

離散数理工学 第 12 回

離散確率論：乱択データ構造とアルゴリズム (発展)

岡本 吉央
okamotoy@uec.ac.jp

電気通信大学

2023 年 1 月 17 日

最終更新：2023 年 1 月 7 日 14:00

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

1 / 30

確率の推定：単純なアルゴリズム

目次

- 1 確率の推定：単純なアルゴリズム
- 2 確率の推定：中央値トリック
- 3 今日のまとめ

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

3 / 30

確率の推定：単純なアルゴリズム

不公平な硬貨

設定

- ▶ 考えている硬貨について

$$\Pr(\text{表}) = p$$

ただし, $0 \leq p \leq 1$

- ▶ **目標**: p を知りたい

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

5 / 30

確率の推定：単純なアルゴリズム

単純なアルゴリズム

単純なアルゴリズム

硬貨を何度も投げてみる

- ▶ n 回投げるとする (独立な試行)
- ▶ 確率変数 X_i を次で定義 ($i \in \{1, \dots, n\}$)

$$X_i = \begin{cases} 0 & (i \text{ 回目に投げたとき裏が出る}) \\ 1 & (i \text{ 回目に投げたとき表が出る}) \end{cases}$$

- ▶ 次の量を出力

$$\frac{X_1 + X_2 + \dots + X_n}{n}$$

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

7 / 30

今日の目標

今日の目標

典型的な乱択アルゴリズムの設計と解析ができるようになる

- ▶ 確率の推定 (中央値トリック)

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

2 / 30

確率の推定：単純なアルゴリズム

不公平な硬貨

設定

- ▶ 硬貨が 1 つある
- ▶ 投げたとき, 表が出る確率はいつも変わらない
- ▶ その確率が分からない
- ▶ **目標**: 表が出る確率を知りたい
- ▶ 可能な操作: 硬貨を投げる (このみ)

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

4 / 30

確率の推定：単純なアルゴリズム

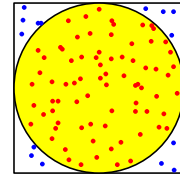
応用例：モンテカルロ法

モンテカルロ法：(実際には用いられない) 例

円周率の計算のために次を行う

- 1 $[-1, 1]^2$ 内の点 (x, y) を一様分布に従って発生させる
 - 2 $x^2 + y^2 \leq 1$ ならば, 1 を出力, そうでなければ 0 を出力
- このとき, この方法が 1 を出力する確率 = $\pi/4$

つまり, $p = \pi/4$ とした不公平な硬貨を考えていることになる



モンテカルロ法は, 次の「単純なアルゴリズム」を実行する

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

6 / 30

確率の推定：単純なアルゴリズム

期待値の解析

- ▶ 出力の期待値は

$$\begin{aligned} E \left[\frac{X_1 + X_2 + \dots + X_n}{n} \right] &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} \sum_{i=1}^n (0 \cdot (1-p) + 1 \cdot p) \\ &= p \end{aligned}$$

期待値は正しい「推測」になっている (不偏推定)

問題点

必ず「 p 」を出力するわけではない \rightsquigarrow 誤差が出る

n を大きくすれば, 誤差は小さくなりそう

岡本 吉央 (電通大)

離散数理工学 (12)

2023 年 1 月 17 日

8 / 30

誤差の解析 (1)

以後、 $X = X_1 + X_2 + \dots + X_n$ とする

- ▶ 真の値 p から出力 $\frac{X}{n}$ がどれだけずれるか？
- ▶ そのずれが ε 未満である確率を知りたい
- ▶ その確率は次のように書ける

$$\Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right)$$

- ▶ 計算

$$\begin{aligned}\Pr\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) &= 1 - \Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right), \\ \Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) &\leq \frac{\mathbb{E}\left[\left|\frac{X}{n} - p\right|\right]}{\varepsilon} \quad (\text{マルコフの不等式})\end{aligned}$$

しかし、 $\mathbb{E}\left[\left|\frac{X}{n} - p\right|\right]$ はどう計算したらいいかわからない

誤差の解析 (3)

$$\mathbb{E}\left[\left|\frac{X}{n} - p\right|^2\right] = \mathbb{E}\left[\left(\frac{X}{n}\right)^2 - 2p\frac{X}{n} + p^2\right] = \frac{1}{n^2}\mathbb{E}[X^2] - \frac{2p}{n}\mathbb{E}[X] + p^2$$

ここで、

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_1 + \dots + X_n] = \sum_{i=1}^n \mathbb{E}[X_i] = pn \\ \mathbb{E}[X^2] &= \mathbb{E}[(X_1 + \dots + X_n)^2] = \sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{i=1}^n \sum_{j=1, (i \neq j)}^n \mathbb{E}[X_i X_j]\end{aligned}$$

誤差の解析 (5)

ここまで、まとめると

$$\begin{aligned}\mathbb{E}\left[\left|\frac{X}{n} - p\right|^2\right] &= \frac{1}{n^2}\mathbb{E}[X^2] - \frac{2p}{n}\mathbb{E}[X] + p^2 \\ &= \frac{1}{n^2}(pn + p^2n(n-1)) - \frac{2p}{n}pn + p^2 \\ &= \frac{p}{n} + \frac{p^2(n-1)}{n} - p^2 \\ &= \frac{p-p^2}{n} \\ &= \frac{p(1-p)}{n}\end{aligned}$$

疑問

単純なアルゴリズム

硬貨を何度も投げてみる

- ▶ n 回投げるとする (独立な試行)
- ▶ 確率変数 X_i を次で定義 ($i \in \{1, \dots, n\}$) (標示確率変数)

$$X_i = \begin{cases} 1 & (i \text{ 回目に投げたとき表が出る}) \\ 0 & (i \text{ 回目に投げたとき裏が出る}) \end{cases}$$

- ▶ 次の量を出力

$$\frac{X_1 + X_2 + \dots + X_n}{n}$$

疑問

この「単純なアルゴリズム」よりもよいアルゴリズムは無いのか？

誤差の解析 (2)

$$\Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) = \Pr\left(\left|\frac{X}{n} - p\right|^2 \geq \varepsilon^2\right) \leq \frac{\mathbb{E}\left[\left|\frac{X}{n} - p\right|^2\right]}{\varepsilon^2}$$

$\mathbb{E}\left[\left|\frac{X}{n} - p\right|^2\right]$ を計算してみる

誤差の解析 (4)

任意の $i \in \{1, \dots, n\}$ に対して

$$\mathbb{E}[X_i^2] = (1-p) \cdot 0^2 + p \cdot 1^2 = p$$

したがって、

$$\sum_{i=1}^n \mathbb{E}[X_i^2] = pn$$

任意の異なる $i, j \in \{1, \dots, n\}$ に対して、 X_i と X_j は独立なので、

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j] = p \cdot p = p^2$$

したがって、

$$\sum_{i=1}^n \sum_{j=1, (i \neq j)}^n \mathbb{E}[X_i X_j] = p^2 n(n-1)$$

誤差の解析 (6)

すなわち、

$$\Pr\left(\left|\frac{X}{n} - p\right| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[\left|\frac{X}{n} - p\right|^2\right]}{\varepsilon^2} = \frac{1}{\varepsilon^2} \frac{p(1-p)}{n}$$

- ▶ この不等式は**チェビシェフの不等式**そのもの
- ▶ この右辺を δ 以下にするには、 $n \geq \frac{1}{\varepsilon^2} \frac{p(1-p)}{\delta}$ とすればよい

結論

誤差が ε 以上になる確率を δ 以下とするためには、

$$n \geq \frac{1}{\varepsilon^2} \frac{p(1-p)}{\delta}$$

とすればよい

目次

- 1 確率の推定：単純なアルゴリズム
- 2 確率の推定：中央値トリック
- 3 今日のまとめ

中央値トリック (median trick)

中央値トリック

- ▶ n 回投げるとする (独立な試行)
- ▶ $n = (2k - 1)t$ とする (k, t は自然数)
- ▶ 確率変数 X_i を次で定義 ($i \in \{1, \dots, n\}$)

$$X_i = \begin{cases} 0 & (i \text{ 回目に投げたとき裏が出る}) \\ 1 & (i \text{ 回目に投げたとき表が出る}) \end{cases}$$

- ▶ 確率変数 Y_j を次で定義 ($j \in \{1, \dots, 2k - 1\}$)

$$Y_j = \frac{X_{(j-1)t+1} + \dots + X_{(j-1)t+t}}{t}$$

- ▶ 次の量を出力

$$Y = \text{med}\{Y_1, \dots, Y_{2k-1}\}$$

med は中央値: $\text{med}\{5, 1, 6, 2, 4\} = 4$

中央値トリック：誤差の解析 (1)

- ▶ 次が成り立つために n が満たす条件を見つけたい

$$\Pr(|Y - p| \geq \varepsilon) \leq \delta$$

- ▶ 今までの議論から, 任意の $j \in \{1, \dots, 2k - 1\}$ に対して

$$\Pr(|Y_j - p| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \frac{p(1-p)}{t}$$

- ▶ $t \geq \frac{8p(1-p)}{\varepsilon^2}$ とすると, $\frac{1}{t} \leq \frac{1}{8} \frac{\varepsilon^2}{p(1-p)}$ なので,

$$\Pr(|Y_j - p| \geq \varepsilon) \leq \frac{1}{8}$$

中央値トリック：誤差の解析 (2)

- ▶ このとき, 合併上界から

$$\Pr(k \text{ 個の } j \text{ に対して, } |Y_j - p| \geq \varepsilon) \leq \binom{2k-1}{k} \left(\frac{1}{8}\right)^k$$

性質：合併上界

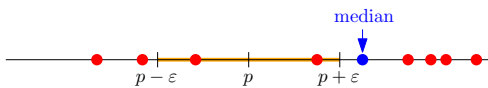
任意の事象 A, B に対して

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$

中央値トリック：誤差の解析 (4)

- ▶ したがって (演習問題),

$$\Pr(|Y - p| \geq \varepsilon) \leq \Pr(k \text{ 個の } j \text{ に対して, } |Y_j - p| \geq \varepsilon) < \left(\frac{3}{4}\right)^k$$



- ▶ $k \geq \log_{3/4} \delta$ とすると

$$\Pr(|Y - p| \geq \varepsilon) < \left(\frac{3}{4}\right)^k \leq \delta$$

実験してみた

- ▶ パラメータ

- ▶ $p = \pi/4 \approx 0.785398$
- ▶ $\varepsilon = 0.1$
- ▶ $\delta = 10^{-3}, 10^{-11}, 10^{-19}$
- ▶ $t = \left\lceil \frac{8p(1-p)}{\varepsilon^2} \right\rceil$
- ▶ $k = \left\lceil \log_{3/4} \delta \right\rceil$
- ▶ $n = (2k - 1)t$

- ▶ 各アルゴリズムを 100 回実行

- ▶ 横軸が単純なアルゴリズム, 縦軸が中央値トリックの結果 (同じ乱数による)

注意：このパラメータ設定はとても恣意的なので, 他のパラメータ設定で追試してみるとよい

中央値トリック：誤差の解析 (3)

- ▶ 二項係数に対する上界を用いて, 右辺を整理すると

$$\binom{2k-1}{k} \left(\frac{1}{8}\right)^k \leq \left(\frac{e(2k-1)}{k}\right)^k \left(\frac{1}{8}\right)^k < \left(\frac{2e}{8}\right)^k < \left(\frac{3}{4}\right)^k$$

二項係数：簡単な評価

(第 1 回講義より)

任意の自然数 $a \geq 1$ と任意の自然数 $b \geq 1$ に対して, $a \geq b$ であるとき,

$$\left(\frac{a}{b}\right)^b \leq \binom{a}{b} \leq \left(\frac{ea}{b}\right)^b$$

中央値トリック：誤差の解析 (まとめ)

- ▶ $t \geq \frac{8p(1-p)}{\varepsilon^2}$, $k \geq \log_{3/4} \delta$ とすると
誤差が ε 以上になる確率を δ 以下にできる
- ▶ このとき, 硬貨を投げる回数 n は

$$n = (2k - 1)t \geq \Omega\left(\frac{p(1-p)}{\varepsilon^2} \log \frac{1}{\delta}\right)$$

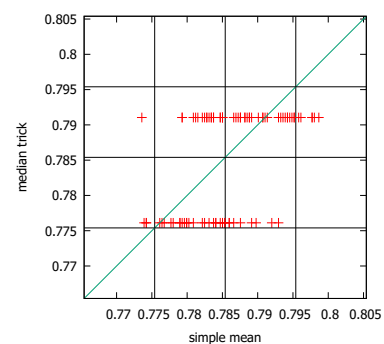
補足：単純なアルゴリズムにて, 硬貨を投げる回数 n は

$$n \geq \frac{p(1-p)}{\varepsilon^2} \frac{1}{\delta}$$

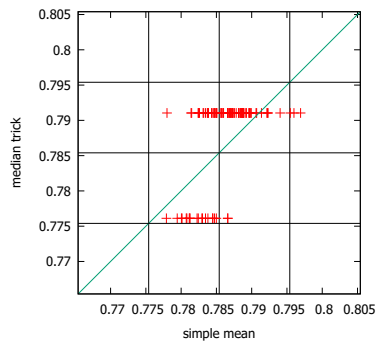
つまり, 中央値トリックにより, 硬貨を投げる回数が減った

実験してみた：結果 1

$$\delta = 10^{-3}, n = 3149$$



$$\delta = 10^{-11}, n = 11725$$



注意

- ▶ 単純なアルゴリズムの出力 X/n に対して,
 $E[X/n] = p$ が成り立つ
- ▶ 中央値トリックの出力 Y に対して,
 $E[Y] = p$ が成り立つとは限らない

X/n は不偏推定量であるが、 Y はそうではない

どういふことか？

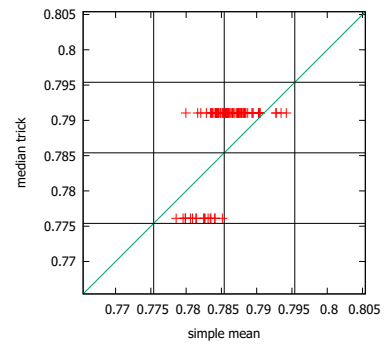
- ▶ n を大きくすると、 X/n は p に近づいていく
- ▶ n を大きくすると、 Y は p の近似値に近づいていく

今日の目標

典型的な乱択アルゴリズムの設計と解析ができるようになる

- ▶ 確率の推定 (中央値トリック)

$$\delta = 10^{-19}, n = 20301$$



- ① 確率の推定：単純なアルゴリズム
- ② 確率の推定：中央値トリック
- ③ 今日のまとめ