UNCG—Mark McConnell's Problems

0.1 Short Vectors in Lattices

Let $L = \{ A \cdot \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right] \mid a, b \in \mathbb{Z} \}$ be the lattice in $\mathbb{R}^2$ generated by the columns of

$$A = \begin{bmatrix} -3070 & 2537 \\ 9910 & -8187 \end{bmatrix}. \tag{1}$$

The columns of A are a *lattice basis* for L.

The first column of A is a vector of length about 10374.63. The second column is shorter, of length 8571.08.

Question. Does L have any non-zero vectors shorter than that second column?

In (0.1.1)–(0.1.3) we walk you through three ways to answer this kind of question in $n = 2$ dimensions. In (0.1.4), we comment on how the methods generalize to higher n.

Remark. Whenever $\vec{v}$ is in a lattice, $-\vec{v}$ is in it also. In the computations outlined in this document, it makes no difference if a result is one or the other.

0.1.1

Pick a bound r. For all integers a, b between $-r$ and r, compute the length of $A \cdot \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$, and see which vector is shortest (not including $\vec{0}$). This is easy to code up. But how big do you have to take r? Is $r = 10$ enough? 20? It isn't, though won't know why not until (0.1.2). To begin to explore why not, use your favorite 2D graphics package to plot the points for a given r. In Sage, for example, let `pts` be a list of the points, and use

```
plot2d(pts).save('mydots.png')
```

The points form an array of dots (a lattice) in a parallelogram with the four corners $A \cdot \left[\begin{smallmatrix} \pm r \\ \pm r \end{smallmatrix}\right]$. You'd think you could just look at the array and see which dot is closest to $\vec{0}$. The trouble with this lattice basis is that the parallelogram is extremely narrow. When the lattice points are plotted as dots of default size, they pile on top of each other, almost like a line segment. You can't see which one is shortest. You certainly can't see whether $r = 20$ has gotten you the very shortest.

The angle between the columns of A (what is it?) is very close to 180°. This is why the parallelogram is so thin. We say the lattice basis A is ill-conditioned, or just plain "bad". In (0.1.3) we will learn how to replace it with a better basis. First, we turn to a method that is guaranteed to work with either a good or a bad basis.

0.1.2

For any column vector $\vec{v}$, the square of its length $\|\vec{v}\|$ is $\|\vec{v}\|^2 = \vec{v}^T\vec{v}$, where ^{T} denotes the transpose. The squared length of $A \cdot \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$ for $a, b \in \mathbb{Z}$ is therefore

$$\begin{bmatrix} a & b \end{bmatrix} A^T A \begin{bmatrix} a \\ b \end{bmatrix} = 107633000a^2 - 177843520ab + 73463338b^2. \tag{2}$$

Let's complete the square. We want to find a number C so that the last expression starts off with a square:

$$107633000(a + Cb)^2 + (?).$$

The (?) will still be a constant times b^2. Show that it works out as

$$107633000\left(a - \frac{2223044}{2690825}b\right)^2 + \frac{287282}{538165}\,b^2. \tag{3}$$

Good news—the coefficient of b^2 is positive. Since (3) is the sum of two squares[1] with positive coefficients, it is always positive unless $a = b = 0$. (Actually, this is not news: formula (2) is the squared length of a vector, which must be positive except at $\vec{0}$.)

The secret to finding short lattice vectors lies in that interesting rational number $\frac{2223044}{2690825}$. Since a and b are integers, they must have absolute value $\geqslant 1$ (unless they're 0). But if we choose a, b carefully, we may be able to get $\left(a - \frac{2223044}{2690825}b\right)$ down to absolute value less than 1, even though a and b are integers. This makes (3) smaller. Of course, there is a price to pay. The "careful" choice of a, b may involve a large b, so that the right-hand term of (3) becomes too large. We have to balance the terms of (3) to make the total small.

Let's write a program to search for the minimal non-zero value of (3). The important thing is that we have a rigorous bound so the program is guaranteed to find the minimum. We've already remarked that the second column of A is shorter than the first. Its squared length is 73463338. We want to beat that, so we only care about a, b where

$$(\text{formula 3}) \leqslant 73463338. \tag{4}$$

Since (3) has positive terms, a necessary condition for (4) is

$$\frac{287282}{538165}\,b^2 \leqslant 73463338,$$

equivalent to

$$|b| < 11731.103. \tag{5}$$

[1] For general n, formula (3) becomes a sum of n squares. It is essentially the *Cholesky decomposition* of the quadratic form $A^T A$. Search for Cholesky in your favorite linear algebra book or software package.

Write a program that iterates over the integers $b = -11731, \ldots, 11731$. For each b, find the integer a that minimizes (3). Show that the shortest non-zero vector in L is

$$\begin{bmatrix} 21 \\ -11 \end{bmatrix}.$$

(How much shorter is this vector than the columns of A?) Show that the shortest vector is attained at

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 19 \\ 23 \end{bmatrix}.$$

To understand what we've just done, temporarily let a, b be real variables rather than integers, and use software to plot the solution set of (4). You get the boundary and interior of an ellipse, but it is an extremely long and narrow ellipse. The ellipse is inscribed in a rectangle (a bounding box) with sides parallel to the a- and b-axes. The height of this bounding box is 11731.103 (why?) What is the width of the box, and why is it so much smaller than 11731.103?

0.1.3

A *lattice basis reduction* algorithm aims to take a bad basis of a lattice and turn it into a good basis. Gauss worked on lattices in dimension $n = 2$, and he came up with an algorithm for a good basis. Not only is the algorithm efficient, but it always finds the shortest vector in the lattice. We emphasize that Gauss's algorithm is only for $n = 2$; see the next section for higher n.

Here is Gauss's algorithm. As input, we are given a lattice basis for $n = 2$. Put the two basis vectors into a set S.

1. Let $\vec{v}$, $\vec{w}$ be the two elements of S. Let $\vec{x} = \vec{v} + \vec{w}$ or $\vec{v} - \vec{w}$, whichever is shorter (but don't let $\vec{x} = \vec{0}$).

2. If $\vec{x}$ is strictly shorter than one of the vectors in S, then remove the longest vector in S, replace it with $\vec{x}$, and go to Step 1. Otherwise, exit, returning S.

As an exercise, apply Gauss's algorithm to the columns of A in (1). Show that the result is the lattice basis given by the columns of

$$A_{\text{best}} = \begin{bmatrix} 21 & 145 \\ -11 & 285 \end{bmatrix}. \tag{6}$$

The first column is the shortest vector, which we found at the end of (0.1.2).

As another exercise, rewrite the program in (0.1.2), using A_{best} instead of A. How do (2)–(5) change? The new bound on $|b|$ is much nicer than 11731. How do you see right away that $a = 1$, $b = 0$ must give the shortest vector?

Show that Gauss's algorithm always produces a lattice basis of the same lattice L that it starts with. *Hint:* the matrix $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, and its inverse (which is what?), are both matrices with entries in $\mathbb{Z}$; everything the algorithm does can be inverted while preserving L.

This last point is important in practice. Since the columns of A and of A_{best} generate the same L, there must be a matrix γ such that γ and γ^{-1} both have entries in $\mathbb{Z}$ and such that

$$A_{\text{best}} = A\,\gamma. \tag{7}$$

This means $\gamma \in \mathrm{GL}_2(\mathbb{Z})$, the group of integer matrices of determinant ± 1. A computationalist often needs to know γ, so please compute it now. If you programmed up an implementation of Gauss's algorithm, consider modifying it so that it computes γ and γ^{-1} as it goes along and returns them at the end.

0.1.4 How the above generalizes to $n > 2$.

Searching over a parallelepiped P of sides $\pm r$ works worse and worse for higher n. The number of lattice points in P is $(2r+1)^n$, which grows exponentially with n.

The algorithm developed in (0.1.2) has been known for a long time. It is presented for general n in Henri Cohen's book[2], the first of two volumes that develop and explain everything that's in the computer algebra system GP/Pari. The algorithm is called `qfminim` in Pari. It also appears in Magma, Sage, etc.

Sadly, the algorithm's running time is still exponential in n. This is because it searches over an ellipsoid inscribed in the parallelpiped P. As a simple illustration, consider the ball of radius r in $\mathbb{R}^3$ inscribed in a cube of sides $\pm r$. The volume of the ball is $\frac{4}{3}\pi r^3$, that of the cube is $8r^3$, and the ratio of these volumes is $\alpha_3 = \frac{4}{3}\pi/8 = \frac{\pi}{6}$. Alpha ($\alpha$) is for Archimedes, who first proved the formula $\frac{4}{3}\pi r^3$. You can compute, or look up on Wikipedia, the volume of the ball of radius r in $\mathbb{R}^n$, and deduce the ball-to-cube volume ratio α_n for all n. The volume of the ellipsoid in `qfminim` is α_n times the volume of the parallelepiped P. One can show $\alpha_n \to 0$ as $n \to \infty$, but for a fixed n it is still "a constant". The running time of `qfminim` is this constant times the time to search over all of P.

Something very important for the performance of `qfminim` is the quality of the lattice basis. With a good basis and reasonable n, `qfminim` often performs fast enough. With a bad basis, `qfminim` can be terrible even for small n.

Finally, here is a crucial principle:

Everything about lattices is too easy in dimension $n = 2$.

Gauss's algorithm, for instance, stops working for n only a little larger than 2. The best general-purpose lattice basis reduction algorithm is Lenstra-Lenstra-Lovász, or LLL. It is described in many places, such as Henri Cohen's book already cited, or this[3] good modern book on all aspects of LLL.

As an exercise, use LLL in your favorite software package to reduce the basis A in (1). You should obtain (6). Here is how to do it in Sage. Sage's

[2] Henri Cohen, *A Course in Computational Algebraic Number Theory*, Grad Texts in Math 138, Springer.

[3] Murray R. Bremner, *Lattice Basis Reduction: An Introduction to the LLL Algorithm and Its Applications*, Chapman and Hall Pure and Appl Math.

matrices are made of row vectors, not column vectors, so we have to take the transpose.

```
sage: A
[-3070  2537]
[ 9910 -8187]
sage: A.transpose().LLL().transpose()
[ -21 -145]
[  11 -285]
```

In Pari, use `qflll` or `qflllgram`.

Lattice-based cryptography is grounded in the belief that no algorithm can find a good basis of an ill-conditioned lattice basis when n is around 500 or 1000, at least not for zillions of years. If and when quantum computers become available, RSA and elliptic-curve cryptography will be broken, by Shor's algorithm, but it seems that lattice-based cryptography will still be secure. For this reason, it's called a form of *post-quantum cryptography*. NTRU, founded by the number theorists Hoffstein, Pipher, and Silverman, was the first commercial version of lattice-based cryptography. See the Wikipedia article on NTRUEncrypt for more information.

0.2 Topology of Modular Curves

These exercises concern $\Gamma(N)$, $\Gamma_0(N)$, and $\Gamma_1(N)$ from section (1.15) of Paul Gunnells' notes. The definition

$$\Gamma(N) = \{\gamma \in \mathrm{SL}_n(\mathbb{Z}) \mid \gamma \equiv I \bmod N\}$$

is valid for any n. Throughout Section 0.2, take any $n \geqslant 2$.

0.2.1

Show that $\Gamma(N)$ is a normal subgroup of $\mathrm{SL}_n(\mathbb{Z})$, but $\Gamma_0(N)$ and $\Gamma_1(N)$ are not.

0.2.2

As in Gunnells' notes (1.15), establish the exact sequence

$$1 \to \Gamma(N) \to \mathrm{SL}_n(\mathbb{Z}) \xrightarrow{\nu} \mathrm{SL}_n(\mathbb{Z}/N\mathbb{Z}) \to 1.$$

Hints. First assume N is prime, so $\mathbb{Z}/N\mathbb{Z}$ is a field. Every matrix of rank n over a field can be row-reduced down to the identity using the *elementary row operations*: add a multiple of one row to another, multiply a row by a scalar, and permute rows. (What minor modifications are needed to preserve determinant 1?) E is an *elementary matrix* if the map $A \mapsto EA$ is one of the elementary row operations on A. Are the elementary matrices in the image of ν?

Imagine you need a subroutine to compute preimages under ν. If you had to code it up in a quick and dirty way, would you use elementary matrices, or something else? What about production code that has to run blindingly fast?

Generalize to composite N.

0.2.3

Find the order of $\mathrm{GL}_n(\mathbb{Z}/N\mathbb{Z})$, the group of all invertible $n \times n$ matrices with entries in $\mathbb{Z}/N\mathbb{Z}$. Go back in time to last year and find the order of $\mathrm{GL}_2(\mathbb{Z}/7\mathbb{Z})$.

Find the order of $\mathrm{SL}_n(\mathbb{Z}/N\mathbb{Z})$. By the previous exercise, this gives the index $[\mathrm{SL}_n(\mathbb{Z}/N\mathbb{Z}) : \Gamma(N)]$.

0.2.4

Let's prove that $\Gamma(N)$ is torsion-free whenever $N \geqslant 3$.

Here are some examples to get started. The polynomial $x^5 - 1$ factors into two pieces $(x-1)(x^4+x^3+x^2+x+1)$ in $\mathbb{Z}[x]$. For x^6-1, it's more complicated: $(x-1)(x+1)(x^2+x+1)(x^2-x+1)$.

(a) For any positive integer k, tell the story of how $x^k - 1$ factors in $\mathbb{Z}[x]$. Do you ever get factors to the 2nd or higher power?

(b) What is the order of the element $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix}$ of $\mathrm{SL}_4(\mathbb{Z})$? Relate this to the factorization of $x^{12} - 1$ in $\mathbb{Z}[x]$.

(c) Review *Jordan canonical form* and *rational canonical form* in your favorite book. What does this have to do with (b)? (As a matter of notation, would your book have used the transpose of that matrix instead?)

(d) Classify all elements of finite order in $\mathrm{SL}_n(\mathbb{C})$ up to conjugacy.

(e) Classify all elements of finite order in $\mathrm{SL}_n(\mathbb{Q})$ up to conjugacy. Show that every conjugacy class has a representative in which all entries are 0 or ± 1.

(f) Classify all elements of finite order in $\mathrm{SL}_n(\mathbb{Z})$ up to conjugacy. Can you have elements of finite order γ, $\gamma' \in \mathrm{SL}_n(\mathbb{Z})$ where $\gamma' = \alpha\gamma\alpha^{-1}$ for some $\alpha \in \mathrm{SL}_n(\mathbb{Q})$ but for no $\alpha \in \mathrm{SL}_n(\mathbb{Z})$?

(g) Conclude that $\Gamma(N)$ is torsion-free whenever $N \geqslant 3$.

0.3 The Well-Rounded Retract for $n = 2$

The standard fundamental domain D for the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathfrak{H}$ is the gray region in Figure 1 of this document. D is cut out by the inequalities $|z| \geqslant 1$ and $|\operatorname{Re} z| \leqslant \frac{1}{2}$. Compare Figure 2 of Gunnells' notes.

0.3.1

Let L be the lattice with the basis (1) from Section 0.1. The lattice also has the good basis (6).

(a) Change L by a rotation and a scaling so that its shortest vector is $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$, which is identified with $1 \in \mathbb{C}$.

Important: we could not even begin to do this if we had not done the work in Section 0.1 to find the shortest vector. This is one reason algorithms like LLL are important.

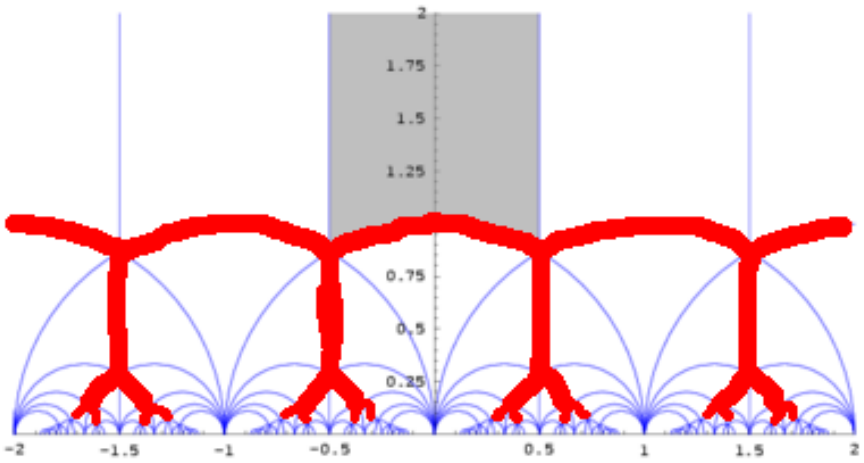

Figure 1: The Well-Rounded Retract for $\mathrm{SL}_2(\mathbb{Z})$

In the good basis, what happens to the second basis vector under this change? Multiply by -1 if necessary so the second vector corresponds to a point $z \in \mathfrak{H}$. Find z. Does it lie in D? Why must it lie there?

Find the point w to which the well-rounded retraction carries z. Draw a picture.

Let γ be as in (7). γ^{-1} acts on z as a linear fractional transformation, carrying it to some z' lying in some far-away translate $\gamma^{-1}D$ of D in $\mathfrak{H}$. Find z'. Find the point w' to which the well-rounded retraction carries z' (*hint:* the well-rounded retraction is $\mathrm{GL}_2(\mathbb{Z})$-equivariant).

0.3.2

Let $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ be a positive definite quadratic form Q. It sends $\begin{bmatrix} x \\ y \end{bmatrix}$ to

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

for $x, y \in \mathbb{Z}$. The *arithmetic minimum* of Q is the minimum value attained for all $x, y \in \mathbb{Z}$, not counting the obvious $(x, y) = (0, 0)$. A *minimal vector* of Q is an $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{Z}^2$ where the arithmetic minimum is attained.

You can always multiply a quadratic form by a positive scalar to make its arithmetic minimum equal to 1. This does not change the minimal vectors.

Whenever $\vec{v}$ is a minimal vector, so is $-\vec{v}$. When we talk about a minimal vector $\vec{v}$, we may not mention $-\vec{v}$, but it is understood.

Consider the family of all $Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ that have arithmetic minimum 1 and

such that $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ are minimal vectors. Show these Q form a one-parameter family, where the parameter is b ranging through a certain closed interval of real numbers. Show that, at one endpoint of the interval, you get a third minimal vector $\begin{bmatrix}1\\1\end{bmatrix}$; at the other endpoint, you get a third minimal vector $\begin{bmatrix}-1\\1\end{bmatrix}$; between the endpoints, the only minimal vectors are the first two, $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$. The two endpoints of the interval are *perfect forms*.

0.3.3

The well-rounded retract W for SL_2 is highlighted in red in Figure 1 above.

Superimposed on top of Figure 1, draw the *Farey triangle* with vertices 0, 1, and ∞. The Farey triangle is near the center of the Farey tesselation in Figure 5 of Gunnells' notes. When you draw the Farey triangle on top of Figure 1, you should see portions of three of the thick red arcs, meeting at the center of the Farey triangle. For these three red arcs, exactly half of the arc lies inside the Farey triangle, and the other half-arc is cut off, since it lies the outside of the Farey triangle. Call the union of these three red half-arcs the *triad in the Farey triangle*.

The triad is important because the well-rounded retraction retracts the Farey triangle onto the triad. All the points above the triad (in the direction of $i\infty$) flow downwards along vertical lines onto the top of the triad. The points near 0 flow up and to the right onto the triad. The points near 1 flow up and to the left onto the triad. This is a homotopy equivalence.

Let's redo some of Gunnells' exercises, getting the triad involved. Show that $Y(3)$ is made from exactly four Farey triangles. Combinatorially, they are laid out like the faces of a regular tetrahedron. $Y(3)$ has four punctures, lying at the vertices of the regular tetrahedron. Draw the triad on each of the four Farey triangles. Show that the union of the four triads is a graph with six edges and four trivalent vertices. (When we walk from the center of one triangle to the center of another triangle, we are tracing out one edge of this graph. Halfway along the walk, we cross an edge of the tetrahedron.) The graph is $\Gamma(3)\backslash W$. The well-rounded retraction makes $Y(3)$ flow onto $\Gamma(3)\backslash W$, as a homotopy equivalence.

The Euler characteristic of a graph with four vertices and six edges is $4 - 6 = -2$. Clearly $Y(3)$ is connected, so $\dim H_0(Y(3)) = 1$. Deduce that $\dim H_1(Y(3)) = 3$. Draw three 1-cycles that generate the homology.

Work out $Y(4)$ in the same way. You get the octahedron in place of the tetrahedron. The graph has 8 trivalent vertices and 12 edges, the Euler characteristic is -4, and $\dim H_1(Y(4)) = 5$. For $Y(5)$, you get the icosahedron in place of the tetrahedron. For $N \geqslant 7$, you no longer get Platonic solids, but surfaces of higher genus (doughnut holes).

0.4 The Well-Rounded Retract for $n = 3$

Consider all positive definite 3×3 quadratic forms

$$Q = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

that have arithmetic minimum 1 and such that

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

are minimal vectors. Show that the family of these Q is a three-dimensional convex body in the space of the three parameters b, c, and e. Find the vertices, edges, and faces of this convex body. Show that, combinatorially, it is a cube where you chop off a neighborhood of four of the eight corners (the pairs of corners that get chopped do not share a common edge). This is called the *Soulé cube*. Mark McConnell can give you hints on this exercise.