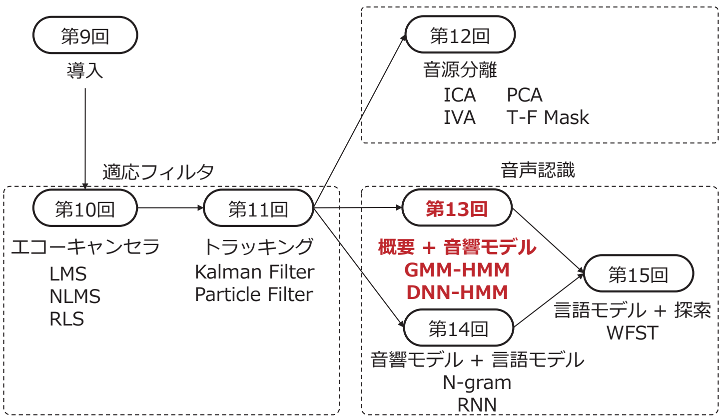
知的情報処理論 第13回

2023年7月11日（火）
武田

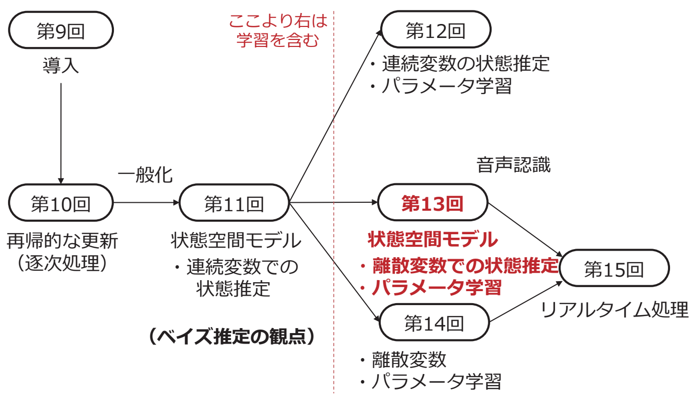
レポート課題

- 8/4 (金) 23:59 提出締め切り
- 詳細は CLE上の pdf を見ること

各回の内容（予定）

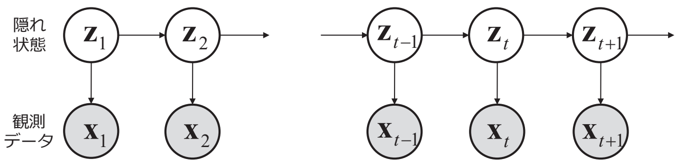


各回の内容（予定）



第11回の内容

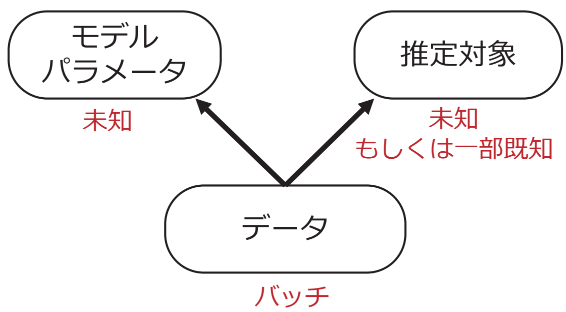
状態空間モデル: state space model



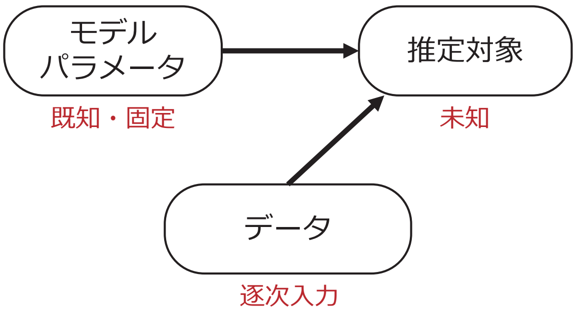
- **「状態」と「観測」の2つの概念**
- **状態に基づいてデータが生成される過程をモデル化**
- **モデルパラメータが既知の元で，隠れ状態を推定**

今回の設定 -- 状態: 離散，観測: 連続

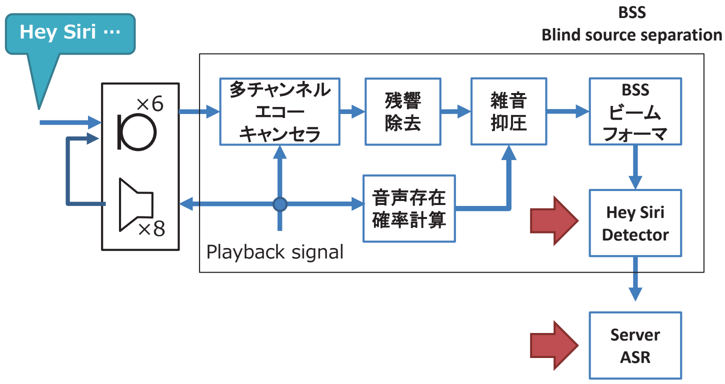
今回の話題: 学習



今回の話題：認識



Apple Home Pod の構成 (SLT2018より)



音声認識

音声信号をテキストに変換

音声認識

音響モデル：(Infinite) GMM

言語モデル：教師なし単語分割

頭の片隅に置いとく方が良い大まかな要素

コスト関数

これが何かわかってないとデバッグできない

- 「良さ」を測る関数
  - 教師あり学習: 二乗誤差, CrossEntropy, etc…
  - 教師なし学習: 尤度, 再構成誤差, etc…
  - 認識・推定: 期待値, 最大事後確率(MAP), etc…
- ロス/目的関数ともいう

モデル

- 入出力の関係や拘束条件を記述
- 方程式や確率モデルで表現
  （コスト関数と一体化 or 切り離せない場合も）

ざっくりとした認識/学習/探索

- コスト関数を{最大化|最小化|良く}する値（集合）を求める手続き（**最適化等の分野**）
- モデル構造や補助変数・関数を利用した手続きも存在

「初期値依存性」 「局所的最適解（局所解）」 「大域的最適解」

本日の内容: 音声認識

第11回 状態空間モデル

0. 補足など
1. 音声認識の概要
2. 信号処理・**特徴抽出**
   ※ これまでの話: 生データに近い
3. 音響モデル I (HMM)

状態: 離散
（記号の世界）
観測: 連続

次回: 音響モデルの続き＋言語モデル

補足など

整理: 前半で少しあった話

識別モデル（逆モデル）

– 事後確率 $p(\mathbf{z}\,|\,\mathbf{x})$ を直接モデル化

(確率的)生成モデル（順モデル）

– ベイズの定理: $p(\mathbf{z}\,|\,\mathbf{x}) \propto p(\mathbf{x}\,|\,\mathbf{z})\,p(\mathbf{z})$

基本的な教師あり学習フェーズ 認識フェーズ

学習: モデルパラメータをチューニング

– 対数尤度基準＋サンプル間で独立の場合

識別モデル

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(\boldsymbol{y}^{(n)}|\boldsymbol{x}^{(n)},\boldsymbol{\theta})$$

生成モデル

$$\hat{\boldsymbol{\theta}}_1 = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(\boldsymbol{x}^{(n)}|\boldsymbol{y}^{(n)},\boldsymbol{\theta}_1)$$

$$\hat{\boldsymbol{\theta}}_2 = \mathrm{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(\boldsymbol{y}^{(n)}|\boldsymbol{\theta}_2)$$

正解付きデータ

$\{(\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})\}_{n=1}^{N}$

パラメータの事前分布
$\ln p(\boldsymbol{\theta})$ を置く場合もある
ガウス → L2ノルム
ラプラス → L1ノルム

認識: 未知入力 $\boldsymbol{x}$ に対する予測

$$\hat{\boldsymbol{y}} = \mathrm{argmax}_{\boldsymbol{y}} \ln p(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{\theta}})$$

$$\hat{\boldsymbol{y}} = \mathrm{argmax}_{\boldsymbol{y}} \ln p(\boldsymbol{x}|\boldsymbol{y},\hat{\boldsymbol{\theta}}_1)p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_2)$$

参考: ノンパラメトリックモデル

データから直接未知入力 x の出力を予測

$$p\left(\boldsymbol{y}\middle|\boldsymbol{x},\{(\boldsymbol{x}^{(n)},\boldsymbol{y}^{(n)})\}_{n=1}^{N}\right)$$

– 1つの方法: パラメータの分布を媒介
– パラメータに関して解析的に周辺化(できるなら)
　→ パラメータ消去が可能
　e.g. パラメータの事後分布で周辺化
　• リッジ回帰+周辺化: ガウス過程
　• ディリクレ分布+周辺化

音声認識の概要

音声認識の応用

音声認識

- 国会の議事録作成支援
- コールセンターの（自動化）モニタリング

音声対話技術

- 対話ロボット
- カーナビ
- 家電（録画予約，．．．）

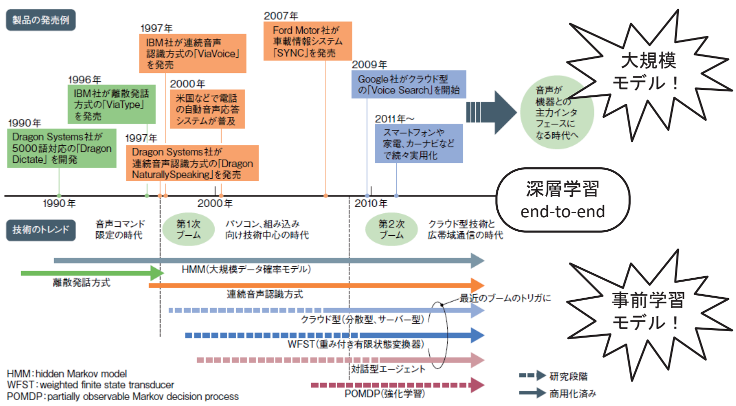
特に「ロボットと話す」はまだまだ発展途上

この30年間の音声認識の進歩

30年前	現在
話者依存	話者非依存
孤立音韻認識，孤立単語認識	連続音声認識
小語彙（数十語）	大語彙（数万語） ⇒ 超大語彙（100万語超）
雑音なし	ノイズ・残響を含んだ音声
オフライン	実時間

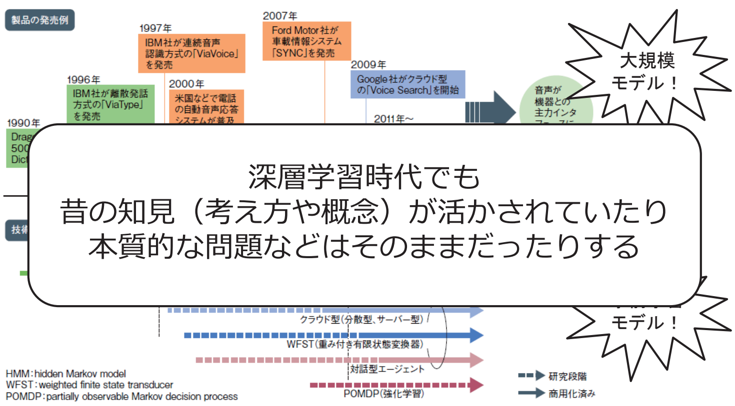
「機械を意識した発声」なら高精度に認識可能
実環境(雑音・残響)での
人間同士の話し言葉の認識は難しい

主に2010年までの技術を取り上げ



日経エレクトロニクス2012年12月24日号 p. 27

主に2010年までの技術を取り上げ



深層学習時代でも
昔の知見（考え方や概念）が活かされていたり
本質的な問題などはそのままだったりする

日経エレクトロニクス2012年12月24日号 p. 27

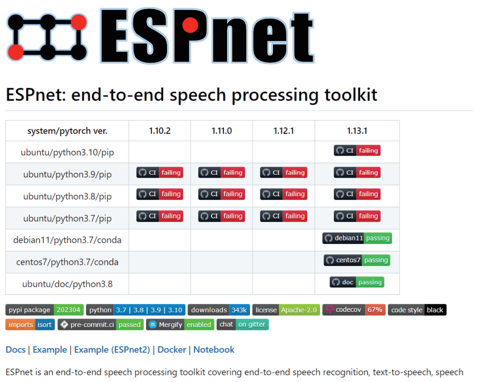
大語彙連続音声認識エンジン Julius

http://julius.sourceforge.jp/

25

End-to-end 音声処理ツール ESPnet

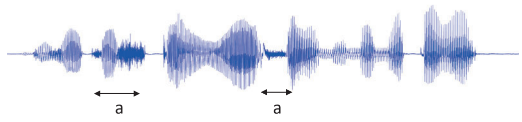
https://github.com/espnet/espnet



26

連続音声認識が難しい理由

1. 入力が区切られていない
2. どの区間が一音素や一文字にあたるか不明
   - **対応関係の伸縮**
   - 草書体の続け文字の認識に近い
   - cf. 郵便番号（数字）の認識：
枠がある／その中には一文字
   - 残響の影響: 画像でいう "ピンボケ", "ブラー"



27

生成モデルに基づく 統計的音声認識の枠組み

認識時の定式化

入力: 系列データ $\mathbf{x}_{1:T} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$

$\mathbf{x}_t$: 時刻 t のデータ/特徴量ベクトル

出力: 事後確率を最大化する文 $\hat{W}$ (シンボル/記号)

$$\hat{W} = \arg\max_W p(W \mid \mathbf{x}_{1:T}) \quad \text{識別モデル}$$

$$= \arg\max_W \underbrace{p(\mathbf{x}_{1:T} \mid W)}_{\text{音響モデル}} \underbrace{p(W)}_{\text{言語モデル}} \quad \text{生成モデル}$$

音響モデル: 候補文W とデータのマッチ度を表すスコア

言語モデル: 候補文W を表すスコア

W を変えてスコアが良いものを選ぶ ＝ 探索

28

具体例

入力波形

どれか？

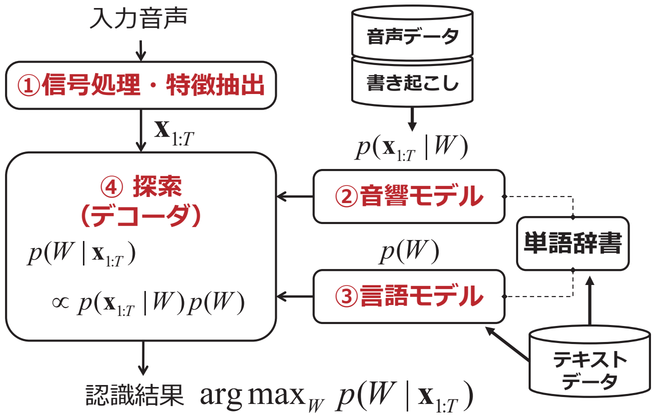
候補	言語モデルスコア $p(W)$	音響モデルスコア $p(\mathbf{X} \mid W)$
明日の天気	**1.0e-3**	**9.4e-6**
今日の天気	**1.0e-3**	***8.1e-8***
アシカの天気	***6.4e-5***	1.9e-6
⋮		

音響・言語モデルの定義が重要

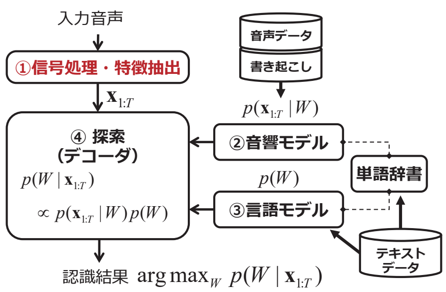
モデルパラメータはデータを用いて事前に学習

29

音声認識の処理フロー



30



①信号処理・特徴抽出

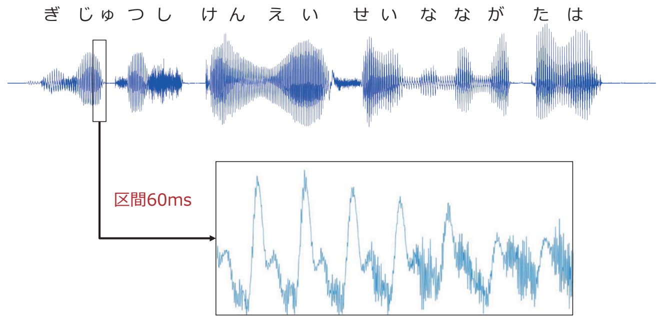
音声信号に含まれる情報

音声認識に用いたい情報
- 音素（発音）を区別する情報
  母音: a i u e o
  子音: k s t n h m y r w, ky, sh, etc…

認識時に影響を吸収したい情報
- 話者に関する情報
  声の高さ，話速, etc…

ひとまず音声信号をよく見てみる

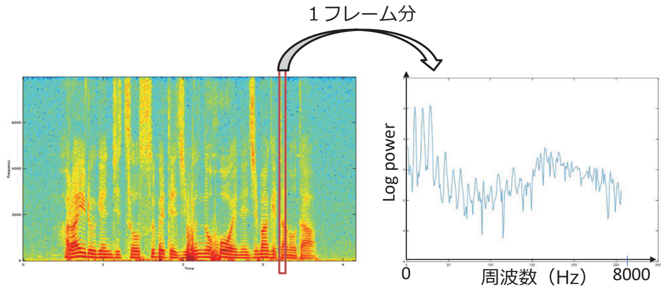
音声信号の拡大



短時間の周期でパターンが変化
→ 短時間周波数解析（STFT）を行う

周波数の時間変化（動画）

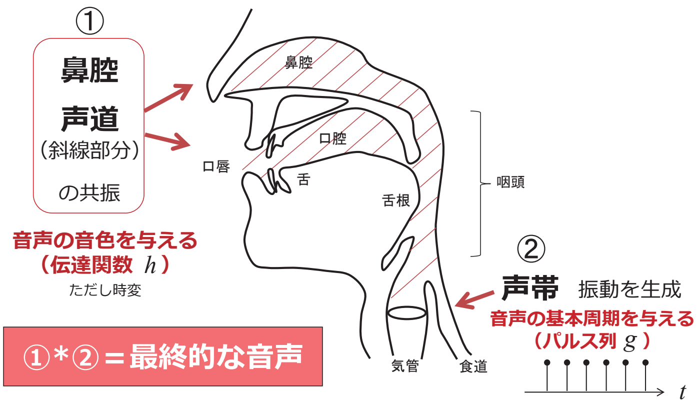
音声のスペクトログラム



ギザギザがうねっている

音声信号の生成過程との関連は？

人間の音声器官の概念図 ＋ 音声生成モデル

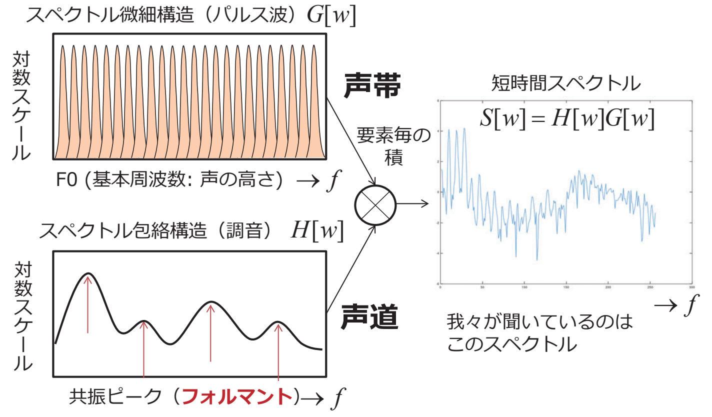


①*②＝最終的な音声

電動式人工喉頭（じんこうこうとう）

https://www.youtube.com/watch?v=4RhS09hXv3U

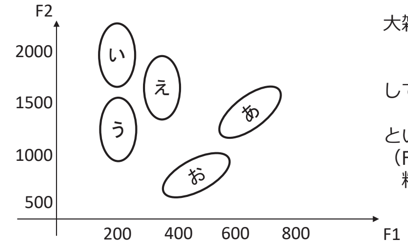
音声のスペクトル構造



フォルマント

スペクトルのピーク

- 低い周波数から第1フォルマント(F1)，第２フォルマント(F2), … と呼ぶ
- 音素を識別するのに有用な情報



大雑把には,
・F1 と F2 を特徴量する
・その分布をあらかじめ学習しておけば,
母音の識別ができる
ということ
（F1, F2の抽出ミスなども含め，精度が良いかは別）

GMM（再掲）

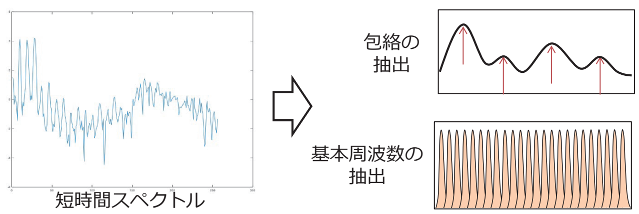
スペクトルからの特徴分離

調音フィルタ（声道の特性）

- 母音や子音などの特徴を含む → 音声認識

基本周波数（声の高さ）

- 話者特徴を含む → 音声合成、韻律分析



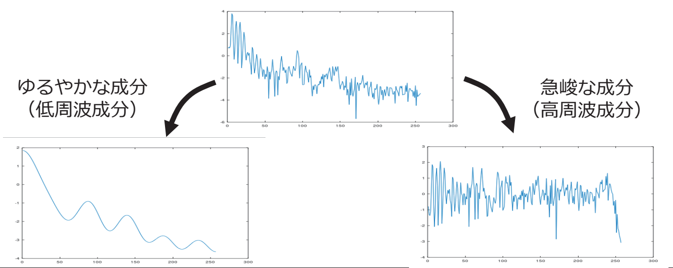
実際の音声特徴量の計算（1）ケプストラム分析

スペクトル包絡と微細構造に分離したい

→ パワースペクトルの周波数分析で分離

包絡: ゆるやかな変化（変化が遅い）

微細構造: 急峻な変化（変化が速い）



実際の音声特徴量の計算（1）ケプストラム分析

1. 音声信号のパワースペクトル

$$|S[w]|=|H[w]||G[w]|$$

2. 対数（要素の足し算）

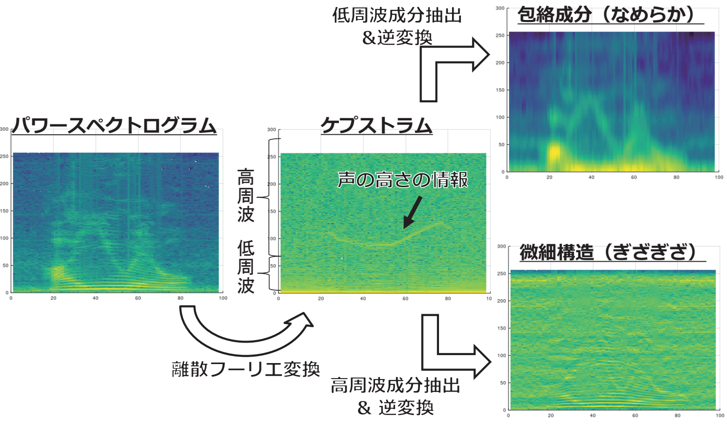
$$\log|S[w]|=\log|H[w]|+\log|G[w]|$$

3. w を時間とみなし離散逆フーリエ変換

値はケプストラム係数と呼ばれる ceps-trum spec-trum

スペクトル包絡の成分 → 低域に集中
微細構造の周期成分 → 高域に集中

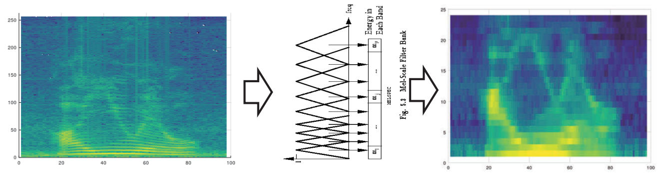
実際の音声特徴量の計算 （1）ケプストラム分析



実際の音声特徴量の計算 （2）メルフィルタバンク

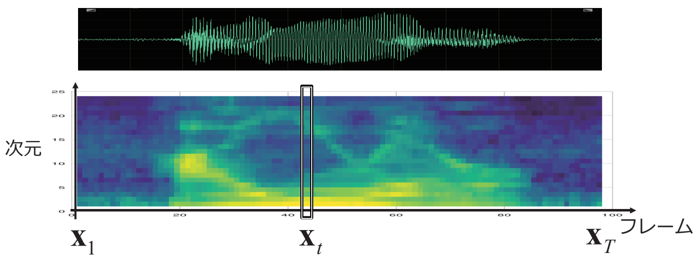
メルフィルタバンク

- メル尺度: 人間の感覚上の音の高さを表す尺度
  - 低い周波数では分解能が高い
- メル尺度上で等間隔になる三角窓を配置
  → パワースペクトルを平滑化 & 低次元化



MFCC: Mel-Frequency Cepstrum Coefficient

音声特徴量のイメージ



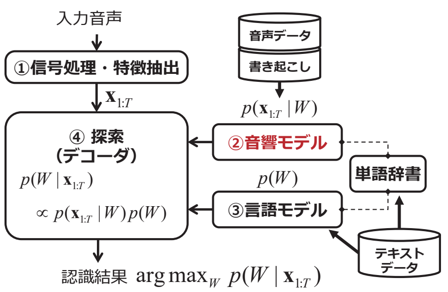
特徴量全体 $\mathbf{x}_{1:T} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$

フレーム間隔 10ms

深層学習ベース: メルフィルタバンク特徴量が主流

Mini Quiz #1

- 話者特有の特徴（話速，声色）などの影響をなるべく受けないように音声認識（文字列・発音）は行いたい．一方で，そのような完全な特徴量を設計するのも難しい．下記に関して適当に考察してみましょう．
  - 話者の特徴も考慮して認識していることはないか
  - 特徴量設計とモデル設計を完全に切り分けることは現実的だろうか？



②音響モデル I (HMM)

話の流れ

前半 – “認識”

- 音響スコア・音声信号の特徴
- 具体的なモデル: 隠れマルコフモデル (HMM)
  = パラメータは既知だと仮定
  Kalman Filter の時と似たような話

※ 演算量などは度外視
  → 実際 – ビタビビームサーチ (第15: リアルタイム処理)

後半 – “学習”: 大枠だけ説明

- パラメータをデータからどう決めるのか
- 状態遷移パラメータ
- 尤度関数のパラメータ（次回）

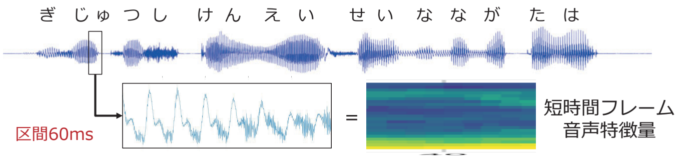
音響モデル

やりたいこと: $p(\mathbf{x}_{1:T} \mid W)$ の定義

– ある文 W が与えられた下での $\mathbf{x}_{1:T}$ の分布

W 既知 ＝「文の発音が既知」として条件付け

– 音声信号: 短時間区間では定常な波形



短時間のパターンなら “クラス化” ができそう

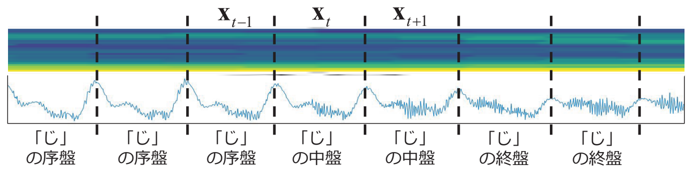
音響モデル

やりたいこと: $p(\mathbf{x}_{1:T} \mid W)$ の定義

– ある文 W が与えられた下での $\mathbf{x}_{1:T}$ の分布

ここでは W 既知 ＝「文の発音が既知」として条件付ける

– 音声信号: 短時間区間では定常な波形



発音と絡めてHMMでモデル化

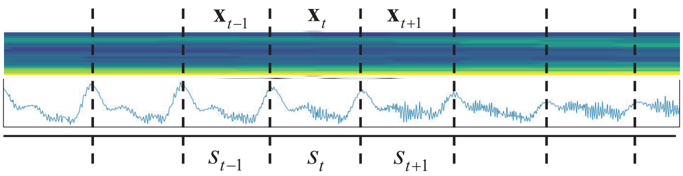
基礎的なモデル: HMM

隠れマルコフモデル(hidden Markov Model; HMM)

– 系列データの生成モデル（状態空間モデル）

– 状態（離散確率変数）s_t とその遷移で表現

状態は直接観測できない＝隠れ/潜在変数



各 $\mathbf{x}_t$ に対し状態（クラス）s_t を割り当て

細かく区分けした「発音ラベル」のようなもの

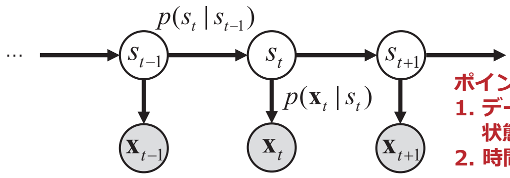
基礎的なモデル: HMM

隠れマルコフモデル(hidden Markov Model; HMM)

– 系列データの生成モデル（状態空間モデル）

– 状態（離散確率変数）s_t とその遷移で表現

状態は直接観測できない＝隠れ/潜在変数



ポイント
1. データ系列と同じ長さの状態系列を設定
2. 時間伸縮(話速)に対応

状態変数列 $s_{1:T} = [s_1, \dots s_T]$ 遷移確率 $p(s_t \mid s_{t-1})$ 尤度 $p(\mathbf{x}_t \mid s_t)$

入力は固定値

基礎的なモデル: HMM

隠れマルコフモデル(hidden Markov Model; HMM)

– 系列データの生成モデル（状態空間モデル）

– 状態（離散確率変数）s_t とその遷移で表現

状態は直接観測できない＝隠れ/潜在変数

$$p(\mathbf{x}_{1:T}) = \sum_{s_{1:T}} p(\mathbf{x}_{1:T} \mid s_{1:T}) p(s_{1:T}) \quad \text{周辺化}$$

系列尤度

$$= \sum_{s_{1:T}} \prod_t p(\mathbf{x}_t \mid s_t) p(s_t \mid s_{t-1}) p(s_1)$$

状態変数列 $s_{1:T} = [s_1, \dots s_T]$ 遷移確率 $p(s_t \mid s_{t-1})$ 尤度 $p(\mathbf{x}_t \mid s_t)$

入力は固定値

HMM に基づく音響モデル

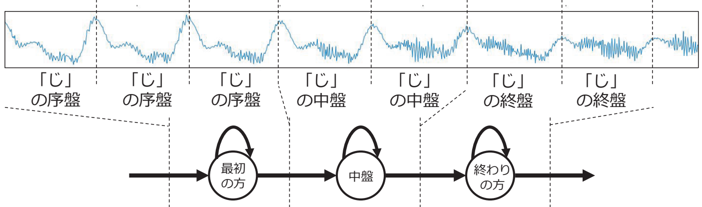
Left-to-right HMM

– 状態遷移図上で一方通行の HMM

– **状態 ＝ 音素（発音）を細分化したクラス**

• 音声信号: 1次元, 音素: 複数フレームにまたがる

• 同じような波形パタン＝同じ状態に留まる

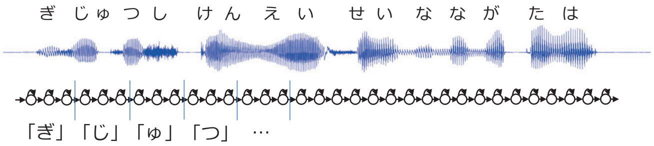


発音「じ」(mono-phone) の 3状態 HMM の状態遷移図

HMM に基づく音響モデル

Left-to-right HMM

- 状態遷移図上で一方通行の HMM
- 状態 = 音素（発音）を細分化したクラス
- 文の HMM は音素 HMM を連結



ある文が与えられたら状態集合が決まる

実質的に発音列のみに状態を依存 → 汎用的

HMM に基づく音響モデル

Left-to-right HMM

- 状態遷移図上で一方通行の HMM
- 状態 = 音素（発音）を細分化したクラス
- 文の HMM は音素 HMM を連結　+パラメータ共有

$$p(\mathbf{x}_{1:T} \mid W) = \sum_{s_{1:T}} p(\mathbf{x}_{1:T} \mid s_{1:T}, W)\, p(s_{1:T} \mid W)$$

$$= \sum_{s_{1:T}} \prod_t \underbrace{p(\mathbf{x}_t \mid s_t, W)}_{\text{音素毎の尤度（発音単位で区別）}}\, \underbrace{p(s_t \mid s_{t-1}, W)}_{\text{音素の状態遷移（自身 or 次状態）}}\, p(s_1 \mid W)$$

以降，式中の W は適当に省略

パラメータ学習後は ある文 Wの音響尤度評価が可能

HMM における $p(\mathbf{x}_{1:T})$ 尤度評価

隠れ状態 s_t の事後分布：再帰推定を利用

Given：状態遷移確率，尤度 （ W は省略）

時刻 t までのデータ $\mathbf{x}_{1:t} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t]$

時刻 t の事後分布　尤度　予測分布

$$p(s_t \mid \mathbf{x}_{1:t}) = \frac{p(\mathbf{x}_t \mid s_t)\, p(s_t \mid \mathbf{x}_{1:t-1})}{\sum_{s_t} p(\mathbf{x}_t \mid s_t)\, p(s_t \mid \mathbf{x}_{1:t-1})}$$

予測分布　状態遷移確率　時刻 t-1 の事後分布

$$p(s_t \mid \mathbf{x}_{1:t-1}) = \sum_{s_{t-1}} p(s_t \mid s_{t-1})\, p(s_{t-1} \mid \mathbf{x}_{1:t-1})$$

「遷移・sum→尤度乗算→正規化」のサイクル

HMM における $p(\mathbf{x}_{1:T})$ 尤度評価

正規化定数を用いた計算

- 離散確率変数：単純な和（not 積分）で計算可

前ページの分母　$$p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}) = \sum_{s_t} p(\mathbf{x}_t \mid s_t)\, p(s_t \mid \mathbf{x}_{1:t-1})$$

- 音響尤度：各時刻の正規化係数の積

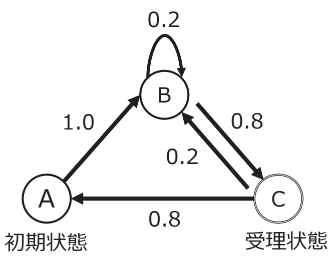
$$\begin{aligned} p(\mathbf{x}_{1:T}) &= p(\mathbf{x}_T \mid \mathbf{x}_{1:T-1})\, p(\mathbf{x}_{1:T-1}) \\ &= p(\mathbf{x}_T \mid \mathbf{x}_{1:T-1})\, p(\mathbf{x}_{T-1} \mid \mathbf{x}_{1:T-2})\, p(\mathbf{x}_{1:T-2}) \\ &= p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1}) \end{aligned}$$

実際はアンダーフロー対策で対数上で計算

補足：状態遷移図による表現

3状態での例

- 状態の関係のみで，時刻は明記されない
  （s_t：{A, B, C} のどれかを取る確率変数）



遷移確率 $p(s_t \mid s_{t-1})$

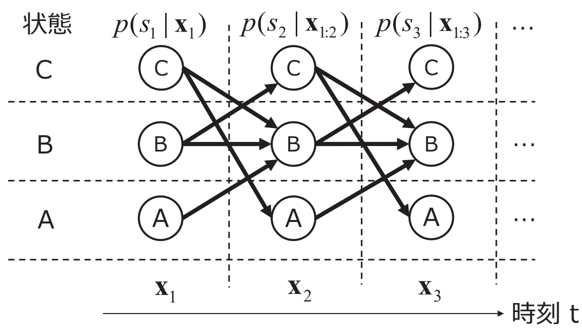
s_{t-1} \ s_t	A	B	C
A	0	1.0	0
B	0	0.2	0.8
C	0.8	0.2	0

（時刻依存なし）

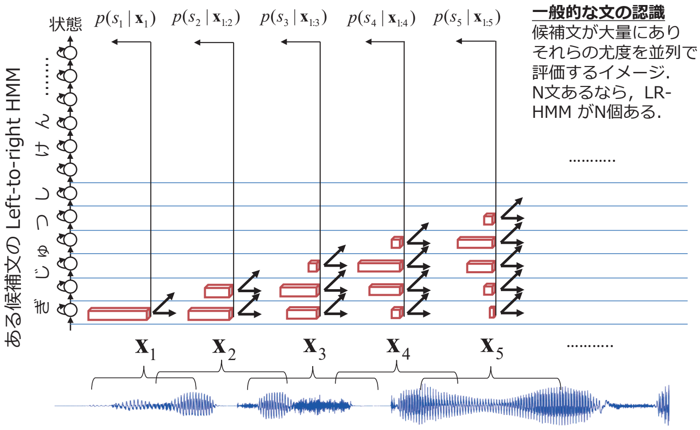
補足：時系列上の表現（トレリス）

3状態での例

- 時間遷移を陽に考慮：推論時のイメージ



Left-to-right HMM における事後分布のイメージ



補足: 「音素」の状態/クラス

調音結合

- 音素の音響的特徴: 周辺の音素の影響で変化
- 例: 「あ」（/a/）
  - 「青い」の「あ」
  - 「間」の「あ」

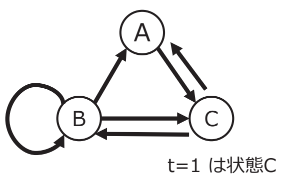
→ 単純な「音素」を状態とすると表現力が弱い

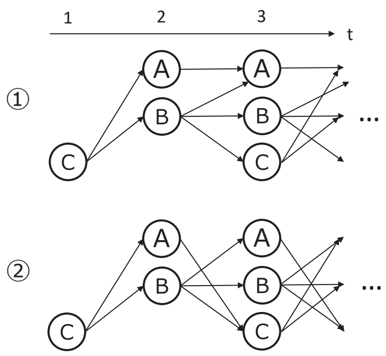
Tri-phone: 音素の3つ組み

- 前後の文脈に依存させたクラス
  - あ+お (a+o), あ-お+い (a-o+i)
- 状態数: 音素数の3乗に比例　　表現力 vs. データ量

Mini Quiz #2

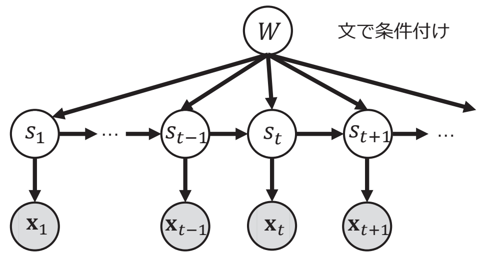
- 左のような状態遷移図に対するトレリス表現はどちらか？





補足: グラフィカルモデル

HMMベースの音響モデル



$$p(\mathbf{x}_{1:T}) = \sum_{s_{1:T}} \prod_t \underbrace{p(\mathbf{x}_t | s_t)}_{\text{尤度関数}} \underbrace{p(s_t | s_{t-1})}_{\text{状態遷移確率}} p(s_1)$$

学習フェーズ

HMM 学習のさわり

学習の前提

- 音声特徴量列 + 書き起こし（音素列）
- 書き起こしデータが given であれば，文毎の HMM の状態集合や状態遷移(図)は定まる

　音素HMMと**尤度関数**のパラメータを学習

HMM におけるパラメータ

- 状態遷移確率と初期状態: $\mathbf{A}$　$\boldsymbol{\pi}$

$$p(s_t = i | s_{t-1} = j) = A_{i,j} \qquad p(s_1 = i) = \pi_i$$

- 尤度関数: $p(\mathbf{x}_t | s_t, \boldsymbol{\theta})$　具体的なモデルは後回し

HMM 学習のさわり

状況の整理

– データ（特徴量列と文HMM構造）は与えられる

ただし，

☹ 各時刻のデータ $\mathbf{x}_t$ がどの状態かはわからない

☹ モデルパラメータ $\mathbf{A}\ \boldsymbol{\pi}\ \boldsymbol{\theta}$ もわからない

cf. Kalman Filter

定義できるのは尤度 $p(\mathbf{x}_{1:T})$ → 最尤推定

– 尤度が最大となるようにパラメータを推定

– 勾配法など通常の方法も使える

HMM 学習のさわり EM アルゴリズム

EM (Expectation Maximization) アルゴリズムを利用 （上界最小化/下界最大化アルゴリズムの一種） majorization minimization algorithm

– 不完全データ（データ $\mathbf{x}$ のみ）に対して，**尤度を単調増加**させるパラメータ更新アルゴリズム

– 「完全データ（潜在変数 $\mathbf{z}$ が隠れてない）の尤度関数最大化は簡単」の場合に有効

- **closed-form のパラメータ更新則**が得られる可能性

sumがややこしい $\log p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta})$

こっちなら計算が簡単 $\log p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta})$

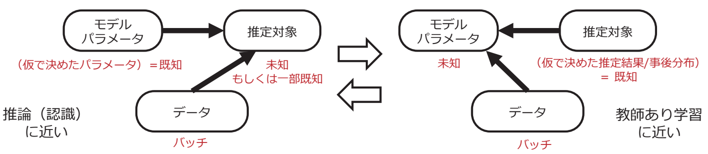
$\mathbf{z}$: HMMだと状態 s

HMM 学習のさわり EM アルゴリズムのイメージ

EM の手続き（ざっくり）

状態推定と**パラメータ更新**を繰り返す

1. 仮のパラメータを用いて状態（=仮の正解ラベルのようなもの）を推定
   - ただし，具体的に1つの値を決めるのではなく，曖昧性を保ったまま推定 = 事後分布を求める
2. 状態推定結果を用いてパラメータ更新



HMM 学習のさわり EM アルゴリズム

EM アルゴリズム

1. パラメータ初期値設定 $\boldsymbol{\theta}^{old}$
2. E-step: $p(\mathbf{z}\,|\,\mathbf{x},\boldsymbol{\theta}^{old})$ を計算: **状態推定**

   完全対数尤度の事後分布による**期待値**を求める

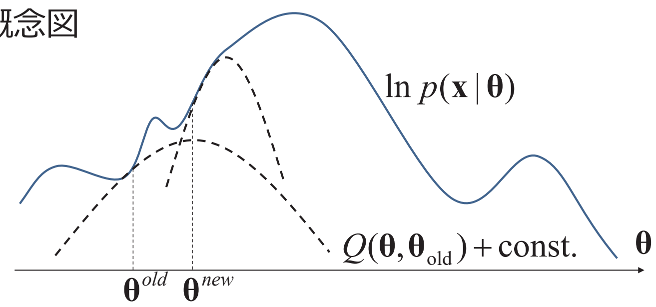
   $$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{\mathbf{z}} p(\mathbf{z}\,|\,\mathbf{x},\boldsymbol{\theta}^{old}) \log p(\mathbf{x},\mathbf{z}\,|\,\boldsymbol{\theta})$$
3. M-step: $\boldsymbol{\theta}^{new}$ を計算: **パラメータ更新**

   $$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$$
4. 収束条件が満たされていなければ step 2へ

※ 何かしらの局所解に収束＝パラメータの初期値が重要

HMM 学習のさわり EM アルゴリズム

概念図



- 解きやすい別の関数（代理関数）を局所的に設定
  - 必ず "同じ" か "下側"にくる必要（下界）
- 良い関数は一般的に問題毎に異なる

HMM 学習の大枠

ある1文に対する状態の事後分布 $p(s_{1:T}\,|\,\mathbf{x}_{1:T},\boldsymbol{\theta}^{old})$

Q関数

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{s_{1:T}} p(s_{1:T}\,|\,\mathbf{x}_{1:T},\boldsymbol{\theta}^{old}) \log p(\mathbf{x}_{1:T},s_{1:T}\,|\,\boldsymbol{\theta})$$

$$= \sum_{s_{1:T}} p(s_{1:T}\,|\,\mathbf{x}_{1:T},\boldsymbol{\theta}^{old})[\log p(\mathbf{x}_{1:T}\,|\,s_{1:T},\boldsymbol{\theta}) + \log p(s_{1:T}\,|\,\mathbf{A},\boldsymbol{\pi})]$$

☺ 尤度と状態遷移のパラメータを分けて計算できる

※1: 尤度関数にも潜在変数が含まれる(入れ子構造の)場合
→ 完全データの尤度関数が変わる点に注意

※2: 大量のデータの場合: 全データの尤度定義からちゃんと行う

73

HMM 学習の大枠

状態遷移関係のパラメータ

$$\sum_{s_{1:T}} p(s_{1:T}\mid \mathbf{x}_{1:T},\boldsymbol{\theta}^{old})\log p(s_{1:T}\mid \mathbf{A})$$

$$=\sum_{s_{1:T}} p(s_{1:T}\mid \mathbf{x}_{1:T},\boldsymbol{\theta}^{old})[\sum_{t>1}\log p(s_t\mid s_{t-1},\mathbf{A})+\log p(s_1)]$$

$$=\sum_{s_1} \underline{p(s_1\mid \mathbf{x}_{1:T},\boldsymbol{\theta}^{old})}\log \pi_{s_1}+\sum_{t>1}\sum_{s_t,s_{t-1}}\underline{p(s_t,s_{t-1}\mid \mathbf{x}_{1:T},\boldsymbol{\theta}^{old})}\log A_{s_t,s_{t-1}}$$

２種類の事後分布が計算に必要

→ フォワードバックワード(Baum-Welch)アルゴリズムで効率的に計算可能

Sum-product algorithm の特別な例

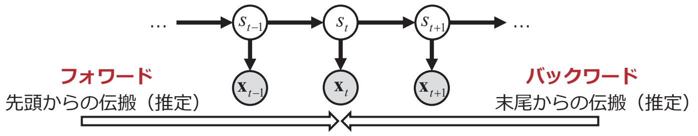
なお，パラメータ $\mathbf{A}$ $\boldsymbol{\pi}$ は拘束条件付きで簡単に解ける

74

HMM 学習の大枠

E-step:

1. 仮のパラメータで事後分布 $p(s_t\mid \mathbf{x}_{1:T})$, $p(s_t,s_{t-1}\mid \mathbf{x}_{1:T})$を求める
   ※フィルタではなくスムージング



2. 完全データの対数尤度の期待値 $\mathrm{E}[\log p(\mathbf{x}_{1:T},s_{1:T})]$ を算出
   各 s_t は（分布として）なんとなく“知っている”

M-step:

$\mathrm{E}[\log p(\mathbf{x}_{1:T},s_{1:T})]$を最大化/良くするようにパラメータ更新.
勾配法でも何でもよい.

75

HMM 学習の大枠

尤度関数のモデル化やパラメータ

– こっちでも事後分布 $p(s_t\mid \mathbf{x}_{1:T},\boldsymbol{\theta}^{old})$ が必要
– 更新ルールは実際のモデルに依存

標準的な2つを取り上げる

1.Gaussian mixture model (GMM)
2.Deep neural network

GMM 単体部分で EMアルゴリズムの具体例も示す

76

補足

EM アルゴリズム

– 基本的には任意の確率モデルに適用可能
  • closed-form の更新則が得られるかは別
  • データセット全体: 全体の同時分布から出発
  • 連続確率変数の場合: sum が 積分にかわるだけ
  • 初期値，局所解の問題はあり
– 概念的に同じような方法
  • majorization minimization algorithm
  • 補助関数法 (auxiliary function)
– 変分ベイズ(Variational Bayes)，Variational Auto Encoder (VAE) との関連

77

EM アルゴリズム

一般的な説明

前提: 観測変数の集合 $\boldsymbol{X}$，潜在変数の集合 $\boldsymbol{Z}$，パラメータ $\boldsymbol{\theta}$，同時分布 $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$

目的: 尤度関数の最大化

$$p(\boldsymbol{X}|\boldsymbol{\theta})=\sum_{\boldsymbol{Z}}p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$$

連続潜在変数を含むなら，その分は積分

想定:

• $p(\boldsymbol{X}|\boldsymbol{\theta})$ の直接的な最適化が困難
• 完全データの対数尤度の最適化は $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ は著しく容易

78

EM アルゴリズム

潜在変数についての分布 $q(\boldsymbol{Z})$ を導入.
すると，分布 $q(\boldsymbol{Z})$ の設定に関わらず以下が成立

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta})=\underline{\mathcal{L}(q,\boldsymbol{\theta})}+\mathrm{KL}(q||p)$$ ← q と事後分布の距離

常に対数尤度の下界 (lower bound)

$$\mathcal{L}(q,\boldsymbol{\theta})=\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}$$

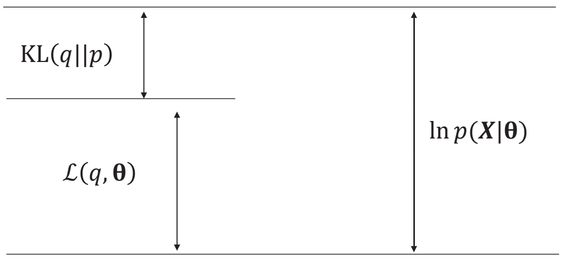
$$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})=\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}|\boldsymbol{\theta})}$$

$$\mathrm{KL}(q||p)=-\sum_{\boldsymbol{Z}}q(\boldsymbol{Z})\ln\frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})}$$

– $\mathcal{L}(q,\boldsymbol{\theta})$: 分布 q の汎関数 +$\boldsymbol{\theta}$ の関数
– $\mathrm{KL}(q||p)$: $q(\boldsymbol{Z})$ と $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ が等しいときに 0

EM アルゴリズム

$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p)$ の図解



$\mathrm{KL}(q||p) \geq 0$ なので，$\mathcal{L}(q, \boldsymbol{\theta})$ は常に対数尤度の下界

EM アルゴリズム

E-step: （パラメータの現在値が $\boldsymbol{\theta}^{\mathrm{old}}$ であるとする）

- 下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ を $\boldsymbol{\theta}^{\mathrm{old}}$ を固定しながら $q(\boldsymbol{Z})$ について最大化する
- $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ は $\boldsymbol{Z}$ に無関係 → $\underbrace{\mathrm{KL}(q||p) = 0}^{\mathcal{L}\text{ の最大化は}}$ と同じ

$q(\boldsymbol{Z})$ が事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ に等しいとき
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ は最大化される（前の図）
→ なのでEMではまず事後分布を計算している

- このとき

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) = \overbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})}}^{\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})}$$

EM アルゴリズム

M-step

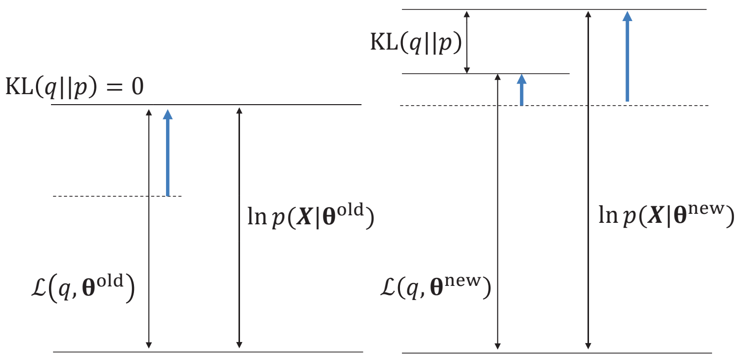
- 分布 $q(\boldsymbol{Z})$ は $\boldsymbol{\theta}^{\mathrm{old}}$ で固定
- 下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を $\boldsymbol{\theta}$ に関して最大化し，新しいパラメータ $\boldsymbol{\theta}^{\mathrm{new}}$ を得る

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})}$$
$$= \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}_{\text{対数同時分布の事後分布による期待値（Q関数）}} + \underbrace{\mathrm{const.}}_{q\text{ のエントロピー（}\boldsymbol{\theta}\text{ と関係なし）}}$$

  - $\mathcal{L}(q, \boldsymbol{\theta})$ を増加 → 対数尤度も増加
  - 分布 q: $\boldsymbol{\theta}^{\mathrm{old}}$ と $\boldsymbol{\theta}^{\mathrm{new}}$ で分布は一致しない

EM アルゴリズム

E-step **M-step**



参考文献

- C.M.ビショップ「パターン認識と機械学習 上下」
- 金森敬文 他「機械学習のための連続最適化」
- 持橋大地，大羽成征「ガウス過程と機械学習」
- 石井健一郎・上田修功「続・わかりやすいパターン認識　教師なし学習入門」
- 河原達也　「音声認識システム」
- 安藤彰男　「リアルタイム音声認識」