

テーマ 4 : 平均時計算量の解析 (その 2)

本日の講義の目的

- NP 困難問題の例
- 最尤推定問題 : 平均的時計算量解析の枠組み

2.1. NP 困難な問題の例

今回は, 3SAT 問題の NP-困難性を紹介した. 結果だけを復習すると次の定理である.

定理 2.1. 3SAT 問題を判定する多項式時間アルゴリズムが存在するならば, すべての NP 問題に対し, それを判定する多項式時間アルゴリズムを作ることができる (つまり, $3SAT \in P \implies P = NP$ である.)

今回は, この 3SAT 問題をもとに, いくつかの問題の NP 困難性を示そう.

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.

定理 2.2. $3NAESAT \in P \implies 3SAT \in P$.

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,

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

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 $\iff D_j$ 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.) $\square$

MAXCUT 問題 (判定版 MAXCUT)

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 BIS)

入力： 無向グラフ $G = (V, E)$ と自然数 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 .

定理 2.3. (1) 判定版 MAXCUT $\in P \implies 3\text{NAESAT} \in P$.

(2) 判定版 BIS $\in P \implies 3\text{NAESAT} \in P$.

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)

□

2.2. 平均時計算量の新たな枠組み

以上のような NP-困難な問題を完全に & 一般的に解く効率よいアルゴリズムは期待できない . しかし、だからと言って手をこまねいている訳にはいかない . そこで最悪時ではなく、平均時での計算量を考えることにしよう . 実際、先週紹介したように、3SAT 問題のように、様々なヒューリスティクス (たとえば各種局所探索法) を用いると、「大抵の場合すんなり解ける」という場合も少なくない .

では、どのような入力分布を考えればよいのか？ 等確率分布がよいのか？ ここでは物理 (統計力学) で発達してきた解析の枠組みにヒントを得た「平均」の考え方を紹介しよう . そのために、導入として、もう一つ NP-困難な問題を考える .

【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 $Ha = 0$ となる 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 n と表し，送信された語が送った符号語 a に対し， $b = a + n$ のようになってしまった場合を想定する．
- 受信した $b = a + n$ に対しては，

$$Hb = H(a + n) = Ha + Hn = 0 + Hn = Hn$$

が成り立つ． Hb 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 Hb からノイズベクトル 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 $c \in \{0, 1\}^m$ ．

仕事： $Hn = c$ となるようなノイズベクトル n で，1 の立っている成分数が最も少ないもの．

注意： $Ha = c$ となる 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，Gallager [1] が考案したアルゴリズム（正確にはヒューリスティクス¹）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 [3]．

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 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 n とし，この H と n から得られる (H, c) が入力である，としたのである．これにより入力例 (H, 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である．しかも，その目標を「最もらしい n を求めること」にすり替えたのである．

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，シードはノイズベクトル n で， $G(n, w)$ は次のように定義できる²．

¹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 [2]．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．

² G が未定義になる確率が非常に高く格好が悪いが，一応，上の枠組みには収まっているので，ここではご容赦願いたい．

$$G_{\text{LDPC}}(n, w) = \begin{cases} (H, c), & \text{もし } w \text{ が } (3, 4)\text{-疎行列 } H \text{ と } n \text{ を生成するとき,} \\ \text{未定義,} & \text{その他のとき.} \end{cases}$$

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 $G_{\text{BIS}(p,r)}((V_1, V_2), w)$ とする．そうすると，この生成方式に関する最尤推定問題は，次の意味で Bisection 問題と同値になる．

事実 1. 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 $G = G_{\text{BIS}(p,r)}((V_1, V_2), w)$ を生成したとき，カット辺数が最小の分割と方式 $G_{\text{BIS}(p,r)}$ 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] R.G. Gallager, Low density parity check codes, *IRE Trans. Inform. Theory*, **IT-8**(21), 21–28, 1962.
- [2] D. MacKay, Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory*, **IT-45**(2), 399–431, 1999.
- [3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., 1988.

テーマ # 4 での課題案（すべてのレポートの〆切は 8 月 4 日（水）とする予定）

次のどちらかを選択してレポートを作成して欲しい（来週，もう一つ，かなり数学的な課題も提案します）．

- 1 . このテーマで議論した例のように，最悪時には難しいかもしれないが，ある種の入力分布を仮定したときには「平均的に」効率良く解ける問題（あるいはそのように言われている問題）を紹介せよ．問題，分布，そして効率良いと言われている解法（アルゴリズムまたはヒューリスティクス）についての解説である．
- 2 . 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．

参考：データならびにサンプルプログラムは

<http://www.is.titech.ac.jp/~watanabe/class/maxlin/> から．