A Note on Floor Computation

Junekey Jeon

July 24, 2024

1 Introduction

This note is an informal writing about an algorithm I developed (and its implementation I wrote in C++) to find the set of pairs $(\xi, \zeta) \in \mathbb{R}^2$ such that the equality

$$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$$

holds for all n in a given interval of integers, where x, y are arbitrarily given real numbers. The intent is that we want to compute $\lfloor nx + y \rfloor$ for some given x, y, but since x, y may be very complicated numbers, its direct computation might be infeasible, so we want to replace x, y by some ξ, ζ specifically chosen to allow us fast computation of $\lfloor n\xi + \zeta \rfloor$. In practice, preferred choices of ξ, ζ would be of the form $\xi = \frac{m}{2^k}$ and $\zeta = \frac{s}{2^k}$ for some integers k, m, s, so that

$$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor = \left\lfloor \frac{nm + s}{2^k} \right\rfloor$$

can be computed by a multiplication, an addition, and a bit-shift of integers.

One of the motivating examples is the problem of fast computation of integer division by a constant divisor. Note that, by setting x, y into appropriate rational numbers, the quotient of the division of $na+b$ by d for some fixed integers a, b, d can be written as $\lfloor nx + y \rfloor$. In modern CPU's, integer division typically has much worse performace characteristics compared to other basic integer arithmetic operations. In this reason, techniques for replacing divisions by multiply-shift's or multiply-add-shift's have been explored at least since 90's, and some of these techniques already have been widely applied into modern optimizing compilers. Some of the well-known earlier works on these techniques include [1], [2] and [3].

However, to my knowledge, there have been not so many formal and rigorous derivations (other than simple exhaustive searches) of sufficient conditions on (ξ, ζ) to allow the equality $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for the case when y is nonzero although it seems to be well-known for long time that a good enough binary approximation (ξ, ζ) of (x, y) should give the identity $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$. For instance, such an observation has already been applied in the implementation of Schubfach algorithm [4], for computing $\left\lfloor n \log_{10} 2 - \log_{10} \frac{4}{3} \right\rfloor$ as one of its substeps.

Yet, somewhat counterintuitively, it turns out that the best choice of $(\xi, \zeta) = \left(\frac{m}{2^k}, \frac{s}{2^k}\right)$ as a replacement for (x, y), that is, the one that allows the equality $\lfloor nx + y \rfloor = \left\lfloor \frac{nm+s}{2^k} \right\rfloor$ to hold on the widest range of n, or the one using the fewest number of bits when the range of n is given, is *not* the one that is closest to the true value of (x, y) among candidates with the same number of bits. It can in fact be quite far from the closest approximation, and this is because the errors in the approximations of x and of y are not just added, rather they can sometimes cancel each other. However, to my knowledge this cancellation phenomenon has not yet been explored in detail, and there is no known simple intuitive way of estimating what would be the optimal choice for m and s.

In the scenario of computing $\left\lfloor \frac{na+b}{d} \right\rfloor$, the case $y \neq 0$ corresponds to the case $b \neq 0$. That is, we want to optimize at once the whole sequence of a multiplication, an addition and a division done consecutively. Note that the classical technique of turning division into multiply-shift can be still applied here so that the computation of $na + b$ divided by d can be done in a multiplication, an addition, another multiplication and a shift. Still, if we can find (ξ, ζ) of the form $\left(\frac{m}{2^k}, \frac{s}{2^k}\right)$, then we can omit the second multiplication.

Even though in practice it is often enough to have a *sufficient* condition on ξ, ζ, the algorithm presented in this note actually computes the precise set of $(\xi, \zeta) \in \mathbb{R}^2$ that satisfies $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for any given real numbers x, y and the interval of n's. That is, it derives a *necessary and sufficient* condition, which is the best possible outcome.

Throughout this note, we let

$$[n_{\min} \colon n_{\max}] \coloneqq [n_{\min}, n_{\max}] \cap \mathbb{Z}$$

for any $n_{\min}, n_{\max} \in \mathbb{Z}$.

First of all, note that having $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for all $n \in [n_{\min} \colon n_{\max}]$ is equivalent to having

$$\max_{n \in [n_{\min} \colon n_{\max}] \cap \mathbb{Z}_{>0}} \frac{\lfloor nx + y \rfloor - \zeta}{n} \leq \xi < \min_{n \in [n_{\min} \colon n_{\max}] \cap \mathbb{Z}_{>0}} \frac{\lfloor nx + y \rfloor + 1 - \zeta}{n},$$

$$\max_{n \in [n_{\min} \colon n_{\max}] \cap \mathbb{Z}_{<0}} \frac{\lfloor nx + y \rfloor + 1 - \zeta}{n} < \xi \leq \min_{n \in [n_{\min} \colon n_{\max}] \cap \mathbb{Z}_{<0}} \frac{\lfloor nx + y \rfloor - \zeta}{n},$$

possibly along with $\lfloor y \rfloor = \lfloor \zeta \rfloor$ if $0 \in [n_{\min} \colon n_{\max}]$, so the problem boils down to the procedure of computing the maximum and the minimum in the above inequalities. Of course, we will not have a closed-form expressions for them, but we will derive an efficient algorithm for computing them. In practical applications, we are often interested in the case when the size of the interval $[n_{\min} \colon n_{\max}]$ is very large, so the exhaustive search may not be desired or even impossible. On the other hand, although I did not try to rigorously prove, the algorithm presented in this note probably has the time complexity of only some polynomial of $\log N$ where N is the size of the interval $[n_{\min} \colon n_{\max}]$. More importantly, the reference implementation I wrote in C++ has no problem running almost instantaneously with $N \sim 2^{64}$ on my laptop I bought 6 years ago (2018) even in the non-optimized debug build.

Before delving into the details, let me note that every algorithm described in this note can run on arbitrary-precision exact rational arithmetic, even though x, y are assumed to be arbitrary real numbers (represented by their *continued fraction expansions*; see Section 3 for details). In particular, there is no floating-point operations involved, and there is no worry of rounding giving imprecise results. Of course, arbitrary-precision exact rational arithmetic can be very slow compared to other alternatives (e.g. floating-point arithmetic), but squeezing the last bit of performance is not our goal since computation of (ξ, ζ) is done as a preparation step (that needs to be done only once) for computing $\lfloor nx + y \rfloor$. The performance may still be a problem if the running time of the algorithm is too long, but as I remarked in the previous paragraph, this is definitely not the case.

I also remark that, although the theory behind the main algorithm we develop does not care too much about rationality of x and y, actual implementation of the algorithm taking possibly irrational x, y as its inputs is fairly more complicated than the implementation only taking rational inputs. The reference implementation I wrote in C++ does allow irrational inputs, but a much simpler implementation should be possible if one cares only about rational inputs. Readers who only care about the rational case can skip reading Theorem 7.5, Section 8, Section 10.3, Section 12 and Section 13.

A paper by Drane et al [5] gives a geometric intuition on the condition that ξ, ζ should satisfy. Note that the equality $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ can be equivalently written as the inequality

$$n(x - \xi) + (y - \zeta) \leq (nx + y) - \lfloor nx + y \rfloor < n(x - \xi) + 1 + (y - \zeta).$$

Note that the left-hand side and the right-hand side are affine functions of n with the same slope $x - \xi$, so what we want to have is to make sure that the "sawtooth function" in the middle is in between two lines, as illustrated in Figure 1.

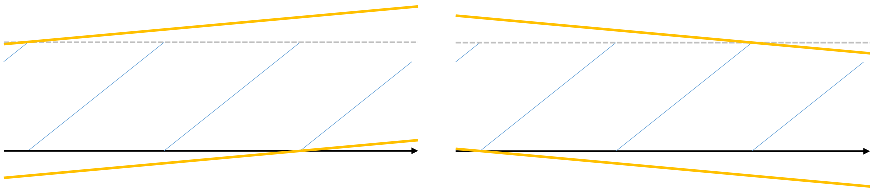

Figure 1: Left: when $x > \xi$, Right: when $x < \xi$

Of course, if n really varied over all real numbers between $n_{\min}$ and $n_{\max}$, then except for fairly special cases, placing the sawtooth function in between two lines is simply impossible unless $x = \xi$, because the height of the sawtooth function is precisely equal to the vertical distance between the lines. However, if n is restricted to integers, the sawtooth function may not actually achieve its top or bottom corners, which leaves a little bit of room for the lines to pass through. This observation leads to the intuition that the maximum and the minimum of the fractional part $(nx + y) - \lfloor nx + y \rfloor$ will play a significant role, because they describe how large these gaps are. Indeed, computation of extremizers of $(nx+y) - \lfloor nx + y \rfloor$ is one of the very first steps in the final algorithm we present in Section 11. Therefore, after developing some fundamental theoretical backgrounds in Section 2, Section 3 and Section 4, we will first describe an algorithm for computing the extremizers of $(nx + y) - \lfloor nx + y \rfloor$ in Section 5.

2 Best rational approximations from below and above

Before moving into the general case, let us first examine a special case when $y = 0$ and $n_{\min} = 1$, that is, we are concerned with the extremizers of $nx - \lfloor nx \rfloor$ when $n = 1, \cdots, n_{\max}$. This special case in fact is the most fundamental building block for solving the general case. First, we reformulate this problem with the following terminology.

Definition 2.1 (Best rational approximations from below/above)**.**
Let x be a real number and $\frac{p}{q}$ a rational number in its reduced form. Then we say $\frac{p}{q}$ is:

1. a *best rational approximation* of x *from below (above, resp.) in the weak sense*, if $\frac{p}{q} \leq x$ ($\frac{p}{q} \geq x$, resp.) holds and for any rational number $\frac{a}{b}$ with $0 < b \leq q$ and $\frac{a}{b} \leq x$ ($\frac{a}{b} \geq x$, resp.), $\frac{a}{b} \leq \frac{p}{q} \leq x$ ($\frac{a}{b} \geq \frac{p}{q} \geq x$, resp.) holds,

2. a *best rational approximation* of x *from below (above, resp.) in the strong sense*, if $\frac{p}{q} \leq x$ ($\frac{p}{q} \geq x$, resp.) holds and for any rational number $\frac{a}{b}$ with $0 < b \leq q$ and $\frac{a}{b} \leq x$ ($\frac{a}{b} \geq x$, resp.), $bx - a \geq qx - p \geq 0$ ($a - bx \geq p - qx \geq 0$, resp.) holds.

Remark 2.2.
Best rational approximation is a standard terminology in mathematical literatures, but they usually use this term to denote a related but different concept: $\frac{p}{q}$ is said to be a best rational approximation of x if it is a best approximation *regardless of the direction* of approximation, from below or above. For example, $\frac{1}{3}$ is a best rational approximation from below of $\frac{3}{7}$ (in the weak sense), but it is not a best rational approximation in the usual, non-directional sense, because $\frac{1}{2}$ is a better approximation with a strictly less denominator. But this concept of non-directional best rational approximations is quite irrelevant to our application, so any usage of the term *best rational approximation* in this note always means the directional ones, either from below or above.

Of course, the terminology is set in the way that "strong" implies "weak".

Lemma 2.3.
Let x be a real number. If a rational number $\frac{p}{q}$ in its reduced form is a best rational approximation of x from below (above, resp.) in the strong sense, then it is a best rational approximation of x from below (above, resp.) in the weak sense as well.

Proof. We only show the statement about best rational approximations from below. The corresponding statement for best rational approximations from above can be done in a similar manner.

Suppose $\frac{p}{q}$ is a best rational approximation of x from below in the strong sense. It is then enough to show that for any rational number $\frac{a}{b}$ with $0 < b \leq q$ and $\frac{a}{b} \leq x$, we must have $\frac{a}{b} \leq \frac{p}{q}$. This follows directly from the assumptions:

$$x - \frac{a}{b} = \frac{bx - a}{b} \geq \frac{qx - p}{b} \geq \frac{qx - p}{q} = x - \frac{p}{q},$$

thus we conclude $\frac{a}{b} \leq \frac{p}{q}$. □

Actually, it turns out that the converse is also true, that is, two concepts are in fact equivalent to each other. This is a very useful fact but unfortunately not entirely trivial as far as I know. We will prove it in Section 4 after developing a sufficient amount of the theory of *continued fractions* in Section 3 and Section 4. Before we prove the equivalence, we will distinguish the weak and the strong notions of best rational approximations from below/above.

Clearly, if $\frac{p}{q}$ is a best rational approximation of x from below, either in the weak or the strong sense, then we must have $p = \lfloor qx \rfloor$, because $p > \lfloor qx \rfloor$ implies $\frac{p}{q} > x$ and $p < \lfloor qx \rfloor$ implies $\frac{p}{q} < \frac{p+1}{q} \leq x$, which both contradict to the assumption. Similarly, if $\frac{p}{q}$ is a best rational approximation of x from above, then we must have $p = \lceil qx \rceil$. From this simple observation, we obtain the following result.

Theorem 2.4 (Extremizers of $nx - \lfloor nx \rfloor$)**.**
Let x be a real number and $n_{\max}$ a positive integer.

1. *If x is rational and its reduced form $\frac{p}{q}$ satisfies $q \leq n_{\max}$, then*

$$\operatorname*{argmin}_{n \in [1:\, n_{\max}]} (nx - \lfloor nx \rfloor) = \{kq \colon k \in \mathbb{Z}_{>0},\ kq \leq n_{\max}\}$$

and

$$\operatorname*{argmax}_{n \in [1:\, n_{\max}]} (nx - \lfloor nx \rfloor) = \{kq - b \colon k \in \mathbb{Z}_{>0},\ kq - b \leq n_{\max}\},$$

where $b \in [0\colon q-1]$ is the modular inverse of p with respect to q.

2. *If x is either irrational or is rational whose reduced denominator is strictly larger than $n_{\max}$, then*
$$\operatorname*{argmin}_{n\in[1:\,n_{\max}]}(nx-\lfloor nx\rfloor)=\{q_*\}$$
and
$$\operatorname*{argmax}_{n\in[1:\,n_{\max}]}(nx-\lfloor nx\rfloor)=\{q^*\}\,,$$
where $\frac{p_*}{q_*}$ and $\frac{p^*}{q^*}$ are best rational approximations in their reduced forms of x from below and above, respectively, in the strong sense, with the largest (reduced) denominators $q_*, q^* \leq n_{\max}$.

Proof. 1. Clearly,
$$nx-\lfloor nx\rfloor=\frac{1}{q}\left(np-\left\lfloor\frac{np}{q}\right\rfloor q\right)=\frac{(np \bmod q)}{q},$$
thus the minimizers of $nx-\lfloor nx\rfloor$ are precisely the multipliers of q. Similarly, the maximizers are precisely the n's such that the remainder of np divided by q is $q-1$. Those n's are precisely the ones with $n\equiv -b \pmod q$, so we arrive at the desired conclusion.

2. Take any $n_0\in\operatorname{argmin}_{n\in[1:\,n_{\max}]}(nx-\lfloor nx\rfloor)$. We claim that $\frac{\lfloor n_0x\rfloor}{n_0}$ is a best rational approximation of x from below in the strong sense and is in its reduced form. Let $\frac{a}{b}$ be any rational number in its reduced form such that $\frac{a}{b}\leq x$ and $b\leq q$. Then since $a\leq bx$ implies $a\leq\lfloor bx\rfloor$, we have
$$bx-a\geq bx-\lfloor bx\rfloor\geq n_0x-\lfloor n_0x\rfloor\geq 0.$$
since n_0 is a minimizer of $nx-\lfloor nx\rfloor$. In particular, if $\frac{a}{b}$ is the reduced form of $\frac{\lfloor n_0x\rfloor}{n_0}$, then $n_0x-\lfloor n_0x\rfloor=k(bx-a)$ for some positive integer k, so we must have either $n_0x=\lfloor n_0x\rfloor$ or $k=1$. The first case is impossible by the assumption on x, and the claim follows.

Next, we claim $n_0=q_*$, the largest denominator of all best rational approximations of x from below in the strong sense whose denominators are no more than n_0. Indeed, since $\frac{p_*}{q_*}$ is a best rational approximation of x from below in the strong sense and since $n_0\leq q_*$ by definition of q_*, we must have
$$n_0x-\lfloor n_0x\rfloor\geq q_*x-\lfloor q_*x\rfloor=q_*x-p_*\geq 0.$$
However, since n_0 is a minimizer of $nx-\lfloor nx\rfloor$, the above should be an equality, thus we obtain
$$(q_*-n_0)x=\lfloor q_*x\rfloor-\lfloor n_0x\rfloor\,.$$
Since the right-hand side is an integer, the assumption on x implies $q_*=n_0$.

Next, we show the statement about the maximizer. First, note that since nx is never an integer, we have $\lceil nx\rceil=\lfloor nx\rfloor+1$ for all $n\in[1:\,n_{\max}]$. Hence,
$$\begin{aligned}\operatorname*{argmax}_{n\in[1:\,n_{\max}]}(nx-\lfloor nx\rfloor)&=\operatorname*{argmin}_{n\in[1:\,n_{\max}]}(\lfloor nx\rfloor-nx)=\operatorname*{argmin}_{n\in[1:\,n_{\max}]}(\lceil nx\rceil-nx)\\&=\operatorname*{argmin}_{n\in[1:\,n_{\max}]}(n(-x)-\lfloor n(-x)\rfloor)\,,\end{aligned}$$

thus by what we have just shown, $\mathrm{argmax}_{n\in[1:\,n_{\max}]}\left(nx-\lfloor nx\rfloor\right)$ must be the singleton set whose unique element is the denominator of the best rational approximation $\frac{p}{q}$ of $-x$ from below in the strong sense with the largest denominator $q \leq n_{\max}$. Then it easily follows from the definition that $\frac{-p}{q}$ is a best rational approximation of x from above in the strong sense with the largest denominator $q \leq n_{\max}$, so we conclude that $q = q^*$ is the unique element in $\mathrm{argmax}_{n\in[1:\,n_{\max}]}\left(nx-\lfloor nx\rfloor\right)$.

□

Note that when $x \notin \mathbb{Z}$, b appearing in the first case is precisely the largest denominator of all best rational approximations of x from below in the strong sense with denominators at most $q-1$, because clearly it is the unique minimizer of $nx - \lfloor nx\rfloor$ for $n \in [1\colon q-1]$. (Similarly, $q-b$ is the largest denominator of all best rational approximations of x from above in the strong sense with denominators at most $q-1$.) Hence, in any case, the problem of finding the extremizers of $nx - \lfloor nx\rfloor$ is equivalent to the problem of finding best rational approximations of x from below/above in the strong sense.

In Section 4, we will derive an efficient algorithm for computing the best rational approximations of x from below/above with the largest denominators bounded by given $n_{\max}$, using the theory of *continued fractions*.

3 Continued fractions

The problem of finding best rational approximations of x from below/above can be solved using the *continued fraction expansion* of x, so let us take a bit of detour and introduce some basic theory of continued fractions first. Many of what follows in this section and the next section are directly copied and only slightly modified from my previous paper on a floating-point formatting algorithm, *Dragonbox* [6], though there are some added materials too. Those from [6] are reproduced here to make this note more self-contained. Contents in this section specifically is standard in the literature, and can be found, for example in [7].

A *continued fraction* means either a finite or infinite sequence of the form

$$a_0,\ a_0 + \frac{b_1}{a_1},\ a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2}},\ a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3}}},\ \cdots .$$

The ith term appearing in the sequence above is called the ith *convergent* of the continued fraction.

Lemma 3.1.
For each $i \geq 0$, inductively define

$$\begin{cases} p_i = b_i p_{i-2} + a_i p_{i-1}, \\ q_i = b_i q_{i-2} + a_i q_{i-1}, \end{cases}$$

where $b_0 \coloneqq 1$, $p_{-1}, q_{-2} \coloneqq 1$ and $q_{-1}, p_{-2} \coloneqq 0$. Then

$$\begin{pmatrix} p_i & p_{i-1} \\ q_i & q_{i-1} \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ b_0 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ b_1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix}$$

holds and $\frac{p_i}{q_i}$ is equal to the ith convergent.

Proof. The recurrence relation and the initial condition immediately gives us

$$\begin{pmatrix} p_i & p_{i-1} \\ q_i & q_{i-1} \end{pmatrix} = \begin{pmatrix} p_{i-1} & p_{i-2} \\ q_{i-1} & q_{i-2} \end{pmatrix} \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix},$$

which then implies the desired matrix identity at once. To show that $\frac{p_i}{q_i}$ is equal to the ith convergent, we use induction on i. For $i = 0$, the conclusion is trivial. Suppose the result holds for some $i \geq 0$, then by applying the induction hypothesis to the continued fraction we obtain by replacing (a_j, b_j) by (a_{j+1}, b_{j+1}) for each j, we conclude that the ith convergent should be equal to

$$a_0 + \frac{b_1}{r_{i-1}/s_{i-1}} = \frac{a_0 r_{i-1} + b_1 s_{i-1}}{r_{i-1}}$$

where $\left((r_j, s_j)\right)_j$ is the sequence defined by the recurrence relation

$$\begin{cases} r_j = b'_{j+1} r_{i-2} + a_{i+1} r_{j-1}, \\ s_j = b'_{j+1} s_{i-2} + a_{i+1} s_{j-1} \end{cases}$$

with the same initial condition as $\left((p_j, q_j)\right)_j$, where $b'_j := b_j$ for all $j > 1$ and $b'_1 := 1$. Then we have

$$\begin{pmatrix} r_{i-1} & r_{i-2} \\ s_{i-1} & s_{i-2} \end{pmatrix} = \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_2 & 1 \\ b_2 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix}.$$

Hence, multiplying

$$\begin{pmatrix} a_0 & 1 \\ b_0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & b_1 \end{pmatrix} = \begin{pmatrix} a_0 & b_1 \\ 1 & 0 \end{pmatrix}$$

to both sides from the left shows that

$$\begin{aligned} \begin{pmatrix} a_0 & b_1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r_{i-1} & r_{i-2} \\ s_{i-1} & s_{i-2} \end{pmatrix} &= \begin{pmatrix} a_0 & 1 \\ b_0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & b_1 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_2 & 1 \\ b_2 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix} \\ &= \begin{pmatrix} a_0 & 1 \\ b_0 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ b_1 & 0 \end{pmatrix} \begin{pmatrix} a_2 & 1 \\ b_2 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix} = \begin{pmatrix} p_i & p_{i-1} \\ q_i & q_{i-1} \end{pmatrix}, \end{aligned}$$

which gives the desired conclusions $p_i = a_0 r_{i-1} + b_1 s_{i-1}$ and $q_i = r_{i-1}$. □

Corollary 3.2.
With the notations from Lemma 3.1,

$$p_i q_{i-1} - p_{i-1} q_i = \prod_{j=0}^{i} (-b_j) \tag{1}$$

and

$$\frac{p_i}{q_i} - \frac{p_{i-1}}{q_{i-1}} = \frac{\prod_{j=0}^{i}(-b_j)}{q_{i-1} q_i}$$

hold for each i.

Proof. The first claim follows from the matrix identity of Lemma 3.1 and $\det \begin{pmatrix} a_i & 1 \\ b_i & 0 \end{pmatrix} = -b_i$. The second claim is a trivial consequence of the first claim. □

In general, we can consider the case when a_i's and b_i's are whatever objects that can make sense of the entries of the continued fraction, but we are mostly concerned with the case when following further assumptions hold:

- all b_i's are equal to 1, and
- all a_i's except for a_0 are positive integers, and
- a_0 is an integer.

In such a case, let us call the continued fraction *regular*.[1]

For any given finite or infinite sequence $a_0, a_1, a_2, \cdots$ of integers with $a_i \geq 1$ for all $i > 0$, there uniquely exists a regular continued fraction which we formally denote as

$$[a_0; a_1, a_2, \cdots] \coloneqq a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots}}}.$$

Here, by *formal* we mean that the right-hand side is just a fancy way of referring to the sequence of convergents, and there is no issue of actual convergence involved here.

However, we will show below that the sequence actually does converge to a real number, thus we make no distinction between the sequence $[a_0; a_1, a_2, \cdots]$ with its limit, although this is strictly speaking an abuse of notation. In particular, for each i such that $a_0, \cdots, a_i$ are well-defined, the ith convergent is denoted as $[a_0; a_1, \cdots, a_i]$. For example, the 5th convergent of $[0; 1, 2, 3, 4, \cdots]$ is denoted as $[0; 1, 2, 3, 4, 5]$ which is equal to

$$0 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{3 + \cfrac{1}{4 + \cfrac{1}{5}}}}} = \frac{157}{225}.$$

Let us now define the sequence $\left(\frac{p_i}{q_i}\right)_i$ as in Lemma 3.1, that is,

$$\begin{cases} p_i = p_{i-2} + a_i p_{i-1}, \\ q_i = q_{i-2} + a_i q_{i-1} \end{cases}$$

with $p_{-1}, q_{-2} \coloneqq 1$ and $q_{-1}, p_{-2} \coloneqq 0$, so that $\frac{p_i}{q_i}$ is equal to the ith convergent. Specializing Corollary 3.2 then gives

$$p_i q_{i-1} - p_{i-1} q_i = (-1)^{i+1} \tag{2}$$

and

$$\frac{p_i}{q_i} - \frac{p_{i-1}}{q_{i-1}} = \frac{(-1)^{i+1}}{q_i q_{i-1}}.$$

[1] It seems terminologies for these concepts are somewhat hazy in the literature. Some refers to regular continued fractions as *simple* continued fractions, and some other refers to them as *standard* continued fractions. In fact, in many of the literatures they just reserve the term *continued fractions* only for the regular ones. In this note, non-regular continued fractions do appear, but only in the context of computing logarithms in Section 13, so we will use the term *continued fractions* to refer to the general case, but we will also mostly only consider the regular ones.

From (2), we deduce that p_i and q_i are coprime, so $\frac{p_i}{q_i}$ is in fact the reduced form of the convergent. (Note that q_i is always strictly positive because $q_{-1} = 0$ and $a_i > 0$ for all $i > 0$.) Also, inductively applying the above formula gives

$$\frac{p_i}{q_i} = a_0 + \sum_{j=0}^{i-1} \frac{(-1)^j}{q_j q_{j+1}}.$$

Since the recurrence relation and positivity of a_i's guarantee at least exponential growth of q_i's, we obtain:

Corollary 3.3.
We have

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \cdots < \frac{p_3}{q_3} < \frac{p_1}{q_1}.$$

Also, if the sequence $(a_i)_i$ is infinite, then the sequence $\left(\frac{p_i}{q_i}\right)_{i=0}^{\infty}$ of convergents converges to a real number.

Conversely, for a given real number x, there is a well-known algorithm for obtaining a continued fraction converging to x. First, take the integer part $a_0 := \lfloor x \rfloor$. Then $x - a_0 \in [0, 1)$. If it is zero, then that means x is an integer, and we terminate the procedure and returns the sequence (a_0) consisting of the single term a_0. If not, then consider the reciprocal $x_1 := \frac{1}{x-a_0} \in (1, \infty)$. Take the integer part $a_1 := \lfloor x_1 \rfloor$, then again we have $x_1 - a_1 \in [0, 1)$. If it is zero, then stop and returns the sequence $(a_0, [a_0; a_1])$ consisting of two terms. Otherwise, take the reciprocal and let $x_2 := \frac{1}{x_1-a_1} \in (1, \infty)$. Again, take the integer part $a_2 := \lfloor x_2 \rfloor$, and if $x_2 - a_2 = 0$, then stop and returns the sequence $(a_0, [a_0; a_1], [a_0; a_1, a_2])$ consisting of three terms. Otherwise, take the reciprocal $x_3 := \frac{1}{x_2-a_2}$ and continue.

To see why the resulting continued fractions converge to x, it is sufficient to see $\frac{p_{2i}}{q_{2i}} \le x$ and $x \le \frac{p_{2i+1}}{q_{2i+1}}$ always holds, because of Corollary 3.3. Note that by induction one can easily see

$$x = [a_0; a_1, \cdots, a_i, x_{i+1}]$$

always holds (here we are allowing non-integers to appear in continued fractions), so it is enough to show that for given positive real numbers $a < b$,

$$[a_0; a_1, \cdots, a_{2i}, a] < [a_0; a_1, \cdots, a_{2i}, b]$$

and

$$[a_0; a_1, \cdots, a_{2i+1}, a] > [a_0; a_1, \cdots, a_{2i+1}, b]$$

holds. These can be easily shown using induction.

Note that if this algorithm terminates in a finite step, then that means the final term in the resulting sequence is precisely the number x we are given with. In particular, x must be a rational number in this case, so if x is irrational then the algorithm produces an infinite sequence.

On the other hand, if $x = \frac{p}{q}$ is rational in its reduced form[2], then what the algorithm does is basically what the Euclid algorithm does to p and q to eventually produce $\gcd(p, q) = 1$. (Details are omitted.) With this correspondence, one can see that whenever x is rational the algorithm must terminate in a finite step. In other words, the algorithm terminates in a finite step if and only if x is rational.

[2] That is, q is a positive integer and p is an integer with $\gcd(p, q) = 1$.

Now, suppose we have $x = [a_0; a_1, \cdots]$. Then, it can be easily seen that

$$a_0 \le x < a_0 + 1$$

must hold unless a_1 is the last term and $a_1 = 1$, in which case we have $x = a_0 + 1$. Also, since $[a_0; a_1, \cdots] = [a_0; b_1, \cdots]$ implies $[a_1; a_2, \cdots] = [b_1; b_2, \cdots]$, we get that two continued fractions $[a_0; a_1, \cdots]$ and $[b_0; b_1, \cdots]$ yield the same real number if and only if:

1. They are identical sequences, or
2. Both are finite sequences with their lengths differ by 1, and assuming $[a_0; a_1, \cdots] = [a_0; a_1, \cdots, a_i]$ is the shorter one, $a_j = b_j$ holds for all $j < i$ and $a_i = b_i + 1$, $b_{i+1} = 1$ hold.

In other words,

1. When x is irrational, then there uniquely exists an infinite continued fraction converging to x, which is generated by the algorithm described above, and
2. When x is rational, then along with the one generated by the algorithm, there is another continued fraction which is obtained by replacing the last coefficient a_i with $[a_i - 1; 1]$. Note that whenever the algorithm terminates in a finite step, the last coefficient must be at least 2.

From now on, we will call the continued fraction $[a_0; a_1, \cdots]$ obtained from the algorithm as *the continued fraction expansion of* x, and the ith convergent of it as *the* i*th convergent of* x. We call this $[a_0; a_1, \cdots]$ the *standard* continued fraction expansion of x, if we need to distinguish it from the other expansion, which we call the *alternative* continued fraction expansion.

Given a continued fraction expansion $x = [a_0; a_1, \cdots]$, whenever we say "the ith coefficient" or "the ith convergent" or anything like that, we are always implicitly assuming that those are well-defined. In particular, the index i will be always assumed to be at least 0. However, sometimes it is convenient to talk about "the -1st convergent" which is conventionally defined as ∞. In a similar spirit, we often view "the coefficient next to the last coefficient" to be ∞, and the one next to it to be 0, and the one next to it to be ∞ again, and so on. Then it is natural to conceptually think that, if $\frac{p_i}{q_i}$ is the last convergent (so that $x = \frac{p_i}{q_i}$), then p_{i+1} and q_{i+1} are both ∞, while at the same time the quotient $\frac{p_{i+1}}{q_{i+1}}$ is considered to be equal to $\frac{p_i}{q_i}$. Then, we have $p_{i+2} = p_i$ and $q_{i+2} = q_i$, and then $p_{i+3} = p_{i+1} = \infty$ and $q_{i+3} = q_{i+1} = \infty$ and $\frac{p_{i+3}}{q_{i+3}} = \frac{p_i}{q_i}$, and then again $p_{i+4} = p_i$ and $q_{i+4} = q_i$, and so on.

4 Semiconvergents

In this section, we will show that best rational approximations of x from below/above in the weak sense and in the strong sense are same, by showing that they are equivalent to *proper semiconvergents* of x, which we define below. This immediately gives an efficient algorithm for computing the best rational approximations from below/above.

Definition 4.1 (Semiconvergents)**.**
Let $\frac{p_i}{q_i}$ denote the ith convergent of a given regular continued fraction $x = [a_0; a_1, \cdots]$. Then a rational number $\frac{p}{q}$ in its reduced form is said to be a *semiconvergent* of the continued fraction $[a_0; a_1, \cdots]$, if either

1. $\frac{p}{q} = x$, or
2. $\frac{p}{q} = \frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ for some consecutive convergents $\frac{p_{i-1}}{q_{i-1}}$, $\frac{p_i}{q_i}$ with $i \geq 0$ and an integer $s \in [0, a_{i+1})$, with the convention $a_{i+1} \coloneqq \infty$ if $x = \frac{p_i}{q_i}$.

A semiconvergent $\frac{p}{q}$ is said to be *proper* if either x is irrational or is rational and its reduced denominator is strictly larger than q.[3]

Remark 4.2.
Semiconvergents are meant to be a monotonic sequence interpolating between two consecutive even/odd convergents (see below). However, there seems to be some ambiguity in how exactly they should be defined, regarding various corner cases occuring for finite continued fractions. For instance, the Wikipedia page on continued fractions (retrieved 2024-07-06) defines semiconvergents as any numbers of the form $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ for consecutive convergents $\frac{p_{i-1}}{p_{i-1}}$, $\frac{p_i}{qi}$ and an integer $s \in [0, a_{i+1}]$. However, it is unclear if they intend to include the cases $i = 0$, or when $\frac{p_i}{q_i}$ is the last convergent (there is no mention at all of how they would define a_{i+1} in this case), or if they allow $i = 0$ then whether they want to include ∞, or things like that. Note that, when x is an integer and the standard expansion is used, then the only possible interpretations of their definition is that, either we count every number of the form $x + \frac{1}{s}$ as semiconvergent but *not* x itself, or no number is counted as a semiconvergent. Or maybe they allow $s = \infty$ for the case of last convergent? When x is an integer and the alternative expansion is used, then x is included if and only if they allow $i = 0$, and $x - 1$ is included if and only if they allow the case of last convergent. They do not clarify which exact option they want to take for this case, though it is not even clear if the alternative expansion is allowed at all. When x is not an integer and $\frac{p_1}{q_1}$ is the last convergent (the standard expansion is used), then x itself is only counted as a semiconvergent if they allow $i = 0$, or maybe $s = \infty$. Again it is hard to know which option they wanted to take. Therefore, I tried my best to clarify these ambiguities as much as possible (because these corner cases do impact the correctness of the algorithm we will describe), and arrived at the definition written above. A few notable special cases are:

- If x is rational, then we *always* count x itself as a semiconvergent.
- We *do allow* the case $i = 0$ *and* the case when $\frac{p_i}{q_i}$ is the last convergent. For the latter case, s can be any nonnegative integer. We explicitly do not include the case $s = \infty$, but $x = \frac{p_i}{q_i}$ is always counted as a semiconvergent anyway, which is effectively equivalent to allowing $s = \infty$.
- We *do not* count $\frac{p_{-1}}{q_{-1}} = \infty$ as a semiconvergent, because semiconvergents are a priori defined to be rational numbers. Hence, when $i = 0$, s should be a positive integer.
- On the other hand, we *do count* every convergent $\frac{p_i}{q_i}$ with ≥ 0 as a semiconvergent. Indeed, if $\frac{p_i}{q_i}$ is not the last convergent, then $\frac{p_i}{q_i} = \frac{p_i + 0 \cdot p_{i+1}}{q_i + 0 \cdot q_{i+1}}$ is a semiconvergent, and if it is the last convergent, then $x = \frac{p_i}{q_i}$ so it is still counted as a semiconvergent.
- It may seem like we disallow the case $s = a_{i+1}$, but this has no impact because $\frac{p_{i-1}+a_{i+1}p_i}{q_{i-1}+a_{i+1}q_i} = \frac{p_{i+1}}{q_{i+1}}$ is a semiconvergent whenever $\frac{p_i}{q_i}$ is not the last convergent, and if it is the last convergent then $a_{i+1} = \infty$.

[3] The term *proper semiconvergent* is not standard.

- There always are exactly two integer semiconvergents, namely a_0 and a_0+1. Indeed, $a_0+1 = \frac{p_{-1}+1\cdot p_0}{q_{-1}+1\cdot q_0}$ is always a semiconvergent. When there are at least two convergents, then $a_0 = \frac{p_0+0\cdot p_1}{q_0+1\cdot q_1}$ is a semiconvergent, and when there is only one convergent, then $a_0 = \frac{p_0}{q_0} = x$ is anyway a semiconvergent. Conversely, if $\frac{p}{1}$ is a semiconvergent, then either $x = \frac{p}{1}$ or $1 = q_{i-1} + sq_i$ for some i and $s \in [0, a_{i+1})$. For the first case, if the standard expansion is used, then $a_0 = x = \frac{p}{1}$, and if the alternative expansion is used, then $a_0 + 1 = x = \frac{p}{1}$. For the second case, we either have $q_{i-1} = 0$ and $s = q_i = 1$ or $q_{i-1} = 1$ and $s = 0$. The first case can happen only when $i = 0$, in which case we have $\frac{p}{1} = \frac{1+p_0}{1} = a_0 + 1$. For the second case, since $(q_j)_j$ is a strictly increasing sequence, we must have $i = 1$, so $\frac{p}{1} = \frac{p_{i-1}}{q_{i-1}} = \frac{p_0}{q_0} = a_0$.
- Hence, if x is not an integer, then $\lfloor x \rfloor = a_0$ and $\lceil x \rceil = a_0 + 1$ are the only integer semiconvergents, and they are both *proper* semiconvergents.
- If x is an integer, then x and $x+1$ are the only integer semiconvergents if the standard expansion is used, while $x-1$ and x are the only integer semiconvergents if the alternative expansion is used. There is no proper semiconvergent in this case.

When i is even, we call $\frac{p_i}{q_i}$ an *even convergent*, and when i is odd, we call it an *odd convergent*. Similarly, by an *even semiconvergent* we mean either x itself when x is an even convergent of itself, or any semiconvergent of the form

$$\frac{p_{2i} + sp_{2i+1}}{q_{2i} + sq_{2i+1}},$$

and by an *odd semiconvergent* we mean either x itself when x is an odd convergent of itself, or any semiconvergent of the form

$$\frac{p_{2i-1} + sp_{2i}}{q_{2i-1} + sq_{2i}}.$$

By *consecutive semiconvergents* we mean two semiconvergents of the form $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ and $\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i}$ with $s < a_{i+1}$. Note that $s+1 = a_{i+1}$ is allowed, in which case the second semiconvergent becomes $\frac{p_{i+1}}{q_{i+1}}$.

Lemma 4.3.
Let $\frac{p_i}{q_i}$ denote the ith convergent of a given regular continued fraction $x = [a_0; a_1, \cdots]$. Then for a given two consecutive convergents $\frac{p_{i-1}}{q_{i-1}}$, $\frac{p_i}{q_i}$, as s varies from 0 to a_{i+1}, the semiconvergent

$$\frac{p_{i-1} + sp_i}{q_{i-1} + sq_i}$$

strictly increases from $\frac{p_{i-1}}{q_{i-1}}$ to $\frac{p_{i+1}}{q_{i+1}}$ when i is odd, and it strictly decreases from $\frac{p_{i-1}}{q_{i-1}}$ to $\frac{p_{i+1}}{q_{i+1}}$ when i is even, where we consider $\frac{p_{i+1}}{q_{i+1}}$ to be equal to x if $\frac{p_i}{q_i}$ is the last convergent.

Proof. Consider the function

$$f(s) := \frac{p_{i-1} + sp_i}{q_{i-1} + sq_i}$$

of real variable s, then from the identity

$$p_i q_{i-1} - p_{i-1} q_i = (-1)^{i+1}$$

it can be easily verified that $f'(s) > 0$ when i is odd and $f'(s) < 0$ when i is even, on $s > 0$. Hence, the claim follows. □

In fact, $s = a_{i+1}$ (assuming $a_{i+1} < \infty$) is the last s such that the order between x and $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ is retained; that is, we have

$$x \leq \frac{p_{i-1} + (a_{i+1}+1)p_i}{q_{i-1} + (a_{i+1}+1)q_i} = \frac{p_i + p_{i+1}}{q_i + q_{i+1}}$$

when i is odd, and

$$x \geq \frac{p_{i-1} + (a_{i+1}+1)p_i}{q_{i-1} + (a_{i+1}+1)q_i} = \frac{p_i + p_{i+1}}{q_i + q_{i+1}}$$

when i is even. To see why, note that if $a_{i+2} = \infty$, then $x = \frac{p_{i+1}}{q_{i+1}}$, and (2) shows

$$\frac{p_i + p_{i+1}}{q_i + q_{i+1}} - \frac{p_{i+1}}{q_{i+1}} = \frac{p_i q_{i+1} - p_{i+1} q_i}{(q_i + q_{i+1})q_{i+1}} = \frac{(-1)^{i+1}}{(q_i + q_{i+1})q_{i+1}},$$

so we indeed get the desired inequalities and they are always strict. If $a_{i+2} < \infty$, then

$$\begin{aligned}\frac{p_i + p_{i+1}}{q_i + q_{i+1}} - \frac{p_{i+2}}{q_{i+2}} &= \frac{(p_i q_{i+2} - p_{i+2} q_i) + (p_{i+1} q_{i+2} - p_{i+1} q_{i+2})}{(q_i + q_{i+1})q_{i+2}} \\ &= \frac{a_{i+2}(p_i q_{i+1} - p_{i+1} q_i) + (-1)^i}{(q_i + q_{i+1})q_{i+2}} = \frac{(-1)^{i+1}(a_{i+2} - 1)}{(q_i + q_{i+1})q_{i+2}}.\end{aligned}$$

Since $x \leq \frac{p_{i+2}}{q_{i+2}}$ when i is odd and $x \geq \frac{p_{i+2}}{q_{i+2}}$ when i is even, we get

$$x \leq \frac{p_i + p_{i+1}}{q_i + q_{i+1}} - \frac{(-1)^{i+1}(a_{i+2} - 1)}{(q_i + q_{i+1})q_{i+2}} \leq \frac{p_i + p_{i+1}}{q_i + q_{i+1}}$$

when i is odd, and

$$x \geq \frac{p_i + p_{i+1}}{q_i + q_{i+1}} - \frac{(-1)^{i+1}(a_{i+2} - 1)}{(q_i + q_{i+1})q_{i+2}} \geq \frac{p_i + p_{i+1}}{q_i + q_{i+1}}$$

when i is even, with the equalities if and only if $x = \frac{p_{i+2}}{q_{i+2}}$ and $a_{i+2} = 1$.

Note that this equality condition means that x is rational and the continued fraction $[a_0; a_1, \cdots]$ ended with 1, which can never happen if we use the standard expansion of x. Hence, we get:

Lemma 4.4.
Let $\frac{p_i}{q_i}$ denote the ith convergent of the standard continued fraction expansion $x = [a_0; a_1, \cdots]$. Then

$$a_{i+1} = \left\lfloor \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \right\rfloor$$

holds for each i.

Proof. Follows directly from the discussion above together with that

$$\frac{p_{2i} + s p_{2i+1}}{q_{2i} + s q_{2i+1}} \leq x$$

is equivalent to

$$s \leq \frac{q_{2i}x - p_{2i}}{p_{2i+1} - q_{2i+1}x}$$

and

$$\frac{p_{2i-1} + sp_{2i}}{q_{2i-1} + sq_{2i}} \geq x$$

is equivalent to

$$s \leq \frac{p_{2i-1} - q_{2i-1}x}{q_{2i}x - p_{2i}}.$$

□

From now on, we will build a theory about how semiconvergents divide the approximation intervals and how rational numbers in between two consecutive semiconvergents should look like. These will lead to the main theorem of this section, which is the characterization of best rational approximations from below/above in terms of semiconvergents.

Lemma 4.5.
Let $\frac{p_i}{q_i}$ denote the ith convergent of a given regular continued fraction $x = [a_0; a_1, \cdots]$. Then for any $y \in [a_0, x)$, there uniquely exist two consecutive even semiconvergents $\frac{p_{2i}+sp_{2i+1}}{q_{2i}+sq_{2i+1}}$ and $\frac{p_{2i}+(s+1)p_{2i+1}}{q_{2i}+(s+1)q_{2i+1}}$ such that

$$\frac{p_{2i} + sp_{2i+1}}{q_{2i} + sq_{2i+1}} \leq y < \frac{p_{2i} + (s+1)p_{2i+1}}{q_{2i} + (s+1)q_{2i+1}}.$$

Similarly, for any $z \in (x, a_0 + 1]$, there uniquely exist two consecutive odd semiconvergents $\frac{p_{2i-1}+sp_{2i}}{q_{2i-1}+sq_{2i}}$ and $\frac{p_{2i-1}+(s+1)p_{2i}}{q_{2i-1}+(s+1)q_{2i}}$ such that

$$\frac{p_{2i-1} + (s+1)p_{2i}}{q_{2i-1} + (s+1)q_{2i}} < z \leq \frac{p_{2i-1} + sp_{2i}}{q_{2i-1} + sq_{2i}}.$$

Proof. By Corollary 3.3, either y lies in between two consecutive even convergents $\frac{p_{2i}}{q_{2i}}$ and $\frac{p_{2i+2}}{q_{2i+2}}$, that is,

$$\frac{p_{2i}}{q_{2i}} \leq y < \frac{p_{2i+2}}{q_{2i+2}},$$

or it lies in between the last even convergent $\frac{p_{2i}}{q_{2i}}$ and $x = \frac{p_{2i+1}}{q_{2i+1}}$, that is,

$$\frac{p_{2i}}{q_{2i}} \leq y < \frac{p_{2i+1}}{q_{2i+1}} = x.$$

Then for either case, Lemma 4.3 shows that there uniquely exists an integer $s \in [0, a_{2i+2})$ such that

$$\frac{p_{2i} + sp_{2i+1}}{q_{2i} + sq_{2i+1}} \leq y < \frac{p_{2i} + (s+1)p_{2i+1}}{q_{2i} + (s+1)q_{2i+1}}$$

holds.

Similarly, with the convention $\frac{p_{-1}}{q_{-1}} = \infty$, z must satisfy either

$$\frac{p_{2i+1}}{q_{2i+1}} < z \leq \frac{p_{2i-1}}{q_{2i-1}}$$

or

$$x = \frac{p_{2i}}{q_{2i}} < z \leq \frac{p_{2i-1}}{q_{2i-1}},$$

and for either case we can find a unique integer $s \in [0, a_{2i+2})$ such that

$$\frac{p_{2i-1} + (s+1)p_{2i}}{q_{2i-1} + (s+1)q_{2i}} < z \leq \frac{p_{2i-1} + sp_{2i}}{q_{2i-1} + sq_{2i}}.$$

□

Lemma 4.6.
Let $\frac{a}{b} < \frac{c}{d}$ be rational numbers in their reduced form. Then any rational number $\frac{x}{y} \in \left(\frac{a}{b}, \frac{c}{d}\right)$ in its reduced form can be uniquely written as

$$\frac{x}{y} = \frac{qa + pc}{qb + pd}$$

for some coprime positive integers p, q. If $bc - ad = 1$, then we have

$$\begin{cases} p = bx - ay, \\ q = cy - dx, \end{cases} \qquad \text{and} \qquad \begin{cases} x = qa + pc, \\ y = qb + pd. \end{cases}$$

Hence, $\frac{qa+pc}{qb+pd}$ is in its reduced form in this case. The exact same result is true even when we allow $\frac{a}{b} = \frac{0}{1} = 0$ or $\frac{c}{d} = \frac{1}{0} = \infty$ or both.

Proof. Consider the function

$$f(t) := \frac{a + tc}{b + td}$$

of real variable t, then

$$f'(t) = \frac{bc - ad}{(b + td)^2} > 0,$$

so f is strictly increasing on $t > 0$. Hence, if we have

$$\frac{q_1 a + p_1 c}{q_1 b + p_1 d} = \frac{q_2 a + p_2 c}{q_2 b + p_2 d}$$

for some positive integers p_1, q_1, p_2, q_2, then we must have

$$\frac{p_1}{q_1} = \frac{p_2}{q_2}.$$

In particular, if $\gcd(p_1, q_1) = \gcd(p_2, q_2) = 1$, then we must have $p_1 = p_2$ and $q_1 = q_2$. This shows the uniquenss part. Also, note that

$$\frac{x}{y} = \frac{(cy - dx)a + (bx - ay)c}{(cy - dx)b + (bx - ay)d}$$

always holds. Indeed, the numerator is equal to $(bc - ad)x$ while the denominator is equal to $(bc - ad)y$. Hence, let $g := \gcd(cy - dx, bx - ay)$, then we have $p = (bx - ay)/g$ and $q = (cy - dx)/g$. Then it only remains to show that $g = 1$ if $bc - ad = 1$.

Since x, y are coprime, we can find integers z, w such that $xz - yw = 1$. Then, the determinant of the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} y & -z \\ -x & w \end{pmatrix} = \begin{pmatrix} ay - bx & bw - az \\ cy - dx & dw - cz \end{pmatrix}$$

must be equal to $bc - ad$. Hence, when $bc - ad = 1$, then $bx - ay$ and $cy - dx$ must be indeed coprime to each other. □

Corollary 4.7.
Let $\frac{p_i}{q_i}$ denote the ith convergent of a given regular continued fraction $[a_0; a_1, \cdots]$. Then

any rational number non-strictly between two consecutive semiconvergents $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ *and* $\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i}$ *is of the form*

$$\frac{bp_{i-1}+(a+bs)p_i}{bq_{i-1}+(a+bs)q_i}$$

for some rational number $\frac{a}{b}\in[0,1]$*. Furthermore, the above expression is always in its reduced form if* $\frac{a}{b}$ *is in its reduced form.*

Proof. It immediately follows from the previous lemma that any rational number between those two semiconvergents must be of the form

$$\frac{qp_{i-1}+pp_i}{qq_{i-1}+pq_i}$$

for some coprime nonnegative integers p, q, because such a rational number must be in between $\frac{p_{i-1}}{q_{i-1}}$ and $\frac{p_i}{q_i}$. Furthermore, the above expression must be of the reduced form. Note that

$$\begin{aligned}(qp_{i-1}+pp_i)(q_{i-1}+sq_i)-(qq_{i-1}+pq_i)(p_{i-1}+sp_i)\\ =(p-qs)(p_iq_{i-1}-p_{i-1}q_i)=(-1)^{i-1}(q-ps),\end{aligned}$$

and similarly the same computation with s replaced by $s+1$ gives $(-1)^{i-1}(q-p(s+1))$. Therefore, $\frac{qp_{i-1}+pp_i}{qq_{i-1}+pq_i}$ lies in between $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ and $\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i}$ if and only if $\frac{p}{q}\in[s,s+1]$. Hence, let $a=p-qs$ and $b=q$, then we get the desired result. □

Corollary 4.8.
Let $\frac{p_i}{q_i}$ *denote the ith convergent of a given regular continued fraction* $x=[a_0;a_1,\ \cdots]$*. Then for any rational number* $\frac{p}{q}$ *in its reduced form between two consecutive semiconvergents, in either* $\left[\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i},\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i}\right)$ *or* $\left(\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i},\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}\right]$ *depending on the parity of i, we always have*

$$|qx-p|\geq|(q_{i-1}+sq_i)x-(p_{i-1}+sp_i)|\,.$$

The equality condition is:

1. *If* $s+1=a_{i+1}$ *and* $x=\frac{p_{i+1}}{q_{i+1}}$ *(thus in particular* $\frac{p_i}{q_i}$ *is not the last convergent and* $x\neq\frac{p_i}{q_i}$*), then the equality holds if and only if*

$$\begin{cases}p=bp_{i-1}+(bs+(b-1))p_i,\\ q=bq_{i-1}+(bs+(b-1))q_i\end{cases}$$

   for a positive integer b.

2. *Otherwise, the equality holds if and only if*

$$\begin{cases}p=p_{i-1}+sp_i,\\ q=q_{i-1}+sq_i.\end{cases}$$

Proof. By Corollary 4.7, we know

$$\begin{cases}p=bp_{i-1}+(a+bs)p_i,\\ q=bq_{i-1}+(a+bs)q_i\end{cases}$$

for some rational number $\frac{a}{b} \in [0, 1)$ in its reduced form. Assume $p_i - q_i x \neq 0$ first, then we get

$$|qx - p| = |b(q_{i-1}x - p_{i-1}) - (a + bs)(p_i - q_i x)| = b\,|p_i - q_i x| \left| \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} - \left(s + \frac{a}{b}\right) \right|,$$

and similarly

$$|(q_{i-1} + sq_i)x - (p_{i-1} + sp_i)| = |p_i - q_i x| \left| \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} - s \right|.$$

Recall from Lemma 4.4 that $\left\lfloor \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \right\rfloor = a_{i+1}$ if $[a_0; a_1, \cdots]$ is the standard continued fraction expansion of x. Otherwise, if $\frac{p_{i+1}}{q_{i+1}}$ is the last convergent, then $\frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} = a_{i+1} = 1$ holds because we know $x = \frac{p_{i-1} + p_i}{q_{i-1} + q_i}$ which shows that the left-hand side is equal to

$$\frac{q_{i-1}(p_{i-1} + p_i) - p_{i-1}(q_{i-1} + q_i)}{p_i(q_{i-1} + q_i) - q_i(p_{i-1} + p_i)} = 1,$$

and if $\frac{p_{i+2}}{q_{i+2}}$ is the last convergent, then $\left\lfloor \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \right\rfloor = a_{i+1} + 1$ holds, and for any other case, $\left\lfloor \frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \right\rfloor = a_{i+1}$ should hold. Therefore, in any case, we always have $\frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \geq a_{i+1} \geq s + 1$. This shows

$$\frac{|qx - p| - |(q_{i-1} + sq_i)x - (p_{i-1} + sp_i)|}{|p_i - q_i x|} = (b - 1)\left(\frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} - \left(s + \frac{a}{b}\right) \right) - \frac{a}{b}.$$

Then since

$$\frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} \geq s + 1 \geq s + \frac{a + 1}{b},$$

we get

$$\frac{|qx - p| - |(q_{i-1} + sq_i)x - (p_{i-1} + sp_i)|}{|p_i - q_i x|} \geq \frac{b - 1}{b} - \frac{a}{b} \geq 0,$$

as desired. Also, the equality holds if and only if either $\frac{a}{b} = 0$ or $a = b - 1$ and

$$\frac{q_{i-1}x - p_{i-1}}{p_i - q_i x} = s + 1,$$

which means $x = \frac{p_{i+1}}{q_{i+1}}$. Hence, the claimed equality condition follows.

The case $p_i - q_i x = 0$ is simpler; in this case, we have

$$|qx - p| = b\,|q_{i-1}x - p_{i-1}|$$

and

$$|(q_{i-1} + sq_i)x - (p_{i-1} + sp_i)| = |q_{i-1}x - p_{i-1}|,$$

so we get the desired inequality and the equality condition is $b = 1$, which implies $a = 0$ and $\frac{p}{q} = \frac{p_{i-1} + sp_i}{q_{i-1} + sq_i}$. □

Lemma 4.9.
Let $\frac{p_i}{q_i}$ denote the ith convergent of a given regular continued fraction $x = [a_0; a_1, \cdots,]$. Then for any two consecutive semiconvergents $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ and $\frac{p_{i-1}+(s+1)p_i}{q_{i-1}+(s+1)q_i}$,

$$|(q_{i-1}+sq_i)x - (p_{i-1}+sp_i)| \geq |(q_{i-1}+(s+1)q_i)x - (p_{i-1}+(s+1)p_i)|.$$

The equality holds if and only if $x = \frac{p_i}{q_i}$.

Proof. This follows from direct computation:

$$((q_{i-1}+sq_i)x - (p_{i-1}+sp_i)) - ((q_{i-1}+(s+1)q_i)x - (p_{i-1}+(s+1)p_i)) = p_i - q_i x.$$

If i odd, then the right-hand side is nonnegative and $(q_{i-1}+sq_i)x - (p_{i-1}+sp_i)$, $(q_{i-1}+(s+1)q_i)x - (p_{i-1}+(s+1)p_i)$ are both nonnegative, so we get the desired inequality. If i is even, all three are nonpositive, so we still get the desired inequality. The equality condition is precisely $x = \frac{p_i}{q_i}$. □

Hence, we get:

Corollary 4.10.
Let $\left(\frac{r_j}{s_j}\right)_j$ be the sequence of all even (odd, resp.) semiconvergents of a regular continued fraction $x = [a_0; a_1, \cdots]$, all in their reduced form, such that $r_j < r_{j+1}$ holds for all j.[4]

1. *If x is rational and is an even (odd, resp.) convergent of itself, then the sequence $(|s_j x - r_j|)_j$ is finite and it strictly decreases until it reaches zero at the final term.*

2. *If x is rational and is an odd (even, resp.) convergent of itself, then the sequence $(|s_j x - r_j|)_j$ is infinite and it strictly decreases until $\frac{r_j}{s_j}$ becomes the last even convergent of x, and it stays constant after that.*

3. *If x is irrational, then the sequence $(|s_j x - r_j|)_j$ is infinite and it strictly decreases forever.*

Finally, we prove our main theorem in this section.

Theorem 4.11 (Characterization of best rational approximations from below/above)**.**
Let $x = [a_0; a_1, \cdots]$ be a regular continued fraction. Then for a rational number $\frac{p}{q} \neq x$ in its reduced form, the followings are equivalent.

1. *$\frac{p}{q}$ is an even (odd, resp.) proper semiconvergent of x.*

2. *$\frac{p}{q}$ is a best rational approximation of x from below (above, resp.) in the strong sense.*

3. *$\frac{p}{q}$ is a best rational approximation of x from below (above, resp.) in the weak sense.*

Proof. Let $\frac{p_i}{q_i}$ denote the ith convergent of $x = [a_0; a_1, \cdots]$.

(*1⇒2*) Let $\frac{p}{q}$ be a proper even semiconvergent. Then we can write $\frac{p}{q} = \frac{p_{2i}+sp_{2i+1}}{q_{2i}+sq_{2i+1}}$ for some consecutive convergents $\frac{p_{2i}}{q_{2i}}$, $\frac{p_{2i+1}}{q_{2i+1}}$ and an integer $s \in [0, a_{2i+2})$. Clearly, $\frac{p}{q} \leq x$ holds. Let $\frac{a}{b}$ be any rational number with $0 < b \leq q$ and $\frac{a}{b} \leq x$. Since $\frac{p}{q}$ is a proper semiconvergent and $0 < b \leq q$, we should have $\frac{a}{b} \neq x$. First, suppose $\frac{a}{b} < a_0$. Then Corollary 4.10 shows

$$bx - a > b(x - a_0) = b(q_0 x - p_0) \geq b(qx - p) \geq qx - p.$$

[4] Note that no two distinct even (odd, resp.) semiconvergents share their denominators.

Next, suppose $\frac{a}{b} \geq a_0$, then by Lemma 4.5, there uniquely exist two consecutive even semiconvergents $\frac{p_{2j}+tp_{2j+1}}{q_{2j}+tq_{2j+1}}$ and $\frac{p_{2j}+(t+1)p_{2j+1}}{q_{2j}+(t+1)q_{2j+1}}$ such that

$$\frac{p_{2j}+tp_{2j+1}}{q_{2j}+tq_{2j+1}} \leq \frac{a}{b} < \frac{p_{2j}+(t+1)p_{2j+1}}{q_{2j}+(t+1)q_{2j+1}}.$$

By Corollary 4.7, we know

$$q_{2j}+tq_{2j+1} \leq b \leq q = q_{2i}+sq_{2i+1},$$

thus Corollary 4.10 shows

$$(q_{2j}+tq_{2j+1})x - (p_{2j}+tp_{2j+1}) \geq qx - p.$$

Hence, from Corollary 4.8, we get

$$bx - a \geq (q_{2j}+tq_{2j+1})x - (p_{2j}+tp_{2j+1}) \geq qx - p,$$

thus we conclude that $\frac{p}{q}$ is a best rational approximation of x from below in the strong sense. In the same way, we can show that if $\frac{p}{q}$ is a proper odd semiconvergent, then it is a best rational approximation of x from above in the strong sense.

(2⇒3) Done in Lemma 2.3.

(3⇒1) Let $\frac{p}{q}$ be a best rational approximation of x from below in the weak sense. Since $\frac{p_0}{q_0} = a_0 \leq x$ and $q_0 = 1 \leq q$ holds, the assumption on $\frac{p}{q}$ implies $a_0 \leq \frac{p}{q} < x$. Hence, by Lemma 4.5, there uniquely exist two consecutive semiconvergents $\frac{p_{2i}+sp_{2i+1}}{q_{2i}+sq_{2i+1}}$ and $\frac{p_{2i}+(s+1)p_{2i+1}}{q_{2i}+(s+1)q_{2i+1}}$ such that

$$\frac{p_{2i}+sp_{2i+1}}{q_{2i}+sq_{2i+1}} \leq \frac{p}{q} < \frac{p_{2i}+(s+1)p_{2i+1}}{q_{2i}+(s+1)q_{2i+1}}.$$

Hence, Corollary 4.7 shows that there exist a rational number $\frac{a}{b} \in [0, 1)$ in its reduced form such that

$$\frac{p}{q} = \frac{bp_{2i}+(a+bs)p_{2i+1}}{bq_{2i}+(a+bs)q_{2i+1}}.$$

We claim that $\frac{a}{b} = \frac{0}{1}$. Suppose otherwise, then $a, b \geq 1$, so $q \geq q_{2i}+(s+1)q_{2i+1}$ holds, but since $\frac{p}{q}$ is a best rational approximation of x from below in the weak sense, this implies $\frac{p_{2i}+(s+1)p_{2i+1}}{q_{2i}+(s+1)q_{2i+1}} \leq \frac{p}{q}$, which is a contradiction. Therefore, we conclude that $\frac{p}{q} = \frac{p_{2i}+sp_{2i+1}}{q_{2i}+sq_{2i+1}}$ is an even semiconvergent. Since $\frac{p_{2i}+(s+1)p_{2i+1}}{q_{2i}+(s+1)q_{2i+1}}$ is also a semiconvergent with a strictly larger denominator, it follows that $\frac{p}{q}$ is a proper semiconvergent. □

This characterization of best rational approximations from below/above immediately leads us to the following conclusion: for a given real number x and a positive integer $n_{\max}$, the best rational approximations $\frac{p_*}{q_*}$ from below and $\frac{p^*}{q^*}$ from above of x with the largest denominators $q_*, q^* \leq n_{\max}$, are either both equal to x, or the even/odd semiconvergents of largest possible denominators bounded by $n_{\max}$ (which are necessarily proper). Therefore, we get the following simple algorithm for computing $\frac{p_*}{q_*}$ and $\frac{p^*}{q^*}$, for any given x of which we can compute the continued fraction expansion:

Algorithm 4.12 (Finding best rational approximations from below and above)**.**

1. Find the last convergent $\frac{p_i}{q_i}$ of x such that $q_i \leq n_{\max}$.
2. If $x = \frac{p_i}{q_i}$, then return $\left(\frac{p_*}{q_*}, \frac{p^*}{q^*}\right) = \left(\frac{p_i}{q_i}, \frac{p_i}{q_i}\right)$.
3. Otherwise, find the last semiconvergent $\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}$ such that $q_{i-1} + sq_i \leq n_{\max}$.
4. If i is even, then return $\left(\frac{p_*}{q_*}, \frac{p^*}{q^*}\right) = \left(\frac{p_i}{q_i}, \frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}\right)$.
5. If i is odd, then return $\left(\frac{p_*}{q_*}, \frac{p^*}{q^*}\right) = \left(\frac{p_{i-1}+sp_i}{q_{i-1}+sq_i}, \frac{p_i}{q_i}\right)$.

Note that all we need to know about x is how to compute its continued fraction expansion. Because of this, the reference implementation is written in a way that all real numbers are supposed to be encoded in terms of their continued fraction expansions. This viewpoint of thinking real numbers and their continued fraction expansions as identical will be retained throughout the whole note, as it is particularly convenient for our purpose.

5 Extremizers of $(nx + y) - \lfloor nx + y \rfloor$

In this section, we now obtain the algorithm for efficiently computing extremizers of $(nx + y) - \lfloor nx + y \rfloor$ with full generality. The algorithm we derive in fact is a simple extension of the classic Euclid's algorithm.

We will actually find the *smallest minimizer* and the *largest maximizer* of $(nx + y) - \lfloor nx + y \rfloor$ for reasons that will become apparent later. Throughout this section, let $x, y \in \mathbb{R}$ be any real numbers and $n_{\min} \leq n_{\max}$ be any integers.

We start with the algorithm for finding the smallest minimizer. In fact, the algorithm follows at once by examining how exactly the fractional part of $nx + y$ would change when n increases. Let us say we start with $n_0 = n_{\min}$ and we are trying to compare the fractional part $(n_0x+y)-\lfloor n_0x + y \rfloor$ of n_0x+y with the fractional part $((n_0+n)x+y)-\lfloor (n_0 + n)x + y \rfloor$ of $(n_0 + n)x + y$. Clearly, if the fractional part of nx is small, then adding it to $n_0x + y$ will just increase the fractional part. Specifically, if the fractional part of nx is strictly less than current the margin of the fractional part of $n_0x + y$, then the fractional parts of them will just get added together. On the other hand, if the fractional part of nx is at least as large as the said margin, then the fractional part will "wrap around" to zero, and the result should be strictly less than the fractional part of $n_0x + y$. Or equivalently, if the fractional part of $n_0x + y$ is greater than or equal to the margin of nx, then adding them together will make the fractional part to strictly decrease, precisely by the amount of said margin. The lemma below describes this rigorously.

Lemma 5.1.
Let $n_0 \in \mathbb{Z}$ and $n \in \mathbb{Z}_{\geq 0}$. Then

$$\begin{aligned}&((n_0 + n)x + y) - \lfloor (n_0 + n)x + y \rfloor \\ &\quad = \begin{cases} ((n_0x + y) - \lfloor n_0x + y \rfloor) - (\lfloor nx \rfloor + 1 - nx) & \text{if } \lfloor nx \rfloor + 1 - nx \leq (n_0x + y) - \lfloor n_0x + y \rfloor, \\ ((n_0x + y) - \lfloor n_0x + y \rfloor) + (nx - \lfloor nx \rfloor) & \text{if } nx - \lfloor nx \rfloor < \lfloor n_0x + y \rfloor + 1 - (n_0x + y). \end{cases}\end{aligned}$$

In particular,

$$((n_0 + n)x + y) - \lfloor (n_0 + n)x + y \rfloor < (n_0x + y) - \lfloor n_0x + y \rfloor$$

holds if and only if

$$\lfloor nx \rfloor + 1 - nx \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor .$$

Proof. If

$$\lfloor nx \rfloor + 1 - nx \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor$$

holds, then

$$(n_0 + n)x + y \geq \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1$$

and if it does not hold, then

$$(n_0 + n)x + y < \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1.$$

Since $\lfloor (n_0 + n)x + y \rfloor$ is equal to either $\lfloor n_0 x + y \rfloor + \lfloor nx \rfloor$ or $\lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1$, we should have

$$\lfloor (n + n_0)x + y \rfloor = \begin{cases} \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1 & \text{if } \lfloor nx \rfloor + 1 - nx \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor, \\ \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor & \text{if } \lfloor nx \rfloor + 1 - nx > (n_0 x + y) - \lfloor n_0 x + y \rfloor. \end{cases}$$

For the first case,

$$\begin{aligned} ((n_0 + n)x + y) - \lfloor (n_0 + n)x + y \rfloor &= ((n_0 x + y) - \lfloor n_0 x + y \rfloor) - (\lfloor nx \rfloor + 1 - nx) \\ &< (n_0 x + y) - \lfloor n_0 x + y \rfloor , \end{aligned}$$

and for the second case,

$$\begin{aligned} ((n_0 + n)x + y) - \lfloor (n_0 + n)x + y \rfloor &= ((n_0 x + y) - \lfloor n_0 x + y \rfloor) + (nx - \lfloor nx \rfloor) \\ &\geq (n_0 x + y) - \lfloor n_0 x + y \rfloor . \end{aligned}$$

□

Therefore, the first $n \in \mathbb{Z}_{>0}$ that makes the fractional part of $(n_0 + n)x + y$ strictly smaller than that of $n_0 x + y$ is the smallest $n \in \mathbb{Z}_{>0}$ such that the $\lfloor nx \rfloor + 1 - nx$ is smaller than or equal to the current fractional part of $n_0 x + y$. Let us call such n as n_1. If we add n_1 to n_0, then the fractional part decreases by $\lfloor n_1 x \rfloor + 1 - n_1 x$. Since it has decreased, no $n \in [1 \colon n_1 - 1]$ can ever satisfy the condition

$$\lfloor nx \rfloor + 1 - nx \leq ((n_0 + n_1)x + y) - \lfloor (n_0 + n_1)x + y \rfloor ,$$

because n_1 was chosen to be the smallest that satisfies a less strict inequality $\lfloor nx \rfloor + 1 - nx \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor$. On the other hand, $n = n_1$ might still work out, that is, we may still have

$$\lfloor n_1 x \rfloor + 1 - n_1 x \leq ((n_0 + n_1)x + y) - \lfloor (n_0 + n_1)x + y \rfloor .$$

If that happens, then still n_1 is the smallest that satisfies the above, so $n = n_0 + 2n_1$ is the next smallest after $n_0 + n_1$ where $(nx + y) - \lfloor nx + y \rfloor$ decreases further than its previous minimum over $[n_0 \colon n - 1]$. Now, note that adding n_1 to $n_0 + n_1$ decreases the fractional part by the exact same amount $\lfloor n_1 x \rfloor + 1 - n_1 x$. Hence, this trend will continue, that is, n_1 continues to be the smallest increment that further decreases the fractional part, until the fractional part finally becomes strictly smaller than $\lfloor n_1 x \rfloor + 1 - n_1 x$. Therefore, the exact number of times that we can reuse n_1 is given as

$$a = \left\lfloor \frac{(n_0 x + y) - \lfloor n_0 x + y \rfloor}{\lfloor n_1 x \rfloor + 1 - n_1 x} \right\rfloor ,$$

and after adding n_1 for a times, finally we will need a different n to make the fractional part even smaller. So we replace n_0 by $n_0 + an_1$ and find the smallest n_1 that satisfies

$$\lfloor n_1 x \rfloor + 1 - n_1 x \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor .$$

Then we can repeat this procedure until n_0 becomes bigger than $n_{\max}$.

Now, so far this feels like a kind of a greedy algorithm, and it is not immediately clear if this approach will really give us a true minimizer, let alone the smallest minimizer. To see it indeed does, let $m \in [n_{\min} \colon n_{\max}]$ be the smallest minimizer of $(nx + y) - \lfloor nx + y \rfloor$ over $[n_{\min} \colon n_{\max}]$. Note that because of Lemma 5.1, we must have either $m = n_0 = n_{\min}$ or

$$\lfloor (m - n_0)x + y \rfloor + 1 - ((m - n_0)x + y) \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor .$$

If $m = n_0$ is the case, then we should have failed to find n_1 (because otherwise the fractional part will strictly decrease when we add n_1 to m), so we would get the correct minimizer n_0. If $m > n_0$, then beause of the minimality of n_1, we must have $m - n_0 \geq n_1$, so m is still greater than or equal to $n_0 + n_1$. In this case, since m is a minimizer, Lemma 5.1 still implies either $m = n_0 + n_1$ or

$$\lfloor (m - n_0 - n_1)x + y \rfloor + 1 - ((m - n_0 - n_1)x + y) \leq ((n_0 + n_1)x + y) - \lfloor (n_0 + n_1)x + y \rfloor .$$

Therefore, $m - n_0 - n_1$ should be greater than or equal to the smallest n such that $\lfloor nx \rfloor + 1 - nx$ is smaller than or equal to the right-hand side of the above inequality. Therefore, our iteration procedure should eventually exhaust the entire gap between m and n_0, and thus reaches m.[5] Consequently, we obtain the following result:

Corollary 5.2 (The smallest minimizer of the fractional part)**.**
Let $n_0^{(0)} := n_{\min}$*. For* $i \in \mathbb{Z}_{>0}$*, inductively define* $n_0^{(i)} := n_0^{(i-1)} + a^{(i)} n_1^{(i)}$*,*

$$n_1^{(i)} := \min \left\{ n \in \mathbb{Z}_{>0} \colon \; \lfloor nx \rfloor + 1 - nx \leq (n_0^{(i-1)} x + y) - \left\lfloor n_0^{(i-1)} x + y \right\rfloor \right\}$$

and

$$a^{(i)} := \left\lfloor \frac{(n_0^{(i-1)} x + y) - \left\lfloor n_0^{(i-1)} x + y \right\rfloor}{\left\lfloor n_1^{(i)} x \right\rfloor + 1 - n_1^{(i)} x} \right\rfloor$$

when the set defining $n_1^{(i)}$ *is not empty, and otherwise* $n_0^{(i)} := n_1^{(i)} := \infty$*. Let* k *be the largest nonnegative integer such that* $n_0^{(k)} \leq n_{\max}$*. Then*

$$\min \operatorname*{argmin}_{n \in [n_{\min} \colon n_{\max}]} ((nx + y) - \lfloor nx + y \rfloor) = n_0^{(k)} + \left\lfloor \frac{n_{\max} - n_0^{(k)}}{n_1^{(k+1)}} \right\rfloor n_1^{(k+1)}$$

with the convention $0 \cdot \infty = 0$*.*

Finding the maximizer can be done in a similar manner. In fact, Lemma 5.1 already gives all needed information. Starting with $n_0 = n_{\min}$, we find the smallest $n_1 \in \mathbb{Z}_{>0}$ satisfying

$$n_1 x - \lfloor n_1 x \rfloor < \lfloor n_0 x + y \rfloor + 1 - (n_0 x + y),$$

[5] Note that this exhaution happens precisely because m is the *smallest* minimizer. Indeed, if m is not the smallest minimizer, then once our iteration procedure reaches to the smallest minimizer, then we can no longer find n that satisfies $\lfloor nx \rfloor + 1 - nx \leq (n_0 x + y) - \lfloor n_0 x + y \rfloor$, because if that were possible, then we can further decrease the fractional part.

and add it to n_0 for a times, where

$$a = \left\lceil \frac{\lfloor n_0x + y\rfloor + 1 - (n_0x + y)}{n_1x - \lfloor n_1x\rfloor} \right\rceil - 1$$

is the smallest integer such that

$$n_1x - \lfloor n_1x\rfloor \geq \lfloor (n_0 + an_1)x + y\rfloor + 1 - ((n_0 + an_1)x + y)$$

holds. Indeed, the right-hand side must be equal to

$$1 - ((n_0x + y) - \lfloor n_0x + y\rfloor + a(n_1x - \lfloor n_1x\rfloor)),$$

so a is the smallest integer satisfying

$$a + 1 \geq \frac{\lfloor n_0x + y\rfloor + 1 - (n_0x + y)}{n_1x - \lfloor n_1x\rfloor}.$$

Then, we replace n_0 by $n_0 + an_1$ and repeat this procedure until n_0 becomes larger than $n_{\max}$.

Of course, this breaks down if n_1x happens to be an integer. We claim that in that case, we have already have reached the smallest maximizer. If the claim holds, then we can simply add as many multiple of n_1 as needed to n_0 to find the largest maximizer. (Note that in this case n_1 must be the smallest positive integer such that n_1x is an integer, and adding n_1 to n_0 should not change the fractional part of $n_0x + y$.)

To see why the claim holds, let $m \in [n_{\min} \colon n_{\max}]$ be the smallest maximizer as in the case of minimizer. Then since m is the smallest, there should be no $n \in [1 \colon m - n_{\min}]$ such that nx is an integer, because once we find such n, then $m - n$ must be a strictly smaller maximizer than m. Hence, the iteration procedure must exhaust the gap between $n_{\min}$ and m before we find n_1 making n_1x into an integer. Consequently, we obtain the following result:

Corollary 5.3 (The largest maximizer of the fractional part)**.**
Let $n_0^{(0)} := n_{\min}$. For $i \in \mathbb{Z}_{>0}$, inductively define $n_0^{(i)} := n_0^{(i-1)} + a^{(i)}n_1^{(i)}$,

$$n_1^{(i)} := \min\left\{n \in \mathbb{Z}_{>0} \colon nx - \lfloor nx\rfloor < \left\lfloor n_0^{(i-1)}x + y\right\rfloor + 1 - (n_0^{(i-1)}x + y)\right\}$$

and

$$a^{(i)} := \left\lceil \frac{\left\lfloor n_0^{(i-1)}x + y\right\rfloor + 1 - (n_0^{(i-1)}x + y)}{n_1^{(i)}x - \left\lfloor n_1^{(i)}x\right\rfloor} \right\rceil - 1$$

when $n_1^{(i)}x$ is not an integer, and otherwise $n_0^{(i)} := \infty$. Let k be the largest nonnegative integer such that $n_0^{(k)} \leq n_{\max}$. Then

$$\max \operatorname*{argmax}_{n\in[n_{\min} \colon n_{\max}]} ((nx + y) - \lfloor nx + y\rfloor) = n_0^{(k)} + \left\lfloor \frac{n_{\max} - n_0^{(k)}}{n_1^{(k+1)}} \right\rfloor n_1^{(k+1)}$$

with the convention $0 \cdot \infty = 0$.

Therefore, the problems of finding the smallest minimizer and the largest maximizer of $(nx + y) - \lfloor nx + y\rfloor$ becomes the problems of finding the smallest $n \in \mathbb{Z}_{>0}$ such that

$$\lfloor nx\rfloor + 1 - nx \leq \tau \quad \text{with} \quad \tau := (n_0x + y) - \lfloor n_0x + y\rfloor \in [0, 1)$$

and

$$nx - \lfloor nx \rfloor < \tau \quad \text{with} \quad \tau \coloneqq \lfloor n_0 x + y \rfloor + 1 - (n_0 x + y) \in (0, 1],$$

respectively.

Recall from Theorem 2.4 that finding extremizers of $nx - \lfloor nx \rfloor$ is equivalent to finding best rational approximations of x from below/above, which, by Theorem 4.11, can be done by enumerating semiconvergents of x with increasingly larger denominators.

Let us give some more details here. As in Section 3 and Section 4, let $x = [a_0; a_1, \cdots]$ and $\frac{p_i}{q_i}$ be the ith convergent. For the problem $\lfloor nx \rfloor + 1 - nx \le \tau$, let $n = n_1$ be the smallest positive integer satisfying the inequality, if any. Then n_1 is the unique maximizer of $nx - \lfloor nx \rfloor$ over $n \in [1\colon n_1]$. By Theorem 2.4 and Theorem 4.11, n_1 must be the denominator of a proper odd semiconvergent.[6]

Hence, we find the first odd convergent $\frac{p_i}{q_i}$ such that $p_i - q_i x \le \tau$ holds. Note that $p_i = \lceil q_i x \rceil = \lfloor q_i x \rfloor + 1$ holds unless $x = \frac{p_i}{q_i}$, thus except for that case, $p_i - q_i x \le \tau$ is equivalent to $\lfloor q_i x \rfloor + 1 - q_i x \le \tau$.

Suppose that we found such an odd convergent $\frac{p_i}{q_i}$ that is not equal to x. Then this means that the smallest n we are looking for must be of the form $n = q_i - s q_{i-1}$ for some integer $s \in [0, a_i)$. By construction, we know

$$(p_i - q_i x) + a_i(q_{i-1}x - p_{i-1}) = p_{i-2} - q_{i-2}x > \tau \ge p_i - q_i x.$$

In this case, s must be the largest integer such that

$$\tau \ge (p_i - q_i x) + s(q_{i-1}x - p_{i-1}),$$

that is,

$$s = \left\lfloor \frac{\tau - (p_i - q_i x)}{q_{i-1}x - p_{i-1}} \right\rfloor.$$

Then $n = q_i - s q_{i-1}$ is the desired output, and additionally we know

$$\lfloor nx \rfloor + 1 - nx = (p_i - q_i x) + s(q_{i-1}x - p_{i-1}).$$

Next, suppose that there is no odd convergent $\frac{p_i}{q_i}$ such that $p_i - q_i x \le \tau$. This can happen if $x = \frac{p_j}{q_j}$ is its own even convergent. In this case, $\frac{p_{j-1}}{q_{j-1}}$ is the proper odd semiconvergent of x with the largest denominator, and since it failed to satisfy $p_{j-1} - q_{j-1}x \le \tau$, we conclude that there is no n satisfying $\lfloor nx \rfloor + 1 - nx \le \tau$.

Next, suppose that $x = \frac{p_i}{q_i}$ itself is the first convergent satisfying $p_i - q_i x \le \tau$. As noted earlier, this does not mean $\lfloor q_i x \rfloor + 1 - q_i x \le \tau$; rather, the left-hand side is equal to 1. In this case, we may need to look at proper odd semiconvergents of the form $\frac{p_i - s p_{i-1}}{q_i - s q_{i-1}}$ for $s \in (0, a_i)$, which are better rational approximations of x from above than the last odd convergent $\frac{p_{i-2}}{q_{i-2}}$. Hence, we try the same formula as before:

$$s = \left\lfloor \frac{\tau - (p_i - q_i x)}{q_{i-1}x - p_{i-1}} \right\rfloor.$$

In this case, if we get $s = 0$, then we conclude there is no n satisfying $\lfloor nx \rfloor + 1 - nx \le \tau$, and otherwise, $n = q_i - s q_{i-1}$ is the desired output and

$$\lfloor nx \rfloor + 1 - nx = (p_i - q_i x) + s(q_{i-1}x - p_{i-1}) = s(q_{i-1}x - p_{i-1}).$$

[6] In particular, n_1 cannot be the denominator of x, if x is rational. Of course, in this case $\lfloor n_1 x \rfloor + 1 - n_1 x = 1 > \tau$.

Now we look at the problem $nx - \lfloor nx \rfloor < \tau$. Similarly, the smallest $n = n_1$ satisfying the inequality should be the unique minimizer of $nx - \lfloor nx \rfloor$ over $n \in [1\colon n_1]$, which, by Theorem 2.4 and Theorem 4.11, is either the denominator of x (if x is rational) or the denominator of a proper even semiconvergent of x.

Hence, we find the first even convergent $\frac{p_i}{q_i}$ such that $q_i x - p_i < \tau$ holds. If there is no such even convergent, then that means $x = \frac{p_j}{q_j}$ is an odd convergent of itself. In this case, the last proper even semiconvergent is $\frac{p_{j-1}}{q_{j-1}}$ which did not satisfy $q_{j-1}x - p_{j-1} < \tau$, thus $n = q_j$ is the smallest n satisfying $nx - \lfloor nx \rfloor < \tau$.

Next, suppose that we found the first even convergent $\frac{p_i}{q_i}$ satisfying $q_i x - p_i < \tau$. If $i = 0$, then $n = 1$ is the desired output. Otherwise, we look at even semiconvergents of the form $\frac{p_i - sp_{i-1}}{q_i - sq_{i-1}}$ for $s \in [0, a_i)$. Note that there is a possibility of $x = \frac{p_i}{q_i}$ so that $\frac{p_i}{q_i}$ is not a proper semiconvergent, but that is okay because $n = q_i$ is the unique minimizer even in that case anyway. Now, we have

$$(q_i x - p_i) + a_i(p_{i-1} - q_{i-1}x) = q_{i-2}x - p_{i-2} \geq \tau > q_i x - p_i,$$

so s must be the largest integer satisfying

$$(q_i x - p_i) + s(p_{i-1} - q_{i-1}x) < \tau,$$

that is,

$$s = \left\lceil \frac{\tau - (q_i x - p_i)}{p_{i-1} - q_{i-1}x} \right\rceil - 1.$$

Then $n = q_i - sq_{i-1}$ is the desired output, and additionally we know

$$nx - \lfloor nx \rfloor = (q_i x - p_i) + s(p_{i-1} - q_{i-1}x).$$

Note that these formulas also work for the case $i = 0$, so in fact we do not need to make a special branch for $i = 0$.

In practice, since one problem only cares about even semiconvergents and the other problem only cares about odd semiconvergents, we can solve both at the same time in a single run. More specifically, we have the following algorithm:

Algorithm 5.4 (Finding extremizers of $(nx + y) - \lfloor nx + y \rfloor$)**.**

1. Set $n_{0,\min} \leftarrow n_{\min}$, $n_{0,\max} \leftarrow n_{\min}$, $\text{found}_{\min} \leftarrow \text{false}$, $\text{found}_{\max} \leftarrow \text{false}$, and $\frac{p}{q} \leftarrow \frac{1}{0}$, the -1st convergent of x.

2. Set terminated $\leftarrow$ false.

3. Set $\tau_{\min} \leftarrow (n_{0,\min}x + y) - \lfloor n_{0,\min}x + y \rfloor$ and $\tau_{\max} \leftarrow \lfloor n_{0,\max}x + y \rfloor + 1 - (n_{0,\max}x + y)$.

4. Check if there are further convergents of x. If so, let $\frac{p}{q}$ be the next convergent of x; in this case, $\frac{p}{q}$ must be an even convergent of x. If not, set terminated $\leftarrow$ true; in this case, x is equal to the last convergent $\frac{p}{q}$ and it is an odd convrgent.

5. If $\text{found}_{\max}$ is false,

    (a) If terminated is false,

        i. Check if $qx - p < \tau_{\max}$ holds. If it doesn't, we move to the next convergent, so go to Step 6.

ii. If it does, then we find the first semiconvergent between the current and the previous even convergents that satisfies the inequality. Set

$$s \leftarrow \left\lceil \frac{\tau_{\max} - (qx - p)}{p' - q'x} \right\rceil - 1, \quad n_1 \leftarrow q - sq' \quad \text{and} \quad \delta \leftarrow (qx - p) + s(p' - q'x),$$

where $\frac{p'}{q'}$ is the previous convergent.

iii. Check if $\delta = 0$. If that is the case, then go to Step 5(b)-i.

iv. Otherwise, set $a \leftarrow \lceil \tau_{\max}/\delta \rceil - 1$.

v. If $n_{0,\max} + an_1 \leq n_{\max}$, then set $n_{0,\max} \leftarrow n_{0,\max} + an_1$, recompute $\tau_{\max}$ accordingly, and then go back to Step 5(a)-i.

vi. Otherwise, go to Step 5(b)-i.

(b) If terminated is true, set $n_1 \leftarrow q$; in this case, n_1 is the smallest n satisfying $nx - \lfloor nx \rfloor < \tau$, and we have $n_1 x - \lfloor n_1 x \rfloor = 0$.

i. Set

$$n_{0,\max} \leftarrow n_{0,\max} + \left\lfloor \frac{n_{\max} - n_{0,\max}}{n_1} \right\rfloor n_1.$$

ii. Set found$_{\max} \leftarrow$ true and go to Step 6.

6. Check if there are further convergents of x. If so, let $\frac{p}{q}$ be the next convergent of x; in this case, $\frac{p}{q}$ must be an odd convergent of x. If not, set terminated $\leftarrow$ true; in this case, x is equal to the last convergent $\frac{p}{q}$ and it is an even convergent.

7. If found$_{\min}$ is false,

(a) If terminated is false,

i. Check if $p - qx \leq \tau_{\min}$ holds. If it doesn't, we move to the next convergent. Go to the Step 8.

ii. If it does, then we find the first semiconvergent between the current and the previous odd convergents that satisfies the inequality. Set

$$s \leftarrow \left\lfloor \frac{\tau_{\min} - (p - qx)}{q'x - p'} \right\rfloor, \quad n_1 \leftarrow q - sq' \quad \text{and} \quad \delta \leftarrow (p - qx) + s(q'x - p')$$

where $\frac{p'}{q'}$ is the previous convergent.

iii. Check if $\delta = 0$. If that is the case, then we conclude there is no n satisfying $\lfloor nx \rfloor + 1 - nx \leq \tau_{\min}$, so go to Step 6(b)-i.

iv. Otherwise, set $a \leftarrow \lfloor \tau_{\min}/\delta \rfloor$.

v. If $n_{0,\min} + an_1 \leq n_{\max}$, set $n_{0,\min} \leftarrow n_{0,\min} + an_1$, recompute $\tau_{\min}$ accordingly, and then go back to Step 7(a)-i.

vi. Otherwise, set

$$n_{0,\min} \leftarrow n_{0,\min} + \left\lfloor \frac{n_{\max} - n_{0,\min}}{n_1} \right\rfloor n_1$$

and go to Step 6(b)-i.

(b) If terminated is true, then there is no n satisfying $\lfloor nx \rfloor + 1 - nx \leq \tau_{\min}$.

i. Set found$_{\min} \leftarrow$ true, and then go to Step 8.

8. If either of $\text{found}_{\min}$ or $\text{found}_{\max}$ is false, then go back to Step 4. Otherwise, $n_{0,\min}$ is the smallest minimizer and $n_{0,\max}$ is the largest maximizer.

At the end of Section 4, I remarked that we will retain the viewpoint of identifying all real numbers with their continued fraction expansions. However, note that in this viewpoint, it is not immediately clear how we can actually compute s and a in the above algorithm unless x is a rational number (so that the computations can be directly done in an arbitrary-precision exact rational arithmetic).

Note that it is not a good idea to use floating-point approximations of real numbers either, because the kind of computations we do are exactly the ones for which working with floating-point numbers can be tricky: subtracting and dividing numbers of comparable size. At least for me, it sounds very challenging to find out the exact error bound that will ensure the correct computation, for any given $[n_{\min} \colon n_{\max}]$.

To resolve this issue, we will develop two algorithms. The first algorithm, called *Gosper's algorithm*, which we develop in Section 12 after giving the explanation of the main algorithm of this whole note for establishing $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$, is an algorithm for computing the continued fraction expansion of any number obtained by combining two numbers x and y, with known continued fraction expansions, in a certain way. The way we combine x and y is general enough to cover all computations we need in Algorithm 5.4, so that those can be done with aribtrary-precision exact rational arithmetic.

However, executing this algorithm everytime in all of the iterations in Algorithm 5.4 would be very inefficient, which arises the desire for having the second algorithm, which computes good enough binary approximations of x and y of the form $\frac{m}{2^k}$ and $\frac{s}{2^k}$, where m, s are integers and k is a nonnegative integer, such that Algorithm 5.4 is guaranteed to give the same output if we replace (x, y) by $\left(\frac{m}{2^k}, \frac{s}{2^k}\right)$. If we can come up with such m, s, k, then arbitrary-precision exact rational arithmetic is enough to execute all the needed computations. This second algorithm is actually relying on the first algorithm, but the first algorithm only needs to be called just several, bounded amount of times.

We will develop this binary approximation algorithm in Section 8 and Section 9. Before we get there, we have to prepare ourselves with some more background works which we do in Section 6 and Section 7.

6 Lemmas for expanding the floor function

We begin with several simple lemmas about extremizers of $(nx+y) - \lfloor nx + y \rfloor$. Throughout this section, let x, y be any real numbers and let $n_{\min} \leq n_{\max}$ be any integers.

Lemma 6.1.
Let $n_0 \in \operatorname{argmin}_{n \in [n_{\min} \colon n_{\max}]} \left((nx + y) - \lfloor nx + y \rfloor \right)$. *Then for any* $n \in [n_{\min} - n_0 \colon n_{\max} - n_0]$,

$$\lfloor (n_0 + n)x + y \rfloor = \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor .$$

Proof. By the assumption on n_0, we have

$$((n_0 + n)x + y) - \lfloor (n_0 + n)x + y \rfloor \geq (n_0 x + y) - \lfloor n_0 x + y \rfloor ,$$

thus

$$\lfloor (n_0 + n)x + y \rfloor \leq \lfloor n_0 x + y \rfloor + nx < \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1.$$

Since $\lfloor (n_0 + n)x + y \rfloor$ should be equal to either $\lfloor n_0 x + y \rfloor + \lfloor nx \rfloor$ or $\lfloor n_0 x + y \rfloor + \lfloor nx \rfloor + 1$, the only possibility is

$$\lfloor (n_0 + n)x + y \rfloor = \lfloor n_0 x + y \rfloor + \lfloor nx \rfloor .$$

□

The intuition behind the above lemma is that, since n_0 achieves the minimum of the fractional part of $nx + y$, adding any nx into $n_0x + y$ should not make its fractional part to "wrap around", because at the exact moment of the wrapping, the fractional part will become strictly smaller than that of $n_0x + y$.

Lemma 6.2.
Let $n_0 \in \operatorname{argmax}_{n \in [n_{\min} :\, n_{\max}]} ((nx + y) - \lfloor nx + y \rfloor)$*. Then for any* $n \in [n_0 - n_{\max} :\, n_0 - n_{\min}]$*,*

$$\lfloor (n_0 - n)x + y \rfloor = \lfloor n_0x + y \rfloor - \lfloor nx \rfloor .$$

Proof. By the assumption on n_0, we have

$$((n_0 - n)x + y) - \lfloor (n_0 - n)x + y \rfloor \le (n_0x + y) - \lfloor n_0x + y \rfloor ,$$

thus

$$\lfloor (n_0 - n)x + y \rfloor \ge \lfloor n_0x + y \rfloor - nx > \lfloor n_0x + y \rfloor - \lfloor nx \rfloor - 1.$$

Since $\lfloor (n_0 - n)x + y \rfloor$ should be equal to either $\lfloor n_0x + y \rfloor - \lfloor nx \rfloor$ or $\lfloor n_0x + y \rfloor - \lfloor nx \rfloor - 1$, the only possibility is

$$\lfloor (n_0 - n)x + y \rfloor = \lfloor n_0x + y \rfloor - \lfloor nx \rfloor .$$

□

Similarly to the previous lemma, subtracting any nx from $n_0x + y$ should not make its fractional part to wrap around, because at the exact moment of the wrapping, the fractional part will become strictly larger than that of $n_0x + y$.

Lemma 6.3.
Let $n_0 = \min \operatorname{argmin}_{n \in [n_{\min} :\, n_{\max}]} ((nx + y) - \lfloor nx + y \rfloor)$*. Then for any* $n \in [1 :\, n_0 - n_{\min}]$*,*

$$\lfloor (n_0 - n)x + y \rfloor = \lfloor n_0x + y \rfloor - \lfloor nx \rfloor - 1.$$

Proof. By Lemma 6.1, we know

$$\lfloor (n_0 - n)x + y \rfloor = \lfloor n_0x + y \rfloor + \lfloor (-n)x \rfloor = \lfloor n_0x + y \rfloor - \lceil nx \rceil .$$

Note that if nx is integer, then we must have

$$((n_0 - n)x + y) - \lfloor (n_0 - n)x + y \rfloor = (n_0x + y) - \lfloor n_0x + y \rfloor ,$$

so by the minimality of n_0, nx is never an integer. Therefore, $\lceil nx \rceil = \lfloor nx \rfloor + 1$, so we get the desired conclusion. □

The intuition for the above lemma is that, if n_0 minimizes the fractional part of $nx + y$, then subtracting any nx from $n_0x + y$ should always make the fractional part wrap around, because otherwise the fractional part of $(n_0 - n)x + y$ will become strictly smaller than that of $n_0x + y$. Except when nx is an integer, in which case the fractional part stays the same, but since we have chosen the *smallest* minimizer, there is no n such that nx is an integer and $n_0 - n$ is still in the range $[n_{\min} :\, n_{\max}]$.

Lemma 6.4.
Let $n_0 = \max \operatorname{argmax}_{n \in [n_{\min} :\, n_{\max}]} ((nx + y) - \lfloor nx + y \rfloor)$*. Then for any* $n \in [1 :\, n_{\max} - n_0]$*,*

$$\lfloor (n_0 + n)x + y \rfloor = \lfloor n_0x + y \rfloor + \lfloor nx \rfloor + 1.$$

Proof. By Lemma 6.2, we know

$$\lfloor (n_0+n)x+y\rfloor = \lfloor n_0x+y\rfloor - \lfloor (-n)x\rfloor = \lfloor n_0x+y\rfloor + \lceil nx\rceil .$$

Note that if nx is integer, then we must have

$$((n_0+n)x+y) - \lfloor (n_0+n)x+y\rfloor = (n_0x+y) - \lfloor n_0x+y\rfloor ,$$

so by the maximality of n_0, nx is never an integer. Therefore, $\lceil nx\rceil = \lfloor nx\rfloor + 1$, so we get the desired conclusion. □

Similarly to the previous lemma, adding any nx to n_0x+y should make the fractional part wrap around (unless nx is an integer), because otherwise the fractional part of $(n_0+n)x+y$ should be larger than that of n_0x+y.

The following special cases of Lemma 6.1 and Lemma 6.4 will be also useful for our purpose.

Lemma 6.5.
Let $n_0 \in \operatorname{argmin}_{n\in[1:\,n_{\max}]}(nx - \lfloor nx\rfloor)$. *Then for any* $k,n\in\mathbb{Z}_{\geq 0}$ *such that* $kn_0+n\leq n_{\max}$,

$$\lfloor (kn_0+n)x\rfloor = k\lfloor n_0x\rfloor + \lfloor nx\rfloor .$$

Proof. We use induction on k. The base case $k=0$ is trivial, so we prove the induction step. By Lemma 6.1, we have

$$\lfloor (kn_0+n)x\rfloor = \lfloor n_0x\rfloor + \lfloor ((k-1)n_0+n)x\rfloor ,$$

thus by applying the induction hypothesis to expand $\lfloor ((k-1)n_0+n)x\rfloor$, we get the desired conclusion. □

Lemma 6.6.
Let $n_0 = \max\operatorname{argmax}_{n\in[1:\,n_{\max}]}(nx - \lfloor nx\rfloor)$. *Then for any* $k,n\in\mathbb{Z}_{\geq 0}$ *such that* $kn_0+n\leq n_{\max}$ *and* $n\neq 0$,

$$\lfloor (kn_0+n)x\rfloor = k(\lfloor n_0x\rfloor + 1) + \lfloor nx\rfloor .$$

Proof. We use induction on k. The base case $k=0$ is trivial, so we prove the induction step. By Lemma 6.4, we have

$$\lfloor (kn_0+n)x\rfloor = \lfloor n_0x\rfloor + \lfloor ((k-1)n_0+n)x\rfloor + 1,$$

thus by applying the induction hypothesis to expand $\lfloor ((k-1)n_0+n)x\rfloor$, we get the desired conclusion. □

7 Extremizers of $\frac{\lfloor nx\rfloor}{n}$ and $\frac{\lfloor nx\rfloor+1}{n}$

In this section, we solve a special case of our main problem: we establish the necessary and sufficient condition for ξ to have $\lfloor nx\rfloor = \lfloor n\xi\rfloor$ for all $n\in[1:\,n_{\max}]$. This is also one of the fundamental building blocks for solving our main problem.

In fact, I already described a solution for this special case in detail in [6], but here I reproduced it for convenience of both the readers and myself. Actually, some of the theorems stated here are slight improvements of the corresponding theorems in [6].

Theorem 7.1 (Extremizers of $\frac{\lfloor nx \rfloor}{n}$ and $\frac{\lfloor nx \rfloor + 1}{n}$)**.**
Let x be a real number and $n_{\max}$ a positive integer. Then we have the following relations between extremizers of $nx - \lfloor nx \rfloor$, $\frac{\lfloor nx \rfloor}{n}$, and $\frac{\lfloor nx \rfloor+1}{n}$:

1. *Let $n_0 \coloneqq \min \operatorname{argmin}_{n \in [1\colon n_{\max}]} (nx - \lfloor nx \rfloor)$, then*
$$\operatorname*{argmax}_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor}{n} = \{kn_0 \colon k \in \mathbb{Z}_{>0},\ kn_0 \le n_{\max}\}$$
and $\frac{\lfloor n_0 x \rfloor}{n_0}$ is a best rational approximation of x from below in its reduced form.

2. *Let $n_0 \coloneqq \max \operatorname{argmax}_{n \in [1\colon n_{\max}]} (nx - \lfloor nx \rfloor)$, then*
$$\operatorname*{argmin}_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n} = \{kn_0 \colon k \in \mathbb{Z}_{>0},\ kn_0 \le n_{\max}\}$$
and $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is in its reduced form. If $nx \notin \mathbb{Z}$ for all $n \in [1\colon n_{\max}]$, then $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is a best rational approximation of x from above.

Proof. 1. Theorem 2.4 immediately shows that $\frac{\lfloor n_0 x \rfloor}{n_0}$ is a best rational approximation of x from below in its reduced form. Choose any
$$n_1 \in \operatorname*{argmax}_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor}{n}$$
and let $\frac{p}{q}$ be the reduced form of $\frac{\lfloor n_1 x \rfloor}{n_1}$. Then for any rational number $\frac{a}{b}$ with $0 < b \le q$ and $\frac{a}{b} \le x$, we have
$$\frac{a}{b} \le \frac{\lfloor bx \rfloor}{b} \le \frac{\lfloor n_1 x \rfloor}{n_1} = \frac{p}{q} \le x,$$
thus $\frac{p}{q}$ is a best rational approximation of x from below.

Hence, if $n_0 \le q$, then
$$qx - p \le n_0 x - \lfloor n_0 x \rfloor$$
holds, but by definition of n_0, this must be an equality. Then by Theorem 2.4, q should be a multiple of n_0.

On the other hand, if $n_0 \ge q$, then
$$\frac{\lfloor n_0 x \rfloor}{n_0} \ge \frac{p}{q}$$
holds since $\frac{\lfloor n_0 x \rfloor}{n_0}$ is a best rational approximation of x from below in its reduced form, thus by definition of $\frac{p}{q}$, this must be an equality. Since both sides are in their reduced form, we conclude $q = n_0$.

Therefore, for any cases, we conclude that n_1 is a multiple of n_0. Converesely, any multiple of n_0 bounded above by $n_{\max}$ is a maximizer, because Lemma 6.5 shows that for any $k \in \mathbb{Z}_{>0}$ such that $kn_0 \le n_{\max}$, $\frac{\lfloor kn_0 x \rfloor}{kn_0} = \frac{\lfloor n_0 x \rfloor}{n_0}$.

2. We claim that $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is in its reduced form and any minimizer of $\frac{\lfloor nx \rfloor + 1}{n}$ is a multiple of n_0.

First, suppose that nx is an integer for some $n \in [1\colon n_{\max}]$, then Theorem 2.4 shows that n_0 is the largest integer in $[1\colon n_{\max}]$ with $n_0 p \equiv -1 \pmod{q}$. Choose any

$$n_1 \in \operatorname*{argmin}_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n}.$$

Then by rearranging the inequality

$$\frac{\lfloor n_1 x \rfloor + 1}{n_1} \leq \frac{\lfloor n_0 x \rfloor + 1}{n_0},$$

we get

$$n_0 \left((\lfloor n_1 x \rfloor + 1)q - n_1 p \right) \leq n_1 \left((\lfloor n_0 x \rfloor + 1)q - n_0 p \right).$$

Since

$$n_0 \left((\lfloor n_1 x \rfloor + 1)q - n_1 p \right) \equiv n_1 \left((\lfloor n_0 x \rfloor + 1)q - n_0 p \right) \pmod{q},$$

there exists a nonnegative integer e such that

$$n_1 \left((\lfloor n_0 x \rfloor + 1)q - n_0 p \right) = n_0 \left((\lfloor n_1 x \rfloor + 1)q - n_1 p \right) + eq.$$

Now, note that

$$(\lfloor n_0 x \rfloor + 1)q - n_0 p = \left(\frac{n_0 p - (q-1)}{q} + 1 \right) q - n_0 p = 1,$$

which in particular shows that $\lfloor n_0 x \rfloor + 1$ and n_0 are coprime. Similarly, we have

$$(\lfloor n_1 x \rfloor + 1)q - n_1 p = \left(\frac{n_1 p - (n_1 p \bmod q)}{q} + 1 \right) q - n_1 p = q - (n_1 p \bmod q) \geq 1,$$

thus we obtain

$$n_1 = (q - (n_1 p \bmod q))\, n_0 + eq \geq n_0 + eq,$$

but by definition of n_0, we must have $n_0 + q > n_{\max}$. Hence, $e = 0$ follows since $n_1 \leq n_{\max}$, which shows $n_1 = (q - (n_1 p \bmod q))\, n_0$.

Next, suppose that nx is not an integer for all $n \in [1\colon n_{\max}]$, then Theorem 2.4 shows that $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is a best rational approximation of x from above in its reduced form. Choose any

$$n_1 \in \operatorname*{argmin}_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n}$$

and let $\frac{p}{q}$ be the reduced form of $\frac{\lfloor n_1 x \rfloor + 1}{n_1}$. Then for any rational number $\frac{a}{b}$ with $0 < b \leq q$ and $\frac{a}{b} \geq x$, we have

$$\frac{a}{b} \geq \frac{\lceil bx \rceil}{b} = \frac{\lfloor bx \rfloor + 1}{b} \geq \frac{\lfloor n_1 x \rfloor + 1}{n_1} = \frac{p}{q} \geq x,$$

thus $\frac{p}{q}$ is a best rational approximation of x from above.

Hence, if $n_0 \leq q$, then

$$p - qx \leq \lfloor n_0 x \rfloor + 1 - n_0 x$$

holds, but since Theorem 2.4 shows that n_0 is the unique minimizer of $\lceil nx \rceil - nx = \lfloor nx \rfloor + 1 - nx = 1 - (nx - \lfloor nx \rfloor)$, we conclude that the above is an equality and $q = n_0$.

On the other hand, if $n_0 \geq q$, then

$$\frac{\lfloor n_0 x \rfloor + 1}{n_0} \leq \frac{p}{q}$$

holds since $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is a best rational approximation of x from above in its reduced form, thus by definition of $\frac{p}{q}$, this must be an equality. Since both sides are in their reduced form, we conclude $q = n_0$.

Therefore, for any cases, we conclude that $\frac{\lfloor n_0 x \rfloor + 1}{n_0}$ is in its reduced form and n_1 is a multiple of n_0, as claimed. Finally, any multiple of n_0 bounded above by $n_{\max}$ is a minimizer, because Lemma 6.6 shows that for any $k \in \mathbb{Z}_{>0}$ such that $kn_0 \leq n_{\max}$,

$$\frac{\lfloor kn_0 x \rfloor + 1}{kn_0} = \frac{(k-1)(\lfloor n_0 x \rfloor + 1) + \lfloor n_0 x \rfloor + 1}{kn_0} = \frac{\lfloor n_0 x \rfloor + 1}{n_0}.$$

□

Remark 7.2.
Somewhat counterintuitively, the set

$$\operatorname*{argmin}_{n \in [1 \colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n} = \{kn_0 \colon k \in \mathbb{Z}_{>0},\ kn_0 \leq n_{\max}\}$$

does not need to be consisting of only one element, regardless of if nx can be an integer or not. For instance, if $x = \frac{1}{7}$ and $n_{\max} = 12$, then the set above is $\{6, 12\}$, because 6 is the only element in $[1 \colon n_{\max}]$ whose remainder is 6 when divided by 7. Indeed, both 6 and 12 minimizes $\frac{\lfloor nx \rfloor + 1}{n}$, with the minimum value $\frac{1}{6}$. For another instance, let $x = \frac{2}{7}$ and $n_{\max} = 6$, then the set above is $\{3, 6\}$. Indeed, both 3 and 6 minimizes $\frac{\lfloor nx \rfloor + 1}{n}$, with the minimum value $\frac{1}{3}$.

As a corollary of the above theorem, we get the following result which is used in Section 10.

Corollary 7.3.
Let x be a real number and $n_{\max}$ be a positive integer. Then we have the followings.

1. *Let $n_0 \in \operatorname{argmax}_{n \in [1 \colon n_{\max}]} \frac{\lfloor nx \rfloor}{n}$, then for any $n \in [0 \colon n_{\max} - n_0]$,*

$$\lfloor (n_0 + n)x \rfloor = \lfloor n_0 x \rfloor + \lfloor nx \rfloor .$$

2. *Let $n_0 \in \operatorname{argmin}_{n \in [1 \colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n}$, then for any $n \in [1 \colon n_{\max} - n_0]$,*

$$\lfloor (n_0 + n)x \rfloor = \lfloor n_0 x \rfloor + \lfloor nx \rfloor + 1.$$

Proof. 1. By Theorem 7.1, we know $n_0 = kn_0'$ for some $k \in \mathbb{Z}_{>0}$ where n_0' is a minimizer of $nx - \lfloor nx \rfloor$ for $n \in [1 \colon n_{\max}]$. Hence, for any $n \in [0 \colon n_{\max} - n_0]$, Lemma 6.5 shows

$$\lfloor (n_0 + n)x \rfloor a = \lfloor (kn_0' + n)x \rfloor = k \lfloor n_0' x \rfloor + \lfloor nx \rfloor = \lfloor kn_0' x \rfloor + \lfloor nx \rfloor = \lfloor n_0 x \rfloor + \lfloor nx \rfloor .$$

2. By Theorem 7.1, we know $n_0 = kn_0'$ for some $k \in \mathbb{Z}_{>0}$ where n_0' is the largest maximizer of $nx - \lfloor nx \rfloor$ for $n \in [1\colon n_{\max}]$. Hence, for any $n \in [1\colon n_{\max} - n_0]$, Lemma 6.6 shows
$$\begin{aligned}\lfloor (n_0 + n)x \rfloor &= \lfloor (kn_0' + n)x \rfloor \\ &= k(\lfloor n_0'x \rfloor + 1) + \lfloor nx \rfloor \\ &= (k-1)(\lfloor n_0'x \rfloor + 1) + \lfloor n_0'x \rfloor + \lfloor nx \rfloor + 1 \\ &= \lfloor kn_0'x \rfloor + \lfloor nx \rfloor + 1 \\ &= \lfloor n_0x \rfloor + \lfloor nx \rfloor + 1.\end{aligned}$$

□

Theorem 7.1 immediately solves the aforementioned special case of our main problem:

Theorem 7.4.
Let x be a real number and $n_{\max}$ a positive integer. Then for a real number ξ, we have the followings.

1. *If x is rational and its reduced form $\frac{p}{q}$ satisfies $q \leq n_{\max}$, then we have*
$$\lfloor nx \rfloor = \lfloor n\xi \rfloor$$
for all $n \in [1\colon n_{\max}]$ if and only if
$$x \leq \xi < x + \frac{1}{vq}$$
holds, where v is the largest integer such that $vp \equiv -1 \pmod{q}$ and $v \leq n_{\max}$.

2. *If x is either irrational or rational whose reduced denominator is strictly larger than $n_{\max}$, then we have*
$$\lfloor nx \rfloor = \lfloor n\xi \rfloor$$
for all $n \in [1\colon n_{\max}]$ if and only if
$$\frac{p_*}{q_*} \leq \xi < \frac{p^*}{q^*}$$
holds, where $\frac{p_*}{q_*}, \frac{p^*}{q^*}$ are the best rational approximations of x from below and above, respectively, with the largest denominators $q_*, q^* \leq n_{\max}$.

Proof. Since having the equality $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ for all $n \in [1\colon n_{\max}]$ is equivalent to having the inequality
$$\max_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor}{n} \leq \xi < \min_{n \in [1\colon n_{\max}]} \frac{\lfloor nx \rfloor + 1}{n},$$
the conclusion immediately follow from Theorem 7.1 and Theorem 2.4. □

One of the important consequences of having $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ for all $n \in [1\colon n_{\max}]$ is that the fractional parts of nx and $n\xi$ should be ordered in the same way. This is also one of the results which I already have shown in [6]. The algorithm for finding a good enough approximation to use in Algorithm 5.4, which we will develop in Section 8, is crucially relying on this result.

Theorem 7.5 (Correspondence of fractional part ordering)**.**
Let x, ξ be real numbers and $n_{\max}$ a positive integer such that

$$\lfloor nx \rfloor = \lfloor n\xi \rfloor$$

holds for all $n \in [1\colon n_{\max}]$. Then we have the followings.

1. *If x is rational and its reduced form $\frac{p}{q}$ satisfies $q \leq n_{\max}$, then for any $n_1, n_2 \in [1\colon n_{\max}]$,*
$$n_1 x - \lfloor n_1 x \rfloor < n_2 x - \lfloor n_2 x \rfloor$$
or
$$n_1 x - \lfloor n_1 x \rfloor = n_2 x - \lfloor n_2 x \rfloor \quad \text{and} \quad n_1 < n_2$$
implies
$$n_1 \xi - \lfloor n_1 \xi \rfloor \leq n_2 \xi - \lfloor n_2 \xi \rfloor ,$$
with possibly the equality only when $\xi = x$.

2. *If x is either irrational or is rational whose reduced denominator is strictly larger than $n_{\max}$, then for any $n_1, n_2 \in [1\colon n_{\max}]$,*
$$n_1 x - \lfloor n_1 x \rfloor < n_2 x - \lfloor n_2 x \rfloor$$
implies
$$n_1 \xi - \lfloor n_1 \xi \rfloor \leq n_2 \xi - \lfloor n_2 \xi \rfloor ,$$
with possibly the equality only when $n_1 < n_2$ and $\xi = \frac{p_*}{q_*}$ is the best rational approximation of x from below with the largest denominator $q_* \leq n_{\max}$.

Note that when x is either irrational or rational with the denominator strictly greater than $n_{\max}$, then the mapping $n \mapsto nx - \lfloor nx \rfloor$ is one-to-one. Indeed, if $n_1 x - \lfloor n_1 x \rfloor = n_2 x - \lfloor n_2 x \rfloor$ holds with $n_1 \neq n_2$, then

$$x = \frac{\lfloor n_2 x \rfloor - \lfloor n_1 x \rfloor}{n_2 - n_1},$$

which is absurd.

Therefore, the theorem establishes the equivalence of two orderings on $[1\colon n_{\max}]$ induced from the embeddings

$$\begin{aligned} n &\mapsto (nx - \lfloor nx \rfloor , n) \quad \text{and} \\ n &\mapsto (n\xi - \lfloor n\xi \rfloor , n) \end{aligned}$$

into $[0,1) \times [1\colon n_{\max}]$ endowed with the lexicographic ordering.

Proof. 1. We want to show the inequality

$$(n_1 - n_2)\xi \leq \lfloor n_1 \xi \rfloor - \lfloor n_2 \xi \rfloor = \lfloor n_1 x \rfloor - \lfloor n_2 x \rfloor ,$$

which can be rewritten as

$$(n_1 - n_2)(\xi - x) \leq (n_2 x - \lfloor n_2 x \rfloor) - (n_1 x - \lfloor n_1 x \rfloor).$$

Since $x = \frac{p}{q}$ and $q \leq n_{\max}$, Theorem 7.4 implies

$$x \leq \xi < x + \frac{1}{vq}$$

where v is the largest integer such that $vp \equiv -1 \pmod{q}$ and $v \leq n_{\max}$. In particular, the inequality we want to show is trivial if $n_1 < n_2$ or $x = \xi$, so assume $n_1 > n_2$ and $x \neq \xi$, then we have

$$(n_1 - n_2)(\xi - x) < \frac{n_1 - n_2}{vq}.$$

Let r_1, r_2 be the remainders of $n_1 p$ divided by q and $n_2 p$ divided by q, respectively, then it suffices to show

$$(n_2 x - \lfloor n_2 x \rfloor) - (n_1 x - \lfloor n_1 x \rfloor) = \frac{r_2 - r_1}{q} \geq \frac{n_1 - n_2}{vq},$$

or equivalently,

$$n_1 - n_2 \leq v(r_2 - r_1),$$

assuming $r_2 > r_1$. Suppose not, so assume $n_1 - n_2 > v(r_2 - r_1)$. Note that

$$v(r_2 - r_1)p \equiv r_1 - r_2 \equiv (n_1 - n_2)p \pmod{q},$$

and since p and q are coprime, we get $v(r_2 - r_1) \equiv n_1 - n_2 \pmod{q}$, thus there exists a positive integer e such that

$$n_1 - n_2 = v(r_2 - r_1) + eq.$$

Since $r_2 > r_1$, this implies

$$n_{\max} \geq n_1 = v(r_2 - r_1) + eq + n_2 > v + q$$

which contradicts to the definition of v. Therefore, we get the desired inequality.

Note that since $n_1 - n_2 \leq v(r_2 - r_1)$ always holds whenever $n_1 > n_2$ and $r_2 > r_1$, in that case it is impossible to achieve the equality $n_1\xi - \lfloor n_1 \xi \rfloor = n_2 \xi - \lfloor n_2 \xi \rfloor$. Hence, the equality might hold only when $n_1 \leq n_2$. Then one can easily see that $\xi = x$ is a necessary condition for having the equality.

2. Let $\frac{p_*}{q_*}$ be the best rational approximation of x from below with the largest denominator $q_* \leq n_{\max}$. Applying Theorem 7.4 with $\xi \leftarrow \frac{p_*}{q_*}$, we immediately get that $\left\lfloor \frac{np_*}{q_*} \right\rfloor = \lfloor nx \rfloor = \lfloor n\xi \rfloor$ holds for all $n \in [1\colon n_{\max}]$. Note that from $n_1 x - \lfloor n_1 x \rfloor < n_2 x - \lfloor n_2 x \rfloor$, we get

$$n_1 \left(x - \frac{p_*}{q_*} \right) + \frac{n_1 p_*}{q_*} - \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor < n_2 \left(x - \frac{p_*}{q_*} \right) + \frac{n_2 p_*}{q_*} - \left\lfloor \frac{n_2 p_*}{q_*} \right\rfloor. \tag{3}$$

We claim that

$$\frac{n_1 p_*}{q_*} - \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor \leq \frac{n_2 p_*}{q_*} - \left\lfloor \frac{n_2 p_*}{q_*} \right\rfloor.$$

This actually follows directly from the first part of the theorem, because if this is not the case, then the first part applied to $x \leftarrow \frac{p_*}{q_*}$ and $\xi \leftarrow x$ implies

$$n_1 x - \lfloor n_1 x \rfloor \geq n_2 x - \lfloor n_2 x \rfloor,$$

directly contradicting to the assumption.

Now, if

$$\frac{n_1 p_*}{q_*} - \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor < \frac{n_2 p_*}{q_*} - \left\lfloor \frac{n_2 p_*}{q_*} \right\rfloor$$

holds, then again we can apply the first part with $x \leftarrow \frac{p_*}{q_*}$ and $\xi \leftarrow \xi$ to conclude

$$n_1\xi - \lfloor n_1\xi \rfloor < n_2\xi - \lfloor n_2\xi \rfloor ;$$

note that the equality cannot hold in this case. Therefore, we only need to consider the case

$$\frac{n_1 p_*}{q_*} - \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor = \frac{n_2 p_*}{q_*} - \left\lfloor \frac{n_2 p_*}{q_*} \right\rfloor .$$

Clearly, in this case (3) shows that we should have $n_1 < n_2$, thus we can again apply the first part to conclude

$$n_1\xi - \lfloor n_1\xi \rfloor \leq n_2\xi - \lfloor n_2\xi \rfloor ,$$

with possibly the equality when $\xi = \frac{p_*}{q_*}$.

□

8 Simultaneously satisfying $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ and $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$

In this section, we now develop an algorithm for coming up with an approximation (ξ, ζ) of (x, y) such that Algorithm 5.4 should return the same outputs for (x, y) and (ξ, ζ). By how the algorithm is derived, one can see that a sufficient condition on (ξ, ζ) guaranteeing this equivalence is that the inequality

$$nx - \lfloor nx \rfloor \geq \lfloor mx + y \rfloor + 1 - (mx + y) \tag{4}$$

holds for some $m, n \in \mathbb{Z}$ with $m \geq n_{\min}$, $n > 0$ and $m + n \leq n_{\max}$ if and only if the same inequality holds with (x, y) replaced by (ξ, ζ). Indeed, recall that the way we iteratively approached to the smallest minimizer and the largest maximizer is to successively add to m the smallest n satisfying/violating the above inequality. Hence, if the set of (m, n) satisfying/violating the inequality is same for (x, y) and (ξ, ζ), then Algorithm 5.4 should return the same outputs for them.

One should however note that this does not mean that the *sets* of minimizers/maximizers for (x, y) and (ξ, ζ) should be equal, rather, it merely says that the *smallest minimizer* and the *largest maximizer* should be equal. Indeed, if nx is an integer for certain n's, then adding or subtracting n to any minimizer/maximizer should still give a minimizer/maximizer, but the corresponding $n\xi$ does not need to be an integer, because the only thing the condition that nx is an integer imposes on (ξ, ζ) is that the inequality $n\xi - \lfloor n\xi \rfloor < \lfloor m\xi + \zeta \rfloor + 1 - (m\xi + \zeta)$ is satisfied for all m, which does not necessarily mean that the left-hand side is zero. This is one of the reasons why we only care about the smallest minimizer and the largest maximizer, rather than the whole sets of minimizers/maximizers. They are much easier to compute than the whole sets.

The inequality (4) is equivalent to

$$(m + n)x + y \geq \lfloor nx \rfloor + \lfloor mx + y \rfloor + 1,$$

which is equivalent to

$$\lfloor (m+n)x + y \rfloor \geq \lfloor nx \rfloor + \lfloor mx + y \rfloor + 1.$$

Therefore, a sufficient condition on (ξ, ζ) we are seeking for is:

1. $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ holds for all $n \in [n_{\min} \colon n_{\max}]$ and
2. $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ holds for all $n \in [0 \colon n_{\max} - n_{\min}]$.

It might be puzzling to some of the readers, because finding a condition for having $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ was our main goal from the first place. However, our main goal was to establish a necessary and sufficient condition for having *only* $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for all $n \in [n_{\min} \colon n_{\max}]$. On the other hand, here it is enough to find a sufficient condition (although we will in fact obtain a necessary and sufficient condition) for having *both* of the above two. They are different problems.

Note that having $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for all $n \in [n_{\min} \colon n_{\max}]$ can be rewritten as

$$\lfloor nx + (n_{\min}x + y) \rfloor = \lfloor n\xi + (n_{\min}\xi + \zeta) \rfloor$$

for all $n \in [0 \colon n_{\max} - n_{\min}]$. Hence, let $y' \coloneqq (n_{\min}x + y) - \lfloor n_{\min}x + y \rfloor$ and $\zeta' \coloneqq (n_{\min}\xi + \zeta) - \lfloor n_{\min}\xi + \zeta \rfloor$, then we can reformulate our problem as finding $\xi \in \mathbb{R}$, $\zeta' \in [0, 1)$ which satisfy

$$\lfloor nx + y' \rfloor = \lfloor n\xi + \zeta' \rfloor$$

for all $n \in [1 \colon n_{\max} - n_{\min}]$ for given $x \in \mathbb{R}$, $y' \in [0, 1)$. Once we find such ξ, ζ', then we can recover ζ from the identity

$$\zeta = \zeta' + \lfloor n_{\min}x + y \rfloor - n_{\min}\xi$$

which also recovers the condition $\lfloor n_{\min}x + y \rfloor = \lfloor n_{\min}\xi + \zeta \rfloor$.

We will solve this reformulated problem using the following result, which is again from [6] with some minor modifications.

Theorem 8.1 (Simultaneously Satisfying $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ and $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$)**.** *Let x, ξ be real numbers and $n_{\max}$ a positive integer such that*

$$\lfloor nx \rfloor = \lfloor n\xi \rfloor$$

holds for all $n \in [1 \colon n_{\max}]$, and let $\frac{p_*}{q_*}, \frac{p^*}{q^*}$ be the best rational approximations of x from below and above, respectively, with the largest denominators $q_*, q^* \leq n_{\max}$. Given $y \in [0, 1]$, define

$$L \coloneqq \{n \in [1 \colon n_{\max}] \colon nx - \lfloor nx \rfloor < 1 - y\},$$
$$R \coloneqq \{n \in [1 \colon n_{\max}] \colon nx - \lfloor nx \rfloor \geq 1 - y\}.$$

Then for a real number ζ, we have the followings.

1. *When $L = \emptyset$,*

   $$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$$

   holds for all $n \in [1 \colon n_{\max}]$ if and only if

   $$p_* + 1 - q_*\xi \leq \zeta < \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + 2 - \mu\xi,$$

   where μ is defined as follows.

(1) *If $x = \frac{p}{q}$ is a rational number with $q \le n_{\max}$, then μ is the largest integer such that $\mu \le n_{\max}$ and $\mu p \equiv -1 \pmod{q}$.*

(2) *If x is either an irrational number or a rational number with the denominator strictly greater than $n_{\max}$, then $\mu = q^*$.*

2. *When $R = \emptyset$,*

$$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$$

holds for all $n \in [1\colon n_{\max}]$ if and only if

$$p_* - q_*\xi \le \zeta < \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + 1 - \mu\xi,$$

where μ is defined as in the case $L = \emptyset$.

3. *When L and R are both nonempty,*

$$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$$

holds for all $n \in [1\colon n_{\max}]$ if and only if

$$\left\lfloor \frac{\nu p_*}{q_*} \right\rfloor + 1 - \nu\xi \le \zeta < \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + 1 - \mu\xi,$$

where μ and ν are defined as follows.

(1) *If x is rational and its reduced denominator is at most $n_{\max}$, then μ is the largest integer such that $\mu \le n_{\max}$ and $\mu p_* \equiv -\lfloor q_* y \rfloor - 1 \pmod{q_*}$, and ν is the smallest positive integer such that $\nu p_* \equiv -\lfloor q_* y \rfloor \pmod{q_*}$.*

(2) *If x is either irrational or is rational whose reduced denominator is strictly larger than $n_{\max}$, then μ is the largest positive integer such that $\mu \le n_{\max}$,*

$$\mu p_* \equiv r \pmod{q_*} \quad \text{and} \quad \mu < \frac{q_*(1-y) - r}{q_* x - p_*},$$

where $r \in [0\colon q_* - 1]$ is chosen to be the largest allowing such μ to exist. Similarly, ν is the smallest positive integer such that $\nu \le n_{\max}$,

$$\nu p_* \equiv r \pmod{q_*} \quad \text{and} \quad \nu \ge \frac{q_*(1-y) - r}{q_* x - p_*},$$

where $r \in [0\colon q_* - 1]$ is chosen to be the smallest allowing such ν to exist.

Proof. Note that

$$\lfloor nx + y \rfloor = \begin{cases} \lfloor nx \rfloor & \text{if } nx - \lfloor nx \rfloor < 1 - y, \\ \lfloor nx \rfloor + 1 & \text{if } nx - \lfloor nx \rfloor \ge 1 - y, \end{cases}$$

so we have the equality

$$\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$$

for all $n \in [1\colon n_{\max}]$ if and only if

$$L = L' \coloneqq \{n \in [1\colon n_{\max}]\colon\ -\zeta \le n\xi - \lfloor n\xi \rfloor < 1 - \zeta\}$$

and

$$R = R' \coloneqq \{n \in [1\colon n_{\max}] \colon 1 - \zeta \le n\xi - \lfloor n\xi \rfloor < 2 - \zeta\} .$$

Now, as $L \cup R = [1\colon n_{\max}]$ and $L' \cap R' = \emptyset$, we have $L = L'$ and $R = R'$ if and only if we have $L \subseteq L'$ and $R \subseteq R'$. Note that $L \subseteq L'$ holds if and only if either $L = \emptyset$ or

$$-\min_{n\in L}(n\xi - \lfloor n\xi \rfloor) \le \zeta < 1 - \max_{n\in L}(n\xi - \lfloor n\xi \rfloor),$$

and $R \subseteq R'$ holds if and only if either $R = \emptyset$ or

$$1 - \min_{n\in R}(n\xi - \lfloor n\xi \rfloor) \le \zeta < 2 - \max_{n\in R}(n\xi - \lfloor n\xi \rfloor).$$

1. When $L = \emptyset$, we only need to verify

   $$\max_{n\in[1\colon n_{\max}]}(n\xi - \lfloor n\xi \rfloor) = \mu\xi - \lfloor \mu\xi \rfloor$$

   and

   $$\min_{n\in[1\colon n_{\max}]}(n\xi - \lfloor n\xi \rfloor) = q_*\xi - \lfloor q_*\xi \rfloor ,$$

   since $\lfloor n\xi \rfloor = \lfloor nx \rfloor = \left\lfloor \frac{np_*}{q_*} \right\rfloor$ holds for all n by Theorem 7.4. These two both follow directly from Theorem 7.5 and Theorem 2.4, because q_* is the smallest minimizer of $nx - \lfloor nx \rfloor$ and μ is the largest maximizer of $nx - \lfloor nx \rfloor$ for $n \in [1\colon n_{\max}]$.

2. Similarly, when $R = \emptyset$, we only need to verify

   $$\max_{n\in[1\colon n_{\max}]}(n\xi - \lfloor n\xi \rfloor) = \mu\xi - \lfloor \mu\xi \rfloor$$

   and

   $$\min_{n\in[1\colon n_{\max}]}(n\xi - \lfloor n\xi \rfloor) = q_*\xi - \lfloor q_*\xi \rfloor ,$$

   which we have already done in the case $L = \emptyset$.

3. Now, suppose L and R are both not empty. Then it suffices to show that

   $$\mu\xi - \lfloor \mu\xi \rfloor = \max_{n\in L}(n\xi - \lfloor n\xi \rfloor)$$

   and

   $$\nu\xi - \lfloor \nu\xi \rfloor = \min_{n\in R}(n\xi - \lfloor n\xi \rfloor).$$

   By Theorem 7.5, it suffices to show that μ is the largest maximizer of $nx - \lfloor nx \rfloor$ for $n \in L$ and ν is the smallest minimizer of $nx - \lfloor nx \rfloor$ for $n \in R$.

   When x is rational whose reduced denominator is at most $n_{\max}$, then we have $x = \frac{p_*}{q_*}$. Then since $n \in R$ if and only if $(np_* \mod q_*) \ge q_* - \lfloor q_*y \rfloor$, the conclusion about ν follows immediately. For μ, note that $\lfloor q_*y \rfloor < q_*$ always holds unless $y = 1$, but $y = 1$ is impossible since it implies $L = \emptyset$. Hence, $q_* - \lfloor q_*y \rfloor - 1 \ge 0$ holds, so we have $n \in L$ if and only if $(np_* \mod q_*) \le q_* - \lfloor q_*y \rfloor - 1$, so the conclusion about μ also follows.

   Next, assume x is either irrational or rational whose reduced denominator strictly larger than $n_{\max}$. In this case, we can apply Theorem 7.5 to see that the largest maximizer of $nx - \lfloor nx \rfloor$ for $n \in L$ is precisely the largest maximizer of $\frac{np_*}{q_*} - \left\lfloor \frac{np_*}{q_*} \right\rfloor$ for $n \in L$. Since

$$\frac{np_*}{q_*} - \left\lfloor \frac{np_*}{q_*} \right\rfloor = \frac{(np_* \mod q_*)}{q_*},$$

the maximum value of $\frac{np_*}{q_*} - \left\lfloor \frac{np_*}{q_*} \right\rfloor$ is obtained at $n \in L$ such that np_* has the largest possible remainder when divided by q_*. In general, for $n \in [1\colon n_{\max}]$, if we write $r := (np_* \mod q_*)$, then

$$nx - \lfloor nx \rfloor = n\left(x - \frac{p_*}{q_*}\right) + \left(\frac{np_*}{q_*} - \left\lfloor \frac{np_*}{q_*} \right\rfloor\right) = n\left(x - \frac{p_*}{q_*}\right) + \frac{r}{q_*},$$

so $x \in L$ if and only if

$$n\left(x - \frac{p_*}{q_*}\right) + \frac{r}{q_*} < 1 - y.$$

or equivalently,

$$n < \frac{q_*(1-y) - r}{q_*x - p_*},$$

which gives the desired desription of μ. We can obtain the description of ν in a similar way.

□

Note that Theorem 7.4 gives us a criterion for having $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ that is easy to verify, so using the above theorem in conjuction with Theorem 7.4 can give us a *necessary and sufficient* condition for having $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ and $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ for all $n \in [1\colon n_{\max}]$ given $y, \zeta \in [0, 1)$, even though we only needed a sufficient condition.

When either of L or R is empty, or when x is rational whose reduced denominator is at most $n_{\max}$, then we can easily compute the complete set of (ξ, ζ) using Theorem 7.4 and Theorem 8.1, given that we know how to compute $\lfloor q_* y \rfloor$. This can be indeed done by using the algorithm we develop in Section 12.

However, when both L and R are not empty and when x is either irrational or is rational but the reduced denominator is strictly larger than $n_{\max}$, then finding μ, ν appearing in the statement of Theorem 8.1 by successively checking all r can be very inefficient, as q_* can be in general as large as $n_{\max}$.[7] Hence, we describe an algorithm for efficiently finding out the optimal r with several bounded number of tries.

We consider the case of μ first. Let q be the modular inverse of $-p_*$ with respect to q_*. In fact, it can be easily shown that q must be the denominator of the odd convergent of x with the largest denominator bounded by q_*. Indeed, since we are considering the case when x is either irrational or is rational whose reduced denominator is strictly larger than $n_{\max}$, q_* should be a proper even semiconvergent of x. Hence, either $q_* = q_0 = 1$ or we have $p_* = p_{2i} + sp_{2i+1}$ and $q_* = q_{2i} + sq_{2i+1}$ for some i and an integer $s \in (0, a_{2i+2}]$, where $\frac{p_i}{q_i}$ is the ith convergent and a_i is the ith continued fraction coefficient of x. The case $q_* = 1$

[7] In practice, I observed that the optimal r in most cases were around $q_* - \lfloor q_* y \rfloor$, often off by small numbers. However, I was not able to prove that this should always be the case, I am suspecting that this was merely because the continued fraction coefficient of x I experimented with were mostly just small numbers. In the description of the algorithm that follows, it becomes apparent that having a big continued fraction coefficient can result in making q_* a lot larger than q, which can shift the optimal r a lot from its obvious rough guess.

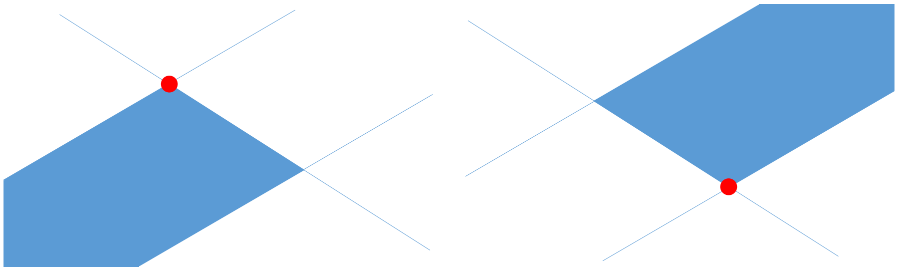

Figure 2: Left: (k, r)-region for μ, Right: (k, r)-region for ν

is actually impossible by our assumption on x, so in particular we deduce $q_* > q_{2i+1}$. Note that

$$p_{2i+1}q_* - q_{2i+1}p_* = p_{2i+1}(q_{2i} + sq_{2i+1}) - q_{2i+1}(p_{2i} + sp_{2i+1}) = 1,$$

so indeed we get that $q = q_{2i+1}$ is the modular inverse of $-p_*$ with respect to q_*.

Now, consier the set of $(k, r) \in \mathbb{Z} \times [0\colon q_* - 1]$ such that

$$0 < q_*k - qr \leq n_{\max},$$

that is, we decompose $[1\colon n_{\max}]$ into set of n's such that $np_* \equiv r \pmod{q_*}$ for each r, which means $n \equiv -qr \pmod{q_*}$, and then let $k := \frac{n+qr}{q_*}$. Then the constraint

$$n < \frac{q_*(1-y) - r}{q_*x - p_*}$$

appearing in the description of μ can be equivalently written as

$$(1 - q(q_*x - p_*))r < q_*(1-y) - q_*(q_*x - p_*)k$$

by substituting $n \leftarrow q_*k - qr$. Since

$$q\left(q_*x - p_*\right) = qq_*\left(x - \frac{p_*}{q_*}\right) < qq_*\left(\frac{p_{2i+1}}{q_{2i+1}} - \frac{p_*}{q_*}\right) = 1,$$

the inequality can be further rewritten as

$$r < \frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)}.$$

Next, let us ignore the constraint $(k, r) \in \mathbb{Z} \times [0\colon q_* - 1]$ for a while and consider the set of $(k, r) \in \mathbb{R} \times \mathbb{R}$ satisfying three inequalities

$$0 < q_*k - qr \leq n_{\max} \quad \text{and} \quad r < \frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)} \tag{5}$$

as shown in the left-hand side of Figure 2.

The point marked in red is the highest possible value of r we can have in this region, once we ignore the constraint $(k, r) \in \mathbb{Z} \times [0\colon q_* - 1]$. We first evaluate the r-coordinate r_T of this point. This can be done by solving the equation

$$\frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)} = \frac{q_*k}{q},$$

for k. The solution is

$$k_T := q(1-y)$$

and as a result we obtain

$$r_T = q_*(1-y).$$

Therefore, the maximum possible value of r is at most $\lceil r_T \rceil - 1$. We start from there and go further down if there is no $k \in \mathbb{Z}$ such that $(k, \lceil r_T \rceil - 1)$ is in the region. Note that r_T is clearly strictly smaller than q_* unless $y = 0$, in which case $L = \emptyset$ contradicting to our assumption. Also, if we cannot find any $(k,r) \in \mathbb{Z} \times \mathbb{Z}$ until we further go down below $r = 0$, then this means that the actual set of (k,r) is empty, and that means $L = \emptyset$ which is again impossible. Hence, we only need to take account of the integer constraint $(k,r) \in \mathbb{Z} \times \mathbb{Z}$, and can ignore the range constraint $0 \le r \le q_* - 1$.

We now describe this search procedure for (k,r). We first decompose the region defined by three inequalities in (5) into two, the top triangular region and the bottom parallelogrammatic region extending indefinitely into the left-bottom direction. These two regions are separated by the horizontal line

$$r = r_B = q_*(1-y) - (q_*x - p_*)n_{\max},$$

where r_B is the r-coordinate of the intersection between the lines

$$q_*k - qr = 0 \quad \text{and} \quad r = \frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)}.$$

We note that the parallelogrammatic region is horizontally wide enough to carry a point of integer k-coordinate for every possible r-coordinate. Indeed, for given r, the horizontal gap between the two bounding lines is

$$\frac{n_{\max} + qr}{q_*} - \frac{qr}{q_*} = \frac{n_{\max}}{q_*}. \tag{6}$$

By the assumption on x, $n_{\max}$ is strictly larger than q_*, so the right-hand side is strictly larger than 1. Hence, there must be a point of the form (k,r) in this region where $k \in \mathbb{Z}$. Assuming that we include the horizontal line $r = r_B$ into the triangular region exclude from the parallelogrammatic region, the largest integer k-coordinate of such points can be always obtained by taking the floor of the k-coordinate of the right-most point in the intersection with $r = r_B$.

There are several cases we have to deal separately.

1. First, suppose $\lceil r_T \rceil - 1 < r_B$, that is, $\lceil r_T \rceil - 1 < \lceil r_B \rceil$. Hence, the horizontal line $r = \lceil r_T \rceil - 1$ crosses the parallelogrammatic region. The k-coordinate of the right-most point of the intersection is

   $$k_{PR} := \frac{n_{\max} + q(\lceil r_T \rceil - 1)}{q_*} = \frac{n_{\max} + q(\lceil r_B \rceil - 1)}{q_*}$$

   and recall that by (6) the intersection is guaranteed to have a point with an integer k-coordinate. Since μ is the largest n satisfying $np_* \equiv r \pmod{q_*}$, we return $(\lfloor k_{PR} \rfloor, \lceil r_B \rceil - 1)$.

2. Next, suppose $\lceil r_T \rceil - 1 \ge \lceil r_B \rceil$, that is, the horizontal line $r = \lceil r_T \rceil - 1$ crosses the triangular region. The k-coordinate of the right-most point of the intersection is

   $$k_{TR} := \frac{q_*(1-y) - (1 - q(q_*x - p_*))(\lceil r_T \rceil - 1)}{q_*(q_*x - p_*)}$$

and the k-coordinate of the left-most point is

$$k_{TL} := \frac{q(\lceil r_T \rceil - 1)}{q_*},$$

both exclusive. If there is a point in the intersection with an integer k-coordinate, then the one with the largest k-coordinate must be the one whose k-coordinate is equal to $\lceil k_{TR} \rceil - 1$. Hence, if $\lceil k_{TR} \rceil - 1 > k_{TL}$, that is, $\lceil k_{TR} \rceil - 1 > \lfloor k_{TL} \rfloor$, then we return $(\lceil k_{TR} \rceil - 1, \lceil r_T \rceil - 1)$.

3. If $\lceil k_{TR} \rceil - 1 \leq \lfloor k_{TL} \rfloor$, then there is no point in the intersection with an integer k-coordinate. In this case we claim that, if there exists any point (k, r) in the triangular region with integer coordinates, then at least one of $(\lceil k_{TR} \rceil, r) = (\lfloor k_{TL} \rfloor + 1, r)$ or $(\lfloor k_{TL} \rfloor, r)$ must be in the triangular region. Indeed, assume $k \geq \lceil k_{TR} \rceil$, then clearly we have

$$r < \frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)} \leq \frac{q_*(1-y) - q_*(q_*x - p_*)\lceil k_{TR} \rceil}{1 - q(q_*x - p_*)}.$$

On the other hand, since we know $k_{TR} > k_T$ from $r_T > \lceil r_T \rceil - 1$ (recall k_T is the k-coordinate of the point marked in red in the lefh-hand side of Figure 2), thus

$$q_* \lceil k_{TR} \rceil - qr > q_* k_T - qr > q_* k_T - q r_T = 0,$$

which shows $(\lceil k_{TR} \rceil, r)$ is in the triangular region. Next, assume $k \leq \lfloor k_{TL} \rfloor$, then clearly we have

$$0 < q_* k - qr \leq q_* \lfloor k_{TL} \rfloor - qr.$$

On the other hand, since we know $k_{TL} < k_T$ from $r_T > \lceil r_T \rceil - 1$, thus

$$\frac{q_*(1-y) - q_*(q_*x - p_*)\lfloor k_{TL} \rfloor}{1 - q(q_*x - p_*)} > \frac{q_*(1-y) - q_*(q_*x - p_*)k_T}{1 - q(q_*x - p_*)} = r_T > r,$$

which shows $(\lfloor k_{TL} \rfloor, r)$ is in the triangular region. Hence, the claim is shown.

Therefore, we see if there is any point in the triangular region with integer coordinates whose k-coordinate is either $\lceil k_{TR} \rceil$ or $\lfloor k_{TL} \rfloor$. If not, then we conclude there is no point with integer coordinates in the triangular region. To do this check, let

$$r_{TR} := \frac{q_*(1-y) - q_*(q_*x - p_*)\lceil k_{TR} \rceil}{1 - q(q_*x - p_*)} \quad \text{and} \quad r_{TL} := \frac{q_* \lfloor k_{TL} \rfloor}{q},$$

the r-coordinates of the intersection of the vertical line $k = \lceil k_{TR} \rceil$ with the right-boundary of the triangular region, and the intersection of the vertical line $k = \lfloor k_{TL} \rfloor$ with the left-boundary of the triangular region, respectively.

If $\lceil r_{TR} \rceil - 1 < \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 < \lceil r_B \rceil$, then we conclude there is no point with integer coordinates in the triangular region. In this case, we simply choose the largest integer r-coordinate $\lceil r_B \rceil - 1$ in the parallelogrammatic region. That is, let

$$k_{PR} := \frac{n_{\max} + q(\lceil r_B \rceil - 1)}{q_*},$$

and we return $(\lfloor k_{PR} \rfloor, \lceil r_B \rceil - 1)$.

4. If $\lceil r_{TR} \rceil - 1 \geq \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 < \lceil r_B \rceil$, then there exists at least one point in the triangular region with integer coordinates, and there must exist a point whose k-coordinate is equal to $\lceil k_{TR} \rceil$ among the ones whose integer r-coordinate is the largest. Hence, we simply choose the one with the largest integer k-coordinate among the points on the intersection of the triangular region and the horizontal line $r = \lceil r_{TR} \rceil - 1$. That is, let
$$k_{TRR} \coloneqq \frac{q_*(1-y) - (1 - q(q_*x - p_*))(\lceil r_{TR} \rceil - 1)}{q_*(q_*x - p_*)}$$
and return $(\lceil k_{TRR} \rceil - 1, \lceil r_{TR} \rceil - 1)$.

5. If $\lceil r_{TR} \rceil - 1 < \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 \geq \lceil r_B \rceil$, then there exists at least one point in the triangular region with integer coordinates, and there must exist a point whose k-coordinate is equal to $\lfloor k_{TL} \rfloor$ among the ones whose integer r-coordinate is the largest. Hence, we simply choose the one with the largest integer k-coordinate among the points on the intersection of the triangular region and the horizontal line $r = \lceil r_{TL} \rceil - 1$. That is, let
$$k_{TLR} \coloneqq \frac{q_*(1-y) - (1 - q(q_*x - p_*))(\lceil r_{TL} \rceil - 1)}{q_*(q_*x - p_*)}$$
and return $(\lceil k_{TLR} \rceil - 1, \lceil r_{TL} \rceil - 1)$.

6. If $\lceil r_{TR} \rceil - 1 \geq \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 \geq \lceil r_B \rceil$, then we choose the larger one between $\lceil r_{TR} \rceil - 1$ and $\lceil r_{TL} \rceil - 1$ and do the same. That is, if $\lceil r_{TR} \rceil - 1 \geq \lceil r_{TL} \rceil - 1$ then return $(\lceil k_{TRR} \rceil - 1, \lceil r_{TR} \rceil - 1)$, and otherwise return $(\lceil k_{TLR} \rceil - 1, \lceil r_{TL} \rceil - 1)$.

After following this procedure, we get a point $(k, r) \in \mathbb{Z} \times [0\colon q_* - 1]$, and we conclude $\mu = q_*k - qr$.

Next, we consider the case of ν. The only difference in this case is that, now the region we are looking at is defined by
$$0 < q_*k - qr \leq n_{\max} \quad \text{and} \quad r \geq \frac{q_*(1-y) - q_*(q_*x - p_*)k}{1 - q(q_*x - p_*)} \tag{7}$$
and we are interested in finding the bottom-most and left-most point (prioritizing the first) in the region with integer coordinates. Again, we start from the point marked in red in the right-hand side of Figure 2. However, in this case, this point, whose coordinates (k_B, r_B) are given as
$$k_B = q(1-y) + \frac{(1 - q(q_*x - p_*))n_{\max}}{q_*} \quad \text{and} \quad r_B = q_*(1-y) - (q_*x - p_*)n_{\max},$$
may have negative r-coordinate. Hence, instead of $\lceil r_B \rceil$ we start with
$$r'_B \coloneqq \max\{\lceil r_B \rceil, 0\}.$$
On the other hand, once the point (k, r) with the smallest-coordinate in the region defined by these constraints is found, then r must be no more than $q_* - 1$, because otherwise we conclude $R = \emptyset$, contradicting the assumption.

Except for that inclusivity of boundary lines are different and now we are seeking for the left-most rather than the right-most point, the procedure of finding (k, r) is largely same as in the case of μ:

1. First, suppose $r'_B \geq r_T$, that is, $r'_B \geq \lceil r_T \rceil$ which actually means $r'_B = \lceil r_T \rceil$ since $r_T = q_*(1-y) \geq 0$. Hence, the horizontal line $r = r'_B$ crosses the parallelogrammatic region. The k-coordinate of the left-most point of the intersection is

$$k_{PL} \coloneqq \frac{qr'_B}{q_*} = \frac{q\lceil r_T \rceil}{q_*},$$

and we return $(\lfloor k_{PL} \rfloor + 1, \lceil r_T \rceil)$.

2. Next, suppose $r'_B < \lceil r_T \rceil$, that is, the horizontal line $r = r'_B$ crosses the triangular region. The k-coordinate of the left-most point of the intersection is

$$k_{BL} \coloneqq \frac{q_*(1-y) - (1 - q(q_*x - p_*))r'_B}{q_*(q_*x - p_*)}$$

and the k-coordinate of the right-most point is

$$k_{BR} \coloneqq \frac{n_{\max} + qr'_B}{q_*},$$

both inclusive. If there is a point in the intersection with an integer k-coordinate, then the one with the smallest k-coordinate must be the one whose k-coordinate is equal to $\lceil k_{BL} \rceil$. Hence, if $\lceil k_{BL} \rceil \leq k_{BR}$, that is, $\lceil k_{BL} \rceil \leq \lfloor k_{BR} \rfloor$, then we return $(\lceil k_{BL} \rceil, r'_B)$.

3. If $\lceil k_{BL} \rceil > \lfloor k_{BR} \rfloor$, then there is no point in the intersection with an integer k-coordinate. By the same argument as in the case of μ, we conclude that if there exists any point (k, r) in the triangular region with integer coordinates, then at least one of $(\lceil k_{BL} \rceil - 1, r) = (\lfloor k_{BR} \rfloor, r)$ or $(\lfloor k_{BR} \rfloor + 1, r)$ must be in the triangular region.

   Therefore, we see if there is any point in the triangular region with integer coordinates whose k-coordinate is either $\lfloor k_{BR} \rfloor$ or $\lceil k_{BL} \rceil$. If not, then we conclude there is no point with integer coordinates in the triangular region. To do this check, let

$$r_{BL} \coloneqq \frac{q_*(1-y) - q_*(q_*x - p_*)\lfloor k_{BR} \rfloor}{1 - q(q_*x - p_*)} \quad \text{and} \quad r_{BR} \coloneqq \frac{q_*\lceil k_{BL} \rceil - n_{\max}}{q},$$

   the r-coordinates of the intersection of the vertical line $k = \lfloor k_{BR} \rfloor$ with the left-boundary of the triangular region, and the intersection of the vertical line $k = \lceil k_{BL} \rceil$ with the right-boundary of the triangular region, respectively (note that "BL" is from "BR" and "BR" is from "BL").

   If $\lceil r_{BL} \rceil \geq \lceil r_T \rceil$ and $\lceil r_{BR} \rceil \geq \lceil r_T \rceil$, then we conclude there is no point with integer coordinates in the triangular region. In this case, we simply choose the smallest integer r-coordinate $\lceil r_T \rceil$ in the parallelogrammatic region. That is, lt

$$k_{PL} \coloneqq \frac{q\lceil r_T \rceil}{q_*},$$

   and we return $(\lfloor k_{PL} \rfloor + 1, \lceil r_T \rceil)$.

4. If $\lceil r_{BL} \rceil < \lceil r_T \rceil$ and $\lceil r_{BR} \rceil \geq \lceil r_T \rceil$, then there exists at least one point in the triangular region with integer coordinates, and there must exist a point whose k-coordinate is equal to $\lfloor k_{BR} \rfloor$ among the ones whose integer r-coordinate is the smallest. Hence, we

simply choose the one with the smallest integer k-coordinate among the points on the intersection of the triangular region and the horizontal line $r = \lceil r_{BL} \rceil$. That is, let

$$k_{BLL} \coloneqq \frac{q_*(1-y) - (1 - q(q_* x - p_*)) \lceil r_{BL} \rceil}{q_*(q_* x - p_*)}$$

and return $(\lceil k_{BLL} \rceil, \lceil r_{BL} \rceil)$.

5. If $\lceil r_{BL} \rceil \geq \lceil r_T \rceil$ and $\lceil r_{BR} \rceil < \lceil r_T \rceil$, then there exists at least one point in the triangular region with integer coordinates, and there must exist a point whose k-coordinate is equal to $\lfloor k_{BR} \rfloor$ among the ones whose integer r-coordinate is the smallest. Hence, we simply choose the one with the smallest integer k-coordinate among the points on the intersection of the triangular region and the horizontal line $r = \lceil r_{BR} \rceil$. That is, let

   $$k_{BRL} \coloneqq \frac{q_*(1-y) - (1 - q(q_* x - p_*)) \lceil r_{BR} \rceil}{q_*(q_* x - p_*)}$$

   and return $(\lceil k_{BRL} \rceil, \lceil r_{BR} \rceil)$.

6. If $\lceil r_{BL} \rceil < \lceil r_T \rceil$ and $\lceil r_{BR} \rceil < \lceil r_T \rceil$, then we choose the smaller one between $\lceil r_{BL} \rceil$ and $\lceil r_{BR} \rceil$ and do the same. That is, if $\lceil r_{BL} \rceil \leq \lceil r_{BR} \rceil$ then return $(\lceil k_{BLL} \rceil, \lceil r_{BL} \rceil)$, and otherwise return $(\lceil k_{BRL} \rceil, \lceil r_{BR} \rceil)$.

After following this procedure, we get a point $(k, r) \in \mathbb{Z} \times [0\colon q_* - 1]$, and we conclude $\nu = q_* k - qr$.

Here we give a summary of all quantities we need, after substituting $y \leftarrow y' = (n_{\min} x + y) - \lfloor n_{\min} x + y \rfloor$ and $n_{\max} \leftarrow n_{\max} - n_{\min}$. Some of them directly involve x, y in their representations, but the algorithm we develop in Section 12 is able to compute all of these within arbitrary-precision exact rational arithmetic. Those quantities are rewritten here in a way that the aforementioned algorithm can be directly applied:

$$\begin{aligned}
\lceil r_T \rceil &= -\lfloor q_* y' - q_* \rfloor = -\lfloor q_* n_{\min} x + q_* y \rfloor + q_*(\lfloor n_{\min} + y \rfloor + 1), \\
\lceil r_B \rceil &= -\lfloor q_*(n_{\max} - n_{\min}) x + q_* y' - (q_* + p_*(n_{\max} - n_{\min})) \rfloor \\
&= -\lfloor q_* n_{\max} x + q_* y \rfloor + (q_*(\lfloor n_{\min} x + y \rfloor + 1) + p_*(n_{\max} - n_{\min})), \\
r'_B &= \max\{\lceil r_B \rceil, 0\}, \\
\lfloor k_{PR} \rfloor &= \left\lfloor \frac{(n_{\max} - n_{\min}) + q(\lceil r_B \rceil - 1)}{q_*} \right\rfloor, \\
\lceil k_{TR} \rceil &= k_{\mathrm{common}}(\lceil r_T \rceil - 1), \\
\lfloor k_{TL} \rfloor &= \left\lfloor \frac{q(\lceil r_T \rceil - 1)}{q_*} \right\rfloor, \\
\lceil r_{TR} \rceil &= r_{\mathrm{common}}(\lceil k_{TR} \rceil), \\
\lceil r_{TL} \rceil &= \left\lceil \frac{q_* \lfloor k_{TL} \rfloor}{q} \right\rceil, \\
\lceil k_{TRR} \rceil &= k_{\mathrm{common}}(\lceil r_{TR} \rceil - 1), \\
\lceil k_{TLR} \rceil &= k_{\mathrm{common}}(\lceil r_{TL} \rceil - 1), \\
\lfloor k_{PL} \rfloor &= \left\lfloor \frac{q \lceil r_T \rceil}{q_*} \right\rfloor,
\end{aligned}$$

$$\begin{aligned}
\lceil k_{BL} \rceil &= k_{\text{common}}(r'_B), \\
\lfloor k_{BR} \rfloor &= \left\lfloor \frac{(n_{\max} - n_{\min}) + qr'_B}{q_*} \right\rfloor, \\
\lceil r_{BL} \rceil &= r_{\text{common}}(\lfloor k_{BR} \rfloor), \\
\lceil r_{BR} \rceil &= \left\lceil \frac{q_* \lceil k_{BL} \rceil - (n_{\max} - n_{\min})}{q} \right\rceil, \\
\lceil k_{BLL} \rceil &= k_{\text{common}}(\lceil r_{BL} \rceil), \quad \text{and} \\
\lceil k_{BRL} \rceil &= k_{\text{common}}(\lceil r_{BR} \rceil)
\end{aligned}$$

where

$$\begin{aligned}
k_{\text{common}} \colon r \mapsto &- \left\lfloor \frac{-qq_*rx + q_*y' + ((qp_* + 1)r - q_*)}{q_*^2 x - q_* p_*} \right\rfloor \\
&= - \left\lfloor \frac{q_* \left(n_{\min} - qr \right) x + q_* y + ((qp_* + 1)r - q_*(\lfloor n_{\min} x + y \rfloor + 1))}{q_*^2 x - q_* p_*} \right\rfloor
\end{aligned}$$

and

$$\begin{aligned}
r_{\text{common}} \colon k \mapsto &- \left\lfloor \frac{q_*^2 kx + q_*y' - (q_*p_*k + q_*)}{-qq_*x + (qp_* + 1)} \right\rfloor \\
&= - \left\lfloor \frac{q_* \left(q_* k + n_{\min} \right) x + q_* y - q_* \left(p_* k + \lfloor n_{\min} x + y \rfloor + 1 \right)}{-qq_*x + (qp_* + 1)} \right\rfloor .
\end{aligned}$$

In addition to these, checking for emptiness of L and R also involves quantities with x and y directly in their representations. That is, we have to compare the minimum and the maximum of $nx - \lfloor nx \rfloor$ for $n \in [1 \colon n_{\max}]$ with $1 - y$. By Theorem 2.4, we know

$$\min_{n \in [1 \colon n_{\max}]} (nx - \lfloor nx \rfloor) = q_* x - p_*,$$

and

$$\max_{n \in [1 \colon n_{\max}]} (nx - \lfloor nx \rfloor) = \frac{q - 1}{q}$$

if $x = \frac{p_*}{q_*}$, and otherwise

$$\max_{n \in [1 \colon n_{\max}]} (nx - \lfloor nx \rfloor) = q^* x - p^* + 1.$$

Therefore, L is empty if and only if

$$q_* x - p_* \geq 1 - y,$$

that is,

$$\lfloor q_* x + y \rfloor > p_*.$$

Similarly, if $x = \frac{p_*}{q_*}$, then R is empty if and only if

$$\frac{q_* - 1}{q_*} < 1 - y,$$

that is,

$$\lfloor q_* y \rfloor = 0,$$

and otherwise,

$$q^*x - p^* + 1 < 1 - y,$$

that is,

$$\lfloor q^*x + y \rfloor < p^*.$$

Again, computation of $\lfloor q_*x + y \rfloor$ and $\lfloor q^*x + y \rfloor$ can be done within arbitrary-precision exact rational arithmetic using the algorithm developed in Section 12. Finally, we obtain the following algorithm for computing a quadruple of integers (a, b, c, d) such that $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ holds for all $n \in [n_{\min} \colon n_{\max}]$ if and only if

$$a - b\xi \leq \zeta < c - d\xi,$$

provided that we know $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ holds for all $n \in [0 \colon n_{\max} - n_{\min}]$. Recall that Theorem 8.1 is supposed to compute the range of $\zeta' \coloneqq (n_{\min}\xi + \zeta) - \lfloor n_{\min}x + y \rfloor$ since y is replaced by $y' \coloneqq (n_{\min}x + y) - \lfloor n_{\min} + y \rfloor$. Hence, when the condition dictated by Theorem 8.1 is of the form

$$a - b\xi \leq \zeta < c - d\xi,$$

then it actually means

$$(a + \lfloor n_{\min}x + y \rfloor) - (b + n_{\min})\xi \leq \zeta < (c + \lfloor n_{\min}x + y \rfloor) - (d + n_{\min})\xi.$$

Algorithm 8.2 (Simultaneously having $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ and $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$)**.**

1. If $n_{\min} = n_{\max}$, we only need to ensure $\lfloor n_{\min}x + y \rfloor = \lfloor n_{\min}\xi + \zeta \rfloor$, which is equivalent to
$$\lfloor n_{\min}x + y \rfloor - n_{\min}\xi \leq \zeta < \lfloor n_{\min}x + y \rfloor + 1 - n_{\min}\xi,$$
thus return $(\lfloor n_{\min}x + y \rfloor, n_{\min}, \lfloor n_{\min}x + y \rfloor + 1, n_{\min})$.

2. Otherwise, compute the best rational approximations $\frac{p_*}{q_*}$ from below and $\frac{p^*}{q^*}$ from above of x with the largest reduced denominators $q_*, q^* \leq n_{\max} - n_{\min}$.

3. Compute the modular inverse q of $-p_*$ with respect to q_*. Here, $q = 1$ if $q_* = 1$, and otherwise it is the largest reduced denominator of all odd semiconvergents of x whose reduced denominator is strictly smaller than q_*; this follows by applying Theorem 2.4 with $x \leftarrow \frac{p_*}{q_*}$ and $n_{\max} \leftarrow q_* - 1$.

4. If $\lfloor (q_* + n_{\min})x + y \rfloor > p_* + \lfloor n_{\min}x + y \rfloor$ holds, then $L = \emptyset$.
   (a) In this case, if $q_* = q^*$, then let $\mu \leftarrow q$, and otherwise, let $\mu \leftarrow q^*$.
   (b) Return $\left(p_* + \lfloor n_{\min}x + y \rfloor + 1, q_* + n_{\min}, \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + \lfloor n_{\min}x + y \rfloor + 2, \mu + n_{\min}\right)$.

5. If $q_* = q^*$ and $\lfloor q_*n_{\min}x + q_*y \rfloor - q_* \lfloor n_{\min}x + y \rfloor = 0$, or $q_* \neq q^*$ and $\lfloor (q^* + n_{\min})x + y \rfloor < p^* + \lfloor n_{\min}x + y \rfloor$, then $R = \emptyset$.
   (a) In this case, if $q_* = q^*$, then let $\mu \leftarrow q$, and otherwise, let $\mu \leftarrow q^*$.
   (b) Return $\left(p_* + \lfloor n_{\min}x + y \rfloor, q_* + n_{\min}, \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + \lfloor n_{\min}x + y \rfloor + 1, \mu + n_{\min}\right)$.

6. If $q_* = q^*$,

(a) In this case, let

$$\begin{aligned}\mu &\leftarrow ((\lfloor q_* n_{\min} x + q_* y \rfloor - q_* \lfloor n_{\min} x + y \rfloor + 1) q \mod q_*)\,, \\ \nu &\leftarrow ((\lfloor q_* n_{\min} x + q_* y \rfloor - q_* \lfloor n_{\min} x + y \rfloor) q \mod q_*)\,.\end{aligned}$$

(b) Return $\left(\left\lfloor \frac{\nu p_*}{q_*} \right\rfloor + \lfloor n_{\min} x + y \rfloor + 1, \nu + n_{\min}, \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + \lfloor n_{\min} x + y \rfloor + 1, \mu + n_{\min} \right)$.

7. Compute μ as follows:

   (a) If $\lceil r_T \rceil - 1 < \lceil r_B \rceil$, then let $(k, r) \leftarrow (\lfloor k_{PR} \rfloor, \lceil r_B \rceil - 1)$.

   (b) Otherwise, check if $\lceil k_{TR} \rceil - 1 > \lfloor k_{TL} \rfloor$ holds. If that is the case, then let $(k, r) \leftarrow (\lceil k_{TR} \rceil - 1, \lceil r_T \rceil - 1)$,

   (c) Otherwise, check if $\lceil r_{TR} \rceil - 1 < \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 < \lceil r_B \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lfloor k_{PR} \rfloor, \lceil r_B \rceil - 1)$.

   (d) Otherwise, check if $\lceil r_{TR} \rceil - 1 \geq \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 < \lceil r_B \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lceil k_{TRR} \rceil - 1, \lceil r_{TR} \rceil - 1)$.

   (e) Otherwise, check if $\lceil r_{TR} \rceil - 1 < \lceil r_B \rceil$ and $\lceil r_{TL} \rceil - 1 \geq \lceil r_B \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lceil k_{TLR} \rceil - 1, \lceil r_{TL} \rceil - 1)$.

   (f) Otherwise,

      i. If $\lceil r_{TR} \rceil - 1 \geq \lceil r_{TL} \rceil - 1$, then let $(k, r) \leftarrow (\lceil k_{TRR} \rceil - 1, \lceil r_{TR} \rceil - 1)$.

      ii. Othrtwise, let $(k, r) \leftarrow (\lceil k_{TLR} \rceil - 1, \lceil r_{TL} \rceil - 1)$.

   (g) Then, let $\mu \leftarrow q_* k - q r$.

8. Compute ν as follows:

   (a) If $r'_B \geq \lceil r_T \rceil$, then let $(k, r) \leftarrow (\lfloor k_{PL} \rfloor + 1, \lceil r_T \rceil)$.

   (b) Otherwise, check if $\lceil k_{BL} \rceil \leq \lfloor k_{BR} \rfloor$ holds. If that is the case, then let $(k, r) \leftarrow (\lceil k_{BL} \rceil, r'_B)$,

   (c) Otherwise, check if $\lceil r_{BL} \rceil \geq \lceil r_T \rceil$ and $\lceil r_{BR} \rceil \geq \lceil r_T \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lfloor k_{PL} \rfloor + 1, \lceil r_T \rceil)$.

   (d) Otherwise, check if $\lceil r_{BL} \rceil < \lceil r_T \rceil$ and $\lceil r_{BR} \rceil \geq \lceil r_T \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lceil k_{BLL} \rceil, \lceil r_{BL} \rceil)$.

   (e) Otherwise, check if $\lceil r_{BL} \rceil \geq \lceil r_T \rceil$ and $\lceil r_{BR} \rceil < \lceil r_T \rceil$ hold. If that is the case, then let $(k, r) \leftarrow (\lceil k_{BRL} \rceil, \lceil r_{BR} \rceil)$.

   (f) Otherwise,

      i. If $\lceil r_{BL} \rceil \leq \lceil r_{BR} \rceil$, then let $(k, r) \leftarrow (\lceil k_{BLL} \rceil, \lceil r_{BL} \rceil)$.

      ii. Othrtwise, let $(k, r) \leftarrow (\lceil k_{BRL} \rceil, \lceil r_{BR} \rceil)$.

   (g) Then, let $\nu \leftarrow q_* k - q r$.

9. Return $\left(\left\lfloor \frac{\nu p_*}{q_*} \right\rfloor + \lfloor n_{\min} x + y \rfloor + 1, \nu + n_{\min}, \left\lfloor \frac{\mu p_*}{q_*} \right\rfloor + \lfloor n_{\min} x + y \rfloor + 1, \mu + n_{\min} \right)$.

9 Lattice points in vertically parallel trapezoidal regions

Combining Theorem 7.4 with Algorithm 8.2, we obtain the full subset in $\mathbb{R}^2$ completely characterizing (ξ, ζ) which satisfies

1. $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ for all $n \in [n_{\min} \colon n_{\max}]$ and
2. $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ for all $n \in [0 \colon n_{\max} - n_{\min}]$,

and the set is always of the form

$$T \coloneqq \left\{(\xi, \zeta) \in \mathbb{R}^2 \colon \xi_{\min} \leq \xi < \xi_{\max}, \quad a - b\xi \leq \zeta < c - d\xi\right\}$$

for some $\xi_{\min}, \xi_{\max} \in \mathbb{Q}$ with $\xi_{\min} < \xi_{\max}$ and $a, b, c, d \in \mathbb{Z}$, which, if drawn in the (ξ, ζ)-plane, is a trapezoid whose two edges are parallel to the vertical axis $\xi = 0$. For use in Algorithm 5.4, we need to find a point in T of the form $\left(\frac{m}{2^k}, \frac{s}{2^k}\right)$ where $m, s \in \mathbb{Z}$ and $k \in \mathbb{Z}_{\geq 0}$.

In this section, we describe how to find such a point. More specifically, we aim to find the smallest k such that a point of the form $\left(\frac{m}{2^k}, \frac{s}{2^k}\right)$ exists in T, and when such k is found, we want to enumerate all points in T of that form.

As the first step, we find the smallest $k \in \mathbb{Z}_{\geq 0}$ such that there exists some $m \in \mathbb{Z}$ satisfying

$$\xi_{\min} \leq \frac{m}{2^k} < \xi_{\max}.$$

This can be done as follows. We start with computing $k_0 \coloneqq \max\left\{-\lfloor \log_2(\xi_{\max} - \xi_{\min}) \rfloor, 0\right\}$. Computation of $\max\left\{\left\lfloor \log_2 \frac{p}{q} \right\rfloor, 0\right\}$ for a rational number $\frac{p}{q}$ can be easily done in integer arithmetic as it is basically just computing the difference of bit-widths between p and q. Then since

$$\xi_{\max} - \xi_{\min} \geq 2^{\lfloor \log_2(\xi_{\max} - \xi_{\min}) \rfloor} \geq \frac{1}{2^{k_0}},$$

there must exist a number of the form $\frac{m_0}{2^{k_0}}$ in $[\xi_{\min}, \xi_{\max})$, and specifically $m_0 = \left\lceil 2^{k_0} \xi_{\min} \right\rceil$ satisfies that. On the other hand, if $k_0 > 0$, then since

$$\xi_{\max} - \xi_{\min} < 2^{\lfloor \log_2(\xi_{\max} - \xi_{\min}) \rfloor + 1} = \frac{1}{2^{k_0 - 1}},$$

there are at most 2 numbers of the form $\frac{m_0}{2^{k_0}}$ in $[\xi_{\min}, \xi_{\max})$. If there is only one element of such form and it is odd, then $k = k_0$ should be the smallest integer we are looking for. Otherwise, there should uniquely an even integer m_0 such that $\frac{m_0}{2^{k_0}} \in [\xi_{\min}, \xi_{\max})$. If we write $m_0 = 2^\tau \sigma$, then $k = \max\{k_0 - \tau, 0\}$ is the smallest integer we are looking for.

Then for any integer $m \in \left[\left\lceil 2^k \xi_{\min} \right\rceil \colon \left\lceil 2^k \xi_{\max} \right\rceil - 1\right]$ (in other words, for any integer m such that $\frac{m}{2^k} \in [\xi_{\min}, \xi_{\max})$), if we set $\xi = \frac{m}{2^k}$, then the corresponding bound on ζ becomes

$$\frac{2^k a - bm}{2^k} \leq \zeta < \frac{2^k c - dm}{2^k}.$$

Therefore, for an integer s, we have $\left(\frac{m}{2^k}, \frac{s}{2^k}\right) \in T$ if and only if $2^k a - bm \leq s < 2^k c - dm$.

However, that it is possible that the range of s we get here is empty. When that happens for every m, then we increase k sufficiently to make sure there is at least one m yielding a nonempty range for s. Note that by Theorem 8.1, the integers a, b, c, d are determined independently to ξ, and Theorem 7.5 and the construction of a, b, c, d ensure that $a - b\xi \leq$

$c - d\xi$ always holds for any $\xi \in [\xi_{\min}, \xi_{\max})$. Hence, the only case when the range of s becomes empty is when $2^k a - bm = 2^k c - dm$ holds.

We claim that this linear equation $2^k a - bm = 2^k c - dm$ on m can have at most 1 solution. Indeed, it has more than 1 solutions if and only if $a = c$ and $b = d$ both hold. However, this means that $a - b\xi = c - d\xi$ holds for *every* ξ, which is clearly impossible because applying Theorem 8.1 to $\xi \leftarrow x$ and $\zeta \leftarrow y$ shows that $a - bx \leq y < c - dx$ holds.

Therefore, all we need to do is to find the smallest k' larger than the current k such that there are at least 2 distinct numbers of the form $\frac{m'}{2^{k'}}$ in $[\xi_{\min}, \xi_{\max})$. In particular, if there already were 2 or more distinct numbers of the form $\frac{m}{2^k}$ in $[\xi_{\min}, \xi_{\max})$, then at least one of them should have yielded a nonempty range for s. Hence, let m be the unique integer such that $\frac{m}{2^k} \in [\xi_{\min}, \xi_{\max})$. If $\frac{m}{2^k} - \xi_{\min} \geq \frac{m}{2^k} - \xi_{\max}$, then we find the smallest $k' > k$ such that

$$\frac{1}{2^{k'}} \leq \frac{m}{2^k} - \xi_{\min},$$

that is,

$$k' := -\left\lfloor \log_2 \left(\frac{m}{2^k} - \xi_{\min} \right) \right\rfloor,$$

so that $\frac{2^{k'-k}m - 1}{2^{k'}} \in [\xi_{\min}, \xi_{\max})$ holds. Otherwise, we find the smallest $k' > k$ such that

$$\frac{1}{2^{k'}} < \xi_{\max} - \frac{m}{2^k},$$

that is,

$$k' := 1 - \left\lceil \log_2 \left(\xi_{\max} - \frac{m}{2^k} \right) \right\rceil,$$

so that $\frac{2^{k'-k}m + 1}{2^{k'}} \in [\xi_{\min}, \xi_{\max})$ holds. Formally writing all of these in details then gives us the following algorithm.

Algorithm 9.1 (Finding lattice points in vertically parallel trapezoidal regions)**.**

```
1. Let $k \leftarrow \max\{-\lfloor \log_2(\xi_{\max} - \xi_{\min}) \rfloor, 0\}$ and $m \leftarrow \lceil 2^k \xi_{\min} \rceil$.
2. If $m$ is even, then factorize it into $m = 2^\tau \sigma$, and then let $m \leftarrow m/2^{\min\{k,\tau\}}$ and
   $k \leftarrow \max\{k - \tau, 0\}$.
3. Otherwise, if $m + 1 < \lceil 2^k \xi_{\max} \rceil$, then factorize $m + 1$ into $m + 1 = 2^\tau \sigma$, and then let
   $m \leftarrow (m+1)/2^{\min\{k,\tau\}}$ and $k \leftarrow \max\{k - \tau, 0\}$.
4. Until $m$ reaches $\lceil 2^k \xi_{\max} \rceil$,
   (a) Let
       $s_{\min} \leftarrow 2^k a - mb$   and   $s_{\max} \leftarrow 2^k c - md - 1$.
   (b) If $s_{\min} \leq s_{\max}$, then record $(k, m, [s_{\min} \colon s_{\max}])$.
   (c) Let $m \leftarrow m + 1$.
5. If there was at least one $m$ that yielded $s_{\min} \leq s_{\max}$, then we are done, so return.
6. Otherwise, let $m \leftarrow m - 1$ (which must be equal to $\lceil 2^k \xi_{\max} \rceil - 1 = \lceil 2^k \xi_{\min} \rceil$).
7. If $m - 2^k \xi_{\min} \geq 2^k \xi_{\max} - m$, then let $k \leftarrow k - \lfloor \log_2 (m - 2^k \xi_{\min}) \rfloor$. Otherwise, let
   $k \leftarrow k + 1 - \lceil \log_2 (2^k \xi_{\max} - m) \rceil$.
8. Let $m \leftarrow \lceil 2^k \xi_{\min} \rceil$ and go back to Step 4. This time Step 5 must succeed to return.
```

10 Extremizers of $\frac{\lfloor nx+y\rfloor-\zeta}{n}$

So far, we have completed our explanation on how to find the smallest minimizer and the largest maximizer of the fractional part $(nx+y)-\lfloor nx+y\rfloor$ (except for the explanation of the algorithm that will be developed in Section 12).

Taking it as granted, in this section we now describe an algorithm for finding extremizers of $\frac{\lfloor nx+y\rfloor-\zeta}{n}$ over an arbitrary interval $[n_{\min}\colon n_{\max}]$ minus 0, provided that ζ is an arbitrary fixed number in addition to x,y. Recall that having $\lfloor nx+y\rfloor=\lfloor n\xi+\zeta\rfloor$ for all $n\in[n_{\min}\colon n_{\max}]$ is equivalent to the inequalities

$$\max_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{>0}}\frac{\lfloor nx+y\rfloor-\zeta}{n}\leq\xi<\min_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{>0}}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}$$

and

$$\max_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{<0}}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}<\xi\leq\min_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{<0}}\frac{\lfloor nx+y\rfloor-\zeta}{n}$$

possibly along with $\lfloor y\rfloor=\lfloor\zeta\rfloor$ if $0\in[n_{\min}\colon n_{\max}]$, so once we now how to find extremizers of $\frac{\lfloor nx+y\rfloor-\zeta}{n}$, we can immediately find a necessary and sufficient condition on ξ for having $\lfloor nx+y\rfloor=\lfloor n\xi+\zeta\rfloor$ for all $n\in[n_{\min}\colon n_{\max}]$, when x,y,ζ are all given real numbers.

In addition to that this problem of finding ξ when ζ is given has its own applications (TODO: elaborate), our main algorithm for finding a necessary and sufficient condition on both ξ and ζ, which we develop in Section 11, is based on it.

Throughout this section, let x,y be real numbers and $n_{\min},n_{\max}$ be integers such that $n_{\min}\leq n_{\max}$. Until Section 10.4, we will also assume $0<n_{\min}$.

10.1 The maximization problem

We consider the maximization problem

$$\max_{n\in[n_{\min}\colon n_{\max}]}\frac{\lfloor nx+y\rfloor-\zeta}{n}$$

first. Since

$$\frac{\lfloor nx+y\rfloor-\zeta}{n}=x-\frac{(nx+y)-\lfloor nx+y\rfloor}{n}-\frac{\zeta-y}{n},$$

our maximization problem can be equivalently written as the minimization problem

$$\min_{n\in[n_{\min}\colon n_{\max}]}\left(\frac{(nx+y)-\lfloor nx+y\rfloor}{n}+\frac{\zeta-y}{n}\right).$$

Since $\zeta-y$ is supposed to be a small constant, it is a reasonable guess that the *smallest minimizer* of $(nx+y)-\lfloor nx+y\rfloor$ will be a good approximation of the maximizer of the above. Hence, we start from there.

Let n_0 be the smallest minimizer of $(nx+y)-\lfloor nx+y\rfloor$ for $n\in[n_{\min}\colon n_{\max}]$, because $\frac{\zeta-y}{n}$ Then, the problem of maximizing $\frac{\lfloor nx+y\rfloor-\zeta}{n}$ can be solved by separately solving two maximization problems

$$\max_{n\in[n_{\min}\colon n_0]}\frac{\lfloor nx+y\rfloor-\zeta}{n}\tag{8}$$

and

$$\max_{n\in[n_0\colon n_{\max}]}\frac{\lfloor nx+y\rfloor-\zeta}{n}.\tag{9}$$

The following two theorems, when combined with Theorem 7.1, Theorem 2.4 and Algorithm 4.12, give complete solutions to these two problems, respectively.

Theorem 10.1 (The smallest maximizer on the left)**.**
Let $n_0^{(0)} := n_0$*. For* $i \in \mathbb{Z}_{>0}$*, inductively define* $n_0^{(i)} := n_0^{(i-1)} - n_1^{(i)}$ *and*

$$n_1^{(i)} := \begin{cases} \max \operatorname*{argmin}_{n\in\left[1:\, n_0^{(i-1)}-n_{\min}\right]} \dfrac{\lfloor nx\rfloor + 1}{n} & \textit{if } n_0^{(i-1)} > n_{\min}, \\ 0 & \textit{otherwise.} \end{cases}$$

Let k *be the smallest nonnegative integer such that either* $n_0^{(k)} = n_{\min}$ *or*

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor + 1}{n_1^{(k+1)}} > \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor - \zeta}{n_0^{(k)}}$$

holds. Then

$$n_0^{(k)} = \min \operatorname*{argmax}_{n\in[n_{\min}\,:\, n_0]} \frac{\lfloor nx + y\rfloor - \zeta}{n}.$$

Proof. First, we claim that

$$\left\lfloor (n_0^{(i)} - n)x + y\right\rfloor = \left\lfloor n_0^{(i)}x + y\right\rfloor - \lfloor nx\rfloor - 1$$

holds for all $i \in \mathbb{Z}_{\geq 0}$ and $n \in \left[1\colon n_0^{(i)} - n_{\min}\right]$. We use induction on i. The base case $i = 0$ is precisely Lemma 6.3, so we only prove the induction step. Since $n_0^{(i)} = n_0^{(i-1)} - n_1^{(i)}$, the induction hypothesis implies

$$\left\lfloor (n_0^{(i)} - n)x + y\right\rfloor = \left\lfloor n_0^{(i-1)}x + y\right\rfloor - \left\lfloor (n_1^{(i)} + n)x\right\rfloor - 1.$$

On the other hand, since

$$n_1^{(i)} + n \leq n_1^{(i)} + n_0^{(i)} - n_{\min} = n_0^{(i-1)} - n_{\min},$$

Corollary 7.3 implies

$$\left\lfloor (n_1^{(i)} + n)x\right\rfloor = \left\lfloor n_1^{(i)}x\right\rfloor + \lfloor nx\rfloor + 1.$$

Again by the induction hypothesis, we know

$$\left\lfloor n_0^{(i)}x + y\right\rfloor = \left\lfloor n_0^{(i-1)}x + y\right\rfloor - \left\lfloor n_1^{(i)}x\right\rfloor - 1, \tag{10}$$

so combining these completes the induction step. Hence, the claim is proved.

Next, we claim that for each $i = 0, \cdots, k$,

$$n_0^{(i)} = \min \operatorname*{argmax}_{n\in\left[n_0^{(i)}\,:\, n_0\right]} \frac{\lfloor nx + y\rfloor - \zeta}{n}$$

holds. We use induction on i. The base case $i = 0$ is trivial, so we show the induction step. By the induction hypothesis, we have

$$\frac{\left\lfloor n_0^{(i-1)}x + y\right\rfloor - \zeta}{n_0^{(i-1)}} \geq \frac{\lfloor nx + y\rfloor - \zeta}{n}$$

for all $n \in \left[n_0^{(i-1)} \colon n_0\right]$. Also, note that

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i)} x + y\right\rfloor - \zeta}{n_0^{(i)}} - \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \zeta}{n_0^{(i-1)}} \\
&\qquad = \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \left\lfloor n_1^{(i)} x\right\rfloor - 1 - \zeta}{n_0^{(i)}} - \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \zeta}{n_0^{(i-1)}} \\
&\qquad = \frac{n_1^{(i)}}{n_0^{(i)}} \left(\frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \zeta}{n_0^{(i-1)}} - \frac{\left\lfloor n_1^{(i)} x\right\rfloor + 1}{n_1^{(i)}} \right),
\end{aligned}$$

and by the definition of k, the right-hand side is nonnegative. Therefore,

$$\frac{\left\lfloor n_0^{(i)} x + y\right\rfloor - \zeta}{n_0^{(i)}} \geq \frac{\lfloor nx + y\rfloor - \zeta}{n}$$

holds for all $n \in \left[n_0^{(i-1)} \colon n_0\right]$.

On the other hand, the previous claim shows that for any $n \in \left[1 \colon n_1^{(i)}\right]$,

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i)} x + y\right\rfloor - \zeta}{n_0^{(i)}} - \frac{\left\lfloor (n_0^{(i-1)} - n) x + y\right\rfloor - \zeta}{n_0^{(i-1)} - n} \\
&\quad = \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \left\lfloor n_1^{(i)} x\right\rfloor - 1 - \zeta}{n_0^{(i)}} - \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \lfloor nx\rfloor - 1 - \zeta}{n_0^{(i-1)} - n} \\
&\quad = \frac{n_0^{(i-1)}(n_1^{(i)} - n)}{n_0^{(i)}(n_0^{(i-1)} - n)} \cdot \frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \zeta}{n_0^{(i-1)}} + \frac{\lfloor nx\rfloor + 1}{n_0^{(i-1)} - n} - \frac{\left\lfloor n_1^{(i)} x\right\rfloor + 1}{n_0^{(i)}}.
\end{aligned}$$

Recall that by the definition of k, we have

$$\frac{\left\lfloor n_0^{(i-1)} x + y\right\rfloor - \zeta}{n_0^{(i-1)}} \geq \frac{\left\lfloor n_1^{(i)} x\right\rfloor + 1}{n_1^{(i)}},$$

thus we get

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i)} x + y\right\rfloor - \zeta}{n_0^{(i)}} - \frac{\left\lfloor (n_0^{(i-1)} - n) x + y\right\rfloor - \zeta}{n_0^{(i-1)} - n} \\
&\qquad \geq \frac{\lfloor nx\rfloor + 1}{n_0^{(i-1)} - n} - \frac{n_1^{(i)}(n_0^{(i-1)} - n) - n_0^{(i-1)}(n_1^{(i)} - n)}{n_0^{(i)}(n_0^{(i-1)} - n)} \cdot \frac{\left\lfloor n_1^{(i)} x\right\rfloor + 1}{n_1^{(i)}} \\
&\qquad = \frac{n}{n_0^{(i-1)} - n} \left(\frac{\lfloor nx\rfloor + 1}{n} - \frac{\left\lfloor n_1^{(i)} x\right\rfloor + 1}{n_1^{(i)}} \right) \geq 0
\end{aligned}$$

by the definition of $n_1^{(i)}$. As a result, we conclude

$$\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\geq\frac{\lfloor nx+y\rfloor-\zeta}{n}$$

for all $n\in\left[n_0^{(i)}\colon n_0\right]$, completing the induction step. Therefore, the claim is proved.

Consequently, it is enough to show that for any $n\in\left[1\colon n_0^{(k)}-n_{\min}\right]$, we have

$$\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}>\frac{\left\lfloor (n_0^{(k)}-n)x+y\right\rfloor-\zeta}{n_0^{(k)}-n}.$$

This is vacuously true if $n_0^{(k)}=n_{\min}$, so suppose otherwise, thus

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor+1}{n_1^{(k+1)}}>\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}$$

holds. By the first claim, for any $n\in\left[1\colon n_0^{(k)}-n_{\min}\right]$, we have

$$\begin{aligned}&\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}-\frac{\left\lfloor (n_0^{(k)}-n)x+y\right\rfloor-\zeta}{n_0^{(k)}-n}\\&\qquad=\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}-\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\lfloor nx\rfloor-1-\zeta}{n_0^{(k)}-n}\\&\qquad=\frac{n}{n_0^{(k)}-n}\left(\frac{\lfloor nx\rfloor+1}{n}-\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}\right),\end{aligned}$$

and by the definition of $n_1^{(k+1)}$, the right-hand side is strictly positive. Therefore, we are done. □

Theorem 10.2 (The largest maximizer on the right)**.**
Let $n_0^{(0)}:=n_0$. For $i\in\mathbb{Z}_{>0}$, inductively define $n_0^{(i)}:=n_0^{(i-1)}+n_1^{(i)}$ and

$$n_1^{(i)}:=\begin{cases}\max\underset{n\in\left[1\colon n_{\max}-n_0^{(i-1)}\right]}{\operatorname{argmax}}\dfrac{\lfloor nx\rfloor}{n} & \text{if } n_0^{(i-1)}<n_{\max},\\ 0 & \text{otherwise.}\end{cases}$$

Let k be the smallest nonnegative integer such that either $n_0^{(k)}=n_{\max}$ or

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor}{n_1^{(k+1)}}<\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}$$

holds. Then

$$n_0^{(k)}=\max\underset{n\in[n_0\colon n_{\max}]}{\operatorname{argmax}}\frac{\lfloor nx+y\rfloor-\zeta}{n}.$$

Proof. First, we claim that

$$\left\lfloor (n_0^{(i)} + n)x + y \right\rfloor = \left\lfloor n_0^{(i)} x + y \right\rfloor + \lfloor nx \rfloor$$

holds for all $i \in \mathbb{Z}_{\geq 0}$ and $n \in \left[0 \colon n_{\max} - n_0^{(i)}\right]$. We use induction on i. The base case $i = 0$ directly follows from Lemma 6.1, so we only prove the induction step. Since $n_0^{(i)} = n_0^{(i-1)} + n_1^{(i)}$, the induction hypothesis implies

$$\left\lfloor (n_0^{(i)} + n)x + y \right\rfloor = \left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor (n_1^{(i)} + n)x \right\rfloor.$$

On the other hand, since

$$n_1^{(i)} + n \leq n_1^{(i)} + n_{\max} - n_0^{(i)} = n_{\max} - n_0^{(i-1)},$$

Corollary 7.3 implies

$$\left\lfloor (n_1^{(i)} + n)x \right\rfloor = \left\lfloor n_1^{(i)} x \right\rfloor + \lfloor nx \rfloor.$$

Again by the induction hypothesis, we know

$$\left\lfloor n_0^{(i)} x + y \right\rfloor = \left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor n_1^{(i)} x \right\rfloor, \tag{11}$$

so combining these completes the induction step. Hence, the claim is proved.

Next, we claim that for each $i = 0, \cdots, k$,

$$n_0^{(i)} = \max \operatorname*{argmax}_{n \in \left[n_0 \colon n_0^{(i)}\right]} \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

holds. We use induction on i. The base case $i = 0$ is trivial, so we show the induction step. By the induction hypothesis, we have

$$\frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor - \zeta}{n_0^{(i-1)}} \geq \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

for all $n \in \left[n_0 \colon n_0^{(i-1)}\right]$. On the other hand, the previous claim shows that for any $n \in \left[0 \colon n_1^{(i)}\right]$,

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i)} x + y \right\rfloor - \zeta}{n_0^{(i)}} - \frac{\left\lfloor (n_0^{(i-1)} + n)x + y \right\rfloor - \zeta}{n_0^{(i-1)} + n} \\
&\quad = \frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor n_1^{(i)} x \right\rfloor - \zeta}{n_0^{(i)}} - \frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor + \lfloor nx \rfloor - \zeta}{n_0^{(i-1)} + n} \\
&\quad = \frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_0^{(i)}} - \frac{\lfloor nx \rfloor}{n_0^{(i-1)} + n} - \frac{n_0^{(i-1)}(n_1^{(i)} - n)}{n_0^{(i)}(n_0^{(i-1)} + n)} \cdot \frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor - \zeta}{n_0^{(i-1)}}.
\end{aligned}$$

Recall that by the definition of k, we have

$$\frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor - \zeta}{n_0^{(i-1)}} \leq \frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}},$$

thus we get

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}-\frac{\left\lfloor (n_0^{(i-1)}+n)x+y\right\rfloor-\zeta}{n_0^{(i-1)}+n}\\
&\qquad\geq \frac{n_1^{(i)}(n_0^{(i-1)}+n)-n_0^{(i-1)}(n_1^{(i)}-n)}{n_0^{(i)}(n_0^{(i-1)}+n)}\cdot\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_1^{(i)}}-\frac{\lfloor nx\rfloor}{n_0^{(i-1)}+n}\\
&\qquad=\frac{n\left\lfloor n_1^{(i)}x\right\rfloor}{n_1^{(i)}(n_0^{(i-1)}+n)}-\frac{\lfloor nx\rfloor}{n_0^{(i-1)}+n}.
\end{aligned}$$

If $n=0$, then the above is equal to zero, and if $n\neq 0$, then it is equal to

$$\frac{n}{n_0^{(i-1)}+n}\left(\frac{\left\lfloor n_1^{(i)}x\right\rfloor}{n_1^{(i)}}-\frac{\lfloor nx\rfloor}{n}\right)\geq 0$$

by the definition of $n_1^{(i)}$. As a result, we conclude

$$\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\geq\frac{\lfloor nx+y\rfloor-\zeta}{n}$$

for all $n\in\left[n_0\colon n_0^{(i)}\right]$, completing the induction step. Therefore, the claim is proved.

Consequently, it is enough to show that for any $n\in\left[1\colon n_{\max}-n_0^{(k)}\right]$, we have

$$\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}>\frac{\left\lfloor (n_0^{(k)}+n)x+y\right\rfloor-\zeta}{n_0^{(k)}+n}.$$

This is vacuously true if $n_0^{(k)}=n_{\max}$, so suppose otherwise, thus

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor}{n_1^{(k+1)}}<\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}$$

holds. By the first claim, for any $n\in\left[1\colon n_{\max}-n_0^{(k)}\right]$, we have

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}-\frac{\left\lfloor (n_0^{(k)}+n)x+y\right\rfloor-\zeta}{n_0^{(k)}+n}\\
&\qquad=\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}-\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor+\lfloor nx\rfloor-\zeta}{n_0^{(k)}+n}\\
&\qquad=\frac{n}{n_0^{(k)}+n}\left(\frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}-\frac{\lfloor nx\rfloor}{n}\right),
\end{aligned}$$

and by the definition of $n_1^{(k+1)}$, the right-hand side is strictly positive. Therefore, we are done. □

10.2 The minimization problem

Next, we consider the minimization problem

$$\min_{n\in[n_{\min}\,:\,n_{\max}]} \frac{\lfloor nx+y\rfloor-\zeta}{n}.$$

Since

$$\frac{\lfloor nx+y\rfloor-\zeta}{n} = x - \frac{(nx+y)-\lfloor nx+y\rfloor}{n} - \frac{\zeta-y}{n},$$

our minimization problem can be equivalently written as the maximization problem

$$\max_{n\in[n_{\min}\,:\,n_{\max}]} \left(\frac{(nx+y)-\lfloor nx+y\rfloor}{n} + \frac{\zeta-y}{n}\right).$$

Again, since $\zeta-y$ is supposed to be a small constant, it is a reasonable guess that the n the *largest maximizer* of $(nx+y)-\lfloor nx+y\rfloor$ will be a good approximation of the minimizer of the above. Hence, we start from there.

Let n_0 be the largest maximizer of $(nx+y)-\lfloor nx+y\rfloor$ for $n\in[n_{\min}\,:\,n_{\max}]$. Then, the problem of minimizing $\frac{\lfloor nx+y\rfloor-\zeta}{n}$ can be solved by separately solving two minimization problems

$$\min_{n\in[n_{\min}\,:\,n_0]} \frac{\lfloor nx+y\rfloor-\zeta}{n} \tag{12}$$

and

$$\min_{n\in[n_0\,:\,n_{\max}]} \frac{\lfloor nx+y\rfloor-\zeta}{n}. \tag{13}$$

The following two theorems, when combined with Theorem 7.1, Theorem 2.4 and Algorithm 4.12, give complete solutions to these two problems, respectively.

Theorem 10.3 (The smallest minimizer on the left)**.**
Let $n_0^{(0)} := n_0$. For $i\in\mathbb{Z}_{>0}$, inductively define $n_0^{(i)} := n_0^{(i-1)} - n_1^{(i)}$ and

$$n_1^{(i)} := \begin{cases} \max \operatorname*{argmax}\limits_{n\in\left[1\,:\,n_0^{(i-1)}-n_{\min}\right]} \dfrac{\lfloor nx\rfloor}{n} & \text{if } n_0^{(i-1)} > n_{\min}, \\ 0 & \text{otherwise.}\end{cases}$$

Let k be the smallest nonnegative integer such that either $n_0^{(k)} = n_{\min}$ or

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor}{n_1^{(k+1)}} < \frac{\left\lfloor n_0^{(k)}x+y\right\rfloor-\zeta}{n_0^{(k)}}$$

holds. Then

$$n_0^{(k)} = \min \operatorname*{argmin}_{n\in[n_{\min}\,:\,n_0]} \frac{\lfloor nx+y\rfloor-\zeta}{n}.$$

Proof. First, we claim that

$$\left\lfloor (n_0^{(i)}-n)x+y\right\rfloor = \left\lfloor n_0^{(i)}x+y\right\rfloor - \lfloor nx\rfloor$$

holds for all $i \in \mathbb{Z}_{\geq 0}$ and $n \in \left[0\colon n_0^{(i)} - n_{\min}\right]$. We use induction on i. The base case $i = 0$ directly follows from Lemma 6.2, so we only prove the induction step. Since $n_0^{(i)} = n_0^{(i-1)} - n_1^{(i)}$, the induction hypothesis implies

$$\left\lfloor (n_0^{(i)} - n)x + y \right\rfloor = \left\lfloor n_0^{(i-1)}x + y \right\rfloor - \left\lfloor (n_1^{(i)} + n)x \right\rfloor .$$

On the other hand, since

$$n_1^{(i)} + n \leq n_1^{(i)} + n_0^{(i)} - n_{\min} = n_0^{(i-1)} - n_{\min},$$

Corollary 7.3 implies

$$\left\lfloor (n_1^{(i)} + n)x \right\rfloor = \left\lfloor n_1^{(i)}x \right\rfloor + \lfloor nx \rfloor .$$

Again by the induction hypothesis, we know

$$\left\lfloor n_0^{(i)}x + y \right\rfloor = \left\lfloor n_0^{(i-1)}x + y \right\rfloor - \left\lfloor n_1^{(i)}x \right\rfloor , \tag{14}$$

so combining these completes the induction step. Hence, the claim is proved.

Next, we claim that for each $i = 0, \cdots, k$,

$$n_0^{(i)} = \min \operatorname*{argmin}_{n \in \left[n_0^{(i)}\colon n_0\right]} \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

holds. We use induction on i. The base case $i = 0$ is trivial, so we show the induction step. By the induction hypothesis, we have

$$\frac{\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \zeta}{n_0^{(i-1)}} \leq \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

for all $n \in \left[n_0^{(i-1)}\colon n_0\right]$. On the other hand, the previous claim shows that for any $n \in \left[0\colon n_1^{(i)}\right]$,

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(i-1)} - n)x + y \right\rfloor - \zeta}{n_0^{(i-1)} - n} - \frac{\left\lfloor n_0^{(i)}x + y \right\rfloor - \zeta}{n_0^{(i)}} \\
&\quad = \frac{\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \lfloor nx \rfloor - \zeta}{n_0^{(i-1)} - n} - \frac{\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \left\lfloor n_1^{(i)}x \right\rfloor - \zeta}{n_0^{(i)}} \\
&\quad = \frac{\left\lfloor n_1^{(i)}x \right\rfloor}{n_0^{(i)}} - \frac{\lfloor nx \rfloor}{n_0^{(i-1)} - n} - \frac{n_0^{(i-1)}(n_1^{(i)} - n)}{n_0^{(i)}(n_0^{(i-1)} - n)} \cdot \frac{\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \zeta}{n_0^{(i-1)}}.
\end{aligned}$$

Recall that by the definition of k, we have

$$\frac{\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \zeta}{n_0^{(i-1)}} \leq \frac{\left\lfloor n_1^{(i)}x \right\rfloor}{n_1^{(i)}},$$

thus we get

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(i-1)} - n)x + y \right\rfloor - \zeta}{n_0^{(i-1)} - n} - \frac{\left\lfloor n_0^{(i)} x + y \right\rfloor - \zeta}{n_0^{(i)}} \\
&\qquad \geq \frac{n_1^{(i)}(n_0^{(i-1)} - n) - n_0^{(i-1)}(n_1^{(i)} - n)}{n_0^{(i)}(n_0^{(i-1)} - n)} \cdot \frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}} - \frac{\lfloor nx \rfloor}{n_0^{(i-1)} - n} \\
&\qquad = \frac{n \left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}(n_0^{(i-1)} - n)} - \frac{\lfloor nx \rfloor}{n_0^{(i-1)} - n}.
\end{aligned}$$

If $n = 0$, then the above is equal to zero, and if $n \neq 0$, then it is equal to

$$\frac{n}{n_0^{(i-1)} - n} \left(\frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}} - \frac{\lfloor nx \rfloor}{n} \right) \geq 0$$

by the definition of $n_1^{(i)}$. As a result, we conclude

$$\frac{\left\lfloor n_0^{(i)} x + y \right\rfloor - \zeta}{n_0^{(i)}} \leq \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

for all $n \in \left[n_0^{(i)} : n_0 \right]$, completing the induction step. Therefore, the claim is proved.

Consequently, it is enough to show that for any $n \in \left[1 \colon n_0^{(k)} - n_{\min} \right]$, we have

$$\frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}} < \frac{\left\lfloor (n_0^{(k)} - n)x + y \right\rfloor - \zeta}{n_0^{(k)} - n}.$$

This is vacuously true if $n_0^{(k)} = n_{\min}$, so suppose otherwise, thus

$$\frac{\left\lfloor n_1^{(k+1)} x \right\rfloor}{n_1^{(k+1)}} < \frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}}$$

holds. By the first claim, for any $n \in \left[1 \colon n_0^{(k)} - n_{\min} \right]$, we have

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(k)} - n)x + y \right\rfloor - \zeta}{n_0^{(k)} - n} - \frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}} \\
&\qquad = \frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \lfloor nx \rfloor - \zeta}{n_0^{(k)} - n} - \frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}} \\
&\qquad = \frac{n}{n_0^{(k)} - n} \left(\frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}} - \frac{\lfloor nx \rfloor}{n} \right),
\end{aligned}$$

and by the definition of $n_1^{(k+1)}$, the right-hand side is strictly positive. Therefore, we are done. □

Theorem 10.4 (The largest minimizer on the right)**.**
Let $n_0^{(0)} := n_0$. For $i \in \mathbb{Z}_{>0}$, inductively define $n_0^{(i)} := n_0^{(i-1)} + n_1^{(i)}$ and

$$n_1^{(i)} := \begin{cases} \max \underset{n \in \left[1 \colon n_{\max} - n_0^{(i-1)}\right]}{\operatorname{argmin}} \dfrac{\lfloor nx \rfloor + 1}{n} & \textit{if } n_0^{(i-1)} < n_{\max}, \\ 0 & \textit{otherwise.} \end{cases}$$

Let k be the smallest nonnegative integer such that either $n_0^{(k)} = n_{\max}$ or

$$\frac{\left\lfloor n_1^{(k+1)} x \right\rfloor + 1}{n_1^{(k+1)}} > \frac{\left\lfloor n_0^{(k)} x + y \right\rfloor - \zeta}{n_0^{(k)}}$$

holds. Then

$$n_0^{(k)} = \max \underset{n \in [n_0 \colon n_{\max}]}{\operatorname{argmin}} \frac{\lfloor nx + y \rfloor - \zeta}{n}.$$

Proof. First, we claim that

$$\left\lfloor (n_0^{(i)} + n)x + y \right\rfloor = \left\lfloor n_0^{(i)} x + y \right\rfloor + \lfloor nx \rfloor + 1$$

holds for all $i \in \mathbb{Z}_{\geq 0}$ and $n \in \left[1 \colon n_{\max} - n_0^{(i)}\right]$. We use induction on i. The base case $i = 0$ is precisely Lemma 6.4, so we only prove the induction step. Since $n_0^{(i)} = n_0^{(i-1)} + n_1^{(i)}$, the induction hypothesis implies

$$\left\lfloor (n_0^{(i)} + n)x + y \right\rfloor = \left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor (n_1^{(i)} + n)x \right\rfloor + 1.$$

On the other hand, since

$$n_1^{(i)} + n \leq n_1^{(i)} + n_{\max} - n_0^{(i)} = n_{\max} - n_0^{(i-1)},$$

Corollary 7.3 implies

$$\left\lfloor (n_1^{(i)} + n)x \right\rfloor = \left\lfloor n_1^{(i)} x \right\rfloor + \lfloor nx \rfloor + 1.$$

Again by the induction hypothesis, we know

$$\left\lfloor n_0^{(i)} x + y \right\rfloor = \left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor n_1^{(i)} x \right\rfloor + 1, \tag{15}$$

so combining these completes the induction step. Hence, the claim is proved.

Next, we claim that for each $i = 0, \cdots, k$,

$$n_0^{(i)} = \max \underset{n \in \left[n_0 \colon n_0^{(i)}\right]}{\operatorname{argmin}} \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

holds. We use induction on i. The base case $i = 0$ is trivial, so we show the induction step. By the induction hypothesis, we have

$$\frac{\left\lfloor n_0^{(i-1)} x + y \right\rfloor - \zeta}{n_0^{(i-1)}} \leq \frac{\lfloor nx + y \rfloor - \zeta}{n}$$

for all $n \in \left[n_0 \colon n_0^{(i-1)}\right]$. Also, note that

$$\begin{aligned}
&\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor-\zeta}{n_0^{(i-1)}}-\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\\
&\qquad=\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor-\zeta}{n_0^{(i-1)}}-\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor+\left\lfloor n_1^{(i)}x\right\rfloor+1-\zeta}{n_0^{(i)}}\\
&\qquad=\frac{n_1^{(i)}}{n_0^{(i)}}\left(\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor-\zeta}{n_0^{(i-1)}}-\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_1^{(i)}}\right),
\end{aligned}$$

and by the definition of k, the right-hand side is nonnegative. Therefore,

$$\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\le\frac{\lfloor nx+y\rfloor-\zeta}{n}$$

holds for all $n \in \left[n_0 \colon n_0^{(i-1)}\right]$.

On the other hand, the previous claim shows that for any $n \in \left[1 \colon n_1^{(i)}\right]$,

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(i-1)}+n)x+y\right\rfloor-\zeta}{n_0^{(i-1)}+n}-\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\\
&\qquad=\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor+\lfloor nx\rfloor+1-\zeta}{n_0^{(i-1)}+n}-\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor+\left\lfloor n_1^{(i)}x\right\rfloor+1-\zeta}{n_0^{(i)}}\\
&\qquad=\frac{n_0^{(i-1)}(n_1^{(i)}-n)}{n_0^{(i)}(n_0^{(i-1)}+n)}\cdot\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor-\zeta}{n_0^{(i-1)}}+\frac{\lfloor nx\rfloor+1}{n_0^{(i-1)}+n}-\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_0^{(i)}}.
\end{aligned}$$

Recall that by the definition of k, we have

$$\frac{\left\lfloor n_0^{(i-1)}x+y\right\rfloor-\zeta}{n_0^{(i-1)}}\ge\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_1^{(i)}},$$

thus we get

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(i-1)}+n)x+y\right\rfloor-\zeta}{n_0^{(i-1)}+n}-\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\\
&\qquad\ge\frac{\lfloor nx\rfloor+1}{n_0^{(i-1)}+n}-\frac{n_1^{(i)}(n_0^{(i-1)}+n)-n_0^{(i-1)}(n_1^{(i)}-n)}{n_0^{(i)}(n_0^{(i-1)}+n)}\cdot\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_1^{(i)}}\\
&\qquad=\frac{n}{n_0^{(i-1)}+n}\left(\frac{\lfloor nx\rfloor+1}{n}-\frac{\left\lfloor n_1^{(i)}x\right\rfloor+1}{n_1^{(i)}}\right)\ge 0
\end{aligned}$$

by the definition of $n_1^{(i)}$. As a result, we conclude

$$\frac{\left\lfloor n_0^{(i)}x + y\right\rfloor - \zeta}{n_0^{(i)}} \leq \frac{\lfloor nx + y\rfloor - \zeta}{n}$$

for all $n \in \left[n_0 \colon n_0^{(i)}\right]$, completing the induction step. Therefore, the claim is proved.

Consequently, it is enough to show that for any $n \in \left[1 \colon n_{\max} - n_0^{(k)}\right]$, we have

$$\frac{\left\lfloor n_0^{(k)}x + y\right\rfloor - \zeta}{n_0^{(k)}} < \frac{\left\lfloor (n_0^{(k)} + n)x + y\right\rfloor - \zeta}{n_0^{(k)} + n}.$$

This is vacuously true if $n_0^{(k)} = n_{\max}$, so suppose otherwise, thus

$$\frac{\left\lfloor n_1^{(k+1)}x\right\rfloor + 1}{n_1^{(k+1)}} > \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor + 1 - \zeta}{n_0^{(k)}}$$

holds. By the first claim, for any $n \in \left[1 \colon n_{\max} - n_0^{(k)}\right]$, we have

$$\begin{aligned}
&\frac{\left\lfloor (n_0^{(k)} + n)x + y\right\rfloor - \zeta}{n_0^{(k)} + n} - \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor - \zeta}{n_0^{(k)}} \\
&\qquad = \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor + \lfloor nx\rfloor + 1 - \zeta}{n_0^{(k)} + n} - \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor - \zeta}{n_0^{(k)}} \\
&\qquad = \frac{n}{n_0^{(k)} + n}\left(\frac{\lfloor nx\rfloor + 1}{n} - \frac{\left\lfloor n_0^{(k)}x + y\right\rfloor - \zeta}{n_0^{(k)}}\right),
\end{aligned}$$

and by the definition of $n_1^{(k+1)}$, the right-hand side is strictly positive. Therefore, we are done. □

10.3 Evaluating the inequalities on ζ

In order to apply the aforementioned theorems for actually computing the extremizers, we should evaluate inequalities of the form

$$\frac{a}{n_1} > \frac{b - \zeta}{c} \quad \text{or} \quad \frac{a}{n_1} < \frac{b - \zeta}{c}$$

where $a, b \in \mathbb{Z}$, $c \in \mathbb{Z}_{>0}$, and $n_1 \in [1 \colon n_{\max} - n_{\min}]$. Assuming that ζ is an arbitrary real number, evaluating these inequalities is not completely trivial.

To explain how to do these computations, we first rewrite them into

$$n_1\zeta > n_1 b - ac \quad \text{and} \quad n_1\zeta < n_1 b - ac.$$

Since the right-hand sides are integers, these inequalities are equivalent to

$$\lceil n_1\zeta\rceil > n_1 b - ac \quad \text{and} \quad \lfloor n_1\zeta\rfloor < n_1 b - ac.$$

Since the only terms involving ζ are of the form $\lfloor n_1\zeta \rfloor$ or $\lceil n_1\zeta \rceil$, we can just Theorem 7.4 here. Let $\frac{p_*}{q_*}, \frac{p^*}{q^*}$ be the best rational approximations of ζ from below and from above, respectively, with the largest denominators $q_*, q^* \leq n_{\max} - n_{\min}$. Then by Theorem 7.4, we know $\lfloor n_1\zeta \rfloor = \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor$, so we know

$$n_1\zeta < n_1 b - ac \quad \text{if and only if} \quad \left\lfloor \frac{n_1 p_*}{q_*} \right\rfloor < n_1 b - ac.$$

Similarly, since $\frac{-p^*}{q^*}$ is the best rational approximation of $-\zeta$ from below with the largest denominator $q^* \leq n_{\max} - n_{\min}$, Theorem 7.4 tells us that

$$\lceil n_1\zeta \rceil = -\lfloor n_1(-\zeta) \rfloor = -\left\lfloor \frac{-np^*}{q^*} \right\rfloor = \left\lceil \frac{np^*}{q^*} \right\rceil.$$

Hence, we know

$$n_1\zeta > n_1 b - ac \quad \text{if and only if} \quad \left\lfloor \frac{n_1 p^*}{q^*} \right\rfloor > n_1 b - ac.$$

Therefore, evaluations of these inequalities can be done in arbitrary-precision exact rational arithmetic.

In fact, there is another inequality involving ζ we need to evaluate: we have to choose between the maximizers/minimizers we get from Theorem 10.1 and Theorem 10.2, and Theorem 10.3 and Theorem 10.4, respectively. These are simply the matter of evaluating an inequality of the form

$$\frac{\lfloor n_L x + y \rfloor - \zeta}{n_L} \leq \frac{\lfloor n_R x + y \rfloor - \zeta}{n_R}$$

for some $n_L, n_R \in [n_{\min} \colon n_{\max}]$ with $n_L \leq n_R$. Here we are assuming that we can already evaluate $\lfloor n_L x + y \rfloor$ and $\lfloor n_R x + y \rfloor$ using Algorithm 8.2. Note that the above inequality is equivalent to

$$n_R \lfloor n_L x + y \rfloor - n_L \lfloor n_R x + y \rfloor \leq (n_R - n_L)\zeta,$$

which is equivalent to

$$n_R \lfloor n_L x + y \rfloor - n_L \lfloor n_R x + y \rfloor \leq \lfloor (n_R - n_L)\zeta \rfloor.$$

Since $n_R - n_L \in [0 \colon n_{\max} - n_{\min}]$, again evaluating the above inequality can be done by evaluating the inequality

$$n_R \lfloor n_L x + y \rfloor - n_L \lfloor n_R x + y \rfloor \leq \left\lfloor \frac{(n_R - n_L)p_*}{q_*} \right\rfloor$$

instead.

10.4 The case $n_{\min} \leq n_{\max} < 0$ and other variations

We have not yet talked about the case $n_{\min} \leq n_{\max} < 0$, but this case can be actually covered by what we have already covered. The idea is simple: we consider $-x$ instead of x. Indeed, if $n_{\min} \leq n_{\max} < 0$, then

$$\max_{n \in [n_{\min} \colon n_{\max}]} \frac{\lfloor nx + y \rfloor - \zeta}{n} = -\min_{n \in [-n_{\max} \colon -n_{\min}]} \frac{\lfloor n(-x) + y \rfloor - \zeta}{n},$$

and similarly

$$\min_{n\in[n_{\min}\colon n_{\max}]} \frac{\lfloor nx+y\rfloor - \zeta}{n} = -\max_{n\in[-n_{\max}\colon -n_{\min}]} \frac{\lfloor n(-x)+y\rfloor - \zeta}{n}.$$

Since we have not put any restrictions on the sign of x, we can apply previous arguments without any problem.

In a similar way, we can also solve the optimization problems

$$\max_{n\in[n_{\min}\colon n_{\max}]} \frac{\lceil nx+y\rceil - \zeta}{n} \quad\text{and}\quad \min_{n\in[n_{\min}\colon n_{\max}]} \frac{\lceil nx+y\rceil - \zeta}{n}.$$

In this case, we note

$$\frac{\lceil nx+y\rceil - \zeta}{n} = \frac{-\lfloor n(-x)-y\rfloor - \zeta}{n} = -\frac{\lfloor n(-x)-y\rfloor + \zeta}{n},$$

thus

$$\max_{n\in[n_{\min}\colon n_{\max}]} \frac{\lceil nx+y\rceil - \zeta}{n} = -\min_{n\in[n_{\min}\colon n_{\max}]} \frac{\lfloor n(-x)-y\rfloor + \zeta}{n},$$

and similarly

$$\min_{n\in[n_{\min}\colon n_{\max}]} \frac{\lceil nx+y\rceil - \zeta}{n} = -\max_{n\in[n_{\min}\colon n_{\max}]} \frac{\lfloor n(-x)-y\rfloor + \zeta}{n}.$$

11 The main algorithm

In this section, we finally develop our main algorithm, which finds the complete set of $(\xi,\zeta)\in\mathbb{R}^2$ such that

$$\lfloor nx+y\rfloor = \lfloor n\xi+\zeta\rfloor$$

holds for all $n\in[n_{\min}\colon n_{\max}]$.

Recall that having $\lfloor nx+y\rfloor = \lfloor n\xi+\zeta\rfloor$ for all $n\in[n_{\min}\colon n_{\max}]$ is equivalent to having

$$\max_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{>0}} \frac{\lfloor nx+y\rfloor - \zeta}{n} \le \xi < \min_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{>0}} \frac{\lfloor nx+y\rfloor + 1 - \zeta}{n},$$

$$\max_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{<0}} \frac{\lfloor nx+y\rfloor + 1 - \zeta}{n} < \xi \le \min_{n\in[n_{\min}\colon n_{\max}]\cap\mathbb{Z}_{<0}} \frac{\lfloor nx+y\rfloor - \zeta}{n},$$

possibly along with $\lfloor y\rfloor = \lfloor\zeta\rfloor$ if $0\in[n_{\min}\colon n_{\max}]$. In Section 10 we described how to find extremizers appearing in the above inequalities. However, as noted earlier, that was provided that ζ is a given fixed number just like x,y. Yet, the way the actual extremizers depend on ζ is very special, and it allows us to obtain a fairly simple description of how the bounds appearing in the above inequalities change as ζ changes.

First, note that when ζ is fixed, the admissible interval for ξ can be described as the intersection of the intervals we obtain by separately solving some optimization problems to which one of the four theorems explained in Section 10 can be directly applied. Let us call such an optimization problem, which is one of the forms

$$\pm\max_{n\in[n_0\colon m]} \frac{\lfloor n(\pm x)+y\rfloor - \zeta}{n}, \quad \pm\max_{n\in[m\colon n_0]} \frac{\lfloor n(\pm x)+y\rfloor - \zeta}{n},$$

$$\pm\min_{n\in[n_0\colon m]} \frac{\lfloor n(\pm x)+y\rfloor + 1 - \zeta}{n}, \quad \pm\min_{n\in[m\colon n_0]} \frac{\lfloor n(\pm x)+y\rfloor + 1 - \zeta}{n},$$

where n_0 is a maximizer/minimizer of the fractional part of $n(\pm x) + y$, as *an elementary optimization problem.* (The case $n = 0$ is special; it only imposes the constraint $\lfloor y \rfloor = \lfloor \zeta \rfloor$, and it does not impose any condition on ξ. For simplicity of the discussion, we will ignore it for the moment.)

Now, an important observation is that, according to the four theorems we proved in Section 10, the way ζ affects the optimizers to these elementary optimization problems is *only through* determining the iteration number k; that is, specifying x, n_0 and m in the above already completely determines the whole sequence $\left(n_0^{(i)}\right)_i$ we see in the aforementioned theorems, and the only thing ζ does is to merely "select" one of the entries of this sequence as the optimizer. More precisely, for each i, there exists an interval J_i of ζ such that the iteration number k equals i if and only if $\zeta \in J_i$. Clearly, these J_i's form a partition of the real line $\mathbb{R}$. In fact, the boundary between J_i and J_{i+1} is solely determined from the inequality at the ith step involving $n_0^{(i)}$ and $n_1^{(i+1)}$. More precisely, the threshold value determined from that inequality is monotone in i.

For instance, let us look at the following inequality from Theorem 10.1 (maximization problem on the left):

$$\frac{\left\lfloor n_1^{(i+1)}x \right\rfloor + 1}{n_1^{(i+1)}} > \frac{\left\lfloor n_0^{(i)}x + y \right\rfloor - \zeta}{n_0^{(i)}}.$$

The threshold value $\zeta^{(i)}$ is

$$\zeta^{(i)} := \frac{n_1^{(i+1)} \left\lfloor n_0^{(i)}x + y \right\rfloor - n_0^{(i)} \left(\left\lfloor n_1^{(i+1)}x \right\rfloor + 1 \right)}{n_1^{(i+1)}}$$

so that the inequality above holds if and only if $\zeta > \zeta^{(i)}$. Then the claim is that $\zeta^{(i-1)} > \zeta^{(i)}$ holds for each $i = 1, \cdots, i_0 - 1$, where i_0 is the first index such that $n_0^{(i_0)} = n_{\min}$ holds. To prove the claim, we simply compute $\zeta^{(i-1)} - \zeta^{(i)}$. In the proof of Theorem 10.1, we saw in (10) that

$$\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \left\lfloor n_0^{(i)}x + y \right\rfloor = \left\lfloor n_1^{(i)}x \right\rfloor + 1$$

holds for all i, regardless of the actual number of iterations determined by ζ. Hence, we get

$$\begin{aligned}
\zeta^{(i-1)} - \zeta^{(i)} &= \frac{n_1^{(i)} \left\lfloor n_0^{(i-1)}x + y \right\rfloor - n_0^{(i-1)} \left(\left\lfloor n_1^{(i)}x \right\rfloor + 1 \right)}{n_1^{(i)}} \\
&\qquad - \frac{n_1^{(i+1)} \left\lfloor n_0^{(i)}x + y \right\rfloor - n_0^{(i)} \left(\left\lfloor n_1^{(i+1)}x \right\rfloor + 1 \right)}{n_1^{(i+1)}} \\
&= \left(\left\lfloor n_0^{(i-1)}x + y \right\rfloor - \left\lfloor n_0^{(i)}x + y \right\rfloor \right) + n_0^{(i)} \left(\frac{\left\lfloor n_1^{(i+1)}x \right\rfloor + 1}{n_1^{(i+1)}} - \frac{\left\lfloor n_1^{(i)}x \right\rfloor + 1}{n_1^{(i)}} \right) \\
&\qquad - \left(n_0^{(i-1)} - n_0^{(i)} \right) \frac{\left\lfloor n_1^{(i)}x \right\rfloor + 1}{n_1^{(i)}} \\
&= n_0^{(i)} \left(\frac{\left\lfloor n_1^{(i+1)}x \right\rfloor + 1}{n_1^{(i+1)}} - \frac{\left\lfloor n_1^{(i)}x \right\rfloor + 1}{n_1^{(i)}} \right).
\end{aligned}$$

Now, since $n_1^{(i)}$ is a minimizer of $\frac{\lfloor nx \rfloor + 1}{n}$ for $n \in \left[1\colon n_0^{(i-1)} - n_{\min}\right]$ and $n_1^{(i+1)}$ is in this range, the right-hand side of the above must be nonnegative. In fact, it must be strictly positive since $n_1^{(i)}$ is the *largest* minimizer. Indeed, from Theorem 7.1 we know that the minimizers of $\frac{\lfloor nx \rfloor + 1}{n}$ are precisely all the multiples of some $m \in \left[1\colon n_0^{(i-1)} - n_{\min}\right]$ contained in $\left[1\colon n_0^{(i-1)} - n_{\min}\right]$. However, since

$$n_1^{(i)} + n_1^{(i+1)} \leq n_1^{(i)} + n_0^{(i)} - n_{\min} = n_1^{(i-1)} - n_{\min}$$

still holds, if both $n_1^{(i)}$ and $n_1^{(i+1)}$ were minimizers of $\frac{\lfloor nx \rfloor + 1}{n}$, then $n_1^{(i)} + n_1^{(i+1)}$ must be a minimizer as well, contradicting to the maximality of $n_1^{(i)}$. Therefore, the claim is proved.

We can apply the same analysis onto other theorems as well. For Theorem 10.2 (maximization problem on the right), we have the inequality

$$\frac{\left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}} < \frac{\left\lfloor n_0^{(i)} x + y \right\rfloor - \zeta}{n_0^{(i)}}$$

if and only if $\zeta < \zeta^{(i)}$, where

$$\zeta^{(i)} := \frac{n_1^{(i+1)} \left\lfloor n_0^{(i)} x + y \right\rfloor - n_0^{(i)} \left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}}.$$

In this case, we must have $\zeta^{(i-1)} < \zeta^{(i)}$ for each $i = 1, \cdots, i_0 - 1$ where i_0 is the first index such that $n_0^{(i_0)} = n_{\max}$ holds. To show that, recall (11) from the proof of Theorem 10.2 that

$$\left\lfloor n_0^{(i)} x + y \right\rfloor = \left\lfloor n_0^{(i-1)} x + y \right\rfloor + \left\lfloor n_1^{(i)} x \right\rfloor$$

holds for all i, which shows

$$\begin{aligned} \zeta^{(i)} - \zeta^{(i-1)} &= \left\lfloor n_1^{(i)} x \right\rfloor + n_0^{(i)} \left(\frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}} - \frac{\left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}} \right) - \left(n_0^{(i)} - n_0^{(i-1)} \right) \frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}} \\ &= n_0^{(i)} \left(\frac{\left\lfloor n_1^{(i)} x \right\rfloor}{n_1^{(i)}} - \frac{\left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}} \right). \end{aligned}$$

In the same way as before, the characterization of maximizers of $\frac{\lfloor nx \rfloor}{n}$ given in Theorem 7.1 shows that the above must be strictly positive.

For Theorem 10.3 (minimization on the left) with $\zeta \leftarrow \zeta - 1$, we have the inequality

$$\frac{\left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}} < \frac{\left\lfloor n_0^{(i)} x + y \right\rfloor + 1 - \zeta}{n_0^{(i)}}$$

if and only if $\zeta < \zeta^{(i)}$, where

$$\zeta^{(i)} := \frac{n_1^{(i+1)} \left(\left\lfloor n_0^{(i)} x + y \right\rfloor + 1 \right) - n_0^{(i)} \left\lfloor n_1^{(i+1)} x \right\rfloor}{n_1^{(i+1)}}.$$

In this case, we must have $\zeta^{(i-1)} < \zeta^{(i)}$ for each $i = 1, \cdots, i_0 - 1$ where i_0 is the first index such that $n_0^{(i_0)} = n_{\min}$ holds. Again, recall that (14) from the proof of Theorem 10.3 tells us

$$\left\lfloor n_0^{(i)}x + y\right\rfloor = \left\lfloor n_0^{(i-1)}x + y\right\rfloor - \left\lfloor n_1^{(i)}x\right\rfloor$$

holds for all i, which shows

$$\begin{aligned}\zeta^{(i)} - \zeta^{(i-1)} &= -\left\lfloor n_1^{(i)}x\right\rfloor + n_0^{(i)}\left(\frac{\left\lfloor n_1^{(i)}x\right\rfloor}{n_1^{(i)}} - \frac{\left\lfloor n_1^{(i+1)}x\right\rfloor}{n_1^{(i+1)}}\right) + \left(n_0^{(i-1)} - n_0^{(i)}\right)\frac{\left\lfloor n_1^{(i)}x\right\rfloor}{n_1^{(i)}} \\ &= n_0^{(i)}\left(\frac{\left\lfloor n_1^{(i)}x\right\rfloor}{n_1^{(i)}} - \frac{\left\lfloor n_1^{(i+1)}x\right\rfloor}{n_1^{(i+1)}}\right),\end{aligned}$$

and again since $n_1^{(i)}$ is the *largest* maximizer, Theorem 7.1 shows that the above must be strictly positive.

Finally, for Theorem 10.4 (minimization on the right) with $\zeta \leftarrow \zeta - 1$, we have the inequality

$$\frac{\left\lfloor n_1^{(i+1)}x\right\rfloor + 1}{n_1^{(i+1)}} > \frac{\left\lfloor n_0^{(i)}x + y\right\rfloor + 1 - \zeta}{n_0^{(i)}}$$

if and only if $\zeta > \zeta^{(i)}$, where

$$\zeta^{(i)} := \frac{n_1^{(i+1)}\left(\left\lfloor n_0^{(i)}x + y\right\rfloor + 1\right) - n_0^{(i)}\left(\left\lfloor n_1^{(i+1)}x\right\rfloor + 1\right)}{n_1^{(i+1)}}.$$

In this case, we must have $\zeta^{(i-1)} > \zeta^{(i)}$ for each $i = 1, \cdots, i_0 - 1$ where i_0 is the first index such that $n_0^{(i_0)} = n_{\max}$ holds. Recall that (15) from the proof of Theorem 10.4 tells us

$$\left\lfloor n_0^{(i)}x + y\right\rfloor = \left\lfloor n_0^{(i-1)}x + y\right\rfloor + \left\lfloor n_1^{(i)}x\right\rfloor + 1$$

holds for all i, which shows

$$\begin{aligned}\zeta^{(i-1)} - \zeta^{(i)} &= -\left(\left\lfloor n_1^{(i)}x\right\rfloor + 1\right) + n_0^{(i)}\left(\frac{\left\lfloor n_1^{(i+1)}x\right\rfloor + 1}{n_1^{(i+1)}} - \frac{\left\lfloor n_1^{(i)}x\right\rfloor + 1}{n_1^{(i)}}\right) \\ &\quad + \left(n_0^{(i)} - n_0^{(i-1)}\right)\frac{\left\lfloor n_1^{(i)}x\right\rfloor + 1}{n_1^{(i)}} \\ &= n_0^{(i)}\left(\frac{\left\lfloor n_1^{(i+1)}x\right\rfloor + 1}{n_1^{(i+1)}} - \frac{\left\lfloor n_1^{(i)}x\right\rfloor + 1}{n_1^{(i)}}\right),\end{aligned}$$

and again since $n_1^{(i)}$ is the *largest* minimizer, Theorem 7.1 shows that the above must be strictly positive.

Therefore, for given elementary optimization problem, we can completely characterize the set of (ξ, ζ) satisfying all the constraints given by the problem. More explicitly, we have the following conclusions:

1. For a maximization problem on the left with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\geq \max_{n\in[m,n_0]}\frac{\lfloor nx+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\geq\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\text{ and }\zeta\in\left(\zeta^{(i)},\zeta^{(i-1)}\right]\right\}$$

where $\zeta^{(-1)}:=+\infty$ and $\zeta^{(i_0)}:=-\infty$.

2. For a maximization problem on the left with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\leq -\max_{n\in[m,n_0]}\frac{\lfloor n(-x)+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\leq\frac{\left\lfloor -n_0^{(i)}x+y\right\rfloor-\zeta}{-n_0^{(i)}}\text{ and }\zeta\in\left(\zeta^{(i)},\zeta^{(i-1)}\right]\right\}$$

where $\zeta^{(-1)}:=+\infty$ and $\zeta^{(i_0)}:=-\infty$.

3. For a maximization problem on the right with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\geq \max_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\geq\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor-\zeta}{n_0^{(i)}}\text{ and }\zeta\in\left[\zeta^{(i-1)},\zeta^{(i)}\right)\right\}$$

where $\zeta^{(-1)}:=-\infty$ and $\zeta^{(i_0)}:=+\infty$.

4. For a maximization problem on the right with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\leq -\max_{n\in[n_0,m]}\frac{\lfloor n(-x)+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi\leq\frac{\left\lfloor -n_0^{(i)}x+y\right\rfloor-\zeta}{-n_0^{(i)}}\text{ and }\zeta\in\left[\zeta^{(i-1)},\zeta^{(i)}\right)\right\}$$

where $\zeta^{(-1)}:=-\infty$ and $\zeta^{(i_0)}:=+\infty$.

5. For a minimization problem on the left with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi< \min_{n\in[m,n_0]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi<\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor+1-\zeta}{n_0^{(i)}}\text{ and }\zeta\in\left[\zeta^{(i-1)},\zeta^{(i)}\right)\right\}$$

where $\zeta^{(-1)}:=-\infty$ and $\zeta^{(i_0)}:=+\infty$.

6. For a minimization problem on the left with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi > -\min_{n\in[m,n_0]}\frac{\lfloor n(-x)+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi>\frac{\left\lfloor -n_0^{(i)}x+y\right\rfloor+1-\zeta}{-n_0^{(i)}}\text{ and }\zeta\in\left[\zeta^{(i-1)},\zeta^{(i)}\right)\right\}$$

where $\zeta^{(-1)}:=-\infty$ and $\zeta^{(i_0)}:=+\infty$.

7. For a minimization problem on the right with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi < \min_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi<\frac{\left\lfloor n_0^{(i)}x+y\right\rfloor+1-\zeta}{n_0^{(i)}}\text{ and }\zeta\in\left(\zeta^{(i)},\zeta^{(i-1)}\right]\right\}$$

where $\zeta^{(-1)}:=+\infty$ and $\zeta^{(i_0)}:=-\infty$.

8. For a minimization problem on the right with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi > -\min_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcup_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi>\frac{\left\lfloor -n_0^{(i)}x+y\right\rfloor+1-\zeta}{-n_0^{(i)}}\text{ and }\zeta\in\left(\zeta^{(i)},\zeta^{(i-1)}\right]\right\}$$

where $\zeta^{(-1)}:=+\infty$ and $\zeta^{(i_0)}:=-\infty$.

We claim that the sets on the right-hand sides are equal to the one we obtain by removing the condition on ζ and replacing the union by the intersection:

1. For a maximization problem on the left with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi \geq \max_{n\in[m,n_0]}\frac{\lfloor nx+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon n_0^{(i)}\xi+\zeta\geq\left\lfloor n_0^{(i)}x+y\right\rfloor\right\}.$$

2. For a maximization problem on the left with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi \leq -\max_{n\in[m,n_0]}\frac{\lfloor n(-x)+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon -n_0^{(i)}\xi+\zeta\geq\left\lfloor -n_0^{(i)}x+y\right\rfloor\right\}.$$

3. For a maximization problem on the right with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2\colon \xi \geq \max_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2\colon n_0^{(i)}\xi+\zeta\geq\left\lfloor n_0^{(i)}x+y\right\rfloor\right\}.$$

4. For a maximization problem on the right with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2 : \xi\le -\max_{n\in[n_0,m]}\frac{\lfloor n(-x)+y\rfloor-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2 :\ -n_0^{(i)}\xi+\zeta\ge\left\lfloor -n_0^{(i)}x+y\right\rfloor\right\}.$$

5. For a minimization problem on the left with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2 : \xi<\min_{n\in[m,n_0]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2 : n_0^{(i)}\xi+\zeta<\left\lfloor n_0^{(i)}x+y\right\rfloor+1\right\}.$$

6. For a minimization problem on the left with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2 : \xi>-\min_{n\in[m,n_0]}\frac{\lfloor n(-x)+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2 :\ -n_0^{(i)}\xi+\zeta<\left\lfloor -n_0^{(i)}x+y\right\rfloor+1\right\}.$$

7. For a minimization problem on the right with positive n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2 : \xi<\min_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2 : n_0^{(i)}\xi+\zeta<\left\lfloor n_0^{(i)}x+y\right\rfloor+1\right\}.$$

8. For a minimization problem on the right with negative n's,

$$\left\{(\xi,\zeta)\in\mathbb{R}^2 : \xi>-\min_{n\in[n_0,m]}\frac{\lfloor nx+y\rfloor+1-\zeta}{n}\right\}$$
$$=\bigcap_{i=0}^{i_0}\left\{(\xi,\zeta)\in\mathbb{R}^2 :\ -n_0^{(i)}\xi+\zeta<\left\lfloor -n_0^{(i)}x+y\right\rfloor+1\right\}.$$

For simplicity, let us denote the one with the union as U and one with the intersection as I. Also, let U_i be the set inside the union so that $U=\bigcup_{i=0}^{i_0}U_i$. Since I clearly contains the left-hand sides of the above list, $U\subseteq I$ follows. On the other hand, for any $(\xi,\zeta)\in I$, ζ must be in one of the intervals between $\zeta^{(i-1)}$ and $\zeta^{(i)}$, so clearly we must have $(\xi,\zeta)\in U_i$. Therefore, $I\subseteq U$ also follows.

All we have done is to find subsets L and U of $[n_{\min}\colon n_{\max}]$ such that $\lfloor nx+y\rfloor=\lfloor n\xi+\zeta\rfloor$ for all $n\in[n_{\min}\colon n_{\max}]$ if and only if

$$n\xi+\zeta\ge\lfloor nx+y\rfloor\ \text{ for all } n\in L\quad\text{and}\quad n\xi+\zeta<\lfloor nx+y\rfloor+1 \text{ for all } n\in U.$$

The point here is that the number of elements in L and U is in general much smaller than that of $[n_{\min}\colon n_{\max}]$. By construction, it appears to me that $|L\cup U|$ is probably of $O(\log N)$

where N is the size of the interval $[n_{\min}\colon n_{\max}]$, but I have not tried to show this rigorously. Also, note that we can now include the case $n = 0$ here by adding 0 to both L and U.

We can even obtain the following generalization without any further complication: given a finite collection $\{(A_i, b_i, [n_{\min,i}\colon n_{\max\colon i}])\}_{i=1}^{r}$ of triples of a linear map $A_i\colon \mathbb{R}^2 \to \mathbb{R}^2$, a vector $b_i \in \mathbb{R}^2$ and an integer interval $[n_{\min,i}\colon n_{\max,i}]$, the set of (ξ, ζ) such that

$$\left\lfloor \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, A_i \begin{pmatrix} x \\ y \end{pmatrix} + b_i \right\rangle \right\rfloor = \left\lfloor \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, A_i \begin{pmatrix} \xi \\ \zeta \end{pmatrix} + b_i \right\rangle \right\rfloor$$

holds for all $n \in [n_{\min,i}\colon n_{\max,i}]$ for all $i = 1, \cdots, r$, can be described as the intersection of finitely many closed half-planes of the form

$$\left\{ (\xi,\zeta) \in \mathbb{R}^2\colon \left\langle A_i^T \begin{pmatrix} n \\ 1 \end{pmatrix}, \begin{pmatrix} \xi \\ \zeta \end{pmatrix} \right\rangle \geq \left\lfloor \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, A_i \begin{pmatrix} x \\ y \end{pmatrix} + b_i \right\rangle \right\rfloor - \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, b_i \right\rangle \right\}$$

or open half-planes of the form

$$\left\{ (\xi,\zeta) \in \mathbb{R}^2\colon \left\langle A_i^T \begin{pmatrix} n \\ 1 \end{pmatrix}, \begin{pmatrix} \xi \\ \zeta \end{pmatrix} \right\rangle < \left\lfloor \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, A_i \begin{pmatrix} x \\ y \end{pmatrix} + b_i \right\rangle \right\rfloor + 1 - \left\langle \begin{pmatrix} n \\ 1 \end{pmatrix}, b_i \right\rangle \right\}.$$

In this case, the $n_0^{(i)}$'s we considered are obtained by applying the previous arguments to $A_j \begin{pmatrix} x \\ y \end{pmatrix} + b_j$ instead of $\begin{pmatrix} x \\ y \end{pmatrix}$.

It is fair to say that at this point we have obtained a complete description of the set of (ξ, ζ) we are seeking for: it is the intersection of some fairly small finite collection of closed or open half-planes which we know the complete description of. Nevertheless, let us try to obtain a simpler description of the intersection, because in general many of the half-planes would be redundant and does not contribute to the intersection. Finding the intersection of closed half-planes is a well-studied problem in computational geometry, and possibility of having open half-planes is just a simple (though quite annoying) extension of it. Among others, one typical strategy is to use duality to turn the problem into the problem of finding the convex hull.

11.1 General outline for computing the intersection

We are now given with a collection $\{(\xi_i^*, \zeta_i^*, \eta_i^*)\}_{i=1}^{k} \subseteq \mathbb{R}^3$, and our next task is to find the intersection of the sets

$$H_i \coloneqq \left\{ (\xi,\zeta) \in \mathbb{R}^2\colon \xi_i^* \xi + \zeta_i^* \zeta + \eta_i^* \geq 0 \right\}.$$

The case of possibly having open half-planes is explained in the next subsection. Before explaining our approach in detail, let us first note what kinds of shapes are possible the intersection can be of and clarify what exactly we want to obtain as a description of the intersection.

- If $k = 0$, of course the intersection is the entire plane $\mathbb{R}^2$.
- If the boundary of every half-plane is parallel to each other, then the intersection is either empty, an infinite parallelogram, an infinite straight line, or a half-plane. In our case, the intersection cannot be empty because we know at least (x, y) belongs to it. It also cannot be a half-plane, because the set we are ultimately looking for is the intersection of infinite parallelograms of the form

$$\left\{ (\xi,\zeta) \in \mathbb{R}^2\colon n\xi + \zeta \geq \lfloor nx + y \rfloor \right\} \cap \left\{ (\xi,\zeta) \in \mathbb{R}^2\colon n\xi + \zeta < \lfloor nx + y \rfloor + 1 \right\}$$

(with an appropriate modification if we consider the generalization taking the affine transform $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \end{pmatrix} + b$ into account; note that such a generalization may turn an infinite parallelogram into the entire $\mathbb{R}^2$, but it can never turn it into a half-plane), although we have separately compactified the list of n's we need to consider for the lower bounds and the upper bounds so that for each i there does not need to exist a j such that $(\xi_i^*, \zeta_i^*) = -(\xi_j^*, \zeta_j^*)$.

- Otherwise, the intersection must be a (possibly degenerate) bounded convex polygon, because the intersection of two non-parallel parallelograms is already bounded, and the intersection of convex sets is convex. The polygon may have nonempty interior, or may degenerate into either a point or a bounded line segment.

Remark 11.1.
The degenerate cases, that the intersection becomes a region of dimension strictly less than 2, can only happen for the generalization involving affine transforms, and it is not possible to happen in the original problem, because as noted in Section 9, the set of (ξ, ζ) *simultaneously* satisfying $\lfloor nx \rfloor = \lfloor n\xi \rfloor$ and $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ already has a nonempty interior.

For the trivial case of the entire plane $\mathbb{R}^2$, of course there is no need of any further description. For the case of single point, of course specifying the point itself is what we would do, and for the case of line segment, we can for example specify a base point b and a vector v so that $\left\{ b + tv \in \mathbb{R}^2 \colon t \in [0, 1] \right\}$ is the intersection.

For the case of infinite parallelogram, the intersection can be described by specifying a normal vector to the boundary and an interval of values that the inner product between the normal vector and (ξ, ζ) can take. The infinite straight line case then becomes a special case when the interval is a single point.

For the case of non-degenerate bounded polygon, there can be many representations (specifying the list of supporting hyperplanes being one of) but we will obtain the list of "vertical slices" of the polygon which are vertically parallel trapezoids, where the slice happens exactly at the vertices of the polygon. This feels to me like a generally useful representation for further processings, for instance, to find lattice points.

Having defined our goal more precisely, now let us see how to achieve it. Our strategy is to divide $(\xi_i^*, \zeta_i^*, \eta_i^*)$'s into those with positive ζ_i^*, negative ζ_i^*, and zero ζ_i^*, compute the intersection of them separately, and then compute the intersection of the resulting intersections. We assume that none of $(\xi_i^*, \zeta_i^*) \in \mathbb{R}^2$ is zero, because in our original problem such a case can happen only if the corresponding H_i is the whole plane $\mathbb{R}^2$ (in other words, it is never empty).

First, let us look at the case of positive ζ_i^*'s. For notational simplicty, suppose that all ζ_i^*'s are positive. In this case, the intersection

$$\bigcap_{i=1}^{k} H_i = \left\{ (\xi, \zeta) \in \mathbb{R}^2 \colon \zeta \geq -\frac{\xi_i^*}{\zeta_i^*}\xi - \frac{\eta_i^*}{\zeta_i^*} \right\}$$

is nothing but the *epigraph* of the continuous convex function

$$\begin{aligned} f \colon \mathbb{R} &\to \mathbb{R} \\ \xi &\mapsto \max_{i \in [1 \colon k]} \left(-\frac{\xi_i^*}{\zeta_i^*}\xi - \frac{\eta_i^*}{\zeta_i^*} \right). \end{aligned}$$

By separating hyperplane theorem, one can show that f is equal to its double *convex conjugate*, that is,

$$f(\xi) = \sup_{\xi^* \in \mathbb{R}} \left(\xi^* \xi - f^*(\xi^*)\right)$$

holds for all $\xi \in \mathbb{R}$ where we define

$$\begin{aligned} f^* \colon \mathbb{R} &\to (-\infty, +\infty] \\ \xi^* &\mapsto \sup_{\xi \in \mathbb{R}} \left(\xi^* \xi - f(\xi)\right). \end{aligned}$$

More generally, given a function $f \colon \mathbb{R} \to (-\infty, +\infty]$, *Fenchel-Moreau theorem* says that f^{**} is equal to the *convex envelope* of f, which is the supremum of every convex lower semicontinuous function bounded above by f. This is a consequence of the separating hyperplane theorem and is a standard result in the literature, which can be found for example in [8].

In general, if $\{f_i \colon \mathbb{R} \to [-\infty, +\infty]\}_{i \in I}$ is any collection of convex lower semicontinuous functions, then Fenchel-Moreau theorem shows that

$$\left(\inf_{i \in I} f_i^*\right)^* (\xi) = \sup_{\xi^*, i} \left(\xi^* \xi - f_i^*(\xi^*)\right) = \sup_{i \in I} f_i^{**}(\xi) = \sup_{i \in I} f_i(\xi),$$

thus

$$\left(\sup_{i \in I} f_i\right)^* = \left(\inf_{i \in I} f_i^*\right)^{**}$$

is the convex envelope of $\inf_{i \in I} f_i^*$.

In our case, let $f_i \colon \xi \mapsto -\frac{\xi_i^*}{\zeta_i^*}\xi - \frac{\eta_i^*}{\zeta_i^*}$, then

$$f_i^*(\xi^*) = \sup_{\xi \in \mathbb{R}} \left(\left(\xi^* + \frac{\xi_i^*}{\zeta_i^*} \right) \xi + \frac{\eta_i^*}{\zeta_i^*} \right) = \begin{cases} \frac{\eta_i^*}{\zeta_i^*} & \text{if } \xi = -\frac{\xi_i^*}{\zeta_i^*}, \\ \infty & \text{otherwise} \end{cases},$$

thus f^* is the convex envelope of the function

$$\xi^* \mapsto \min_{i \in [1\colon k]} f_i^*(\xi^*) = \begin{cases} \min_{j; \frac{\xi_i^*}{\zeta_i^*} = \frac{\xi_j^*}{\zeta_j^*}} \frac{\eta_j^*}{\zeta_j^*} & \text{if } \xi = -\frac{\xi_i^*}{\zeta_i^*} \text{ for some } i \in [1\colon k], \\ \infty & \text{otherwise.} \end{cases}$$

The convex envelope g^{**} of a function $g \colon \mathbb{R} \to (-\infty, \infty]$ is the unique function having the closed convex hull of the epigraph of g as its epigraph, so any convex hull algorithm can be utilized for finding the convex envelope. In the reference implementation, I used a variant of *Andrew's monotone chain algorithm*. Roughly speaking, we first sort the points $\left\{ \left(-\frac{\xi_i^*}{\zeta_i^*}, \frac{\eta_i^*}{\zeta_i^*} \right) \right\}_{i=1}^k$ in ascending order according to the first coordinate, and while looping over all points from left to right, add each new point to the stack after removing enough previously added points to make a counterclockwise turn. The original algorithm by A. M. Andrew is two-phased because it has to march to right from left and then come back to left from right, but in our case the second phase is not needed. We omit details here.

After running the algorithm, we find a subset of $\left\{ \left(-\frac{\xi_i^*}{\zeta_i^*}, \frac{\eta_i^*}{\zeta_i^*} \right) \right\}_{i=1}^k$ so that f^* is a piecewise affine function interpolating between those selected points. To ease notational overhead, let us again write this subset as just $\left\{\left(-\frac{\xi_i^*}{\zeta_i^*}, \frac{\eta_i^*}{\zeta_i^*}\right)\right\}_{i=1}^k$, and furthermore assume $\left(-\frac{\xi_i^*}{\zeta_i^*}\right)_{i=1}^k$ is strictly increasing. Then $f^*(\xi^*)$ turns out to be equal to

$$h_i(\xi^*) \coloneqq \frac{\eta_i^*\zeta_{i+1}^*}{\xi_{i+1}^*\zeta_i^* - \xi_i^*\zeta_{i+1}^*}\left(\xi^* + \frac{\xi_{i+1}^*}{\zeta_{i+1}^*}\right) - \frac{\eta_{i+1}^*\zeta_i^*}{\xi_{i+1}^*\zeta_i^* - \xi_i^*\zeta_{i+1}^*}\left(\xi^* + \frac{\xi_i^*}{\zeta_i^*}\right)$$

whenever $\xi^* \in \left[-\frac{\xi_{i+1}^*}{\zeta_{i+1}^*}, -\frac{\xi_i^*}{\zeta_i^*}\right]$ for some $i = 1, \cdots, k-1$, and otherwise $f^*(\xi^*) = \infty$. Or, we can write f^* as the minimum of functions $g_i \colon \mathbb{R} \to (-\infty, +\infty]$ defined as

$$g_i \colon \xi^* \mapsto \begin{cases} h_i(\xi^*) & \text{if } \xi^* \in \left[-\frac{\xi_{i+1}^*}{\zeta_{i+1}^*}, -\frac{\xi_i^*}{\zeta_i^*}\right]. \\ \infty & \text{otherwise,} \end{cases}$$

thus $f = (f^*)^* = \max_{i\in[1:\,k]} g_i^*$. Note that $g_i\left(-\frac{\xi_i^*}{\zeta_i^*}\right) = -\frac{\eta_i^*}{\zeta_i^*}$ and $g_i\left(-\frac{\xi_{i+1}^*}{\zeta_{i+1}^*}\right) = -\frac{\eta_{i+1}^*}{\zeta_{i+1}^*}$, so for each $i = 1, \cdots, k$, let

$$\xi_i \coloneqq \frac{\eta_i^*\zeta_{i+1}^* - \eta_{i+1}^*\zeta_i^*}{\xi_{i+1}^*\zeta_i^* - \xi_i^*\zeta_{i+1}^*},$$

then it follows that

$$\begin{aligned} g_i^* \colon \xi \mapsto & \sup_{\xi^* \in \left[-\frac{\xi_{i+1}^*}{\zeta_{i+1}^*}, -\frac{\xi_i^*}{\zeta_i^*}\right]} \left(\xi^*\left(\xi - \xi_i\right) + \frac{\eta_{i+1}^*\xi_i^* - \eta_i^*\xi_{i+1}^*}{\xi_{i+1}^*\zeta_i^* - \xi_i^*\zeta_{i+1}^*}\right) \\ &= \begin{cases} f_i(\xi) & \text{if } \xi \geq \xi_i, \\ f_{i+1}(\xi) & \text{if } \xi \leq \xi_i. \end{cases} \end{aligned}$$

Since f^* is convex and $\left(-\frac{\xi_i^*}{\zeta_i^*}\right)_{i=1}^k$ is strictly increasing, the slope ξ_i of h_i should be increasing in i. By merging neighboring h_i's with the same slope if necessary, we can even assume that $(\xi_i)_{i=1}^k$ is strictly increasing. Then it follows that $f(\xi) = f_i(\xi)$ if $\xi \in [\xi_{i-1}, \xi_i]$ for some $i = 1, \cdots, k$, where $\xi_0 \coloneqq -\infty$ and $\xi_n \coloneqq +\infty$. This gives a direct piecewise affine description of f, which then immediately gives a similar description of the epigraph $\bigcap_{i=1}^k H_i$ of f. Let us call these ξ_i's the *turning points* of f.

Computing the intersection for the case of negative ζ_i^*'s can be done similarly. In this case, the intersection

$$\bigcap_{i=1}^k H_i = \left\{(\xi, \zeta) \in \mathbb{R}^2 \colon \zeta \leq -\frac{\xi_i^*}{\zeta_i^*}\xi - \frac{\eta_i^*}{\zeta_i^*}\right\}$$

is the *hypograph* of the concanve function

$$\begin{aligned} f \colon & \mathbb{R} \to \mathbb{R} \\ & \xi \mapsto \min_{i\in[1:\,k]}\left(-\frac{\xi_i^*}{\zeta_i^*}\xi - \frac{\eta_i^*}{\zeta_i^*}\right), \end{aligned}$$

or equivalently, inversion of the epigraph of the convex function

$$\begin{aligned} -f \colon & \mathbb{R} \to \mathbb{R} \\ & \xi \mapsto \max_{i\in[1:\,k]}\left(\frac{\xi_i^*}{\zeta_i^*}\xi + \frac{\eta_i^*}{\zeta_i^*}\right). \end{aligned}$$

Then all the arguments we made for the positive case apply here too without any problem.

The case of zero ζ_i^*'s is the simplest. In this case, the intersection

$$\bigcap_{i=1}^{k} H_i = \left\{(\xi, \zeta) \in \mathbb{R}^2 \colon \xi_i^* \xi \geq -\eta_i^*\right\}$$

can be found by finding the largest $\frac{-\eta_i^*}{\xi_i^*}$ among i's such that $\xi_i^* > 0$, and finding the smallest $\frac{-\eta_i^*}{\xi_i^*}$ among i's such that $\xi_i^* < 0$. (Recall that we already have exlcluded the case $\xi_i^* = 0$.)

Next, once we find the intersections for $\zeta_i^* > 0$, $\zeta_i^* < 0$, and $\zeta_i^* = 0$ separately, each denoted as P, N, and Z, we have to find the intersection $P \cap N \cap Z$. Note that computing the intersection with Z is simply a matter of cutting the region vertically, so we will focus only on how to compute the intersection of P and N, and omit details about taking the intersection with Z at the final stage.

Recall that P is the epigraph of a continuous convex piecewise affine function, and N is the hypograph of a continuous concave piecewise affine function. Let f, g denote these functions, respectively, then the region we are looking for is

$$P \cap N = \left\{(\xi, \zeta) \in \mathbb{R}^2 \colon f(\xi) \leq \zeta \leq g(\xi)\right\}.$$

We will refer to each of the affine piece of f and g defined on their respective intervals as *affine segments* of f and g. For instance, if f itself is an affine function, then its unique affine segment is f itself. On the other hand, if f is given as

$$f \colon \xi \mapsto \begin{cases} 0 & \text{if } \xi \leq 0, \\ \xi & \text{if } \xi \geq 0, \end{cases}$$

then its affine segments are functions $\xi \mapsto 0$ and $\xi \mapsto \xi$ defined on $(-\infty, 0]$ and $[0, \infty)$, respectively.

Consider the projection $\pi_\xi[P \cap N]$ of $P \cap N$ onto the ξ-axis. Note that this is precisely the set of $\xi \in \mathbb{R}$ such that $(f - g)(\xi) \leq 0$. Since $f - g$ is convex and continuous, this set must be closed and convex, that is, a closed interval. Since we know $(x, y) \in P \cap N$, $\pi_\xi[P \cap N]$ is never empty.

Suppose that $\pi_\xi[P \cap N] = \mathbb{R}$. This means $f - g \leq 0$ holds on all of $\mathbb{R}$, which implies that $f - g$ is a constant. Indeed, if not, then there exists a supporting hyperplane of $f - g$ with a nonzero slope, enforcing either $\lim_{\xi \to +\infty}(f - g)(\xi) = +\infty$ or $\lim_{\xi \to -\infty}(f - g)(\xi) = +\infty$. Hence, we have $g = f + c$ for some constant $c \geq 0$, so g should be convex since f is convex. However, since g is concave, the only possibility is that g is affine. Hence, $f = g - c$ is also affine. Therefore, we conclude that $\pi_\xi[P \cap N] = \mathbb{R}$ holds if and only if f and g are parallel affine functions.

Then $\pi_\xi[P \cap N]$ can be either of the forms $[\xi_{\min}, \xi_{\max}]$, $(-\infty, \xi_{\min}]$, or $[\xi_{\max}, \infty)$. To distinguish between these cases, it is enough to compare the slopes of the left-most and the right-most segments of f and g. More precisely, we claim that the slope of the left-most affine segment of f is strictly smaller than that of g if and only if $\inf \pi_\xi[P \cap N] > -\infty$. Indeed, if the slope of the segment of f is strictly smaller, then the left-most affine segment of $f - g$ has a strictly negative slope, so we have $\lim_{\xi \to -\infty}(f - g)(\xi) = +\infty$, which shows $\inf \pi_\xi[P \cap N] > -\infty$. On the other hand, if the slope of the segment of f is strictly larger, then similarly we conclude $\lim_{\xi \to -\infty}(f - g)(\xi) = -\infty$ so $\inf \pi_\xi[P \cap N] = -\infty$. If two slopes are the same, then $(f - g)(\xi) = c$ for some constant $c \in \mathbb{R}$ whenever ξ is small enough. Note that this c must be nonpositive; otherwise, choose any $\xi < x$ small enough so that

$(f-g)(\xi) = c$, then the slope of the line segment joining $(\xi, (f-g)(\xi))$ and $(x, (f-g)(x))$ is strictly negative, contradicting to the convexity of $f-g$. Hence, we get $\inf \pi_\xi[P \cap N] = -\infty$ in this case as well. In the same way, one can see that the slope of the right-most affine segment of f is strictly larger than that of g if and only if $\sup \pi_\xi[P \cap N] < \infty$.

In summary,

1. If the slope of the left-most affine segment of f is strictly smaller than that of g,
   (a) If the slope of the right-most affine segment of f is strictly larger than that of g, then $\pi_\xi[P \cap N]$ is of the form $[\xi_{\min}, \xi_{\max}]$ for some $\xi_{\min}, \xi_{\max} \in \mathbb{R}$ with $\xi_{\min} \leq \xi_{\max}$.
   (b) Otherwise, $\pi_\xi[P \cap N]$ is of the form $[\xi_{\min}, +\infty)$ for some $\xi_{\min} \in \mathbb{R}$.
2. Otherwise,
   (a) If the slope of the right-most affine segment of f is strictly larger than that of g, then $\pi_\xi[P \cap N]$ is of the form $(-\infty, \xi_{\max}]$ for some $\xi_{\max} \in \mathbb{R}$.
   (b) Otherwise, f and g are parallel affine functions.

We can detect and return early for the case 2(b). In this case, if the intersection Z is not the entire $\mathbb{R}^2$, then we get a bounded parallelogram, and otherwise we get an infinite parallelogram.

For all other cases, the graph of f and the graph of g should meet at some point, so we want to find their first intersection. To do so, we start from the left-most segments of f and g, and see if they intersect. If not, then compare the right endpoint of their domains, and choose the one that is smaller, and move to the next segment. If the domains have the same right endpoint, then we proceed both f and g.

When the intersection of the graphs is found for the first time, either the segments of f and g we are looking at can be parallel (that is, they are the same functions except for their domains) or not. In fact, they can be parallel only if they are the left-most segments of f and g, because otherwise the intersection should happen at the left endpoint of the domain of either of the segments, but in that case the intersection should have been already found from the previous segment. Hence, let us suppose that f and g share the same left-most segment except for their domains. This can happen only when $\pi_\xi[P \cap N]$ is of the form $(-\infty, \xi_{\max}]$, and $\xi_{\max}$ is precisely where one of the left-most segments of f and g ends, because after the segment ends the next segment must have a strictly larger slope if it were for f and a strictly smaller slope if it were for g. Hence, the convex function $f-g$ is strictly positive at some $\xi > \xi_{\max}$ if ξ close enough to $\xi_{\max}$, and by convexity $f-g$ can never hit zero again. Therefore, the shared part of the left-most segments of f and g is precisely the intersection $P \cap N$. In this case, Z must be nontrivial and the intersection $P \cap N \cap Z$ is either a line segment or a point, and for either case it is easy to get the desired description of the resulting set.

Next, suppose that the first intersection happens on non-parallel segments. In this case, what matters is how the graphs of f and g will continue immediately after the intersection. There are three possibilities: f goes strictly above g, f goes strictly below g, or f and g proceed together. These possibilities can be disinguished by comparing the slopes of the affine segments of f and g. Since we are interested in what happens *after* the intersection, we have to choose the next segment instead if the intersection happens at the right endpoint of the domain of the current segment.

If the slope for f is strictly larger, then f goes strictly above g. If we know $\inf \pi_\xi[P \cap N]$ is finite, then since this is the first intersection we found, it should coincide with $\inf \pi_\xi[P \cap N]$. By convexity, $f-g$ can never hit zero again, thus in this case $P \cap N$ should be a single point and we know exactly where the point is. On the other hand, if $\inf \pi_\xi[P \cap N]$ is infinite, then this firstly found intersection must be exactly $\xi_{\max} \coloneqq \sup \pi_\xi[P \cap N] < \infty$ because again convexity ensures that $f - g$ can never hit zero again. Then by vertically slicing the domains of the segments of f and g at all turning points of f and g before $\xi_{\max}$, we obtain a description of $P \cap N$ in terms of a union of essentially disjoint vertically parallel trapezoids. Since the set we get is not bounded, in this case Z must be not trivial, and by taking the intersection with Z, we get either a single point (if the left boundary of Z passes through $P \cap N$ exactly at $\xi_{\max}$) or a non-degenerate bounded polygon with the desired description.

If the slope of f is either same as that of g or is strictly smaller, then f either proceeds with g together or goes strictly below g, at least for a while. Note that in this case we can never have $\inf \pi_\xi[P \cap N] = -\infty$, because if that were the case, then $(f - g)(\xi)$ is nonpositive for all sufficiently small ξ while $f - g$ becomes zero at some point and is nonpositive on a nontrivial interval immediately following that point, which by convexity is only possible if $(f - g)(\xi) = 0$ for all small enough ξ. However, that means f and g share their left-most segments, which cannot be the case as we are looking at the first intersection of their graphs. Therefore, $\xi_{\min} \coloneqq \inf \pi_\xi[P \cap N]$ is finite, so the intersection we found must be exactly at $\xi_{\min}$.

Specializing further, if the slope of f is same as that of g, then they remain to be same immediately after the intersection at least for a while. Since the convexity guarantees that f can never go below g after that, either f and g separates and never meet again, or they continue to be together forever. For both cases, after taking the intersection with Z, we get either a single point or a line segment, and it is easy to obtain their desired descriptions.

If the slope for f is strictly smaller, then f goes strictly below g. If $\sup \pi_\xi[P \cap N] = \infty$, then it must be of the form $[\xi_{\min}, \infty)$. Hence, by slicing the domains of the segments of f and g at all turning points of f and g after $\xi_{\min}$, we obtain a description of $P \cap N$ in terms of a union of essentially disjoint vertically parallel trapezoids. Since the set we get is not bounded, in this case Z must be not trivial, and by taking the intersection with Z, we get either a single point (if the right boundary of Z passes through $P \cap N$ exactly at $\xi_{\min}$) or a non-degenerate bounded polygon with the desired description.

Finally, suppose $\xi_{\max} \coloneqq \sup \pi_\xi[P \cap N] < \infty$. In this case, we must find the second intersection which again can be done by successively moving onto the next segments of f and g while checking for the intersection. Meanwhile, we record the turning points of f and g and the corresponding vertically parallel trapezoidal slices. Once the second intersection is found, it must be $\xi_{\max}$, because $f - g$ is strictly negative right before the intersection and we assumed that its slope at $\xi_{\min}$ is strictly negative. Hence, the slope of $f - g$ at the intersection should be strictly positive and $f - g$ never hits zero again. Hence, we found $[\xi_{\min}, \xi_{\max}]$, and $P \cap N$ is a non-degenerate bounded polygon. Again, taking the intersection with Z can result in a single point (if either the left boundary of Z is at $\xi_{\max}$ or the right boundary of Z is at $\xi_{\min}$) or a bounded polygon.

11.2 Considertaion of open half-planes

The only issue remaining is that some of the half-planes we consider are open rather than closed. What we need to do in this case is mostly just keeping track of whether an affine segment is coming from a closed half-plane or an open half-plane, and we will not go into all details about where this can matter.

Instead, I will point out one particular issue that might be considered nontrivial. Recall that when we obtained the piecewise affine representation of the function f by computing the convex envelope of a function, we assumed that the turning points $(\xi_i)_{i=1}^k$ are strictly increasing, and this was done by merging neighboring affine segments of f^* with the same slope. In terms of how the monotone chain algorithm works, such a merging happens if we detect that points we are looking at are colinear. If all half-planes are closed, we can simply drop intermediate points and only take the extreme points. However, if some half-planes are open, then we may loss some information from it. Indeed, three points in the dual domain being colinear means that the corresponding three affine functions in the original domain having a common intersection. If, in the dual domain the two outer points are closed but the inner one is open, then this common intersection point should not be included into the intersection even though both of the visible affine segments around it are from closed half-planes. Hence, when colinear points are detected, we have to keep track of whether some of them are from open half-planes.

12 Gosper's algorithm

I completed explaining all of the main algorithm of this note, except for the algorithm for computing the continued fraction expansion of numbers we obtain by combining other numbers with known continued fraction expansions. This section is for that missing piece, called *Gosper's algorithm*, invented by Bill Gosper around 1970's.

Apparently Gosper did not publish his work into any academic journal (probably he felt it is not that significant), instead he wrote some informal notes [9] about the algorithm. Unfortunately these notes are somewhat "hidden" on the Internet behind lots of so-called "tutorials" with *wrong* explanation of the algorithm. I uploaded the original notes written by Gosper I found on the Internet into the Github repository for my C++ implementation of our main algorithm. Interested readers can take a look at it. (There probably are more notes by Gosper on this topic, but I did not feel obliged to collect all of them.)

In fact, the algorithm explained in this section is not exactly what Gosper invented. Rather, it is a somewhat more complete, but also more complicated form of it. This itself is probably a novel work, though I did not put a lot of effort to try to find existing literatures. I felt some need of this "completed" version for a fully automated implementation of our main algorithm we saw in Section 11 when I first wrote it, but it probably just works fine with the simpler, original form by Gosper. Since the algorithm explained in this section is fairly complicated, I recommend readers to just read Gosper's original notes if they want a quick implementation. I also want to remind them that Gosper's algorithm is not needed at all if all they care are rational inputs.

There are two versions of Gosper's algorithm that we want to use: unary and binary. Binary version is strictly more general, but since it is much more complicated we begin with the unary version first.

12.1 Unary Gosper's algorithm

Given a real number x, *unary Gosper's algorithm* is an algorithm for computing the continued fraction expansion of $\frac{ax+b}{cx+d}$ using the continued fraction expansion of x, for any given $a, b, c, d \in \mathbb{Z}$. For the purpose of this section, it is much more convenient to work with the projective line $\mathbb{RP}^1$ rather than the real line $\mathbb{R}$, so we will assume x is an element in $\mathbb{RP}^1$.[8]

[8] In case you are not familiar, $\mathbb{RP}^1$ is the set of all equivalence classes of $\mathbb{R}^2 \setminus \{0\}$ by the equivalence relation $x \sim y$ if and only if $y = kx$ for some $k \in \mathbb{R} \setminus \{0\}$. $\mathbb{RP}^1$ can also be regarded as the set of all 1-dimensional linear subspaces of $\mathbb{R}^2$. We can regard $\mathbb{R}$ as a subset of $\mathbb{RP}^1$, by identifying $\mathbb{R}$ with the affine subspace $\mathbb{R} \times \{1\} = \{(x,1) \in \mathbb{R}^2 : x \in \mathbb{R}\}$ of $\mathbb{R}^2$, with the obvious mapping $x \mapsto (x,1)$. Then each element $(x,1)$ in $\mathbb{R} \times \{1\}$ gives a unique equivalence class in $\mathbb{RP}^1$ namely the linear subspace spanned by $(x,1)$. On the other hand, $\mathbb{RP}^1$ contains one more element than $\mathbb{R}$, the subspace spanned by $(1,0)$. One can regard $\mathbb{RP}^1$ as the real line with this "infinity" added.

For a moment we let a, b, c, d live in $\mathbb{R}$, although we eventually want them to be in $\mathbb{Z}$. Also, it is more convenient for our purpose to work with the matrix $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ rather than the individual components.

Note that the linear map $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ induces a map $T_A\colon \operatorname{dom}(T_A) \subseteq \mathbb{RP}^1 \to \mathbb{RP}^1$ given as

$$T_A\colon [u] \mapsto [Au],$$

where $[\,\cdot\,]\colon \mathbb{R}^2 \setminus \{0\} \to \mathbb{RP}^1$ is the natural map that identifies $\mathbb{RP}^1$ as a quotient space of $\mathbb{R}^2 \setminus \{0\}$, and $\operatorname{dom}(T_A)$ is the set of $[u] \in \mathbb{RP}^1$ such that $Au \neq 0$; note that $\operatorname{dom}(T_A)$ is a well-defined subset of $\mathbb{RP}^1$. To be more precise,

1. If $\operatorname{rank} A = 0$, i.e., $A = 0$, then $\operatorname{dom}(T_A) = \emptyset$.
2. If $\operatorname{rank} A = 1$, then $\ker A$ corresponds to a unique element in $\mathbb{RP}^1$, so $\operatorname{dom}(T_A)$ is $\mathbb{RP}^1$ minus the single point set $\{\ker A\}$. Note that in this domain T_A is a constant function, whose unique value is the image $\operatorname{im} A$ of A.
3. If $\operatorname{rank} A = 2$, i.e., A is invertible, then $\operatorname{dom}(T_A) = \mathbb{RP}^1$.

For the case $\operatorname{rank} A = 1$, since T_A is constant on all of its domain, it is natural to think of its extension onto the whole $\mathbb{RP}^1$ defined to take the unique constant value even at the singularity $\ker A \in \mathbb{RP}^1$. In more fancy terms, it admits the unique *continuous extension* onto $\mathbb{RP}^1$. Let us denote this unique extension as $\overline{T}_A$ from now on. For all practical purposes, it is this extension $\overline{T}_A$ that matters so we can safely forget about the original T_A, but for the sake of developing the theory, let us be clear about this distinction.

In general, for linear maps A, B on $\mathbb{R}^2$, we have

$$\overline{T}_{AB} = \overline{T}_A \circ \overline{T}_B$$

provided that $AB \neq 0$. Indeed, it immediately follows from the definition of $T_{\cdot}$ that $T_{AB} = T_A \circ T_B$ holds on the intersection of their domains. If A is not invertible, then the right-hand side is just a constant function whose unique value is the image of A. At the same time, the left-hand side is not invertible, and since $AB \neq 0$, AB is of rank 1, so $\operatorname{im}(AB) = \operatorname{im} A$ should hold, which means that the left-hand side is also the constant function whose unique value is $\operatorname{im} A$. If B is not invertible, then since $AB \neq 0$, $\operatorname{im}(AB) = A[\operatorname{im} B]$ is not zero, so it should be a 1-dimensional subspace which must be the unique value of both sides.

With the usual identification $\mathbb{RP}^1 = \mathbb{R} \sqcup \{\infty\}$ where $(u_1 : u_2) := [(u_1, u_2)] \in \mathbb{RP}^1$ belongs to $\mathbb{R}$ if and only if $u_2 \neq 0$, in which case it corresponds to $\frac{u_1}{u_2} \in \mathbb{R}$, one can easily see that for $x = (x_1 : x_2) \in \operatorname{dom}(T_A)$,

$$T_A(x) = \frac{ax_1 + bx_2}{cx_1 + dx_2},$$

so in particular when $x \in \mathbb{R}$,

$$T_A(x) = \frac{ax + b}{cx + d}.$$

Of course, we define $T_A(x) = \infty$ if $x \in \operatorname{dom}(T_A)$ and $cx + d = 0$. If $\operatorname{rank} A = 1$, the unique constant value that T_A takes is $\frac{a}{c}$ if one of a, c is not zero, or $\frac{b}{d}$ otherwise.

The reason for considering this particular class of functions is that the transforms we perform for computing the continued fraction expansion do not leave this class of functions, and as a result the whole problem of computing the continued fraction expansion only boils down to the computing the floor of $\overline{T}_A(x)$. Indeed, suppose that we know how to compute $\left\lfloor \overline{T}_A(x) \right\rfloor$ for any $A \neq 0$, assuming $\overline{T}_A(x) \in \mathbb{R}$. When $\overline{T}_A(x) = \infty$, we can regard its continued fraction is “empty”, that is, the -1st convergent, which is always defined to be $\frac{1}{0} = \infty$ by convention, is the last convergent.

Then, we consider a new matrix

$$A' := \begin{pmatrix} c & d \\ a - \lfloor \overline{T}_A(x) \rfloor c & b - \lfloor \overline{T}_A(x) \rfloor d \end{pmatrix} = SA \tag{16}$$

where

$$S := \begin{pmatrix} 0 & 1 \\ 1 & -\lfloor \overline{T}_A(x) \rfloor \end{pmatrix}.$$

Then $\overline{T}_{A'} = \overline{T}_S \circ \overline{T}_A$ holds (even for the case $A = 0$ as both sides are just empty function in that case), thus

$$\overline{T}_{A'}(x) = \overline{T}_S \left(\overline{T}_A(x) \right) = \frac{1}{\overline{T}_A(x) - \lfloor \overline{T}_A(x) \rfloor}.$$

Hence, the 0th continued fraction coefficient of $\overline{T}_A(x)$ is $\lfloor \overline{T}_A(x) \rfloor$, the 1st coefficient of it is $\lfloor \overline{T}_{A'}(x) \rfloor$, and the 2nd coefficient can be obtained by repeating the same procedure with $A \leftarrow A'$, and so on. Note that since S is always invertible, we never get $A' = 0$ unless $A = 0$.

The main idea for computing $\lfloor \overline{T}_A(x) \rfloor$ is to use a good enough rational approximation of x. If our approximation is good enough, then we may be able to narrow down the possible range of $\lfloor \overline{T}_A(x) \rfloor$ into a single point so that we know what $\lfloor \overline{T}_A(x) \rfloor$ is. If the approximation is not good enough for that, then we refine the approximation and retry until we get a good enough one.

Since the viewpoint we want to adapt here is that it is more natural to put the problem into the framework of $\mathbb{RP}^1$ rather than of $\mathbb{R}$, we need a notion of “good approximation” in $\mathbb{RP}^1$. To incorporate that idea, we adapt the notion of *cyclic order*, which is a natural replacement for $\mathbb{RP}^1$ of the usual linear order on $\mathbb{R}$.

Definition 12.1 (Cyclic order)**.**
Given a set X, a *cyclic order* on X is a ternary relation $[\,\cdot\,,\,\cdot\,,\,\cdot\,]$ satisfying the following axioms:

1. *Cyclicity*: if $[a, b, c]$, then $[b, c, a]$.

2. *Asymmetry*: if $[a, b, c]$, then not $[c, b, a]$.

3. *Transitivity*: if $[a, b, c]$, and $[a, c, d]$, then $[a, b, d]$.

4. *Connectedness*: if a, b, c are distinct, then either $[a, b, c]$ or $[c, b, a]$.

One can think a cyclically ordered set as anything that can be arranged on a circle. Then $[a, b, c]$ means a, b, c are arranged on the circle in the counterclockwise direction. Note that $[a, b, c]$ is always false if a, b, c are not distinct. Indeed, suppose $[a, a, b]$ holds, then by the cyclicity axiom we get $[a, b, a]$, but then the asymmetry axiom contradicts itself by saying that $[a, b, a]$ should not hold. Similarly, $[a, b, b]$ is also impossible.

Identifying $\mathbb{RP}^1$ with $\mathbb{R} \sqcup \{\infty\}$, i.e., we “wrap” $\mathbb{R}$ onto a circle exactly once, with $-\infty$ and $+\infty$ being attached at a single point on the circle, then it is natural to define a cyclic order on $\mathbb{RP}^1$ as for $a, b, c \in \mathbb{RP}^1$, $[a, b, c]$ holds if and only if:

1. $a, b, c \in \mathbb{R}$ and one of $a < b < c$, $b < c < a$, and $c < a < b$ holds, or

2. $a = \infty$, $b, c \in \mathbb{R}$ and $b < c$ holds, or

3. $b = \infty$, $c, a \in \mathbb{R}$ and $c < a$ holds, or

4. $c = \infty$, $a, b \in \mathbb{R}$ and $a < b$ holds.

In fact, there is a more natural way of writing this in terms of the representation of $\mathbb{RP}^1$ as a quotient space of $\mathbb{R}^2 \setminus \{0\}$ [9]: for $a, b, c \in \mathbb{RP}^1$ with the representation $a = (a_1 : a_2)$, $b = (b_1 : b_2)$, $c = (c_1 : c_2)$, we have

$$[a, b, c] \quad \text{if and only if} \quad (a_1 b_2 - a_2 b_1)(b_1 c_2 - b_2 c_1)(c_1 a_2 - c_2 a_1) > 0. \tag{17}$$

(Note that the cyclic order on $\mathbb{RP}^1$ depends on how we embed $\mathbb{R}$ into $\mathbb{RP}^1$, i.e., whether we identify $(a_1 : a_2)$ with $\frac{a_1}{a_2}$ or $\frac{a_2}{a_1}$. The above equivalence is only true assuming the first convention. We should invert the inequality if we use the other convention.)

Definition 12.2 (Cyclic intervals)**.**
Let X be a set and $[\,\cdot\,,\,\cdot\,,\,\cdot\,]$ a cyclic order on X. A *cyclic open interval* in X is a subset of X which is either $\emptyset$, X, or of the form

$$(s, t) \coloneqq \{x \in X : [s, x, t]\}$$

for some distinct $s, t \in X$. Similarly, a *cyclic closed interval* in X is a subset of X which is either $\emptyset$, X, a singleton set, or of the form

$$[s, t] \coloneqq (s, t) \cup \{s, t\}$$

for some distinct $s, t \in X$. We similarly define the notion of *cyclic half-open intervals* and the notations $(s, t]$, $[s, t)$ in the obvious way. (Due to ambiguity, we never use the notation (s, t), $(s, t]$, $[s, t)$, or $[s, t]$ when $s = t$.)

For the cyclic order on $\mathbb{RP}^1$, the cyclic open interval (s, t) is given as follows:

1. If $s, t \in \mathbb{R}$ and $s < t$, then (s, t) is the usual open interval (s, t).
2. If $s, t \in \mathbb{R}$ and $t < s$, then (s, t) is the union of $(s, \infty) \cup (-\infty, t)$ and $\{\infty\}$.
3. If $s = \infty$ and $t \in \mathbb{R}$, then (s, t) is the usual open interval $(-\infty, t)$.
4. If $s \in \mathbb{R}$ and $t = \infty$, then (s, t) is the usual open interval (s, ∞).

Note that for $\mathbb{RP}^1$, cyclic open intervals are precisely connected open subsets, while cyclic closed intervals are precisely connected closed subsets.[10]

Proposition 12.3.
Let $A \in \mathbb{R}^{2\times 2} \setminus \{0\}$ and $[s, t] \subseteq \mathbb{RP}^1$ be any cyclic closed interval. Then:

$$\overline{T}_A\,[[s, t]] = \begin{cases} [\overline{T}_A(s), \overline{T}_A(t)] & \textit{if } \det A > 0, \\ [\overline{T}_A(t), \overline{T}_A(s)] & \textit{if } \det A < 0, \\ \{\overline{T}_A(s)\} & \textit{if } \det A = 0. \end{cases}$$

Furthermore, for the first two cases, the endpoints are mapped into the endpoints.

Proof. If $\operatorname{rank} A = 1$ then $\overline{T}_A$ is a constant function, so we may assume that A is invertible.

Then $T_A = \overline{T}_A$ is a homeomorphism of $\mathbb{RP}^1$, so $T_A\,[(s, t)]$ must be a nonempty connected open proper subset of $\mathbb{RP}^1$, thus is a cyclic open interval (u, v). Similarly, $T_A\,[[s, t]]$ is a

[9] Seung uk Jang taught me about this.

[10] This can be easily proven by using the fact that $\mathbb{RP}^1$ minus any point is homeomorphic to $\mathbb{R}$.

cyclic closed interval, and by continuity this must be $[u, v]$. Since T_A is injective, we also have $\{T_A(s), T_A(t)\} = \{u, v\}$. Hence, it is enough to show that $T_A(x) \in (T_A(s), T_A(t))$ holds for all $x \in (s, t)$ if and only if $\det A > 0$.

In fact, note that for any $(u_1 : u_2)$ and $(v_1 : v_2)$ in $\mathbb{RP}^1$,

$$u_1 v_2 - u_2 v_1 = \det \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \end{pmatrix},$$

so if $T_A((u_1 : u_2)) = (u_1' : u_2')$ and $T_A((v_1 : v_2)) = (v_1' : v_2')$, then

$$u_1' v_2' - u_2' v_1' = (\det A)(u_1 v_2 - u_2 v_1).$$

Therefore, the sign of $u_1 v_2 - u_2 v_1$ is preserved by T_A if and only if $\det A > 0$, so (17) gives the desired conclusion. $\square$

Therefore, if we know in advance that our input $x \in \mathbb{RP}^1$ belongs to a nonempty cyclic closed interval I, then we can compute $\overline{T}_A[I]$ by knowing the sign of $\det A$ and evaluating T_A at the endpoints of I. Based on this idea, we now present Gosper's algorithm in detail.

The algorithm is based on the sole assumption that we have a way to obtain a decreasing sequence of cyclic intervals $(I_n)_{n=1}^{\infty}$ in $\mathbb{RP}^1$ such that $x \in I_n$ holds for all n and $\{x\} = \bigcap_{n=1}^{\infty} I_n$. (Recall that we do allow $x = \infty$.) In other words, we will only rely on a sequence of approximations of x and that is the only information available to us.

Suppose we know that x is in a cyclic interval $I \subseteq \mathbb{RP}^1$. For simplicity, let us assume I is closed.[11] Then $\overline{T}_A[I]$ should be either a single point, a cyclic closed interval of the form $[u, v]$, or $\mathbb{RP}^1$. When $\overline{T}_A[I]$ is a single point, which happens if and only if I itself is a single point or A is singular, then obviously either $\overline{T}_A(x) = \infty$ and $\overline{T}_A[I] = \{\infty\}$, or $\overline{T}_A(x) \in \mathbb{R}$ and the floor of the unique point in $\overline{T}_A[I]$ is the floor of $\overline{T}_A(x)$. So we are done.

Suppose now that $\overline{T}_A[I]$ is not a single point. If $\infty \in \overline{T}_A[I]$, then there is not very much we can know about $\lfloor \overline{T}_A(x) \rfloor$, so in this case we just refine the interval I and retry. If $\infty \notin \overline{T}_A[I]$, that is, when $\overline{T}_A[I] = [u, v]$ is a compact subinterval of $\mathbb{R}$, then we must have

$$\lfloor u \rfloor \leq \lfloor \overline{T}_A(x) \rfloor \leq \lfloor v \rfloor.$$

Therefore, if the left-hand side and the right-hand side coincide, then we can compute $\lfloor \overline{T}_A(x) \rfloor$. If $\lfloor u \rfloor \neq \lfloor v \rfloor$, then we refine the interval I and retry.

Now, note that when $\overline{T}_A(x)$ happens to be an integer, then this procedure may not succeed no matter how small I is unless $I = \{x\}$. Since $\overline{T}_A$ is a continuous function, for small enough I, $\overline{T}_A[I]$ should be contained in any small neighborhood around $\overline{T}_A(x)$, but whenever $\overline{T}_A[I]$ includes an element that is strictly below $\overline{T}_A(x)$, then $\lfloor u \rfloor = \lfloor v \rfloor$ can never happen, so in practice the computation procedure will trap inside an infinite loop. Similarly, when $\overline{T}_A(x) = \infty$, then that fact is discoverable through this procedure only if $\overline{T}_A(x)$ eventually becomes of the form $[\infty, v)$ or $\{\infty\}$. (The first case is only valid if $\lfloor \overline{T}_A(x) \rfloor$ is not the first continued fraction coefficient.)

Of course, these can happen if and only if:

1. Either $I_n = \{x\}$ for all large enough n, or
2. $\det A = 0$, or
3. $\det A > 0$ and x is the left endpoint of I_n for all large enough n, or

[11] It should be possible to improve the algorithm by allowing other types of interval, but the gain would be marginal at best at the expense of complicating the implementation a lot.

4. $\det A < 0$ and x is the right endpoint of I_n for all large enough n.

Assuming that coefficients a, b, c, d of A are all integers, when we repeatedly apply the transform $A \mapsto A'$ given in (16), $\overline{T}_A(x)$ eventually becomes an integer or ∞ if and only if $\overline{T}_A(x) \in \mathbb{Q} \cup \{\infty\}$. Actually, it can become ∞ only after $\overline{T}_A(x)$ becomes an integer, unless $\overline{T}_A(x) = \infty$ initially holds. Hence, $\overline{T}_A(x) \in \mathbb{Q} \cup \{\infty\}$ is the only potentially problematic case. Note that this implies there exists $(p : q) \in \mathbb{QP}^1$ such that

$$q(ax_1 + bx_2) = p(cx_1 + dx_2),$$

so

$$(qa - pc)x_1 + (qb - pd)x_2 = 0,$$

where $x = (x_1 : x_2)$. Since we assume $a, b, c, d \in \mathbb{Z}$, this is possibly the case only when $\det A = 0$ or $x \in \mathbb{Q} \cup \{\infty\}$. As pointed out before, the case $\det A = 0$ is actually fine as it always yield $\overline{T}_A[I_n] = \left\{\overline{T}_A(x)\right\}$.

Recall that our interval estimate I_n would be in practice based on rational approximations of x, and we may assume that I_n is always with rational enpoints, and $I_n = \{x\}$ holds for large enough n if and only if x is rational. Hence, we always end up with either $\det A = 0$ or $I_n = \{x\}$ whenever $\overline{T}_A(x) \in \mathbb{Q} \cup \{\infty\}$, so this infinite loop problem does not actually exist. Nevertheless, I want to point out that other than this potential infinite loop problem, everything should work out without any problem even if we do not impose any further assumptions on I_n's other than that it is decreasing and $\bigcap_{n=1}^{\infty} I_n = \{x\}$.

In summary, if we assume that I_n is always with rational enpoints and $I_n = \{x\}$ holds for large enough n if and only if x is rational, then the described algorithm always either succeeds in computing the continued fraction expansion of $\overline{T}_A(x)$ indefinitely.

Remark 12.4.
In Gosper's note, he recognizes that the interval estimate for $\overline{T}_A$ can be done a little bit better by observing that, except for the first coefficient matrix A, we always have $\overline{T}_A(x) \in (1, \infty]$ because $\overline{T}_{A'}(x) = \left(\overline{T}_A(x) - \left\lfloor \overline{T}_A(x) \right\rfloor\right)^{-1}$. Therefore, we can further impose the condition $\overline{T}_A(x) \in (1, \infty]$ when computing $\left\lfloor \overline{T}_A(x) \right\rfloor$ which may narrow down the list of possible values of $\left\lfloor \overline{T}_A(x) \right\rfloor$ a little bit. However, this actually does not improve the interval estimate. Indeed, note that $\overline{T}_{A'}[I]$ is necessarily equal to $\overline{T}_S\left[\overline{T}_A[I]\right]$ where

$$S \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & -\left\lfloor \overline{T}_A(x) \right\rfloor \end{pmatrix}.$$

That we were able to find $\left\lfloor \overline{T}_A(x) \right\rfloor$ means that we found out that $\overline{T}_A[I]$ is contained in the interval $\left[\left\lfloor \overline{T}_A(x) \right\rfloor, \left\lfloor \overline{T}_A(x) \right\rfloor + 1\right)$, which is equivalent to say that $\overline{T}_{A'}[I]$ is contained in $(1, \infty]$. Therefore, imposing $\overline{T}_{A'}(x) \in (1, \infty]$ does not help.

12.2 A special case of unary Gosper's algorithm

The special case of the algorithm explained in the previous section, when the cyclic closed interval I_n is chosen to be the interval formed by the $(n-1)$th and the nth convergents of x, is often called *Gosper's algorithm*. I am not sure if this special case was his original proposal before he ended up writing the notes I cited [9], but it seems like this version is more widely recognized as *Gosper's algorithm* than the one actually written in the cited notes. This seemingly wider exposure is the sole reason why I included this special case here even though I never needed it in the reference implementation.

This version of Gosper's algorithm is somewhat special in that we do not actually need to compute the endpoints of the cyclic intervals. Rather, we just need to update the coefficients a, b, c, d in an appropriate way when a new convergent is supposed to be evaluated.

Suppose that x admits the continued fraction expansion

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

Recall the recurrence relation

$$P_n := \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} = \begin{pmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{pmatrix} \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix} = \cdots = \begin{pmatrix} p_{-1} & p_{-2} \\ q_{-1} & q_{-2} \end{pmatrix} \prod_{i=0}^{n} \begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix} \tag{18}$$

where we define

$$\begin{pmatrix} p_{-1} & p_{-2} \\ q_{-1} & q_{-2} \end{pmatrix} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $C_n := \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}$ so that $P_n = P_{n-1}C_n$. Since x is in between $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_n}{q_n}$, we want to evaluate $\overline{T}_A$ at those two rational numbers. However, since $P_n = \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix}$, we see that $\overline{T}_A(p_n/q_n)$ and $\overline{T}_A(p_{n-1}/q_{n-1})$ are nothing but (the equivalence classes in $\mathbb{RP}^1$ of) two columns of the matrix AP_n. Hence, all we need to do is to keep updating the coefficient matrix AP_n by successivley multiplying C_n's.

The procedure of keep multiplying C_n to the previous coefficient matrix at each iteration can be interpreted in another way too. Note that if we define

$$x_0 := x, \quad x_{i+1} := T_{C_i^{-1}}(x_i)$$

for each i, then

$$\overline{T}_A(x) = \overline{T}_{AC_0}(x_1) = \overline{T}_{AC_0C_1}(x_2) = \cdots = \overline{T}_{AP_n}(x_{n+1}).$$

Since

$$C_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -a_i \end{pmatrix},$$

in fact we have

$$x_{i+1} = \frac{1}{x_i - a_i},$$

so inductively we get $a_{i+1} = \lfloor x_{i+1} \rfloor$ for each i. Note that this is exactly the procedure of getting the continued fraction expansion of x. Then x_n is nothing but the nth complete fraction of x, so we can interpret the computation of AP_n from AP_{n-1} as the replacement of x_{n-1} by its one-level deeper nested fraction of the expansion of x.

Since each x_n for $n > 0$ should be in $(1, \infty]$, one can interpret the procedure of comparing $T_A(p_n/q_n)$ and $T_A(p_{n-1}/q_{n-1})$ as the estimation of $\frac{at+b}{ct+d}$ from $\frac{a}{c}$ and $\frac{b}{d}$. Note that for $t \in (0, \infty)$, $\frac{at+b}{ct+d}$ would be in between $\frac{a}{c}$ and $\frac{b}{d}$, so we may say $\left\lfloor \frac{at+b}{ct+d} \right\rfloor$ is equal to $\left\lfloor \frac{a}{c} \right\rfloor$ or $\left\lfloor \frac{b}{d} \right\rfloor$ when the latter two are equal. However, this naïve idea subtly fails if we allow negative values for x, c, d, because then $\frac{at+b}{ct+d}$ no longer needs to be in between $\frac{a}{c}$ and $\frac{b}{d}$. The core reason of the failure is because the denominator of $t \mapsto \frac{at+b}{ct+d}$ can pass through 0, which means its value abruptly changes from $\pm\infty$ to $\mp\infty$ when t passes that. However, this is a natural phenomena if one works with $\mathbb{RP}^1$ instead of $\mathbb{R}$.

It seems that many sources available freely on the Internet are not so careful about this issue and suggest that inspecting the equality $\left\lfloor \frac{a}{c} \right\rfloor = \left\lfloor \frac{b}{d} \right\rfloor$ alone would be sufficient. And indeed that is wrong. For example, consider

$$T_A(x) = \frac{1-3x}{1-2x}$$

with $x = \frac{1}{4}$. Then we have $\left\lfloor \frac{a}{c} \right\rfloor = \left\lfloor \frac{b}{d} \right\rfloor = 1$ but $T_A(x) = \frac{1/4}{1/2} = \frac{1}{2}$ so $\lfloor T_A(x) \rfloor = 0$. Note that this failure persists even if we proceed one step further by replacing x by $x_1 = \frac{1}{1-\lfloor x \rfloor} = \frac{1}{x} = 4$, so that

$$T_{AC_0}(x_1) = \frac{x_1 - 3}{x_1 - 2}.$$

To avoid this trap, one must take into account of the orientation of the cyclic interval between $\overline{T}_A(p_n/q_n)$ and $\overline{T}_A(p_{n-1}/q_{n-1})$. Note that this cannot be done by just looking at these two values, because regardless of which one is bigger, there always exist two possible orientations: the one that avoids ∞ and the one that includes ∞. It is only the former one that allows us to do the floor analysis. Note that, nevertheless, the correct tracking of the orientation is in fact very simple: one just needs to remember the determinant of A. If it is positive, then the direction is preserved. If it is negative, then the direction is reversed.

12.3 Binary Gosper's algorithm

We wish to extend our argument from the previous sections into the case when there are two unknowns x, y involved. It turns out, the natural class of combinations of x, y we may consider are numbers of the following form:

$$F(x,y) := \frac{axy + bx + cy + d}{exy + fx + gy + h},$$

where $a, b, c, d, e, f, g, h \in \mathbb{Z}$. Our goal is thus to find a way to compute $\lfloor F(x,y) \rfloor$. Once we are able to do so for generic a, b, c, d, e, f, g, h, then by replacing the above f by

$$F'(x,y) = \frac{exy + fx + gy + h}{(a-ke)xy + (b-kf)x + (c-kg)y + (d-kh)} \tag{19}$$

with $k := \lfloor F(x,y) \rfloor$ and keep repeating the same procedure, we can compute the continued fraction expansion of $F(x,y)$.

Again, it is much more convenient to think x, y as elements in $\mathbb{RP}^1$ rather than $\mathbb{R}$. Let

$$A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad B := \begin{pmatrix} e & f \\ g & h \end{pmatrix}$$

and view $\mathbb{RP}^1$ as a quotient space of $\mathbb{R}^2 \setminus \{0\}$. Let $u, v \in \mathbb{R}^2 \setminus \{0\}$ and write $u = (u_1, u_2)$, $v = (v_1, v_2)$, then we can see that

$$F(u_1/u_2, v_1/v_2) = \frac{\langle u, Av \rangle}{\langle u, Bv \rangle}$$

provided that the denominators are not zero. Hence, we can naturally regard F as a function

$$F(u,v) = (\langle u, Av \rangle : \langle u, Bv \rangle) \in \mathbb{RP}^1,$$

provided that $\langle u, Av\rangle$ and $\langle u, Bv\rangle$ are not simultaneously zero. (This representation motivates us to call such a function F a *bilinear fractional mapping.*)

To be more precise, define

$$\mathrm{dom}(F) := \left\{(\langle u\rangle, \langle v\rangle) \in \mathbb{RP}^1 \times \mathbb{RP}^1 \colon \langle u, Av\rangle \neq 0 \text{ or } \langle u, Bv\rangle \neq 0\right\},$$

and

$$\begin{aligned} F\colon\ &\mathrm{dom}(F) \to \mathbb{RP}^1 \\ &(\langle u\rangle, \langle v\rangle) \mapsto (\langle u, Av\rangle : \langle u, Bv\rangle)\,. \end{aligned}$$

As in the unary case, we want to build an algorithm for computing the range of possible values of $F(x, y)$ assuming we know x, y are from some cyclic intervals $I, J \subseteq \mathbb{RP}^1$, respectively, and based on that, come up with a way to compute $\lfloor F(x, y)\rfloor$. More precisely, as in the unary case we assume we have decreasing sequences $\{I_n\}_{n=1}^{\infty}$, $\{J_n\}_{n=1}^{\infty}$ of cyclic intervals in $\mathbb{RP}^1$ such that $\{x\} = \bigcap_{n=1}^{\infty} I_n$ and $\{y\} = \bigcap_{n=1}^{\infty} J_n$, and given n, we compute the range $F[I \times J]$. Since F is continuous and $I \times J$ is connected, $F[I \times J]$ is a connected subset of $\mathbb{RP}^1$, that is, a cyclic interval. Once we figure out this cyclic interval, if that allows only one possible value of $F(x, y)$, then we output that value. To keep things not overly complicated, we will always assume in this section that the cyclic intervals I, J are closed. Hence, in particular, $I \times J$ is compact, so $F[I \times J]$ is compact as well, thus it must be a closed cyclic interval.

Before actually getting into the algorithm, we have to clarify an issue: $\mathrm{dom}(F)$ may be a proper subset of $\mathbb{RP}^1 \times \mathbb{RP}^1$. Note that it is not clear what we even mean by computing the continued fraction expansion of $F(x, y)$ when $F(x, y)$ is not even well-defined. Recall that in the unary case, this was not a big issue, because the domain of the linear fractional mapping T_A is a proper subset of $\mathbb{RP}^1$ only if the matrix A is singular, and unless A is entirely zero (which is an obvious degenerate case that can and should be rejected upfront), the mapping T_A is nothing but a constant function whenever A is singular. Therefore, it is not at all a problem to extend T_A onto $\mathbb{RP}^1$ and deal with the extension instead of the partial function T_A itself.

Similarly, we may wish to see if we can continuously extend F to a larger domain, hopefully $\mathbb{RP}^1 \times \mathbb{RP}^1$. However, unlike the unary case, this is impossible in general. A prototypical counterexample (aside from the obvious degenerate case $A = B = 0$) is $F(x, y) = xy$. Of course, there is no meaningful way of defining $F(0, \infty) = 0 \cdot \infty$.

Note that in the binary case, the function in general is much likely to posses a singularity than the unary case. In the unary case, $x \in \mathbb{RP}^1$ hits the singularity only if it is the solution of *two* equations. Since $\mathbb{RP}^1$ is only 1-dimensional, this is very unlikely. However, in the binary case, the domain is now 2-dimensional, so having at least one solution to *two* equations is quite likely. Indeed, for almost every practically meaningful examples of A, B, the domain of F, or even the extension of F is a proper subset of $\mathbb{RP}^1 \times \mathbb{RP}^1$.

Nevertheless, $\mathrm{dom}(F)$ is already a "large" subset; it is a *dense open subset*:

Proposition 12.5.
Let $\tilde{F}\colon \left(\mathbb{R}^2 \setminus \{0\}\right)^2 \to \mathbb{R}^2$ *be the bilinear mapping defined as*

$$\tilde{F}\colon (u, v) \mapsto (\langle u, Av\rangle, \langle u, Bv\rangle)$$

for some $A, B \in \mathbb{R}^{2\times 2}$*. Then the projection onto* $\mathbb{RP}^1 \times \mathbb{RP}^1$ *of the zero set* $\tilde{F}^{-1}[\{0\}]$ *is compact. Therefore,* $\mathrm{dom}(F)$ *is an open subset of* $\mathbb{RP}^1 \times \mathbb{RP}^1$*. Furthermore, if* A *and* B *are not both zero, then* $\mathrm{dom}(F)$ *is dense in* $\mathbb{RP}^1 \times \mathbb{RP}^1$*.*

Proof. Clearly, $\tilde{F}$ is continuous so $\tilde{F}^{-1}[\{0\}]$ is closed. Note that whenever $(u,v) \in \tilde{F}^{-1}[\{0\}]$, we also have $\left(\frac{u}{|u|}, \frac{v}{|v|}\right) \in \tilde{F}^{-1}[\{0\}]$. Therefore, let S^1 be the unit circle in $\mathbb{R}^2$, then the projection of $\tilde{F}^{-1}[\{0\}]$ onto $\mathbb{RP}^1 \times \mathbb{RP}^1$ and that of $\tilde{F}^{-1}[\{0\}] \cap (S^1 \times S^1)$ are same. Clearly, $\tilde{F}^{-1}[\{0\}] \cap (S^1 \times S^1)$ is compact, thus its projection onto $\mathbb{RP}^1 \times \mathbb{RP}^1$ is compact, thus is closed.

Now, suppose $\operatorname{dom}(F)$ is not dense, that is, this projection of $\tilde{F}^{-1}[\{0\}]$ has a nonempty interior. By continuity, this implies that there is an open set $U \subseteq \left(\mathbb{R}^2 \setminus \{0\}\right)^2$ such that $\tilde{F}[U]$ is contained in $\{0\}$. Hence, for given $(u,v) \in U$, the kernel of the linear functional $x \mapsto \langle u, Ax\rangle$ contains an open neighborhood of v in $\mathbb{R}^2$, which means $A^T u = 0$. On the other hand, the kernel of A^T also contains an open neighborhood of u, thus we get $A = 0$. By the same reason, $B = 0$ holds as well, as desired. □

To construct the extension of F defined on the largest possible domain, we use the following theorem by J. L. Kelley [10]: a continuous function $f\colon A \to Y$ from a dense subspace $A \subseteq X$ of a topological space X into a regular Hausdorff space Y admits a unique continuous extension onto X, if and only if for each $x \in X$, there exists $y \in Y$ such that for any net $(x_\alpha)_{\alpha\in D}$ in A converging to x, the corresponding net $(f(x_\alpha))_{\alpha\in D}$ in Y converges to y. Since $\mathbb{RP}^1$ is clearly regular and Hausdorff, we can apply this theorem to find the largest continuous extension of $F\colon \operatorname{dom}(f) \to \mathbb{RP}^1$.

Lemma 12.6.
Let $A, B \in \mathbb{R}^{2\times 2}$ be not both zero and define $F\colon (x,y) \mapsto (\langle x, Ay\rangle : \langle x, By\rangle)$ as before. Suppose $([x],[y]) \in (\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$. Then F admits a unique continuous extension onto $\operatorname{dom}(F) \cup \{([x],[y])\}$ if and only if one of the following conditions hold:

1. *A and B are linearly dependent, or*
2. *$Ay = By = 0$, or*
3. *$A^T x = B^T x = 0$.*

Proof. Let $R \coloneqq \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ be the counterclockwise 90°-rotation matrix. Then we can write

$$\begin{aligned} A &= axy^T + bx(Ry)^T + c(Rx)y^T + d(Rx)(Ry)^T, \\ B &= exy^T + fx(Ry)^T + g(Rx)y^T + h(Rx)(Ry)^T \end{aligned}$$

for some $a,b,c,d,e,f,g,h \in \mathbb{R}$. Since $([x],[y])$ is not in the domain of f, we know $a = e = 0$.

Now, suppose that $Ay = By = 0$, which means $c = g = 0$. Hence,

$$\begin{aligned} A &= (bx + dRx)(Ry)^T, \\ B &= (fx + hRx)(Ry)^T. \end{aligned}$$

Thus, take any net $(([x_\alpha],[y_\alpha]))_{\alpha\in D}$ in $\operatorname{dom}(F)$ converging to $([x],[y])$, then

$$\begin{aligned} F([x_\alpha],[y_\alpha]) &= (\langle x_\alpha, bx + dRx\rangle \langle Ry, y_\alpha\rangle : \langle x_\alpha, fx + hRx\rangle \langle Ry, y_\alpha\rangle) \\ &= (\langle x_\alpha, bx + dRx\rangle : \langle x_\alpha, fx + hRx\rangle) \end{aligned}$$

since $\langle x_\alpha, bx + Rx\rangle$ and $\langle x_\alpha, fbx + hRx\rangle$ are not simultaneously zero and $\langle Ry, y_\alpha\rangle \neq 0$ by the assumption $([x_\alpha],[y_\alpha]) \in \operatorname{dom}(F)$. If b, f are not simultaneously zero, then d, h cannot be simultaneously zero and

$$F([x_\alpha],[y_\alpha]) = (d : h)$$

for any α. On the other hand, if b, f are not simultaneously zero, then we have

$$F([x_\alpha], [y_\alpha]) \to (b : f)\,.$$

Therefore, $(F([x_\alpha], [y_\alpha]))_{\alpha\in D}$ always converges and the limit is determined independently of the choice of the net. Similarly, if $A^T x = B^T x = 0$, then one can show that $(F([x_\alpha], [y_\alpha]))_{\alpha\in D}$ converges to either $(c : g)$ or $(d : h)$, determined independently of the choice of the net.

Next, suppose that b, f are not simultaneously zero and also c, g are not simultaneously zero. If A, B are linearly dependent, that is, if $tA = sB$ for some $(s, t) \in \mathbb{R}^2 \setminus \{0\}$, then clearly F is constantly equal to $(s : t)$, so suppose otherwise. Then there are two cases: (1) $(d : h) \in \mathbb{RP}^1$ and $(b : f) = (c : g) \neq (d : h)$, or (2) $(b : f) \neq (c : g)$.

For the first case, take a nonzero vector $v = (v_1, v_2) \in \mathbb{R}^2$ such that $(b, f) = sv$ and $(c, g) = tv$ for some $s, t \in \mathbb{R}$. For each $n \in \mathbb{Z}_{>0}$, define

$$x_n \coloneqq x - \frac{s}{n}Rx, \qquad y_n \coloneqq y + \frac{t}{n}Ry.$$

The clearly, $(([x_n], [y_n]))_{n=1}^\infty$ converges to $([x], [y])$. Note that

$$\langle x_n, Ay_n\rangle = \frac{bt}{n}\left|x\right|^2\left|y\right|^2 - \frac{cs}{n}\left|x\right|^2\left|y\right|^2 - \frac{dst}{n^2}\left|x\right|^2\left|y\right|^2 = -\frac{dst\left|x\right|^2\left|y\right|^2}{n^2}$$

since $bt = cs$, and similarly

$$\langle x_n, By_n\rangle = -\frac{hst\left|x\right|^2\left|y\right|^2}{n^2}.$$

Therefore, $([x_n], [y_n]) \in \operatorname{dom}(F)$ and

$$F([x_n], [y_n]) = (d : h)$$

holds for all n. On the other hand, for each $n \in \mathbb{Z}_{>0}$, define

$$z_n \coloneqq x + \frac{t}{n}Rx, \qquad w_n \coloneqq y + \frac{s}{n}Ry.$$

The clearly, $(([z_n], [w_n]))_{n=1}^\infty$ converges to $([x], [y])$, and

$$\begin{aligned}\langle z_n, Aw_n\rangle &= \frac{bs}{n}\left|x\right|^2\left|y\right|^2 + \frac{ct}{n}\left|x\right|^2\left|y\right|^2 + \frac{dst}{n^2}\left|x\right|^2\left|y\right|^2\\ &= \frac{(s^2+t^2)v_1\left|x\right|^2\left|y\right|^2}{n} + \frac{dst\left|x\right|^2\left|y\right|^2}{n^2}\end{aligned}$$

and

$$\langle z_n, Bw_n\rangle = \frac{(s^2+t^2)v_2\left|x\right|^2\left|y\right|^2}{n} + \frac{hst\left|x\right|^2\left|y\right|^2}{n^2}.$$

Hence, $([x_n], [y_n]) \in \operatorname{dom}(F)$ whenever n is large enough and

$$F([z_n], [w_n]) = \left(v_1 + \frac{dst}{n(s^2+t^2)} : v_2 + \frac{hst}{n(s^2+t^2)}\right)$$

for such n. Hence, $F([z_n], [w_n]) \to [v] \neq (d : h)$. Thus, F does not admit a continuous extension onto $\operatorname{dom}(F) \cup \{([x], [y])\}$ in this case.

For the second case, for each $n \in \mathbb{Z}_{>0}$, define

$$(x_n, y_n) \coloneqq \left(x, y + \frac{1}{n}Ry\right), \qquad (z_n, w_n) \coloneqq \left(x + \frac{1}{n}Rx, y\right).$$

Clearly, $([x_n],[y_n]),([z_n],[w_n]) \to ([x],[y])$ as $n \to \infty$. Note that

$$\begin{aligned}\langle x_n, Ay_n\rangle &= \frac{b}{n}\,|x|^2\,|y|^2\,,\\ \langle x_n, By_n\rangle &= \frac{f}{n}\,|x|^2\,|y|^2\,,\end{aligned}$$

and since b and f are not simultaneously zero, we get $([x_n],[y_n]) \in \operatorname{dom}(F)$ and

$$F([x_n],[y_n]) = (b : f).$$

Similarly, one can see that

$$F([z_n],[w_n]) = (c : g) \neq (b : f),$$

thus F does not admit a continuous extension onto $\operatorname{dom}(F) \cup \{([x],[y])\}$ in this case. □

Theorem 12.7 (Characterization of singularities of bilinear fractional mappings)**.** *Let $A, B \in \mathbb{R}^{2\times 2}$ be not both zero and define $F\colon (x,y) \mapsto (\langle x, Ay\rangle : \langle x, By\rangle)$ as before. Let $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ be the $90°$-counterclockwise rotation matrix and L the symmetric bilinear form on $\mathbb{R}^2$ defined as*

$$L\colon (u,v) \mapsto \left\langle u, (A^TRB - B^TRA)v\right\rangle.$$

1. *If A and B are linearly dependent, then $L = 0$ and F admits a unique continuous extension onto $\mathbb{RP}^1 \times \mathbb{RP}^1$, and it is a constant function. In this case, if A, B are both scalar multiples of a nonzero matrix $C \in \mathbb{R}^{2\times 2}$, then*

$$\operatorname{dom}(F) = \left\{([x],[y]) \in \mathbb{RP}^1 \times \mathbb{RP}^1 \colon y \notin \ker C \text{ and } [x] \neq [RCy]\right\}.$$

2. *If $\det L > 0$, then $\operatorname{dom}(F) = \mathbb{RP}^1 \times \mathbb{RP}^1$.*

3. *If $\det L < 0$, then the complement $(\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$ of $\operatorname{dom}(F)$ has exactly two elements whose second coordinates are precisely the only two distinct elements $[y_1],[y_2] \in \mathbb{RP}^1$ satisfying $L(y_1,y_1) = L(y_2,y_2) = 0$. Furthermore, F does not admit any proper continuous extension onto a larger subset of $\mathbb{RP}^1 \times \mathbb{RP}^1$.*

4. *If $\det L = 0$ and $L \neq 0$, then the second coordinate of any element in the complement $(\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$ of $\operatorname{dom}(F)$ is the unique element $[y_0] \in \mathbb{RP}^1$ satisfying $L(y_0,y_0) = 0$.*

   (a) *If $Ay_0 = By_0 = 0$, then F admits a unique continuous extension onto $\mathbb{RP}^1 \times \mathbb{RP}^1$. In this case, F is independent to its second variable on its domain, thus its extension is actually a linear fractional mapping of only one variable. Furthermore, A, B should be of the form $A = uv^T$, $B = wv^T$ for some $u, v, w \in \mathbb{R}^2 \setminus \{0\}$ and u, w are linearly independent. Also,*

$$\operatorname{dom}(F) = \left\{([x],[y]) \in \mathbb{RP}^1 \times \mathbb{RP}^1 \colon \langle v, y\rangle \neq 0\right\}.$$

   (b) *If $Ay_0 \neq 0$ or $By_0 \neq 0$, then $(\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$ consists of a single element $([x_0],[y_0])$ for some $[x_0] \in \mathbb{RP}^1$. In this case, F does not admit any proper continuous extension onto a larger subset of $\mathbb{RP}^1 \times \mathbb{RP}^1$.*

5. *If $L = 0$, then F admits a unique continuous extension onto $\mathbb{RP}^1 \times \mathbb{RP}^1$. In this case, F is independent to its first variable on its domain, thus its extension is actually a linear fractional mapping of only one variable. If A, B are linearly independent, then A, B should be of the form $A = uv^T$, $B = uw^T$ for some $u, v, w \in \mathbb{R}^2 \setminus \{0\}$ and v, w are linearly independent. Also,*

$$\mathrm{dom}(F) = \left\{([x], [y]) \in \mathbb{RP}^1 \times \mathbb{RP}^1 \colon \langle x, u \rangle \neq 0\right\}.$$

Proof. Note that for any $y \in \mathbb{R}^2$, $L(y, y) = 0$ if and only if

$$\left\langle y, A^T RBy \right\rangle = \left\langle y, B^T RAy \right\rangle$$

if and only if

$$\langle Ay, RBy \rangle = \langle By, RAy \rangle = - \langle Ay, RBy \rangle$$

if and only if Ay and By are parallel. Now we show each of the claims.

1. If A and B are linearly dependent, then clearly $L = 0$. Also, in this case, F is clearly a constant function on $\mathrm{dom}(F)$, so by density it admits a unique continuous extension which is a constant function. The claim about $\mathrm{dom}(F)$ is also trivial.

2. Suppose $\det L > 0$. This means that L is either a positive-definite form or a negative definite form, so in particular $L(u, u) = 0$ holds if and only if $u = 0$. Therefore, there is no $y \in \mathbb{R}^2 \setminus \{0\}$ such that Ay and By are parallel, thus there is no $x \in \mathbb{R}^2 \setminus \{0\}$ simultaneously satisfying $\langle x, Ay \rangle = 0$ and $\langle x, By \rangle = 0$. This shows $\mathrm{dom}(F) = \mathbb{RP}^1 \times \mathbb{RP}^1$.

3. Suppose $\det L < 0$. This means that L has two distinct eigenvalues with different signs. Hence, if $\lambda_1 > 0 > -\lambda_2$ are eigenvalues with associated normalized eigenvectors $u_1, u_2 \in \mathbb{R}^2 \setminus \{0\}$, then for any $a, b \in \mathbb{R}$,

   $$L(au_1 + bu_2, au_1 + bu_2) = \lambda_1 a^2 - \lambda_2 b^2 = 0$$

   if and only if $\sqrt{\lambda_1} a = \pm\sqrt{\lambda_2} b$. Therefore, there are exactly two $[y_1], [y_2] \in \mathbb{RP}^1$ such that $L(y_1, y_1) = L(y_2, y_2) = 0$, which are

   $$y_1 \coloneqq \sqrt{\lambda_2} u_1 + \sqrt{\lambda_1} u_2 \quad \text{and} \quad y_2 \coloneqq \sqrt{\lambda_2} u_1 - \sqrt{\lambda_1} u_2.$$

   Now, since Ay_1 and By_1 are parallel to each other, there exists a nonzero vector $u \in \mathbb{R}^2 \setminus \{0\}$ and $s, t \in \mathbb{R}$ such that $Ay_1 = su$ and $By_1 = tu$. Note that s, t cannot be both zero because that implies

   $$(A^T RB - B^T RA) y_1 = tA^T Ru - sB^T Ru = 0,$$

   but since $\det L \neq 0$, the left-hand side cannot be zero. Therefore, define $x_1 \coloneqq Ru$, then $[x_1]$ is the unique element in $\mathbb{RP}^1$ such that $([x_1], [y_1]) \notin \mathrm{dom}(F)$. Note that we have

   $$(A^T RB - B^T RA) y_1 = tA^T Ru - sB^T Ru = tA^T x_1 - sB^T x_1,$$

   but again since the left-hand side is not zero, we cannot have $A^T x_1 = B^T x_1 = 0$. Therefore, the lemma shows that F cannot be continuously extended to $\mathrm{dom}(F) \cup \{([x_1], [y_1])\}$. The same argument for y_2 shows that there uniquely exists $[x_2] \in \mathbb{RP}^1$ such that $([x_2], [y_2]) \notin \mathrm{dom}(F)$ and F cannot be continuously extended to $\mathrm{dom}(F) \cup \{([x_2], [y_2])\}$.

4. Suppose $\det L = 0$ and $L \neq 0$. This means that L has two distinct eigenvalues and one of them is zero. Then by representing L in the eigenbasis, it becomes immediately clear that the unique element $[y_0] \in \mathbb{RP}^1$ such that $L(y_0, y_0) = 0$ is the equivalence class of eigenvectors associated to the eigenvalue 0. Hence, we not only have $L(y_0, y_0) = 0$ but also have $A^T RBy_0 = B^T RAy_0$.

   (a) Suppose $Ay_0 = By_0 = 0$. Then there exist $u, v \in \mathbb{R}^2$ such that $A = u(Ry_0)^T$ and $B = v(Ry_0)^T$. Since $L \neq 0$, A, B are linearly independent, thus u, v are linearly independent. Now, let $x, y \in \mathbb{R}^2$ be arbitrary. Then

   $$\langle x, Ay\rangle = \langle x, u\rangle \langle Ry_0, y\rangle \quad \text{and} \quad \langle x, By\rangle = \langle x, v\rangle \langle Ry_0, y\rangle,$$

   thus in particular if $([x], [y]) \in (\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$, then

   $$\langle x, u\rangle \langle Ry_0, y\rangle = \langle x, v\rangle \langle Ry_0, y\rangle = 0.$$

   Since u, v are linearly independent, $\langle x, u\rangle$ and $\langle x, v\rangle$ cannot be simultaneously zero, thus we must have $\langle Ry_0, y\rangle = 0$. This shows the claim about the forms of A, B, and $\operatorname{dom} F$. Also, if $\langle Ry_0, y\rangle = 0$, then $A^T x_0 = B^T x_0 = 0$ follows, thus the lemma shows that F admits a unique continuous extension onto $\operatorname{dom}(F) \cup \{([x], [y])\}$. Since $([x], [y])$ is arbitrary, we conclude that F admits a unique continuous extension onto all of $\mathbb{RP}^1 \times \mathbb{RP}^1$. Also, since

   $$F([x], [y]) = \frac{\langle x, u\rangle \langle Ry_0, y\rangle}{\langle x, v\rangle \langle Ry_0, y\rangle} = \frac{\langle x, u\rangle}{\langle x, v\rangle}$$

   whenever $([x], [y]) \in \operatorname{dom}(F)$, the unique extension $\overline{F}$ should be given as

   $$\overline{F} \colon ([x], [y]) \mapsto \frac{\langle x, u\rangle}{\langle x, v\rangle},$$

   which is a linear fractional mapping of $[x]$.

   (b) Next, suppose $Ay_0 \neq 0$ or $By_0 \neq 0$. Without loss of generality, let us assume $Ay_0 \neq 0$. Note that $([x], [y]) \notin \operatorname{dom}(F)$ if and only if $\langle x, Ay\rangle = \langle x, By\rangle = 0$ and it implies that Ay and By are parallel, thus we must have $L(y, y) = 0$. Since $[y_0]$ is the unique element such that $L(y_0, y_0) = 0$, $[y] = [y_0]$ follows. Since $Ay_0 \neq 0$, $[x] = [RAy_0]$ also follows. Hence, define $x_0 := RAy_0$, then the complement of $\operatorname{dom}(F)$ is the singleton set $\{([x_0], [y_0])\}$.

   Since $Ay_0 \neq 0$ and A, B are linearly independent, the lemma shows that F admits a continuous extension onto $\mathbb{RP}^1 \times \mathbb{RP}^1$ if and only if $A^T x_0 = B^T x_0 = 0$. Note that

   $$RA^T x_0 = RA^T RAy_0,$$

   so using the 2×2-matrix identity

   $$RA^T R = -(\operatorname{adj} A),$$

   we get

   $$R(A^T x_0) = -(\operatorname{adj} A)Ay_0 = -(\det A)y_0.$$

   Therefore, $A^T x_0 = 0$ holds if and only if $\det A = 0$, that is, A is not invertible.

If $By_0 \neq 0$ in addition to $Ay_0 \neq 0$, then we must have $[Ay_0] = [By_0]$, thus we can perform the same computation to verify that $B^T x_0 = 0$ holds if and only if B is not invertible. On the other hand, if $By_0 = 0$, then we always have

$$B^T x_0 = (B^T RA)y_0 = (A^T RB)y_0 = 0.$$

Of course, in this case B must be not invertible. Therefore, F admits a unique continuous extension onto $\mathbb{RP}^1 \times \mathbb{RP}^1$ if and only if A, B are both not invertible. Since $A = 0$ or $B = 0$ yields $L = 0$, this means $\operatorname{rank} A = \operatorname{rank} B = 1$, so there exist $u_1, v_1, u_2, v_2 \in \mathbb{R}^2 \setminus \{0\}$ such that $A = u_1 v_1^T$ and $B = u_2 v_2^T$. However, in this case we have

$$L = v_1 u_1^T R u_2 v_2^T - v_2 u_2^T R u_1 v_1^T = \langle u_1, Ru_2 \rangle \left(v_1 v_2^T + v_2 v_1^T \right).$$

By the $2{\times}2$ determinant sum formula $\det(A+B) = \det(A) + \det(B) + \operatorname{tr}(A)\operatorname{tr}(B) - \operatorname{tr}(AB)$, we have

$$\det \left(v_1 v_2^T + v_2 v_1^T \right) = \langle v_1, v_2 \rangle^2 - |v_1|^2 |v_2|^2 ,$$

so

$$\det L = \langle u_1, Ru_2 \rangle^2 \left(\langle v_1, v_2 \rangle^2 - |v_1|^2 |v_2|^2 \right).$$

Since we assumed $\det L = 0$ and $L \neq 0$, we conclude that u_1, u_2 are not parallel and v_1, v_2 are parallel. Then since $L(y_0, y_0) = 0$, the only possibility is that v_1, v_2 are scalar multiples of Ry_0. However, this contradicts to the assumption that either of Ay_0 or By_0 is nonzero.

5. Suppose $L = 0$. This means that Ay and By are parallel for every $y \in \mathbb{R}^2$. Suppose first that there exists $y_0 \in \mathbb{R}^2 \setminus \{0\}$ such that $Ay_0 = By_0 = 0$. Then there exist $u, v \in \mathbb{R}^2$ such that $A = u(Ry_0)^T$ and $B = v(Ry_0)^T$. Since $A(Ry_0)$ and $B(Ry_0)$ should be parallel as well, it follows that u and v are parallel. Therefore, A and B are linearly dependent in this case.

   Next, suppose there exists $y_0 \in \mathbb{R}^2 \setminus \{0\}$ such that $Ay_0 \neq 0$ and $By_0 = 0$. Again, there exists $v \in \mathbb{R}^2$ such that $B = v(Ry_0)^T$. Let $x_0 := RAy_0$, then

$$B^T x_0 = B^T RAy_0 = A^T RBy_0 = 0.$$

   Since $B^T x_0 = \langle v, x_0 \rangle Ry_0$, we get $\langle v, x_0 \rangle = \langle v, RAy_0 \rangle = 0$, thus v is parallel to Ay_0. Hence, there exists $\lambda \in \mathbb{R}$ such that $B = \lambda (Ay_0)(Ry_0)^T$. Clearly, if $\lambda = 0$ then A and B are linearly dependent, so we may assume $\lambda \neq 0$. Since By is parallel to Ay for all $y \in \mathbb{R}^2$, and By is parallel to Ay_0 for all $y \in \mathbb{R}^2$, and By is not zero for all y not parallel to y_0, it follows that $A = (Ay_0)u^T$ for some $u \in \mathbb{R}^2$. If u and Ry_0 are linearly dependent, then A and B are linearly dependent, so suppose they are linearly independent.

   Now, take arbitrary $x, y \in \mathbb{R}^2$. Then by the representations of A, B we got, we know

$$\langle x, Ay \rangle = \langle x, Ay_0 \rangle \langle u, y \rangle \quad \text{and} \quad \langle x, By \rangle = \langle x, Ay_0 \rangle \langle Ry_0, y \rangle ,$$

   thus in particular if $([x], [y]) \in (\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \operatorname{dom}(F)$, then

$$\langle x, Ay_0 \rangle \langle u, y \rangle = \langle x, Ay_0 \rangle \langle Ry_0, y \rangle = 0.$$

Since u and Ry_0 are linearly independent, $\langle u, y\rangle$ and $\langle Ry_0, y\rangle$ cannot be simultaneously zero. Therefore, we must have $\langle x, Ay_0\rangle = 0$. Note that this means $A^T x = \langle x, Ay_0\rangle u$ and $B^T x = \langle x, Ay_0\rangle Ry_0$ are both zero, thus by the lemma we conclude that F extends to $\operatorname{dom}(F) \cup \{([x], [y])\}$. Since $([x], [y])$ is arbitrary, it follows that F admits a unique continuous extension to all of $\mathbb{RP}^1 \times \mathbb{RP}^1$. Also, since

$$F([x], [y]) = \frac{\langle x, Ay_0\rangle \langle u, y\rangle}{\langle x, Ay_0\rangle \langle Ry_0, y\rangle} = \frac{\langle u, y\rangle}{\langle Ry_0, y\rangle}$$

whenever $([x], [y]) \in \operatorname{dom}(F)$, the unique extension $\overline{F}$ should be given as

$$\overline{F}\colon ([x], [y]) \mapsto \frac{\langle u, y\rangle}{\langle Ry_0, y\rangle},$$

which is a linear fractional mapping of $[y]$. By symmetry, we get the same conclusion if we assume there exists $y_0 \in \mathbb{R}^2 \setminus \{0\}$ such that $Ay_0 = 0$ and $By_0 \neq 0$.

Finally, assume Ay, By are both nonzero for all $y \in \mathbb{R}^2 \setminus \{0\}$. This means that A and B are both invertible. Since $A^{-1}By$ is parallel to y for all $y \in \mathbb{R}^2$, every vector in $\mathbb{R}^2$ is an eigenvector of $A^{-1}B$. This is only possible if $A^{-1}B$ is a scalar matrix, so A and B are linearly dependent. Hence, if A and B are linearly independent, then there exists $y_0 \in \mathbb{R}^2 \setminus \{0\}$ with $Ay_0 \neq 0$ and $By_0 = 0$, or $Ay_0 = 0$ and $By_0 \neq 0$. We have already shown that in this case A, B and $\operatorname{dom}(F)$ are of the claimed forms.

□

Definition 12.8.
We denote the unique maximal continuous extension of $F\colon \operatorname{dom}(F) \to \mathbb{RP}^1$ appearing in the above theorem as $\overline{F}\colon \operatorname{dom}(\overline{F}) \to \mathbb{RP}^1$. The complement of $\operatorname{dom}(\overline{F})$ in $\mathbb{RP}^1 \times \mathbb{RP}^1$ is called the *indeterminacy locus* of F.

Corollary 12.9.
Let $A, B \in \mathbb{R}^{2\times 2}$ be not both zero.

1. *If $A = sC$, $B = tC$ for some $C \in \mathbb{R}^{2\times 2} \setminus \{0\}$, then $\overline{F}$ is constantly equal to $(s : t)$.*
2. *If A, B are linearly independent and there exist $u, v, w \in \mathbb{R}^2 \setminus \{0\}$ such that $A = uv^T$, $B = uw^T$, then*
$$\overline{F}(\,\cdot\,,\,\cdot\,)\colon (x, y) \mapsto (\langle v, y\rangle, \langle w, y\rangle)$$
and is defined everywhere on $\mathbb{RP}^1 \times \mathbb{RP}^1$.
3. *If A, B are linearly independent and there exists $u, v, w \in \mathbb{R}^2 \setminus \{0\}$ such that $A = uv^T$, $B = wv^T$, then*
$$\overline{F}(\,\cdot\,,\,\cdot\,)\colon (x, y) \mapsto (\langle x, u\rangle, \langle x, w\rangle)$$
and is defined everywhere on $\mathbb{RP}^1 \times \mathbb{RP}^1$.
4. *For all other cases, $\overline{F} = F$.*

Proof. The first three cases follow easily from the definition of $\overline{F}$ as a continuous extension of F. Suppose those are not the case, then Theorem 12.7 shows that we must have $F = \overline{F}$. □

Consequently, $\overline{F}$ either:

1. is a constant function, or

2. is a linear fractional mapping of just one variable, or
3. does not add any new point to the domain of F.

Obviously, we can just do what we did for the unary case if $\overline{F}$ falls into the first two cases, so from now on let us suppose that $\overline{F} = F$.

Note that $\mathrm{dom}(F)$ is either the entire $\mathbb{RP}^1 \times \mathbb{RP}^1$ or it avoids only one or two points. Still, if we are unlucky, we may do hit the singularity, in which case it does not even make sense to talk about the continued fraction expansion of $F(x, y)$. Hence, we want to detect if (x, y) is located at singularities or not.

Recall that the only information on our hand is the cyclic intervals $I, J \subseteq \mathbb{RP}^1$ to which x, y are known to belong, respectively. Hence, the only way we can tell if (x, y) belongs to $\mathrm{dom}(F)$ or not is to see if $I \times J$ is either completely contained in $\mathrm{dom}(F)$ or completely disjoint from it. Since $\mathrm{dom}(F)$ is open, if (x, y) indeed belongs to $\mathrm{dom}(F)$ then any sufficiently fine interval estimate I, J should yield $I \times J \subseteq \mathrm{dom}(F)$. However, since the complement of $\mathrm{dom}(F)$ is discrete, it is never possible to conclude $I \times J \subseteq (\mathbb{RP}^1 \times \mathbb{RP}^1) \setminus \mathrm{dom}(F)$ and correctly reject the input (x, y), unless (the sequence of) $I \times J$ shrinks down to $\{(x, y)\}$ in finite steps.

Example 12.10.
Consider $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix}$, that is,

$$F(x, y) = \frac{xy + 1}{3x + y}.$$

In this case,

$$L = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -2 & 0 \\ 0 & 6 \end{pmatrix}.$$

Hence, $\det L < 0$, and there are two distinct y such that $L(y, y) = 0$, namely

$$y = \begin{pmatrix} \sqrt{3} \\ \pm 1 \end{pmatrix},$$

and corresponding x's are given as

$$x = \begin{pmatrix} \mp 1 \\ \sqrt{3} \end{pmatrix}.$$

That is, consider $(x, y) = \left(\mp \frac{1}{\sqrt{3}}, \pm\sqrt{3}\right)$. Then we can easily see that these (x, y)'s eliminates both the numerator and the denominator of $F(x, y)$. As shown in Theorem 12.7, these (x, y)'s are "genuine" singularities of F, and there is no meaningful way of defining F at them. In practice, if we ignore that (x, y) is a singularity and just run the algorithm explained below anyway, then we will get meaningless outputs that totally depends on how exactly the approximating sequence $((I_n, J_n))_{n=1}^{\infty}$ is chosen.

Note that the indeterminacy locus can contain points with irrational coordinates even if the coefficients of A, B are all integers. This is because the point y is determined as a solution to a quadratic equation. Therefore, even if we assume that the interval estimates do shrink down to a single point when the number is rational, we can still trap in an infinite loop while checking for existence of an indeterminacy locus inside $I \times J$, because an indeterminacy locus can be located at irrational points. However, this is rather a fundamentally unsolvable issue unless we make many more assumptions on what kinds of numbers x, y can be, which we do not want to.

Fortunately, under the assumption that entries of A, B are all integers, this problem can only occur if $\{xy, x, y, 1\}$ is $\mathbb{Q}$-linearly dependent, as we require an equation of the form $axy + bx + cy + d = 0$ to hold. (More precisely, since we are working on the projective space $\mathbb{RP}^1$, the equation we have is of the form $ax_1y_1 + bx_1y_2 + cx_2y_1 + dx_2y_2 = 0$, so $\{x_1y_1, x_1y_2, x_2y_1, x_2y_2\}$ is $\mathbb{Q}$-linearly dependent, where $x = (x_1 : x_2)$ and $y = (y_1 : y_2)$.)

Now, let us forget about this issue for a while and focus on how to actually check for the existence of an indeterminacy locus in given $I \times J$. Again, since an indeterminacy locus may be located at irrational points and we want to only rely on arbitrary-precision exact rational arithmetic, it is not the best idea to directly solve the equation $L(y, y) = 0$ and then check if the root is in the current interval estimate J of y. Rather, there is a way to check its presence in J without actually solving the equation.

Proposition 12.11.
Let $A, B \in \mathbb{R}^{2\times 2}$, and $R = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ the rotation matrix previously mentioned. Let L be the symmetric bilinear form on $\mathbb{R}^2$ defined as

$$L\colon (u, v) \mapsto \left\langle u, (A^T RB - B^T RA)v \right\rangle .$$

Then for linearly independent vectors $u, v \in \mathbb{R}^2 \setminus \{0\}$, there exists $[y] \in [[u], [v]]$ such that Ay and By are parallel if and only if:

1. *$\det L \leq 0$ and*
2. *either $L(u, u)L(v, v) \leq 0$ or $L(u, u)L(u, v) < 0$.*

Proof. As noted in the proof of Theorem 12.7, Ay and By are parallel if and only if $L(y, y) = 0$, and if $\det L > 0$ then such y does not exist.

Now, suppose $\det L \leq 0$. We want to see if $L(y, y) = 0$ admits a solution in $[[u], [v]]$. Consider the matrix

$$P = \begin{pmatrix} v_1 & u_1 \\ v_2 & u_2 \end{pmatrix}$$

so that P sends $(0, 1)$ to u and $(1, 0)$ to v. By linear independence of u, v, P must be invertible. Also, by replacing v by $-v$ if necessary, we can assume $\det P > 0$. Hence, by Proposition 12.3, T_P maps the interval $[[0 : 1], [1 : 0]] = [0, \infty]$ onto $[[u], [v]]$. Using (17), it can be easily seen that the interval $[[0 : 1], [1 : 0]]$ is precisely the line segment $t \mapsto [t : 1 - t]$. Define a function $h\colon \mathbb{R} \to \mathbb{R}$ as

$$h\colon t \mapsto L\left(P\begin{pmatrix} t \\ 1-t \end{pmatrix}, P\begin{pmatrix} t \\ 1-t \end{pmatrix}\right) = t^2 L(u, u) + 2tL(u, v) + (1 - t)^2 L(v, v)$$

so that $h(0) = L(u, u)$ and $h(1) = L(v, v)$. Clearly, if $h(0)h(1) \leq 0$, then there exists $t \in [0, 1]$ such that $h(t) = 0$, so suppose otherwise. In particular, h never becomes zero at $t = 0, 1$, so consider the change of variable $\tau := \frac{t}{1-t}$ so that

$$g(\tau) := \frac{h(t)}{(1 - t)^2} = \tau^2 L(u, u) + 2\tau L(u, v) + L(v, v).$$

Then $g(\tau) = 0$ if and only if

$$\tau = \frac{-L(u, v) \pm \sqrt{L(u, v)^2 - L(u, u)L(v, v)}}{L(u, u)},$$

and since we assumed $L(u,u)L(v,v) > 0$, both roots are of the same sign as $-\frac{L(u,v)}{L(u,u)}$. Therefore, g has a positive root if and only if $L(u,u)L(u,v) < 0$, so we get the desired conclusion. □

Remark 12.12.
Using the 2×2 matrix identity

$$RA^T R = -\operatorname{adj} A,$$

we have

$$\det(A^T RB - B^T RA) = \det(RA^T RB - RB^T RA) = \det\left((\operatorname{adj} B)A - (\operatorname{adj} A)B\right).$$

By the sum identity for 2×2 matrix determinant, the above is equal to

$$\begin{aligned} &\det\left((\operatorname{adj} B)A\right) + \det\left((\operatorname{adj} A)B\right) + \left(\operatorname{tr}\left((\operatorname{adj} B)A\right)\right)\left(\operatorname{tr}\left((\operatorname{adj} A)B\right)\right) - \operatorname{tr}\left((\operatorname{adj} B)A(\operatorname{adj} A)B\right) \\ &\quad = 2(\det A)(\det B) - \left(\operatorname{tr}\left((\operatorname{adj} B)A\right)\right)^2 + (\det A)(\det B)\operatorname{tr}(I) \\ &\quad = 4(\det A)(\det B) - \left(\operatorname{tr}\left((\operatorname{adj} B)A\right)\right)^2, \end{aligned}$$

so

$$\det L = 4(\det A)(\det B) - \left(\operatorname{tr}\left((\operatorname{adj} B)A\right)\right)^2.$$

Or, we can apply another identity $\operatorname{adj} A = (\operatorname{tr} A)I - A$ to get

$$\det L = 4(\det A)(\det B) - \left((\operatorname{tr} A)(\operatorname{tr} B) - \operatorname{tr}(AB)\right)^2,$$

so the condition $\det L \leq 0$ is equivalent to

$$\left((\operatorname{tr} A)(\operatorname{tr} B) - \operatorname{tr}(AB)\right)^2 \geq 4(\det A)(\det B).$$

Using Theorem 12.7 and Proposition 12.11, we can derive an algorithm that checks existence of an indeterminacy locus in $I \times J$ for given cyclic interval estimates I, J of x, y, assuming $F = \overline{F}$. Note, however, that Proposition 12.11 only gives a way to check if J contains the y-coordinate of the indeterminacy locus. We need to check if the associated x-coordinate also is contained in I. To rescue ourselves from all possible weird corner cases, let us first form the intersection of J with I after appropriately transforming it. Then by applying Proposition 12.11 to the intersection, we can check the condition for I and J at the same time.

Still, forming the intersecion of two cyclic intervals in $\mathbb{RP}^1$ is more complicated than taking the intersection of two intervals in $\mathbb{R}$. Intuitively, it sounds clear that when we form the intersection of two cyclic intervals $[a,b]$ and $[c,d]$, then the result can be either empty, a cyclic interval, or union of two disjoint cyclic intervals. However, I did not felt very insured about relying on this intuition, so I tried to prove that claim formally, which required me to delve a bit more into the thery of cyclic order. More specifically, I needed to know that the standard ternary cyclic order always extends to a cyclic order of arbitrarily many finite points, which is not entirely obvious just be looking at the definition. So here I wrote the proof in detail.

Definition 12.13 (Monotone functions)**.**
A function $f\colon X \to Y$ between cyclically ordered sets X, Y is said to be *monotone* if $[f(a), f(b), f(c)]$ implies $[a,b,c]$ for any $a,b,c \in X$. We say f an *embedding* if f is monotone and injective, and an *isomorphism* if it is monotone and bijective.

Proposition 12.14.
Let X, Y be cyclically ordered sets and assume $|X| \geq 3$. Then a function $f\colon X \to Y$ is an embedding if and only if $[a, b, c]$ implies $[f(a), f(b), f(c)]$ for any $a, b, c \in X$.

Proof. ($\Rightarrow$) Let a, b, c be distinct elements in X satisfying $[a, b, c]$. Then since f is injective, we have either $[f(a), f(b), f(c)]$ or $[f(c), f(b), f(a)]$. Since f is monotone, the second case is impossible as it implies $[c, b, a]$, thus we get $[f(a), f(b), f(c)]$.

($\Leftarrow$) Let $a, b, c \in X$. Suppose $[f(a), f(b), f(c)]$, then a, b, c must be distinct, so either $[a, b, c]$ or $[c, b, a]$ should hold. The latter is thus impossible because it implies $[f(c), f(b), f(a)]$, contradicting to $[f(a), f(b), f(c)]$. Hence, we obtain $[a, b, c]$, so f is monotone. To show that f is injective, let a, b be distinct elements in X. Then since $|X| \geq 3$, we can find another distinct element $c \in X$. Then either $[a, b, c]$ or $[c, b, a]$ should hold, thus by the assumption we must have either $[f(a), f(b), f(c)]$ or $[f(c), f(b), f(a)]$. For any cases, we have that $f(a) \neq f(b)$, thus concluding that f is an injection. □

Proposition 12.15.
For a positive integer n, consider the group $\mathbb{Z}/n$ of integers modulo n. We define a cyclic order on $\mathbb{Z}/n$ as $[a, b, c]$ if and only if there exists $k \in \mathbb{Z}/n$ with

$$(a + k \bmod n) < (b + k \bmod n) < (c + k \bmod n).$$

Then a bijection $\sigma\colon \mathbb{Z}/n \to \mathbb{Z}/n$ is an isomorphism (of cyclically ordered sets) if and only if there exists $k \in \mathbb{Z}/n$ such that $\sigma(i) = i + k$ for all $i \in \mathbb{Z}/n$.

Proof. Clearly, if such k exists, then σ is an isomorphism. For the other direction, we use induction on n. The conclusion we want to draw is vacuously true for $n \leq 2$, so take $n \geq 2$ and suppose that it is true for that n. Let $\sigma\colon \{0, \cdots, n\} \to \{0, \cdots, n\}$ be an isomorphism (when we identify $\{0, \cdots, n\}$ with $\mathbb{Z}/(n+1)$), then by adding an appropriate element in $\mathbb{Z}/(n+1)$, we can assume without loss of generality that $\sigma(n) = n$. Then since $\sigma|_{\{0, \cdots, n-1\}}$ is an isomorphism, the induction hypothesis shows that there exists $k = 0, \cdots, n-1$ such that $\sigma(i) + k \equiv i \pmod{n}$ holds for all $i = 0, \cdots, n-1$.

Now we show k must be zero. Suppose not, then $n - k$, n, and 0 are distinct elements in $\{0, \cdots, n\}$. Adding $k + 1$ and then taking the modulo operation by $n + 1$ then maps them into 0, k, $k + 1$, respectively, thus by definition of the cyclic order on $\mathbb{Z}/(n+1)$, we must have $[n - k, n, 0]$. However, since σ is an isomorphism, this implies $[\sigma(n-k), \sigma(n), \sigma(k)]$, which means $[0, n, k]$ with $k \leq n - 1$. This contradicts to the definition of the cyclic order on $\mathbb{Z}/(n+1)$, so we must have $k = 0$, so σ is the identity map on $\mathbb{Z}/(n+1)$. □

Definition 12.16.
Let X be a cyclically ordered set. Then for any positive integer n, we define an n-ary relation $[\,, \cdots, \,]$ on X as $[x_0, \cdots, x_{n-1}]$ (or $[x_i]_{i\in\mathbb{Z}/n}$ as an abbreviation) if and only if the function $\mathbb{Z}/n \to X$ given as $i \mapsto x_i$ is an embedding. In particular, we always have $[x_0]$, and $[x_0, x_1]$ holds if and only if $x_0 \neq x_1$.

Lemma 12.17.
Let X be a cyclically ordered set.

1. *When $n = 3$, the ternary relation defined above is precisely the original cyclic order.*

2. *For any $\{x_i\}_{i\in\mathbb{Z}/n} \subseteq X$, $[x_i]_{i\in\mathbb{Z}/n}$ implies $[x_{i+1}]_{i\in\mathbb{Z}/n}$.*

3. *For any $x_1, \cdots, x_{n+1}$, if $[x_1, \cdots, x_n]$ and $[x_n, x_{n+1}, x_1]$ hold, then $[x_1, \cdots, x_n, x_{n+1}]$ holds.*

4. *Let $x_0, \cdots, x_n \in X$ be distinct. Then there uniquely exists a permutation $\sigma \in S_n$ such that $\left[x_0, x_{\sigma(1)}, \cdots, x_{\sigma(n)}\right]$ holds.*

Proof. 1. To distinguish the two relations, let us denote the newly defined relation as $\langle \cdot, \cdot, \cdot \rangle$ for a moment. For given $x_0, x_1, x_2 \in X$, suppose $[x_0, x_1, x_2]$. Then x_0, x_1, x_2 are necessarily distinct, so the mapping $i \mapsto x_i$ from $\mathbb{Z}/3$ is injective. Then it can be easily checked that it is monotone, thus is an embedding. Hence, we conclude $\langle x_0, x_1, x_2 \rangle$. Conversely, suppose $\langle x_0, x_1, x_2 \rangle$. Note that this implies $|X| \geq 3$, so we get $[x_0, x_1, x_2]$ from the relation $[0, 1, 2]$ on $\mathbb{Z}/3$. Therefore, the desired equivalence is established.

2. Follows from that $i \mapsto i - 1$ is an isomorphism on $\mathbb{Z}/n$.

3. Let $i, j, k \in \{1, \cdots, n+1\}$, $i < j < k$. Then it is enough to show that $[x_i, x_j, x_k]$ holds. This is clear if $k \leq n$, so let $k = n+1$. If $j = n$, then we can apply the transitivity axiom to $[x_n, x_{n+1}, x_1]$ and $[x_n, x_1, x_i]$ to conclude $[x_n, x_{n+1}, x_i]$. Inductively, if we have $[x_i, x_j, x_{n+1}]$ and $i < j - 1$, then $[x_i, x_{j-1}, x_j]$ and $[x_i, x_j, x_{n+1}]$ together gives $[x_i, x_{j-1}, x_{n+1}]$. Hence, we have $[x_i, x_j, x_k]$ for all $i < j < k$, thus $[x_1, \cdots, x_{n+1}]$ follows.

4. Uniqueness follows directly from Proposition 12.15. To show existence, we use induction on n. The base case $n = 1$ is trivial. Suppose the conclusion is true for some $n \geq 1$, and let $x_0, \cdots, x_{n+1} \in X$ be any distinct elements. Then there exists a permuation $\sigma \in S_n$ such that $\left[x_0, x_{\sigma(1)}, \cdots, x_{\sigma(n)}\right]$ holds. Without loss of generality, we may assume σ is the identity. We claim that at least one of $[x_0, x_{n+1}, x_1]$, $[x_1, x_{n+1}, x_2]$, $\cdots$, $[x_{n-1}, x_{n+1}, x_n]$, $[x_n, x_{n+1}, x_0]$ holds.

   Suppose not, then we have all of $[x_1, x_{n+1}, x_0]$, $[x_2, x_{n+1}, x_1]$, $\cdots$, $[x_n, x_{n+1}, x_{n-1}]$, $[x_0, x_{n+1}, x_n]$. From $[x_{n+1}, x_0, x_1]$ and $[x_{n+1}, x_1, x_2]$, applying the transitivity axiom gives us $[x_{n+1}, x_0, x_2]$. Then again applying the transitivity axiom to it together with $[x_{n+1}, x_2, x_3]$ gives $[x_{n+1}, x_0, x_3]$. Thus, inductively, we obtain $[x_{n+1}, x_0, x_n]$, which contradicts to $[x_0, x_{n+1}, x_n]$. This shows the claim.

   Therefore, there exists $j = 0, \cdots, n$ such that $[x_j, x_{n+1}, x_{(j+1) \bmod (n+1)}]$ holds. Hence, applying *3* shows $[x_{(j+1) \bmod (n+1)}, \cdots, x_n, x_0, \cdots, x_j, x_{n+1}]$, or equivalently, $[x_0, \cdots, x_j, x_{n+1}, x_{j+1}, \cdots, x_n]$, thus we have found a desired permutation σ.

□

Therefore, for given two cyclic closed intervals $[a, b], [c, d]$ with a, b, c, d all dinstinct, there are exactly six possibilities: $[a, b, c, d]$, $[a, b, d, c]$, $[a, c, b, d]$, $[a, c, d, b]$, $[a, d, b, c]$, and $[a, d, c, b]$. This gives us the following result:

Proposition 12.18 (Intersection of cyclic intervals)**.**
Let $[a, b]$, $[c, d]$ be cyclic closed intervals in a cyclically ordered set X with $a \neq b$, $c \neq d$.

Then

$$[a,b] \cap [c,d] = \begin{cases} \emptyset & \textit{if } [a,b,c,d], \\ [a,b] & \textit{if } [a,b,d,c], \\ [c,b] & \textit{if } [a,c,b,d], \\ [c,d] & \textit{if } [a,c,d,b], \\ [a,d] & \textit{if } [a,d,b,c], \\ [a,d] \sqcup [c,b] & \textit{if } [a,d,c,b], \\ [a,b] & \textit{if } a = c \textit{ and } [a,b,d], \\ [a,d] & \textit{if } a = c \textit{ and } [a,d,b], \\ \{a\} & \textit{if } a = d \textit{ and } [a,b,c], \\ \{a\} \sqcup [c,b] & \textit{if } a = d \textit{ and } [a,c,b], \\ \{b\} & \textit{if } b = c \textit{ and } [a,b,d], \\ [a,d] \sqcup \{b\} & \textit{if } b = c \textit{ and } [a,d,b], \\ [a,b] & \textit{if } b = d \textit{ and } [a,b,c], \\ [c,b] & \textit{if } b = d \textit{ and } [a,c,b], \\ [a,b] & \textit{if } a = c \textit{ and } b = d, \\ \{a\} \sqcup \{b\} & \textit{if } a = d \textit{ and } b = c. \end{cases}$$

In a similar way, we can also make a table of intersection of three intervals $[a,b]$, $[c,d]$, $[e,f]$. In that case, the maximum possible number of components is 3. For exmaple, if $[a,f,e,d,c,b]$ holds, then the intersections becomes three disjoint intervals $[a,f]$, $[e,d]$, and $[c,b]$. Three is indeed the maximum number, because there are only 6 possible endpoints and to have disjoint intervals each interval should have at least two distinct points as its endpoint. This trend continues when the number of intervals increases.

Taking the procedure of computing the intersection of two cyclic intervals as granted, we obtain the following algorithm for checking for the existence of the indeterminacy locus in $I \times J$.

Algorithm 12.19 (Check for existence of indeterminacy locus)**.**

1. Compute $L := A^T RB - B^T RA$.
2. If one of the following conditions holds, then we conclude that the indeterminacy locus is empty, so return immediately.
   - $L = 0$, or
   - $\det L > 0$, or
   - $\det L = \det A = \det B = 0$.
3. For any other cases, any $[y_0] \in \mathbb{RP}^1$ such that $L(y_0, y_0) = 0$ is associated to a point in the indeterminacy locus.
   (a) If A and B are both of rank 1, let y_1, y_2 be generators of their kernels, respectively, then $L(y_1, y_1) = L(y_2, y_2) = 0$ holds. In this case, y_1 and y_2 should be linearly independent because otherwise $\det L = \det A = \det B = 0$. Then indeterminacy locus is consisting of exactly two points, and their y-coordinates are precisely y_1, y_2. Hence, there is an indeterminacy locus in $I \times J$ if and only if either of $([RBy_1], [y_1]) \in I \times J$ or $([RAy_2], [y_2]) \in I \times J$ holds.

(b) Otherwise, at least one of A, B is invertible. If A is invertible, then let $K \coloneqq T_{A^{-1}R}[I] \cap J$, and otherwise, let $K \coloneqq T_{B^{-1}R}[I] \cap J$.

(c) If $K = \mathbb{RP}^1$, then we conclude there is an indeterminacy locus in $I \times J$.

(d) If $K = \emptyset$, then we conclude there is no indeterminacy locus in $I \times J$.

(e) If K is a singleton set, then we conclude there is an indeterminacy locus in $I \times J$ if and only if the bilinear form $y \mapsto L(y, y)$ evaluated at the unique element of K is zero.

(f) If $K = [[u], [v]]$ for some $u, v \in \mathbb{R}^2 \setminus \{0\}$, then compute $L(u, u)L(v, v)$ and $L(u, u)L(u, v)$. Then we conclude there is an indeterminacy locus in $I \times J$ if and only if $L(u, u)L(v, v) \leq 0$ and $L(u, u)L(u, v) > 0$ hold.

(g) If K is a union of two disjoint cyclic intervals, then we check both of the components as above.

Note that this indeterminacy locus check needs to be done only once, because once we conclude that the input is not in the indeterminacy locus for the mapping $F\colon (x, y) \mapsto (\langle x, Ay\rangle : \langle x, By\rangle)$, then it stays to be outside of the indeterminacy locus of the mapping

$$F'\colon (x, y) \mapsto (\langle x, By\rangle : \langle x, (A - kB)\, y\rangle)$$

as well for any $k \in \mathbb{R}$. Indeed, clearly $\operatorname{dom}(F) = \operatorname{dom}(F')$. Furthermore, for any $(x, y) \in \operatorname{dom}(F)$, we have $F'(x, y) = T_C \circ F(x, y)$ where

$$C \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & -k \end{pmatrix}.$$

Since T_C is a homeomorphism from $\mathbb{RP}^1$ onto $\mathbb{RP}^1$, it easily follows that $T_C \circ \overline{F}$ is the unique maximal continuous extension of F'. Therefore, $\operatorname{dom}(\overline{F}) = \operatorname{dom}(\overline{F}')$ holds.

Now, let us see how to compute the image $F[I \times J]$ for given cyclic interval estimates I, J of x, y, respectively. Note that $F[I \times J]$ is a cyclic closed interval as we assume I, J are closed in this section. Hence, finding $F[I \times J]$ is to figure out the endpoints of $F[I \times J]$, or showing $F[I \times J] = \mathbb{RP}^1$ or a singleton set.

The main idea is that because of the "basically monotone" behavior of F (quoting Bill Gosper), the image $F[I \times J]$ should be the union of images of F on each of the four edges of the rectangle $I \times J$. Then, we observe that the restrictions of F onto those edges are linear fractional mappings, which we already know how to compute the image of an interval. The proposition below allows us to compute the image $F\,[I \times J]$ by separately computing the images of four edges of $I \times J$ using Proposition 12.3 and then joining them.

Proposition 12.20.
Let $[u_1, u_2], [v_1, v_2]$ be cyclic closed intervals in $\mathbb{RP}^1$ such that $[u_1, u_2] \times [v_1, v_2] \subseteq \operatorname{dom}(F)$. Then

$$\begin{aligned} F\,[[u_1, u_2] \times [v_1, v_2]] = F\,[[u_1, u_2] \times \{v_1\}] &\cup F\,[\{u_2\} \times [v_1, v_2]] \\ \cup F\,[[u_1, u_2] \times \{v_2\}] &\cup F\,[\{u_1\} \times [v_1, v_2]]\,. \end{aligned}$$

Proof. Clearly, the right-hand side is contained in the left-hand side. To show the other direction, note that we can assume without loss of generality that $[u_1, u_2] = [v_1, v_2] = [0, \infty]$, by performing an appropriate change of basis on A and B. Then any $(x, y) \in [u_1, u_2] \times [v_1, v_2]$ can be uniquely written as $([1 - s\colon s], [1 - t\colon t])$ for $(s, t) \in [0, 1] \times [0, 1]$.

Now, consider the function $\tilde{F}\colon [0,1] \times [0,1] \to \mathbb{R}^2$ given as

$$\tilde{F}\colon (s,t) \mapsto \left(\begin{pmatrix} 1-s & s \end{pmatrix} A \begin{pmatrix} 1-t \\ t \end{pmatrix}, \begin{pmatrix} 1-s & s \end{pmatrix} B \begin{pmatrix} 1-t \\ t \end{pmatrix} \right)$$

so that $F([1-s\colon s],[1-t\colon t]) = \left[\tilde{F}(s,t)\right]$. Clearly, we have

$$\tilde{F}(s,t) = (1-s)(1-t)\tilde{F}(0,0) + (1-s)t\tilde{F}(0,1) + s(1-t)\tilde{F}(1,0) + st\tilde{F}(1,1)$$

for all $(s,t) \in [0,1]\times[0,1]$, thus $F\left[[u_1,u_2]\times[v_1,v_2]\right]$ is the image under the natural projection $[\,\cdot\,]\colon \mathbb{R}^2 \setminus \{0\} \to \mathbb{RP}^1$ of the convex hull K of four points $\tilde{F}(0,0)$, $\tilde{F}(0,1)$, $\tilde{F}(1,0)$, and $\tilde{F}(1,1)$ in $\mathbb{R}^2$. Then it is enough to show that any straight line passing through a point in K should intersect with one of the line segments $[\tilde{F}(0,0),\tilde{F}(1,0)]$, $[\tilde{F}(1,0),\tilde{F}(1,1)]$, $[\tilde{F}(1,1),\tilde{F}(0,1)]$, and $[\tilde{F}(0,1),\tilde{F}(0,0)]$. This is proved in the lemma below. □

Lemma 12.21.
Let $\{x_i\}_{i\in\mathbb{Z}/n} \subseteq \mathbb{R}^2$ be any finite collection of points and let y be any point in the convex hull of $\{x_i\}_{i\in\mathbb{Z}/n}$. Then any straight line passing through y must intersect with at least one of the line segments in $\{[x_i,x_{i+1}]\}_{i\in\mathbb{Z}/n}$.

Proof. Without loss of generality, we may let $y = 0$ and the given straight line is the horizontal axis. Define

$$\begin{aligned} P &\coloneqq \{x_i\colon i\in\mathbb{Z}/n,\ x_{i2} > 0\}, \\ N &\coloneqq \{x_i\colon i\in\mathbb{Z}/n,\ x_{i2} < 0\}, \\ Z &\coloneqq \{x_i\colon i\in\mathbb{Z}/n,\ x_{i2} = 0\}, \end{aligned}$$

where x_{i2} is the vertical coordinate of x_i. Clearly, we cannot have $P = Z = \emptyset$ because in that case any point in the convex hull of $\{x_i\}_{i\in\mathbb{Z}/n}$ should have strictly negative vertical coordinate, contradicting to $y = 0$. In the same reason, we cannot have $N = Z = \emptyset$.

Now, if $Z \neq \emptyset$, then the horizontal axis must pass through any line segment $[x_i,x_{i+1}]$ such that either $x_i \in Z$ or $x_{i+1} \in Z$, so we may assume $Z = \emptyset$, so that P and N are both nonempty. Hence, there exists i such that $x_i \in P$ and $x_{i+1} \in N$. Then the horizontal line must pass through the line segment $[x_i,x_{i+1}]$. □

We could use the following general result for computing this union:

Proposition 12.22 (Union of cyclic intervals)**.**
Let $[a,b]$, $[c,d]$ be cyclic closed intervals in a cyclically ordered set X with $a \neq b$, $c \neq d$.

Then

$$[a,b]\cup[c,d]=\begin{cases}[a,b]\sqcup[c,d] & \textit{if } [a,b,c,d],\\ [c,d] & \textit{if } [a,b,d,c],\\ [a,d] & \textit{if } [a,c,b,d],\\ [a,b] & \textit{if } [a,c,d,b],\\ [c,b] & \textit{if } [a,d,b,c],\\ X & \textit{if } [a,d,c,b],\\ [a,d] & \textit{if } a=c \textit{ and } [a,b,d],\\ [a,b] & \textit{if } a=c \textit{ and } [a,d,b],\\ [c,b] & \textit{if } a=d \textit{ and } [a,b,c],\\ X & \textit{if } a=d \textit{ and } [a,c,b],\\ [a,d] & \textit{if } b=c \textit{ and } [a,b,d],\\ X & \textit{if } b=c \textit{ and } [a,d,b],\\ [c,b] & \textit{if } b=d \textit{ and } [a,b,c],\\ [a,b] & \textit{if } b=d \textit{ and } [a,c,b],\\ [a,b] & \textit{if } a=c \textit{ and } b=d,\\ X & \textit{if } a=d \textit{ and } b=c.\end{cases}$$

Or, we could simplify this procedure a little bit as explained below. First, we compute the four "endpoints" $F(u_1,v_1)$, $F(u_2,v_1)$, $F(u_2,v_2)$, and $F(u_1,v_2)$. Next, by computing the determinant of the corresponding linear fractional transforms, we record whether the current line segment is going right or left from the last endpoint. For example, if the determinants of four linear fractional transforms are all positive, then the first two ($F(\,\cdot\,,v_1)$ and $F(u_2,\,\cdot\,)$) appends the interval to right of their respective first endpoints ($F(u_1,v_1)$ and $F(u_2,v_1)$), and the other two ($F(\,\cdot\,,v_2)$ and $F(u_1,\,\cdot\,)$) moves the cursor back to the starting point $F(u_1,v_1)$. In this specific scenario, there are two possibilities. The first case is that $F(u_1,v_1)$ is the left endpoint and $F(u_2,v_2)$ (the last endpoint of the second segment) is the right endpoint, which is natrual because $F(u_2,u_2)$ is the point where it "turns back". Note that this is not the only possibility, because the interval can "cross" the beginning point $F(u_1,v_1)$ at some point while proceeding to $F(u_2,v_2)$. When this happens, the union should be the whole $\mathbb{RP}^1$.

To make this idea more precise, let us write the sign of four determinants in a row like $(+,+,+,+)$ or $(+,-,-,+)$, where $+$ means that the endpoint is moving right and $-$ means it is moving left, when we are travelling the boundary of the rectangle $[u_1,u_2]\times[v_1,v_2]$ counterclockwise, starting at (u_1,v_1). Note that thus the signs at the third and the fourth positions are the opposite signs to the determinants.

There are 16 possible sign patterns. We can in fact reduce this into 8 by observing that given a sign pattern $\sigma := (\sigma_1,\sigma_2,\sigma_3,\sigma_4)$, the completely opposite sign pattern to σ can be treated in the same way as σ while we just switch the left and the right end points in the last step. We can further reduce these by observing the cyclic symmetry; for example, $(+,+,-,-)$ and $(+,-,-,+)$ can be treated in the same way by merely changing the starting point, and $(+,-,-,-)$ and $(+,+,+,-)$ can be treated in the same way by first rotating $(+,-,-,-)$ into $(-,-,-,+)$ and then changing the sign. We thus analyze resulting 4 possible patterns below. For simplicity, we did not count the case when a segment is degenerate (that is, the image of the segment is a single point).

1. $\sigma=(+,+,+,+)$.

In this case, $F[I \times J]$ should be the entire $\mathbb{RP}^1$.

2. $\sigma = (+,+,+,-)$.

   In this case, $F[I \times J]$ should be either $[F(u_1, v_1), F(u_1, v_2)]$ or $\mathbb{RP}^1$. The first case happens if and only if the second and the third segments never pass through $F(u_1, v_1)$.

3. $\sigma = (+,+,-,-)$.

   In this case, $F[I \times J]$ should be either $[F(u_1, v_1), F(u_2, v_2)]$ or $\mathbb{RP}^1$. The first case happens if and only if the second and the third segments never pass through $F(u_1, v_1)$.

4. $\sigma = (+,-,+,-)$.

   In this case, $F[I \times J]$ should be either $[F(u_1, v_1), F(u_2, v_1)]$, $[F(u_1, v_1), F(u_1, v_2)]$, $[F(u_2, v_2), F(u_2, v_1)]$, $[F(u_2, v_2), F(u_1, v_2)]$, or $\mathbb{RP}^1$:

   (a) If the second and the third segments pass through $F(u_1, v_1)$ and the third segment also passes through $F(u_2, v_1)$, then it is $[F(u_2, v_2), F(u_1, v_2)]$.

   (b) If the second and the third segments pass through $F(u_1, v_1)$ but the third segment never passes through $F(u_2, v_1)$, then it is $[F(u_2, v_2), F(u_2, v_1)]$.

   (c) If the second segment passes through $F(u_1, v_1)$ and the third segment never passes through $F(u_1, v_1)$, then it is $\mathbb{RP}^1$.

   (d) If the second segment never passes through $F(u_1, v_1)$ and the third segment passes through both $F(u_2, v_1)$ and $F(u_1, v_1)$, then it is $\mathbb{RP}^1$.

   (e) If the second segment never passes through $F(u_1, v_1)$ and the third segment passes through $F(u_2, v_1)$ but not $F(u_1, v_1)$, then it is $[F(u_1, v_1), F(u_1, v_2)]$.

   (f) If the second segment never passes through $F(u_1, v_1)$ and the third segment never passes through $F(u_2, v_1)$, then it is $[F(u_1, v_1), F(u_2, v_1)]$.

Note that the resulting union must be either a single point, a cyclic closed interval, or the whole $\mathbb{RP}^1$ since the union must be connected. For the first case, we can compute $\lfloor F(x, y) \rfloor$ by taking the floor of that single point. For the second case, provided that the cyclic interval happens to be a compact interval in $\mathbb{R}$, we write it as $[u, v]$ and see if $\lfloor u \rfloor = \lfloor v \rfloor$ holds. If that is the case, then their common value must be the value of $\lfloor F(x, y) \rfloor$. For all the other cases, we cannot reliably compute $\lfloor F(x, y) \rfloor$, so we refine the domain $I \times J$ and then retry.

Now, we have to ask if this procedure will always succeed. Since $\mathbb{R}$ is an open subset of $\mathbb{RP}^1$, if (x, y) is such that $F(x, y) \in \mathbb{R}$, then any sufficiently fine estimate $I \times J$ should satisfy $F[I \times J] \subseteq \mathbb{R}$ as well by continuity of F, provided that $I \times J \subseteq \mathrm{dom}(F)$. However, like in the unary case, ensuring $\lfloor u \rfloor = \lfloor v \rfloor$ for the endpoints u, v turns out to be quite tricky when $F(x, y)$ happens to be an integer. It sounds likely that, for such a case, unless either I or J converges down to a single point, we will never be able to compute $\lfloor F(x, y) \rfloor$ and will trap inside an infinite loop. In a similar vein, if $F(x, y) = \infty$ were true, then it would be likely that we will never be able to tell that $F(x, y) = \infty$ is really the case.

We want to characterize exactly when this can happen. As a preparation, we first state some definition.

Definition 12.23 (Supporting vectors)**.**
Let $(I_n \times J_n)_{n=1}^{\infty}$ be a decreasing sequence of sets such that I_n, J_n's are cyclic closed intervals

satisfying $(x, y) \in I_n \times J_n$ for all n and

$$\{(x, y)\} = \bigcap_{n=1}^{\infty} I_n \times J_n.^{12}$$

Then we call a nonzero tangent vector $X \in T_{(x,y)}(\mathbb{RP}^1 \times \mathbb{RP}^1)$ a *supporting vector* if for all n, there exists a smooth curve $\gamma\colon (-\epsilon, \epsilon) \to \mathbb{RP}^1 \times \mathbb{RP}^1$ such that $\gamma(0) = (x, y)$, $\gamma'(0) = X$, and $\gamma\,[[0, \epsilon)] \subseteq I_n \times J_n$.

In other words, we call X a supporting vector if every $I_n \times J_n$ contains a small line segment starting from (x, y) into the direction X. Due to the simple geometry of $I_n \times J_n$, there are only a few possibilities:

1. If $I_n \times J_n$ shrinks down to a single point (x, y) in a finite step, then there is no supporting vector.
2. If I_n shrinks down to a single point in a finite step while J_n stays to be nontrivial, then there are three possibilities:
   (a) y eventually becomes the "bottom" endpoint of J_n. In this case, supporting vectors are the vectors along the positive vertical direction.
   (b) y eventually becomes the "top" endpoint of J_n. In this case, supporting vectors are the vectors along the negative vertical direction.
   (c) y stays to be inside the interior of J_n. In this case, supporting vectors are the vectors along the vertical direction.
3. If J_n shrinkgs down to a single point in a finite step while I_n stays to be nontrivial, then there are three possibilities:
   (a) x eventually becomes the "left" endpoint of I_n. In this case, supporting vectors are the vectors along the positive horizontal direction.
   (b) x eventually becomes the "right" endpoint of I_n. In this case, supporting vectors are the vectors along the negative horizontal direction.
   (c) x stays to be inside the interior of I_n. In this case, supporting vectors are the vectors along the horizontal direction.
4. If both I_n and J_n stays to be nontrivial, there are nine possibilities:
   (a) (x, y) eventually becomes the "left-bottom" corner of $I_n \times J_n$. In this case, supporting vectors are the vectors in the first quadrant.
   (b) (x, y) eventually becomes the "right-bottom" corner of $I_n \times J_n$. In this case, supporting vectors are the vectors in the second quadrant.
   (c) (x, y) eventually becomes the "right-top" corner of $I_n \times J_n$. In this case, supporting vectors are the vectors in the third quadrant.
   (d) (x, y) eventually becomes the "left-top" corner of $I_n \times J_n$. In this case, supporting vectors are the vectors in the fourth quadrant.

[12] Note that this condition is equivalent to that $I_n \times J_n$ is eventually contained in any neighborhood of (x, y): if not, then for each n we can find $(x_n, y_n) \in (I_n \times J_n) \setminus U$ for some fixed neighborhood U of (x, y). By compactness, $((x_n, y_n))_{n=1}^{\infty}$ admits a convergent subsequence, but by the construction this subsequence must be eventually contained in every $I_n \times J_n$, thus is convergent to x, which is a contradiction.

(e) (x, y) eventually stays in the interior of the "bottom" edge of $I_n \times J_n$. In this case, supporting vectors are the vectors with positive vertical coordinate.

(f) (x, y) eventually stays in the interior of the "right" edge of $I_n \times J_n$. In this case, supporting vectors are the vectors with negative horizontal coordinate.

(g) (x, y) eventually stays in the interior of the "top" edge of $I_n \times J_n$. In this case, supporting vectors are the vectors with negative vertical coordinate.

(h) (x, y) eventually stays in the interior of the "left" edge of $I_n \times J_n$. In this case, supporting vectors are the vectors with positive horizontal coordinate.

(i) (x, y) stays in the interior of $I_n \times J_n$. In this case, every nonzero tangent vector is supporting.

Proposition 12.24.
Suppose that $(x, y) \in \mathrm{dom}(F)$ and $F(x, y) \in \mathbb{Z} \cup \{\infty\}$. Let $L_1, L_2 \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be symmetric bilinear forms given as

$$L_1 \colon (u, v) \mapsto \left\langle u, (A^T RB - B^T RA)v \right\rangle,$$
$$L_2 \colon (u, v) \mapsto \left\langle u, (ARB^T - BRA^T)v \right\rangle.$$

Let $(I_n \times J_n)_{n=1}^{\infty}$ be a decreasing sequence of sets such that I_n's, J_n's are cyclic closed intervals satisfying $(x, y) \in I_n \times J_n \subseteq \mathrm{dom}(F)$ for all n and

$$\{(x, y)\} = \bigcap_{n=1}^{\infty} (I_n \times J_n).$$

Let $(u, v) \in (\mathbb{R}^2 \setminus \{0\})^2$ be any vector such that $x = [u]$ and $y = [v]$ and $z \in \mathbb{RP}^1 \setminus \{F(x, y)\}$. Then

$$F[I_n \times J_n] \subseteq [F(x, y), z) \tag{20}$$

holds for large enough n if and only if one of the followings hold:

1. *$I_n \times J_n = \{x\} \times \{y\}$ for large enough n, or*

2. *$I_n = \{x\}$ for large enough n and $L_2(u, u) = 0$, or*

3. *$I_n = \{x\}$, y is one of the endpoints of J_n for large enough n and $L_2(u, u) > 0$ if y is the bottom endpoint and $L_2(u, u) < 0$ if y is the top endpoint, or*

4. *$J_n = \{y\}$ for large enough n and $L_1(v, v) = 0$, or*

5. *$J_n = \{y\}$, x is one of the endpoints of I_n for large enough n and $L_1(v, v) > 0$ if x is the left endpoint and $L_1(v, v) < 0$ if x is the right endpoint, or*

6. *(x, y) is one of the corners of $I_n \times J_n$ for large enough n, $\sigma_h L_1(v, v) \geq 0$ and $\sigma_v L_2(u, u) \geq 0$ hold with at least one of the inequalities being strict, where*

$$(\sigma_h, \sigma_v) = \begin{cases} (+1, +1) & \text{if } (x, y) \text{ is the left-bottom corner,} \\ (-1, +1) & \text{if } (x, y) \text{ is the right-bottom corner,} \\ (-1, -1) & \text{if } (x, y) \text{ is the right-top corner,} \\ (+1, -1) & \text{if } (x, y) \text{ is the left-top corner, or} \end{cases}$$

7. *(x, y) is one of the corners of $I_n \times J_n$ for large enough n, $L_1(v,v) = L_2(u,u) = 0$ and $\sigma\left(\langle Ru, ARv\rangle \langle u, Bv\rangle - \langle Ru, BRv\rangle \langle u, Av\rangle\right) < 0$ where*

$$\sigma = \begin{cases} +1 & \textit{if } (x,y) \textit{ is the left-bottom or the right-top corner,} \\ -1 & \textit{if } (x,y) \textit{ is the right-bottom or the left-top corner, or} \end{cases}$$

8. *(x, y) is on one of the edges of $I_n \times J_n$ for large enough n and*

$$\begin{cases} L_1(v,v) = 0, L_2(u,u) > 0 & \textit{if } (x,y) \textit{ is on the bottom edge,} \\ L_1(v,v) < 0, L_2(u,u) = 0 & \textit{if } (x,y) \textit{ is on the right edge,} \\ L_1(v,v) = 0, L_2(u,u) < 0 & \textit{if } (x,y) \textit{ is on the top edge,} \\ L_1(v,v) > 0, L_2(u,u) = 0 & \textit{if } (x,y) \textit{ is on the left edge} \end{cases}$$

holds, or

9. *A, B are linearly dependent.*[13]

Similarly,

$$F[I_n \times J_n] = \{F(x,y)\} \tag{21}$$

holds for large enough n if and only if one of 1, 2, 4, and 9 holds.

Proof. Consider the bilinear mapping

$$\begin{aligned} \tilde{F} \colon \mathbb{R}^2 \times \mathbb{R}^2 &\to \mathbb{R}^2 \\ (u,v) &\mapsto \left(\langle u, Av\rangle, \langle u, Bv\rangle\right). \end{aligned}$$

Let $\pi\colon (\mathbb{R}^2 \setminus \{0\}) \to \mathbb{RP}^1$ be the canonical projection. Take any $(u,v) \in (\mathbb{R}^2 \setminus \{0\})^2$ with $(\pi(u), \pi(v)) = (x,y)$, and take any unit vector $(a,b) \in \mathbb{R}^2 \times \mathbb{R}^2$ such that $d\pi_{(u,v)}(a,b)$ is a supporting vector. Note that for any $w \in \mathbb{R}^2 \setminus \{0\}$, we have $\ker d\pi_w = \langle w\rangle$, so we can identify the tangent space $T_w\mathbb{RP}^1$ with $\langle Rw\rangle \subseteq \mathbb{R}^2$. Note that in this identification, the direction corresponding to the positive direction on $\mathbb{R}$, with the usual embedding $r \mapsto [r\colon 1]$ of $\mathbb{R}$ into $\mathbb{RP}^1$, is the direction along $-Rw$. Therefore, the condition

$$F[I_n \times J_n] \subseteq [F(x,y), z)$$

for large enough n is equivalent to

$$\tau(t) := \left\langle \tilde{F}\left((u,v) + t(a,b)\right), R\tilde{F}(u,v)\right\rangle \leq 0$$

for all $t \geq 0$ small enough and for all (a,b). Note that τ is a quadratic function of t, whose constant coefficient is zero and the linear coefficient is equal to

$$\begin{aligned} &\left\langle \left(\langle u, Ab\rangle + \langle a, Av\rangle, \langle u, Bb\rangle + \langle a, Bv\rangle\right), RF(u,v)\right\rangle \\ &\qquad = \det\left(\begin{pmatrix} a^T \\ u^T \end{pmatrix} \begin{pmatrix} Av & Bv \end{pmatrix}\right) + \det\left(\begin{pmatrix} b^T \\ v \end{pmatrix} \begin{pmatrix} A^T u & B^T u \end{pmatrix}\right), \end{aligned}$$

which by applying the identity $\det\begin{pmatrix} u_1 & u_2 \end{pmatrix} = \langle u_1, Ru_2\rangle$ can be shown to be equal to

$$\langle a, Ru\rangle \left\langle v, A^T RBv\right\rangle + \langle b, Rv\rangle \left\langle u, ARB^T u\right\rangle = \frac{1}{2}\left(\langle a, Ru\rangle L_1(v,v) + \langle b, Rv\rangle L_2(u,u)\right).$$

[13] This is in fact never the case as we already assumed $\overline{F} = F$.

Finally, the leading coefficient is equal to

$$\left\langle \left(\langle a, Ab \rangle, \langle a, Bb \rangle \right), R\tilde{F}(u,v) \right\rangle = \langle a, Ab \rangle \langle u, Bv \rangle - \langle a, Bb \rangle \langle u, Av \rangle .$$

Hence, $F[I_n \times J_n] \subseteq F[(x,y), z)$ holds for all large enough n if and only if for any (a,b) such that $d\pi_{(u,v)}(a,b)$ is a supporting vector, either

$$\langle a, Ru \rangle L_1(v,v) + \langle b, Rv \rangle L_2(u,u) < 0$$

or

$$\begin{aligned} \langle a, Ru \rangle L_1(v,v) + \langle b, Rv \rangle L_2(u,u) = 0 \quad \text{and} \\ \langle a, Ab \rangle \langle u, Bv \rangle - \langle a, Bb \rangle \langle u, Av \rangle \leq 0 \end{aligned}$$

holds.

Note that the "horizontal coordinate" of the supporting vector $d\pi_{(u,v)}(a,b)$ is $-\langle a, Ru \rangle < 0$ while the "vertical coordinate" is $-\langle b, Rv \rangle$, according to our identification of $T_u\mathbb{RP}^1$ with $\langle Ru \rangle$. Then considering how the set of supporting vectors look like, it can be easily seen that $L_1(v,v) \geq 0$ should hold if there is a supporting vector of positive vertical coordinate, $L_1(v,v) \leq 0$ should hold if there is a supporting vector of negative vertical coordinate, and similarly $L_2(u,u) \geq 0$ or $L_2(u,u) \leq 0$ should hold according to existence of a supporting vector with positive or negative horizontal coordinate.

Similarly, $F[I_n \times J_n] = \{F(x,y)\}$ holds for all large enough n if and only if for any (a,b) such that $d\pi_{(u,v)}(a,b)$ is a supporting vector,

$$\langle a, Ru \rangle L_1(v,v) + \langle b, Rv \rangle L_2(u,u) = \langle a, Ab \rangle \langle u, Bv \rangle - \langle a, Bb \rangle \langle u, Av \rangle = 0$$

holds. In particular, we must have $L_1(v,v) = 0$ if there is a supporting vector with a nonzero horizontal coordinate while we must have $L_2(u,u) = 0$ if there is one with a nonzero vertical coordinate. Based on this, we consider each possible location of (x,y) in $I_n \times J_n$.

- When $I_n \times J_n = \{x\} \times \{y\}$ for large enough n.

  In this case, we always have $F[I_n \times J_n] = \{F(x,y)\}$, so there is nothing to prove.

- When $I_n = \{x\}$ for large enough n, but J_n is not a single point for all n.

  Note that $L_2(u,u) = 0$ is a necessary condition for having (21). To see sufficiency, note that $\tau(t)$ has no quadratic term in t if $a = 0$ or $b = 0$, so the linear term being zero is same as it being identically zero. Hence, whenever J_n is small enough, $F[I_n \times J_n]$ should be a singleton set.

  On the other hand, (20) certainly implies $L_2(u,u) = 0$ if y is in the interior of J_n, $L_2(u,u) \geq 0$ if it is the bottom endpoint, or $L_2(u,u) \leq 0$ if it is the top endpoint. As noted above, $L_2(u,u) = 0$ is equivalent to $\tau \equiv 0$, so $L_2(u,u) = 0$ is certainly a sufficient condition for all cases. When $L_2(u,u) \neq 0$, the sign of $\left\langle \tilde{F}\left((u,v) + t(0,b)\right), R\tilde{F}(u,v) \right\rangle$ for small enough t is determined by the sign of its first derivative at $t = 0$, so $L_2(u,u) > 0$ or $L_2(u,u) < 0$ should be a sufficient condition as well, respectively for when y is the bottom or the top endpoint of J_n.

- When $J_n = \{y\}$ for large enough n, but I_n is not a single point for all n.

  Exactly the same analysis applies, so we omit this case.

- When neither I_n nor J_n is a singleton set for all n.

  We first claim that in this case, $F|_{I_n \times J_n}$ being constant is equivalent to A, B being linearly dependent. This in particular shows that when neither I_n nor J_n is a singleton set, then (21) is equivalent to *9*. Since we already assumed $I_n \times J_n \subseteq \operatorname{dom}(F)$, clearly A, B being linearly dependent implies that $F|_{I_n \times J_n}$ is constant.

  For the converse, note that since we can apply the same analysis for any point in the interior of $I_n \times J_n$, not necessarily just at (x, y), we must have $L_1(u_1, u_1) = L_2(v_1, v_1) = 0$ for all (u_1, v_1) with $\pi(u_1, v_1) \in I_n \times J_n$. Since L_1, L_2 are bilinear forms, this implies $L_1 \equiv L_2 \equiv 0$, in other words,

  $$B^T RA = ARB^T = 0.$$

  Multiplying R on left or right and then applying the identity $RB^T R = -\operatorname{adj} B = -(\operatorname{tr} B)I + B$ then shows

  $$(\operatorname{tr} B)A = AB = BA,$$

  and by symmetry, the above must be equal to $(\operatorname{tr} A)B$ as well. Therefore, whenever at least one of $\operatorname{tr} A$ and $\operatorname{tr} B$ is not zero, then A, B should be linearly dependent, so suppose that $\operatorname{tr} A = \operatorname{tr} B = 0$. If A is invertible, then

  $$B = A^{-1}AB = (\operatorname{tr} B)A^{-1}A = 0,$$

  and similarly if B is invertible then we conclude $A = 0$. Hence, we may assume that A, B are both singular and not zero, i.e., they are both of rank 1. Hence, there exist nonzero vectors $u_1, u_2, v_1, v_2 \in \mathbb{R}^2$ such that $A = u_1 v_1^T$ and $B = u_2 v_2^T$. Then the condition $\operatorname{tr} A = \operatorname{tr} B = 0$ enforces $\langle u_1, v_1 \rangle = \langle u_2, v_2 \rangle = 0$, so we can write without loss of generality that $A = \pm u_1 u_1^T R$ and $B = \pm u_2 u_2^T R$. Then, the condition $AB = 0$ gives

  $$\langle u_1, Ru_2 \rangle (u_1 u_2^T R) = 0,$$

  thus u_1 and u_2 should be parallel, so A and B are linearly dependent as claimed.

  Now to analyze (20), we further divide the cases.

  - When (x, y) is one of the corners of $I_n \times J_n$ for large enough n.

    Clearly, $\sigma_h L_1(v, v) \geq 0$ and $\sigma_v L_2(u, u) \geq 0$ together are necessary conditions for (20). Then it is enough to show three claims: (1) if $L_1(v, v) = L_2(u, u) = 0$, then (20) is equivalent to

    $$\sigma \left(\langle Ru, ARv \rangle \langle u, Bv \rangle - \langle Ru, BRv \rangle \langle u, Av \rangle \right) \leq 0, \tag{22}$$

    (2) if the above is zero in addition to $L_1(v, v) = L_2(u, u) = 0$, then A, B should be linearly dependent, and (3) having a strict inequality for one of $\sigma_h L_1(v, v) \geq 0$ and $\sigma_v L_2(u, u) \geq 0$ is enough to conclude (20).

    For the first claim, suppose $L_1(v, v) = L_2(u, u) = 0$. Then (20) is equivalent to having

    $$\langle a, Ab \rangle \langle u, Bv \rangle - \langle a, Bb \rangle \langle u, Av \rangle \leq 0$$

    for all (a, b) such that $d\pi_{(u,v)}(a, b)$ is a supporting vector. Recall from Proposition 12.11 that $L_1(v, v) = 0$ is equivalent to that Av and Bv are parallel. In the same way, $L_2(u, u) = 0$ is equivalent to that $A^T u$ and $B^T u$ are parallel. Then, decomposing a into u-component and Ru-component and b into v-component and Rv-component shows that the above equality is equivalent to

$$\langle a, Ru\rangle \langle b, Rv\rangle (\langle Ru, ARv\rangle \langle u, Bv\rangle - \langle Ru, BRv\rangle \langle u, Av\rangle) \leq 0.$$

Then it is easy to see that σ is precisely the sign of $\langle a, Ru\rangle \langle b, Rv\rangle$, so the first claim is proved.

For the second claim, note that having all three being zero means that the function τ is identically zero. Therefore, $F|_{I_n\times J_n}$ should be a constant in a neighborhood of (x, y), which shows that A, B are linearly dependent.

The third claim directly follows from that the sign of $\langle a, Ru\rangle$ and $\langle b, Rv\rangle$ is the negative of σ_h and σ_v, respectively.

- When (x, y) is on one of the edges of $I_n \times J_n$ for large enough n.

  In this case, sufficiency of

$$\begin{cases} L_1(v, v) = 0, L_2(u, u) > 0 & \text{if } (x, y) \text{ is on the bottom edge,}\\ L_1(v, v) < 0, L_2(u, u) = 0 & \text{if } (x, y) \text{ is on the right edge,}\\ L_1(v, v) = 0, L_2(u, u) < 0 & \text{if } (x, y) \text{ is on the top edge,}\\ L_1(v, v) > 0, L_2(u, u) = 0 & \text{if } (x, y) \text{ is on the left edge} \end{cases}$$

  for having (20) is clear. Then it is enough to show that $L_1(v, v) = L_2(u, u) = 0$ can happen only when A, B are linearly dependent. This follows from that (22) should hold for both positive and negative σ, because $\langle a, Ru\rangle \langle b, Rv\rangle$ can have two different signs. Therefore, τ is identically zero, thus $F|_{I_n\times J_n}$ is a constant in a neighborhood of (x, y), which can happen only when A, B to be linearly dependent.

- When (x, y) stays inside the interior of $I_n \times J_n$.

  In this case, (20) implies $L_1(v, v,) = L_2(u, u) = 0$. Now, note that for any point u' and v' in small neighborhoods around u and v, respectively, we must have $L_2(u', u') = 0$ and $L_2(v', v') = 0$ as well because (20) is still true when we replace (x, y) by $([u'], y)$ or by $(x, [v'])$, since $L_1(v, v) = 0$ implies that $F(\,\cdot\,, y)$ is constant around x, and similarly $L_2(u, u) = 0$ implies that $F(x, \,\cdot\,)$ is constant around y. Therefore, by the claim above, F is constant on a small neighborhood of (u, v), thus A, B must be linearly dependent, which already have been shown to be equivalent to (21).

□

First, note that when we repeatedly apply the transform $F \mapsto F'$ given in (19), $F(x, y)$ eventually becomes an integer or ∞ if and only if $F(x, y) \in \mathbb{Q}\cup\{\infty\}$. (Actually, it can become ∞ only after $F(x, y) \in \mathbb{Z}$ holds, unless $F(x, y) = \infty$ were the case from the beginning.) Hence, the problem may exist only when $F(x, y) \in \mathbb{Q} \cup \{\infty\}$.

Note that if we assume that the interval estimates I_n, J_n are always with rational endpoints and they converge down to single points if and only if x, y are rational, respectively, then this infinite loop problem can only happen when x, y are both irrational numbers, in which case none of the items from the proposition above applies since (x, y) is always in the interior of $I_n \times J_n$. Indeed, let $x = [u]$, $y = [v]$ for $u, v \in \mathbb{R}^2 \setminus \{0\}$ and suppose x is irrational while y is rational (i.e., $y \in \mathbb{Q} \cup \{\infty\}$), then

$$F(x, y) = [\langle u, Av\rangle : \langle u, Bv\rangle] \in \mathbb{Q} \cup \{\infty\}$$

implies that there exists $[p : q] \in \mathbb{QP}^1$ such that

$$\langle u, (qA - pB)v \rangle = 0,$$

and since $[u]$ is irrational and $[(qA - pB)v]$ is rational, we must have

$$qAv = pBv.$$

Therefore, as noted in Theorem 12.7, $L_1(v, v) = 0$ holds, thus the condition *4* from Proposition 12.24 is satisfied.

Similarly, if $F(x, y) \in \mathbb{Q} \cup \{\infty\}$ then $\{xy, x, y, 1\}$ must be $\mathbb{Q}$-linearly dependent. Indeed, from the equation

$$\langle u, (qA - pB)v \rangle = 0,$$

assuming $\mathbb{Q}$-linear independence of $\{xy, x, y, 1\}$ (which by definition means $\mathbb{Q}$-linear independence of $\{u_1v_1, u_1v_2, u_2v_1, u_2v_2\}$) shows $qA - pB = 0$, contradicting to $\overline{F} = F$, which implies that A, B are linearly independent.

Finally, note that if (x, y) stays inside the interior, then (20) holds if and only if A, B are linearly dependent. In other words, if A, B are linearly independent and x, y are both irrational, then having $F(x, y) \in \mathbb{Q} \cup \{\infty\}$ guarantees the procedure to fall into an infinite loop.

In summary, assuming $A, B \in \mathbb{Z}^{2\times 2}$ and that I, J are always with rational endpoints and they converge down to single points in a finite step if and only if x, y are rational, respectively, then the described procedure either succeeds in computing the continued fraction expansion of $\overline{F}(x, y)$ indefinitely, or rejects the input (x, y) for successfully identifying $(x, y) \notin \mathrm{dom}(\overline{F})$, if and only if one of the following conditions hold:

1. A, B are linearly dependent, or
2. F is independent of x and y is rational, or
3. F is independent of y and x is rational, or
4. $\overline{F} = F$ and at least one of x, y is rational, or
5. $(x, y) \in \mathrm{dom}(\overline{F})$ and $\overline{F}(x, y) \notin \mathbb{Q} \cup \{\infty\}$.

Also, assuming that A, B are $\mathbb{Q}$-linearly independent, the last condition automatically holds whenever $\{xy, x, y, 1\}$ is $\mathbb{Q}$-linearly independent.

Unfortunately, generically speaking, if x, y are both irrational but $\overline{F}(x, y)$ is rational, then the procedure is almost guaranteed to trap into an infinite loop, but as previously pointed out this is rather a fundamentally unsolvable issue.

Remark 12.25.
There are two notable differences of the explained algorithm from what Gosper originally proposed.

1. There is no mention of singularity in Gosper's original proposal. I did not really carefully think about this, but I think his original proposal will not run correctly if $(x, y) \notin \mathrm{dom}(F)$. It of course cannot run correctly if $(x, y) \notin \mathrm{dom}(\overline{F})$ because I consider "the correct behavior" is to simply reject the input, but I am not sure if it will just run fine or not if $(x, y) \in \mathrm{dom}(\overline{F}) \setminus \mathrm{dom}(F)$, though I guess it will not at least for certain examples.

2. Computation of $F[I \times J]$ is much simpler in Gosper's original proposal, because it gives up and try to refine I, J immediately whenever any of the images of four edges by F contains ∞. Since $F[I \times J]$ is the union of those four images and we cannot compute $\lfloor F(x,y) \rfloor$ anyway if $F[I \times J]$ contains ∞, this is totally fine, and it significantly simplifies the procedure of computing the union because all we need to do is to simply find the minimum and the maximum of endpoints once every interval is contained in $\mathbb{R}$. However, I considered the unconditional computation of $F[I \times J]$ might be useful in certain scenarios, for example when we want to feed $F(x,y)$ as another input to Gosper's algorithm, e.g. to compute $F(F(x,y),z)$. My viewpoint is that Gosper's algorithm is, in essence, an algorithm for finding the corresponding interval estimate from given interval estimates for the input, rather than an algorithm for computing the continued fraction expansion. The latter just immediately follows from the former as a trivial application.

13 Continued fractions for logarithms

Some of the (in fact, *the only*) motivating examples I had for computing $\lfloor nx + y \rfloor$ that involved irrational x, y was when x, y are given by some general logarithms with base 2 or 10. Hence, I devoted a section for explaining how to obtain continued fraction expansions of logarithms.

When we restrict ourselves to the case when the base and the input are both rational, there is a very simple algorithm for computing the continued fraction expansion. Let us say that a, b are positive rational numbers with $a \neq 1$, and we want to compute the continued fraction expansion of $\log_a b$. Then once we know $k := \lfloor \log_a b \rfloor$, then

$$(\log_a b - k)^{-1} = \left(\log_a (b/a^k)\right)^{-1} = \log_{(b/a^k)} a$$

is still a general logarithm with rational base and input. Therefore, it is enough to see how to compute $\lfloor \log_a b \rfloor$. For that, we can use the definition of logarithms: we find the largest integer k such that

$$a^k \leq b < a^{k+1}$$

if $a > 1$, or

$$a^k \geq b > a^{k+1}$$

if $a < 1$.

However, no matter what specific algorithm we use for finding k, this way of computing the continued fraction expansion of $\log_a b$ does not work in general, not because it is wrong, rather because the required precision of involved numbers grows exponentially. The computation very quickly becomes totally infeasible.

Instead, we compute the continued fraction expansion of $\log_a b$ using the identity $\log_a b = \frac{\ln b}{\ln a}$: once we know how to compute the continued fraction expansions of $\ln a$ and $\ln b$, we can apply Gosper's algorithm. In fact, we do not really need to compute the continued fraction expansions for $\ln a$ and $\ln b$. What we really need are the interval estimates $(I_n)_{n=1}^{\infty}$, $(J_n)_{n=1}^{\infty}$ of $\ln a$ and $\ln b$, because that is all we need to run Gosper's algorithm.

Therefore, any method of computing successively better approximations of the natural logarithm should work out. Nevertheless, when I started writing the implementation, I did not really realize this, so was trying to find a way to compute the continued fraction expansions of natural logarithms. Thus, I searched for existing literatures, and found two papers by Isao Makino and Takeshi Aoyama on computing natural logarithm using its continued fraction expansion [11][12].[14] After I realized there probably are other methods that fit my purpose better, I already had implemented the method explained in these papers and it did its job good enough, so I just moved on. Following subsections explain some details of this method I implemented.

13.1 Continued fractions for natural logarithm

As far as I am aware, there is no known formula for the regular continued fraction expansion of natural logarithms. The method I will explain in this subsection does not compute such; rather, it computes a *generalized* continued fraction that converges to the natural logarithm. However, this is fine because all we need is just a sequence of successively better interval estimates. In the end of the day, coming up with an error bound is the only thing we need to do.

We start from the following expansion of the inverse hyperbolic tangent:

$$\log\left(\frac{1+z}{1-z}\right) = 2\tanh^{-1} z = \cfrac{2z}{1 - \cfrac{z^2}{3 - \cfrac{4z^2}{5 - \cfrac{9z^2}{7 - \cdots}}}}$$

In particular, when $z = \frac{p}{q}$, we can write it as

$$2\tanh^{-1}\left(\frac{p}{q}\right) = \cfrac{2p}{q - \cfrac{p^2}{3q - \cfrac{4p^2}{5q - \cfrac{9p^2}{7q - \cdots}}}} \tag{23}$$

Recall from Section 3 that given a generalized continued fraction

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3 + \cfrac{b_4}{a_4 + \cdots}}}} \tag{24}$$

the recurrence relation for the convergents is given as

$$\begin{cases} p_{n+1} = a_{n+1}p_n + b_{n+1}p_{n-1} \\ q_{n+1} = a_{n+1}q_n + b_{n+1}q_{n-1} \end{cases}, \quad \begin{cases} (p_{-1}, p_0) = (1, a_0) \\ (q_{-1}, q_0) = (0, 1) \end{cases}$$

where $\frac{p_n}{q_n}$ is the nth convergent. Specializing this for (23), we obtain

$$\begin{cases} p_{n+1} = (2n+1)qp_n - n^2p^2p_{n-1} \\ q_{n+1} = (2n+1)qq_n - n^2p^2q_{n-1} \end{cases}, \quad \begin{cases} (p_0, p_1) = (0, 2p) \\ (q_0, q_1) = (1, q) \end{cases}$$

for $n \geq 1$.

[14] These papers are quite poorly written and especially contain a lot of typos.

At this point, it is not clear if the sequence $\left(\frac{p_n}{q_n}\right)_{n=1}^{\infty}$ even converges.[15]

Recall from Corollary 3.2 the identity

$$\frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n} = \frac{\prod_{i=1}^{n}(-a_i)}{q_{n-1}q_n},$$

so in the case of (23), this becomes

$$\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{2p^{2n+1}(n!)^2}{q_n q_{n+1}} \tag{25}$$

for $n \geq 0$. Therefore, if q_n grows sufficiently faster than $p^n n!$, then $\left(\frac{p_n}{q_n}\right)_{n=1}^{\infty}$ converges to some real number α and the convergence is monotone.

Now, recall that $\tanh^{-1}$ is not well-defined for real numbers of modulus greater than or equal to 1, so our $z = \frac{p}{q}$ should be such that $|p| < q$. Note that, because of this, if we had only the first term in the recurrence relation

$$q_{n+1} = (2n+1)qq_n - n^2p^2q_{n-1},$$

then we indeed have an enough growth of q_n needed to conclude the convergence. Of course, the problem is that the second term is of the opposite sign so it is not obvious if we can still ensure a similar rate of growth. In this reason, Makino and Aoyama tried to compare a smaller multiple of $(2n+1)qq_n$ with $n^2p^2q_{n-1}$, and obtained the result below. To be precise, they claimed they have a more general result (which they called the "generalized error evaluation equation" or *GEEE* in short), but I seriously doubt the credibility of that result.

Theorem 13.1 (Makino-Aoyama, 2000)**.**
Let $z = \frac{p}{q}$ be any rational number with $|z| < 1$. Then

$$qq_n > np^2q_{n-1}$$

holds for all $n \geq 1$, where $\frac{p_n}{q_n}$ is the nth convergent from the generalized continued fraction (23)*.*

Proof. We use induction on n. When $n = 1$,

$$qq_1 - p^2q_0 = q^2 - p^2 > 0$$

by the assumption on z. Now, suppose

$$qq_n > np^2q_{n-1}$$

holds for some $n \geq 1$, then

$$\begin{aligned} qq_{n+1} - (n+1)p^2q_n &= q\left((2n+1)qq_n - n^2p^2q_{n-1}\right) - (n+1)p^2q_n \\ &= \left((2n+1)q^2 - (n+1)p^2\right)q_n - qn^2p^2q_{n-1} \\ &> (n+1)(p^2 - q^2)q_n \end{aligned}$$

where the last inequality follows from the induction hypothesis. By the assumption on z, the right-hand side is positive, so we get the conclusion. □

[15] It seems that (23) is a specific instance of the so called *Gauss' continued fractions*, which seem to have a general theory of convergence already developed. However, I think that is not extremely relevant here. I will probably look into it later.

Corollary 13.2.
In the setting of Theorem 13.1, we have followings:

1. *$q_{n+1} \geq (n+1)qq_n$ holds for all $n \geq 0$ where the equality holds only when $n = 0$, and*
2. *The sequence $\left(\frac{p_n}{q_n}\right)_{n=0}^{\infty}$ converges to some $\alpha \in \mathbb{R}$ and*
$$\left|\frac{p_n}{q_n} - \alpha\right| \leq \frac{2\left|z\right|^{2n+1}}{(n+1)(1-\left|z\right|^2)} \tag{26}$$
holds for all $n \geq 0$. Furthermore, the sequence $\left(\frac{p_n}{q_n}\right)_{n=0}^{\infty}$ strictly increases to the limit α if $z > 0$ and it strictly decreases to α if $z < 0$.

Proof. 1. Directly follows from the recurrence relation $q_{n+1} = (2n+1)qq_n - n^2p^2q_{n-1}$ and Theorem 13.1.

2. Since $q_n \geq (n!)q^n$ follows from *1*, (25) yields
$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \leq \frac{2\left|z\right|^{2n+1}}{n+1}.$$
Hence, for any $m > n$,
$$\left|\frac{p_m}{q_m} - \frac{p_n}{q_n}\right| \leq \frac{2\left|z\right|^{2n+1}}{n+1}\left(1 + \left|z\right|^2 + \ \cdots \ + \left|z\right|^{2(m-n-1)}\right) \leq \frac{2\left|z\right|^{2n+1}}{(n+1)(1-\left|z\right|^2)}$$
holds. This shows the convergence and the inequality (26). Note that in (25), the sign of the right-hand side is equal to the sign of z, so the last statement also follows.

□

Therefore, we can use the convergent $\frac{p_n}{q_n}$ to approximate $\log\left(\frac{1+z}{1-z}\right)$ as precise as desired, where the error bound can be estimated using (26). However, the convergence speed estimated by (26) is not very great, so we derive a better bound.

Proposition 13.3.
In the setting of Theorem 13.1, let $\alpha := \lim_{n\to\infty} \frac{p_n}{q_n}$, then
$$\left|\frac{p_0}{q_0} - \alpha\right| \leq \frac{2\left|p\right|q}{q^2 - p^2}$$
and
$$\left|\frac{p_n}{q_n} - \alpha\right| \leq \frac{2\left|p\right|^{2n+1}(n-1)!(n!)}{(n+1)(q^2-p^2)q_{n-1}q_n}$$
holds for all $n \geq 1$.

Proof. Recall that for all $n \geq 0$, we have $q_{n+1} \geq (n+1)qq_n$. Hence,
$$\begin{aligned}\left|\frac{p_{n+2}}{q_{n+2}} - \frac{p_{n+1}}{q_{n+1}}\right| &= \frac{2\left|p\right|^{2n+3}((n+1)!)^2}{q_{n+1}q_{n+2}} \leq \frac{\left|p\right|^2(n+1)}{q^2(n+2)} \cdot \frac{2\left|p\right|^{2n+1}(n!)^2}{q_nq_{n+1}} \\ &= \frac{n+1}{n+2}\left|z\right|^2\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right|\end{aligned} \tag{27}$$

holds for all n. Therefore, inductively we get

$$\begin{aligned}\left|\frac{p_{n+k+1}}{q_{n+k+1}} - \frac{p_{n+k}}{q_{n+k}}\right| &\leq \frac{n+k}{n+k+1} \cdot \frac{n+k-1}{n+k} \cdot \cdots \cdot \frac{n+1}{n+2} |z|^{2k} \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \\ &= \frac{n+1}{n+k+1} |z|^{2k} \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right|\end{aligned}$$

for any $k \geq 0$. Thus,

$$\left|\frac{p_{n+k+1}}{q_{n+k+1}} - \frac{p_n}{q_n}\right| \leq \frac{1}{1-|z|^2} \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right|$$

holds for all $k \geq 0$, so taking the limit $k \to \infty$ gives

$$\left|\frac{p_n}{q_n} - \alpha\right| \leq \frac{1}{1-|z|^2} \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right|.$$

Specializing this with $n = 0$ then gives

$$\left|\frac{p_0}{q_0} - \alpha\right| \leq \frac{q^2}{q^2-p^2} \cdot \frac{2|p|}{q} = \frac{2|p|q}{q^2-p^2}.$$

On the other hand, if $n \geq 1$, then (27) shows

$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \leq \frac{n}{n+1} |z|^2 \left|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right|,$$

thus we get

$$\left|\frac{p_n}{q_n} - \alpha\right| \leq \frac{|z|^2}{1-|z|^2} \cdot \frac{n}{n+1} \cdot \frac{2|p|^{2n-1}((n-1)!)^2}{q_{n-1}q_n} = \frac{2|p|^{2n+1}(n!)^2}{n(n+1)(q^2-p^2)q_{n-1}q_n}$$

as desired. □

Having established an error bound formula for $2\tanh^{-1}\left(\frac{p}{q}\right)$, we compute that for $\ln x$ by doing the change of variable $z = \frac{x-1}{x+1}$. Note that $z \in (-1, 1)$ holds if and only if $x \in (0, \infty)$ holds. Then when x is rational, we can find p, q by finding the reduced form of $\frac{x-1}{x+1}$.

13.2 When the logarithm is rational

Once we have a way to compute successively better interval estimates for natural logarithm of positive rational numbers, we can immediately compute the continued fraction expansion of $\log_a b$ for any positive rational numbers a, b (even including the case $a = 1$, if we interpret $\log_a b$ as ∞ in that case) by applying Gosper's algorithm to $\ln b/\ln a$. However, recall that Gosper's algorithm traps in an infinite loop if $\log_a b$ is rational.[16] Therefore, we have to detect this case upfront, and use the continued fraction expansion algorithm for rational numbers if $\log_a b$ is detected to be rational.

Essentially, this can be in fact done by going back to our first, naïve attempt of computing the continued fraction expansion of $\log_a b$. Roughly speaking, what we do is basically to

[16] Note that $\ln x$ is never rational if x is rational. Indeed, if $e^{p/q} = x$ holds, then e is in a finite field extension of $\mathbb{Q}(x)$, but since e is transcendental, this is impossible if x is rational.

just follow the naïve algorithm, and conclude that $\log_a b$ should be irrational if the numbers involved grow at all.

To be precise, suppose if we want to see $\log_a b = \frac{m}{n}$ holds where $\frac{m}{n}$ is a rational number in its reduced form, which means $a^m = b^n$. If $a = \prod_{i=1}^{d} p_i^{e_i}$ is the prime factorization, then $a^m = \prod_{i=1}^{d} p_i^{me_i}$ must be the prime factorization of b^n, but since m and n are coprime, it follows that n divides every e_i. Therefore, there exists a rational number c such that $a = c^n$, which shows $b = c^m$.

Now, let $a = \frac{p}{q}$, $b = \frac{r}{s}$, and $c = \frac{t}{u}$ be their reduced forms. Then since $\frac{p}{q}$ and $\frac{t^n}{u^n}$ are both of their reduced forms, we conclude $p = t^n$ and $q = u^n$. Similarly, we have $r = t^m$ and $s = u^m$ if $m \geq 0$, or $r = u^{-m}$ and $s = t^{-m}$ if $m < 0$.

For simplicity, so we first detect the sign of m and replace b by $1/b$ to make m nonnegative if necessary. Since $\log_a b \geq 0$ if and only if either $a > 1$ and $b \geq 1$ or $a < 1$ and $b \leq 1$, this can be easily done by inspecting these inequalities. Hence, without loss of generality, we can assume $m \geq 0$, so $r = t^m$ and $s = u^m$.

Next, we ensure $m \geq n$ by swapping a and b if necessary. Note that $\frac{m}{n} \geq 1$ if and only if either $a > 1$ and $b \geq a$ or $a < 1$ and $b \leq a$, so by simply inspect these inequalities, we can assume without loss of generality that $m \geq n$.

Then, now we see if the equations $p = t^n$, $r = t^m$ have a solution. To do so, note that if it ever had a solution, then p must divide r since $m \geq n$ is assumed. If this is not the case, then we immediately conclude that $\log_a b$ is irrational. Otherwise, write $r = p^d p'$ where d is the highest integer exponent such that p^d divides r. If $\log_a b$ is rational, then this d must be equal to $\left\lfloor \frac{m}{n} \right\rfloor$ and p' must be equal to $t^{(m \mod n)}$. Hence, we run Euclid's algorithm at this point. That is, we replace (p, r) by (p', p) and repeat this procedure. If $\log_a b$ is rational, then at some point p' must become 1. Conversely, if we reach to the point where p' becomes 1, then we conclude that $p = t^n$ and $r = t^m$ have a solution.

In addition to that, by tracking the relevant coefficients, we can also figure out m, n from this procedure. Concretely, we start from

$$\begin{pmatrix} \log_t p \\ \log_t r \end{pmatrix} = \begin{pmatrix} m \\ n \end{pmatrix},$$

and note that

$$\begin{pmatrix} \log_t p' \\ \log_t p \end{pmatrix} = \begin{pmatrix} -d & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \log_t p \\ \log_t r \end{pmatrix} = \begin{pmatrix} -d & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} m \\ n \end{pmatrix},$$

so

$$\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & d \end{pmatrix} \begin{pmatrix} \log_t p' \\ \log_t p \end{pmatrix}.$$

Hence, by repeating the procedure until p' becomes 1, we get

$$\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & d_1 \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & d_k \end{pmatrix} \begin{pmatrix} 0 \\ \log_t r_k \end{pmatrix}$$

where (p_{i+1}, r_{i+1}) is the pair (p', p) obtained from (p_i, r_i), $(p_0, r_0) = (p, r)$ and d_{i+1} is the highest integer exponent such that $p_i^{d_{i+1}}$ divides r_i. Since m and n are coprime, r_k should be equal to t, so the second entry of the last matrix in the right-hand side is actually 1. Therefore, by keeping track of the product of $\begin{pmatrix} 0 & 1 \\ 1 & d_i \end{pmatrix}$'s, we can compute m, n.

Note that this only solves the half of the problem: we also need to see the existence of a solution to the equations $q = u^n$ and $s = u^m$. However, the exactly the same numbers should appear for d_i's, and if not, then we immediately conclude that $\log_a b$ is irrational. When p' reaches 1, the corresponding q' also must be 1 if $\log_a b$ is rational. Otherwise, we conclude $\log_a b$ is irrational. If they are both 1, then we conclude $\log_a b$ is rational.

14 TODO List

- Think of an algorithm that determines the exact range of $n \geq 0$ such that $\lfloor nx + y \rfloor = \lfloor n\xi + \zeta \rfloor$ holds, when x, y, ξ, ζ are all given. Will it be possible to get rid of the assumption $n_{\min} = 0$?
- Applications to James Anhalt's algorithm.
- Application to log computations in Dragonbox.

15 Acknowledgements

Seung uk Jang and Seok-Hyeong Lee helped me a lot on several things. Seok-Hyeong Lee reminded me of the observation that the set of (ξ, ζ)'s must be convex, which led me to the idea of using convex duality and convex hull algorithm for computing the intersection. This idea was originally revolving around using the duality of convex cones, but it was Seung uk Jang who pointed out that the duality of convex functions might be easier to work with. They also helped me on the understanding of the indeterminacy locus of bilinear fractional mappings. In addition to that, the rational detection algorithm explained in Section 13.2 is just a slight refinement of Seok-Hyeong Lee's suggestion. My original attempt was something much more naïve, based on prime factorization.

References

[1] T. Granlund and P. L. Montgomery, *Division by invariant integers using multiplication*, ACM SIGPLAN Notices, vol.29, no.6, 1994, pp.61-72.

[2] H. S. Warren Jr, *Hacker's Delight (1st edition)*, Addison-Wesley, Boston.

[3] D. Lemire, C. Bartlett and O. Kaser, *Integer division by constants: optimal bounds*, Heliyon, vol.7, no.6, 2021, doi:10.1016/j.heliyon.2021.e07442.

[4] R. Giulietti, *The Schubfach way to render doubles*, `https://drive.google.com/file/d/1IEeATSVnEE6TkrHlCYNY2GjaraBjOT4f/edit`, retrieved 2024-07-15.

[5] T. Drane, W. -c. Cheung and G. Constantinides, *Correctly rounded constant integer division via multiply-add*, 2012 IEEE International Symposium on Circuits and Systems (ISCAS), Seoul, Korea (South), 2012, pp.1243-1246, doi:10.1109/ISCAS.2012.6271461.

[6] J. Jeon, *Dragonbox: A New Floating-Point Binary-to-Decimal Conversion Algorithm*, `https://github.com/jk-jeon/dragonbox/blob/master/other_files/Dragonbox.pdf`, retrieved 2024-07-15.

[7] M. Einsiedler and T. Ward, *Ergodic Theory with a view towards Number Theory*, Graduate Texts in Mathematics 259, Springer.

[8] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific.

[9] B. Gosper, *Continued Fractions*, Informal notes. Accessible at `https://github.com/jk-jeon/idiv/blob/main/docs/pdf/cont-frac-gosper-1.pdf`, retrieved 2024-07-15.

[10] J. L. Kelley, *General Topology*, Springer Verlag.

[11] I. Makino and T. Aoyama, *Calculation of logarithmic function with Continued Fraction Expansions*, Josai Mathematical Monographs, 2, 2000, pp.43-68.

[12] I. Makino and T. Aoyama, *The arbitrary precision calculation of logarithms with continued fraction expansions*, RIMS Kôkyûroku, no.1138, 2000, pp.240-246.