

6 Diophantine Approximation and Lattice Reduction

In Section 2, we presented a polynomial-time algorithm for solving a linear diophantine matrix equation. In this section, we discuss how we can find approximate solutions to linear diophantine equations with some prescribed accuracy. Our presentation is mainly based on the excellent book “Geometric Algorithms and Combinatorial Optimization” by Grötschel, Lovász and Schrijver [5, Chapter 5].

Two closely related optimization problems we discuss here are (1) the closest vector problem and (2) the shortest vector problem that were briefly presented in Section 1.

6.1 Diophantine Approximation

Dirichlet’s Theorems

For a given rational number a , a positive integer N , and a positive rational number ϵ , the *diophantine approximation problem* is to decide whether there exists a rational number a' with denominator at most N such that $|a - a'| < \epsilon$.

Theorem 6.1 (Dirichlet (1842)) *For a given real number a and $0 < \epsilon < 1$, there exist integers p and q such that $1 \leq q \leq 1/\epsilon$ and $|qa - p| \leq \epsilon$.*

Proof. A short elegant non-polynomial proof omitted. See, for example, [5]. ■

For any rational a , there is a polynomial-time algorithm to find integers p and q required in Theorem 6.1. This uses the notion of continued fractions and the Euclidean algorithm. This will be discussed briefly in Section 6.1.

For given rational numbers a_1, \dots, a_n , a positive integer N , and a positive rational number ϵ , the *simultaneous diophantine approximation problem* is to decide whether there exist integers p_1, \dots, p_n , and a positive integer $q \leq N$ such that $|qa_i - p_i| < \epsilon$ for all i .

The following classical theorem gives a sufficient condition for the existence of such an approximation.

Theorem 6.2 (Dirichlet (1842)) *For given real numbers a_1, \dots, a_n and $0 < \epsilon < 1$, there exist integers p_1, \dots, p_n , and a positive integer $q \leq \epsilon^{-n}$ such that $|qa_i - p_i| < \epsilon$ for all i .*

Proof. A short elegant nonconstructive proof omitted. See, e.g., [5]. ■

Unfortunately, no polynomial algorithm is known for the simultaneous diophantine approximation problem nor to find an existing approximation in Theorem 6.2. We shall present a polynomial-time algorithm (Theorem 6.12) which finds an approximate solution, that satisfies substantially weakened conditions. Namely, we need to add an additional factor $2^{n(n+1)/4}$ before ϵ^{-n} for the upper bound of q .

Continued Fractions

For a sequence of integers, $a_0, a_1, \dots, a_i, \dots$ all positive except for possibly a_0 , the expression

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\ddots + \frac{1}{a_{i-1} + \frac{1}{a_i}}}}} \tag{6.1}$$

is called a *continued fraction*, and is denoted as $\langle a_0, a_1, \dots, a_i \rangle$.

For a given $a \in \mathbb{R}$, the sequence $a_0, a_1, a_2, a_3, \dots$ defined by

$$a_0 := \lfloor a \rfloor, \quad b_0 := a - \lfloor a \rfloor; \tag{6.2}$$

$$a_i := \left\lfloor \frac{1}{b_{i-1}} \right\rfloor, \quad b_i := \frac{1}{b_{i-1}} - \left\lfloor \frac{1}{b_{i-1}} \right\rfloor, \quad \text{as long as } b_{i-1} \neq 0 \tag{6.3}$$

is called the *continued expansion* of the number a .

For any $a \in \mathbb{Q}$, its continued expansion is a finite sequence a_0, a_1, \dots, a_k . In fact, this sequence can be computed in polynomial time by using the Euclidean algorithm described in Section 2.2.

Lemma 6.3 *Let $a \in \mathbb{Q}$, let $a_0, a_1, a_1, a_2, \dots$ be the continued expansion of a , and let p_i/q_i denote the canonical (coprime) representation of $\langle a_0, a_1, \dots, a_i \rangle$. Let i be an odd index such that a_{i+1} exists. Then,*

- (a) $\frac{p_i}{q_i} < a \leq \frac{p_{i+1}}{q_{i+1}}$,
- (b) $p_{i+1} q_i - p_i q_{i+1} = 1$.

Now we apply the Euclidean Algorithm to the pair of rational numbers a and 1, with a slight modification. In matrix form, the Euclidean algorithm generates a finite sequence of unimodular matrices T^1, T^2, \dots, T^k , α_i 's and β_i 's such that

$$\begin{aligned} [a, 1] T^1 &= [\alpha_1, \beta_0 = 1], \\ [a, 1] T^2 &= [\alpha_1, \beta_2], \\ &\vdots \\ [a, 1] T^i &= \begin{cases} [\alpha_i, \beta_{i-1}], & \text{if } i \text{ is odd,} \\ [\alpha_{i-1}, \beta_i], & \text{if } i \text{ is even,} \end{cases} \\ &\vdots \\ [a, 1] T^k &= [0, \beta_{k-1}] \text{ or } [\alpha_{k-1}, 0]. \end{aligned}$$

Here we assume that the algorithm starts by subtracting a multiple of the second (pivot) column with value 1 from the first and thus $\beta_0 = 1$, even if $a < 1$ (in this case, no change takes place). Also, we do not use the swap operation and thus a pivot alternates between the first and the second columns.

Theorem 6.4 *The unimodular matrices T^i 's generated by the Euclidean algorithm applied to $[a, 1]$ satisfy*

$$T^i = \begin{cases} \begin{bmatrix} q_i & -q_{i-1} \\ -p_i & p_{i-1} \end{bmatrix}, & \text{if } i \text{ is odd,} \\ \begin{bmatrix} q_{i-1} & -q_i \\ -p_{i-1} & p_i \end{bmatrix}, & \text{if } i \text{ is even.} \end{cases} \quad (6.4)$$

Therefore, the algorithm generates the continued expansion of the number a in polynomial time, in particular, with $q_k a - p_k = 0$ at the termination.

Now we show how Theorem 6.1 can be proved constructively by the polynomial algorithm for any rational a and $0 < \epsilon < 1$. First, compute the continued expansion of a , find the largest index i such that $q_i \leq 1/\epsilon$ and set $q := q_i$. If i is the last index, then p_i/q_i coincides with a . Otherwise, by Lemma 6.3, we have

$$\left| a - \frac{p_i}{q_i} \right| < \left| \frac{p_{i+1}}{q_{i+1}} - \frac{p_i}{q_i} \right| = \frac{|p_{i+1} q_i - p_i q_{i+1}|}{q_i q_{i+1}} = \frac{1}{q_i q_{i+1}} < \frac{\epsilon}{q_i}.$$

6.2 Lattice Reduction

Gram-Schmidt Orthogonalization

For an ordered basis $B = (b_1, b_2, \dots, b_n)$ of \mathbb{R}^n , the *Gram-Schmidt orthogonalization* (abbreviated by GSO) is the following procedure to compute an orthogonal basis $B^* = (b_1^*, b_2^*, \dots, b_n^*)$:

$$\begin{aligned} b_1^* &:= b_1; \\ b_i^* &:= b_i - \sum_{j=1}^{i-1} \frac{b_i^T b_j^*}{\|b_j^*\|^2} b_j^*, \quad i = 1, \dots, n. \end{aligned} \quad (6.5)$$

When b_1, b_2, \dots, b_n are rational vectors, the running time of the GSO can be bounded by a polynomial function of the input size. The analysis is similar to that of the Gaussian elimination (Theorem 2.4), because the vectors b_i^* 's are unique solutions to certain linear equations that are defined by the input vectors b_1, b_2, \dots, b_n .

Lemma 6.5 *The ordered vectors $(b_1^*, b_2^*, \dots, b_n^*)$ computed by the GSO satisfy the following properties:*

- (a) $(b_1^*, b_2^*, \dots, b_n^*)$ forms an orthogonal basis of \mathbb{R}^n with $\|b_i^*\| \leq \|b_i\|$ for each i ;
- (b) $\|b_1^*\| \|b_2^*\| \cdots \|b_n^*\| = |\det[b_1^*, b_2^*, \dots, b_n^*]| = |\det[b_1, b_2, \dots, b_n]|$;

(c) $b_i = \sum_{j=1}^i \mu_{ij} b_j^*$ for some μ_{ij} with $\mu_{ii} = 1$ for $i = 1, \dots, n$.

Now we have an application of the GSO to lattices.

Lemma 6.6 *Let $L(B)$ be the lattice generated by an ordered basis $B = (b_1, \dots, b_n)$ of \mathbb{R}^n and let $(b_1^*, b_2^*, \dots, b_n^*)$ be the orthogonal basis computed by the GSO. Then, for any lattice point $b \in L \setminus \{\mathbf{0}\}$,*

$$\|b\| \geq \min\{\|b_1^*\|, \|b_2^*\|, \dots, \|b_n^*\|\}. \quad (6.6)$$

Proof. Assume all the assumptions are met. Let b be any lattice point in $L \setminus \{\mathbf{0}\}$. Then, $b = \sum_{i=1}^n \lambda_i b_i$ for some integer λ_i 's. Let k be the largest index i such that $\lambda_i \neq 0$. By Lemma 6.5 (c), there are μ_i 's such that $b = \sum_{i=1}^k \mu_i b_i^*$ and $\mu_k = \lambda_k$ is a nonzero integer. Now,

$$\|b\|^2 = \sum_{i=1}^k \mu_i^2 \|b_i^*\|^2 \geq \mu_k^2 \|b_k^*\|^2 \geq \|b_k^*\|^2.$$

This completes the proof. ■

Recall the *shortest vector problem* which is to find a shortest nonzero vector in the lattice generated by $B = (b_1, \dots, b_n)$,

$$\min\{\|Bx\| : x \in \mathbb{Z}^n \setminus \{\mathbf{0}\}\}. \quad (6.7)$$

Here we identify an ordered basis B with the $n \times n$ matrix $[b_1, \dots, b_n]$. We will use this convention whenever no confusion may arise.

The following theorem provides a bounded region in which a shortest vector exists.

Theorem 6.7 *Let $L(B)$ be the lattice generated by a basis $B = (b_1, \dots, b_n)$ of \mathbb{R}^n and let the value $\alpha(B)$ be defined by*

$$\alpha(B) := \|b_1\| \|b_2\| \cdots \|b_n\|. \quad (6.8)$$

Then for any solution x^0 to

$$\min\{\|Bx\| : |x_j| \leq \frac{\alpha(B)}{|\det(B)|} \text{ for all } j, x \in \mathbb{Z}^n \setminus \{\mathbf{0}\}\}, \quad (6.9)$$

the vector Bx^0 is a shortest vector in the lattice $L(B)$.

Proof. Let v be a shortest vector in the lattice $L(B)$. Since $v = Bx$ for some $x \in \mathbb{Z}^n$, by Cramer's rule, $x_j = \det(B_j) / \det(B)$, where $B_j = [b_1, \dots, b_{j-1}, v, b_{j+1}, \dots, b_n]$. By Hadamard's inequality, $|\det(B_j)| \leq \|b_1\| \cdots \|b_{j-1}\| \|v\| \|b_{j+1}\| \cdots \|b_n\|$. Since v is a shortest vector, we have $\|v\| \leq \|b_j\|$ and thus $|\det(B_j)| \leq \alpha(B)$. It follows that $|x_j| \leq \alpha(B) / |\det(B)|$. This completes the proof. ■

Reduced Bases and the LLL Algorithm

As we learned in Section 2.4, $|\det(B)|$ has the same value for all bases of a full-dimensional lattice L , which we denote by $\det L$. Yet, there are bases that are “more orthogonal” than others in the sense that the value $\alpha(B)/|\det(B)|$ is reasonably small. Such a basis will help us, for example, find a smaller region to search for a shortest vector, see Theorem 6.6. Note that the base of \mathbb{R}^n computed by the GSO from an ordered basis of a lattice is typically not a lattice basis.

A ordered basis $B = (b_1, \dots, b_n)$ of a full-dimensional lattice L is called *reduced* if the following two conditions are satisfied:

$$|\mu_{ij}| \leq \frac{1}{2}, \quad \forall 0 \leq j < i \leq n; \tag{6.10}$$

$$\|b_{i+1}^* + \mu_{i+1,i}b_i^*\|^2 \geq \frac{3}{4}\|b_i^*\|^2, \quad \forall i = 1, \dots, n, \tag{6.11}$$

where $B^* = (b_1^*, b_2^*, \dots, b_n^*)$ is the orthogonal basis computed by the GSO and μ_{ij} are as defined in Lemma 6.5.

The first condition (6.10) says that the vectors b_1, \dots, b_n are nearly orthogonal. The second condition (6.11) is not very intuitive. One can interpret that when one exchanges the order of b_i and b_{i+1} , the vector $(b_{i+1}^* + \mu_{i+1,i}b_i^*)$ represents the new b_i^* and this condition says the new b_i^* does not get shortened too much.

Example 6.1 Let us take a small example. The lattice shown in Figure 6.1 is generated by (the columns of) a random 2×2 matrix $A = \begin{bmatrix} 46 & 48 \\ 48 & 23 \end{bmatrix}$. A reduced basis computed by the function *LatticeReduce* of *Mathematica 5.2* is $B = \begin{bmatrix} 2 & 50 \\ -25 & -2 \end{bmatrix}$. The alpha value $\alpha(A)$ is roughly 2.83 times larger than that $\alpha(B)$ of the reduced basis. The critical value $\frac{\alpha(B)}{|\det(B)|}$ in Lemma 6.7 is about 1.00722. From this, one can easily conclude (without watching the figure) that $(2, -25)^T$ is a shortest vector.

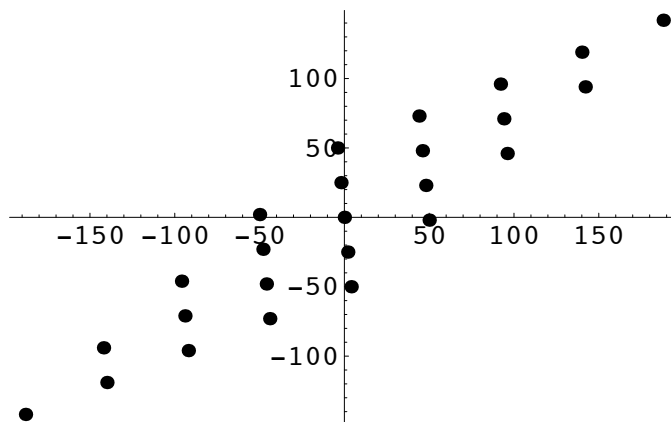


Figure 6.1: A Randomly Generated Lattice.

The following theorem provides some of the most important properties of reduced bases.

Theorem 6.8 *Let $B = (b_1, \dots, b_n)$ be a reduced basis of a full-dimensional lattice L and $B^* = (b_1^*, b_2^*, \dots, b_n^*)$ the orthogonal basis by the GSO. The following statements hold.*

- (a) $\|b_{i+1}^*\|^2 \geq 1/2\|b_i^*\|^2$,
- (b) $\|b_1\| \leq 2^{(n-1)/4}(\det L)^{1/n}$,
- (c) $\|b_1\| \leq 2^{(n-1)/2} \min\{\|b\| : b \in L \setminus \{\mathbf{0}\}\}$,
- (d) $\alpha(B) \leq 2^{n(n-1)/4} \det L$.

Proof.

- (a) Because b_j^* 's are orthogonal, by (6.11), we have

$$\|b_{i+1}^* + \mu_{i+1,i} b_i^*\|^2 = \|b_{i+1}^*\|^2 + \mu_{i+1,i}^2 \|b_i^*\|^2 \geq 3/4 \|b_i^*\|^2.$$

By (6.10), we have $\mu_{i+1,i}^2 \leq 1/4$, and thus $\|b_{i+1}^*\|^2 \geq 1/2\|b_i^*\|^2$.

- (b) By (a), we have $\|b_i^*\|^2 \geq 2^{-(i-1)}\|b_1^*\|^2 = 2^{-(i-1)}\|b_1\|^2$. By Lemma 6.5 (b),

$$\det L = \|b_1^*\| \|b_2^*\| \cdots \|b_n^*\| \geq \prod_{i=1}^n 2^{-(i-1)/2} \|b_1\| = 2^{-n(n-1)/4} \|b_1\|^n.$$

Hence, we have $\|b_1\| \leq 2^{(n-1)/4}(\det L)^{1/n}$.

- (c) By the proof of (b), we have $\|b_i^*\| \geq 2^{-(n-1)/2}\|b_1\|$. By Lemma 6.6, for any nonzero vector b in L ,

$$\|b\| \geq \min\{\|b_1^*\|, \|b_2^*\|, \dots, \|b_n^*\|\} \geq 2^{-(n-1)/2}\|b_1\|.$$

- (d) Because $b_i = \sum_{j=1}^i \mu_{ij} b_j^*$ with $(\mu_{ii} = 1)$, by (6.10) and by (a), we have

$$\|b_i\|^2 = \sum_{j=1}^i \mu_{ij}^2 \|b_j^*\|^2 \leq \|b_i^*\|^2 + \frac{1}{4} \sum_{j=1}^{i-1} \|b_j^*\|^2 \quad (6.12)$$

$$\leq \|b_i^*\|^2 + \frac{1}{4} \sum_{j=1}^{i-1} 2^{i-j} \|b_i^*\|^2 \quad (6.13)$$

$$\leq 2^{i-1} \|b_i^*\|^2. \quad (6.14)$$

Multiplying all $\|b_i^*\|^2$ together for $i = 1, \dots, n$, we obtain

$$\alpha(B)^2 = \|b_1\|^2 \cdots \|b_n\|^2 \leq 2^{n(n-1)/2} \|b_1^*\|^2 \cdots \|b_n^*\|^2 = 2^{n(n-1)/2} (\det L)^2. \quad (6.15)$$

This completes the proof. ■

Theorem 6.9 *There is a polynomial-time algorithm to find a reduced basis $B = (b_1, \dots, b_n)$ of the lattice $L(A)$, for a given rational nonsingular matrix $A \in \mathbb{Q}^{n \times n}$.*

We do not give a proof of Theorem 6.9, because it is quite technical. However, we briefly present the polynomial-time algorithm due to Lenstra, Lenstra and Lovász (1982).

Without loss of generality we assume that the input matrix A is integral. Set $B := A$. The algorithm consists of following two stages (I) and (II).

- (I) Compute the orthogonal basis B^* of B using the GSO. For each $i = 1, \dots, n$ (in this order), and replace b_i by $b_i - \sum_{j=1}^{i-1} \lceil \mu_{ij} \rceil b_j$, where μ_{ij} is as defined in Lemma 6.5, and $\lceil x \rceil$ denotes the (smaller) nearest integer to x . Just to clarify this notation, $\lceil 2.5 \rceil = 2$ and $\lceil 2.51 \rceil = 3$. Note that the GSO basis B^* stays unchanged because of the fact that the subtraction of any linear combination of b_1, \dots, b_{i-1} from b_i has no effect on the GSO result. One can show that at the end of this stage, the new μ_{ij} 's satisfy $|\mu_{ij}| \leq 1/2$ and thus the first condition of reduced basis (6.10).
- (II) Now we only have to worry about the second condition (6.11). For that, find any index i such that the condition (6.11) is violated. If no such index exists, stop. Otherwise, swap b_i and b_{i+1} , and go back to the stage (I).

It can be shown that this algorithm terminates in polynomial-time.

6.3 Applications of Lattice Reduction

Theorem 6.10 (Shortest Vector in Fixed Dimension) *For fixed n , there is a polynomial-time algorithm to find a shortest vector in $L(A)$ for a given rational nonsingular matrix $A \in \mathbb{Q}^{n \times n}$.*

Proof. This follows directly from Theorem 6.7 and Theorem 6.9. Namely, once a reduced basis B is computed, one only needs to enumerate all vectors $x \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $|x_j| \leq 2^{n(n-1)/4}$ for all j , and take a vector Bx with the smallest $\|Bx\|$. ■

Theorem 6.11 (Approximate Closest Vector) *There is a polynomial-time algorithm, for given rational nonsingular matrix $A \in \mathbb{Q}^{n \times n}$ and a given rational vector $b \in \mathbb{Q}^n$, to compute a vector $y \in L(A)$ such that*

$$\|b - y\| \leq 2^{(n/2)} \min\{\|b - v\| : v \in L(A)\}. \quad (6.16)$$

Proof. Let $A \in \mathbb{Q}^{n \times n}$, and let $b \in \mathbb{Q}^n$. First we find a reduced basis $B = (b_1, \dots, b_n)$ of $L(A)$ and compute the GSO basis $B^* = (b_1^*, b_2^*, \dots, b_n^*)$. Then, we find a lattice vector $y \in L(A)$ such that

$$b - y = \sum_{i=1}^n \mu_i b_i^* \quad \text{with } |\mu_i| \leq \frac{1}{2} \quad \text{for all } i = 1, \dots, n. \quad (6.17)$$

To find such y , use the fact that B^* is a basis, $b = \sum_{i=1}^n \lambda_i^0 b_i^*$ for some λ^0 . Then,

$$b - \lceil \lambda_n^0 \rceil b_n = \sum_{i=1}^n \lambda_i^1 b_i^*, \quad (6.18)$$

for some λ^1 with $|\lambda_n^1| \leq 1/2$. Continuing with $(n-1)$ st component by subtracting $\lceil \lambda_{n-1}^1 \rceil b_{n-1}$ from both sides, we obtain λ^2 with $|\lambda_i^2| \leq 1/2$ for $i = n-1, n$, and do this all the way down to 1st component to obtain λ^n . By setting $y := \sum_{i=1}^n \lceil \lambda_i^{n-i} \rceil b_i$ and $\mu := \lambda^n$, we have a lattice vector y we are looking for.

Now we claim that such a vector y satisfies the condition (6.16), which completes the proof. To prove the claim, let z be any vector in $L(A) \setminus \{y\}$. We can write $b - z = \sum_{i=1}^n \zeta_i b_i^*$ for some ζ_i 's. Let k be the largest index such that $\mu_i \neq \zeta_i$. Since both y and z are lattice points, $z - y = \sum_{i=1}^k (\zeta_i - \mu_i) b_i^*$ is also a lattice point. This means that $(\zeta_k - \mu_k)$ is integer by Lemma 6.5 (c). Because $|\mu_k| \leq 1/2$, we have $|\zeta_k| \geq 1/2$. Therefore,

$$\|b - z\|^2 \geq \sum_{i=k+1}^n \zeta_i^2 \|b_i^*\|^2 + \frac{1}{4} \|b_k^*\|^2 = \sum_{i=k+1}^n \mu_i^2 \|b_i^*\|^2 + \frac{1}{4} \|b_k^*\|^2. \quad (6.19)$$

On the other hand,

$$\begin{aligned} \|b - y\|^2 &\leq \sum_{i=k+1}^n \mu_i^2 \|b_i^*\|^2 + \frac{1}{4} \sum_{i=1}^k \|b_i^*\|^2 \\ &\leq \sum_{i=k+1}^n \mu_i^2 \|b_i^*\|^2 + \frac{1}{4} \|b_k^*\|^2 \left(\sum_{i=1}^k 2^{k-i} \right) \quad (\text{by Theorem 6.8 (a)}) \\ &< 2^k \|b - z\|^2 \leq 2^n \|b - z\|^2. \end{aligned}$$

■

Theorem 6.12 (Simultaneous Diophantine Approximation) *There is a polynomial-time algorithm, for given rational numbers a_1, \dots, a_n and $0 < \epsilon < 1$, to compute integers p_1, \dots, p_n and q such that*

$$1 \leq q \leq 2^{n(n+1)/4} \epsilon^{-n} \quad \text{and} \quad |qa_i - p_i| < \epsilon \quad \forall i = 1, \dots, n. \quad (6.20)$$

Proof. Let a_1, \dots, a_n be rational numbers and $0 < \epsilon < 1$. Let e_i be the i th unit vector in \mathbb{R}^{n+1} , and let $a = (a_1, \dots, a_n, 2^{-n(n+1)/4} \epsilon^{n+1})^T$. Let L be the lattice generated by $[e_1, e_2, \dots, e_n, a]$. It follows that $\det L = 2^{-n(n+1)/4} \epsilon^{n+1}$. We can compute a reduced basis $B = (b_1, \dots, b_n)$ of L in polynomial-time. By Theorem 6.8 (b),

$$\|b_1\| \leq 2^{n/4} (\det L)^{1/(n+1)} = \epsilon. \quad (6.21)$$

The vector b_1 can be represented as $p_1 e_1 + \dots + p_n e_n - qa$ for some integer p_i 's and q . Since $\epsilon < 1$, $q \neq 0$ and we may assume $q > 0$. By (6.21), we have

$$|p_i - qa_i| < \epsilon, \quad (6.22)$$

$$2^{-n(n+1)/4} \epsilon^{n+1} q \leq \epsilon, \quad \text{equivalently, } q \leq 2^{n(n+1)/4} \epsilon^{-n}. \quad (6.23)$$

This completes the proof. ■

Compare this result with Dirichlet's non-constructive theorem, Theorem 6.2. In particular, there is a huge extra factor $2^{n(n+1)/4}$ for the upper bound for q . Yet, this constructive theorem is strong enough for various applications.