トピックモデルによる統計的潜在意味解析 読書会 2章「Latent Dirichlet Allocation」(前半)

日付：2015/06/04
発表者：@_kobacky
場所：株式会社 ALBERT セミナールーム

【Amazon】
http://www.amazon.co.jp/dp/4339027588/

【注意】
この資料は @_kobacky が本を読んで理解した内容を記載したものであり、本には存在しない記述や本とは異なる表現をしている部分があります。

2章「Latent Dirichlet Allocation」

2.1：概要
2.2：多項分布と Dirichlet 分布
2.3：LDA の生成過程

2.1 概要

- LDA：文書の確率的生成モデル
- Bag of Words (BoW)
  - 文書を単語と出現頻度のペアの集合として表現
  - 単語の順序は無視
- 各文書には潜在トピックがあると仮定
  - 統計的に共起しやすい単語集合が生成される要因を潜在トピックという観測できない確率変数を用いて定式化
  - LDA では一つの文書には複数の潜在トピックが存在すると仮定
- 潜在トピックモデル
  - LDAの改良モデルの総称
  - 共起情報が大量にあるようなデータ全般に適用可能
    - Bag of Items (ユーザーの購買履歴)や、その他 Bag of XXX

2.2 多項分布とDirichlet分布

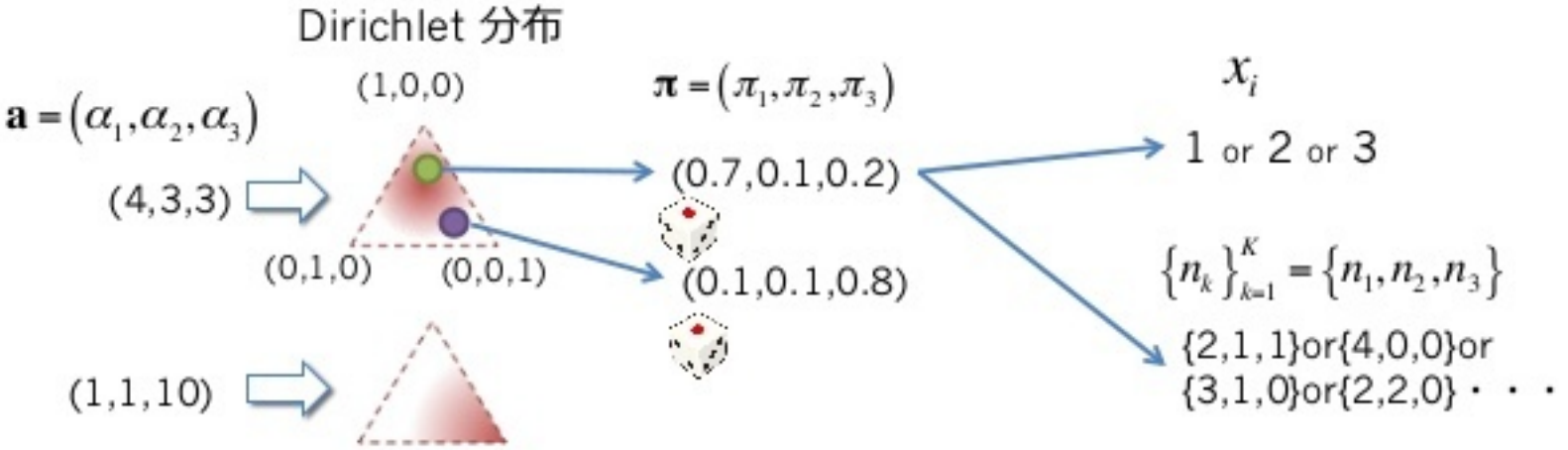
- LDA の説明の準備として、サイコロのイメージを用いて多項分布と Dirichlet 分布について説明
- 多項分布
  - 各目が出る確率が $\boldsymbol{\pi}$ で表されるサイコロを n 回振った時に各目が出る回数の分布
- Dirichlet 分布
  - サイコロの形状(出る目の種類 K と出やすさ $\boldsymbol{\pi}$) の分布

サイコロの形状
(Dirichlet 分布により生成)

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_K) \qquad \left(\sum_{k=1}^{K} \pi_k = 1\right)$$

「1」が出る確率　　出る目の種類数

2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ



2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ①

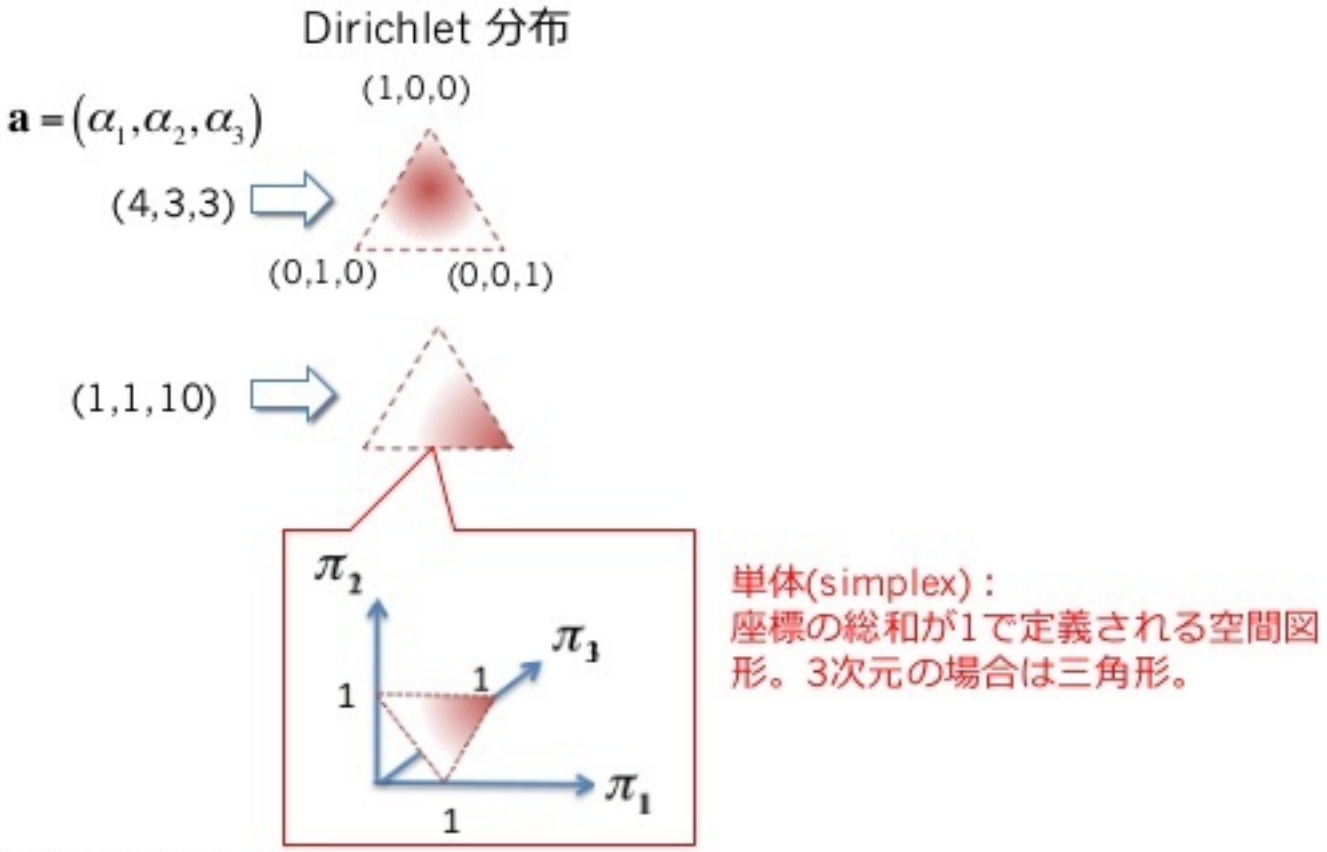
(K=3)

$\mathbf{a} = (\alpha_1, \alpha_2, \alpha_3)$

(4,3,3)

(1,1,10)

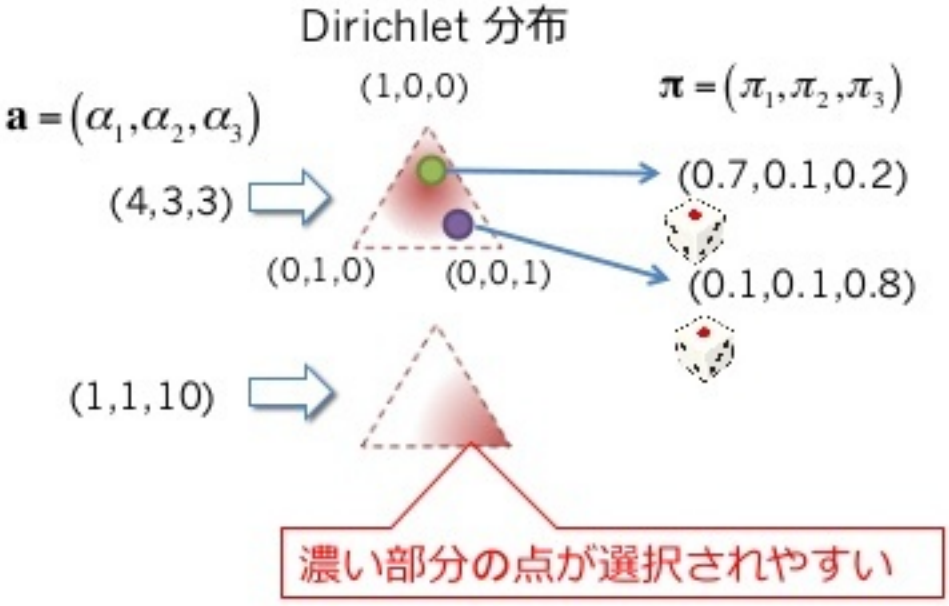
2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ②



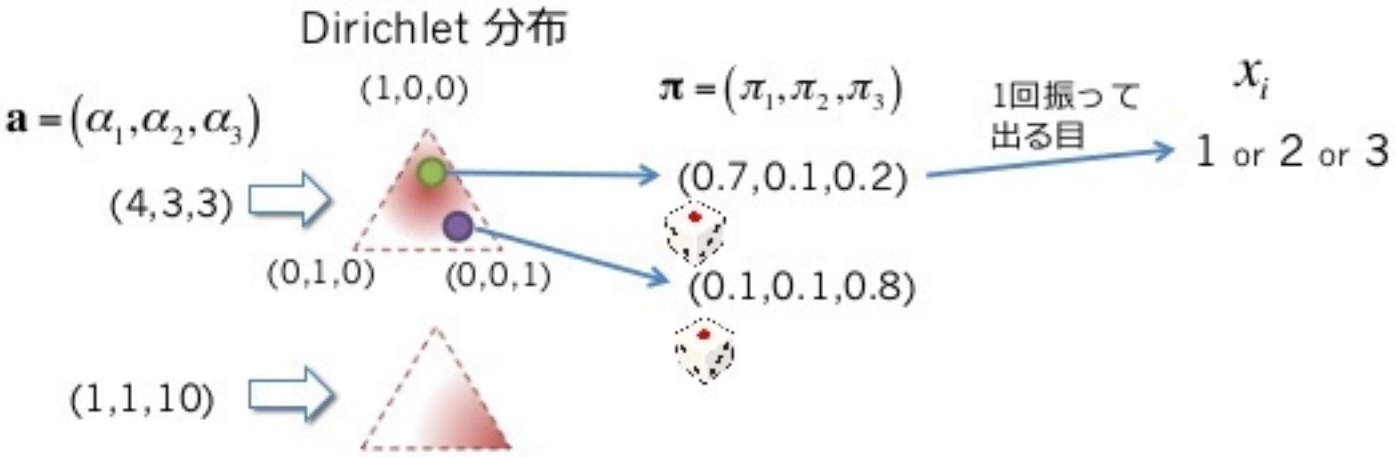
単体(simplex)：
座標の総和が1で定義される空間図形。3次元の場合は三角形。

※αの値と Dirichlet 分布の図の対応はイメージです。

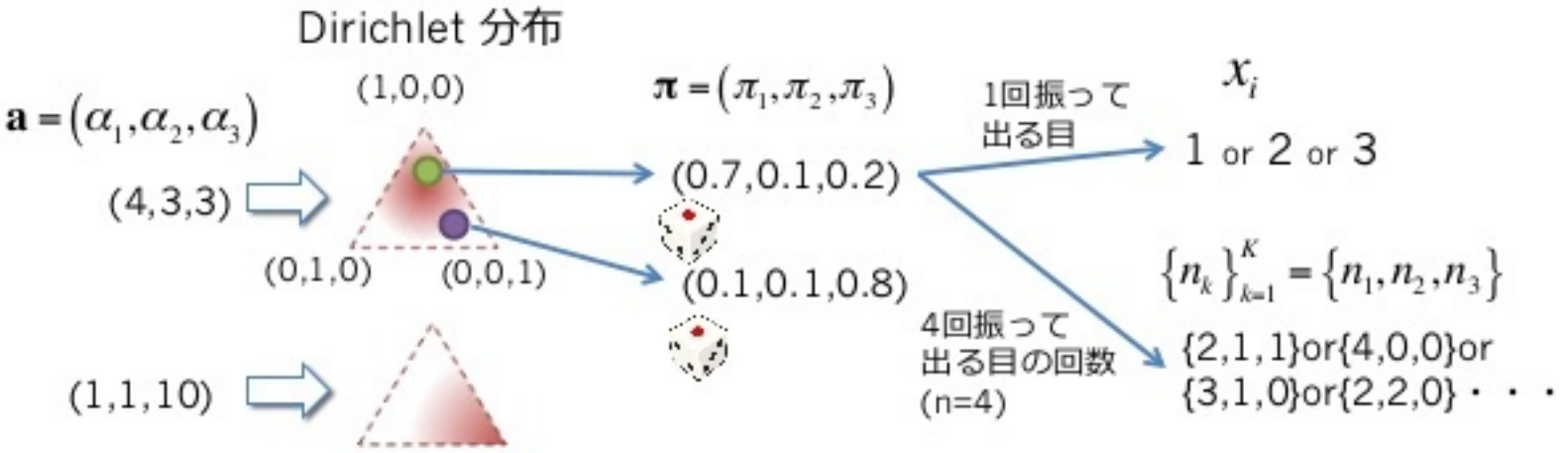
2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ③



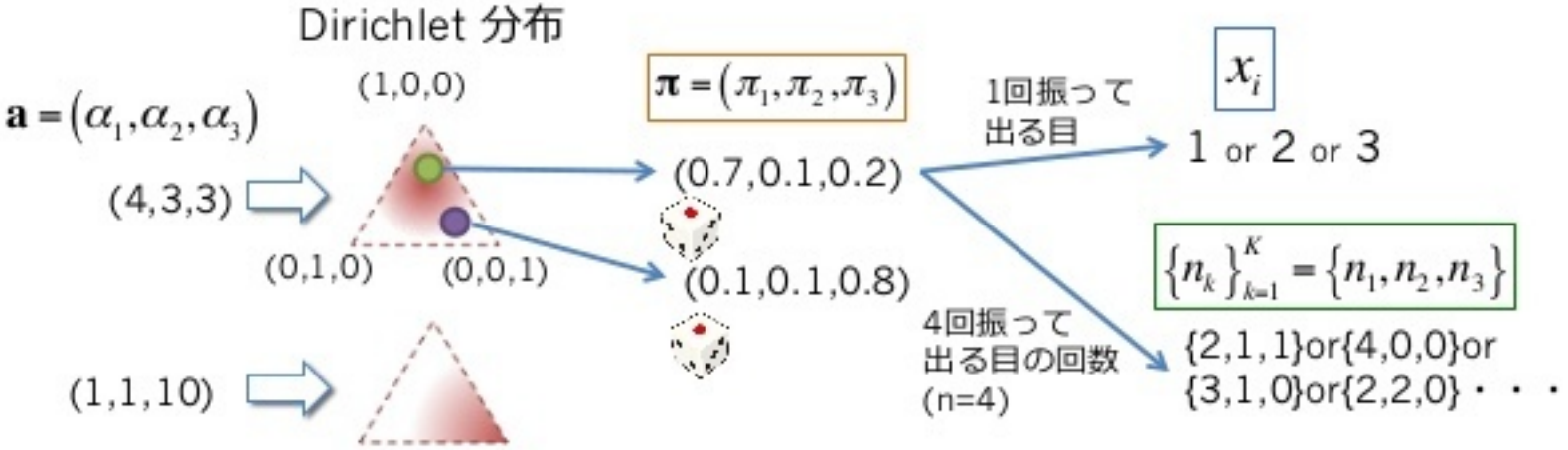
2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ④



2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ⑤

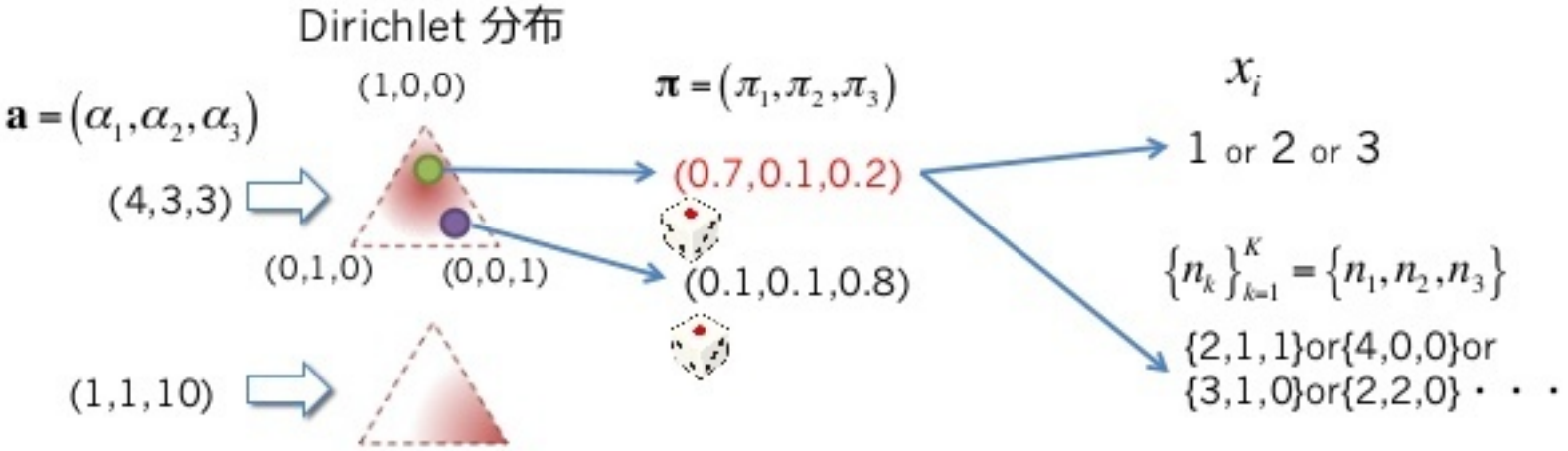


2.2 多項分布とDirichlet分布
サイコロの目が生成されるイメージ⑥



- **π (ベクトル)**：サイコロの目の出やすさ
  - **α**で特徴付けられる Dirichlet 分布に従う
- x_i ：1回の試行で出る目
  - **π** の分布に従う (n=1 の多項分布に従う)
- $\{n_k\}$ (集合)：各目が出る回数
  - **π** と **n** で特徴付けられる多項分布に従う

2.2 多項分布とDirichlet分布
式(2.1)：x の生起確率



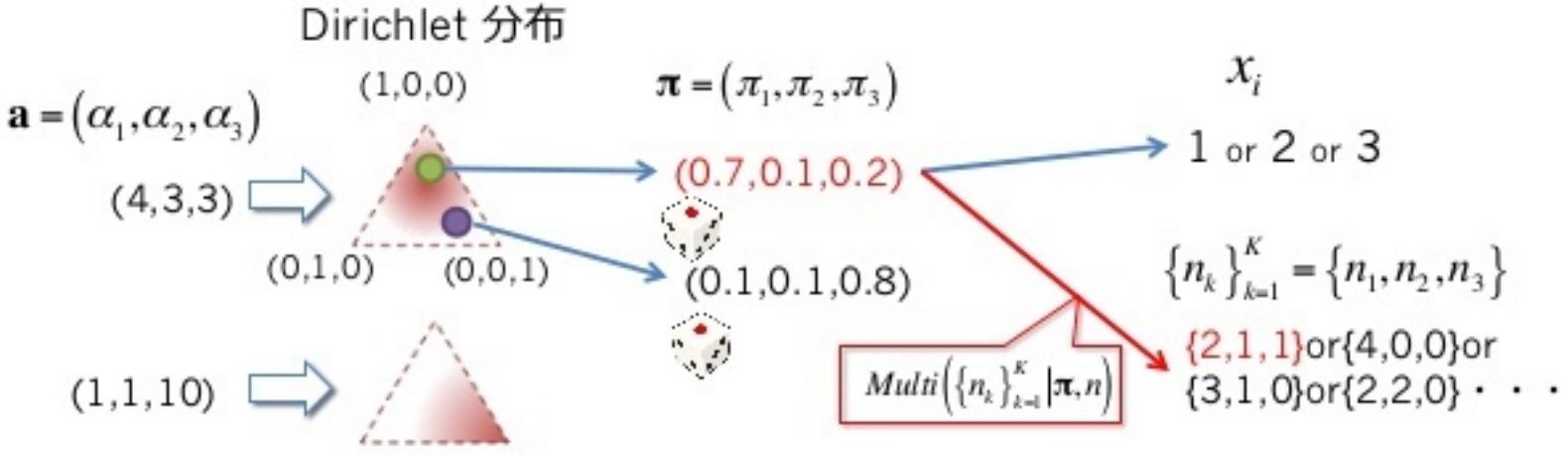
$\mathbf{x}=(x_1,x_2,x_3,\cdots,x_n)$ の生起確率

$$p(\mathbf{x}|\boldsymbol{\pi})=\prod_{i=1}^{n}p(x_i|\boldsymbol{\pi})=\prod_{k=1}^{K}\pi_k^{n_k}$$ 式(2.1)

【例】x=(1,2,1,3)の場合

$$p((1,2,1,3)|\boldsymbol{\pi})=0.7\times0.1\times0.7\times0.2=0.7^2\times0.1^1\times0.2^1$$

2.2 多項分布とDirichlet分布
式(2.2)：多項分布による出目回数の生成

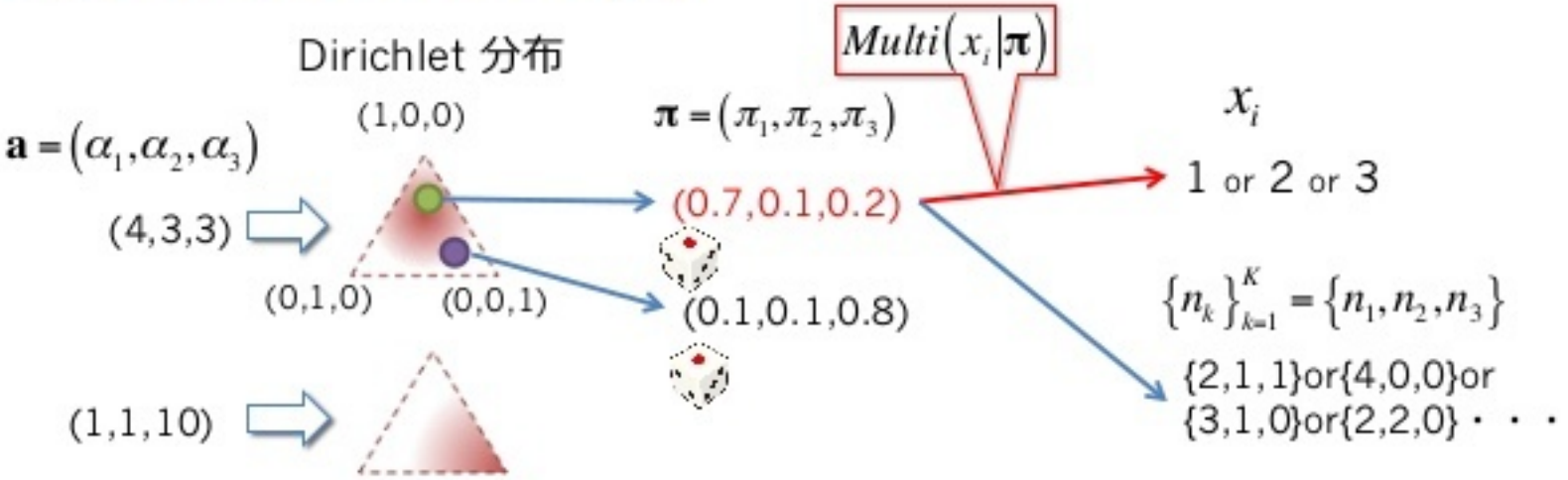


$$p\left(\{n_k\}_{k=1}^K|\boldsymbol{\pi},n\right)=Multi\left(\{n_k\}_{k=1}^K|\boldsymbol{\pi},n\right)\equiv\frac{n!}{\prod_{k=1}^K n_k!}\prod_{k=1}^K\pi_k^{n_k}$$ 式(2.2)

【例】π=(0.7,0.1,0.2)のサイコロで「1が2回、2が1回、3が1回」の場合

$$p\left(\{2,1,1\}|\boldsymbol{\pi},4\right)={}_4C_2\times{}_2C_1\times{}_1C_1\times\left(0.7^2\times0.1^1\times0.2^1\right)=\frac{4!}{2!\times1!\times1!}\left(0.7^2\times0.1^1\times0.2^1\right)$$

2.2 多項分布とDirichlet分布
式(2.3)：xi の生成



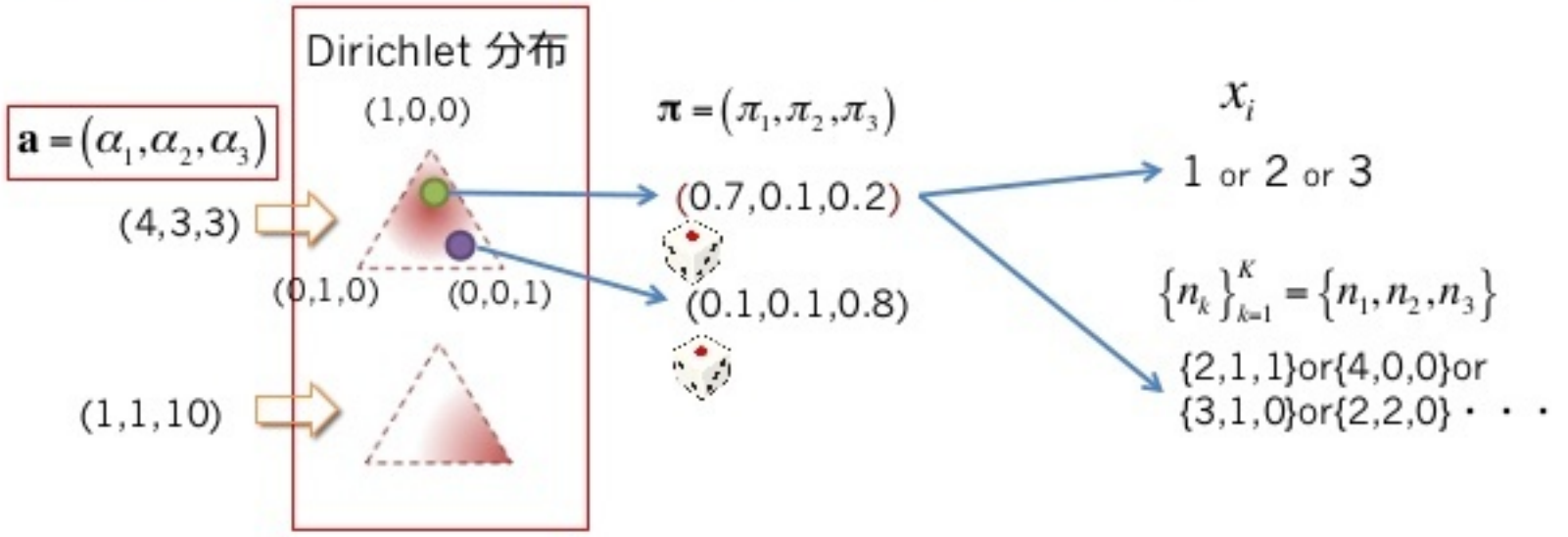
$$p(x_i=k|\boldsymbol{\pi})=Multi(n_k=1|\boldsymbol{\pi},1)=\frac{1}{\prod_{k=1}^{K}n_k!}\prod_{k=1}^{K}\pi_k^{n_k}=\pi_k$$

$$p(x_i|\boldsymbol{\pi})=Multi(x_i|\boldsymbol{\pi})$$

$$x_i \sim Multi(x_i|\boldsymbol{\pi}) \quad \text{式(2.3)}$$

kの目が出た回数は1
k以外の目が出た回数は0
であるため

2.2 多項分布とDirichlet分布
式(2.4)～式(2.6)：Dirichlet 分布



式(2.4)：Dirichlet 分布

$$p(\boldsymbol{\pi}|\mathbf{a})=Dir(\boldsymbol{\pi}|\mathbf{a})\equiv\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$$

正規化項
(a が定まった元では定数)

式(2.5)：ガンマ関数

$$\Gamma(1)=1$$
$$\Gamma(n)=(n-1)\Gamma(n-1)$$
$$\Gamma(\alpha+n)=(\alpha+n-1)\Gamma(\alpha+n-1)$$

式(2.6)：期待値・分散

$$E[\pi_k]=\frac{\alpha_k}{\alpha_0}$$
$$V[\pi_k]=\frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0^2(1+\alpha_0)}$$
$$\alpha_0=\sum_{k=1}^{K}\alpha_k$$

2.2 多項分布とDirichlet分布
共役事前分布

p(x)をp(y|x)の事前分布とする。
事後分布p(x|y)∝p(y|x)p(x)の分布がp(x)と同じ分布になるとき、p(x)はp(y|x)の共役事前分布であるという。

x の生成過程　$x_i \sim Multi(x_i|\boldsymbol{\pi})$　を考える。

式(2.4) **π** の事前分布　$p(\boldsymbol{\pi}|\mathbf{a}) = Dir(\boldsymbol{\pi}|\mathbf{a}) \equiv \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$

x が観測される前の π の分布

事前分布・
事後分布が共に Dirichlet 分布

式(2.9) **π** の事後分布　$p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) = \frac{\Gamma\left(\sum_{k=1}^{K}n_k+\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(n_k+\alpha_k)}\prod_{k=1}^{K}\pi_k^{n_k+\alpha_k-1}$

x が観測された後の π の分布

※導出は後ほど・・

α_k は k の仮想的頻度の意見を持つと見れる。

2.2 多項分布とDirichlet分布
式(2.7)：事後分布の導出(前半)

P(x|a) が定数である事に注目

$$p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) = \frac{p(\mathbf{x},\boldsymbol{\pi}|\mathbf{a})}{p(\mathbf{x}|\mathbf{a})} \propto p(\mathbf{x},\boldsymbol{\pi}|\mathbf{a}) = p(\mathbf{x}|\boldsymbol{\pi})\,p(\boldsymbol{\pi}|\mathbf{a})$$

$$= \prod_{i=1}^{n} p(x_i|\boldsymbol{\pi})\,p(\boldsymbol{\pi}|\mathbf{a}) = \prod_{i=1}^{n}\prod_{k=1}^{K} \pi_k^{\delta(x_i=k)}\,p(\boldsymbol{\pi}|\mathbf{a})$$

$$= \prod_{i=1}^{n} \pi_k^{\sum_{i=1}^{n}\delta(x_i=k)}\,p(\boldsymbol{\pi}|\mathbf{a}) = \prod_{i=1}^{n} \pi_k^{n_k}\,p(\boldsymbol{\pi}|\mathbf{a})$$

$$\propto \prod_{i=1}^{n} \pi_k^{n_k} \prod_{i=1}^{n} \pi_k^{\alpha_k-1} = \prod_{i=1}^{n} \pi_k^{n_k+\alpha_k-1}$$

P(π|a) の正則化項が定数である事に注目

2.2 多項分布とDirichlet分布
式(2.8)～(2.9)：事後分布の導出(後半)

$$p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) \propto \prod_{i=1}^{n} \pi_k^{n_k+\alpha_k-1}$$

式の形からp(π| x,a)がDirichlet 分布に従うことがわかる

$$\int p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) d\pi = 1$$

$$p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) = \frac{\prod_{k=1}^{K} \pi_k^{n_k+\alpha_k-1}}{\int \prod_{k=1}^{K} \pi_k^{n_k+\alpha_k-1} d\pi}$$

式(2.8)

p(π| x,a)がDirichlet 分布に従うから

$$p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a}) = \frac{\Gamma\left(\sum_{k=1}^{K} n_k + \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(n_k + \alpha_k\right)} \prod_{k=1}^{K} \pi_k^{n_k+\alpha_k-1}$$

式(2.9)

2.2 多項分布とDirichlet分布
式(2.10)

学習アルゴリズムの導出時に必要な計算
(おそらく後で使われる)

$$E_{p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a})}[\pi_k] = \int \pi_k p(\boldsymbol{\pi}|\mathbf{x},\mathbf{a})\, d\boldsymbol{\pi} = \frac{n_k + \alpha_k}{\sum_{k'=1}^{K} n_{k'} + \alpha_{k'}}$$

事後分布の期待値

式(2.6) より

2.2 多項分布とDirichlet分布
α の値による Dirichlet 分布の変化

- 詳細は教科書 p.30 参照
  - 要素の値が小さいと単体の縁周辺の密度が高くなる。
  - 要素の値が大きいと単体の中央周辺の密度が高くなる。
  - 要素間の値に差がある場合、値の大きい要素と k が対応する π の要素値が高い領域の密度が高くなる。
- Dirichlet 分布によって単語の分布の偏りを表現できる。

2.3 LDAの生成過程

- 文書集合の解析では単語の出現分布を分析
- 文書集合の背景に存在するトピックを仮定
  - トピック毎に単語の出現頻度が異なると仮定
  - 文書集合にはトピックの情報は明示的に与えられていない
  - 観測できない潜在トピック(潜在変数)として抽出できるようモデル化

$\phi_{k,v}$ ：トピックk における単語 v の出現確率

$v = \{1,2,3,\cdots,V\}$ ：単語のインデックス集合

$\phi_k = (\phi_{k,1},\cdots,\phi_{k,V})$ ：トピック k における単語の出現分布

$w_{d,i}$ ：文書 d の i 番目の単語

$z_{d,i} \in \{1,2,3,\cdots,K\}$ ：$w_{d,i}$ に対応する潜在変数

2.3 LDAの生成過程
TASA 文書コーパスの LDA 分析例(前半)

文書1

単語	潜在変数
slope	071
music	077
concert	077
play	077
jazz	077
・・・	・・・

文書2

単語	潜在変数
periods	078
audiences	082
play	082
play	082
read	254
・・・	・・・

文書3

単語	潜在変数
game	166
comes	040
Don	180
play	166
boys	020
・・・	・・・

- 「play」は文脈によって意味が異なる多義語
  - 各文書の「paly」に対して異なる潜在変数が付与されていて、「play」の多義性が捉えられていることがわかる。

2.3 LDAの生成過程
TASA 文書コーパスの LDA 分析例(後半)

トピック77 $\phi_{77,v}$

単語	確率
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
・・・	・・・

トピック82 $\phi_{82,v}$

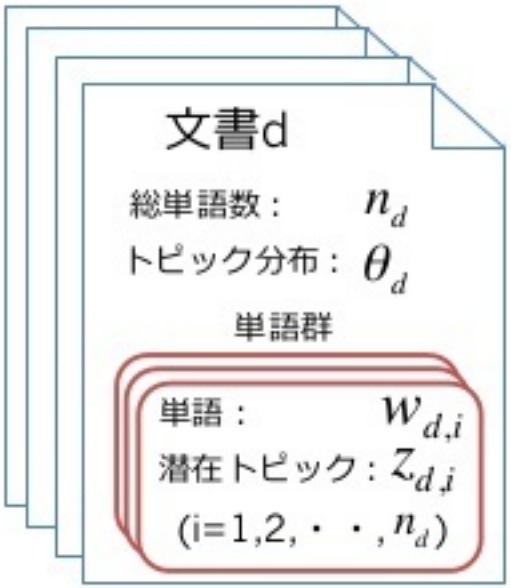
単語	確率
LITERATURE	.031
POEM	.028
・・・	・・・
PLAY	.015
LITERARY	.013
・・・	・・・

トピック166 $\phi_{166,v}$

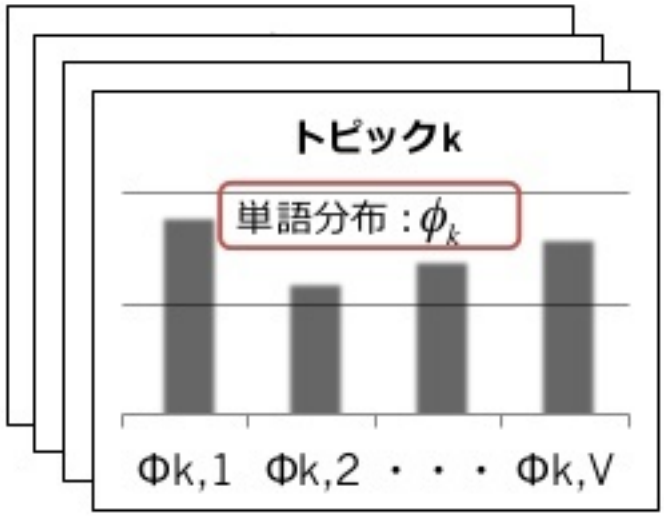
単語	確率
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
・・・	・・・

- 各トピックの意味は、そのトピックにおける単語の出現分布によって特徴付けられる。

2.3 LDAの生成過程
定式化



文書数：M

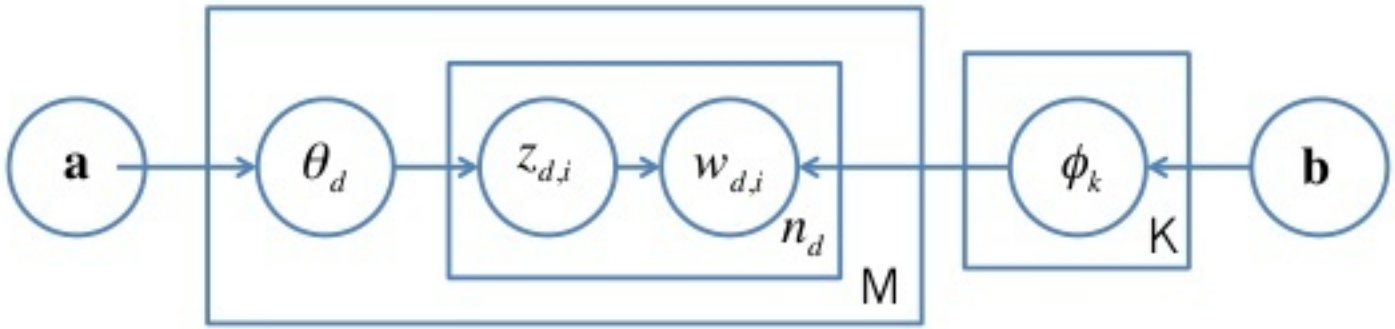


トピック数：K
(総単語数：V)

$$\theta_d = (\theta_{d,1}, \cdots, \theta_{d,K}) \sim Dir(\mathbf{a}) \quad (d=1,2,\cdot\cdot,M) \qquad z_{d,i} \sim Multi(\theta_d)$$

$$\phi_k = (\phi_{k,1}, \cdots, \phi_{k,V}) \sim Dir(\mathbf{b}) \quad (k=1,2,\cdot\cdot,K) \qquad w_{d,i} \sim Multi(\phi_{z_{d,i}})$$

2.3 LDAの生成過程
グラフィカルモデル



ありがとうございました！