September 2, 2022 10:38 AM

Convolution / FFT : More Applications

Appl'n 4. String matching with mismatches

Given $p_1 \cdots p_m \in \Sigma^*$
$t_1 \cdots t_n \in \Sigma^*$ and k,

decide $\exists i$ s.t. Hamming distance between $p_1 \cdots p_m$ and $t_{i+1} \cdots t_{i+m}$ $\leq k$.

i.e. $|\{j : p_j \neq t_{i+j}\}| \leq k$

e.g. Hamming dist
$$\begin{matrix} 1&0&1&1&0&1&0 \\ 1&1&0&1&0&0&0 \end{matrix} = 3$$

text: "algorithmisfun"
pattern: "rythmic" $k=2$

trivial algm: $O(mn)$ time

Algm 1. Suffice to compute $\mu_i = |\{j : p_j = t_{i+j}\}|$ for all i

(yes if $\mu_i \geq m-k$)

How? for each $c \in \Sigma$,

compute $\mu_i^{(c)} = |\{j : p_j = t_{i+j} = c\}|$

$$= \sum_j \underbrace{[p_j = c]}_{A_j} \cdot \underbrace{[t_{i+j} = c]}_{B_{i+j}}$$

Convolution! (after reversing B)

let $\mu_i = \sum_{c \in \Sigma} \mu_i^{(c)}$

$\Rightarrow \boxed{O(|\Sigma|\, n \log n)}$ time

What if $|\Sigma|$ large?

Abrahamson's Alg'm '87:

idea - divide into high vs. low frequency case

let $\Sigma_H = \{ c \in \Sigma : c \text{ occurs} > \Delta \text{ times in pattern} \}$

$\Sigma_L = \{ c \in \Sigma : c \text{ occurs} \le \Delta \text{ times in pattern} \}$

$|\Sigma_H| \le \frac{m}{\Delta}$

1. High Case: for each $c \in \Sigma_H$,
   compute $\mu_i^{(c)}$ for all i by convolution

   let $\mu_i = \sum_{c \in \Sigma_H} \mu_i^{(c)}$

   $$\Rightarrow O(|\Sigma_H| \, n \log n)$$
   $$= O\left(\frac{m}{\Delta} n \log n\right)$$

2. Low case: for $\ell = 1$ to n
   - if $t_\ell \in \Sigma_L$ ← $\le \Delta$ choices for j
     - for each j with $p_j = t_\ell$
       - increment $\mu_{\ell - j}$

   $$\Rightarrow O(\Delta n)$$

total time $O\left(\frac{m}{\Delta} n \log n + \Delta n\right)$

choose $\Delta = \sqrt{m \log n} \Rightarrow \boxed{O(n\sqrt{m \log n} + n \log n)}$

$$= \boxed{\tilde{O}(n\sqrt{m})}$$

[illegible] Amir Lewenstein-Porat '04: $\tilde{O}(n\sqrt{k})$

(note - Amir-Lewenstein-Porat '04: $\tilde{O}(n\sqrt{k})$

Gawrychowski-Uzanski '18: $\tilde{O}\left(n + \frac{nk}{\sqrt{m}}\right)$

⋮

approx $O(n)$)

App'n 5: Subset Sum

Given set S of n positive integers & target number T,

decide $\exists$ subset $R \subseteq S$ that sums to T.

$(a \le T)$

Known: $\sim O(2^{n/2})$

or $O(nT)$ by DP

(say $S = \{a_1, \dots, a_n\}$

Define $C[\ell, i] = \text{true}$ iff $\exists R \subseteq \{a_1, \dots, a_\ell\}$ that sums to i

$(\ell = 0, \dots, n, \quad i = 0, \dots, T)$

Then $C[\ell, i] = C[\ell-1, i] \vee C[\ell-1, i-a_\ell]$)

Note/obs - if duplicates are allowed (i.e. R is a multiset)

can do better

Define $C^{(j)}[i] = \text{true}$ iff $\exists$ multiset $R \subseteq S$ with $\le j$ elems

that sums to i.

Then $$C^{(j)}[i] = \bigvee_{i'=0}^{i} \left(C^{(j/2)}[i'] \wedge C^{(j/2)}[i-i'] \right)$$

Boolean Convolution!

of array of size T

$\Rightarrow$ can compute $C^{(j)}$ from $C^{(j/2)}$ in $O(T \log T)$ time

$\Rightarrow$ total time $\boxed{O(T \log^2 T)}$

$= \boxed{\tilde{O}(T)}$

But original problem... doesn't work for

Koiliaris & Xu '17: $\tilde{O}(T\sqrt{n})$ deterministic

$\rightarrow$ Bringmann '17: $\tilde{O}(T)$ randomized.

Bringmann's Alg'm:

Lemma Suppose $S \subseteq [U]$ & sol'n uses $\leq k$ elems.

Then there is rand. alg'm with $\tilde{O}(k^2 U)$ time.

Pf:

Fact Put k balls in k^2 bins randomly.

With prob $\geq \frac{1}{2}$, every bin contains ≤ 1 ball.

Pf: $\Pr(\exists$ 2 balls in same bin$)$

$$\leq \binom{k}{2} \cdot \frac{1}{k^2} = \frac{1}{2}. \quad \square$$

idea - Partition S into k^2 subsets $S_1, \ldots, S_{k^2}$