SITA2025

情報は確率変数 + 情報の生成メカニズムも確率変数

Shannon Modeled Information as a Random Variable + What if We also Modeled the Information Generation Mechanism as a Random Variable?

早稲田大学　松嶋敏泰

最近のAI･機械学習の学会のタイトルの流行り

AISTATS 2025　タイトルで用いられる用語の傾向

Neural /Net：３４件
Statistical：１１件

Information Theory：１３件
Entropy：７件

Bayes：３９件
Posterior：８件

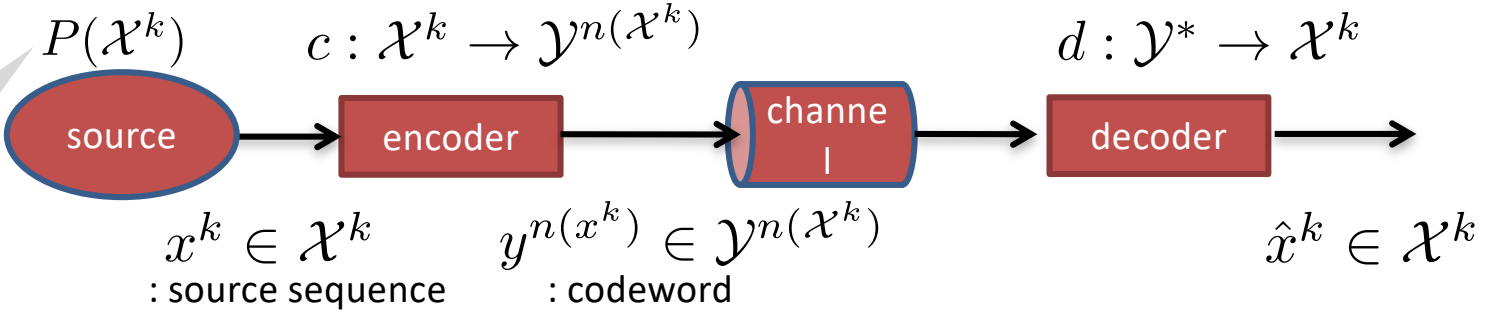
内容については？ですが

情報理論(Information Theory)

C.E.Shannon (1948)

“A Mathematical Theory of Communication”

例）FV ブロック符号



情報を確率変数として定式化

例）FV source coding theorem

FV 符号化定理：

無歪みの FV 符号化で以下の期待符号長を満たすものが存在する。

$$E[n(X)] < \frac{H(X)}{\log |Y|} + 1$$

又は $$\frac{1}{k} E[n(X^k)] < \frac{H(X)}{\log |Y|} + \frac{1}{k}$$ （k 次拡大符号）

基本的には
分布 $P(\mathcal{X}^k)$ は既知

大数の法則
大偏差原理
中心極限定理 等
により美しい理論を
構築

FV 符号化逆定理：

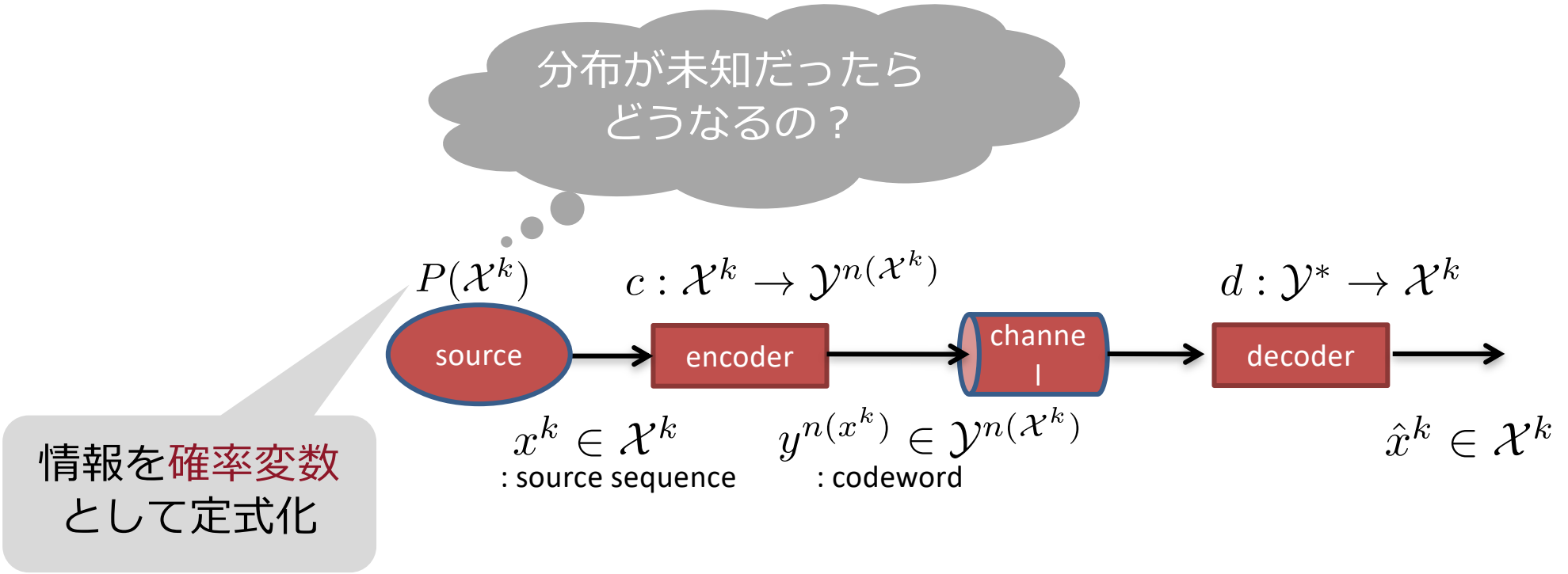
無歪みの FV 符号化ならば、期待符号長は以下を満たさなければならない。

$$E[n(X)] \geq \frac{H(X)}{\log |Y|}$$

情報理論(Information Theory) 分布未知の場合

C.E.Shannon (1948)

“A Mathematical Theory of Communication”



パラメトリック分布でのパラメータ θ 未知の場合の対応

Known ← a parametric distribution: $P(x^n|\theta)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

eg) アルファベット数 k+1 のカテゴリカル分布

最尤推定量 $\hat{\theta}$ をプラグインした確率分布 $P(x^n|\hat{\theta})$ を用いて
分布既知の場合と同じ符号化すれば良い

最尤推定量の漸近一致性や漸近正規性の性質を用い期待
符号長などのエントロピーへの収束が議論できる

未知のパラメータ θ も一階層上の情報なので確率変数と考えればいい（メタ情報）

Θ：未知の定数　-> 最尤推定量（頻度論的統計学）
Θ：未知の確率変数　-> ベイズ決定理論からの推定量（ベイズ統計学）

パラメータ θ も送りたい情報 X より
一階層上の情報(メタ情報)と考えれば
確率変数として扱ったら

全称限定子付：$\forall x\ Q(x) \rightarrow R(x)$
アトム：$Q(a), R(b)$

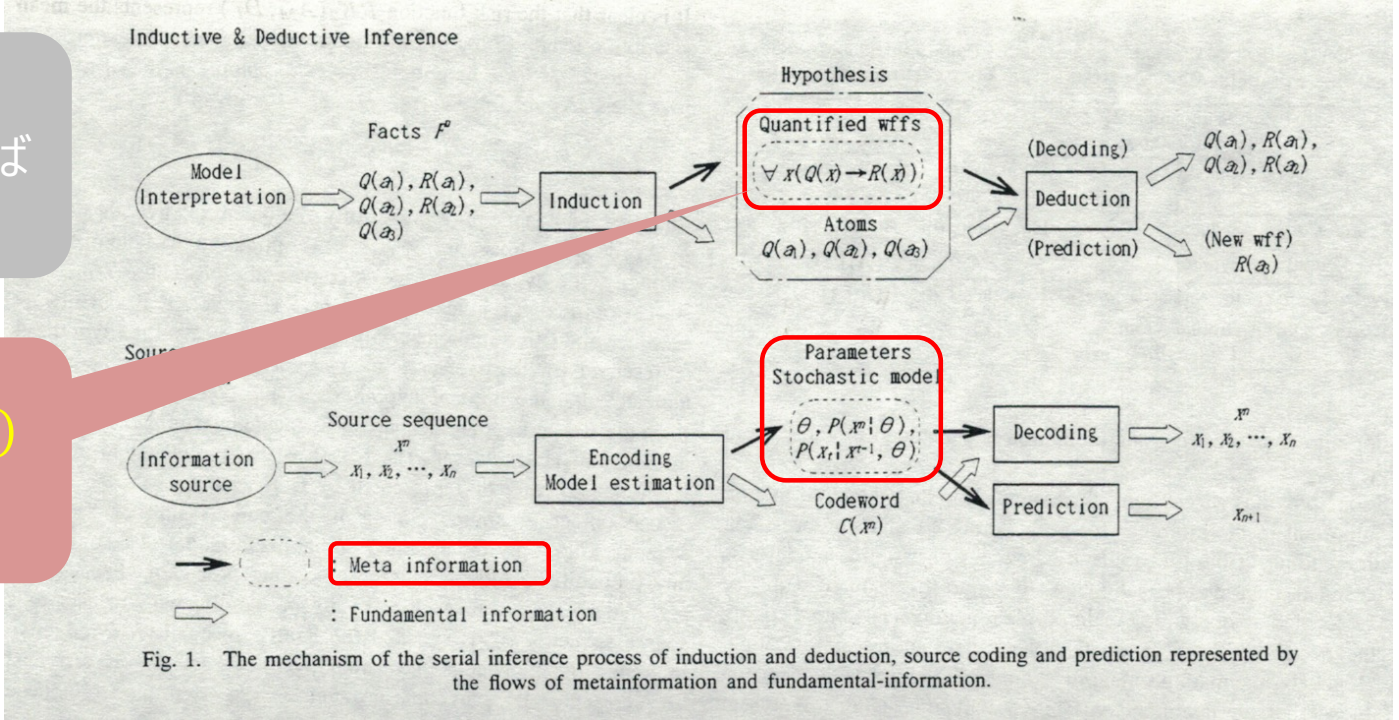
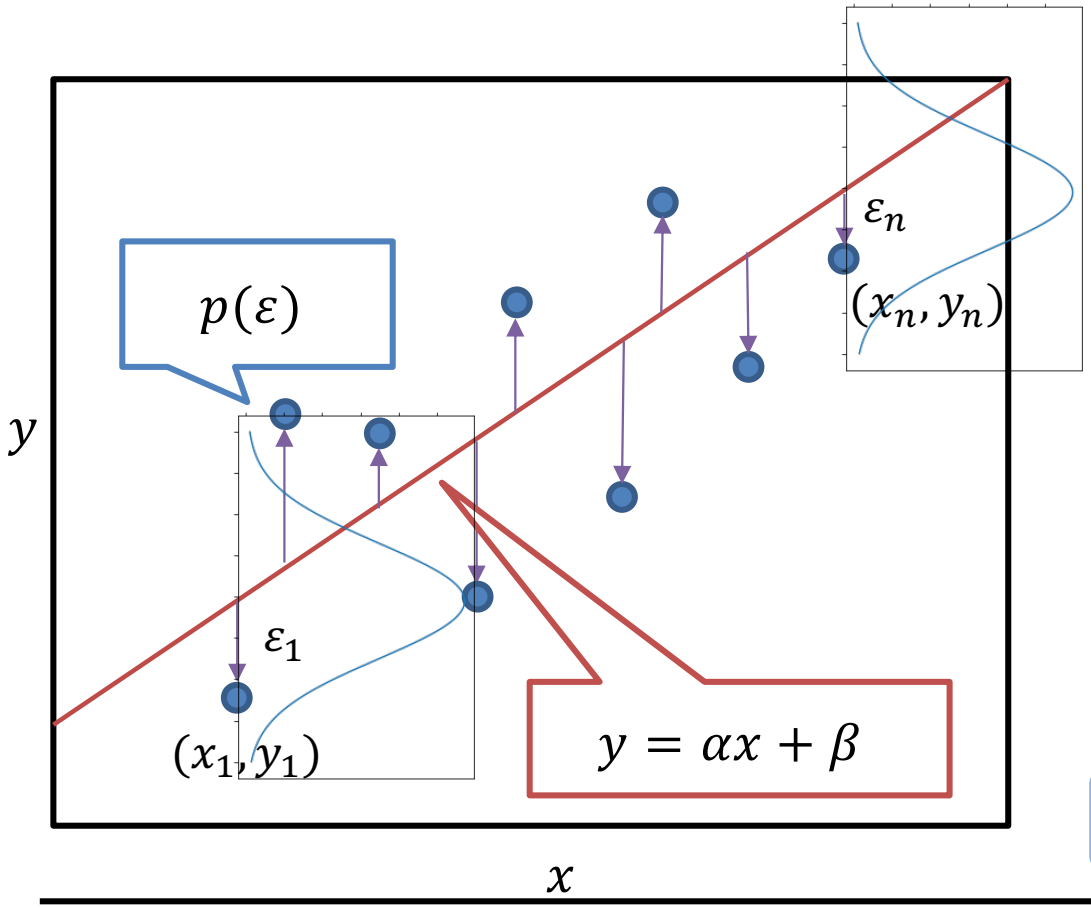


Fig. 1. The mechanism of the serial inference process of induction and deduction, source coding and prediction represented by the flows of metainformation and fundamental-information.

[Matsushima et al. 88] Representation of Logic Model and Meta-Information, SITA 88.
[Matsushima et al. 89] Applications of Information Theory into Predicate Logic for AI systems, IEICE TRANSACTIONS.
[松嶋,平澤 92] 解説：ＭＤＬの帰納推論への応用，人工知能学会誌

メタ情報のパラメータ θ は未知の定数か確率変数か（回帰分析を例に）

- 各(x_i, y_i)が次のような過程で生成されると仮定



- 更に，$p(\varepsilon)$が平均0，分散σ_ε^2の正規分布であると仮定する

$$\varepsilon_i = y_i - (\alpha x_i + \beta)$$

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma_\varepsilon^2}\right)$$

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(y_i - (\alpha x_i + \beta))^2}{2\sigma_\varepsilon^2}\right)$$

確率変数

構造推定（パラメータ推定）の意思決定写像

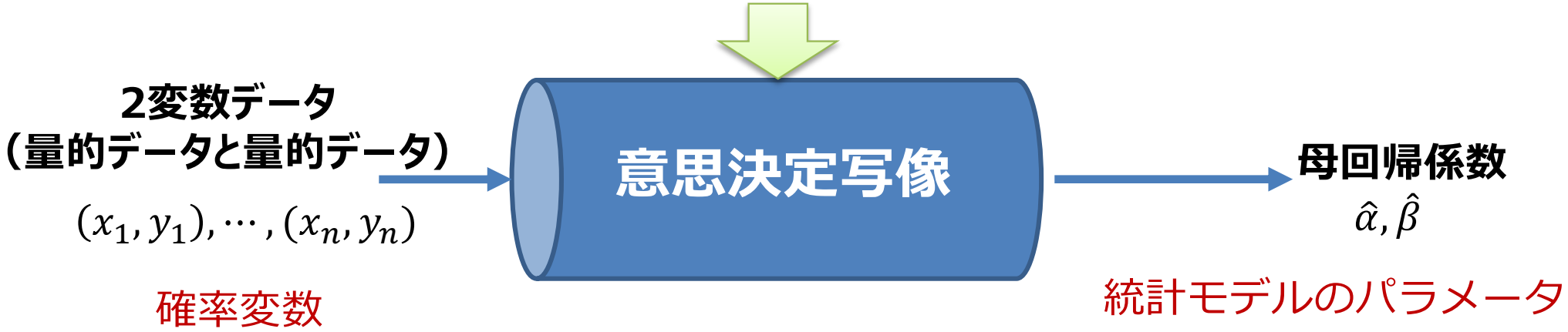
目的：　説明変数xと目的変数yの生成メカニズムを線形関係で表現する

$$y = \hat{\alpha}x + \hat{\beta}$$

を求める

設定：　$y_i = \alpha x_i + \beta + \varepsilon_i$,　ε_iは独立に分布$p(\varepsilon)$に従う　α, β は未知の定数

評価基準：　α, β の推定量の評価基準（例えば尤度最大，不偏性，一致性）



母回帰係数の最尤推定

- 尤度関数の値を最大にするα, βを求める ⇒α, βの最尤推定

尤度関数：$L(\alpha, \beta, \sigma_\varepsilon^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(y_i - (\alpha x_i + \beta))^2}{2\sigma_\varepsilon^2}\right)$

対数尤度関数：$\ln L(\alpha, \beta, \sigma_\varepsilon^2) = -\frac{n}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - (\alpha x_i + \beta))^2$

- 対数尤度関数の値を最大にするα, β：赤線部分を最小にするα, β

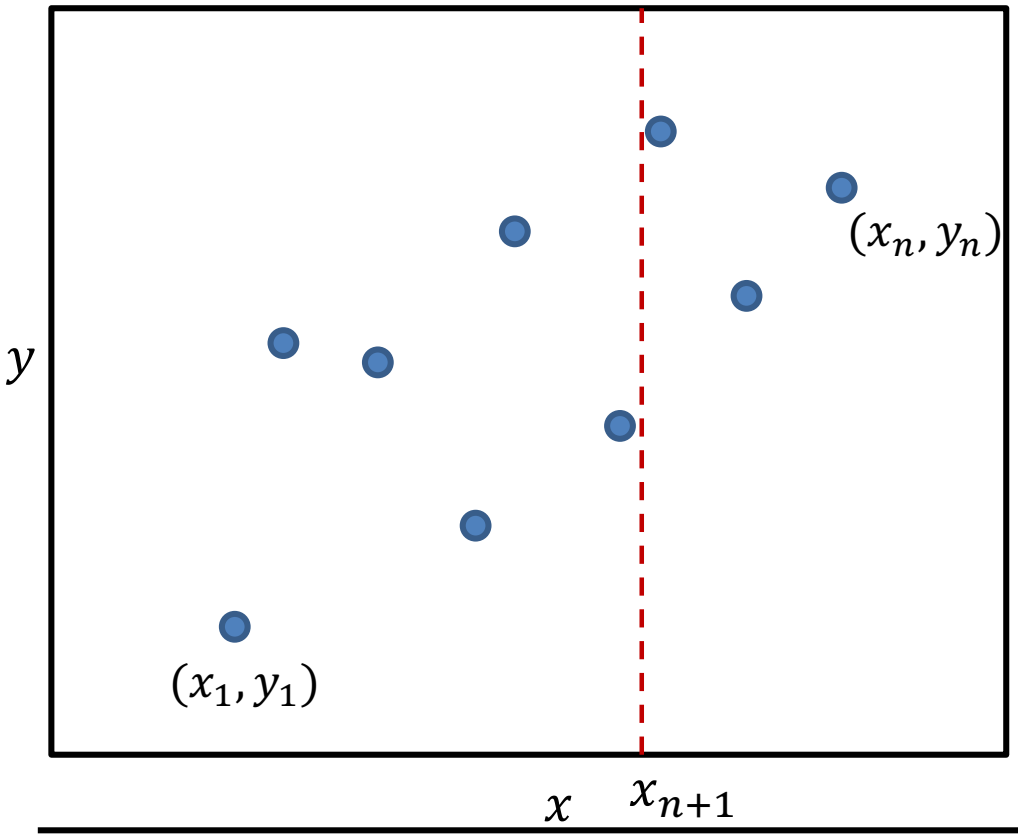
最尤推定量

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x}$$

予測をするという意思決定写像

- n組のデータ$(x_1, y_1), \cdots, (x_n, y_n)$



No.	身長	体重
1	181.1	95.3
2	169.0	58.2
・ ・ ・	・ ・ ・	・ ・ ・
n	168.0	51.3
$n+1$	**174.9**	**?**

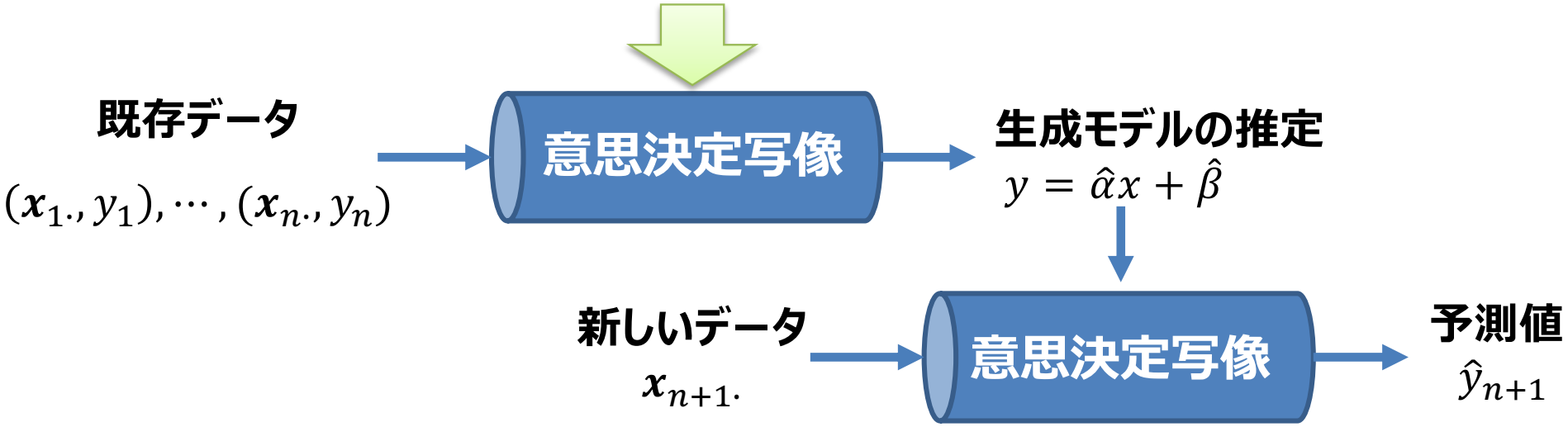
- 与えられたx_{n+1}に対してy_{n+1}を予測する

間接予測の意思決定写像

目的：　データの生成分布パラメータ α, β をまず推定してから $\hat{y}_{n+1}$ を予測

設定：　$y_i = \alpha x_i + \beta + \varepsilon_i$,　ε_iは独立に分布$p(\varepsilon)$に従う

評価基準：α, β の推定の良さに対する様々な評価基準（例えば尤度最大）



直接予測の意思決定写像

目的：　説明変数x_{n+1}に対する目的変数y_{n+1}の値を予測したい

設定：　$y_i = \alpha x_i + \beta + \varepsilon_i, \varepsilon_i$は分布$p(\varepsilon)$に従う

評価基準：　予測値 $\hat{y}_{n+1}$に対する損失関数，危険関数等

[松嶋 98] 帰納・演繹推論と予測－決定理論による学習モデル, 第１回情報論的学習理論ワークショップ予稿集（IBIS1998）pp.1-8

一階述語論理の間接予測（帰納推論+演繹推論）と直接予測

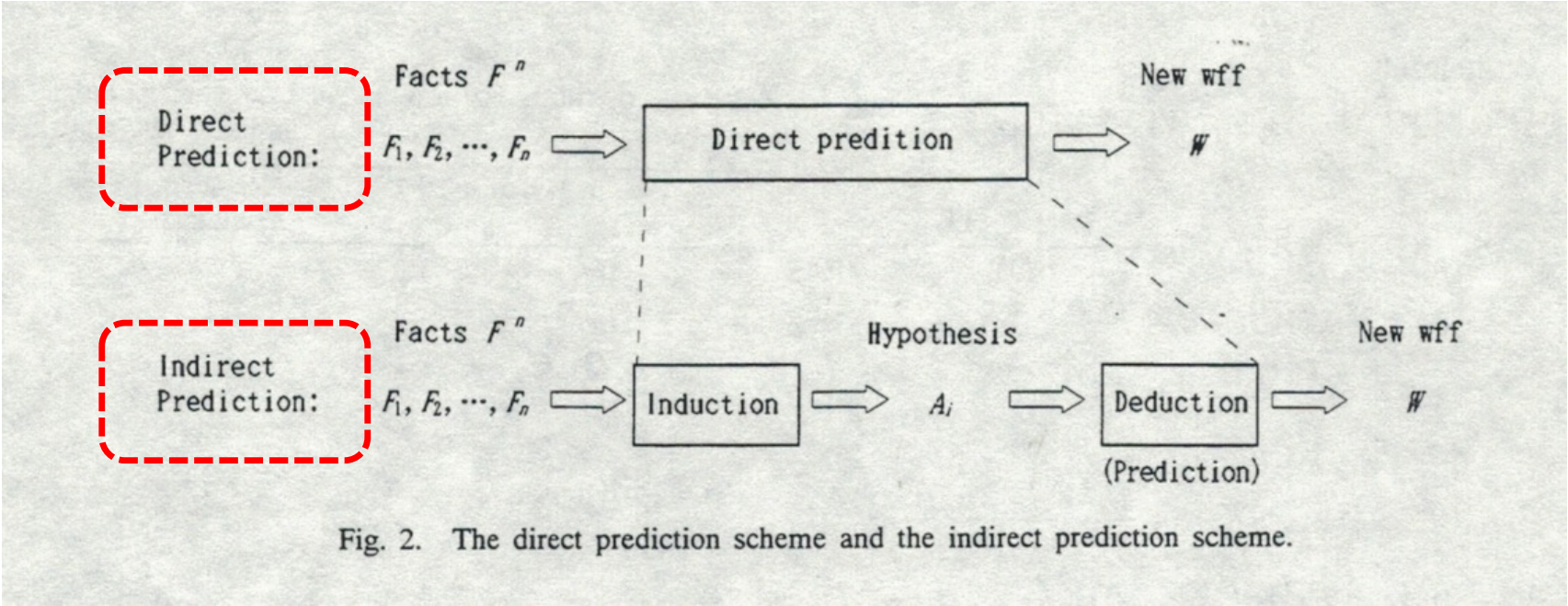


Fig. 2. The direct prediction scheme and the indirect prediction scheme.

[Matsushima et al. 90] Inductive Inference Scheme at a Finite Stage of Process from View Point of Source Coding, IEICE TRANSACTIONS.
[Matsushima et al. 93] Inductive Inference Procedure to Minimize Prediction Error, IEEE Transaction on System, Man & Cybernetics.

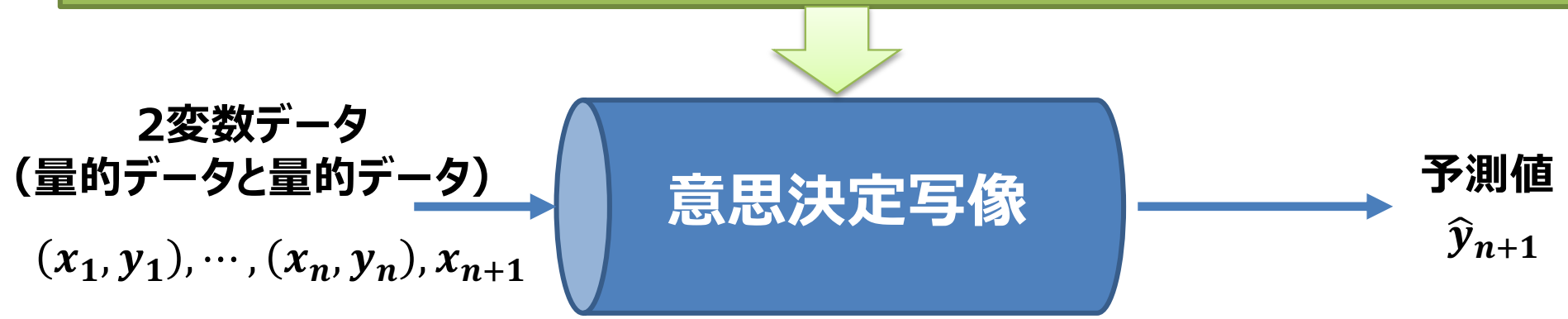
メタ情報 α,β を確率変数とした直接予測の意思決定写像（ベイズ決定理論）

目的：　説明変数x_{n+1}に対する目的変数y_{n+1}の値を予測したい

設定：　$y_i = \alpha x_i + \beta + \varepsilon_i, \varepsilon_i$は分布$p(\varepsilon)$に従う　α,β は未知の確率変数

メタ情報 α,β は確率変数で事前分布 $p(\alpha), p(\beta)$ に従う

評価基準：　予測値 $\hat{y}_{n+1}$に対する損失関数，危険関数，ベイズ危険関数等



[松嶋 98] 帰納・演繹推論と予測－決定理論による学習モデル, 第１回情報論的学習理論ワークショップ予稿集（IBIS1998）pp.1-8

ベイズ最適な予測と予測分布

例）二乗誤差損失：$\ell'(y_{n+1}, d) = (y_{n+1} - d)^2$

定理：
損失$\ell'(y_{n+1}, d)$の損失関数：

$$\ell(Y_{n+1}, d) = \int \ell'(y_{n+1}, d) p(y_{n+1}; \alpha, \beta, \sigma_\varepsilon^2) \mathrm{d}y_{n+1}$$

とした場合，ベイズ最適（ベイズ危険関数最小）なdは以下で与えられる．

$$d^*\left((x_1, y_1), \cdots (x_n, y_n), x_{n+1}\right) = \arg\min_d \int \ell'(y_{n+1}, d) \underbrace{p(y_{n+1}|y_1, \cdots, y_n)}_{y_{n+1}\text{の}\textbf{予測分布}} \mathrm{d}y_{n+1}$$

$$p(y_{n+1}|y_1, \cdots, y_n) = \int\int p(y_{n+1}|\alpha, \beta) p(\alpha, \beta|y_1, \cdots, y_n) \mathrm{d}\alpha \mathrm{d}\beta$$

メタ情報パラメータ θ を確率変数とする利点

Θ：未知の定数　-> 最尤推定量（頻度論的統計学）
Θ：未知の確率変数　-> ベイズ決定理論からの推定量（ベイズ統計学）

パラメータ θ も一階層上の情報
(メタ情報)なので
確率変数として扱ったら

目的：　説明変数x_{n+1}に対する目的変数y_{n+1}の値を予測したい

設定：　$y_i = \alpha x_i + \beta + \varepsilon_i, \varepsilon_i$は分布$p(\varepsilon)$に従う　α, β は未知の確率変数
メタ情報 α, β は確率変数で事前分布 $p(\alpha), p(\beta)$ に従う

評価基準：　予測値 $\hat{y}_{n+1}$に対する損失関数，危険関数，ベイズ危険関数等

メタ情報は確率変数 ＋ ベイズ決定理論 ＋ 直接予測

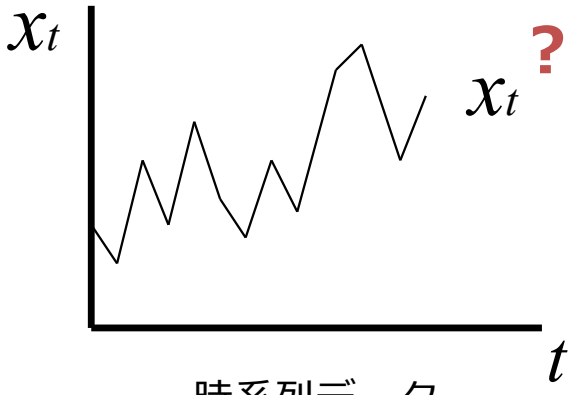
[松嶋 98] 帰納・演繹推論と予測－決定理論による学習モデル, 第１回情報論的学習理論ワークショップ予稿集（IBIS1998）pp.1-8
[Matsushima et al. 88] Representation of Logic Model and Meta-Information, SITA 88.
[Matsushima et al. 89] Applications of Information Theory into Predicate Logic for AI systems, IEICE TRANSACTIONS.
[松嶋,平澤 92] 解説：ＭＤＬの帰納推論への応用，人工知能学会誌

時系列予測：予測値（意思決定写像の出力）として分布も

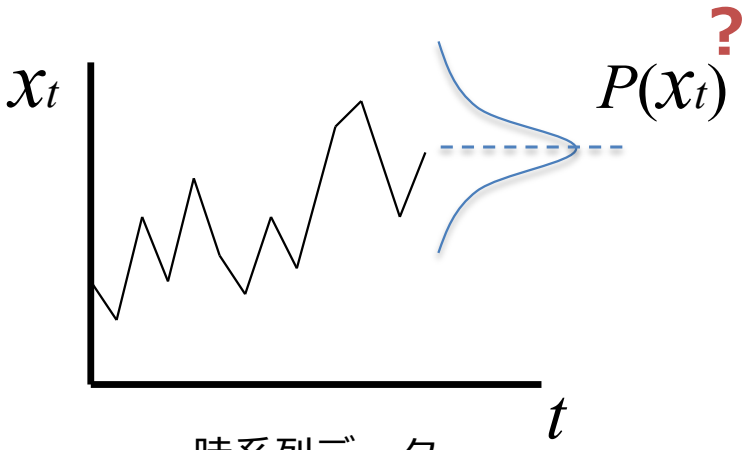
$$X = \{a,b,c\}$$

$$aabcba \rightarrow x_t\,?$$

次に出現するシンボルの予測



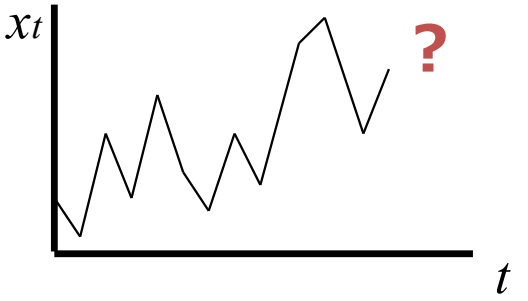
時系列データ



時系列データ

情報理論の問題に戻りましょう：FV 逐次符号と時系列予測

時系列予測： $x_1, x_2, \cdots, x_{t-1} \Rightarrow x_t$



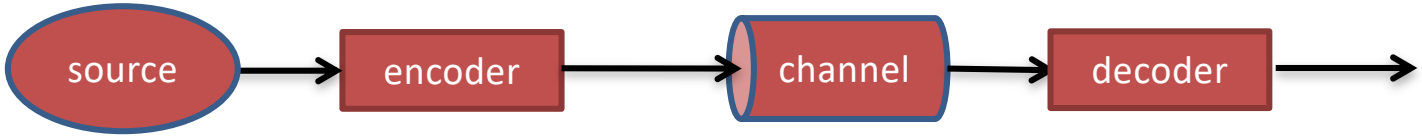
予測値： $\hat{P}(X_t|x^{t-1})$

同値な問題

危険関数(K–L情報量)：

$$l_{KL}(\hat{P}(X_t|x^{t-1}), P(X_t|x^{t-1})) = \sum_{X_t} P(x_t|x^{t-1}) \log \frac{P(x_t|x^{t-1})}{\hat{P}(x_t|x^{t-1})}$$

FV逐次符号化：



時刻 t

$$c_1 : \quad \mathcal{X}_1 \to \mathcal{Y}^{n(\mathcal{X}_1)} \qquad n(x_1) = -\log P(x_1)$$
$$c_2 : \quad \mathcal{X}_2 \to \mathcal{Y}^{n(\mathcal{X}_2)} \qquad n(x_2) = -\log P(x_2|x_1)$$
$$\vdots$$
$$c_k : \quad \mathcal{X}_k \to \mathcal{Y}^{n(\mathcal{X}_k)} \qquad n(x_k) = -\log P(x_k|x^{k-1})$$

理想逐次符号長

情報源の構造が未知の場合の対応：パラメトリック分布でのパラメータ未知の場合

Known ← a parametric distribution: $P(x^n|\theta)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

最尤推定量 $\hat{\theta}$ をプラグインした確率分布 $P(x^n|\hat{\theta})$ を用いて
分布既知の場合と同じ符号化すれば良い

最尤推定量の漸近一致性や漸近正規性の性質を用い期待
符号長などのエントロピーへの収束が議論できる

Θ：未知の定数　-> 最尤推定量（頻度論的統計学）
Θ：未知の確率変数　-> ベイズ決定理論からの推定量（ベイズ統計学）

シャノン的？には

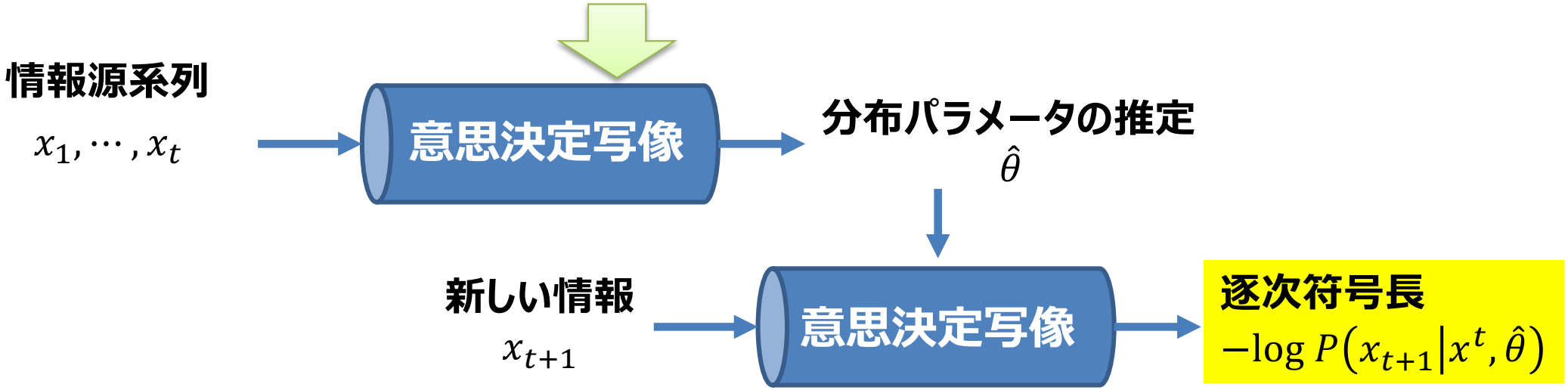
情報 X より一階層上の情報(メタ情報)であるので確率変数として考えればいい

間接符号化の意思決定写像

目的：　パラメータ θ をまず推定してから逐次符号長 $-\log P(x_{t+1}|x^t,\theta)$を決定

設定：　情報源は確率分布$P(x_{t+1}|x^t,\theta)$に従う　メタ情報 θは未知定数

評価基準：θ の推定の良さに対する様々な評価基準（例えば尤度最大）



直接符号化のベイズ決定理論のよる意思決定写像

目的：　情報源系列 $x_1, \cdots, x_t$から逐次符号長 $-\log P(x_{t+1}|x^t)$を直接決定

設定：　情報源は確率分布$P(x_{t+1}|x^t, \theta)$に従う　メタ情報 θは未知確率変数

評価基準：期待符号長最小（ベイズ危険関数）

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on Information Theory, vol.37

ベイズ決定理論の直接符号化：Bayes codes

Known ← a parametric distribution: $P(x^n|\theta)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

Under the condition that the **Arithmetic Coding** is used for coding, the decision of **encoding probability** $\hat{P}(x^n)$ is the main problem in universal coding.

The optimal solution from the view point of Bayes strategy

ブロック符号も逐次符号も符号長は同じ

Block code: $\hat{P}_B(x^n) = P(x^n) = \int_\Theta P(x^n|\theta)dP(\theta),$

A block sequence x^n is encoded to a code word $C(x^n)$ simultaneously

Predictive code: $\hat{P}_P(x_t|x^{t-1}) = P(x_t|x^{t-1}) = \int_\Theta P(x_t|x^{t-1}),\theta))dP(\theta|x^{t-1})$

A symbol x_t given x^{t-1} is encoded to a code word in order

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on Information Theory, vol.37

The conjugate prior distribution of multinomial (Categorical) distributions

We assume the prior probability of the $(|X|-1)$-dimensional parameter vector θ_s given a state s to be a Direchlet distribution, which is the conjugate prior of multinomial distributions, as follows:

$$P(\theta_s|s) = \frac{\Gamma(\sum_{i=0}^{|X|-1} \beta(i|s))}{\prod_{i=0}^{|X|-1} \Gamma(\beta(i|s))} \prod_{i=0}^{|X|-1} \theta_{i|s}^{\beta(i|s)-1},$$

where $\Gamma(\cdot)$ is the Gamma function.

Lemma 1:

Let $n(i|x^n, s)$ be the number of the occurrences of i under a state s in the source sequence x^n. The predictive Bayes coding probability $P(x_t|x^{t-1}, s)$ is calculated by

$$P(x_t|x^{t-1}, s) = \frac{n(x_t|x^{t-1}, s)) + \beta(x_t|s)}{\sum_{i=0}^{|X|-1} n(i|x^{t-1}, s) + \beta(s)},$$

where $\beta(s) = \sum_{i=0}^{|X|-1} \beta(i|s)$.

例）符号化確率（直接）をベルヌーイ分布で考えてみる

ベータ・ベルヌーイ分布

符号化確率
（直接符号化）：$\hat{P}(x_{t+1}|x^t) = \dfrac{\alpha + n(x_{t+1})}{\alpha + \beta + t}$

$x_{t+1} = 1, \alpha = 1/2, \beta = 1/2$

$\hat{P}(1|x^t) = \dfrac{1/2 + n(x_{t+1})}{1 + t}$

一見 間接符号化
に見えるが

本来 Krichevski-Trofimov 推定量
はこっちの流れ
（CTWで使用）

符号化確率
（間接符号化）：$P\left(x_{t+1}\middle|x^t, \hat{\theta}\right)$

ベルヌーイ分布

$\widehat{\theta_{\mathrm{KT}}} = \dfrac{1/2 + n(x_{t+1})}{1 + t}$

$\widehat{\theta_{\mathrm{ML}}} = \dfrac{n(x_{t+1})}{t}$

最尤推定量

[Krichevsky, Trofimov 81] The Performance of Universal Encoding, IEEE Trans. on Information Theory, vol.27
[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Trans. on Information Theory, vol.37
[Willems et al. 95] The context tree weighting method: basic properties, IEEE Trans. on Information Theory, vol. 41.

未知の確率構造も一階層上の情報（メタ情報）なので確率変数と考えればいい

メタ情報は確率変数 ＋ ベイズ決定理論 ＋ 直接予測

情報の発生メカニズムの情報はパラメータだけでは無い

そう考えるとここまでで説明した上記の概念の
重要性や有用性がさらに明らかになる

確率モデルが未知の設定：重回帰モデルの推定（選択）

統計的モデルとしての回帰：データを発生する構造を仮定（母回帰式）

ここまでの設定：回帰パラメータは未知，母回帰式の構造は既知

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$ の関係（**構造/モデル**）：既知

$\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]^\top$：未知　　→ データから $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p]^\top$を決定

ここからの設定：母回帰式の構造mが未知 (説明変数の一部だけが使われる)

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (m1)\\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (m2)\\
y_i &= \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (m3)\\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (m4)\\
&\cdots\\
y_i &= \beta_0 + \varepsilon_i \quad (mM)
\end{aligned}
$$

n組のデータ

$(y_1, \boldsymbol{x}_{1.}), (y_2, \boldsymbol{x}_{2.}), \cdots (y_n, \boldsymbol{x}_{n.})$

の重回帰**モデル** m も未知

偏回帰係数 $\boldsymbol{\beta}_m$ も未知.

確率モデルの推定（モデル選択）の意思決定写像

設定： $y_i = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$,　　ε_iは分布$p(\varepsilon)$に従う

（ただし、データを生成する重回帰モデル m は未知定数）

多変数データ
（量的データと量的データ）

$(\boldsymbol{x_1}., y_1), \cdots, (\boldsymbol{x_n}., y_n)$

意思決定写像

重回帰モデルの推定

$y_i = \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}^{\mathsf{T}}\boldsymbol{x}_{i\cdot}$ or $(\hat{m}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{m}})$

一致性を持つ推定法（モデル選択）：

$$\mathrm{BIC}(k) = -\ln p\left(y_1, \cdots, y_n; m, \hat{\beta}_0, \cdots, \hat{\beta}_k\right) + \frac{k+1}{2}\ln n$$

尤度だけだと最大次数の回帰モデルが選択されてしまう

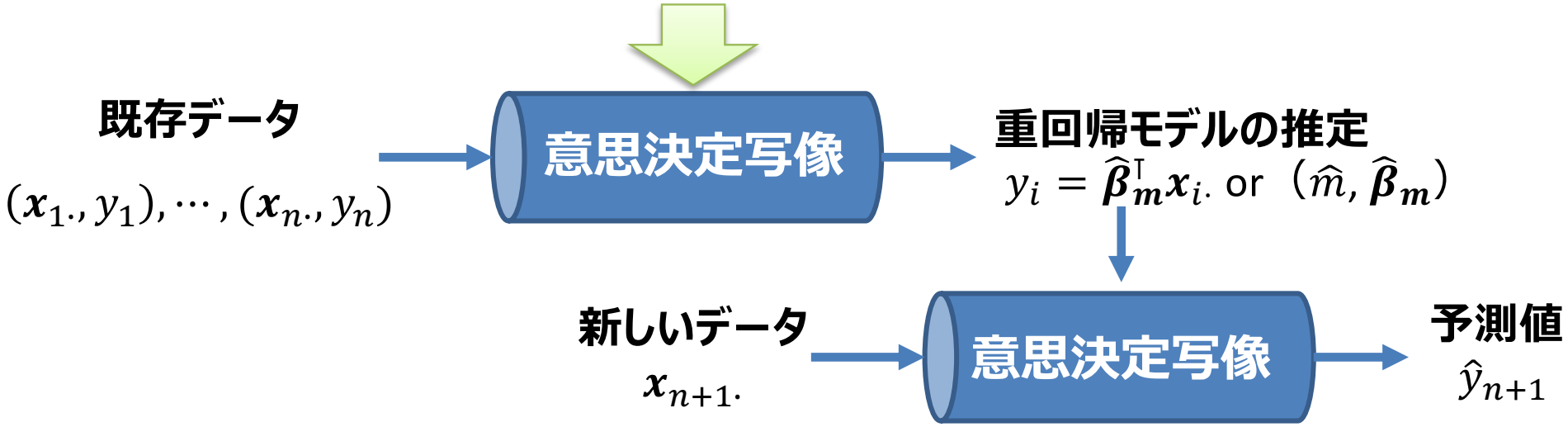
間接予測の意思決定写像

目的：　重回帰モデル（$m, \boldsymbol{\beta}_m$）をまず推定してから予測

設定：$y_i = \boldsymbol{\beta}_m^{\mathsf{T}} \boldsymbol{x}_i + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$,　ε_iは分布$p(\varepsilon)$に従う

（ただし、データを生成する重回帰モデル m は未知定数）

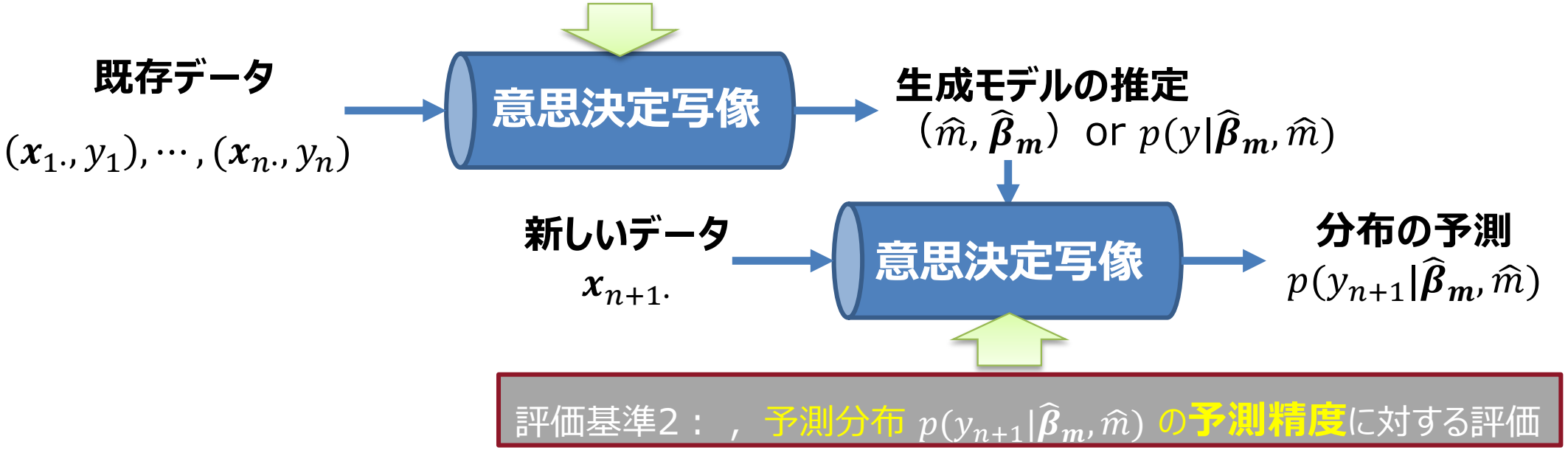
評価基準：$m, \boldsymbol{\beta}_m$ 等のモデルの推定に関する評価



間接予測で予測の精度を評価基準とする

目的：　重回帰モデル（m, $\boldsymbol{\beta}_m$ ）をまず推定してから予測
設定：$y_i = \boldsymbol{\beta}_m^{\mathsf{T}} \boldsymbol{x}_i + \varepsilon_i$,　ε_iは分布$p(\varepsilon)$に従う
（ただし、データを生成する重回帰モデル m は未知定数）

評価基準1：$\widehat{\boldsymbol{\beta}}_m$ は**最尤推定量**(尤度最大）とする



評価基準2：，予測分布 $p(y_{n+1}|\widehat{\boldsymbol{\beta}}_{\boldsymbol{m}}, \hat{m})$ の**予測精度**に対する評価

AICとBICとL0正則化

赤池情報量基準（Akaike Information Criterion）

$$\mathrm{AIC}(m) = -\ln p\left(y_1, \cdots, y_n; m, \hat{\beta}_0, \cdots, \hat{\beta}_k\right) + (k+1)$$

期待値の偏りの補正項

MDLもこの考え方と似ている（後で詳しく）

L0正則化： $\sum_{i=0}^{k} \left|\beta_{i(m)}\right|^0 = k+1$ （回帰係数の数）

$$\mathrm{BIC}(m) = -\ln p\left(y_1, \cdots, y_n; m, \hat{\beta}_0, \cdots, \hat{\beta}_k\right) + \frac{k+1}{2}\ln n$$

AIC規準による推定モデルは
漸近的に母回帰構造と一致する保証はない
(予測を目的とした選択方法であるため)
BIC基準は一致性がある

直接予測の意思決定

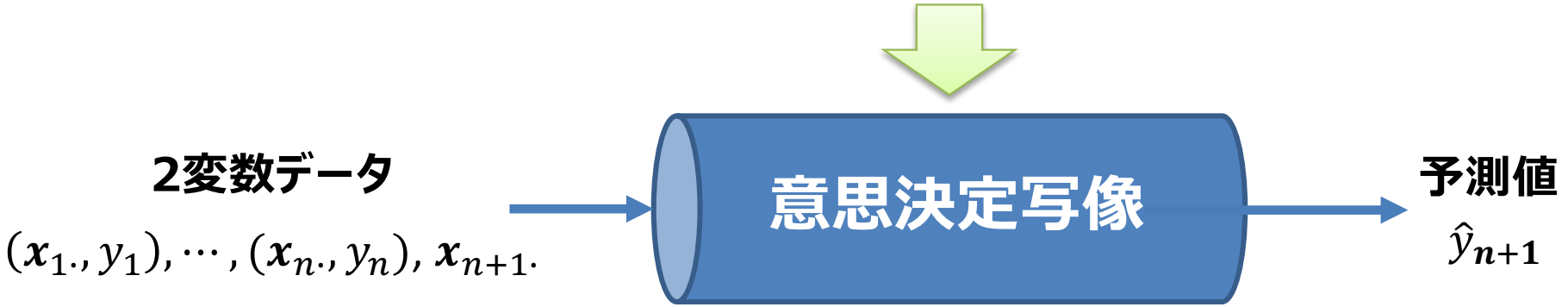
モデル m もメタ情報なので確率変数と考えたら

目的： 説明変数$\boldsymbol{x}_{n+1}$に対する目的変数y_{n+1}の値を予測したい

設定： $y_i = \boldsymbol{\beta}_{\boldsymbol{m}}^{\mathrm{T}}\boldsymbol{x}_i + \varepsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$, ε_iは分布$p(\varepsilon)$に従う

（ただし、データを生成する重回帰構造モデル m は未知確率変数）

評価基準： $\ell\left(Y_{n+1}, d\left((x_1, y_1), \cdots (x_n, y_n), x_{n+1}\right)\right)$等， 予測値 $\hat{y}_{n+1}$ の精度に対する評価



ベイズ決定理論からの直接予測の最適予測値

損失関数として二乗誤差を考えた場合，例えば次のような予測値が得られる（$\boldsymbol{x}$ は省略）：

$$\hat{y}_{n+1} = \int y_{n+1} \sum_m \int p(y_{n+1}|m, \boldsymbol{\beta}_m) p(\boldsymbol{\beta}_m|y_1, \cdots, y_n, m) d\boldsymbol{\beta}\, p(m|y_1, \cdots, y_n)\, dy_{n+1}$$

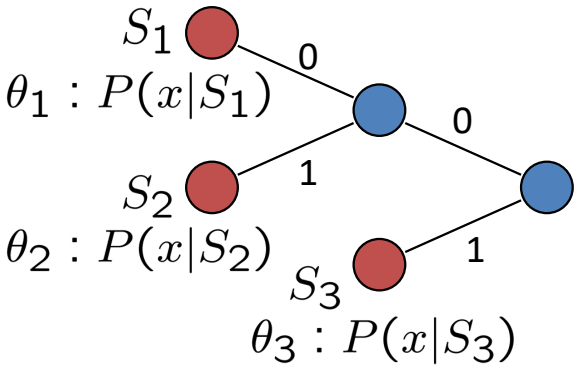
構造モデル m と $\boldsymbol{\beta}_m$ が決まった元での y_{n+1} の生起確率

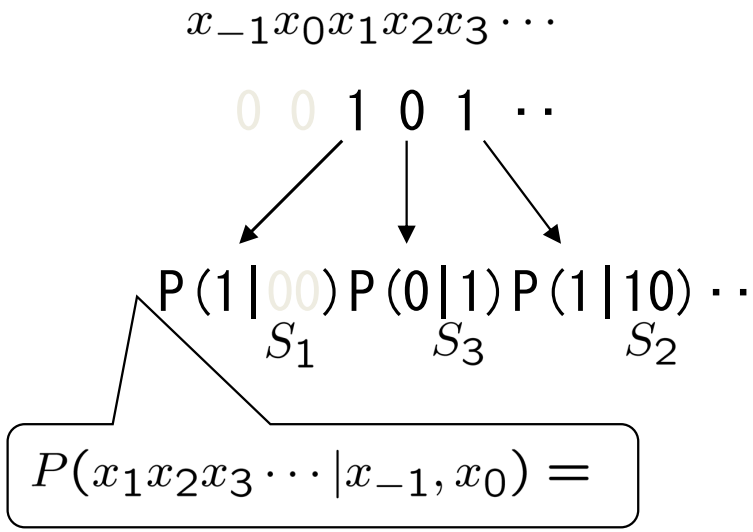
$y_1, \cdots, y_n$ を得た元で構造 m 上で $\boldsymbol{\beta}_m$ のの事後確率

$y_1, \cdots, y_n$ を得た元での構造 m の事後確率

それでは情報理論に戻りましょう：情報源モデルとして文脈木モデルを考える

The context tree model is a finite state source whose state at t is determined by the context of a source sequence x^t





The **contexts** of a model are represented by **a context tree**

In a context tree model, the probability of a symbol is represented by a conditional probability of the symbol given the context.

文脈木モデルを未知の定数と考えると

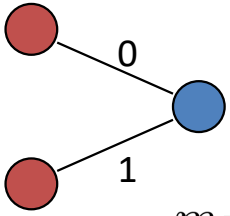
Unknown model $m_i = \{S_1, S_2, S_3 \cdots\} \in M$

Unknown parameters (Random variables) $\theta(m) = \{\theta_1, \theta_2, \theta_3 \cdots\} \in \Theta(m)$

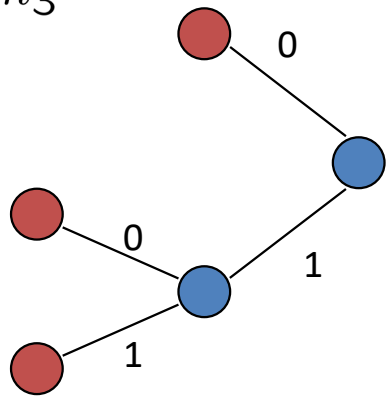
M: the class of context trees up to depth 2

m_1: i.i.d

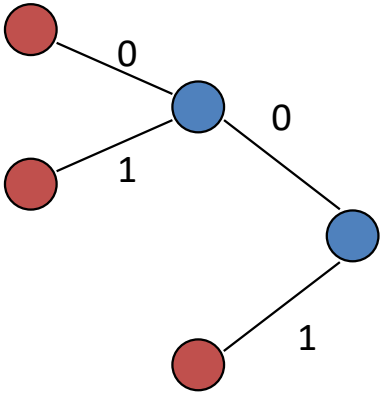
m_2: simple Markov model



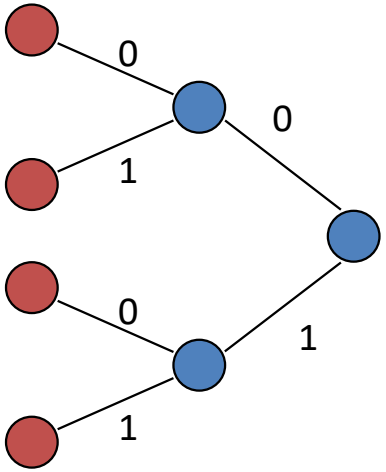
m_3



m_4



m_5



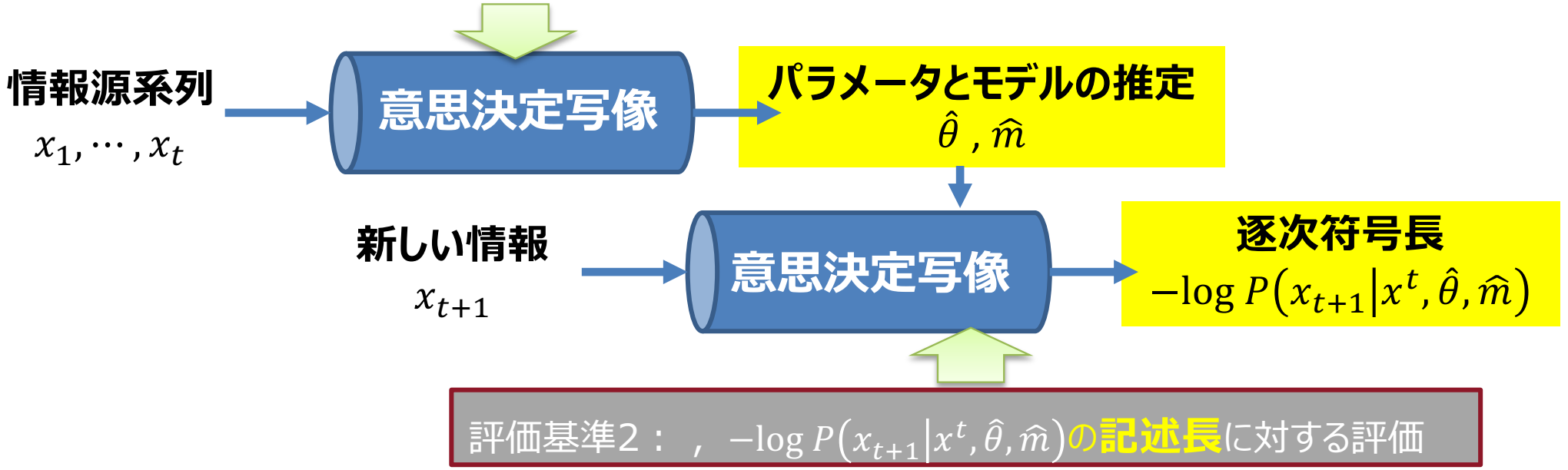
間接符号化で記述長を評価基準とする（MDL基準）

目的：パラメータ θ とモデル m をまず推定してから逐次符号長 $-\log P(x_{t+1}|x^t,\theta,m)$を決定

設定： $y_i = \boldsymbol{\beta}_m^{\mathsf{T}}\boldsymbol{x}_i + \varepsilon_i$,　　ε_iは分布$p(\varepsilon)$に従う

（ただし、情報を生成する確率モデル m が未知）

評価基準1：$\hat{\theta}$ は**最尤推定量**とする

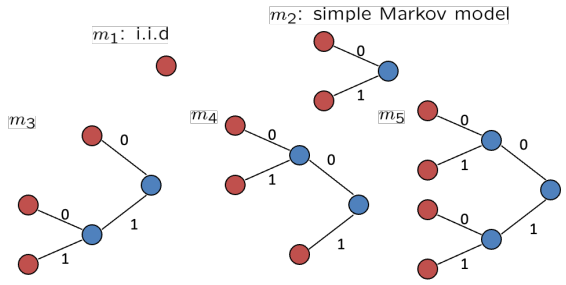


[Rissanen 78] Modeling by Shortest Data Description, IEEE Transactions on Automatic Control, Vol. 23, No. 4

文脈木モデルを未知の確率変数と考えると

Unknown model (Random variables) $m_i = \{S_1, S_2, S_3 \cdots\} \in M$

Unknown parameters (Random variables) $\theta(m) = \{\theta_1, \theta_2, \theta_3 \cdots\} \in \Theta(m)$



[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on IT.
[Willems et al. 93] Context Tree Weighting: Basic Properties, Proceeding of ISIT.
[Matsushima, Hirasawa 94] A Bayes Coding Algorithm Using Context Trees, Proceeding of ISIT.
[Matsushima, Hirasawa 09] Reducing the Space Complexity of a Bayes Coding Algorithm Using an Expanded Context Tree, Proceedings of ISIT.
[Nakahara et al. 20] Bayes Code for 2-dimensional Auto-regressive Hidden Markov Model and Its Application to Lossless Image Compression, Proceedings of 2020 International Workshop on Advanced Image Technology.
[Shimada et al. 21] An Efficient Bayes Coding Algorithm for the Non-Stationary Source in Which Context Tree Model Varies from Interval to Interval, Proceedings of ITW.
[Nakahara et al. 22] Probability Distribution on Rooted Trees, Proceedings of ISIT.
[Kontoyiannis et al.22] Bayesian context trees: Modelling and exact inference for discrete time series, Journal of the Royal Statistical Society Series B 1

モデル m をメタ情報とした直接符号化の意思決定写像（ベイズ決定理論による）

目的：　情報源系列 $x_1, \cdots, x_t$から逐次符号長 $-\log P(x_{t+1}|x^t)$を直接決定

設定：　情報源は確率分布$P(x_{t+1}|x^t, \theta)$に従う　メタ情報の θ と m は未知確率変数

評価基準：期待符号長最小（ベイズ危険関数）

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on Information Theory, vol.37
[Matsushima, Hirasawa 94] A Bayes Coding Algorithm Using Context Trees, Proceeding of ISIT

Bayes codes (unknown model)

Known ← a parametric distribution class: $P(x^n|\theta(m), m)$

Unknown ← k-dimensional real parameter vector: $\theta \in \Theta \subset \Re^k$

Unknown ← parametric model: $m \in M$

Under the condition that the **Arithmetic Coding** is used for coding, the decision of **encoding probability** $\hat{P}(x^n)$ is the main problem in universal coding.

ブロック符号も逐次符号も
符号長は同じ

Block coding probability:

$$\hat{P}(x^n) = \sum_{m \in M} \int P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta$$

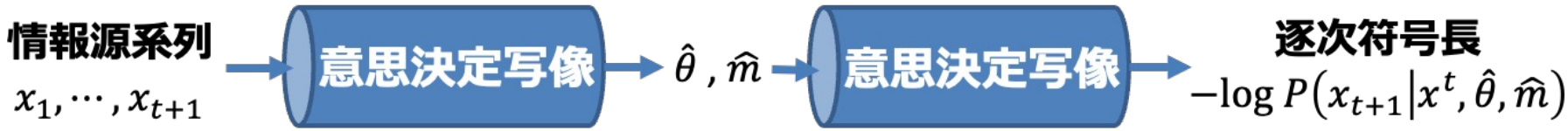
Conditional coding probability:

$$\hat{P}(x_t|x^{t-1}) = \sum_{m \in M} \int P(x_t|x^{t-1}, \theta_m, m) P(\theta_m|m, x^{t-1}) P(m|x^{t-1}) d\theta_m$$

Expectation is taken all over the model class (Weighting all models)

間接符号化に対する直接符号化の優位性：データ処理定理

間接符号化（例 MDL）



直接符号化（例 ベイズ符号）

VI

[Rissanen 78] Modeling by Shortest Data Description, IEEE Transactions on Automatic Control, Vol. 23, No. 4
[後藤 他 99] 事前分布が異なる場合のMDL原理に基づく符号とベイズ符号の符号長に関する解析,電子情報通信学会論文誌, vol.J82-A

ベイズ符号（直接符号化）のデメリット

The complexity of the coding calculation

Assuming the conjugate prior as the prior probability of θ, the integral does not need for calculating the coding probability.

Integral is taken all over the parameter space

$$\hat{P}_B(x^n) = \sum_{m \in M} \int_{\Theta} P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta.$$

Summation is taken all over the model class

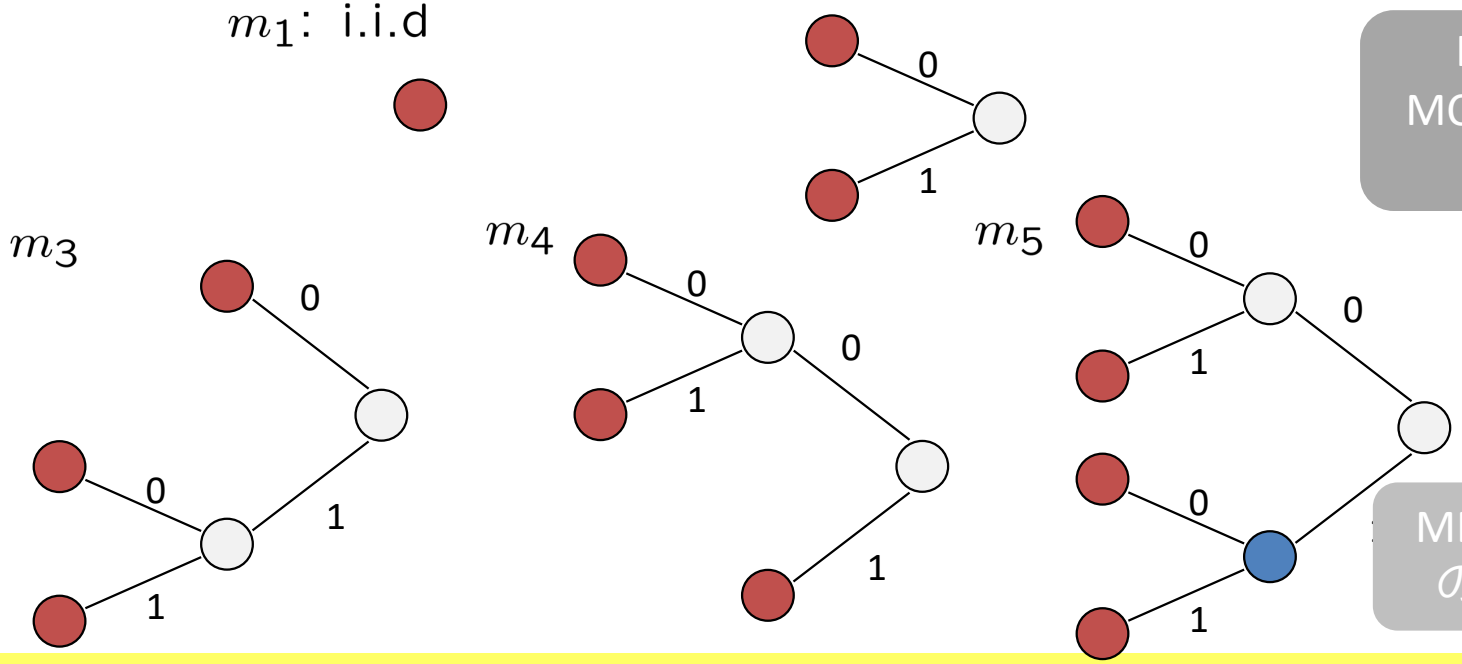
For calculation of the Bayes codes, the expectation is taken all over the context tree models whose number is $O(2^{|X|^{d-1}})$

Bayes codes for Context tree models

M: the class of context trees up to depth 2

m_1: i.i.d

m_2: simple Markov model



Kontoyianisらは
MCMCで計算のため
計算量膨大

MDLもモデル選択な
ので同等の計算量

For calculation of the Bayes codes, the expectation is taken all over the context tree models whose number is $O(2^{|X|^{d-1}})$.

[Kontoyiannis et al.22] Bayesian context trees: Modelling and exact inference for discrete time series, Journal of the Royal Statistical Society Series B

ベイズ符号（直接符号化）のデメリットの解消

The complexity of the coding calculation

Assuming the conjugate prior as the prior probability of θ, the integral does not need for calculating the coding probability.

Integral is taken all over the parameter space

$$\hat{P}_B(x^n) = \sum_{m \in M} \int_{\Theta} P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta.$$

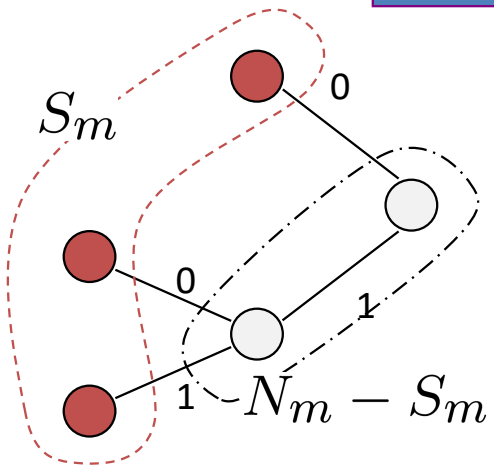
Summation is taken all over the model class

We propose a special prior distribution on the context tree models and construct an efficient algorithm assuming the prior.

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on Information Theory, vol.37
[Matsushima, Hirasawa 94] A Bayes Coding Algorithm Using Context Trees, Proceeding of ISIT

Special prior distribution for Context tree models

A special class of the prior distribution of *m*



Definition 1

We define the decomposable prior distribution $P(m|q)$ as the following formula:

$$P(m|q) = \prod_{s \in N_m - S_m} (1 - q(s)) \prod_{s_L \in S_m} q(s_L), \quad (1)$$

where N_m denotes the set of all nodes on the context tree model T_m and

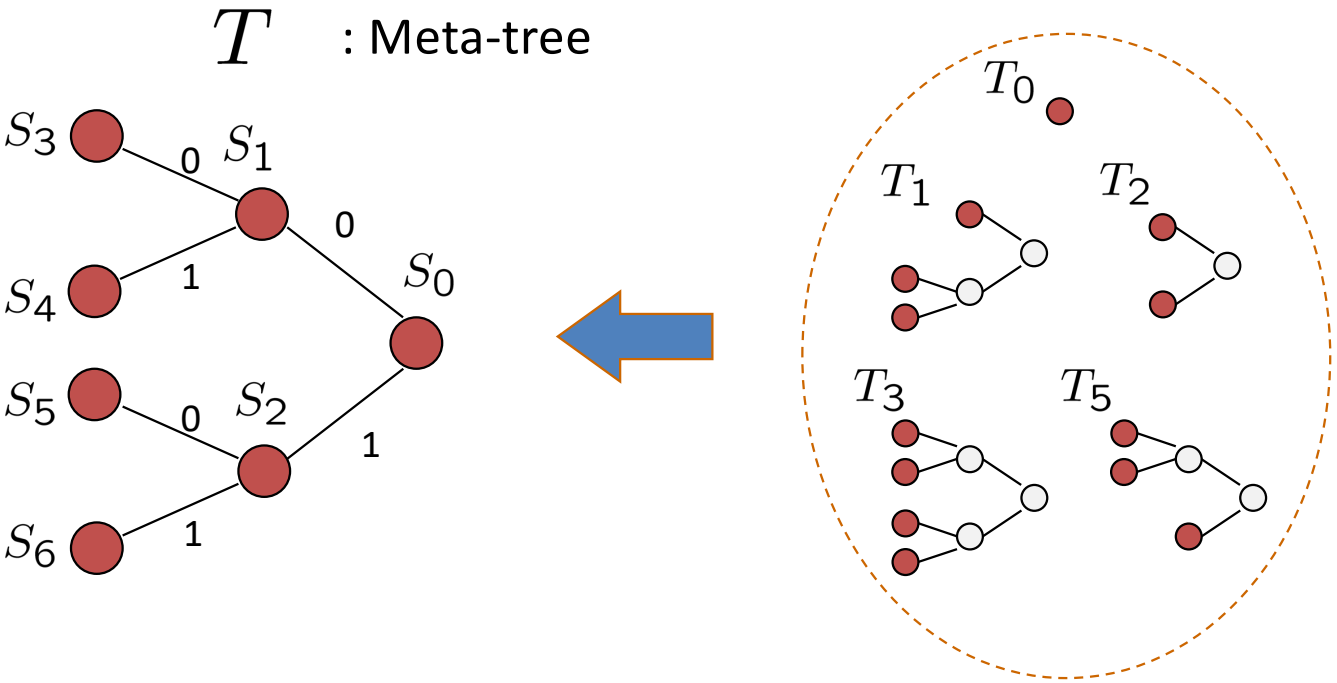
$$q(s) = \frac{\sum_{\{m|s \in S_m\}} P(m|q)}{\sum_{\{m|s \in N_m\}} P(m|q)}. \quad (2)$$

The prior probability of m is represented by $\{q(s)|s \in N_m\}$. Moreover, the posterior probability $P(m|x^t)$ is also represented by $\{q(s|x^t)|s \in N_m\}$.

Bayes codes algorithm using the special prior (1)

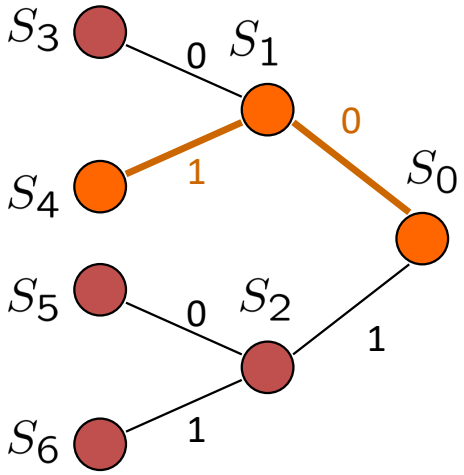
The efficient coding algorithm is given by using **the special prior** and **the superposed tree** to which all context tree models in M are merged.

Let the superposed context tree be $T = \bigcup_{m \in M} T_m$ and the set of all leaf nodes in T be L.



Bayes codes algorithm using the special prior (2)

By using the special prior and a superposed tree, we don't need the summation that is taken all over the model class.

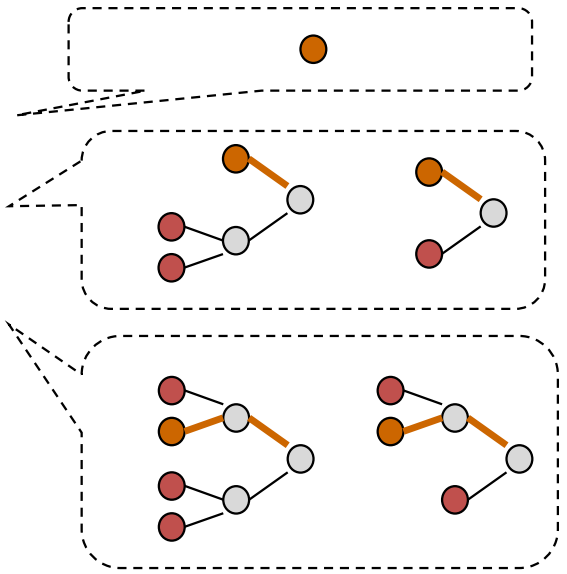


$$\cdots x_{t-1}x_t x_{t+1}$$
$$\cdots \quad 1 \quad 0 \quad 0$$

Posterior probability of models

$$P(0|10) = \hat{P}(0|S_0)\hat{q}(S_0|x^t) + \hat{P}(0|S_1)\hat{q}(S_1|x^t) + \hat{P}(0|S_4)\hat{q}(S_4|x^t)$$

$$\hat{P}(0|S_i) = \frac{n(0|S_i) + \alpha/2}{n(S_i) + \alpha}$$



By using *q(s)*, the complexity is deduced to *O(d)*.

ポイントは：
特殊な事前分布
& モデルの集約

Bayes codes algorithm using the special prior (3)

Under the condition that the prior distribution of a class of context tree models is assumed as Definition 1, the Bayes predictive coding probability $P_C(x_t|x^{t-1})$ is calculated by the following recursive formulas.

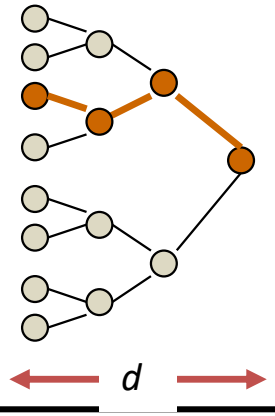
$$P_C(x_t|x^{t-1}) = P_C(x_t|s = s_0),$$

$$P_C(x_t|s) = \begin{cases} P(x_t|x^{t-1}, s) & \text{if } s \in L \\ q(s|x^{t-1})P(x_t|x^{t-1}, s) + (1 - q(s|x^{t-1}))P_C(x_t|s_c), & \text{otherwise}, \end{cases}$$

where $s_c = \{xs \in \bigcup_{m \in M}\{s_m(x^{t-1})\}\}$, s_0 is the root node of T.

葉ノードから根ノードへの再帰的計算で符号化確率が計算可能

The calculation is done through the path from the leaf to the root on the gathering tree.



By using *q(s),* the complexity is deduced to *O(d).*

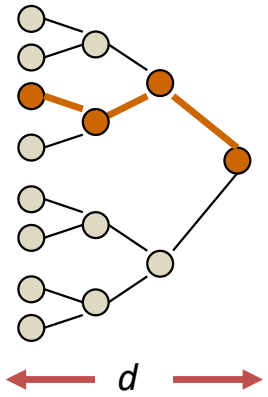
Bayes codes algorithm using the special prior (4)

The posterior probability $P(m|x^t)$ is also represented by $\{q(s|x^t)|s \in N_m\}$.

The posterior probability $q(s|x^t)$ is updated by the following formula:

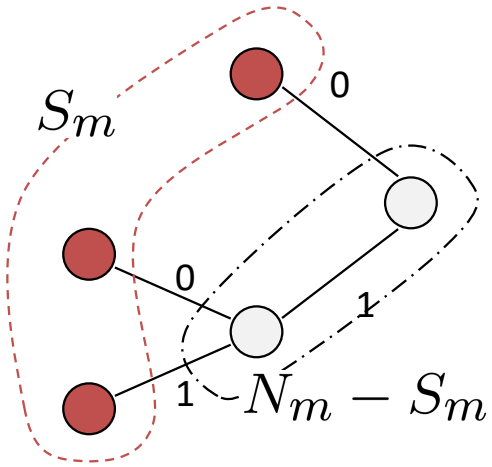
$$q(s|x^t) = \frac{P(x_t|x^{t-1}, s)q(s|x^{t-1})}{P_C(x_t|s)}.$$

The update can be calculated simultaneously with the calculation of the conditional coding probability $P_C(x_t|x^{t-1})$ through the path from the leaf to the root.



The time complexity for updating $q(s|x^t)$ is also $O(d)$

Bayes codes algorithm using the special prior (5)



A special class of the prior distribution of *m*

We define the decomposable prior distribution $P(m|q)$ as the following formula:

$$P(m|q) = \prod_{s \in N_m - S_m} (1 - q(s)) \prod_{s_L \in S_m} q(s_L), \quad (1)$$

where N_m denotes the set of all nodes on the context tree model T_m and

$$q(s) = \frac{\sum_{\{m|s \in S_m\}} P(m|q)}{\sum_{\{m|s \in N_m\}} P(m|q)}. \quad (2)$$

Block coding probability:

$$\hat{P}(x^n) = \sum_{m \in M} \int P(x^n|\theta_m, m) p(\theta_m|m) P(m) d\theta$$

Conditional coding probability:

$$\hat{P}(x_t|x^{t-1}) = \sum_{m \in M} \int P(x_t|x^{t-1}, \theta_m, m) P(\theta_m|m, x^{t-1}) P(m|x^{t-1}) d\theta_m$$

CTW は $q(s) = \frac{1}{2}$ として
ブロック符号に
メタ木を用いたアルゴリズム
と解釈される

[Willems et al. 93] Context Tree Weighting: Basic Properties, Proceeding of ISIT.

The mean code length of Bayes codes for Context tree models

One of the advantages of Bayes codes is that we can evaluate their code length precisely.

Theorem 1:

Let m^* be a minimal true model and $\theta^*_{m^*}$ the true parameters of the model. If $P(m^*) > 0$, the mean code length of both the predictive and non-predictive Bayes codes is given by

$$E_{X^n}[-\log P_c(X^n)] = H(X^n|\theta^*_{m^*}, m^*) - \log P(m^*) + \frac{k_{m^*}}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta^*_{m^*}|m^*)}}{p(\theta^*_{m^*}|m^*)} + o(1), \quad (1)$$

where $I(\theta_m|m)$ is the Fisher information of θ_m and k_{m^*} is the number of the parameters in m^*.

パラメータだけでなく
モデルも未知の場合の
様々な符号長が
解析できる

[Gotoh et al. 98] A Generalization of B.S.Clarke and A.R.Barron's Asymptotics of Bayes Codes for FSMX Sources, IEICE Trans. on Fundamentals.

The mean code length of Bayes codes for Context tree models

Theorem 1:

Let m^* be a minimal true model and $\theta^*_{m^*}$ the true parameters of the model. If $P(m^*) > 0$, the mean code length of both the predictive and non-predictive Bayes codes is given by

$$E_{X^n}[-\log P_c(X^n)] = H(X^n|\theta^*_{m^*}, m^*) - \log P(m^*) + \frac{k_{m^*}}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta^*_{m^*}|m^*)}}{p(\theta^*_{m^*}|m^*)} + o(1), \quad (1)$$

where $I(\theta_m|m)$ is the Fisher information of θ_m and k_{m^*} is the number of the parameters in m^*.

その他 様々な良い性質を有していることが証明されている

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on IT.
[Matsushima, Hirasawa 94] A Bayes Coding Algorithm Using Context Trees, Proceeding of ISIT.
[Gotoh et al. 98] A Generalization of B.S.Clarke and A.R.Barron's Asymptotics of Bayes Codes for FSMX Sources, IEICE Trans. on Fundamentals.
[Matsushima, Hirasawa 09] Reducing the Space Complexity of a Bayes Coding Algorithm Using an Expanded Context Tree, Proceedings of ISIT.
[Nakahara et al. 21] Hyperparameter Learning of Stochastic Image Generative Models with Bayesian Hierarchical Modeling and Its Effect on Lossless Image Coding, Proceedings of ITW.
[Nakahara et al. 22] Probability Distribution on Rooted Trees, Proceedings of ISIT.

中間のまとめ：考え方の流れ

メタ情報は確率変数 ：情報の生成メカニズムの情報は確率モデルや確率パラメータ

＋

ベイズ決定理論 ：予測や符号化のベイズ危険関数を最小化

＋

直接予測（符号化） ：間接予測に対する直接予測の優位性

＋

効率的アルゴリズム ：ベイズ最適法のデメリットを事前分布とモデルの集約で効率化

[Matsushima et al. 91] A Class of Distortionless Codes Designed by Bayes Decision Theory, IEEE Transaction on Information Theory, vol.37

[CMです] 独自の統一的視点からの教科書シリーズ「ライブラリ データ科学」

松嶋敏泰(早稲田大学教授) 監修
早稲田大学データ科学教育チーム 著

データ科学は，人間の知的活動を対象とする根源的な学問分野であると共に，人間の活動や社会の変革・発展に直接的に寄与する分野であり，情報・通信での理論・技術の進歩やインフラの発展により，今後益々重要となっていくものであるともいえます．
そこで，本ライブラリは，データ科学を統一的視点から体系化し，総合的かつ体系的に学べるよう配慮された構成としました．
「データ科学とはなにか」「なぜ重要であるか」という問いに対して科学的な視点から考え，統計学や機械学習等の個別の学問領域からではなく，データ科学を統一的視点から一つの体系として扱うことを大きな特徴としたテキスト群となっております．

5.データ科学実践
6.回帰と分類のデータ科学
7.時系列構造のデータ科学
8.潜在構造のデータ科学
9.空間構造のデータ科学
10.因果構造のデータ科学
11.データ科学のためのモデリング

書評：データ科学は，まさに変容・進化しており，新しい手法や理論が次々登場している．（中略）このような時代におけるデータ科学の学習においては，統計学・情報科学を含めた基本的な概念の学習が求められている．そのような意味で，統計学と機械学習の架け橋となる本書は，名著だと考える．
（統計数理研究所 水田正弘）横幹第19号第1巻

メタ情報を確率変数で扱う考え方の　発展系　横への展開

メタ情報は確率変数

＋

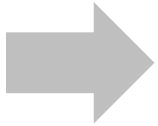
ベイズ決定理論

＋

直接予測（符号化）

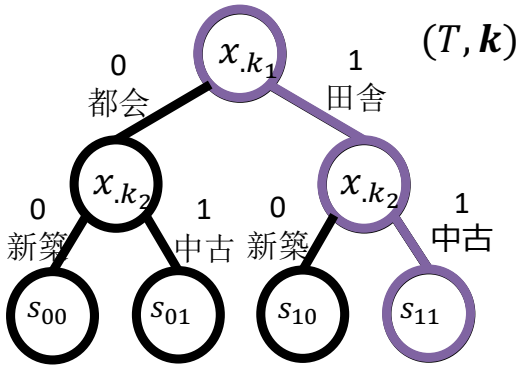
＋

効率的アルゴリズム



様々な情報を
扱う分野へ

データサイエンス分野での木構造モデル：決定木（回帰分類問題）



$\boldsymbol{k} = (1,2,2)$ $\boldsymbol{x_i}$
k_1,k_2,k_3 $= (1,1,0)^\top$

- $T \in \mathcal{T}$:最大深さD_{max}の木構造
- $\boldsymbol{k} \coloneqq (k_s)_{s \in \mathcal{S}} \in \mathcal{K}$: 説明変数割当ベクトル
- $s(\boldsymbol{x})$: $\boldsymbol{x}$に対応する葉ノード
- $\mathcal{S}$：Tのノード集合

関数木(予測関数)→機械学習分野で主流

$\boldsymbol{x}_{n+1} \rightarrow f(\boldsymbol{x}_{n+1}; (T, \boldsymbol{k})) \rightarrow \hat{y}_{n+1}$

葉に予測値

$s_{11} \rightarrow \hat{y}_{s_{11}}$

- データの生成モデルの仮定なし
  - 解に統計的な最適性を与えることが困難
  - 学習データだけに適合した$T, \boldsymbol{k}$

モデル木 (生成メカニズム)

$\boldsymbol{x}_{n+1} \rightarrow p(y_{n+1}|\boldsymbol{x}_{n+1}, T, \boldsymbol{k}, \boldsymbol{\theta}) \rightarrow \hat{y}_{n+1}$

確率パラメータ

葉に確率分布

$s_{11} \rightarrow \theta_{11}$

- データの生成モデルを仮定する
  - $T, \boldsymbol{k}, \boldsymbol{\theta}$を確率変数として，統計的な最適性を検討できる

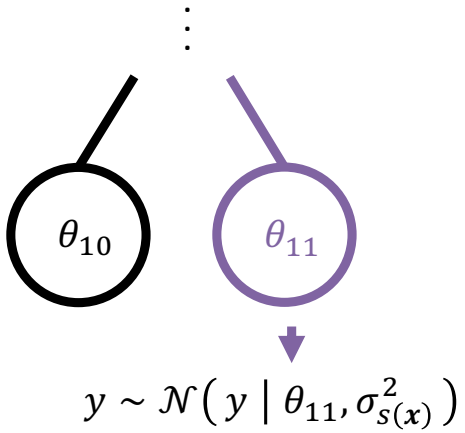
決定木を生成メカニズムを表現している情報とみなす（メタ情報）

本研究では，$p(y|\boldsymbol{x})$を表現するデータ生成モデルを以下で定義する

xのもとで,yが生成される確率構造を表すデータ生成モデルを次式で定義する

$$p(y|\boldsymbol{x},\boldsymbol{\theta},T,\boldsymbol{k}) := \mathcal{N}(y|\theta_{s(\boldsymbol{x})},\sigma^2_{s(\boldsymbol{x})}) \qquad (1)$$

- $\mathcal{N}(y|\theta_{s(\boldsymbol{x})},\sigma^2_{s(\boldsymbol{x})})$は平均$\theta_{s(\boldsymbol{x})}$，分散$\sigma^2_{s(\boldsymbol{x})}$の正規分布の確率密度関数を表す.
- $\boldsymbol{\theta} := (\theta_s)_{s\in\mathcal{S}} \in \boldsymbol{\Theta}$: Tの各ノードsのパラメータ
- $(\boldsymbol{\theta},\boldsymbol{T},\boldsymbol{k})$の組を一つの生成メカニズム情報（メタ情報）とみなす



メタ情報を確率変数としてベイズ最適な推定，計算量的困難性も

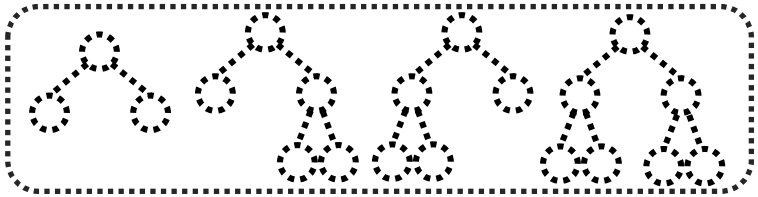
ベイズリスク関数を最小にする決定関数は以下で与えられる

$$\delta^*(\mathcal{D}, \boldsymbol{x}_{n+1}) = \int_{\mathbb{R}} y_{n+1} \sum_{\boldsymbol{k}\in\mathcal{K}} p(\boldsymbol{k} \mid \mathcal{D}) \sum_{T\in\mathcal{T}} p(T \mid \mathcal{D}, \boldsymbol{k}) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid \mathcal{D}, T, \boldsymbol{k})\, p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\theta}, T, \boldsymbol{k}) d\boldsymbol{\theta} dy_{n+1} \quad (5)$$

$|\mathcal{T}|$と$|\mathcal{K}|$は木の深さに応じて，指数的に増加で計算が困難

これまでの決定木アルゴリズムはモデル選択とみなせる

最大深さ$D_{\max} = 2,\ k_s \in \{1,2,3\}$



木構造に対する和計算

メタツリー(MT)利用した効率的計算法で解決

[峯松 他 96] 確率的規則の学習アルゴリズム, 人工知能学会全国大会（第10回）論文集.
[須子 他 03] 決定木モデルにおける予測アルゴリズムについて，電子情報通信学会技術研究報告コンピュテーション, Vol.103.
[Dobashi et al. 20] Probabilistic Data Generating Process on Tree Structure Model: Bayes Optimal Prediction and Sub-Optimal Algorithm, Japanese Federation of Statistical Science Associations.
[Dobashi et al. 21] Meta-Tree Random Forest: Probabilistic Data-Generative Model and Bayes Optimal Prediction, Entropy vol. 23.

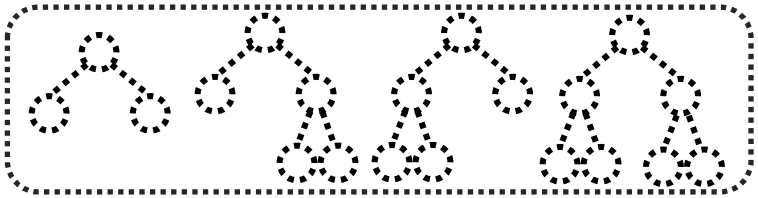
説明変数配置のメタ情報 $\boldsymbol{k}$ に関する計算効率化の問題が残る

木構造のメタ情報 T に関しては効率化可能
説明変数割当のメタ情報 $\boldsymbol{k}$ に関する計算効率化が問題

$$\delta^*(\mathcal{D}, \boldsymbol{x}_{n+1}) = \int_{\mathbb{R}} y_{n+1} \sum_{\boldsymbol{k}\in\mathcal{K}} p(\boldsymbol{k} \mid \mathcal{D}) \sum_{T\in\mathcal{T}} p(T \mid \mathcal{D}, \boldsymbol{k}) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid \mathcal{D}, T, \boldsymbol{k})\, p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\theta}, T, \boldsymbol{k}) d\boldsymbol{\theta} dy_{n+1} \quad (5)$$

$|\mathcal{T}|$と$|\mathcal{K}|$は木の深さに応じて，指数的に増加で計算が困難

最大深さ$D_{\max} = 2,\ k_s \in \{1,2,3\}$



木構造に対する和計算 ⇒ メタツリー(MT)利用した効率的計算法で解決

$\boldsymbol{k} = (1,\bullet,\bullet)$
$\boldsymbol{k} = (2,\bullet,\bullet)$
$\boldsymbol{k} = (3,\bullet,\bullet)$

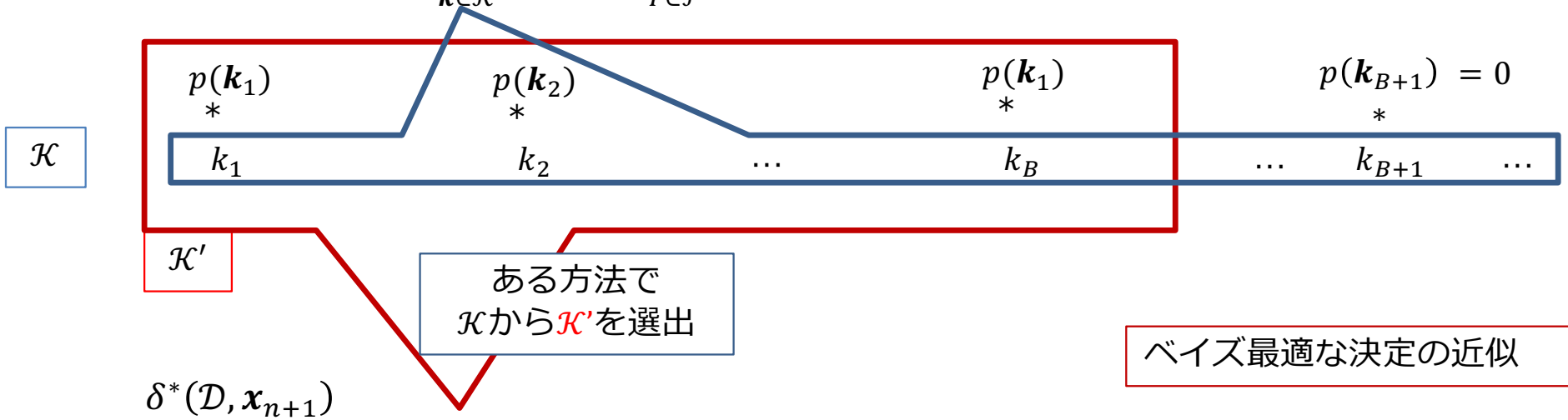
$\boldsymbol{k} = (1,2,3)$
$\boldsymbol{k} = (1,3,2)$
$\boldsymbol{k} = (2,1,3)$
$\boldsymbol{k} = (2,3,1)$
$\boldsymbol{k} = (3,1,2)$
$\boldsymbol{k} = (3,2,1)$

説明変数に対する和計算 ⇒ 効率的な計算法はないため，近似の計算法を提示

説明変数割当のメタ情報 k に関する計算効率化（部分集合 $\mathcal{K}'$ を用いる近似）

$|\mathcal{K}|$に関する問題は以下のアプローチで近似する

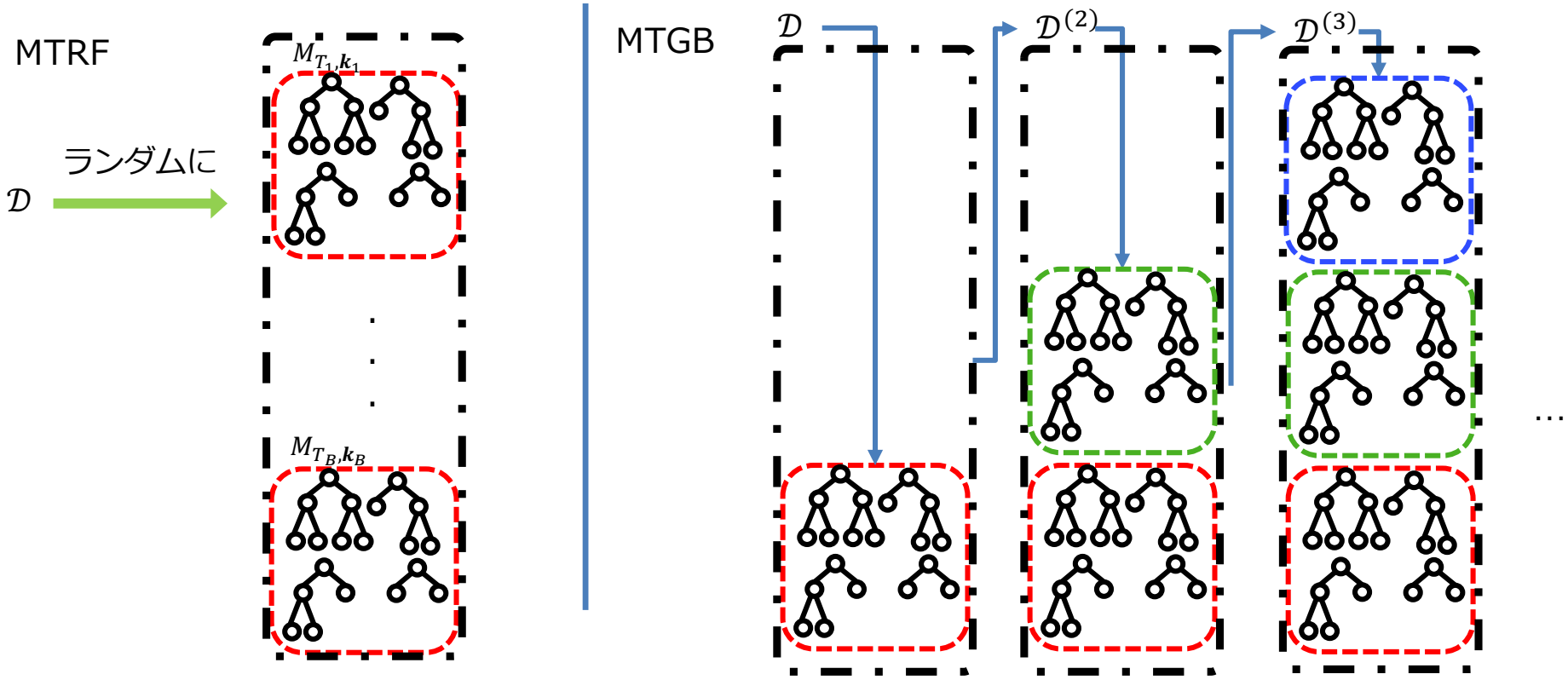
(8) $$\delta^*(\mathcal{D}, \boldsymbol{x}_{n+1}) = \int_{\mathbb{R}} y_{n+1} \sum_{\boldsymbol{k}\in\mathcal{K}} p(\boldsymbol{k} \mid \mathcal{D}) \sum_{T\in\mathcal{T}} p(T \mid \mathcal{D}, \boldsymbol{k}) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid \mathcal{D}, T, \boldsymbol{k})\, p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\theta}, T, \boldsymbol{k}) d\boldsymbol{\theta} dy_{n+1}$$



ベイズ最適な決定の近似

$$\delta^*(\mathcal{D}, \boldsymbol{x}_{n+1})$$
$$\approx \int_{\mathbb{R}} y_{n+1} \sum_{\boldsymbol{k}\in\mathcal{K}'} p(\boldsymbol{k} \mid \mathcal{D}) \sum_{T\in\mathcal{T}} p(T \mid \mathcal{D}, \boldsymbol{k}) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid \mathcal{D}, T, \boldsymbol{k})\, p(y_{n+1} \mid \boldsymbol{x}_{n+1}, \boldsymbol{\theta}, T, \boldsymbol{k}) d\boldsymbol{\theta} dy_{n+1} \quad (7)$$

$|\mathcal{K}'| \subset \mathcal{K}$の構成法：従来の関数木としての複数の決定木の構成法を利用



[Dobashi et al. 21] Meta-Tree Random Forest: Probabilistic Data-Generative Model and Bayes Optimal Prediction, Entropy vol. 23.
[Ichijo et al. 25] Meta-Tree: Bayesian Approach to Avoid Overfitting in Decision Trees and Analysis on the Application to Boosting, IEEE International Workshop on Machine Learning for Signal Processing (MLSP).

MTMCMC：メタツリーの説明変数割当の重み付けにマルコフ連鎖モンテカルロ法

- MCMC法をメタツリーの積分に適用

$$\delta^*(\mathcal{D}, x_{n+1})$$

$$= \arg\max_{y_{n+1}} \sum_{\boldsymbol{k}\in\mathcal{K}} p(\boldsymbol{k}|\mathcal{D}) \sum_{T\in\mathcal{T}} p(T|\mathcal{D},\boldsymbol{k}) \int p(y_{n+1}|\boldsymbol{x}_{n+1},\boldsymbol{\theta},T,\boldsymbol{k}) p(\boldsymbol{\theta}|\mathcal{D},T,\boldsymbol{k}) \mathrm{d}\boldsymbol{\theta}$$

$$= \arg\max_{y_{n+1}} \frac{1}{t_{\mathrm{end}}} \sum_{t=1}^{t_{\mathrm{end}}} \sum_{T\in\mathcal{T}} p\left(T\middle|\mathcal{D},\boldsymbol{k}^{(t)}\right) \int p\left(y_{n+1}\middle|\boldsymbol{x}_{n+1},\boldsymbol{\theta},T,\boldsymbol{k}^{(t)}\right) p\left(\boldsymbol{\theta}\middle|\mathcal{D},T,\boldsymbol{k}^{(t)}\right) \mathrm{d}\boldsymbol{\theta}$$

  - MCMCサンプル$\left\{k^{(1)}, k^{(2)}, \ldots, k^{(t_{\mathrm{end}})}\right\}$上の標本平均で近似

- 提案分布$q\left(k^{(t+1)}\middle|k^{(t)}\right)$をどのように設計するか？
  - 説明変数割当ベクトルはカテゴリカルなので分散の調整という考え方は使えない.
    → 木の事後分布$p\left(T\middle|\mathcal{D},k^{(t)}\right)$を利用し，固定する変数の数を調整

[Nakahara et al. 25] Bayesian Decision Theory on Decision Trees: Uncertainty Evaluation and Interpretability, Proceedings of International Conference on Artificial Intelligence and Statistics.
[中原悠太 他 22] 決定木モデルにおけるメタツリーに対するマルコフ連鎖モンテカルロ法，SITA2022

画像処理分野での木構造モデル：セグメンテーション

- **画像セグメンテーションとは？**

画像をオブジェクトの領域ごとに分割して，どこに何が写っているのかを推定する画像処理.

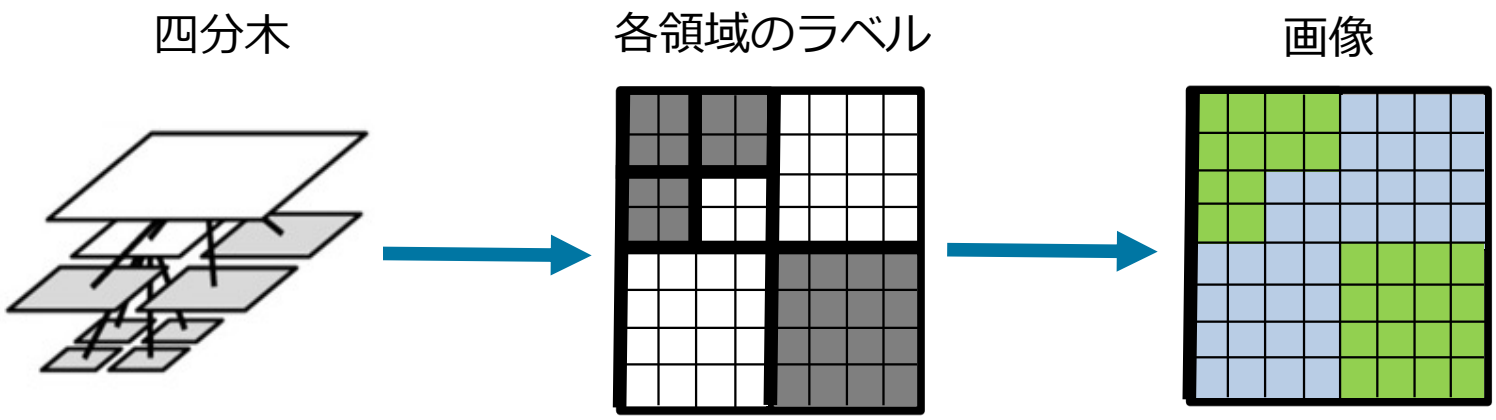
衛星画像から農地の検出
[X.Gao et.al, 2020]

MR画像から腫瘍の検出
[S.Pereira et.al, 2016]

メタ情報として領域分割の木構造モデル

メタ情報としての画像のモデル:

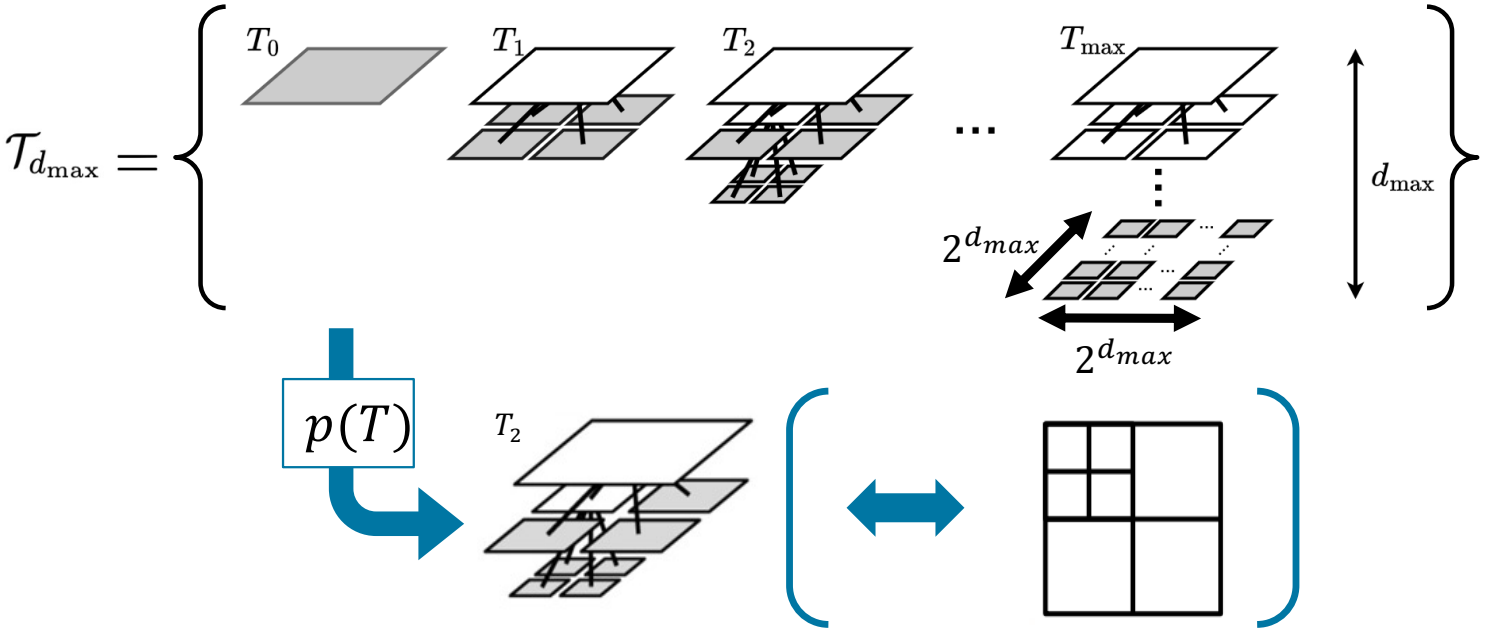
1. 画像内は四分木構造の分割領域の組み合わせで表現されるモデル
2. 各領域内のピクセル値は確率分布から生成されるモデル



[Wang et al. 24] An Efficient Image Segmentation Algorithm Based on Bayes Decision Theory and Its Application for Region Detection, The International Symposium of Information Theory and Its Applications (ISITA).
[Nakahara, Matsushima 21] A Stochastic Model for Block Segmentation of Images Based on the Quadtree and the Bayes Code for It, Entropy.

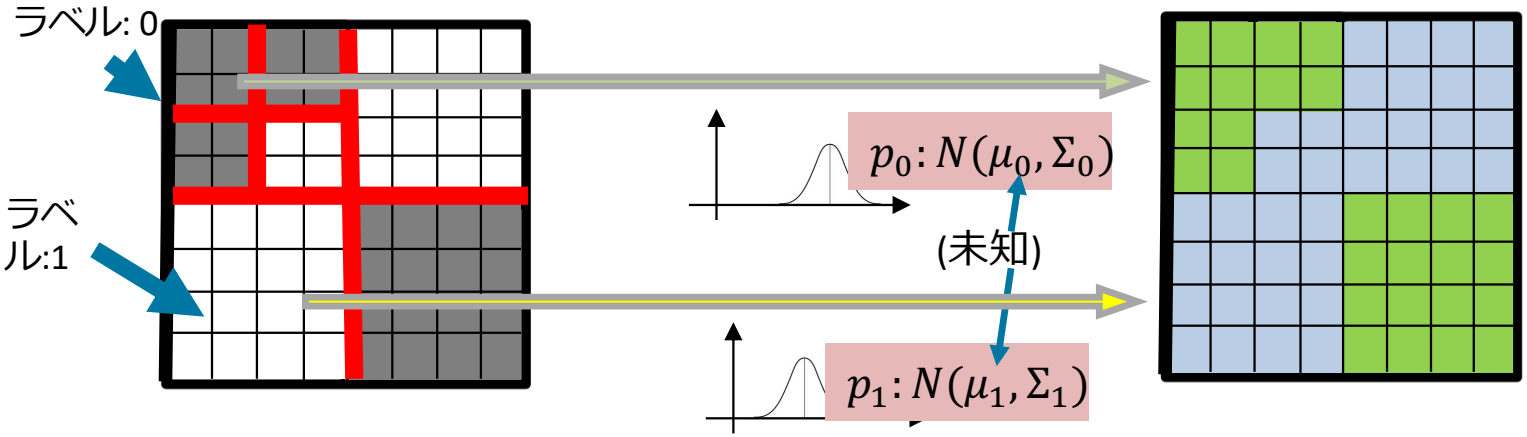
メタ情報の四分木領域分割モデル T を確率変数とし，事前分布を仮定する

- メタ情報 T をは四分木モデル， $\mathcal{T}_{d_{max}}$ を深さが depth ≤ d_{max} であるような四分木の集合であるとする．
- 四分木Tは確率分布 $p(T)$に従って生成される



メタ情報：各領域のピクセル値発生モデル

- 四分木の各葉ノード内は，オブジェクトの特徴に応じて様々な確率モデルを仮定することができる．
  (確率分布のパラメータは未知である．)
- 本発表では，簡単のために，それぞれの領域内ではピクセル値は一つの正規分布に従っていると仮定する．



ベイズ最適解の計算上の問題点と解決策

ベイズ最適な決定関数

決定$\delta^*_{(i,j)}$ において，損失関数を事前分布で重み付けしたベイズリスク関数を最小にするような決定関数は以下の式で導かれる:

$$\delta^*_{(i,j)}(Y) = \underset{x_{(i,j)} \in X}{\operatorname{argmax}} \sum_{T \in \mathcal{T}_{d_{max}}} p\left(x_{(i,j)} \middle| T, Y\right) p(T|Y).$$

計算上の問題点：
1. 事後分布 $p(T|Y)$の計算　→解決策: 特別な事前分布$p(T)$を設定.
2. Tの総和計算: 〖|$\mathcal{T}$〗_(d_max) | は四分木最大深さd_maxに対して指数関数的に増大する. (|$\mathcal{T}$_(d+1) |=|$\mathcal{T}$_d |^4+1)

→解決策: ・ラベルの事後分布が同じ値になる四分木のサブセットを考える.
・四分木の再帰構造を利用.

メタ情報の四分木モデル T に仮定する事前分布 $p(T)$

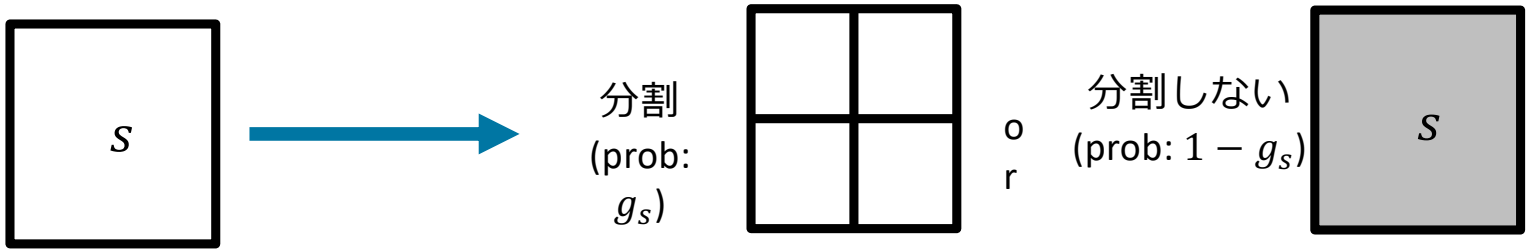
仮定 1. 四分木に仮定する事前分布 $p(T)$

パラメータ $g_s \in [0,1]$ を四分木の各ノード sに設定する. この時，T の事前分布は以下の確率分布として仮定する:

$$p(T;g) = \prod_{s \in L(T)} (1 - g_s) \prod_{s \in I(T)} g_s \tag{5}$$

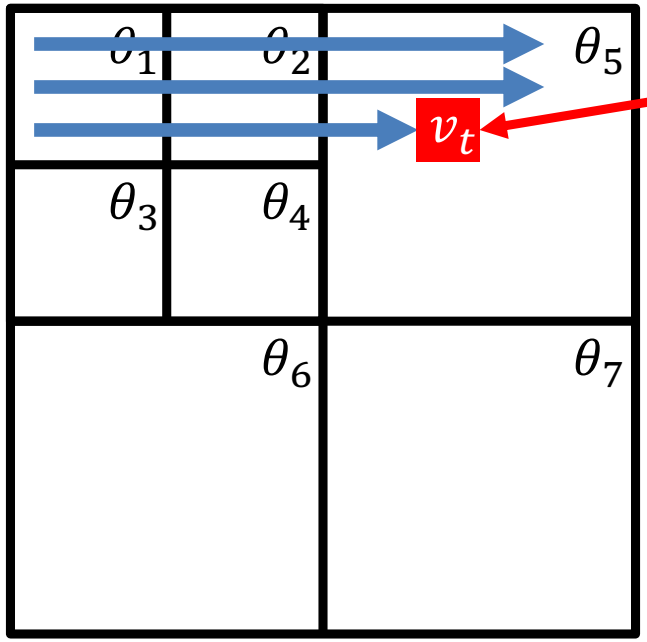
($L(T)$: 四分木Tの葉ノードの集合，$I(T)$: 四分木Tの内部ノードの集合.)

- パラメータ g_s はノード s での枝分かれ確率を表す. (ただし，ノードs が最大深さに位置する場合，$g_s = 0$である．)

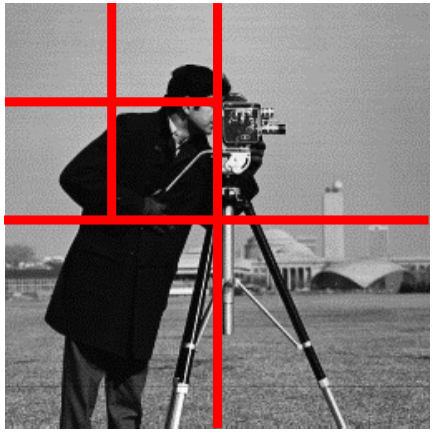


より一般化した画像生成モデル：領域分割モデル＋自己回帰モデル

Pixel value v_t at block s is generated in order of the raster scan with probability $p(v_t|v^{t-1}, \theta_s, m)$.



$p(v_t|v^{t-1}, \theta_5, m)$



v_t depends only on
- The past sequence v^{t-1}
- The parameter θ_s of the block s which contains v_t

[Nakahara et al. 20] Bayes Code for 2-dimensional Auto-regressive Hidden Markov Model and Its Application to Lossless Image Compression, Proceedings of 2020 International Workshop on Advanced Image Technology
[Nakahara et al. 22] Probability Distribution on Rooted Trees, Proceedings of ISIT.

メタ情報 θ^m, m を確率変数としてベイズ最適な符号化確率 $\hat{p}_c(v_t|v^{t-1})$ を決定

  - We cannot use $p(v_t|v^{t-1}, \theta^m, m)$ because true m and θ^m are unknown.
  - We estimate it by $\hat{p}_c(v_t|v^{t-1})$ in Bayesian manner.
- Optimal coding probability $p_c^*(v_t|v^{t-1})$ for our model

$$p_c^*(v_t|v^{t-1}) = \sum_{m\in\mathcal{M}} p(m|v^{t-1}) \int p(v_t|v^{t-1}, \boldsymbol{\theta}^m, m) p(\boldsymbol{\theta}^m|v^{t-1}, m)\, \mathrm{d}\boldsymbol{\theta}^m$$

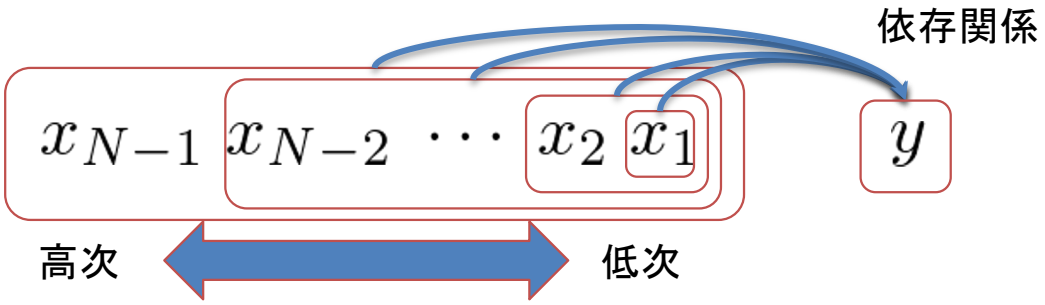
- Three computationally hard parts
  - 1. The summation w.r.t. m ← A recursive structure of quadtree
  - 2. The posterior $p(m|v^{t-1})$ ← Special prior (Detailed in the paper)
  - 3. The integral w.r.t. θ^m ← Conjugate prior

69

自然言語処理分野でのマルコフモデル：Nグラムモデル

メタ情報：Nグラムモデルの次数決定問題

Nグラムモデルの特徴：入れ子型の階層モデル族



RNN（リカレントニューラル）

LSTM（長・短期記憶）

Transformer

高次のモデルの方がより広いクラスの確率モデルが表現可能
高次のモデルの方がよりパラメータの数が多い

同数のサンプルから各パラメータの推定に用いられるサンプル数が相対的に少なくなるために，推定誤差は大きくなる傾向にある．

従来研究の次数モデルごとの重み付け法

ニーザー・ナイ法 [KN][Chen]

階層的に重み付けし単語の出現頻度を調整するパラメータを設定する予測式

$$D_{kn}(x_p^m, \boldsymbol{\theta}_m) = \frac{\max\{c(y|x_p^m) - d, 0\}}{\sum_{y \in Y} c(y|x_p^m)}$$

高次の項

低次の項

$$+ \frac{d}{\sum_{y \in Y} c(y|x_p^m)} \left|y : c(y|x_p^m) > 0\right| D_{kn}(x_p^{m-1}, \boldsymbol{\theta}_{m-1})$$

d　割引係数，$c(y|x^m)$単語列 x^mに後続するyの出現頻度

$\left|y : c(y|x^m) > 0\right|$　y　の種類に比例する重み

単語予測の精度が高いことが報告されている[Chen]
上記で設定されたパラメータの算出方法について
理論的な意味が明確でない

メタ情報（マルコフモデル）を確率変数としてベイズ最適に

メタ情報の次数モデルを確率変数として，
予測誤り率の損失関数をベイズ基準で最適化し決定関数を導出

$$\hat{y} = \arg\max_{y} \sum_{m=1}^{N} p(m|(x^{N-1}, y)^n) \int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1}, \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|(x^{N-1}, y)^n, m) d\boldsymbol{\theta}_m$$

次数mの事後確率

次数mの予測分布

Nの次数決定問題に対して，
ある次数を一つに決めたモデルで予測するのではなく，
各次数の事後確率で重み付けて予測を行うことが
ベイズ基準のもとで最適な予測法であることと言える

[末永 他 13] ベイズ決定理論に基づく階層Nグラムを用いた最適予測法, 情報処理学会論文誌 数理モデル化と応用 vol.6.

従来法の解釈

導出された予測法を漸化式で書き換え

$$F_{\mathrm{r}}(m) = F(m) + \frac{G(m-1)}{G(m)} F_{\mathrm{r}}(m-1). \qquad (3.18)$$

ただし，

$$F(m) = \int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1}, \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|(x^{N-1}, y)^n, m) d\boldsymbol{\theta}_m$$

$$G(m) = p(m|(x^{N-1}, y)^n)$$

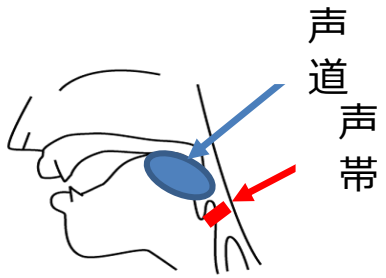
従来法であるニーザー・ナイ法の解釈

重みパラメータは，
モデルの事後確率の比 $G(m-1)/G(m)$ を
近似したものと解釈可能

音声処理分野の音響モデル：隠れマルコフモデル

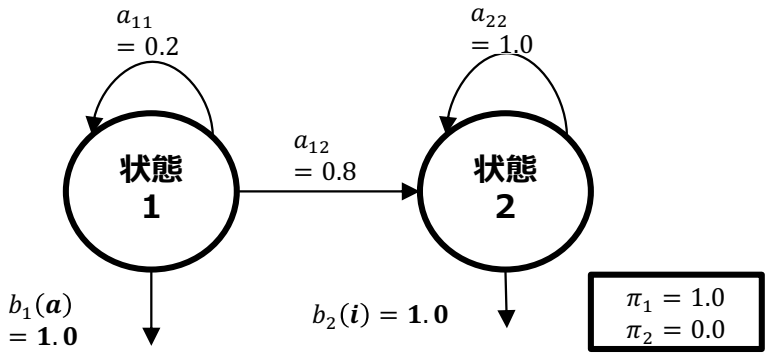
音声から言葉を推測するのに重要な音声の特徴

声道の特徴
・声道の形は，舌の位置や顎の構えによって時刻とともに様々に変化する
・同じ言葉を発声する時の口の形や舌の位置は，いつもおおよそ同じである



◆**HMMは音響確率モデルの代表的モデルで以下のパラメータから構成されている**

- **初期確率π_i：時点1で状態iにいる確率**
- **遷移確率a_{ij}：状態iから状態jに遷移する確率**
- **出力確率 $b_j(\boldsymbol{o_t})$：状態jにおいて特徴ベクトル$\boldsymbol{o_t}$が出力される確率分布**



◆**出力確率には，正規分布や混合正規分布などの確率分布が仮定されることが多い.**

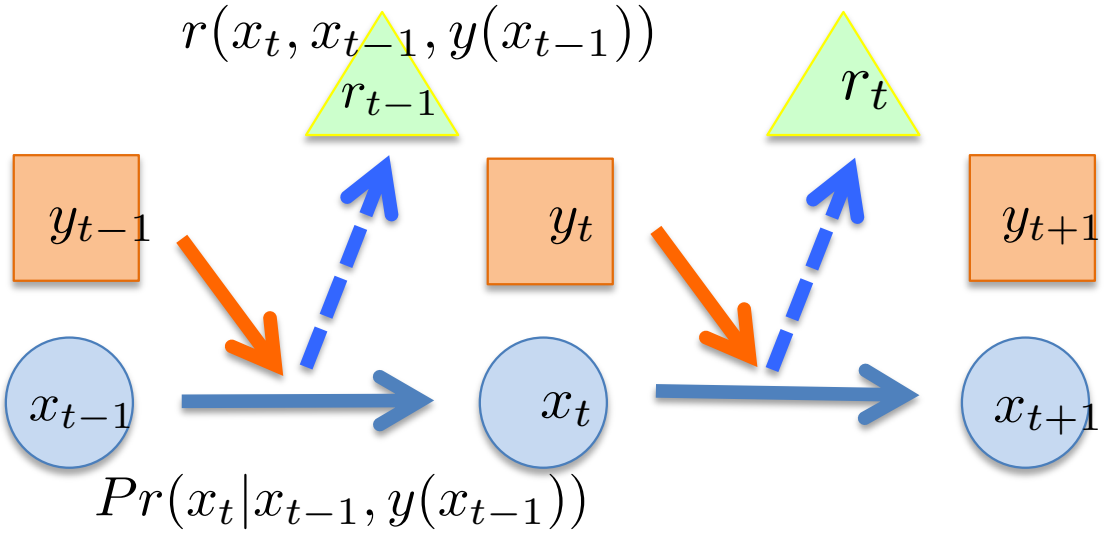
[山岡 他 21] 複数の隠れマルコフモデルの重み付けによるベイズ基準のもとでの最適な音素の予測, 音声研究会,2021

制御分野のMDP (Markov Decision Processes)モデル：強化学習

Reward:

Action:
(Decision)

State:



Target: Expectation of Total Rewards:

$$R^T(y) = \sum_{t=1}^{T} E[r(X_t, X_{t-1}, y(X_{t-1})]$$

Optimal Decision:

$$y^* = \arg\max_y R^T(y)$$

Optimal Decision is calculated by DP

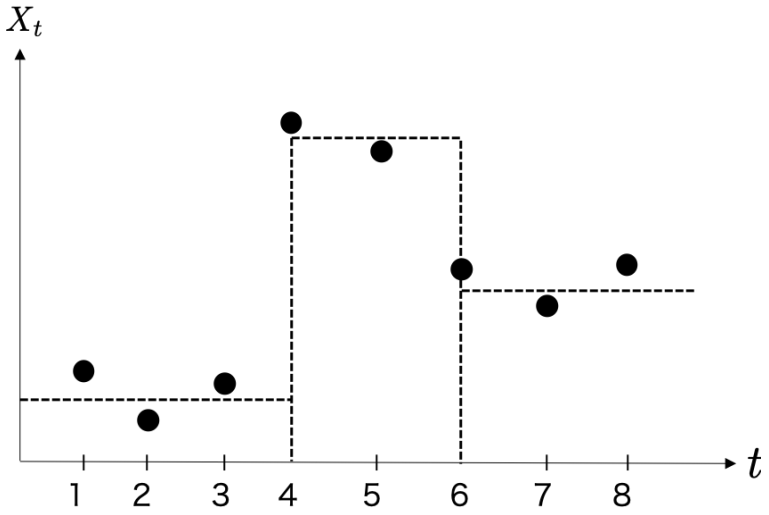
[桑田 他 13] 推薦システムのための状態遷移確率の構造を未知としたマルコフ決定過程, 情報処理学会論文誌 数理モデル化と応用 vol.6.
[MAEDA et al. 10] Applying Markov Decision Processes to Selective-repeat ARQ with Finite Receiver Buffer, International Symposium on Multimedia and Communication Technology(ISMAC2010).
[塚本 他 98] 適応型ARQにおける制御パラメータ決定方式について, 電子情報通信学会論文誌 vol.J81-B-1.

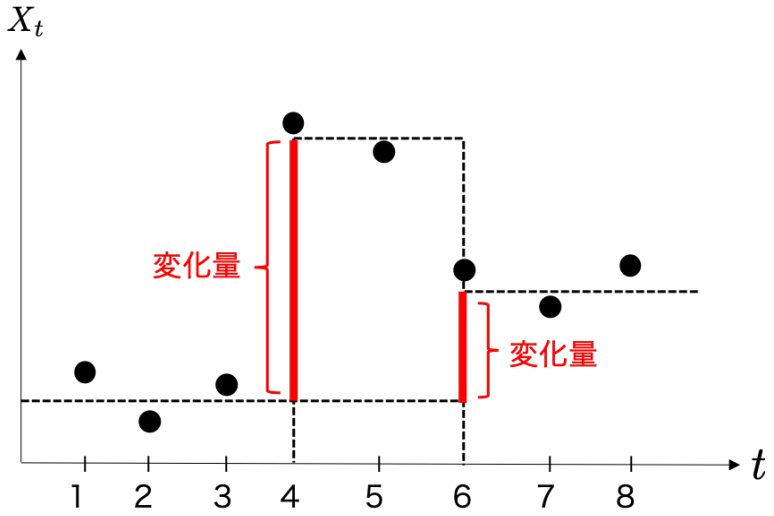
時系列解析分野の変化点モデル：変化点検出，区間回帰モデル

下図のようなデータが得られたときに，

- 変化が起こった回数を推定する
- 変化が起きた時点（変化点）を推定する
- 変化量を推定する
- 変化量が大きい（Relevant な）変化のみを検出する

問題を総じて**変化点検出問題**と呼ぶ．
品質管理，サイバー攻撃の検出，薬効実験など実問題でも重要な問題．

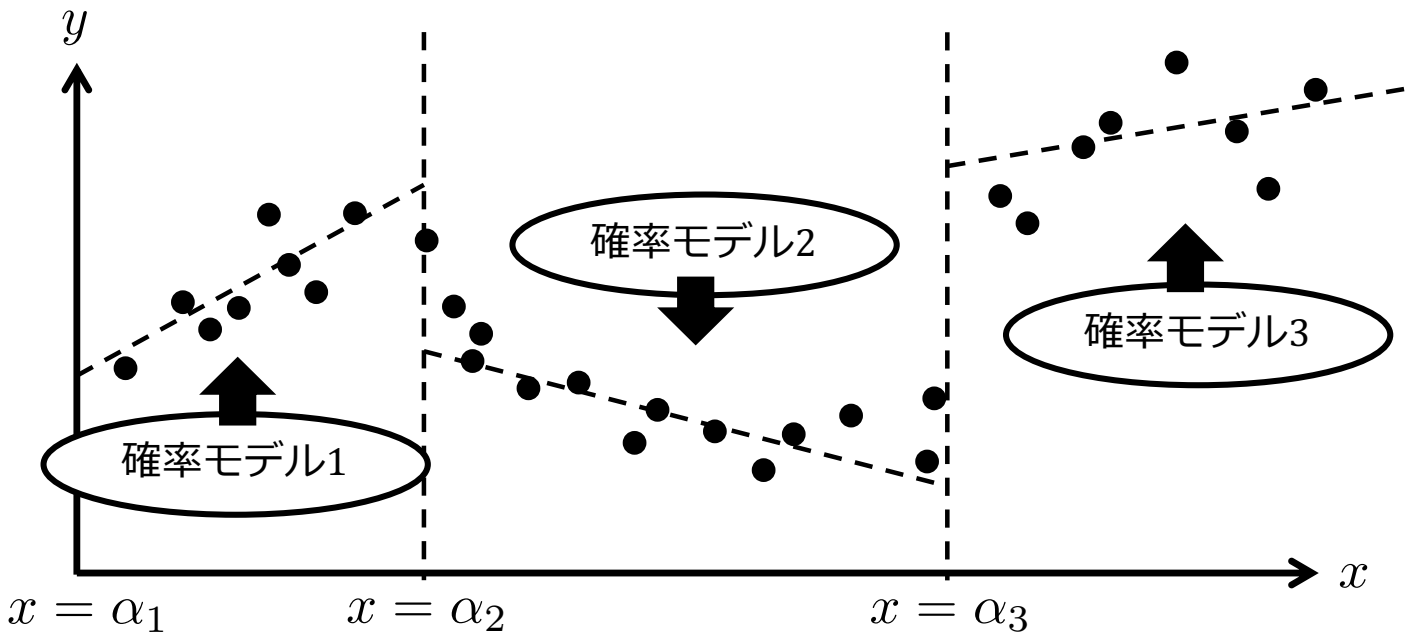




発展系として区分回帰モデル＝変化点モデル＋回帰モデル

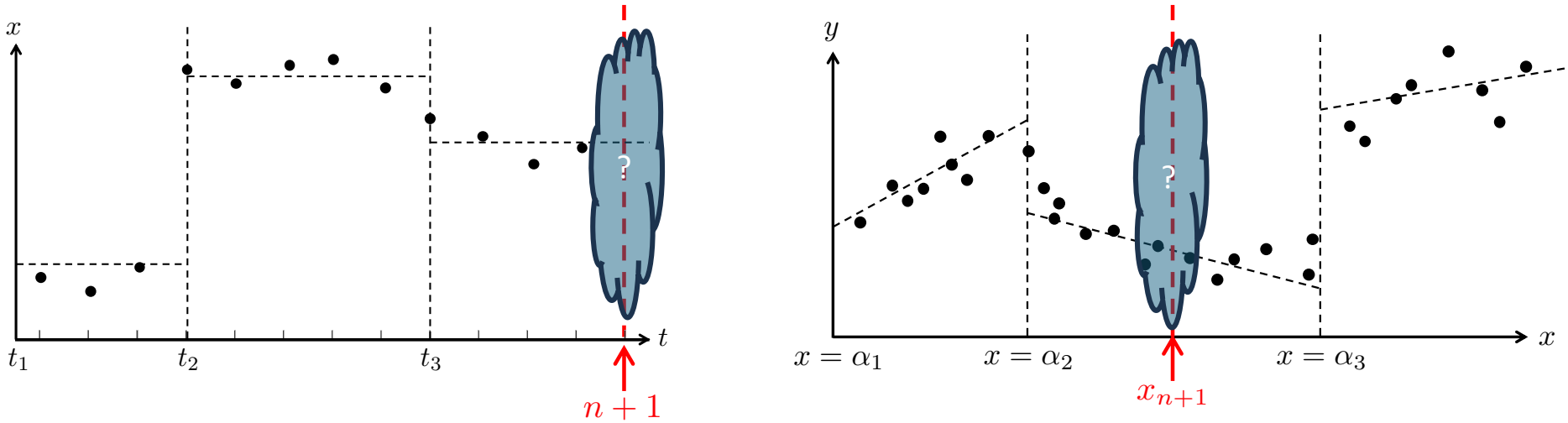
（例2：関係のある2変数のモデル）

- 関係のある2変数(x, y)、$x \in \mathbb{R}$：$y \in \mathbb{R}$
- 説明変数xの区間ごとに異なる確率モデルからデータが発生



1.3 本論文の特徴2：本論文で考えている情報処理（予測問題）

- $x_1, x_2, \cdots, x_n$を得たもとでx_{n+1}に関する情報を得たい
- $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$を得たもとで$x_{n+1}$に対応する$y_{n+1}$に関する情報を得たい



[Suko et al. 08] Asymptotic Property of Universal Lossless Coding for Independent Piecewise Identically Distributed Sources, Proc. of Pre-ICM International Convention on Mathematical Sciences.
[K. Suzuki et al.20] A bayesian decision-theoretic change-point detection for i.p.i.d. sources,IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E103.A.
[K. Suzuki et al.19] Optimal Estimating of the Magnitude of the change for Sources with Piecewise Constant Parameters under Bayesian Criterion, Bayes on the Beach

ベイズ決定理論に基づく最適な予測

ベイズ最適な予測：区間ごとに異なる確率モデルに対するベイズ最適な予測分布

$$\mathrm{AP}^*(x_n|x^{n-1}) = \sum_{c=1}^{N} \sum_{1=t_1<\cdots<t_c\le N} \pi(t^c|x^{n-1})$$

$$\times \left[\sum_{m^c \in \mathcal{M}^c} P\left(m^c \middle| x^{n-1}, t^c\right) \int p\left(x_n \middle| x^{n-1}, \Theta^{m^c}, m^c, t^c\right) \right.$$

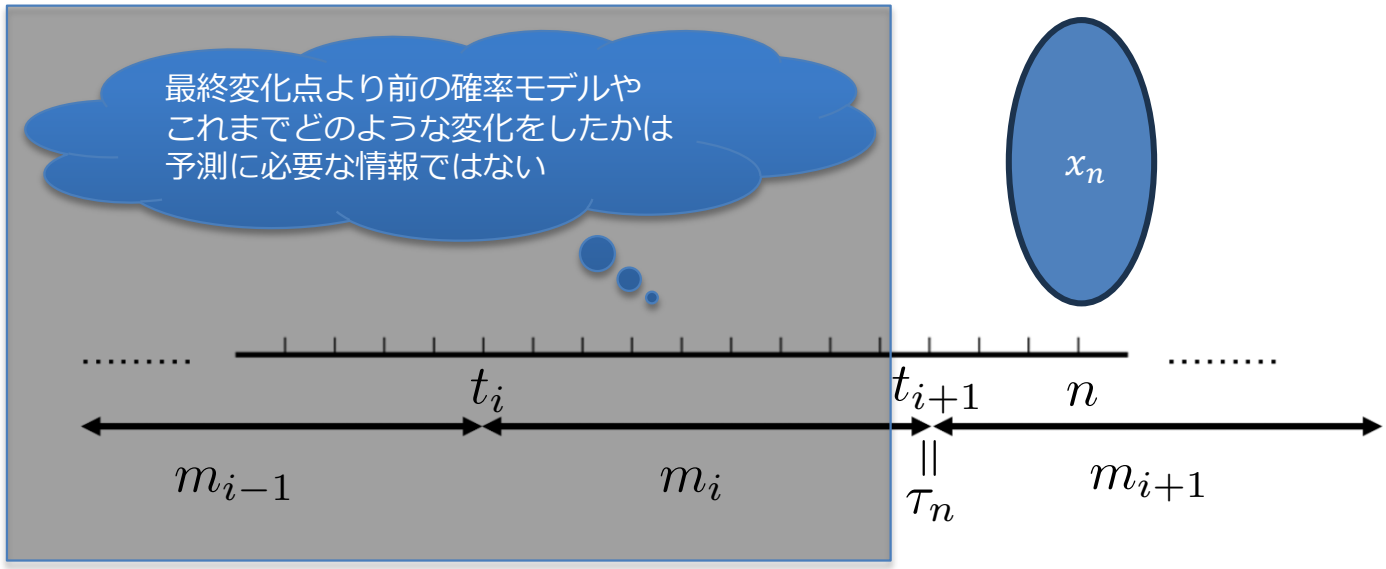
$$\left. \times\ w\left(\Theta^{m^c} \middle| x^{n-1}, m^c, t^c\right) d\Theta^{m^c} \right]$$

指数個存在する
変化点列のパターン
による重み付け

指数オーダーの計算量を要する
→計算量を削減した効率的アルゴリズム

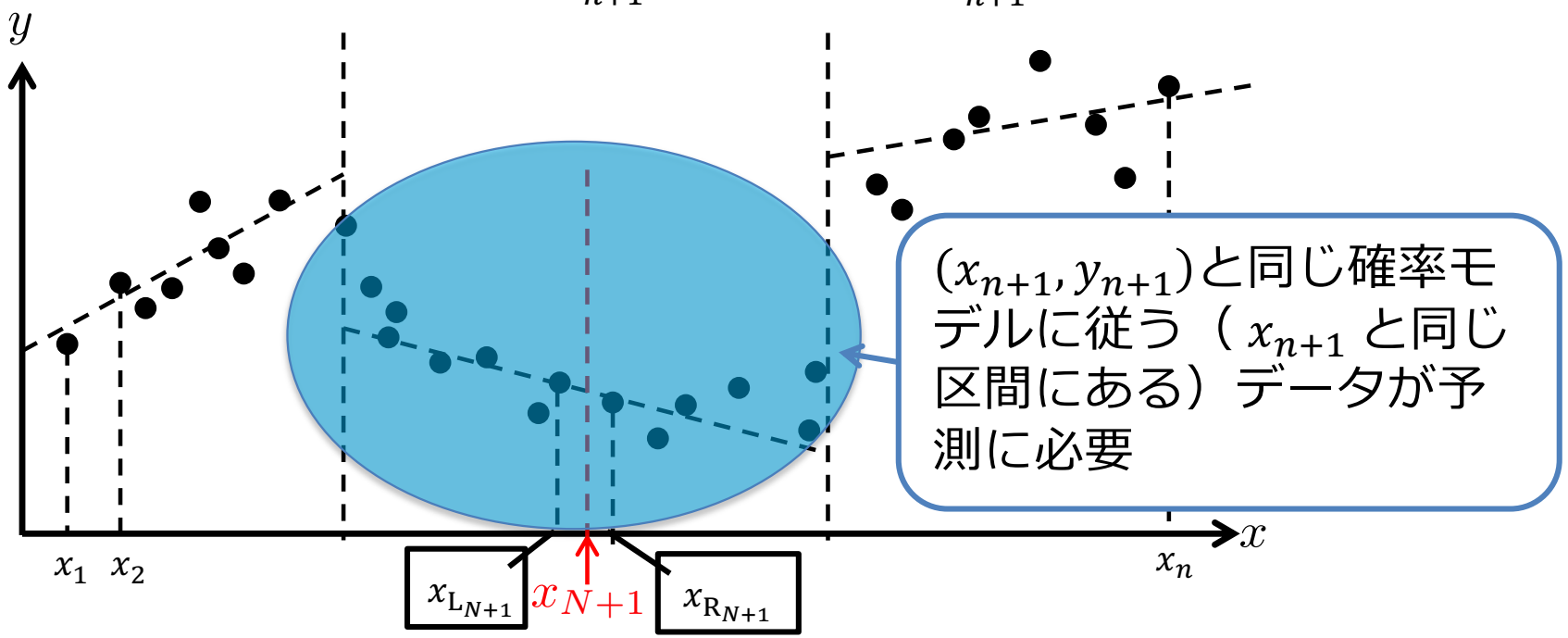
予測アルゴリズムの効率化：区間の統合例１

- x_nの予測をするとき、x_nの直前で確率モデルが変化した時点をτ_nとおくと、$t = \tau_n$より前がどのような確率モデルだったかはx_nの予測に関係がない



予測アルゴリズムの効率化：区間の統合例２

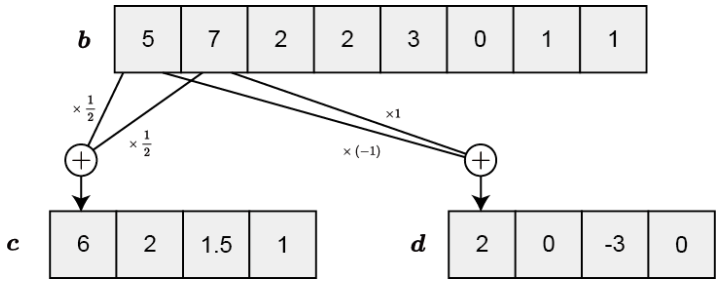
- 便宜上、x_1からx_nは小さい順に添え字を振っているとする
- x軸上で見てx_{n+1}の左隣の値を$x_{L_{n+1}}$、右隣の値を$x_{R_{n+1}}$とおく



信号処理分野での変換モデル：メタ情報としてのウェーブレットパケット木

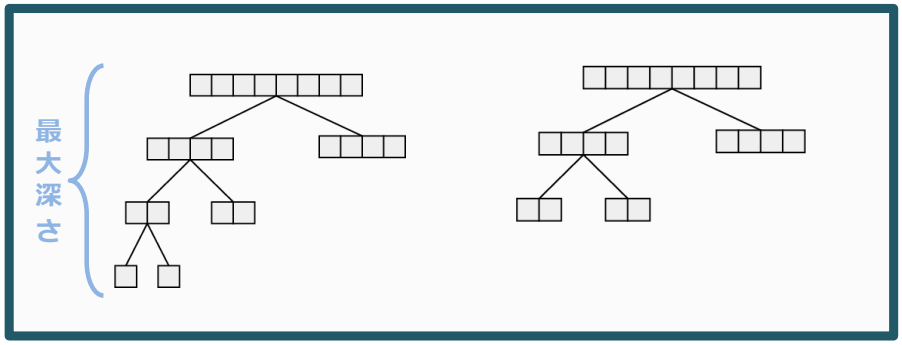
- **ウェーブレット変換（以降WT）**
  - 低周波数領域（左のノード）のみが分解可能
  - WT木は一種類のみ（モデルは固定）

例

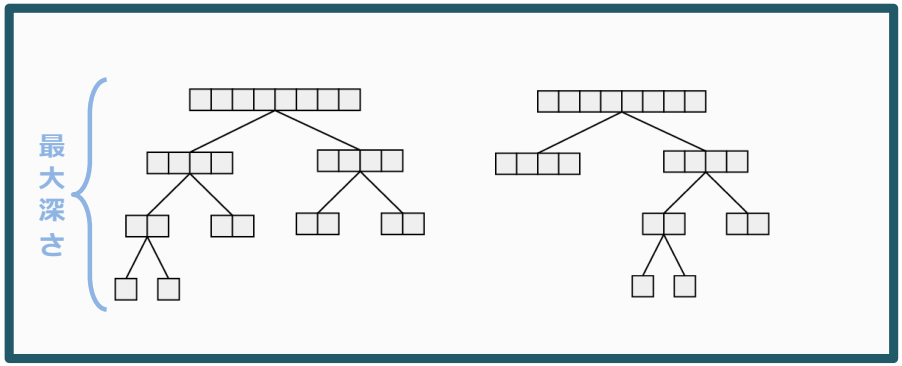


- **ウェーブレットパケット変換（以降WPT）**
  - すべてのノードに対して分解可能
  - WPT木のバリエーションは最大深さに対して**指数的に増加する**.
- **WPTはWTを含む.**

ウェーブレット変換（WT）



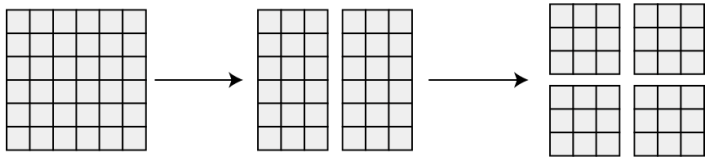
ウェーブレットパケット変換（WPT）

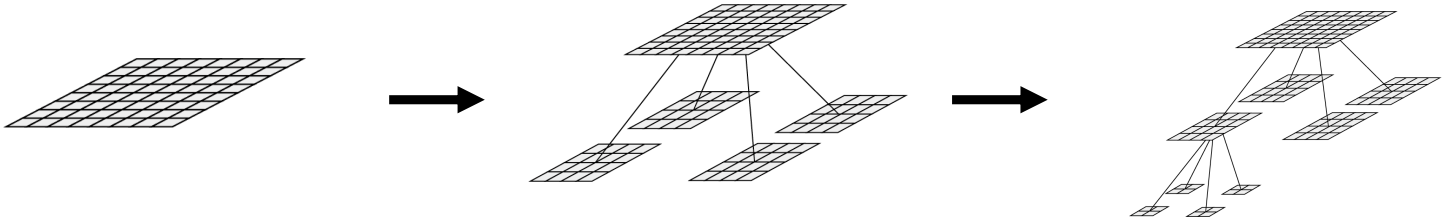


メタ情報としてウェーブレットパケット木モデル　画像

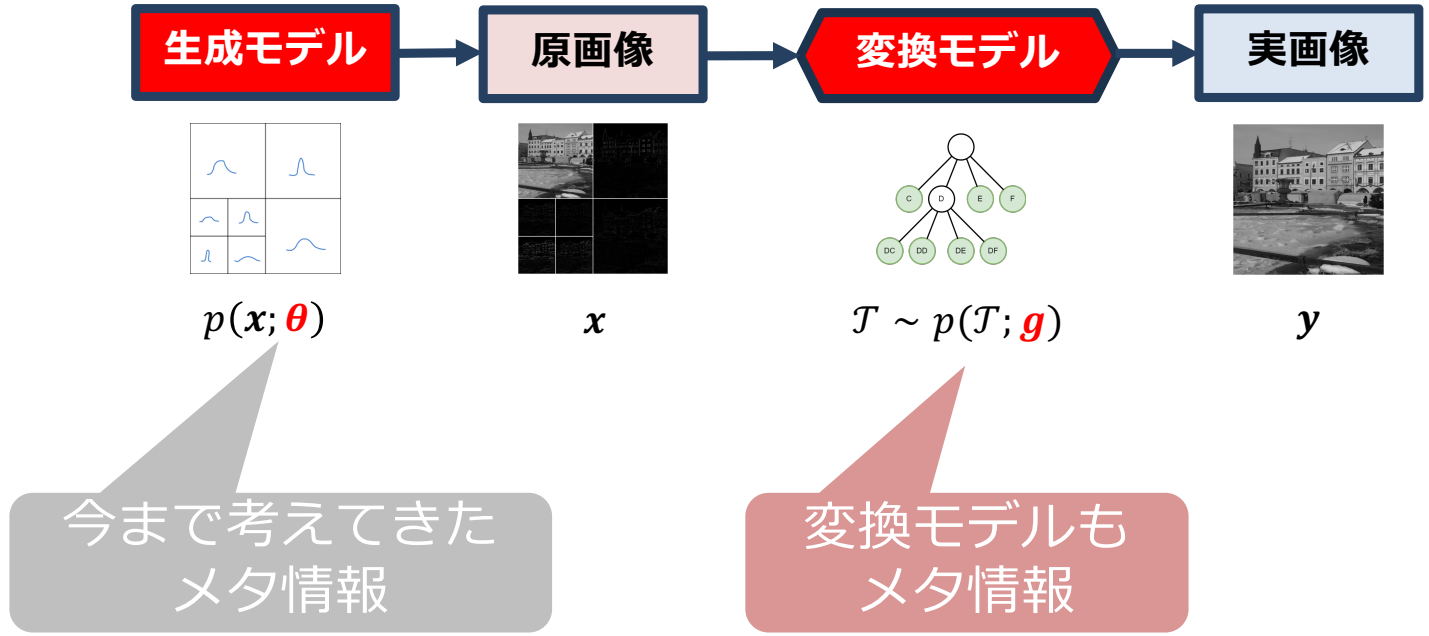
画像（二次元デジタル信号）の場合

　　画像をWPで変換すると，四分木構造になる





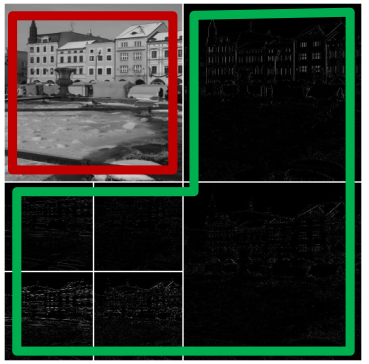
実画像の生成モデル＝原画像生成モデル＋変換モデル



[北原 他 02] ウェーブレットパケット基底を用いた信号推定におけるベイズ決定理論の適用に関する一考察, 電子情報通信学会論文誌 vol.J85-A .
[岡 他 22] 二次元離散ウェーブレットパケット変換の基底が未知の場合のベイズ基準のもと最適なノイズ除去アルゴリズム, SITA2021).

メタ情報：変換モデル（WPT木モデル）と原画生成モデル(Spike and Slab)

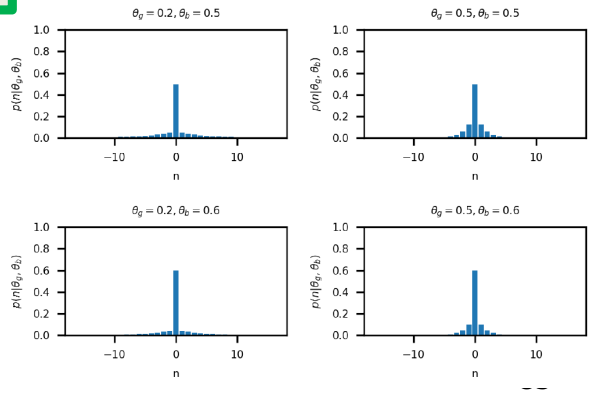
- 変換モデル：$p(\mathcal{T};g) \coloneqq \prod_{s\in\mathcal{I}(\mathcal{T})} g^{(s)} \prod_{s'\in\mathcal{L}(\mathcal{T})}\left(1-g^{(s')}\right)$
- 生成モデル：数値実験より低周波数領域（赤）は一様分布，高周波係数領域（緑）はSpike and Slab分布が正規分布より原画像に適する分布となった.



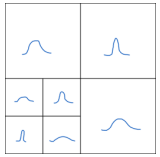
$$p(x) \coloneqq \frac{1}{256}\cdot I_{\{0\le x\le 255\}}$$

$$p_{\mathrm{ss}}(x|\theta_{\mathrm{g}},\theta_{\mathrm{b}}) \coloneqq \begin{cases} \theta_{\mathrm{b}}, & n=0 \\ \frac{1}{2}(1-\theta_{\mathrm{b}})\theta_{\mathrm{g}}(1-\theta_{\mathrm{g}})^{|x|-1}, & n\neq 0 \end{cases}$$

$$p(\theta_{\mathrm{g}};\alpha_{\mathrm{g}},\beta_{\mathrm{g}}) \coloneqq \mathrm{Beta}(\theta_{\mathrm{g}};\alpha_{\mathrm{g}},\beta_{\mathrm{g}})$$
$$p(\theta_{\mathrm{b}};\alpha_{\mathrm{b}},\beta_{\mathrm{b}}) \coloneqq \mathrm{Beta}(\theta_{\mathrm{b}};\alpha_{\mathrm{b}},\beta_{\mathrm{b}})$$

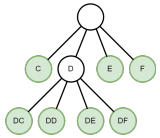


生成モデル → 原画像 → **変換モデル** → 実画像



$p(\boldsymbol{x}|\boldsymbol{\theta}), p(\boldsymbol{\theta};\boldsymbol{h})$

x

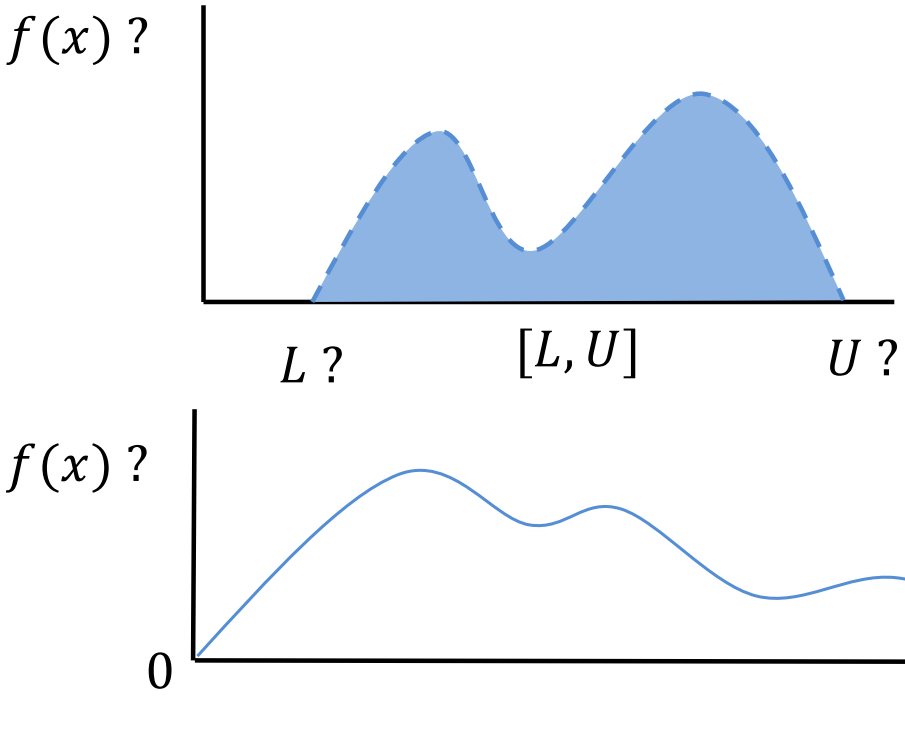


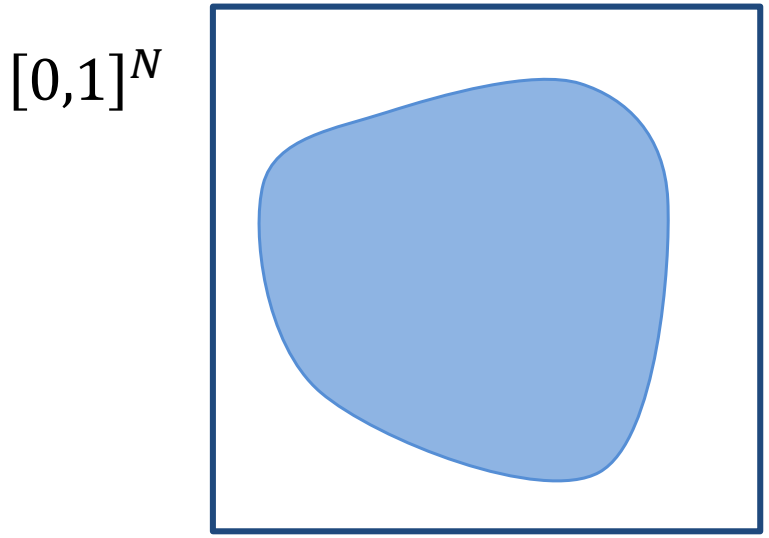
$\mathcal{T} \sim p(\mathcal{T};\boldsymbol{g})$

y

統計学分野のサポート未知な問題

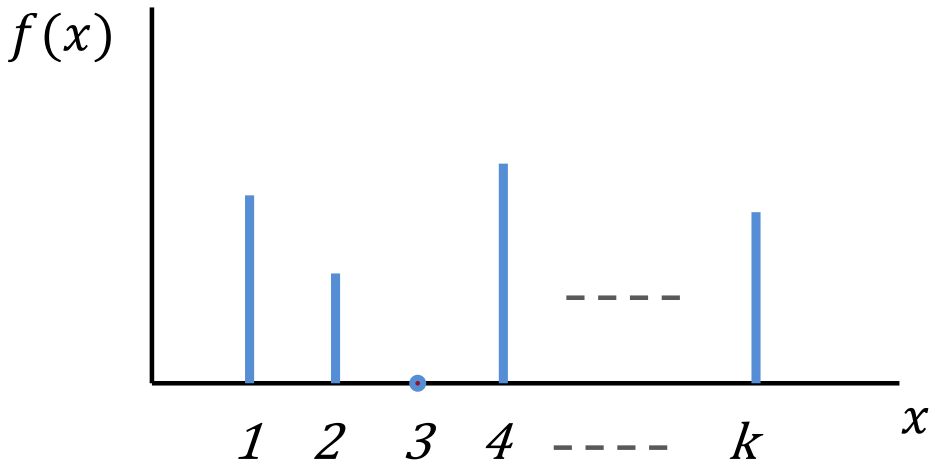
サポートと密度関数が未知 ー＞両方を推定問題　[Kazemi eta. 17]
多次元空間上で滑らかな境界をもつサポート(boundary fragment) の推定　[Gayrand 97]
Lifetime分布のサポートの推定　[Efromovic 24]





情報理論のおけるサポート未知な問題

Assuming that the range of values of the discrete random variable is $A = \{0, 1, \cdots, k\}$, its support is defined as $a = \{i | \theta_i > 0, i \in A\}$, where θ_i is the occurrence probability of a symbol i. The problems with unknown support arise when a is unknown.



情報理論では unknown alphabet 問題

eg)

・エントロピー推定 [Schober eta.12]

・ユニバーサル情報源符号化 [Shtarkov95][Aberg98]

真のアルファベットが未知な情報源符号化

真のアルファベットが未知な情報源符号化

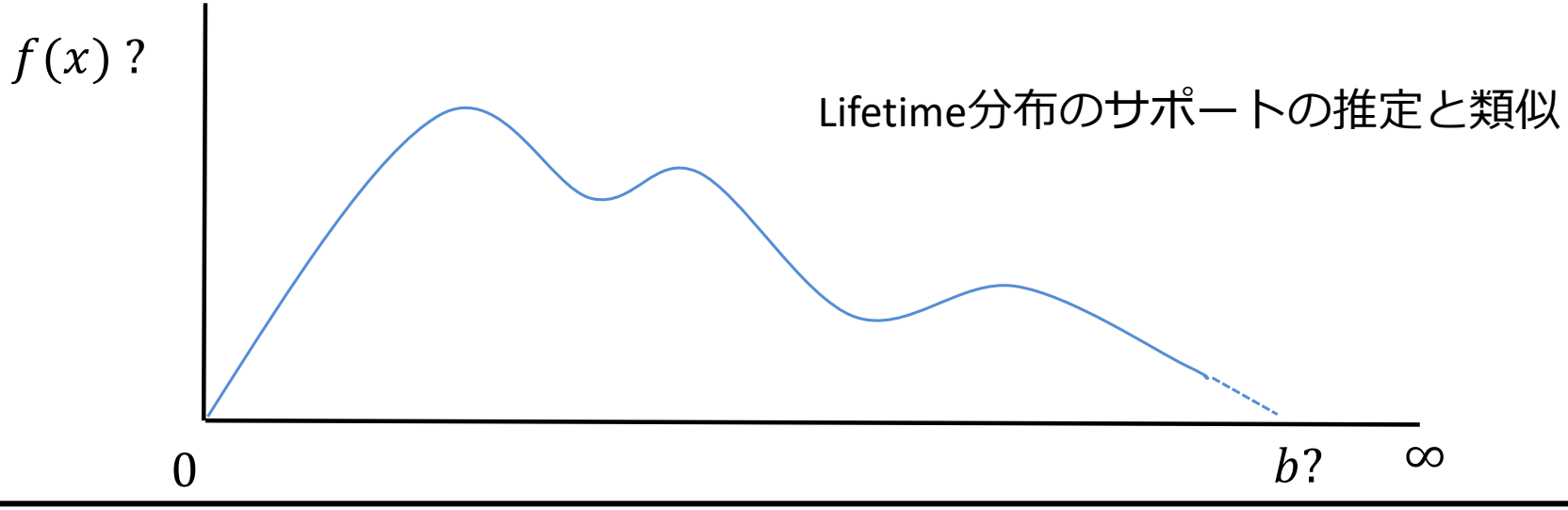
Information sequence: $x^n = x_1, x_2, \cdots, x_n (x_i \in A)$

Alphabet: $A = \{0, 1, \cdots, \infty\}$

True alphabet: A^*?

アルファベット数の増加が
データ数 *n* の *sublinear*

[Shtarkov95][Aberg98][Orlitsky04][Shamir06]



定理：サポート未知の問題のベイズ最適予測

Theorem 1 *$P(\theta_a^{k_a}|a)$ is the prior distribution of $\theta_a^{k_a}$ given a support a, and $P(a)$ is the prior distribution of a, where k_a is the number of the parameters in a. When a support $a \in 2^A$ and its parameters $\theta_a^{k_a} \in \Theta_a^{k_a}$ are unknown, the optimal prediction of decision $d_2(x^{t-1}) = \hat{P}(x_t|x^{t-1})$ under Bayes risk function $BR_3(d_2(x^n))$ is given by*

$$\begin{aligned}\hat{P}_B^A(x_t|x^{t-1}) \\ = \sum_{a \in 2^A} \int P(x_t|x^{t-1}, \theta_a^{k_a}, a) P(\theta_a^{k_a}|a, x^{t-1}) \\ P(a|x^{t-1}) d\theta_a^{k_a}, \end{aligned} \tag{15}$$

where $P(\theta_a^{k_a}|a, x^{t-1})$ is the posterior distribution of $\theta_a^{k_a}$ given a and x^{t-1}, and $P(a|x^{t-1})$ is the posterior distribution of a given x^{t-1}.

特殊な事前分布の仮定

アルファベットモデルの事前分布とパラメータの事前分布に特殊な分布を仮定することで計算量が削減される

Condition 1 *The prior distribution of supports* $a \in 2^A$*:*

$$\forall a \in \{a | k_a = j\}\ P(a) = P(j). \tag{16}$$

Condition 2 *The prior distribution of the probabilistic parameter vector* $P(\theta_a^{k_a}|a)$ *given support* a*:*

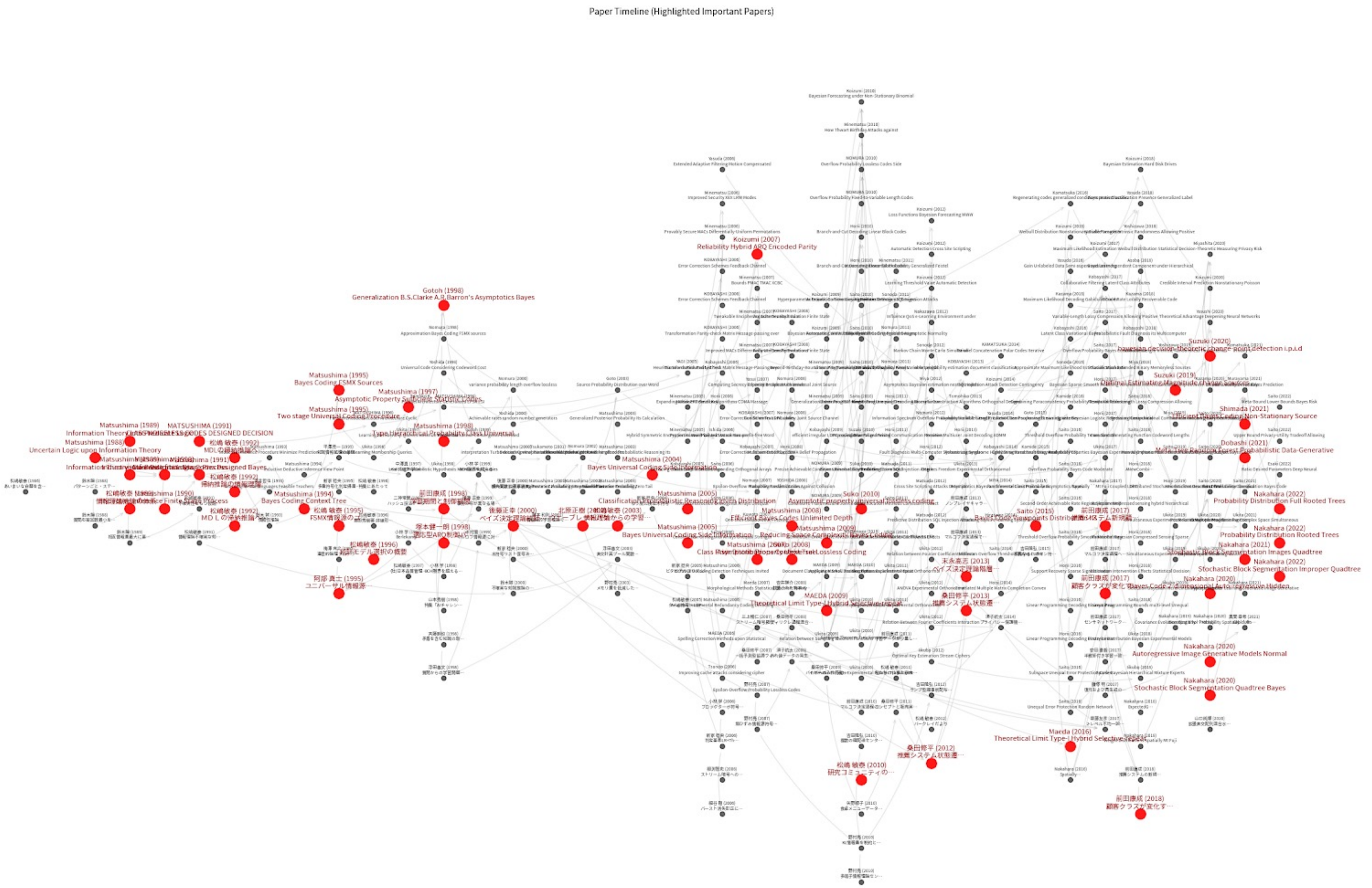
$$P(\theta_a^{k_a}|a) = \frac{\Gamma(k_a\beta)}{\Gamma(\beta)^{k_a}} \prod_{i \in a} \theta_i^{\beta-1}.$$

アルファベット数で事前分布が決まる

パラメータの事前分布に対称性を仮定

[Matsushima 24] Sequential Prediction Problems under Unknown Support, SITA2024

まとめとして？



まとめとして：情報の生成メカニズム（メタ情報）も確率変数で考えたら

AI・データサイエンス系学会の傾向：
Information Theory と Bayes

統合して論じられている
訳ではない

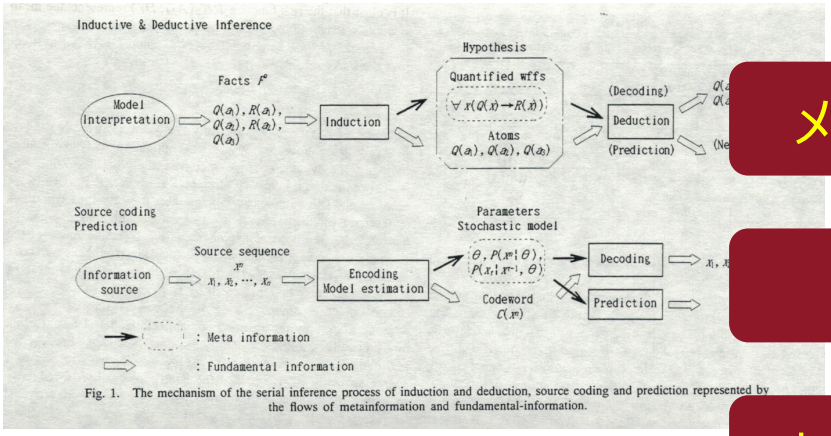


Fig. 1. The mechanism of the serial inference process of induction and deduction, source coding and prediction represented by the flows of metainformation and fundamental-information.

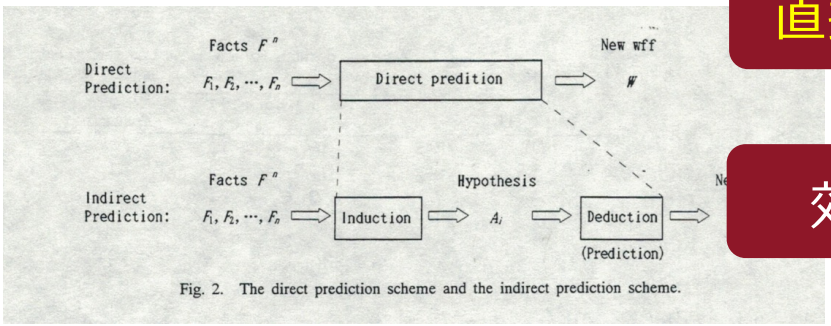


Fig. 2. The direct prediction scheme and the indirect prediction scheme.

メタ情報は確率変数

＋

ベイズ決定理論

＋

直接予測（符号化）

＋

効率的アルゴリズム

様々な情報
を扱う分野

「落ち？」として：情報の生成メカニズム（メタ情報）も確率変数で考えたら

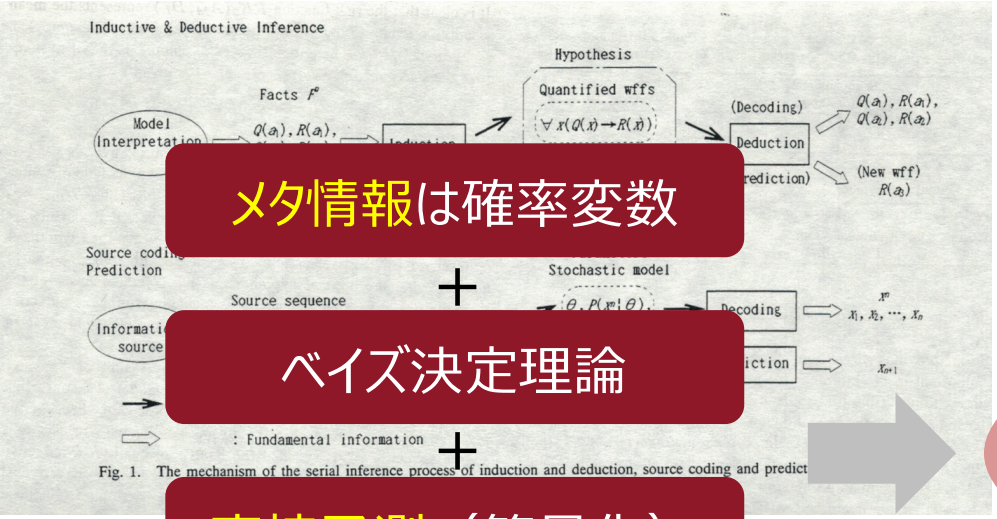


Fig. 1. The mechanism of the serial inference process of induction and deduction, source coding and predict

メタ情報は確率変数

＋

ベイズ決定理論

＋

直接予測（符号化）

＋

効率的アルゴリズム

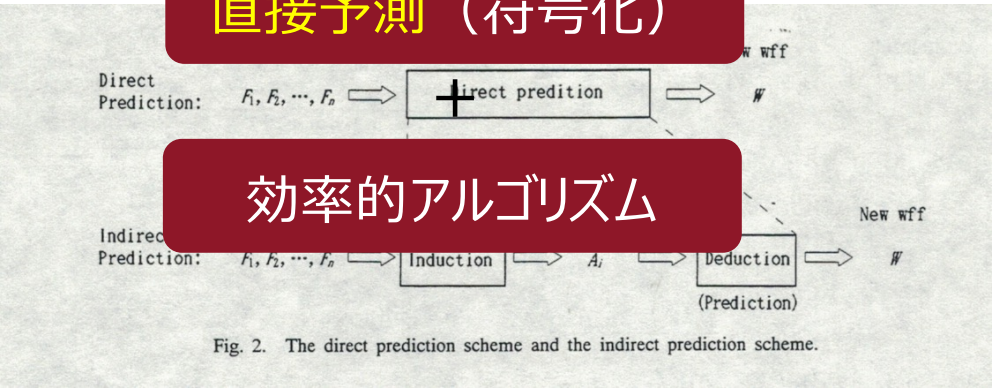


Fig. 2. The direct prediction scheme and the indirect prediction scheme.

Meta-Information Theory ?