Title	定数時間アルゴリズム：性質検査と定数時間近似
Author(s)	伊藤, 大雄
Citation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.287-290.
Issue Date	2011-06
Doc URL	http://hdl.handle.net/2115/48416
Type	conference presentation
Note	ERATO 湊離散構造処理系プロジェクト春のワークショップ（キックオフシンポジウム）. 2010年5月28日（金）～29日（土）. ERATO湊プロジェクト研究室.
File Information	02.ito_06.pdf

定数時間アルゴリズム
ー性質検査と定数時間近似ー

伊藤大雄
京都大学 大学院 情報学研究科

2010.5.28
第１回 湊離散構造処理系シンポジウム

決定問題と最適化問題

- 決定問題の例：連結性判定問題
  - 入力：グラフG
  - 要請：Gが連結であるか判定せよ（YesかNoか）
- 最適化問題の例：最大マッチング問題
  - 入力：グラフG
  - 要請：Gのマッチングの最大サイズを求めよ（数値）
- 一般的要請：入力長nに対し、計算時間$T(n)$がなるべく小さいアルゴリズムを設計せよ。

定数時間アルゴリズムとは

- これまでの常識：入力をすべて見なければ話にならない。つまり $T(n)=\Omega(n)$.
- しかし、一部を見るだけで、判断できないか？
  - e.g., モンテカルロ法、サンプリング調査、料理の味見・・・
- nに無関係の定数個のデータを見るだけで判断！
  - つまり $T(n)=O(1)$.
  - 理論的保証。（ヒューリスティックスでは無い。）
  - 確率アルゴリズム。
  - ウェブグラフ、ゲノムなど超巨大情報が簡単に扱える。

性質検査(property testing)

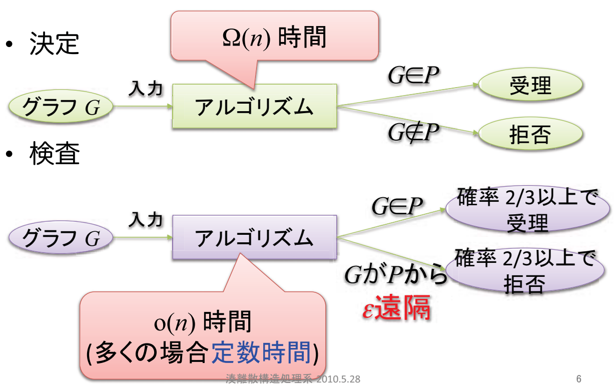
- 判定問題(YesかNoか)の緩和→性質検査(property testing)
- 例えば、n個の節点、m本の枝からなるグラフが連結であることを、n,mに無関係な定数個のデータを見るだけで検査する。
- 直感的には「不可能」。（見ていない所に枝が１本あるだけで、結果が変わる。）
- →多少の不正確さを許容することで、可能にする。

性質検査の考え方

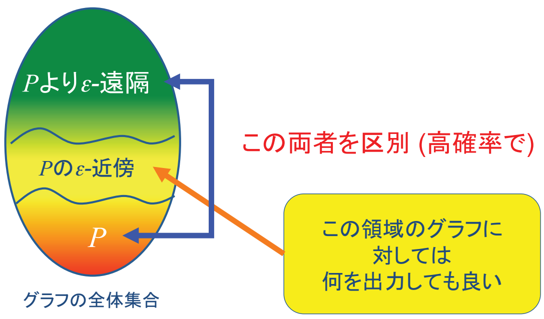
- 考え方１：大まかな区別がつけば良い

- 考え方２：確率的に判定
  - (任意の入力に対し)確率2/3以上で正解を出す。

決定アルゴリズムと検査アルゴリズム



検査アルゴリズムの目的



検査アルゴリズムの実用上の意味

- 一部を見て全体の性質を推定することは、実用上多い。
  - プログラムのチェック、視聴率調査、料理の味見。
- 大規模な対象物の性質を調べる時、事前に検査アルゴリズムを実行する。その結果、
  - ε遠隔⇒終了。
  - その他⇒（通常の）判定アルゴリズムを適用。

大幅な時間削減が可能！

グラフGと性質Pの距離

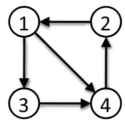
- Γ_n: 考慮するグラフの全体集合（ただしn節点）
- $G, G' \in \Gamma_n$の距離：
- dist(G, G'):=GをG'に変形するために、操作(追加または削除)しなければならない枝の割合（$\therefore 0 \le \text{dist}(G, G') \le 1$）。
- グラフGと性質Pの距離：
- $\text{dist}(G, P) := \min\{\text{dist}(G, G') \mid G' \in P\}$
- GがPよりε遠隔⇔$\text{dist}(G,P) \ge \varepsilon$.　(ただし$0<\varepsilon\le 1$)

グラフの表現法１：隣接行列モデル(Adjacency Matrix Model)

- 密なグラフに使用
- グラフは隣接行列$g(i,j)$, $i,j \in \{1,..,n\}$で表現

$$g(i,j) = \begin{cases} 1 & ((i,j) \in E) \\ 0 & ((i,j) \notin E) \end{cases} \quad (i,j \in \{1,\ldots,n\})$$

- 例：

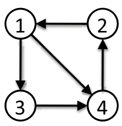


	1	2	3	4
1	0	0	1	1
2	1	0	0	0
3	0	0	0	1
4	0	1	0	0

グラフの表現法２：次数制限モデル(Bounded Degree Model)

- 疎なグラフに使用。
- Γ_nは次数上限dの制限を持つ（dは整定数）。
- グラフは節点iのp番目の隣接節点を表す行列

  $g(i,p)$, $i \in \{1,..,n\}$, $p \in \{1,..,d\}$
- で表現。該当節点が存在しないときは0を返す。
- 例：（d=2）



	1	2
1	3	4
2	1	0
3	4	0
4	2	0

次数制限モデルの下でのグラフ間の距離

- dist(G, G'):=GをG'に変形するために、操作(追加または削除)しなければならない枝の割合
- :=(GをG'に変形するために、操作しなければならない枝の数)/(dn/2)

- Gが性質Pからε遠隔⇔Gを（次数上限を満たしつつ）Pを満足するようにするために少なくとも$\varepsilon dn/2$本の枝の操作が必要。

連結性の検査

定理1: 次数制限モデルの下で、与えられたグラフGが連結であることを検査する, i.e.,

- Gが連結ならば確率$2/3$以上で受理し、
- 連結からε遠隔ならば、確率$2/3$以上で拒否する

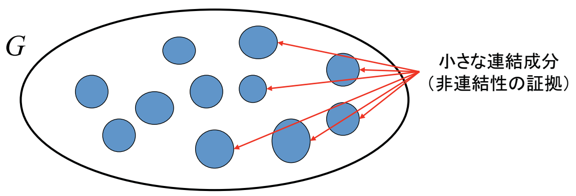
$O(1/\varepsilon^2 d)$時間アルゴリズムが存在する。

連結からε遠隔であるグラフの性質

- 補題2: 連結からε遠隔なグラフには、少なくとも$\varepsilon dn/6$個の連結成分が存在する。□
- 補題3: 連結からε遠隔なグラフには、サイズが高々$12/\varepsilon d$の連結成分が少なくとも$\varepsilon dn/12$個存在する。（証明：補題2+連結成分は互い素であること。）□

非連結性の証拠

- 補題3より、ε遠隔なグラフには、小さな$(\leq 12/\varepsilon d)$連結成分が多数$(\geq \varepsilon dn/12)$存在する。
- そのうちの一つでも見つけることができれば、Gが非連結であることが判明する。



連結成分を如何に見つけるか

- 無作為に節点vを選べば、
- $\Pr(v \in$小さな連結成分$) \geq (\varepsilon dn/12)/n = \varepsilon d/12$
- 無作為に $24/\varepsilon d$ 個節点を選べば、少なくとも一つが小さな連結成分に入っている確率は高い。
- 節点の回りを探索することで、その節点が小さな$(\leq 12/\varepsilon d)$連結成分に入っているか否かを、定数時間$(O(1/\varepsilon d) \cdot d = O(1/\varepsilon))$で確認できる。

$$1-(1-\varepsilon d/12)^{24/\varepsilon d} > 1-(e^{-\varepsilon d/12})^{24/\varepsilon d} = 1-e^{-2} > 6/7$$

連結性検査アルゴリズム

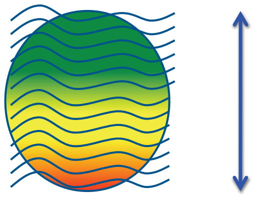
- 無作為に $24/\varepsilon d$ 個節点を選ぶ。
- 選ばれた各節点vのまわりを (BFS等で)$12/\varepsilon d+1$歩まで探索する。
- どれか一つのvでも、上記の探索を終えることができなければ、「拒否」を出力し、終了。そうでなければ「受理」を出力し終了。
- 計算時間：一つのBFSは$O(d/\varepsilon d)=O(1/\varepsilon)$時間でできる。従ってアルゴリズム全体で$O(1/\varepsilon^2 d)$時間。
- 注意：連結なグラフは常に(=確率1で)受理される。（1-sided error）

定数(準線形)時間近似

- 最適化問題の検査版（近似解を算出）
- 最小化問題の(α，εn)近似とは、最適解の値をOPTとすると
- OPT ≦ ALG ≦ αOPT+εn
- である解ALGを求める。
- 最大化問題なら
- OPT/α-εn ≦ ALG ≦ OPT

なぜ定数時間で近似できるのか？

- 解の範囲が{1,...,n}の場合、それらをすべて区別するためにはΩ(log n)の計算が必要。
- しかしεnの絶対誤差を許せば、解の範囲をεn毎に刻むことになり、n/εn=1/ε通りの区別ができれば良い。→定数時間でできる可能性。



グラフの全体集合

定数時間近似の枠組み

- 関数 $f(v)$に対して $f(G)=\sum_{v\in V} f(v)$ を計算したい。
- （例：枝の数 $m=\sum_{v\in V}\frac{d_v}{2}$）

- 頂点を一様ランダムにs個サンプリング→S
- $\forall v\in S$ に対して $f(v)$を計算。
- $\tilde f(G)=\frac{n}{s}\sum_{v\in S} f(v)$ を出力。

- $0\le|f(v)|\le w$であれば $s=O(w^2/\varepsilon^2)$ とすると、高い確率で $|f(G)-\tilde f(G)|\le \varepsilon n$ となる。
- （Hoeffdingの不等式より）

Hoeffding's Inequality

- $X_1, \ldots, X_N$:
  - indep. rand. vars.,
  - $\Pr(X_i - E[X_i]\in[a_i,b_i])=1$ for $1\le i\le N$, i.e., a. s. bound.
  - Let $S=X_1+\ldots+X_N$.
- Then for any $t>0$ we have

$$\Pr(|S-\mathrm{E}[S]|\ge t)\le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^N (b_i-a_i)^2}\right).$$

- If $t=\varepsilon N, N=cd^2/2\varepsilon^2$, then

$$\Pr(|S-\mathrm{E}[S]|\ge N\varepsilon)\le 2\exp\left(-\frac{2\varepsilon^2N^2}{Nd^2}\right)=2\exp(-c).$$

定数時間近似の枠組み

- 頂点を一様ランダムにs個サンプリング→S
- $\forall v\in S$ に対して $f(v)$を計算。
- $\tilde f(G)=\frac{n}{s}\sum_{v\in S} f(v)$ を出力。

- この枠組み自体は定数時間で作動する。
- よって、もし $f(v)$が高い確率で定数時間で十分な精度で計算可能ならば、全体として定数時間で近似できる。
- ローカルサーチ、貪欲算法、枝刈り…解析法

我々の主な成果

- 性質検査
  - k-点連結性検査：$O\left(d\left(\frac{ck}{\varepsilon d}\right)^k \log\left(\frac{k}{\varepsilon d}\right)\right)$ 時間 ICALP 2008
  - グラフの外平面性検査：$\tilde O\left(\frac{1}{\varepsilon^{13}d^6}+\frac{d}{\varepsilon^2}\right)$ 時間 [submitted]
  - 3辺彩色、3次元マッチング、有向・無向ハミルトン(閉)路、Schaeferの3SATがΩ(n)時間必要。[IEICE2010]
- 定数時間近似
  - 最小頂点被覆、最大マッチングの(2, ε)近似：$O(d^4/\varepsilon^2)$時間 STOC 2009
  - 最大マッチングの(1, ε)近似：$d^{O(1/\varepsilon^2)}(1/\varepsilon)^{O(1/\varepsilon)}$ 時間

まとめ

- データの一部（定数個）しか見ずに対象の性質を検査したり、解を近似したりする手法がある。
- 確率的な解答（確率2/3以上で正答する）と「εx（解の範囲）」の絶対誤差を許すことで実現している（ε>0は任意の定数）。
- ウェブグラフやゲノム情報などの大規模なデータを扱う際に有効。
- 近年のホットな研究テーマであり、様々な性質や関数が検査・近似可能であることが分かってきている。