区間定常無記憶情報源 (PSMS) の学習アルゴリズム

金澤 宏紀　山西 健司
東京大学大学院 情報理工学系研究科

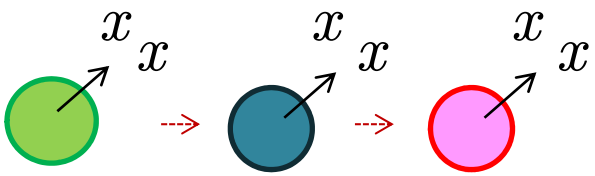
2012 年 9 月 24 日
第 3 回 Latent Dynamics Workshop

研究の動機

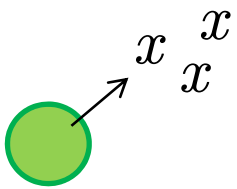
- 非定常情報源の変化検出

  時系列データ $x^n = x_1 \ldots x_n$

情報源が時間変化 または
複数の情報源がスイッチする



cf.) 定常情報源



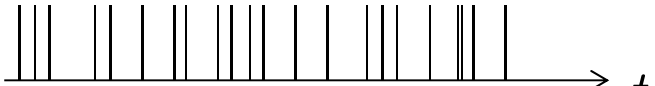
変化検出 { いつ変化したか（変化点検知）
どの程度変化したか（変化量検知）

をデータのみから行う

PSMS 学習 ～ パラメータ変化検出アルゴリズム

研究の応用先

- 例
  - 脳波の時系列データ
    - 非自明な変化点を探る



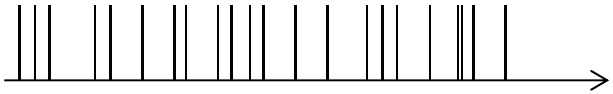
点過程データ

  - 事故件数の時系列データ

1期	2期	3期	4期	5期	…
12	8	18	21	22	…

研究の応用先

- 例
  - 脳波の時系列データ
    - 非自明な変化点を探る



点過程データ

インターバルをモデリング
指数分布

$$f(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}$$

ガンマ分布

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\beta x}$$

  - 事故件数の時系列データ

1期	2期	3期	4期	5期	…
12	8	18	21	22	…

Poisson 分布

$$f(x;\lambda) = \frac{\lambda^{x}\mathrm{e}^{-\lambda}}{x!}$$

✓ 状況の同定
✓ パラメトリックモデルの適切性を評価

モデル選択規準

- 記述長最小原理 (MDL principle) [Rissanen 1978]-

$$\operatorname*{argmin}_{M \in \mathcal{M}} \mathcal{L}(x^n, M) = \operatorname*{argmin}_{M \in \mathcal{M}} \big(\mathcal{L}(x^n \mid M) + \mathcal{L}(M) \big)$$

  - モデル M 下でのデータ列記述長 + モデル M の記述長を最小にするモデルを選ぶ
  - 記述長: $\mathcal{L}(A) = -\log \Pr(A)$ → 対数損失

- 動的モデル選択 (DMS) [Yamanishi and Maruyama 2005, 2007]

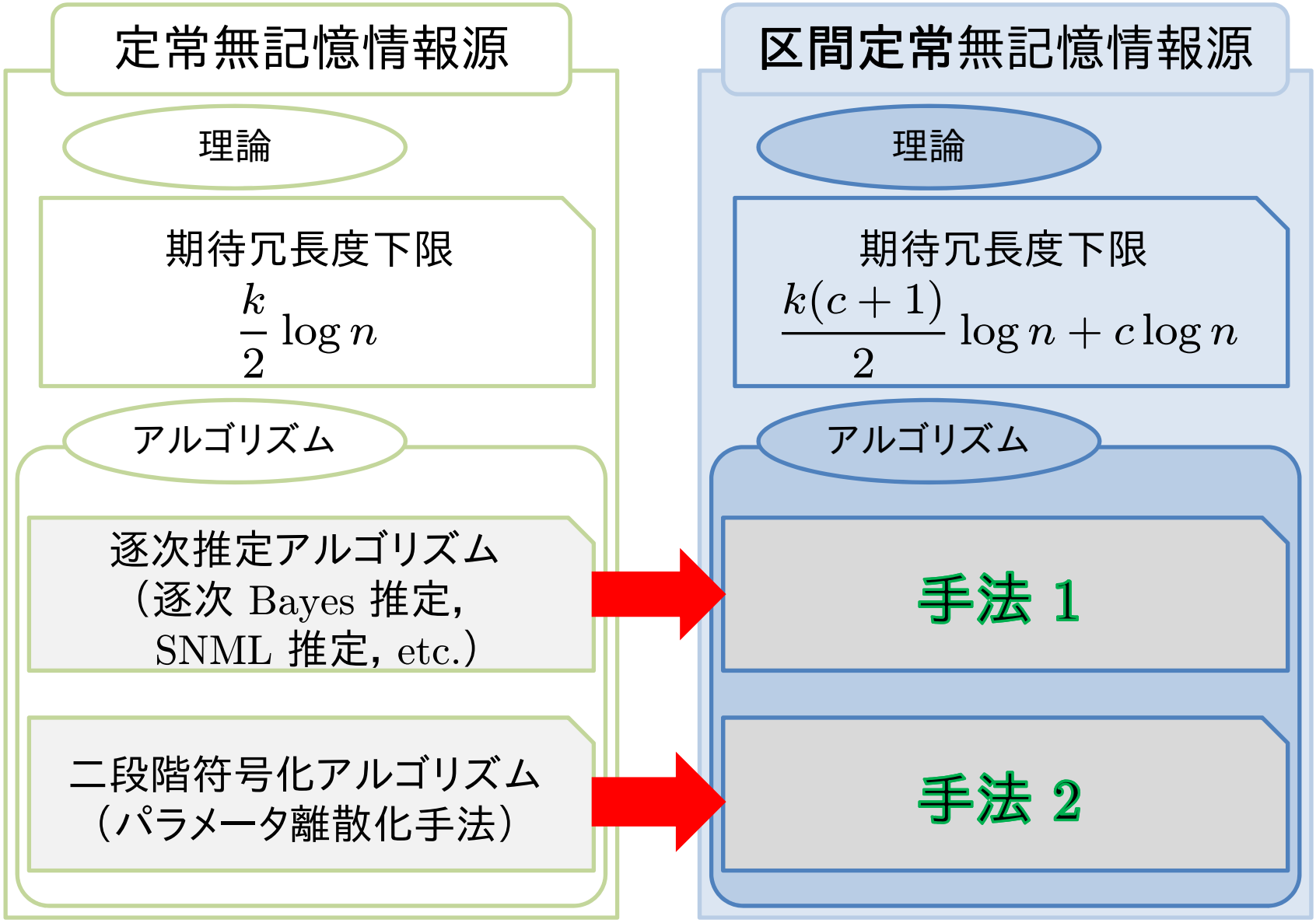
$$\operatorname*{argmin}_{M^n \in \mathcal{M}^n} \mathcal{L}(x^n, M^n) = \operatorname*{argmin}_{M^n \in \mathcal{M}^n} \big(\mathcal{L}(x^n \mid M^n) + \mathcal{L}(M^n) \big)$$

$$\downarrow$$

$$\textstyle\sum_{t=1}^n \mathcal{L}(x_t \mid M_t) + \sum_{t=1}^n \mathcal{L}(M_t \mid M^{t-1})$$

  - モデル列を逐次的に選ぶ
  - モデル列に遷移確率が入る

分野全体の概観


定常無記憶情報源
理論
期待冗長度下限
$\frac{k}{2}\log n$
アルゴリズム
逐次推定アルゴリズム
（逐次 Bayes 推定，
SNML 推定，etc.）
二段階符号化アルゴリズム
（パラメータ離散化手法）
区間定常無記憶情報源
理論
期待冗長度下限
$\frac{k(c+1)}{2}\log n + c\log n$
アルゴリズム
手法 1
手法 2

問題設定

- 区間定常無記憶情報源（PSMS）
  - c , $\theta(p)$ $(p = 0, \ldots, c)$ および $m(p)/n$ $(p = 1, \ldots, c)$ の組によって特徴づけられる非定常情報源の列

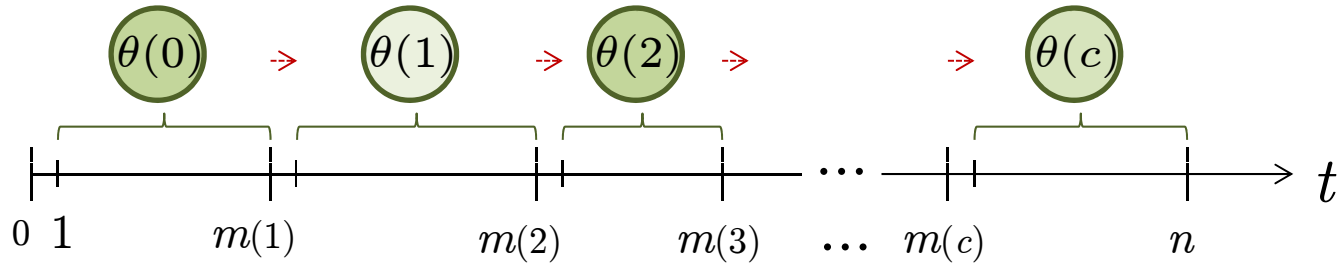
$$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\} \text{ for } x \in \mathcal{X}$$

データ列 $x^n = x_1 \ldots x_n$ が以下の確率分布にしたがう

$$\begin{cases} x_t \sim f\big(x;\theta(0)\big) & \big(1 \le t \le m(1)\big), \\ x_t \sim f\big(x;\theta(1)\big) & \big(m(1)+1 \le t \le m(2)\big), \\ \quad\vdots & \\ x_t \sim f\big(x;\theta(c)\big) & \big(m(c)+1 \le t \le n\big) \end{cases}$$

各 x_t 独立

c ：変化回数（定数）



アルゴリズムの評価指標

- PSMS に対する期待冗長度
  - PSMS 符号化アルゴリズムの性能評価指標の一つ

$$\mathcal{R}_{\mathcal{A}}^{(n)} = \mathbb{E}\left[\mathcal{L}_{\mathcal{A}}(x^n) - \sum_{p=0}^{c} \sum_{t=m(p)+1}^{m(p+1)} \left(-\log f(x_t;\theta(p))\right)\right]$$

期待冗長度
(expected redundancy)

アルゴリズム $\mathcal{A}$
でのデータ記述長

PSMS の情報をすべて
得ているときの記述長
（基準となるライン）

cf.) 定常情報源の場合

$$\mathcal{R}_{\mathcal{A}}^{(n)} = \mathbb{E}_{\theta}\left[\mathcal{L}_{\mathcal{A}}(x^n) - \sum_{t=1}^{n} \left(-\log f(x_t;\theta)\right)\right]$$

アルゴリズムの評価指標

- 期待冗長度の理論的な下限

  定理 [Merhav 1993]

  $\mathcal{X}$: 有限離散, 他いくつかの仮定をおく

  ほとんどすべての PSMS に対し

  $$^{\forall}\varepsilon > 0,\ ^{\exists}n_0 \in \mathbb{N},\ n \geq n_0 \Rightarrow$$

  $$\inf_{\mathcal{A}} \mathcal{R}_{\mathcal{A}}^{(n)} \geq (1-\varepsilon)\left(\frac{k(c+1)}{2}\log n + c\log n\right)$$

  $\inf_{\mathcal{A}}$: すべての無歪符号化アルゴリズム
  k : パラメータの次元
  c : (真の)変化回数

  cf.) 定常情報源の場合

  $$\inf_{\mathcal{A}} \mathcal{R}_{\mathcal{A}}^{(n)} \geq (1-\varepsilon)\frac{k}{2}\log n$$

  [Rissanen 1984]

→ 理論下限によりアルゴリズムの性能を評価できる

手法 1

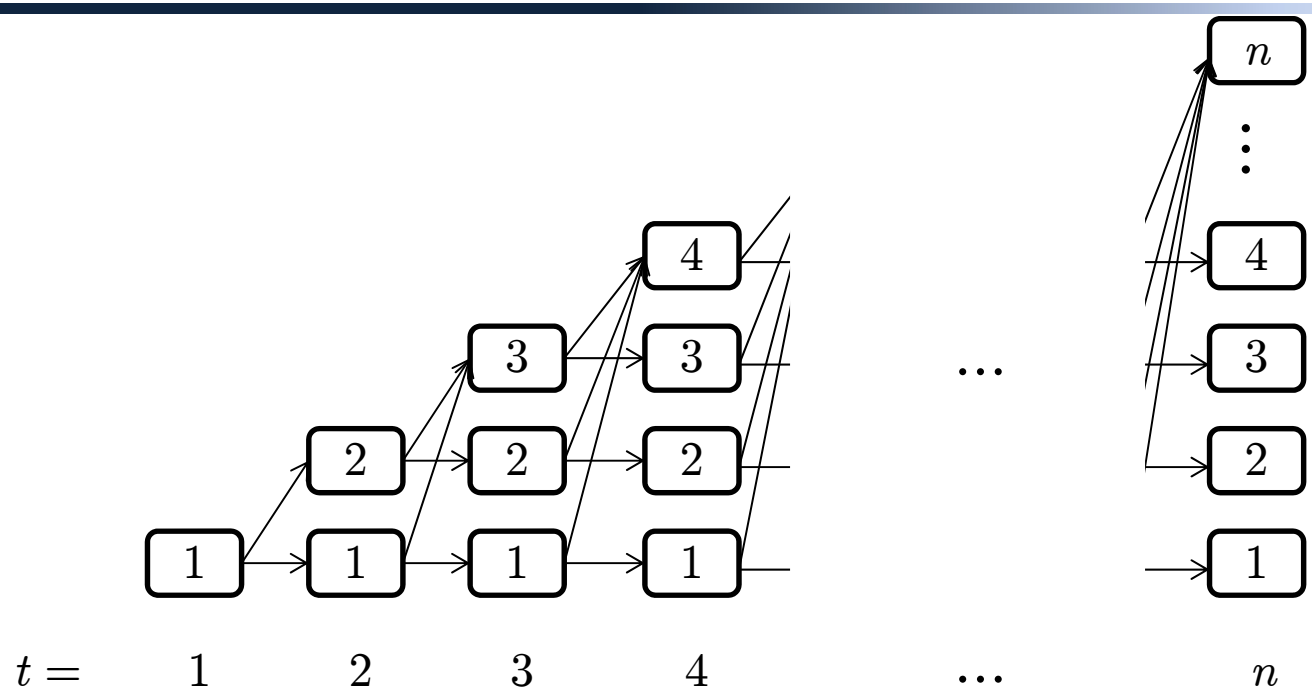
- 逐次推定法に基づく変化検出手法
  - 確率分布の逐次推定
    - 例 : 逐次 Bayes 推定

$$P_{\mathsf{B}}(x \mid \emptyset) = f(x; \theta_0),$$

$$P_{\mathsf{B}}(x \mid x_{t_1}^{t_2}) = \frac{\int p(\theta) f(x;\theta) \prod_{t=t_1}^{t_2} f(x_t;\theta) \mathrm{d}\theta}{\int p(\theta) \prod_{t=t_1}^{t_2} f(x_t;\theta) \mathrm{d}\theta}$$

- 関連研究
  - [Merhav 1993]
  - [Willems 1996], [Shamir and Merhav 1999]
    - 逐次 Bayes 推定 + リセット確率
    - 対象: 有限離散分布
  - [Sakurai and Yamanishi 2012]
    - SNML 推定 + リセット確率
    - 対象: 正規分布

手法 1: 概要

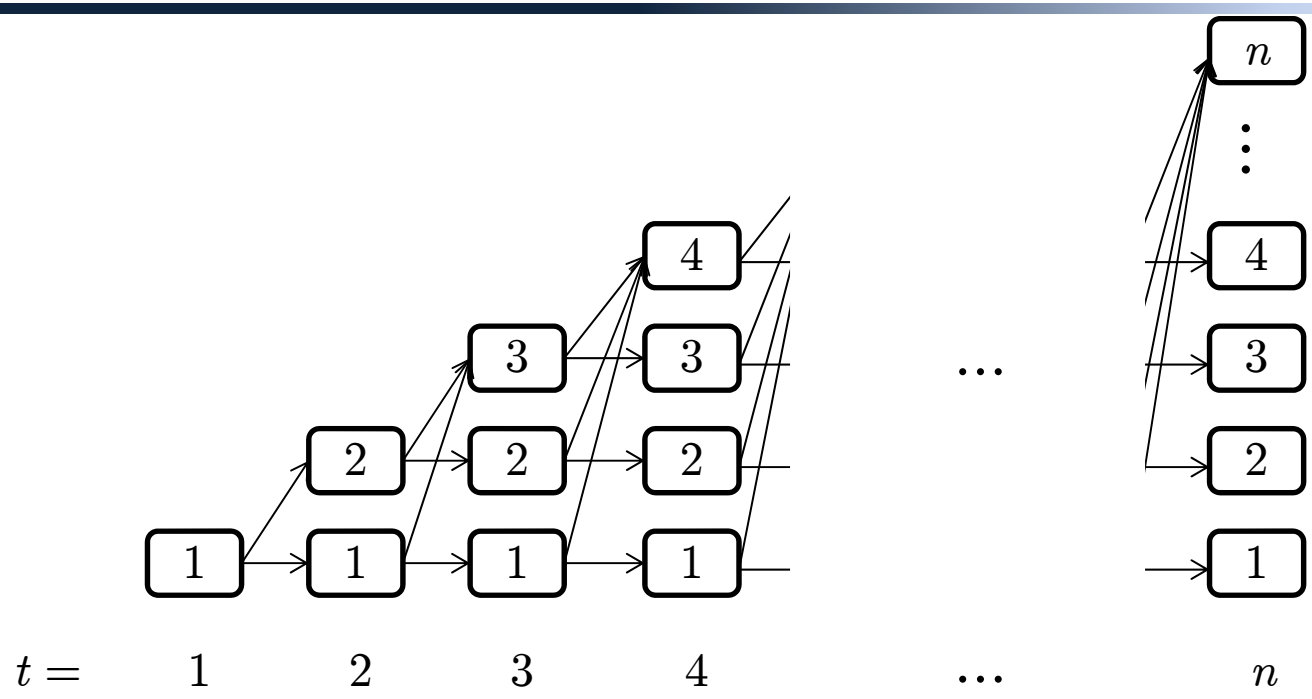


リセット確率付き逐次推定

$$P_{\mathcal{A}}(s_t \mid s_{t-1}) = \begin{cases} P_{\text{reset}}(t \mid s_{t-1}) & s_t = t \\ 1 - P_{\text{reset}}(t \mid s_{t-1}) & s_t = s_{t-1} \end{cases}$$

$$P_{\mathcal{A}}(x_t \mid s_t) = \begin{cases} P_{\text{update}}(x_t \mid \emptyset) & s_t = t \\ P_{\text{update}}(x_t \mid x_{s_t}^{t-1}) & s_t = s_{t-1} \end{cases}$$

手法 1: 概要



DMS を適用すると

$$-\log P_{\mathcal{A}}(x^n, s^n)$$

$$= \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(x_t \mid s_t) \big) + \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(s_t \mid s_{t-1}) \big)$$

$\rightarrow$ minimize w.r.t. $s^n = s_1 \ldots s_n$

手法 1: 性能

- 手法 1: リセット確率付き逐次推定 の期待冗長度
  - リセット確率
    - Willems 推定量

$$P_{\mathrm{reset}}(t \mid s^{t-1}) = \frac{1/2}{t - s_{t-1}}$$

    - Shamir and Merhav 推定量

$$P_{\mathrm{reset}}(t \mid s^{t-1}) = \frac{\pi(t - s_{t-1})}{Z_\infty - Z_{s_{t-1}-1}}$$

$$\begin{cases} \pi(j) = 1/j^{1+\varepsilon} \\ Z_\infty = \sum_{j=1}^{\infty} \pi(j) \\ Z_t = \sum_{j=1}^{t} \pi(j) \end{cases}$$

  - 逐次推定
    - 逐次 Bayes 推定
      - 事前分布に Jeffreys 分布を用いる
      - 有限離散分布では簡単に求まる(Krichevsky and Trofimov 推定)

$$P_{\mathrm{update}}(x = a \mid x_{t_1}^{t_2}) = \frac{N_a(x_{t_1}^{t_2}) + 1/2}{t_2 - t_1 + |\mathcal{X}|/2}$$

手法 1: 性能

- 手法 1: リセット確率付き逐次推定 の期待冗長度
  - リセット確率
    - Willems 推定量

$$P_{\mathrm{reset}}(t \mid s^{t-1}) = \frac{1/2}{t - s_{t-1}}$$

    - Shamir and Merhav 推定量

$$P_{\mathrm{reset}}(t \mid s^{t-1}) = \frac{\pi(t -}{Z_\infty - Z_{s_t}}$$

  - 逐次推定
    - 逐次 Bayes 推定
      - 事前分布に Jeffreys 分布を用いる
      - 有限離散分布では簡単に求まる(Krichevsky and Trofimov 推定)

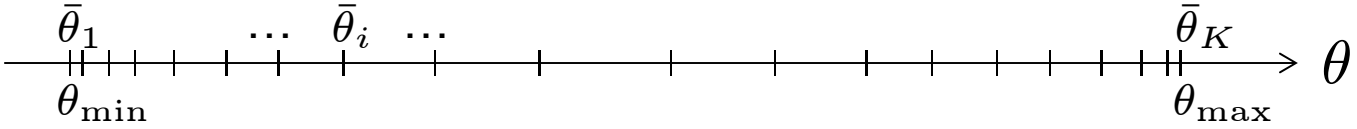
$$P_{\mathrm{update}}(x = a \mid x_{t_1}^{t_2}) = \frac{N_a(x_{t_1}^{t_2}) + 1/2}{t_2 - t_1 + |\mathcal{X}|/2}$$

(理論上)
期待冗長度として最適
Merhav の下限に
(漸近的に)一致

手法 2

- パラメータ空間離散化に基づく変化検出手法
  - パラメータ空間離散化

1 次元パラメータ空間の離散化イメージ

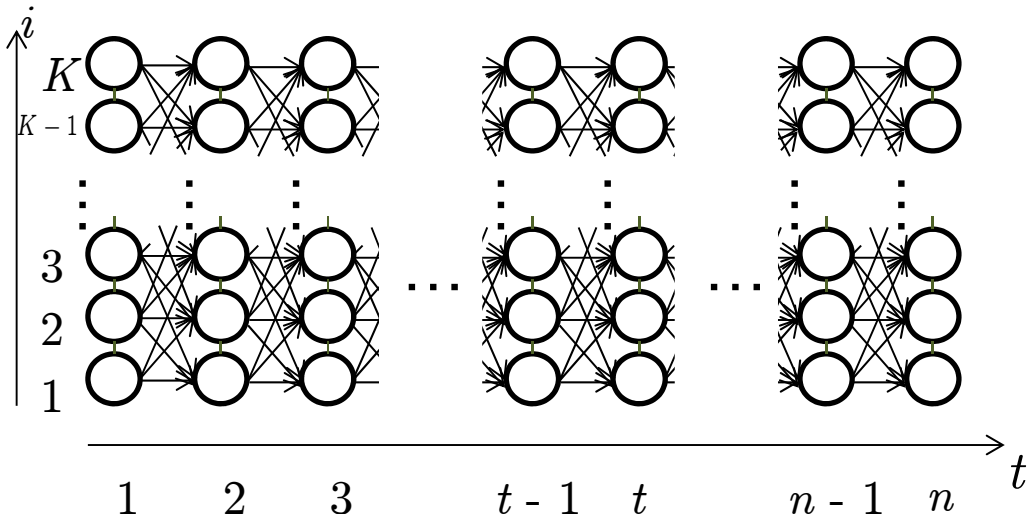


- 離散化点数 K が過多でも過小でも期待冗長度を上げる

- 関連研究
  - [Kleinberg 2003]
    - テキストデータマイニングでの研究（MDL 分野の研究ではない）
  - [Kanazawa and Yamanishi 2012], [金澤・山西 2012]

手法 2: 概要

- コンパクトなパラメータ空間を離散化
- 状態遷移図



- DMS を適用

$$
\begin{aligned}
& -\log P_{\mathcal{A}}(x^n \mid i^n) \\
&= \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(x_t \mid i_t) \big) + \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(i_t \mid i_{t-1}) \big) \\
&= \sum_{t=1}^{n} \big(-\log f(x_t; \bar{\theta}_{i_t}) \big) + \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(i_t \mid i_{t-1}) \big)
\end{aligned}
$$

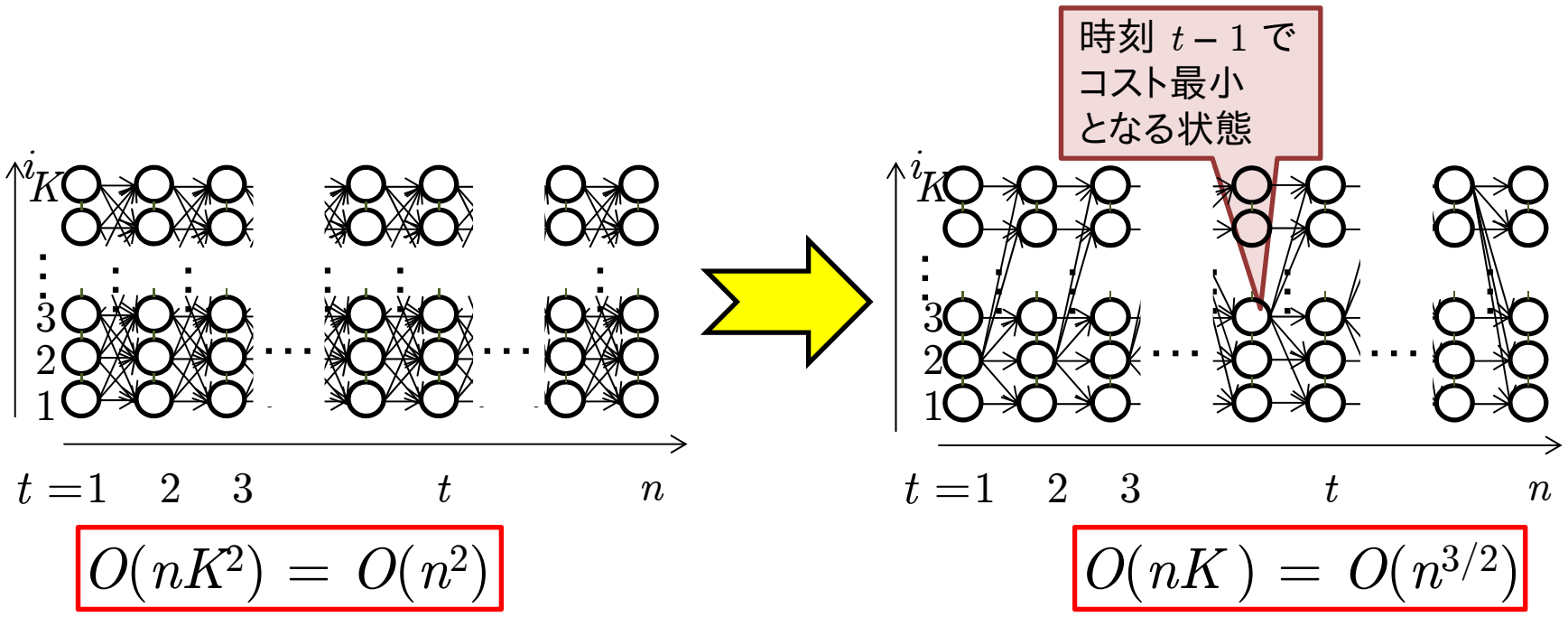
$$
\rightarrow \text{minimize w.r.t. } i^n = i_1 \dots i_n
$$

手法 2: 詳細と議論

- 遷移確率の設定
  - 一様遷移確率を入れると計算時間を削減できる

$$P_{\mathcal{A}}(i_t \mid i_{t-1}) = \begin{cases} \alpha/(K-1) & i_t \neq i_{t-1} \\ 1-\alpha & i_t = i_{t-1} \end{cases}$$

$$\alpha = 1/n$$



手法 2: 詳細と議論

- 離散化について
  - 1 次元では Fisher 情報量を用いて

$$\hookrightarrow I(\theta) = \mathbb{E}[-\partial^2 \log f(x;\theta)/\partial\theta^2]$$

離散化幅 $\delta_I = \int_{\theta_{\min}}^{\theta_{\max}} \sqrt{I(\theta)}\mathrm{d}\theta/(K-1)$

離散化点 $\bar{\theta}_i$: $\int_{\theta_{\min}}^{\bar{\theta}_i} \sqrt{I(\theta)}\mathrm{d}\theta = (i-1)\delta_I \quad (i = 1, \ldots, K)$

とし，$K = O(\sqrt{n})$ とすると最適な離散化

→ 期待冗長度を抑える

  - 多次元パラメータ空間に対しては難しい
    - 直交しているパラメータを取ることができるか否か
    - 直交パラメータであれば上の離散化法の直積をとった離散化が可能
    - 2 次元指数型分布族ならば e-, m- 接続のトリックが使える

手法 2: 詳細と議論

- 手法 2 の期待冗長度評価
  ⇔ 離散化手法の最適性

$$\mathcal{R}_{\mathcal{A}}^{(n)} = \mathbb{E}\Bigg[\sum_{t=1}^{n} \big(-\log f(x_t; \bar{\theta}_{i_t}) \big) + \sum_{t=1}^{n} \big(-\log P_{\mathcal{A}}(i_t \mid i^{t-1}) \big) - \sum_{p=0}^{c} \sum_{t=m(p)+1}^{m(p+1)} \big(-\log f(x_t; \theta(p)) \big) \Bigg]$$

上限

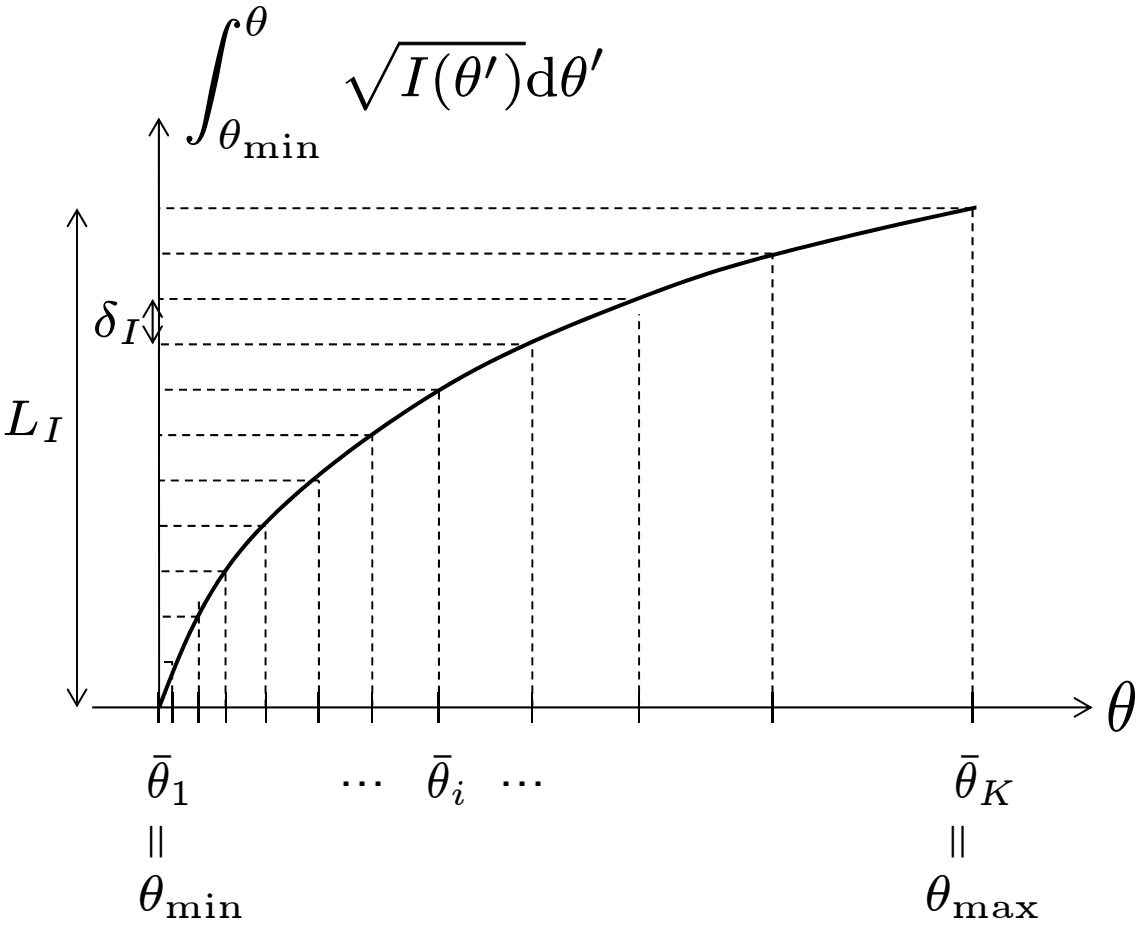
$$\sum_{p=0}^{c} \sum_{t=m(p)+1}^{m(p+1)} \underbrace{\mathbb{E}_{\theta(p)}\big[-\log f(x_t; \bar{\theta}_{i_t}) - \big(-\log f(x_t; \theta(p)) \big) \big]}_{= D\left(\theta(p) \| \bar{\theta}_{i_t}\right)} + \log K + c \log(K-1) - c \log \alpha - (n-c) \log(1-\alpha)$$

ポイント

- Kullback—Leibler ダイバージェンスを小さくする離散化
- (準)最適な離散化点数 $K = O(\sqrt{n})$ (1 次元)

手法 2: 詳細と議論

- 手法 2 の期待冗長度評価
  ⇔ 離散化手法の最適性



$$\max_{\theta\in[\theta_{\min},\theta_{\max}]} \min_{\bar{\theta}_i} D(\theta\|\bar{\theta}_i)$$

$$\frac{I(\theta)}{2}(\theta-\bar{\theta}_i)^2 + O(|\theta-\bar{\theta}_i|^3)$$

$$\& \sqrt{I(\theta)}|\theta-\bar{\theta}_i| < \delta_I$$

$$< \frac{(\delta_I)^2}{2} + O\big((\delta_I)^3\big)$$

真のパラメータから
離散化点までの
KL ダイバージェンスが
一様に抑えられる
→ **最適 K 点離散化**

手法 2: 詳細と議論

- 手法 2 の期待冗長度評価
  - 離散化手法 $\bar{\theta}_i$: $\int_{\theta_{\min}}^{\bar{\theta}_i} \sqrt{I(\theta)}\mathrm{d}\theta = (i-1)\delta_I$
  - 離散化点数 $K = \lfloor \sqrt{n} \rfloor$
  - 遷移パラメータ $\alpha = 1/n$

$$\Rightarrow \mathcal{R}_{\mathcal{A}}^{(n)} < \frac{c+1}{2}\log n + c\log n + \frac{\left(\int_{\theta\in\Theta}\sqrt{I(\theta)}\mathrm{d}\theta\right)^2}{2} + \log \mathrm{e} + O(n^{-1/2})$$

→ Merhav の下限（1 次元）に漸近的に一致

  - 課題:
    - 離散化点数 $K = O(\sqrt{n/c})$
    - 遷移パラメータ $\alpha \approx c/n$

が理想

数値実験

- Bernoulli 分布 $\mathcal{F}_{\mathrm{Ber}} = \{\theta^x(1-\theta)^{1-x}\}$ for $x = 0, 1$
  - $n = 500, 1000, 5000$ で実験

$$\theta = \begin{cases} 0.8 & (1 \le t \le 0.2n) \\ 0.2 & (0.2n+1 \le t \le 0.6n) \\ 0.1 & (0.6n+1 \le t \le 0.8n) \\ 0.4 & (0.8n+1 \le t \le n) \end{cases}$$

  - 各 n に対し 200 回のシミュレーション
    - 手法 1-W: Jeffreys 事前分布 逐次 Bayes 推定 + Willems 推定量
    - ~~手法 1-SM: Jeffreys 事前分布 逐次 Bayes 推定 + SM 推定量~~
    - 手法 2-KY: 前ページの離散化手法・離散化点数・遷移パラメータ
  - 記述長・FAR・benefit を見る

FAR → 誤報率

benefit → 各 真の変化点について $\max\{1 - |t_c - t_c^*|/R, 0\}$ をとった平均（今回は $R = 0.01n$ ）

数値実験：結果

$n = 500$

	記述長	FAR	benefit
手法 1-W	275.7656	32.23 %	0.3210
手法 2-KY	270.5700	22.91 %	0.3520

$n = 1000$

	記述長	FAR	benefit
手法 1-W	532.7412	27.87 %	0.4763
手法 2-KY	525.8352	16.51 %	0.4848

$n = 5000$

	記述長	FAR	benefit
手法 1-W	2550.0200	39.11 %	0.7778
手法 2-KY	2539.3914	8.33 %	0.7825

結論

- 問題設定：区間定常無記憶情報源 (PSMS)

- PSMS に対する変化検出手法の紹介
  - 手法 1: 逐次推定法 + リセット確率
  - 手法 2: パラメータ離散化 + 動的計画法

- 数値実験による比較
  - 手法 1 は FAR が多くなる
    - 変化直後の変化確率が大きいため
  - その他は 手法 2-KY のほうが 手法 1-W に比べやや優れている

参考文献 (1/2)

- Clarke, B.S., and Barron, A.R. “Information-theoretic asymptotics of Bayes methods,” *IEEE Trans. Inform. Theory*, vol. 36, pp. 453-471, May 1990.
- Kanazawa, H., and Yamanishi, K., “An MDL-based change-detection algorithm with its applications to learning piecewise stationary memoryless sources,” IEEE ITW2012, Sept. 2012, pp. 562-566.
- Kleinberg, J. “Bursty and hierarchical structure in streams,” *D. M. K. D.*, vol. 7, pp. 373-397, Nov. 2003.
- Krichevsky and Trofimov “The performance of universal encoding,” *IEEE Trans. Inform. Theory*, vol. 27, pp. 199-207, Mar. 1981.
- Merhav, N. “On the minimum description length principle for sources with piecewise constant parameters,” *IEEE Trans. Inform. Theory*, vol. 39, pp. 1962-1967, Nov. 1993.
- Rissanen, J. “Universal coding, information, prediction, and estimation,” *IEEE Trans. Inform. Theory*, vol. 30, pp. 629-636, July 1984.
- Rissanen, J. *Information and complexity in statistical modeling.* Springer, New York, 2007.

参考文献 (2/2)

- Sakurai, E., and Yamanishi, K. “Comparison of dynamic model selection with infinite HMM for statistical model change detection,” IEEE ITW2012, Sept. 2012, pp. 302-306.
- Shamir, G.I., and Merhav, N. “Low complexity sequential lossless coding for piecewise stationary memoryless sources,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 1498-1519, July 1999.
- Willems, F.M.J. “Coding for a binary independent piecewise-identically-distributed source,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 2210-2217, Nov. 1996.
- Yamanishi, K., and Maruyama, Y. “Dynamic syslog mining for network failure monitoring,” KDD2005, Aug. 2005, pp. 499-508.
- Yamanishi, K., and Maruyama, Y. “Dynamic model selection with its applications to novelty detection,” *IEEE Trans. Inform. Theory*, vol. 53, pp. 2180-2189, June 2007.
- 金澤 宏紀, 山西 健司 “多次元パラメータを有する区間定常無記憶情報源に対しての MDL 原理に基づく変化検出アルゴリズム, ” 第 9 回 IBISML 研究会, 2012 年 6 月.