テーマ２：平均時計算量の解析

本日の講義の目的

- NP 困難問題の例
- 平均時計算量解析の枠組み

2.1. NP-困難な問題の例

今回のテーマでは充足性判定・充足解計算問題の形で表される NP-問題を対象とする．たとえば，次の 3SAT 問題が最も有名な例だろう．

3SAT 問題
入力： n 変数の 3-積和標準形 命題論理式 $F(X_1, ..., X_n)$.
仕事： F が 充足可能 か否かの判定.
補足： 以下では積和標準形を CNF (Conjunctive Normal Form) と略記する．

この問題にはいろいろな変種を考えることができる．ただし，そのすべてが NP-完全（もしくは NP-困難）であるとは限らない．たとえば，論理式として，3-積和標準形ではなく，2-積和標準形を考えた場合，対応する充足性判定問題は（項の数の）線形時間時間で解くことができる．また，次のような XORSAT 型の問題も（結局は線形連立方程式の解を求める問題なので）多項式時間に計算可能である．

kXORSAT 問題（判定版，解探索版）
入力： kXOR-CNF 論理式 F（変数数を n とする）.
仕事： F が充足可能か否かの判定（もしくは充足解の計算）.

しかしながら，問題を少し変えて，次のような最適解探索問題（２種類ある）を考えると，そのどちらもが NP-困難になってしまうのである．

MAX-2XORSAT 問題
入力： 2XOR-CNF 論理式 F（変数数を n とする）.
仕事： 最も多くの節 clause を充足させる解の計算.

MINW-kXORSAT 問題
入力： kXOR-CNF 論理式 F（変数数を n とする）.
仕事： 充足解のうちで 1 の数が最も少ない解の計算.
注）$k = 3$ で NP-困難.（$k = 2$ でも NP-困難だと思います.）

これらの NP-困難性の証明には，たとえば，次のような（多少人工的な）充足性判定問題が用いられる．

3NAESAT 問題
入力： n 変数の 3-CNF 論理式 F（変数数を n とする）．
仕事： F が Not-All-Equal の意味で充足可能か否かの判定．

これを使うと次のようなグラフの問題の NP-困難性も示すことができる．後の議論で出てくる関係上，これらについても解説しておく．

MAXCUT 問題（判定版）
入力： 無向グラフ $G = (V, E)$ と自然数数 K（グラフの頂点数を n とする）．
仕事： G の 頂点分割 (V_1, V_2) で，カット辺数 が K 以上のものが存在するかの判定．

Bisection 問題（判定版）
入力： 無向グラフ $G = (V, E)$ と自然数数 K（グラフの頂点数を n とする）．
仕事： G の 等数頂点分割 (V_1, V_2) で，カット辺数が K 以下のものが存在するかの判定．

2.2. 平均時計算量解析の枠組み

3SAT 問題は NP-困難であり，一般的に解くのは難しいかもしれない．しかし，一方で様々なヒューリスティクス[1]やアルゴリズムが提案されている．その中には「平均的には」うまく動くと言われているものもあるし，実際に証明されているものもある．こうしたヒューリスティクス的アルゴリズムの平均的な振る舞いの解析が今回のテーマである．

もちろん，アルゴリズムの平均的な振る舞いを解析する技法の紹介が目標だが，もう一つ，重要なポイントがある．それは「平均」を議論するうまい枠組み，それにそれを作り出した功績である．今回は，この枠組みの話しをしよう．

まず，3SAT 問題に対しては，命題論理式をランダムに生成したときに，それを解くヒューリスティクスの研究が人工知能系の研究者を中心に行われ，理論的にも論文 [1] のような解析が示された．その後，3SAT の phase transition 現象（閾値現象，臨界現象？）が，物理学系の研究者の注目を集め，この実験的解析のためにも「ランダム 3SAT 例題」に対するヒューリスティクスの提案や，その理論解析が行われ，たとえば論文 [2] のような結果が得られている．

それに対して，まったく異なった文脈で，やはり物理系の研究者が中心となって最適問題型の充足解計算問題に対する平均時計算が議論されるようになった．そのために導入として，符号理論の話と，それに関係する NP-困難な問題（ほぼ MINW-kXORSAT と同様）について説明する．

[1]アルゴリズム理論の分野では，正しく動くことが証明されていないもの，もしくはその性能が実験的にしか示されていないものを，すべて「ヒューリスティクス」と呼ぶ場合が多い．

【用語解説：低密度パリティ検査行列を用いた線形符号】

線形符号の一つに低密度パリティ検査行列 Low Density Parity Check matrix (LDPC) を用いた符号がある．これに関する用語を簡単にまとめておく．

- 線形符号で符号語をチェックする行列を パリティ検査行列 という．以下ではパリティ検査行列として，$n \times m$ の 0,1-行列のみを考える（一般には $n > m$）．
- パリティ検査行列を以下では H で表す．この行列に対し，符号語は $H\boldsymbol{a} = \mathbf{0}$ となる $\boldsymbol{a}$ の集合である．
- 受信した語（元々は符号語）にはノイズが含まれている可能性が高い．ここでは，ノイズベクトルを $\boldsymbol{n}$ と表し，送信された語が送った符号語 $\boldsymbol{a}$ に対し，$\boldsymbol{b} = \boldsymbol{a} + \boldsymbol{n}$ のようになってしまった場合を想定する．
- 受信した $\boldsymbol{b} = \boldsymbol{a} + \boldsymbol{n}$ に対しては，

$$H\boldsymbol{b} \;=\; H(\boldsymbol{a}+\boldsymbol{n}) \;=\; H\boldsymbol{a} + H\boldsymbol{n} \;=\; \mathbf{0} + H\boldsymbol{n} \;=\; H\boldsymbol{n}$$

が成り立つ．$H\boldsymbol{b}$ のことを シンドローム という．したがって，目標は，このシンドローム $H\boldsymbol{b}$ からノイズベクトル $\boldsymbol{n}$ を求めることになる．

LDPC 符号復号問題（$(3,4)$-疎行列の場合）
入力： $(3,4)$-疎な $n \times m$ 行列 H とシンドローム $\boldsymbol{c} \in \{0,1\}^m$.
仕事： $H\boldsymbol{n} = \boldsymbol{c}$ となるようなノイズベクトル $\boldsymbol{n}$ で，1 の立っている成分数が最も少ないもの．
注）$H\boldsymbol{a} = \boldsymbol{c}$ となる $\boldsymbol{a}$ を求めるだけなら難しくない．

この問題も NP-困難である．しかし，Gallagar [3] が考案したアルゴリズム（正確にはヒューリスティクス[2]）は，多くの場合非常に効率良く働く．そのことを統計力学の研究者たちが再発見・再評価したため，大きな注目を集めた [4].

そこでの議論の枠組みが，ここでは重要だ．彼らは，パリティ検査行列 H はランダムに与えられるとし，さらにノイズベクトル $\boldsymbol{n}$ も，ある確率分布でランダムに与えられるものとした．具体的には，各成分ごとに独立に，低い確率 p で 1，確率 $1-p$ で 0 という値を取るベクトルをノイズベクトル $\boldsymbol{n}$ とし，この H と $\boldsymbol{n}$ から得られる $(H, \boldsymbol{c})$ が入力である，としたのである．これにより入力例 $(H, \boldsymbol{c})$ に対する確率分布が定義できる．その分布のもとで「平均的な」効率や性能を議論したのだ．しかも，その目標を「最もらしい $\boldsymbol{n}$ を求めること」にすり替えたのである．また，最適解探索問題が，問題例を作る際に使った解を発見する問題「埋め込み解探索問題」にすり替わった点も非常に重要である．

つまり，あるパラメータの範囲内では，「最もらしい解を求めること」と「解埋め込み型問題の埋め込んだ解を求めること」が同値なので，目標を「最もらしい解を求めること」と言うことで多くの人々が納得のいくストーリーを作ることができたのである．

[2] このヒューリスティクスは，人工知能の分野で独立に発見され（より一般的な枠組みで）研究されていたものであることがわかった [4]．そこで人工知能で呼ばれていた名を利用して，確率伝播法 (Belief Propagation) という名称を使うのが一般的である．

この枠組みを一般化すると次のような最尤推定問題となる．

生成方式 G に対する最尤推定問題
入力： シード $s \in \{0,1\}^n$ と一様ランダム列 $w \in \{0,1\}^r$ から作られた入力例 $x = G(s,w)$.
仕事： $\Pr_{w:\{0,1\}^r}[\, x = G(s,w)\,]$ を最大にするような s を求める．

LDPC 符号復号問題では，シードは特に無く，ノイズベクトル $\boldsymbol{n}$ がランダム列 w に対応するため，生成モデル G は次のように定義できる（実際には，H もランダムに生成することを考えるのだが，ここでは簡単のためノイズだけのランダムネスで議論する）．

ランダムな $(3,4)$-疎行列 H と ランダムノイズ列 $\boldsymbol{n}$ に対し，
$G_{\mathrm{LDPC}}(\boldsymbol{n}; H) \;=\; \boldsymbol{c}\ (= H\boldsymbol{n})$
ただし $\boldsymbol{n}$ の各ビットは独立に確率 p で 1 となるものとする．

ここで重要となるのは，最適解が実は埋め込んだ解である，という点だ．つまり次の性質である．

定理 2.1. 上の枠組みのもとで与えらえれたシンドローム $\boldsymbol{c}$（とパリティ検査行列 H）に対して，$H\boldsymbol{n} = \boldsymbol{c}$ となるようなノイズベクトル $\boldsymbol{n}$ で，1 の立っている成分数が最も少ないものが最尤解であり，それは高い確率で $\boldsymbol{c}$ を作るときに使った解である．

Bisection 問題も同様の枠組みで考えることができる．この場合には，シードは頂点集合 V の二等分割 (V_1, V_2) である．一方，グラフを生成する際には，パラメータ $0 < r < p < 1$ を用い，同じ分割の頂点同士は確率 p で辺を結び，異なる分割に所属する頂点同士は確率 r で辺を結ぶようにする．このような生成方式を（一様ランダム列 w を用いて）実現する関数を $G_{\mathrm{BIS}(p,r)}((V_1,V_2),w)$ とする．そうすると，この生成方式に関する最尤解計算問題は，次の意味で Bisection 問題と同値になる．

定理 2.2. 与えられた頂点集合 V の二等分割 (V_1, V_2) に対して，一様ランダムな $w \in \{0,1\}^r$ を用いてグラフ $G = G_{\mathrm{BIS}(p,r)}((V_1,V_2),w)$ を生成したとき，カット辺数が最小の分割と方式 $G_{\mathrm{BIS}(p,r)}$ で G を最も高い確率で生成する分割（これを最尤分割と呼ぼう）が，高い確率で一致する．さらに，$p > 3r$ で $p \geq \Omega(\log n/n)$ のときには，(V_1, V_2) が最尤分割である確率が高い．

参考文献

[1] E. Koutsoupias and C.H. Papadimitriou, On the greedy algorithm for satisfiability,

[2] D. Achlioptas, Lower bounds for random 3-SAT via differential equations, Theoretical Computer Science 265, 159–185, 2001.

[3] R.G. Gallager, Low density parity check codes, *IRE Trans. Inform. Theory*, **IT-8**(21), 21–28, 1962.

[4] D. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory*, **IT-45**(2), 399–431, 1999.

テーマ＃２での課題（〆切：以後のレポートの〆切りはすべて７月２７日（金）４時とします）

次のどれかを選択してレポートを作成して欲しい.（3 については来週詳しく説明します. また追加の課題を出すかもしれません.）

1. 定理 2.1 を証明せよ．この定理は正確には次のような定理である.（ヒント（？）：そう簡単ではない．ランダムな H と $\boldsymbol{n}_*$ に対して，$\boldsymbol{c}$ となるような別のノイズベクトルの重みが（確率的に）どうなるかを議論すればよいと思う．論文 [4] がその手助けになるだろう.）

定理 2.1 与えれれた n に対して，一様ランダムに $(3,4)$-疎行列 H を生成することを仮定する．また（あらかじめ決められている）パラメータ $p << 1$ に対して，ノイズベクトル $\boldsymbol{n}_* \in \{0,1\}^n$ を各ビットごと独立に確率 p で 1 に，$1-p$ で 0 となるように生成する．こうしたランダムモデルのもとで

$$\Pr_{H,\boldsymbol{n}_*}[\ \boldsymbol{n}_* \text{ が } \boldsymbol{c} = G(\boldsymbol{n}_*; H) \text{ のただ一つの最尤解である }]$$

という確率が n を大きくしたときに 0 に収束する.

2. 定理 2.2 を証明せよ.（ヒント：これも正確に定理を書くことから始めよう．証明はそう難しくない．グラフの expander 性などの解析手法を使うと証明できる.）

3. LDPC 問題も，Bisection 問題も，mod 2 上の線形連立方程式のある種の解を求める問題として定式化できる．そのような問題に対しては，先週説明したような局所探索法や来週説明するようなメッセージ伝播法（確率伝播法もその一つ）が平均的に有効に働くことが知られている．どのような場合に，どのように働くかを実験し，その結果・解析・考察を述べよ.

付録：NP-困難性の証明

定理 2.3. $\mathrm{3NAESAT} \in \mathrm{P} \Longrightarrow \mathrm{3SAT} \in \mathrm{P}$.

証明： 3SAT 問題の入力例 F に対し，次の関係を満たす 3NAESAT 問題の入力例 F' を構成することができる.

$$F \text{ が充足可能} \iff F' \text{ が NAE-充足可能}$$

具体的には，各節 $C_j = (X_1 \vee X_2 \vee X_3)$ を,

$$D_j = (X_1 \vee X_2 \vee Y_j) \wedge (\overline{Y}_j \vee X_3 \vee Z)$$

と変換すればよい．ただし，Y_j は節 C_j の変換のためだけに導入された変数．それに対し Z はすべての節で共通に使うために導入された変数．こうすると，C_j が充足可能 $\Longleftrightarrow$ D_j が NAE-充足可能，が成り立つ.（注：ポイントは，ある式 G の NAE-充足解の真偽を反転させたものも，やはり G の NAE-充足解である，という点.） □

定理 2.4. (1) 判定版 MAXCUT $\in$ P $\Longrightarrow$ 3NAESAT $\in$ P.
(2) 判定版 BIS $\in$ P $\Longrightarrow$ 3NAESAT $\in$ P.

証明： 3NAESAT 問題に対する問題例（すなわち 3CNF-論理式）F から，MAXCUT 問題に対する問題例（すなわちグラフ G_F とカット下界 K）を構成し，それが次の関係を満たしていることを示せばよい.

$$F \text{ が NAE-充足可能} \iff G_F \text{ が } K \text{ 以上のカットを持つ}$$

（黒板で説明）