Дървета на ван Емде Боас

20.11.2020 г. (van Emde Boas Trees)

Дадено: $U \in \mathbb{N}$.

Динамични заявки към множество $S \subseteq \{0,1,...,U-1\}$

- 1. $pred_{S}(x)$
- 2. $succ_{S}(x)$
- 3. $search_S(x) : "x \in S"$ $x \in \{0,...,U-1\}:$
- 4. $insert(S, x) : S \leftarrow S \cup \{x\}$
- 5. $delete(S, x) : S \leftarrow S \setminus \{x\}$

Решение:

Наивно:

Сортиран списък за S

Сортиран масив за S

Балансирани наредени дървета $O(\log |S|)$ за всяка една от заявките.

Ако x_1, x_2, \ldots, x_n са произволни (различни) числа може да ги сортираме по следния начин:

$$S \leftarrow \emptyset$$

for $i = 1$ **to** n **do**
 $S \leftarrow Insert(S, x_i)$
 $y_1 \leftarrow succ_S(0)$
for $i = 1$ **to** $n - 1$ **do**
 $y_{i+1} \leftarrow succ_S(y_i + 1)$

Ако числата са естествени и различни, то горната процедура ще ги сортира. Операциите Insert и succ може да ги ограничим с $\log(n)$.

В общия случай не може да осигурим нищо по-добро от $O(\log |S|)$ за заявка.

С ограничението обаче за U с дърветата на ван Емде Боас получаваме ограничение за всяка една от двете заявки от $O(\log \log U)$.

$$x = \overline{x_1 x_2 \dots x_u}$$

В алгоритъма на Willard направихме двоично търсене по думата образувана от двоичния запис на числото x, за да намерим най-дългия ѝ префикс, който се съдържа в дървото от префикси (се среща в нашето множество). Но нека сега не избързваме толкова с това търсене, а разгледаме само първата му стъпка:

$$m = \left[\frac{l+r}{2}\right]$$

1 сл. $x_1x_2\dots x_m$ се среща в дървото, тогава търсим продължение x_{m+1},\dots,x_r и 2 сл. $x_1x_2\dots x_m$ не се среща в дървото и тогава търсим съкращаване което се съдържа в x_1,\dots,x_m

Това разделяне на висши и нисши битове ще ни послужи за изграждане на идеята. То както видяхме съответства на делене с остатък.

Основното наблюдение в алгоритъма на ван Емде Боас е следното аритметично наблюдение:

<u>Твърдение</u>: Ако $0 \le x < 2^u$, то x се представя по единствен начин като $x = 2^{u_1}x_a + x_b$, като

- 1) $u_1 = [u/2]$
- 2) $0 \le x_b < 2^{u_1}$
- 3) $0 \le x_a < 2^{u-u_1}$

<u>Доказателство</u>: Делим x с частно и остатък на 2^{u_1} . Тогава $x=2^{u_1}x_a+x_b$, като $0\leq x_b<2^{u_1}$. Накрая, ако $x_a\geq 2^{u-u_1}$, то $x\geq 2^{u_1}x_a+0\geq 2^{u_1}2^{u-u_1}=2^u$. \Rightarrow противоречие с $x<2^{u_1}$.

Идея за търсене $x \in S$ и $0 \le x < 2^u$.

- 1. Представяме $x=2^{u_1}a+b$. За всяко $0 \le a < 2^{u-u_1}$: $R[a]=\{0 \le b < 2^{u_1} | 2^{u_1}a+b \in S\}$
- 2. $S' = \{0 \le a < 2^{u-u_1} | R[a] \ne \emptyset\}$ Търсим x_b в множеството $R[x_a]$, като параметъра в този случай ще е $u_1 = [u/2]$

Тоест, ако успеем да организираме нашата структура от данни по начин, по който да имаме тези множества подготвени по удобен начин, така че да може рекурсивно да може да задаваме тези заявки - на всяка стъпка параметъра u, който съответства на броя битове намалява два пъти. В резултат на което имаме логаритмичен процес за числата по броя битове в x, но x ще бъде в зададения интервал, то спрямо y ще имаме y по y п

Разклонеността на дърветата на дърветата на ван Емде Боас е различна във всяко ниво.

Естествения въпрос е каква ще е големината на тази дървовидна структура.

S е множество с елементи от $\{0,1,...,2^{u}-1\}$

Разбира се при u=0,1 поддържаме елементите на S експлицитно в масив с 4 елемента $arr[0..3] // arr[i] = 0 \Leftrightarrow i \notin S, arr[i] = 1 \Leftrightarrow i \in S.$

Твърдение:

За всяко u, празно дърво на ван Емде боас с параметри u, $U=2^u$ заема памет O(U) и може да се построи за такова време .

Доказателство:

 $Space(2^u) \leq c + Space(2^{u-[u/2]}) + 2^{u-[u/2]}Space(2^{[u/2]})$. Целта е да покажем, че тази памет е линейна.

С индукция по u може да покажем, че $Space(2^{u}) \leq c_0(2^{u}-2)$.

Базата е тривиална, защото лесно може да си изберем c_0 , така, че да ни бъде удобно.

Индуктивна стъпка: нека $u_1 = [u/2]$. Тогава

$$\begin{array}{l} \underset{hypothesys}{induction} \\ Space(2^{u}) & \leq c + c_{0}(2^{u-u_{1}} - 2) + 2^{u-u_{1}}c_{0}(2^{u_{1}} - 2) = \\ & = c_{0}2^{u} - 2c_{0}2^{u-u_{1}} + c_{0}2^{u-u_{1}} - 2c_{0} + c = c_{0}2^{u} - 2c_{0} + \underbrace{c - c_{0}2^{u-u_{1}}}_{u_{1} < u} < \\ & < c_{0}2^{u} - 2c_{0} = c_{0}(2^{u} - 2). \end{array}$$

$U \in \mathbb{N}$

динамично $S \subseteq \{0,1,...,U-1\}$

- $search(S, x) : x \in S$?
- $pred(S, x) : y = max\{z \in S | z \le x\}$
- $succ(S, x) : y = min\{z \in S | z \ge x\}$
- $insert(S, x) : S \leftarrow S \cup \{x\}$
- $delete(S, x) : S \leftarrow S \setminus \{x\}$

В основата на алгопритъма на ван Емде Боас е следната фина идея (разделяй и владей относно двоичния запис на числото $x \in \{0,1,...,U-1\}$):

$$U \leq 2^u$$
 $x = a_x 2^{[u/2]} + b_x$, където $0 \leq b_x < 2^{[u/2]}$ и $0 \leq a_x < 2^{u-[u/2]}$

За
$$S:0\leq a<2^{u-[u/2]}$$
 дефинираме $R[a]=\{b<2^{[u/2]}\,|\,a2^{[u/2]}+b\in S\}$ и $S'=\{a<2^{u-[u/2]}\,|\,R[a]\neq\emptyset\}$

Допълнително имахме размера (броя на елементите на множество S):

• size(S)

И също така в интерес на ефективността за делението с частно и остатък:

- $u; U = 2^u; U_1 = 2^{[u/2]}$
- $\max(S)$; $\min(S)$

Това е представянето за u > 2, докато за $u \le 2$ правим наивно представяне с масив.

$$U = 64$$
, $u = 6$
 $S = \{0, 1, 8, 9, 10, 11, 19, 22, 54, 62\}$

$$u_1 = [u/2] = 3$$

 $U_1 = 2^{u_1} = 8$

$$0 = 0 \times 8 + 1$$

$$1 = 0 \times 8 + 1$$

$$8 = 1 \times 8 + 0$$

$$9 = 1 \times 8 + 1$$

$$10 = 1 \times 8 + 2$$

$$11 = 1 \times 8 + 3$$

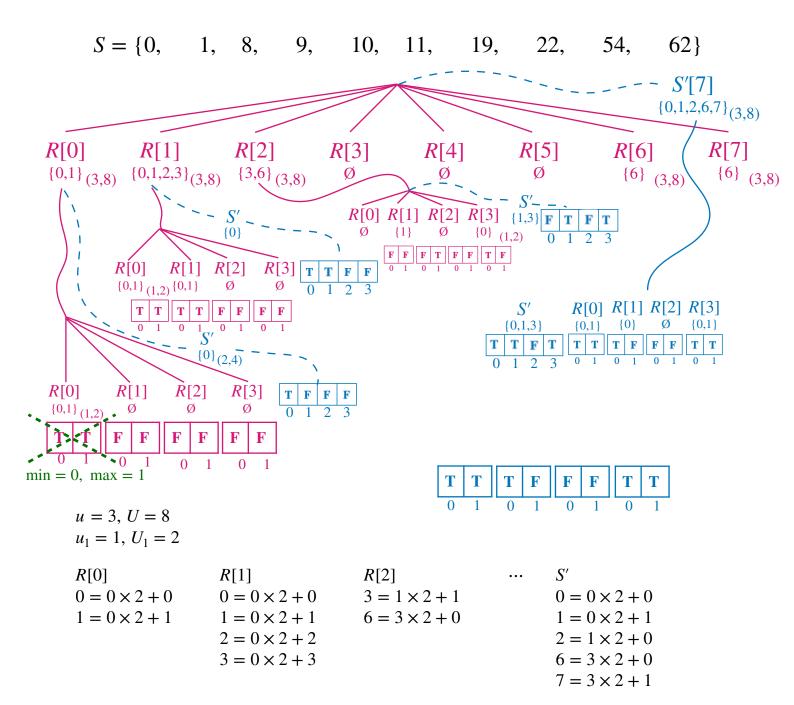
 $19 = 2 \times 8 + 3$

$$19 = 2 \times 8 + 3$$

 $22 = 2 \times 8 + 6$

$$54 = 6 \times 8 + 6$$

$$62 = 7 \times 8 + 6$$



<u>Забележка</u>: Височината на дърво на ван Емде Боас за $U=2^u$ (макс. бр. ел.) е $O(\log u) = O(\log \log U)$, т.е. дървото е доста плитко. Това се дължи на разклоняването с високи темпове по етажите на дървото.

```
 \begin{aligned} \mathbf{procedure} \ search(S,x) \ \{ \\ \mathbf{if} \ u(S) & \leq 2 \ \mathbf{then} \\ \mathbf{return} \ arr_S[x] \\ a_x & \leftarrow [X/U_1(S)] \\ b_x & \leftarrow x \mod (U_1(S)) \\ \mathbf{return} \ search(R[a_x],b_x) \\ \} \\ \mathbf{C} \text{Ложност:} \ O(\log u(S)) & = O(\log \log U(S)) = O(\log \log U) \end{aligned}
```

Ясно е че двете заявки съотв. за предшественик и наследник са интересни само когато елемента, за който ги търсим не е в дървото, тъй като тогава отговора ще е тривиален и реално на него ще е също толкова ефективно да се отговори и със заявката за search/member.

```
Пример: 13 = 1 \times 8 + 5
pred_{S}(13)
5 = 2 \times 2 + 1
2 + към S'
Търсим предшественика във S' на 2: pred_{S'}(2) = 1
pred_{R[0]}(5) = 1 \times 2 + \max R[1] = 1 \times 2 + 1 = 3
                                     lower R[1]
 \Rightarrow 1 \times 8 + 3 = 11.
   pred(S, x)
• pred_naive(S, x)
          U(S) < 4
 procedure pred(S, x) {
         if size(S) = 0 then
                   return - 1
         if x \ge \max(S) then
                   return max(S)
         if x < \min(S) then
                   return - 1
         if U(S) < 4 then
                   return pred_naive(S, x)
[x/U_1(S)]
3Haem, че:
\min(S) \le x < \max(S)
x = a_x U_1(S) + b_x
         a_x \leftarrow [x/U_1(S)]
         b_x \leftarrow x \mod U_1(S)
         if size(S.R[a_x]) \neq 0 and min(S.\widetilde{R}[a_x]) \leq b_x then
                  b' \leftarrow prev(S.\widetilde{R}[a_r], b_r)
         a' \leftarrow prev(S.\widetilde{S'}, a_x - 1)
b' \leftarrow \max(\widetilde{R}[a']) \quad \text{if } a' = -1 \text{ then}
\text{return } a'.U_1(S) + b' \quad \text{return } \min(S)
```

Сложност: Сложността на алгоритама е очевидно толкова колкото е дълбочината на дървото, тъй като имаме толкова рекурсивни извиквания (най-много едно за всяко ниво) колкото е височината на дървото, а останалите съпровождащи операции се изпълняват за константно време. Броят на битовете в u(S) намалява с единица след всяко извикване на рекурсивната функция.

Следователно сложността на процедурата за намиране на предшественик на x е $O(\log u(S)) = O(\log \log U(S)) = O(\log \log U(S))$.

Коректност: Ще подходим индуктивно по нивата на дървото.

<u>1 сл</u>. Да допуснем, че $x < \max(S)$ и $x \ge \min(S)$

$$R[a_x] \neq \emptyset$$

 $b_x \ge \min(R[a_x])$

От индукционната хипотеза имаме, че $b = pred_{R[a_x]}(b_x)$.

Твърдим, че $a_x U_1(S) + b' = a' U_1(S) + b' = pred_S(x)$.

От
$$b' \le b_x \Rightarrow a_x U_1(S) + b' \le x$$
.

Да допуснем, че има елемент в нашата структура межди горните два, т.е.:

$$a_x U_1(S) + b' \le a'' U_1(S) + b'' \le x = a_x U_1(S) + b_x$$
 и това което знаем е, че

 $0 \leq b',\, b_{\scriptscriptstyle X} < U_1(S) \Rightarrow\,$ делейки целочислено на $U_1(S)$ получаваме, че

 $a_{\scriptscriptstyle X} \leq a'' \leq a_{\scriptscriptstyle X} \Rightarrow a'' = a_{\scriptscriptstyle X}$ и ако $a''U_1(S) + b'' \in S \Leftrightarrow b'' \in R[a_{\scriptscriptstyle X}].$ Сега вече е ясно, че

$$a_x U_1(S) + b' \le a'' U_1(S) + b'' \le a_x U_1(S) + b_x \Leftrightarrow b' \le b'' \le b_x$$
. Ho

$$b' = pred_{R[a_x]}(b) \Rightarrow b'' = b'$$
if $a''U_1(S) + b'' \in S$

2 сл. Отново може да предполагаме, че $\min(S) \leq x < \max(S)$, но този път се е случило така, че $R[a_x] = \emptyset$ или $b_x < \min(R[a_x])$

$$a' = pred_{S'}(a_x - 1)$$

$$b' = \max(R[a'])$$

Твърдим, че $a'U_1(S) + b' = pred_S(x)$

1).
$$a' < a_x$$
 u $b' < U_1(S) \Rightarrow$

$$a'U_1(S) + b' \le (a_r - 1)U_1(S) + b' < (a_r - 1)U_1(S) + U_1(S) = a_rU_1(S) \le x.$$

Нека
$$y \in S, y = a''U_1(S) + b'', 0 \le b'' < U_1(S)$$
 и

$$a'U_1(S) + b' \le y \le x \Rightarrow a'U_1(S) + b' \le a''U_1(S) + b'' \le a_xU_1(S) + b_x \Rightarrow$$

$$\begin{array}{c} \text{divide to } U_1(S) \\ \Rightarrow \\ b', b'', b_x < U_1(S) \end{array} a' \leq a'' \leq a_x \text{. Ho } a' = pred_{S'}(a_x - 1) \text{ in } a'' \in S' \Rightarrow \\ \end{array}$$

$$\Rightarrow$$
1.) $a'' = a' (a'' \le a_x - 1)$ или 2.) $a'' = a_x$

1 сл.
$$a'' = a'$$
: $b'' \in R[a''] = R[a'] \Rightarrow$

$$\Rightarrow b'' \le \max(R[a'']) = b' \Rightarrow a''U_1(x) + b'' \le a'U_1(x) + b' \Rightarrow$$

$$=R[a']$$

$$\Rightarrow$$
 y = $a'U_1(x) + b'$.

2 сл. $a''=a_x$: $b''\in R[a_x]$. Тъй като $y\leq x\Rightarrow b''\leq b_x<\min(R[a_x])\Rightarrow$ ч противоречие с това, че $b_x\in R[a_x]$!

• Добавяне на елемент

$$x \notin S$$
; $x = a_x U_1(S) + b_x$.

$$R[a] = \{b < U_1(S) \mid aU_1(S) + b \in S\}$$

 $S' = \{a \mid R[a] \neq \emptyset\}$

- ако $R[a_x] \neq \emptyset$, добавяме b_x към $R[a_X]$;
- ако $R[a_x] = \emptyset$; първо добавяме a_x към S' и второ добавяме b_x към $R[a_x]$.

На пръв поглед може да си кажем, че вместо едно добавяне правим две добавяния и промяната е дребна. Но тези добавяния са рекурсивни операции и не са атомарни, което означава, че ако запишем рекурентното уравнение което съответства на времевата сложност във втория случай, бихме получили:

$$Timeig(U(S)ig)=2Timeig(rac{U(S)}{2}ig)$$
, което е ралично от $Timeig(U(S)ig)=Timeig(rac{U(S)}{2}ig)+1$,

кактото е при първия случай и там получаваме логаритмична сложност, а в първото уравнение ще получим линейна сложност.

Как ще се справим с този проблем? Справянето с този проблем е свързан с едно простичко наблюдение: ако се налага да добавяме $\mathbf{a}_{\mathbf{x}}$, то $\mathbf{R}[\mathbf{a}_{\mathbf{x}}] = \emptyset$ е празно. При това положение b_x ще бъде единствения елемент в $R[a_x]$ след добавянето и той ще бъде и \max и \min .

$$\tilde{R}[a] = \{b < U_1(S) \mid aU_1(S) + b \in S \setminus \min(S) \& \max(S)\}$$

$$\tilde{S} = \{a \mid \tilde{R}[a] \neq \emptyset\}$$

Може да се наложи да добавим x в S : $x < \min(S)$ или $x > \max(S)$

$$S \cup \{x\} = S. \left(S \setminus \{y\} \cup \{x\} \right) \cup \{y\}$$

В случая когато $y = \min(S)$, оперативно това математическо равенство може да се изрази по следния начин:

$$insert(S, x) \{ \\ y \leftarrow \min(S) \\ \min(S) \leftarrow x \\ insert(S, y) \}$$
 (S\{y\} \cup \{x\})

В общия случай, когато min(S) < x < max(S):

$$x = a_x U_1(S) + b_x$$
• $\tilde{R}[a_x] \neq \emptyset$; $\tilde{R}[a_x] \cup \{b_x\}$
• $\tilde{R}[a_x] = \emptyset$; $\tilde{S}' \cup \{a_x\}$

$$\max(R[a_x]) = \min(\tilde{R}[a_x]) = b_x$$

Предлагаме:

 $insert_naive(S, x)$ sa $U(S) \le 4$

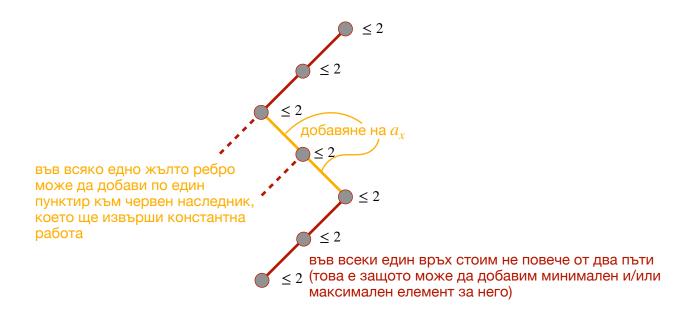
```
procedure insert(S, x) {
           if U(S) \leq 4 then
                   return inert_naive(S, x)
           if size(S) < 2 then {
                   if size(S) = 0 then {
                           \min(S) \leftarrow x
                           \max(S) \leftarrow x
                    \{ |S| = 1, min(S) = max(S) \}
                           if min(S) < x then
O(1)
                                    \max(S) \leftarrow x
                           else
                                    \min(S) \leftarrow x
                   size(S) = size(S) + 1
                   return
           if max(S) < x then
   min(S) < y < max(S)
                   y \leftarrow \max(S)
                   \max(S) \leftarrow x
                   return insert(S, y)
           if min(S) > x then
                   y \leftarrow \min(S)
                   \min(S) \leftarrow x
                   return insert(S, y)
           a_r \leftarrow [x/U_1(S)]
           b_x \leftarrow x \mod U_1(S)
           if size(S.\tilde{R}[a_x]) = 0 then
                   insert(S.\tilde{S}, a_x)
           insert(S, \tilde{R}[a_r], b_r)
           size(S) \leftarrow size(S) + 1
   }
```

Анализ на времето: Основния инвариант, който наблюдаваме е u(S). Да видим какво се случва с него. В първите няколко \mathbf{if} -а (първите 2) не се случва нищо и директно приключваме с процедурата за константно време. В третия и четвъртия \mathbf{if} оператори, напървите редове имаме O(1), но както отбелязахме и в двата случая $\min(S) < y < \max(S)$. Следователно, въпреки че броя на битовете не се променя при тези стъпки, рекурсивното извикване на функцията с y ще прескочи първата и втората част от програмата и ще се засили да изпълнява последната част от тази процедура. Така, че:

• за всяко S има най-много две последователни извиквания $insert(S,\cdot)$

$$u(\tilde{S}) \le \left\lceil \frac{u(S) + 1}{2} \right\rceil$$

Рекурсивния процес може да се разглежда като разходка в дървото, като преминаваме от връх на връх. В един връх се задържаме най-много два поредни пъти след което продължаваме надолу в поддървото и изминаваме точно един път от корен до листо В някой случай този път може да бъде разклонен, но тогава пътя който ще получим ще изглежда по следния начин:



Общия брой преходи, които ще извърви нашия алгоритъм е еквивалентен на времевата сложност и тя е не повече от три пъти дължината на пътя от корена до листо в това дърво. А този път както видяхме има дължина от $O(\log u(S)) = O(\log \log U)$. Така, че най-много $3 \times \log \log U$ е сложността на алгоритъма, който описахме. Основната печалба идва от това, че минимума и максимума на нашето множество не се пазят експлицитно два пъти както беше в първоначалната реализация, а се пазят само по веднъж в самия връх.

Изтриването е дуална операция на вмъкването. Трудността ще дойде от това, че от време на време ще трябва да актуализираме минималния и максималния елемент в нашето множество.

Изтриване на елемент.

Предполагаме, че $x \in S$

• $delete_naive(S,x) // U(S) \le 4$ Отново представяме x по следния начин:

$$x = a_x U_1(S) + b_x$$

• $\min(S) < x < \max(S)$

$$S \setminus x \sim \tilde{R}[a_x] \setminus \{b_x\}$$

 \cdot b_x - може да се окаже единствен в $\tilde{R}[a_x]$ и ако това се случи, тогава трябва да изтрием a_x от S': $S' \setminus \{a_x\}$

• $x = \min(S)$

 $y = \min(S \setminus \{x\})$ - в общия случай тази функция ще отнеме $\log \log U$ и ще получим две рекирсивни извиквания, което е лошо.

Но ако за момент си мислим, че имаме у, тогава:

$$S \setminus \{x\} = S \setminus \{x, y\} \cup \{y\} = (S \setminus \{y\}) \setminus \{x\} \cup \{y\}$$

$$delete(S, x)$$

$$delete(S, y)$$

$$min(S) \leftarrow y // \setminus \{x\} \cup \{y\}$$

```
procedure fast\_succ\_min(S) { // предполагаме, че |S| \ge 2 if size(S) = 2 then /\!/ S = \{\min(S), \max(S)\} return \max(S) if size(S) \ge 3 then /\!/ |S| \ge 3 a \leftarrow \min(S.\tilde{S}) /\!/ \tilde{R}[a] \ne \emptyset b \leftarrow \min(S.\tilde{R}[a]) return aU_1(S) + b }
```

Коректност: $\min(S) = a_x U_1(S) + b_x$. Останалите елементи на S ще имат или по-голямо a или същото a, но по-голямо b. Намирайки най-малкото a в множеството \tilde{S} - намираме всъщност най-малкото a на елемент в множеството S, който е различен от минималния и е различен и от максималния, тъй като те са единствените, които не са представени по рекурсивен начин. Тъй като множеството S има поне три елемента - намираме най-малкото a, което е по-голямо или равно от a_x . И от тези, които имат едно и съшо a взимаме това с най-малкото b. Ясно е, че елемента, който ще получим ще бъде строго по-голям от минималния в множеството и други между него и минималния няма да има, защото те или биха имали по-малко a, което ще доведе до противоречие.

Дуално за предшественик на максимален елемент:

```
procedure delete(S, x) {
        if U(S) \leq 4 then
                 delete_naive(S, x)
                 return
        if U(S) \le 2 then //S = \{\min, \max\}
                if x = \min(S)
                         \min(S) \leftarrow \max(S)
                 else
                         \max(S) \leftarrow \min(S)
                 size(S) \leftarrow size(S) - 1
                return
        // if size(S) = 0 then max(S) = -1; min(S) \leftarrow +\infty
        //|S| \ge 3
        if max(S) = x then
                y \leftarrow fast\_pred\_max(S) // min(S) < y < max(S)
                delete(S, y)
                \max(S) \leftarrow y
                return
        if min(S) = x then
                y \leftarrow fast\_succ\_min(S)
                delete(S, y)
                \min(S) \leftarrow y
                return
         // ако сме стигнали до тук, чначи \min(S) < x < \max(S)
        a_x \leftarrow [x/U_1(S)]
        b_x \leftarrow x \mod U_1(S)
        // b_x \in S.\tilde{R}[a_x]; a_x \in S.\tilde{S}'
        delete(S.\tilde{R}[a_x], b_x) // (\star)
        if size(S.\tilde{R}[a_x]) = 0 then
                delete(S.\,\tilde{S}',a_r) // ако това се случи, то в ( \star ) сме имали времева сл. O(1)
        size(S) \leftarrow size(S) - 1
        return
}
S \subseteq \{0,1,...,U-1\}
U = U(S)
u(S) = \log U(S)
U_1(S) = 2^{[u(S)/2]}
min(S)
max(S)
\tilde{R}[a] = \{b < U_1(S) \mid aU_1(S) + b \in S \setminus \{\min(S), \max(S)\}\}\
\tilde{S}' = \{ a \mid \tilde{R}[a] \neq \emptyset \}
```