Decomposition of Masses

CGC Pre-Doc Project

Michael Eisenring*

January 7, 2002

1 Problem Setting

In this project we consider the following problem:

> [1]Proteins are large molecules made up of smaller molecules called amino-acids. The amino-acids bind together in a certain order to form the protein; thus, a protein can be viewed as a string over a finite alphabet (this is its so-called primary structure). Each amino-acid has a specific molecular weight (or mass), which can be specified very exactly. Given a substring of a protein (a peptide), its molecular weight is simply the sum of the masses of the individual amino-acids. However, the converse is less obvious: Given a mass $M \in \mathbb{R}^+$, is there a multiset of amino-acids such that their mass equals M? Or put differently, is there a peptide whose mass is exactly M? We refer to such a multiset of amino-acids as a *decomposition* of M.

Example: The four amino-acids threonine (T), serine (S), alanine (A), and arginine (R) have the weights

$$\begin{aligned} m(\mathrm{T}) &= 101.04768, \\ m(\mathrm{S}) &= 87.03203, \\ m(\mathrm{A}) &= 71.03711, \\ m(\mathrm{R}) &= 156.10111. \end{aligned}$$

So the weight of the string TTSAR is 516.26561. We are interested in the converse case, where for a given mass M, e.g. $M = 516.26561$, we look for decompositions of M.

*supervised by Mark Cieliebak, Zsuzsanna Lipták and Emo Welzl

[1]written by Zsuzsanna Lipták

In general, we would like to find, for a given mass M and an error tolerance ε, all multisets of letters with weights $M \pm \varepsilon$. In reality we are given twenty amino-acids with weights between 57 and 186 and we want to determine substrings of 20-100 amino-acids. In the laboratory we can weigh with a precision of ± 0.5.

2 Mathematical formulation of the problem

We are given an alphabet of constant size $d \in \mathbb{N}$ and weights $m_1, \dots m_d \in \mathbb{R}^+$ for the different letters. For a given weight $M \in \mathbb{R}^+$ and an error tolerance $\varepsilon > 0$, our question is if there exist multiplicities $k_1, \dots k_d \in \mathbb{N}_0$ such that

$$k_1 m_1 + \dots + k_d m_d \ \in \ [M - \varepsilon, M + \varepsilon].$$

In vector formulation we define the weight vector $m := (m_1, \dots m_d)$ and then ask for a grid point $k = (k_1, \dots k_d) \in \mathbb{N}_0^d$ such that

$$\langle k, m \rangle \ \in \ [M - \varepsilon, M + \varepsilon]. \tag{1}$$

Because we are interested in all possible decompositions of M, we would like to characterize the set of all grid points $k \in \mathbb{N}_0^d$ satisfying (1):

$$\Gamma := \{k \in \mathbb{N}_0^d \mid \langle k, m \rangle \in [M - \varepsilon, M + \varepsilon]\}.$$

Geometrically, Γ is the subset of $\mathbb{N}_0^d$ lying between the hyperplanes h^+ and h^- defined by

$$h^+ := \{x \in \mathbb{R}^d \mid \langle x, m \rangle = M + \varepsilon\},$$
$$h^- := \{x \in \mathbb{R}^d \mid \langle x, m \rangle = M - \varepsilon\}.$$

3 Algorithm

As the masses of the amino-acids should be measurable within sufficient exactness, we can assume that ε is much smaller than the mass of the lightest amino-acid:

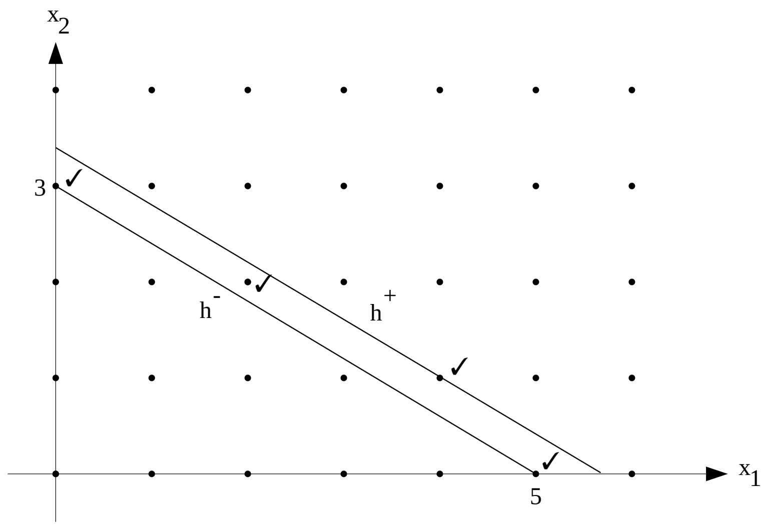
$$\varepsilon \ll \min\{m_1, \dots m_d\}. \tag{2}$$

In the following we mean by equation (2) that $\varepsilon > 0$ is chosen sufficiently small such that every hyperplane parallel to one of the axes intersects with at most one point of Γ, or in other terms:

$$\|k - k'\|_1 \geq 2 \quad \forall\, k, k' \in \Gamma,\ k \neq k'.$$

For the intuition we first investigate the case of having a binary alphabet, that is $d = 2$: we look for all grid points $(k_1, k_2) \in \mathbb{N}_0^2$ which lie between the two lines $h^{\pm} := \{(x_1, x_2) \in \mathbb{R}^2 \mid m_1 x_1 + m_2 x_2 = M \pm \varepsilon\}$.

Example: $m_1 = 3, m_2 = 5, M = 16, \varepsilon = 1$.



From the figure or by calculating we get $\Gamma = \{(5,0),(4,1),(2,2),(0,3)\}$.

To get all possible decompositions lying in the interval $[M-\varepsilon, M+\varepsilon]$, we could proceed as follows: we start by looking at the two points where h^- and h^+ intersect with the x_1-axis. If there is a grid point on the x_1-axis lying between these two points, it is an element of Γ. Then we go backwards on the x_1-axis to the next smaller grid point, always checking if the parallel to the x_2-axis contains a grid point lying above h^- and below h^+.

Algorithm $\textsc{Test}(m_1, m_2, M, \varepsilon)$
Input: $m_1, m_2, M, \varepsilon \in \mathbb{R}^+$,
    $\varepsilon \ll \min\{m_1, m_2\}$,
    $\mathfrak{L} = \emptyset$
Output: $\mathfrak{L} = \Gamma$
    $k_1 \leftarrow \lceil \frac{M-\varepsilon}{m_1} \rceil$
    if $k_1 = \lfloor \frac{M+\varepsilon}{m_1} \rfloor$ **add** $(k_1, 0)$ **to** $\mathfrak{L}$
    while $k_1 > 0$
        do $k_1 \leftarrow k_1 - 1$
            $k_2 \leftarrow \lceil \frac{M-\varepsilon-k_1 m_1}{m_2} \rceil$
            if $k_2 = \lfloor \frac{M+\varepsilon-k_1 m_1}{m_1} \rfloor$ **add** (k_1, k_2) **to** $\mathfrak{L}$
    return $\mathfrak{L}$

The number of if-tests is $\lceil \frac{M-\varepsilon}{m_1} \rceil + 1$, hence we can reduce it by choosing $m_1 \geq m_2$. In the example, interchanging m_1 and m_2 would reduce the number of if-tests from 6 to 4.

The following algorithm IMPLICIT is a generalization of the above algorithm TEST to any dimension d:

Algorithm IMPLICIT$(d, m_1, \ldots m_d, M, \varepsilon)$
Input: $d \in \mathbb{N}, m_1, \ldots m_d, M, \varepsilon \in \mathbb{R}^+$,
$\varepsilon \ll \min\{m_1, \ldots m_d\}$,
$\mathfrak{L} = \emptyset$
Output: $\mathfrak{L} = \Gamma$
$k_1 \leftarrow \lceil \frac{M-\varepsilon}{m_1} \rceil$
if $k_1 = \lfloor \frac{M+\varepsilon}{m_1} \rfloor$ **add** $(k_1, 0, \ldots 0)$ **to** $\mathfrak{L}$
if $d > 1$
  while $k_1 > 0$
   do $k_1 \leftarrow k_1 - 1$
    $\mathfrak{L}' \leftarrow$ IMPLICIT$(d-1, m_2, \ldots m_d, M - k_1 m_1, \varepsilon)$
    add $(k_1, \mathfrak{L}')$ **to** $\mathfrak{L}$
return $\mathfrak{L}$

By $(k_1, \mathfrak{L}')$ we mean the set of grid points $\{(k_1, k_2, \ldots, k_d) \mid (k_2, \ldots, k_d) \in \mathfrak{L}'\}$. Analogous to the two-dimensional case the number of if-tests is

$$\prod_{i=1}^{d-1} \left(\left\lceil \frac{M-\varepsilon}{m_1} \right\rceil + 1 \right)$$

and to get it as small as possible we choose $m_d = \min\{m_1, \ldots m_d\}$. Thus the number of if-tests is polynomial of order $O(M^{d-1})$ with leading term $1/(m_1 \cdots m_{d-1})$.

In the problem setting of the twenty amino-acids we have $m_1 \cdot \ldots \cdot m_{19} = 2.75 \cdot 10^{39}$. An amino-acid has an average weight of about 120, thus determining a string of length 50 and weight $50 \cdot 120$ costs about

$$\frac{(50 \cdot 120)^{19}}{2.75 \cdot 10^{39}} = 2.22 \cdot 10^{32}$$

if-tests. Hence calculating all possible decompositions can become quite expensive!

4 Lattice points of tetrahedra

Suppose that we are interested only in the *number* of all possible decompositions of M and not in the concrete solutions. Then a formula for the number of all grid points $k \in \mathbb{N}_0^d$ such that

$$\langle k, m \rangle \leq \lambda, \quad \lambda \in \mathbb{R}^+,$$

would lead us close to the solution of the problem: denoting the number of all these points by $N_d(\lambda; m)$, we get

$$\begin{aligned} &N_d(M+\varepsilon; m) - N_d(M-\varepsilon; m) \\ =\ &\#\{k \in \mathbb{N}_0^d \mid \langle k, m\rangle \in (M-\varepsilon, M+\varepsilon]\} = \#(\Gamma - h^-). \end{aligned} \tag{3}$$

Unfortunately, there are no exact formulas for $N_d(\lambda; m)$. Lehmer [1] showed the existence of polynomials $P_d(\lambda; m), Q_d(\lambda; m)$ of degree d in λ such that

$$P_d(\lambda; m) < N_d(\lambda; m) < Q_d(\lambda; m).$$

Furthermore $P_d(\lambda; m)$ and $Q_d(\lambda; m)$ have the same leading term, namely

$$\frac{\lambda^d}{d! m_1 m_2 \cdot \ldots \cdot m_d},$$

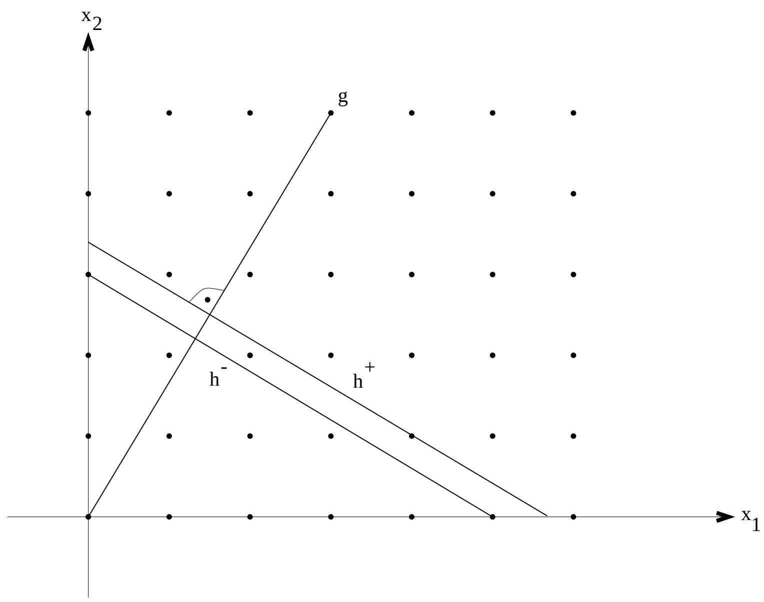
which yields an error of $O(\lambda^{d-1})$. Thus, we can estimate the number (3) with an error of $O(M^{d-1})$. Spencer [3] showed that, if $\frac{m_i}{m_j} \notin \mathbb{Q}$ for some i, j, then there exists a polynomial $R_d(\lambda)$ of degree d in λ such that

$$N_d(\lambda; m) = R_d(\lambda) + o(\lambda^{d-1}).$$

Since we measure only rational quotients $\frac{m_i}{m_j}$ in the laboratory, Spencer's result doesn't help for our problem setting.

5 Orthogonal projection

We consider the line $g := \{t(m_1, \ldots m_d) \mid t \in \mathbb{R}\}$ orthogonal to the hyperplanes h^+ and h^-. Then a grid point $(k_1, \ldots k_d) \in \mathbb{N}_0^d$ is an element of Γ if and only if its orthogonal projection onto g has distance $\lambda \in [M-\varepsilon, M+\varepsilon]$ from the origin.



This leads us to the following idea: we project the grid $\mathbb{N}_0^d$ orthogonally onto g, getting a point set $\mathfrak{P}$, and then investigate how the set $\mathfrak{P}$ is distributed on g.

If the masses $m_1, \ldots m_d$ are rational, I conjecture that the points of $\mathfrak{P}$ have minimal positive distance from each other:

$$\delta := \inf\{|p - q| \mid p, q \in \mathfrak{P}, p \neq q\} > 0.$$

For the two-dimensional case there is a simple formula for δ:

Proposition 1. *If $m_1 : m_2 = u : v$ for relatively prime numbers $u, v \in \mathbb{N}$, then*

$$\delta = \frac{1}{\sqrt{u^2 + v^2}}.$$

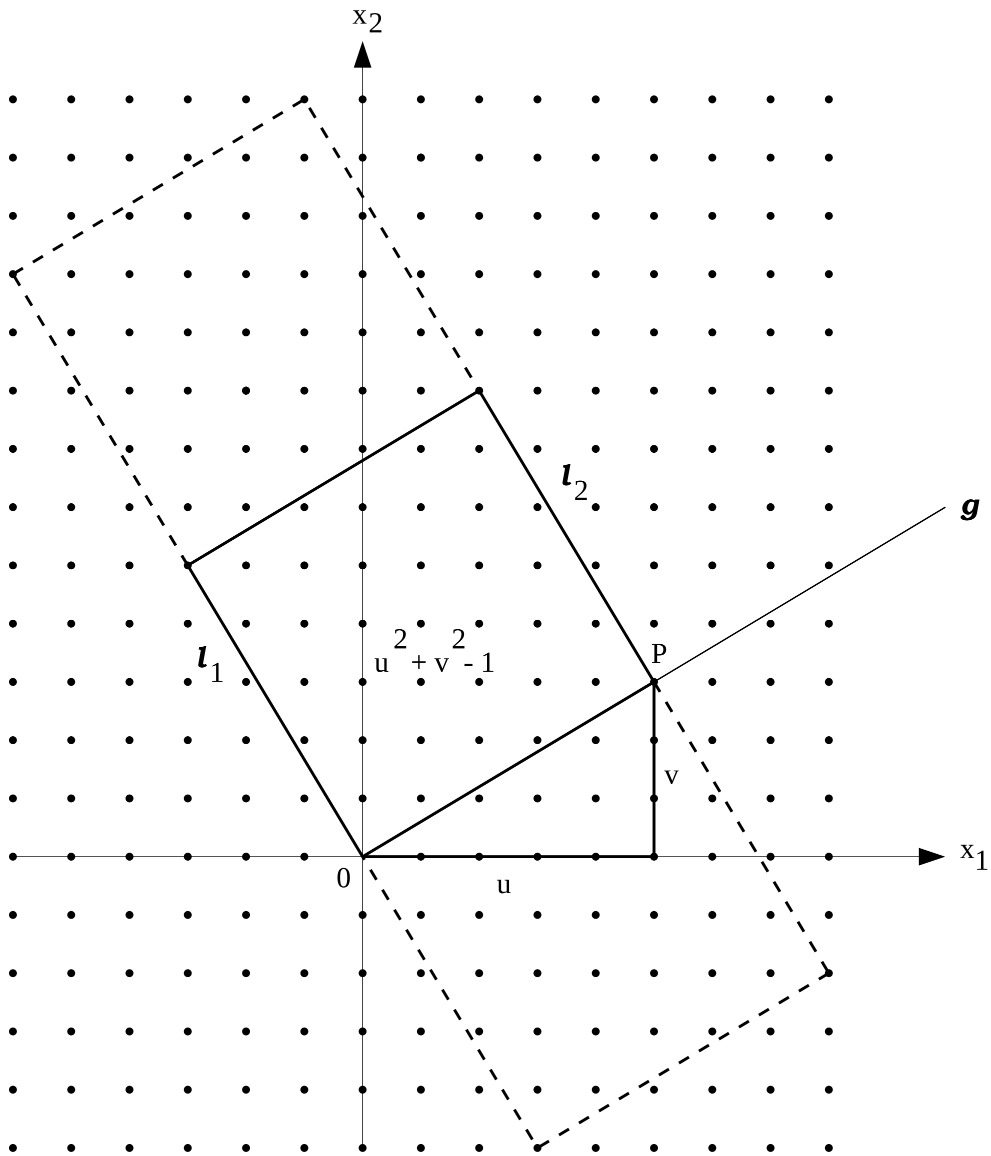
For the proof we need Pick's Theorem:

Lemma 2 ([2]). *Let A be the area of a simply closed lattice polygon*[2]*. Let B denote the number of lattice points on the edges and I the number of points in the interior of the polygon. Then*

$$A = I + \frac{1}{2}B - 1.$$

Proof. Replace $\mathbb{N}_0^2$ by the whole grid $\mathbb{Z}^2$ and observe the square spanned over the rectangular triangle with sides u and v:

[2] A lattice polygon is a polygon with all its corners on the grid.



Because we assumed u and v to be relatively prime, the only lattice points on the edges of the square are the four corners. With Pick's Theorem, $A = u^2 + v^2$ and $B = 4$ we get the number I of interior grid points in the square to be $I = u^2 + v^2 - 1$. The projections of these interior points onto g divide the segment $\overline{0P}$ into $u^2 + v^2$ pieces, and by the intercept theorem (in German: Strahlensatz) they all have the same length $\frac{\sqrt{u^2+v^2}}{u^2+v^2} = \frac{1}{\sqrt{u^2+v^2}}$.

Now we look at all grid points between the two parallel lines l_1 and l_2: dividing this area into squares as in the figure it is easy to see that for every grid point there is a unique grid point in the interior of the original square which has the same image under the projection, or in other words: projecting only the interior grid points of the square yields already all image points $\mathfrak{P}$. $\square$

References

[1] Lehmer, D. H., The lattice-points of an n-dimensional tetrahedron, Duke Math. Journ. 7, p. 341-353 (1940).

[2] Coxeter, H. S. M., Introcuction to Geometry, 2nd ed. New York: Wiley, p. 209 (1969).

[3] Spencer, D. C., The lattice points of tetrahedra, Journ. Math. Phys., Mass. Inst. Tech., 21, p. 189-197 (1942).