July 26, 2013 @COSS

Exploring the Limits of Computation

確率と計算

来嶋秀治

九州大学 大学院システム情報科学研究院 情報学部門

本講演のねらい

「乱択アルゴリズムにおいて、乱数に真に求める性質は何か？」

1. 乱択の威力: ストリーム中の頻出アイテム検知
2. 高度な乱択技法: 組合せ的対象のランダム生成
3. 脱乱択化: ランダムウォークの脱乱択化

2. 組合せ的対象のランダム生成

マルコフ連鎖モンテカルロ法

(MCMC: Markov chain Monte Carlo)

joint with 松井知己 (中央大学)

例: 2行分割表のランダム生成

2元分割表

- ✓ 各セルには非負整数が入る
- ✓ （与えられた）周辺和を満たす

						12
						18
5	4	3	7	5	6	30

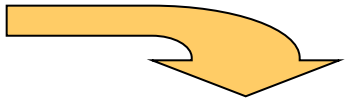
与えられた周辺和を満たす2行分割表の**個数を求める問題**

⇒#P完全 (NP困難) ['97 Dyer, Kannan, & Mount]

問題

Given: 周辺和

出力: 分割表の一様ランダム生成

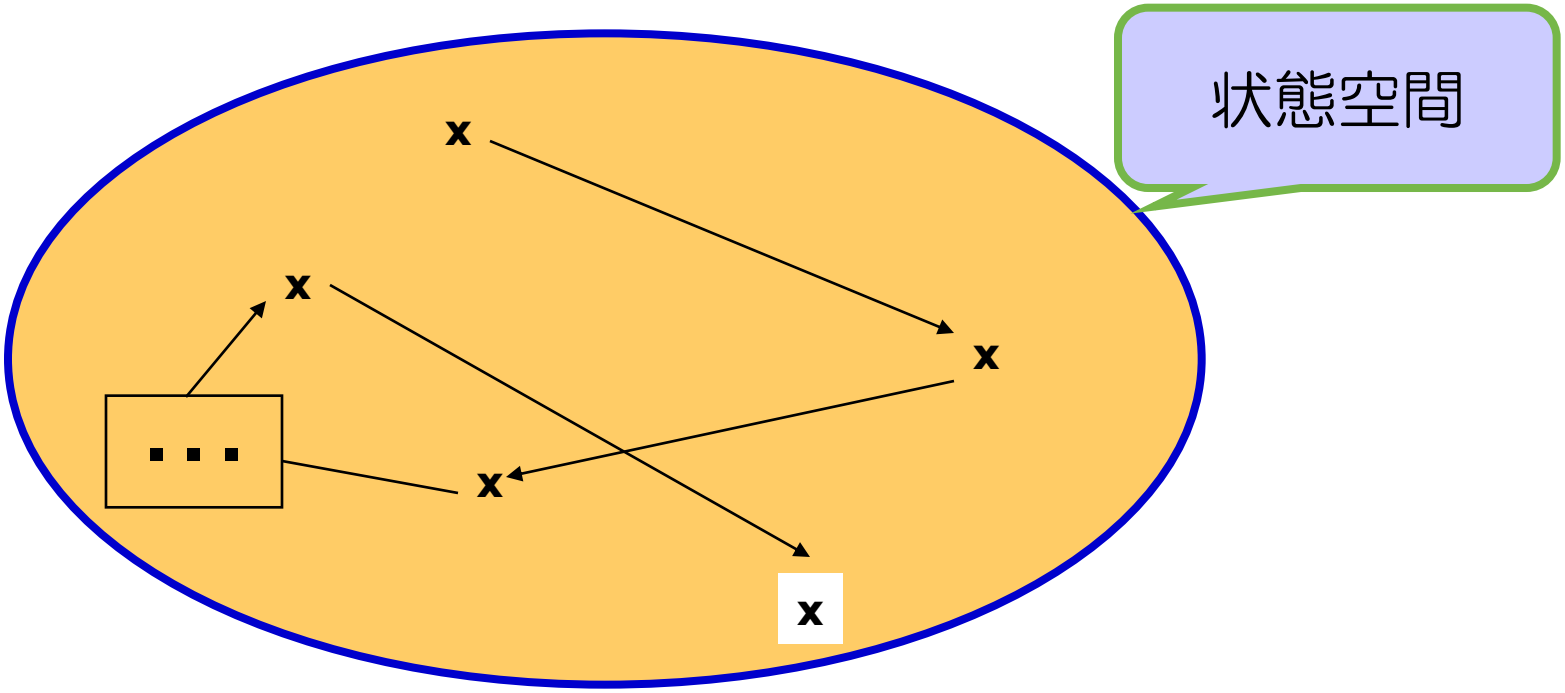


マルコフ連鎖を用いた

サンプリング法

MCMC法のアイデア

1. 所望の分布を定常分布にもつマルコフ連鎖を設計する.
2. 十分な回数推移させて，定常分布からサンプリングする.



⇒ (漸近的に)所望の分布に従うサンプリングを実現

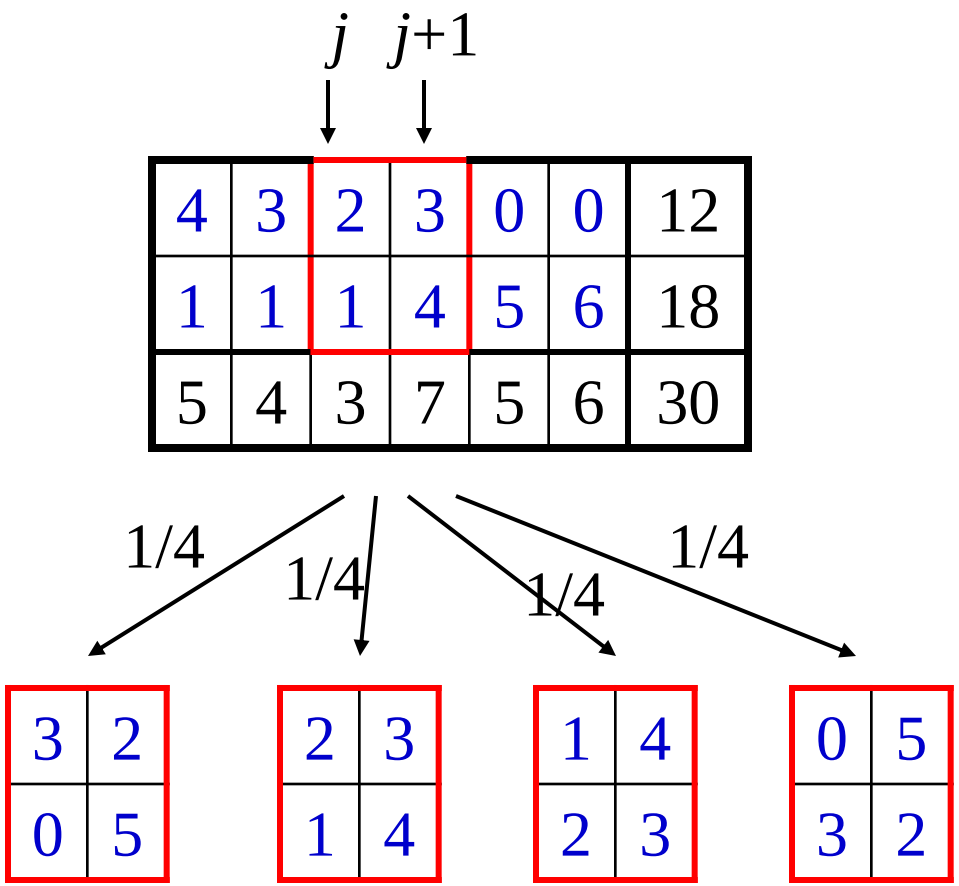
例: 2行分割表に対するマルコフ連鎖 [K & Matsui '06]

- j 列目（とj+1列目）を $1/(n-1)$の確率で選ぶ。
- j 列目と j+1列目に対して推移可能な状態に等確率で推移する。

2	3	5
1	4	5
3	7	10

+

$+k$	$-k$
$-k$	$+k$



提案するマルコフ連鎖の特徴

定理

提案したマルコフ連鎖 の定常分布は一様分布である.

略証: $\forall(X, Y),\ P(X, Y) > 0 \Rightarrow P(Y, X) > 0$ かつ $P(X, Y) = P(Y, X)$

X

4	3	2	3	0	0	12
1	1	1	4	5	6	18
5	4	3	7	5	6	30

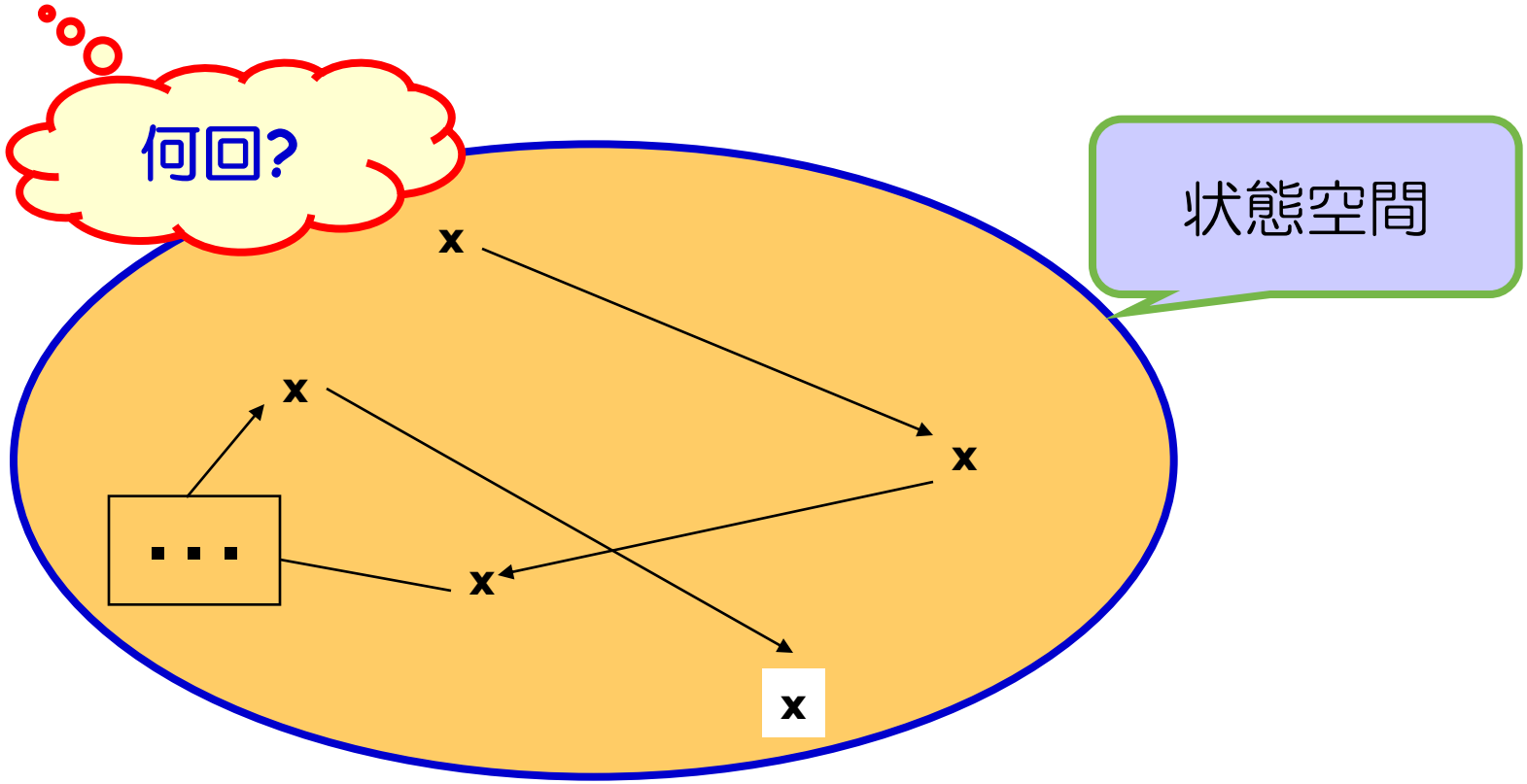
Y

4	3	0	5	0	0	12
1	1	3	2	5	6	18
5	4	3	7	5	6	30

(detailed balance equation $\pi(X)\, P(X, Y) = \pi(Y)\, P(Y, X)$)

MCMC法のアイデア

1. 所望の分布を定常分布にもつマルコフ連鎖を設計する.
2. 十分な回数推移させて，定常分布からサンプリングする.



⇒ (漸近的に)所望の分布に従うサンプリングを実現

問題点は何か？

「何回推移させれば十分か？」

近似サンプリング法

- ✓ 収束スピードの算定
  - **mixing time**, total variation distance

完璧サンプリング法

- ✓ マルコフ連鎖の推移シミュレーションを工夫
- ✓ 無限回の推移の結果を出力 (⇒ 定常分布に厳密に従う)
  - **Coupling from the past** [Propp & Wilson 1996]

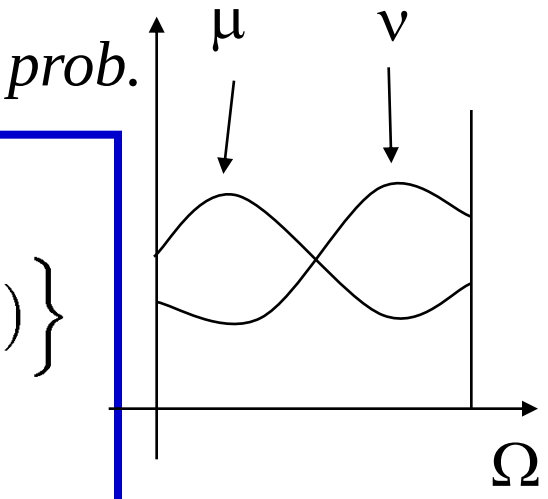
・・・

単調マルコフ連鎖の設計で効率化

Mixing time

μ, ν: 状態空間 Ω 上の分布

Error



Total variation distance

$$d_{\mathrm{TV}}(\mu,\nu) \stackrel{\text{def.}}{=} \max_{Q\subseteq\Omega}\left\{\sum_{x\in Q}(\mu(x)-\nu(x))\right\} \equiv \frac{1}{2}\sum_{x\in\Omega}|\mu(x)-\nu(x)|$$

マルコフ連鎖 M (状態空間 Ω, 推移確率 P, 定常分布 π)

Mixing time

$$\tau(\varepsilon) \stackrel{\text{def.}}{=} \max_{x\in\Omega}\{\min\{t \mid \forall s\geq t,\ d_{\mathrm{TV}}(P_x^t,\pi)\leq\varepsilon\}\}$$

- rapidly mixing if $\tau(\varepsilon) \leq \mathrm{poly.}(\log\Omega, \varepsilon^{-1})$

mixing timeの直観

P^t を計算する => 対角化 $\Lambda = Q^{-1} P Q$

$$\Lambda^t = (Q^{-1}PQ)^t$$
$$= (Q^{-1}PQ)(Q^{-1}PQ)(Q^{-1}P)\ldots(Q^{-1}PQ)$$
$$= Q^{-1}P^tQ$$

すなわち

$$P^t = Q\Lambda^tQ^{-1}$$
$$= Q\begin{pmatrix} \lambda_1^t & & & O \\ & \lambda_2^t & & \\ & & \ddots & \\ O & & & \lambda_n^t \end{pmatrix}Q^{-1}$$

確率行列の固有値 --- Perron-Frobeniusの定理

Perron-Frobeniusの定理 (cf. Gershgorinの定理)

n×n正行列$A=(a_{ij})$の固有値を$\lambda_1,\ldots,\lambda_n \in C$とすると

$$\min_i \sum_j a_{ij} \leq \max_k |\lambda_k| \leq \max_i \sum_j a_{ij}$$

が成り立つ.

さらに，絶対値最大の固有値は単根で実数値をとる.

系

確率行列Pが既約で非周期的の時,

- Pの最大固有値は1，かつ
- それ以外の固有値の絶対値は1より小さい.

mixing timeの直観

命題

Pの固有値1に対する右固有ベクトルは$\mathbf{1}$.

(Because $(P\mathbf{1})_i = \sum_{j=1}^{n} p_{ij} = 1$.)

$$P^t = Q\Lambda^t Q^{-1}$$

$$= Q\begin{pmatrix} 1^t & & & O \\ & \lambda_2^t & & \\ & & \ddots & \\ O & & & \lambda_n^t \end{pmatrix} Q^{-1} \xrightarrow{t\to\infty} Q\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} Q^{-1}$$

$$= \begin{pmatrix} 1 & q_{21} & \cdots & q_{n1} \\ 1 & q_{22} & \cdots & q_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & q_{2n} & \cdots & q_{nn} \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} Q^{-1}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} Q^{-1}$$

i.e., $\forall \boldsymbol{x}$:初期分布, $\boldsymbol{x}P^t \xrightarrow{t\to\infty} \boldsymbol{\pi}$,

where $\boldsymbol{\pi}$はQ^{-1}の1行目 (=固有値1の左固有ベクトル)

mixing timeの直観

第二固有値が大事！

- カップリング法
- コンダクタンス法
- スペクトル解析

数え上げ, MCMC, random walkに関連する研究

1979, Valiant, #P完全の提唱

1982, Aldous, coupling法

1986, Jerrum, Valiant, Vazirani, 数え上げとサンプリング

1989, Jerrum & Sinclair, コンダクタンス (expander)

1989, Toda, $PH \subseteq P^{\#P}$

1991, Dyer, Frieze, Kannan, 凸体のFPRAS

1996, Propp & Wilson, 完璧サンプリング法

1997, Bubley & Dyer, path coupling法

2003, Ikeda, Kubo, Okumoto, Yamashita, βランダムウォーク

2004, Jerrum, Sinclair, Vigoda, パーマネントのFPRAS

近似数え上げ決定性アルゴリズム

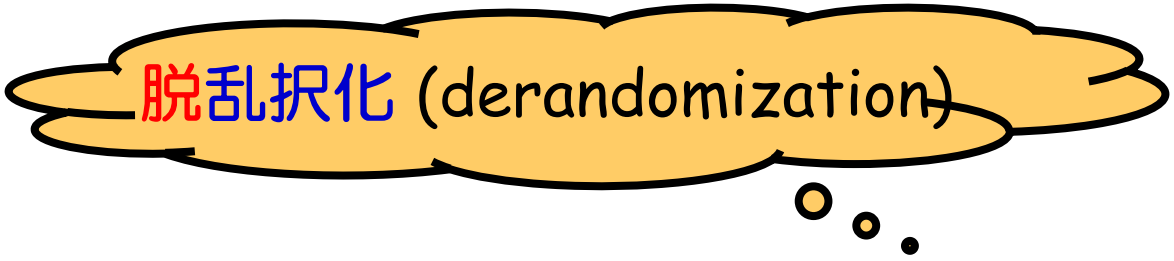
Stefankovic, Vempala, Vigoda, A Deterministic Polynomial-Time Approximation Scheme for Counting Knapsack Solutions. SIAM J. Comput. 41(2): 356-366 (2012)

Gopalan, Klivans, Meka, Polynomial-Time Approximation Schemes for Knapsack and Related Counting Problems using Branching Programs. CoRR abs/1008.3187 (2010) (cf. FOCS2011)

Yitong Yin, Chihao Zhang, Approximate Counting via Correlation Decay on Planar Graphs. SODA 2013: 47-66

Liang Li, Pinyan Lu, Yitong Yin, Correlation Decay up to Uniqueness in Spin Systems. SODA 2013: 67-84

「乱択アルゴリズムにおいて、乱数に真に求める性質は何か？」



3. ランダムウォークの脱乱択化

- ✓ 来嶋, 古賀 健太郎 (FANUC), 牧野 和久 (東大)
- ✓ 梶野 洸 (東大), 来嶋, 牧野 和久 (東大)
- ✓ 白髪 丈晴, 山内 由紀子, 来嶋, 山下 雅史 (九大)

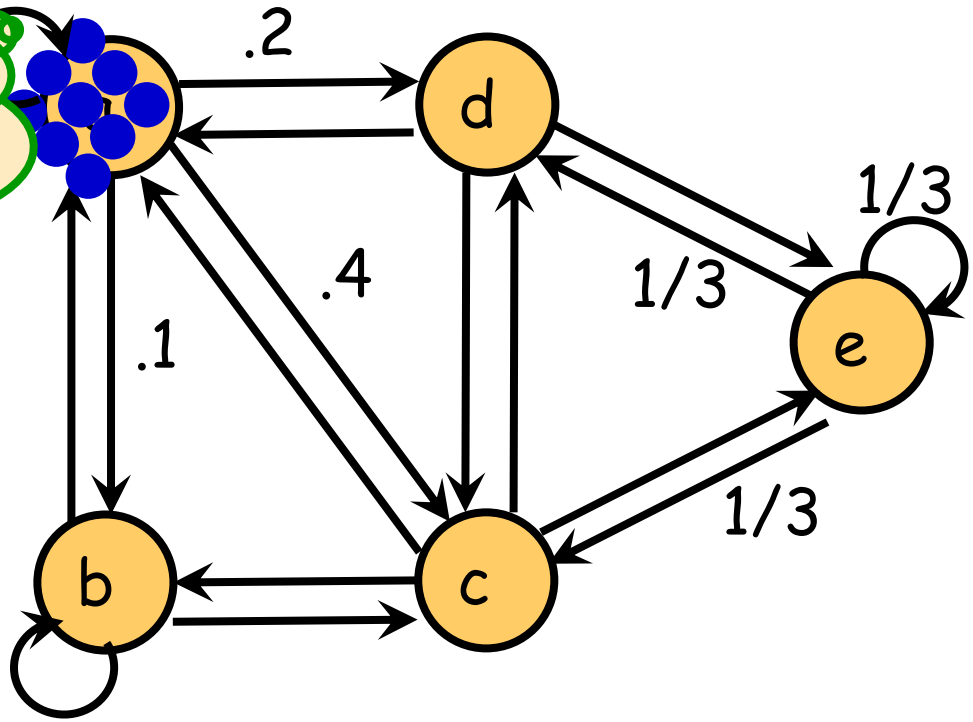
ランダムウォーク (複数トークンによる分布の近似)

N個のトークンがグラフ上を独立にランダムウォークする.

- ✓ μ^0: 初期配置 ($\pi^0 \approx \mu^0/N$)
- ✓ P: 推移確率行列 (確率P_{uv}で頂点u頂点vに移動)
- ✓ 時刻tの期待配置$\mu^t := \mu^0 P^t$ ($\pi^t \approx \mu^t/N$)

(Nが非常に大きければ,)
トークンを約2:4:1:3の割合で
隣接点にばらまいている.

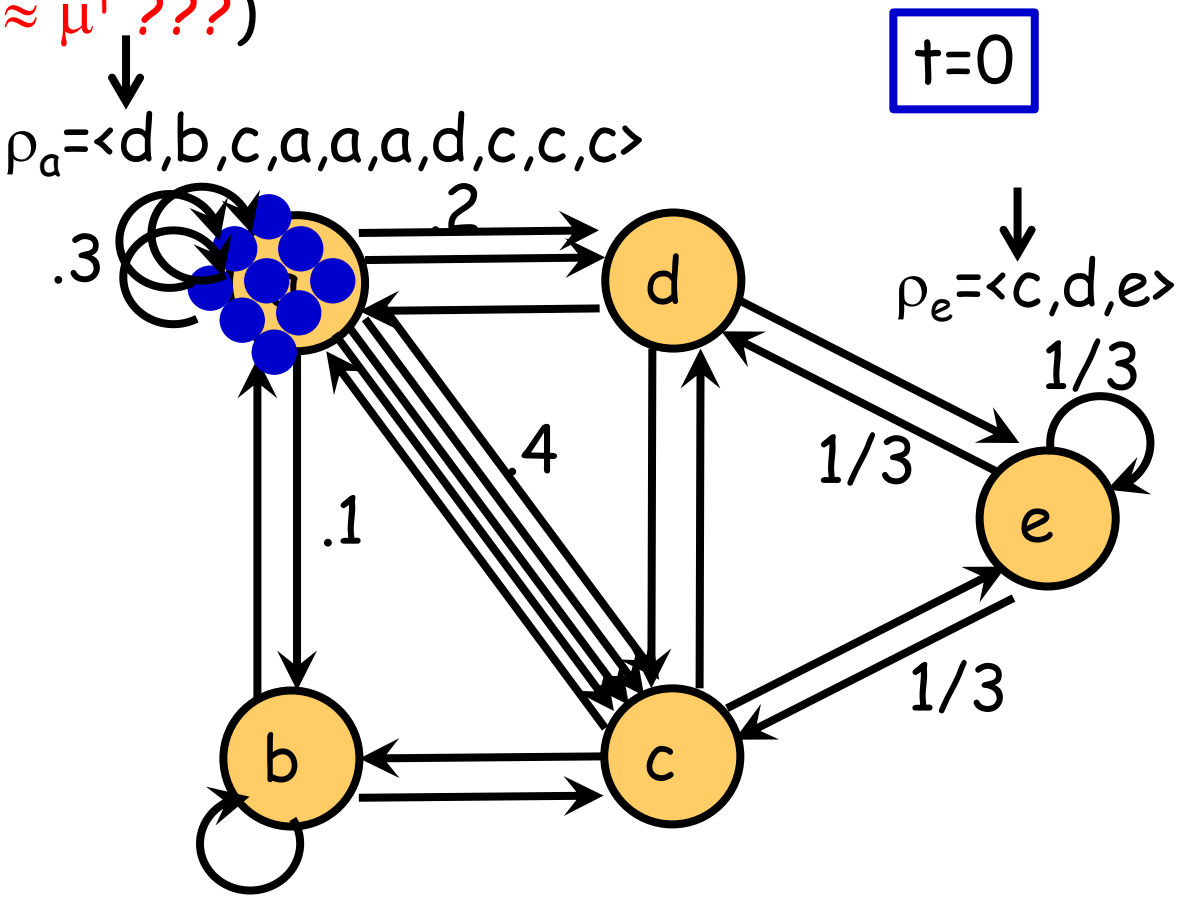
決定的過程でも大差ない?



deterministic RW (Propp機械; rotor-router)

N個のトークンがグラフ上を移動する.

- ✓ χ^0: 初期配置 ($\chi^0 = \mu^0$)
- ✓ ρ: "**rotor router**" (比率P_{uv}で頂点u頂点vに"順次"移動)
- ✓ 時刻tの配置χ^t ($\chi^t \approx \mu^t$???)



誤差の下界

定理 [K, Koga, Makino 10+]

ある多重有向グラフ$\mathcal{G}$=(V,$\mathcal{E}$),

ある初期状態, あるrotor-routerが存在して,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \geq \Omega(m)$$

但し, mは多重グラフの頂点数,枝数.

誤差の下界

定理 [K, Koga, Makino 10+]

ある多重有向グラフ$\mathcal{G}$=(V,$\mathcal{E}$),

ある初期状態, あるrotor-routerが存在して,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \geq \Omega(m)$$

但し, mは多重グラフの頂点数,枝数.

主結果 2

定理 [K, Koga, Makino 10+]

対応する推移確率行列Pの固有値がすべて非負ならば,
任意の多重有向グラフ, 任意の初期状態, 任意のrotor-router,
任意の頂点w, 任意の時刻tについて,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \leq (2m-n)\left(\max_{i\in\{1,\ldots,\kappa\}} n_i + n + 3\right) \leq 4mn + \mathrm{O}(m)$$

但し, n,mは多重グラフの頂点数,枝数. n_iはJordan cellのサイズ.

Remark

「推移確率行列Pの固有値がすべて非負」という条件.

- reversible lazy Markov chainはこの条件を満たす.
  - ✓ MCMC法で使われるマルコフ連鎖
- +Pが対称 ⇒ Pは半正定値行列

チップ数に依存しない

先行研究と本研究の成果

定理 [Cooper & Spencer 2006]

$\mathbb{Z}^d$上で, 任意の初期状態, 任意のrotor-router,
任意の頂点w, 任意の時刻tについて,
(dのみに依存する)定数c_dが存在して,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \leq C_d$$

単一頂点誤差に関する研究(1/2)

2006	Cooper, Spencer	Z^d上のロータールーター ➢ 誤差 $\leq C_d$
2007	Cooper, Doerr, Spencer, Tardos	Z^1上のロータールーター ➢ $C_1 \leq 2.29$
2008	Cooper, Doerr, Friedrich, Spencer	無限のk正則木上のロータールーター ➢ 誤差 $> \Omega(\sqrt{kT})$ at time T
2009	Doerr, Friedrich	Z^2上のロータールーター ➢ $C_2 \leq 7.83$ (上右下左) ➢ $C_2 \leq 7.29$ (上下左右)
2012	Kijima, Koga, Makino	有限多重有向グラフ G上のロータールーター ➢ 誤差$\leq 4mn+O(m)$ $\{0,1\}^d$上のPropp機械 ➢ 誤差$\leq O(d^3)$ (頂点数のpoly log)
2012+	Kajino et al.	(後述)
2012+	Shiraga et al.	(後述)

単一頂点誤差に関する研究(2/2)

2012 (2010)	Kijima, Koga, Makino	有限多重有向グラフ G上のロータールーター ➢ 誤差 $\leq 4m^*n+O(m^*)$ (P: 有理数 + 既約 + 非周期 + 可逆 + lazy) $\{0,1\}^d$上のPropp機械 ➢ 誤差 $\leq O(d^3)$ (頂点数のpoly log)
2012+ (2011)	Kajino, Kijima, Makino	有限多重有向グラフ G上のロータールーター ➢ 誤差 $\leq 0\left(\alpha \frac{m^*n^2}{1-\lambda}\right)$ (P: 有理数 + 既約) $\{0,1\}^d$上のPropp機械 ➢ 誤差 $\leq O(d^2)$ (頂点数のpoly log)
2012+	Shiraga, Yamauchi, Kijima, Yamashita	有限有向グラフ G上の関数ルーター ➢ 誤差 $\leq 0\left(\sqrt{\frac{\pi_{\max}}{\pi_{\min}}}\frac{mn}{1-\lambda}\log M\right)$ (P: 実数 + 既約 + 非周期 + 可逆)

K., Koga, Makino

定理 [K, Koga, Makino 10+]

対応する推移確率行列Pの固有値がすべて非負ならば,
任意の多重有向グラフ, 任意の初期状態, 任意のrotor-router,
任意の頂点w, 任意の時刻tについて,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \leq (2m-n)\left(\max_{i\in\{1,\ldots,\kappa\}} n_i + n + 3\right) \leq 4mn + \mathrm{O}(m)$$

但し, n,mは多重グラフの頂点数,枝数. n_iはJordan cellのサイズ.

Remark

「推移確率行列Pの固有値がすべて非負」という条件.

- reversible lazy Markov chainはこの条件を満たす.
  - ✓ MCMC法で使われるマルコフ連鎖
- +Pが対称 ⇒ Pは半正定値行列

K., Koga, Makino

定理 [K, Koga, Makino 10+]

$\{0,1\}^d$超立方体の稜線グラフに対して,
任意の初期状態, 任意のrotor-router,
任意の頂点w, 任意の時刻tについて,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \leq \frac{3}{2}d^3 + \mathrm{O}(d^2)$$

定理 [K, Koga, Makino 10+]

Johnsonグラフ$J(d,c)$に対して,
任意の初期状態, 任意のrotor-router,
任意の頂点w, 任意の時刻tについて,

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| \leq 2c^3 \cdot (d-c)^2 + \mathrm{O}(c^3 \cdot (d-c))$$

Johnsonグラフ$J(d,c) = (V_J, E_J)$
$V_J = \{S \subset \{1,\ldots,d\} \mid |S| = c\}$,
$E_J = \{\{S,T\} \in V_J^2 \mid |S \oplus T| = 2\}$

頂点数に対して
対数多項式上界

一般グラフに対する定理の証明

$$X_v^{(t)} := \sum_{s=0}^{t} \chi_v^{(s)}$$

$$s_v(i) := \min\left\{ t \geq 0 \mid i < \sum X_v^{(t)} \right\},$$

Cooper & Spencer 2006の手法を
推移確率行列Pで理解

補題 1

$$\chi_w^{(T)} - \mu_w^{(T)} = \sum_{v \in V} \sum_{i=0}^{X_v^{(T-1)}-1} \left(P^{T-s_v(i)-1}(\rho_v(i), w) - P^{T-s_v(i)}(v, w) \right).$$

「時刻†以前Propp機械, 時刻†以後RW」過程. (初期配置はχ^0(と同一)とする.)

ζ(w;t,T): 時刻Tの期待トークン配置.

Remark
- ✓ $\zeta(w;T,T) = \chi^T$
- ✓ $\zeta(w;0,T) = \mu^T$

$$\chi_w^{(T)} - \mu_w^{(T)} = \sum_{t=0}^{T-1} (\zeta(w; t+1, T) - \zeta(w; t, T)).$$

$$\zeta(w; t+1, T) - \zeta(w; t, T) = \sum_{v \in V} \sum_{i=X_v^{(t-1)}}^{X_v^{(t)}-1} \left(P^{T-t-1}(\rho_v(i), w) - P^{T-t}(v, w) \right)$$

関連研究

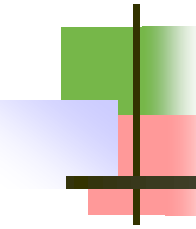
- ✓ IDLA (Internal Diffusion-Limited Aggregation)
  - Levine & Peres 2005
- ✓ Information Spreading
  - Doerr, Friedrich, & Sauerwald 2008
  - Doerr, Friedrich, Kunnemann, & Sauerwald 2009
- ✓ Hitting time, Cover time
  - Friedrich & Sauerwald 2010
  - Holroyd & Propp 2010+

cf. [K, Koga, Makino 10+]
$|\chi^t_v/N - \pi^t_v| \leq O(mn/N)$

定理 [Holroyd & Propp 2010+]
単一トークンのPropp機械を考える.
F^t_v := 時刻0からtまでに頂点vを訪れた回数.
任意の有限グラフについて,
$|F^t_v/t - \pi^*_v| \leq O(mn/t)$
ただしπ^*は対応するマルコフ連鎖の定常分布.

今後の課題

- ✓ 上下界の一致. (O(mn), Ω(m))
- ✓ 組合せ構造に由来するグラフに対するpolylogの上界.
- ✓ Blanket time vs Mixing time.
- ✓ MCMC法の脱乱択化.
- ✓ 乱数とは？
  - ➢ 乱択アルゴリズムにおける「乱数」の持つべき性質は？
    - ◆ 準モンテカルロ (quasi Monte Carlo)
    - ◆ カオス系列 (Chaos time series)



3. Functional-router model

We propose a new deterministic process
which imitates irrational transition probabilities,
in a similar fashion to the rotor-router model.

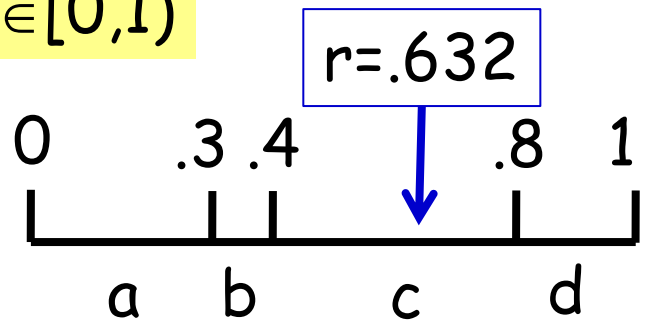
ランダムウォーク

トークンがグラフ上をランダムウォークする.

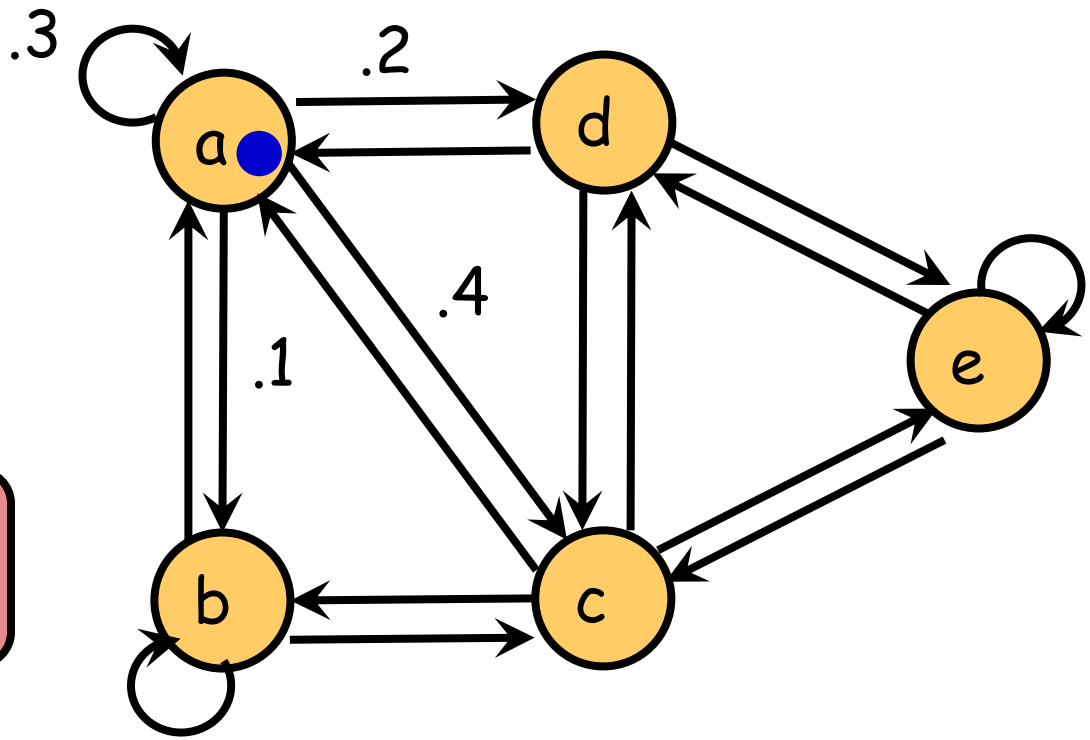
- ✓ π^0: 初期分布 (トークンは確率π^0_vで頂点vに居る)
- ✓ P: 推移確率行列 (確率P_{uv}で頂点u頂点vに移動)
- ✓ 時刻tの確率分布$\pi^t := \pi^0 P^t$ (頂点vに居る確率 $(\pi^t)_v$)

ランダムウォークを
乱数を使って実現

$r \in [0,1)$



乱数の代わりに
超一様分布列を使う.



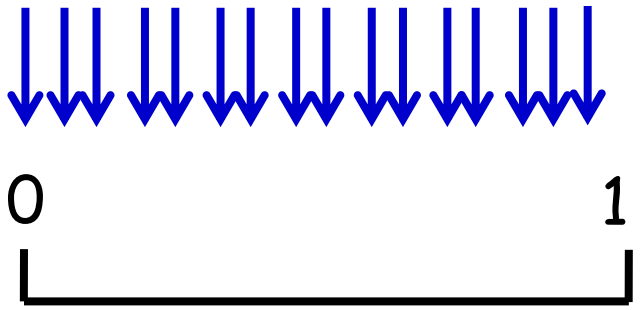
Van der Corput列

自然数 $i = \sum_{j=0}^{\lfloor \lg i \rfloor} \beta_j(i) 2^j$

ただし $\beta_j(i) \in \{0,1\}$ $(j = 0,1, \ldots, \lfloor \lg i \rfloor)$

$$\psi(i) \coloneqq \sum_{j=0}^{\lfloor \lg i \rfloor} \beta_j(i) 2^{-(j+1)}$$

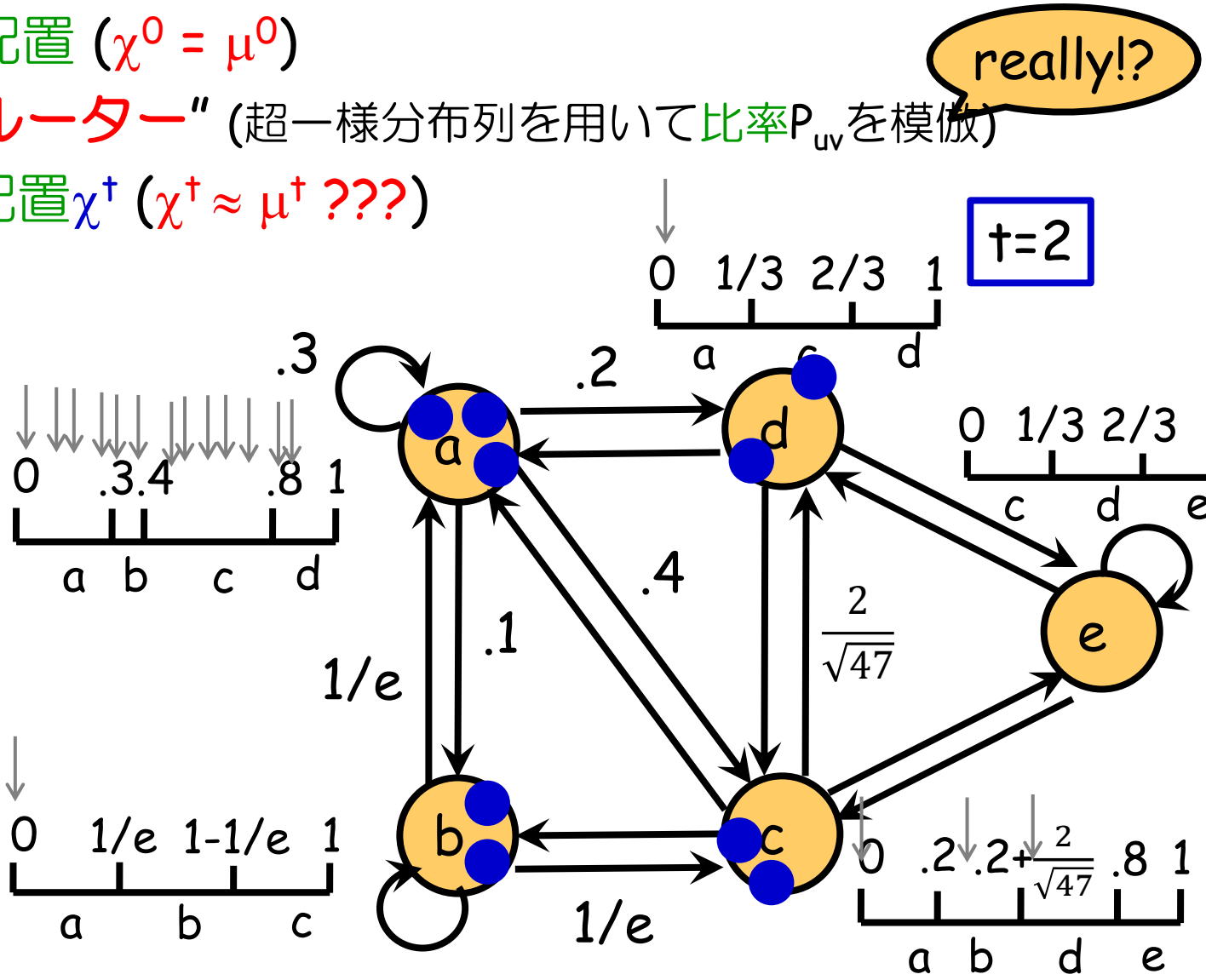
i	$(i)_2$	$(\psi(i))_2$	$\psi(i)$
0	0	0	0
1	1	0.1	1/2
2	10	0.01	1/4
3	11	0.11	3/4
4	100	0.001	1/8
5	101	0.101	5/8
6	110	0.011	3/8
…	…	…	…



関数ルーターモデル

N個のトークンが決定的にグラフ上を移動.

- ✓ χ^0: 初期配置 ($\chi^0 = \mu^0$)
- ✓ σ: "**関数ルーター**" (超一様分布列を用いて比率P_{uv}を模倣)
- ✓ 時刻tの配置χ^t ($\chi^t \approx \mu^t$???)



van der Corput列の誤差

$\left|I_{u,v}[0,z)\right|$: 頂点uから発射された総トークン数がz個の時，vに発射されたトークン数

定理

任意の行列Pに対して，

$$\left|\frac{\left|I_{u,v}[0,z)\right|}{z} - P(u,v)\right| < \frac{2\lfloor \lg z \rfloor + 2}{z} = O\left(\frac{\log z}{z}\right)$$

が任意の$u, v \in V$ および$z \in \mathbb{Z}_{z\geq 0}$について成り立つ．

Unfortunately this bound is tight.

命題

ある行列 P に対して，ある $u, v \in V$で，

$$\left|\frac{\left|I_{u,v}[0,z)\right|}{z} - P(u,v)\right| > \frac{\lg\left(\frac{3}{4}z\right)}{3z} = \Omega\left(\frac{\log z}{z}\right)$$

が無数の$z \in \mathbb{Z}_{\geq 0}$について成り立つ．

下界の例

命題

無数の$M \in \mathbb{Z}_{>0}$に対して $\left|\chi_w^{(T)} - \mu_w^{(T)}\right| > \lg\left(\frac{3}{4}M\right)$

の成り立つ例が存在する.

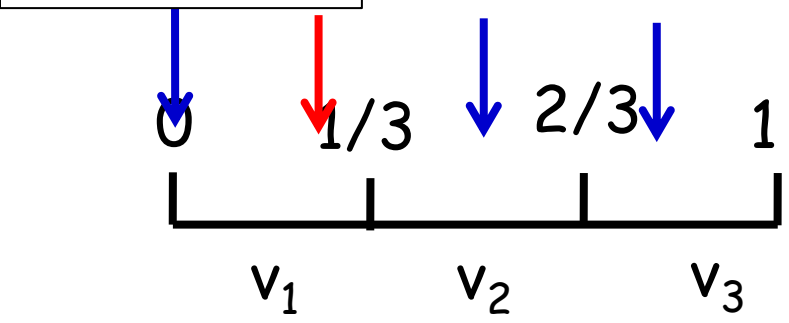
K_3上の単純ランダムウォークを考える.

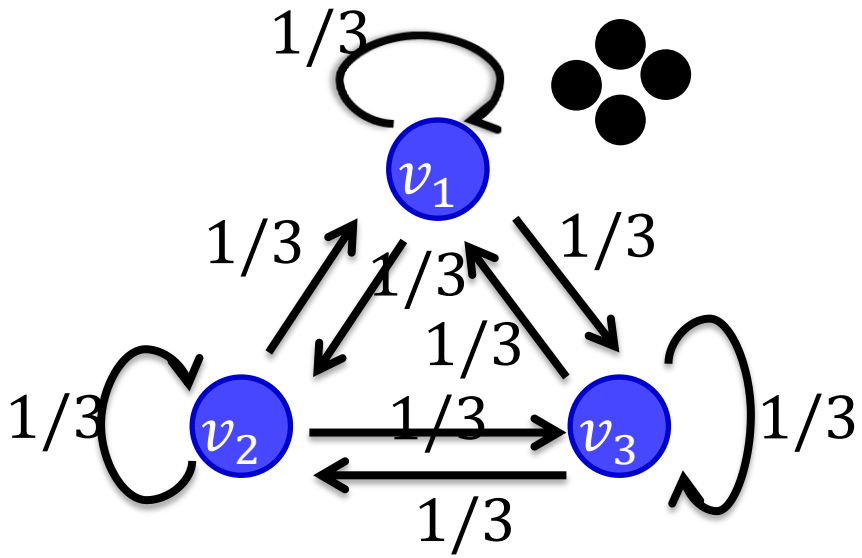
総トークン数を$M := \sum_{l=1}^{k} 4^l$ ($k \in \mathbb{Z}_{>0}$)とする.

このとき, $4^l \equiv 1 \pmod 3$

($4^l - 1 = 3(4^{l-1} + \cdots + 1)$)

直観的には





主定理

定理[白髪, 山内, K., 山下12+]

Pがエルゴード的で可逆の時,

任意の初期配置状態, 任意のw∈V, 任意の時刻 t に対して

$$\left|\chi_w^{(T)} - \mu_w^{(T)}\right| < \sqrt{\frac{\pi(w)}{\pi_{\min}} \cdot \frac{m(n-1)}{1-\lambda^*} \cdot 2(\lg M + 1)}$$

が成り立つ. ただし, n=|V|, m=|E|,

λ*はPの第２固有値, πはPの定常分布, $\pi_{\min} = \min_{v \in V} \pi(v)$ とする.

MCMC法では,
しばしば仮定される.

P が可逆とは詳細釣合の式

$$\pi(u)\, P(u,v) = \pi(v)\, P(v,u)$$

が任意の u,v ∈Vに成り立つことをいう

注意

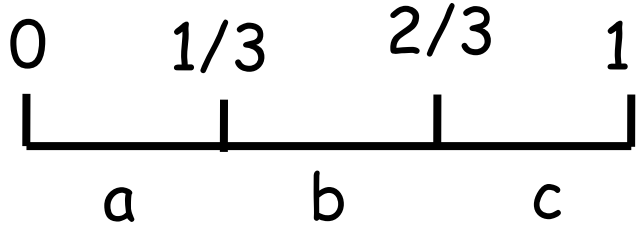
"棄却サンプリング"のアイデアに基づいて,

数ルーター (の変種)がロータールーターに一致する

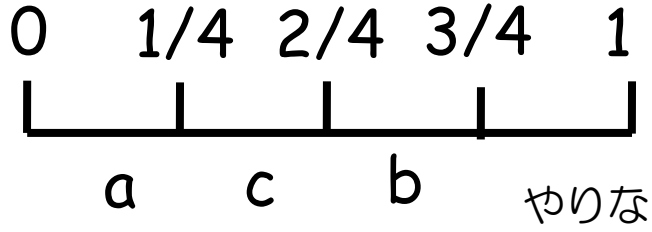
(ただし，推移確率行列は有理数とする．)

ロータールーター: $\rho=\langle a,b,c\rangle$

(素朴な) 関数ルーター



(変形) 関数ルーター



やりなおし

"関数ルーター" はロータールーターの一般化モデル

1. ストリーム中の頻出アイテム検知

緒方 正虎，山内 由紀子，来嶋 秀治，山下 雅史

九州大学

θ: "頻出度"パラメータ

Σ: アイテム集合(有限)

ストリームデータ中の頻出アイテム検知

問題: 頻出アイテム検知

Input: $\theta \in (0,1)$ $\boldsymbol{x} = (x_1,\ldots,x_N) \in \Sigma^N$ (順々に)

Find: all $s \in \Sigma$ s.t. $f(s) \geq \theta \cdot N$ ●○○

但し $f(s)$ は s が $\boldsymbol{x}$ 中で出現した回数

事前には、
N (or log Nの近似値)も
わからない。

例1. 1日のPOSデータ(@果物屋)

Σ ={ 🍎, 🍈, 🍌, …, 🍇 }

$\boldsymbol{x}$ = 🍎, 🍌, 🍇, 🍎, 🍎, 🍈, 🍎, 🍌, 🍌, 🍎, 🍌, ….

例. 1日のアクセスIPアドレス

$\Sigma \subseteq$ {0.0.0.0,…,255.255.255.255}

$\boldsymbol{x}$ = 123.45.67.89, 111.11.1.1., 123.45,67,89, 122.122.12.12…

would like to find items appearing w/ frequency more than θ=1% of N.

頻出アイテム検知の領域複雑度

定理 [Karp, Shenker, Papadimitriou '03]

頻出アイテム検知を厳密に行うには,

$\Omega(|\Sigma| \log (N/|\Sigma|))$ bits が必要.

($N \gg |\Sigma| \gg 1/\theta$ とする)

定理 [Karp, Shenker, Papadimitriou '03]

頻出アイテム検知に対する

$O((1/\theta) \log N)$ bits の偽陽性(近似)アルゴリズムが存在.

($N \gg |\Sigma| \gg 1/\theta$ とする)

決定的

$o(\log N)$ bitsアルゴリズム？

- e.g. $O(\log \log N)$ bits?

単純化: o(log N) bitsで要素数を数えられるか？

問題: 要素数え上げ

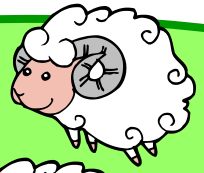
Input: $x = (a,\ldots,a) \in \Sigma^N$ (順々に)

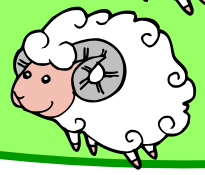
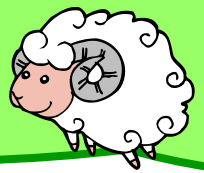
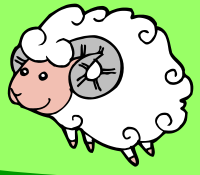
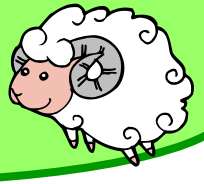
Find: N

事前には、
N (or log Nの近似値)も
わからない。

アルゴリズム: 数え上げ

0. **Set** n:= 0.
1. **Read** an input. **If** no more input, **goto** 3.
2. n++, **Goto** 1.
3. **Output** n (as N = n).

Σ = {  }

X =  ,  ,  ,  , … …

単純化: o(log N) bitsで要素数を数えられるか？

問題: 要素数え上げ

Input: $x = (a,\ldots,a) \in \Sigma^N$ (順々に)

Find: N

事前には、
N (or log Nの近似値)も
わからない。

Remark

- NはO(log N) bitsで表現可能.
  - ✓ N = 1,351,127,649,213
- Nの近似はO(log log N) bitsで表現可能
  - ✓ $N \approx 1.351 \times 10^{12}$

つまり、

指数部(=log N)がo(log N) bits領域で近似計算できるか？ということ。

表現はO(log log N) bitsで可能

確率的数え上げ

問題: 要素数え上げ

Input: $x = (a,\dots,a) \in \Sigma^N$ (順々に)

Find: N

事前には、
N (or log Nの近似値)も
わからない

⇒ key point

"w.p. $1/2^k$ " using
$O(\log K)$ bits on PTM.

アルゴリズム: 確率的数え上げ

0. **Set** k:=0.
1. **Read** an input. **If** no more input, **goto** 3.
2. k++, w.p. $1/2^k$. ~~**Goto**~~ 1.
3. **Output** k (as $N \approx 2^k$).

$O(\log \log N)$ bits

Thm. [Morris '78, Flajolet '85]

$E[2^k] \approx N+1$

because
$N \approx 1+2+4+8+16+\dots+2^k$

確率的数え上げ(改良型)

問題: 要素数え上げ

Input: $x = (a,\ldots,a) \in \Sigma^N$ (順々に)

Find: N

事前には、
N (or log Nの近似値)も
わからない。

アルゴリズム: 確率的数え上げ(改良型)

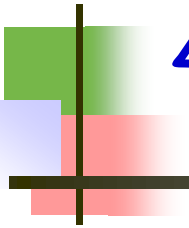
0. **Set** k:=0, l:=0.
1. **Read** an input. **If** no more input, **goto** 4.
2. l++, w.p. $1/2^k$.
3. If l==2^b, k++, and **set** l := l' w.p. $\binom{2^b}{l'} \cdot 2^{-2^b}$. **Goto** 1.
4. **Output** l and k (as $N \approx l * 2^k$).

O(log log N) bits

Thm. [Ogata, Yamauchi, K., Yamashita '11]

$E[l*2^k] \approx N$.

"改良型" を使うと、頻出アイテム検知も可能. (詳細略)



4. まとめ

1. 乱択の威力: ストリーム中の頻出アイテム検知
   - ✓ $O(\log \log N)$領域計算法
2. 高度な乱択技法: 組合せ的対象のランダム生成
3. 脱乱択化: ランダムウォークの脱乱択化