# テーマ１：ストリームアルゴリズム（その３）

本日の講義の目的

- 最頻出回数近似問題のストリーミング計算困難性の証明

### 1.4. 最頻出回数近似問題とそのストリーミング計算困難性

問題と記号の復習から．以下では次のような記法を用いる．

$$
\begin{array}{lcl}
\boldsymbol{a} & = & a_1, a_2, ..., a_t, ..., a_m \ = \ a'_1, a'_2, ..., a'_j, ..., a'_n \ = \ \text{与えられる入力列} \\
\boldsymbol{a}_t & = & a_1, ..., a_t \ \ \text{時刻 } t \text{ までの列} \\
D & = & \{0, ..., N-1\} \ = \ \text{入力要素 } a_i \text{ の定義域} \\
n_{j,t} & = & a'_j \text{ が } \boldsymbol{a}_t \text{ に出現する回数} \\
F_{\infty,t} & = & \max_{1 \le j \le n} n_{j,t}.
\end{array}
$$

この $F_{\infty,t}$ が（時刻 $t$ までにおける）最頻出データの頻出回数であり，これを簡単に最頻出回数と呼ぶ．また，ここでは，最頻出回数 $F_{\infty,t}$ に対し，$((1-\varepsilon)F_{\infty,t}, (1+\varepsilon)F_{\infty,t})$ の範囲に入る値（$F_{\infty,t}$ の $\varepsilon$-近似）を求めることを考える．この問題に対し，$t = m$（最後まで読んだ時）のみを考えたとしても，少ない領域量でストリーミング計算することは不可能であることを示す．つまり，次の定理の証明が今回の目標である（なお証明の出典は [1,4]）．

**定理 3.11.** 任意の与えられた列 $\boldsymbol{a} = a_1, ..., a_m$ （ただし，$m \le 2N$ で $a_i \in \{0, ..., N-1\}$）に対し，$F_\infty$ の相対誤差 1/3-近似を 0.8 以上の確率で計算するストリームアルゴリズムの（平均）領域量は $cN$ 以上である．

### 準備：通信計算量 communication complexity

ある問題の計算量を評価する際に，その計算の本質を保ったまま，問題をもっと議論のしやすい計算モデルの上の問題に変換して考えることが少なくない．そのような変換を一般に還元という．今回の証明は，その還元がうまくいく代表的な例の一つである．

議論のしやすい計算モデルとして，ここでは Yao [2] が提唱した通信計算を考え，そのときの通信計算量を用いて計算の限界を議論する．「通信計算」と言っても，実際の通信とはかなり異なったもので，ゲームのようなものと考えた方がよいだろう．

ゲームは計算したい関数を $f$ ごとに決まる．関数 $f$ が通信計算問題を 1 つ決めると考えてよい．関数 $f(x, y)$ は $0, 1$ の列 $x, y$ を入力とし，0 もしくは 1 を返す関数とする．ルールとして $x, y$ は同じ長さ[1]の列とする．ゲームは二人のプレーヤー $A$, $B$ により行わ

[1] ここでは慣例に従い列の長さを $n$ と表すが，ストリーム計算の下界を議論するときは $N$ に対応する．

れる．二人には，それぞれ $x, y$ の一方が与えられる．それに対して二人が協力して $f(x, y)$ を計算するのがゲームである．その際，二人で交換する通信メッセージの総ビット数を，なるべく小さくする戦略（プロトコル）を考えたいのである．

$A, B$ は互いに計算の「戦略」を決めておく．つまり，通信の方法や計算のアルゴリズムである．これらを総称して「プロトコル」という．プロトコルではランダムネスを使ってもよい．したがって，間違える場合もあるが，誤り率を $\gamma$ 以下にしなければならないことにする．そのように誤り率 $\gamma$ 以下のプロトコルを二人で定めたときに，そのプロトコルに対して最も都合の悪い $x, y$ を与えたとき，何ビットの通信が行われたかで，そのプロトコルの通信量がきまる．関数 $f$ を計算する通信問題の <u>通信計算量</u> は（直感的には）最適なプロトコルの通信量である．正確には，関数 $f$ の通信計算量 $C_\gamma(f)$ は，次のように定義される．

**定義 3.12.** 関数 $f$ の通信計算量 $C_\gamma(f)$ に対して，ある $f$ を誤り率 $\gamma$ 以下で計算する通信計算するプロトコルで，長さ $n$ の入力列の対に対し，つねに平均で $c(n)$ ビットの通信で計算が終了するものが存在するとき「$C_\gamma(f) \leq c(n)$ である」という．逆に，そうならないとき，すなわち，誤り率 $\gamma$ 以下の通信計算するプロトコルでは，どのようなものを考えても，無限個の $n$ で，その長さで何らかの（意地悪な）入力列の対が存在して，平均通信総ビット数が $c(n)$ 以上である場合に「$C_\gamma(f) \geq c(n)$ である」という．

我々は対象の関数として，排反性 (disjointness) を調べる次の関数 $f_{\mathrm{dis}}$ を用いる．

$$f_{\mathrm{dis}}(x, y) \;=\; \begin{cases} 0, & \text{if } x_i = y_i = 1 \text{ for some } i, \\ 1, & \text{otherwise.} \end{cases}$$

この関数の通信計算量に関して以下のことが知られている [3].

**定理 3.13.** 任意の $\gamma < 0.5$ に対して，ある定数 $c > 0$ が存在して

$$C_\gamma(f_{\mathrm{dis}}) \;\geq\; cn.$$

この証明をする際に乱択アルゴリズムの動きを考えるのは少々面倒である．Yao は,「乱択」の解析を「平均」の解析を使って行ううまい方法を考えた．

乱択を使わない（通常の）アルゴリズムだけを考える．ただし，今度は最悪の入力列対を考えるのではなく，与えられた長さ $n$ の（ある列の集合の）中から入力列対を等確率[2]にとってきた場合を考える．誤り率はその場合の誤り率であり，通信コスト（総通信ビット数）は，そのときの平均の総通信ビット数である．それを使って $C_\gamma(f)$ と同様に定義したものを $D_\gamma(f)$ とする[3].

[2]実際には等確率でなくてもよいし，定理の証明 [3] ではその確率を決めるところに工夫がある．しかし，授業では簡単のため等確率として説明する．

[3]本当はもう少し正確な定義の方がよいが，ここでは概略にとどめる．講義ではより正確な定義を紹介する．

そうすると Min-Max 定理のような比較的簡単な議論で，次の関係を示すことができる．

**補題 3.14.** $C_\gamma(f) \geq D_{2\gamma}(f)/2$.

あとは，この $D_\gamma(f)$ を解析すればよい．(以下，もう少し簡単な [4] の結果を口頭で説明する．)

**$F_\infty$ の計算限界**

定理 3.13 を用いると，$F_\infty$ のストリーム計算の限界に対して，今回示したい定理を導くことができる．以下にその概略を述べる．

入力列として，排反性判定から定義されるものだけを考える．つまり，$\boldsymbol{a} = a_1, ..., a_{m_1}, b_1, ..., b_{m_2}$ という列を考える．ただし，$a_1, ..., a_{m_1}$ はすべて異なる $D$ の要素で，同様に $b_1, ..., b_{m_2}$ もすべて異なる $D$ の要素であるとする．したがって，列の長さ $m = m_1 + m_2 \leq 2N$ である．このような入力列に対し，$F_\infty = F_{\infty,m}$ の 1/3-相対近似を，誤り率 0.2 以下で計算する乱択ストリームアルゴリズム SA があったとする．この SA を用いれば，二つの集合 $A = \{a_1, ..., a_{m_1}\}$ と $B = \{b_1, ..., b_{m_2}\}$ の排反性判定が（誤り率 0.2 以下で）可能である．というのも，もし $A \cap B \neq \emptyset$ ならば $F_\infty = 2$ なので，SA が答える近似値は少なくとも $\widehat{F}_{\infty,m} > 4/3$ である．逆に，もし $A \cap B = \emptyset$ ならば $F_\infty = 1$ なので，SA が答える近似値は $\widehat{F}_{\infty,m} < 4/3$ となる．つまり，SA の出力する近似値を見て，排反性を判定できるのである．

一方，このストリームアルゴリズムの領域計算量が $o(N)$ であったとすると，時刻 $t = m_1$ での使用領域も $o(N)$ ビットである．したがって，そのときにメモリにある値を Alice が Bob に送れば，通信計算量 $o(N)$ ビットの排反性計算プロトコルとなる．これは，排反性計算の乱択通信計算量の下界に反する．

## References


[1] N. Alon, Y. Matias, and M. Szegedy, The space complexity of approximating the frequency moments, in *Proc. ACM Sympos. on Theory of Computing*, 1996, 20–29.

[2] A.C. Yao, Lower bounds by probabilistic arguments, in *Proc. IEEE Foundations of Computer Sci.*, 1983, 420–428.

[3] B. Kalyanasundaram and G. Schnitger, The probabilistic communication compexity of set intersection, in *Proc. 2nd Structure in Complexity Theory Conf.*, 1987, 41–49.

[4] L. Babai, P. Frankl, and J. Simon, Complexity classes in communication complexity theory, in *Proc. IEEE Foundations of Computer Sci.*, 1986, 337–347.