universität wien

MASTERARBEIT

Titel der Masterarbeit

Geometry and Gabor Frames

Verfasst von

Markus Faulhuber BSc.

angestrebter akademischer Grad

Master of Science (MSc.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 066 821
Studienrichtung lt. Studienblatt:	Masterstudium Mathematik
Betreut von:	Prof. Dr. Hans Georg Feichtinger

Acknowledgement

I want to thank Prof. Feichtinger for coming up with the idea for this work as well as for his supervision and for providing support on several other occasions. I want to thank Prof. de Gosson for supporting me and this work by employing me in the FWF funded project P23902, "Hamiltonian Deformation of Gabor Frames". I also want to thank all members of "NuHAG" for inspiring me and helping me out whenever necessary. In this context I especially want to thank Andreas, Angelika, Christoph, Gero, Jose–Luis, Karlheinz, Martin, Monika, Nicki, Norbert, Peter, Radu, Sebastian and Severin.

I want to thank my whole family for supporting me in all situations, especially my parents Gabi and Walter for never putting any pressure on me during the long period of my studies. I also want to thank my sister Katrin for helping me out with her verbal skills several times.

I want to thank my friends who stayed with me over the years and cheered me up whenever necessary. Thank you Angie, Bernhard, David, Dominik, Harald, Josef, Lukas, Liesi, Roli, Sabi, Sabrina and Yuuzuki.

I want to thank my former colleagues of "UHC Gänserndorf" for spending a very long time of my life with me. Alexander, Andreas, Christian, Gernot, Gustav, Martin, Maximilian, Michael, Philipp, Thomas it was great.

My greatest thanks go to Marie–Thérèse for being the inspiring example I needed to find my path and for encouraging me to take on these studies. You give me hold whenever needed and bear my tics with patience. I thank you from the bottom of my heart.

Contents

Abstract

In the underlying work we will point out conjectured connections between Gabor frames and geometric properties of lattices.

The concept of Gabor frame is a certain kind of time–frequency representation method and as such underlying the uncertainty principles. This means that the product of a signal's length and its bandwidth cannot be arbitrarily small. The heart of a Gabor representation is found in the short–time Fourier transform (STFT), which includes the Fourier transform as special case. Therefore, the signals will usually be functions of finite energy, i.e. will be elements of the Hilbert space $L^2(\mathbb{R}^d)$, though we will note that the setting for the STFT can be extended by using so–called Banach–Gelfand triples.

The classical Fourier transform does not provide any information of the points in time at which certain frequencies occur. The STFT tries to solve this problem by using a so–called window function. Due to the uncertainty principles mentioned before, good localisation in time yields to worse localisation in frequency. The canonical window function therefore is a Gaussian function as it uniquely minimises the uncertainty principle.

The STFT with respect to a given window function is a continuous time–frequency representation, however, $L^2(\mathbb{R}^d)$ is a separable Hilbert space and hence, it should be sufficient to find a discrete way of representing the signal. This issue is taken care of by the concept of Gabor frames. The idea is to find a generating system for $L^2(\mathbb{R}^d)$, consisting of translated and modulated versions of the window function. The choice of the translations and modulations creates a pattern in the so–called time–frequency plane $\mathbb{R}^d \times \mathbb{R}^d$, a concept similar to the concept of phase space know from the theory of ordinary differential equations. This pattern is usually chosen to be a lattice in the time–frequency plane, which can be represented by a $2d \times 2d$ matrix.

A question still unanswered is whether good geometric properties of the underlying lattice already provide better frame properties, i.e. stable and fast reconstruction of the signal from its coefficients measured at the lattice points. This is not even clear for the 1–dimensional Gaussian function and 2–dimensional lattices yet. It is conjectured that for a 1–dimensional Gaussian window the so–called hexagonal lattice provides the best choice for a Gabor system, as the essential support of the ambiguity function, is a disc and the hexagonal lattice uniquely provides the best setting of arranging discs in 2 dimensions.

We will start with some classical packing problems, which seem to be a good choice for measuring geometric properties of a lattice and will return to them at the end of this work, when we investigate Hamiltonian deformations of lattices.

Packing and Covering Problems

Arranging sets in regular patterns has struggled mathematicians for centuries. In 1611 Johannes Kepler came up with the conjecture that in 3–dimensional Euclidean space, there is no arrangement of equally sized balls, which exceeds a density of $\frac{\pi}{\sqrt{18}}$. This means that approximately 74% of 3–dimensional Euclidean space are used up by this arrangement. The conjecture became part of the 18^{th} of Hilbert's famous 23 problems [20] and remained unsolved until 2005 when Thomas Hales' proof was published [19]. In this first part we have a closer look at the classical sphere packing problem in 2 dimensions and will ask for an optimal arrangement of ellipses in the plane.

1 The classical Problems

The following section mainly relies on [7] and [30]. We start with the definition of a lattice in d–dimensional Euclidean space and will explain what a packing and a covering is.

Definition 1.1. A *lattice* Λ in d–dimensional Euclidean space is a discrete subgroup of $\mathbb{R}^d$. An invertible matrix $A \in \mathbb{R}^d \times \mathbb{R}^d$ is called a *generator matrix* for the lattice Λ, if

$$\Lambda = \Lambda_A = A \cdot \mathbb{Z}^d.$$

The generator matrix A is non-unique, in fact there are countably many different matrices, generating the same lattice. An example in 2 dimensions is given by the so-called *integer lattice* Λ_{I_2}, which is generated by the identity matrix $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, but also by the matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 3 \\ 1 & 4 \end{pmatrix}$ to name two more of the infinitely many possibilities.

Definition 1.2. A family of subsets $(\Omega_i)_{i \in I}$, with I being a countable set of indices, is called a packing of $\mathbb{R}^d$, if no point of $\mathbb{R}^d$ belongs to the interior of two sets Ω_i, Ω_j when $i \neq j$, i.e.

$$\Omega_i^\circ \cap \Omega_j^\circ = \emptyset \quad \forall i \neq j.$$

A packing is called lattice packing, if it is of the form $(\Omega + \lambda)_{\lambda \in \Lambda}$, where Λ is a lattice.

Definition 1.3. A family of subsets $(\Omega_i)_{i\in I}$, with I being a countable set of indices, is called a covering of $\mathbb{R}^d$, if each point of $\mathbb{R}^d$ belongs to at least one of the sets $\overline{\Omega_i}$, i.e.

$$\mathbb{R}^d = \bigcup_{i\in I} \overline{\Omega_i}\ .$$

A covering is called lattice covering, if it is of the form $(\Omega + \lambda)_{\lambda\in\Lambda}$, where Λ is a lattice.

Remark. The definitions given above do not necessarily ask for *open* sets in order to have a packing or *closed* sets in order to have a covering. Let Ω be the following subset of $\mathbb{R}^2$

$$\Omega = \{(x,y)^T \in \mathbb{R}^2 \mid -1/2 \le x < 1/2 \ \wedge\ -1/2 \le y < 1/2\}$$

and let Λ_{I_2} be the integer lattice. Then

$$(\Omega + \lambda)_{\lambda\in\Lambda_{I_2}}$$

is a lattice packing as well as a lattice covering.

In the given example, the packing and the covering are equal to each other because of the choice of the set. If, instead, we want our sets to be spheres, then a packing of the integer lattice is given by $\left(\mathcal{B}_{1/2}(\lambda)\right)_{\lambda\in\Lambda_{I_2}}$, balls of radius $1/2$ centred at the lattice points and a covering is given by $\left(\mathcal{B}_{\sqrt{2}/2}(\lambda)\right)_{\lambda\in\Lambda_{I_2}}$, balls of radius $\sqrt{2}/2$ centred at the lattice points. These are actually the "best" possible packing and covering that we can achieve for the integer lattice using spheres. This brings us to some definitions which will allow us to measure the quality of a lattice packing or lattice covering.

Definition 1.4. Let $\Lambda = A\mathbb{Z}^d$ be a d–dimensional lattice, generated by the non-singular matrix $A \in \mathbb{R}^d \times \mathbb{R}^d$. The volume of a lattice is defined by

$$vol(\Lambda) = |\det(A)|.$$

Let $(\Omega_p + \lambda)_{\lambda\in\Lambda}$ be a lattice packing, then

$$\rho = \rho_{\Lambda_A,\Omega_p} := \frac{vol(\Omega_p)}{vol(\Lambda)}$$

is called the packing density of Λ. Let $(\Omega_c + \lambda)_{\lambda\in\Lambda}$ be a lattice covering of $\mathbb{R}^d$, then

$$\Delta = \Delta_{\Lambda_A,\Omega_c} := \frac{vol(\Omega_c)}{vol(\Lambda)}$$

is called the covering density of Λ.

Subsequently we ill usually drop some of the indices if the context allows for it.

Note that the packing and covering densities are only defined for a lattice packing and a lattice covering. From Definitions 1.2, 1.3 and 1.4 it is clear that the packing density fulfils

$$\rho_\Lambda \leq 1$$

and that the covering density fulfils

$$\Delta_\Lambda \geq 1.$$

We want the packing as well as the covering density being close to 1.

As already mentioned above, for the integer lattice the radius r_p leading to the *densest* sphere packing is $\frac{1}{2}$ and the radius r_c giving the *thinnest* covering is $\frac{\sqrt{2}}{2}$. The corresponding densities are

$$\rho_{I_2} = \frac{\pi}{4} \approx 0,7854, \qquad \delta_{I_2} = \frac{\pi}{2} \approx 1,5708.$$

For the so-called *hexagonal* lattice, which can be generated by the matrix

$$Hex = \begin{pmatrix} 1 & \cos(\pi/3) \\ 0 & \sin(\pi/3) \end{pmatrix},$$

the densities are closer to 1, as they are

$$\rho_{Hex} = \frac{\pi}{\sqrt{12}} \approx 0,9069, \qquad \Delta_{Hex} = \frac{2\pi}{3\sqrt{3}} \approx 1,2092.$$

It is well-known and has already been proven by Gauss, that for dimension $d = 2$ the latter densities are optimal and can only be achieved for the hexagonal lattice [7]. For a given lattice the sphere packing problem is optimally solved by finding the shortest vector. A first attempt to find the shortest vector is to compare the lengths of the vectors given by the generating matrix. For the integer lattice, deduced from the generating matrix $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, we quickly find out and verify that the shortest vector has length $l = 1$. From this we can conclude that the largest possible radius for spheres which will provide a packing is $r_p = 1/2$. However, the matrix

$$A = \begin{pmatrix} 1 & 4 \\ 1 & 3 \end{pmatrix},$$

is also a generator matrix for the integer lattice, as $(4,3)^T - 3 \cdot (1,1)^T = (1,0)^T$ and $4 \cdot (1,1)^T - (4,3)^T = (0,1)^T$. This time we cannot read off the optimal packing radius directly from the generator matrix. In this case it was rather easy to find the shortest vector, but in general this problem is not easily solvable at all. In order to find the optimal covering radius for dimension $d = 2$, we would need to find the two shortest vectors and additionally take care of the angle in between them. In other words we need to find two points as close as possible to the origin with the restriction that the angle in between the first point, the origin and the second point is acute. The sphere that contains these 3 points has the optimal covering radius.

Remark. As we are interested in the density of a lattice, Definition 1.4 already shows the scale invariance of the solution to the packing and covering problem. Therefore, we will use lattices of volume $vol(\Lambda) = 1$ and even more, we will only use elements of $SL_2(\mathbb{R}) = \{A \in \mathbb{R}^2 \times \mathbb{R}^2 \mid det(A) = 1\}$ as the desired solutions are invariant under changing the orientation of the coordinate system as well as rotations and reflections.

Remark. We see that already in the case of two dimensions the problem is not always easy to solve, depending on how the generator matrix of the lattice is given. The only thing easily computed is the volume of the lattice, which is given by the absolute value of the determinant of the generator matrix and it remains invariant under a change of basis. Anything else requires algorithms which run-times grow exponentially in the dimension. Indeed, the problem of finding the covering radius of an arbitrary lattice is known to be in the class of *NP-hard* problems and the problem of finding the packing radius is conjectured to be in this class [7].

A positive result, especially for low dimensions, is given by Lemma 1.6.

Definition 1.5. Let $\Lambda_A \subset \mathbb{R}^d$ be a d–dimensional lattice. If for any arbitrary vector $v \in \Lambda_A$, $v \notin \{v_1, \ldots, v_d\}$ the inequality $\|v_i\| \leq \|v\|$ for $i = 1, \ldots, d$ and $\langle v_i, v_j \rangle \geq 0$ for all $(i,j) \in \{1, \ldots, d\} \times \{1, \ldots, d\}$ hold true, we call the matrix $A = (v_1, \ldots, v_d)$ a reduced basis for the resulting lattice.

Lemma 1.6. *For every lattice $\Lambda_A \subset \mathbb{R}^d$ generated by a matris A there exists a matrix $\widetilde{A}$ such that $\widetilde{A}$ is a reduced basis for $\Lambda_A = \Lambda_{\widetilde{A}}$.*

Proof. See [7, p.40 ff]. □

Especially in low dimensions such a basis is usually found by using the LLL-algorithm described by Lenstra, Lenstra and Lovász [22].

Remark. For further investigation and the rest of this work we are usually interested in 2–dimensional lattices and may assume that the generator matrix $A = (v_1, v_2)$ is a reduced basis and that $\|v_1\| \leq \|v_2\|$.

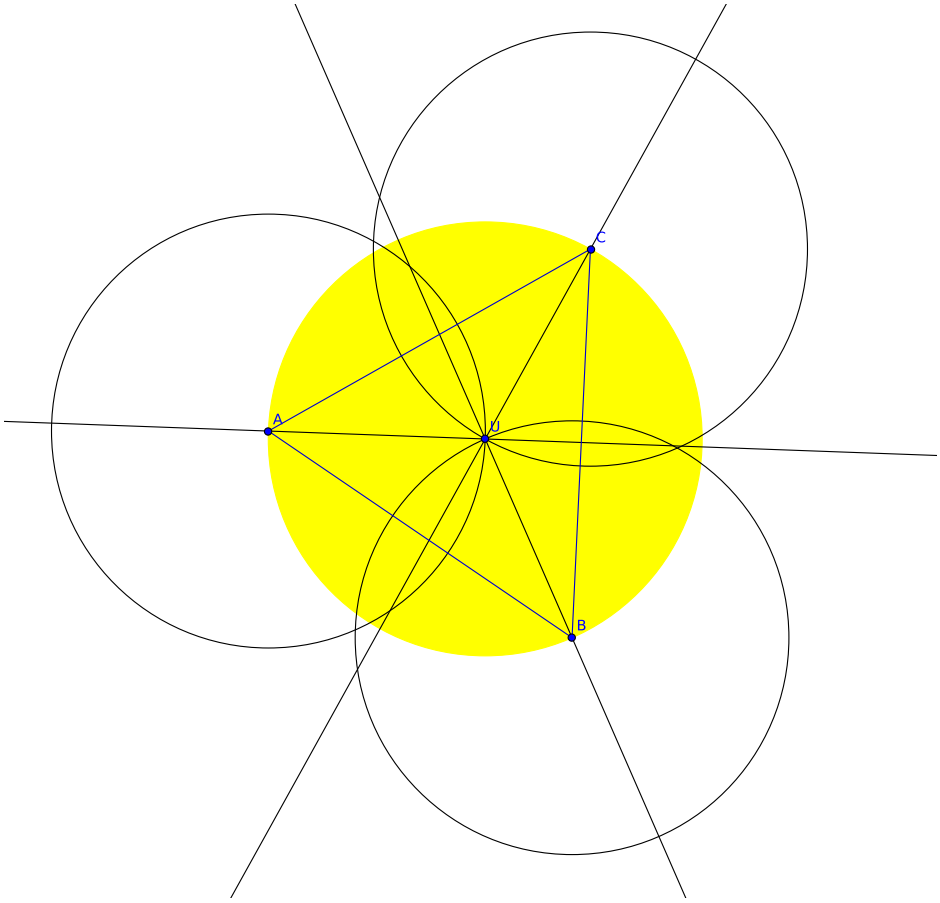


Figure 1: If copies of the circumsphere are centred at the points A, B and C then they will intersect in exactly one point, namely U.

2 Packing and Covering of $\mathbb{R}^2$ with Ellipses

In the upcoming section we restrict ourselves to dimension $d = 2$. The issue for the packing as well as the covering problem seems to be that the circle does not fit as well into the integer lattice as it does into the hexagonal lattice. We will describe construction rules for arbitrary lattices, that lead to an optimal packing using ellipses. We will also see that the ellipse packing problem already solves the ellipse covering problem and hence, we will only be interested in the ellipse packing problem after that point. The idea is based on the singular value decomposition, which is illustrated in Figure 2

The representation of the packing circle for the integer lattice is

$$[I_2 * v]^T * [I_2 * v] = r_{I_2}^2,$$

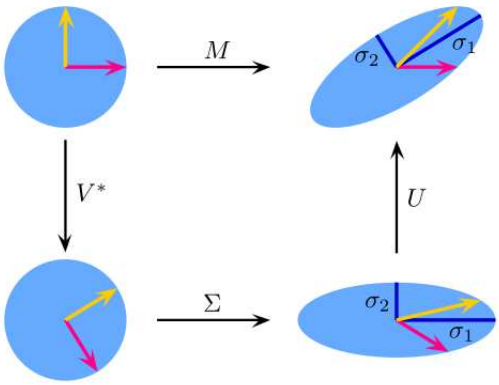


Figure 2: Geometric illustration of the singular value decomposition. Source: http://commons.wikimedia.org/w/index.php?title=File:Singular-Value-Decomposition.svg&oldid=70177732, [13]

where $v = \begin{pmatrix} x \\ y \end{pmatrix}$ and $r_{I_2} = 1/2$.

Changing our integer lattice to the hexagonal lattice also changes our coordinate system and the circle would be represented as

$$[Hex * I_2 * v]^T * [Hex * I_2 * v] = r_{Hex}^2, \tag{2.1}$$

where $Hex = (v_1, v_2)$ is a reduced basis for the hexagonal lattice and $r_{Hex} = 1/2\|v_1\|$. The ellipse we retrieve from (2.1) has the same packing density as the circle has for the integer lattice.

As our problem was the integer lattice as starting point, we now see how to deal with this issue. We want the hexagonal lattice to have a circle as best fitting ellipse. All we have to do is substituting I_2 in (2.1) by Hex^{-1}.

$$\left[Hex * Hex^{-1} * v\right]^T * \left[Hex * Hex^{-1} * v\right] = r_{Hex}^2 \tag{2.2}$$

At first it may seem strange to write down (2.1) and (2.2) in that way. But let us consider an arbitrary lattice Λ_A. Then the best fitting ellipse is retrieved by again simply substituting Hex^{-1} by A^{-1}.

Theorem 2.1. *Given any arbitrary lattice Λ_A generated by the reduced basis A, there exist ellipses $\mathcal{E}_0$ and $\widetilde{\mathcal{E}}_0$, such that*

$$\rho_{A,\mathcal{E}_0} = \frac{\pi}{\sqrt{12}} \approx 0,9069, \qquad \Delta_{A,\widetilde{\mathcal{E}}_0} = \frac{2\pi}{3\sqrt{3}} \approx 1,2092.$$

and such that $\mathcal{E}_0$ provides a lattice packing and $\widetilde{\mathcal{E}}_0$ provides a lattice covering for $\mathbb{R}^2$.

Proof. There are two alternative approaches to prove this statement. At first let us have a look at (2.2), but this time for an arbitrary lattice Λ_A where

$$A = \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = (v_1, v_2)$$

is a reduced basis for Λ_A and $Hex = \sqrt{\frac{2}{\sqrt{3}}} \cdot \begin{pmatrix} 1 & \cos(\pi/3) \\ 0 & \sin(\pi/3) \end{pmatrix}$ is a reduced basis for the hexagonal lattice such that the volume of the hexagonal lattice is equal to 1 and the optimal packing radius is given by $r_{Hex} = \frac{1}{2}\sqrt{\frac{2}{\sqrt{3}}}$. This yields to

$$\left[Hex * A^{-1} * v\right]^T * \left[Hex * A^{-1} * v\right] = r_{Hex}^2 \tag{2.3}$$

which is equivalent to

$$\begin{aligned} & v^T * {A^{-1}}^T * Hex^T * Hex * A^{-1} * v = r_{Hex}^2 \\ \Leftrightarrow \quad & v^T * {A^{-1}}^T * \begin{pmatrix} \frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{3}} \end{pmatrix} * A^{-1} * v = \frac{1}{4}\frac{2}{\sqrt{3}} \\ \Leftrightarrow \quad & v^T * {A^{-1}}^T * \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} * A^{-1} * v = 1 \end{aligned} \tag{2.4}$$

An alternative approach is achieved by explicitly constructing the packing ellipse for the integer lattice. We rotate the whole lattice by 45 degrees, find the ellipse which is centred at the origin, containing the points $\left(0; \pm 1/\sqrt{2}\right)$, $\left(\pm 1/(2\sqrt{2}); 1/(2\sqrt{2})\right)$ and $\left(1/(2\sqrt{2}); \pm 1/(2\sqrt{2})\right)$. The problem of finding a rotated ellipse in the integer case has been transformed to finding an non-rotated ellipse in the so-called *quincunx* case as illustrated in Figure 3.

The equation of our ellipse is as follows.

$$b^2 \cdot x^2 + a^2 \cdot y^2 = a^2 \cdot b^2$$

or

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

where $a = 1/\sqrt{2}$ and $b = 1/\sqrt{6}$. Written as matrix-vector equation we have

$$\left[\begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6} \end{pmatrix} * \begin{pmatrix} x \\ y \end{pmatrix}\right]^T * \left[\begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6} \end{pmatrix} * \begin{pmatrix} x \\ y \end{pmatrix}\right] = 1$$

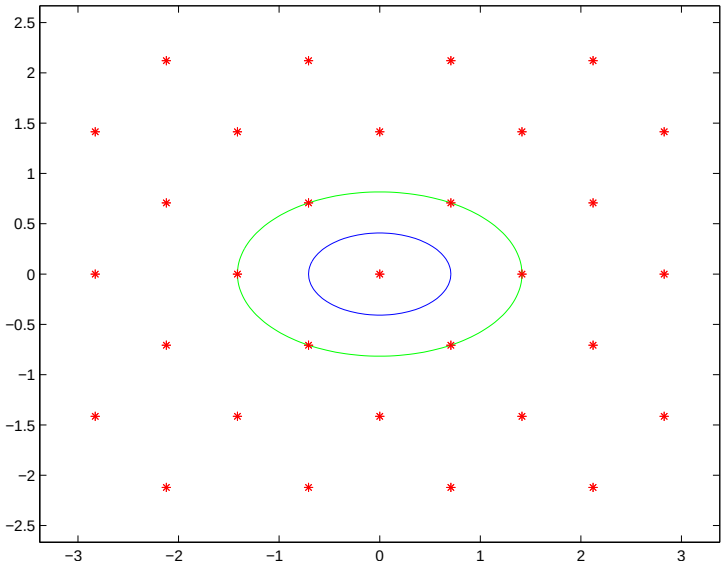


Figure 3: We find the six points closest to the origin and construct an ellipse which contains these points. Scaling the ellipse by 1/2 yields the desired packing ellipse.

We define the dilation matrix $Dil = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6} \end{pmatrix}$ and the rotation matrix $Rot = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. Then our equation for the rotated ellipse in the integer lattice reads as follows.

$$\left[Rot^{-1} * Dil * Rot * v\right]^T * \left[Rot^{-1} * Dil * Rot * v\right] = 1$$

By using the same idea as for equation (2.3), we retrieve

$$\left[Rot^{-1} * Dil * Rot * A^{-1} * v\right]^T * \left[Rot^{-1} * Dil * Rot * A^{-1} * v\right] = 1 \quad (2.5)$$

plugging in the values, we get

$$v^T * A^{-1^T} * \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} * A^{-1} * v = 1 \quad (2.6)$$

We have seen two equivalent approaches on how to construct an ellipse which provides a candidate for an optimal packing. Next we want to show, that the retrieved ellipse from Equation (2.6) really provides a packing by using translates centred at the lattice points of Λ_A.

It is enough to solve this problem locally for 3 neighbouring points, which are w.l.o.g. the origin and the points that are reached from the origin via the

vectors v_1 and v_2 of which we assume that they provide a reduced basis for Λ_A. If we set $\mathcal{E} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ then our sets centred at the origin, $(x_1, y_1)^T$ and $(x_2, y_2)^T$ have the following forms

$$\mathcal{E}_0 = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, v^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} v \leq 1 \right\} \tag{2.7}$$

$$\mathcal{E}_{v_1} = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, (v-v_1)^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} (v-v_1) \leq 1 \right\} \tag{2.8}$$

$$\mathcal{E}_{v_2} = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, (v-v_2)^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} (v-v_2) \leq 1 \right\} \tag{2.9}$$

Solving the system consisting of (2.7), (2.8), (2.9) leads to

$$\begin{aligned} \mathcal{E}_0 \cap \mathcal{E}_{v_1} &= \left(\frac{x_1}{2}, \frac{y_1}{2} \right)^T \\ \mathcal{E}_{v_1} \cap \mathcal{E}_{v_2} &= \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right)^T \\ \mathcal{E}_{v_2} \cap \mathcal{E}_0 &= \left(\frac{x_2}{2}, \frac{y_2}{2} \right)^T . \end{aligned}$$

By using translations we conclude that $(\mathcal{E}_0 + \lambda)_{\lambda \in \Lambda_A}$ provides a lattice packing with ellipses for $\mathbb{R}^2$.

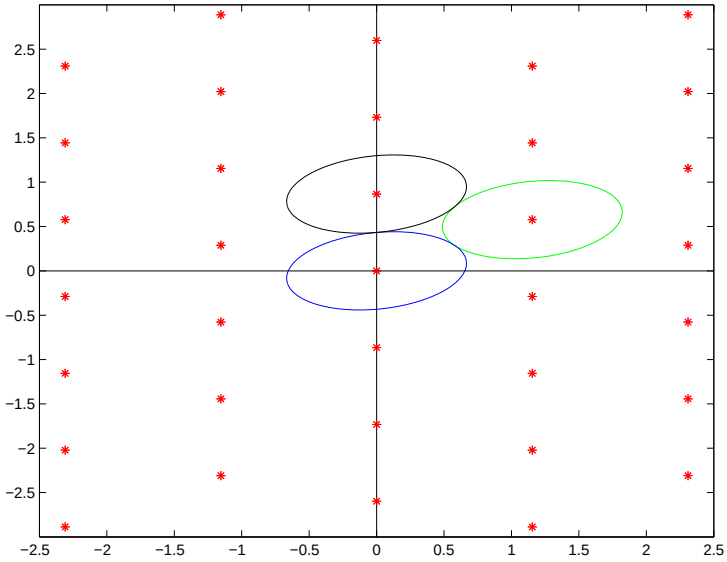


Figure 4: A lattice with the packing ellipses from (2.7), (2.8) and (2.9). By shifting the origin into other points we can repeat the whole process and after countably many steps arrive at a lattice packing for $\mathbb{R}^2$.

Next we will show that the packing density always is

$$\rho_{A,\mathcal{E}_0} = \frac{vol(\mathcal{E}_0)}{vol(\Lambda_A)} = \frac{\pi}{\sqrt{12}} \approx 0,9069.$$

First we have to compute the area of the ellipse $\mathcal{E}_0$. If the ellipse is given as the following set

$$\mathcal{E}_0 = \left\{(x,y)^T \in \mathbb{R}^2 \,|\, v^T \left(A^1\right)^T \mathcal{E}\, A^{-1} v \leq 1\right\}$$

then the area is given by $vol(\mathcal{E}_0) = \det\left(\left(A^1\right)^T \mathcal{E}\, A^{-1}\right)^{-1/2} \pi$. First we compute

$$\det\left(\left(A^1\right)^T \mathcal{E}\, A^{-1}\right) = \det\left(A^{-1}\right)^2 \det\left(Hex\right)^2 = \frac{\det(Hex)^2}{\det(A)^2} = \frac{12}{\det(A)^2}$$

as $\mathcal{E} = Hex^T\, Hex$. This leads to the packing density

$$\rho = \rho_{A,\mathcal{E}_0} = \frac{Vol(\mathcal{E}_0)}{Vol(\Lambda_A)} = \frac{\frac{\det(A)}{\det(Hex)}\pi}{\det(A)} = \frac{\pi}{\sqrt{12}} \approx 0,9069 \quad \forall A \in GL_2(\mathbb{R}).$$

Having done all the work for the ellipse packing, the covering is actually for free. If we compare

$$\rho_{Hex} = \frac{\pi}{\sqrt{12}} \text{ to} \qquad \Delta_{Hex} = \frac{2\pi}{3\sqrt{3}}.$$

from the classical packing and covering problem using spheres, we see that not only the densities are the most economic, but also that the quotient

$$q_{Hex} = \frac{\Delta_{Hex}}{\rho_{Hex}} = \frac{4}{3}$$

has the lowest value for all possible lattices. Knowing that the packing density for ellipses does not change, we can expect the same result for the covering density. Taking into account that q_{Hex} actually also is the ratio of the area of our spheres, namely

$$q_{Hex} = \frac{\frac{Vol(B_R)}{Vol(\Lambda_{Hex})}}{\frac{Vol(B_r)}{Vol(\Lambda_{Hex})}} = \frac{Vol(B_R)}{Vol(B_r)}$$

where R and r are the covering and packing radii respectively for the hexagonal lattice. We expect our ellipses to behave in the same way as our spheres

do in the hexagonal case. Therefore we take equation (2.6) and on the right-hand side we plug in 4/3 instead of 1. The same we do for equations (2.7), (2.8) and (2.9), we multiply the right-hand side only by 4/3. We end up with the system

$$\widetilde{\mathcal{E}}_0 = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, v^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} v \leq \frac{4}{3} \right\} \tag{2.10}$$

$$\widetilde{\mathcal{E}}_{v_1} = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, (v-v_1)^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} (v-v_1) \leq \frac{4}{3} \right\} \tag{2.11}$$

$$\widetilde{\mathcal{E}}_{v_2} = \left\{ (x,y)^T \in \mathbb{R}^2 \,|\, (v-v_2)^T \left(A^{-1}\right)^T \mathcal{E}\, A^{-1} (v-v_2) \leq \frac{4}{3} \right\} \tag{2.12}$$

which has the unique solution

$$\widetilde{\mathcal{E}}_0 \cap \widetilde{\mathcal{E}}_{v_1} \cap \widetilde{\mathcal{E}}_{v_2} = \left(\frac{x_1 + x_2}{3}, \frac{y_1 + y_2}{3} \right)^T.$$

By using translation we get a full lattice covering of $\mathbb{R}^2$, which has covering density

$$\Delta = \Delta_{A,\widetilde{\mathcal{E}}_0} = \frac{2\pi}{3\sqrt{3}}.$$

□

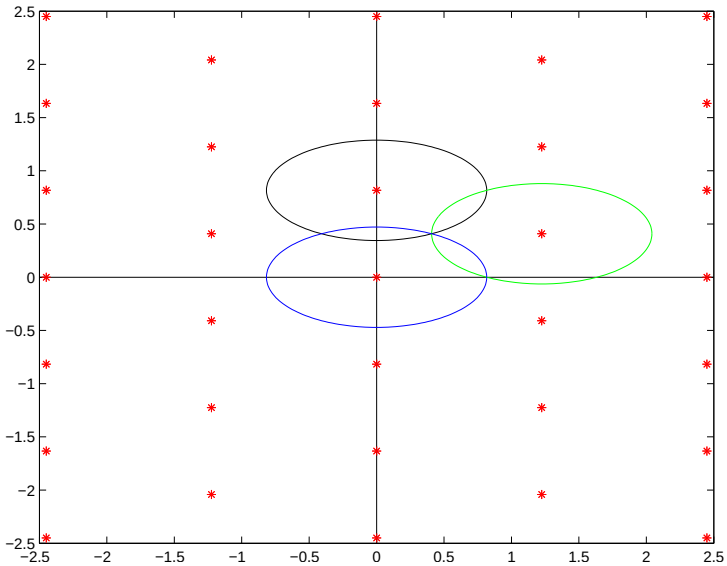


Figure 5: An ellipse lattice covering can be constructed by scaling the packing ellipses by the factor $2/\sqrt{3}$.

Lemma 2.2. *The optimal density of a lattice packing $(\mathcal{E}+\lambda)_{\lambda\in\Lambda}$ using ellipses is $\rho = \rho_{\Lambda,\mathcal{E}} = \frac{\pi}{\sqrt{12}}$ and the optimal density of a lattice covering $(\widetilde{\mathcal{E}} + \lambda)_{\lambda\in\Lambda}$ using ellipses is given by $\Delta = \Delta_{\Lambda,\widetilde{\mathcal{E}}} = \frac{2\pi}{3\sqrt{3}}$. Hence, the optimal ratio is given by $q = \frac{\Delta}{\rho} = \frac{4}{3}$.*

Proof. A proof can be found in [34] and [29]. □

Remark. Comparing (2.4) and (2.6) we have seen that both approaches lead to the same result. From Equations (2.4) and (2.6) we conclude that

$$\begin{aligned}&\left[Rot^{-1} * Dil * Rot * A^{-1}\right]^T * \left[Rot^{-1} * Dil * Rot * A^{-1}\right]\\ =&\left[Dil * Rot * A^{-1} * v\right]^T * \left[Dil * Rot * A^{-1} * v\right]\\ =&\begin{pmatrix} 4 & 2\\ 2 & 4\end{pmatrix},\end{aligned}$$

which in combination with Equations (2.3) with (2.5) shows

$$Hex = r_{Hex} \cdot \widetilde{Rot}^{-1} * Dil * Rot, \tag{2.13}$$

with $\widetilde{Rot} = \begin{pmatrix} \cos(\pi/3) & -\sin(\pi/3)\\ \sin(\pi/3) & \cos(\pi/3)\end{pmatrix}$, so we have found the singular value decomposition of the generator matrix of the hexagonal lattice, up to a scaling factor which is actually the packing radius. If we use any other orthogonal matrix in Equation (2.13) instead of $\widetilde{Rot}$ we have a rotated version of the hexagonal lattice, but still end up at Equation (2.6). In other words, in order to retrieve the hexagonal lattice out of the integer lattice, we have to rotate the integer lattice by 45 degrees, dilate our lattice with the dilation matrix $\begin{pmatrix} \sqrt{2} & 0\\ 0 & \sqrt{6}\end{pmatrix}$ and rotate everything back by -60 degrees or any other arbitrary angle in order to end up with a rotated version. In order to preserve the volume we need to find a scaling factor, which gives the packing radius.

Remark. The solution of the (simultaneous) ellipse packing and covering problem, for a given lattice, is not unique. The use of a reduced basis leads to the least elliptic ellipses. In other words, the condition number of $Hex\, A^{-1}$, $A \in GL_2(\mathbb{R})$, $\Lambda = \Lambda_A$ is closest to 1 if A is a reduced basis. Using any other basis $\widetilde{A}$ also solves the packing as well as the covering problem in an optimal way, but the the condition of the matrix $Hex\, \widetilde{A}^{-1}$ will be greater than the condition number of $Hex\, A^{-1}$. This fact is illustrated in Figure 6. We will use the condition number of $Hex\, A^{-1}$ as a measure of quality of a lattice.

The idea of using quadratic forms has often been used when investigating the quality of lattices [7]. For a positive definite quadratic form Q, using

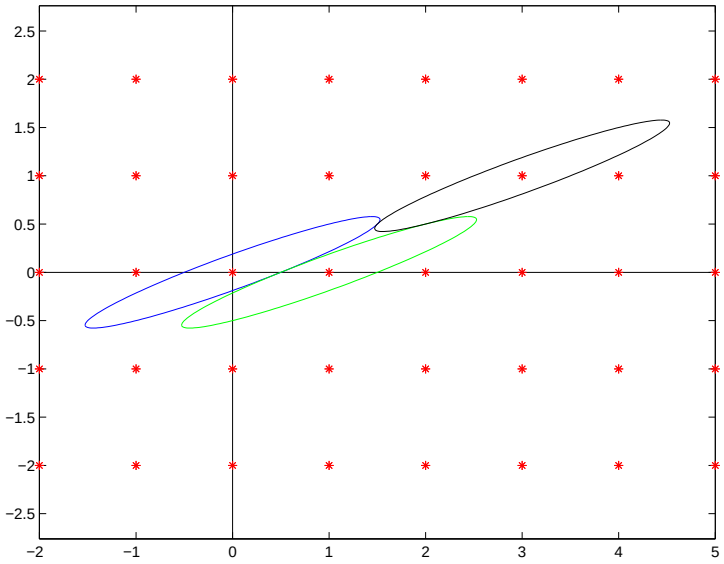


Figure 6: An Ellipse packing for the integer lattice using the vectors $v_1 = (1,0)^T$ and $v_2 = (3,1)^T$ as basis.

the inner product $\langle x, y\rangle_Q = x^T Q\, y$ instead of the standard Euclidean inner product is an essential part in Voronoi's reduction theory, finding Voronoi polytopes and Delauney triangulation [30].

Remark. Assume $A = (v_1, v_2)$ is a reduced basis for the lattice Λ_A and $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(\mathbb{R})$. Then

$$v_1^T \left(A^{-1}\right)^T M\, A^{-1} v_1 = (1,0) M (1,0)^T = a \tag{2.14}$$

and

$$v_2^T \left(A^{-1}\right)^T M\, A^{-1} v_2 = (0,1) M (0,1)^T = d \tag{2.15}$$

and

$$(v_1 - v_2)^T \left(A^{-1}\right)^T M\, A^{-1} (v_1 - v_2) = (1,-1) M (1,-1)^T = a - b - c + d. \tag{2.16}$$

If $M = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$, then the results of (2.14), (2.15) and (2.16) are all equal to 1. This illustrates that using the reduced basis of the hexagonal lattice is an essential part of Theorem 2.1.

3 Diagonal Distortion

In this section we want to investigate the method of diagonal distortion presented by Edelsbrunner and Kerber [8]. It is a method to systematically create new lattices out of the integer lattice. Each point of the d–dimensional integer lattice $\Lambda_{I_d} \subset \mathbb{R}^d$ is mapped onto a new point in the resulting lattice $\Lambda_\delta \subset \mathbb{R}^d$. The diagonal distortion is given by the mapping

$$\begin{aligned} T_\delta \colon \quad & \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \\ & (x, \delta) \mapsto x + \frac{\delta - 1}{d} \cdot \sum_{i=1}^{d} x_i * \overrightarrow{1}, \end{aligned} \tag{3.1}$$

with $x \in \mathbb{R}^d$, $\delta \in \mathbb{R}$ and $\overrightarrow{1} \in \mathbb{R}^d$ is the vector with 1 in each entry. A proof that $\Lambda_\delta = T_\delta(\mathbb{Z}^n)$ is a lattice is given in [8]. As we are only interested in the 2–dimensional case, we can write our distortion mapping as

$$\begin{aligned} {}_2T_\delta \colon \quad & \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2 \\ & (x, y, \delta) \mapsto \begin{pmatrix} x \\ y \end{pmatrix} + \frac{\delta - 1}{2} \cdot (x + y) \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{aligned} \tag{3.2}$$

$x, y, \delta \in \mathbb{R}$.

Lemma 3.1. *Let ${}_2T_\delta$ be the diagonal distortion as described in* (3.2)*. Let $\Lambda_\delta = {}_2T_\delta(\mathbb{Z}^2) \subset \mathbb{R}^2$ and let $(\mathcal{B}_\delta + \lambda)_{\lambda \in \Lambda_\delta}$ be the optimal lattice packing with spheres for the lattice Λ_δ. Then the packing density is given by*

$$\rho(\delta) = \begin{cases} \frac{\pi\delta}{2}, & \text{for } 0 \le \delta \le \frac{1}{\sqrt{3}} \\ \frac{\pi}{8}\left(\delta + \frac{1}{\delta}\right), & \text{for } \frac{1}{\sqrt{3}} \le \delta \le \sqrt{3} \\ \frac{\pi}{2\delta}, & \text{for } \sqrt{3} \le \delta \end{cases} \tag{3.3}$$

Proof. The lemma as well as its proof can be found in [8]. □

We want to describe a lattice Λ of volume $vol(\Lambda) = 1$ via one of its generator matrices. Therefore, we reformulate (3.2) in the following, canonical way.

$$\begin{aligned} \mathcal{D}_\delta \colon \quad & GL_2(\mathbb{R}) \times \mathbb{R} \to GL_2(\mathbb{R}) \\ & (v_1, v_2) \mapsto \left({}_2T_\delta(v_1), {}_2T_\delta(v_2)\right). \end{aligned} \tag{3.4}$$

The results derived in [8] are for integer lattices of arbitrary dimension as initial set of points. It is left as an open question to extend the approach to more general initial sets of points.

We want to investigate the above mentioned method for 2–dimensional lattices and start with the 2–dimensional integer lattice Λ_{I_2}. We define the matrix

$$\mathcal{M}_\delta := \frac{1}{\sqrt{\delta}} \cdot \mathcal{D}_\delta\left(I_2\right) = \frac{1}{\sqrt{\delta}} \cdot \begin{pmatrix} \frac{1+\delta}{2} & \frac{-1+\delta}{2} \\ \frac{-1+\delta}{2} & \frac{1+\delta}{2} \end{pmatrix} \tag{3.5}$$

which we will call *distortion matrix* in the subsequent paragraphs.

It is easy to see that

$$\det\left(\mathcal{M}_\delta\right) = 1$$

as we use $1/\sqrt{\delta}$ as normalising factor. From (3.5) we also see that $M_1 = I_2$. For $\delta = \sqrt{3}$ and $\delta = \frac{1}{\sqrt{3}}$ we get rotated versions of the hexagonal lattice. Indeed, we have $\mathcal{M}_{\sqrt{3}} * \mathcal{M}_{1/\sqrt{3}} = I_2$. This motivates the following definition and lemma.

Definition 3.2. The *distortion group* $DG_2(\mathbb{R})$ is defined as

$$DG_2(\mathbb{R}) := \left\{ \mathcal{M}_\delta \in SL_2(\mathbb{R}) \,|\, \mathcal{M}_\delta = \frac{1}{\sqrt{\delta}} \cdot \begin{pmatrix} \frac{1+\delta}{2} & \frac{-1+\delta}{2} \\ \frac{-1+\delta}{2} & \frac{1+\delta}{2} \end{pmatrix}, \delta > 0 \right\}. \tag{3.6}$$

Lemma 3.3. *$DG_2(\mathbb{R})$ is a group and has the following properties. For $\mathcal{M}_\delta \in DG_2(\mathbb{R})$ the following holds:*

$$\mathcal{M}_1 = I_2, \tag{3.7}$$
$$\mathcal{M}_\delta^{-1} = \mathcal{M}_{1/\delta}, \tag{3.8}$$
$$\mathcal{M}_{\delta_1} * \mathcal{M}_{\delta_2} = \mathcal{M}_{\delta_1 \cdot \delta_2} \tag{3.9}$$

This means that

$$\begin{aligned} \Phi\colon \quad & \mathbb{R}^+ \to DG_2(\mathbb{R}) \\ & \delta \mapsto \mathcal{M}_\delta \end{aligned}$$

is a group representation of the multiplicative group $(\mathbb{R}^+, \cdot)$ on $(DG_2(\mathbb{R}), *)$.

Proof. The proof is by straight forward computation and hence, is left to the interested reader. □

Still, we want to give a geometric, very intuitive explanation which is illustrated in Figures 7 and 8. The way of optimally packing the integer lattice with ellipses is not unique as the integer lattice remains invariant under rotations by 90 degrees. This is not true for the (optimal) packing ellipse. We see that taking the reciprocal distortion parameter $1/\delta$ is the same as using the distortion parameter δ followed by a rotation of 90 degrees. This is the geometric interpretation of Equation (3.8).

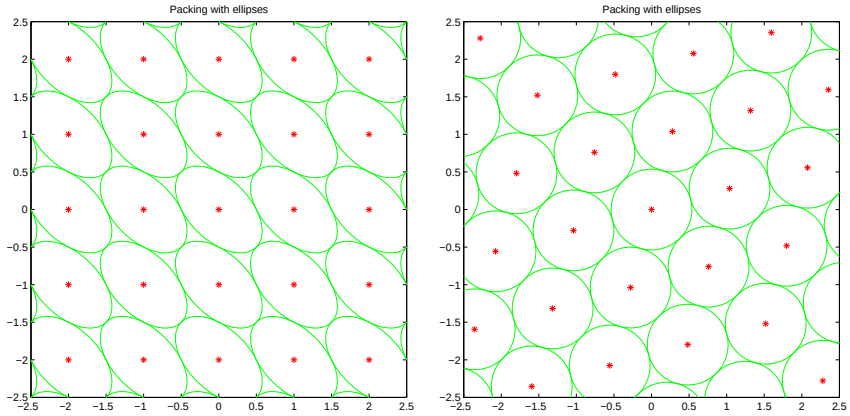


Figure 7: One possible arrangement of packing ellipses. Diagonal distortion by the factor $\sqrt{3}$ leads to a rotated version of the hexagonal lattice.

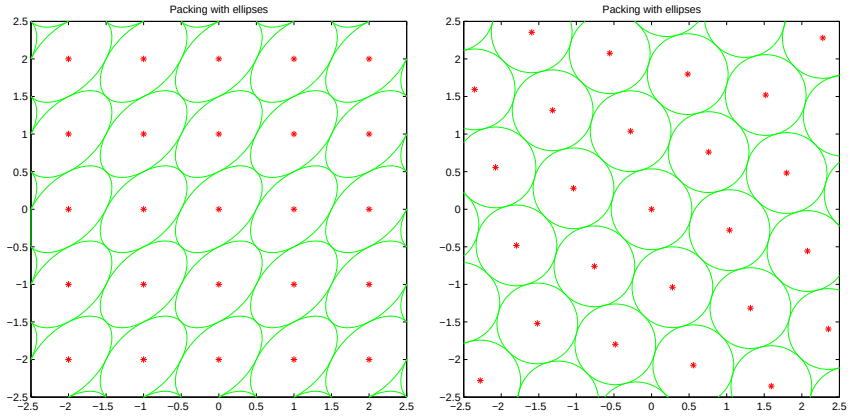


Figure 8: Another possible arrangement of the packing ellipses rotated by 90. The distortion factor $1/\sqrt{3}$ leads to another rotated version of the hexagonal lattice.

Remark. Definition 3.2 can be extended in the following way.

$$\widetilde{DG}_2(\mathbb{R}) = \left\{ \mathcal{M}_\delta \in SL_2(\mathbb{R}) \,|\, \mathcal{M}_\delta = \frac{1}{\sqrt{|\delta|}} \cdot \begin{pmatrix} \frac{1+\delta}{2} & \frac{-1+\delta}{2} \\ \frac{-1+\delta}{2} & \frac{1+\delta}{2} \end{pmatrix}, \delta \in \mathbb{R}\backslash\{0\} \right\}.$$

Then, Lemma 3.3 can be extended by replacing the multiplicative group $(\mathbb{R}^+, \cdot)$ by $(\mathbb{R}\backslash\{0\}, \cdot)$. The geometric interpretation for negative distortion parameter δ is that we first reflect the integer lattice with respect to the second median and apply the diagonal distortion afterwards or vice versa. It is easy to see that Definition 3.2 is still meaningful and that Lemma 3.3 still holds. However, there is no need for this more general setting as $\Lambda_{\mathcal{M}_\delta} = \Lambda_{\mathcal{M}_{-\delta}}$.

Theorem 3.4. *Let $\delta \geq 1$, $\mathcal{M}_\delta$ be defined as above, $\mathcal{E} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$. The eigenvalues of $\left(\mathcal{M}_\delta^{-1}\right)^T \mathcal{E} \mathcal{M}_\delta^{-1}$ are*

$$\lambda_1 = 2\delta, \quad \lambda_2 = \frac{6}{\delta}.$$

The corresponding eigenvectors σ_1, σ_2 are

$$\sigma_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The ratio of the eigenvalues is

$$\lambda = \frac{\lambda_1}{\lambda_2} = \frac{\delta^2}{3}$$

and hence, the condition number of $Hex * \mathcal{M}_\delta^{-1}$ is given by $cond(Hex * \mathcal{M}_\delta = \frac{\delta}{\sqrt{3}}$

Proof. $\left(\mathcal{M}_\delta^{-1}\right)^T = \mathcal{M}_\delta^{-1} = \frac{1}{\sqrt{\delta}} \cdot \begin{pmatrix} \frac{1+\delta}{2} & \frac{1-\delta}{2} \\ \frac{1-\delta}{2} & \frac{1+\delta}{2} \end{pmatrix}$, hence

$$\left(\mathcal{M}^{-1}\right)^T \mathcal{E} \mathcal{M}_\delta^{-1} = \frac{1}{\delta} \cdot \begin{pmatrix} 3+\delta^2 & 3-\delta^2 \\ 3-\delta^2 & 3+\delta^2 \end{pmatrix}$$

and therefore

$$\det(\mathcal{E} - \lambda I_2) = \frac{(\lambda - 2\delta)(\lambda\delta - 6)}{\delta}$$

$$\Rightarrow \ \lambda_1 = 2\delta, \, \lambda_2 = \frac{6}{\delta}, \quad \lambda = \frac{\lambda_1}{\lambda_2} = \frac{\delta^2}{3}.$$

It is easy to verify that

$$\left(\mathcal{M}^{-1}\right)^T \mathcal{E}\mathcal{M}_\delta^{-1} * \sigma_1 = \lambda_1 \cdot \sigma_1 = \begin{pmatrix} 2\delta \\ -2\delta \end{pmatrix}$$

and

$$\left(\mathcal{M}^{-1}\right)^T \mathcal{E}\mathcal{M}_\delta^{-1} * \sigma_2 = \lambda_2 \cdot \sigma_2 = \begin{pmatrix} 6/\delta \\ 6/\delta \end{pmatrix}.$$

Hence, $\lambda = \frac{\lambda_1}{\lambda_2} = \frac{\delta^2}{3}$ and the condition number is $cond(Hex * \mathcal{M}_\delta^{-1}) = \frac{\delta}{\sqrt{3}}$. □

Remark. Assuming that $\delta \geq 1$ is not a real restriction in Theorem 3.4. If we replace δ by $\frac{1}{\delta}$, then the case $0 < \delta \leq 1$ is covered and hence, the condition number of $Hex * \mathcal{M}_\delta^{-1}$ is given by $cond(Hex * \mathcal{M}_\delta^{-1}) = 1/(\sqrt{3}\,\delta)$. Theorem 3.4 together with the result from this remark shows that ellipse packing on a lattice generated by a matrix $\mathcal{M}_\delta \in DG_2(\mathbb{R})$ is optimally solved for $\delta = \sqrt{3}$ and $\delta = \frac{1}{\sqrt{3}}$. Furthermore, Theorem 3.4 shows that the behaviour of the optimal packing density of a distortion lattice described in Lemma 3.1 can be geometrically explained by the packing ellipse. For 2–dimensional lattices that are generated from the integer lattice by diagonal distortion a high packing density for spheres also means that the ellipse packing consists of ellipses with little eccentricity and vice versa. This fact is illustrated in Figure 9.

4 Distortion of the rectangular Lattice

Throughout this section let us assume that $\delta, k \in \mathbb{R}$, $\delta, k \geq 1$. In the last sections we created lattices by diagonal distortion, starting with the unit vectors. Therefore, our resulting lattices inherited the property that the generating matrix contained vectors of equal length. Hence, the next step is to apply the method of diagonal distortion to a rectangular lattice $\Lambda = \frac{1}{\sqrt{k}} \cdot \mathbb{Z} \times \sqrt{k} \cdot \mathbb{Z}$ with $Vol(\Lambda) = 1$. We define the distortion matrix for the rectangular lattice in the following way.

$$\mathcal{M}_{\delta,k} := \mathcal{M}_d * \begin{pmatrix} \frac{1}{\sqrt{k}} & 0 \\ 0 & \sqrt{k} \end{pmatrix} = \frac{1}{2\sqrt{k}} \begin{pmatrix} \sqrt{\delta} + \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} - \frac{1}{\sqrt{\delta}}) \\ \sqrt{\delta} - \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}) \end{pmatrix}.$$

Properties. For $(\delta, k) \in \mathbb{R}^+ \times \mathbb{R}^+$ we state the following properties of $\mathcal{M}_{\delta,k}$, which are easily verified.

$$\mathcal{M}_{1,1} = I_2 \tag{4.1}$$

$$\mathcal{M}_{\delta_1 \cdot \delta_2, k_1 \cdot k_2} = (\mathcal{M}_{\delta_1,1} * \mathcal{M}_{\delta_2,1}) * (\mathcal{M}_{1,k_1} * \mathcal{M}_{1,k_2}) \tag{4.2}$$

$$\mathcal{M}_{1/\delta,1/k} = \mathcal{M}_{\delta,1}^{-1} * \mathcal{M}_{1,k}^{-1} \tag{4.3}$$

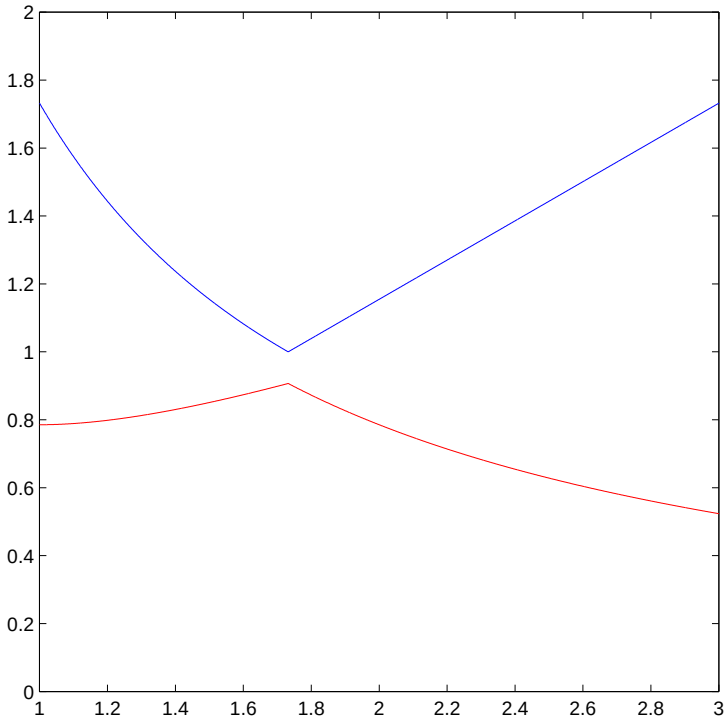


Figure 9: Starting from the integer lattice the packing density first gets better until we are in the hexagonal case and decreases afterwards as δ grows (red). The ratio of the axis of the packing ellipse behaves conversely.

We note that the order in Equations (4.2) and (4.3) is important. In (4.2) the order can be changed within the brackets.

5 The problem of finding the shortest Vector

The main problem for the packing and covering problem always is to find a reduced basis, i.e. finding vectors of shortest length. This problem is rather easy to solve for $\mathcal{M}_{\delta,1}$, but if we have $k > 1$ we get into some trouble. The big issue is that all of a sudden, arbitrary linear combinations of our generating vectors may become shorter than the vectors contained in the matrix $\mathcal{M}_{\delta,k>1}$. Thus let us compare the norm of the first vector (which is by assumption not longer than the second vector) to an arbitrary linear combination of the two generating vectors.

For further investigations we set

$$v_1 = v_1(\delta, k) = \frac{1}{2\sqrt{k}} \begin{pmatrix} \sqrt{\delta} + \frac{1}{\sqrt{\delta}} \\ \sqrt{\delta} - \frac{1}{\sqrt{\delta}} \end{pmatrix}, \; v_2 = v_2(\delta, k) = \frac{1}{2\sqrt{k}} \begin{pmatrix} k(\sqrt{\delta} - \frac{1}{\sqrt{\delta}}) \\ k(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}) \end{pmatrix}$$

and

$$\mathcal{M}_{\delta,k} = (v_1, v_2).$$

Let $m \in \mathbb{Z}\backslash\{0\}$, $n \in \mathbb{Z}$, $k \geq 1$. For growing δ there will be integer linear combinations $\|-m \cdot v_1 + n \cdot v_2\|$ which will provide the shortest vector in the lattice. The minus sign is chosen on purpose and as by the following computations we will see that we only need to consider points in the second and fourth quadrant. The inequality

$$\|-m \cdot v_1 + n \cdot v_2\| \leq \|v_1\|$$

has the solutions

$$\delta \geq \frac{\sqrt{(kn+m)^2 - 1}}{\sqrt{1-(kn-m)^2}} \tag{5.1}$$

and

$$\delta \leq -\frac{\sqrt{(kn+m)^2 - 1}}{\sqrt{1-(kn-m)^2}}.$$

As we want the distortion parameter $\delta > 0$, we are only interested in the positive solution given by equation (5.1). The term under the root of the numerator is non-negative if $(kn+m)^2 \geq 1$ and the denominator is defined if $(kn-m)^2 < 1$. This results in

$$\frac{m}{n} - \frac{1}{n} < k < \frac{m}{n} + \frac{1}{n} \quad \text{for } n > 0$$

respectively

$$\frac{m}{n} + \frac{1}{n} < k < \frac{m}{n} - \frac{1}{n} \quad \text{for } n < 0.$$

As $k \geq 1$ by assumption, we see that the signs of m and n have to be the same, which means that only lattice points in the second and fourth quadrant have the potential to get closer to the origin than the point which is reached from the origin by v_1. This seems quite reasonable, as we always push the points further into the first and third quadrant. Hence, by symmetry with respect to the origin, we may only have a look at points in the second quadrant.

We have seen, that for $k \geq 1$ there are integers m, n such that

$$\|-m \cdot v_1 + n \cdot v_2\| \leq \|v_1\|,$$

but it is not guaranteed that there are no other linear combinations which are shorter, i.e. there may exist integers j, i such that

$$\|-i \cdot v_1 + j \cdot v_2\| \leq \|-m \cdot v_1 + n \cdot v_2\| \leq \|v_1\|,$$

for suitable δ.

For further investigations, it will be useful to think of m, n, i and j as pairs of numbers, i.e. (m, n) and (i, j).

Definition 5.1. Let $v_1, v_2 \in \mathbb{R}^2$. We say that an integer linear combination has *symbol* $(m,n) \in \mathbb{Z} \times \mathbb{Z}$ and identify $-m \cdot v_1 + n \cdot v_2$ with (m,n). If

$$\|-i \cdot v_1 + j \cdot v_2\| < \|-m \cdot v_1 + n \cdot v_2\|.$$

holds, we call (i,j) *shorter* than (m,n) and will write $\|(i,j)\| \leq \|(m,n)\|$.

For $\delta = 1$ the shortest symbol is always given by $(1,0)$ and $(-1,0)$ respectively. Actually, setting $(m,n) = (1,0)$ in Equation (5.1) leads to an indefinite expression. However, by rewriting (5.1) as

$$\delta \sqrt{1-(kn-m)^2} = \sqrt{(kn+m)^2 - 1} \tag{5.2}$$

and setting $\delta = 1$, we see that the last equation holds for $(m,n) = (1,0)$ and that (5.2) is a reinterpretation of (5.1). Alternatively, we could use $(1,\varepsilon)$ in Equation (5.1) and let ε tend to 0 and come up with $\delta = 1$.

We also know that $sgn\left(\frac{m}{n}\right) = 1$ is a necessary condition for $\|(m,n)\| \leq \|(1,0)\|$. A simple calculation shows that $\|(m,n)\| \leq \|(t \cdot m, t \cdot n)\|$ for $|t| \geq 1$. From this we can conclude, that if we think of the pair (m,n) as the fraction $\frac{m}{n}$, this fraction has to be fully reduced, i.e. $gcd(m,n) = 1$. This means that our problem has now turned into a number theoretical problem. Hence, in the upcoming section we will take a short excursion into the field of number theory, but before that we compare $\|(i,j)\|$ to $\|(m,n)\|$.

Lemma 5.2. *Let*

$$v_1 = \frac{1}{2\sqrt{k}} \begin{pmatrix} \sqrt{\delta} + \frac{1}{\sqrt{\delta}} \\ \sqrt{\delta} - \frac{1}{\sqrt{\delta}} \end{pmatrix}, \; v_2 = \frac{1}{2\sqrt{k}} \begin{pmatrix} k(\sqrt{\delta} - \frac{1}{\sqrt{\delta}}) \\ k(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}) \end{pmatrix}.$$

Let $m \in \mathbb{N}\backslash\{0\}$, $n \in \mathbb{N}$, $k \geq 1$ *and* $\delta \geq 1$. *If*

$$\|(m,n)\| < \|(1,0)\|$$

then

$$\delta > 1$$

and for $i \in \mathbb{N}$, $j \in \mathbb{N}$ *and* $\frac{i}{j} \neq \frac{m}{n}$

$$\|(i,j)\| = \|(m,n)\|$$

if, and only if

$$\delta^2 = \frac{(i+kj)^2 - (m+kn)^2}{(m-kn)^2 - (i-kj)^2}.$$

Also,

$$\|(i,j)\| < \|(m,n)\| \Leftrightarrow \delta^2 > \frac{(i+kj)^2-(m+kn)^2}{(m-kn)^2-(i-kj)^2}$$

and

$$\|(i,j)\| > \|(m,n)\| \Leftrightarrow \delta^2 < \frac{(i+kj)^2-(m+kn)^2}{(m-kn)^2-(i-kj)^2}.$$

Proof. The first part of the proof is obvious. If $\delta = 1$, then $\|(1,0)\| \leq \|(m,n)\|$ for all $(m,n) \in \mathbb{Z} \times \mathbb{Z}$, $(m,n) \neq (0,0)$ and the statement follows. The rest follows by direct calculation and is left to the interested reader. □

Remark. In [8, p.13] Edelsbrunner and Kerber state that the one-parameter family introduced in Equation (3.1), contains the optimal lattices for covering in dimension 2,3,4 and 5 and the optimal lattices for packing in dimension 2 and 3 but misses the optimal lattices in higher dimensions. They state the open problem of extending their one-parameter family to a two or more independent parameter family. For dimension $d = 2$ the matrices $\mathcal{M}_{\delta,k}$ produce such a two-parameter family of lattices by identifying a lattice via the generator matrix $\mathcal{M}_{\delta,k}$.

$$\begin{aligned} \Phi\colon \quad & \mathbb{R}^+ \times \mathbb{R}^+ \to SL_2(\mathbb{R}) \\ & (\delta,k) \mapsto \mathcal{M}_{\delta,k} \end{aligned}$$

This family has the properties

$$\begin{aligned} &\Phi(1,1) = I_2 \\ &\Phi(\delta_1\delta_2, k_1k_2) = \Big(\Phi(\delta_1,1)\Phi(\delta_2,1)\Big)\Big(\Phi(1,k_1)\Phi(1,k_2)\Big) \\ &\Phi(1/\delta, 1/k) = \Phi^{-1}(\delta,1)\Phi^{-1}(1,k) \end{aligned}$$

which are equivalent to the properties described in (4.1), (4.2) and (4.3).

Remark. We state it as an open problem whether in dimension higher than 2 the problem can be solved in a similar way. In 3 dimensions for example, it might be enough to divide the space into the x-y-plane, the y-z-plane and the z-x-plane and have a distortion parameter in each of the planes. In addition, using two more parameters (k_1, k_2) to cover the cases where the generating vectors are of different lengths might already describe all possible geometric aspects of three dimensional lattices.

We recall that the distortion matrix is given by

$$\mathcal{M}_{\delta,k} = \frac{1}{2\sqrt{k}} \begin{pmatrix} \sqrt{\delta} + \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} - \frac{1}{\sqrt{\delta}}) \\ \sqrt{\delta} - \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}) \end{pmatrix}. \tag{5.3}$$

As already discussed, we are in the uncomfortable situation that as δ grows the vectors contained in our matrix are no longer a reduced basis for our lattice. In order to solve the packing problem, we need to find the shortest vector in our lattice and the packing radius is given by half of the length of this vector. Hence, we need to find integer linear combinations which produce this vector from our two starting vectors v_1 and v_2.

Lemma 5.3. *Let $\mathcal{M}_{\delta,k}$ be the lattice generating matrix with $0 < k < \infty$ fixed and let $\widetilde{v} = \widetilde{v}(\delta)$ be the shortest vector in the lattice. Then the packing density is given by*

$$\rho_k(\delta) = \frac{\pi}{4}\|\widetilde{v}\|^2. \tag{5.4}$$

Proof. The packing radius is given by half of the Euclidean norm of the shortest possible vector [7]. □

Lemma 5.4. *Let $\mathcal{M}_{\delta,k}$ be the lattice generating matrix with $0 < k < \infty$ fixed and let $\rho_k(\delta)$ be the sphere packing density of the lattice, then*

$$\lim_{\delta\to\infty} \rho_k(\delta) = 0 \Leftrightarrow k \in \mathbb{Q}. \tag{5.5}$$

Proof. Let $k = \frac{p'}{q'}$ with $p', q' \in \mathbb{Z}\backslash\{0\}$ fixed and $sgn(\frac{p'}{q'}) = 1$, then

$$\begin{aligned}
\lim_{\delta\to\infty} \rho_k(\delta) &= \lim_{\delta\to\infty} \frac{\pi}{4}\|(p,q)\|^2 \\
&= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{1}{4k}\left\|-p\begin{pmatrix}\sqrt{\delta}+\frac{1}{\sqrt{\delta}}\\ \sqrt{\delta}-\frac{1}{\sqrt{\delta}}\end{pmatrix} + qk\begin{pmatrix}\sqrt{\delta}-\frac{1}{\sqrt{\delta}}\\ \sqrt{\delta}+\frac{1}{\sqrt{\delta}}\end{pmatrix}\right\|^2 \\
&= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{1}{4k}\left[\left((-p+qk)\sqrt{\delta} + (-p-qk)\frac{1}{\sqrt{\delta}}\right)^2\right. \\
&\qquad\qquad \left. + \left((-p+qk)\sqrt{\delta} - (-p-qk)\frac{1}{\sqrt{\delta}}\right)^2\right] \\
&= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{1}{2k}\left[(-p+q\cdot k)^2\delta + (-p-q\cdot k)^2\frac{1}{\delta}\right] \\
&= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{q}{2p}\left[(-p+q\frac{p'}{q'})^2\delta + (-p-q\frac{p}{q})^2\frac{1}{\delta}\right]
\end{aligned}$$

If we pick $p = t\cdot p'$ and $q = t\cdot q'$, $t \in \mathbb{R}$ we get

$$\begin{aligned}
\lim_{\delta\to\infty} \rho_k(\delta) &= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{q}{2p}\left[(-2p)^2\frac{1}{\delta}\right] \\
&= \lim_{\delta\to\infty} \frac{\pi}{4}\frac{2pq}{\delta} \\
&= 0.
\end{aligned}$$

□

Corollary 5.5. *The norm of a linear combination of the vectors given by the matrix $\mathcal{M}_{\delta,k}$ tends to zero if, and only if, the symbol of the linear combination represents the parameter k with $0 < k < \infty$, i.e. if the symbol of the linear combination is given by (p, q), then the norm of this linear combination tends to zero if, and only if $k = \frac{p}{q}$.*

$$\lim_{\delta \to \infty} \|(p, q)\| = 0 \Leftrightarrow k = \frac{p}{q}$$

Proof. See proof of Lemma 5.4. □

Lemma 5.6. *Let $k \in \mathbb{N}$, $k > 0$ and $\delta \geq 1$. Then the packing density $\rho_k(d)$ of the lattice generated by the matrix $\mathcal{M}_{\delta,k}$ is given by*

$$\rho_k(\delta) = \begin{cases} \frac{\pi}{4}\|(1,0)\|^2, & \delta \leq \sqrt{4k^2 - 1} \\ \frac{\pi}{4}\|(k,1)\|^2, & \sqrt{4k^2 - 1} \leq \delta \end{cases} . \tag{5.6}$$

Proof. We will show that

$$\delta^2 < 4k^2 - 1 \Rightarrow \|(1,0)\| < \|(m,n)\|$$

for all $m, n \in N$ and that

$$\delta^2 > 4k^2 - 1 \Rightarrow \|(k,1)\| < \|(m,n)\|$$

for all $m, n \; in \mathbb{N}$ with $\frac{m}{n} \neq \frac{k}{1}$.

We assume $\|(1,0)\|^2 > \|(m,n)\|^2$. Then we have

$$\begin{aligned} & (-1 + 0 \cdot k)^2\delta + (1 + 0 \cdot k)^2\frac{1}{\delta} > (-m + n \cdot k)^2\delta + (m + n \cdot k)^2\frac{1}{\delta} \\ \Leftrightarrow \quad & \delta^2 + 1 > (nk - m)^2\delta^2 + (m + nk)^2 \\ \Leftrightarrow \quad & \delta^2\left(1 - (nk - m)^2\right) > (m + nk)^2 - 1 \end{aligned}$$

We have 3 cases to consider.

Case 1. If $1 - (nk - m)^2 < 0$, we get

$$\delta^2 < \frac{(m + nk)^2 - 1}{1 - (nk - m)^2} < 0$$

which is not possible.

Case 2. If $1 - (nk - m)^2 > 0$ we get

$$(nk - m)^2 = 0$$

which is only possible for $k = \frac{m}{n}$ which implies that $m = k$ and $n = 1$ as $gcd(m, n) = 1$ is necessary and $k \in \mathbb{N}$. By Lemma 5.2 we know that

$$\|(k, 1)\| < \|(1, 0)\| \Leftrightarrow \delta^2 > \sqrt{4k^2 - 1}.$$

Case 3. If $1 - (nk - m)^2 = 0$ is not possible as

$$\delta^2 \left(1 - (nk - m)^2\right) > (m + nk)^2 - 1$$

would imply that

$$0 > (m + nk)^2 - 1.$$

But, we know that $k \in \mathbb{N}$, $k \geq 1$ and $nk = m + 1$. Thus, $0 > (2m + 1)^2 - 1$.

This shows that

$$\delta^2 < 4k^2 - 1 \Rightarrow \|(1, 0)\| < \|(m, n)\|.$$

Next, we show that

$$\delta^2 > 4k^2 - 1 \Rightarrow \|(k, 1)\| < \|(m, n)\|.$$

for all $\frac{m}{n} \neq \frac{k}{1}$. We assume that

$$\|(k, 1)\| \geq \|(m, n)\|$$

which is equivalent to

$$4k^2 - 1 \geq \frac{4k^2 - (m + nk)^2}{(nk - m)^2} \geq \delta^2$$

as $(m + nk)^2 \geq 1$ and $(nk - m)^2 \geq 1$. Hence, the statement follows and the packing density is given by Equation (5.6). □

Remark. The author conjectures that Farey sequences and continued fractions will help us to understand, how the sphere packing density will behave for the lattices generated by $\mathcal{M}_{\delta,k}$ for $0 < k < \infty$ fixed and $\delta \geq 1$ growing. Therefore, we will give some definitions from the field of number theory and provide an example.

Definition 5.7. Let $a_0 \in \mathbb{Z}$, $a_1, \ldots, a_n \in \mathbb{N} \setminus \{0\}$. We call

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \frac{1}{a_n}}}}$$

a *finite regular continued fraction* and will write $[a_0; a_1, \ldots, a_n]$.

By $[a_0; a_1, a_2, \ldots]$ we denote *infinite regular continued fractions*. We call $[a_0; a_1, \ldots, a_k]$ the *k-th convergent*. In the finite case we have the restriction $k \leq n$.

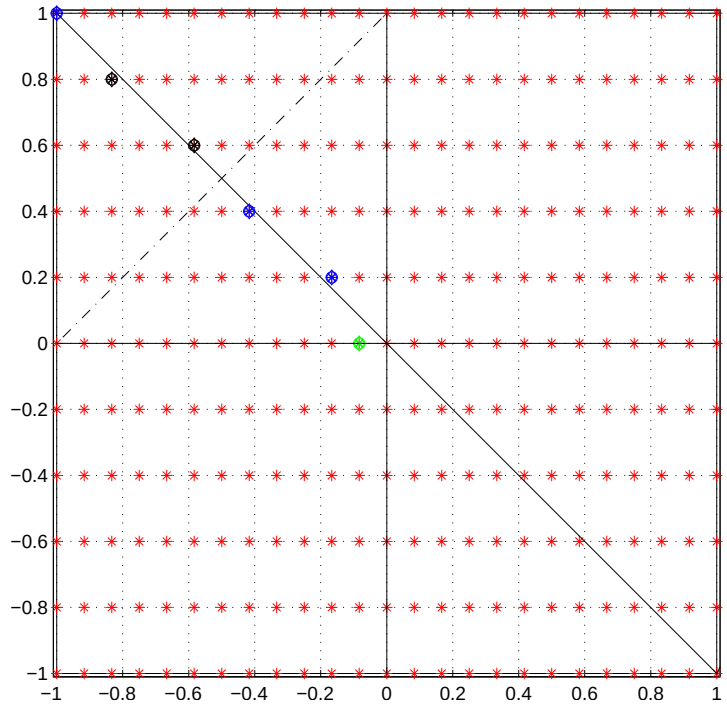


Figure 10: The relevant points of the lattice $\mathcal{M}_{0,12/5}$ are marked in different colors. The points marked black are "counterparts" of the two points marked blue. For the plot we have scaled the lattice to have determinant $\frac{1}{60} = \frac{1}{12 \cdot 5}$.

Definition 5.8. The *Farey sequence* $\mathcal{F}_n$ of order n is the ascending sequence of fully reduced fractions between 0 and 1 whose denominators do not exceed n.

Definition 5.9. The *mediant* $\frac{p''}{q''}$ of the fractions $\frac{p}{q}$ and $\frac{p'}{q'}$ is defined as

$$\frac{p''}{q''} := \frac{p + p'}{q + q'}.$$

Example. We give the following example without proofing the validity of the statements. Assuming the correctness of the example, the packing density for lattices derived by distorting a rectangular lattice is strongly connected with continued fraction representation and convergents and hence, Farey sequences.

We set $k = 12/5 = [2; 2, 2]$ and inspect the lattice $\mathcal{M}_{\delta,k}$, i.e. we are interested in the shortest vector.

As we can see in Figures 10 and 11 we only have to take points into account which are relatively close to the second median. This is due to

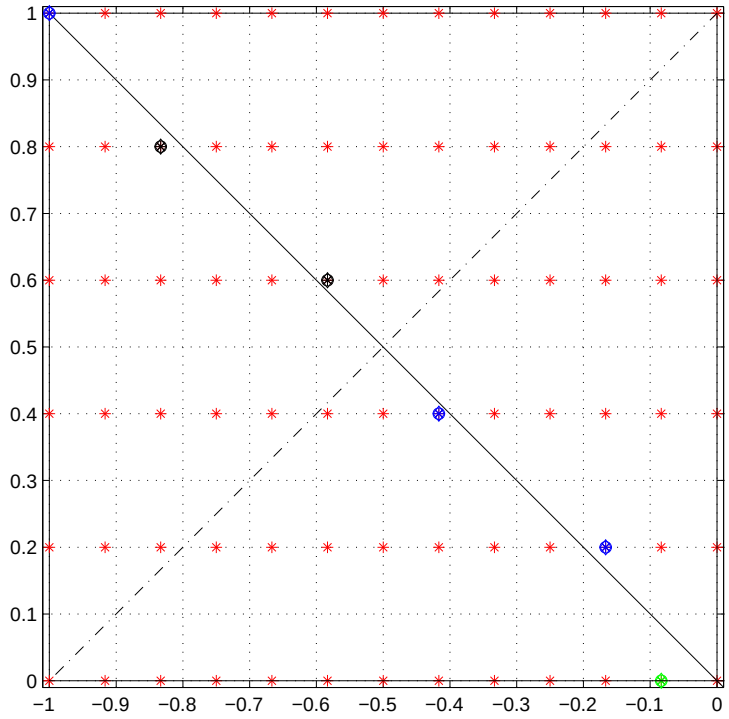


Figure 11: Zoom into the second quadrant of the lattice $\mathcal{M}_{1,12/5}$.

the fact that we push our points away from the diagonal. As we always normalise our lattice to have determinant 1, the points move towards the first median as they slide away from the second median. The green marker gives us the first reference for the packing radius. The two blue markers are the linear combinations $(2, 1)$ and $(5, 2)$ which interpreted as fractions have the continuous fraction representation $[2]$ and $[2; 2]$, hence they are the two convergents of $k = 12/5 = [2; 2, 2]$. The two points marked black have the same distance from the second median as the their blue "counterpart", but their distance to the origin is of course greater, hence they do not play a role for the packing problem. Still we should note that the symbols of these linear combinations are $(10, 4)$ (which is not fully reduced) and $(7, 3)$. We see that k is the mediant of $(2, 1)$ and $(10, 4)$, as well as of $(5, 2)$ and $(7, 3)$. Now as δ grows, we will find that the linear combination $(2, 1)$ will provide the shortest vector. Distorting the lattice even more will lead to the result that $(5, 2)$ will be closest to the origin and finally $(12, 5)$ will slide along the second mediant and approach the origin, leaving no other points a chance to provide a shorter vector. The packing density will behave as shown in Figure 12. As done in [8] we can express the packing density as a piecewise

function, depending on δ.

$$\rho_k(\delta) = \begin{cases} \frac{\pi}{4} \cdot \|(m_0, n_0)\|^2, & \delta \leq \sqrt{\frac{(m_1+kn_1)^2-(m_0+kn_0)^2}{(m_0-kn_0)^2-(m_1-kn_1)^2}} \\ \frac{\pi}{4} \cdot \|(m_\iota, n_\iota)\|^2, & A_\iota \leq \delta \leq B_\iota \\ \frac{\pi}{4} \cdot \|(m_3, n_3)\|^2, & else \end{cases}$$

where

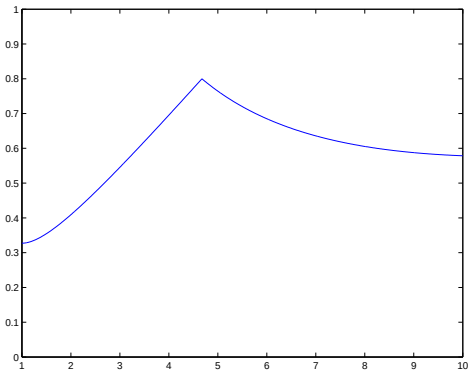
$$A_\iota = \sqrt{\frac{(m_\iota + kn_\iota)^2 - (m_{\iota-1} + kn_{\iota-1})^2}{(m_{\iota-1} - kn_{\iota-1})^2 - (m_\iota - kn_\iota)^2}}$$

$$B_\iota = \sqrt{\frac{(m_{\iota+1} + kn_{\iota+1})^2 - (m_\iota + kn_\iota)^2}{(m_\iota - kn_\iota)^2 - (m_{\iota+1} - kn_{\iota+1})^2}}$$

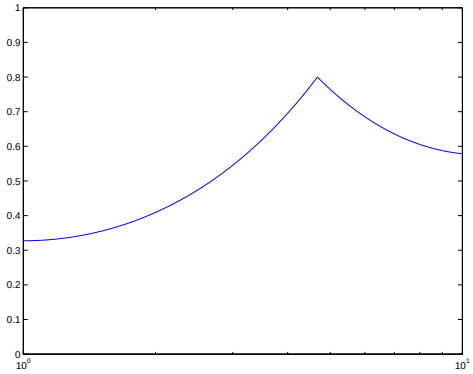
$k = 12/5$, $\iota \in \{1, 2\}$ and

$$\begin{aligned} (m_0, n_0) &= [\,] = (1, 0) \\ (m_1, n_1) &= [2] = (2, 1) \\ (m_2, n_2) &= [2; 2] = (5, 2) \\ (m_3, n_3) &= [2; 2, 2] = (12, 5). \end{aligned}$$

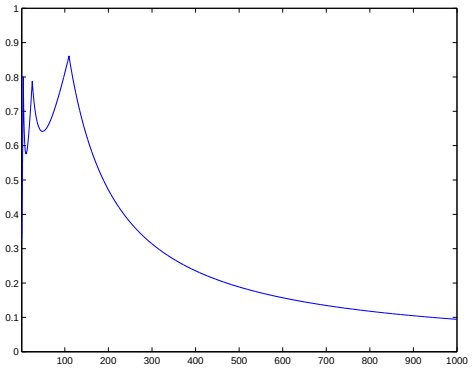
Remark. A strong indicator that Farey sequences and continued fraction representation will be needed in order to understand which linear combination of the vectors v_1 and v_2 provided by the distortion matrix $\mathcal{M}_{\delta,k}$ will provide the shortest possible vector is that both concepts provide a way of labelling fully reduced fractions. Hence, this is also a proper way of labelling pairs of numbers with greatest common divisor 1, which is a necessary condition for solving the packing problem using integer linear combinations of v_1 and v_2.



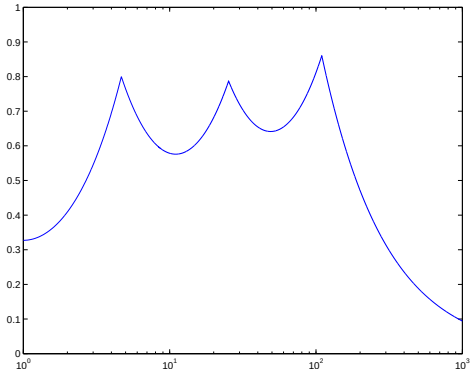
(a) The packing radius for growing δ.



(b) The packing radius for growing δ with logarithmic scale on the δ-axis.



(c) Same as above but with larger range of δ.



(d) Same as above but with larger range of δ. Again logarithmic scale on abscissa.

Figure 12: On the abscissa we have δ running and on the ordinate we have the corresponding packing radius.

Frame Theory

The following part mainly relies on [17]. The upcoming section 6 should be seen as a short motivation for time-frequency analysis and the subsequent sections, where we want to introduce Gabor frames.

6 The Short-Time Fourier Transform

The classical *Fourier Transform* tells us which frequencies occur in a signal (function) in total, but we have no information at what time they occur and at which amplitude. This is due to the fact, that we transform the whole signal at once. The *short-time Fourier transform* (STFT) only considers short time intervals and performs a Fourier transform of the signal in this shorter interval. This is done by choosing an appropriate *window function*, sometimes only called window, which cuts out pieces of the signal. Some qualities one may expect from the window could be smoothness, fast decay or even compact support, i.e. the window function should be well localised.

Remark. The Fourier transform has become a standard tool in mathematics as well as in most natural sciences such as physics, chemistry (see e.g. [14]) and not least in engineering. Hence, we do not introduce the Fourier transform and its practical properties. For details on the classical Fourier transform see e.g. [15], [17] or [21]. The only thing we clarify about the Fourier transform is the normalisation we use, i.e.

$$\mathcal{F}f(\omega) = \widehat{f}(\omega) = \int_{\mathbb{R}^d} f(t)\mathrm{e}^{-2\pi i\omega\cdot t}\,dt$$

and hence, *Plancherel's theorem* takes the following form

$$\|f\|_2 = \|\widehat{f}\|_2.$$

Definition 6.1 (Translation and Modulation Operators)**.** For $x, \omega \in \mathbb{R}^d$ we define the operators

$$T_x f(t) = f(t-x)$$

and

$$M_\omega f(t) = \mathrm{e}^{2\pi i\omega\cdot t} f(t),$$

which we will call *translation* and *modulation* operator respectively. The compositions of the form $T_x M_\omega$ and $M_\omega T_x$ are called *time-frequency shifts*.

Note that the translation and modulation operator do not commute in general, but that we have the following relation

$$T_x M_\omega = \mathrm{e}^{-2\pi i \omega \cdot x} M_\omega T_x. \tag{6.1}$$

From this we can immediately conclude that T_x and M_ω commute if and only if $x \cdot \omega \in \mathbb{Z}$. Subsequently we will often use the notation $\pi(\lambda) = \pi(x, \omega) := M_\omega T_x$ with $\lambda = (x, \omega) \in \mathbb{R}^{2d}$.

Definition 6.2 (Short-Time Fourier Transform)**.** For a fixed, non-zero function $g \in L^2(\mathbb{R}^d)$, called the *window function*, we can define the *short-time Fourier transform* (STFT) of a function $f \in L^2(\mathbb{R}^d)$ with respect to the window function g as

$$V_g f(x, \omega) = \int_{\mathbb{R}^d} f(t) \overline{g(t-x)} \mathrm{e}^{-2\pi i \omega \cdot t} \, dt, \quad \text{for } x, \omega \in \mathbb{R}^d. \tag{6.2}$$

The definition was given for $L^2(\mathbb{R}^d)$, but we will introduce a more general setting for time-frequency analysis right away. For the following definitions see also [9] and [10].

Definition 6.3. A *weight function* v is a non-negative, locally integrable function on $\mathbb{R}^{2d}$. A weight function v on $\mathbb{R}^{2d}$ is called *submultiplicative*, if

$$v(z_1 + z_2) \leq v(z_1) v(z_2)$$

for all $z_1, z_2 \in \mathbb{R}^{2d}$. A weight function m on $\mathbb{R}^{2d}$ is called *v-moderate*, if

$$m(z_1 + z_2) \leq C v(z_1) m(z_2)$$

for all $z_1, z_2 \in \mathbb{R}^{2d}$, $C < \infty$ and a weight function v.

Definition 6.4 (Modulation Space)**.** Let $m(x, \omega)$ be a v-moderate weight function on $\mathbb{R}^{2d}$, let $1 \leq p, q \leq \infty$ and let $g \in \mathcal{S}(\mathbb{R}^d)$ be a fixed, non-zero window in the Schwartz space, then we can define a norm via

$$\|f\|_{M^{p,q}_m} = \left(\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |V_g f(x, \omega)|^p m(x, \omega)^p \, dx \right)^{q/p} d\omega \right)^{1/q}.$$

An element $f \in \mathcal{S}'(\mathbb{R}^d)$ belongs to the *modulation space* $M^{p,q}_m(\mathbb{R}^d)$ if its norm is finite, hence, $\|f\|_{M^{p,q}_m} < \infty$.

Remark. Standard weights on $\mathbb{R}^{2d}$ are typically of polynomial type, e.g. $(1 + |z|)^s$ where $z = (x, \omega)$ or the equivalent weights $(1 + |x| + |\omega|)^s$ or $(1 + |z|^2)^{s/2}$.

Most often in this work we will use the *standard Gaussian*

$$g_0 = 2^{d/4} \mathrm{e}^{-\pi x^2}$$

as our window function, as $\|g_0\|_{L^2} = 1$.

Definition 6.5 (Feichtinger's Algebra)**.** Let g be the standard Gaussian and consider the unweighted space $M_1^{1,1}(\mathbb{R}^d) = M^1(\mathbb{R}^d)$. Then

$$\|f\|_{\mathcal{S}_0} := \|f\|_{M^1} = \|V_g f\|_{L^1}.$$

The space of all functions f for which this norm is finite is called *Feichtinger's Algebra* and is usually denoted by $\mathcal{S}_0(\mathbb{R}^d)$. Hence we define $\mathcal{S}_0(\mathbb{R}^d)$ in the following way.

$$\mathcal{S}_0(\mathbb{R}^d) := \{f \in L^2(\mathbb{R}^d) \mid \|V_g f\|_{L^1} < \infty\}$$

Remark. Once $\mathcal{S}_0$ is defined in the above sense, the space remains invariant under the choice of any other non-zero window $g \in \mathcal{S}_0$. The space is also invariant under the Fourier transform, meaning that if $f \in \mathcal{S}_0$ then also $\widehat{f} \in \mathcal{S}_0$ with $\|f\|_{\mathcal{S}_0} = \|\widehat{f}\|_{\mathcal{S}_0}$. The same is true for time-frequency shifts of f, meaning $M_\omega T_x f \in \mathcal{S}_0$ for all $(x, \omega) \in \mathbb{R}^{2d}$. Further on the Schwartz space $\mathcal{S}$ is contained densely in $\mathcal{S}_0$. Consequently the dual space of $\mathcal{S}_0$ denoted by $\mathcal{S}_0'$ is contained in $\mathcal{S}'$. By using so-called *Banach-Gelfand* triples, i.e. the Banach-Gelfand triple $(\mathcal{S}_0, L^2, \mathcal{S}_0')$, we can extend the STFT and the inner product from L^2 to $\mathcal{S}_0'$, which is a (the) suitable space to do time-frequency analysis. For further notes on Banach-Gelfand triples see ?? (reference needed).

Remark. Having extended the STFT and the inner product to $\mathcal{S}_0'$ by means of Banach-Gelfand triples, we can write the STFT as defined in (6.2) in the following compact form for all $f \in \mathcal{S}_0'$ and $g \in \mathcal{S}_0$ (w.l.o.g. we may think of g as the standard Gaussian as mentioned above).

$$V_g f(\lambda) = \langle f, \pi(\lambda) g \rangle, \quad \lambda = (x, \omega) \in \mathbb{R}^{2d}. \tag{6.3}$$

This generalises the definition of the STFT as given in Definition 6.2, where we only defined the STFT for $f, g \in L^2(\mathbb{R}^d)$.

Remark. Feichtinger's Algebra $\mathcal{S}_0$ is the smallest meaningful setting for the STFT whereas its dual space is the largest possible setting, meaning that the STFT exists for all $f \in \mathcal{S}_0'$ if we choose a window $g \in \mathcal{S}_0$. However, we only wanted to briefly mention this most general setting, but as we do not want to go into technical details, e.g. extending the inner product, we will use a slightly smaller space for further investigations, namely $L^2 \subset \mathcal{S}_0'$ (w*-dense) and its dual which of course is also L^2.

Our next aim is to gain an inversion formula for the STFT. We start with some preparations.

Theorem 6.6 (Orthogonality relations for the STFT)**.** *Let $f_1, f_2, \in L^2(\mathbb{R}^d)$ and $g_1, g_2 \in L^2(\mathbb{R}^d)$, then $V_{g_j} f_j \in L^2(\mathbb{R}^{2d})$ for $j = 1, 2$ and the following relation holds.*

$$\langle V_{g_1} f_1, V_{g_2} f_2 \rangle = \langle f_1, f_2 \rangle \overline{\langle g_1, g_2 \rangle}$$

Proof. A detailed proof can be found in [17, 3.2. p.42]. It makes use of the fact that $L^1 \cap L^\infty(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$ densely and picks the windows from this dense subspace. Then Parseval's formula can be applied and Fubini's Theorem and the density argument from before complete the proof. □

Corollary 6.7 (Inversion Formula for the STFT)**.** *Let $g, \gamma \in L^2(\mathbb{R}^d)$ and let $\langle g, \gamma \rangle \neq 0$. Then for all $f \in L^2(\mathbb{R}^d)$ we have*

$$f = \frac{1}{\langle g, \gamma \rangle} \int_{\mathbb{R}^{2d}} V_g f(x, \omega) M_\omega T_x \gamma \, d\omega dx. \tag{6.4}$$

Proof. Again we will follow the proof as given in [17]. As a consequence of Theorem 6.6 we know that $\|V_g f\|_{L^2(\mathbb{R}^{2d})} = \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}$, i.e. $V_g f \in L^2(\mathbb{R}^{2d})$. Consequently

$$\widetilde{f} = \frac{1}{\langle g, \gamma \rangle} \int_{\mathbb{R}^{2d}} V_g f(x, \omega) M_\omega T_x \gamma \, d\omega dx$$

is well-defined in $L^2(\mathbb{R}^d)$. Using 6.6 once more, we see that

$$\begin{aligned} \langle \widetilde{f}, h \rangle &= \frac{1}{\langle g, \gamma \rangle} \int_{\mathbb{R}^{2d}} V_g f(x, \omega) \overline{\langle h, M_\omega T_x \gamma \rangle} \, d\omega dx \\ &= \frac{1}{\langle g, \gamma \rangle} \langle V_g f, V_\gamma h \rangle = \langle f, h \rangle. \end{aligned}$$

Thus we know that $\widetilde{f} = f$ and the proof is finished. □

Remark. Once more one can think of both g and γ as the standard Gaussian.

7 Discrete Time-Frequency Representations: Gabor Frames

This section mainly relies on [17] again as well as on [6].

In the previous section we have discussed how to recover a given function from an expansion over an uncountable set of time-frequency shifts of given

window functions g and γ. However, $L^2(\mathbb{R}^d)$ is a separable Hilbert space (see e.g. [31]) and hence, it is admissible to expect that it suffices to find a countable set of time-frequency shifted windows in order to represent a function $f \in L^2(\mathbb{R}^d)$. A first attempt to achieve a discrete representation of f could be to replace the integrals in the inversion formula (6.4) by a Riemannian sum like

$$f = \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} c_{k,n} T_{\alpha k} M_{\beta n} g$$
$$\sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha k} M_{\beta n} \gamma \rangle T_{\alpha k} M_{\beta n} g$$

for suitable windows (e.g. the Standard Gaussian) $g, \gamma \in L^2(\mathbb{R}^d)$ and $\alpha, \beta > 0$. The special case $g = \gamma = g_0 = \mathrm{e}^{-\pi x^2}$ and $\alpha = \beta = 1$ was investigated by D. Gabor as early as 1946 (see [12]). Therefore expansions as above are often referred to as *Gabor expansions* and the coefficients are called *Gabor coefficients*. We will call the elements of the set $\{T_{\alpha k} M_{\beta n} g\}$ time-frequency atoms and a first observation yields, that these atoms are not orthogonal in general. Hence the question arises in which sense the above series converges. Another main part of this work will deal with the choice of the parameters α and β, which are called the *lattice parameters* of the *separable* lattice $\alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d$.

Remark. Although many statements are formulated for $\mathbb{R}^d$, we will mostly focus on the special case $d = 1$. Hence, whenever this leads to simplifications in the proofs, we will simply drop the dimension, even if we could formulate the statement for dimensions higher than 1.

Definition 7.1. Let $\mathcal{H}$ be a separable Hilbert space. A labelled family $\{e_j \in \mathcal{H} \,|\, j \in J\}$, with J being an index set, is called a *frame*, if there exist constants $0 < A \leq B$ such that for all $f \in \mathcal{H}$

$$A\|f\|^2 \leq \sum_{j\in J} |\langle f, e_j \rangle|^2 \leq B\|f\|^2. \tag{7.1}$$

Any two constants A, B satisfying (7.1) are called *frame bounds*. If we can choose $A = B$ the frame is called *tight*.

Example.

- Any orthonormal basis is a tight frame with frame bounds $A = B = 1$.
- The union of two (not necessarily different) orthonormal bases is again a tight frame, but with frame bounds $A = B = 2$.

- The set resulting from adding L arbitrary unit vectors to an orthonormal basis is a frame (not tight) with frame bounds $A = 1$ and $B = L+1$. Note that if the L unit vectors form an orthonormal basis again, then the frame is tight as $A = B = 2$, but still, according to the definition, $A = 1$ and $B = L + 1$ are also frame bounds.

As we can see already from these rather trivial examples, frames are a generalisation of the concept of orthonormal bases. In general frames lack orthogonality as well as linear independence. In fact we can think of a frame as a basis with additional elements added. One can easily check, that $E = \{e_j \mid j \in J\}$ has to span the separable Hilbert space $\mathcal{H}$ because otherwise there exists an element $f \in E^{\perp}$, thus A would have to be 0 in order to fulfil (7.1) and hence E cannot be a frame by definition.

What we also see from the examples above is that the frame bounds are not unique. The supremum over all lower frame bounds is usually called *optimal lower frame bound* and analogously the infimum over all upper frame bounds is called *optimal upper frame bound*. They are indeed frame bounds and are the objects of interest further on. Hence, when we talk about upper and lower frame bound, we can think of the optimal upper and lower frame bound. The ratio $Q = B/A$ of (optimal) upper and lower frame bound is called the *frame condition number*. Clearly $Q \geq 1$ with equality if and only if the frame is tight.

Definition 7.2. Let $\mathcal{H}$ be a Hilbert space, $\{e_j \in \mathcal{H} \mid j \in J\}$ be a labelled family for some index set J. We define the *analysis operator* (or *coefficient operator*) C by

$$Cf = \{\langle f, e_j \rangle \mid j \in J\}, \quad f \in \mathcal{H}.$$

Let $I \subset J$ finite. For a finite sequence $c = (c_i)_{i \in I}$, $c_i \in \mathbb{C}$ we define the *synthesis operator* (or *reconstruction operator*) D by

$$Dc = \sum_{i \in I} c_i e_i \in \mathcal{H}.$$

The *frame operator* S is then defined on $\mathcal{H}$ by

$$Sf = \sum_{j \in J} \langle f, e_j \rangle e_j.$$

Remark. In the upcoming section we will construct frames using window functions. We will sometimes write $S_{g,\gamma}^{\alpha,\beta}$ or $S_g^{\alpha,\beta} = S_{g,g}^{\alpha,\beta}$ or use parts of this notation whenever we want to emphasize the dependence of the frame operator on the window or lattice parameters.

We state the following lemma summing up important properties of the frame operator S without proof.

Proposition 7.3. *Let $\{e_j \,|\, j \in J\}$ be a frame for $\mathcal{H}$. Then the following holds.*

(a) The analysis operator C is a bounded operator from $\mathcal{H}$ to $\ell^2(J)$ with closed range.

(b) The analysis operator C and the synthesis operator D are adjoint to each other. Hence, D extends to a bounded operator from $\ell^2(J)$ to $\mathcal{H}$.

(c) The frame operator $S = C^*C = DD^* = DC$ maps $\mathcal{H}$ onto $\mathcal{H}$. It is a positive invertible operator. Let $0 < A \leq B$ be the frame bounds, then

$$A\mathbb{I}_{\mathcal{H}} \leq S \leq B\mathbb{I}_{\mathcal{H}}$$

where $\mathbb{I}_{\mathcal{H}}$ is the identity operator on $\mathcal{H}$. For the inverse operator S^{-1} the following inequality holds

$$B^{-1}\mathbb{I}_{\mathcal{H}} \leq S^{-1} \leq A^{-1}\mathbb{I}_{\mathcal{H}}.$$

(d) The optimal frame bounds are given by

$$B = \|S\|_{Op} \quad \textit{and} \quad A = \|S^{-1}\|_{Op}^{-1}$$

where $\|.\|_{Op}$ is the operator norm.

Proof. As mentioned we omit the proof. A proof can be found in [17, Prop. 5.1.1. p.86-87]. □

Remark. From 7.3 (c) we see that $\{e_j \,|\, j \in J\}$ is a tight frame if and only if $S = A\mathbb{I}_{\mathcal{H}}$, a multiple of the identity.

Corollary 7.4. *Let $\{e_j \,|\, j \in J\}$ be a frame with frame bounds $0 < A \leq B$, then the set $\{S^{-1}e_j \,|\, j \in J\}$ is a frame with frame bounds $0 < B^{-1} \leq A^{-1}$. We call the latter frame the dual frame of the first frame (and vice versa). For every $f \in \mathcal{H}$ we can find an expansion of the form*

$$f = \sum_{j \in J} \langle f, S^{-1}e_j \rangle e_j = \sum_{j \in J} \langle f, e_j \rangle S^{-1}e_j. \tag{7.2}$$

The convergence in equation (7.2) is unconditional in $\mathcal{H}$ for both sums.

Proof. See [17, Prop. 5.1.4., p.89]. □

Note that the convergence in the above corollary is unconditional. This is quite useful as we do not have to care about convergence questions when relabelling our frame elements as well as when interchanging summation and action of linear operators.

We are now equipped with the tools that we need in order to begin discretising our time-frequency representation and get a bit closer to the very heart of this work.

Definition 7.5. Let $g \in L^2(\mathbb{R}^d)$ be a (non-zero) window function and let $\alpha, \beta > 0$ be lattice parameters (for the separable lattice $\alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d$). The set of time-frequency shifted version of g,

$$\mathcal{G}(g, \alpha, \beta) = \{T_{\alpha k} M_{\beta n} g \,|\, k, n \in \mathbb{Z}^d\},$$

is called a *Gabor system*. If $\mathcal{G}(g, \alpha, \beta)$ is a frame for $L^2(\mathbb{R}^d)$, it is called a *Gabor frame* or *Weyl-Heisenberg frame*. The *Gabor frame operator* is then given by

$$\begin{aligned} Sf &= \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha k} M_{\beta n} g\rangle T_{\alpha k} M_{\beta n} g \\ &= \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} V_g f(\alpha k, \beta n) M_{\beta n} T_{\alpha k} g. \end{aligned} \tag{7.3}$$

In equation (7.3) we are in the comfortable situation that the order of the time-frequency shifts is not important as the additional phase factor $\mathrm{e}^{-2\pi i x\omega}$ resulting from $T_x M_\omega = \mathrm{e}^{-2\pi i x\omega} M_\omega T_x$ appears linearly and conjugate-linearly and hence cancels. Also, due to the unconditional convergence of the frame operator, the order of summation is of no importance. In Corollary 7.4 we have seen how to expand $f \in L^2(\mathbb{R}^d)$ using a frame and its dual. The next proposition will do the same for Gabor frames and we will get an insight on how the dual window looks.

Proposition 7.6. *Let $\mathcal{G}(g, \alpha, \beta)$ be a frame for $L^2(\mathbb{R}^d)$. Then there exists a window function $\gamma \in L^2(\mathbb{R}^d)$ called the dual window to g, such that $\mathcal{G}(\gamma, \alpha, \beta)$ is the dual frame to $\mathcal{G}(g, \alpha, \beta)$ and hence for $f \in L^2(\mathbb{R}^d)$ we have the following expansion*

$$\begin{aligned} f &= \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha k} M_{\beta n} g\rangle T_{\alpha k} M_{\beta n} \gamma \\ &= \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha k} M_{\beta n} \gamma\rangle T_{\alpha k} M_{\beta n} g. \end{aligned}$$

The convergence is unconditional in $L^2(\mathbb{R}^d)$ and further on we have the following inequalities

$$A\|f\|_{L^2}^2 \leq \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} |V_g f(\alpha k, \beta n)|^2 \leq B\|f\|_{L^2}^2,$$

$$B^{-1}\|f\|_{L^2}^2 \leq \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} |V_\gamma f(\alpha k, \beta n)|^2 \leq A^{-1}\|f\|_{L^2}^2$$

Proof. By straight forward computation we see that the Gabor frame operator commutes with time-frequency shifts. Let $f \in L^2(\mathbb{R}^d)$, $r, s \in \mathbb{Z}^d$, then

$$\begin{aligned}
&(T_{\alpha r}M_{\beta s})^{-1} S\, T_{\alpha r}M_{\beta s}\, f \\
&\qquad = (T_{\alpha r}M_{\beta s})^{-1} \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} \langle T_{\alpha r}M_{\beta s} f, T_{\alpha k}M_{\beta n} g\rangle T_{\alpha k}M_{\beta n} g \\
&\qquad = \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} \langle f, (T_{\alpha r}M_{\beta s})^{-1} T_{\alpha k}M_{\beta n} g\rangle\, (T_{\alpha r}M_{\beta s})^{-1} T_{\alpha k}M_{\beta n} g \\
&\qquad = \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha(k-r)}M_{\beta(n-s)} g\rangle T_{\alpha(k-r)}M_{\beta(n-s)} g \\
&\qquad = Sf
\end{aligned}$$

after rearranging the indices which is not a problem because of the unconditional convergence of the sum. Note that the phase factor that should appear following (6.1) as a result of interchanging translation and modulation cancels as it appears both linearly and conjugate linearly (we already used this argument once in (7.3)). From this we can conclude that the inverse frame operator S^{-1} also commutes with time-frequency shifts and thus the elements of the dual frame are

$$S^{-1}\left(T_{\alpha k}M_{\beta n} g\right) = T_{\alpha k}M_{\beta n} S^{-1} g.$$

We set $\gamma = S^{-1}g$ and call it the dual window to g. For the rest we refer to Corollary 7.4 and [17]. □

As a corollary of the previous results we get

Corollary 7.7. *If $\mathcal{G}(g, \alpha, \beta)$ is a frame for $L^2(\mathbb{R}^d)$ then the inverse frame operator is given by*

$$S_g^{-1} f = S_\gamma f = \sum_{k\in\mathbb{Z}^d}\sum_{n\in\mathbb{Z}^d} \langle f, T_{\alpha k}M_{\beta n}\gamma\rangle T_{\alpha k}M_{\beta n}\gamma,$$

where $\gamma = S^{-1}g \in L^2(\mathbb{R}^d)$ is the dual window to g.

Proof. The proof follows by applying Proposition 7.6 and Lemma 7.3 (c). □

As a conclusion from $S_g^{-1} = S_\gamma$ we get that the inverse frame operator is already determined by the dual window γ. Combining that with Proposition 7.6 we know that a series expansion of f is given by

$$f = \sum_{k\in\mathbb{Z}^d} \sum_{n\in\mathbb{Z}^d} V_g f(\alpha k, \beta n) M_{\beta n} T_{\alpha k} \gamma. \tag{7.4}$$

This means that we can reconstruct our function f from *samples* of its STFT with respect to g and the dual window γ on a discrete and countable set $\alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \subset \mathbb{R}^{2d}$ with coefficients in $\ell^2(\mathbb{Z}^{2d})$. Furthermore, we can conclude that it is enough to solve the equation $S\gamma = g$ in order to gain the reconstruction in (7.4) which saves us a lot of trouble, as we do not have to solve the general equation $Sf = \widetilde{f}$ for each f we want to reconstruct via samples of the STFT.

Remark. As a final remark in this section we want to mention that the frame operator depends on the window g as well as on the lattice parameters α and β. So far we only considered separable lattices of the form $\alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d$. The corresponding lattice generating matrix hence would be a diagonal matrix. Consequently in the upcoming part of this work we will start working with non-separable lattices. That are lattices which generating matrix is a shear matrix (upper/lower triangle matrix). We will also restrict ourselves to the case $d = 1$ and $g = g_0$ mostly.

8 A Remark on the Frame Constants

As the heart of this work will be the experimental comparison of the Gabor frame constants with window g_0 and the geometry of the lattice in the time-frequency plane, we should briefly mention why the frame constants are of such big interest. The purpose of Gabor frames is to have a reconstruction method which is more stable under the influence of disturbances compared to orthonormal basis. This is due to the overcompleteness of the system, hence losing coefficients (information) does not necessarily imply, that we are not able to reconstruct our function (signal) up to a bearable error. We will introduce the frame algorithm which can be found in [17]. It is described as a more convenient and efficient way of iterative construction compared to series expansions, which needs the calculation of the dual window, which is not always easy.

Algorithm. Given a relaxation parameter $0 < \lambda < 2/B$, we set $\delta = \max\{|1 - \lambda A|, |1 - \lambda B|\} < 1$. We start with $f_0 \equiv 0$ and define the recursion

formula

$$f_{n+1} = f_n + \lambda S(f - f_n),$$

where S denotes the frame operator and f the function (signal) we want to reconstruct. Then $\lim_{n\to\infty} f_n = f$ with a geometric rate of convergence,

$$\|f - f_n\| \leq \delta^n \|f\|. \tag{8.1}$$

Remark. The reason why the reconstruction works is that $f_1 = \lambda S f$ and hence contains the frame coefficients as input.

We omit the proof of the convergence of the above algorithm, and refer the interested reader to [17]. We want to mention the following. If λ is small, then δ is rather large (close to 1), which results in slow convergence as can be seen from (8.1). If we want to choose the parameter λ in an optimal way, we have to set $\lambda = \lambda_{opt} = \frac{2}{A_{opt}+B_{opt}}$, which yields to

$$\delta = \frac{B_{opt} - A_{opt}}{B_{opt} + A_{opt}} = \frac{\frac{B_{opt}}{A_{opt}} - 1}{\frac{B_{opt}}{A_{opt}} + 1} = \frac{Q-1}{Q+1},$$

where Q is the frame condition number as introduced in section 7. This means that the convergence speed of the above algorithm also relies on good estimates on the frame bounds, which is a cumbersome problem. Consequently in [17] it is suggested to combine the algorithm with acceleration methods from numerical analysis found in e.g. [16] or [18]. Note that for a tight frame, the iteration process is finished within one step. As we will see in the following sections, there have already been observations that better properties in the geometry lead to better frame bounds and hence, faster and more stable reconstruction of signals. Nevertheless, there is not much theory yet, describing the connection between geometric properties and frame constants.

From Gabor Frames to Geometry

9 The Ambiguity Function

The following definitions can be looked up in [17].

Definition 9.1. The *cross-ambiguity function* of a function $f \in L^2(\mathbb{R}^d)$ with respect to a function $g \in L^2(\mathbb{R}^d)$ is defined as

$$\begin{aligned} A_g f(x,\omega) &= \int_{\mathbb{R}^d} f\left(t+\frac{x}{2}\right)\overline{g\left(t-\frac{x}{2}\right)}\mathrm{e}^{-2\pi i\omega t}\,dt \\ &= \mathrm{e}^{\pi i\omega x}\,V_g f(x,\omega). \end{aligned}$$

If $f = g$ we speak of the *ambiguity function* and write $Af(x,\omega)$.

Definition 9.2. We say a function $f \in L^2(\mathbb{R}^d)$ is ε-concentrated on a measurable set $\Omega \subset \mathbb{R}^d$, if

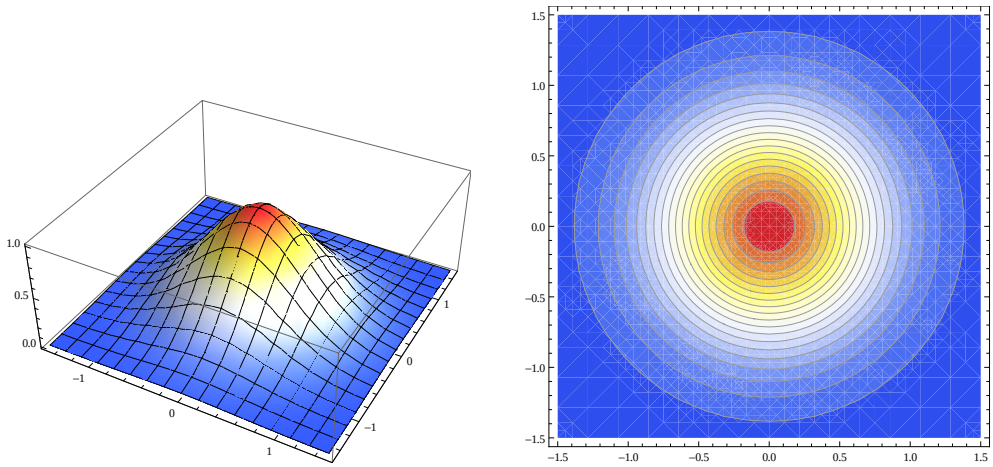
$$\|f - f\cdot \mathbb{1}\!\!\!\!\mathbb{1}_\Omega\| = \left(\int_{\mathbb{R}^d\setminus\Omega} |f(x)|^2\,dx\right)^{1/2} \leq \|f\|\cdot\varepsilon.$$

If $0 \leq \varepsilon \leq 1/2$ we call $supp_\varepsilon(f) = \Omega$ the *essential support* of f. Ω is called *exact support* if $\varepsilon = 0$ can be achieved.

Example. As an example we will compute the ambiguity function of the 1-d standard Gaussian $g_0 = 2^{1/4}\mathrm{e}^{-\pi x^2}$.

$$\begin{aligned} Ag_0(x,\omega) &= \int_{\mathbb{R}} g_0\left(t+\frac{x}{2}\right)\overline{g_0\left(t-\frac{x}{2}\right)}\mathrm{e}^{-2\pi i\omega t}\,dt \\ &= \int_{\mathbb{R}} 2^{1/4}\mathrm{e}^{-\pi(t+x/2)^2}2^{1/4}\mathrm{e}^{-\pi(t-x/2)^2}\mathrm{e}^{-2\pi i\omega t}\,dt \\ &= \sqrt{2}\int_{\mathbb{R}} \mathrm{e}^{-2\pi(t^2+x^2/4)}\mathrm{e}^{-2\pi i\omega t}\,dt \\ &= \sqrt{2}\,\mathrm{e}^{-\pi\frac{x^2}{2}}\int_{\mathbb{R}} \frac{1}{\sqrt{2}}\mathrm{e}^{-\pi t^2}\mathrm{e}^{-2\pi i\frac{\omega}{\sqrt{2}}t}\,dt \\ &= \mathrm{e}^{-\pi\frac{x^2}{2}}\cdot 2^{-1/4}\cdot\widehat{g_0}\left(\omega/\sqrt{2}\right) \\ &= \mathrm{e}^{-\pi\frac{x^2}{2}}\cdot 2^{-1/4}\cdot g_0\left(\omega/\sqrt{2}\right) \\ &= \mathrm{e}^{-\pi\frac{x^2}{2}}\mathrm{e}^{-\pi(w/\sqrt{2})^2} \\ &= \mathrm{e}^{-\frac{\pi}{2}\left(x^2+\omega^2\right)} \end{aligned}$$

Here $\widehat{g_0}$ denotes the Fourier transform of the standard Gaussian, which is again the standard Gaussian, i.e. $\widehat{g_0}(\omega) = g_0(\omega)$. A proof for the Fourier invariance of g_0 can be found in [17, Lemma 1.5.1., p.17]. As can be seen from Figure 13, the ambiguity function Ag_0 is radial symmetric and the essential support can be chosen as a disc for some $\varepsilon > 0$.



(a) 3-d plot of the ambiguity function of g_0.

(b) Contour plot of the ambiguity function of g_0.

Figure 13: The ambiguity function of g_0.

Remark. The ambiguity function determines a function up to a phase factor, meaning that $Af = A(cf)$ for some constant c with $|c| = 1$. This property also shows up when we try to recover the function from its ambiguity function. Similar to the above example, we interpret the ambiguity function as the Fourier transform of $t \mapsto f(t + x/2)\overline{f(t - x/2)}$, where $x \in \mathbb{R}^d$ is fixed. Then, by the inversion formula for the Fourier transform, we can write

$$f(t + x/2)\overline{f(t - x/2)} = \int_{\mathbb{R}^d} Af(x, \omega) \mathrm{e}^{2\pi i t\omega} \, d\omega.$$

For $t = x/2$ we have

$$f(x) = \frac{1}{\overline{f(0)}} \int_{\mathbb{R}^d} Af(x, \omega) \mathrm{e}^{\pi i x\omega} \, d\omega \tag{9.1}$$

for suitable f, i.e. for all $f \in \mathcal{S}(\mathbb{R}^d)$, $f(0) \neq 0$. As mentioned $Af = A(cf)$, hence we actually do not have a unique solution as cf with $|c| = 1$ also is a solution of (9.1) (see [17, p. 61-62]).

We note from Definition 9.1 that the ambiguity function is closely related to the STFT. Indeed they only differ by a phase factor. Most often one is interested in the energy density which we call the *spectrogram* of the function f (with respect to the window $g \in L^2(\mathbb{R}^d)$ with $\|g\| = 1$). The spectrogram is then defined as $spec_g f(x, \omega) = |V_g f|^2$. Hence $spec_g f = |A_g f|^2$ and the essential supports of $V_g f$ and $A_g f$ are the same, as by Theorem 6.6 $\|V_g f\| = \|A_g f\| = \|f\|\|g\|$. Consequently we conclude from the earlier example that the essential support of the STFT of the standard Gaussian in the time-frequency domain is a disc.

The *Heisenberg uncertainty principle* states that a function $f \in L^2(\mathbb{R}^d)$ cannot be arbitrarily well-concentrated in the time and the frequency domain at the same time. It is also well-known that the (possibly translated and modulated) Gaussian minimises the Heisenberg uncertainty relation uniquely. Hence, the conclusion is admissible that discs are the right domain to measure concentration in the time-frequency plane for the Gaussian.

Another import criterion for a Gabor system is the *redundancy*. The redundancy describes the overcompleteness of the Gabor system and will be exactly defined in the subsequent work. In [1] Dörfler and Abreu also conjecture that for given *redundancy* $red > 1$ the condition number of a Gabor frame with Gaussian window g_0 is optimal for the hexagonal lattice. This has already (partially) been observed by Strohmer and Beaver (see [28]). In their paper they show up some relation between the essential support of the ambiguity function of a signal f and classical lattice packing in applications. They state a problem for *orthogonal frequency-division multiplexing* (OFDM) an efficient technology for wireless data transmission (see [35]). Back in 2003 this method had applications in audio broadcasting, digital terrestrial TV broadcasting and broadband indoor wireless systems (see [5] and [23]). In [28] it is stated that the OFDM on rectangular time-frequency lattices is not optimal and a more general approach, the lattice-OFDM (LOFDM) is proposed, using general time-frequency lattices. Their numerical investigations show that the hexagonal lattice provides a better frame condition number than a rectangular lattice when using a Gaussian window. However, at the time of this work, no proof is known for this conjecture.

10 Connecting Frames and Geometry

From here on we restrict ourselves to the case with dimension $d = 1$. In the following we will investigate some connections between Gabor frames and

the related lattice given by a matrix

$$M = \begin{pmatrix} a & 0 \\ s & b \end{pmatrix}$$

which already provides a reduced basis for $0 \leq s \leq b$.

First we have to extend the definition of the frame operator (Definition 7.5) a bit, as we now want to deal with non-separable lattices.

Definition 10.1. Given a window function $g \in L^2(\mathbb{R})$, a lattice $\Lambda = M\mathbb{Z}^2$ with $M = \begin{pmatrix} a & 0 \\ s & b \end{pmatrix}$ and lattice parameters $a, b, s > 0$ and $h, k \in \mathbb{Z}$, the set of time-frequency shifted versions of g,

$$\mathcal{G}(\Lambda) = \mathcal{G}(g, a, b, s) = \{T_{ah} M_{bk+sh} g\},$$

is called a Gabor system. If $\mathcal{G}(g, a, b, s)$ is a frame for $L^2(\mathbb{R}^d)$ we call it a Gabor frame. The Gabor frame operator is of the form

$$\begin{aligned} Sf &= \sum_h \sum_k \langle f, T_{ah} M_{bk+sh} g \rangle T_{ah} M_{bk+sh} g \\ &= \sum_h \sum_k V_g f(ah, bk + sh) M_{bk+sh} T_{ah} g. \end{aligned}$$

In order to be a frame, the system has to fulfil the frame condition

$$A \|f\|^2 \leq \sum_h \sum_k |V_g f(ah, bk + sh)|^2 \leq B \|f\|^2$$

for some constants $A, B > 0$.

That was the easy part of changes we have to make if we want to investigate non-separable Gabor frames numerically. Even though we want to think of our windows and signals as continuous functions in $L^2(\mathbb{R})$, there is no possibility to implement algorithms to investigate these signals without discretising everything. First of all we cannot implement algorithms for signals of infinite signal length. Hence we have to set a maximal signal length $L \in \mathbb{N}$, which means that our signals are no longer in $L^2(\mathbb{R})$, but in $L^2([0, L))$, hence periodic with period length L. In a next step we need to sample our function as well as the window, meaning we can only consider information at certain points (in time). We have to take numerical incorrectness and machine precission into account, e.g. in MATLAB R2012 the numerical precission, found via the command
>> eps

is approximately $2.2204 \cdot 10^{-16}$. Hence, we decide for an $\varepsilon \in [10^{-16}, 10^{-12}]$ and set our signal and window function equal to 0 outside their essential supports. Depending on the choice of ε the essential support for the standard Gaussian g_0 lies within $[-3,3] \subset supp_\varepsilon(g_0) \subset [-3.5,3.5]$. If we want to use a Gaussian window for implementation, we now need to periodise and sample it. A ready-to-use toolbox which can deal with all the above mentioned issues when implementing time-frequency analysis is the LTFAT (see [27]). Depending on the chosen signal length, our window as well as the signal are no longer functions within $L^2(\mathbb{R})$, but vectors in $\mathbb{C}^L$. Consequently we also need a discrete form of the Fourier transform. We will now introduce the DFT (Discrete Fourier Transform) very quickly. For more details see [24], [25] or [26].

Definition 10.2. The DFT of a vector $f \in \mathbb{C}^L$ is given by

$$\widehat{f}(k) = \frac{1}{\sqrt{L}} \sum_{l=0}^{L-1} f(l) e^{-2\pi i k \cdot l/L},\ k = 0, \ldots, L-1. \tag{10.1}$$

Similarly the inverse DFT is given by

$$\check{f}(k) = \frac{1}{\sqrt{L}} \sum_{l=0}^{L-1} f(l) e^{2\pi i k \cdot l/L},\ k = 0, \ldots, L-1. \tag{10.2}$$

Remark. Actually the above defined transform is the *unitary* DFT. The factor $\frac{1}{\sqrt{L}}$ is usually not in the definition of the DFT, instead there is a factor $\frac{1}{L}$ in the formula for the inverse DFT. However, using this normalisation is appropriate for our goals, as we want Fourier invariance of the standard Gaussian.

As the nodes in the DFT are chosen equidistant and the signals are of finite length L, they can be mapped to the torus and hence the nodes correspond to the unit roots. Using this fact, the DFT can be implemented in a way such that the number of operations is reduced from the order of $O(n^2)$ to the order of $O(n \log(n))$ (see [26]). Algorithms using this fact are called *Fast Fourier Transform* (FFT). The FFT is already a built-in function in MATLAB R2012.

Definition 10.3. The *redundancy* of a (discrete) Gabor system $\mathcal{G}(\Lambda)$ is given by $L/\det(\Lambda)$.

The redundancy of a Gabor system measures how many times more coefficients are computed, than are at least necessary in order to reconstruct a signal properly. Choosing a lattice in the time-frequency plane leading

to the critical density $L/det(\Lambda) = 1$ cannot provide a Gabor systems with good time-frequency concentration, i.e. not with a Gaussian window, by the *Balian-Low Theorem* (see [11] and [17]). Hence good time-frequency concentration leads to redundant systems. The more redundant the system, the easier is the reconstruction, but the greater is the computational effort. This means we have to choose between reconstruction precision and computational effort. For our computations we will usually use a redundancy between 1.5 and 4.

We have now discussed almost all tools we need in order to start our experiments. There is one final remark we want to make. When working with discrete signals of finite length L, there are some restrictions to our lattice Λ we have to make, namely that Λ has to be a subgroup of the cyclic group $\mathbb{Z}_L \times \mathbb{Z}_L$ as a necessary condition for $\mathcal{G}(\Lambda)$ being a frame. I.e. we have the following proposition which can be found in [32].

Proposition 10.4. *For every subgroup of $\mathbb{Z}_L^2$ there exist unique $a, b|L$ and $0 \leq s \leq b$ with $s \in \frac{ab}{gcd(ab,L)}\mathbb{Z}$, such that*

$$\Lambda = \begin{pmatrix} a & 0 \\ s & b \end{pmatrix}$$

with $\Lambda\, \mathbb{Z}_L^2 \subset \mathbb{Z}_L^2$.

Loosely speaking, for a given signal length L and a chosen redundancy red, there are only finitely many possible lattices, but not all of them will lead to a proper Gabor system.

Example. We choose $L = 144$ and $red \in \{1.5, 2, 4\}$. For redundancy $red = 1.5$ we have 60 possible lattices. There are only 8 separable lattices ($s = 0$), with

$$(a, b) \in \{(2, 48), (4, 24), (6, 16), (8, 12), (12, 8), (16, 6), (24, 4), (48, 2)\}.$$

The 52 non-separable lattices are generated by shearing the separable ones. For $red = 2$ there are 195 possible lattices, 12 of them are separable and for $red = 4$ we have 91 lattices, where 9 of them are separable. A MATLAB routine generating all possible lattices for given signal length and redundancy can be found at the database section at www.nuhag.eu.

Example. Let $L = 144$ and $red = 1.5$. We will now look at the 8 separable lattices only, consisting of $\Lambda = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$, with

$$(a, b) \in \{(2, 48), (4, 24), (6, 16), (8, 12), (12, 8), (16, 6), (24, 4), (48, 2)\}.$$

Clearly the last four lattices have the same geometric properties for sphere packing and covering, as well as for ellipse packing, as their counterparts within the first four lattices, as they can be retrieved from them by rotation of 90 degrees and reflections. We start with comparing the ratio of the axis of the best fitting packing ellipses. The ratio takes the following values

$$\{27.7249,\ 6.9773,\ 3.1997,\ 2.0025,\ 2.0025,\ 3.1997,\ 6.9773,\ 27.7249\}.$$

The routine used for computing these values can be found in the database section at www.nuhag.eu. Next we compute the frame condition numbers, which take the following values

$$\{4.11 \cdot 10^{10},\ 267.7458,\ 8.1616,\ 2.4379,\ 2.4379,\ 8.1616,\ 267.7458,\ 4.11 \cdot 10^{10}\}$$

The frame condition numbers have been computed with a routine from the LTFAT. In this very simple example, we can clearly see that bad geometric properties of the lattice, result in bad frame constants. Hence we have to ask ourselves whether this is a coincidence because of the simplicity of this example or not. As a next step we take some shears into account.

We pick one of the two good lattices and fix $a = 8$ and $b = 12$ and take the different possible shears $s = 0, 2, 4, 6, 8, 10, 12$ into account (note that a shear by 12 is the same as not shearing the lattice). We start with the ratio of the axis of the best fitting ellipses. They are as follows

$$\{2.0025,\ 1.6582,\ 1.4054,\ 1.2990,\ 1.4054,\ 1.6582,\ 2.0025\}.$$

The frame condition numbers are as follows

$$\{2.4379,\ 2.2345,\ 1.8511,\ 1.6625,\ 1.8511,\ 2.2345,\ 2.4379\}.$$

We see that the condition number as well as the ratio of the axis becomes better, in the sense that they get smaller, when s starts running and after we have reached $s = b/2$ they become worse again.

We will now proof some geometric properties, of which we will see in the following examples, that they seem to carry over to the Gabor frame properties. First we need the following lemma.

Lemma 10.5. *Let $f \in C^1(\mathbb{R})$ be positive and $g \in C^1(\mathbb{R})$ be non-negative and $f(x) - g(x) \neq 0\, \forall x \in \mathbb{R}$, then the following equivalence holds true.*

$$\left(\frac{f+g}{f-g}\right)' = 0 \Longleftrightarrow \left(\frac{g}{f}\right)' = 0.$$
$$\left(\frac{f+g}{f-g}\right)' < 0 \Longleftrightarrow \left(\frac{g}{f}\right)' < 0.$$
$$\left(\frac{f+g}{f-g}\right)' > 0 \Longleftrightarrow \left(\frac{g}{f}\right)' > 0.$$

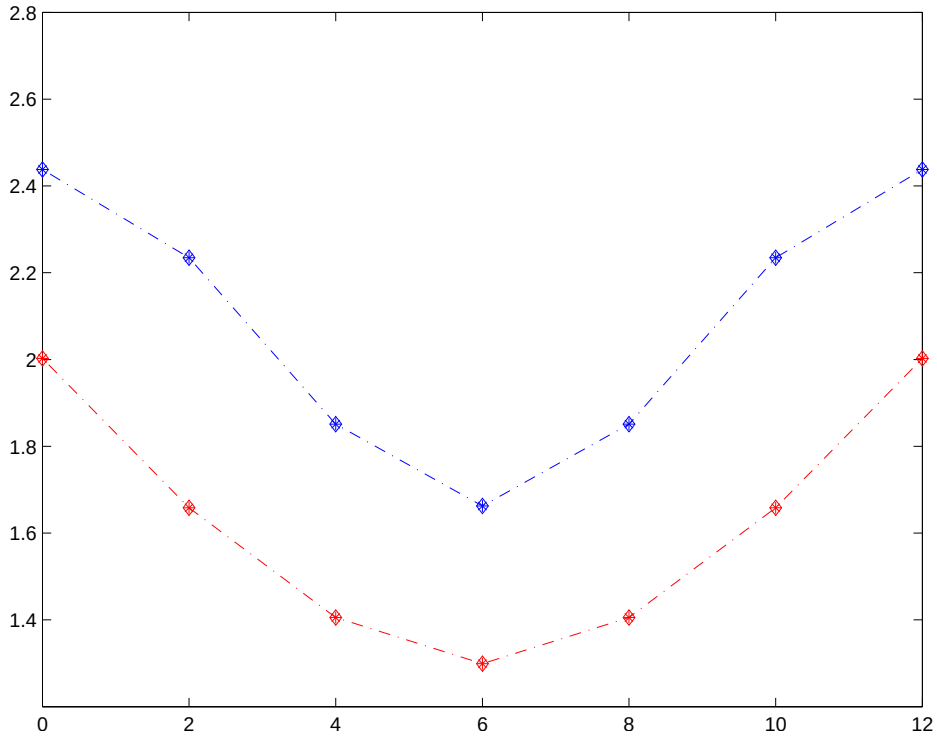


Figure 14: We fix $a = 8$ and $b = 12$. On the abscissa we have s running. Blue: Frame condition number. Red: Ratio of axis of best fitting ellipse.

Proof.

$$
\begin{aligned}
& 0 = \left(\frac{f+g}{f-g}\right)' \\
\Leftrightarrow \quad & 0 = \frac{(f'+g')(f-g)-(f+g)(f'-g')}{(f-g)^2} \\
\Leftrightarrow \quad & 0 = g'f - f'g \\
\Leftrightarrow \quad & 0 = \frac{g'f - gf'}{f^2} \\
\Leftrightarrow \quad & 0 = \left(\frac{g}{f}\right)'
\end{aligned}
$$

The same arguments can be used for the inequalities. □

Remark. We could strengthen the assumption of $g \in C^1(\mathbb{R})$ being non-negative to $g \in C^1(\mathbb{R})$ being positive and $f(x) - g(x) \neq 0 \, \forall x \in \mathbb{R}$. Then we

would have the equivalences

$$\left(\frac{f+g}{f-g}\right)' = 0 \Longleftrightarrow \left(\frac{g}{f}\right)' = 0 \Longleftrightarrow \left(\frac{f}{g}\right)' = 0.$$

The same holds true for negative functions. However, we will only use the weaker statement of g being non-negative.

Lemma 10.6. *Let* $0 < a \in \mathbb{R}$ *be fixed,* $0 < b \in \mathbb{R}$, $M = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$. *Let* $E = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ *and define*

$$\mathcal{E} := \left(M^{-1}\right)^T * E * M^{-1}$$

as in Theorem 3.4. Let λ_1, λ_2 *be the eigenvalues of* $\mathcal{E}$, *such that* $\lambda_1 \geq \lambda_2$. *Then* $\lambda = \frac{\lambda_1}{\lambda_2}$ *takes the minimum for* $b = a$.

Proof. $\mathcal{E} = \begin{pmatrix} \frac{4}{a^2} & \frac{2}{ab} \\ \frac{2}{ab} & \frac{4}{b^2} \end{pmatrix}$ and the characteristic polynomial of $\mathcal{E}$ is of the form

$$p_{\mathcal{E}}(\lambda) = \lambda^2 - \text{trace }(\mathcal{E}) \cdot \lambda + \det(\mathcal{E}).$$

The eigenvalues can now be computed by simply setting $p_{\mathcal{E}}(\lambda) = 0$. This equation has the solutions

$$\lambda_{1,2} = \frac{\text{trace}(\mathcal{E}) \pm \sqrt{\text{trace}(\mathcal{E})^2 - 4 \cdot \det(\mathcal{E})}}{2}. \tag{10.3}$$

In order to minimise $\lambda = \frac{\lambda_1}{\lambda_2}$ we make use of Lemma 10.5. We set

$$f(b) = \text{trace}(\mathcal{E})$$

and

$$g(b) = \sqrt{\text{trace}(\mathcal{E}) - 4 \cdot \det(\mathcal{E})}.$$

We leave it to the interested reader to verify that the assumptions of Lemma 10.5 are fulfilled. Hence we can find the critical points by setting $\left(\frac{g}{f}\right)' = 0$. After simplification we have

$$\frac{g(b)}{f(b)} = \frac{\sqrt{a^4 + b^4 - a^2b^2}}{a^2 + b^2}$$

and consequently we get (again after simplification)

$$\left(\frac{g(b)}{f(b)}\right)' = \frac{3a^2b(b-a)(b+a)}{(a^2+b^2)\sqrt{a^4+b^4-a^2b^2}} = 0.$$

By assumption we have $0 < a$ and $0 < b$, so the above equation reduces to

$$b^2 = a^2.$$

It is easy to see that $\left(\frac{g}{f}\right)'$ is continuous on $\mathbb{R}^+$ (even on $\mathbb{R}$). From the last two equations we also see that $\left(\frac{g}{f}\right)' < 0$ for $b < a$ and $\left(\frac{g}{f}\right)' > 0$ for $b > a$ which carries over to $\left(\frac{f+g}{f-g}\right)'$ by Lemma 10.5. This shows that we have a global minimum for $b = a$. □

Lemma 10.7. *Let $0 < a, b \in \mathbb{R}$ be fixed, $0 \leq s \leq b$, $M = \begin{pmatrix} a & 0 \\ s & b \end{pmatrix}$. Let $E = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ and define*

$$\mathcal{E} := \left(M^{-1}\right)^T * E * M^{-1}$$

as in Theorem 3.4. Let λ_1, λ_2 be the eigenvalues of $\mathcal{E}$, such that $\lambda_1 \geq \lambda_2$. Then $\lambda = \frac{\lambda_1}{\lambda_2}$ takes the minimum for $s = b/2$.

Proof. $\mathcal{E} = \begin{pmatrix} \frac{4(b^2-bs+s^2)}{a^2b^2} & \frac{2a(b-2s)}{a^2b^2} \\ \frac{2a(b-2s)}{a^2b^2} & \frac{4a^2}{a^2b^2} \end{pmatrix}$ and the characteristic polynomial of $\mathcal{E}$ is of the form

$$p_{\mathcal{E}}(\lambda) = \lambda^2 - \text{trace}\,(\mathcal{E}) \cdot \lambda + \det(\mathcal{E}).$$

The eigenvalues can now be computed by simply setting $p_{\mathcal{E}}(\lambda) = 0$. This equation has the solutions

$$\lambda_{1,2} = \frac{\text{trace}(\mathcal{E}) \pm \sqrt{\text{trace}(\mathcal{E})^2 - 4 \cdot \det(\mathcal{E})}}{2}. \tag{10.4}$$

In order to minimise $\lambda = \frac{\lambda_1}{\lambda_2}$ we make use of Lemma 10.5. We set

$$f(s) = \text{trace}(\mathcal{E})$$

and

$$g(s) = \sqrt{\text{trace}(\mathcal{E}) - 4 \cdot \det(\mathcal{E})}.$$

We leave it to the interested reader to verify that the assumptions of Lemma 10.5 are fulfilled. Hence we can find the critical points by setting $\left(\frac{g}{f}\right)' = 0$. After simplification we have

$$\frac{g(s)}{f(s)} = \frac{\sqrt{a^4 - a^2(b+s)^2 - 3a^2s^2 + (b^2 - bs + s^2)^2}}{a^2 + b^2 - bs + s^2}$$

and consequently we get (again after simplifying)

$$\left(\frac{g(s)}{f(s)}\right)' = \frac{-3a^2b^2(b-2s)}{(a^2 + b^2 - bs + s^2)^2 \sqrt{a^4 - a^2(b+s)^2 + 3a^2s^2 + (b^2 - bs + s^2)^2}}. \tag{10.5}$$

Hence the equation

$$\left(\frac{g(s)}{f(s)}\right)' = 0$$

has the solution

$$s = b/2.$$

The denominator of $\left(\frac{g}{f}\right)'$ is positive for all $s \in \mathbb{R}$, hence $\left(\frac{g}{f}\right)'$ is continuous on $\mathbb{R}$. From equation (10.5) we compute that $\left(\frac{g}{f}\right)' < 0$ for $s < b/2$ and $\left(\frac{g}{f}\right)' > 0$ for $s > b/2$. By Lemma 10.5 the same is true for $\left(\frac{f+g}{f-g}\right)'$, which shows that we have a global minimum for $s = b/2$. □

Remark. We call a lattice generated by a matrix $M = \begin{pmatrix} a & 0 \\ b/2 & b \end{pmatrix}$ a quincunx. The last lemma shows that the quincunx is preferable over the separable (rectangular) lattice with same a and b from the geometric point of view. The example above suggests, that it would be worth investigating whether the same is true for Gabor frames.

Lemma 10.8. *Let* $0 < a \in \mathbb{R}$ *be fixed,* $0 < b$, $0 \leq s \leq b$, $M = \begin{pmatrix} a & 0 \\ s & b \end{pmatrix}$. *Let* $E = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$ *and define*

$$\mathcal{E} := {M^{-1}}^T * E * M^{-1}$$

as in Theorem 3.4. Let λ_1, λ_2 *be the eigenvalues of* $\mathcal{E}$, *such that* $\lambda_1 \geq \lambda_2$. *Then* $\lambda = \frac{\lambda_1}{\lambda_2}$ *takes the minimum for the hexagonal lattice, meaning* $b = \frac{2a}{\sqrt{3}}$ *and* $s = b/2 = \frac{a}{\sqrt{3}}$ *and* $\lambda = 1$.

Proof. $\mathcal{E} = \begin{pmatrix} \frac{4(b^2-bs+s^2)}{a^2b^2} & \frac{2a(b-2s)}{a^2b^2} \\ \frac{2a(b-2s)}{a^2b^2} & \frac{4a^2}{a^2b^2} \end{pmatrix}$ and the characteristic polynomial of $\mathcal{E}$ is of the form

$$p_{\mathcal{E}}(\lambda) = \lambda^2 - \text{trace}\,(\mathcal{E}) \cdot \lambda + \det(\mathcal{E}).$$

The eigenvalues can now be computed by simply setting $p_{\mathcal{E}}(\lambda) = 0$. This equation has the solutions

$$\lambda_{1,2} = \frac{\text{trace}(\mathcal{E}) \pm \sqrt{\text{trace}(\mathcal{E})^2 - 4 \cdot \det(\mathcal{E})}}{2}. \tag{10.6}$$

Analogously to the previous proofs of Lemmas 10.6 and 10.7 we set

$$f(b,s) = \text{trace}(\mathcal{E})$$

and

$$g(b,s) = \sqrt{\text{trace}(\mathcal{E}) - 4 \cdot \det(\mathcal{E})}.$$

It is easy to see that Lemma 10.5 holds for functions in more than 1 variable for the partial derivatives of order 1, hence we can make use of it. As in the proof of Lemma 10.7 we get

$$\frac{g(b,s)}{f(b,s)} = \frac{\sqrt{a^4 - a^2(b+s)^2 - 3a^2s^2 + \left(b^2 - bs + s^2\right)^2}}{a^2 + b^2 - bs + s^2},$$

but now we have to take a look at the partial derivatives with respect to b and s.

$$\begin{aligned} \frac{\partial}{\partial b}\left(\frac{g(b,s)}{f(b,s)}\right) = &\frac{-3a^2b\left(a^2 - b^2 + s^2\right)}{\left(a^2 + b^2 - bs + s^2\right)^2} \\ &\cdot \frac{1}{\sqrt{a^4 - a^2(b+s)^2 + 3a^2s^2 + \left(b^2 - bs + s^2\right)^2}} \end{aligned} \tag{10.7}$$

$$\begin{aligned} \frac{\partial}{\partial s}\left(\frac{g(b,s)}{f(b,s)}\right) = &\frac{-3a^2b^2(b-2s)}{\left(a^2 + b^2 - bs + s^2\right)^2} \\ &\cdot \frac{1}{\sqrt{a^4 - a^2(b+s)^2 + 3a^2s^2 + \left(b^2 - bs + s^2\right)^2}}. \end{aligned} \tag{10.8}$$

We already know the zeros of the expression in (10.8) as they are the same as for (10.5) and so the only solution to the equation $\frac{\partial}{\partial s}\left(\frac{g(b,s)}{f(b,s)}\right) = 0$ is given

by $s = b/2$. The zeros of (10.7) lie on the hyperbola $b^2 - s^2 = a^2$. In order to have a local minimum (10.7) and (10.8) have to be 0 simultaneously. This means we find a local minimum at $\left(\frac{2a}{\sqrt{3}}, \frac{a}{\sqrt{3}}\right)$. By assumption $\lambda_1 = \lambda_1(b,s) \geq \lambda_2 = \lambda_2(b,s)$ and

$$\lambda(2a/\sqrt{3}, a/\sqrt{3}) = \frac{\lambda_1(2a/\sqrt{3}, a/\sqrt{3})}{\lambda_2(2a/\sqrt{3}, a/\sqrt{3})} = 1$$

and so this minimum has to be global. We also see that this solution is unique for $(b,s) \in \mathbb{R}^+ \times [0,b]$. □

Remark. As the axis ratio of the best fitting ellipse for the hexagonal lattice is 1 it is actually a sphere. By the above lemmas we also see that the hexagonal lattice is the only lattice which can achieve a circle as best packing or covering ellipse. However, this does not proof the optimality of the hexagonal lattice for the sphere packing problem.

After Lemmas 10.6, 10.7 and 10.8 we know that for a fixed rectangular lattice the condition number of the matrix describing the best fitting ellipse becomes better if we start shearing the lattice and becomes optimal for $s = b/2$. A shear $\widetilde{s} > b/2$ gives the same condition number as the shear $b - \widetilde{s}$. The optimal lattice would be given by the hexagonal lattice which we cannot achieve in the discrete case, as $a \cdot b$ is never rational and hence can only be achieved by an irrational signal length L which is not possible in the discrete case. However, the longer the signal length L is, the closer we can get to the hexagonal lattice, as we have more possibilities to choose a and b and can find a ratio closer to $\sqrt{3}/2$.

Example. For this example we will choose a signal length $L = 7200$ and redundancy $red = 1.5$. This gives us 3844 possible lattices. For each lattice we compute the frame condition number and the ratio of the axes of the packing ellipse. The MATLAB routines used for the computation can be found in the database section at www.nuhag.eu.

We have computed all admissible lattices of signal length $L = 7200$ with redundancy 1.5. For the comparison in Figure 15 the lattices which lead to frames with condition number greater than 100000 have been discarded. This leaves 3548 lattices out of 3844, so approximately 92.3% of all possible lattices. For each lattice we plot a point in the plane, where the abscissa coordinates are the solutions of the ellipse packing problem and the ordinate coordinates are the corresponding frame condition numbers. In the figure we put the ordinate to logarithmic scale. Let I be an index set and let Λ_i be an admissible lattice for $i \in I$. For $i \in I$ let q_i be the solution of the packing problem for lattice Λ_i and let Q_i be the frame condition for the

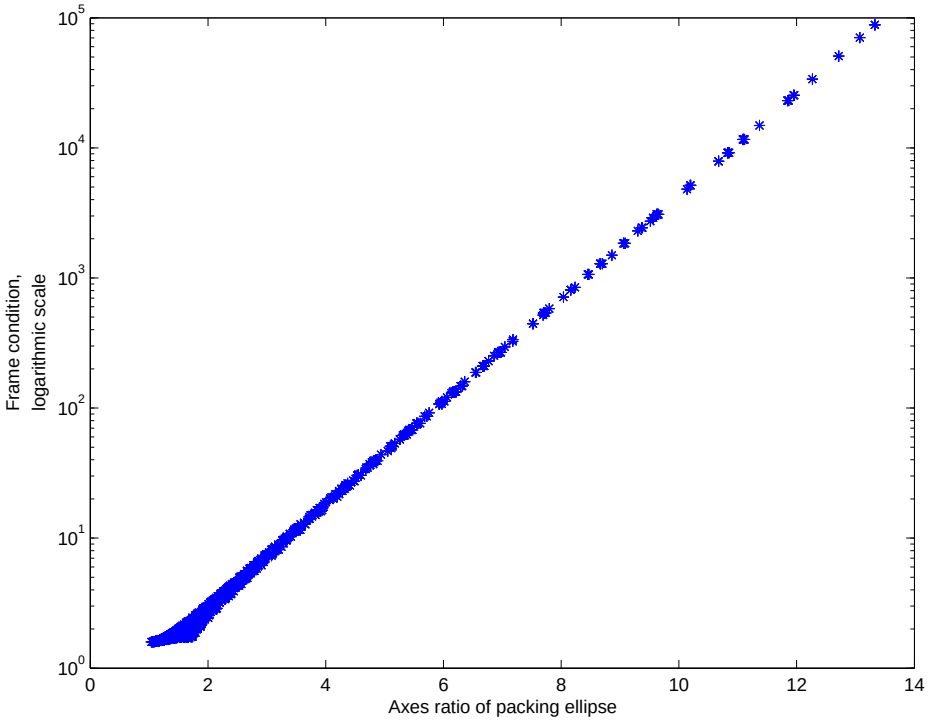


Figure 15: Comparison of the axes ratio of the packing ellipse on the abscissa and the frame condition number for admissible lattices of signal length 7200 with redundancy 1.5 and frame condition less than 100.

corresponding Gabor frame $\mathcal{G}_i = \{\Lambda_i, g_0\}$. Then Figure 15 suggests that our data points $(q_i, \log(Q_i))$ correlate in an almost linear way. Indeed the correlation coefficient is 0.9992 and a hypothesis test done with the MATLAB built-in routine suggests that the probability of getting a correlation as large as this at random is 0.

For Figures 16 and 17 we have only taken a look at lattices which give us a Gabor frame with condition number less than or equal to 10. Unfortunately, it seems that the behaviour of the frame condition for relatively small numbers cannot be described by axes ratio of the packing ellipse as easily as in the case for relatively large condition numbers. If we denote the frame condition by Q and the axes ratio of the packing ellipse by q, the example suggests that

$$\log(Q) = \mathcal{O}(q).$$

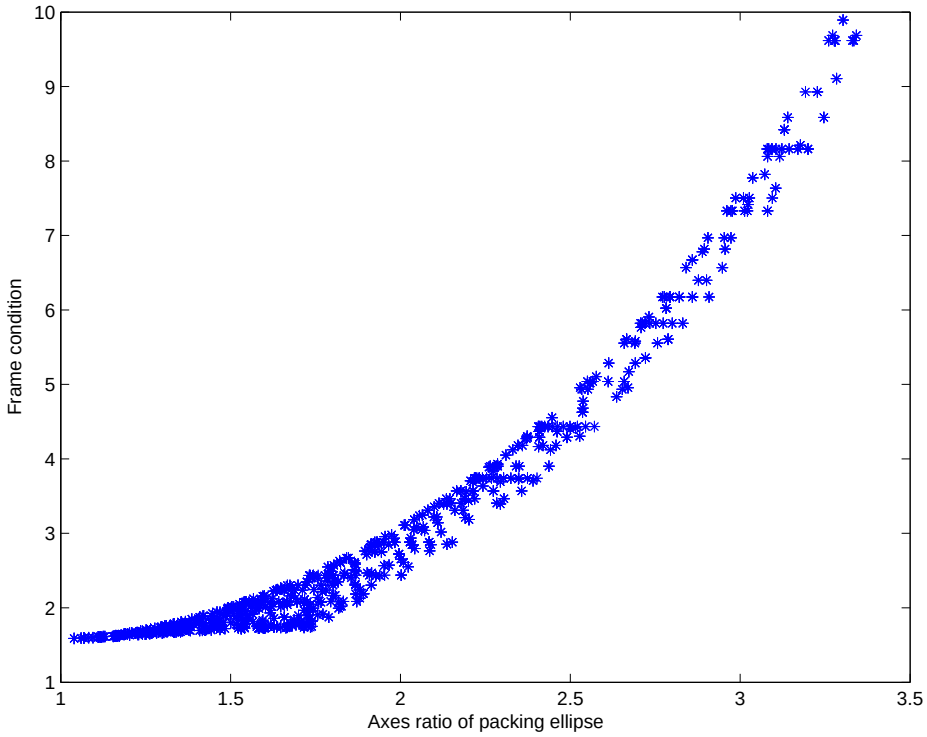


Figure 16: Comparison of the axes ratio of the packing ellipse on the abscissa and the frame condition number for admissible lattices of signal length 7200 with redundancy 1.5 and frame condition less than 10.

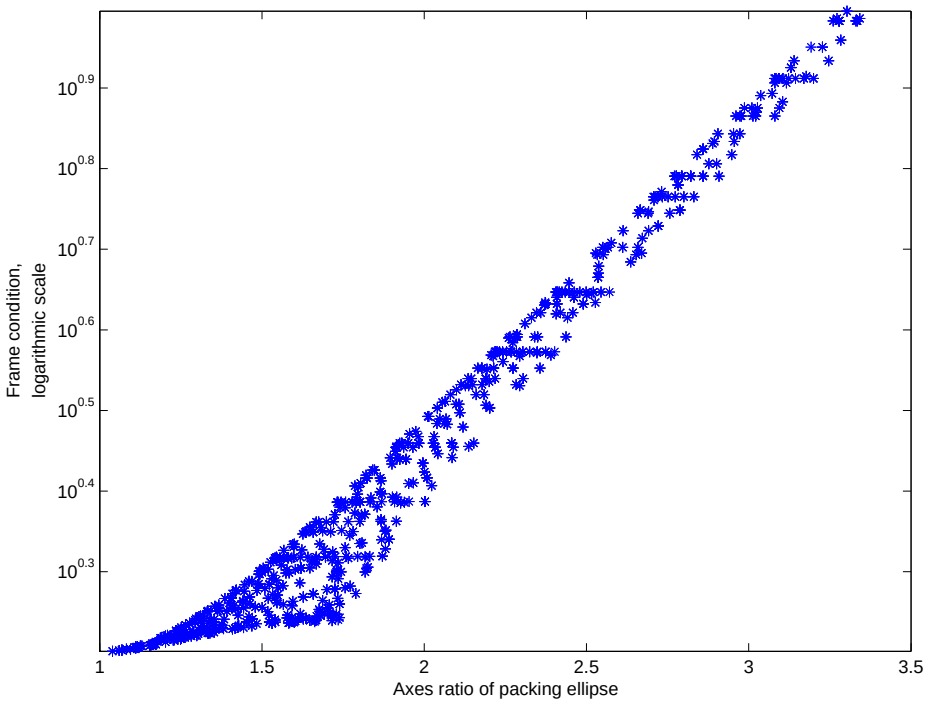


Figure 17: Comparison of the axes ratio of the packing ellipse on the abscissa and the frame condition number for admissible lattices of signal length 7200 with redundancy 1.5 and frame condition less than 10, logarithmic scale on ordinate.

Hamiltonian Mechanics

11 Classical Mechanics

The following part mainly relies on [2]. In this part, whenever we write x we imply that x depends on time denoted by t, hence $x = x(t)$. Also x is usually a vector in $\mathbb{R}^d$. Furthermore, we will make use of physicists notation $\dot{x} = \frac{d}{dt}x$ and $\ddot{\cdot}\, x = \frac{d^2}{dt^2}x$ in the following part.

In *classical* (or *Newtonian*) *mechanics* the *harmonic oscillator* in its simplest form is defined via

$$\ddot{x} = -x. \tag{11.1}$$

It describes the motion of a pendulum of mass $m = 1$ with only small oscillations or a mass fixed to a spring hanging loosely from a ceiling, moving gently up and down. More generally the equation for the harmonic oscillator is given by

$$m \cdot \ddot{x} = -k \cdot x \tag{11.2}$$

which is derived by using a different scale. The number of cases where classical mechanics can be applied reduces to *conservative systems*. These are systems where a *potential* $U(x)$ exists, such that

$$F = m \cdot \ddot{x} = -\frac{\partial U}{\partial x}. \tag{11.3}$$

Here F denotes the *force* acting on the particle x. The *kinetic energy* is defined as

$$T = \frac{1}{2}\langle \dot{x}, \dot{x} \rangle \tag{11.4}$$

and the *total energy* is defined as $E = U + T$. The most important role in classical mechanics is certainly the conservation of energy.

Theorem 11.1 (Conservation of energy)**.** *Let $x = x(t) \in \mathbb{R}^d$ and let the total energy of the system be described by*

$$E(x, \dot{x}, t) = T(\dot{x}) + U(x). \tag{11.5}$$

Then

$$\frac{d}{dt}E = 0 \tag{11.6}$$

Proof.

$$\begin{aligned}\frac{d}{dt}E(x,\dot{x},t) &= \frac{d}{dt}\left(T(\dot{x})+U(x)\right)\\ &= \frac{1}{2}\langle\dot{x},\ddot{x}\rangle + \frac{1}{2}\langle\ddot{x},\dot{x}\rangle + \langle\frac{\partial}{\partial t}U,\dot{x}\rangle\\ &= \langle\dot{x},\ddot{x}\rangle - \langle\dot{x},\ddot{x}\rangle\\ &= 0\end{aligned}$$

□

Another way of describing the motion of a particle in a mechanical system is by means of *Lagrangian mechanics*. The motion is then describe in the so-called *configuration space* or *phase space*, which is a differentiable manifold, on which its group of diffeomorphisms acts. A mechanical system in Lagrangian mechanics is then given by a manifold and the so-called *Lagrangian function*, a function on the tangent bundle of configuration space. The Newtonian mechanics follow as a special case of Lagrangian mechanics. The configuration space is then the Euclidean space and the Lagrangian function is given by $T - U$, the difference between kinetic and potential energy.

12 Calculus of Variations

Let γ be a curve in $\mathbb{R}^d$. By $\mathcal{C}^n([t_0,t_1],\mathbb{R}^d)$ we denote the space of of n times continuously differentiable curves in $\mathbb{R}^d$, starting in $\gamma(t_0)$ and ending in $\gamma(t_1)$. A mapping Φ from the space of curves into the field of real numbers is called a functional. An example of a such a functional is the length functional.

$$\begin{aligned}\Phi\colon \mathcal{C}^1([t_0,t_1],\mathbb{R}^d) &\to \mathbb{R}\\ \gamma &\mapsto \Phi(\gamma) = \int_{t_o}^{t_1} \|\dot{\gamma}(t)\|\,dt\end{aligned}$$

Let h be a curvein Rd. A functional Φ is called differentiable if

$$\Phi(\gamma+h) - \Phi(\gamma) = F(h) + R(h,\gamma),$$

where F depends linearly on h and $R(h,\gamma) = \mathcal{O}(h^2)$, meaning $|h| < \varepsilon$, $|\frac{d}{dt}h| < \varepsilon \Rightarrow |R| < C\,\varepsilon^2$. $F(h)$ is called the differential or variation of Φ and h is called the variation of γ. An extremal of a function $\Phi(\gamma)$ is a curve γ such that $F(h) = 0$ for all h. As a next step we want to derive a criterion under which a curve γ is an extremal of a functional Φ.

Theorem 12.1. *Let* $L = L(x, \dot{x}, t) \in \mathcal{C}^2(\mathbb{R}^{2d+1})$ *and* $\gamma = \{(t, x) : x = x(t),\, t_0 \le t \le t_1\}$. *If we define*

$$\Phi(\gamma) = \int_{t_0}^{t_1} L(x, \dot{x}, t)\, dt$$

then Φ *is differentiable and its derivative is given by*

$$F(h) = \int_{t_0}^{t_1} \left[\frac{\partial}{\partial x} L - \frac{d}{dt}\frac{\partial}{\partial \dot{x}} L\right] h\, dt + \left(\frac{\partial}{\partial \dot{x}} L\right) h \Big|_{t_0}^{t_1}.$$

Proof. By direct calculation we have

$$\Phi(\gamma + h) - \Phi(\gamma) = \int_{t_0}^{t_1} \Big[L(x + h, \dot{x} + \dot{h}, t) - L(x, \dot{x}, t)\Big]\, dt$$

$$= \underbrace{\int_{t_0}^{t_1} \left[\frac{\partial}{\partial x} L\, h - \frac{\partial}{\partial \dot{x}} L\, \dot{h}\right] dt}_{F(h)} + \underbrace{\mathcal{O}(h^2)}_{R(h,\gamma)}.$$

Integrating by parts leads to

$$F(h) = \int_{t_0}^{t_1} \left[\frac{\partial}{\partial x} L\, h - \frac{d}{dt}\frac{\partial}{\partial \dot{x}} L\, h\right] dt + \frac{\partial}{\partial \dot{x}} L\, h \Big|_{t_0}^{t_1}$$

and the proof is complete. □

The part $\left[\frac{\partial}{\partial x} L - \frac{d}{dt}\frac{\partial}{\partial \dot{x}} L\right]$ takes an important role in calculus of variations as we will see in the next theorem.

Theorem 12.2. *Let* $L = L(x, \dot{x}, t) \in \mathcal{C}^2(\mathbb{R}^{2d+1}$ *and* $\gamma = \{(t, x) : x = x(t),\, t_0 \le t \le t_1\}$ *with* $x(t_0) = x_0$ *and* $x(t_1) = x_1$ *fixed. Then* γ *is an extremal of* $\Phi(\gamma) = \int_{t_0}^{t_1} L(x, \dot{x}, t)\, dt$ *if and only if*

$$\left[\frac{\partial}{\partial x} L - \frac{d}{dt}\frac{\partial}{\partial \dot{x}} L\right] = 0. \tag{12.1}$$

Equation (12.1) *is called Euler-Lagrange equation of the functional* $\Phi(\gamma)$.

Proof. We have to solve $F(h) = 0$ for all possible h that leave the endpoints of the curve γ fixed. This means we have to solve

$$F(h) = \int_{t_0}^{t_1} \left[\frac{\partial}{\partial x} L - \frac{d}{dt}\frac{\partial}{\partial \dot{x}} L\right] h\, dt + \left(\frac{\partial}{\partial \dot{x}} L\right) h \Big|_{t_0}^{t_1} = 0.$$

The part outside the integral is 0 as x_0 and x_1 are fixed, hence $h(t_0) = h(t_1) = 0$. As the integral must vanish for all possible h the the Euler-Lagrangian equation (12.1) must be fulfilled. □

13 The Legendre Transform

We have introduced Newtonian as well as Lagrangian mechanics and have mentioned that the former one is a special case of the latter one. Besides these two formalisms we want to introduce a third one, which for classical mechanics again predicts the same outcome as the former formalisms. In order to come up with this formalism, which is called Hamiltonian mechanics, we first have to introduce the Legendre transform of a convex function f.

Definition 13.1. Let $f : I \to \mathbb{R}$ be a (strictly) convex, differentiable function, where I denotes an interval. Let p be a real number and $y(x) = p \cdot x$ the straight line with slope p. We introduce the new function

$$F(x, p) := y(x) - f(x) = p \cdot x - f(x),$$

where $x = x(p)$. Then the function $g(p) = F(x(p), p)$ with the additional condition that $\frac{\partial}{\partial x} F = 0$ is called the Legendre transform of $f(x(p))$.

The Legendre transform describes a function not via pairs of points, but via the slope and interception values of its tangent lines.

Example. Let $f(x) = \frac{x^\alpha}{\alpha}$ and $F(x, p) = p \cdot x - f(x)$, where $x = x(p)$. In order to have $\frac{\partial}{\partial x} F = 0$ it is necessary that $\frac{\partial}{\partial x} f \equiv p$. This means that $x^{\alpha-1} = p$ which implies that $x(p) = p^{\frac{1}{\alpha-1}}$. Then the Legendre transform of f is given by

$$\begin{aligned} g(p) &= p \cdot p^{\frac{1}{\alpha-1}} - \frac{p^{\frac{\alpha}{\alpha-1}}}{\alpha} \\ &= p^{\frac{\alpha-1+1}{\alpha-1}} - \frac{p^{\frac{\alpha}{\alpha-1}}}{\alpha} \\ &= p^{\frac{\alpha}{\alpha-1}}(1 - \frac{1}{\alpha}) \\ &= p^{\frac{\alpha}{\alpha-1}}(\frac{\alpha-1}{\alpha}) \\ &= \frac{p^\beta}{\beta} \end{aligned}$$

where $\beta = \frac{\alpha}{\alpha-1}$ or equivalently $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

Remark. This examples illustrates that the Legendre transform is involutive (which we will not prove), meaning that if g is the Legendre transform of f, then f is the Legendre transform of g. Any two functions which are Legendre transforms of one another are called dual in the sense of Young.

14 Hamilton's Equations

The Lagrange function if given by $L = T - U$ where U is a potential as defined in (11.3) and T is the kinetic energy as defined in (11.4). We want to describe the evolution of system of n mass points by means of Lagrangian mechanics. We introduce the *generalised coordinates* $q = (q1, \ldots, q_n)$ in configuration space and have $\dot{q} = (\dot{q}_1, \ldots, \dot{q}_n)$ as the *generalised velocities*. The functional $\int_{t_0}^{t_1} L(q, \dot{q}, t)\, dt$ is called the *action* and further on the *generalised momenta* are given by $p = \frac{\partial}{\partial \dot{q}} L$ and the *generalised forces* are given by $\frac{\partial}{\partial q} L$.

Theorem 14.1. *The motions in a conservative field of the mechanical system described in* (11.3) *coincide with the extrema of the functional*

$$\Phi(\gamma) = \int_{t_0}^{t_1} L(x, \dot{x}, t)\, dt,$$

where $L = T - U$ is the Lagrange function and $T = \frac{1}{2} m \langle \dot{x}, \dot{x} \rangle$ and $U = U(x)$ fulfils (11.3).

Proof. By Theorem 12.2 we know that extrema of the functional Φ have to fulfil the Euler-Lagrange equation (12.1). The Euler-Lagrangian equation takes the following form

$$\begin{aligned} & \frac{d}{dt} \frac{\partial}{\partial \dot{x}} \left[\frac{1}{2} m \langle \dot{x}, \dot{x} \rangle - U(x) \right] - \frac{\partial}{\partial x} \left[\frac{1}{2} m \langle \dot{x}, \dot{x} \rangle - U(x) \right] = 0 \\ \Leftrightarrow \quad & \frac{d}{dt} m \dot{x} - \left(-\frac{\partial}{\partial x} U(x) \right) = 0 \\ \Leftrightarrow \quad & m \ddot{x} = -\frac{\partial}{\partial x} U(x). \end{aligned}$$

□

In the above theorem we have used classical coordinates in Euclidean space. If we use the generalised coordinates and velocities then the Euler-Lagrange equation has the following form

$$\frac{d}{dt} \frac{\partial}{\partial \dot{q}} L - \frac{\partial}{\partial q} L = 0$$

or using generalised momenta and forces we have

$$\frac{d}{dt} p - \dot{p} = 0.$$

The Lagrange equations are

$$\begin{aligned} \dot{p} =& \frac{\partial}{\partial q} L \\ p =& \frac{\partial}{\partial \dot{q}} L, \end{aligned} \tag{14.1}$$

where

$$\begin{aligned} &L \colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R} \\ &L = L(q, \dot{q}, t) \end{aligned}$$

is the Lagrange function.

We have now prepared all necessary tools in order to introduce Hamilton's formalism. Hamilton's equations also find applications in quantum mechanics and convert the system (14.1) of Lagrange's equations into a symmetric or more precisely an antisymmetric form.

Theorem 14.2. *Let the Lagrange function $L = L(q, \dot{q}, t)$ be convex with respect to $\dot{q}$. The system* (14.1) *is equivalent to the system of Hamilton's equations*

$$\begin{aligned} \dot{p} =& -\frac{\partial}{\partial q} H \\ \dot{q} =& \quad \frac{\partial}{\partial p} H \end{aligned} \tag{14.2}$$

with $H = H(q, p, t) = \langle p, \dot{q} \rangle - L(q, \dot{q}, t)$ being the Legendre transform of L. The function H is called Hamiltonian or Hamilton's function.

Proof. See [2]. □

15 Liouville's Theorem

Let $(q, p) = (q_1, \ldots, q_n, p_1, \ldots p_n)$ be generalised coordinates in phase space. The *phase flow* is the one-parameter group

$$\varphi^t \colon (q(0), p(0)) \mapsto (q(t), p(t))$$

of transformations on the phase space with the following properties

$$\varphi^0 = Id \tag{15.1}$$

$$\varphi^{t+s} = \varphi^t \circ \varphi^s \tag{15.2}$$

$$(\varphi^t)^{-1} = \varphi^{-t}, \tag{15.3}$$

where $(q(0), p(0)), (q(t), p(t))$ are solutions of Hamilton's system of equations (14.2).

In the case of Hamiltonian mechanics we call the phase flow a *Hamiltonian flow*. This flow has the property of preserving volume when applied on a set in phase space.

Theorem 15.1. *For the vector field $\dot{x} = f(x)$ with divergence $div(f) = 0$ and a set Ω with volume $vol(\Omega)$ and phase flow φ^t we have*

$$vol(\varphi^t(\Omega)) = vol(\Omega).$$

The divergence of the vector field is $div(f) = \sum_{i=1}^{n} \frac{\partial}{\partial x_1} f_i$.

Proof. We look at the Taylor expansion of the flow (around 0).

$$\varphi^t(x) = x + f(x) \cdot t + \mathcal{O}(t^2)$$

The volume of the set Ω under the phase flow is given by

$$vol(\varphi^t(\Omega)) = \int_\Omega \det\left(\frac{\partial}{\partial x}\varphi^t(x)\right) dx.$$

For any matrix it is true that $\det(I - t \cdot A) = 1 + trace(A) \cdot t + \mathcal{O}(t^2)$ for $t \to 0$.

$$\begin{aligned} \Rightarrow vol(\varphi^t(\Omega)) &= \int_\Omega \left(\det\left(I + \frac{\partial}{\partial x} f(x) \cdot t\right) + \mathcal{O}(t^2)\right) dx \\ &= \int_\Omega \left(1 + trace\left(\frac{\partial}{\partial x} f(x)\right) \cdot t + \mathcal{O}(t^2)\right) dx \end{aligned}$$

Therefore we get

$$\frac{d}{dt} vol(\varphi^t(\Omega)) = \int_\Omega div(f)\, dx = 0$$

and hence,for divergence free vector fields the volume is preserved under the flow. □

Theorem 15.2 (Liouville)**.** *The Hamiltonian flow is volume preserving.*

Proof. Hamilton's equation in the compact form reads

$$\begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = f = \begin{pmatrix} -\frac{\partial}{\partial q} H \\ \frac{\partial}{\partial p} H \end{pmatrix}$$

The divergence of f is given by

$$\begin{aligned} div(f) &= \sum_{i=1}^{n} \frac{\partial}{\partial p_i} f_i + \sum_{i=1}^{n} \frac{\partial}{\partial q_i} f_{i+n} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial p_i} \left(-\frac{\partial}{\partial q_i} H \right) + \sum_{i=1}^{n} \frac{\partial}{\partial q_i} \left(\frac{\partial}{\partial p_i} H \right) \\ &= 0. \end{aligned}$$

Hence, volume is preserved under a Hamiltonian flow by Theorem 15.1. □

We want to state two more properties about Hamilton's function.

Theorem 15.3. *Let $L = T - U$ be the Lagrangian function and let T be a quadratic form, i.e. $T = \frac{1}{2} \sum_{i=1}^{n} a_{ij} \dot{q}_i \dot{q}_j$, with $a_{ij} = a_{ij}(q,t)$ and $U = U(q)$. Then the Hamiltonian gives the full energy, i.e. $H = T + U$.*

Proof. The proof uses Euler's Lemma on homogeneous functions of order α and the fact that the values of a quadratic form $f(x)$ and its Legendre transform $g(p)$ coincide at corresponding points, i.e. $f(x(p)) = g(p)$. For a full proof see [2]. □

Corollary 15.4. *If H does not explicitly depend on t, then H is constant for all t, or in other words, energy is conserved over time.*

Proof.

$$\frac{d}{dt} H = \frac{\partial}{\partial p} H \cdot \dot{p} + \frac{\partial}{\partial q} H \cdot \dot{q} + \frac{\partial}{\partial t} H = \dot{q}\dot{p} - \dot{p}\dot{q} + \frac{\partial}{\partial t} H = \frac{\partial}{\partial t} H.$$

□

Hamiltonian Deformation of Lattices

16 The Harmonic Oscillator and its Hamiltonian Flow

In section 11 we already introduced the harmonic oscillator via equation (11.2) which was $m \cdot \ddot{x} = -k \cdot x$. We have done all necessary preparations in order to write the equations for the harmonic oscillator in the sense of Hamiltonian mechanics. The Hamiltonian H is given by

$$H(q,p,t) = \frac{p^2}{2\,m} + \frac{m\,\omega^2 q^2}{2} \tag{16.1}$$

and Hamilton's equations read as follows

$$\begin{aligned} \dot{p} &= -\frac{\partial}{\partial q} H = -m\omega^2 q \\ \dot{q} &= \frac{\partial}{\partial p} H = \frac{p}{m} \end{aligned} \tag{16.2}$$

For simplification we assume $m = 1$. From Theorem 15.3 and Corollary 15.4 we already know that the Hamiltonian gives the full energy and that in the case of the harmonic oscillator the energy does not change over time, as t does not enter explicitly in the Hamiltonian. Another way of writing equation (16.1) is the following

$$\frac{q^2}{a^2} + \frac{p^2}{b^2} = 1, \tag{16.3}$$

where $a^2 = \frac{2H}{\omega^2}$ and $b^2 = 2H$, with H being interpreted as the total energy of the system. It is easy to see that equation (16.3) describes an ellipse in phase space with axis ratio $a/b = 1/\omega$. A flow is given by

$$\begin{aligned} &\varphi : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2 \\ &\varphi(q,p,t) = \varphi^t(q,p) = \begin{pmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{pmatrix} \cdot \begin{pmatrix} q \\ p \end{pmatrix} \end{aligned} \tag{16.4}$$

We start with the initial solution $(q(0), p(0))^T = (q_0, p_0)^T = (0, b)^T$. It is clear that after a period of $T = 2\pi$ we are back at our initial solution as the trajectory is the closed ellipse defined by (16.3). We split the period

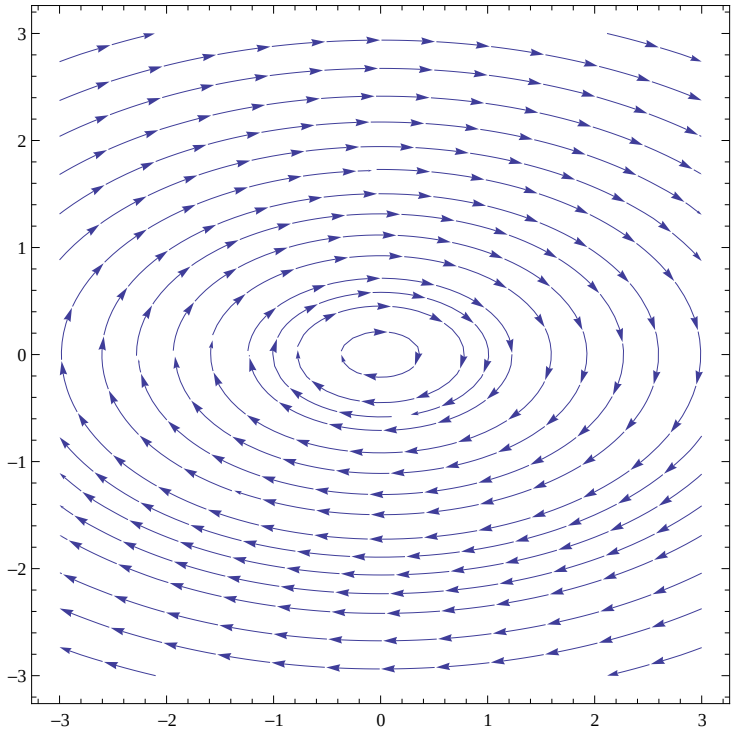


Figure 18: The flow of the harmonic oscillator

into 6 equally distributed parts and look at the solutions. The next solution after $t = \pi/3$ is then given by $\varphi^{\pi/3}(q_0, p_0)^T = (q_{\pi/3}, p_{\pi/3})^T = (\frac{\sqrt{3}}{2}a, \frac{1}{2}b)^T$. By defining

$$M := \left(\begin{array}{cc} \varphi^{\pi/3}(q_0, p_0), & \varphi^0(q_0, p_0) \end{array} \right) = \left(\begin{array}{cc} \frac{\sqrt{3}}{2}a & 0 \\ \frac{1}{2}b & b \end{array} \right)$$

we have a matrix defining a lattice. As we want our lattice to have volume 1, we have the obstruction that

$$a \cdot b = \frac{2}{\sqrt{3}}.$$

It is easy to verify that $\varphi^{k \cdot \frac{\pi}{3}}(p, q)$ are lattice points of the lattice $M\mathbb{Z}^2$ for all $k \in \mathbb{Z}$. For the special case $a = b = 1$ we get the hexagonal lattice generated by the matrix

$$Hex = \left(\begin{array}{cc} \frac{\sqrt{3}}{2} & 0 \\ \frac{1}{2} & 1 \end{array} \right)$$

with volume $\sqrt{3}/2$. We define

$$\mathcal{E} = \left(Hex \cdot M^{-1}\right)^T \cdot \left(Hex \cdot M^{-1}\right)$$

which simplifies to

$$\mathcal{E} = \begin{pmatrix} \frac{3}{4}b^2 & 0 \\ 0 & \frac{3}{4}a^2 \end{pmatrix}.$$

By the obstruction on $a \cdot b = \frac{2}{\sqrt{3}}$ we get

$$\mathcal{E} = \begin{pmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{pmatrix}$$

and this gives the quadratic form which defines the ellipse

$$(q,p)\,\mathcal{E}\,(q,p)^T = 1$$

or

$$\frac{q^2}{a^2} + \frac{p^2}{b^2} = 1.$$

By construction we know that the six lattice points closest to the origin lie on this ellipse. By scaling this ellipse properly we get a packing for $M \cdot \mathbb{Z}^2$. The packing ellipses have to meet half way between the lattice points, which means that we have to scale the vectors of the generating matrix by $1/2$. This means that the area of the ellipse shrinks down to $1/4$ of the original size. By substituting 1 by $1/4$ in the last equation we get the desired result. Alternatively we can scale the left hand side by 4, which is the same as multiplying $\mathcal{E}$ with a factor of 4.

$$\mathcal{P} = 4 \cdot \mathcal{E} = \left(2 \cdot Hex \cdot M^{-1}\right)^T \cdot \left(2 \cdot Hex \cdot M^{-1}\right) = {M^{-1}}^T \cdot \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \cdot M^{-1}.$$

By Theorem 2.1 this is the matrix used for the construction of the packing ellipse for the lattice $M\mathbb{Z}^2$. This proofs that the above construction leads to a packing.

Next we investigate the set $\Omega_0 = \{(q,p) \in \mathbb{R}^2 \mid \frac{4q^2}{a^2} + \frac{4p^2}{b^2} \leq 1\}$ centred at the origin and its translates Ω_{ξ_t,ν_t} centred at $(\xi_t, \nu_t) = \varphi^t(q_0, p_0)$ for $t = k \cdot \frac{2\pi}{6}$, $k = 0, \ldots, 5$. By Theorem 15.2 we know that volume is preserved under a Hamiltonian flow. As this holds true for the ellipses as well as for each fundamental domain of the lattice, we know that the density of the sets Ω_{ξ_t,ν_t} does not change under the flow defined by (16.4). Also we know from the theory of ordinary differential equations that trajectories cannot cross under the given circumstances. This means that distinct points of distinct sets stay distinct under the flow. We use the flow properties (15.1), (15.2) and (15.3) in order to show that the ellipses keep their shape. We set

$$M_t = \begin{pmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{pmatrix}$$

and can then write the flow as $\varphi^t(q,p) = M_t \cdot \begin{pmatrix} q \\ p \end{pmatrix} = \widetilde{M}_t$. We can write $M = \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)$ and $\mathcal{E} = \left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1}\right)^T \left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1}\right)$ and hence, for $m = 1$, equations (16.1) and (16.3) are equivalent to the following equation

$$\begin{aligned} &\left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right)^T \cdot \\ &\left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right) = 1. \end{aligned} \tag{16.5}$$

The same holds true for

$$\begin{aligned} &\left(Hex \cdot \left(\widetilde{M}_{\pi/3+s}, \widetilde{M}_{0+s}\right)^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right)^T \cdot \\ &\left(Hex \cdot \left(\widetilde{M}_{\pi/3+s}, \widetilde{M}_{0+s}\right)^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right) = 1, \end{aligned} \tag{16.6}$$

as $\left(\widetilde{M}_{\pi/3+s}, \widetilde{M}_{0+s}\right)^{-1} \begin{pmatrix} q \\ p \end{pmatrix} = \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1} M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}$ and $M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}$ is a solution to equations (16.1) and (16.3) by the definition of the flow. Therefore, we see that equation (16.6) is equivalent to

$$\begin{aligned} &\left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1} M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right)^T \cdot \\ &\left(Hex \cdot \left(\widetilde{M}_{\pi/3}, \widetilde{M}_0\right)^{-1} M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right) = 1, \end{aligned} \tag{16.7}$$

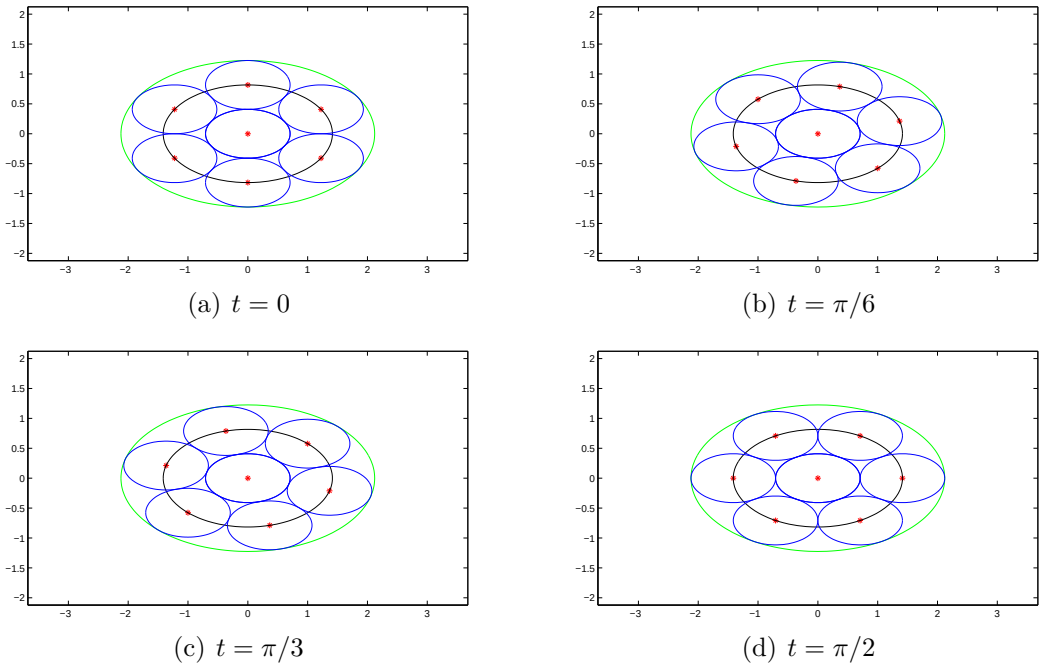
or written in compact form

$$\left(M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right)^T \cdot \mathcal{E} \cdot \left(M_s^{-1} \begin{pmatrix} q \\ p \end{pmatrix}\right) = 1. \tag{16.8}$$

Hence, for $m = 1$, we have an equivalence between equations (16.1) and (16.8), which proves that the packing ellipses keep their shape under the given circumstances.

Figure 19 illustrates the process of the Hamiltonian deformation of a lattice. We clearly see that the packing ellipses are preserved and that any arrangement is optimal in the sense of the packing density.

Summing up the above results we get the following theorem.



(a) $t = 0$ (b) $t = \pi/6$

(c) $t = \pi/3$ (d) $t = \pi/2$

Figure 19: Hamiltonian deformation of a lattice via a Hamiltonian flow. The red marks are lattice points, the blue ellipses give a packing. The black ellipse is a flow line for the harmonic oscillator from which the blue ellipses are derived by the scaling factor 1/2. The green ellipse is a flow line for a different energy level of the harmonic oscillator, derived by the scaling the black ellipse by the factor 3/2.

Theorem 16.1. *For any given ellipse there are uncountably many possible arrangements of translated copies of this ellipse leading to an optimal lattice packing. Hence, there are uncountably many lattices which can be optimally packed with the given ellipse and translated copies of it and these lattices can be derived from a single lattice.*

Proof. W.l.o.g. we may assume that the axes of our ellipse coincide with the Cartesian axes. We denote the length of the first axis by $\widetilde{a}$ and the length of the second axis by $\widetilde{b}$ and the ratio $\frac{\widetilde{a}}{\widetilde{b}} = \frac{1}{\omega}$ and the coordinate are denoted by (q, p). Then the equation describing the ellipse is given by $\frac{q^2}{\widetilde{a}^2} + \frac{p^2}{\widetilde{b}^2} = 1$. Setting $a = 2\widetilde{a} = \sqrt{2H}/\omega$ and $b = 2\widetilde{b} = \sqrt{2H}$ we get back to equation (16.1) for $m = 1$. The rest follows by the above given arguments of section 16. □

17 The Inverted Harmonic Oscillator

In section 11 we described the harmonic oscillator via the equation $m\ddot{x} = -kx$ (11.2), but did not specify the range of k. We implied that k is positive from equation (11.1) and we mentioned that k is a scale factor. However, in equations (16.1) and (16.2) we replaced k by ω^2 and everything was clear, but still, if we assume that k could take negative values then the scale would be inverted and we get the equations for the *inverted harmonic oscillator*. Speaking in terms of Hamiltonian mechanics the frequency ω is replaced by $i\omega$ and then the Hamiltonian H is given by

$$H(q,p,t) = \frac{p^2}{2m} - \frac{m\omega^2 q^2}{2} \tag{17.1}$$

and Hamilton's equations read as follows

$$\begin{aligned} \dot{p} &= -\frac{\partial}{\partial q} H = m\omega^2 q \\ \dot{q} &= \frac{\partial}{\partial p} H = \frac{p}{m} \end{aligned} \tag{17.2}$$

For a more precise physical description of the inverted harmonic oscillator and further applications see [3], [4] and [33]. We are now interested in the flow of the inverted harmonic oscillator. For simplicity reasons we set $m = 1$. The flow is then given by

$$\begin{aligned} &\varphi : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2 \\ &\varphi(q,p,t) = \varphi^t(q,p) = \begin{pmatrix} \cosh(\omega t) & \frac{1}{\omega}\sinh(\omega t) \\ \omega \sinh(\omega t) & \cosh(\omega t) \end{pmatrix} \cdot \begin{pmatrix} q \\ p \end{pmatrix}. \end{aligned} \tag{17.3}$$

Instead of ellipses the trajectories are now hyperbolas. For the special case $\omega = 1$ the flow is given by

$$\varphi(q,p,t) = \varphi^t(q,p) = \begin{pmatrix} \cosh(t) & \sinh(t) \\ \sinh(t) & \cosh(t) \end{pmatrix} \cdot \begin{pmatrix} q \\ p \end{pmatrix}. \tag{17.4}$$

We are now interested in the matrix given in equation (17.4). If we rewrite the matrix using exponential functions instead of hyperbolic functions we have $\begin{pmatrix} \frac{e^t+e^{-t}}{2} & \frac{e^t-e^{-t}}{2} \\ \frac{e^t-e^{-t}}{2} & \frac{e^t+e^{-t}}{2} \end{pmatrix}$ and by setting $\sqrt{\delta} = e^t$ this gives us the distortion matrix

$$\mathcal{M}_{\delta,k} = \frac{1}{2\sqrt{k}} \begin{pmatrix} \sqrt{\delta} + \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} - \frac{1}{\sqrt{\delta}}) \\ \sqrt{\delta} - \frac{1}{\sqrt{\delta}} & k(\sqrt{\delta} + \frac{1}{\sqrt{\delta}}) \end{pmatrix}.$$

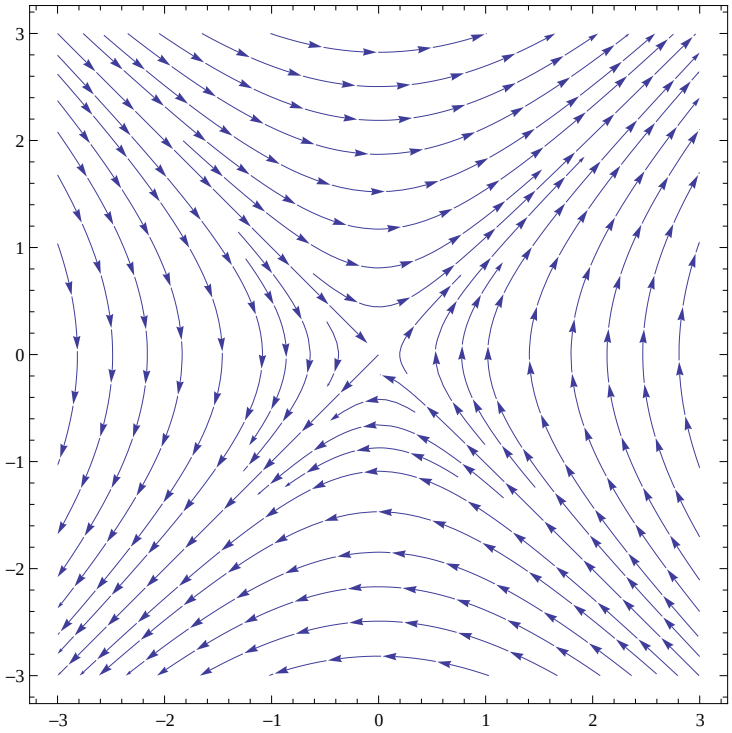


Figure 20: The flow of the inverted harmonic oscillator with parameters $m = 1$ and $\omega = 1$.

as defined in (5.3) with $k = 1$. This shows the equivalence of the distortion method presented in [8] and the Hamiltonian deformation of a lattice via the flow of the inverted harmonic oscillator for $m = 1$ and $\omega = 1$.

References

[1] Luis Daniel Abreu and Monika Dörfler. An inverse problem for localization operators. *Inverse Problems*, 28, 2012.

[2] V.I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 1989.

[3] R. K. Bhaduri, A. Khare, and J. Law. The Phase of the Riemann Zeta Function and the Inverted Harmonic Oscillator. *Phys.Rev. E52 (1995) 486*, June 1994.

[4] R. K. Bhaduri, A. Khare, and J. Law. The Riemann Zeta Function and the Inverted Harmonic Oscillator. *Annals of Physics*, 254(1), February 1997.

[5] D. Catelin, B. Le Floch, and R. Halbert Lassalle. Digital sound broadcasting to mobile receivers. *IEEE Trans. Consumer Electron.*, 73:30–34, 1989.

[6] Ole Christensen. *Frames and Bases. An Introductory Course.* Applied and Numerical Harmonic Analysis. Basel Birkhäuser, 2008.

[7] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups*, volume 290. Springer, third edition, 1999.

[8] Herbert Edelsbrunner and Michael Kerber. Covering and packing with spheres by diagonal distortion in $\mathbb{R}^n$. In *Rainbow of computer science*, volume 6570 of *Lecture Notes in Comput. Sci.*, pages 20–35. Springer, Heidelberg, 2011.

[9] Hans G. Feichtinger. On a new Segal algebra. *Monatsh. Math.*, 92:269–289, 1981.

[10] Hans G. Feichtinger. Modulation spaces on locally compact Abelian groups. Technical report, January 1983.

[11] Hans G. Feichtinger and Thomas Strohmer. *Advances in Gabor Analysis*. Birkhäuser, Basel, 2003.

[12] D. Gabor. Theory of communication. *J. IEE*, 93(26):429–457, 1946.

[13] Georg-Johann. Singular-value-decomposition.svg. Wikimedia Commons, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license., URL: http://creativecommons.org/licenses/by-sa/3.0/deed.en.

[14] C. Giacovazzo, H. L. Monaco, G. Artioli, D. Viterbo, M. Milanesio, G. Ferraris, G. Gilli, P. Gilli, G. Zanotti, and M. Catti. *Fundamentals of Crystallography*. IUCr Texts on Crystallography, 15. Oxford University Press, Great Clarendon Street, Oxford, UK, 2011.

[15] L. Grafakos. *Classical Fourier Analysis (Second Edition)*. Springer, 2008.

[16] Karlheinz Gröchenig. Acceleration of the frame algorithm. *IEEE Trans. SSP*, 41/12:3331–3340, 1993.

[17] Karlheinz Gröchenig. *Foundations of Time-Frequency Analysis*. Appl. Numer. Harmon. Anal. Birkhäuser Boston, Boston, MA, 2001.

[18] Wolfgang Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme. (Iterative Solution of Large Sparse Systems of Equations)*. 1991.

[19] Thomas Hales. A proof of the Kepler conjecture. *Annals of Mathematics*, 162, 2005.

[20] David Hilbert. Mathematische Probleme. *Archiv der Mathematik und Physik*, Reihe 3, Band 1, 1901.

[21] Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge University Press, 2003.

[22] A.K. Lenstra, H.W.jun. Lenstra, and László Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515 – 534, 1982.

[23] A.S. Macedo and E.S. Sousa. Coded OFDM for broadcast indoor wireless systems. *Proc. IEEE Int. Conf. Communications (ICC)*, 2:934–938, 1997.

[24] Kurt Meyberg and Peter Vachenauer. *Hoehere Mathematik 2. Differentialgleichungen, Funktionentheorie, Fourier-Analysis, Variationsrechnung. (Higher mathematics 2. Differential equations, function theory, Fourier analysis, calculus of variations). 4., korrig. Aufl.* Springer, Berlin, 2001.

[25] K. Rao, D.N. Kim, and J.J. Hwang. *Fast Fourier Transform - Algorithms and Applications: Algorithms and Applications*. Signals and communication technology. Springer, 2011.

[26] Robert Schaback and Holger Wendland. *Numerical Mathematics. (Numerische Mathematik) 5th Completely new Revised ed.* Springer-Verlag, 2004.

[27] Peter L. Søndergaard, Bruno Torrésani, and Peter Balazs. The Linear Time Frequency Analysis Toolbox. *International Journal of Wavelets, Multiresolution Analysis and Information Processing*, 10(4), 2012.

[28] Thomas Strohmer and Scott Beaver. Optimal OFDM system design for time-frequency dispersive channels. *IEEE Trans. Comm.*, 51(7):1111–1122, July 2003.

[29] L.F. Tóth. Packungs- und Deckungswirtschaftlichkeit einer Scheibenfolge. In *Lagerungen in der Ebene auf der Kugel und im Raum*, volume 65 of *Die Grundlehren der mathematischen Wissenschaften*, pages 99–113. Springer Berlin Heidelberg, 1972.

[30] F. Vallentin. *Sphere Coverings, Lattices and Tilings (in low dimension).* PhD thesis, 2003.

[31] Dirk Werner. *Funktionalanalysis 7., korrigierte und erweiterte Auflage.* Springer-Verlag, Berlin Heidelberg, 2011.

[32] Christoph Wiesmeyr, Nicki Holighaus, and Peter L. Søndergaard. Efficient Algorithms for Discrete Gabor Transform on a Nonseparable Lattice. *IEEE Transaction on Signal Processing*, 2013.

[33] C. Yuce, A. Kilic, and A. Coruh. Inverted Oscillator. *Phys. Scr. 74 114 (2006)*, March 2006.

[34] Chuanming Zong. Simultaneous Packing and Covering in the Euclidean Plane. *Monatshefte für Mathematik*, 134(3):247–255, 2002.

[35] William Y. Zou and Yiyan Wu. COFDM: An overview. *IEEE Transactions on Broadcasting*, 41(1):1–8, Mar 1995.

Deutsche Zusammenfassung

In der zu Grunde liegenden Arbeit sollen vermutete Verbindungen zwischen Gabor Frames und geometrischen Eigenschaften von Gittern aufgezeigt werden.

Das Konzept der Gabor Frames ist eine spezielle Zeit-Frequenz Darstellungsmethode und unterliegt als solche Unschärferelationen. Dies bedeutet, dass das Produkt aus Signallänge und Bandbreite nicht beliebig klein werden kann. Das Kernstück der Gabor Darstellungsmethode ist die Kurzzeit-Fourier-Transformation, welche die Fourier-Transformation als Spezialfall abdeckt. Die zu untersuchenden Signale werden daher für gewöhnlich endliche Energie haben, sind also Elemente des Hilbertraums $L^2(\mathbb{R}^d)$. Durch die Verwendung von Banach-Gelfand Tripeln ist es allerdings möglich den Definitionsbereich der Kurzzeit-Fourier-Transformation zu erweitern.

Die klassische Fourier-Transformation bietet keine Information darüber, zu welchen Zeitpunkten welche Frequenzen in einem Signal vorkommen. Die Kurzzeit-Fourier-Transformation versucht diese Problem zu lösen, indem sogenannte Fensterfunktionen verwendet werden. Auf Grund der erwähnten Unschärferelationen führt gute Konzentration im Zeitbereich, zu einer schlechteren Konzentration im Frequenzbereich. Die kanonische Wahl der Fensterfunktion fällt daher auf eine Gauss Funktion, da diese allein die Unschärferelation minimiert.

Die Kurzzeit-Fourier-Transformation zu einer gegebenen Fensterfunktion liefert eine stetige Zeit-Frequenz Darstellungsmethode, nachdem $L^2(\mathbb{R}^d)$ aber ein separabler Hilbertraum ist, sollte es eine diskrete Methode geben um ein Signal darzustellen. Darum geht es beim Konzept der Gabor Frames. Die Idee ist es ein Erzeugendensystem für $L^2(\mathbb{R}^d)$ zu finden, welches aus modulierten Translaten der Fensterfunktion besteht. Die Auswahl der Translate und Modulationen generiert ein Muster in der sogenannten Zeit-Frequenz Ebene $\mathbb{R}^d \times \mathbb{R}^d$, ein Konzept ähnlich dem des Phasenraums aus dem Gebiet der gewöhnlichen Differentialgleichungen. Das Muster wird für gewöhnlich in Form eines Gitters gewählt, welches durch eine $2d \times 2d$ Matrix dargestellt werden kann.

Es ist ungeklärt, ob gute geometrische Eigenschaften des verwendeten Gitters zu guten Frames Eigenschaften führen, wie beispielsweise stabiler und schneller Rekonstruktion des Signals aus Koeffizienten, welche an den Gitterpunkten gemessen wurden. Das Problem ist bis heute noch nicht einmal für ein eindimensionales Gauss Fenster und zweidimensionale Gitter gelöst. Es wird vermutet, dass für das eindimensionale Gauss Fenster ein hexagonales Gitter die beste Wahl für ein Gabor System bieten sollte, da der ε-Träger

der Ambiguitätsfunktion eine Kreisscheibe ist und eine hexagonales Gitter das ökonomischte Arrangement für Kreisscheiben bietet.

Als Maß für die geometrische Güte eines Gitters werden Lösungen von Packungsproblemen verwendet und es werden Überlegungen angestellt, wie sich diese unter Hamiltonschen Deformationen verhalten.

Curriculum Vitae

Markus Faulhuber

Personal Details

Date of Birth:	March 23, 1985
Place of Birth:	Vienna, Austria
Nationality:	Austria
E-Mail:	markus.faulhuber@aon.at

Education

2012 - 2014	Master's degree programme in Mathematics, University of Vienna
2006 - 2012	Bachelor's degree programme in Mathematics, University of Vienna
2004 - 2006	Bachelor's degree programme in Computer Sciences, University of Vienna and Technical University of Vienna
1995 - 2003	Bundesrealgymnasium Gaenserndorf, Lower Austria

Academic Qualifications

03/2012	Bachelor of Science (BSc.) in Mathematics from the University of Vienna

---

Professional Career *(Selection)*

since 10/2012	**NuHAG, Faculty for Mathematics, University of Vienna** Project: P23902, "Hamiltonian Deformation of Gabor Frames" Head of Project: Prof. Maurice de Gosson
since 10/2012	**DIBB, University of Natural Resources and Life Sciences, Vienna** Tutor in Mathematics for

- Food- and Biotechnology
- Environmental and Bioresources Management
- Forestry
- Wood and natural Fibre Engineering

Scientific Talks

January 21, 2014 **IST Austria, Klosterneuburg** (invited)
Lattice Deformations and Packing Problems