ノンパラメトリックベイズに基づく統計的機械学習

Takaki Makino

Division of Project Coordination
University of Tokyo

2010.6.15 IBISML 第1回研究会にて

本発表の目標

- 誰でもノンパラベイズがわかるように
  - 背景知識、用語の知識が必要
  - この本 → のおかげで、ベイズについての知識はだいぶ普及した?
  - わからない人はとことんわからない
- できるだけ背景知識を含めてお話しします
  - Introductory な内容なので、ご存じの方にはつまらないかもしれませんがご了承ください

Introduction: ベイジアンとは何か?

- Quiz: 10万人に1人がかかる病気 θ がある。
  検査 D は、θ 病の人を99.8%陽性と判定するが、
  θ 病でない人も0.1%陽性となってしまう。
- あなたが D 陽性のとき、θ 病である確率は?

$p(D\|\theta)$	D陽性	D陰性
$p(D\|\theta$病$)$	0.998	0.002
$p(D\|$健康$)$	0.001	0.999

$$p(\theta|D) = ?$$

Introduction: ベイジアンとは何か?

- Quiz: 10万人に1人がかかる病気 θ がある。検査 D は、θ 病の人を99.8%陽性と判定するが、θ 病でない人も0.1%陽性となってしまう。
- あなたが D 陽性のとき、θ 病である確率は?

$p(D\|\theta)$	D陽性	D陰性
$p(D\|\theta$病$)$	0.998	0.002
$p(D\|$健康$)$	0.001	0.999

	$p(\theta)$
θ病	0.00001
健康	0.99999

Bayes' Theorem

$$p(\theta \mid D) = \frac{p(D \mid \theta)\,p(\theta)}{p(D)}$$

Posterior | Likelihood | Prior

Introduction: ベイジアンとは何か?

- Quiz: 10万人に1人がかかる病気 θ がある。
  検査 D は、θ 病の人を99.8%陽性と判定するが、
  θ 病でない人も0.1%陽性となってしまう。
- あなたが D 陽性のとき、θ 病である確率は?

$p(D\|\theta)$	D陽性	D陰性
$p(D\|\theta$病$)$	0.998	0.002
$p(D\|$健康$)$	0.001	0.999

	$p(\theta)$
θ病	0.00001
健康	0.99999

$p(\theta\|D)$	$p(\theta\|D$陽性$)$	$p(\theta\|D$陰性$)$
θ病	≈0.01	≈0.00000002
健康	≈0.99	≈0.99999998

Prior (事前分布)
Likelihood (尤度)
Posterior (事後分布)

Introduction: ベイジアンとは何か?

- ベイジアン確率論: 仮説の「確からしさ」を確率で表現し、ベイズの定理で推論する
  - θ: 「宇宙人がいる」or「いない」という仮説
    D: 「UFOを目撃する」or「しない」という事象
  - $p(\theta)$: 宇宙人の存在はどのぐらい確からしいか?
    $p(D|\theta)$: 宇宙人がいる時にUFOを目撃する確率
    $p(\theta|D)$: UFOを目撃したあとでは、
    宇宙人の存在はどのぐらい確からしいか?

Bayes' Theorem

$$\text{Posterior}\quad p(\theta \mid D) = \frac{\text{Likelihood}\ p(D \mid \theta)\ \text{Prior}\ p(\theta)}{p(D)}$$

Introduction: ベイジアンとは何か?

- ベイジアン確率論: 仮説の「確からしさ」を確率で表現し、ベイズの定理で推
  - θ: 「宇宙人がいる」or「いない
    D: 「UFOを目撃する」or「しな
  - $p(\theta)$: 宇宙人の存在はどのぐらい確からしいか?
    $p(D|\theta)$: 宇宙人がいる時にUFOを目撃する確率
    $p(\theta|D)$: UFOを目撃したあとでは、宇宙人の存在はどのぐらい確からしいか?

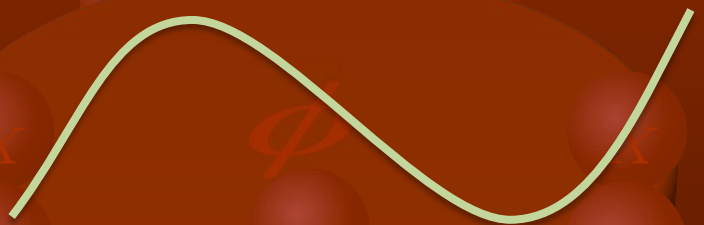
一般の (頻度主義的) 考え方からすると、確率では書けない

Bayes' Theorem

$$\text{Posterior}\quad p(\theta|D) = \frac{\text{Likelihood}\ p(D|\theta)\ \text{Prior}\ p(\theta)}{p(D)}$$

Introduction: 機械学習とベイズ

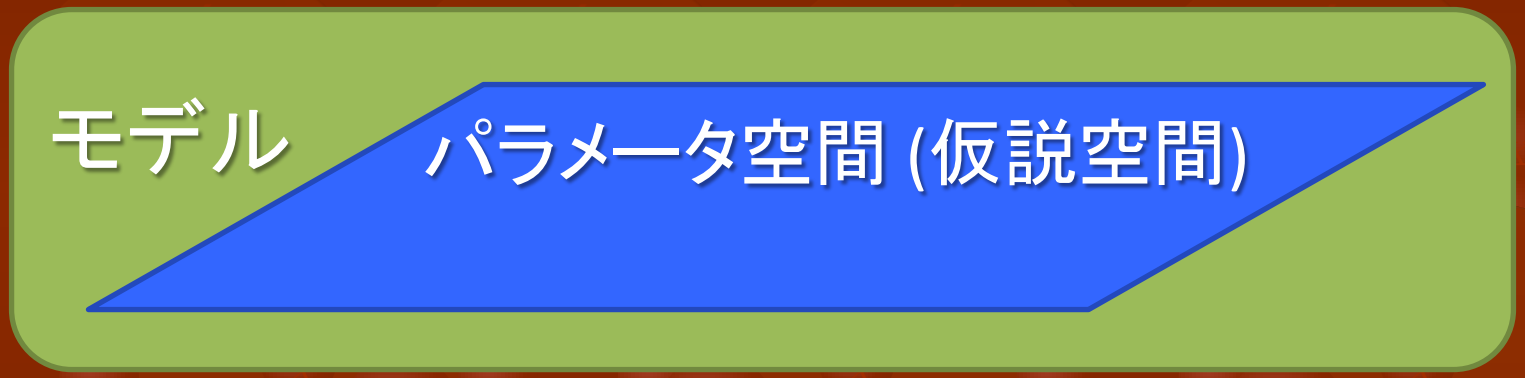
- 機械学習 = パラメータ決定 = 仮説の選択



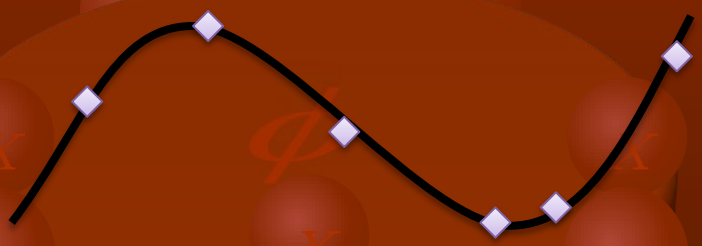
真の分布

Introduction: 機械学習とベイズ

- 機械学習 = パラメータ決定 = 仮説の選択



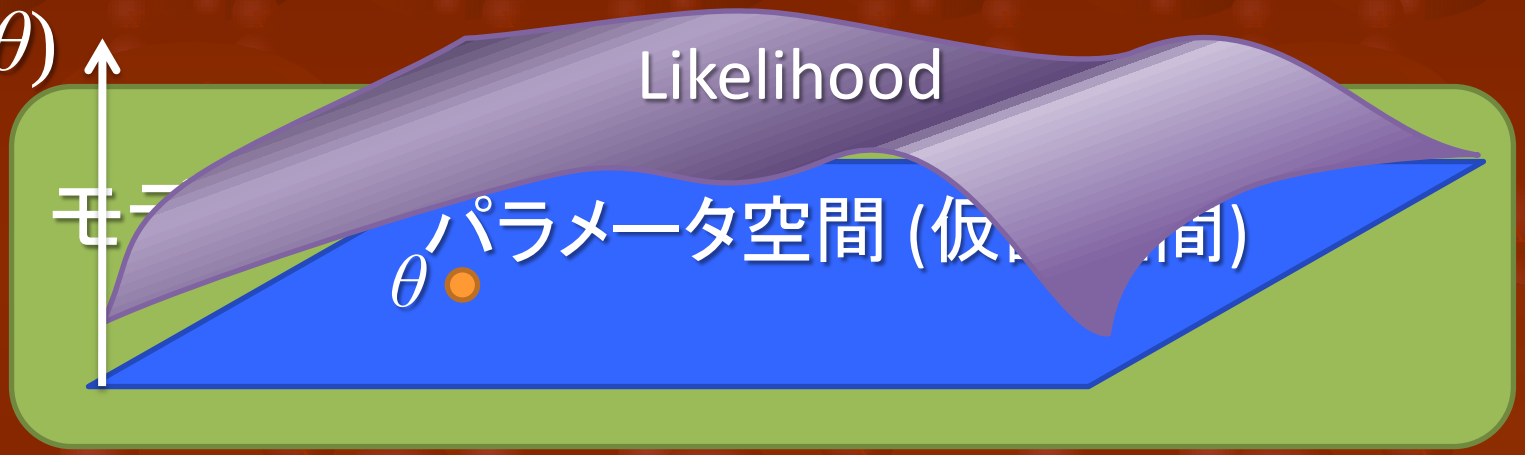
観測したデータ D

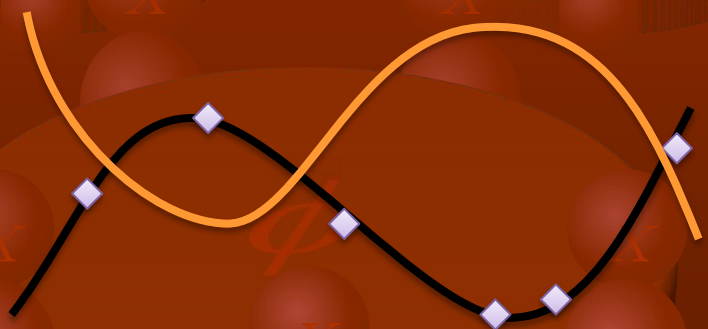


真の分布

Introduction: 機械学習とベイズ

- 機械学習 = パラメータ決定 = 仮説の選択





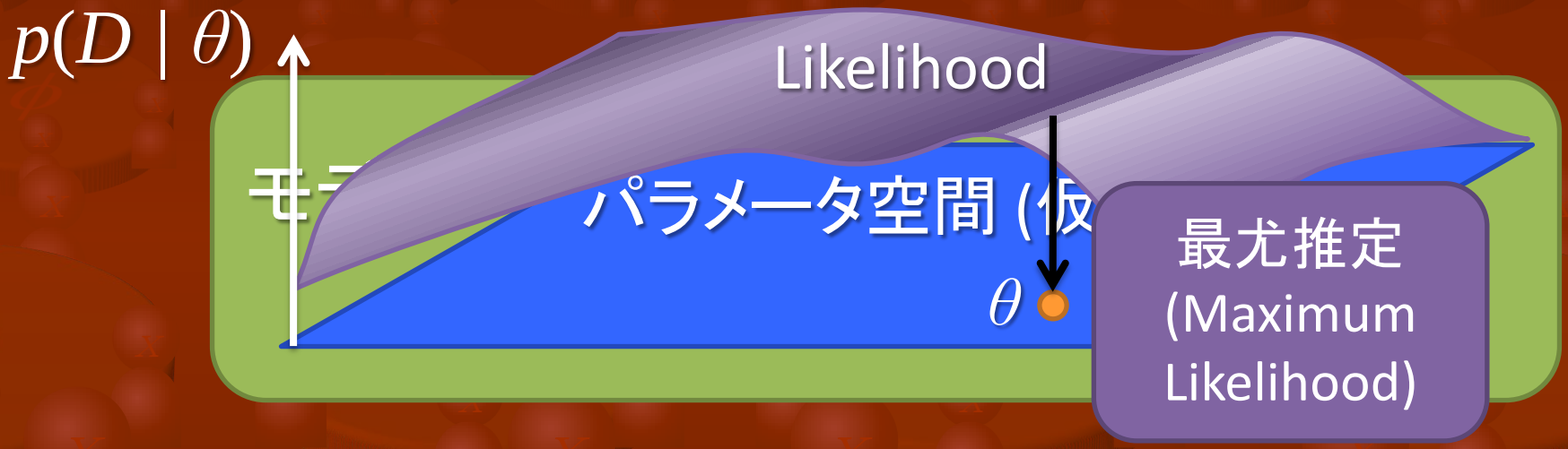
観測したデータ D

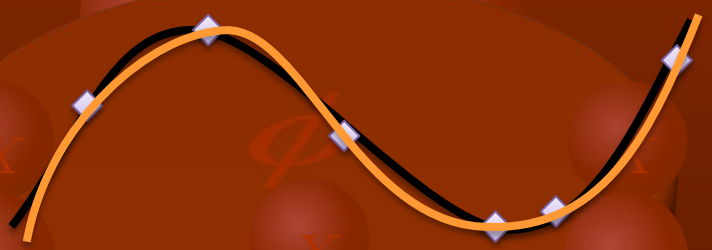
分布の仮説

真の分布

Introduction: 機械学習とベイズ

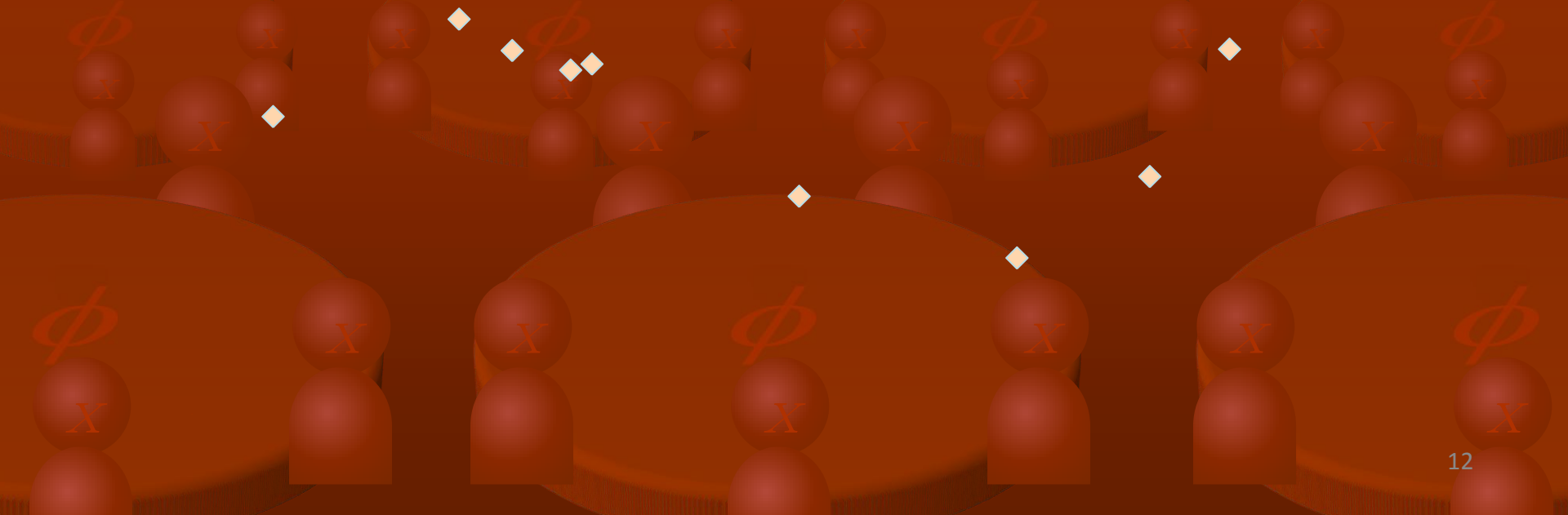
- 機械学習 = パラメータ決定 = 仮説の選択





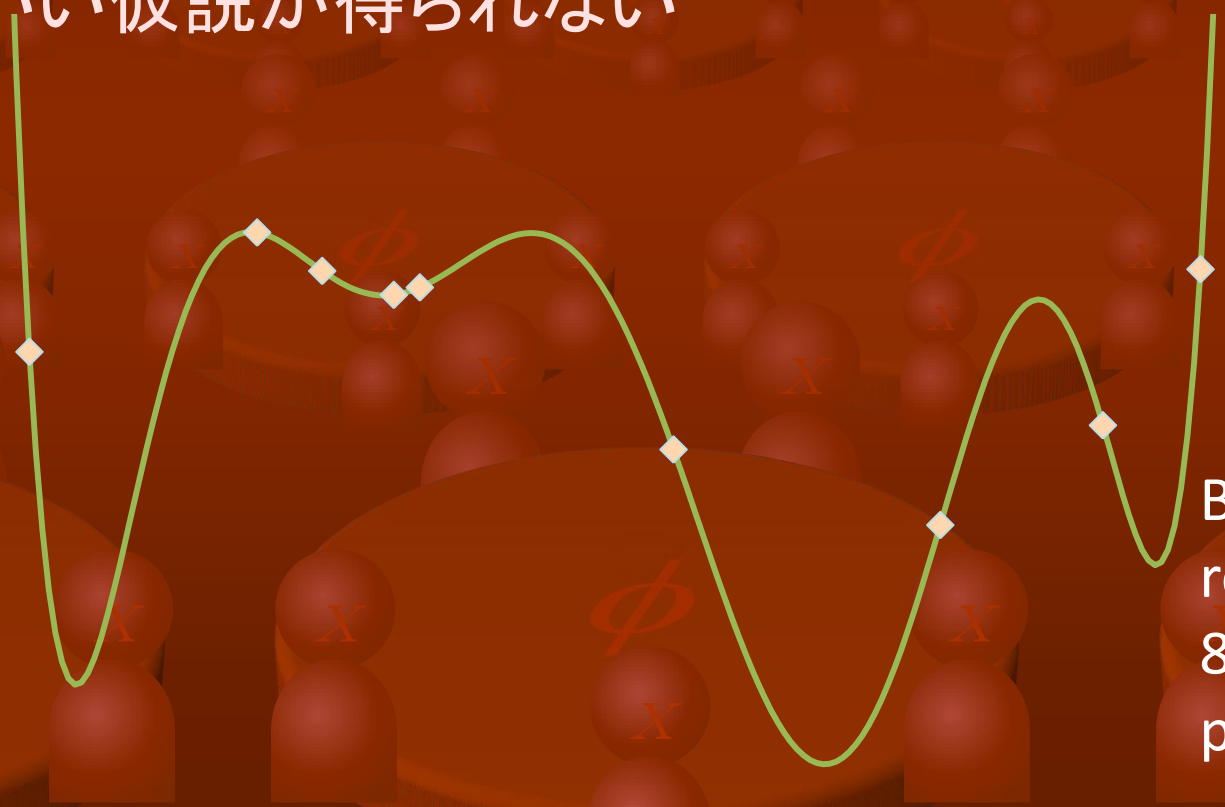
Introduction: 機械学習とベイズ

- 過学習問題
  - Likelihood 最大のパラメータを選んでも、いい仮説が得られない



Introduction: 機械学習とベイズ

- 過学習問題
  - Likelihood 最大のパラメータを選んでも、いい仮説が得られない



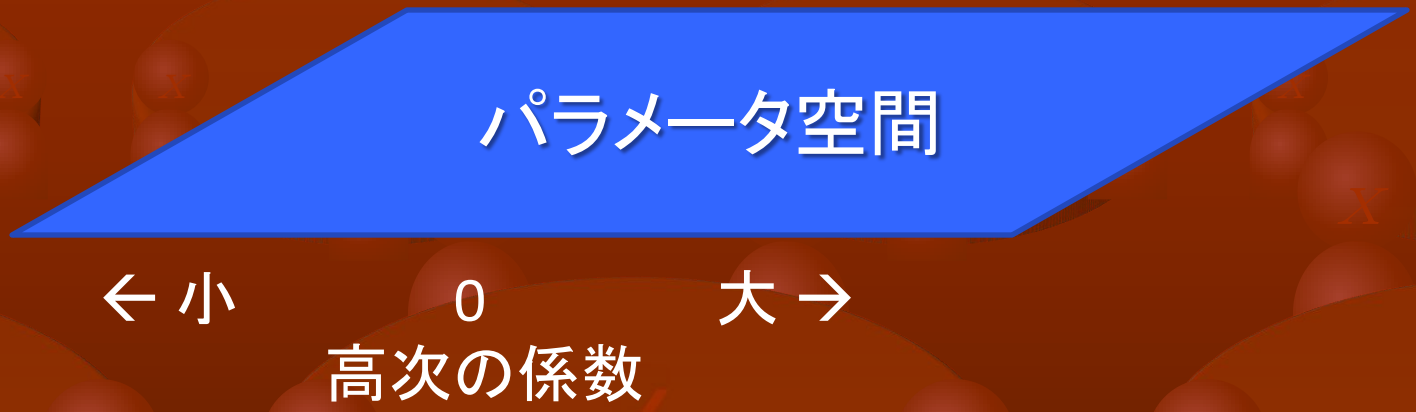
Best-fit regression by 8th-degree polynomial

Introduction: 機械学習とベイズ

- 非ベイジアンの考え方:
  - 複雑すぎるモデルを選んだのが原因
  - 適切な複雑さのモデルを選ぶべし
- ベイジアンの考え方:
  - Likelihood $p(D|\theta)$ だけで考えているのが原因
  - Posterior $p(\theta|D)$ を考えることが重要
  - そのために必要なのは Prior $p(\theta)$、つまり「あるパラメータ θ がどれだけ確からしいか」に関する (データ D 観測前の) 知識

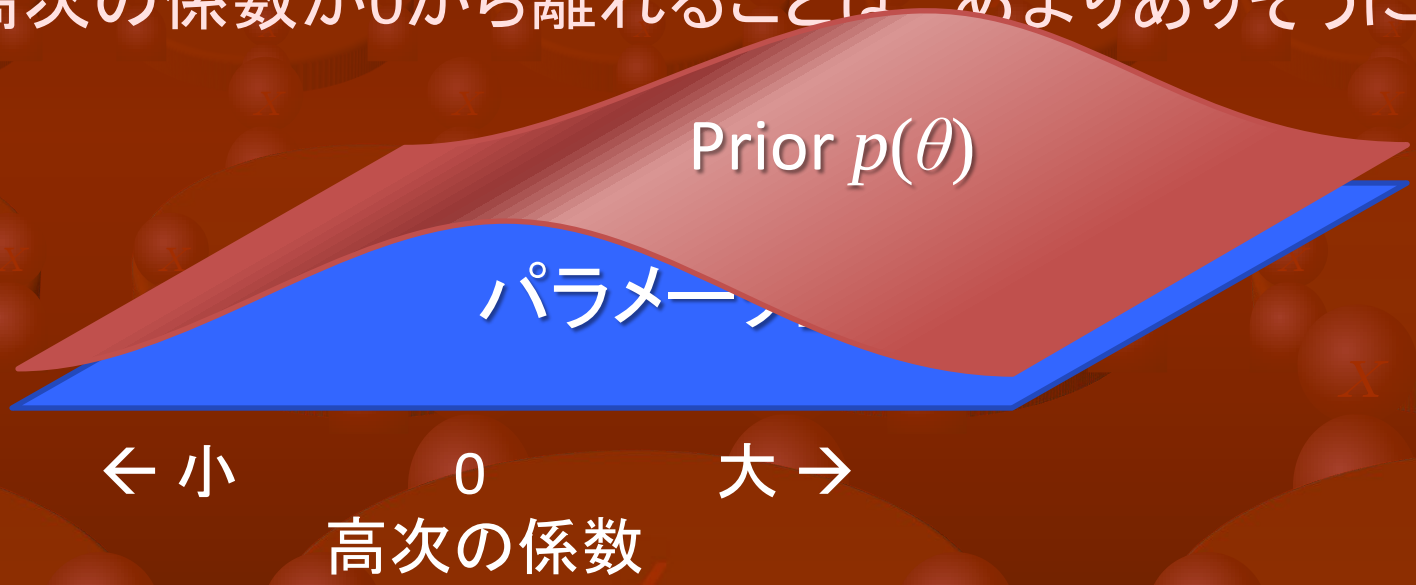
Introduction: 機械学習とベイズ

- Prior: データ観測前の (主観的な) 仮説の確からしさを、確率分布の形で表現したもの
  - 高次の係数が0から離れることは、あまりありそうにない



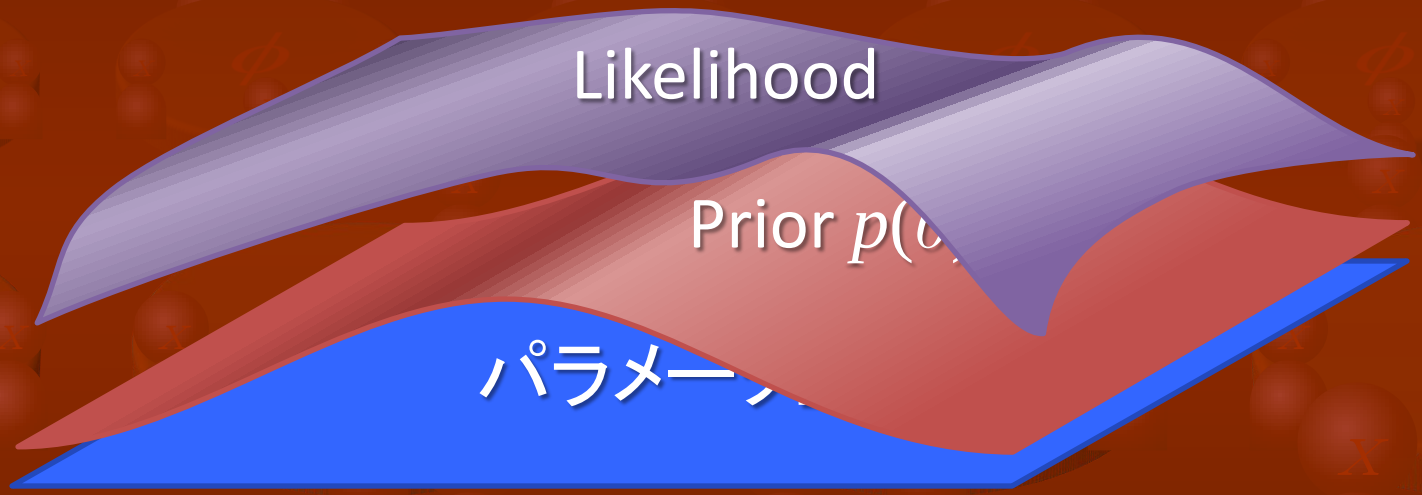
Introduction: 機械学習とベイズ

- Prior: データ観測前の (主観的な) 仮説の確からしさを、確率分布の形で表現したもの
  - 高次の係数が0から離れることは、あまりありそうにない



Introduction: 機械学習とベイズ

- Posterior: データ観測後の確からしさ



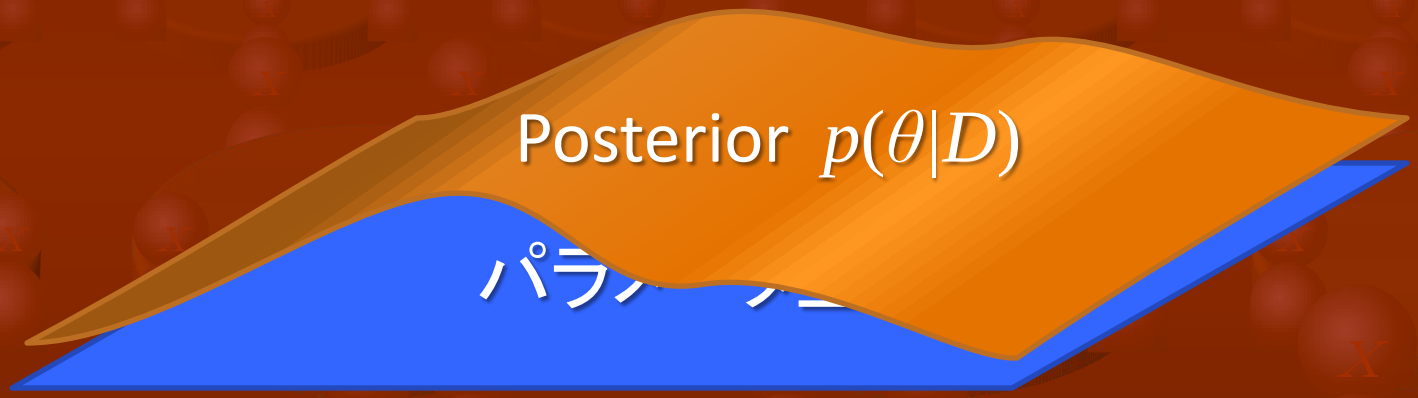
観

Bayes' Theorem

$$\text{Posterior}\quad p(\theta \mid D) = \frac{\text{Likelihood}\ p(D \mid \theta)\ \text{Prior}\ p(\theta)}{p(D)}$$

Introduction: 機械学習とベイズ

- Posterior: データ観測後の確からしさ



観

Bayes' Theorem

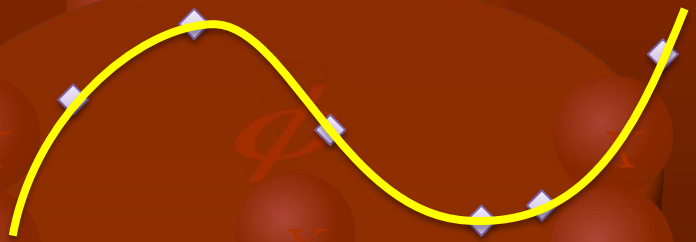
$$p(\theta \mid D) = \frac{p(D \mid \theta)\,p(\theta)}{p(D)}$$

Posterior | Likelihood | Prior

Introduction: 機械学習とベイズ

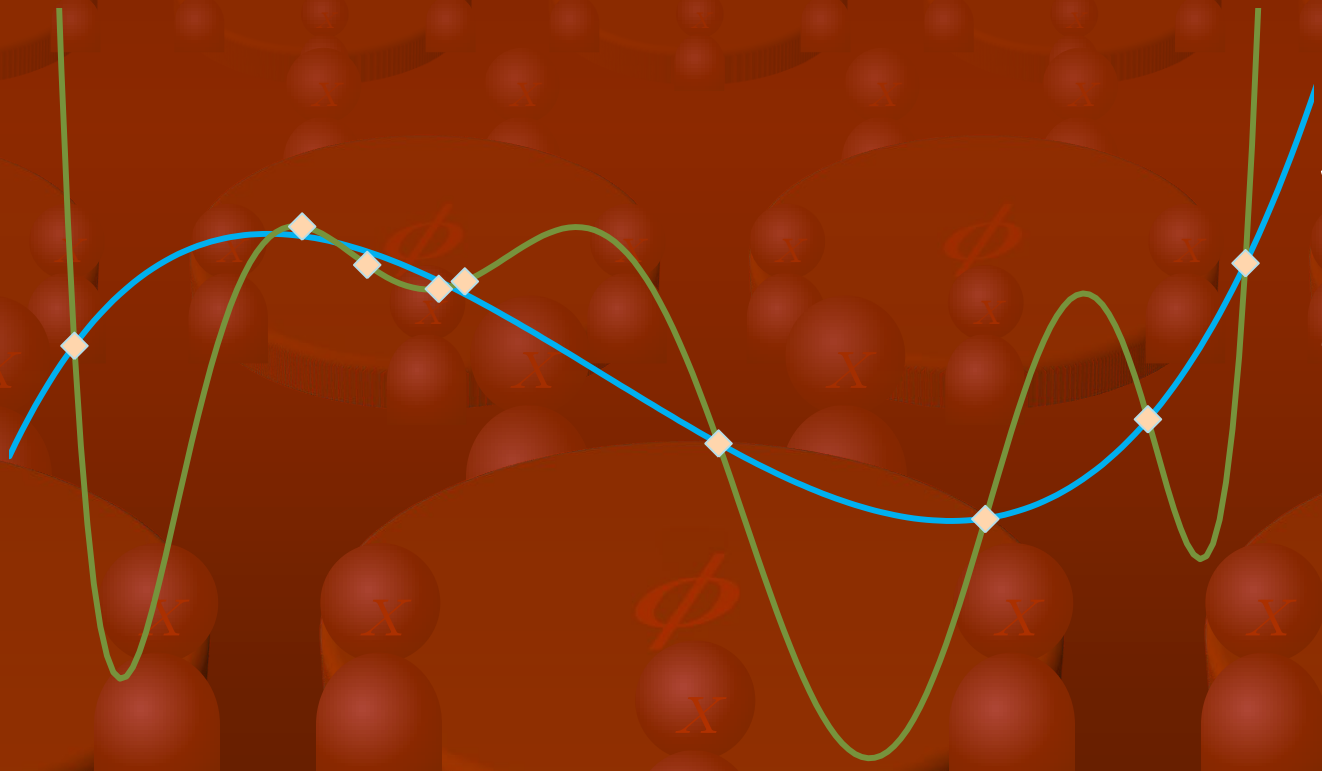
- Posterior: データ観測後の確からしさ

観測したデータ D



Introduction: 機械学習とベイズ

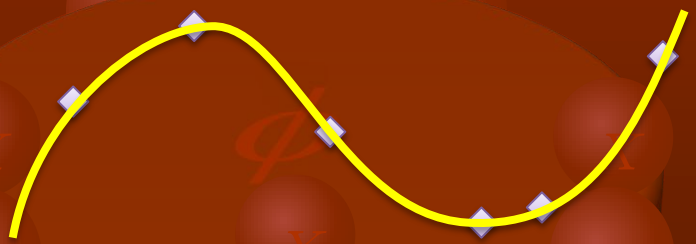
- 適切な prior があれば、過学習を起こしにくい



Regression with prior using 8^{th}-degree polynomial

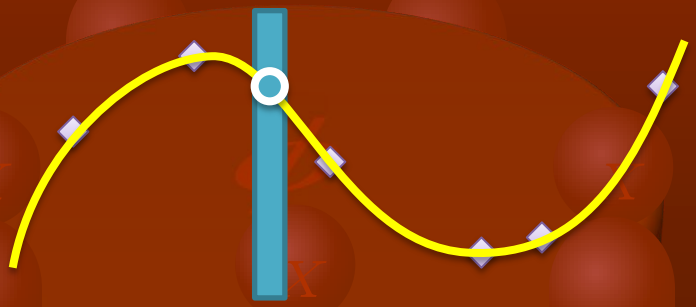
Introduction: 機械学習とベイズ

- なぜパラメータ θ の同定にこだわる?



Introduction: 機械学習とベイズ

- なぜパラメータ θ の同定にこだわる?
- 欲しいものは、モデルに基づく何らかの予測



未知のケース x に対応する予測

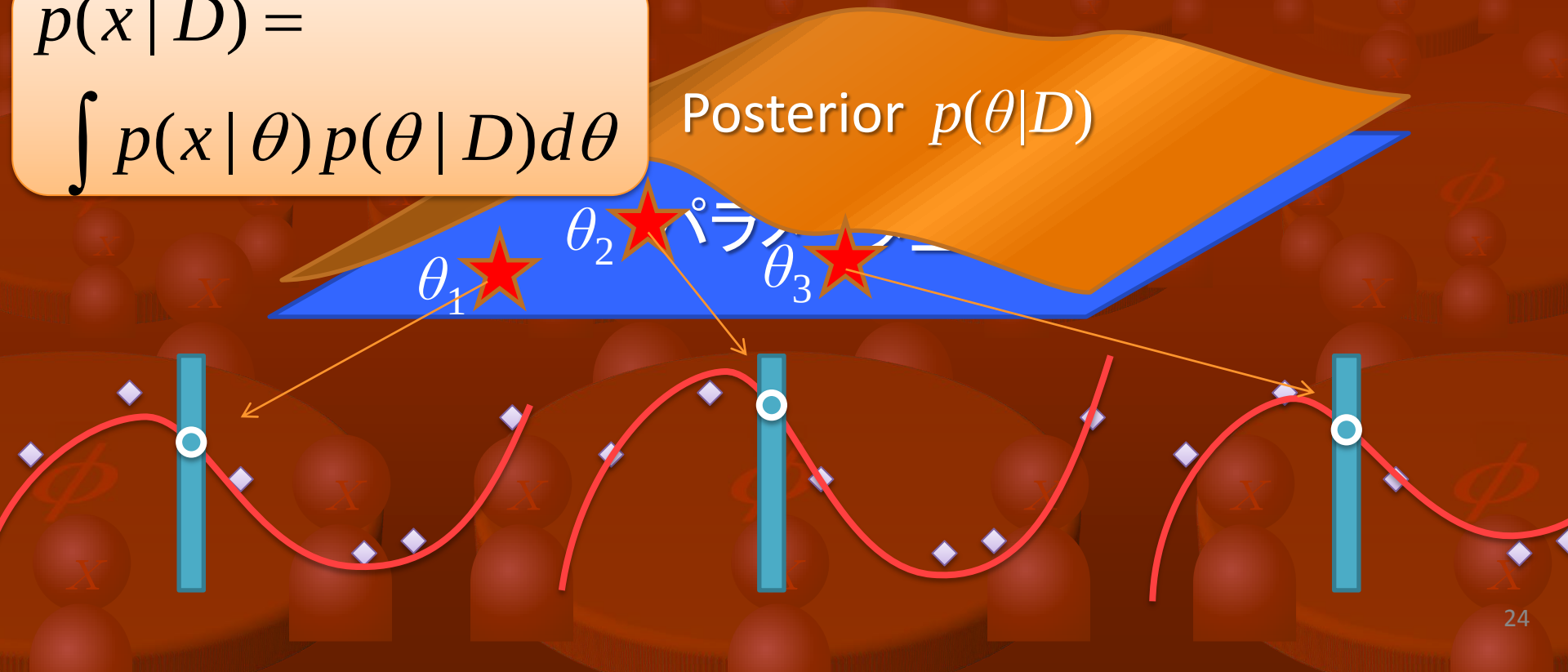
Introduction: 機械学習とベイズ

- なぜパラメータ θ の同定にこだわる?
- 欲しいものは、モデルに基づく何らかの予測
- θ をひとつ選ぶ = posterior が含む多くの情報を利用しない

Introduction: 機械学習とベイズ

- Posterior 全体から予測をしたい!
- パラメータ θ を積分消去 → 予測分布 $p(x|D)$

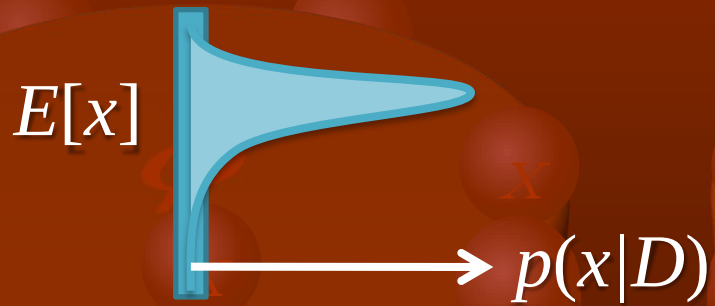
$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$



Introduction: 機械学習とベイズ

- Posterior 全体から予測をしたい!
- パラメータ θ を積分消去 → 予測分布 $p(x|D)$

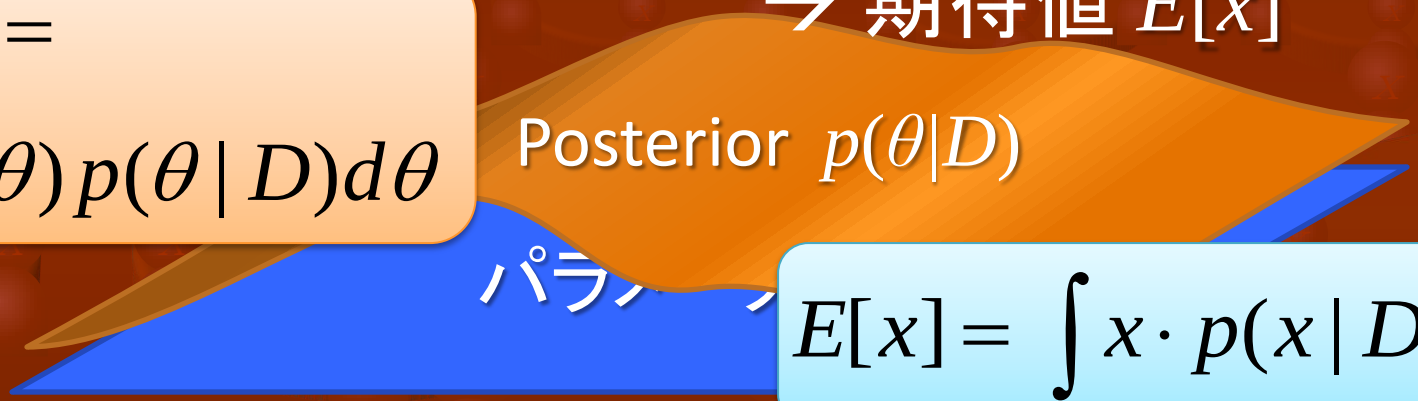
$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$



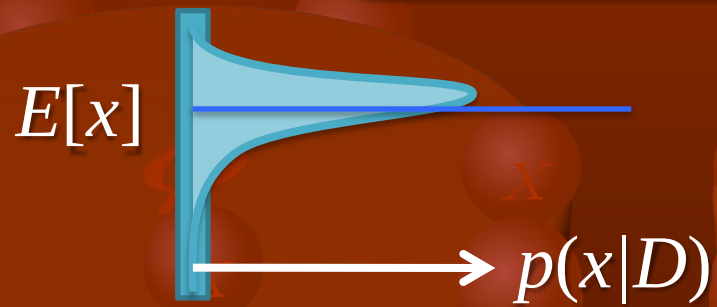
Introduction: 機械学習とベイズ

- Posterior 全体から予測をしたい!
- パラメータ θ を積分消去 → 予測分布 $p(x|D)$
  → 期待値 $E[x]$

$$p(x|D) = \int p(x|\theta)\,p(\theta|D)\,d\theta$$

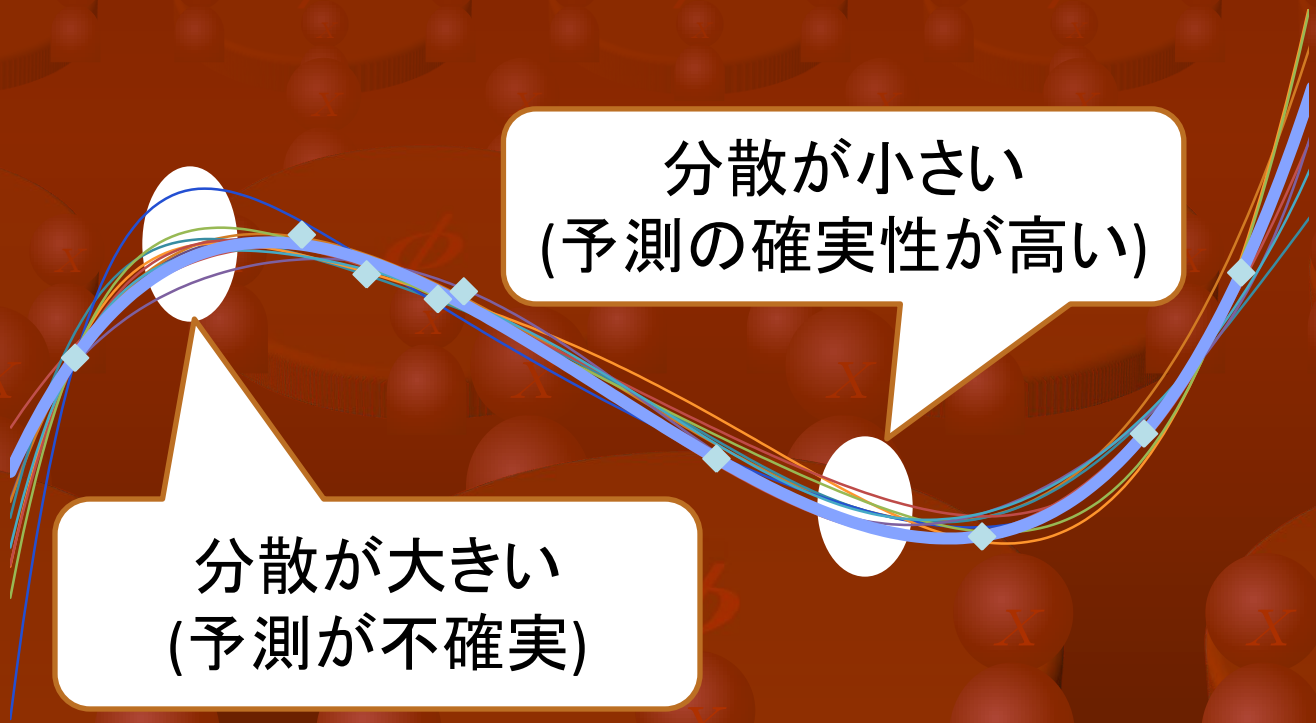


$$E[x] = \int x \cdot p(x|D)\,dx$$



Introduction: 機械学習とベイズ

- 予測の分布 $p(x|D)$ から、予測の確からしさ(不確かさ) を知ることができる



Introduction: ベイジアン機械学習のまとめ

- パラメータ=モデル仮説の「確からしさ」を確率分布で考える
  - ベイズの定理で、データからパラメータ分布を推論
- 過学習を起こしにくい
  - 適切な prior が必要 (モデル選択よりは簡単)
  - 自動モデル選択とみなすこともできる
- Posterior 全体を使って予測分布を作れる
  - パラメータ空間の座標の取り方に依存しない
  - 予測の不確かさが明示的に表現される
- 学習が理論的に、きれいに記述できる
  - 主観的要素 (prior) と推論の分離

Talk Outline

- Introduction
  - ベイジアンとは何か?
  - 機械学習とベイズ
- ベイズHMMモデル
  - Dirichlet 分布
  - Bayesian HMM
- ノンパラメトリックベイズHMMモデル
  - Dirichlet process
  - Hierarchical Dirichlet process
  - Infinite Hidden Markov Model
- HMMの状態の階層的クラスタリング

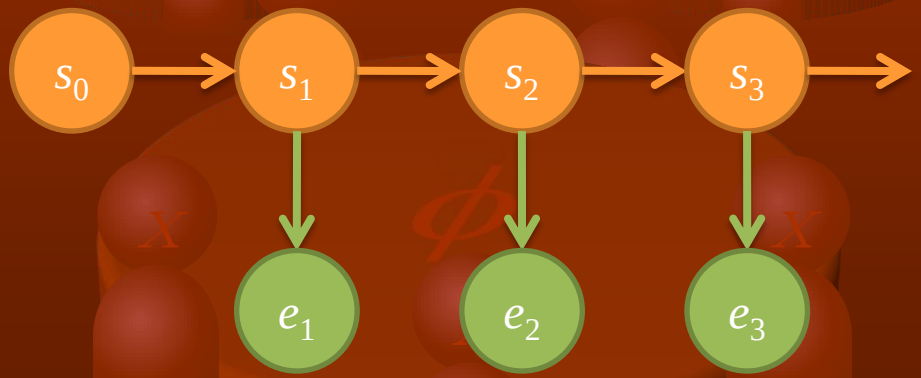
隠れマルコフモデル (HMM)

- 時系列の確率分布を表わすモデル
- 状態 s_t は直前の状態 s_{t-1} にのみ依存

$$p(s_{t+1} \mid s_1, s_2, \ldots, s_t) = p(s_{t+1} \mid s_t)$$

- 出力記号 e_t は s_t にのみ依存

$$p(e_t \mid s_1, \ldots, s_t, e_1, \ldots, e_{t-1}) = p(e_t \mid s_t)$$



HMMの行列表現

- 状態を整数 $\{1,\ldots,N\}$ で、出力を$\{1,\ldots,M\}$で表わす
  → パラメータは2つの行列で表わされる

$$a_{ji} = p(s_{t+1} = j \mid s_t = i)$$

$$b_{ki} = p(e_t = k \mid s_t = i)$$

列 i は、状態 i の後の状態の分布を表わす

列 i は、状態 i から出力される記号の分布を表わす

Transition Matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1N} \\ a_{21} & \cdots & a_{2i} & \cdots & a_{2N} \\ \vdots & & \vdots & & \vdots \\ a_{N1} & \cdots & a_{Ni} & \cdots & a_{NN} \end{pmatrix}$$

Emission Matrix

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1i} & \cdots & b_{1N} \\ b_{21} & \cdots & b_{2i} & \cdots & b_{2N} \\ \vdots & & \vdots & & \vdots \\ b_{M1} & \cdots & b_{Mi} & \cdots & b_{MN} \end{pmatrix}$$

HMMのベイズ的扱い方

- Prior、Posterior を表わすために、HMMのパラメータ空間の上の確率分布を記述したい
  → Dirichlet 分布を組み合わせて使う
  (多項分布のパラメータ上の確率分布)

Transition Matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1N} \\ a_{21} & \cdots & a_{2i} & \cdots & a_{2N} \\ \vdots & & \vdots & & \vdots \\ a_{N1} & \cdots & a_{Ni} & \cdots & a_{NN} \end{pmatrix}$$

Emission Matrix

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1i} & \cdots & b_{1N} \\ b_{21} & \cdots & b_{2i} & \cdots & b_{2N} \\ \vdots & & \vdots & & \vdots \\ b_{M1} & \cdots & b_{Mi} & \cdots & b_{MN} \end{pmatrix}$$

多項分布
(multinomial distribution)

- N値離散確率変数の確率分布

$$p(X=j\,|\,\boldsymbol{\theta})=\theta_j$$

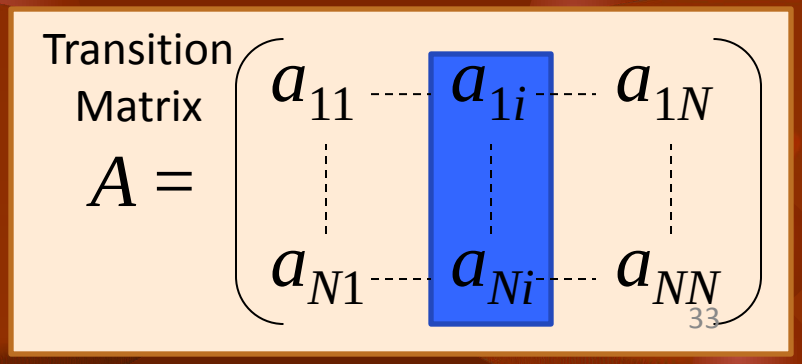
  - N 個のパラメータで表わされる

$$\boldsymbol{\theta}=(\theta_1,\ldots,\theta_N)$$

  - 確率分布としての制約を満たす

$$\theta_i \geq 0 \qquad \sum_{i=1}^{N}\theta_i = 1$$

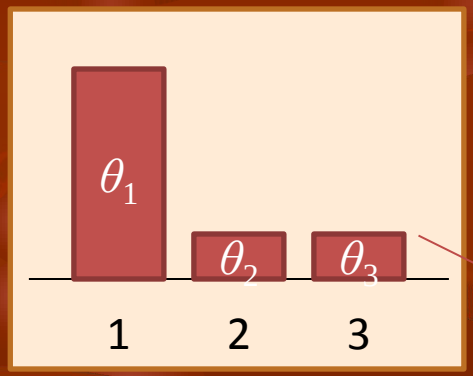
- HMM の遷移行列の
  各列は多項分布である



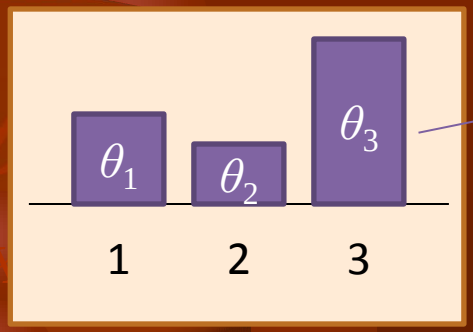
Dirichlet 分布 (1)

- 多項分布の prior を表わすのに便利な分布
  - 多項分布のパラメータ $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ 上の確率分布
  - ハイパーパラメータ $\alpha_1 \ldots \alpha_N$ が分布の形を表わす

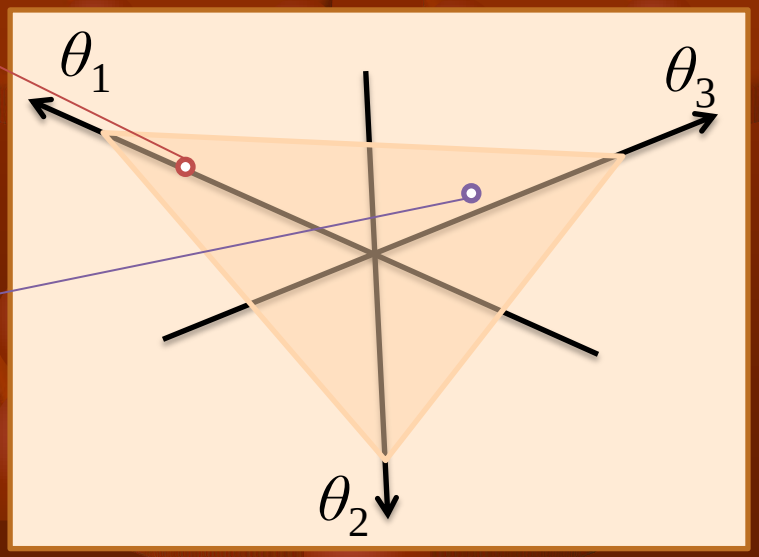
1,
1,
2,
1,
3,
...

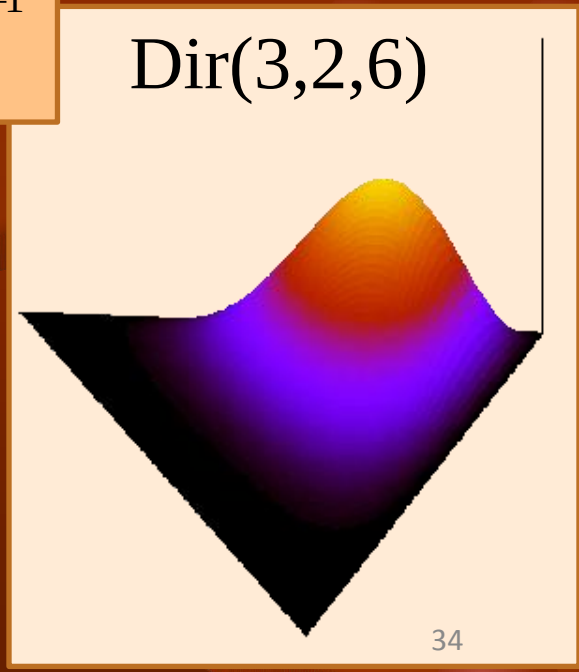


3,
1,
3,
2,
3,
...



$$\mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_N) \propto \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}$$





Dirichlet 分布 (2)

- 多項分布の共役事前分布 (conjugate prior)
  - データ観測後のposteriorも、priorと同じ形式で記述できる

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_N) \propto \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}$$

$$p(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_N, X = j) \propto p(X = j \mid \boldsymbol{\theta}) \cdot \mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_N)$$

$$= \theta_j \cdot \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}$$

$$\propto \mathrm{Dir}(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_N)$$

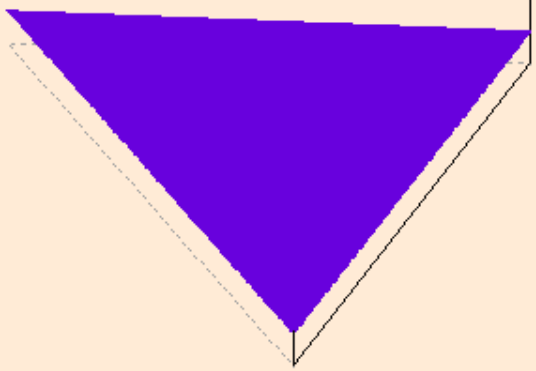
ハイパーパラメータ α_j は、事前に j を観測していた回数に相当

Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

Prior

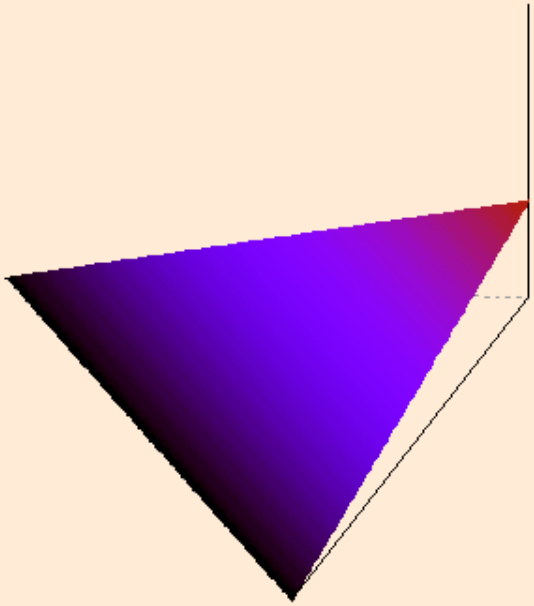
$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,1,1,1)$$



$\times$

Likelihood

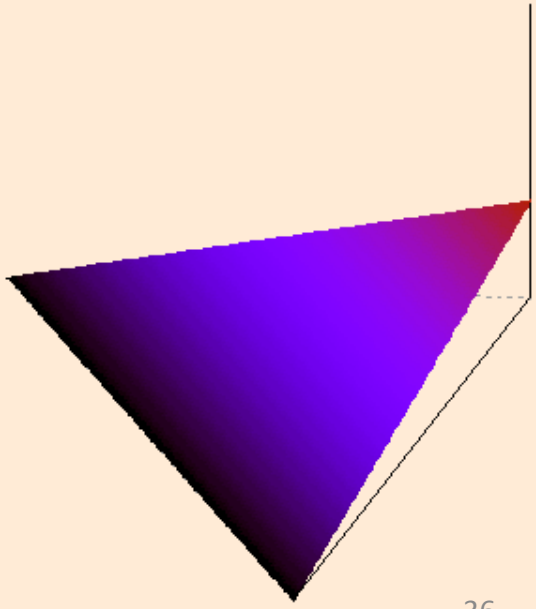
$$p(X=3\,|\,\boldsymbol{\theta})=\theta_3$$



$\propto$

Posterior

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,1,1,2)$$

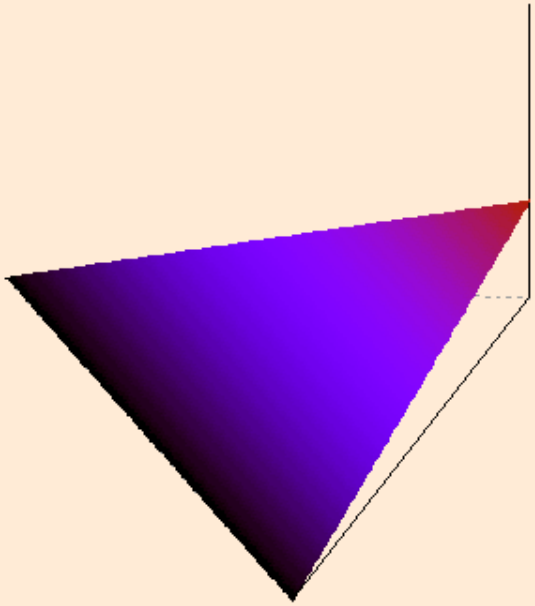


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

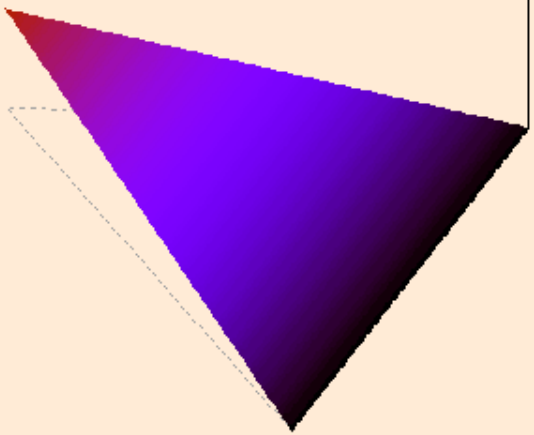
Prior

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,1,1,2)$$



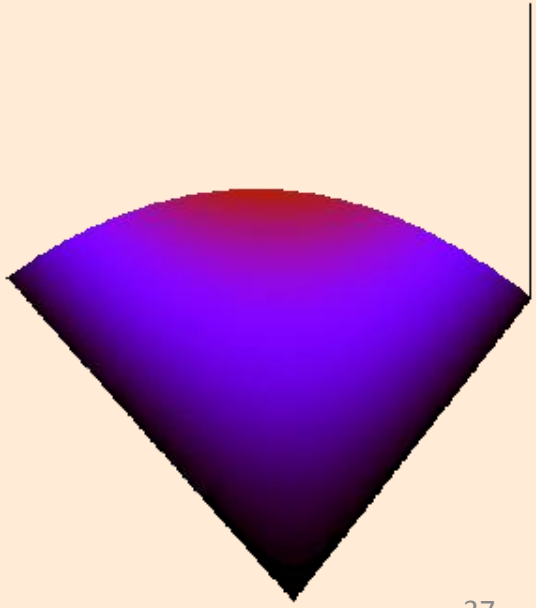
$\times$ Likelihood

$$p(X=1\,|\,\boldsymbol{\theta})=\theta_1$$



$\propto$ Posterior

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,2,1,2)$$

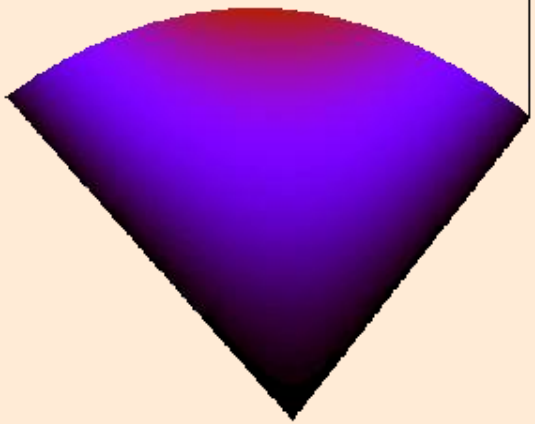


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

Prior

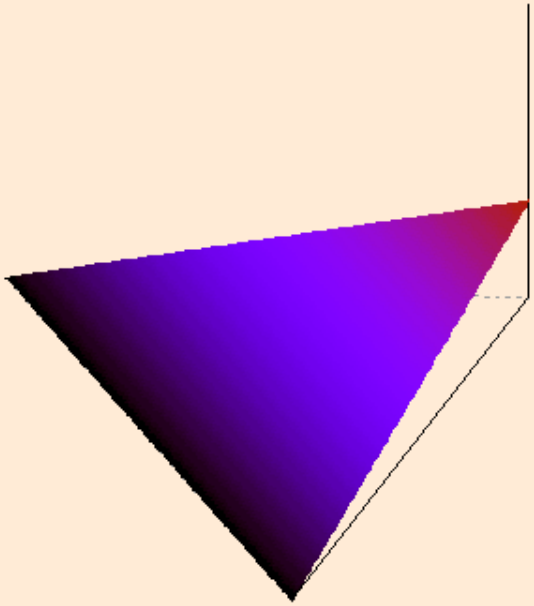
$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 1, 2)$$



×

Likelihood

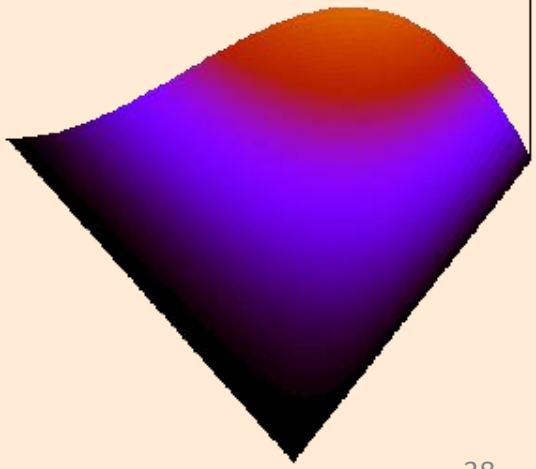
$$p(X = 3 \mid \boldsymbol{\theta}) = \theta_3$$



∝

Posterior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 1, 3)$$

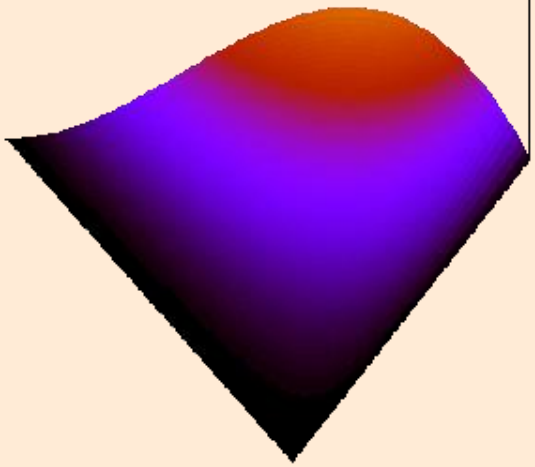


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

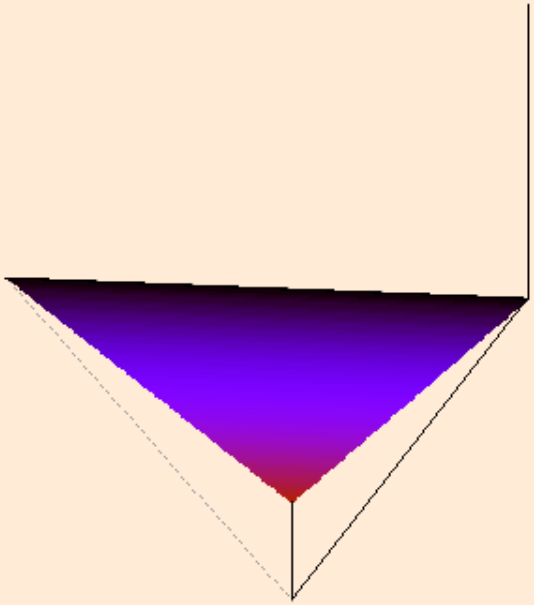
Prior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 1, 3)$$



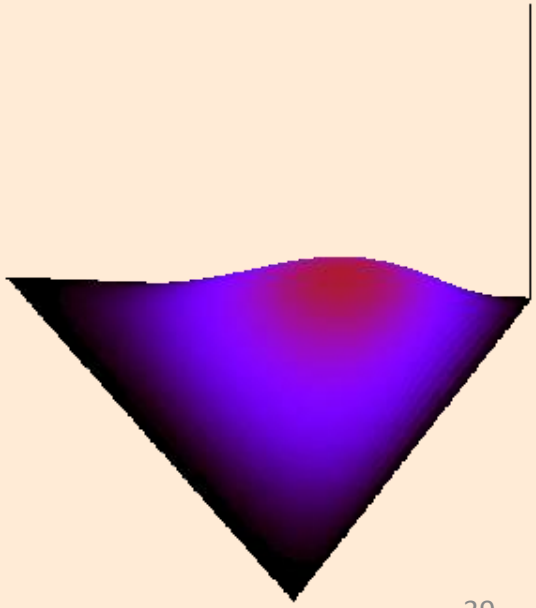
$\times$ Likelihood

$$p(X = 2 \mid \boldsymbol{\theta}) = \theta_2$$



$\propto$ Posterior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 2, 3)$$

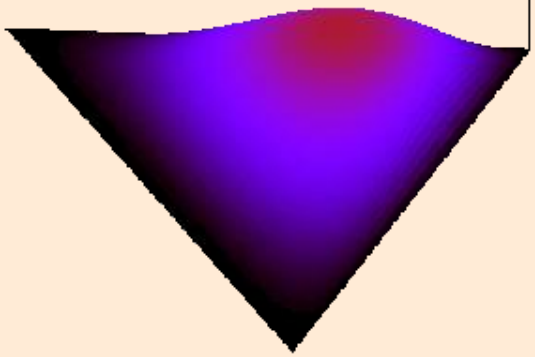


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

Prior

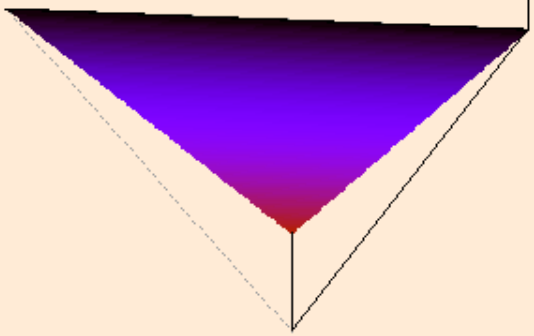
$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 2, 3)$$



$\times$

Likelihood

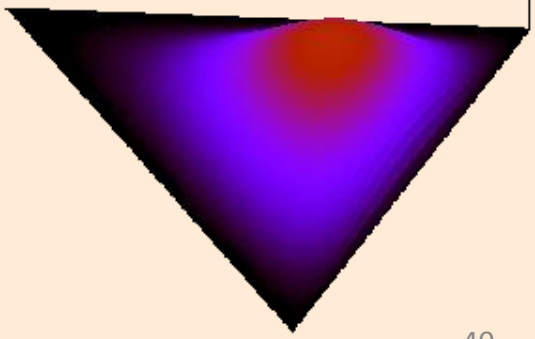
$$p(X = 2 \mid \boldsymbol{\theta}) = \theta_2$$



$\propto$

Posterior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 3, 3)$$

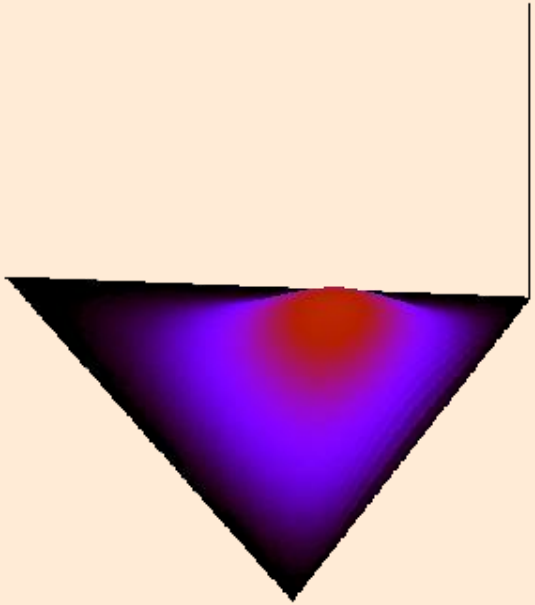


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

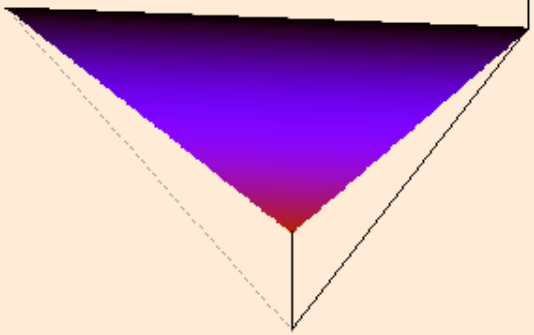
Prior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 3, 3)$$



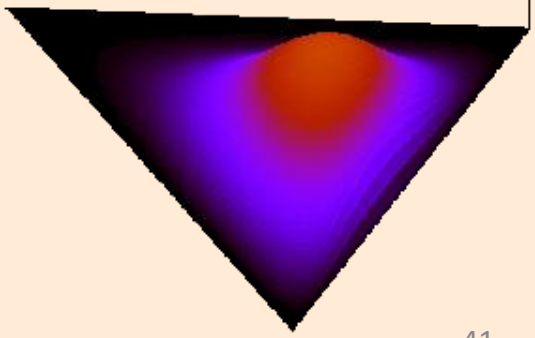
$\times$ Likelihood

$$p(X = 2 \mid \boldsymbol{\theta}) = \theta_2$$



$\propto$ Posterior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 4, 3)$$

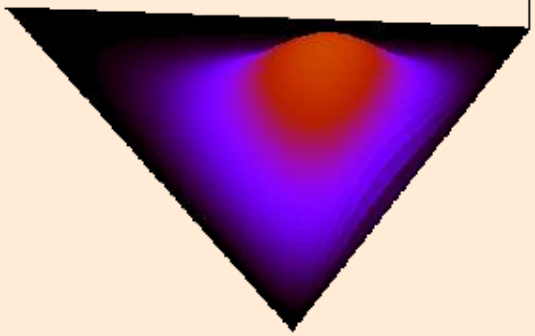


Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

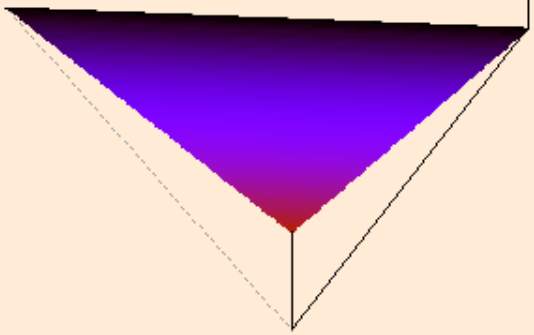
Prior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 4, 3)$$



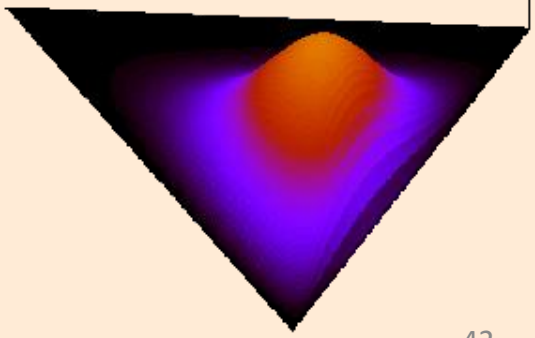
$\times$ Likelihood

$$p(X = 2 \mid \boldsymbol{\theta}) = \theta_2$$



$\propto$ Posterior

$$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 5, 3)$$



Dirichlet 分布 (3)

- Dirichlet 分布の例
  - 観測のたびに、ハイパーパラメータが更新される

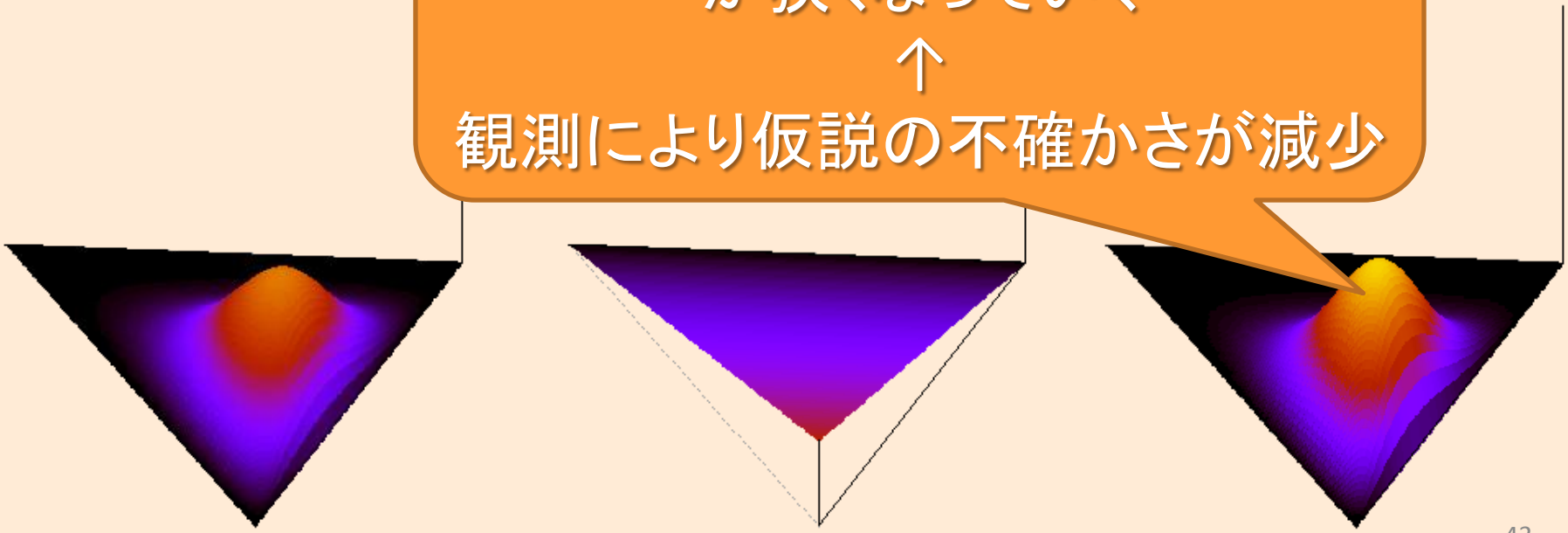
Prior

$\mathrm{Dir}(\boldsymbol{\theta} \mid 2, 5, 3)$

観測数が増えると、事後分布の幅が狭くなっていく

↑

観測により仮説の不確かさが減少



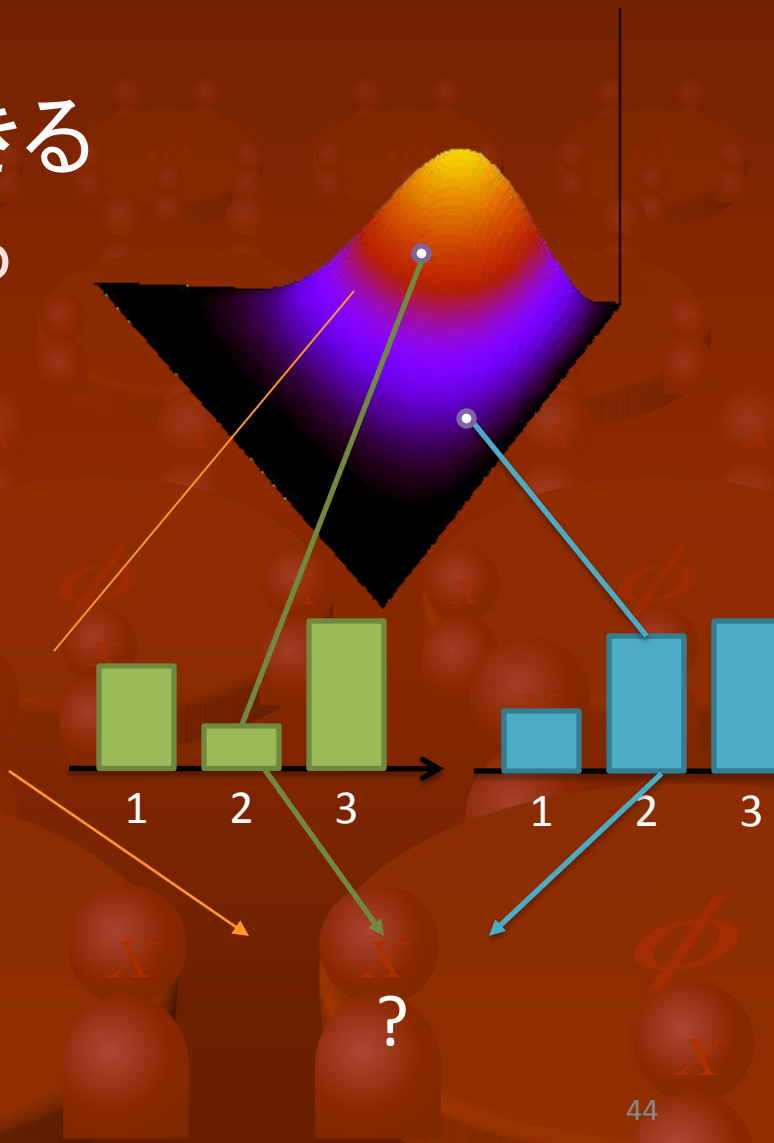
Dirichlet 分布 (4)

- 予測分布も簡単に構成できる
  - パラメータ θ を積分消去する

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,\alpha_1,\ldots,\alpha_N) \propto \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}$$

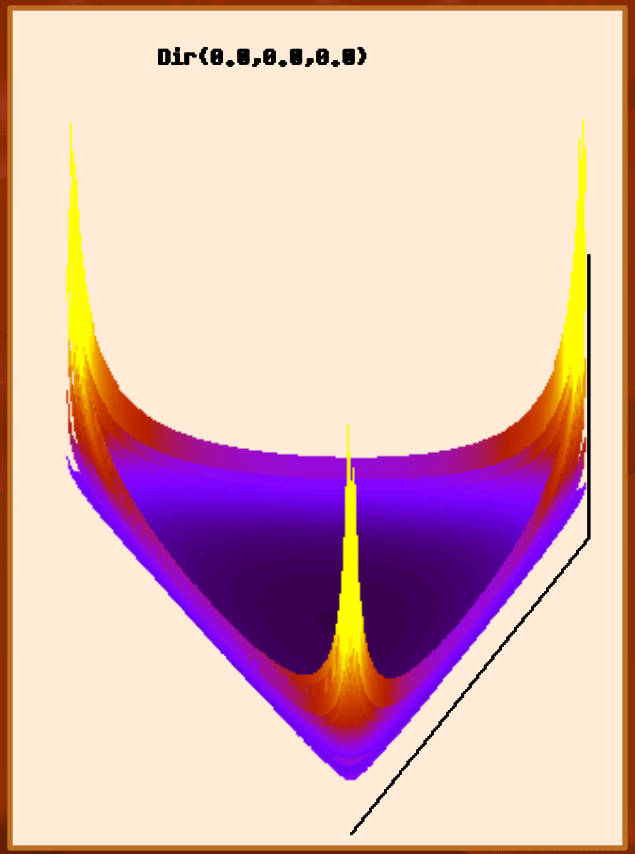
$$p(X = j\,|\,\boldsymbol{\theta}) = \theta_j$$

$$\begin{aligned} p(X = j\,|\,\alpha_1,\ldots,\alpha_N) &= \int p(X = j\,|\,\boldsymbol{\theta})\,\mathrm{Dir}(\boldsymbol{\theta}\,|\,\alpha_1,\ldots,\alpha_N)\,d\boldsymbol{\theta} \\ &= \frac{\alpha_j}{\alpha_1 + \cdots + \alpha_N} \end{aligned}$$



Dirichlet 分布 (5)

- 1未満の α = 疎な分布の事前分布
  - 確率が端に集中…出る確率がほぼ0の値がある



HMM パラメータの分布

- パラメータ: 遷移行列と出力行列
- 行列の各列が多項分布
  → Dirichlet 分布で prior を表現できる
- 状態列 S が与えられれば、posterior もまた Dirichlet 分布で表現できる
  - S 中の各状態遷移の回数分、ハイパーパラメータを更新

Transition Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1i} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2i} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{Ni} & \cdots & a_{NN} \end{pmatrix}$$

$$a_{ji} = p(s_{t+1} = j \mid s_t = i)$$

HMMのベイズ推定

- 出力列 D 観測後の posterior $p(\theta|D)$ が欲しい
  - 状態列 S が与えられれば、$p(\theta|S,D)$ は解析的に解ける (Dirichlet 分布を使う)
  - 通常は S が不明 → $p(\theta|D)$ は解析的に解けない

➔状態列 S を隠れ変数とみなし、
状態列 S とパラメータ θ の同時分布を推定

  - 状態列 S をパラメータ θ から推定: $p(S|\theta,D)$
  - パラメータ θ を状態列 S から推定: $p(\theta|S,D)$

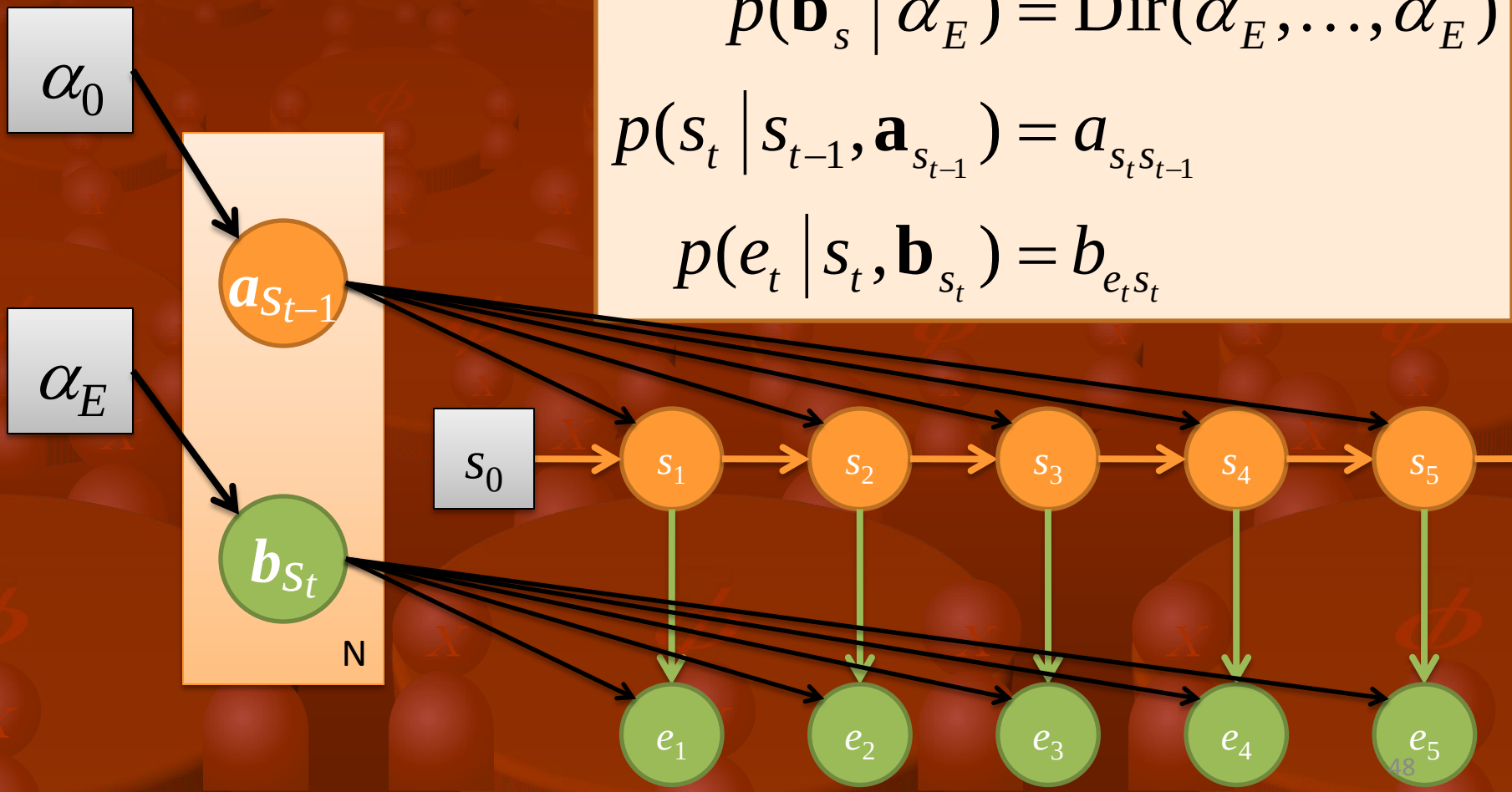
HMM Bayesian Model

$$p(\mathbf{a}_s \mid \alpha_0) = \mathrm{Dir}(\alpha_0, \ldots, \alpha_0)$$

$$p(\mathbf{b}_s \mid \alpha_E) = \mathrm{Dir}(\alpha_E, \ldots, \alpha_E)$$

$$p(s_t \mid s_{t-1}, \mathbf{a}_{s_{t-1}}) = a_{s_t s_{t-1}}$$

$$p(e_t \mid s_t, \mathbf{b}_{s_t}) = b_{e_t s_t}$$

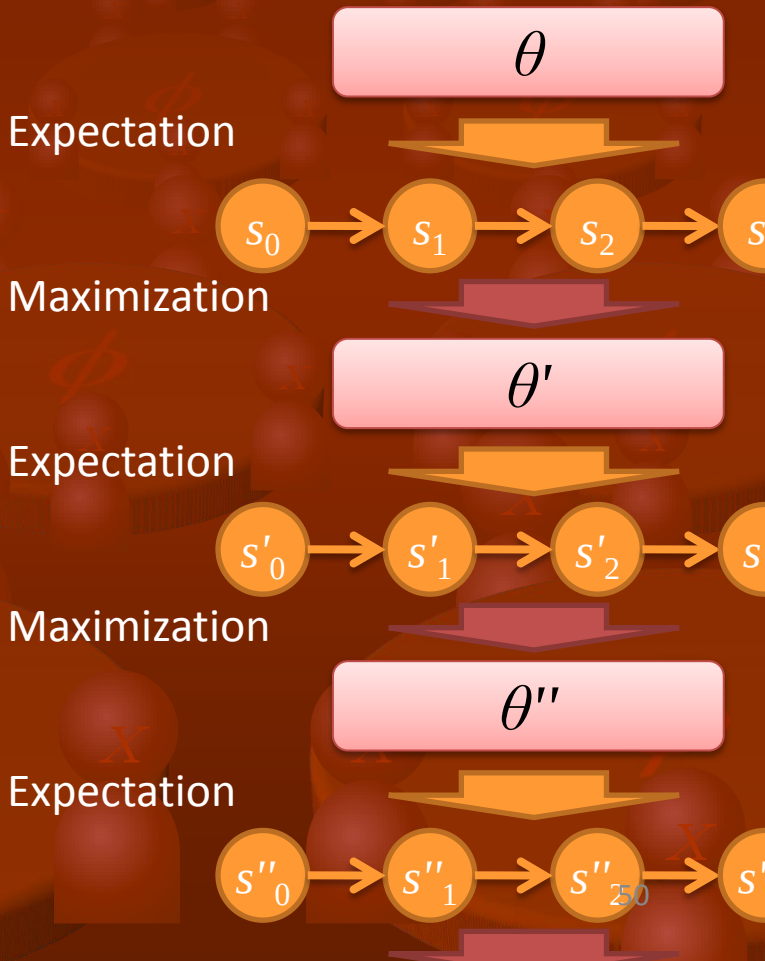


隠れ変数があるモデルのベイズ推定

- 変分ベイズ
  - 分布の形状を、変分法を使って近似
  - 変数間の独立を仮定
  - 熱力学モデルの平均場近似に相当
  - 比較的高速だが、近似誤差が残る
- マルコフ連鎖モンテカルロ (MCMC) 法
  - サンプリングにより分布を推定
  - 多くのサンプルを集めることで、誤差を減らせる
  - 遅い

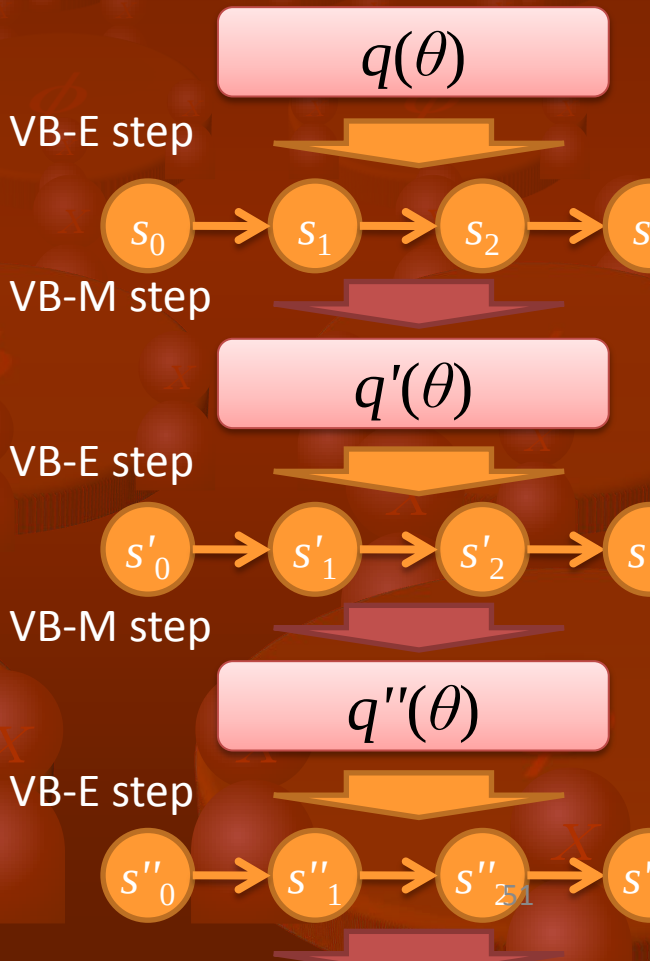
(非ベイズ) EM アルゴリズム

1. 適当なパラメータ θ からスタート
2. 状態列の分布確率を計算
   $q(S) \propto p(S|\theta,D)$
3. Likelihood 最大になる
   パラメータ θ を計算
   $\theta = \mathrm{argmax}_\theta \mathbf{E}_{q(S)}[p(D|\theta,S)]$
4. 収束してなければ2に戻る



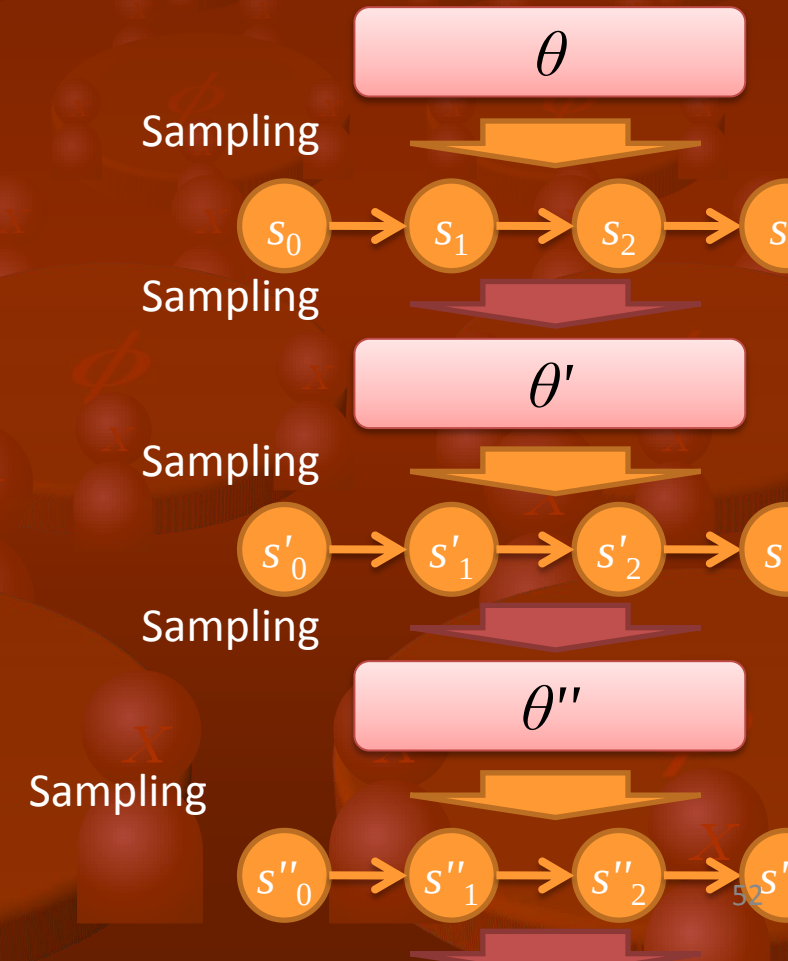
変分ベイズ法

1. 適当なパラメータ分布 $q(\theta)$ からスタート
2. 状態の期待分布を計算
   $q(S) \propto \exp \mathbf{E}_{q(\theta)}[\log p(D,S|\theta)]$
3. パラメータの期待分布を計算
   $q(\theta) \propto p(\theta) \exp \mathbf{E}_{q(S)}[\log p(D,S|\theta)]$
4. 収束してなければ2に戻る



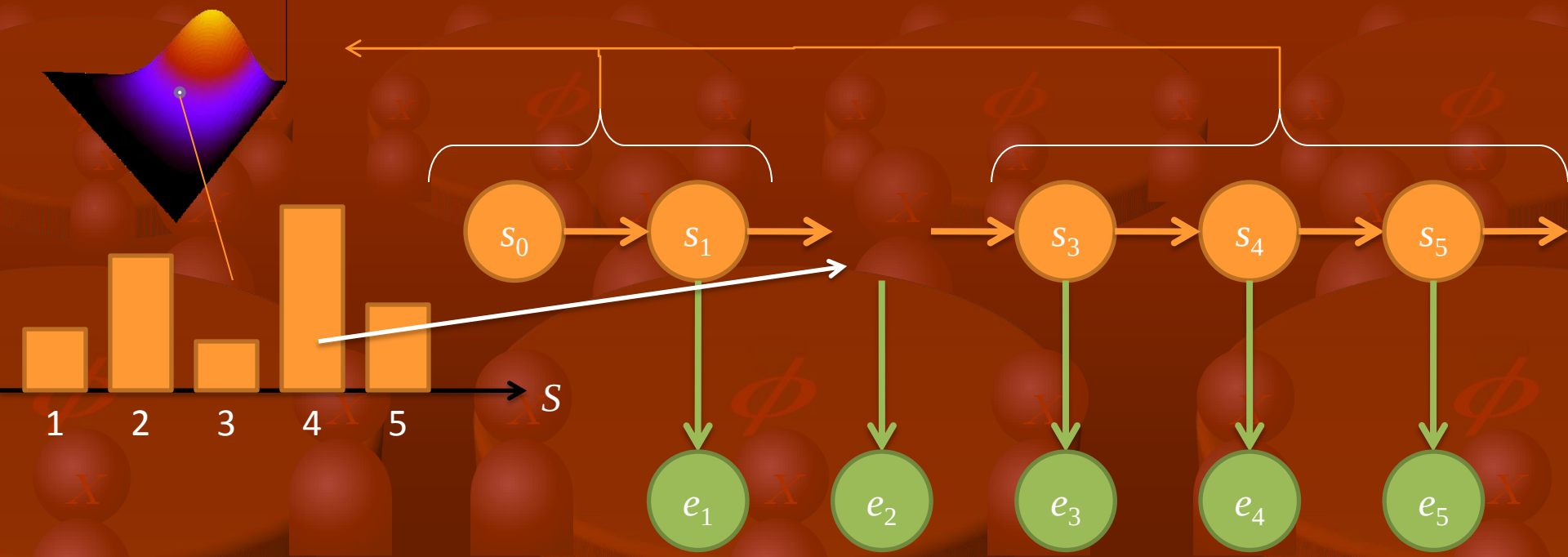
マルコフ連鎖モンテカルロ法 (1)
Blocked Gibbs Sampling

1. 適当なパラメータ θ からスタート
2. 期待分布から状態列 S をサンプリング
   $q(S) \propto p(S,E|\theta)$
3. 期待分布からパラメータ θ をサンプリング
   $q(\theta) \propto p(\theta)\, p(S,E|\theta)$
4. 2に戻る (収束しない)



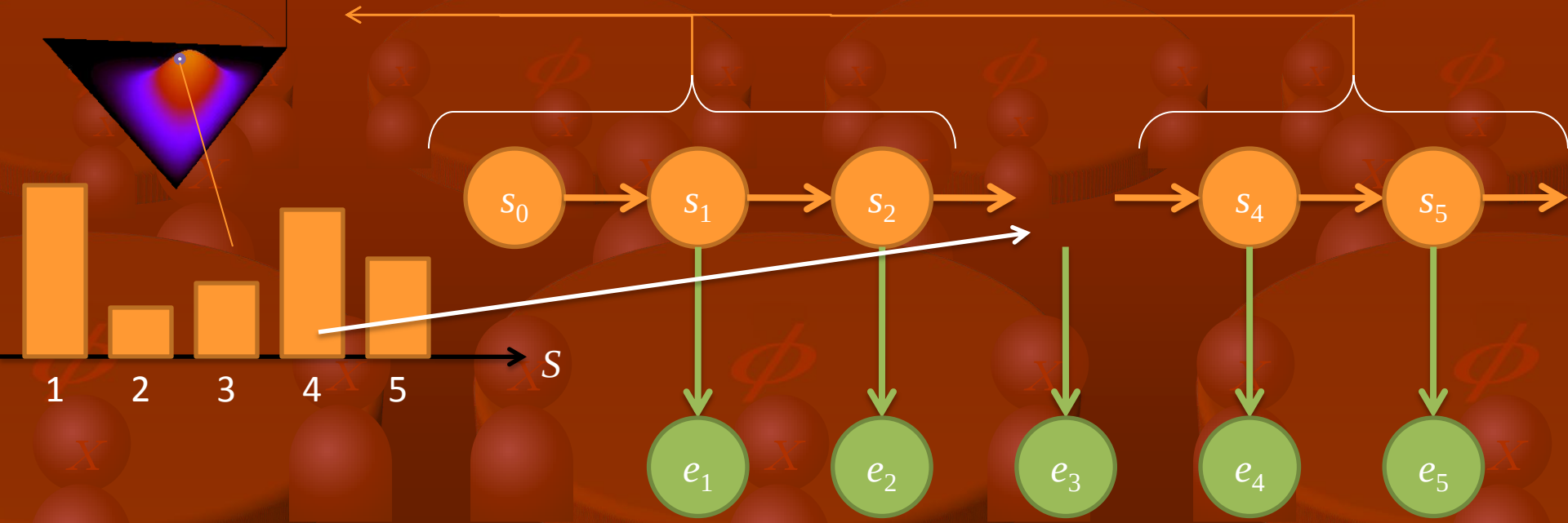
マルコフ連鎖モンテカルロ法 (2)
Collapsed Gibbs Sampling

- 状態列 S を各々の時刻ごとに Gibbs Sampling
- 予測分布を使うことでパラメータ θ を積分消去
  - Dirichlet 分布の exchangeability を利用



マルコフ連鎖モンテカルロ法 (2)
Collapsed Gibbs Sampling

- 状態列 S を各々の時刻ごとに Gibbs Sampling
- 予測分布を使うことでパラメータ θ を積分消去
  - Dirichlet 分布の exchangeability を利用



Talk Outline

- Introduction
  - ベイジアンとは何か?
  - 機械学習とベイズ
- ベイズHMMモデル
  - Dirichlet 分布
  - Bayesian HMM
- ノンパラメトリックベイズHMMモデル
  - Dirichlet process
  - Hierarchical Dirichlet process
  - Infinite Hidden Markov Model
- HMMの状態の階層的クラスタリング

ノンパラメトリックベイズ

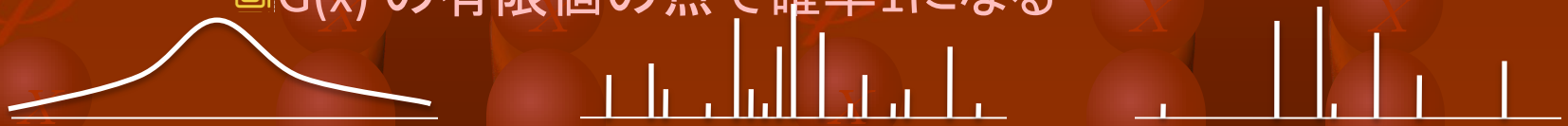
- 複数のパラメータ空間に対する prior を導入
  - HMM であれば、それぞれの N (隠れ状態数) に対して prior 確率 p(N) を割り振る
  - N が異なる複数のモデルの予測を統合できる (Nを積分消去)
- 適切な prior を与えれば、Nの異なるモデルを無限個同時に考えることができる
  → ノンパラメトリックベイズモデル
  - “ノンパラメトリック” とは、パラメータ空間が固定されていない、とか、データに応じて変わる、という意味

Dirichlet Process

- $G \sim DP(\alpha_0, G_0)$
- Dirichlet 分布を N → ∞ に拡張したもの

$$\mathrm{Dir}(\alpha_1, \ldots, \alpha_N) \qquad \alpha_0 = \sum_i \alpha_i \quad G_0(i) = \frac{\alpha_i}{\alpha_0}$$

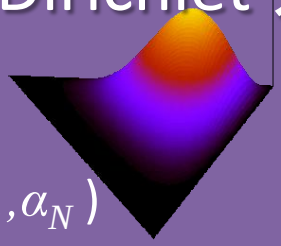
- G_0: 基底測度 (base measure)
  - G_0 が連続であれば、DP は無限次元の離散分布
- α_0: ハイパーパラメータ
  - $\alpha_0 \to \infty$ の極限では、$G \to G_0$
  - α_0 が小さくなると、より疎な分布になる
    - G(x) の有限個の点で確率1になる



G_0　　α_0 Large　　α_0 Small

ハイパーパラメータ
$(\alpha_1, \ldots, \alpha_N)$

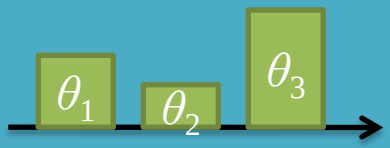
N項の Dirichlet 分布



$\mathrm{Dir}(\alpha_1, \ldots, \alpha_N)$

パラメータ
$(\theta_1, \ldots, \theta_N)$

N項の多項分布



N値の離散確率変数
1, 3, 1, 2, 3, 3, 3, …

ハイパーパラメータ α_0

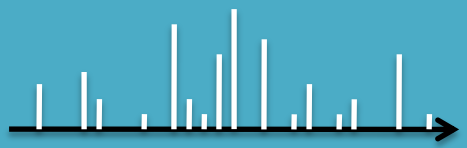
基底測度 G_0

Dirichlet Process

$\mathrm{DP}(\alpha_0, G_0)$

無限個のパラメータ
$G = (\theta_1, \theta_2, \ldots,)$

無限項の多項分布



無限値の離散確率変数
-0.24, 0.93, 0.07, -0.24, …

基底測度 G_0

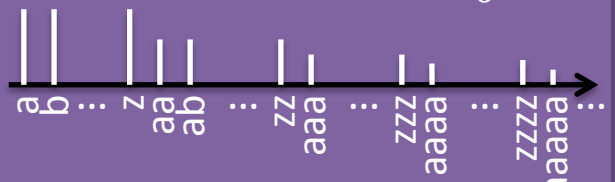
実数空間上の分布

$\mathrm{DP}(\alpha_0, G_0)$

分布 G

実数の離散分布

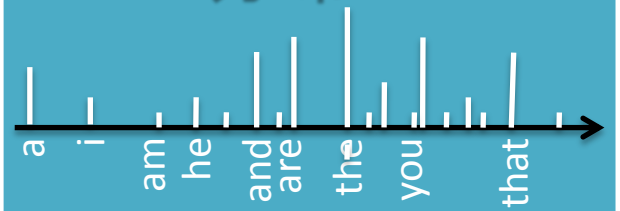
基底測度 G_0



アルファベット列の分布

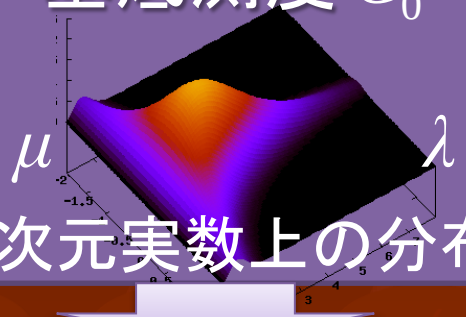
$\mathrm{DP}(\alpha_0, G_0)$

分布 G



アルファベット列の離散分布

基底測度 G_0



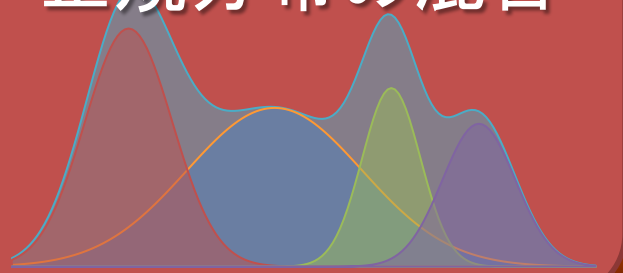
2次元実数上の分布

$\mathrm{DP}(\alpha_0, G_0)$

分布 G

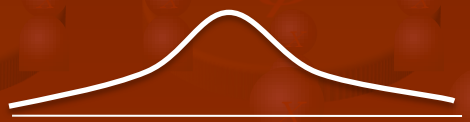
実数の組の離散分布
→正規分布の分布

正規分布の混合

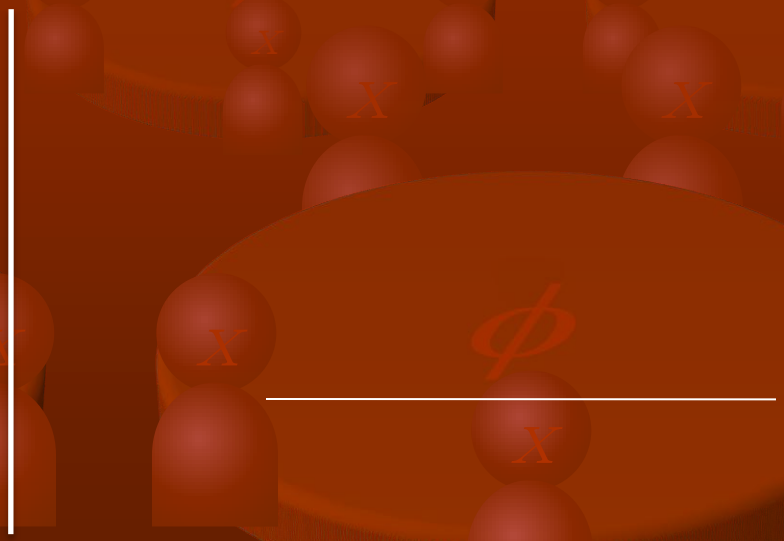


Stick-breaking construction

- 分布 G を明示的にサンプリングする方法
  (Sethuraman, 1994; Ishwaran & James , 2001)
  - 長さ1の棒を $\nu_i \sim \mathrm{Beta}(1, \alpha_0)$ の比率で折りながら、G_0 から選んだ点 θ_i に確率を振る

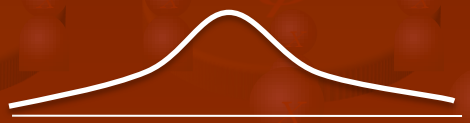


G_0: Prior Distribution of English words

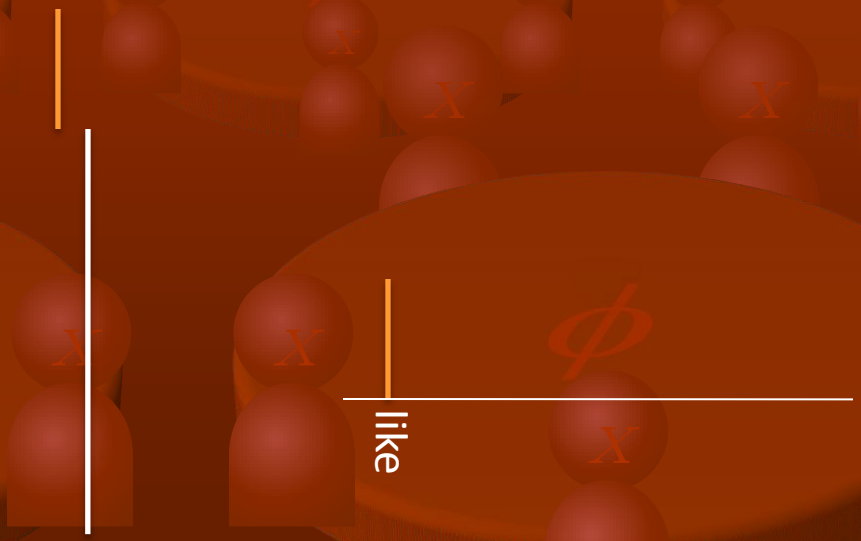


Stick-breaking construction

- 分布 G を明示的にサンプリングする方法
  (Sethuraman, 1994; Ishwaran & James , 2001)
  - 長さ1の棒を $\nu_i \sim \mathrm{Beta}(1, \alpha_0)$ の比率で折りながら、G_0 から選んだ点 θ_i に確率を振る

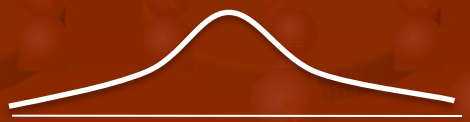


G_0: Prior Distribution of English words

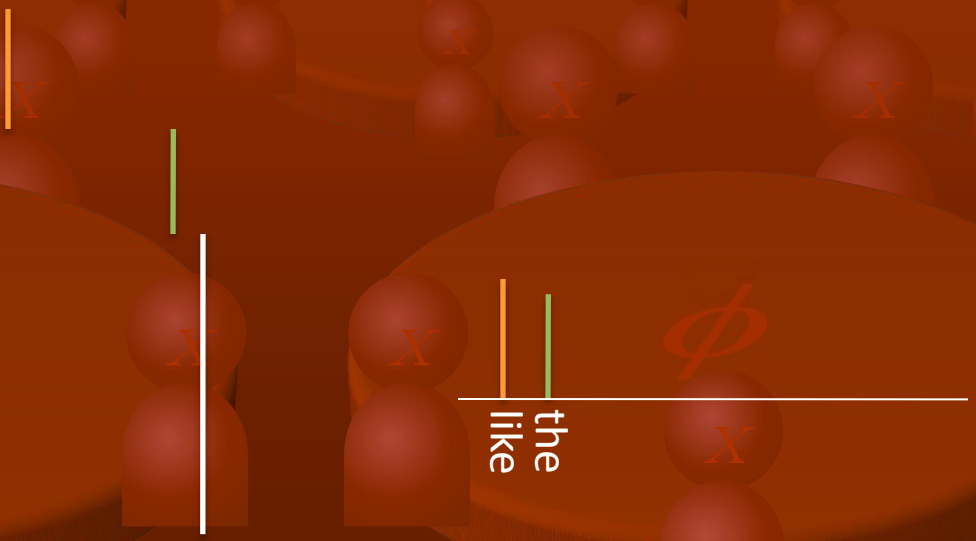


Stick-breaking construction

- 分布 G を明示的にサンプリングする方法
  (Sethuraman, 1994; Ishwaran & James , 2001)
  - 長さ1の棒を $\nu_i \sim \mathrm{Beta}(1, \alpha_0)$ の比率で折りながら、G_0 から選んだ点 θ_i に確率を振る

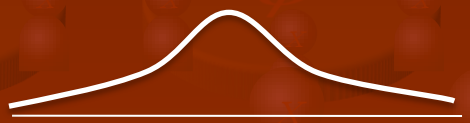


G_0: Prior Distribution of English words

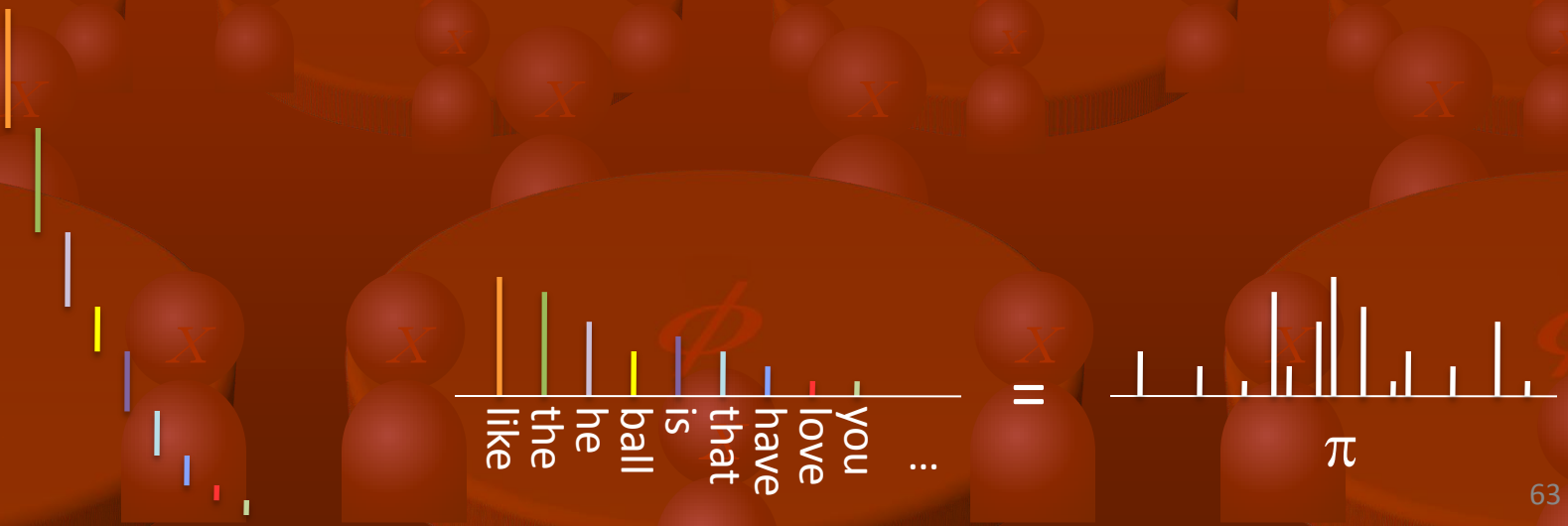


Stick-breaking construction

- 分布 G を明示的にサンプリングする方法
  (Sethuraman, 1994; Ishwaran & James , 2001)
  - 長さ1の棒を $\nu_i \sim \mathrm{Beta}(1, \alpha_0)$ の比率で折りながら、G_0 から選んだ点 θ_i に確率を振る

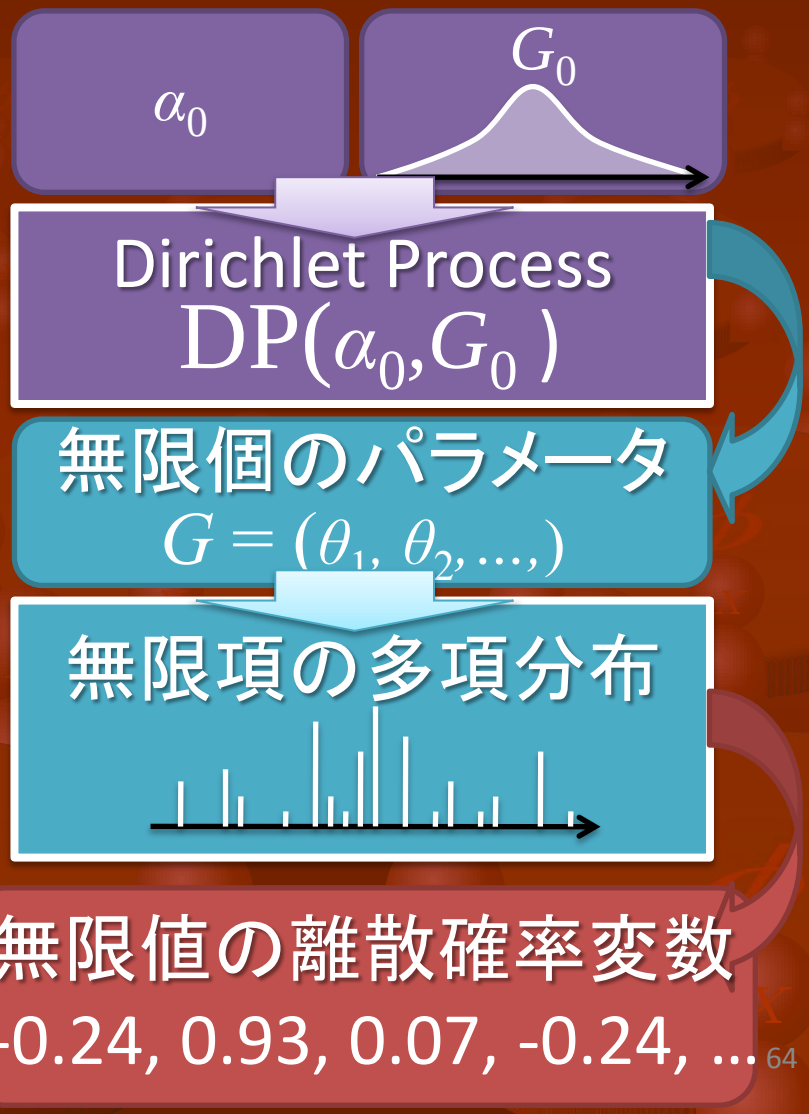


G_0: Prior Distribution of English words



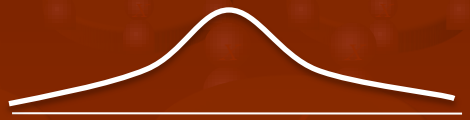
Chinese Restaurant Process

- Dirichlet Process のよい性質を活用したいが、無限個のパラメータは扱いづらい
→ パラメータを積分消去して、確率変数の予測分布を直接計算する

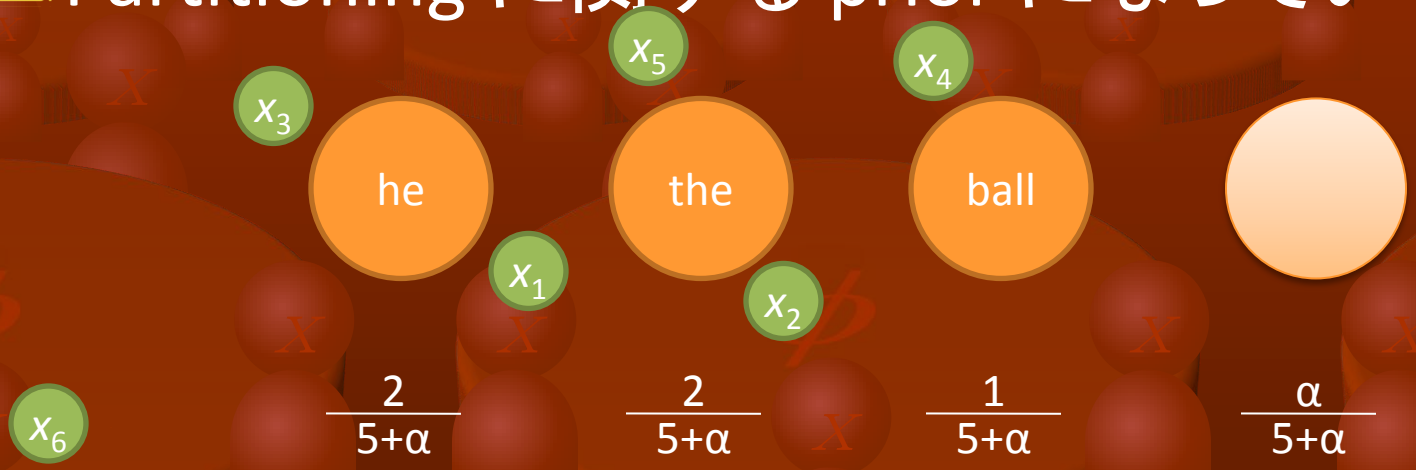


Chinese Restaurant Process

- G を計算せず、多項分布の予測分布を求める
  (Aldous, 1983; Pitman, 1995)
  - 客 = 確率変数
  - 皿 = 確率変数に割り当てられた値
  - 卓 = 客と皿の組み合わせかた
- Partitioning に関する prior になっている

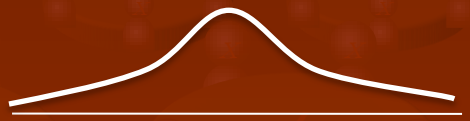


G_0: 英単語の分布

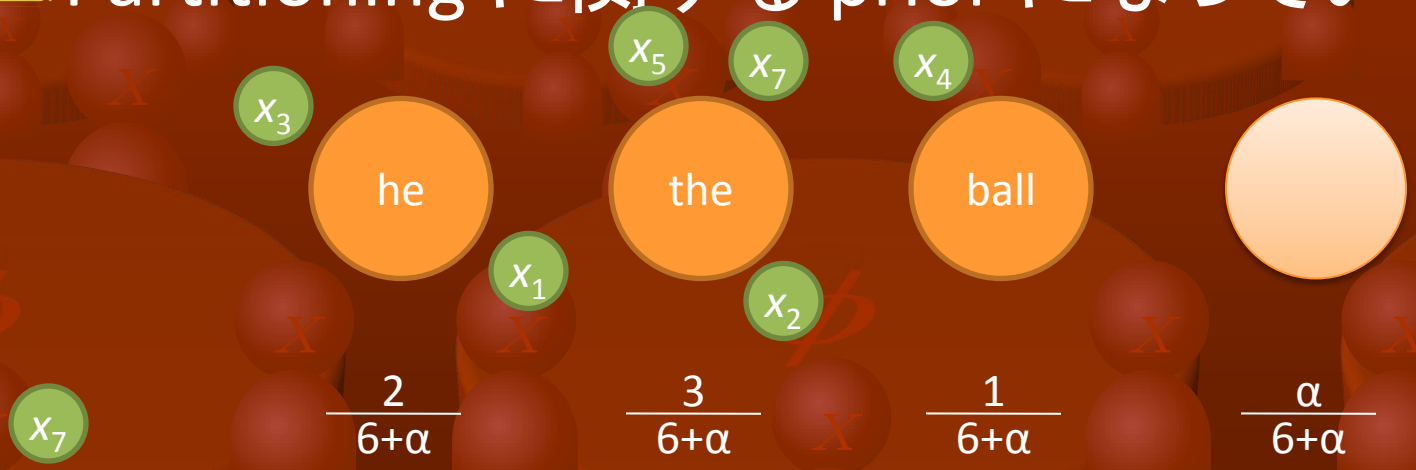


Chinese Restaurant Process

- G を計算せず、多項分布の予測分布を求める
  (Aldous, 1983; Pitman, 1995)
  - 客 = 確率変数
  - 皿 = 確率変数に割り当てられた値
  - 卓 = 客と皿の組み合わせかた
- Partitioning に関する prior になっている

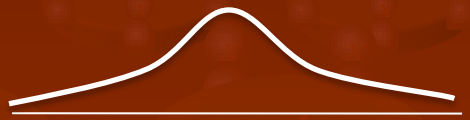


G_0: 英単語の分布

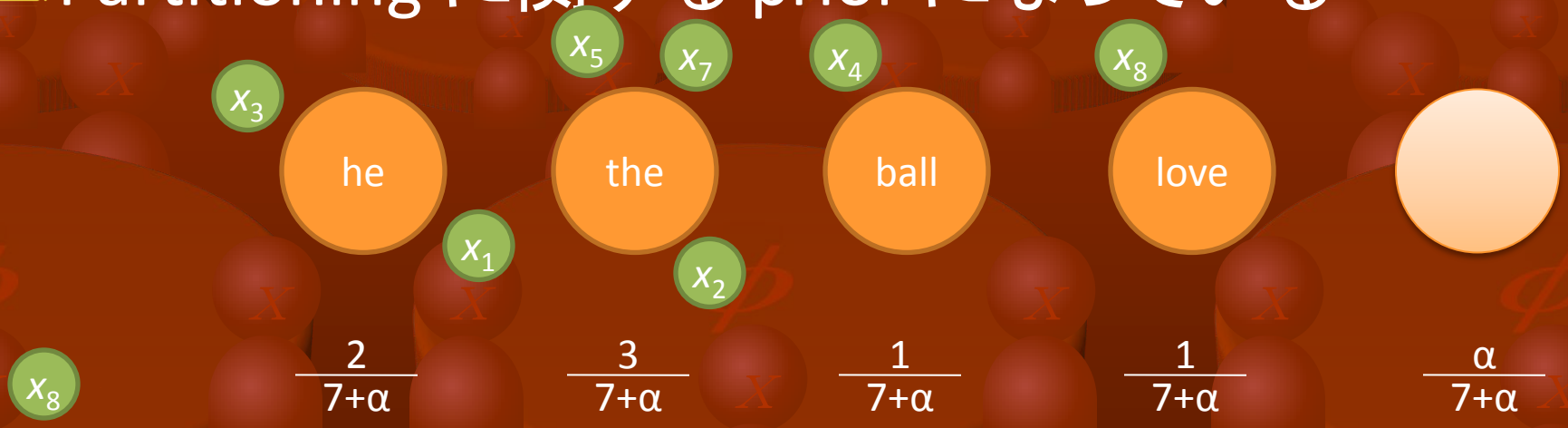


Chinese Restaurant Process

- G を計算せず、多項分布の予測分布を求める
  (Aldous, 1983; Pitman, 1995)
  - 客 = 確率変数
  - 皿 = 確率変数に割り当てられた値
  - 卓 = 客と皿の組み合わせかた
- Partitioning に関する prior になっている

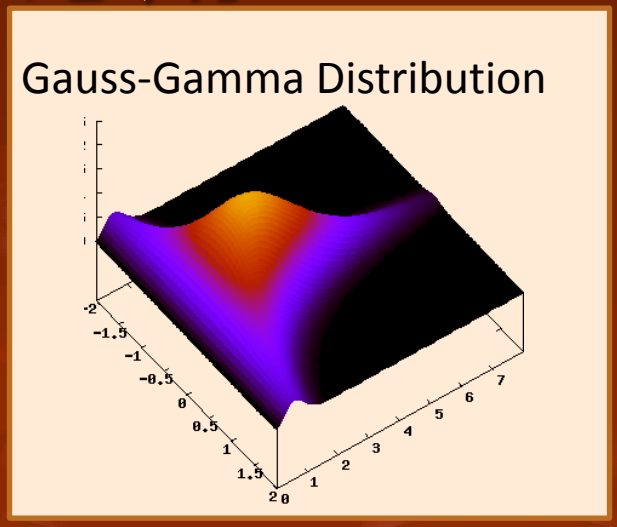


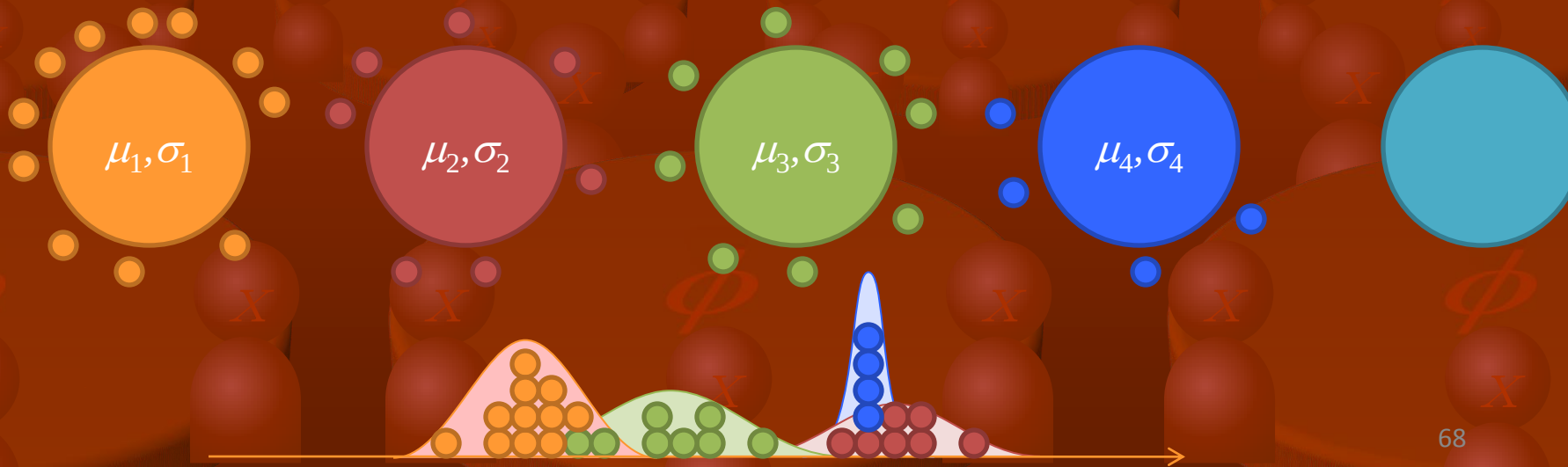
G_0: 英単語の分布



Dirichlet Process の応用

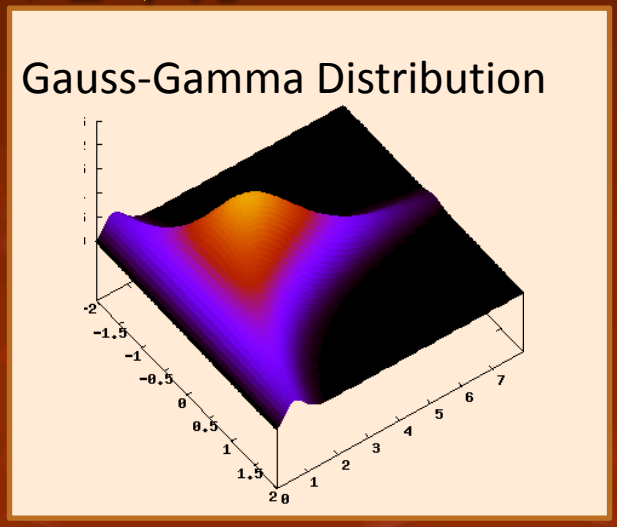
- 混合正規分布
  - それぞれの皿がひとつの正規分布 (クラスタ) を表わす
  - 予測分布を使うことで、各データが所属するクラスタを再計算 = collapsed Gibbs sampling



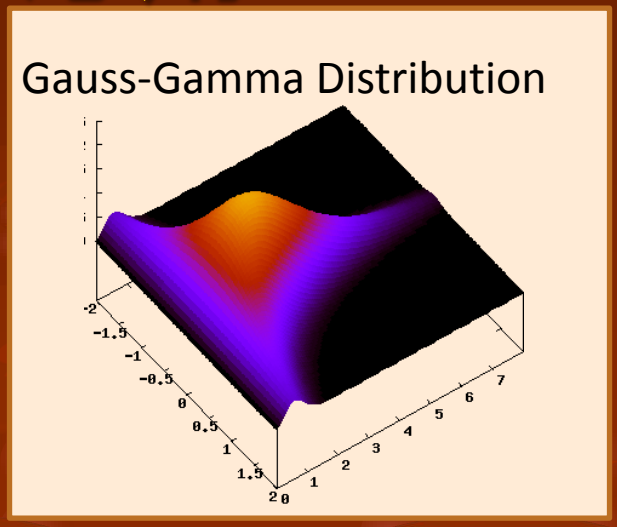


Dirichlet Process の応用

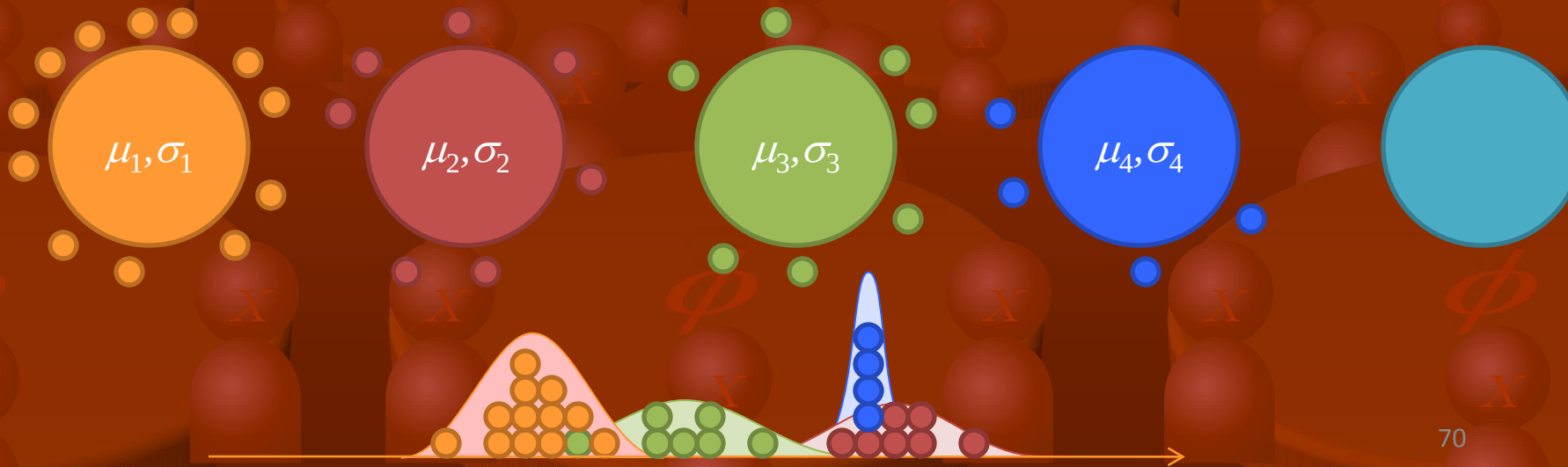
- 混合正規分布
  - それぞれの皿がひとつの正規分布 (クラスタ) を表わす
  - 予測分布を使うことで、各データが所属するクラスタを再計算 = collapsed Gibbs sampling



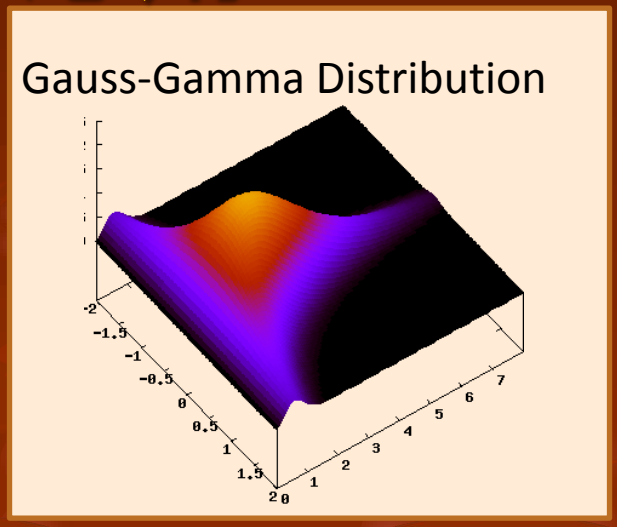
Dirichlet Process の応用



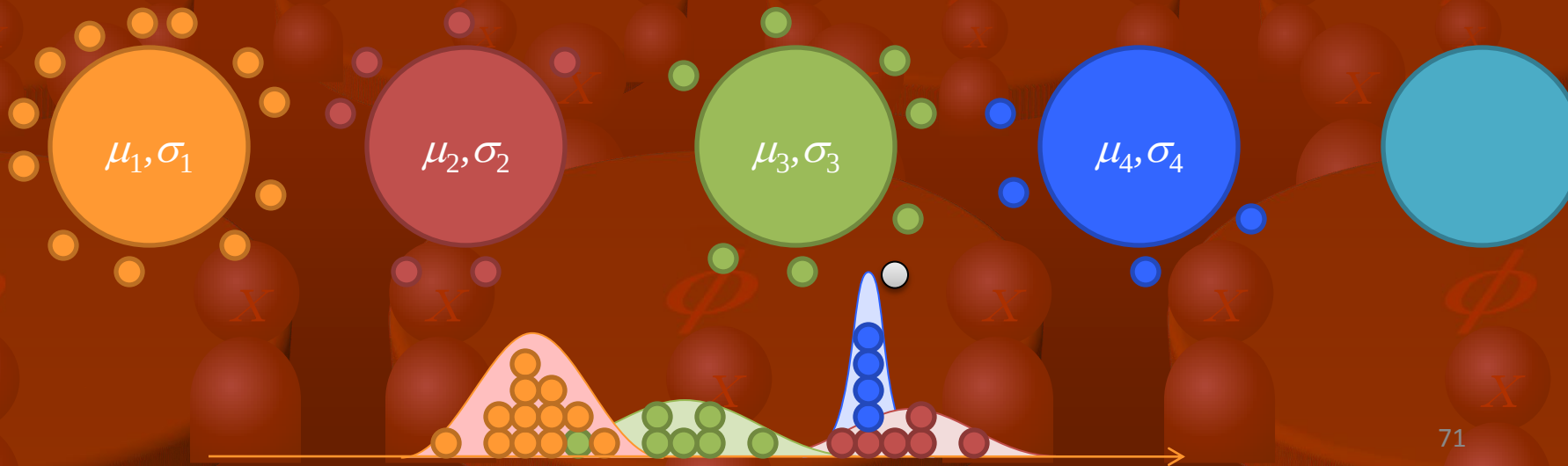
- 混合正規分布
  - それぞれの皿がひとつの正規分布 (クラスタ) を表わす
  - 予測分布を使うことで、各データが所属するクラスタを再計算 = collapsed Gibbs sampling



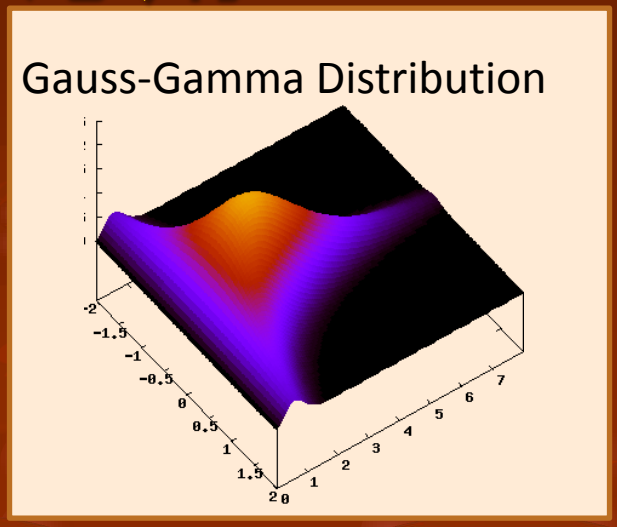
Dirichlet Process の応用



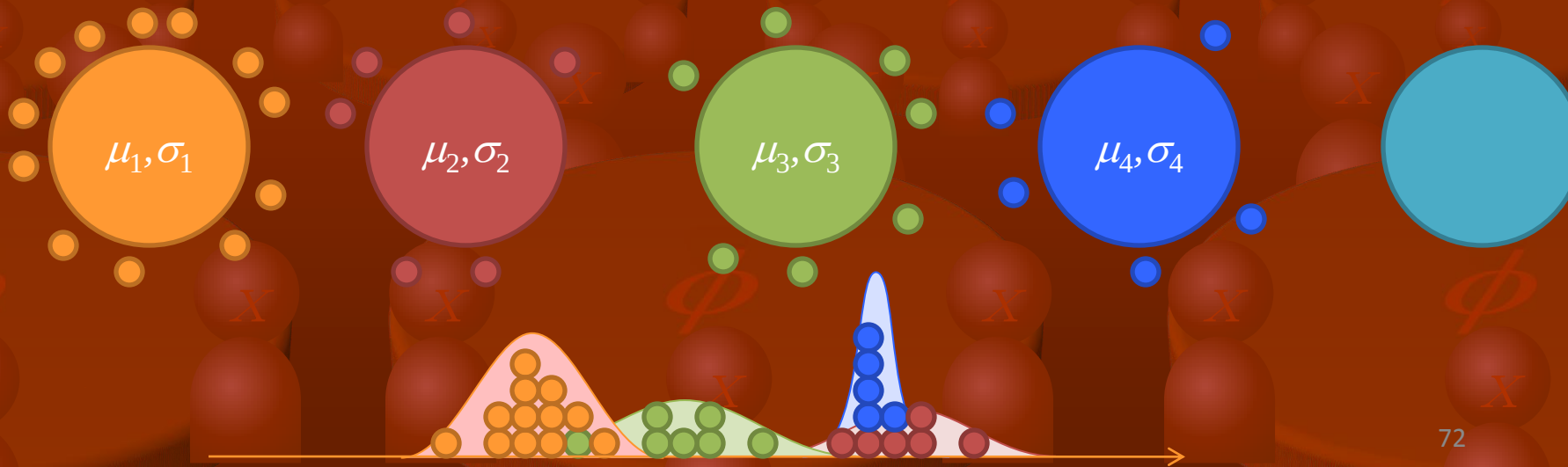
- 混合正規分布
  - それぞれの皿がひとつの正規分布 (クラスタ) を表わす
  - 予測分布を使うことで、各データが所属するクラスタを再計算 = collapsed Gibbs sampling



Dirichlet Process の応用

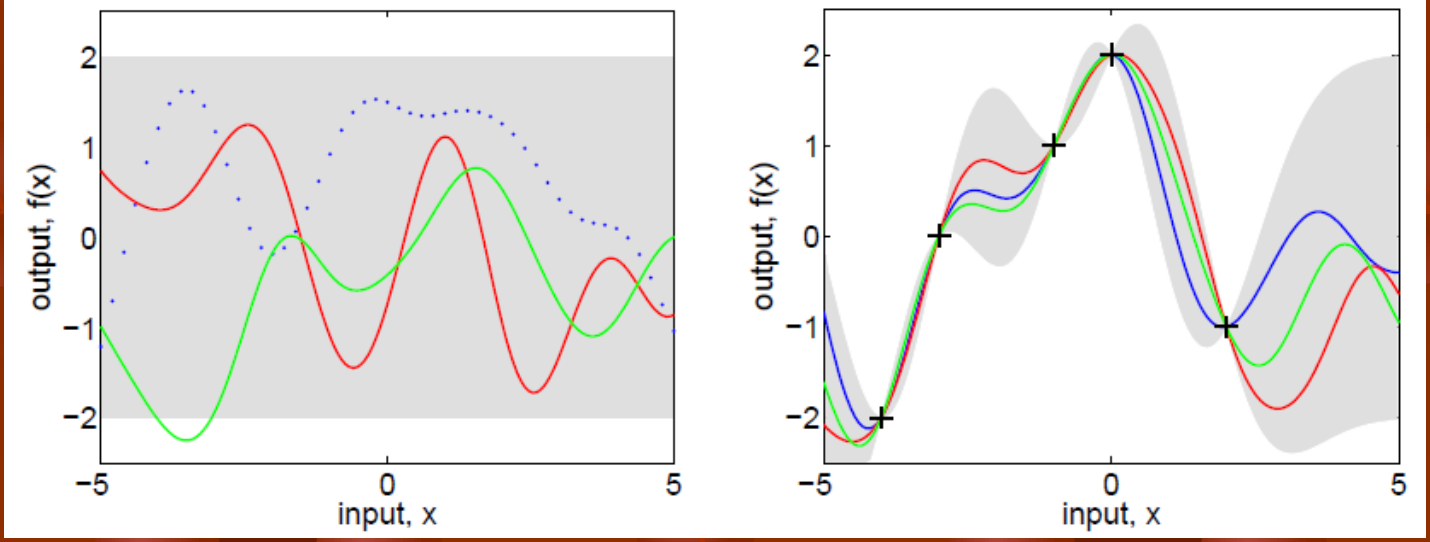


- 混合正規分布
  - それぞれの皿がひとつの正規分布 (クラスタ) を表わす
  - 予測分布を使うことで、各データが所属するクラスタを再計算 = collapsed Gibbs sampling



そのほかの主な ノンパラメトリックベイズモデル (1)

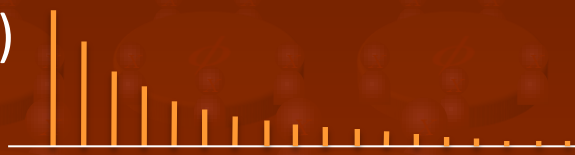
- Gaussian Process (O'Hagen, 1978)
  - 多変数ガウス分布の無限変数版
  - 関数回帰のノンパラメトリックモデル
  - 双対表現、カーネルトリックを利用



(Figure from Rasmussen&Williams, 2005)

そのほかの主な ノンパラメトリックベイズモデル (2)

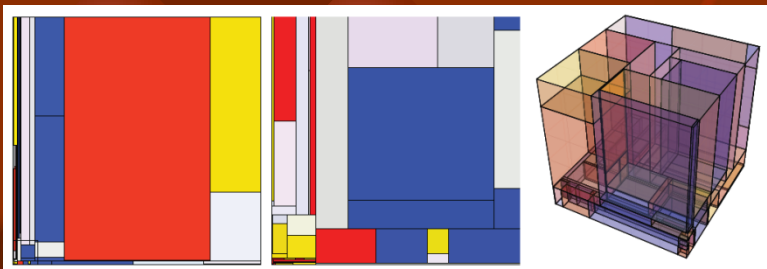
- Pitman-Yor Process (Pitman & Yor, 1997)
  - Dirichlet Process の拡張
  - k 番目のStick を折る比率が $\mathrm{Beta}(1-d, \alpha_0+kd)$
  - Power-law 分布の prior =「ロングテール」を表現



- Beta Process, Indian Buffet Process (Griffiths&Ghahramani, 2005; Thibaux&Jordan, 2007)
  - データの背後にある隠れ変数が単一のカテゴリではなく、feature の組み合わせ

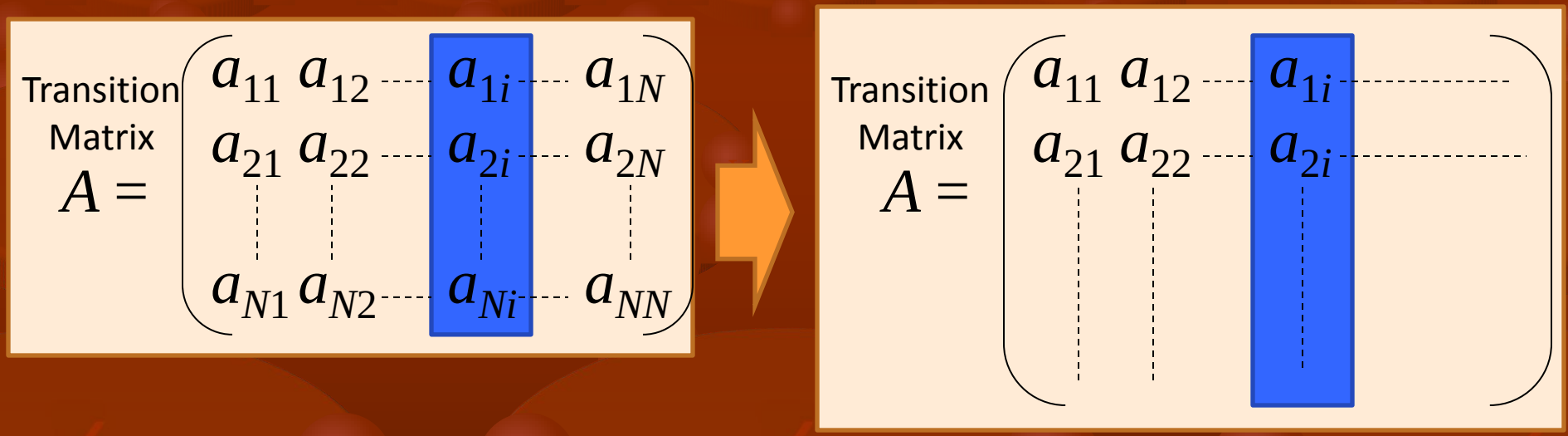
(Figure from Ghahramani, 2005 / Roy&Teh, 2008)

- Mondorian Process (Roy&Teh, 2008)
  - k次元treeに対する prior



Nonparametric Bayesian HMM (1)

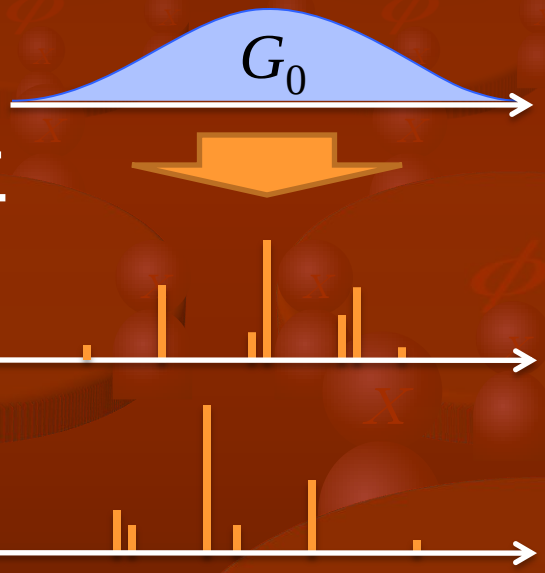
- 直観的には、遷移行節の各列のpriorを Dirichlet 分布から Dirichlet Process に置き換えればよさそうに思える



- しかし、これはうまくいかない

Nonparametric Bayesian HMM (2)

- Dirichlet Process からサンプルした多項分布においては、有限個の点だけが非0確率を持つ
- → 列iと列jは確率1でdisjoint
  → 遷移先は常に未知の状態になってしまう



Transition Matrix

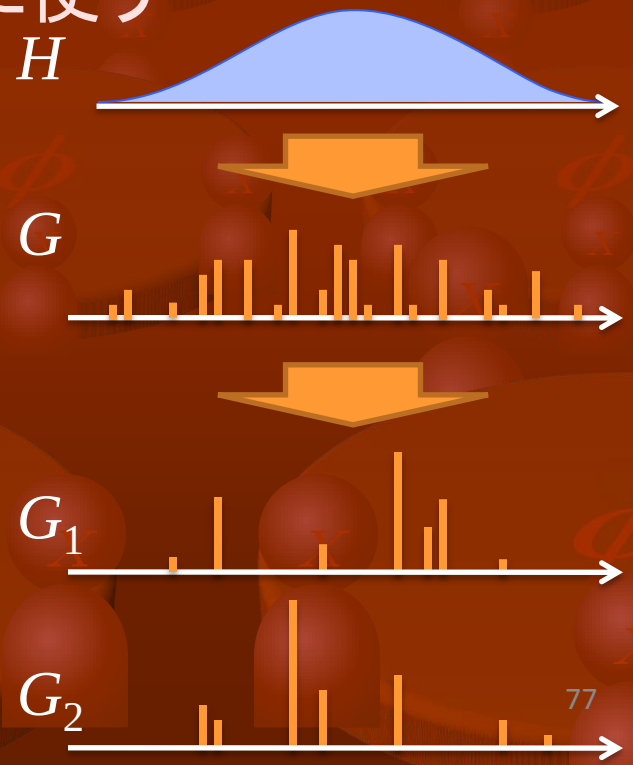
$$A = \begin{pmatrix} \cdots & 0 & 0 & \cdots \\ & 0 & a_{j'i'} & \\ & \vdots & \vdots & \\ & a_{ji} & 0 & \\ & \vdots & \vdots & \end{pmatrix}$$

Nonparametric Bayesian HMM (3)

- Hierarchical Dirichlet Process を使う
  (Beal+, 2002; Teh+, 2006)
  - Dirichlet Process からサンプルした分布を、別の Dirichlet Process の基底測度に使う

$$G \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_i \mid \alpha_0, G \sim \mathrm{DP}(\alpha_0, G)$$



  - G が離散なので、多数の分布がサポートを共有するようにできる

Infinite Hidden Markov Models

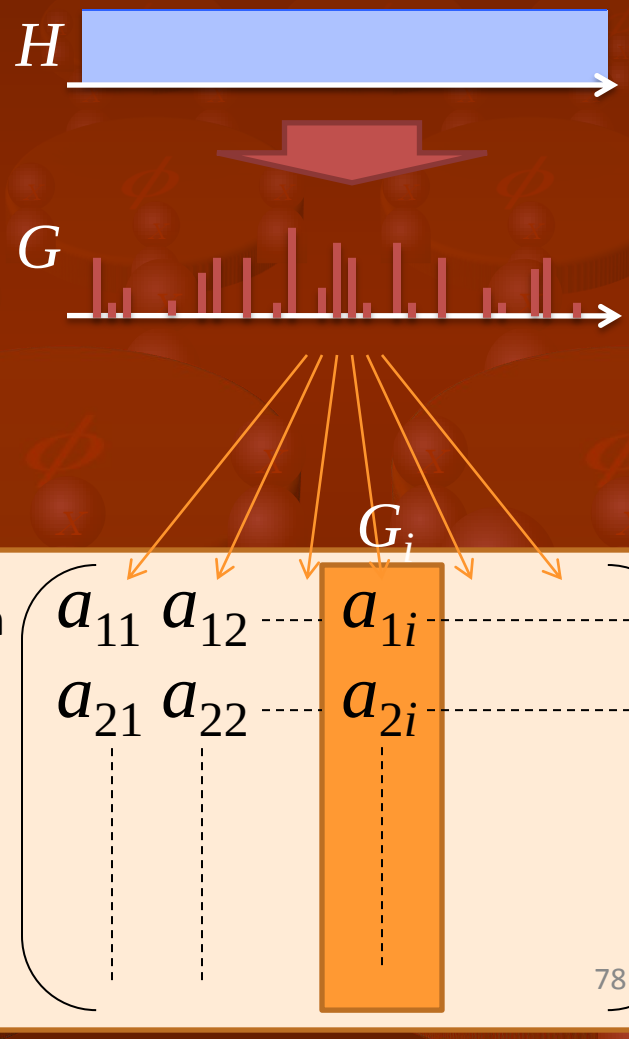
HMM で使うことのできるすべての状態に対する (一様) 分布

この HMM で利用される状態の確率分布

$$G \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_i \mid \alpha_0, G \sim \mathrm{DP}(\alpha_0, G)$$

遷移行列の i 列目
=状態 i からの遷移先の状態の確率分布



Infinite Hidden Markov Models

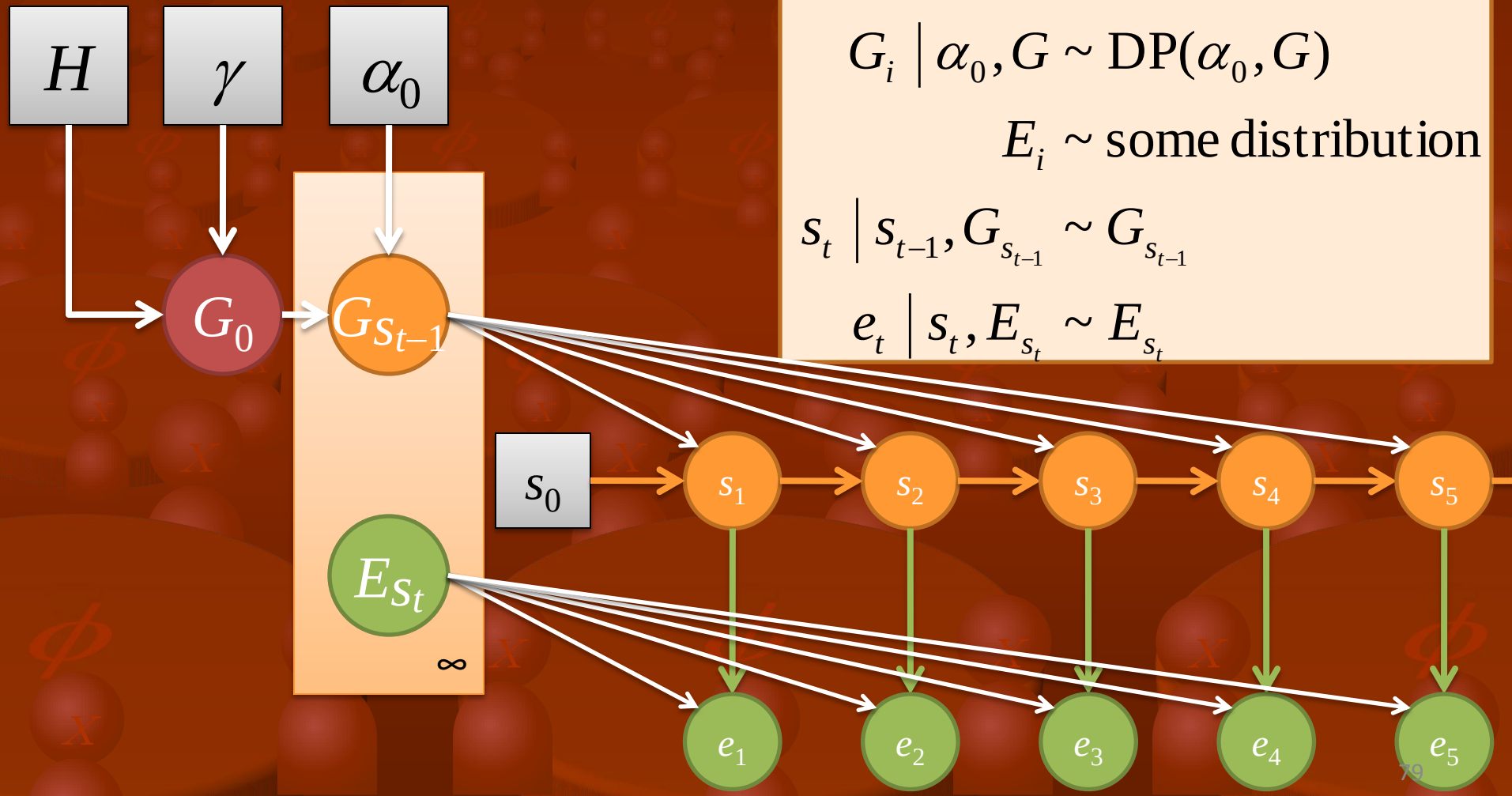
$$G \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_i \mid \alpha_0, G \sim \mathrm{DP}(\alpha_0, G)$$

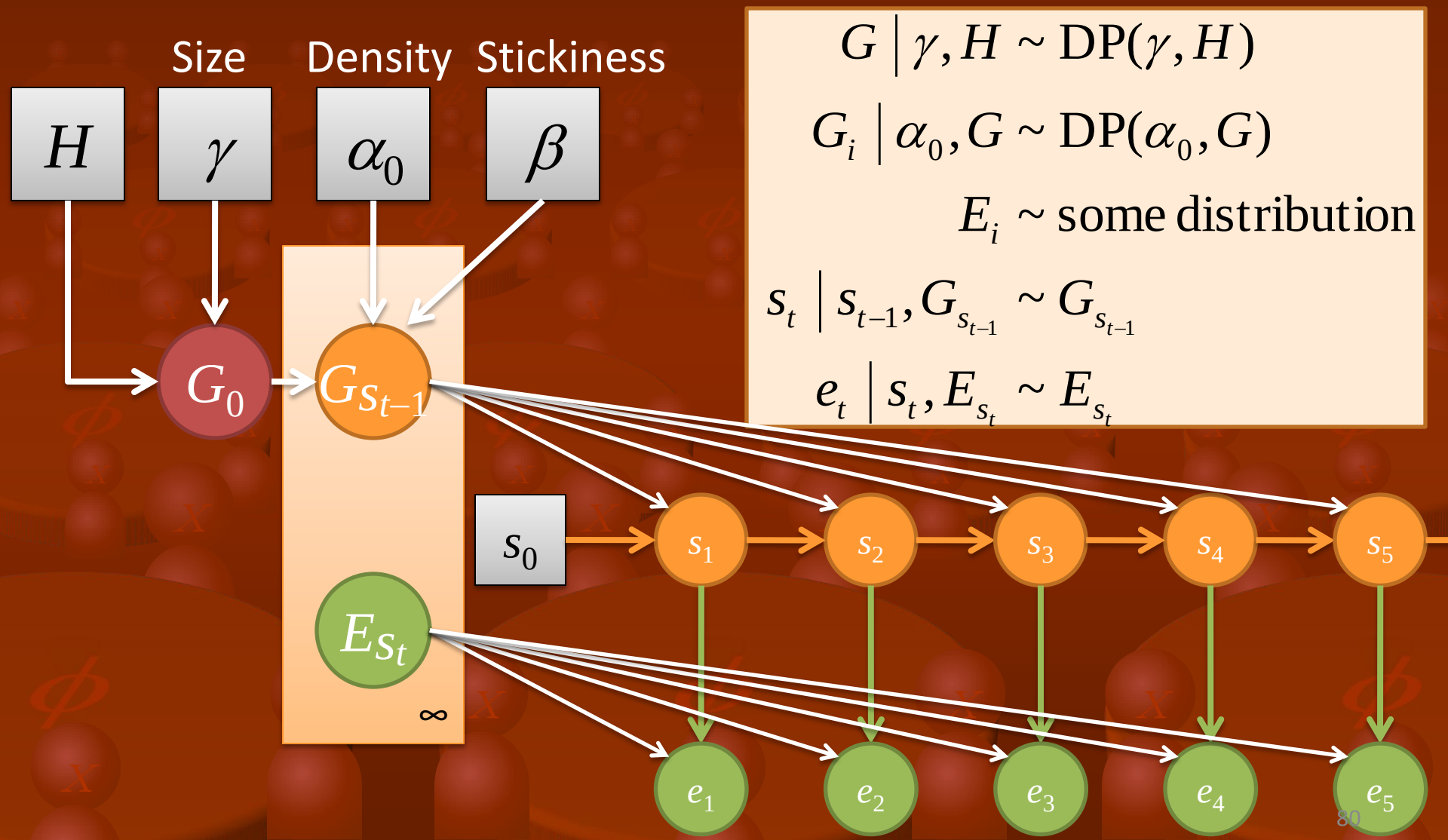
$$E_i \sim \text{some distribution}$$

$$s_t \mid s_{t-1}, G_{s_{t-1}} \sim G_{s_{t-1}}$$

$$e_t \mid s_t, E_{s_t} \sim E_{s_t}$$



Infinite Hidden Markov Models



$$G \mid \gamma, H \sim \mathrm{DP}(\gamma, H)$$

$$G_i \mid \alpha_0, G \sim \mathrm{DP}(\alpha_0, G)$$

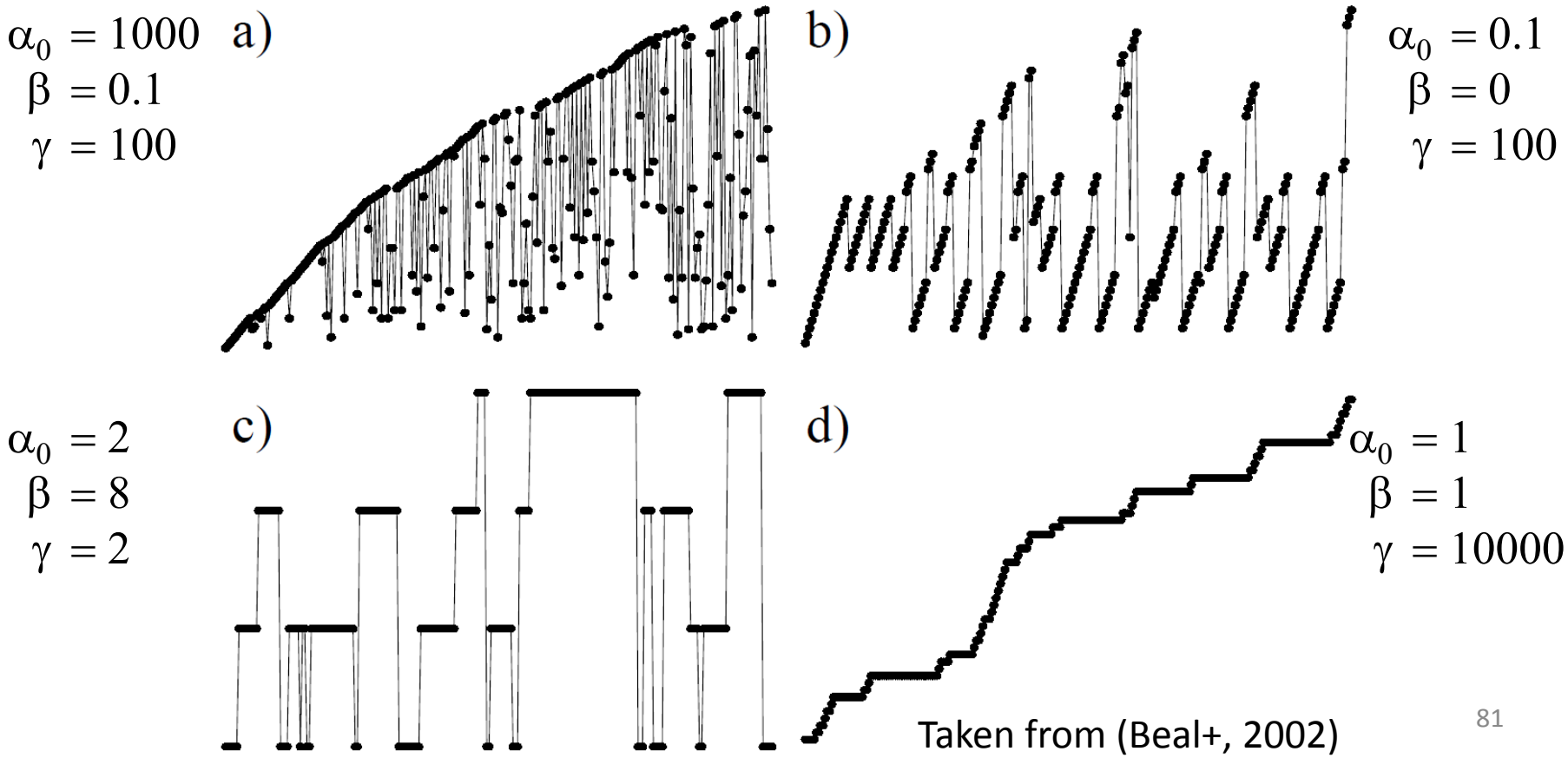
$$E_i \sim \text{some distribution}$$

$$s_t \mid s_{t-1}, G_{s_{t-1}} \sim G_{s_{t-1}}$$

$$e_t \mid s_t, E_{s_t} \sim E_{s_t}$$

State Sequence Produced by iHMMs

- ハイパーパラメータを設定することで、様々な系列が生成される



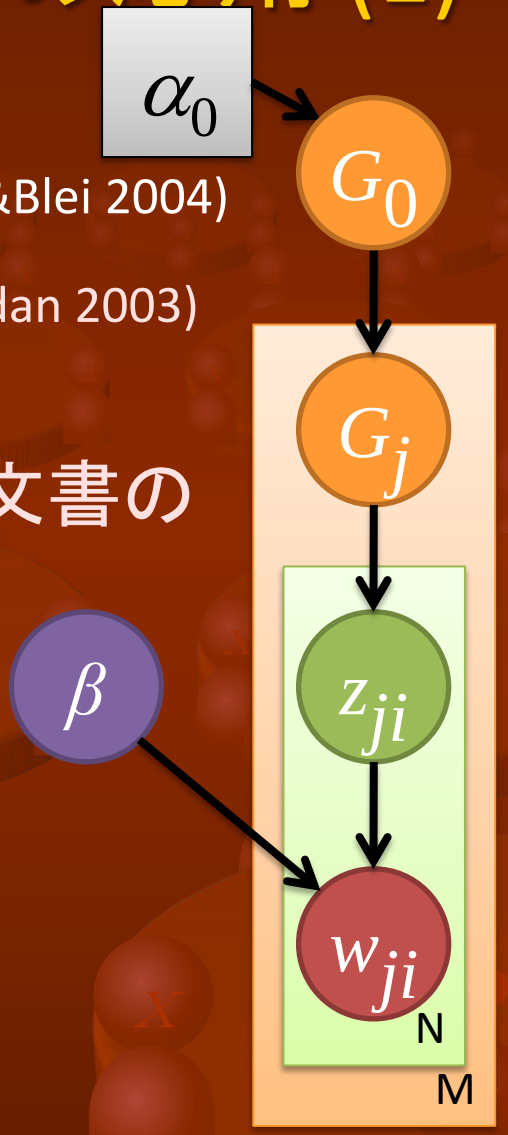
Taken from (Beal+, 2002)

iHMM の学習

- 変分ベイズ法 (Beal, 2003)
  - 状態数を有限個 (多めに) 置いて近似する
- Blocked Gibbs Sampler (Teh+, 2006)
  - Stick-Breaking process を使い、遷移行列を sampling
  - 状態が無限個あるため、Forward-backward が利用不能 → Beam Sampler (Van Gael+, 2008)
- Collapsed Gibbs Sampler
  - Chinese Restaurant Process を使い、遷移行列の期待分布を計算
  - Conditional Simultaneous Draws Sampler (Makino+, 2009)

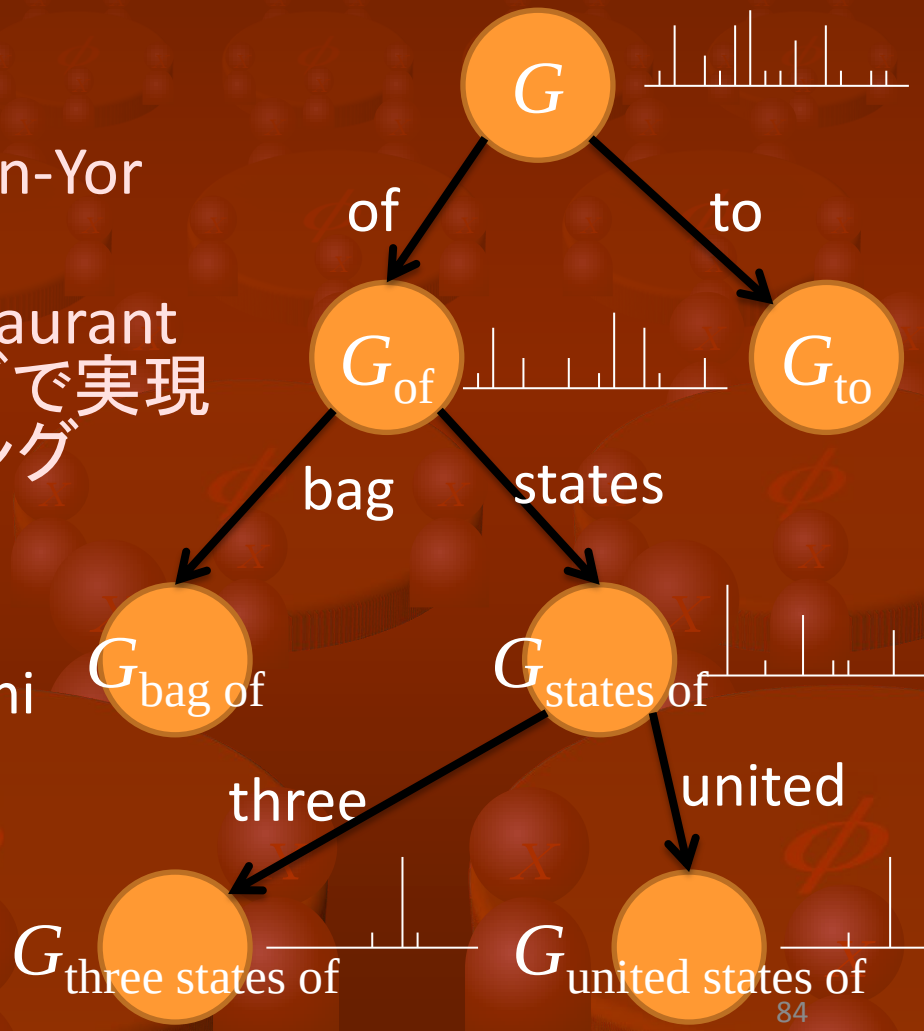
Hierarchical Dirichlet Process の応用 (1)

- 文書のトピックモデル (Teh, Jordan, Beal&Blei 2004)
  - Latent Dirichlet Allocation (Blei, Ng&Jordan 2003) のノンパラメトリック版
  - トピック数 K を明示的に与えず、各文書のトピック分布を生成する
  - データ (文書集合) 観測後の事後分布を計算することで、最適なトピック数の分布とともに、各トピックの単語分布を推定できる



Hierarchical Dirichlet Process の応用 (2)

- N-gram model (Teh, 2006)
  - Suffix tree の各 node に、単語分布を表わす Pitman-Yor Process を割り当てる
  - 学習は階層 Chinese Restaurant process 上のサンプリングで実現 → Kneser-Ney スムージングが帰着できる
- さらに N を積分消去した ∞-gram モデル (Mochihashi &Sumita, 2007)、Sequence Memoizer (Wood+, 2009) …



iHMM の状態の階層クラスタリング

(Makino+, 2009; Makino+, in submission)

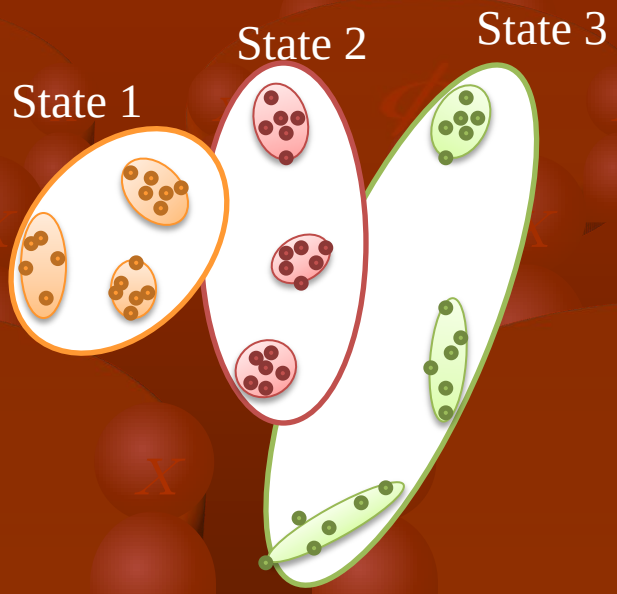
- HMM: 各々の隠れ状態は独立
- 現実には、状態間の類似性がある場合が多い
- Dirichlet Process は、クラスタリングのためにも利用できる

→ うまく組み合わせて、
隠れ状態をクラスタリングする

adjective	noun	noun	noun	noun	verb	noun	aux verb	verb
superlative	common	common	common	proper	transitive	pronoun	present	transitive
	time	title	title	name	past			bare form
				family				
Last	**night**	**Prime**	**Minister**	**Koizumi**	**said**	**he**	**will**	**make**

iHMM with Dirichlet Process Mixtures
(Fox+, 2008)

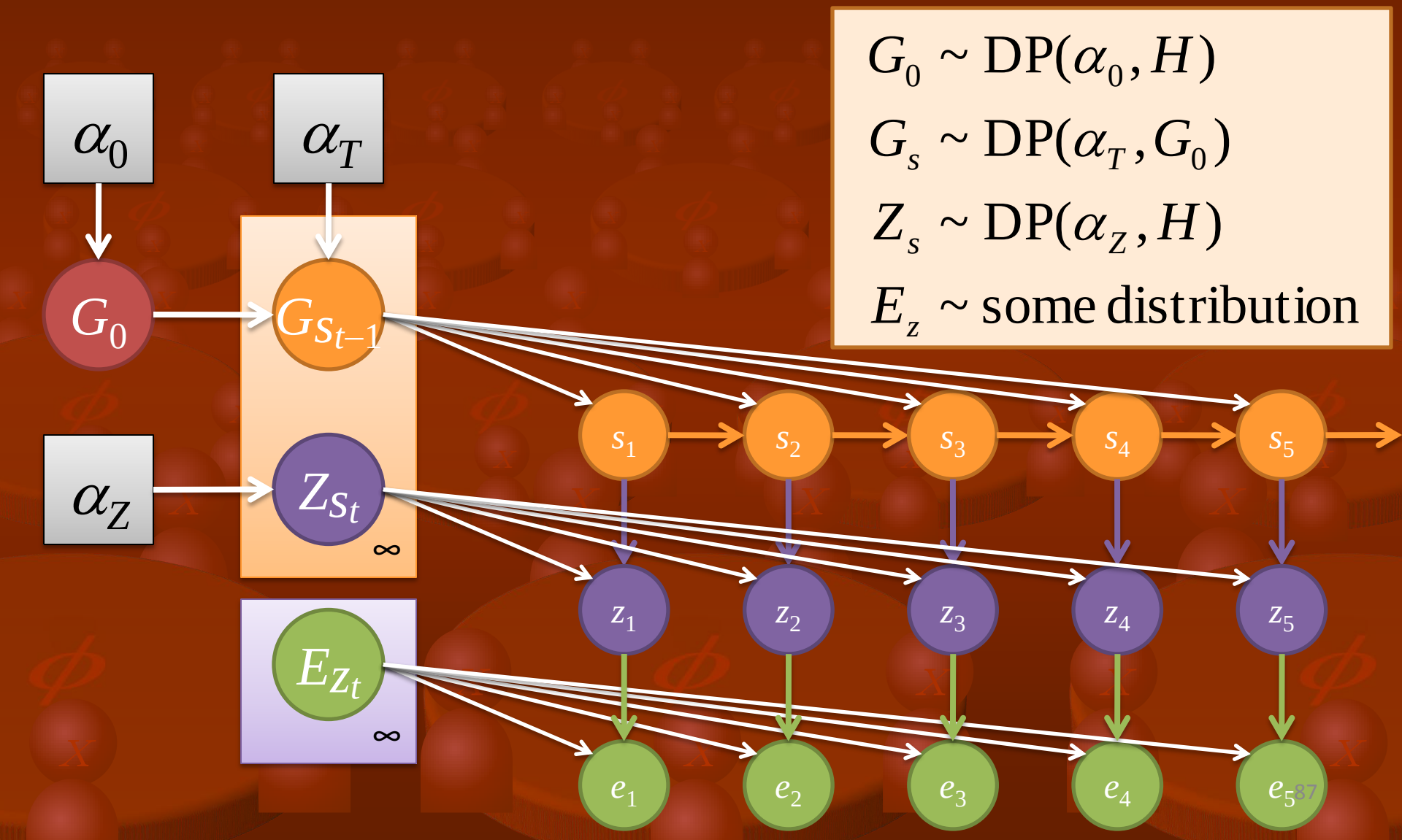
- HMMからの出力をクラスタリングする
  - Nested Chinese Restaurant Process (Blei+, 2004) を HMM に拡張したものに相当



$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_s \sim \mathrm{DP}(\alpha_T, G_0)$$
$$Z_s \sim \mathrm{DP}(\alpha_Z, H)$$
$$E_z \sim \text{some distribution}$$

e.g. Gauss-Wishart

iHMM with Dirichlet Process Mixtures



$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_s \sim \mathrm{DP}(\alpha_T, G_0)$$
$$Z_s \sim \mathrm{DP}(\alpha_Z, H)$$
$$E_z \sim \text{some distribution}$$

提案手法

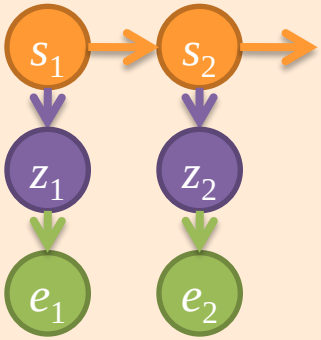
- クラスタ構造が、遷移に影響しない
- 欲しいのは、むしろ遷移確率に関するクラスタ
  - クラスタ内の各状態への遷移が類似
  - クラスタ内の各状態からの遷移が類似

- z_t も状態の一部となるようにできないか?
  → Hierarchical Nested infinite HMM (HN-iHMM)

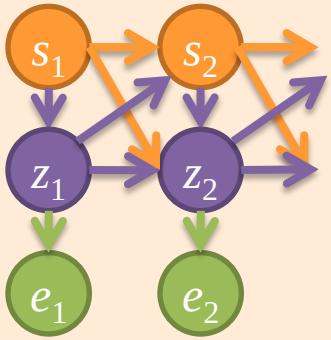
2階層 HN-iHMM

- z_t が状態の一部となるためには:
  - z_t が s_{t+1} の確率に影響する
  - z_t の確率が s_{t-1} (と z_{t-1}) に依存する

$$s_t \sim p(s_t \mid s_{t-1})$$
$$z_t \sim p(z_t \mid s_t)$$
$$e_t \sim p(e_t \mid z_t)$$

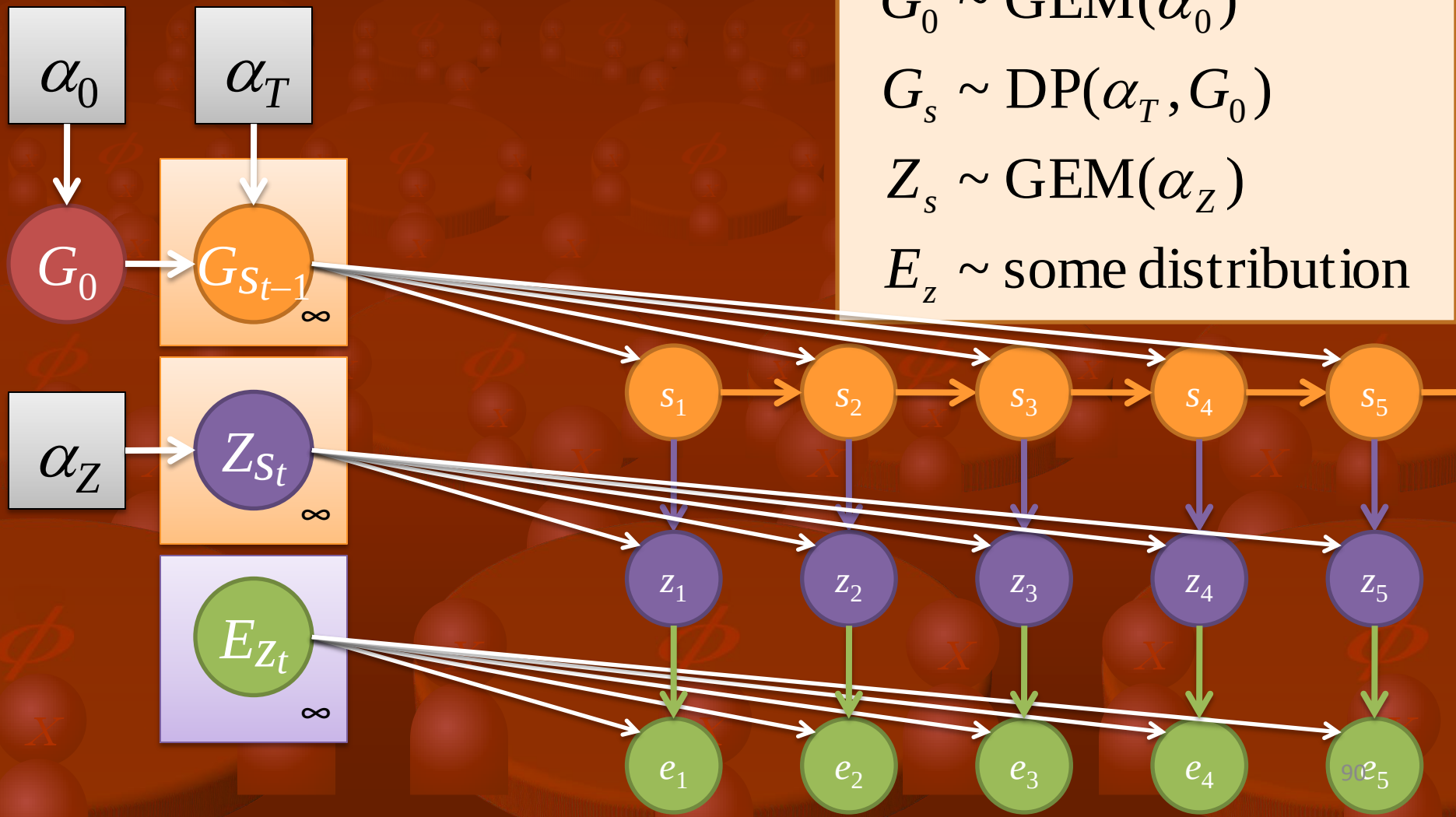


$$s_t \sim p(s_t \mid s_{t-1}, z_{t-1})$$
$$z_t \sim p(z_t \mid s_t, s_{t-1}, z_{t-1})$$
$$e_t \sim p(e_t \mid z_t)$$

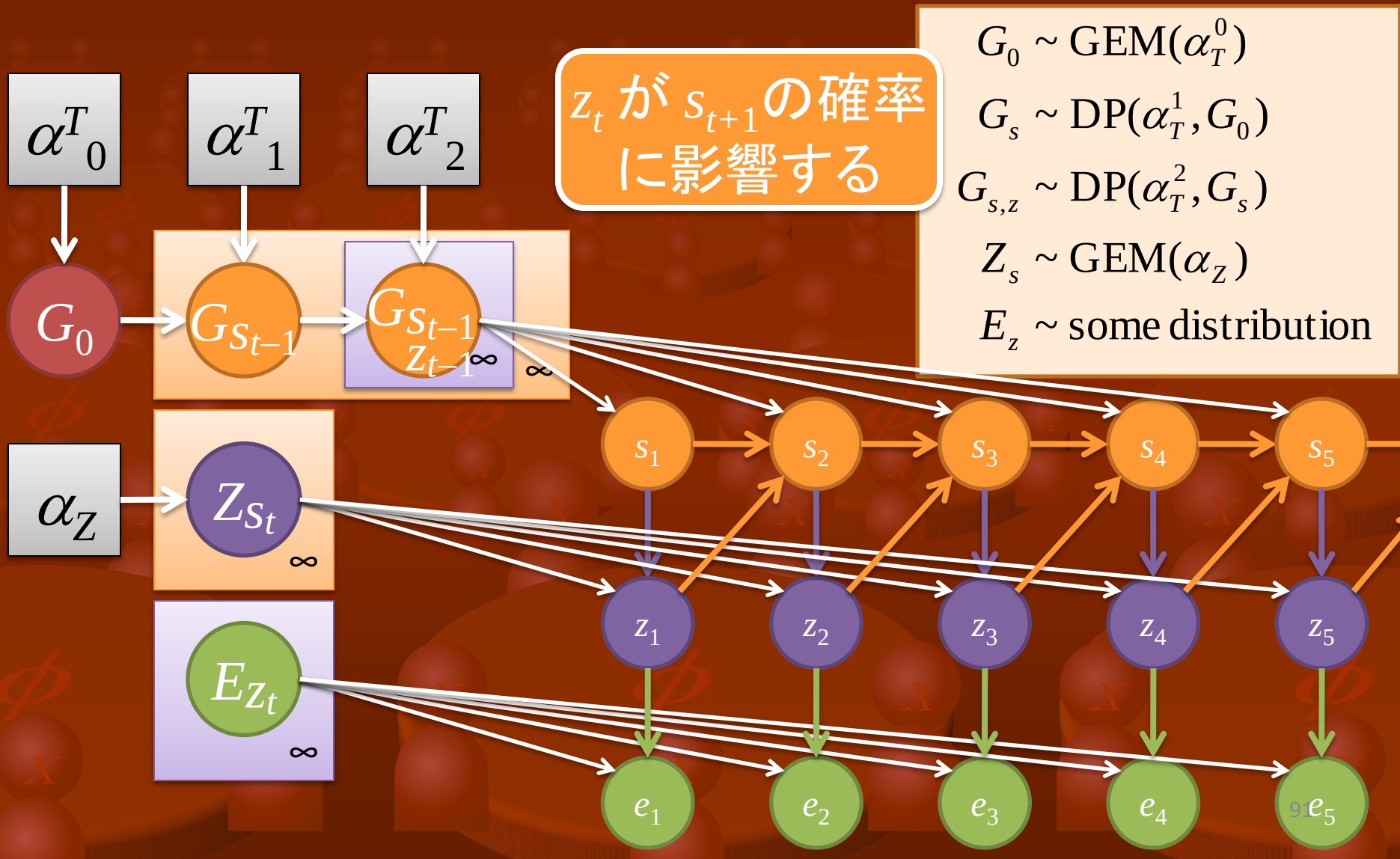


2階層HN-iHMMの構成 (1)

$$G_0 \sim \mathrm{GEM}(\alpha_0)$$
$$G_s \sim \mathrm{DP}(\alpha_T, G_0)$$
$$Z_s \sim \mathrm{GEM}(\alpha_Z)$$
$$E_z \sim \text{some distribution}$$



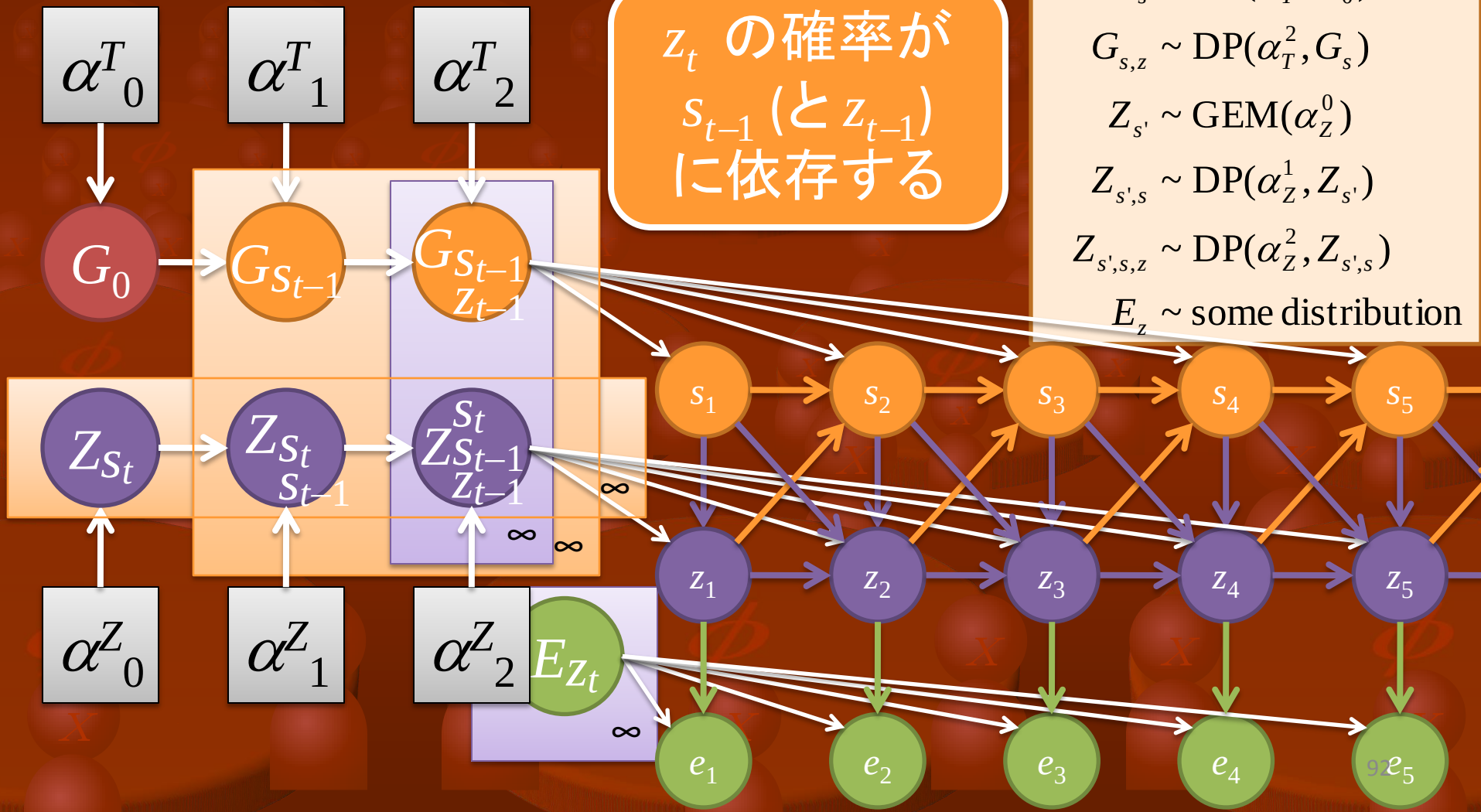
2階層HN-iHMMの構成 (2)



2階層HN-iHMMの構成 (3)

z_t の確率が
s_{t-1} (と z_{t-1})
に依存する

$$G_0 \sim \mathrm{GEM}(\alpha_T^0)$$
$$G_s \sim \mathrm{DP}(\alpha_T^1, G_0)$$
$$G_{s,z} \sim \mathrm{DP}(\alpha_T^2, G_s)$$
$$Z_{s'} \sim \mathrm{GEM}(\alpha_Z^0)$$
$$Z_{s',s} \sim \mathrm{DP}(\alpha_Z^1, Z_{s'})$$
$$Z_{s',s,z} \sim \mathrm{DP}(\alpha_Z^2, Z_{s',s})$$
$$E_z \sim \text{some distribution}$$



2階層以上のHN-iHMM

- z_t が状態の一部であるように表記を変更
  - $s_t \rightarrow s_t^1$
  - $z_t \rightarrow s_t^2$
- n-階層状態の HMM に自然に拡張できる

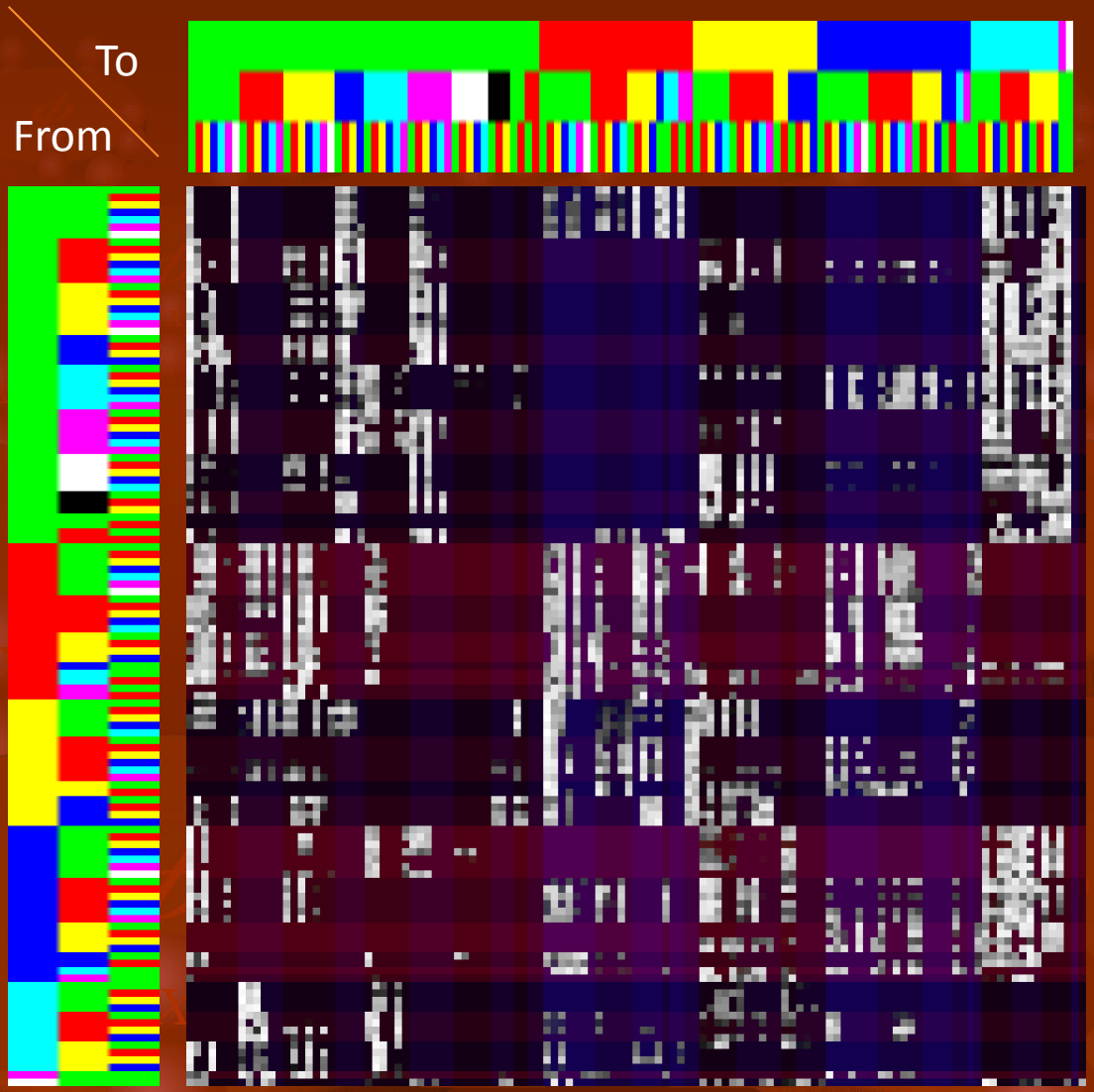
$$s_t^k \sim p(s_t^k \mid s_t^{1:k-1}, s_{t-1}^{1:n}) \qquad G_{s'^{1:k-1}}^k \sim \mathrm{GEM}(\alpha_0^k)$$

$$e_t \sim p(e_t \mid s_t^n) \qquad G_{s'^{1:k-1}, s^{1:l}}^k \sim \mathrm{DP}(\alpha_l^k, G_{s'^{1:k-1}, s^{1:l-1}}^k)$$

$$E_{s^{1:n}} \sim \text{some distribution}$$

HN-iHMMの例

- 生成された遷移行列の例
  - クラスタ構造による、遷移確率の類似性が実現されている

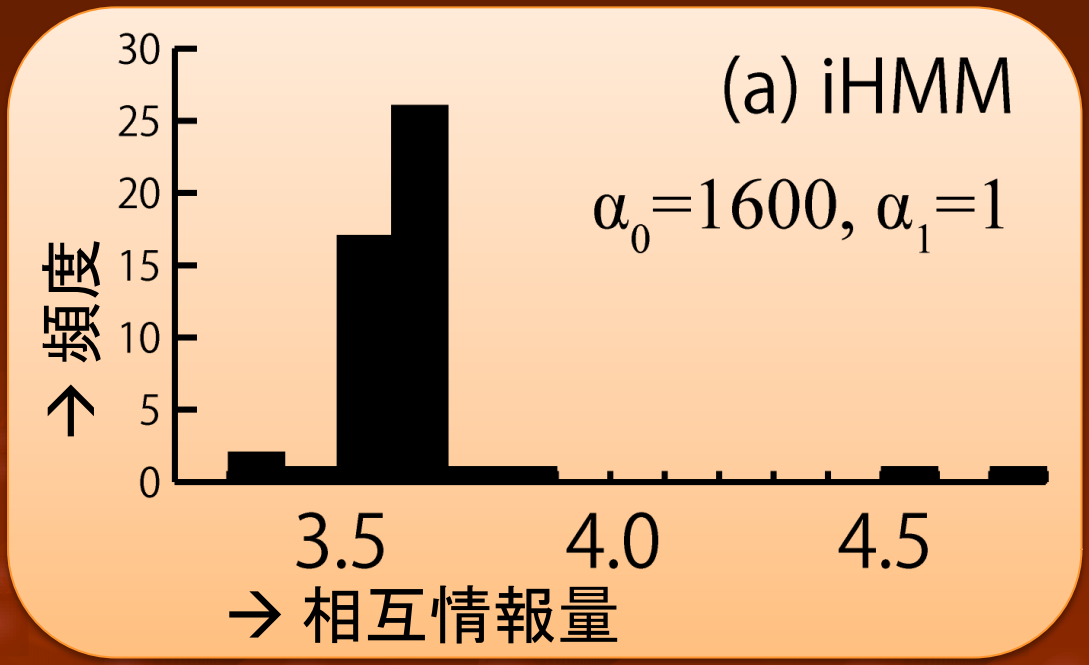


実験 (1)

- 人工HMMから生成したシーケンスをもとに、もとの HMM を推定する
  - 一見、ノイズの多い A-B-C の繰り返しに見えるが、少しずつ性質の異なる多数のシーケンスの組み合わせで構成されている
- Beam Sampling 50,000 step 後の状態列と、真の状態列との間の相互情報量で評価
  - データ長 N=500

結果 (1)

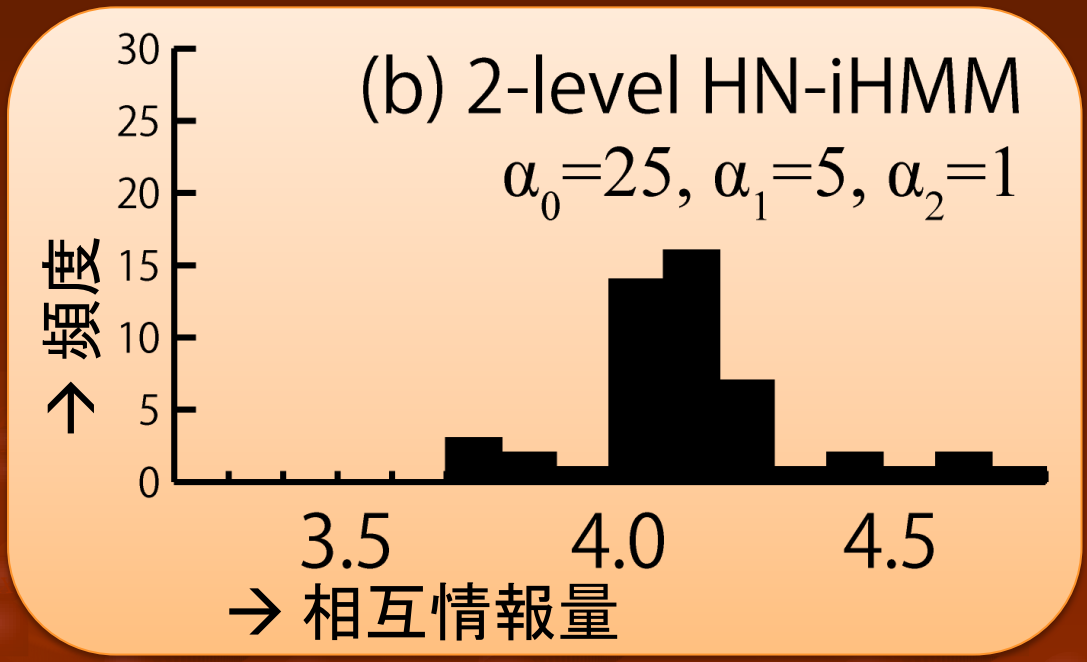
- 従来の iHMM では、正しく分離して隠れ状態を学習することができない
  - MCMC サンプリングの系列が local maximum に陥ってしまい、脱出できない



このときの
相互情報量
MI = 3.61

結果 (2)

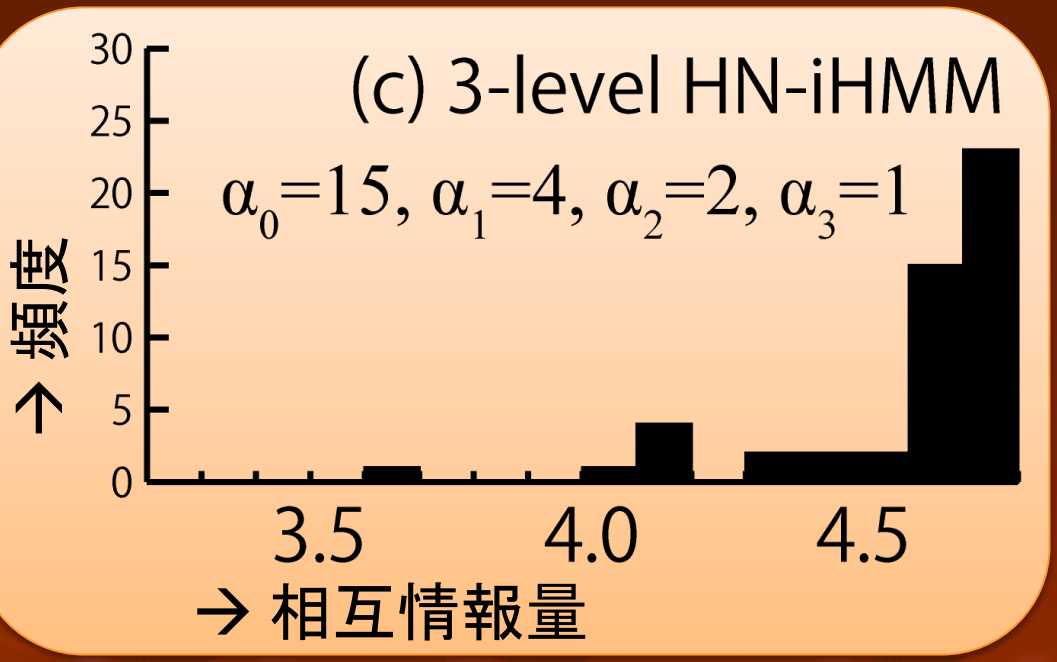
- 2階層 HN-iHMM を使うと、より精度よく推定できる
  - しかし、手掛かりの少ない状態については、分離できないことが多い

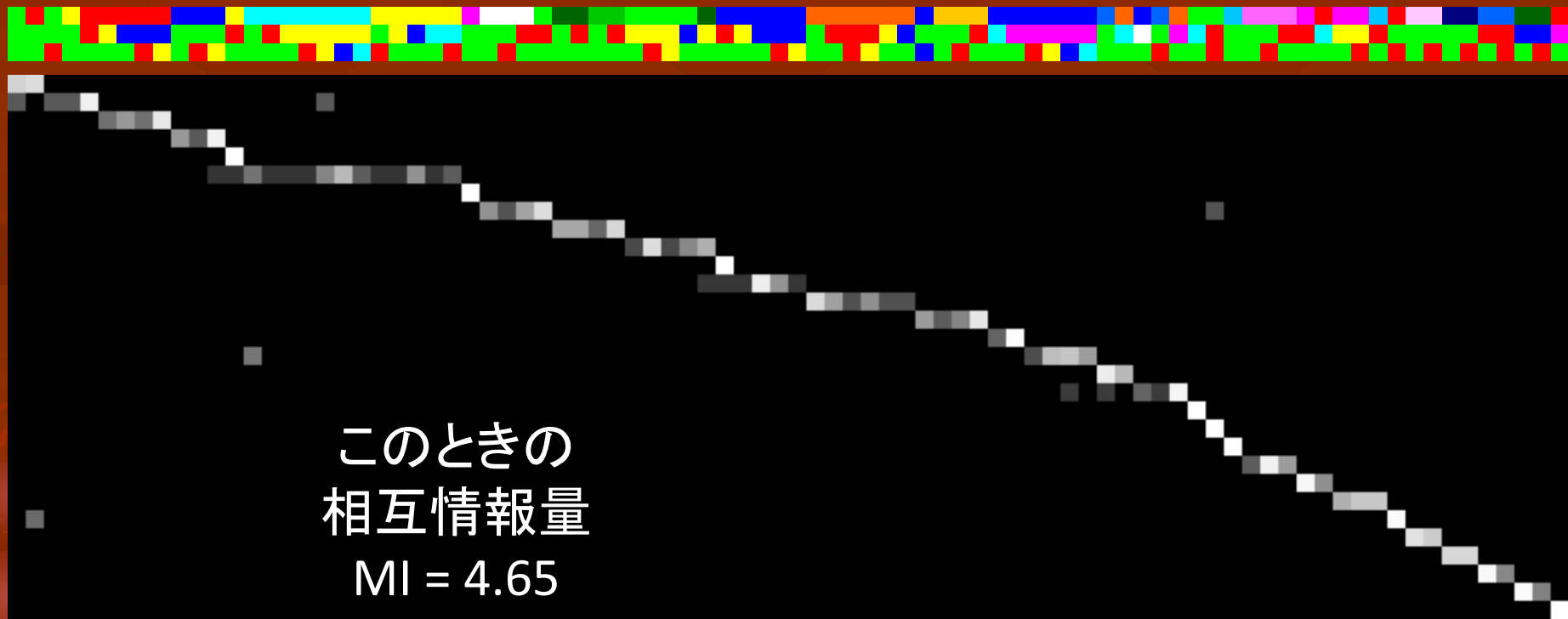


このときの
相互情報量
MI = 4.18

結果 (3)

- 3階層 HN-iHMM を使うと、かなり精度よく推定可能
  - → サンプリングされた状態





実験1の考察

- HN-iHMM では、余分な state が多数生成されるが、精度が下がらない
  - 通常の HMM (や iHMM) では、余分な state があると精度が低下する
  - HN-iHMM では、クラスタ中の他状態の情報が使えるので、精度が下がりにくい
  - 逆に、学習における local maximum からの脱出を助ける効果がある

実験 (2)

- Alice's Adventure in Wonderland のテキストを利用した実験 (Makino+, in submission)
  - 2階層状態 HMM を利用
  - 28,120 語 (End-Of-Sentence 含む)
  - 2回以下しか登場しない単語は UNK に置換 → 異なり単語数 1,078 語
  - Conditional Simultaneous Draws Sampler (Makino+, 2009) によりサンプリング
  - 19,000 回目の sweep の後では、34 個のクラスタ中に 774 個の状態があった

結果 (2)

Level 1 (clusters)	Level 2 (states)		emission symbols with occurrence counts				
A	a	714	457–she	96–they	63–it	55–he	30–alice
3994	b	467	61–queen	61–king	55–gryphon	53–hatter	38–duchess
[42]	c	324	180–it	67–there	29–she	21–this	16–that
	d	303	283–and	8–while	6–even		
	e	279	267–and				
	f	229	79–when	69–as	30–if	17–that	15–because
B	a	1577	1412–EOS	71–and	66–but	11–why	9–UNK
3716	b	675	653–the	19–a			
[33]	c	400	137–her	75–his	57–its	46–their	40–my
	d	180	174–alice	5–five			
	e	155	110–that	22–what	11–whether	9–though	
	f	144	99–very	20–quite	18–rather		
C	a	460	459–to				
3396	b	417	412–of				
[48]	c	322	232–in	63–with	19–on		
	d	179	64–up	59–down	22–out	17–back	6–on
	e	147	107–at	13–after	11–among	7–before	5–like
	f	147	53–off	39–down	24–UNK	22–away	
D	a	557	497–the	20–this	17–her	7–these	6–those
2631	b	449	441–the				
[41]	c	438	436–a				
	d	123	94–a	16–another	5–such		
	e	115	71–no	18–any	14–some		
	f	86	65–one	8–each	7–his		
E	a	378	334–said	15–cried	7–added	6–shouted	
2314	b	356	252–you	50–i	29–they	14–we	5–cats
[34]	c	337	321–i	10–we			
	d	198	194–as				
	e	172	43–don't	28–can't	25–could	20–can	20–shall
	f	107	50–and	20–now	12–just	11–not	5–dinah
F	a	218	169–UNK	27–put	10–have		
1705	b	175	148–be	10–take	8–feel		
[40]	c	132	50–go	32–come	22–look	8–sit	6–stand
	d	111	23–do	18–say	17–UNK	13–begin	12–speak
	e	96	28–tell	15–let	14–UNK	10–make	7–fetch
	f	93	81–went	5–swam			
G	a	211	172–was	16–UNK	7–is		
1504	b	182	149–was	33–were			
[40]	c	102	32–came	26–began	11–UNK	10–kept	6–stood
	d	101	33–UNK	24–made	10–left	8–called	8–got
	e	72	29–heard	15–gave	5–noticed		
	f	64	26–got	19–tried	10–set		
H	a	197	57–would	44–must	18–will	16–doesn't	15–could
1443	b	167	161–had				
[41]	c	86	51–know	10–like	9–were	6–mean	5–do
	d	84	66–have	7–never	5–seem		
	e	82	30–think	20–wish	10–suppose	6–believe	6–say
	f	77	42–did	21–could	13–would		
I	a	252	247–UNK				
1003	b	114	65–little	13–UNK	9–great	6–sharp	5–sudden
[30]	c	92	48–UNK	14–curious	12–good		
	d	57	57–turtle				
	e	55	15–sort	12–UNK	7–tone	5–pair	
	f	48	32–tone	14–voice			
J	a	86	29–eyes	15–UNK	14–face	9–feet	8–sister
949	b	77	43–way	9–dormouse	6–UNK	6–lobsters	5–hookah
[41]	c	74	30–door	18–house			
	d	72	38–time	17–moment	6–question	5–morning	
	e	56	36–thing	6–day	5–question		
	f	52	19–UNK	17–words	–	–	

結果 (2)

Level 1 (clusters)	Level 2 (states)		emission symbols with occurrence counts				
C	a	460	459–to				
3396	b	417	412–of				
[48]	c	322	232–in	63–with	19–on		
	d	179	64–up	59–down	22–out	17–back	6–on
	e	147	107–at	13–after	11–among	7–before	5–like
	f	147	53–off	39–down	24–UNK	22–away	
D	a	557	497–the	20–this	17–her	7–these	6–those
2631	b	449	441–the				
[41]	c	438	436–a				
	d	123	94–a	16–another	5–such		
	e	115	71–no	18–any	14–some		
	f	86	65–one	8–each	7–his		
E	a	378	334–said	15–cried	7–added	6–shouted	
2314	b	356	252–you	50–i	29–they	14–we	5–cats
[34]	c	337	321–i	10–we			
	d	198	194–as				

結果 (2)

Level 1 (clusters)	Level 2 (states)		emission symbols with occurrence counts				
F	a	218	169–UNK	27–put	10–have		
1705	b	175	148–be	10–take	8–feel		
[40]	c	132	50–go	32–come	22–look	8–sit	6–stand
	d	111	23–do	18–say	17–UNK	13–begin	12–speak
	e	96	28–tell	15–let	14–UNK	10–make	7–fetch
	f	93	81–went	5–swam			
G	a	211	172–was	16–UNK	7–is		
1504	b	182	149–was	33–were			
[40]	c	102	32–came	26–began	11–UNK	10–kept	6–stood
	d	101	33–UNK	24–made	10–left	8–called	8–got
	e	72	29–heard	15–gave	5–noticed		
	f	64	26–got	19–tried	10–set		
H	a	197	57–would	44–must	18–will	16–doesn't	15–could
1443	b	167	161–had				
[41]	c	86	51–know	10–like	9–were	6–mean	5–do
	d	84	66–have	7–never	5–seem		
	e	82	30–think	20–wish	10–suppose	6–believe	6–say

まとめ

- 機械学習におけるベイジアンの位置づけ
  - パラメータの分布を考えることで、パラメータ空間の取り方に依存せず、過学習しにくい推定を可能に
  - 不確かさを明示的に扱える
  - パラメータを積分消去 → 予測分布の利用
- Infinite HMM: HMMのノンパラメトリックベイズ版
  - 隠れ状態数 N を積分消去し、理論上、無限個の隠れ状態を扱える学習モデル
  - Hierarchical Dirichlet Process を利用
- HN-iHMM: HMM隠れ状態の階層クラスタリング
  - 遷移確率の類似に基づいて隠れ状態を階層化
  - 類似する系列が多数現れる場合の学習・予測の実現