we get using Fubini's theorem

$$\begin{aligned}(\mathcal{F}g)(\omega) =& \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \left(\prod_{k=1}^{d} g^1(x_k)\right) e^{-ix\cdot\omega} \,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \prod_{k=1}^{d} \left(g^1(x_k) e^{-ix_k\omega_k}\right) \mathrm{d}x \\ =& \prod_{k=1}^{d} (\mathcal{F}g^1)(\omega_k).\end{aligned}$$

Thus, it is sufficient to show $g^1(t) = (\mathcal{F}g^1)(t)$. Hence, let $d = 1$ and thus $g = g^1$. We have

$$g'(t) = -tg(t).$$

By applying the Fourier transform to the equation, it follows from Lemma 3.17 that

$$i\omega(\mathcal{F}g)(\omega) = -i(\mathcal{F}g)'(\omega) \Rightarrow (\mathcal{F}g)'(\omega) = -\omega(\mathcal{F}g)(\omega).$$

Moreover, we have

$$(\mathcal{F}g)(0) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{\mathbb{R}} g(x) e^{-ix\cdot 0} \,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{\mathbb{R}} g(x) \,\mathrm{d}x = \frac{1}{(2\pi)^{\frac{1}{2}}} = g(0).$$

Thus, $g$ and $\mathcal{F}g$ solve the same linear homogeneous differential equation of first order with the same initial condition. Since the solution of such ODEs is unique, $g$ and $\mathcal{F}g$ coincide. □

**Proposition 3.21** (Fourier inversion theorem). *Let $f \in L^1(\mathbb{R}^d, \mathbb{C})$ with $(\mathcal{F}f) \in L^1(\mathbb{R}^d, \mathbb{C})$. If $f$ is continuous in $x \in \mathbb{R}^d$, we have*

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} (\mathcal{F}f)(\omega) e^{ix\cdot\omega} \,\mathrm{d}\omega\,.$$

**Proof.** Noting that

$$\left|(\mathcal{F}f)(\omega)e^{-\frac{1}{2}\epsilon^2|\omega|^2+ix\cdot\omega}\right| = |(\mathcal{F}f)(\omega)| \left|e^{-\frac{1}{2}\epsilon^2|\omega|^2}\right| \left|e^{ix\cdot\omega}\right| \leq |(\mathcal{F}f)(\omega)|$$

and since by assumption $(\mathcal{F}f) \in L^1$, the dominated convergence theorem (Theorem B.11) with majorant $|\mathcal{F}f|$ implies

$$\frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\mathcal{F}f)(\omega)e^{ix\cdot\omega}\,\mathrm{d}\omega = \lim_{\epsilon\to 0}\frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\mathcal{F}f)(\omega)e^{-\frac{1}{2}\epsilon^2|\omega|^2+ix\cdot\omega}\,\mathrm{d}\omega\,.$$

Let $\psi(\omega) := \frac{1}{(2\pi)^{\frac{d}{2}}}e^{-\frac{1}{2}|\omega|^2}$ (Gaussian kernel with filter width 1). Then,

$$\frac{1}{(2\pi)^{\frac{d}{2}}}e^{-\frac{1}{2}\epsilon^2|\omega|^2+ix\cdot\omega} = (M_x D_{\epsilon\,\mathbb{1}}\psi)(\omega)$$

and using Lemmas 3.4 and 3.20, we get

$$\mathcal{F}(M_x D_{\epsilon\,\mathbb{1}}\psi) = T_{-x}(\mathcal{F}D_{\epsilon\,\mathbb{1}}\psi) = \epsilon^{-d}T_{-x}D_{\epsilon^{-1}\,\mathbb{1}}(\mathcal{F}\psi) \overset{\text{Lemma 3.20}}{=} \epsilon^{-d}T_{-x}D_{\epsilon^{-1}\,\mathbb{1}}(\psi).$$

$$\Rightarrow \mathcal{F}(M_x D_{\epsilon\,\mathbb{1}}\psi)(\omega) = \frac{1}{\epsilon^d}\psi\left(\frac{\omega-x}{\epsilon}\right) = \psi_\epsilon(x-\omega).$$

Using this and Lemma 3.8, we get

$$\begin{aligned}\frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\mathcal{F}f)(\omega)e^{-\frac{1}{2}\epsilon^2|\omega|^2+ix\cdot\omega}\,\mathrm{d}\omega &= \int_{\mathbb{R}^d}(\mathcal{F}f)(\omega)(M_x D_{\epsilon\,\mathbb{1}}\psi)(\omega)\,\mathrm{d}\omega \\ \overset{\text{Lemma 3.8}}{=} \int_{\mathbb{R}^d} f(\omega)(\mathcal{F}(M_x D_{\epsilon\,\mathbb{1}}\psi))(\omega)\,\mathrm{d}\omega &= \int_{\mathbb{R}^d} f(\omega)\psi_\epsilon(x-\omega)\,\mathrm{d}\omega \\ = (f*\psi_\epsilon)(x) \to f(x) \text{ for } \epsilon \to 0.\end{aligned}$$

Here, we used among other things $\int \psi\,\mathrm{d}x = 1$ and the continuity of $f$ in $x$. Different from Proposition B.10 (iii), one can make use of $\psi \in \mathcal{S}(\mathbb{R}^d,\mathbb{C})$ to go without the boundedness of $f$ (exercise). Finally, we get

$$\frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d}(\mathcal{F}f)(\omega)e^{ix\cdot\omega}\,\mathrm{d}\omega = f(x). \qquad \square$$

*Remark* 3.22. If the continuity assumption of $f$ in Proposition 3.21 is dropped, the inversion theorem holds for almost all $x$. The proof works analogously, just in the last step, instead of $(f*\psi_\epsilon)(x) \to f(x)$ for $\epsilon \to 0$, one uses $\|f*\psi_\epsilon - f\|_{L^1} \to 0$. For the latter, $f \in L^1(\mathbb{R}^d,\mathbb{C})$ is sufficient. Since $\|f*\psi_\epsilon - f\|_{L^1} \to 0$, a subsequence $f*\psi_{\epsilon_n}$ converges pointwise a.e. to $f$. Then, the claim follows from the uniqueness of the limit.

**Corollary 3.23.** $\mathcal{F} : L^1(\mathbb{R}^d,\mathbb{C}) \to C(\mathbb{R}^d,\mathbb{C})$ *is injective.*

**Proof.** Since $\mathcal{F}$ is linear, it is sufficient to show that $\mathcal{F}f = 0$ implies $f = 0$. Therefore, let $f \in L^1(\mathbb{R}^d,\mathbb{C})$ with $\mathcal{F}f = 0$. Thus, $\mathcal{F}f \in L^1(\mathbb{R}^d,\mathbb{C})$ and the inversion theorem immediately variant from Remark 3.22 implies $f = 0$ a.e. $\square$

**Corollary 3.24.** *For $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, it holds that*

$$(\mathcal{F}\mathcal{F}f)(x) = f(-x) \text{ for all } x \in \mathbb{R}^d.$$

**Proof.** By Proposition 3.18, we have $\mathcal{F}f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$. Since $\mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset L^1(\mathbb{R}^d, \mathbb{C})$, this implies $f, \mathcal{F}f \in L^1(\mathbb{R}^d, \mathbb{C})$. Moreover, $f$ is continuous, thus we can apply Proposition 3.21 and get

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} (\mathcal{F}f)(\omega) e^{ix\cdot\omega} \,\mathrm{d}\omega = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} (\mathcal{F}f)(\omega) e^{-i\omega\cdot(-x)} \,\mathrm{d}\omega = (\mathcal{F}\mathcal{F}f)(-x)$$

for all $x \in \mathbb{R}^d$, which shows the claim. □

**Proposition 3.25.** *The Fourier transform is a bijection from $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$ to itself. The inverse is $\mathcal{F}^{-1} = \mathcal{F}^3$ and fulfills*

$$(\mathcal{F}^{-1}f)(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(\omega) e^{ix\cdot\omega} \,\mathrm{d}\omega \text{ for all } x \in \mathbb{R}^d \text{ and } f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C}).$$

*Moreover,*

$$(f, g)_{L^2} = (\mathcal{F}f, \mathcal{F}g)_{L^2} \text{ for all } f, g \in \mathcal{S}(\mathbb{R}^d, \mathbb{C}).$$

**Proof.** For all $x \in \mathbb{R}^d$ and $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, Corollary 3.24 implies

$$(\mathcal{F}^4 f)(x) = \mathcal{F}^2 f(-x) = f(x) \text{ for all } x \in \mathbb{R}^d.$$

Thus, $\mathcal{F}^4$ is the identity on $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$. As shown below, this implies the bijectivity of $\mathcal{F}$ on $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$. For $f, g \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$ with $\mathcal{F}f = \mathcal{F}g$, we get

$$f = \mathcal{F}^3(\mathcal{F}f) = \mathcal{F}^3(\mathcal{F}g) = g,$$

which shows the injectivity (already known from Corollary 3.23). Moreover, for $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, we have $f = \mathcal{F}(\mathcal{F}^3 f)$, which shows the surjectivity.

Moreover, $\mathcal{F}^4 = \mathrm{id}_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})}$ implies $\mathcal{F}^{-1} = \mathcal{F}^3$. In particular, we get

$$(\mathcal{F}^{-1}f)(x) = (\mathcal{F}^3 f)(x) = (\mathcal{F}^2(\mathcal{F}f))(x) \overset{\text{Corollary 3.24}}{=} (\mathcal{F}f)(-x),$$

which is the claimed representation of the inverse. Now let $f, g \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$. Then,

$$\begin{aligned}(\mathcal{F}f, \mathcal{F}g)_{L^2} &= \int_{\mathbb{R}^d} (\mathcal{F}f)(x)\overline{(\mathcal{F}g)(x)} \,\mathrm{d}x \overset{\text{Lemma 3.8}}{=} \int_{\mathbb{R}^d} f(x)(\mathcal{F}\overline{(\mathcal{F}g)})(x) \,\mathrm{d}x \\ &\overset{\text{Lemma 3.4}}{=} \int_{\mathbb{R}^d} f(x)\overline{(\mathcal{F}(\mathcal{F}g))(-x)} \,\mathrm{d}x \overset{\text{Corollary 3.24}}{=} \int_{\mathbb{R}^d} f(x)\overline{g(x)} \,\mathrm{d}x = (f, g)_{L^2}.\end{aligned}$$ □

**Lemma 3.26.** *Let $X$ be a normed vector space, $Y$ be a Banach space (i. e. a complete, normed vector space) and denote the corresponding norms with $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. Moreover, let $V \subset X$ be a dense subspace of $X$ and $F : V \to Y$ be linear isometry, i. e. $F$ is linear and $\|F(x)\|_Y = \|x\|_X$ for all $x \in V$. Then, there is a unique linear isometry $\hat{F} : X \to Y$ with $\hat{F}|_V = F$.*

**Proof.** Let $x \in X$ be arbitrary but fixed. Since $V$ is dense in $X$, there is a sequence $(x_n) \in V^{\mathbb{N}}$ such that $\|x - x_n\|_X \to 0$ for $n \to \infty$. Then,

$$\|F(x_n) - F(x_m)\|_Y \overset{F \text{ linear}}{=} \|F(x_n - x_m)\|_Y \overset{F \text{ isometric}}{=} \|x_n - x_m\|_X \text{ for all } m, n \in \mathbb{N}.$$

Note that this shows that any linear isometry is Lipschitz continuous with constant 1. Since $x_n$ is a Cauchy sequence in $X$, $F(x_n)$ is a Cauchy sequence in $Y$. Since $Y$ is complete, $F(x_n)$ converges to a $y \in Y$. Now we define $\hat{F}$ via $\hat{F}(x) := y$. To show that $\hat{F}$ is well-defined, we need to show that $y$ is independent of the specific choice of the sequence. Let $(\tilde{x}_n) \in V^{\mathbb{N}}$ be another sequence with $\|x - \tilde{x}_n\|_X \to 0$ for $n \to \infty$. Then,

$$\begin{aligned}\|F(\tilde{x}_n) - y\|_Y &\le \|F(\tilde{x}_n) - F(x_n)\|_Y + \|F(x_n) - y\|_Y = \|\tilde{x}_n - x_n\|_X + \|F(x_n) - y\|_Y \\ &\le \underbrace{\|\tilde{x}_n - x\|_X}_{\to 0} + \underbrace{\|x - x_n\|_X}_{\to 0} + \underbrace{\|F(x_n) - y\|_Y}_{\to 0} \to 0 \text{ for } n \to \infty.\end{aligned}$$

Thus, both $F(\tilde{x}_n)$ and $F(x_n)$ converge to $y$, which shows that $\hat{F}$ is well defined.

The linearity of $\hat{F}$ follows from the linearity of $F$ and the linearity of the limit of convergent sequences. Using the reverse triangle inequality, we get

$$\begin{aligned}&0 \le |\|F(x_n)\|_Y - \|y\|_Y| \le \|F(x_n) - y\|_Y \to 0 \;\Rightarrow\; \lim_{n\to\infty} \|F(x_n)\|_Y = \|y\|_Y \\ &\Rightarrow\; \|\hat{F}(x)\|_Y = \|y\|_Y = \lim_{n\to\infty} \|F(x_n)\|_Y = \lim_{n\to\infty} \|x_n\|_X = \|x\|_X .\end{aligned}$$

Thus, $\hat{F}$ is a linear isometry.

For $x \in V$, the constant sequence $x_n := x$ converges to $x$ and we get

$$\hat{F}(x) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} F(x) = F(x),$$

which means $\hat{F}|_V = F$.

Finally, consider any linear isometry $\tilde{F} : X \to Y$ with $\tilde{F}|_V = F$. Let $x \in X$ be arbitrary but fixed and $(x_n) \in V^{\mathbb{N}}$ such that $\|x - x_n\|_X \to 0$ for $n \to \infty$. Since $\tilde{F}$ is a linear isometry, $\tilde{F}$ is continuous (even Lipschitz continuous with constant 1, see above). Thus, we get

$$\tilde{F}(x) = \lim_{n\to\infty} \tilde{F}(x_n) = \lim_{n\to\infty} F(x_n) = \hat{F}(x).$$

Thus, $\tilde{F} = \hat{F}$, which shows the claimed uniqueness. □

**Theorem 3.27** (Plancherel's theorem)**.** *There is a unique, bijective, linear isometry*

$$\mathcal{F}_2 : L^2(\mathbb{R}^d, \mathbb{C}) \to L^2(\mathbb{R}^d, \mathbb{C}) \textit{ with } \mathcal{F}_2|_{\mathcal{S}(\mathbb{R}^d,\mathbb{C})} = \mathcal{F}.$$

*Moreover, the* Plancherel formula *holds:*

$$(f, g)_{L^2} = (\mathcal{F}_2 f, \mathcal{F}_2 g)_{L^2} \textit{ for all } f, g \in L^2(\mathbb{R}^d, \mathbb{C})$$

**Proof.** By Proposition 3.25, we have

$$\|f\|_{L^2}^2 = (f, f)_{L^2} = (\mathcal{F}f, \mathcal{F}f)_{L^2} = \|\mathcal{F}f\|_{L^2}^2 \text{ for all } f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C}).$$

Due to the linearity of $\mathcal{F}$, Proposition 3.18 and Lemma 3.14

$$\mathcal{F} : \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \to \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset L^2(\mathbb{R}^d, \mathbb{C})$$

is a linear isometry (with respect to the $L^2$-norm). Since $C_c^\infty(\mathbb{R}^d, \mathbb{C})$ is dense in $L^2(\mathbb{R}^d, \mathbb{C})$ (Proposition 2.7 can be applied to the real and imaginary part separately) and since $C_c^\infty(\mathbb{R}^d, \mathbb{C}) \subset \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$ is dense in $L^2(\mathbb{R}^d, \mathbb{C})$. Thus, we can apply Lemma 3.26 with $V = \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, $X = Y = L^2(\mathbb{R}^d, \mathbb{C})$ and $F = \mathcal{F}$ to get a unique linear isometry

$$\mathcal{F}_2 : L^2(\mathbb{R}^d, \mathbb{C}) \to L^2(\mathbb{R}^d, \mathbb{C}) \text{ with } \mathcal{F}_2|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} = \mathcal{F}.$$

Let $f, g \in L^2(\mathbb{R}^d, \mathbb{C})$. Since $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$ is dense in $L^2(\mathbb{R}^d, \mathbb{C})$, there are sequences $(f_n), (g_n) \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})^{\mathbb{N}}$ with $\|f_n - f\|_{L^2} \to 0$ and $\|g_n - g\|_{L^2} \to 0$. This implies (exercise)

$$(f, g)_{L^2} = \lim_{n\to\infty} (f_n, g_n)_{L^2}.$$

Since $\mathcal{F}_2$ is a linear isometry, we have $\|\mathcal{F}_2 f_n - \mathcal{F}_2 f\|_{L^2} = \|f_n - f\|_{L^2} \to 0$ and $\|\mathcal{F}_2 g_n - \mathcal{F}_2 g\|_{L^2} = \|g_n - g\|_{L^2} \to 0$. Thus, like above, we get

$$(\mathcal{F}_2 f, \mathcal{F}_2 g)_{L^2} = \lim_{n\to\infty} (\mathcal{F}_2 f_n, \mathcal{F}_2 g_n)_{L^2}.$$

Combined with $(f_n, g_n)_{L^2} = (\mathcal{F}_2 f_n, \mathcal{F}_2 g_n)_{L^2}$ from Proposition 3.25, this proves the Plancherel formula. It remains to show that $\mathcal{F}_2$ is bijective.

Like $\mathcal{F}$, the mapping

$$\mathcal{F}^3 : \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \to \mathcal{S}(\mathbb{R}^d, \mathbb{C}) \subset L^2(\mathbb{R}^d, \mathbb{C})$$

is also a linear isometry. Analogously to the above, there is a unique linear isometry

$$\mathcal{G}_2 : L^2(\mathbb{R}^d, \mathbb{C}) \to L^2(\mathbb{R}^d, \mathbb{C}) \text{ with } \mathcal{G}_2|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} = \mathcal{F}^3.$$

Since $\mathcal{F}^3 = \mathcal{F}^{-1}$ on $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$ (Proposition 3.25), we immediately get

$$\mathcal{F}_2\mathcal{G}_2|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} = \mathcal{G}_2\mathcal{F}_2|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} = \mathrm{id}_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} \;\Rightarrow\; \mathcal{F}_2^{-1}|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})} = \mathcal{G}_2|_{\mathcal{S}(\mathbb{R}^d, \mathbb{C})}.$$

Since $\mathcal{F}_2$ and $\mathcal{G}_2$ are continuous and $S(\mathbb{R}^d, \mathbb{C})$ is dense in $L^2(\mathbb{R}^d, \mathbb{C})$, this implies $\mathcal{F}_2^{-1} = \mathcal{G}_2$ on $L^2(\mathbb{R}^d, \mathbb{C})$: For $f \in L^2(\mathbb{R}^d, \mathbb{C})$, there is a sequence $(f_n) \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})^{\mathbb{N}}$ with $\|f_n - f\|_{L^2} \to 0$. Thus,

$$\begin{aligned} 0 &\le \|\mathcal{F}_2\mathcal{G}_2 f - f\|_{L^2} \le \|\mathcal{F}_2\mathcal{G}_2 f - \mathcal{F}_2\mathcal{G}_2 f_n\|_{L^2} + \|\mathcal{F}_2\mathcal{G}_2 f_n - f\|_{L^2} \\ &= \underbrace{\|\mathcal{F}_2\mathcal{G}_2 f - \mathcal{F}_2\mathcal{G}_2 f_n\|_{L^2}}_{\to 0 \text{ since } \mathcal{F}_2\mathcal{G}_2 \text{ is continuous}} + \|f_n - f\|_{L^2} \to 0 \text{ for } n \to \infty. \end{aligned}$$

This shows that $\mathcal{F}_2\mathcal{G}_2 f = f$. Thus, $\mathcal{F}_2$ is bijective. □

**Proposition 3.28.**

*(i) For $f \in L^1(\mathbb{R}^d, \mathbb{C}) \cap L^2(\mathbb{R}^d, \mathbb{C})$, we have*

$$(\mathcal{F}_2 f)(\omega) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\omega} \,\mathrm{d}x \;\; \textit{for a.e. } \omega \in \mathbb{R}^d.$$

*(ii) For $f \in L^2(\mathbb{R}^d, \mathbb{C})$, we have*

$$\|\mathcal{F}_2 f - \phi_R\|_{L^2} \to 0 \textit{ for } R \to \infty, \textit{ where } \phi_R(\omega) := \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{B_R(0)} f(x) e^{-ix\cdot\omega} \, \mathrm{d}x \, .$$

**Proof.**

(i) One can show that there is a sequence $f_n \in C_c^\infty(\mathbb{R}^d, \mathbb{C})$ with $\|f - f_n\|_{L^1(\mathbb{R}^d,\mathbb{C})} \to 0$ and $\|f - f_n\|_{L^2(\mathbb{R}^d,\mathbb{C})} \to 0$. This can be done using that $\|f - \psi_\epsilon * f\|_{L^p(\mathbb{R}^d,\mathbb{C})} \to 0$ for $\epsilon \to 0$, if $f \in L^p(\mathbb{R}^d, \mathbb{C})$, $p \in [1, \infty)$ and $\psi$ is a mollifier (cf. Proposition 2.7 for the corresponding property of the correlation). Proposition 3.2 implies

$$\|\mathcal{F}f - \mathcal{F}f_n\|_{C(\mathbb{R}^d,\mathbb{C})} = \|\mathcal{F}(f - f_n)\|_{C(\mathbb{R}^d,\mathbb{C})} \leq C \, \|f - f_n\|_{L^1(\mathbb{R}^d,\mathbb{C})} \, ,$$

which means $\|\mathcal{F}f - \mathcal{F}f_n\|_{L^\infty(\mathbb{R}^d,\mathbb{C})} = \|\mathcal{F}f - \mathcal{F}f_n\|_{C(\mathbb{R}^d,\mathbb{C})} \to 0$ for $n \to \infty$. Combined with

$$0 \leq \int_{B_R(0)} |\mathcal{F}f_n(\omega) - \mathcal{F}f(\omega)|^2 \, \mathrm{d}\omega \leq \mathrm{Vol}(B_R(0)) \, \|\mathcal{F}f - \mathcal{F}f_n\|^2_{L^\infty(\mathbb{R}^d,\mathbb{C})} \quad \text{for all } R > 0,$$

we get

$$\lim_{n\to\infty} \int_{B_R(0)} |\mathcal{F}f_n(\omega) - \mathcal{F}f(\omega)|^2 \, \mathrm{d}\omega = 0 \text{ for all } R > 0.$$

Since $f_n \in C_c^\infty(\mathbb{R}^d, \mathbb{C}) \subset \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, we have $\mathcal{F}_2 f_n = \mathcal{F}f_n$ by Theorem 3.27. Combined with the fact that $\mathcal{F}_2$ is an isometry, we get

$$\|\mathcal{F}f_n - \mathcal{F}_2 f\|_{L^2(\mathbb{R}^d,\mathbb{C})} = \|\mathcal{F}_2(f_n - f)\|_{L^2(\mathbb{R}^d,\mathbb{C})} = \|f_n - f\|_{L^2(\mathbb{R}^d,\mathbb{C})} \to 0,$$

which, combined with $0 \leq \|\mathcal{F}f_n - \mathcal{F}_2 f\|_{L^2(B_R(0),\mathbb{C})} \leq \|\mathcal{F}f_n - \mathcal{F}_2 f\|_{L^2(\mathbb{R}^d,\mathbb{C})}$, implies

$$\lim_{n\to\infty} \int_{B_R(0)} |\mathcal{F}f_n(\omega) - \mathcal{F}_2 f(\omega)|^2 \, \mathrm{d}\omega = 0 \text{ for all } R > 0.$$

In other words, on each $B_R(0)$, $\mathcal{F}f_n$ converges in the $L^2$-norm to both $\mathcal{F}f$ and $\mathcal{F}_2 f$. Thus, $\mathcal{F}f = \mathcal{F}_2 f$ almost everywhere.

(ii) Since $L^2(B_R(0), \mathbb{C}) \subset L^1(B_R(0), \mathbb{C})$ and $f \in L^2(\mathbb{R}^d, \mathbb{C})$, we have $\chi_{B_R(0)} f \in L^1(\mathbb{R}^d, \mathbb{C})$. From (i), we get $\phi_R = \mathcal{F}(\chi_{B_R(0)} f) = \mathcal{F}_2(\chi_{B_R(0)} f)$ a.e. By Corollary B.13 with $|f| \in L^2(\mathbb{R}^d)$ as majorant, we get $\left\|f - \chi_{B_R(0)} f\right\|_{L^2} \to 0$ for $R \to \infty$. Combined with

$$\left\|f - \chi_{B_R(0)} f\right\|_{L^2} = \left\|\mathcal{F}_2(f - \chi_{B_R(0)} f)\right\|_{L^2} = \left\|\mathcal{F}_2 f - \mathcal{F}_2(\chi_{B_R(0)} f)\right\|_{L^2} = \|\mathcal{F}_2 f - \phi_R\|_{L^2} \, ,$$

the claim follows. □

*Remark* 3.29. Proposition 3.28 also holds for $\mathcal{F}_2^{-1}$ when replacing "$e^{-ix\cdot\omega}$" by "$e^{ix\cdot\omega}$". Moreover, Lemma 3.4, Corollary 3.6, the convolution theorem 3.7 and Corollary 3.10 also hold for $\mathcal{F}_2$. While the integrals in our definition of $\mathcal{F}$ do not necessarily exist for $L^2$-functions, Proposition 3.28 (ii) shows that we can still represent it as a limit of integrals of the same structure. For the sake of simplicity, hereafter, we still use the $L^1$-integral representation $\mathcal{F}$ also for $L^2$-functions, but point out that one actually has to use the exact representation. In particular, $\mathcal{F}$ has to be understood as $\mathcal{F}_2$.

*Remark* 3.30 (Frequency domain representation). For $f \in L^2(\mathbb{R}^d, \mathbb{C})$, we have $f = \mathcal{F}^{-1}\mathcal{F}f$. In particular,

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} (\mathcal{F}f)(\omega) e^{ix\cdot\omega} \, \mathrm{d}\omega \,.$$

In this sense, an $L^2$-function can be expressed as superposition of complex exponential functions. Here, the Fourier transform of the function in $\omega$ specifies the weight of $e^{ix\cdot\omega}$. Thus, $\mathcal{F}f$ is also called *frequency domain representation* of $f$. $f$ is called *space domain representation* or, in case of signals ($d = 1$), *time domain representation.*

*Remark* 3.31 (Frequency domain representation of linear filters). Let $f \in L^2(\mathbb{R}^d, \mathbb{C})$ be an image and $\psi \in L^2(\mathbb{R}^d, \mathbb{C})$ a convolution kernel. By the convolution theorem, we get

$$\mathcal{F}(\psi * f) = (2\pi)^{\frac{d}{2}} \mathcal{F}(\psi)\mathcal{F}(f).$$

Note: $\psi * f$ is not necessarily in $L^2$. The equation still holds, but needs an even more general definition of $\mathcal{F}$, cf. Remark 3.35. The Fourier transform of the kernel $\mathcal{F}\psi$ thus specifies how the convolution with $\psi$ amplifies or lowers the frequencies of the frequency domain representation of $f$. In this context, $\mathcal{F}\psi$ is also called *transfer function.* A convolution kernel $\psi$ with $(\mathcal{F}\psi)(\omega) \approx 0$ for "high" $\omega$ (i. e. for $|\omega|$ large) is called *low-pass filter*, since it lets only "low" frequencies ($|\omega|$ small) pass. Analogously, a *high-pass filter* is a filter with $(\mathcal{F}\psi)(\omega) \approx 0$ for "small" $\omega$.

Low-pass filters reduce noise, since noise mostly consists of high frequency components. However, edges also have high frequency components and thus are "smeared out" by a low-pass filter. This is not surprising, since from Section 2, we already know that linear filters cannot preserve edges.

Since $\mathcal{F}(\psi \star f) = (2\pi)^{\frac{d}{2}} \overline{\mathcal{F}(\psi)} \mathcal{F}(f)$, the same holds for the correlation. Here, one just has to additionally take into account the complex conjugation of the Fourier transform of the filter kernel.

*Example* 3.32. Let $\psi$ be a filter with $\mathcal{F}\psi = \frac{1}{(2\pi)^{\frac{d}{2}}} \chi_{B_r(0)}$. Since

$$\mathcal{F}(\psi \star f) = (2\pi)^{\frac{d}{2}} \overline{\mathcal{F}(\psi)} \mathcal{F}(f) = \chi_{B_r(0)} \mathcal{F}(f),$$

such a filter is called *perfect low-pass filter* with radius $r > 0$. With it, an image $f$ can be decomposed into a low and a high frequency component:

$$f_{\text{low}} = (\psi \star f), \quad f_{\text{high}} = f - f_{\text{low}}.$$

$f_{\text{high}}$ corresponds to a high-pass filtered image, since we have

$$\mathcal{F}f_{\text{high}} = \mathcal{F}f - \mathcal{F}f_{\text{low}} = \mathcal{F}f - \chi_{B_r(0)}\mathcal{F}f = \chi_{\mathbb{R}^d \setminus B_r(0)} \mathcal{F}f.$$

*Example* 3.33 (Deconvolution with the convolution theorem). The convolution theorem allows undoing a convolution, e. g. to correct motion blur (convolution with a box filter) or undoing Gaussian blur (convolution with the Gaussian kernel). This is called *deconvolution* or *deblurring.* We consider an image $f$ that is the result of a convolution of an image $f_0$ with a kernel $\psi$, i. e. $f = \psi * f_0$. If $\mathcal{F}\psi(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$, we get

$$\mathcal{F}^{-1}\left(\frac{\mathcal{F}f}{(2\pi)^{\frac{d}{2}}\mathcal{F}\psi}\right) = \mathcal{F}^{-1}\left(\frac{(2\pi)^{\frac{d}{2}}\mathcal{F}\psi\mathcal{F}f_0}{(2\pi)^{\frac{d}{2}}\mathcal{F}\psi}\right) = f_0.$$

Thus, if $f$ and $\psi$ are known exactly (and the denominator is never zero), $f_0$ can be reconstructed exactly. If $f$ is not known exactly but only approximately, there are noticeably artifacts in the estimate of $f_0$ obtained with this approach. Even an 8-bit quantization of $f$, which is commonly used when storing digital images, has a visible effect on the reconstruction. Filters for which $\mathcal{F}\psi$ is very small are also problematic, since these can arbitrarily amplify errors in $\mathcal{F}f$ due to the division by $\mathcal{F}\psi$.

Numerically, the division has to be approximated, for instance, by regularizing the complex division with as follows $\epsilon > 0$:

$$\frac{a+bi}{c+di} \approx \frac{(ac+bd)+(bc-ad)i}{c^2+d^2+\epsilon}.$$

Even today, with an unknown convolution kernel, the problem of deconvolution is not totally solved, but still subject of current research.

*Remark* 3.34. With our current definition, the Fourier transform cannot by applied to constant functions: For $c \in \mathbb{R} \setminus \{0\}$ and $d = 1$, we have

$$(\mathcal{F}c)(\omega) = \frac{c}{(2\pi)^{\frac{1}{2}}} \int_{\mathbb{R}} e^{-ix\omega} \, \mathrm{d}x \, .$$

This integral does not exist, since

$$\int_a^b e^{-ix\omega} \, \mathrm{d}x \overset{\text{(exercise)}}{=} \begin{cases} b-a & \omega = 0 \\ \frac{i}{\omega}\left(e^{-ib\omega} - e^{-ia\omega}\right) & \omega \neq 0 \end{cases}$$

diverges for $a \to -\infty$ and $b \to \infty$. It diverges even for $a = -R$, $b = R$ and $R \to \infty$, i. e. $\phi_R$ from Proposition 3.28 diverges. Nevertheless, $\mathcal{F}c$ can be defined implicitly using the inversion theorem. Let $d$ be arbitrary but fixed. For $\mathcal{F}c$ and all $x$ it must hold that

$$\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} (\mathcal{F}c)(\omega) e^{ix\cdot\omega} \, \mathrm{d}\omega = c.$$

For this, it is not sufficient to describe $\mathcal{F}c$ as a function, but as a measure. Let $\delta_0$ be the Dirac measure at 0 (Dirac delta function at 0) and $\mathcal{F}c = (2\pi)^{\frac{d}{2}} c\delta_0$. Then,

$$\frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{ix\cdot\omega} \mathrm{d}(\mathcal{F}c)(\omega) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{ix\cdot 0} c (2\pi)^{\frac{d}{2}} = c.$$

In this sense, the Fourier transform of the constant function $c$, which is maximally "smooth" and has maximal support, is the (scaled) Dirac measure and thus maximally concentrated.

*Remark* 3.35. The Fourier transform can be extended to even more general spaces than $L^2$, for instance, to the space of tempered distributions (dual space of $\mathcal{S}(\mathbb{R}^d, \mathbb{C})$, see Definition A.10). We will not discuss this in detail, but just exemplify it with the Dirac delta function $\delta_x$.

The Fourier transform of the distribution $\delta_x$ is defined by its evaluation at $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$:

$$(\mathcal{F}\delta_x)(f) \coloneqq \delta_x(\mathcal{F}f) = (\mathcal{F}f)(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(y) e^{-iy\cdot x} \, \mathrm{d}y \, .$$

Thus, $\mathcal{F}\delta_x$ is a so-called *regular distribution*, i. e. it can be expressed as integral with a density $y \mapsto \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-iy\cdot x}$. In particular, $\mathcal{F}\delta_0$ coincides with the constant function $\frac{1}{(2\pi)^{\frac{d}{2}}}$. Thus, the Fourier transform of this maximally concentrated "function" has maximal support and is maximally "smooth".

This generalized definition is consistent with our definition of $\mathcal{F}$ on $L^2$: Let $g \in L^2(\mathbb{R}^d, \mathbb{C})$. Through $g(f) := \int_{\mathbb{R}^d} gf \, \mathrm{d}x$ for $f \in \mathcal{S}(\mathbb{R}^d, \mathbb{C})$, $g$ is a (tempered) distribution and we get

$$(\mathcal{F}g)(f) = g(\mathcal{F}f) = \int_{\mathbb{R}^d} g(\mathcal{F}f) \, \mathrm{d}x \overset{\text{Lemma 3.8}}{=} \int_{\mathbb{R}^d} (\mathcal{F}g) f \, \mathrm{d}x \, .$$

## 3.2. Orthogonal expansions

**Definition 3.36.** Let $X$ be a pre-Hilbert space, i. e. a $\mathbb{K}$-vector space with a scalar product and $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. C : Complex Numbers

(i) A sequence $(e_n)_n \in X^N$ with $N \subset \mathbb{N}$ is called *orthonormal system (ONS) in $X$*, if

$$(e_i, e_j)_X = \delta_{ij} \text{ for all } i, j \in N.$$

(ii) If $(e_n)_n \in X^N$ is an ONS and $x \in X$, then, for $k \in N$, $x_k := (x, e_k)_X$ is called *$k$-th Fourier coefficient* (with regard to the ONS).

*Example* 3.37. Let $(f_n)_n \in L^2((0,\pi))^{\mathbb{N}}$ be given by $f_n(x) = \sin(nx)$. Then,

$$(f_i, f_j)_{L^2((0,\pi))} = \frac{\pi}{2}\delta_{ij} \text{ for all } i, j \in \mathbb{N}.$$

For $i \neq j$, we get using $\sin(x)\sin(y) = \frac{1}{2}(\cos(x-y) - \cos(x+y))$ that

$$\begin{aligned}(f_i, f_j)_{L^2((0,\pi))} &= \int_0^\pi \sin(ix)\sin(jx) \, \mathrm{d}x = \frac{1}{2}\int_0^\pi \cos((i-j)x) - \cos((i+j)x) \, \mathrm{d}x \\ &= \frac{1}{2}\left(-\frac{1}{i-j}\sin((i-j)x) + \frac{1}{i+j}\sin((i+j)x)\right)\Big|_0^\pi = 0.\end{aligned}$$

Moreover, we get

$$\begin{aligned}(f_i, f_i)_{L^2((0,\pi))} &= \frac{1}{2}\int_0^\pi \cos((i-i)x) - \cos((i+i)x) \, \mathrm{d}x = \frac{1}{2}\int_0^\pi 1 - \cos(2ix) \, \mathrm{d}x \\ &= \frac{1}{2}\left(x + \frac{1}{2i}\sin(2ix)\right)\Big|_0^\pi = \frac{\pi}{2}.\end{aligned}$$

Thus, $(\sqrt{\frac{2}{\pi}} f_n)_n$ is an ONS in $L^2((0,\pi))$.

**Proposition 3.38** (Best approximation by Fourier coefficients)**.** *Let $X$ be a pre-Hilbert space, $K \in \mathbb{N}$ and $(e_n)_{n=1}^K \subset X$ an orthonormal system. Then, for $x \in X$ and $c_1, \ldots, c_K \in \mathbb{K}$,*

*(i)* $\left\| x - \sum_{k=1}^K x_k e_k \right\|_X^2 = \|x\|_X^2 - \sum_{k=1}^K |x_k|^2.$

*(ii)* $\left\| x - \sum_{k=1}^K c_k e_k \right\|_X^2 = \|x\|_X^2 - \sum_{k=1}^K |x_k|^2 + \sum_{k=1}^K |c_k - x_k|^2.$

**Proof.** (i) is shown as follows (omitting the index $X$ from the norm and the scalar product):

$$\begin{aligned}
\left\| x - \sum_{k=1}^{K} x_k e_k \right\|^2 &= \left( x - \sum_{k=1}^{K} x_k e_k, x - \sum_{k=1}^{K} x_k e_k \right) \\
&= (x,x) - \left( \sum_{k=1}^{K} x_k e_k, x \right) - \left( x, \sum_{k=1}^{K} x_k e_k \right) + \left( \sum_{k=1}^{K} x_k e_k, \sum_{l=1}^{K} x_l e_l \right) \\
&= \|x\|^2 - \sum_{k=1}^{K} x_k \underbrace{(e_k, x)}_{=\overline{x_k}} - \sum_{k=1}^{K} \overline{x_k} \underbrace{(x, e_k)}_{=x_k} + \sum_{k=1}^{K} \sum_{l=1}^{K} x_k \overline{x_l} \underbrace{(e_k, e_l)}_{=\delta_{kl}} \\
&= \|x\|^2 - \sum_{k=1}^{K} |x_k|^2 .
\end{aligned}$$

For (ii): We have

$$\left\| x - \sum_{k=1}^{K} c_k e_k \right\|^2 = \left\| \underbrace{x - \sum_{k=1}^{K} x_k e_k}_{=:Y} + \underbrace{\sum_{k=1}^{K} (x_k - c_k) e_k}_{=:Z} \right\|^2 = (Y,Y) + \underbrace{(Y,Z) + \overline{(Y,Z)}}_{=2\mathrm{Re}(Y,Z)} + (Z,Z).$$

Combined with

$$\begin{aligned}
&(Y,Z) = \left( x - \sum_{k=1}^{K} x_k e_k, \sum_{l=1}^{K} (x_l - c_l) e_l \right) \\
&= \sum_{l=1}^{K} \overline{(x_l - c_l)} \underbrace{(x, e_l)}_{=x_l} - \sum_{k=1}^{K} x_k \sum_{l=1}^{K} \overline{(x_l - c_l)} \underbrace{(e_k, e_l)}_{\delta_{kl}} = \sum_{l=1}^{K} \overline{(x_l - c_l)} x_l - \sum_{k=1}^{K} x_k \overline{(x_k - c_k)} = 0,
\end{aligned}$$

it follows that

$$\begin{aligned}
\left\| x - \sum_{k=1}^{K} c_k e_k \right\|^2 &= (Y,Y) + (Z,Z) = \left\| x - \sum_{k=1}^{K} x_k e_k \right\|^2 + \left\| \sum_{k=1}^{K} (x_k - c_k) e_k \right\|^2 \\
&\overset{(i)}{=} \|x\|^2 - \sum_{k=1}^{K} |x_k|^2 + \left( \sum_{k=1}^{K} (x_k - c_k) e_k, \sum_{l=1}^{K} (x_l - c_l) e_l \right) \\
&= \|x\|^2 - \sum_{k=1}^{K} |x_k|^2 + \sum_{k=1}^{K} \sum_{l=1}^{K} (x_k - c_k) \overline{(x_l - c_l)} \underbrace{(e_k, e_l)}_{=\delta_{kl}} \\
&= \|x\|^2 - \sum_{k=1}^{K} |x_k|^2 + \sum_{k=1}^{K} |x_k - c_k|^2 .
\end{aligned}$$

□

*Remark* 3.39. Proposition 3.38 (ii) quantifies how well an element $x \in X$ is approximated in the $X$-norm by a finite linear combination of an ONS. Here, the error is apparently minimal if the Fourier coefficients are used as coefficients, i. e. $c_k = x_k$. In this sense, the Fourier coefficients result in the best approximation.

**Corollary 3.40** (Bessel's inequality). *Let $X$ be a pre-Hilbert space, $K \in \mathbb{N}$ and $(e_n)_{n=1}^K \subset X$ an orthonormal system. Then, for $x \in X$,*

$$\sum_{k=1}^{K} |x_k|^2 \leq \|x\|_X^2 .$$

This part should be smaller than the right side
and the both same if and only if the below zero

*Here, equality holds, if and only if* $\left\| x - \sum_{k=1}^K x_k e_k \right\|_X = 0.$

*Moreover, if $(e_n)_n \in X^{\mathbb{N}}$ is an orthonormal system, then*

$$\sum_{k=1}^{\infty} |x_k|^2 \leq \|x\|_X^2 .$$

*Here, equality holds, if and only if*

$$\lim_{K\to\infty} \left\| x - \sum_{k=1}^{K} x_k e_k \right\|_X^2 = 0.$$

**Proof.** The first part is an immediate consequence of Proposition 3.38 (i):

$$\|x\|_X^2 - \sum_{k=1}^{K} |x_k|^2 = \left\| x - \sum_{k=1}^{K} x_k e_k \right\|_X^2 \geq 0.$$

By going to the limit $K \to \infty$, we get $\sum_{k=1}^\infty |x_k|^2 \leq \|x\|_X^2$. In particular, the series converges absolutely. In case $\sum_{k=1}^\infty |x_k|^2 = \|x\|_X^2$, going to the limit in Proposition 3.38 (i) implies

$$\lim_{K\to\infty} \left\| x - \sum_{k=1}^{K} x_k e_k \right\|_X^2 = \|x\|_X^2 - \sum_{k=1}^{\infty} |x_k|^2 .$$

Thus, $\sum_{k=1}^\infty |x_k|^2 = \|x\|_X^2$ holds, if and only if $\lim\limits_{K\to\infty} \left\| x - \sum_{k=1}^K x_k e_k \right\|_X^2 = 0.$ □

*Remark* 3.41. Corollary 3.40 means that if $\sum_{k=1}^\infty |x_k|^2 = \|x\|_X^2$, the sequence $(\sum_{k=1}^K x_k e_k)_K$ converges (strongly, i. e. in the $X$-norm) to $x$. This motivates the following definition.

**Definition 3.42.** Let $X$ be a pre-Hilbert space, $N \subset \mathbb{N}$ and $(e_n)_n \in X^N$ an orthonormal system. The ONS is called *complete* or a *complete orthonormal system (CONS) of $X$*, if

$$\sum_{k\in N} |x_k|^2 = \|x\|_X^2 \text{ for all } x \in X.$$

A CONS is also called *orthonormal basis*.

*Remark* 3.43. Due to Corollary 3.40, Definition 3.42 implies: If $(e_n)_n \in X^{\mathbb{N}}$ is a CONS in a pre-Hilbert space $X$, $\sum_{k=1}^K x_k e_k$ converges (strongly) to $x$ for $K \to \infty$.

**Corollary 3.44.** *Let $X$ be a pre-Hilbert space, $N \subset \mathbb{N}$ and $(e_n)_n \in X^N$ an ONS. The ONS is complete, if and only if*

$$\underset{x\in X}{\forall}\ \underset{\epsilon>0}{\forall}\ \underset{K\in\mathbb{N}}{\exists}\ \underset{i_1,\dots,i_K\in N}{\exists}\ \underset{c_{i_1},\dots,c_{i_K}\in\mathbb{K}}{\exists}\ \left\| x - \sum_{k=1}^{K} c_{i_k} e_{i_k} \right\|_X < \epsilon$$

*In other words, the ONS is complete, if and only if* $\overline{\operatorname{span}\{e_k : k \in N\}} = X$.

**Proof.** WLOG $N = \{1, \dots, M\}$ for $M \in \mathbb{N}$ or $N = \mathbb{N}$ (else re-index the $e_n$).

"$\Rightarrow$": Follows from $\sum_{k\in N} |x_k|^2 = \|x\|_X^2$ with Corollary 3.40 and $i_k = k$, $c_k = x_k$.

"$\Leftarrow$": From Proposition 3.38 (ii) and Corollary 3.40, we get

$$\begin{aligned}\epsilon^2 > \left\| x - \sum_{k=1}^{K} c_{i_k} e_{i_k} \right\|_X^2 &\overset{3.38\,\text{(ii)}}{=} \|x\|_X^2 - \sum_{k=1}^{K} |x_{i_k}|^2 + \sum_{k=1}^{K} |c_{i_k} - x_{i_k}|^2 \\ &\geq \|x\|_X^2 - \sum_{k=1}^{K} |x_{i_k}|^2 \geq \|x\|_X^2 - \sum_{k\in N} |x_k|^2 \overset{\text{Corollary 3.40}}{\geq} 0.\end{aligned}$$

From this, the claim follows by going to the limit $\epsilon \to 0$. ☐

**Proposition 3.45** (Parseval's identity)**.** *Let $X$ be a pre-Hilbert space and $(e_n)_n \in X^{\mathbb{N}}$ a CONS in $X$. Then,*

$$(x, y)_X = \sum_{k=1}^{\infty} (x, e_k)_X \overline{(y, e_k)_X} \quad \textit{for all } x, y \in X.$$

**Proof.** Let $x_k := (x, e_k)$ and $y_k := (y, e_k)$. For $K \in \mathbb{N}$, we get

$$\sum_{k=1}^{K} |x_k \overline{y_k}| = \overset{\text{Cauchy Schwartz}}{\sum_{k=1}^{K}} |x_k|\,|y_k| \overset{\text{CS}}{\leq} \left(\sum_{k=1}^{K} |x_k|^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{K} |y_k|^2\right)^{\frac{1}{2}} \overset{\text{Corollary 3.40}}{\leq} \|x\|\,\|y\| < \infty.$$

Thus, the series $\sum_{k=1}^{\infty} x_k \overline{y_k}$ is absolutely convergent. We show the claimed identity separately for the real and the imaginary part. First, note that

$$\|x + y\|^2 = (x + y, x + y) = (x, x) + (x, y) + (y, x) + (y, y) = \|x\|^2 + \|y\|^2 + 2\mathrm{Re}(x, y). \quad (*)$$

With

$$x_k + y_k = (x, e_k) + (y, e_k) = (x + y, e_k) = (x + y)_k,$$

we get

$$\begin{aligned}\|x + y\|^2 &\overset{\text{CONS}}{=} \sum_{k=1}^{\infty} |(x + y)_k|^2 = \sum_{k=1}^{\infty} |x_k + y_k|^2 = \sum_{k=1}^{\infty} \left(|x_k|^2 + |y_k|^2 + 2\mathrm{Re}(x_k \overline{y}_k)\right) \\ &\overset{\text{CONS}}{=} \|x\|^2 + \|y\|^2 + 2\mathrm{Re}\left(\sum_{k=1}^{\infty} x_k \overline{y}_k\right).\end{aligned}$$

Combined with $(*)$, it follows that $\mathrm{Re}(x, y) = \mathrm{Re}\left(\sum_{k=1}^{\infty} x_k \overline{y}_k\right)$. Since $\mathrm{Re}(iz) = -\mathrm{Im} z$ for $z \in \mathbb{C}$, replacing $x$ by $ix$ in the calculation above, shows $-\mathrm{Im}(x, y) = -\mathrm{Im}\left(\sum_{k=1}^{\infty} x_k \overline{y}_k\right)$ and the claim follows. ☐

## 3.3. Fourier series

Since digital images are typically given on bounded hyperrectangles $[a_1, b_1] \times \cdots \times [a_d, b_d]$, we are not only interested in the Fourier transform on the whole $\mathbb{R}^d$, but also in analogous transforms on such hyperrectangles.

First, we just consider the case $d = 1$ and $[a_1, b_1] = [-B, B]$ with $B > 0$. Using scaling and translation, this case can be generalized to all bounded intervals and then, using the tensor product, to hyperrectangles in any dimension.

Since $[-B, B]$ is bounded, we can directly consider $L^2$-functions.

**Definition 3.46.** For $B > 0$ and $f, g \in L^2([-B, B], \mathbb{C})$, we define the normalized scalar product

$$(f, g)_{[-B,B]} = \frac{1}{2B} \int_{-B}^{B} f(x)\overline{g(x)} \,\mathrm{d}x\,.$$

**Proposition 3.47.** *The functions* $(e_k)_{k\in\mathbb{Z}}$ *given by* $e_k(x) = e^{ikx}$ *are an orthonormal basis of* $L^2([-\pi, \pi], \mathbb{C})$ *with respect to the scalar product* $(\cdot, \cdot)_{[-\pi,\pi]}$.

$(e_i, e_j)_{[-\pi,\pi]} = \delta_{i,j}$ is shown by a straightforward computation (exercise). The basis property can be shown using Corollary 3.44 and the Weierstrass approximation theorem for trigonometric polynomials together with the denseness of $C([-\pi, \pi], \mathbb{C})$ in $L^2([-\pi, \pi], \mathbb{C})$ (cf. [1, Theorem 4.33]). Note that we have to re-index $\mathbb{Z}$ with $\mathbb{N}$ to use results from Section 3.2. This can be done since absolute convergence implies unconditional convergence in complete spaces.

reindex? small example?

*Remark* 3.48. From the basis property, we get that $f \in L^2([-\pi, \pi], \mathbb{C})$ can be expressed as the series

$$f = \sum_{k\in\mathbb{Z}} (f, e_k)_{[-\pi,\pi]} e_k,$$

where the equality is to be understood as convergence of the series in the $L^2$-norm. This series is called *Fourier series* and the coefficients

$$(f, e_k)_{[-\pi,\pi]} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{-ikx} \,\mathrm{d}x$$

are the Fourier coefficients from Definition 3.36 (here we used $\overline{e^{ikx}} = e^{-ikx}$).

By transforming the integral with $\varphi_B : [-\pi, \pi] \to [-B, B], x \mapsto \frac{B}{\pi}x$, the series representation can be generalized to $L^2([-B, B], \mathbb{C})$ for $B > 0$. Here, the orthonormal basis is

$$e_k^B(x) = e^{ik\frac{\pi}{B}x}.$$

*Example* 3.49. Let $f : [-\pi, \pi] \to \mathbb{R}, x \mapsto x(\pi - |x|)$. Then, $f(-x) = -x(\pi - |-x|) = -f(x)$, i. e. $f$ is odd. Moreover, we have

$$\begin{aligned}
(f, e_k)_{[-\pi,\pi]} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)e^{-ikx} \,\mathrm{d}x = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)(\cos(-kx) + i\sin(-kx)) \,\mathrm{d}x \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} \underbrace{f(x)\cos(kx)}_{\text{odd}} \,\mathrm{d}x - \frac{i}{2\pi} \int_{-\pi}^{\pi} \underbrace{f(x)\sin(kx)}_{\text{even}} \,\mathrm{d}x \\
&= -\frac{i}{\pi} \int_{0}^{\pi} f(x)\sin(kx) \,\mathrm{d}x = -\frac{i}{\pi} \int_{0}^{\pi} (x\pi - x^2)\sin(kx) \,\mathrm{d}x
\end{aligned}$$

and

$$\int_0^\pi (x\pi - x^2)\sin(kx)\,\mathrm{d}x = (x\pi - x^2)\left(-\frac{1}{k}\cos(kx)\right)\Big|_0^\pi - \int_0^\pi (\pi - 2x)\left(-\frac{1}{k}\cos(kx)\right)\mathrm{d}x$$
$$= 0 + \frac{1}{k}\int_0^\pi (\pi - 2x)\cos(kx)\,\mathrm{d}x = \frac{\pi}{k}\int_0^\pi \cos(kx)\,\mathrm{d}x - \frac{2}{k}\int_0^\pi x\cos(kx)\,\mathrm{d}x$$
$$= \frac{\pi}{k^2}\sin(kx)\Big|_0^\pi - \frac{2}{k}\left(\frac{x}{k}\sin(kx)\big|_0^\pi - \int_0^\pi \frac{1}{k}\sin(kx)\,\mathrm{d}x\right)$$
$$= 0 - \frac{2}{k}\left(0 + \frac{1}{k^2}\cos(kx)\Big|_0^\pi\right) = -\frac{2}{k^3}((-1)^k - 1).$$

In total, we get

$$(f, e_k)_{[-\pi,\pi]} = \frac{i}{\pi}\frac{2}{k^3}((-1)^k - 1) = \begin{cases} -\frac{i}{\pi}\frac{4}{k^3} & k \text{ odd} \\ 0 & k \text{ even} \end{cases}.$$

**Theorem 3.50** (Sampling theorem, Shannon-Whitakker). *Let $B > 0$ and $f \in L^2(\mathbb{R}, \mathbb{C})$ with $\mathcal{F}f(\omega) = 0$ for $|\omega| > B$. Then, $f$ is continuous (more precisely, has a continuous representative) and is uniquely determined by the values $(f(\frac{k\pi}{B}))_{k\in\mathbb{Z}}$. Moreover,*

$$f(x) = \sum_{k\in\mathbb{Z}} f\left(\frac{k\pi}{B}\right)\operatorname{sinc}\left(\frac{B}{\pi}\left(x - \frac{k\pi}{B}\right)\right).$$

record a sample depending on time, if you know what the
the bandwidth of the sample and how often the recording is done,
then we do not have to get the whole signal

**Proof.** Since $L^2([-B,B],\mathbb{C}) \subset L^1([-B,B],\mathbb{C})$ and $\hat{f}(\omega) = \mathcal{F}f(\omega) = 0$ for all $|\omega| > B$, we have $\hat{f} \in L^1(\mathbb{R},\mathbb{C}) \cap L^2(\mathbb{R},\mathbb{C})$. Thus, by Remark 3.29, we have $f = \mathcal{F}^{-1}\hat{f} = D_{-\mathbb{1}}\mathcal{F}\hat{f}$, which is continuous by Proposition 3.2. More precisely, $f$ has a continuous representative, which we also denote by $f$. In particular, using $f = D_{-\mathbb{1}}\mathcal{F}\hat{f}$ and the continuity, we get

$$f\left(\tfrac{k\pi}{B}\right) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \hat{f}(\omega)e^{i\frac{k\pi}{B}\omega}\,\mathrm{d}\omega = \frac{1}{\sqrt{2\pi}}\int_{-B}^{B} \hat{f}(\omega)e^{i\frac{k\pi}{B}\omega}\,\mathrm{d}\omega = \sqrt{\frac{2}{\pi}}B(\hat{f}, e^B_{-k})_{[-B,B]}.$$

Therefore, all Fourier coefficients $(\hat{f}, e^B_{-k})_{[-B,B]}$ are solely determined by the values $f\left(\frac{k\pi}{B}\right)$. By Remark 3.48, the values uniquely determine $\hat{f} \in L^2([-B,B],\mathbb{C})$. Due to the bijectivity of $\mathcal{F} : L^2(\mathbb{R},\mathbb{C}) \to L^2(\mathbb{R},\mathbb{C})$ and the assumption $\hat{f} \equiv 0$ on $\mathbb{R} \setminus [-B,B]$, also $f \in L^2(\mathbb{R},\mathbb{C})$ is uniquely determined by these values.

In particular, we have

$$\hat{f} = \chi_{[-B,B]}\sum_{k\in\mathbb{Z}}(\hat{f}, e^B_k)_{[-B,B]}e^B_k = \sqrt{\frac{\pi}{2}}\frac{1}{B}\sum_{k\in\mathbb{Z}} f\left(\tfrac{-k\pi}{B}\right)e^B_k\chi_{[-B,B]}.$$

Applying $\mathcal{F}^{-1}$ to this equation and using the linearity and the continuity of $\mathcal{F}^{-1}$, we get

$$f = \sqrt{\frac{\pi}{2}}\frac{1}{B}\sum_{k\in\mathbb{Z}} f\left(\tfrac{-k\pi}{B}\right)\mathcal{F}^{-1}(e^B_k\chi_{[-B,B]}).$$

Using Lemma 3.4 and Example 3.3, we get

$$\begin{aligned}\mathcal{F}^{-1}(e_k^B\chi_{[-B,B]})(x) &= \overline{\mathcal{F}^{-1}(e_k^B\chi_{[-B,B]})(x)} = \overline{\mathcal{F}(\overline{e_k^B\chi_{[-B,B]}})(x)} = \overline{\mathcal{F}(\overline{M_{k\frac{\pi}{B}}\chi_{[-B,B]}})(x)}\\ &= (D_{-\mathbb{1}}\mathcal{F}(M_{k\frac{\pi}{B}}\chi_{[-B,B]}))(x) = (D_{-\mathbb{1}}T_{-k\frac{\pi}{B}}\mathcal{F}(\chi_{[-B,B]}))(x)\\ &= (D_{-\mathbb{1}}T_{-k\frac{\pi}{B}}\sqrt{\tfrac{2}{\pi}}B\operatorname{sinc}\left(\tfrac{B}{\pi}\cdot\right))(x) = \sqrt{\frac{2}{\pi}}B\operatorname{sinc}\left(\frac{B}{\pi}\left(-x-k\frac{\pi}{B}\right)\right)\\ &= \sqrt{\frac{2}{\pi}}B\operatorname{sinc}\left(\frac{B}{\pi}\left(x+k\frac{\pi}{B}\right)\right).\end{aligned}$$

In total, we have

$$f = \sum_{k\in\mathbb{Z}} f\left(\frac{-k\pi}{B}\right)\operatorname{sinc}\left(\frac{B}{\pi}\left(x+k\frac{\pi}{B}\right)\right).$$

After substituting $k$ with $-k$, the claim follows. $\square$

*Remark* 3.51. The constant $B$ from the sampling theorem is called *bandwidth* of the signal $f$ and specifies the highest (possible) frequency occurring in the signal. Colloquially, the sampling theorem means:

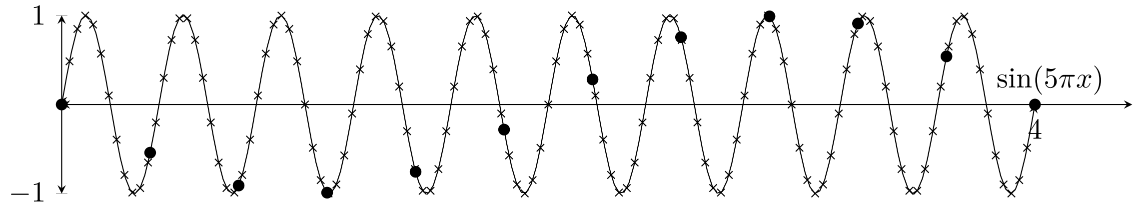
"If a signal with bandwidth $B$ is sampled with a sampling rate $\frac{\pi}{B}$, one gets all information contained in the signal".

Attention: As mentioned before, there are multiple definitions of the Fourier transform in the literature. This is the same for the term frequency. In engineering, frequency often refers to the so-called *circular frequency* that corresponds to a definition of the Fourier transform where $e^{-2\pi i x\cdot\omega}$ is used instead of $e^{-ix\cdot\omega}$. A signal with bandwidth $B$ has a maximal circular frequency of $\nu = \frac{B}{2\pi}$. In this terminology, the sampling theorem means:

"If a signal with frequencies up to a maximal circular frequency $\nu$ is sampled with a sampling rate $\frac{1}{2\nu}$, one gets all information contained in the signal".

The sampling rate $\frac{1}{2\nu}$ (or equivalently $\frac{\pi}{B}$) is called *Nyquist rate*.

*Remark* 3.52. As evident, for instance, for the sine function, it is not possible to identity functions when using a sampling rate that is too low. Sampling with a sampling rate smaller




than the Nyquist rate is called *undersampling*, correspondingly *oversampling* refers to a sampling rate larger than the Nyquist rate. Undersampling leads to an effect called *aliasing*, i. e. the sampled signal contains frequencies that are not in the original signal. In a way, these are an alias for frequencies that are actually present in the original signal. Aliasing can be prevented by pre-processing with a suitable low-pass filtering that removes all frequencies from the original signal that cannot be identified with the used sampling rate.

## 3.4. Discrete Fourier transform

complex vectors with N entries

In this section, we consider $\mathbb{C}^N$ indexed from 0 to $N-1$, i. e. $F \in \mathbb{C}^N$ is interpreted as mapping $F : \{0, \dots, N-1\} \to \mathbb{C}$. Analogously, both components of $\mathbb{C}^{N \times M}$ are indexed starting with 0.

**Definition 3.53.** Let $F = (f_n)_{n=0}^{N-1} \in \mathbb{C}^N$, then $\hat{F} = (\hat{f}_k)_{k=0}^{N-1} \in \mathbb{C}^N$, the *(one-dimensional) discrete Fourier transform (DFT)* of $F$, is defined by

$$\mathrm{DFT}(F)_k := \hat{f}_k := \frac{1}{N} \sum_{n=0}^{N-1} f_n \exp\left(\frac{-2\pi i n k}{N}\right), \quad k = 0, \dots, N-1.$$

*Remark* 3.54. Since the DFT is linear, it can be expressed as matrix. Let $B = [b^0 \dots b^{N-1}]$ be the matrix with the columns

$$b^n = \left(\exp\left(\frac{-2\pi i n k}{N}\right)\right)_{k=0}^{N-1} \in \mathbb{C}^N.$$

Then, we have $\hat{F} = \frac{1}{N} B F$, so $\frac{1}{N} B$ is the matrix representation. The matrix $B$ is dense and thus is not assembled in practice. However, the matrix can be used to easily derive the inverse DFT.

**Proposition 3.55.** *The vectors $b^0, \dots, b^{N-1}$ are an orthogonal basis of $\mathbb{C}^N$ with respect to the unitary scalar product: We have,*

$$(b^k, b^l) = N \delta_{k,l}.$$

*In particular, $(\frac{1}{N} B)^{-1} = \overline{B}^T$, i. e. the inverse of the DFT is*

$$f_n = \sum_{k=0}^{N-1} \hat{f}_k \exp\left(\frac{2\pi i n k}{N}\right), \quad n = 0, \dots, N-1.$$

**Proof.** As in Proposition 3.47, the orthogonality is shown by a straightforward computation (exercise). From $(b^k, b^l) = N \delta_{k,l}$, we get $(\overline{B}^T B) = N \mathbb{1}$ and thus $(\frac{1}{N} B)^{-1} = \overline{B}^T$. For the adjoint matrix $\overline{B}^T$, we have

$$(\overline{B}^T)_{k,l} = \overline{b}_{l,k} = \exp\left(\frac{2\pi i k l}{N}\right).$$

From this, the expression of the inverse directly follows. □

*Remark* 3.56. In other words, the DFT is a change of basis from the canonical basis to the basis $\overline{b}^0, \dots, \overline{b}^{N-1}$, since we have

$$f = \sum_{k=0}^{N-1} \hat{f}_k \overline{b}^k.$$

We denote the inverse transform of $F$ by $\check{F}$.

*Remark* 3.57. We have

$$b^n_{k+N} = \exp\left(\frac{-2\pi i n(k+N)}{N}\right) = \exp\left(\frac{-2\pi i n k}{N} - 2\pi i n\right) = \exp\left(\frac{-2\pi i n k}{N}\right) = b^n_k,$$

i. e. $b^n$ is $N$-periodic if the definition is extended to the index set $\mathbb{Z}$. Analogously, one gets $b^{n+N} = b^n$. $\hat{f}_k = \frac{1}{N}(F, \overline{b}^k)$ depends (independent of $k$) only on $f_0, \ldots, f_{N-1}$ (and $N$). Thus, with the definition of $b^n$ for $n \in \mathbb{Z}$, we can also extend $\hat{f}_k$ to $\mathbb{Z}$. Due to the periodicity of $b^n$, also the DFT is $N$-periodic, i. e. $\hat{f}_{k+N} = \hat{f}_k$. Thus, for the DFT, all signals are $N$-periodic.

The periodicity allows interpreting the coefficients of the DFT as high and low frequencies: Since $b^{N-k} = b^{-k}$ and $\hat{f}_{N-k} = \hat{f}_{-k}$, for small $k$, $\hat{f}_{N-k}$ (and $\hat{f}_k$) corresponds to a low frequency component while, for $k$ close to $\frac{N}{2}$, $\hat{f}_{N-k}$ (and $\hat{f}_k$) corresponds to a high frequency component.

*Remark* 3.58 (2D DFT). Let $F \in \mathbb{C}^{N\times M}$, then $\hat{F} \in \mathbb{C}^{N\times M}$, the two-dimensional DFT, is defined by

$$\hat{f}_{k,l} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f_{n,m} \exp\left(\frac{-2\pi i n k}{N}\right) \exp\left(\frac{-2\pi i m l}{M}\right).$$

The inverse is

$$f_{n,m} = \sum_{l=0}^{M-1} \sum_{k=0}^{N-1} \hat{f}_{k,l} \exp\left(\frac{2\pi i n k}{N}\right) \exp\left(\frac{2\pi i m l}{M}\right).$$

The 2D DFT is separable, thus it can be computed by first computing the 1D DFT of all columns (this gives an element of $\mathbb{C}^{N\times M}$) and then, on the result, computing the 1D DFT of all rows. Alternatively, one can first transform the rows and then the columns.

**Definition 3.59.** Let $F, G \in \mathbb{C}^N$, then $(F \circledast G)$, the *periodic convolution* of $F$ with $G$, is defined as

$$(F \circledast G)_n = \sum_{k=0}^{N-1} g_k f_{(n-k) \bmod N}, \; n = 0, \ldots, N-1.$$

**Proposition 3.60** (Discrete convolution theorem)**.** *Let $F, G \in \mathbb{C}^N$. Then,*

$$(\widehat{F \circledast G})_n = N \hat{f}_n \hat{g}_n.$$

**Proof.** The claim follows from the $N$-periodicity of $\exp\left(\frac{-2\pi i n k}{N}\right)$ in $k$:

$$\begin{aligned}
(\widehat{F \circledast G})_n &= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \left( g_l f_{(k-l) \bmod N} \exp\left(\frac{-2\pi i n k}{N}\right)\right) \\
&= \frac{1}{N} \sum_{l=0}^{N-1} \left( g_l \exp\left(\frac{-2\pi i n l}{N}\right) \sum_{k=0}^{N-1} \left( f_{(k-l) \bmod N} \exp\left(\frac{-2\pi i n (k-l)}{N}\right)\right)\right) \\
&= \frac{1}{N} \sum_{l=0}^{N-1} \left( g_l \exp\left(\frac{-2\pi i n l}{N}\right) \sum_{k=0}^{N-1} \left( f_k \exp\left(\frac{-2\pi i n k}{N}\right)\right)\right) \\
&= N \hat{g}_n \hat{f}_n.
\end{aligned}$$

□

*Remark* 3.61 (Fast Fourier transform). The canonical evaluation of the DFT (exactly like the discrete convolution) needs $O(N^2)$ operations. However, using symmetries, it is possible to reduce the number of necessary operations to $O(N \log_2 N)$. A DFT implemented like this is called *fast Fourier transform* (FFT). Using Proposition 3.60, this also allows for a fast convolution.

In case $N = 2^n$, an algorithm for the FFT can be constructed with reasonable effort (exercise). A truly efficient implementation for arbitrary grid sizes requires considerable effort though. Thus, one should use an existing implementation. In MATLAB, the function `fft` provides such an implementation. In Python, `fft` is found in `numpy.fft`. If the FFT is needed in C or Fortran, the free library "Fastest Fourier Transform in the West" (FFTW) can be used.

*Example* 3.62 (Multiplication of large integers). The discrete convolution theorem can be used for the multiplication of large integers. The sign of the product is trivial to compute, so it is sufficient to consider natural numbers. Let $a, b \in \mathbb{N}$. Then, $a$ and $b$ can be expressed in the decimal system, i. e.

$$a = \sum_{i=0}^{m} a_i 10^i, \quad b = \sum_{j=0}^{n} b_j 10^j,$$

with $a_0, \dots, a_m, b_0, \dots, b_n \in \{0, \dots, 9\}$. Setting $a_i = 0$ for $i > m$ and $b_j = 0$ for $j > n$ and $j < 0$, we get

$$\begin{aligned} ab &= \left(\sum_{i=0}^{m} a_i 10^i\right)\left(\sum_{j=0}^{n} b_j 10^j\right) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j 10^{i+j} = \sum_{k=0}^{n+m} \sum_{i+j=k} a_i b_j 10^{i+j} \\ &= \sum_{k=0}^{n+m} 10^k \sum_{l=0}^{k} a_l b_{k-l} = \sum_{k=0}^{n+m} 10^k \sum_{l=0}^{n+m} a_l b_{k-l}. \end{aligned}$$

Thus, the digits of the product $ab$ are given by the (linear) convolution of the digit vectors of $a$ and $b$, which can be computed very efficiently using the FFT.

*Remark* 3.63. Similarly to $\mathcal{F}(\overline{f}) = \overline{D_{-\mathbb{1}}(\mathcal{F}f)}$ (cf. Lemma 3.4), i.e. $(\mathcal{F}f)(\omega) = \overline{(\mathcal{F}(\overline{f}))(-\omega)}$, the DFT fulfills the symmetry

$$\hat{f}_k = \frac{1}{N}\sum_{n=0}^{N-1} f_n b_k^n = \overline{\frac{1}{N}\sum_{n=0}^{N-1} \overline{f}_n \overline{b_k^n}} = \overline{\frac{1}{N}\sum_{n=0}^{N-1} \overline{f}_n b_{N-k}^n} = \overline{\hat{\overline{f}}}_{N-k}.$$

If $F$ is real-valued, this symmetry implies $\hat{f}_k = \overline{\hat{f}}_{N-k}$. Thus, in this case, the DFT is partially redundant: $\hat{F}$ is already completely characterized by $\hat{f}_k$ for $k = 0, \dots, ((N+1) \operatorname{div} 2) - 1$. FFT implementations like FFTW can avoid the redundant computations when working with real-valued input data.

The computation of the unnecessary coefficients is also avoided by the *discrete cosine transform* (DCT). For $F \in \mathbb{R}^N$, the DCT is defined as

$$\mathrm{DCT}(F)_k = \frac{2\lambda_k}{N}\sum_{n=0}^{N-1} f_n \cos\left(\tfrac{k\pi}{N}(n + \tfrac{1}{2})\right),$$

where $\lambda_k = \begin{cases} \frac{1}{\sqrt{2}} & k = 0 \\ 1 & k \neq 0 \end{cases}$. We have $\mathrm{DCT}(F)_k \in \mathbb{R}$. Like the DFT (Recall: $(\overline{B}^T B) = N\,\mathbb{1}$), the DCT is an orthogonal mapping (exercise) and its inverse transform is

$$f_n = \mathrm{IDCT}(\mathrm{DCT}(F))_n = \lambda_n \sum_{k=0}^{N-1} \mathrm{DCT}(F)_k \cos\left(\tfrac{k\pi}{N}(n + \tfrac{1}{2})\right).$$

Like the DFT, the DCT can be computed with $O(N \log_2 N)$ operations. Attention: The DCT discussed here ist just one of four commonly used variants.

*Remark* 3.64. The DCT is an integral component of JPEG compression. Essentially, the image to be JPEG compressed is decomposed into blocks of size $8 \times 8$ and the DCT is computed on each block. The DCT coefficients are saved after quantization, resorting and entropy encoding. In a pre-processing step, low-pass filtering is used.

*Example* 3.65 (Inpainting of signals with limited bandwidth). Let $f_0 \in L^2(\mathbb{R}^d, \mathbb{C})$ and $f$ be a measurement of $f_0$ where the part on $\Omega \subset \mathbb{R}^d$ is missing, i. e. we only know the function

$$f : \mathbb{R}^d \to \mathbb{C}, x \mapsto \begin{cases} f_0(x) & x \notin \Omega \\ 0 & x \in \Omega \end{cases}.$$

The aim is to estimate the function $f_0$ based on $f$ and $\Omega$. This problem is called *inpainting*.

If $\Omega$ is not too big and there is $B > 0$ with $\mathcal{F} f_0(\omega) = 0$ for $|\omega| \geq B$ (and $B$ is minimal), we can approximately reconstruct $f_0$ using its frequency domain representation in a fixed point iteration. Let $\psi$ be a perfect low-pass filter with radius $B$. Then, the iteration is given by

$$f^{n+1}(x) = \begin{cases} (\psi \star f^n)(x) & x \in \Omega \\ f(x) & x \notin \Omega \end{cases} \text{ for } n \in \mathbb{N},$$

where $f^1 = f$. By assumption, we have $\psi \star f_0 = f_0$. Thus, $f_0$ is a fixed point of this iteration.

In other words, the method consists of repeated low-pass filtering where the known values are re-inserted after each application of the filter. The low-pass filtering should be implemented using the fast convolution idea, i.e. in Fourier space. Attention: An accurate estimate of $B$ is required for the method to work. In particular, it is important to ensure that the selected $B$ is neither too big nor too small!

**Definition 3.66.** Let $F, G \in \mathbb{C}^N$, then $(F \circledast G)$, the *periodic correlation* of $F$ and $G$, is defined as

$$(F \circledast G)_n = \sum_{k=0}^{N-1} \overline{f}_k g_{(n+k) \bmod N}, \; n = 0, \ldots, N-1.$$

*Remark* 3.67. Like in the continuous setting, there is a counterpart of the discrete convolution theorem for the periodic correlation. By slightly adjusting the proof of Proposition 3.60, one can show (exercise) that

$$(\widehat{F \circledast G})_n = N \overline{\hat{f}_n} \hat{g}_n \text{ for } F, G \in \mathbb{C}^N.$$

This is used to build a fast correlation.

*Example* 3.68 (Registration of images using the fast correlation). Let $\Omega := (0,1)^2 \subset \mathbb{R}^2$ be the unit square and $f, g \in L^2(\Omega)$ be a given image pair. Assuming that $f$ and $g$ show the same or a similar scene, the task is the align the two images, i.e. we want to find a deformation $\phi : \Omega \to \mathbb{R}^2$ such that

$$(g \circ \phi)(x) = g(\phi(x)) \approx f(x) \text{ for all } x \in \Omega, \text{ i.e. } g \circ \phi \approx f \text{ on } \Omega.$$

This task is called *image registration.* The easiest variant of this is not to consider all kinds of deformations $\phi$, but only translations, i.e. $\phi(x) = x + b$ for all $x \in \Omega$, where $b \in \mathbb{R}^2$ is a translation vector. Then, we have $g \circ \phi = T_b g$. To find the translation vector $b$ that aligns the images, we look for the $b$ that minimizes the $L^2$-difference of $f$ and $T_b g$, i.e. we want find a minimizer of

$$J : \mathbb{R}^2 \to \mathbb{R}, b \mapsto \frac{1}{2} \int_\Omega |f(x) - g(x+b)|^2 \,\mathrm{d}x = \frac{1}{2} \|T_b g - f\|_{L^2}^2 .$$

Note that $g$ has to be extended appropriately to evaluate $T_b g(x) = g(x+b)$ for all $x \in \Omega$. Using periodic extension leads to

$$\frac{1}{2}\|T_b g - f\|_{L^2}^2 = \frac{1}{2}\|T_b g\|_{L^2}^2 - (T_b g, f)_{L^2} + \frac{1}{2}\|f\|_{L^2}^2 = \frac{1}{2}\|g\|_{L^2}^2 - (T_b g, f)_{L^2} + \frac{1}{2}\|f\|_{L^2}^2 ,$$

where we used that $\|T_b g\|_{L^2} = \|g\|_{L^2}$ due to the periodic extension. Thus, minimizing $J$ is equivalent to maximizing

$$(T_b g, f)_{L^2} = \int_\Omega f(x) g(x+b) \,\mathrm{d}x = (f \star g)(b).$$

Thus, $f$ and $g$ can be aligned by maximizing their correlation.

Given discrete images $F$ and $G$ and only considering translations $b$ that shift by integer multiples of the pixel size, $F$ and $G$ can be aligned by computing the index pair $m, n$ that maximizes $(F \circledast G)_{m,n}$ (where $\circledast$ is canonically extended to 2D). For this, it is sufficient to compute the entire convolution $F \circledast G$ (which can be done in $O(N \log N)$) and then simply to select the index pair corresponding to the maximal element of the 2D array $F \circledast G$.

*Remark* 3.69 (Relation of DFT and FT). Let $f \in L^1(\mathbb{R}, \mathbb{C})$ be Riemann-integrable with $(\mathcal{F}f) \in L^1(\mathbb{R}, \mathbb{C})$. Moreover, let $\epsilon > 0$. Then, there are $L, B \in \mathbb{N}$ with

$$\int_{\mathbb{R}\setminus[-L,L]} |f(t)| \,\mathrm{d}t < \epsilon \text{ and } \int_{\mathbb{R}\setminus[-B,B]} |\mathcal{F}f(t)| \,\mathrm{d}t < \epsilon.$$

For $\omega \in \mathbb{R}$, we get

$$\left| \mathcal{F}f(\omega) - \frac{1}{\sqrt{2\pi}} \int_{-L}^{L} f(t) e^{-it\cdot\omega} \,\mathrm{d}t \right| \le \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}\setminus[-L,L]} |f(t)| \underbrace{\left|e^{-it\cdot\omega}\right|}_{=1} \mathrm{d}t \le \frac{\epsilon}{\sqrt{2\pi}}.$$

Since $\cos(\omega\cdot)$ and $\sin(\omega\cdot)$ are Lipschitz continuous with constant $\omega$ and $f$ is Riemann-integrable, one can show (exercise) that there is $N \in \mathbb{N}$ such that for all $\omega \in [-B, B]$

$$\left| \int_{-L}^{L} f(t) e^{-it\cdot\omega} \,\mathrm{d}t - \frac{2L}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f(t_n) \exp(-it_n \cdot \omega) \right| \le \epsilon,$$

where $t_n := \frac{2Ln}{N}$. For $\omega = \frac{\pi k}{L}$, we get

$$\begin{aligned}
\sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f(t_n)\exp(-it_n\cdot\omega) &= \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} f(t_n)\exp\left(\frac{-2\pi ink}{N}\right) \\
&= \sum_{n=0}^{N-1} f\left(t_{n-\frac{N}{2}}\right)\exp\left(\frac{-2\pi i\left(n-\frac{N}{2}\right)k}{N}\right) \\
&= \exp(\pi ik)\sum_{n=0}^{N-1} f\big(t_{n-\frac{N}{2}}\big)\exp\left(\frac{-2\pi ikn}{N}\right) = (-1)^k N\mathrm{DFT}\big(\widetilde{f}\big)_k
\end{aligned}$$

where $\widetilde{f} = \left(f\big(t_{n-\frac{N}{2}}\big)\right)_{n=0}^{N-1} \in \mathbb{C}^N$. Therefore, (if $\omega = \frac{\pi k}{L} \in [-B, B]$)

$$\left|\frac{1}{\sqrt{2\pi}}\int_{-L}^{L} f(t)e^{-it\cdot\omega}\,\mathrm{d}t - \frac{2L}{\sqrt{2\pi}}(-1)^k\mathrm{DFT}\big(\widetilde{f}\big)_k\right| \le \frac{\epsilon}{\sqrt{2\pi}}$$

and in total we get

$$\left|\mathcal{F}f(\tfrac{\pi k}{L}) - \frac{2L}{\sqrt{2\pi}}(-1)^k\mathrm{DFT}\big(\widetilde{f}\big)_k\right| \le 2\frac{\epsilon}{\sqrt{2\pi}}.$$

Thus, except for scaling and sign, the DFT is an approximation of the FT.

## 3.5. Trigonometric interpolation

The DFT is not only an approximation of the FT, it also solves a trigonometric interpolation problem, which we will show in this section.

**Definition 3.70.** For $m \in \mathbb{N}$, let

$$\mathcal{T}_m := \operatorname{span}\{e_j :\ 0 \le j < m\}$$

with $e_j$ from Proposition 3.47. The elements of $\mathcal{T}_m$ are called *(complex) trigonometric polynomials of degree* $m-1$.

*Remark* 3.71. Since $(e_j)_j$ is an orthonormal system (Proposition 3.47), we have $\dim \mathcal{T}_m = m$. Moreover,

$$\mathcal{T}_m = \left\{[-\pi,\pi] \to \mathbb{C}, x \mapsto \sum_{j=0}^{m-1} c_j e^{ijx} :\ c_0,\dots,c_{m-1} \in \mathbb{C}\right\}.$$

With $z = e^{ix}$, each element of $\mathcal{T}_m$ has the form

$$\sum_{j=0}^{m-1} c_j z^j,$$

i. e. the elements of $\mathcal{T}_m$ are complex polynomials, which are restricted to the unit circle of the complex plane, since $|e^{ix}| = 1$. Combined with the identity $e^{ix} = \cos x + i\sin x$, this is the reason why they are called complex trigonometric polynomials. Moreover, since $m-1$ is the highest power of $z$, their degree is $m-1$.

Note that $z = e^{ix}$ also allows us to draw a connection between the complex Taylor series of a function $f : \mathbb{C} \to \mathbb{C}$ and the Fourier series of $\tilde{f} : [-\pi, \pi] \to \mathbb{C}, x \mapsto f(e^{ix})$. Under very strong assumptions, the coefficients of these series are the same. In general, however, Taylor series and Fourier series are very different. The (complex) Taylor series is local, needs $f$ only to be known locally at the point of expansion, but also to be infinitely differentiable. The Fourier series is global, needs $\tilde{f}$ to be known everywhere, but does not require differentiability, only square integrability.

**Proposition 3.72.**

*(i) Let $p_1 \in \mathcal{T}_m$ and $p_2 \in \mathcal{T}_n$. Then, $p_1 p_2 \in \mathcal{T}_{m+n-1}$.*

*(ii) Let $p \in \mathcal{T}_m \setminus \{0\}$. Then, $p$ has at most $m-1$ roots in $[-\pi, \pi)$.*

**Proof.**

(i) Since $p_1 \in \mathcal{T}_m$ and $p_2 \in \mathcal{T}_n$, there are coefficients $c_j$ and $d_k$ such that

$$p_1(x) = \sum_{j=0}^{m-1} c_j e^{ijx} \text{ and } p_2(x) = \sum_{k=0}^{n-1} d_k e^{ikx}.$$

$$\Rightarrow \ p_1(x)p_2(x) = \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} c_j d_k e^{i(j+k)x} \in \mathcal{T}_{m+n-1}.$$

(ii) Let $p \in \mathcal{T}_m \setminus \{0\}$. Then, there are coefficients $c_j$ such that $p(x) = \sum_{j=0}^{m-1} c_j e^{ijx}$. Assume $p$ has $m$ different roots $x_1, \dots, x_m \in [-\pi, \pi)$. We consider the complex polynomial

$$q : \mathbb{C} \to \mathbb{C}, z \mapsto \sum_{j=0}^{m-1} c_j z^j.$$

Then,

$$q(e^{ix_k}) = p(x_k) = 0 \text{ for all } k \in \{1, ..., m\}.$$

Moreover, due to $x_k \in [-\pi, \pi)$, the $e^{ix_k}$ are different, which means that the complex polynomial $q$ has $m$ different roots, but is of degreee $m-1$. This implies $q = 0$, which means that all coefficients $c_j$ have to be zero. Thus, $p = 0$, which is a contradiction to $p \in \mathcal{T}_m \setminus \{0\}$. □

**Lemma 3.73.** *For $N \in \mathbb{N}$ and $m \in \{-N+1, \dots, N-1\}$, it holds that*

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{-i2\pi mj/N} = \delta_{m,0}.$$

**Proof.** For $m = 0$, we immediately get

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{-i2\pi 0j/N} = \frac{1}{N}\sum_{j=0}^{N-1} 1 = 1.$$

In case $m \neq 0$, we recall recall the sum formula of the geometric series, i. e.

$$\sum_{j=0}^{N-1} q^j = \frac{1-q^N}{1-q} \text{ for all } q \in \mathbb{C} \setminus \{1\}.$$

With $q = e^{-i2\pi m/N}$, we get

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{-i2\pi mj/N} = \frac{1}{N}\frac{1-e^{-i2\pi m}}{1-e^{-i2\pi m/N}} = 0. \qquad \square$$

*Remark* 3.74. Note that Lemma 3.73 allows to show the orthogonality of the DFT claimed in Proposition 3.55 almost immediately.

**Proposition 3.75.** *Let $y = (y_k)_{k=0}^{N-1} \in \mathbb{C}^N$ and $x_k = \frac{2\pi k}{N}$ ($k = 0, \ldots, N-1$). Then, the trigonometric polynomial with the coefficient vector $\hat{y}$, i. e.*

$$T_N(y; x) := \sum_{j=0}^{N-1} \hat{y}_j e^{ijx}$$

*fulfills the interpolation conditions*

$$T_N(y; x_k) = y_k \quad (k = 0, \ldots, N-1).$$

*Moreover, $T_N(y; \cdot)$ is the only element of $\mathcal{T}_N$ that fulfills the interpolation conditions.*

**Proof.** With Lemma 3.73, we get

$$\begin{aligned} T_N(y; x_k) &= \sum_{j=0}^{N-1} \hat{y}_j e^{i2\pi jk/N} = \frac{1}{N}\sum_{j=0}^{N-1}\left(\sum_{l=0}^{N-1} y_l e^{-2\pi ilj/N}\right) e^{i2\pi jk/N} \\ &= \sum_{l=0}^{N-1} y_l \frac{1}{N}\sum_{j=0}^{N-1} e^{-i2\pi j(l-k)/N} \overset{\text{Lemma 3.73}}{=} \sum_{l=0}^{N-1} y_l \delta_{l-k,0} = y_k. \end{aligned}$$

Assume $T \in \mathcal{T}_N$ also fulfills the interpolation conditions. Then, $T - T_N(y, \cdot) \in \mathcal{T}_N$ has $N$ different roots: For $k \in \{0, \ldots, N-1\}$, we have

$$T(x_k) - T_N(y; x_k) = y_k - y_k = 0.$$

Thus, $T - T_N(y, \cdot)$ has to be the zero polynomial, i. e. $T = T_N(y, \cdot)$. $\square$

*Remark* 3.76. Proposition 3.75 shows which trigonometric interpolation problem the DFT solves: Given the equidistant points $x_k = \frac{2\pi k}{N}$ and values $y_k \in \mathbb{C}$ ($k = 0, \ldots, N-1$), the DFT determines the trigonometric polynomial that interpolates the values $y_k$ at the points $x_k$, i. e. the DFT determines the element $T \in \mathcal{T}_N$ with $T(x_k) = y_k$ ($k = 0, \ldots, N-1$).

# 4. The wavelet transform

As we have seen before, the Fourier transform is a global operator: For each $\omega \in \mathbb{R}^d$, $\mathcal{F}(\omega)$ depends on the entire function $f : \mathbb{R} \to \mathbb{C}$. This is not always desirable. Thus, one is also interested in more localized transforms. It will turn out that the wavelet transform is such a transform.

## 4.1. The continuous wavelet transform

**Definition 4.1.** Let $f, \psi \in L^2(\mathbb{R})$. Then, the *wavelet transform* of $f$ with $\psi$, *scale parameter* $a > 0$ and *spatial parameter* $b \in \mathbb{R}$ is defined as

$$L_\psi f(a,b) \coloneqq \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \mathrm{d}x\,.$$

*Remark* 4.2. Unlike the Fourier Transform, which maps a function $f : \mathbb{R} \to \mathbb{C}$ to a function $\mathcal{F}f : \mathbb{R} \to \mathbb{C}$, the wavelet transform maps $f : \mathbb{R} \to \mathbb{R}$ to a function $L_\psi f : (0,\infty) \times \mathbb{R} \to \mathbb{R}$. In particular, the transformed function has a two dimensional domain even though the original function had a one dimension domain. There are two different ways to represent the transform,

(i) as scalar product, i.e. $L_\psi f(a,b) = \frac{1}{\sqrt{a}}(f, T_{-b} D_{\frac{1}{a}} \psi)_{L^2}$,

(ii) and as convolution, i.e. $L_\psi f(a,b) = \frac{1}{\sqrt{a}}(f * D_{-\frac{1}{a}} \psi)(b)$,

which will be useful in the following. The second representation is related to Remark 2.16, where we considered

$$u(x,\sigma) = (g_{\sqrt{2\sigma}} * f)(x).$$

to define a hierarchy of edges.

To first step to understand the wavelet transform is to understand to which space it maps. To this end, we define

$$L^2([0,\infty) \times \mathbb{R}, \tfrac{\mathrm{d}a\,\mathrm{d}b}{a^2}) \coloneqq \left\{ F : [0,\infty) \times \mathbb{R} \to \mathbb{R} \ \mu\text{-measurable} : \int_{\mathbb{R}} \int_0^\infty |F(a,b)|^2 \frac{\mathrm{d}a\,\mathrm{d}b}{a^2} < \infty \right\}$$

and equip it with the same equivalence relation used for the Lebesgue spaces, cf. Definition B.3, and the scalar product

$$(F,G)_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} \coloneqq \int_{\mathbb{R}} \int_0^\infty F(a,b) G(a,b) \frac{\mathrm{d}a\,\mathrm{d}b}{a^2}.$$

**Proposition 4.3.** *Let $\psi \in L^2(\mathbb{R})$ such that*

$$0 < c_\psi \coloneqq 2\pi \int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega} \mathrm{d}\omega < \infty.$$

*Then,*

$$L_\psi : L^2(\mathbb{R}) \to L^2([0,\infty) \times \mathbb{R}, \tfrac{\mathrm{d}a\,\mathrm{d}b}{a^2})$$

*is a linear mapping and we have*

$$(L_\psi f, L_\psi g)_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} = c_\psi (f,g)_{L^2(\mathbb{R})} \textit{ for all } f,g \in L^2(\mathbb{R}).$$

**Proof.** Let $f \in L^2(\mathbb{R})$. We use the scalar product representation of the transform, the Plancherel formula (Theorem 3.27) and Lemma 3.4 to get

$$\begin{aligned} L_\psi f(a,b) &= \tfrac{1}{\sqrt{a}}(f, T_{-b}D_{\frac{1}{a}}\psi)_{L^2} \overset{\text{Theorem 3.27}}{=} \tfrac{1}{\sqrt{a}}(\mathcal{F}f, \mathcal{F}(T_{-b}D_{\frac{1}{a}}\psi))_{L^2} \\ &\overset{\text{Lemma 3.4}}{=} \tfrac{1}{\sqrt{a}}(\hat{f}, aM_{-b}D_a\hat{\psi})_{L^2} = \tfrac{a}{\sqrt{a}}\int_{\mathbb{R}} \hat{f}(\omega)\overline{e^{-ib\omega}\hat{\psi}(a\omega)}\,\mathrm{d}\omega = \sqrt{a}\int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{\psi}(a\omega)}e^{ib\omega}\,\mathrm{d}\omega \\ &= \sqrt{a2\pi}\mathcal{F}^{-1}(\hat{f}\overline{D_a\hat{\psi}})(b). \end{aligned} \tag{$*^1$}$$

Since $L_\psi f(a,b) \in \mathbb{R}$, this also implies

$$L_\psi f(a,b) = \overline{\sqrt{a2\pi}\mathcal{F}^{-1}(\hat{f}\overline{D_a\hat{\psi}})(b)}. \tag{$*^2$}$$

Plugging $(*^1)$ for $f$ and $(*^2)$ for $g$ into the definition of the special scalar product leads to

$$\begin{aligned} (L_\psi f, L_\psi g)_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} &= \int_{\mathbb{R}}\int_0^\infty L_\psi f(a,b)L_\psi g(a,b)\frac{\mathrm{d}a\,\mathrm{d}b}{a^2} \\ &= \int_{\mathbb{R}}\int_0^\infty \sqrt{a2\pi}\mathcal{F}^{-1}(\hat{f}\overline{D_a\hat{\psi}})(b)\overline{\sqrt{a2\pi}\mathcal{F}^{-1}(\hat{g}\overline{D_a\hat{\psi}})(b)}\frac{\mathrm{d}a\,\mathrm{d}b}{a^2} \\ &\overset{\text{Fubini}}{=} 2\pi\int_0^\infty a\int_{\mathbb{R}} \mathcal{F}^{-1}(\hat{f}\overline{D_a\hat{\psi}})(b)\overline{\mathcal{F}^{-1}(\hat{g}\overline{D_a\hat{\psi}})(b)}\,\mathrm{d}b\,\frac{\mathrm{d}a}{a^2} \\ &\overset{\text{Theorem 3.27}}{=} 2\pi\int_0^\infty a\int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{\psi}(a\omega)}\,\overline{\hat{g}(\omega)\overline{\hat{\psi}(a\omega)}}\,\mathrm{d}\omega\,\frac{\mathrm{d}a}{a^2} \\ &\overset{\text{Fubini}}{=} 2\pi\int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{g}(\omega)}\int_0^\infty \overline{\hat{\psi}(a\omega)}\hat{\psi}(a\omega)\frac{\mathrm{d}a}{a}\,\mathrm{d}\omega \\ &= 2\pi\int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{g}(\omega)}\int_0^\infty \frac{|\hat{\psi}(a\omega)|^2}{a}\,\mathrm{d}a\,\mathrm{d}\omega\,. \end{aligned}$$

Since $\psi$ is real-valued, $\hat{\psi}$ is Hermitian (Corollary 3.6), which implies $|\hat{\psi}(\omega)| = |\hat{\psi}(-\omega)|$ for all $\omega \in \mathbb{R}$. Thus, using this and the substitution rule, we get

$$\int_0^\infty \frac{|\hat{\psi}(a\omega)|^2}{a}\,\mathrm{d}a = \int_0^\infty \frac{|\hat{\psi}(a|\omega|)|^2}{a}\,\mathrm{d}a \overset{\phi(a)=|\omega|a}{=} \int_0^\infty \frac{|\hat{\psi}(\phi(a))|^2}{\phi(a)}\phi'(a)\,\mathrm{d}a = \int_0^\infty \frac{|\hat{\psi}(a)|^2}{a}\,\mathrm{d}a = \frac{c_\psi}{2\pi}.$$

Combined with the above, this leads to

$$(L_\psi f, L_\psi g)_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} = 2\pi\int_{\mathbb{R}} \hat{f}(\omega)\overline{\hat{g}(\omega)}\frac{c_\psi}{2\pi}\,\mathrm{d}\omega = c_\psi(\hat{f},\hat{g})_{L^2(\mathbb{R})} \overset{\text{Theorem 3.27}}{=} c_\psi(f,g)_{L^2(\mathbb{R})}. \square$$

**Definition 4.4.** The condition

$$0 < c_\psi = 2\pi\int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega}\,\mathrm{d}\omega < \infty$$

from Proposition 4.3 is called *admissibility condition.* Any $\psi \in L^2(\mathbb{R})$ that fulfills this condition is called *wavelet.*

*Remark* 4.5. Since

$$\begin{aligned} \|L_\psi f\|_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} &= \sqrt{(L_\psi f, L_\psi f)_{L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})}} \overset{\text{Proposition 4.3}}{=} \sqrt{c_\psi(f,f)_{L^2(\mathbb{R})}} \\ &= \sqrt{c_\psi}\,\|f\|_{L^2(\mathbb{R})} \text{ for all } f \in L^2(\mathbb{R}), \end{aligned}$$

Proposition 4.3 shows that the wavelet transform is continuous in case $c_\psi < \infty$. Since

$$\int_0^1 \frac{1}{\omega}\,\mathrm{d}\omega = \lim_{a\searrow 0}\int_a^1 \frac{1}{\omega}\,\mathrm{d}\omega = \lim_{a\searrow 0}\log\omega\Big|_a^1 = \log 1 - \lim_{a\searrow 0}\log a = \infty,$$

it is necessary that $|\hat{\psi}(\omega)|^2$ decays to 0 quickly enough for $\omega \to 0$ in a suitable sense to get $c_\psi < \infty$. In case $\hat{\psi}$ is continuous (at least in 0), this implies $\hat{\psi}(0) = 0$ is necessary. In other words, it is necessary (but not sufficient) that the mean value of $\psi$ is 0 for $\psi$ to be a wavelet.

The condition $c_\psi > 0$ is necessary to guarantee the invertibility of the transform on its image:

**Corollary 4.6.** *Let $\psi \in L^2(\mathbb{R})$ be a wavelet. Then, for all $f \in L^2(\mathbb{R})$, it holds that*

$$f(x) = \frac{1}{c_\psi}\int_{\mathbb{R}}\int_0^\infty L_\psi f(a,b)\frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)\frac{\mathrm{d}a\,\mathrm{d}b}{a^2} \text{ for a.e. } x \in \mathbb{R}.$$

**Proof.** Let $f \in L^2(\mathbb{R})$ and $F \in L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2})$. Then,

$$\begin{aligned}(L_\psi f, F)_{L^2([0,\infty)\times\mathbb{R},\frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} &= \int_{\mathbb{R}}\int_0^\infty\int_{\mathbb{R}} f(x)\frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)\mathrm{d}x\,F(a,b)\frac{\mathrm{d}a\,\mathrm{d}b}{a^2}\\ &\overset{\text{Fubini}}{=} \int_{\mathbb{R}} f(x)\underbrace{\int_{\mathbb{R}}\int_0^\infty \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)F(a,b)\frac{\mathrm{d}a\,\mathrm{d}b}{a^2}}_{=:G(x)}\mathrm{d}x = \int_{\mathbb{R}} f(x)G(x)\,\mathrm{d}x = (f,G)_{L^2(\mathbb{R})}.\end{aligned}$$

Thus, we have

$$(f,G)_{L^2(\mathbb{R})} = (L_\psi f, F)_{L^2([0,\infty)\times\mathbb{R},\frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} = (f, L_\psi^* F)_{L^2(\mathbb{R})},$$

for all $f \in L^2(\mathbb{R})$, where $L_\psi^* : L^2([0,\infty)\times\mathbb{R}, \frac{\mathrm{d}a\,\mathrm{d}b}{a^2}) \to L^2(\mathbb{R})$ denotes the Hilbertian adjoint of $L_\psi$ (see the corresponding exercise for the definition). This implies

$$L_\psi^* F(x) = G(x) = \int_{\mathbb{R}}\int_0^\infty \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)F(a,b)\frac{\mathrm{d}a\,\mathrm{d}b}{a^2} \text{ for a.e. } x \in \mathbb{R}.$$

By Proposition 4.3, we have

$$(f,g)_{L^2(\mathbb{R})} = \frac{1}{c_\psi}(L_\psi f, L_\psi g)_{L^2([0,\infty)\times\mathbb{R},\frac{\mathrm{d}a\,\mathrm{d}b}{a^2})} = (\tfrac{1}{c_\psi}L_\psi^* L_\psi f, g)_{L^2(\mathbb{R})} \text{ for all } g \in L^2(\mathbb{R}).$$

This means, we have

$$f = \tfrac{1}{c_\psi}L_\psi^* L_\psi f \text{ a.e.}$$

Combined with the representation of $L_\psi^*$ shown above, the statement follows. □

*Example* 4.7.

(i) Let $\psi$ be given by $\psi(x) := xe^{-\frac{x^2}{2}} = -\frac{\mathrm{d}}{\mathrm{d}x}g(x)$ where $g(x) := e^{-\frac{x^2}{2}}$. Then,

$$\begin{aligned}\hat{\psi}(\omega) = \mathcal{F}\psi(\omega) = \mathcal{F}(-\tfrac{\mathrm{d}}{\mathrm{d}x}g)(\omega) &\overset{\text{Lemma 3.17}}{=} -i\omega\mathcal{F}g(\omega) \overset{\text{Lemma 3.20}}{=} -i\omega g(\omega)\\ &= -i\omega e^{-\frac{\omega^2}{2}} = -i\psi(\omega).\end{aligned}$$

With this, we get

$$\begin{aligned}
c_\psi &= 2\pi \int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega}\,\mathrm{d}\omega = 2\pi \int_0^\infty \frac{|\omega e^{-\frac{\omega^2}{2}}|^2}{\omega}\,\mathrm{d}\omega = 2\pi \int_0^\infty \frac{\omega^2 e^{-2\frac{\omega^2}{2}}}{\omega}\,\mathrm{d}\omega \\
&= \pi \int_0^\infty 2\omega e^{-\omega^2}\,\mathrm{d}\omega = \pi \left(-e^{-\omega^2}\right)\Big|_0^\infty = \pi\,(-0-(-1)) = \pi.
\end{aligned}$$

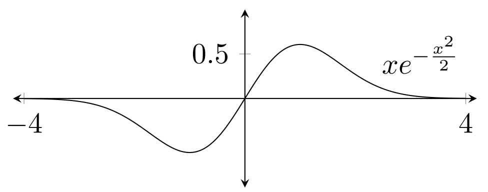
Thus, $\psi$ is a wavelet. Note that the mean value of $g$ is not zero, thus $g$ itself cannot be a wavelet.

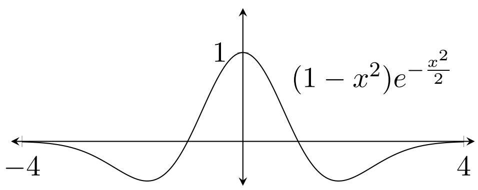
With the convolution representation of the wavelet transform, we get

$$\begin{aligned}
L_\psi f(a,b) &= \tfrac{1}{\sqrt{a}}(f * D_{-\frac{1}{a}}\psi)(b) = \tfrac{-1}{\sqrt{a}}(f * D_{-\frac{1}{a}}g')(b) = \tfrac{a}{\sqrt{a}}(f * (D_{-\frac{1}{a}}g)')(b) \\
&= \sqrt{a}(f * (D_{\frac{1}{a}}g))'(b).
\end{aligned}$$

Similarly to Remark 2.16, the input $f$ is convolved with a scaled Gaussian for different scales. Unlike Remark 2.16, a derivative is taken after the convolution.

(ii) Let $\psi$ be given by $\psi(x) := (1-x^2)e^{-\frac{x^2}{2}} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}g(x)$ with $g$ as before. Due to its shape, this $\psi$ is called *Mexican hat*. One can show that (exercise) $\hat{\psi}(\omega) = \omega^2 e^{-\frac{\omega^2}{2}}$ and that $c_\psi = \pi$. Thus, the Mexican hat is also a wavelet.




(iii) Another example is the *Haar wavelet* given by

$$\psi(x) := \begin{cases} 1 & 0 \le x < \frac{1}{2} \\ -1 & \frac{1}{2} \le x < 1 \\ 0 & \text{else} \end{cases}.$$

Using $\psi = \chi_{[0,\frac{1}{2})} - \chi_{[\frac{1}{2},1)}$ to represent $\psi$, one can use the properties of the Fourier transform to show that

$$\hat{\psi}(\omega) = \tfrac{1}{\sqrt{2\pi}} i e^{-i\frac{\omega}{2}} \sin(\tfrac{\omega}{4}) \operatorname{sinc}(\tfrac{\omega}{4\pi}).$$

Based on this, one can show that $c_\psi = \ln 2$, which means that $\psi$ is indeed a wavelet. Note that $c_\psi = \ln 2$ also follows from Proposition 4.16 combined with Example 4.17 (i).

Unlike the derivatives of the Gaussian we considered before that are very smooth and have non-compact support, the Haar wavelet is discontinuous and has compact support.

## 4.2. The discrete wavelet transform

The (continuous) wavelet transform of a function $f \in L^2(\mathbb{R})$ is apparently a redundant representation of $f$ ($f$ is univariate, $L_\psi f$ is bivariate). The question arises whether it is sufficient to know $L_\psi f$ on a subset of $[0, \infty) \times \mathbb{R}$. It will turn out that this is indeed the case for certain discrete subsets, but showing this requires several preparations.

**Definition 4.8** (Multiresolution analysis)**.** For $j \in \mathbb{Z}$, let $V_j$ be a closed subspace of $L^2(\mathbb{R})$. Then, $(V_j)_{j\in\mathbb{Z}}$ is called *multiresolution analysis (MRA)* or *multiscale approximation (MSA)*, if it fulfills the following conditions:

$$\begin{aligned} \text{translational invariance} \quad & \forall j, k \in \mathbb{Z} : f \in V_j \Leftrightarrow T_{2^j k} f \in V_j \\ \text{inclusion} \quad & \forall j \in \mathbb{Z} : V_{j+1} \subset V_j \\ \text{scaling} \quad & \forall j \in \mathbb{Z} : f \in V_j \Leftrightarrow D_{\frac{1}{2}} f \in V_{j+1} \\ \text{trivial intersection} \quad & \bigcap_{j\in\mathbb{Z}} V_j = \{0\} \\ \text{completeness} \quad & \overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}) \\ \text{orthonormal basis} \quad & \exists \phi \in V_0 : \{T_k \phi \ : \ k \in \mathbb{Z}\} \ \text{ is a CONS of } V_0 \end{aligned}$$

The function $\phi$ from the orthonormal basis property is called *generator* or *scaling function* of the MRA.

*Example* 4.9. For $j \in \mathbb{Z}$, let

$$V_j := \left\{ f \in L^2(\mathbb{R}) \ : \ f|_{[k2^j,(k+1)2^j)} \text{ is constant for all } k \in \mathbb{Z} \right\},$$

i. e. $V_j$ is the space of square integrable functions that are constant on the dyadic intervals $[k2^j, (k+1)2^j)$. Then, $(V_j)_{j\in\mathbb{Z}}$ is a MRA: The $V_j$ are obviously closed subspaces of $L^2(\mathbb{R})$. Translational invariance, inclusion and scaling are obviously fulfilled from the corresponding properties of the dyadic intervals. A function that is in the intersection of all $V_j$ has to be constant on $[0, \infty)$ and on $(-\infty, 0)$. The only $L^2$-function that fulfills this is the zero function. The completeness follows from the fact $C_c(\mathbb{R})$ is dense in $L^2(\mathbb{R})$ (Proposition B.8), continuous functions on compact sets are uniformly continuous and that uniformly continuous functions can be approximated in the supremum norm with functions constant on dyadic intervals. Finally, $\phi = \chi_{[0,1)}$ is a generator. The ONS property is obviously fulfilled and the completeness is shown based on the fact that the squared $L^2$-norm of an element of $V_0$ has to be the sum of the squared values of the element on each of the intervals $[k, k+1)$.

*Remark* 4.10. Let $(V_j)_{j\in\mathbb{Z}}$ be a MRA. Moreover, let $P_{V_j}$ denote the orthogonal projection from $L^2(\mathbb{R})$ to $V_j$. Note that the orthogonal projection exists since $L^2(\mathbb{R})$ is a Hilbert space and $V_j$ a nonempty, closed subspace of $L^2(\mathbb{R})$.

Let $f \in L^2(\mathbb{R})$. Then, we have

$$\lim_{j\to-\infty} P_{V_j} f = f \text{ and } \lim_{j\to\infty} P_{V_j} f = 0.$$

This means that we can use the nested spaces $V_j$ to approximate $f$, the smaller $j$, the better the approximation.

We will show the limit $j \to \infty$ in Remark 4.14. The limit $j \to -\infty$ is shown as follows. Since the orthogonal projection fulfills

$$\left\|f - P_{V_j} f\right\|_{L^2} = \inf_{u \in V_j} \|f - u\|_{L^2} \text{ for all } j \in \mathbb{Z},$$

the inclusion property implies, for all $j \in \mathbb{Z}$, that

$$\left\|f - P_{V_{j+1}} f\right\|_{L^2} = \inf_{u \in V_{j+1}} \|f - u\|_{L^2} \geq \inf_{u \in V_j} \|f - u\|_{L^2} = \left\|f - P_{V_j} f\right\|_{L^2} \geq 0.$$

Thus, $(\left\|f - P_{V_{-j}} f\right\|_{L^2})_{j \in \mathbb{N}}$ is a decreasing bounded sequence and thus convergent. Let $\epsilon > 0$ be arbitrary but fixed. The completeness and inclusion properties imply

$$L^2(\mathbb{R}) = \overline{\bigcup_{j \in \mathbb{Z}} V_j} = \overline{\bigcup_{j \leq -1} V_j}.$$

Therefore, there is a $u \in \bigcup_{j \leq -1} V_j$ with $\|f - u\|_{L^2} < \epsilon$. Since $u \in \bigcup_{j \leq -1} V_j$, there is a $k \geq 1$ such that $u \in V_{-k}$. Hence,

$$0 \leq \lim_{j \to \infty} \left\|f - P_{V_{-j}} f\right\|_{L^2} \leq \left\|f - P_{V_{-k}} f\right\|_{L^2} \leq \|f - u\|_{L^2} \leq \epsilon.$$

Since $\epsilon > 0$ was arbitrary, we get

$$\lim_{j \to \infty} \left\|f - P_{V_{-j}} f\right\|_{L^2} = 0 \;\Rightarrow\; P_{V_j} f \to f \text{ for } j \to -\infty.$$

*Remark* 4.11. Let $(V_j)_{j \in \mathbb{Z}}$ be a MRA with generator $\phi$. Moreover, let

$$\phi_{j,k}(x) := 2^{-\frac{j}{2}} \phi(2^{-j} x - k) \text{ for all } j, k \in \mathbb{Z} \text{ and } x \in \mathbb{R}.$$

The scaling property and the orthonormal basis property together imply that, for all $j \in \mathbb{Z}$,

$$\{\phi_{j,k} \,:\, k \in \mathbb{Z}\}$$

is an orthonormal basis of $V_j$ (exercise). Due to the inclusion property, we have $\phi \in V_0 \subset V_{-1}$, i. e. the generator is not only in $V_0$, but also in $V_{-1}$. Since $\{\phi_{-1,k} \,:\, k \in \mathbb{Z}\}$ is a CONS of $V_{-1}$, we have

$$\phi = \sum_{k \in \mathbb{Z}} h_k \phi_{-1,k} = \sum_{k \in \mathbb{Z}} h_k \sqrt{2} \phi(2 \cdot -k),$$

where $h_k = (\phi, \phi_{-1,k})_{L^2}$ for $k \in \mathbb{Z}$, cf. Remark 3.43 (note that this does not mean that the series is converging pointwise a.e.). This equation is called *scaling equation* and the reason why $\phi$ is also called scaling function (cf. Definition 4.8).

For $f \in L^2(\mathbb{R})$, we have

$$(f, \phi_{j,k})_{L^2} = \int_{\mathbb{R}} f(x) 2^{-\frac{j}{2}} \phi(2^{-j} x - k) \,\mathrm{d}x = \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{2^j}} \phi\left(\frac{x - 2^j k}{2^j}\right) \mathrm{d}x = L_\phi f(2^j, 2^j k).$$

Thus, $(f, \phi_{j,k})_{L^2}$ is the evaluation of the continuous wavelet transform $L_\phi f$ for $a = 2^j$ and $b = 2^j k$.

**Lemma 4.12.** *Let $h_k$ be the coefficients of the scaling equation of an MRA. Then,*

$$\sum_{k\in\mathbb{Z}} h_k h_{k+2m} = \delta_{0,m} \text{ for all } m \in \mathbb{Z}.$$

**Proof.** Using the orthonormal basis property and the scaling equation, we get

$$\begin{aligned}\delta_{0,m} = (\phi, \phi(\cdot + m))_{L^2} &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} h_k h_l (\sqrt{2}\phi(2\cdot -k), \sqrt{2}\phi(2\cdot -l+2m))_{L^2} \\ &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} h_k h_l (\phi_{-1,k}, \phi_{-1,l-2m})_{L^2} = \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} h_k h_l \delta_{k,l-2m} = \sum_{k\in\mathbb{Z}} h_k h_{k+2m}.\end{aligned}$$

□

**Definition 4.13.** Let $(V_j)_{j\in\mathbb{Z}}$ be a MRA. Moreover, let $W_j$ be such that

$$V_{j-1} = V_j \oplus W_j \text{ and } V_j \perp W_j,$$

i. e. $W_j$ is the orthogonal complement of $V_j$ in $V_{j-1}$. Then, $V_j$ is called *approximation space to the scale* $j$ and $W_j$ is called *detail space* or *wavelet space to the scale* $j$.

*Remark* 4.14. The definition of $W_j$ implies by iteration that

$$V_j = V_{j+1} \oplus W_{j+1} = V_{j+2} \oplus W_{j+2} \oplus W_{j+1} = \ldots = V_{j+M} \oplus \left(\bigoplus_{m=j+1}^{j+M} W_m\right) \text{ for all } M \geq 1.$$

This immediately implies

$$V_j \supset \bigoplus_{m\geq j+1} W_m.$$

Assume there is $v \in V_j \setminus \left(\bigoplus_{m\geq j+1} W_m\right)$. Since

$$V_j \setminus \left(\bigoplus_{m\geq j+1} W_m\right) = \left(V_{j+M} \oplus \left(\bigoplus_{m=j+1}^{j+M} W_m\right)\right) \setminus \left(\bigoplus_{m\geq j+1} W_m\right) = V_{j+M} \setminus \left(\bigoplus_{m\geq j+1+M} W_m\right),$$

we have, $v \in V_{j+M}$ for all $M \geq 1$. Due to the inclusion property, this implies $v \in V_M$ for all $M \in \mathbb{Z}$, which implies $v = 0$ due to the trivial intersection property. This is a contradiction to $v \notin \bigoplus_{m\geq j+1} W_m \supset \{0\}$. Thus, we have shown

$$V_j = \bigoplus_{m\geq j+1} W_m.$$

Combined with the completeness, we have

$$L^2(\mathbb{R}) = \overline{\bigcup_{j\in\mathbb{Z}} V_j} = \overline{\bigcup_{j\in\mathbb{Z}} \bigoplus_{m\geq j+1} W_m} = \overline{\bigoplus_{m\in\mathbb{Z}} W_m}.$$

Since $V_j \oplus W_j$ is an orthogonal decomposition of $V_{j-1}$, we have

$$P_{V_{j-1}} = P_{V_j} + P_{W_j} \;\Rightarrow\; P_{W_j} = P_{V_{j-1}} - P_{V_j}.$$

Due to

$$L^2(\mathbb{R}) = \overline{\bigoplus_{m\in\mathbb{Z}} W_m} = \overline{\bigoplus_{m\geq j+1} W_m \oplus \bigoplus_{m\leq j} W_m} = \overline{V_j \oplus \bigoplus_{m\leq j} W_m},$$

$f \in L^2(\mathbb{R})$ can be expressed by

$$f = \sum_{m\in\mathbb{Z}} P_{W_m} f = P_{V_j} f + \sum_{m\leq j} P_{W_m} f.$$

This implies that $P_{V_j} f \to 0$ for $j \to \infty$ holds, which we claimed earlier.

*Example* 4.15. Let $(V_j)_{j\in\mathbb{Z}}$ the piecewise constant MRA from Example 4.9 and let $f \in L^2(\mathbb{R})$. Since the elements of $V_j$ are constant on the intervals $[k2^j, (k+1)2^j)$ and $\left\|f - P_{V_j} f\right\|_{L^2} = \inf_{u\in V_j} \|f-u\|_{L^2}$, we get for $x \in [k2^j, (k+1)2^j)$ that

$$\begin{aligned}(P_{V_j} f)(x) &= 2^{-j} \int_{k2^j}^{(k+1)2^j} f(y)\,\mathrm{d}y = 2^{-j} \int_{\mathbb{R}} \chi_{[k2^j,(k+1)2^j)}(y) f(y)\,\mathrm{d}y \\ &= 2^{-j} \int_{\mathbb{R}} \chi_{[k,k+1)}(2^{-j}y) f(y)\,\mathrm{d}y = 2^{-\frac{j}{2}} \int_{\mathbb{R}} 2^{-\frac{j}{2}} \chi_{[0,1)}(2^{-j}y - k) f(y)\,\mathrm{d}y \\ &= 2^{-\frac{j}{2}} \int_{\mathbb{R}} \phi_{j,k}(y) f(y)\,\mathrm{d}y = (f, \phi_{j,k})_{L^2} \phi_{j,k}(x).\end{aligned}$$

Noting that for any $x \in \mathbb{R}$, we have $\phi_{j,k}(x) = 0$ for all $k \in \mathbb{Z}$ but one, we get

$$P_{V_j} f = \sum_{k\in\mathbb{Z}} (f, \phi_{j,k})_{L^2} \phi_{j,k}.$$

Moreover, we have for this MRA that

$$\begin{aligned}h_k = (\phi, \phi_{-1,k})_{L^2} &= \int_0^1 2^{\frac{1}{2}} \phi(2x-k)\,\mathrm{d}x = \int_0^1 \sqrt{2}\chi_{[k/2,(k+1)/2)}(x)\,\mathrm{d}x = \begin{cases} \frac{\sqrt{2}}{2} & k \in \{0,1\} \\ 0 & \text{else} \end{cases} \\ &= 2^{-\frac{1}{2}}(\delta_{k,0} + \delta_{k,1}).\end{aligned}$$

Thus, in this case, the scaling equation simplifies to

$$\phi = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\phi(2\cdot -k) = \phi(2\cdot) + \phi(2\cdot -1).$$

Composing this with $x \mapsto 2^{-j}x - k$, this shows

$$\phi(2^{-j}\cdot -k) = \phi(2^{-(j-1)}\cdot -2k) + \phi(2^{-(j-1)}\cdot -(2k+1)).$$

Multiplying both sides with $2^{-\frac{j}{2}}$ shows

$$\phi_{j,k} = \frac{1}{\sqrt{2}}(\phi_{j-1,2k} + \phi_{j-1,2k+1}). \tag{$*$}$$

This allows us to compute the projection to $W_{j+1}$. We get

$$P_{W_{j+1}} f = P_{V_j} f - P_{V_{j+1}} f = \sum_{k\in\mathbb{Z}} (f, \phi_{j,k})_{L^2} \phi_{j,k} - \sum_{k\in\mathbb{Z}} (f, \phi_{j+1,k})_{L^2} \phi_{j+1,k}$$

Splitting the first sum in even and odd indices and using $(*)$ for the second sum, leads to

$$
\begin{aligned}
P_{W_{j+1}} f = & \sum_{k\in\mathbb{Z}} (f, \phi_{j,2k})_{L^2} \phi_{j,2k} + \sum_{k\in\mathbb{Z}} (f, \phi_{j,2k+1})_{L^2} \phi_{j,2k+1} \\
& - \frac{1}{2} \sum_{k\in\mathbb{Z}} (f, \phi_{j,2k} + \phi_{j,2k+1})_{L^2} (\phi_{j,2k} + \phi_{j,2k+1}) \\
= & \frac{1}{2} \sum_{k\in\mathbb{Z}} (f, \phi_{j,2k} - \phi_{j,2k+1})_{L^2} (\phi_{j,2k} - \phi_{j,2k+1}),
\end{aligned}
$$

where, for the last equality, we have used that

$$(c, a)_X a + (c, b)_X b - \tfrac{1}{2}(c, a + b)_X (a + b) = \tfrac{1}{2}(c, a - b)_X (a - b)$$

holds for any scalar product, as a generalization of $a^2 + b^2 - \frac{1}{2}(a + b)^2 = \frac{1}{2}(a - b)^2$. Introducing

$$\psi(x) \coloneqq \phi(2x) - \phi(2x - 1) \text{ and } \psi_{j,k}(x) \coloneqq 2^{-\frac{j}{2}} \psi(2^{-j} x - k),$$

we get $\psi_{j+1,k} = \frac{1}{\sqrt{2}} (\phi_{j,2k} - \phi_{j,2k+1})$ and the representation of $P_{W_{j+1}} f$ simplifies to

$$P_{W_{j+1}} f = \sum_{k\in\mathbb{Z}} (f, \psi_{j+1,k})_{L^2} \psi_{j+1,k}.$$

The function $\psi$ is the Haar wavelet we have seen before in Example 4.7 (iii). Moreover, we have

$$
\begin{aligned}
(\psi_{j,k}, \psi_{j,m})_{L^2} &= \tfrac{1}{2}((\phi_{j-1,2k} - \phi_{j-1,2k+1}), (\phi_{j-1,2m} - \phi_{j-1,2m+1}))_{L^2} \\
&= \tfrac{1}{2}(\delta_{k,m} + 0 + 0 + \delta_{k,m}) = \delta_{k,m}.
\end{aligned}
$$

Combined with the representation of $P_{W_{j+1}}$ we derived above and that $P_{W_{j+1}}$ is the identity on $W_{j+1}$, this means that $\{\psi_{j,k} : k \in \mathbb{Z}\}$ is an orthonormal basis of $W_j$.

The existence of a wavelet that generates an orthonormal basis for $W_j$ not only holds in the special case here, but more generally.

**Proposition 4.16.** *Let $(V_j)_{j\in\mathbb{Z}}$ be a MRA with generator $\phi$ and let $h_k$ be the coefficients of the scaling equation. Moreover, let $\psi \in V_{-1}$ be defined by*

$$\psi \coloneqq \sum_{k\in\mathbb{Z}} (-1)^k h_{1-k} \phi_{-1,k}.$$

*Then,*

*(i)* $\{\psi_{j,k} : k \in \mathbb{Z}\}$ *is an orthonormal basis of* $W_j$.

*(ii)* $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$ *is an orthonormal basis of* $L^2(\mathbb{R})$.

*(iii)* $\psi$ *is a wavelet with* $c_\psi = \ln 2$.

**Proof.** We first show that $\psi \in W_0$. Using the definition of $\psi$ and the scaling equation, we get

$$\begin{aligned}
(\psi, \phi(\cdot - n))_{L^2} &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}}(-1)^k h_{1-k}h_l(\sqrt{2}\phi(2\cdot - k), \sqrt{2}\phi(2\cdot - l - 2n))_{L^2} \\
&= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}}(-1)^k h_{1-k}h_l(\phi_{-1,k}, \phi_{-1,2n+l})_{L^2} = \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}}(-1)^k h_{1-k}h_l\delta_{k,2n+l} \\
&= \sum_{l\in\mathbb{Z}}(-1)^{2n+l}h_{1-2n-l}h_l = \sum_{l\in\mathbb{Z}}(-1)^l h_{1-2n-l}h_l \\
&\overset{\text{even/odd split}}{=} \sum_{l\in\mathbb{Z}} h_{1-2n-2l}h_{2l} - \sum_{l\in\mathbb{Z}} h_{1-2n-2l-1}h_{2l+1} = \sum_{l\in\mathbb{Z}} h_{1-2(n+l)}h_{2l} - \sum_{l\in\mathbb{Z}} h_{-2(n+l)}h_{2l+1} \\
&\overset{k=-(n+l)}{=} \sum_{k\in\mathbb{Z}} h_{1+2k}h_{-2(n+k)} - \sum_{l\in\mathbb{Z}} h_{-2(n+l)}h_{2l+1} = 0.
\end{aligned}$$

Since $(T_k\phi)$ is a CONS of $V_0$, this show that $\psi \in (V_0)^\perp$. Combined with $\psi \in V_{-1} = V_0 \oplus W_0$, this shows $\psi \in W_0$. Using a similar computation and Lemma 4.12, one shows that

$$(\psi(\cdot - k), \psi(\cdot - l))_{L^2} = \delta_{k,l}.$$

Thus, $\{\psi_{0,k} \,:\, k \in \mathbb{Z}\}$ is an ONS. Noting that

$$(\psi(\cdot - m), \phi(\cdot - n))_{L^2} = (\psi, \phi(\cdot - (n - m)))_{L^2} = 0,$$

shows that $\{\phi_{0,k}, \psi_{0,k} \,:\, k \in \mathbb{Z}\}$ is also an ONS.

Since $\psi \in W_0$, $\phi \in V_0$ and $V_{-1} = V_0 \oplus W_0$, it is sufficient to show that $\{\phi_{0,k}, \psi_{0,k} \,:\, k \in \mathbb{Z}\}$ is complete in $V_{-1}$ to show that $\{\psi_{0,k} \,:\, k \in \mathbb{Z}\}$ is complete in $W_0$. To show the former, it is sufficient to show that $\phi_{-1,0}$ can be expressed with $\{\phi_{0,k}, \psi_{0,k} \,:\, k \in \mathbb{Z}\}$. Using the scaling equation and the definition of $\psi$ gives

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} \left(|(\phi_{-1,0}, \phi_{0,k})_{L^2}|^2 + |(\phi_{-1,0}, \psi_{0,k})_{L^2}|^2\right) \\
&= \sum_{k\in\mathbb{Z}} \left( \Big| \sum_{l\in\mathbb{Z}} h_l \underbrace{(\phi_{-1,0}, \phi_{-1,l+2k})_{L^2}}_{=\delta_{0,l+2k}} \Big|^2 + \Big| \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} \underbrace{(\phi_{-1,0}, \phi_{-1,l+2k})_{L^2}}_{\delta_{0,l+2k}} \Big|^2 \right) \\
&= \sum_{k\in\mathbb{Z}} \left( |h_{-2k}|^2 + |(-1)^{-2k}h_{1+2k}|^2 \right) = \sum_{k\in\mathbb{Z}} \left(|h_{-2k}|^2 + |h_{1+2k}|^2\right) = \sum_{k\in\mathbb{Z}} |h_k|^2 \overset{\text{Lemma 4.12}}{=} 1 \\
&= \|\phi_{-1,0}\|_{L^2}^2 .
\end{aligned}$$

Thus, Corollary 3.40 implies

$$\lim_{K\to\infty} \left\| \sum_{k=-K}^{K} ((\phi_{-1,0}, \phi_{0,k})_{L^2}\phi_{0,k} + (\phi_{-1,0}, \psi_{0,k})_{L^2}\psi_{0,k}) - \phi_{-1,0} \right\|_{L^2} = 0.$$

Thus, we have shown that, for $j = 0$, $\{\psi_{j,k} \,:\, k \in \mathbb{Z}\}$ is an orthonormal basis of $W_j$. Fortunately, the proof for $j = 0$ can be reused to prove the statement also for $j \in \mathbb{Z}$. First, we note that (as we will show in Lemma 4.18)

$$\psi_{j,0} = \sum_{k\in\mathbb{Z}} (-1)^k h_{1-k}\phi_{j-1,k}.$$

Moreover, we have

$$
\begin{aligned}
h_k = (\phi, \phi_{-1,k})_{L^2} &= \int_{\mathbb{R}} \phi(x) 2^{\frac{1}{2}} \phi(2x-k)\,\mathrm{d}x \\
&= \int_{\mathbb{R}} \phi(\varphi(2^{-j}x)) 2^{\frac{1}{2}} \phi(\varphi(2^{-j+1}x)-k) 2^{-j} \varphi'(x)\,\mathrm{d}x \ \text{ with } \varphi(x) \coloneqq 2^j x \\
&= \int_{\mathbb{R}} 2^{-\frac{j}{2}} \phi(2^{-j}x) 2^{\frac{-j+1}{2}} \phi(2^{-j+1}x-k)\,\mathrm{d}x = \int_{\mathbb{R}} \phi_{j,0}(x)\phi_{j-1,k}(x)\,\mathrm{d}x = (\phi_{j,0}, \phi_{j-1,k})_{L^2}.
\end{aligned}
$$

Finally, we recall that $\{\phi_{j,k} : k \in \mathbb{Z}\}$ is an orthonormal basis of $V_j$. Thus, for any $j \in \mathbb{Z}$, we can consider the MRA given by $\tilde{V}_i \coloneqq V_{i+j}$ with generator $\tilde{\phi} \coloneqq \phi_{j,0}$ to conclude from the above that $\{\tilde{\psi}_{0,k} : k \in \mathbb{Z}\} = \{\psi_{j,k} : k \in \mathbb{Z}\}$ is an orthonormal basis of $\tilde{W}_0 = W_j$, which shows the first statement. The second statement follows from the first and $L^2(\mathbb{R}) = \overline{\bigoplus_{m\in\mathbb{Z}} W_m}$.

The proof for the last statement, i. e. $c_\psi = \ln 2$, can be found in the book of Louis, Maaß and Rieder[3], cf. “Satz 2.2.10” in there. Note that the continuous wavelet transform and thus $c_\psi$ is defined slightly differently there, i. e. using $a \in \mathbb{R}$ instead of $a \in (0, \infty)$. This causes the $c_\psi$ used in that book to be bigger by a factor of 2. □

*Example* 4.17.

(i) Let $(V_j)_{j\in\mathbb{Z}}$ the piecewise constant MRA from Example 4.9. Recalling, $h_k = 2^{-\frac{1}{2}}(\delta_{k,0}+\delta_{k,1})$, we get

$$
\begin{aligned}
\psi &= \sum_{k\in\mathbb{Z}} (-1)^k h_{1-k} \phi_{-1,k} = \sum_{k\in\mathbb{Z}} (-1)^k 2^{-\frac{1}{2}} (\delta_{1-k,0} + \delta_{1-k,1}) \phi_{-1,k} \\
&= (-1)^1 2^{-\frac{1}{2}} \phi_{-1,1} + (-1)^0 2^{-\frac{1}{2}} \phi_{-1,0} = -\phi(2\cdot -1) + \phi(2\cdot) \\
&= -\chi_{[0,1)}(2\cdot -1) + \chi_{[0,1)}(2\cdot) = -\chi_{[\frac{1}{2},1)} + \chi_{[0,\frac{1}{2})},
\end{aligned}
$$

which is again the Haar wavelet.

(ii) Constructing wavelets that both have compact support and have some smoothness is highly non-trivial. The discovery of compact, regular and orthogonal wavelets by Ingrid Daubechies in 1988 is one of the reasons why wavelets became very popular in the 90s-era. The Daubechies wavelets are a family of orthogonal wavelets that are characterized by a maximal number $k$ of vanishing moments, i. e.

$$
\int_{\mathbb{R}} x^l \psi(x)\,\mathrm{d}x = 0 \text{ for } l \in \{0, \dots, k-1\},
$$

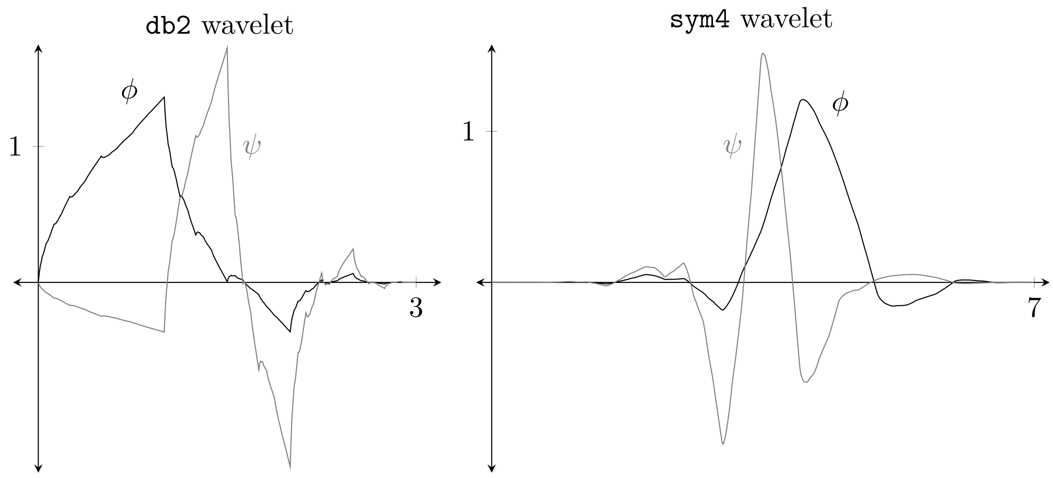
for a given support. Perhaps the most famous wavelet pair $\phi$, $\psi$ is the Daubechies wavelet with two vanishing moments, denoted by `db2`. Interestingly, the Daubechies wavelet with one vanishing moment, `db1`, is equivalent to the Haar wavelet.

The values of the coefficients $h_k$ are available as tabulated numerical values. For `db2`, the values are also available analytically:

$$
h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}},\ h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}},\ h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}},\ h_4 = \frac{1-\sqrt{3}}{4\sqrt{2}},\ h_k = 0 \text{ for } k \in \mathbb{Z}\backslash\{0,1,2,3\}
$$

[3]A. K. Louis, P. Maaß, and A. Rieder. *Wavelets - Theorie und Anwendungen.* Vieweg+Teubner Verlag, 1998.

(iii) A modification of the Daubechies wavelets are the *symlets*, which are sometimes also called Daubechies' least-asymmetric wavelets. As the name suggests, these wavelets are more symmetric than the Daubechies wavelets and also have vanishing moments. For instance, the symlet with four vanishing moments is denoted by `sym4`. Note that `db2` and `sym2` are equivalent.




## 4.3. The fast wavelet transform

Due to Proposition 4.16, we know that $\psi$ constructed from an MRA is a wavelet and $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. In particular, we have

$$f = \sum_{j,k\in\mathbb{Z}} (f, \psi_{j,k})_{L^2} \psi_{j,k}.$$

Thus, we need to compute the coefficients $(f, \psi_{j,k})_{L^2}$ to compute the wavelet transform of $f$. The key to a fast wavelet transform is to use that those coefficients can be computed recursively, which is a consequence of the scaling equation and the definition of $\psi$:

**Lemma 4.18.** *Let $(V_j)_{j\in\mathbb{Z}}$ be a MRA with generator $\phi$, let $h_k$ be the coefficients of the scaling equation and let $\psi \in V_{-1}$ be the wavelet from Proposition 4.16. Then, for all $j,k \in \mathbb{Z}$, we have*

$$\phi_{j,k} = \sum_{l\in\mathbb{Z}} h_l \phi_{j-1,l+2k},$$
$$\psi_{j,k} = \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} \phi_{j-1,l+2k}.$$

*Moreover, for all $f \in L^2(\mathbb{R})$ and all $j,k \in \mathbb{Z}$, we have*

$$(f, \phi_{j,k})_{L^2} = \sum_{l\in\mathbb{Z}} h_l (f, \phi_{j-1,l+2k})_{L^2},$$
$$(f, \psi_{j,k})_{L^2} = \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} (f, \phi_{j-1,l+2k})_{L^2}.$$

**Proof.** Using the definition of $\phi_{j,k}$ and the scaling equation, we get

$$\begin{aligned}\phi_{j,k} &= 2^{-\frac{j}{2}}\phi(2^{-j}\cdot - k) = 2^{-\frac{j}{2}}\sum_{l\in\mathbb{Z}} h_l\sqrt{2}\phi(2(2^{-j}\cdot - k) - l)\\ &= \sum_{l\in\mathbb{Z}} h_l 2^{-\frac{j+1}{2}}\phi(2^{-j+1}\cdot - 2k - l) = \sum_{l\in\mathbb{Z}} h_l\phi_{j-1,l+2k}.\end{aligned}$$

Moreover, with the definition of $\psi_{j,k}$ and $\psi$, we get

$$\begin{aligned}\psi_{j,k} &= 2^{-\frac{j}{2}}\psi(2^{-j}\cdot - k) = 2^{-\frac{j}{2}}\sum_{l\in\mathbb{Z}}(-1)^l h_{1-l}\phi_{-1,l}(2^{-j}\cdot - k)\\ &= \sum_{l\in\mathbb{Z}}(-1)^l h_{1-l}2^{\frac{-j+1}{2}}\phi(2^{-j+1}\cdot - 2k - l) = \sum_{l\in\mathbb{Z}}(-1)^l h_{1-l}\phi_{j-1,l+2k}.\end{aligned}$$

The equations for the scalar products follow directly from the equations for $\phi_{j,k}$, and $\psi_{j,k}$. □

*Remark* 4.19. Starting with the values $(f,\phi_{j,k})_{L^2}$, the lemma above supplies recursion formulas to compute the coefficients on coarser scales, i. e. scales larger than $j$. Abbreviating the coefficient sequences by

$$c^j := ((f,\phi_{j,k})_{L^2})_{k\in\mathbb{Z}} \text{ and } d^j := ((f,\psi_{j,k})_{L^2})_{k\in\mathbb{Z}} \text{ for all } j\in\mathbb{Z},$$

we get

$$c_k^{j+1} = \sum_{l\in\mathbb{Z}} h_l c_{l+2k}^j \text{ and } d_k^{j+1} = \sum_{l\in\mathbb{Z}}(-1)^l h_{1-l}c_{l+2k}^j.$$

The fast wavelet transform now uses this to compute from $P_{V_j}f$ (given in form of the coefficient vector $c^j$) the coarser projection $P_{V_{j+1}}f$ (as the coefficient vector $c^{j+1}$) and the wavelet component $P_{W_{j+1}}f$ (as the coefficient vector $d^{j+1}$). Note that the sums are finite in case the scaling equation coefficient sequence only has finitely many nonzero entries. If only few are nonzero (e. g. for the Haar wavelet only $h_0$ and $h_1$ are nonzero), only few computations have to be done when going from $j$ to $j+1$.

Finally, we need to be able to efficiently invert the (fast) wavelet transform, i. e. we need to be able to compute $P_{V_j}f$ from the coarser projection $P_{V_{j+1}}f$ and the details $P_{W_{j+1}}f$. The next lemma answers how this can be done.

**Lemma 4.20.** *Assumptions and notations are the same as in Lemma 4.18. Moreover, let $f\in L^2(\mathbb{R})$. Then, for $c^j = ((f,\phi_{j,k})_{L^2})_{k\in\mathbb{Z}}$ and $d^j = ((f,\psi_{j,k})_{L^2})_{k\in\mathbb{Z}}$, we have*

$$c_k^j = \sum_{l\in\mathbb{Z}} c_l^{j+1}h_{k-2l} + \sum_{l\in\mathbb{Z}} d_l^{j+1}(-1)^{k-2l}h_{1-(k-2l)}.$$

**Proof.** Recalling Remark 4.14, we have $P_{V_j} = P_{V_{j+1}} + P_{W_{j+1}}$. Moreover, since $\{\phi_{j,k} : k\in\mathbb{Z}\}$

is an orthonormal basis of $V_j$ and $\{\psi_{j,k} : k \in \mathbb{Z}\}$ is an orthonormal basis of $W_j$, we get

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} c_k^j \phi_{j,k} &= P_{V_j} f = P_{V_{j+1}} f + P_{W_{j+1}} f = \sum_{k\in\mathbb{Z}} c_k^{j+1} \phi_{j+1,k} + \sum_{k\in\mathbb{Z}} d_k^{j+1} \psi_{j+1,k} \\
&\overset{\text{Lemma 4.18}}{=} \sum_{k\in\mathbb{Z}} c_k^{j+1} \sum_{l\in\mathbb{Z}} h_l \phi_{j,l+2k} + \sum_{k\in\mathbb{Z}} d_k^{j+1} \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} \phi_{j,l+2k} \\
&= \sum_{k\in\mathbb{Z}} c_k^{j+1} \sum_{l\in\mathbb{Z}} h_{l-2k} \phi_{j,l} + \sum_{k\in\mathbb{Z}} d_k^{j+1} \sum_{l\in\mathbb{Z}} (-1)^{l-2k} h_{1-(l-2k)} \phi_{j,l} \\
&= \sum_{l\in\mathbb{Z}} \left( \sum_{k\in\mathbb{Z}} c_k^{j+1} h_{l-2k} + \sum_{k\in\mathbb{Z}} d_k^{j+1} (-1)^{l-2k} h_{1-(l-2k)} \right) \phi_{j,l}
\end{aligned}$$

Now, by comparing the coefficients of the ONS elements $\phi_{j,l}$, the statement follows. □

*Remark* 4.21. To get a compact notation of the transform using the recursion formulas from Remark 4.19, we introduce the mappings

$$\begin{aligned}
&H : l^2(\mathbb{R}) \to l^2(\mathbb{R}), (Hc)_k = \sum_{l\in\mathbb{Z}} h_l c_{l+2k}, \\
&G : l^2(\mathbb{R}) \to l^2(\mathbb{R}), (Gc)_k = \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} c_{l+2k}.
\end{aligned}$$

Here, $l^2(\mathbb{R})$ is the space of square summable sequences, i. e.

$$l^2(\mathbb{R}) = \left\{ (x_k)_{k\in\mathbb{Z}} \subset \mathbb{R} : \sum_{k\in\mathbb{Z}} |x_k|^2 < \infty \right\}.$$

Then, the fast wavelet transform is described by the following function:

```
function FWT(c^0 ∈ l^2(R), decomposition depth M)
    for m = 1 to M do
        compute d^m = G c^{m-1}, c^m = H c^{m-1}.
    end for
    return d^1, ..., d^M, c^M
end function
```

For $c, d \in l^2(\mathbb{R})$, we have

$$\begin{aligned}
\sum_{l\in\mathbb{Z}} (H^*c)_l d_l &= (H^*c, d)_{l^2} = (c, Hd)_{l^2} = \sum_{k\in\mathbb{Z}} c_k (Hd)_k = \sum_{k\in\mathbb{Z}} c_k \sum_{l\in\mathbb{Z}} h_l d_{l+2k} \\
&= \sum_{k\in\mathbb{Z}} c_k \sum_{l\in\mathbb{Z}} h_{l-2k} d_l = \sum_{l\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} h_{l-2k} c_k d_l.
\end{aligned}$$

Thus, the adjoint of $H$ is

$$H^* : l^2(\mathbb{R}) \to l^2(\mathbb{R}), (H^*c)_l = \sum_{k\in\mathbb{Z}} h_{l-2k} c_k.$$

Analogously, one shows that the adjoint of $G$ is

$$G^* : l^2(\mathbb{R}) \to l^2(\mathbb{R}), (G^*d)_l = \sum_{k\in\mathbb{Z}} (-1)^{l-2k} h_{1-(l-2k)} d_k.$$

Using this and Lemma 4.20, the inverse fast wavelet transform is described by the following function:

**function** IFWT($d^1, \dots, d^M, c^M \in l^2(\mathbb{R})$)
  **for** $m = M-1$ to $0$ **do**
    compute $c^m = H^*c^{m+1} + G^*d^{m+1}$
  **end for**
  **return** $c^0$
**end function**

*Remark* 4.22. $H$ and $G$ can be expressed as convolutions. For $c \in l^2(\mathbb{R})$, we have

$$(Hc)_k = \sum_{l\in\mathbb{Z}} h_l c_{l+2k} = \sum_{l\in\mathbb{Z}} h_{l-2k} c_l = (c * D_{-1}h)_{2k}$$

where

$$* : l^2(\mathbb{R}) \times l^2(\mathbb{R}) \to l^2(\mathbb{R}), (c,d) \mapsto (c*d) := \left(\sum_{l\in\mathbb{Z}} c_l d_{k-l}\right)_{k\in\mathbb{Z}}$$

is the convolution on $l^2(\mathbb{R})$ and

$$D_{-1} : l^2(\mathbb{R}) \to l^2(\mathbb{R}), c \mapsto (c_{-k})_{k\in\mathbb{Z}}.$$

Moreover, we have

$$(Gc)_k = \sum_{l\in\mathbb{Z}} (-1)^l h_{1-l} c_{l+2k} = \sum_{l\in\mathbb{Z}} (-1)^{l-2k} h_{1-(l-2k)} c_l = (c * D_{-1}g)_{2k},$$

where $g_l := (-1)^l h_{1-l}$. Note that only the even entries of the convolutions $c * D_{-1}h$ and $c * D_{-1}g$ are needed. This is effectively a downsampling by a factor of 2.

Also $H^*$ and $G^*$ can be expressed as convolutions. Instead of a downsampling, we need an upsampling though. To this end, let

$$U : l^2(\mathbb{R}) \to l^2(\mathbb{R}), (Uc)_k = \begin{cases} c_{k/2} & k \text{ even} \\ 0 & \text{else} \end{cases}.$$

Then, we have

$$(H^*c)_l = \sum_{k\in\mathbb{Z}} h_{l-2k} c_k = \sum_{k\in\mathbb{Z}} h_{l-k} (Uc)_k = (Uc * h)_l.$$

Analogously, we get

$$(G^*d)_l = \sum_{k\in\mathbb{Z}} (-1)^{l-2k} h_{1-(l-2k)} d_k = \sum_{k\in\mathbb{Z}} (-1)^{l-k} h_{1-(l-k)} (Ud)_k = (Ud * g)_l,$$

with $g$ as above.

*Remark* 4.23. To compute the FWT, we still need the coefficients $c^j = ((f, \phi_{j,k})_{L^2})_{k\in\mathbb{Z}}$ for one $j \in \mathbb{Z}$, i. e. the representation of our signal $f$ on an initial scale $j$. Usually, $f$ is not available analytically, so we cannot just compute $c^j$. Thus, we need a way to estimate $c^j$ from available data.

Assuming $f \in C_c^1(\mathbb{R})$, we get

$$\begin{aligned}
(f, \phi_{j,k})_{L^2} &= \int_{\mathbb{R}} f(x) 2^{-\frac{j}{2}} \phi(2^{-j}x - k)\,\mathrm{d}x \\
&= \int_{\mathbb{R}} f(\varphi(2^j x + 2^j k)) \phi(\varphi(x)) 2^{\frac{j}{2}} \varphi'(x)\,\mathrm{d}x && (\varphi(x) := 2^{-j}x - k) \\
&= 2^{\frac{j}{2}} \int_{\mathbb{R}} f(2^j x + 2^j k)\phi(x)\,\mathrm{d}x \overset{\text{Taylor}}{=} 2^{\frac{j}{2}} \int_{\mathbb{R}} \left( f(2^j k) + f'(\zeta(x)) 2^j x \right) \phi(x)\,\mathrm{d}x \\
&= 2^{\frac{j}{2}} f(2^j k) \int_{\mathbb{R}} \phi(x)\,\mathrm{d}x + 2^j \int_{\mathbb{R}} f'(\zeta(x)) x \phi(x)\,\mathrm{d}x\,.
\end{aligned}$$

Thus, using the notation $C := \int_{\mathbb{R}} \phi(x)\,\mathrm{d}x$, we get

$$\begin{aligned}
\left| (f, \phi_{j,k})_{L^2} - 2^{\frac{j}{2}} C f(2^j k) \right| &\le 2^j \int_{\mathbb{R}} \left| f'(\zeta(x)) x \phi(x) \right| \mathrm{d}x \overset{\text{Hölder}}{\le} 2^j \sqrt{\int_{\mathbb{R}} \left| f'(\zeta(x)) x \right|^2 \mathrm{d}x}\, \|\phi\|_{L^2} \\
&\le 2^j \left\| f' \right\|_{L^\infty} \sqrt{\int_{\operatorname{supp} f} |x|^2\,\mathrm{d}x} \le 2^j \left\| f' \right\|_{L^\infty} \sqrt{2} R^{\frac{3}{2}},
\end{aligned}$$

where $R > 0$ is such that $\operatorname{supp} f \subset [-R, R]$, which exists since $f$ has compact support. In other words, we have shown

$$(f, \phi_{j,k})_{L^2} = 2^{\frac{j}{2}} C f(2^j k) + \mathcal{O}(2^j).$$

Thus, for $-j$ large enough, $2^{\frac{j}{2}} C f(2^j k)$ is a good approximation of $(f, \phi_{j,k})_{L^2}$. This means it is sufficient, if we can sample our signal $f$ at equidistant points as long as the distance between neighboring sample points is sufficiently small.

In practice, discrete signals are vectors of finite lengths. Thus, we have to replace the convolution on $l^2(\mathbb{R})$ with a convolution on $\mathbb{R}^N$, which requires an extension of our input vectors. To this end, we can use one of the extensions we have used before in this context, e. g. periodic extension, zero extension, symmetric extension, etc.

Let $N = 2^M$. If $c \in \mathbb{R}^N$ and $h \in \mathbb{R}^n$, then computing a single coefficient $(Hc)_k = (c * D_{-1}h)_{2k}$ is $\mathcal{O}(n)$ (just computing this straight with the definition of the convolution, not with the fast convolution for which we would need to extend $h$ to length $N$). The same is true for the coefficients $(Gc)_k$. Both $Hc$ and $Gc$ are (due to the downsampling) vectors of length $N/2$ and thus can be computed with $\mathcal{O}(nN)$. To compute the coefficients for the next coarser scale, we only have an input vector of length $N/2$. Since

$$\sum_{j=0}^{M-1} \frac{1}{2^j} = \frac{1 - \frac{1}{2^M}}{1 - \frac{1}{2}} < \frac{1}{\frac{1}{2}} = 2,$$

the FWT for decomposition depth $M$ can be computed with $\mathcal{O}(nN)$, which is even faster than the FFT, if $n$ is small. Similarly, one can show that the IFWT is also $\mathcal{O}(nN)$.

*Remark* 4.24 (2D discrete wavelet transform). There are numerous ways to construct a two-dimensional wavelet transform. The easiest way is to extend the one-dimensional wavelet transform by using the tensor product. Let $(V_j)_{j \in \mathbb{Z}}$ be a MRA with generator $\phi$ and let $\psi$ be the corresponding wavelet from Proposition 4.16. By taking the tensor product of $V_j$ with itself, i. e.

$$V_j^2 = V_j \otimes V_j \subset L^2(\mathbb{R}^2),$$

we get an MRA of $L^2(\mathbb{R}^2)$. Note that a 2D MRA is defined like the 1D MRA expect that we need to consider $k \in \mathbb{Z}^2$ instead of $k \in \mathbb{Z}$. To show that $(V_j^2)_{j\in\mathbb{Z}}$ is a MRA, one shows that

$$\left\{\Phi_{j,k} \,:\, k \in \mathbb{Z}^2\right\}$$

is an orthonormal basis of $V_j^2$, where

$$\Phi_{j,k} : \mathbb{R}^2 \to \mathbb{R}, (x_1, x_2) \mapsto \phi_{j,k_1}(x_1)\phi_{j,k_2}(x_2) \text{ for all } j \in \mathbb{Z}, k \in \mathbb{Z}^2.$$

The decomposition of $V_j^2$ into an approximation space and a detail space has more structure than in 1D. Using the 1D decomposition of $V_j$, we get

$$\begin{aligned} V_{j-1}^2 &= V_{j-1} \otimes V_{j-1} = (V_j \oplus W_j) \otimes (V_j \oplus W_j) \\ &= \underbrace{(V_j \otimes V_j)}_{=V_j^2} \oplus \underbrace{(V_j \otimes W_j)}_{=:H_j^2} \oplus \underbrace{(W_j \times V_j)}_{=:S_j^2} \oplus \underbrace{(W_j \otimes W_j)}_{=:D_j^2} \end{aligned}$$

Thus, the detail space in 2D consists of three components: The horizontal details $H_j^2$, the vertical details $S_j^2$ and the diagonal details $D_j^2$. Using the 1D scaling function $\phi$ and the 1D wavelet $\psi$, we can construct orthonormal bases of these detail spaces. Let

$$\psi^1(x_1, x_2) \coloneqq \phi(x_1)\psi(x_2),\ \psi^2(x_1, x_2) \coloneqq \psi(x_1)\phi(x_2),\ \psi^3(x_1, x_2) \coloneqq \psi(x_1)\psi(x_2)$$

and

$$\psi_{j,k}^m(x_1, x_2) = 2^{-j}\psi^m(2^{-j}x_1 - k_1, 2^{-j}x_2 - k_2) \text{ for } m \in \{1, 2, 3\}.$$

Then, $\left\{\psi_{j,k}^1 \,:\, k \in \mathbb{Z}^2\right\}$ is an orthonormal basis of $H_j^2$, $\left\{\psi_{j,k}^2 \,:\, k \in \mathbb{Z}^2\right\}$ is an orthonormal basis of $S_j^2$ and $\left\{\psi_{j,k}^3 \,:\, k \in \mathbb{Z}^2\right\}$ is an orthonormal basis of $D_j^2$. Thus,

$$\left\{\psi_{j,k}^m \,:\, m \in \{1, 2, 3\}, k \in \mathbb{Z}^2, j \in \mathbb{Z}\right\}$$

is an orthonormal basis of $L^2(\mathbb{R}^2)$.

Like in 1D, the FWT in 2D starts with coefficients $c^j$ given on a certain scale. The only difference is that $c^j$ is now indexed over $\mathbb{Z}^2$ instead of $\mathbb{Z}$. To compute $c^{j+1}$, we can apply the 1D $H$ first along the columns and then along the rows, similarly to how the 2D FFT can be computed from the 1D FFT along the columns and then the rows. $d^{j+1}$ now consists of three components, $d^{1,j+1}$, $d^{2,j+1}$ and $d^{3,j+1}$, corresponding to the three detail spaces $H_{j+1}^2$, $S_{j+1}^2$ and $D_{j+1}^2$, respectively. To compute $d^{1,j+1}$, we first compute $H$ along the columns and then (the 1D) $G$ along the rows. Similarly, first computing $G$ along the columns and then $H$ along the rows results in $d^{2,j+1}$. Finally, $d^{3,j+1}$ is computed by computing $G$ along the columns and then $G$ along the rows. Since $G$ and $H$ both downsample by a factor of 2, $c^{j+1}$, $d^{1,j+1}$, $d^{2,j+1}$ and $d^{3,j+1}$ are each downsampled by a factor of 2 horizontally and vertically. The IFWT is computed similarly using $G^*$ and $H^*$ from 1D.

While the tensor product approach is very easy to handle, it only has a very limited directional resolution. Due to its construction, it is more or less limited to the detection of horizontal, vertical or diagonal structures. There are different ways to overcome this limitation by directly constructing a 2D transformation that not only takes into account scaling and translation, but also some directional parameters. Two well known examples of such approaches are curvelets and shearlets.

*Remark* 4.25. Wavelets have countless of applications, both in theoretical and in practical work. Here, we just confine to briefly mention one practical application. Like the DCT, the FWT can be used for image compression by only storing the most important wavelet coefficients. In JPEG2000, the block-wise DCT is replaced by a wavelet transform, which results in significantly improved compression rates, i. e. better image quality using the same amount of data or same image quality using less data. While the classical JPEG format is still much more popular in digital photography than JPEG2000, digital cinema uses JPEG2000 as basis for video compression.

# 5. The heat equation

*Remark* 5.1 (Derivation of the heat equation in 1D). We consider a wire of length 1 and denote the temperature of the wire at position $x \in \mathbb{R}^3$ and at time $t \geq 0$ by $u(x,t)$.

For the modelling, we make the following assumptions on the wire:

- Constant cross sectional area $A$.
- Heat flux only along the wire, i. e. no in- or outflow.
- The wire is so thin that the temperature is constant orthogonal to the wire direction.
- Boundary effects are not considered at first.

This leads to a 1D model of the wire temperature, i. e. we consider $u : [0,1] \times [0,\infty) \to \mathbb{R}$.

Two parameters of the wire are important to model its heat conduction:

(i) The specific heat capacity $c$ (amount of energy required to heat 1 kg by 1 K)

(ii) The density $\rho$.

We assume that $c$ and $\rho$ are independent from temperature and time, but not from the location, i. e. $c = c(x)$ and $\rho = \rho(x)$. Then, the amount of heat $Q$ in a piece of the wire of length $h$ starting at position $x \in [0, 1-h]$ is given by

$$Q(x,h,t) = \int_x^{x+h} Au(y,t)\rho(y)c(y)\,\mathrm{d}y\,.$$

By *Fourier's law* (1822), the heat flux $q(x,t)$, i. e. the transport of heat at position $x$ and time $t$, is proportional to the cross sectional area and the spatial gradient of the temperature, i. e.

$$q(x,t) = -K(x)A\partial_x u(x,t),$$

where $K(x)$ is the heat conductivity of the material at position $x$. The negative sign means that heat flows in direction of decreasing temperature. Thus, temperature differences even up with time. Since $-\partial_x u$ is the temperature decrease in direction of the $x$-axis, $q$ describes the flux in direction of the $x$-axis.

The temporal change of the amount of heat $Q$ in a piece of the wire of length $h$ has to be equal to the difference of in- and outflow at the boundary of the piece of wire, i. e.

$$\begin{aligned}\partial_t Q(x,h,t) &= q(x,t) - q(x+h,t)\\ &= -K(x)A\partial_x u(x,t) + K(x+h)A\partial_x u(x+h,t)\\ &= A(K(x+h)\partial_x u(x+h,t) - K(x)\partial_x u(x,t)).\end{aligned}$$

Additionally, we have (if $u \in C^1([0,1]\times[0,\infty))$ and $\rho, c \in L^\infty([0,1])$)

$$\begin{aligned}\partial_t Q(x,h,t) &= \partial_t \int_x^{x+h} Au(y,t)\rho(y)c(y)\,\mathrm{d}y = \int_x^{x+h} A\partial_t u(y,t)\rho(y)c(y)\,\mathrm{d}y\\ &= hA\partial_t u(\zeta,t)\rho(\zeta)c(\zeta)\end{aligned}$$

for $\zeta \in [x, x+h]$ by the first mean value theorem for definite integrals (if $\rho, c \in C([0,1])$). Combined, we get

$$A(K(x+h)\partial_x u(x+h,t) - K(x)\partial_x u(x,t)) = hA\partial_t u(\zeta,t)\rho(\zeta)c(\zeta)$$
$$\Rightarrow \frac{K(x+h)\partial_x u(x+h,t) - K(x)\partial_x u(x,t)}{h} = \partial_t u(\zeta,t)\rho(\zeta)c(\zeta).$$

By passing to the limit $h \to 0$, we have

$$\partial_x \left(K(x)\partial_x u(x,t)\right) = \partial_t u(x,t)\rho(x)c(x).$$

This leads to the *(general) heat equation* without boundary conditions

$$\partial_t u(x,t) = a(x)\partial_x \left(K(x)\partial_x u(x,t)\right) \text{ for all } (x,t) \in \Omega \times (0,\infty),$$

where $a(x) = \frac{1}{\rho(x)c(x)}$ and $\Omega = (0,1)$. The case $c = \rho = K \equiv 1$ is the *specific heat equation*

$$\partial_t u(x,t) = \partial_x^2 u(x,t) \text{ for all } (x,t) \in \Omega \times (0,\infty).$$

*Remark* 5.2 (Initial and boundary conditions). To describe the problem uniquely, we are still missing initial conditions (otherwise $u \equiv 0$ is a solution) as well as boundary conditions. Let $f : \Omega \to \mathbb{R}$ be the temperature at time $t = 0$. Then, a solution $u : \overline{\Omega} \times [0,\infty) \to \mathbb{R}$ has to fulfill the initial condition

$$u(x,0) = f(x) \text{ for all } x \in \Omega.$$

Depending on the physical situation to be modelled, different boundary conditions are necessary.

- Dirichlet boundary condition:

  $$u(x,t) = g(x,t) \text{ for all } (x,t) \in \partial\Omega \times (0,\infty)$$

  for a given $g : \partial\Omega \times (0,\infty) \to \mathbb{R}$. This means that the temperature at the boundary is prescribed by exterior influences (e. g. heating).

- Neumann boundary condition:

  $$\partial_{\nu_x} u(x,t) = g(x,t) \text{ for all } (x,t) \in \partial\Omega \times (0,\infty)$$

  for a given $g : \partial\Omega \times (0,\infty) \to \mathbb{R}$. Here, the flux at the boundary is controlled by exterior influences. For instance, a perfect thermal insulation of the wire at the boundary corresponds to Neumann boundary conditions with $g = 0$.

- Periodic boundary condition:

  $$u(0,t) = u(1,t) \text{ for all } t \in (0,\infty).$$

  This corresponds to a ring-shaped wire.

**Proposition 5.3.** *Let $u \in C([0,1] \times [0,\infty))$ be twice continuously differentiable with respect to the first variable $x$ and continuously differentiable with respect to the second variable $t$. If $u$ is a solution of the specific heat equation with zero Dirichlet or Neumann boundary conditions and intial condition $f$, we have*

$$\int_0^1 u^2(x,t)\,\mathrm{d}x \leq \int_0^1 f^2(x)\,\mathrm{d}x \text{ for all } t \geq 0.$$

**Proof.** We show that $E(t) \coloneqq \int_0^1 u^2(x,t)\,\mathrm{d}x$ is decreasing. Then, we have $E(t) \le E(0)$ for all $t \ge 0$. Combined with $u(x,0) = f(x)$, the claim follows. The monotonicity follows from

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}E(t) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\left(u^2(x,t)\right)\mathrm{d}x = \int_0^1 2u(x,t)\partial_t u(x,t)\,\mathrm{d}x = 2\int_0^1 u(x,t)\partial_x^2 u(x,t)\,\mathrm{d}x \\ &= 2\underbrace{u(x,t)\partial_x u(x,t)\Big|_0^1}_{=0\ \text{(BC)}} - 2\int_0^1 \underbrace{(\partial_x u(x,t))^2}_{\ge 0}\,\mathrm{d}x \le 0.\end{aligned}$$ $\square$

**Corollary 5.4.** *Let $u \in C([0,1]\times[0,\infty))$ be twice continuously differentiable with respect to the first variable $x$ and continuously differentiable with respect to the second variable $t$. If $u$ is a solution of the specific heat equation with Dirichlet or Neumann boundary conditions $g$ and intial condition $f$, $u$ is determiend uniquely.*

**Proof.** Let $u_1$ and $u_2$ fulfill the assumptions, i. e. be solutions of the specific heat equation with the same boundary and intial conditions. Then, $u_1 - u_2$ is a solution of the specific heat equation with zero boundary conditions and initial value 0. From Proposition 5.3, we get

$$0 \le \int_0^1 (u_1(x,t) - u_2(x,t))^2\,\mathrm{d}x \le \int_0^1 0^2\,\mathrm{d}x = 0 \text{ for all } t \ge 0$$

Thus, $u_1 = u_2$. $\square$

*Remark* 5.5. As we have already seen in Remark 2.16, the solution of the specific heat equation is the convolution with the Gaussian kernel. Thus, the specific heat equation is suitable to remove noise, but cannot preserve edges in the process.

The general heat equation (with $a \equiv 1$)

$$\begin{aligned}\partial_t u(x,t) &= \partial_x\left(K(x)\partial_x u(x,t)\right) && \text{for all } (x,t) \in \Omega \times (0,\infty),\\ u(x,0) &= f(x) && \text{for all } x \in \Omega\end{aligned}$$

however allows adjusting the strength of the smoothing locally with $K(x)$:

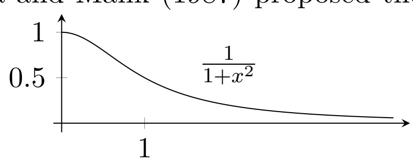
$$K \equiv 1 \Rightarrow \partial_t u(x,t) = \partial_x^2 u(x,t) \Rightarrow u(x,t) = (g_{\sqrt{2t}} \star f)(x) \text{ (strong smoothing)}$$

$$K \equiv 0 \Rightarrow \partial_t u(x,t) = 0 \Rightarrow u(x,t) = f(x) \text{ (no smoothing)}$$

Idea: Choose $K$ such that $K(x) \approx 0$ if $u$ has an edge near $x$ and $K(x) \approx 1$ otherwise.

Since edges are associated with large derivatives, Perona and Malik (1987) proposed the following approach:

$$K(x,t) = c_\lambda(|\partial_x u(x,t)|^2), \text{ where } c_\lambda\left(s^2\right) = \frac{1}{1+\frac{s^2}{\lambda^2}}.$$




The parameter $\lambda > 0$ controls the size of the derivative necessary for a position to be classified as edge. This $K$ is even time-dependent. If $K$ is supposed not to depend on $t$, one can use $K(x) = c_\lambda(|f'(x)|^2) = K(x,0)$.

To transfer this approach to images, we need the heat equation in higher dimensions. Generalized to a domain $\Omega \subset \mathbb{R}^d$ and with a time-dependent $K$, the equation is

$$\begin{aligned}\partial_t u(x,t) &= a(x)\operatorname{div}_x\left(K(x,t)\nabla_x u(x,t)\right) && \text{for all } (x,t) \in \Omega \times (0,\infty),\\ u(x,0) &= f(x) && \text{for all } x \in \Omega,\end{aligned}$$

where $K : \Omega \times (0, \infty) \to \mathbb{R}$ or more generally $K : \Omega \times (0, \infty) \to \mathbb{R}^{d \times d}$. The generalization can be obtained by doing the derivation directly with a higher spatial dimension or it can be motivated by "$\Delta = \operatorname{div}\nabla$" and that, for $d = 1$, the second derivative coincides with the Laplace operator, i. e. $\partial_x^2 = \Delta$. The real-valued generalization of $K$ to preserve edges for arbitrary $d$ is

$$K(x, t) = c_\lambda(|\nabla_x u(x, t)|^2).$$

With a matrix-valued $K$, it is even possible to preserve edges while smoothing along the edges (anisotropic diffusion).

**Proposition 5.6.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary and $u \in C(\overline{\Omega} \times [0, \infty))$ be twice continuously differentiable with respect to the first variable $x$ and continuously differentiable with respect to the second variable $t$. If $u$ is a solution of*

$$\partial_t u(x, t) = \operatorname{div}_x \Big( c_\lambda(|\nabla_x u(x, t)|^2) \nabla_x u(x, t) \Big) \text{ for all } (x, t) \in \Omega \times (0, \infty)$$

*with zero Dirichlet or Neumann boundary conditions and intial condition $f$, we have*

$$\|u(\cdot, t)\|_{L^p(\Omega)} \leq \|f\|_{L^p(\Omega)} \text{ for all } t \geq 0 \text{ and } p \in [2, \infty].$$

*In case of zero Neumann boundary conditions, we additionally have*

$$\int_\Omega u(x, t) \,\mathrm{d}x = \int_\Omega f(x) \,\mathrm{d}x \text{ for all } t \geq 0.$$

**Proof.** Like in Proposition 5.3, it is sufficient to show that $E(t) := \int_\Omega |u(x, t)|^p \,\mathrm{d}x$ is decreasing to show the first statement. Let $p \in [2, \infty)$. Then, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} E(t) &= \int_\Omega \frac{\mathrm{d}}{\mathrm{d}t} |u(x, t)|^p \,\mathrm{d}x = \int_\Omega p \,|u(x, t)|^{p-2} \,u(x, t) \partial_t u(x, t) \,\mathrm{d}x \\
&= \int_\Omega p \,|u(x, t)|^{p-2} \,u(x, t) \operatorname{div}_x \Big( c_\lambda(|\nabla_x u(x, t)|^2) \nabla_x u(x, t) \Big) \,\mathrm{d}x \\
&= \int_{\partial\Omega} p \,|u(x, t)|^{p-2} \underbrace{u(x, t) c_\lambda(|\nabla_x u(x, t)|^2) \partial_{\nu_x} u(x, t)}_{=0 \text{ (BC)}} \,\mathrm{d}A(x) \\
&\quad - p \int_\Omega \underbrace{\nabla_x \Big( |u(x, t)|^{p-2} \,u(x, t) \Big)}_{\substack{=(p-2)|u(x,t)|^{p-4}(u(x,t))^2 \nabla_x u(x,t) \\ +|u(x,t)|^{p-2} \nabla_x u(x,t)}} \cdot \Big( c_\lambda(|\nabla_x u(x, t)|^2) \nabla_x u(x, t) \Big) \,\mathrm{d}x \\
&= -p(p-1) \int_\Omega |u(x, t)|^{p-2} \,c_\lambda(|\nabla_x u(x, t)|^2) \,\|\nabla_x u(x, t)\|_2^2 \,\mathrm{d}x \leq 0.
\end{aligned}$$

This shows the necessary monotonicty and thus the first statement for $p \in [2, \infty)$. The case $p = \infty$ follows by passing to the limit $p \to \infty$ and the fact that

$$\lim_{p \to \infty} \|g\|_{L^p(\Omega)} = \|g\|_{L^\infty(\Omega)} \text{ for all } g \in L^\infty(\Omega),$$

which is true since $\Omega$ is bounded.

To show the second statement, it is sufficient to show that $F(t) := \int_\Omega u(x,t)\,\mathrm{d}x$ is constant. This follows from

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}F(t) &= \int_\Omega \partial_t u(x,t)\,\mathrm{d}x = \int_\Omega \mathrm{div}_x\left(c_\lambda(|\nabla_x u(x,t)|^2)\nabla_x u(x,t)\right)\mathrm{d}x \\
&= \int_{\partial\Omega} c_\lambda(|\nabla_x u(x,t)|^2)\underbrace{\partial_{\nu_x} u(x,t)}_{=0\text{ (BC)}}\,\mathrm{d}A(x) \\
&\quad - \int_\Omega \underbrace{\nabla_x(1)}_{=0}\cdot\left(c_\lambda(|\nabla_x u(x,t)|^2)\nabla_x u(x,t)\right)\mathrm{d}x = 0.
\end{aligned}$$

□

*Remark* 5.7. To better understand the effects of the Perona–Malik model, it is helpful to represent the right hand side of the PDE in a different coordinate system. For $|\nabla_x u| \neq 0$, we have

$$\begin{aligned}
&\mathrm{div}_x\left(c_\lambda(|\nabla_x u|^2)\nabla_x u\right) = \sum_{i=1}^d \partial_{x_i}\left(c_\lambda(|\nabla_x u|^2)\partial_{x_i}u\right) \\
&= \sum_{i=1}^d \left(c_\lambda'(|\nabla_x u|^2)(2\nabla_x u\cdot\partial_{x_i}\nabla_x u)\partial_{x_i}u + c_\lambda(|\nabla_x u|^2)\partial_{x_i}^2 u\right) \\
&= 2c_\lambda'(|\nabla_x u|^2)(\nabla_x u\cdot\nabla_x^2 u\nabla_x u) + c_\lambda(|\nabla_x u|^2)\Delta_x u \\
&= (c_\lambda(|\nabla_x u|^2) + 2\,|\nabla_x u|^2\,c_\lambda'(|\nabla_x u|^2) - c_\lambda(|\nabla_x u|^2))\frac{\nabla_x u\cdot\nabla_x^2 u\nabla_x u}{|\nabla_x u|^2} + c_\lambda(|\nabla_x u|^2)\Delta_x u.
\end{aligned}$$

In case of images, i.e. $d = 2$, using the notation $u_{x_j} := \partial_{x_j}u$, $u_{x_ix_j} := \partial_{x_i}\partial_{x_j}u$, $\eta := \frac{\nabla_x u}{|\nabla_x u|}$, $u_{\eta\eta} := \eta\cdot\nabla_x^2 u\eta$ and the function $h(s) := c_\lambda(s) + 2sc_\lambda'(s)$, this leads to

$$\begin{aligned}
&\mathrm{div}_x\left(c_\lambda(|\nabla_x u|^2)\nabla_x u\right) \\
&= h(|\nabla_x u|^2)\frac{\nabla_x u\cdot\nabla_x^2 u\nabla_x u}{|\nabla_x u|^2} - c_\lambda(|\nabla_x u|^2)\frac{u_{x_1}^2 u_{x_1x_1} + 2u_{x_1}u_{x_2}u_{x_1x_2} + u_{x_2}^2 u_{x_2x_2}}{|\nabla_x u|^2} \\
&\quad + c_\lambda(|\nabla_x u|^2)(u_{x_1x_1} + u_{x_2x_2}) \\
&= h(|\nabla_x u|^2)u_{\eta\eta} + c_\lambda(|\nabla_x u|^2)\underbrace{\left(u_{x_1x_1} + u_{x_2x_2} - \frac{u_{x_1}^2 u_{x_1x_1} + 2u_{x_1}u_{x_2}u_{x_1x_2} + u_{x_2}^2 u_{x_2x_2}}{|\nabla_x u|^2}\right)}_{=:(*)}.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
|\nabla_x u|^2\,(*) &= \left(u_{x_1}^2 + u_{x_2}^2\right)(u_{x_1x_1} + u_{x_2x_2}) - u_{x_1}^2 u_{x_1x_1} - 2u_{x_1}u_{x_2}u_{x_1x_2} - u_{x_2}^2 u_{x_2x_2} \\
&= u_{x_1}^2 u_{x_2x_2} + u_{x_2}^2 u_{x_1x_1} - 2u_{x_1}u_{x_2}u_{x_1x_2} = \begin{pmatrix}-u_{x_2} & u_{x_1}\end{pmatrix}\nabla_x^2 u\begin{pmatrix}-u_{x_2}\\ u_{x_1}\end{pmatrix}.
\end{aligned}$$

Finally, using the notation $\xi := \frac{1}{|\nabla_x u|}\begin{pmatrix}-u_{x_2} & u_{x_1}\end{pmatrix}^T$ and $u_{\xi\xi} = \xi\cdot\nabla_x^2 u\xi$, we have shown

$$\mathrm{div}_x\left(c_\lambda(|\nabla_x u|^2)\nabla_x u\right) = h(|\nabla_x u|^2)u_{\eta\eta} + c_\lambda(|\nabla_x u|^2)u_{\xi\xi}.$$

Thus, $h(|\nabla_x u|^2)$ controls the behavior in direction of $\nabla u$ (orthogonal to edges) and $c_\lambda(|\nabla_x u|^2)$ controls the behavior in direction orthogonal to $\nabla u$ (along edges). Since $c_\lambda$ is always positive, we always have a forward diffusion orthogonal to $\nabla u$. To understand the behavior in direction of $\nabla u$, we need the sign of $h(|\nabla_x u|^2)$. For $s \in [0, \infty)$, we have

$$h(s) = c_\lambda(s) + 2sc'_\lambda(s) = \frac{1}{1+\frac{s}{\lambda^2}} + 2s\frac{-\frac{1}{\lambda^2}}{(1+\frac{s}{\lambda^2})^2} = \frac{1+\frac{s}{\lambda^2}-\frac{2s}{\lambda^2}}{(1+\frac{s}{\lambda^2})^2} = \frac{1-\frac{s}{\lambda^2}}{(1+\frac{s}{\lambda^2})^2}.$$

Thus, we get

$$h(|\nabla_x u|^2) > 0 \;\Leftrightarrow\; 1 - \frac{|\nabla_x u|^2}{\lambda^2} > 0 \;\Leftrightarrow\; |\nabla_x u| < \lambda.$$

Analogously, we have $h(|\nabla_x u|^2) < 0 \;\Leftrightarrow\; |\nabla_x u| > \lambda$. Thus, we have backward diffusion in direction $\nabla u$ at strong edges. This means that Perona–Malik can enhance edges, but also that it is unstable in this case, since it is a backward diffusion.

*Remark* 5.8 (Instability of the derivative). The computation of $K$ directly from the derivative of the signal or image that we want to denoise is problematic in practice. To illustrate this, we consider the following example. Let

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto 0 \text{ and } f_n : \mathbb{R} \to \mathbb{R}, x \mapsto \tfrac{\sin(n^2 x)}{n} \text{ for } n \in \mathbb{N}.$$

Then, we have $\|f - f_n\|_{L^\infty} = \frac{1}{n} \to 0$ for $n \to \infty$ but $\|f' - f'_n\|_{L^\infty} = n \to \infty$ for $n \to \infty$. In other words, marginal, smooth changes in $f$ can cause unbounded changes of $f'$. This problem can be resolved exactly like in Remark 2.16, that is by a convolution with the Gaussian kernel, i. e. we consider the following $K$:

$$K(x,t) = c_\lambda(|\nabla_x (g_\sigma \star u)(x,t)|^2).$$

Here, a fixed filter width $\sigma > 0$ is chosen corresponding to the strength of the noise in the input data. Note that the filtering is only done with respect to the variable $x$ for fixed $t$.

*Remark* 5.9 (Anisotropic diffusion). The anisotropic diffusion is based on

$$J_0[f_\sigma](x) = (\nabla f_\sigma(x) \otimes \nabla f_\sigma(x)),$$

the *structure tensor* of $f_\sigma \coloneqq g_\sigma \star f$. If $\nabla f_\sigma(x) \neq 0$, then $J_0[f_\sigma](x)$ has two eigenvectors $v_1$ and $v_2$ with $v_1 || \nabla f_\sigma$ and $v_2 \perp \nabla f_\sigma$ with associated eigenvalues $\mu_1 = \|\nabla f_\sigma\|^2$ and $\mu_2 = 0$. This can be shown as follows. For $v \in \mathbb{R}^d$ and $1 \le i \le d$, we have

$$(J_0[f_\sigma]v)_i = \sum_{j=1}^d \partial_i f_\sigma \partial_j f_\sigma v_j = (\nabla f_\sigma \cdot v)\partial_i f_\sigma = (\nabla f_\sigma \cdot v)(\nabla f_\sigma)_i \Rightarrow J_0[f_\sigma]v = (\nabla f_\sigma \cdot v)\nabla f_\sigma.$$

From this, we immediately get

$$J_0[f_\sigma]\nabla f_\sigma = \|\nabla f_\sigma\|^2 \nabla f_\sigma \text{ and } J_0[f_\sigma]v = 0 \text{ for all } v \in \operatorname{span}(\nabla f_\sigma)^\perp.$$

In practice, $J_0$ is also smoothed, i. e. one considers

$$J_\rho[f_\sigma](x) = (g_\rho \star (\nabla f_\sigma \otimes \nabla f_\sigma))(x).$$

Let $v_1$ and $v_2$ be the eigenvectors of $J_\rho[f_\sigma](x)$ for the eigenvalues $\mu_1 \geq \mu_2 \geq 0$. Then, the anisotropic diffusion is given by the heat equation with

$$K(x) = \underbrace{\frac{1}{1+\mu_1}(v_1 \otimes v_1)}_{\text{orthogonal to the edge}} + \underbrace{\left[\lambda + (1-\lambda)\exp\left(\frac{-1}{(\mu_1-\mu_2)^2}\right)\right](v_2 \otimes v_2)}_{\text{tangential to the edge}}.$$

Here, $\lambda \in (0,1)$ is a small parameter to ensure that $K(x)$ is uniformly positive definite, i. e. there is $c > 0$, such that $K(x)\xi \cdot \xi > c\,\|\xi\|_2^2$ for all $x \in \Omega$ and $\xi \in \mathbb{R}^d$. By considering $u(\cdot,t)$ instead of $f$ in $J_\rho[f_\sigma]$, one gets a time-dependent $K$.

*Remark* 5.10. To solve the general heat equation numerically, one has to discretize in space and time. Here, we only consider the case $a \equiv 1$ and $K(x,t) = K(x)$. In time, we use the backward Euler method with fixed step size $\tau > 0$, i. e. we consider the solution $u$ at times $t_n = n\tau$ for $n \in \mathbb{N}_0$. With $u_n(x) \coloneqq u(x,t_n)$, this leads to the time-discrete problem

$$\begin{aligned}\frac{u^{n+1}(x) - u^n(x)}{\tau} &= \operatorname{div}\left(K(x)\nabla u^{n+1}(x)\right) \text{ for all } x \in \Omega, n \in \mathbb{N}_0,\\ u^0(x) &= f(x) \text{ for all } x \in \Omega.\end{aligned}$$

For the spatial discretization, it is useful to switch to a weak formulation of the PDE, which makes it possible to use the finite element method for the spatial discretization. Using the fundamental lemma and integration by parts, it follows that the PDE (with zero Neumann boundary conditions) is equivalent to

$$\int_\Omega u^{n+1}(x)\vartheta(x)\,\mathrm{d}x = \int_\Omega u^n(x)\vartheta(x)\,\mathrm{d}x - \tau\int_\Omega K(x)\nabla u^{n+1}(x)\cdot\nabla\vartheta(x)\,\mathrm{d}x \text{ for all } \vartheta \in C_c^\infty(\Omega).$$

Let $\mathcal{V}$ be a finite element space on $\Omega = (0,1)^2$, for instance, bilinear finite elements on the grid with the nodes $\left\{(\frac{i}{N},\frac{j}{N}) : i,j \in \{0,\dots,N\}\right\}$. Let $\varphi_1,\dots,\varphi_L$ be the canonical basis of $\mathcal{V}$, where $L \coloneqq \dim\mathcal{V}$. If we consider $u^n \in \mathcal{V}$, we have $u^n(x) = \sum_{i=1}^L u_i^n\varphi_i(x)$ with $u_i^n \in \mathbb{R}$ and the weak form of the PDE can be expressed as

$$\int_\Omega \sum_{i=1}^L u_i^{n+1}\varphi_i(x)\varphi_j(x)\,\mathrm{d}x = \int_\Omega \sum_{i=1}^L u_i^n\varphi_i(x)\varphi_j(x)\,\mathrm{d}x - \tau\int_\Omega K(x)\nabla\sum_{i=1}^L u_i^{n+1}\varphi_i(x)\cdot\nabla\varphi_j(x)\,\mathrm{d}x$$

for all $j \in \{1,\dots,L\}$. This equation is equivalent to

$$\sum_{i=1}^L\left(\int_\Omega \varphi_i(x)\varphi_j(x)\,\mathrm{d}x + \tau\int_\Omega K(x)\nabla\varphi_i(x)\cdot\nabla\varphi_j(x)\,\mathrm{d}x\right)u_i^{n+1} = \sum_{i=1}^L\left(\int_\Omega\varphi_i(x)\varphi_j(x)\,\mathrm{d}x\right)u_i^n.$$

Using the matrices $M = (\int_\Omega \varphi_i\varphi_j\,\mathrm{d}x)_{ji}$ and $L[K] = (\int_\Omega K\nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x)_{ji}$, as well as the vectors $U^n = (u_i^n)_i$, we get the linear system

$$(M + \tau L[K])U^{n+1} = MU^n.$$

$M$ and $L$ are sparse, which has to be taken into account when solving the system numerically. The most simple way to solve the system is to use an iterative solver like the conjugate gradient method (if $K$ is symmetric). At least in case $d \leq 2$, it is usually more efficient though to use a direct solver for sparse matrices, e. g. Sparse Cholesky.

*Remark* 5.11. We already know that $\partial_t(g_{\sqrt{2t}}) = \partial_x^2 g_{\sqrt{2t}}$ for all $t > 0$. Thus,

$$\phi : \mathbb{R} \times (0, \infty) \to \mathbb{R}, (x,t) \mapsto g_{\sqrt{2t}}(x) = \frac{1}{\sqrt{4\pi t}} \exp\left(\frac{-x^2}{4t}\right)$$

is called *fundamental solution of the (specific) heat equation.* This function can be inferred from the scaling behavior of the heat equation: Let $u(x,t)$ be a solution of this equation. Then, for $\lambda > 0$ and $u_\lambda(x,t) \coloneqq u(\lambda x, \lambda^2 t)$, we have

$$\partial_t u_\lambda(x,t) = \lambda^2 \partial_t u(\lambda x, \lambda^2 t) = \lambda^2 \partial_x^2 u(\lambda x, \lambda^2 t) = \partial_x^2 u_\lambda(x,t),$$

i. e. $u_\lambda$ is also a solution. This motivates to look for solutions with the scaling behavior

$$u(x,t) = \lambda^\alpha u(\lambda x, \lambda^2 t)$$

for a suitable $\alpha \in \mathbb{R}$. For $\lambda = \frac{1}{\sqrt{t}}$, we get

$$u(x,t) = t^{-\frac{\alpha}{2}} u\left(\tfrac{x}{\sqrt{t}}, 1\right) = t^{-\frac{\alpha}{2}} v\left(\tfrac{x}{\sqrt{t}}\right),$$

where $v(y) \coloneqq u(y,1)$. Moreover, we obtain

$$\partial_t u(x,t) = -\frac{\alpha}{2} t^{-\frac{\alpha}{2}-1} v\left(\tfrac{x}{\sqrt{t}}\right) + t^{-\frac{\alpha}{2}} v'\left(\tfrac{x}{\sqrt{t}}\right)\left(-\frac{x}{2} t^{-\frac{3}{2}}\right)$$
$$\partial_x^2 u(x,t) = t^{-\frac{\alpha}{2}-1} v''\left(\tfrac{x}{\sqrt{t}}\right)$$

Thus, $v$ has to fulfill

$$-\frac{\alpha}{2} t^{-\frac{\alpha}{2}-1} v\left(\tfrac{x}{\sqrt{t}}\right) - \frac{x}{2} t^{-\frac{\alpha}{2}-\frac{3}{2}} v'\left(\tfrac{x}{\sqrt{t}}\right) = t^{-\frac{\alpha}{2}-1} v''\left(\tfrac{x}{\sqrt{t}}\right) \qquad \Big| \cdot t^{\frac{\alpha}{2}+1}$$
$$\Rightarrow -\frac{\alpha}{2} v\left(\tfrac{x}{\sqrt{t}}\right) - \frac{x}{2} t^{-\frac{1}{2}} v'\left(\tfrac{x}{\sqrt{t}}\right) = v''\left(\tfrac{x}{\sqrt{t}}\right)$$

With $s = \frac{x}{\sqrt{t}}$, we get

$$\frac{\alpha}{2} v(s) + \frac{s}{2} v'(s) + v''(s) = 0.$$

For $\alpha = 1$, we have

$$\left(\frac{s}{2} v(s)\right)' + v''(s) = 0.$$

Integrating both sides leads to

$$\frac{s}{2} v(s) + v'(s) = c.$$

If we require $v(s) \to 0$ and $v'(s) \to 0$ for $s \to \infty$, we get $c = 0$ and $v$ is a solution of the ODE

$$v'(s) = -\frac{s}{2} v(s).$$

The solution of this is $v(s) = be^{-\frac{s^2}{4}}$. In total, we get $u(x,t) = t^{-\frac{\alpha}{2}} v\left(\frac{x}{\sqrt{t}}\right) = \frac{b}{\sqrt{t}} e^{-\frac{x^2}{4t}}$. If we additional require $\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = 1$, we get $b = \frac{1}{\sqrt{4\pi}}$ and have constructed the fundamental solution.

# 6. Outlook - Variational image processing

Fundamental idea of a variational approach:

- rephrase task as conditions on the solution
- find a function measuring how well the conditions are fulfilled
- by computing a minimizer of the function, the task is solved

*Denoising* Given a noisy image/signal $f : \Omega \to \mathbb{R}$, i. e. $f = f_0 + n$, find $f_0$.

$$J[u] = \underbrace{\int_\Omega (u - f)^2 \,\mathrm{d}x}_{\text{data term}} + \lambda \underbrace{\int_\Omega \|\nabla u(x)\| \,\mathrm{d}x}_{\text{regularizer}} \quad \text{(Rudin-Osher-Fatemi)}$$

*Deconvolution/Deblurring* Given a blurry image/signal $f : \Omega \to \mathbb{R}$, i. e. $f = Af_0$, find $f_0$.

$$J[u] = \int_\Omega (Au - f)^2 \,\mathrm{d}x + \lambda \int_\Omega \|\nabla u(x)\| \,\mathrm{d}x$$

*Registration* Given two images $f, g : \Omega \to \mathbb{R}$, find a deformation $\phi : \Omega \to \Omega$, such that $f \approx g \circ \phi$ holds and $\phi$ is smooth.

$$J[\phi] = \int_\Omega |f(x) - g(\phi(x))|^2 \,\mathrm{d}x + \lambda \int_\Omega \|D(\phi(x) - x)\|^2 \,\mathrm{d}x$$

*Segmentation* Decompose an image $f : \Omega \to \mathbb{R}$ in foreground $\mathcal{O}$ (color $c_1$) and background $\Omega \setminus \mathcal{O}$ (color $c_2$).

$$J[\mathcal{O}] = \int_\mathcal{O} (f - c_1)^2 \,\mathrm{d}x + \int_{\Omega\setminus\mathcal{O}} (f - c_2)^2 \,\mathrm{d}x + \lambda \operatorname{Per}(\mathcal{O}) \quad \text{(binary Mumford–Shah functional)}$$

# A. Some elements of measure theory

**Definition A.1** (Measurable space)**.** Let $\Omega$ be a nonempty set, $\mathcal{P}(\Omega)$ the power set of $\Omega$ and $\mathcal{A} \subset \mathcal{P}(\Omega)$. Then, $\mathcal{A}$ is called *$\sigma$-algebra*, if the following conditions hold:

(i) $\Omega \in \mathcal{A}$.

(ii) $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$.

(iii) $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

The pair $(\Omega, \mathcal{A})$ is called *measurable space*. The elements of $\mathcal{A}$ are called *measurable*. If $\Omega$ is a normed vector space[4], the smallest $\sigma$-algebra that contains the open sets of the space $\Omega$ is called *Borel algebra* and denoted by $\mathcal{B}(\Omega)$.

**Definition A.2** (Positive measure)**.** Let $(\Omega, \mathcal{A})$ be a measurable space. A function $\mu : \mathcal{A} \to [0, \infty]$ is called *(positive) measure*, if it fulfills the following conditions:

(i) $\mu(\emptyset) = 0$.

(ii) If $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

The triple $(\Omega, \mathcal{A}, \mu)$ is called *measure space*. If $\mathcal{A} = \mathcal{B}(\Omega)$, then $\mu$ is called *Borel measure*. If additionally $\mu(K) < \infty$ holds for all compact[5] sets $K \in \mathcal{B}(\Omega)$, then $\mu$ is called *positive Radon measure*. Note that there are also other definitions for term Radon measure in the literature.

If $(\Omega, \mathcal{A}, \mu)$ is a measure space and it holds that

- there is a sequence $(\Omega_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ with $\mu(\Omega_n) < \infty$ and $\bigcup_{n=1}^{\infty} \Omega_n = \Omega$,

$\mu$ is called *$\sigma$-finite*.

*Example* A.3.

(i) Let $\Omega = \mathbb{R}$ and $\mathcal{A} = \mathcal{B}(\mathbb{R})$. Then, the *counting measure*

$$\mu : \mathcal{B}(\mathbb{R}) \to [0, \infty], A \mapsto \begin{cases} \#(A) & \text{if } A \text{ contains finitely many elements} \\ \infty & \text{else} \end{cases}$$

is a Borel measure, but no positive Radon measure, since $[0, 1]$ is compact but $\mu([0, 1]) = \infty$. It is also not $\sigma$-finite.

(ii) Let $\Omega \subset \mathbb{R}^d$ and $x \in \Omega$. Then,

$$\delta_x : \mathcal{B}(\Omega) \to [0, \infty], A \mapsto \begin{cases} 1 & x \in A \\ 0 & \text{else} \end{cases}$$

is a positive Radon measure and called *Dirac measure at $x$*. Moreover, $\delta_x$ is $\sigma$-finite ($\Omega_n = \Omega$).

[4] Technically, a topological space is sufficient, but this generality will not be needed.

[5] In general, a set $K$ is called *compact*, if every sequence in $K$ has a converging subsequence with limit in $K$. For $K \subset \mathbb{R}^d$, compactness of $K$ is equivalent to $K$ being closed and bounded. This equivalence is known as Heine–Borel theorem.

(iii) For half-open cuboids $[a_1, b_1) \times \ldots \times [a_d, b_d)$, let

$$\mu([a_1, b_1) \times \ldots \times [a_d, b_d)) = \prod_{i=1}^{d}(b_i - a_i).$$

This can be uniquely extended to a positive Radon measure on $\mathcal{B}(\mathbb{R}^d)$. This measure is called *(d-dimensional) Lebesgue measure* and is $\sigma$-finite ($\Omega_n = B_n(0)$[6]). For $d = 1, 2$ and 3, the Lebesgue measure coincides with the standard notion of length, area and volume, respectively.

**Definition A.4.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.

- Let $N \subset \Omega$. If there is a set $A \in \mathcal{A}$ with $N \subset A$ and $\mu(A) = 0$, then $N$ is called *$\mu$-null set*.
- Let $P(x)$ be a property (e. g. convergence of a sequence of functions in a point $x$). If $P(x)$ holds for all $x \in A \subset \Omega$ and $\Omega \setminus A$ is a $\mu$-null set, then we say $P(x)$ *$\mu$-almost everywhere in $\Omega$*.
- The $\sigma$-algebra

  $$\mathcal{A}_\mu = \{B \subset \Omega \,:\, B = A \cup N \text{ with } A \in \mathcal{A}, N \ \mu\text{-null set}\}$$

  is called *completion of $\mathcal{A}$ with respect to $\mu$*. The elements of $\mathcal{A}_\mu$ are called *$\mu$-measurable*.
- The measure $\mu$ can be extended to $\mathcal{A}_\mu$ by $\mu(A \cup N) \coloneqq \mu(A)$ and then is a measure on $\mathcal{A}_\mu$.

**Definition A.5.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $X$ a Banach space. A function $f : \Omega \to X$ is called *step function*, if $f(\Omega)$ contains only finitely many elements. A step function $f$ is called *$\mu$-measurable*, if $f^{-1}(x) \in \mathcal{A}_\mu$ for all $x \in X$. If $f$ is a $\mu$-measurable step function, then its *integral* is defined as

$$\int_\Omega f \mathrm{d}\mu \coloneqq \sum_{x \in f(\Omega)} \mu(f^{-1}(x))x.$$

An arbitrary mapping $f : \Omega \to X$ is called *$\mu$-measurable*, if there is a sequence of measurable step functions $f_n$ that converges pointwise $\mu$-almost everywhere to $f$. The integral of a $\mu$-measurable mapping $f$ is defined as

$$\int_\Omega f \mathrm{d}\mu \coloneqq \lim_{n\to\infty} \int_\Omega f_n \mathrm{d}\mu.$$

If $\int_\Omega \|f\| \, \mathrm{d}\mu < \infty$, then $f$ is called *$\mu$-integrable*.

*Remark* A.6. One can show that the integral is independent of the choice of the sequence $f_n$. Thus, the integral is well defined. Moreover, it fulfills the typical properties one would expect from an integral, e. g. linearity, monotonicity and that the norm of an integral is smaller than the integral of the point-wise norm of a function.

**Definition A.7.** Let $(\Omega, \mathcal{A})$ and $(\Sigma, \mathcal{B})$ be a measurable spaces. A function $f : \Omega \to \Sigma$ is called *$\mathcal{A}$-$\mathcal{B}$ measurable* or just *measurable*, if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$.

[6]For $x \in \mathbb{R}^d$ and $r > 0$, $B_r(x)$ is the open ball with center $x$ and radius $r$, i.e. $B_r(x) \coloneqq \{y \in \mathbb{R}^d : |y - x| < r\}$.

**Proposition A.8** (Push-forward measure)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $(\Sigma, \mathcal{B})$ be a measurable space. If $f : \Omega \to \Sigma$ is $\mathcal{A}$-$\mathcal{B}$ measurable,*

$$\nu : \mathcal{B} \to [0, \infty], B \mapsto \mu(f^{-1}(B))$$

*is a measure on $\mathcal{B}$. $\nu$ is called* push-forward measure *or* image measure *and denoted by $f(\mu)$.*

*Proof.* Since $f$ is $\mathcal{A}$-$\mathcal{B}$ measurable, $\nu$ is well defined. Moreover, we have

$$\nu(\emptyset) = \mu(f^{-1}(\emptyset)) = \mu(\emptyset) = 0.$$

Let $B_1, B_2, \ldots \in \mathcal{B}$ be pairwise disjoint. Since the preimage fulfills

$$f^{-1}(M \cap N) = f^{-1}(M) \cap f^{-1}(N) \text{ for all } M, N \subset \Sigma,$$

the $f^{-1}(B_1), f^{-1}(B_2), \ldots \in \mathcal{A}$ are also pairwise disjoint. We get

$$\nu\Big(\bigcup_{n\in\mathbb{N}} B_n\Big) = \mu\Big(f^{-1}\Big(\bigcup_{n\in\mathbb{N}} B_n\Big)\Big) = \mu\Big(\bigcup_{n\in\mathbb{N}} f^{-1}(B_n)\Big) = \sum_{n\in\mathbb{N}} \mu(f^{-1}(B_n)) = \sum_{n\in\mathbb{N}} \nu(B_n). \quad \square$$

**Proposition A.9.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $(\Sigma, \mathcal{B})$ be a measurable space. If $f : \Omega \to \Sigma$ is $\mathcal{A}$-$\mathcal{B}$ measurable and $g : \Sigma \to X$ is $f(\mu)$-integrable, then*

$$\int_\Sigma g \mathrm{d}f(\mu) = \int_\Omega (g \circ f) \mathrm{d}\mu.$$

*Proof.* Let $B \in \mathcal{B}$. Then,

$$\begin{aligned}\int_\Sigma \chi_B \mathrm{d}f(\mu) &= \sum_{x\in\chi_B(\Sigma)} f(\mu)(\chi_B^{-1}(x))x = f(\mu)(\chi_B^{-1}(1)) = f(\mu)(B) = \mu(f^{-1}(B))\\ &= \mu(\chi_{f^{-1}(B)}^{-1}(1)) = \int_\Omega \chi_{f^{-1}(B)} \mathrm{d}\mu = \int_\Omega (\chi_B \circ f) \mathrm{d}\mu.\end{aligned}$$

Thus, the statement is true for characteristic functions. Due to the linearity of the integral, it also follows for step functions. Now let $g_n : \Sigma \to X$ be a sequence of measurable step functions that converges pointwise $f(\mu)$-almost everywhere to $g$. Then, we get

$$\int_\Sigma g \mathrm{d}f(\mu) = \lim_{n\to\infty} \int_\Sigma g_n \mathrm{d}f(\mu) = \lim_{n\to\infty} \int_\Omega (g_n \circ f) \mathrm{d}\mu = \int_\Omega (g \circ f) \mathrm{d}\mu.$$

Here, we have used that $g_n \circ f : \Omega \to X$ is a sequence of measurable step functions that converges pointwise $\mu$-almost everywhere to $g \circ f$. $\square$

**Definition A.10** (Distributions)**.** Let $X$ be a normed vector space with norm $\|\cdot\|$. The set

$$X' := \big\{x' : X \to \mathbb{R} \,:\, x' \text{ linear and continuous wrt. } \|\cdot\|\big\}$$

is called *(topological) dual space* of $X$.

Let $\Omega \subset \mathbb{R}^d$. The elements of $(C_c^\infty(\Omega))'$, i. e. the continuous, linear mappings from $C_c^\infty(\Omega)$ (cf. Definition B.1) to $\mathbb{R}$, are called *distributions*.

*Remark* A.11. Every function $g \in L^1(\Omega)$ can be interpreted as distribution using

$$C_c^\infty(\Omega) \to \mathbb{R}, f \to \int_\Omega fg \, \mathrm{d}x \,.$$

This distribution is called *induced* by $g$. In this sense, distributions are sometimes also called *generalized functions*. Distributions that are induced by functions are called *regular distributions*. Moreover, every positive Radon measure $\mu$ induces a distribution by

$$C_c^\infty(\Omega) \to \mathbb{R}, f \to \int_\Omega f \mathrm{d}\mu.$$

In general, such distributions are not regular.

The distribution induced by the Dirac measure $\delta_x$ is called *Dirac delta function centered at* $x$. For $f \in C_c^\infty(\Omega)$, we have (exercise)

$$\delta_x(f) = \int_\Omega f \mathrm{d}\delta_x = f(x).$$

**Theorem A.12** (Riesz representation theorem)**.** *Let $X$ be a Hilbert space. For all $x' \in X'$ there is a $y \in X$ such that*

$$x'(x) = (x, y)_X \text{ for all } x \in X$$

*(cf. [1, Theorem 2.35]).*

# B. Important spaces and statements

**Definition B.1.** Let $\Omega \subset \mathbb{R}^d$ and $g : \Omega \to \mathbb{R}$ be a function. Then,

$$\operatorname{supp}(g) \coloneqq \overline{\{x \in \Omega \,:\, g(x) \neq 0\}}$$

is called the *(closed) support* of $g$.

*Remark* B.2. Let $\Omega \subset \mathbb{R}^d$. The spaces of continuous and continuously differentiable functions are denoted by

$$\begin{aligned} C^0(\Omega) &\coloneqq C(\Omega) \coloneqq \{f : \Omega \to \mathbb{R} \,:\, f \text{ continuous}\}\,, \\ C^m(\Omega) &\coloneqq \{f : \Omega \to \mathbb{R} \,:\, f \text{ } m\text{-times continuously differentiable}\}\,, \\ C^\infty(\Omega) &\coloneqq \bigcap_{m=1}^{\infty} C^m(\Omega). \end{aligned}$$

The space of continuous functions with compact support is denoted by

$$C_c^0(\Omega) \coloneqq C_c(\Omega) \coloneqq \left\{f \in C^0(\Omega) \,:\, \operatorname{supp} f \subset \Omega \text{ and } \operatorname{supp} f \text{ compact}\right\}.$$

$C_c^m(\Omega)$ and $C_c^\infty(\Omega)$ are defined analogously.

For $f \in C(\Omega)$, we define

$$\|f\|_{C(\Omega)} \coloneqq \sup_{x\in\Omega} |f(x)|\,.$$

If $\Omega$ is compact, this defines a norm on $C(\Omega)$.

**Definition B.3** (Lebesgue spaces)**.** For $\Omega \subset \mathbb{R}^d$ $\mu$-measurable and $1 \leq p \leq \infty$, the Lebesgue spaces $L^p(\Omega)$ is defined as

$$L^p(\Omega) \coloneqq \left\{ y : \Omega \to \mathbb{R} \,:\, y \ \mu\text{-measurable } \wedge \|y\|_{L^p(\Omega)} < \infty \right\}$$

with the equivalence relation

$$f = g \text{ in } L^p(\Omega) :\Leftrightarrow f = g \ \mu\text{-almost everywhere.}$$

Here,

$$\|y\|_{L^p(\Omega)} \coloneqq \left( \int_\Omega |y|^p \,\mathrm{d}\mu \right)^{\frac{1}{p}} \text{ for } 1 \leq p < \infty, \quad \|y\|_{L^\infty(\Omega)} \coloneqq \inf_{\mu(N)=0} \sup_{x \in \Omega \setminus N} |y(x)|^{7}$$

and $\mu$ is the Lebesgue measure on $\mathbb{R}^d$.

*Remark* B.4. With the usual addition and scaling of mappings, $L^p(\Omega)$ is a vector space and $\|\cdot\|_{L^p(\Omega)}$ is a norm on $L^p(\Omega)$. The triangle inequality for the $L^p$-norm is also called *Minkowski inequality*. On can show that $L^p$ is complete with respect to this norm (Riesz–Fischer theorem).

**Lemma B.5** (Hölder's inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be $\mu$-measurable and $1 \leq p, q \leq \infty$ with $1 = \frac{1}{p} + \frac{1}{q}$. If $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$, we have $fg \in L^1(\Omega)$ and*

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)} .$$

*In particular, for $1 < p, q < \infty$, we have*

$$\int_\Omega |f(x)g(x)| \,\mathrm{d}x \leq \left( \int_\Omega |f(x)|^p \,\mathrm{d}x \right)^{\frac{1}{p}} \left( \int_\Omega |g(x)|^q \,\mathrm{d}x \right)^{\frac{1}{q}} .$$

**Theorem B.6** (Fubini's theorem)**.** *Let $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ both be $\mu$-measurable. If $f \in L^1(X \times Y)$, then*

$$\int_X \left( \int_Y f(x,y) \,\mathrm{d}y \right) \mathrm{d}x = \int_Y \left( \int_X f(x,y) \,\mathrm{d}x \right) \mathrm{d}y = \int_{X \times Y} f(x,y) \mathrm{d}(x,y)$$

*(cf. [1, Theorem 2.66 2.]).*

**Theorem B.7** (Substitution rule for multiple variables)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and $\phi : \Omega \to \mathbb{R}^d$ be a diffeomorphism, i. e. a bijective, continuously differentiable mapping whose inverse is also continuously differentiable. Let $f : \phi(\Omega) \to \mathbb{R}$ be measurable. Then $f$ is integrable on $\phi(\Omega)$ if and only if the mapping*

$$\Omega \to \mathbb{R}, x \mapsto f(\phi(x)) \,|\det(D\phi(x))|$$

*is integrable on $\Omega$. In this case, we have*

$$\int_{\phi(\Omega)} f(y) \,\mathrm{d}y = \int_\Omega f(\phi(x)) \,|\det(D\phi(x))| \,\mathrm{d}x$$

*(cf. [1, Theorem 2.69]).*

[7] It holds that $\|y\|_{L^\infty(\Omega)} = \operatorname{ess\,inf} y \coloneqq \inf \{a \in \mathbb{R} : \lambda(\{y > a\}) = 0\}$, i.e. the $L^\infty$-norm is equivalent to the essential supremum.

**Proposition B.8.** *$C_c(\Omega)$ is dense in $L^p(\Omega)$ for $1 \le p < \infty$, i. e. for all $f \in L^p(\Omega)$ and $\epsilon > 0$, there is $g \in C_c(\Omega)$ with $\|f - g\|_{L^p(\Omega)} < \epsilon$ (cf. [1, Theorem 2.55]).*

**Definition B.9.** For measurable $\psi : \Omega \to \mathbb{R}$ and $f : \mathbb{R}^d \to \mathbb{R}$,

$$(\psi * f)(x) = \int_\Omega \psi(y) f(x-y)\,\mathrm{d}y$$

is called *convolution* of $f$ and $\psi$ at position $x \in \mathbb{R}^d$.

**Proposition B.10** (Properties of the convolution)**.**

*(i)* *Let $1 \le p, q, r \le \infty$ with $\frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$, $f \in L^p(\Omega)$ and $g \in L^q(\mathbb{R}^d)$. Then, $(f * g) \in L^r(\mathbb{R}^d)$ and*

$$\|f * g\|_{L^r(\mathbb{R}^d)} \le \|f\|_{L^p(\Omega)} \|g\|_{L^q(\mathbb{R}^d)}.$$

*This is called Young's convolution inequality.*

*If $\Omega = \mathbb{R}^d$, we have $(f * g)(x) = (g * f)(x)$ for all $x \in \mathbb{R}^d$ and*

$$\|f * g\|_{L^r(\mathbb{R}^d)} = \|g * f\|_{L^r(\mathbb{R}^d)}.$$

*(ii)* *Let $\psi \in C_c^k(\mathbb{R}^d)$ and $f \in L^p(\Omega)$ with $1 \le p \le \infty$. Then, $(f * \psi) \in C^k(\mathbb{R}^d)$. If $\alpha$ is a multi-index with $|\alpha| \le k$, we have*

$$\frac{\partial^\alpha}{\partial x^\alpha}(f * \psi) = f * \frac{\partial^\alpha}{\partial x^\alpha}\psi.$$

*If $\Omega = \mathbb{R}^d$, we also have*

$$\frac{\partial^\alpha}{\partial x^\alpha}(\psi * f) = \frac{\partial^\alpha}{\partial x^\alpha}\psi * f.$$

*(iii)* *Let $\psi \in L^1(\mathbb{R}^d)$ be such that $\psi \ge 0$ and $\int_{\mathbb{R}^d} \psi(x)\,\mathrm{d}x = 1$. Moreover, for $\epsilon > 0$, let*

$$\psi_\epsilon : \mathbb{R}^d \to \mathbb{R}, x \mapsto \tfrac{1}{\epsilon^d}\psi(\tfrac{x}{\epsilon})$$

*and let $f \in L^\infty(\mathbb{R}^d)$. If $f$ is continuous in $x \in \mathbb{R}^d$, we have*

$$\lim_{\epsilon \to 0}(\psi_\epsilon * f)(x) = f(x).$$

*If $f$ is uniformly continuous, $\psi_\epsilon * f$ converges to $f$ on each compact subset of $\mathbb{R}^d$ uniformly.*

The properties of the convolution can be shown analogously to the properties cross-correlation in Proposition 2.3. There are just some minor sign differences.

**Theorem B.11** (Dominated convergence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be measurable. Moreover, let $f_n : \Omega \to \mathbb{C}$ be a sequence of $\mu$-measurable functions that converges pointwise almost everywhere in $\Omega$ to a measurable function $f : \Omega \to \mathbb{C}$. If there is $g \in L^1(\Omega)$ such that for all $n \in \mathbb{N}$, it holds that*

$|f_n(x)| \le g(x)$ *for almost all* $x \in \Omega$, *then* $f_n, f \in L^1(\Omega, \mathbb{C})$,

$$\lim_{n\to\infty} \int_\Omega |f_n - f| \,\mathrm{d}x = 0 \text{ and } \lim_{n\to\infty} \int_\Omega f_n \,\mathrm{d}x = \int_\Omega f \,\mathrm{d}x$$

*(cf. [1, Theorem 2.47]). Roughly speaking, if there is an integrable majorant for a pointwise a.e. convergent sequence, integration and limit can be interchanged.*

**Corollary B.12.** *Let* $X \subset \mathbb{R}$ *be open and let* $\Omega \subset \mathbb{R}^d$ *be measurable. Moreover, let* $f : X \times \Omega \to \mathbb{C}$ *be such that*

*(i)* $\omega \mapsto f(x, \omega)$ *is in* $L^1(\Omega, \mathbb{C})$ *for each* $x \in X$

*(ii)* $x \mapsto f(x, \omega)$ *is differentiable for almost all* $\omega \in \Omega$

*(iii) there is a* $g \in L^1(\Omega)$ *with* $|\partial_x f(x, \omega)| \le g(\omega)$ *for all* $x \in X$ *and almost all* $\omega \in \Omega$,

*then, for all* $x \in X$,

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_\Omega f(x, \omega) \,\mathrm{d}\omega = \int_\Omega \partial_x f(x, \omega) \,\mathrm{d}\omega \,.$$

*Roughly speaking, if the derivative is uniformly bounded by an integrable function, differentiation and integration may be interchanged.*

*Proof.* Let $x \in X$ be arbitrary, but fixed. We consider the sequence

$$f_n : \Omega \to \mathbb{C}, \omega \mapsto \begin{cases} n(f(x + \frac{1}{n}, \omega) - f(x, \omega)) & x + \frac{1}{n} \in X \\ 0 & \text{else} \end{cases} .$$

Since $X$ is open, there is $N \in \mathbb{N}$ such that $x + \frac{1}{n} \in X$ for all $n \ge N$. Thus, we have

$$\lim_{n\to\infty} f_n(\omega) = \lim_{n\to\infty} n(f(x + \tfrac{1}{n}, \omega) - f(x, \omega)) = \partial_x f(x, \omega)$$

for almost all $\omega \in \Omega$.

By the mean value theorem, for every $x \in X$, $n \ge N$ and almost all $\omega \in \Omega$, there is a $\zeta(x, n, \omega) \in (x, x + \frac{1}{n}) \subset X$ such that

$$n(f(x + \tfrac{1}{n}, \omega) - f(x, \omega)) = \partial_x f(\zeta(x, n, \omega), \omega).$$

Thus, we get

$$|f_n(\omega)| = |\partial_x f(\zeta(x, n, \omega), \omega)| \le g(\omega).$$

That means we have an integrable majorant for a pointwise a.e. convergent sequence, such we can apply the dominated convergence theorem (note that it is sufficient if the conditions on $f_n$ hold for $n \ge N$), which implies

$$\begin{aligned} \int_\Omega \partial_x f(x, \omega) \,\mathrm{d}\omega &= \lim_{n\to\infty} \int_\Omega f_n(x, \omega) \,\mathrm{d}\omega = \lim_{n\to\infty} n \left( \int_\Omega f(x + \tfrac{1}{n}, \omega) \,\mathrm{d}\omega - \int_\Omega f(x, \omega) \,\mathrm{d}\omega \right) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} \int_\Omega f(x, \omega) \,\mathrm{d}\omega \,. \end{aligned}$$

□

**Corollary B.13** (Dominated convergence in $L^p$)**.** *Let $\Omega \subset \mathbb{R}^d$ be measurable. Moreover, let $f_n : \Omega \to \mathbb{C}$ be a sequence of $\mu$-measurable functions that converges pointwise almost everywhere in $\Omega$ to a measurable function $f : \Omega \to \mathbb{C}$. If there is $g \in L^p(\Omega)$ for $p \in 1 \leq p < \infty$ such that for all $n \in \mathbb{N}$, it holds that*

$$|f_n(x)| \leq g(x) \text{ for almost all } x \in \Omega,$$

*then $f_n, f \in L^p(\Omega, \mathbb{C})$,*

$$\lim_{n\to\infty} \int_\Omega |f_n - f|^p \, \mathrm{d}x = 0.$$

*Proof.* We consider the sequence $g_n \coloneqq |f_n - f|^p$. Since $f_n$ converges pointwise a.e. to $f$, $g_n$ converges pointwise a.e. to 0. Noting that $|f_n(x)| \leq g(x)$ for all $n \in \mathbb{N}$ and the pointwise convergence of $f_n$ to $f$ implies $|f(x)| \leq g(x)$ almost everywhere, we get

$$|f_n(x) - f(x)| \leq |f_n(x)| + |f(x)| \leq 2g(x) \;\Rightarrow\; g_n(x) = |f_n(x) - f(x)|^p \leq (2g(x))^p.$$

Thus, $(2g(x))^p \in L^p(\Omega)$ is an integrable majorant for $g_n$, and Theorem B.11 gives

$$0 = \lim_{n\to\infty} \int_\Omega |g_n - 0| \, \mathrm{d}x = \lim_{n\to\infty} \int_\Omega |f_n - f|^p \, \mathrm{d}x \,. \qquad \square$$

**Proposition B.14** (Integrals of rotationally symmetric functions)**.** *Let $f : [0, \infty) \to \mathbb{R}$ be measurable. Then, the rotationally symmetric function*

$$\mathbb{R}^d \to \mathbb{R}, x \mapsto f(\|x\|_2)$$

*is in $L^1(\mathbb{R}^d)$, if and only if the function*

$$[0, \infty) \to \mathbb{R}, r \mapsto f(r)r^{d-1}$$

*is in $L^1([0, \infty))$. If this is the case, then*

$$\int_{\mathbb{R}^d} f(\|x\|_2) \, \mathrm{d}x = d\tau_d \int_{[0,\infty)} f(r)r^{d-1} \mathrm{d}r.$$

*Here, $\tau_d = \mathrm{Vol}(B_1(0))$ is the volume of the d-dimensional unit ball.*