Ch.4 近似Bayes推論

- $\mathcal{X}$：観測データ.

  $\mathcal{Z}$：非観測変数（パラメータ，潜在変数など）

〈Bayes推論の流れ〉

1. 確率モデル $p(\mathcal{X}, \mathcal{Z})$ の設計.
2. 事後分布 $p(\mathcal{Z}|\mathcal{X})$ の計算，パラメータ更新（学習）
3. 予測分布の計算（予測）

- Ch.3で扱ったモデルは単純だったので，事後分布 $p(\mathcal{Z}|\mathcal{X})$ が厳密に得られた.

→ 複雑なモデルでは，厳密に $p(\mathcal{Z}|\mathcal{X})$ を得るのは困難！

→ 近似手法を用いて何とかしたい.

§4.1 サンプリングに基づく推論手法.

- $p(\mathcal{Z}|\mathcal{X})$ を求める代わりに，$p(\mathcal{Z}|\mathcal{X})$ から複数のサンプルを得ることで事後分布の特性を調べることにする.

4.1.1 単純モンテカルロ法.

〈問題設定〉

$$\mathbb{E}_{p(\mathbf{z})}[f(\mathbf{z})] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$$ を求める.

[仮定]
- 積分計算は困難.
- 分布 $p(\mathbf{z})$ からのサンプリングができる.

サンプリングが容易なことは少ない.

Th. (大数の強法則)

独立に同分布に従う

$\{X_i\}$: d-次元 i.i.d. 確率変数列. $\mathbb{E}[X_i] = \mu$.

$\Rightarrow P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1$. ← 確率1で標本平均は期待値に収束. $\square$

・この定理を利用して積分を計算するのが

単純モンテカルロ法 (simple Monte Carlo method):

$z^{(1)}, \dots, z^{(T)} \overset{i.i.d.}{\sim} p(z)$ とT個サンプリングして,

$\int f(z)p(z)\,dz \approx \frac{1}{T}\sum_{t=1}^{T} f(z^{(t)})$ と近似する.

→ 大数の法則により 標本平均 $\frac{1}{T}\sum_{t=1}^{T} f(z^{(t)})$ は

期待値 $\mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)\,dz$ に収束するので, 確率1で.

Tを十分大きくとるようにする.

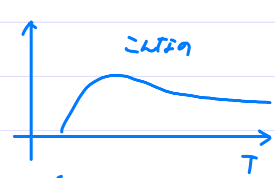
ex) N個の観測データ $\mathcal{X} = \{x_1, \dots, x_N\}$: given.

パラメータ θ をもつモデル $p(\mathcal{X}, \theta) = p(\mathcal{X}|\theta)p(\theta)$.

周辺尤度 $p(\mathcal{X}) = \mathbb{E}_{p(\theta)}[p(\mathcal{X}|\theta)] = \int p(\mathcal{X}|\theta)p(\theta)\,d\theta$ を評価.

$p(\mathcal{X}) = \int p(\mathcal{X}|\theta)p(\theta)\,d\theta \approx \frac{1}{T}\sum_{t=1}^{T} p(\mathcal{X}|\theta^{(t)})$. $\left(\theta^{(t)} \overset{i.i.d.}{\sim} p(\theta)\right)$

と近似すればよい.



収束オーダーは θの次元には依っていないので 高次元であっても問題ない

→ ただし, この近似は収束が遅め.

(重複対数の法則より, 収束のオーダーは $O\left(\sqrt{\frac{\log\log T}{T}}\right)$.)

$X_i \overset{i.i.d.}{\sim} p$, $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \sigma^2 < \infty$, $S_n = \sum_{i=1}^{n} X_i$

$\Rightarrow P\left(\limsup_{n\to\infty} \frac{S_n - n\mu}{\sqrt{2n\log\log n}} = \sigma\right) = 1$, $P\left(\liminf_{n\to\infty} \frac{S_n - n\mu}{\sqrt{2n\log\log n}} = -\sigma\right) = 1$.

4.1.2 棄却サンプリング.

棄却サンプリング (rejection sampling):

単純モンテカルロ法のサンプリングに利用できる.

密度の計算が困難な目標分布 $p(\mathbf{z})$ からのサンプルを得る方法

次の定理に基づく.

<u>Th.</u> (棄却サンプリング)

$p(\mathbf{z}), q(\mathbf{z})$: pdf.

$kq(\mathbf{z})$ を $p(\mathbf{z})$ の包絡函数という

$k>0$: 定数を $kq(\mathbf{z}) \geq p(\mathbf{z}) \quad (^\forall \mathbf{z})$ となるように定める.

$\mathbf{z} \sim q(\mathbf{z})$ と $u \sim Uni(0,1)$ は独立とする.

$\Rightarrow u \leq \dfrac{p(\mathbf{z})}{kq(\mathbf{z})}$ のもとでの $\mathbf{z}$ の条件付き pdf は $p(\mathbf{z})$ となる. $\square$

<u>pf.</u> Bayes の定理より,

$$q\left(\mathbf{z} \,\middle|\, u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})}\right) = \frac{P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})} \,\middle|\, \mathbf{z}\right) q(\mathbf{z})}{P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})}\right)}.$$

ここで,

Unif(0,1) の pdf.

$$P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})} \,\middle|\, \mathbf{z}\right) = \int_0^{\frac{p(\mathbf{z})}{kq(\mathbf{z})}} 1\, du = \frac{p(\mathbf{z})}{kq(\mathbf{z})}.$$

$$P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})}\right) = \int P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})} \,\middle|\, \mathbf{z}\right) q(\mathbf{z})\, d\mathbf{z} = \int \frac{p(\mathbf{z})}{k\cancel{q(\mathbf{z})}} \cancel{q(\mathbf{z})}\, d\mathbf{z}$$

$$= \frac{1}{k} \underbrace{\int p(\mathbf{z})\, d\mathbf{z}}_{=1} = \frac{1}{k}.$$

$$\therefore q\left(\mathbf{z} \,\middle|\, u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})}\right) = \frac{p(\mathbf{z})/k}{1/k} = p(\mathbf{z}).$$ ▨

→ 提案分布 (proposal distribution) $q(\mathbf{z})$ をサンプリングしやすい分布にしておくことで, $p(\mathbf{z})$ のサンプルを次のように得ることができる.

Algo. (棄却サンプリング)

Input：目標分布 $p(\mathbf{z})$, 提案分布 $q(\mathbf{z})$, 定数 k : $kq(\mathbf{z}) \geq p(\mathbf{z})$.

$T \in \mathbb{Z}_{>0}$

Output：$\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)}\}$: $p(\mathbf{z})$からのサンプル.

1 for $t = 1, 2, \dots, T$ do:

2 　repeat

3 　　$\mathbf{z} \sim q(\mathbf{z})$;　$u \sim \mathrm{Uni}(0,1)$

4 　until $u \leq \dfrac{p(\mathbf{z})}{kq(\mathbf{z})}$;　← この条件がみたされるまで、サンプル $\mathbf{z}$ を棄却しリサンプリング.

5 　$\mathbf{z}^{(t)} \leftarrow \mathbf{z}$　← サンプル $\mathbf{z}$ を受容.

6 end for;

7 return $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)}\}$.　□

・目標分布 $p(\mathbf{z})$ の正規化定数が分からなくてもサンプリングできる.

$$p(\mathbf{z}) = \frac{1}{Z_p}\tilde{p}(\mathbf{z}), \quad Z_p = \int \tilde{p}(\mathbf{z})d\mathbf{z}$$

計算が難しいことがある

のとき, $kq(\mathbf{z}) \geq \frac{1}{Z_p}\tilde{p}(\mathbf{z}) \Leftrightarrow kZ_p q(\mathbf{z}) \geq \tilde{p}(\mathbf{z})$ なので、改めて $k \leftarrow kZ_p$ とすれば $p(\mathbf{z})$ のかわりに $\tilde{p}(\mathbf{z})$ を使っても $p(\mathbf{z})$ からのサンプリングができる.

・棄却サンプリングの効率：2〜4行目の繰り返し回数による.

受容確率は $P\left(u \leq \frac{p(\mathbf{z})}{kq(\mathbf{z})}\right) = \frac{1}{k}$. 各回の試行は独立.

→ 受容されるまでの繰り返し回数は幾何分布 $Ge(\frac{1}{k})$ に従う.

($Ge(p)$ の pmf : $p(n) = p(1-p)^{n-1}$.)

繰り返し回数の期待値は k.

$p(\boldsymbol{z})$ を近似できるようなもの

・ k はできるだけ小さい方がよい. そうなるように, $q(\boldsymbol{z})$ をうまく選ぶとよいが

そういった $q(\boldsymbol{z})$ を見つけるのは容易ではない.

・ $\boldsymbol{z}$ の次元が大きくなってくると, k が大きくなる傾向にあり効率が悪くなる.

(次元の呪い)

cf.) 効率を良くするために, $p(\boldsymbol{z})$ を区分的に近似する $q(\boldsymbol{z})$ を使った

適応的棄却サンプリングといったものもある.

4.1.3 自己正規化重点サンプリング

・ 期待値 $\mathbb{E}_{p(\boldsymbol{z})}[f(\boldsymbol{z})] = \int f(\boldsymbol{z})p(\boldsymbol{z})\,d\boldsymbol{z}$ を計算するのに,

単純MCでは $\int f(\boldsymbol{z})p(\boldsymbol{z})\,d\boldsymbol{z} \approx \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{z}^{(t)})$, $\boldsymbol{z}^{(t)} \sim p(\boldsymbol{z})$

と $p(\boldsymbol{z})$ からのサンプリングができる必要があった.

・ $p(\boldsymbol{z})$ からのサンプリングが容易でないときは, 上述の棄却サンプリングを

利用できた. サンプリングされた $\boldsymbol{z} \sim q(\boldsymbol{z})$ に対して

・ $p(\boldsymbol{z})$ が小さいとき $\boldsymbol{z}$ は棄却されやすいので, $p(\boldsymbol{z})$ の大きいサンプルが得られやすい.

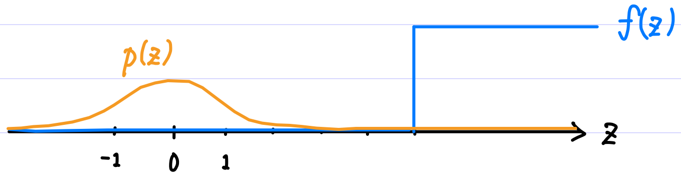
・ $|f(\boldsymbol{z})|$ が小さいとき 和 $\sum_{t=1}^{T} f(\boldsymbol{z}^{(t)})$ への寄与が小さいため,

$\mathbb{E}_{p(\boldsymbol{z})}[f(\boldsymbol{z})]$ への収束に T をかなり大きくする必要がある.

ex) 極端だが、$p(z) = \mathcal{N}(z|0,1)$,

$$f(z) = I(\{z \geq 5\}) = \begin{cases} 1 & (z \geq 5) \\ 0 & (z < 5) \end{cases}$$

とすると下図のような状況になる.



積分 $\int f(z)p(z)dz \neq 0$ だが、 2.8665×10⁻⁷ ぐらい

$z \geq 5$ のサンプルを得る確率がほぼ 0 のため、$z^{(t)} \sim p(z)$ としても

$\frac{1}{T}\sum_{t=1}^{T} f(z^{(t)})$ はぴったり 0 になってしまうことがほとんど.

(近似はできているが …)

$p(z)$ が大きくなる z は高次元空間のごく一部.
そこでの $|f(z)|$ が小さいととても計算効率が悪くなる.

以上のような問題は、高次元だとさらに悪化する(次元の呪い).

→ $|f(z)|p(z)$ の大きい領域を重点的にサンプリングすることで、

計算効率を上げる(Tをなるべく小さくする)ことができると良い.

◦ 重点サンプリング (importance sampling).

サンプリングが容易にできる提案分布 $q(z)$ を考える.

$f(z)p(z) \neq 0 \Rightarrow q(z) > 0$ と仮定. 計算したい期待値は,

$\mathrm{supp}(q) \supseteq \mathrm{supp}(f \cdot p)$.
($\mathrm{supp}(f) := \{z \mid f(z) \neq 0\}$: f の台 (support))

$$\mathbb{E}_{p(z)}[f(z)]$$

$$= \int f(z)p(z)\,dz$$ ← 積分は $f(z)q(z) \neq 0$ のところを考えれば十分.
このとき 仮定より $q(z) > 0$.

$$= \int f(z)\frac{p(z)}{q(z)}q(z)\,dz.$$

ここで，$w(\mathbb{z}) := \frac{p(\mathbb{z})}{q(\mathbb{z})}$ （重要度重み，importance weight）とおくと，

$$\mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})] = \int w(\mathbb{z}) f(\mathbb{z}) q(\mathbb{z}) d\mathbb{z} = \mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z}) f(\mathbb{z})].$$

よって，単純MCのように

$$\mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})] \approx \frac{1}{T}\sum_{t=1}^{T} w(\mathbb{z}^{(t)}) f(\mathbb{z}^{(t)}) \quad \left(\mathbb{z}^{(t)} \overset{i.i.d.}{\sim} q(\mathbb{z})\right)$$

と近似することで $T \to \infty$ のとき右辺は $\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z}) f(\mathbb{z})] = \mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})]$

almost surely

に収束する．

- 近似値の分散について考えてみる．

$I := \mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})]$，$\hat{I} := \frac{1}{T}\sum_{t=1}^{T} w(\mathbb{z}^{(t)}) f(\mathbb{z}^{(t)})$ とおく

$\hat{I}$はIの不偏推定量

$\mathbb{E}_{q(\mathbb{z})}[\hat{I}] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z}) f(\mathbb{z})] = I$ なので，

$$
\begin{aligned}
&\mathbb{V}_{q(\mathbb{z})}[\hat{I}] \\
&= \mathbb{E}_{q(\mathbb{z})}[\hat{I}^2] - I^2 \\
&= \frac{1}{T^2}\mathbb{E}_{q(\mathbb{z})}\left[\sum_{t=1}^{T} w(\mathbb{z}^{(t)})^2 f(\mathbb{z}^{(t)})^2 + 2\sum_{s<t} w(\mathbb{z}^{(s)}) f(\mathbb{z}^{(s)}) w(\mathbb{z}^{(t)}) f(\mathbb{z}^{(t)})\right] - I^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] + \frac{2}{T^2}\sum_{s<t}\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z}) f(\mathbb{z})]^2 - I^2 \\
&= \frac{1}{T}\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] + \frac{2}{T^2}\cdot\frac{T(T-1)}{2} I^2 - I^2 \\
&= \frac{1}{T}\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] - \frac{1}{T} I^2 \\
&= \frac{1}{T}\left(\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] - I^2\right). \quad \left(= \frac{1}{T}\mathbb{V}_{q(\mathbb{z})}[w(\mathbb{z}) f(\mathbb{z})]\right)
\end{aligned}
$$

- $\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] < \infty$ でないと，近似値 $\hat{I}$ の分散が ∞ になる．

$w(\mathbb{z})$の分母は$q(\mathbb{z})$なので，$q(\mathbb{z})$が$p(\mathbb{z})$よりはやく0に近づくと$w(\mathbb{z})$が

大きくなり分散が発散するおそれがある．

→ $q(\mathbb{z})$ は $p(\mathbb{z})$ より "裾の厚い" ものを選ぶとよい．

・$\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2] < \infty$ のとき，Jensenの不等式より，

$I \subseteq \mathbb{R}$：区間，$h: I \to \mathbb{R}$：凸函数．
X：I-値確率変数．$\mathbb{E}[X] < \infty$．
$\Rightarrow \mathbb{E}[h(X)] \geq h(\mathbb{E}[X])$．
等号成立は $X = \mathbb{E}[X]$, a.s. のとき

$$\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})^2 f(\mathbb{z})^2]$$
$$\geq \left(\mathbb{E}_{q(\mathbb{z})}[w(\mathbb{z})|f(\mathbb{z})|]\right)^2.$$

等号成立は，$|f(\mathbb{z})|w(\mathbb{z}) = \mathbb{E}_{q(\mathbb{z})}[|f(\mathbb{z})|w(\mathbb{z})]$, a.s.

$\Leftrightarrow q(\mathbb{z}) \propto |f(\mathbb{z})|w(\mathbb{z})$, a.s. のとき．

→ q で $|f(\mathbb{z})|w(\mathbb{z})$ の大きい範囲を重点的にサンプリングすると良い

ことが確かめられた

◦ 自己正規化重点サンプリング (self-normalized importance sampling).

上では p, q とも pdf としてきた．ここでは，

$$p(\mathbb{z}) = \frac{1}{Z_p}\tilde{p}(\mathbb{z}), \quad q(\mathbb{z}) = \frac{1}{Z_q}\tilde{q}(\mathbb{z})$$

として，$\tilde{p}(\mathbb{z}), \tilde{q}(\mathbb{z})$ のみが計算できる状況を考える．

（もちろん $q(\mathbb{z})$ からのサンプリングはできるものとする）

$$\mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})]$$
$$= \int f(\mathbb{z})\frac{p(\mathbb{z})}{q(\mathbb{z})}q(\mathbb{z})d\mathbb{z}.$$
$$= \frac{Z_q}{Z_p}\int f(\mathbb{z})\frac{\tilde{p}(\mathbb{z})}{\tilde{q}(\mathbb{z})}q(\mathbb{z})d\mathbb{z}.$$

ここで，$\tilde{w}(\mathbb{z}) := \dfrac{\tilde{p}(\mathbb{z})}{\tilde{q}(\mathbb{z})}$ とすると，

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(\mathbb{z})d\mathbb{z} = \frac{1}{Z_q}\int \frac{\tilde{p}(\mathbb{z})}{\tilde{q}(\mathbb{z})}\tilde{q}(\mathbb{z})d\mathbb{z}$$
$$= \int \tilde{w}(\mathbb{z})\, q(\mathbb{z})\, d\mathbb{z} = \mathbb{E}_{q(\mathbb{z})}[\tilde{w}(\mathbb{z})]$$

なので，

$$\mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})] = \frac{\mathbb{E}_{q(\mathbb{z})}[\tilde{w}(\mathbb{z})f(\mathbb{z})]}{\mathbb{E}_{q(\mathbb{z})}[\tilde{w}(\mathbb{z})]}.$$

よって，T個のサンプル $\mathbb{z}^{(1)}, \dots, \mathbb{z}^{(T)} \overset{i.i.d}{\sim} q(\mathbb{z})$ を用いて

$$\mathbb{E}_{q(\mathbb{z})}[\tilde{w}(\mathbb{z})f(\mathbb{z})] \approx \frac{1}{T}\sum_{t=1}^{T}\tilde{w}(\mathbb{z}^{(t)})f(\mathbb{z}^{(t)}),$$
$$\mathbb{E}_{q(\mathbb{z})}[\tilde{w}(\mathbb{z})] \approx \frac{1}{T}\sum_{t=1}^{T}\tilde{w}(\mathbb{z}^{(t)})$$

と近似することにする，i.e.,

$$\mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})] \approx \frac{\cancel{\frac{1}{T}}\sum_{t=1}^{T}\tilde{w}(\mathbb{z}^{(t)})f(\mathbb{z}^{(t)})}{\cancel{\frac{1}{T}}\sum_{t=1}^{T}\tilde{w}(\mathbb{z}^{(t)})} = \sum_{t=1}^{T}\underbrace{\frac{\tilde{w}(\mathbb{z}^{(t)})}{\sum_{t'=1}^{T}\tilde{w}(\mathbb{z}^{(t')})}}_{=: w^*(\mathbb{z}^{(t)}) \text{ とおく}} f(\mathbb{z}^{(t)})$$

と近似する.

- $\sum_{t=1}^{T} w^*(\mathbb{z}^{(t)}) = 1$ なので"自己正規化".

Algo. (自己正規化重点サンプリング")

Input: 目標分布 p の主要項 $\tilde{p}(\mathbb{z})$, 提案分布 $q(\mathbb{z}) = \frac{1}{Z_q}\tilde{q}(\mathbb{z})$,

　　関数 $f(\mathbb{z})$, $T \in \mathbb{Z}_{>0}$.

Output: 期待値 $I = \mathbb{E}_{p(\mathbb{z})}[f(\mathbb{z})]$ の近似値 $\hat{I}$.

1. for $t = 1, \dots, T$ do:
2. 　$\mathbb{z}^{(t)} \sim q(\mathbb{z})$; $\tilde{w}(\mathbb{z}^{(t)}) \leftarrow \dfrac{\tilde{p}(\mathbb{z}^{(t)})}{\tilde{q}(\mathbb{z}^{(t)})}$
3. end for;

4. $\mathcal{S} \leftarrow \sum_{t=1}^{T} \tilde{w}(\boldsymbol{z}^{(t)})$;

5. for $t = 1, \dots, T$ do:

6. $\quad w^*(\boldsymbol{z}^{(t)}) \leftarrow \frac{1}{\mathcal{S}} \tilde{w}(\boldsymbol{z}^{(t)})$

7. end for;

8. return $\sum_{t=1}^{T} w^*(\boldsymbol{z}^{(t)}) f(\boldsymbol{z}^{(t)})$ as $\hat{I}$.

(実装の際は $\tilde{w}(\boldsymbol{z}^{(t)})$ や $f(\boldsymbol{z}^{(t)})$ の値を保存していなくても、$\tilde{w}(\boldsymbol{z}^{(t)})$ や $f(\boldsymbol{z}^{(t)})\tilde{w}(\boldsymbol{z}^{(t)})$ をサンプル生成の都度足しこんでいけばよい。)

Remark (1) $\mathbb{E}_{q(\boldsymbol{z})}\left[\sum_{t=1}^{T} w^*(\boldsymbol{z}^{(t)}) f(\boldsymbol{z}^{(t)})\right] \neq I$ であり、$\sum_{t=1}^{T} w^*(\boldsymbol{z}^{(t)}) f(\boldsymbol{z}^{(t)})$ は I の不偏推定量ではない しかし、$\mathbb{E}_{q(\boldsymbol{z})}\left[\sum_{t=1}^{T} w^*(\boldsymbol{z}^{(t)}) f(\boldsymbol{z}^{(t)})\right] \xrightarrow[T\to\infty]{} I$, a.s. であり、漸近不偏推定量になっている

(2) $\hat{I}$ の分散が最小になるのは $q(\boldsymbol{z}) \propto |f(\boldsymbol{z}) - I| p(\boldsymbol{z})$ のとき (Hesterberg, 1988)

(3) 棄却サンプリングは重点サンプリングの特別な場合とみなせる。

cf.) 古澄『ベイズ計算統計学』

(4) 重点サンプリングでは、$\boldsymbol{z}^{(t)} \overset{i.i.d.}{\sim} q(\boldsymbol{z})$ としていたので、$p(\boldsymbol{z})$ からのサンプルは得られないが、サンプリング-重点リサンプリング (sampling/importance resampling : SIR) を利用すると近似的に $p(\boldsymbol{z})$ に従ったサンプルを得ることはできる。

cf.) Bishop : PRML, 古澄『ベイズ計算統計学』など。

4.1.4. Markov連鎖モンテカルロ法

・Markov連鎖モンテカルロ法 (Markov Chain Monte Carlo : MCMC) :

高次元の分布 $p(\boldsymbol{z})$ から効率的にサンプリングを行うための方法.

Def. (Markov連鎖)

$\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$: 確率変数列.

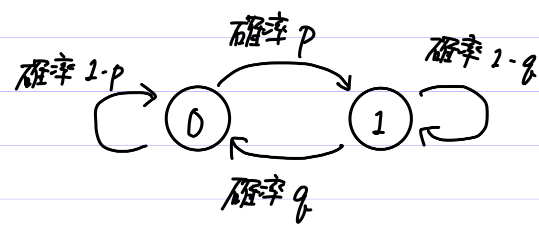
一つ前の"状態"にしか影響されない.

$$p(\boldsymbol{z}^{(t)} \mid \boldsymbol{z}^{(t-1)}, \dots, \boldsymbol{z}^{(1)}) = p(\boldsymbol{z}^{(t)} \mid \boldsymbol{z}^{(t-1)})$$

が成り立つとき, $\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ を (1次の) Markov連鎖 という.

・$T(\boldsymbol{\zeta}, \boldsymbol{z}) := p(\boldsymbol{z} \mid \boldsymbol{\zeta})$: 遷移確率密度 という.

ex) (有限状態の Markov連鎖)



状態 0, 1 があって,

右図のように状態が遷移していく.

$z^{(t)} = 0$ のときは $p(z^{(t+1)} \mid z^{(t)}) = \mathrm{Bern}(z^{(t+1)} \mid p)$,

$z^{(t)} = 1$ のときは $p(z^{(t+1)} \mid z^{(t)}) = \mathrm{Bern}(z^{(t+1)} \mid 1-q)$

$z^{(t+1)}$ は $z^{(t)}$ のみに依存. ⇒ 1次の Markov連鎖.

$T(z^{(t)}, z^{(t+1)}) = \mathrm{Bern}(z^{(t+1)} \mid p^{1-z^{(t)}}(1-q)^{z^{(t)}})$ と表せる.

ふつうは行列の形で $\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ と表す

遷移確率行列

ex) $\boldsymbol{z}^{(i+1)} = \boldsymbol{z}^{(i)} + \boldsymbol{\varepsilon}_i$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, I_2)$

$\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ は Markov連鎖.

$T(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(i+1)}) = \mathcal{N}(\boldsymbol{z}^{(i+1)} \mid \boldsymbol{z}^{(i)}, I_2)$.

〈やりたいこと〉

$\mathcal{T}(\xi, \mathbf{z})$ をうまく設計して、サンプリングしたい分布 $p(\mathbf{z})$ に従うような

Markov連鎖 $\{\mathbf{z}^{(i)}\}_{i=1}^{\infty}$, $\mathbf{z}^{(i)} \sim p(\mathbf{z})$ を得たい。

$\mathbf{z}^{(i)}$ たちは同分布に従うが、独立ではないことに注意。

・ $\mathbf{z}^{(i)} \sim p(\mathbf{z}) \Rightarrow \mathbf{z}^{(i+1)} \sim p(\mathbf{z})$ となるような分布を"定常分布"という：

Def. (定常分布)

確率分布 p が、全ての $\mathbf{z}$ に対して

$$p(\mathbf{z}) = \int p(\xi)\mathcal{T}(\xi, \mathbf{z})\, d\xi$$

が成り立つとき、p を定常分布 (stationary distribution) という。

・定常分布はいつでも存在するとは限らないし、存在しても一意とは限らない。

定常分布が存在するための十分条件として、次がある。

Th. (詳細釣り合い条件)

分布 p と遷移確率密度 $\mathcal{T}$ が、全ての $\mathbf{z}, \xi$ に対して

$$p(\mathbf{z})\mathcal{T}(\mathbf{z}, \xi) = p(\xi)\mathcal{T}(\xi, \mathbf{z})$$

マスター方程式から出てくる式。

← $\mathbf{z}$ から ξ に移る"確率"と ξ から $\mathbf{z}$ に移る"確率"が等しいということ。

が満たされるとき、p は Markov連鎖の定常分布。

pf. 両辺を ξ について積分。

$$\int p(\mathbf{z})\mathcal{T}(\mathbf{z}, \xi)\, d\xi = p(\mathbf{z})\underbrace{\int \mathcal{T}(\mathbf{z}, \xi)\, d\xi}_{=1} = p(\mathbf{z}).$$

$$\therefore\ p(\mathbf{z}) = \int p(\xi)\mathcal{T}(\xi, \mathbf{z})\, d\xi$$

∎

・MCMCでは $\mathcal{T}$ が詳細釣合い条件をみたすようになっている。

これができていれば苦労しない……

・ p が定常になる $\mathcal{T}$ が設計できたとしても、初期値 $\boldsymbol{z}^{(1)}$ が $p(\boldsymbol{z})$ に従うようにサンプリングできなければ、p に従う変数からなる Markov連鎖 $\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ は得られない。

どのぐらい後かはものによる

遷移をくり返した後の"極限分布"が定常分布と一致する、ということ。

どんな $\boldsymbol{z}^{(1)}$ から始めても、十分ステップ後の変数 $\boldsymbol{z}^{(T)}$ が $p(\boldsymbol{z})$ に従うようになってくれればよい。($\boldsymbol{z}^{(T)}$ を改めて $\boldsymbol{z}^{(1)}$ とおけば p に従う変数からなる Markov連鎖 $\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ が得られることになる (burn-in))

さらに独立な変数をサンプルしたければ、これから間引きすればよい。

→ エルゴード性 (ergodicity)

- 既約 : どの変数 $\boldsymbol{z}$ も、任意の $\boldsymbol{z}'$ へ有限回で遷移可能。
- 正再帰的 : どの変数 $\boldsymbol{z}$ も、有限回の遷移で戻ってこれる。
- 非周期的 : どの変数 $\boldsymbol{z}$ も、周期的に現れない

⇒ 定常分布の存在と一意性

な Markov連鎖のことをエルゴード的という。

Fact. (1) エルゴード的な Markov連鎖 $\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ は唯一つの定常分布 p をもつ。

(2) 任意の初期値 $\boldsymbol{z}^{(1)}$ に対し、$\boldsymbol{z}^{(i)}$ の従う分布は $i \to \infty$ の極限で p に法則収束する。

(3) $\mathbb{E}_{p(\boldsymbol{z})}[f(\boldsymbol{z})] < \infty$ のとき、

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{z}^{(i)}) = \mathbb{E}_{p(\boldsymbol{z})}[f(\boldsymbol{z})]$$ ← これの計算をする分には burn-in とかやる必要はない

が成立する (Markov連鎖に対するエルゴード定理)

→ MCMC では基本的にエルゴード的な Markov連鎖になってくれる!

4.1.5 Metropolis-Hastings法.

- $\mathcal{T}$をうまく設計するために、次のように仮定してみる：

$$\mathcal{T}(\boldsymbol{\zeta}, \boldsymbol{z}) = A(\boldsymbol{\zeta}, \boldsymbol{z}) q(\boldsymbol{z}|\boldsymbol{\zeta})$$

- $q(\boldsymbol{z}|\boldsymbol{\zeta})$：遷移確率密度，提案分布.
- $A(\boldsymbol{\zeta}, \boldsymbol{z}) \in [0, 1]$：採択率

→ $\boldsymbol{\zeta}$から$\boldsymbol{z}$へと仮に遷移したうち，確率$A(\boldsymbol{\zeta}, \boldsymbol{z})$で$\boldsymbol{z}$を採択し，

確率$1 - A(\boldsymbol{\zeta}, \boldsymbol{z})$で$\boldsymbol{z}$を棄却して$\boldsymbol{\zeta}$に戻る．ということ．

このとき，詳細釣合い式は，$\forall \boldsymbol{z}, \boldsymbol{\zeta}$で

$$p(\boldsymbol{z}) A(\boldsymbol{z}, \boldsymbol{\zeta}) q(\boldsymbol{\zeta}|\boldsymbol{z}) = p(\boldsymbol{\zeta}) A(\boldsymbol{\zeta}, \boldsymbol{z}) q(\boldsymbol{z}|\boldsymbol{\zeta})$$

$$\Leftrightarrow \frac{A(\boldsymbol{\zeta}, \boldsymbol{z})}{A(\boldsymbol{z}, \boldsymbol{\zeta})} = \frac{p(\boldsymbol{z}) q(\boldsymbol{\zeta}|\boldsymbol{z})}{p(\boldsymbol{\zeta}) q(\boldsymbol{z}|\boldsymbol{\zeta})} =: \rho(\boldsymbol{\zeta}, \boldsymbol{z}).$$

Lem. 任意の$p(\boldsymbol{z})$, $q(\boldsymbol{z}|\boldsymbol{\zeta})$に対し，次の$A(\boldsymbol{\zeta}, \boldsymbol{z})$は詳細釣合い式をみたす：

$$A(\boldsymbol{\zeta}, \boldsymbol{z}) = \min\left(1, \rho(\boldsymbol{\zeta}, \boldsymbol{z})\right)$$

pf. $\rho(\boldsymbol{z}, \boldsymbol{\zeta}) = \rho(\boldsymbol{\zeta}, \boldsymbol{z})^{-1}$に注意する．

$$\frac{A(\boldsymbol{\zeta}, \boldsymbol{z})}{A(\boldsymbol{z}, \boldsymbol{\zeta})} = \frac{\min(1, \rho(\boldsymbol{\zeta}, \boldsymbol{z}))}{\min(1, \rho(\boldsymbol{\zeta}, \boldsymbol{z})^{-1})} = \frac{\min(\rho(\boldsymbol{\zeta}, \boldsymbol{z}), \rho(\boldsymbol{\zeta}, \boldsymbol{z})^2)}{\min(\rho(\boldsymbol{\zeta}, \boldsymbol{z}), 1)}$$

$$= \begin{cases} \dfrac{\rho(\boldsymbol{\zeta}, \boldsymbol{z})}{1} & (\rho(\boldsymbol{\zeta}, \boldsymbol{z}) \geq 1) \\ \dfrac{\rho(\boldsymbol{\zeta}, \boldsymbol{z})^2}{\rho(\boldsymbol{\zeta}, \boldsymbol{z})} & (\rho(\boldsymbol{\zeta}, \boldsymbol{z}) < 1) \end{cases}$$

$$= \rho(\boldsymbol{\zeta}, \boldsymbol{z})$$

$\blacksquare$

→ $q(\mathbb{z}|\zeta)$ をうまく選べば, $\mathcal{T}(\zeta,\mathbb{z}) = q(\mathbb{z}|\zeta)\min\left(1, \rho(\zeta,\mathbb{z})\right)$

でMarkov連鎖のエルゴード性が保証される.

$q(\mathbb{z}|\zeta) = \mathcal{N}(\mathbb{z}|\zeta, I_d)$ などとしておけばOK.

cf.) Robert and Casella "Monte Carlo Statistical Methods".

<u>Algo.</u> (Metropolis-Hastings)

Input: 目標分布 $p(\mathbb{z})$, 提案分布 $q(\mathbb{z}|\zeta)$. 繰り返し回数 $N \in \mathbb{Z}_{>0}$

Output: Markov連鎖 $\{\mathbb{z}^{(i)}\}_{i=1}^{N}$

```
z^(1) をランダムに初期化;
for i = 1, ..., N-1 do:
    z* ~ q(z | z^(i));   u ~ Unif(0, 1);
    r ← ρ(z^(i), z*);
    if u ≤ min(1, r):
        z^(i+1) ← z*          ← 採択
    else:
        z^(i+1) ← z^(i)       ← 棄却
return {z^(i)}_{i=1}^N .
```

Remark p のかわりに, 正規化されていない $\tilde{p}$ を用いても良い ($\rho(\zeta,\mathbb{z})$ は等しい)

$q(\mathbb{z}|\zeta) = q(\zeta|\mathbb{z})$ (対称) のとき Metropolis法 という.

4.1.6 ハミルトニアン・モンテカルロ法

- q を Gauss分布とした Metropolis-Hastings法：
  - 分散を大きくする
    - ✓ $z^{(t)}$ と $z^{(t+1)}$ の相関 低.　1回の遷移での平均移動キョリが大きくなるから
    - ✗ 目標分布から外れて棄却率大.
  - 分散を小さくする
    - ✓ 採択率 大
    - ✗ $z^{(t)}$ と $z^{(t+1)}$ の相関 高 ⇒ 目標分布からのサンプルを得るまで時間がかかる.

→ 高次元だと分散の調整は困難. 特に採択率の低下が著しくなる.
高次元空間で平均移動距離が長くとれて, 採択率を高くすることはできるのか…?

- ハミルトニアン モンテカルロ法 / ハイブリッド・モンテカルロ法
(Hamiltonian / Hybrid Monte Carlo : HMC)：
解析力学的な物体の軌道のシミュレーション + Metropolis-Hastings.
→ 効率のよいサンプリングができる.

4.1.6.1 Hamiltonian のシミュレーション.

- 少しだけ物理の復習.

・質量 $m=1$ の物体を考える.

$\boldsymbol{z}=\boldsymbol{z}(\tau)$: 時刻 τ での位置ベクトル

$\boldsymbol{p}=\boldsymbol{p}(\tau)=\frac{d}{d\tau}\boldsymbol{z}$: 時刻 τ での運動量ベクトル

$\boldsymbol{z}$ で微分できるとする

・$\mathcal{U}(\boldsymbol{z})$: 位置 $\boldsymbol{z}$ における物体の位置エネルギー（ポテンシャル）

$\mathcal{K}(\boldsymbol{p})=\frac{1}{2}\boldsymbol{p}^T\boldsymbol{p}$: 運動量 $\boldsymbol{p}$ のときの物体の運動エネルギー.

・今考える運動では, ポテンシャルによる力のみが働くとする.

$\rightarrow \underbrace{\frac{d}{d\tau}\boldsymbol{p}}_{\text{加速度}}=\underbrace{-\frac{\partial}{\partial \boldsymbol{z}}\mathcal{U}(\boldsymbol{z})}_{\text{力}}$: ポテンシャルの下での運動方程式.

$\mathcal{H}(\boldsymbol{z},\boldsymbol{p}):=\mathcal{U}(\boldsymbol{z})+\mathcal{K}(\boldsymbol{p})$: 系の全エネルギー（Hamiltonian）

・$(\boldsymbol{z},\boldsymbol{p})\in\mathbb{R}^d\times\mathbb{R}^d$ と考えて, 空間 $\mathbb{R}^d\times\mathbb{R}^d$ を相空間 (phase space) という.

Lem. (Hamiltonの運動方程式)

$$\frac{d}{d\tau}\boldsymbol{z}=\frac{\partial}{\partial\boldsymbol{p}}\mathcal{H}(\boldsymbol{z},\boldsymbol{p}),\quad \frac{d}{d\tau}\boldsymbol{p}=-\frac{\partial}{\partial\boldsymbol{z}}\mathcal{H}(\boldsymbol{z},\boldsymbol{p}).$$

pf.
$$\frac{\partial}{\partial\boldsymbol{p}}\mathcal{H}(\boldsymbol{z},\boldsymbol{p})=\frac{\partial}{\partial\boldsymbol{p}}\mathcal{K}(\boldsymbol{p})=\boldsymbol{p}=\frac{d}{d\tau}\boldsymbol{z}.$$
$$\frac{\partial}{\partial\boldsymbol{z}}\mathcal{H}(\boldsymbol{z},\boldsymbol{p})=\frac{\partial}{\partial\boldsymbol{z}}\mathcal{U}(\boldsymbol{z})=-\frac{d}{d\tau}\boldsymbol{p}.$$
□

〈ポイント〉

1. Hamiltonianは保存する. 実際,

$$\frac{d\mathcal{H}}{d\tau}=\sum_{i=1}^{d}\left(\frac{\partial\mathcal{H}}{\partial z_i}\frac{dz_i}{d\tau}+\frac{\partial\mathcal{H}}{\partial p_i}\frac{dp_i}{d\tau}\right)=\sum_{i=1}^{d}\left(\frac{\partial\mathcal{H}}{\partial z_i}\frac{\partial\mathcal{H}}{\partial p_i}-\frac{\partial\mathcal{H}}{\partial p_i}\frac{\partial\mathcal{H}}{\partial z_i}\right)=0.$$

$\rightarrow$ 物体は $\mathcal{H}$ を一定に保ったまま運動する.

相空間内の運動の軌道は, Hamiltonの運動方程式を解けば求まる.

2. 可逆性 (reversibility) :

$T_\sigma(\mathbb{z}(\tau), \mathbb{p}(\tau)) := (\mathbb{z}(\tau+\sigma), \mathbb{p}(\tau+\sigma))$ とすると,

変換 T_σ は逆変換 $T_\sigma^{-1} = T_{-\sigma}$ をもつ.

3. 体積保存 (volume preservation) :

相空間内の領域 $D \subseteq \mathbb{R}^d \times \mathbb{R}^d$ について, D の体積と

$T_\sigma(D) = \{ T_\sigma(\mathbb{z}, \mathbb{p}) \mid (\mathbb{z}, \mathbb{p}) \in D \}$ の体積は等しい (Liouville の定理).

(i.e., $\mathbb{v} = \left(\frac{d\mathbb{z}}{d\tau}, \frac{d\mathbb{p}}{d\tau} \right)$ とすると $\operatorname{div} \mathbb{v} = 0$.)

領域の体積拡大率

→ T_σ で $(\mathbb{z}, \mathbb{p})$ を変数変換しても, 変換の Jacobian は 1 になる.

・ T_σ による変換を求めるには, Hamilton の運動方程式

$$\frac{d}{d\tau}\mathbb{z} = \frac{\partial}{\partial \mathbb{p}}\mathcal{H}(\mathbb{z}, \mathbb{p}), \quad \frac{d}{d\tau}\mathbb{p} = -\frac{\partial}{\partial \mathbb{z}}\mathcal{H}(\mathbb{z}, \mathbb{p})$$

を解けばよい. 数値的に解く.

$\varepsilon > 0$: 微小時間幅. $\mathbb{z}^\tau = (\mathbb{z}(\tau)$ の計算値$)$ などのように書く

◦ (前進) Euler法では $\frac{d}{d\tau}\mathbb{z} \approx \frac{\mathbb{z}(\tau+\varepsilon) - \mathbb{z}(\tau)}{\varepsilon}$ などと近似して,

$$\mathbb{z}^{\tau+\varepsilon} = \mathbb{z}^\tau + \varepsilon \frac{\partial}{\partial \mathbb{p}}\mathcal{H}(\mathbb{z}, \mathbb{p})\Big|_{\mathbb{z}=\mathbb{z}^\tau, \mathbb{p}=\mathbb{p}^\tau} = \mathbb{z}^\tau + \varepsilon \underbrace{\frac{d}{d\tau}\mathbb{z}}_{=\mathbb{p}(\tau)}\Big|_{\mathbb{z}=\mathbb{z}^\tau, \mathbb{p}=\mathbb{p}^\tau}$$

$$= \mathbb{z}^\tau + \varepsilon \mathbb{p}^\tau$$

$$\mathbb{p}^{\tau+\varepsilon} = \mathbb{p}^\tau - \varepsilon \frac{\partial}{\partial \mathbb{z}}\mathcal{H}(\mathbb{z}, \mathbb{p})\Big|_{\mathbb{z}=\mathbb{z}^\tau, \mathbb{p}=\mathbb{p}^\tau} = \mathbb{p}^\tau + \varepsilon \frac{d}{d\tau}\mathbb{p}\Big|_{\mathbb{z}=\mathbb{z}^\tau, \mathbb{p}=\mathbb{p}^\tau}$$

$$= \mathbb{p}^\tau - \varepsilon \frac{\partial}{\partial \mathbb{z}}\mathcal{U}(\mathbb{z})\Big|_{\mathbb{z}=\mathbb{z}^\tau}$$

というスキームで ε 時刻先の相空間での座標を計算.

誤差が $O(\varepsilon)$

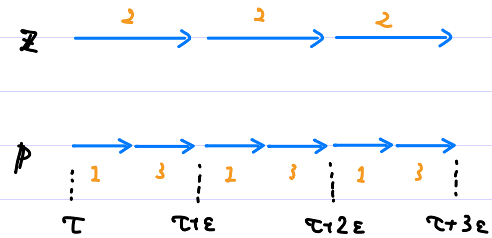
→ Euler法は離散化の数値誤差が大きい（1次精度）

・リープフロッグ法では、運動量の離散化を $\frac{\varepsilon}{2}$ 刻みにしておく。

1. $$p^{\tau+\frac{\varepsilon}{2}} = p^{\tau} - \frac{\varepsilon}{2}\frac{\partial}{\partial z}U(z)\Big|_{z=z^{\tau}}$$

2. $$z^{\tau+\varepsilon} = z^{\tau} + \varepsilon p^{\tau+\frac{\varepsilon}{2}}$$ ← zの計算でpの $\tau, \tau+\varepsilon$での平均を使う感じ。

3. $$p^{\tau+\varepsilon} = p^{\tau+\frac{\varepsilon}{2}} - \frac{\varepsilon}{2}\frac{\partial}{\partial z}U(z)\Big|_{z=z^{\tau+\varepsilon}}$$



$L \in \mathbb{Z}_{>0}$ ：ステップ数として、上の更新をL回行い、時刻 εL 先の物体の位置と運動量 $(z^{\tau+\varepsilon L}, p^{\tau+\varepsilon L})$ を計算する。

→数値誤差がEuler法よりも小さくなる（2次精度）

また、このスキームは可逆性があり、$(z^{\tau+\varepsilon L}, p^{\tau+\varepsilon L})$ から $-\tau$ 方向にリープフロッグ法をL回行うと、もとの (z^{τ}, p^{τ}) が得られる。

さらに、シンプレクティック性を有しており、$(z^{\tau}, p^{\tau}) \mapsto (z^{\tau+\varepsilon L}, p^{\tau+\varepsilon L})$ の変換で Hamiltonian が保存される。 ← もちろん誤差はでるが、振動するだけ

cf.) シンプレクティック積分法 (Symplectic Integrator)

・T_{σ} のシミュレートをする上では、リープフロッグ法を利用するのが良い。

4.1.6.2 サンプリングアルゴリズムへの適用。

・サンプリングしたい分布を $p(z) = \frac{1}{Z}\tilde{p}(z)$ とする。 $(z \in \mathbb{R}^d)$

$p \in \mathbb{R}^d$ を導入して、同時分布を $p(z,p) = p(z)p(p)$ とする。

（p と z は独立だと仮定している）

(z,p) を $p(z,p)$ からサンプリングすることで，z は p と無関係に $p(z)$ から サンプリングされることになる．

・ $p(p) = \mathcal{N}(p|0, I_d) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2} p^T p\right)$ とする．

$p(z) = \frac{1}{Z} \tilde{p}(z) = \frac{1}{Z} \exp(-\mathcal{U}(z))$ と表す． ← $\tilde{p}(z) > 0$ なのでこのように表せる．

同時分布は，

$$p(z,p) = \frac{1}{Z}\frac{1}{\sqrt{(2\pi)^d}} \exp\Big(-\mathcal{U}(z) - \underbrace{\frac{1}{2} p^T p}_{= \mathcal{K}(p)}\Big) \propto \exp(-\mathcal{H}(z,p))$$

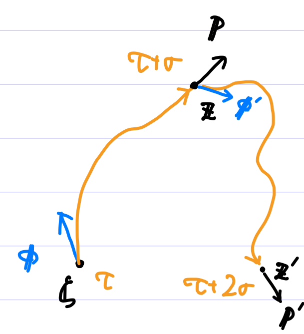
となる．

・ Metropolis-Hastings法では，提案分布 $q(z|\zeta)$ からのサンプリングをしていた．

Hamiltonian Monte Carlo では，代わりに ϕ を $\mathcal{N}(\phi|0, I_d)$ からサンプリングし，

(ζ, ϕ) の位置から σ だけ時間発展させて得られる

$(z, p) = T_\sigma(\zeta, \phi)$ への遷移を考える． 本当はこれの leapfrog 離散近似を考えている．



T_σ の体積保存性から，推移確率密度が

$$\mathcal{T}((\zeta,\phi),(z,p)) = \exp(-\mathcal{H}(z,p) + \mathcal{H}(\zeta,\phi))$$

と表せ，更に可逆性より同様に

$$\mathcal{T}((z,p),(\zeta,\phi)) = \exp(-\mathcal{H}(\zeta,\phi) + \mathcal{H}(z,p))$$

も成り立つ．すると，

$$p(z,p)\mathcal{T}((z,p),(\zeta,\phi)) = \exp(-\mathcal{H}(\zeta,\phi)),$$

$$p(\zeta,\phi)\mathcal{T}((\zeta,\phi),(z,p)) = \exp(-\mathcal{H}(z,p))$$

Hamiltonian は時間発展で不変なので,

$$p(z,p)\mathcal{T}((z,p),(\xi,\phi)) = p(\xi,\phi)\mathcal{T}((\xi,\phi),(z,p))$$

と詳細釣合い条件が成立し, $p(z,p)$ はこの Markov 連鎖の定常分布.

- 実際は T_σ を適用するかわりに leapfrog によるシミュレートを行うので, 数値誤差のために Hamiltonian は完全に保存されるわけではない.

→ Metropolis 法での採択率

$$A((\xi,\phi),(z,p)) := \min\left(1, \frac{p(z,p)}{p(\xi,\phi)}\right) = \min\left(1, \exp(-\mathcal{H}(z,p)+\mathcal{H}(\xi,\phi))\right)$$

を利用する. 数値誤差があっても, leapfrog 法では $\mathcal{H}(z,p) \simeq \mathcal{H}(\xi,\phi)$

となるので, 採択率は 1 に近くなる!

Algo. (Hamiltonian Monte Carlo)

Input : $\tilde{p}(z)$, $\varepsilon > 0$, $L \in \mathbb{Z}_{>0}$, $N \in \mathbb{Z}_{>0}$

Output : Markov 連鎖 $\{z^{(i)}\}_{i=1}^N$

1. $z^{(1)}$ をランダムに初期化 ;
2. for $i = 1, \dots, N-1$ do:
3. $\quad p \sim \mathcal{N}(p \mid 0, I_d)$; $u \sim \mathrm{Unif}(0,1)$;
4. $\quad (z^*, p^*) \leftarrow \mathrm{leapfrog}_{\varepsilon,L}(z^{(i)}, p)$; $r \leftarrow \dfrac{p(z^*,p^*)}{p(z^{(i)},p)}$;
5. $\quad z^{(i+1)} \leftarrow z^*$ if $u \le \min(1,r)$ else $z^{(i)}$
6. return $\{z^{(i)}\}_{i=1}^N$.

4.1.6.3 Langevin動力学法.

- Langevin動力学法 (Langevin dynamics method):

$L=1$ とした HMC.

leapfrog の式をまとめて,

$$\mathbb{Z}^{\tau+\varepsilon} = \mathbb{Z}^{\tau} + \varepsilon p^{\tau+\frac{\varepsilon}{2}} = \mathbb{Z}^{\tau} + \varepsilon \left(p^{\tau} - \frac{\varepsilon}{2} \frac{\partial}{\partial \mathbb{Z}} \mathcal{U}(\mathbb{Z}) \Big|_{\mathbb{Z}=\mathbb{Z}^{\tau}} \right)$$

$$= \mathbb{Z}^{\tau} - \frac{\varepsilon^2}{2} \frac{\partial}{\partial \mathbb{Z}} \mathcal{U}(\mathbb{Z}) \Big|_{\mathbb{Z}=\mathbb{Z}^{\tau}} + \varepsilon p^{\tau}.$$

により $\mathbb{Z}^*$ を計算する. $\mathbb{Z}^*$ をそのまま採択してしまうことで計算を簡便にする.

- 少しランダムウォーク的な挙動がある.

4.1.7 Gibbsサンプリング.

- 分布 $p(\mathbb{Z})$ から直接 $\mathbb{Z}$ 全体をサンプリングするのが難しいとき,

$\mathbb{Z} = (\mathbb{Z}_1, \ldots, \mathbb{Z}_M)$ と M個に分けて, 次のように逐次的に

Gibbs分布からサンプリングするのに使われた.

サンプリングする (Gibbsサンプリング)

Algo. (Gibbsサンプリング)

Input: $p(\mathbb{Z}) = p(\mathbb{Z}_1, \ldots, \mathbb{Z}_M)$, $N \in \mathbb{Z}_{>0}$

Output: サンプル $\{\mathbb{Z}^{(i)}\}_{i=1}^{N}$

$\mathbb{Z}^{(1)}$: ランダムに初期化;

for $i = 1, \ldots, N-1$ do:

　for $m = 1, \ldots, M$ do:

4. $$\mathbb{Z}_m^{(i+1)} \sim p(\mathbb{Z}_m \mid \mathbb{Z}_1^{(i+1)}, \dots, \mathbb{Z}_{m-1}^{(i+1)}, \mathbb{Z}_{m+1}^{(i)}, \dots, \mathbb{Z}_M^{(i)})$$

return $\{(\mathbb{Z}_1^{(i)}, \dots, \mathbb{Z}_M^{(i)})\}_{i=1}^N$

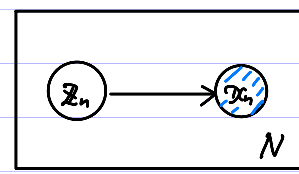
- サンプルを得たい変数の数が膨大なとき，複数の確率モデルが組み合わさった巨大な確率モデルからのサンプリングに利用するとよい．
- Gibbsサンプリングは，採択率が常に1のMetropolis法．

§4.2 最適化に基づく推論手法

・サンプリング法	最適化に基づく手法
・確率的な近似.	・決定的な近似.
✓ サンプルサイズを増やせば厳密解に近づく	✓ 収束性能が良い
✗ 必要なサンプルサイズが不明.	✗ 厳密解は決して求まらない.
✗ 計算コストが膨大.	

〈問題設定(再確認)〉

$\mathcal{X} = \{x_1, \dots, x_N\}$: 観測データ.



$\mathcal{Z} = \{z_1, \dots, z_N\}$: 潜在変数 (非観測)

$p(\mathcal{X}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Z}) = \prod_{n=1}^{N} p(x_n|z_n)p(z_n)$: 確率モデル

$z_n \overset{i.i.d.}{\sim} p(z)$.

事後分布 $p(\mathcal{Z}|\mathcal{X})$ を求めたい.

4.2.1 変分推論法

・変分推論 (variational inference):

分布の集合 $\mathcal{D}$ に属する分布のうち, 分布 $p(z)$ を"うまく近似する"分布 $q^*(z)$ を(何らかの)変分問題 $q^*(z) = \underset{q \in \mathcal{D}}{\operatorname{argmin}}\, \mathcal{S}[q]$ を解くことで求める手法.

"函数の函数"

$\mathcal{S} : \mathcal{D} \to \mathbb{R}$ を汎函数 (functional) という.

Remark ふつうは変分問題を解く際には変分法を用いるが，ここで扱うものは変分法を明示的に使わなくても解ける.

・事後分布について，Bayesの定理より

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = p(x,z)p(x)^{-1}$$

$$\Leftrightarrow \log p(z|x) = -\log p(x) + \log p(x,z).$$

$p(x,z)$は確率モデル（これは自分で設定したものになっている）. しかし，周辺尤度 $p(x) = \int p(x,z)dz$ は，モデルが複雑になると解析的に計算できない.

・zについての適当な分布 $q(z)$ を考える

$$\log p(x) = \log p(x,z) - \log p(z|x)$$

の両辺 $q(z)$ について期待値をとると，

$$\int q(z)\log p(x)dz = \int q(z)\log p(x,z)dz - \int q(z)\log p(z|x)dz$$

$$(\text{l.h.s.}) = \log p(x)\underbrace{\int q(z)dz}_{=1} = \log p(x)$$

(r.h.s.)

$$= \int q(z)\log \frac{p(x,z)}{q(z)}\cdot\frac{q(z)}{p(z|x)}dz$$

$$= \mathbb{E}_{q(z)}\left[\log\frac{p(x,z)}{q(z)}\right] + \mathbb{E}_{q(z)}\left[\log\frac{q(z)}{p(z|x)}\right]$$

$$= \mathbb{E}_{q(z)}\left[\log\frac{p(x,z)}{q(z)}\right] + D_{KL}[q(z)\|p(z|x)].$$

$\mathcal{L}[q] := \mathbb{E}_{q(z)}\left[\log\frac{p(x,z)}{q(z)}\right]$ とおくと，

$\log p(x) = \mathcal{L}[q] + D_{KL}[q(z) \| p(z|x)]$ を得る.

Kullback-Leibler divergence は非負なので,

$\log p(x) \geq \mathcal{L}[q]$.

エビデンス モデルが与えられたときのデータxの出現するもっともらしさ.

$\mathcal{L}[q]$ は 対数周辺尤度 の下界を与える. これを

エビデンス下界 (evidence lower bound : ELBO), あるいは

変分下界 (variational lower bound) という.

cf.) $\log p(x|\theta)$ を θ について最大化するのは 最尤法の枠組み.

$\log p(x)$ が求まらなくても, $\mathcal{L}[q]$ を最大化することで

$\log p(x)$ の (下からの) よい評価が得られる.

Remark

- $\mathcal{F}[q] := -\mathcal{L}[q]$ を変分エネルギー (variational energy) という.
- $\mathcal{L}[q]$ の設計の仕方は他にもある.

  cf.) Rényi Lower Bound.

- $q^*(z) := \underset{q}{\operatorname{argmax}}\ \mathcal{L}[q]$

  qによらず一定.

$$= \underset{q}{\operatorname{argmax}} \left(\underline{\log p(x)} - D_{KL}[q(z) \| p(z|x)] \right)$$

$$= \underset{q}{\operatorname{argmin}}\ D_{KL}[q(z) \| p(z|x)].$$

→ $q(z)$ が $p(z|x)$ をうまく近似するように最適化している.

- $D_{KL}[q(z) \| p(z|x)] \geq 0$ で, 等号成立は $q(z) = p(z|x)$ のとき.

$p(z|x)$ は $p(x)$ のために求まらなかったので, これでは意味がない...

→ 考える分布 q を ある範囲 $\mathcal{D}$ に制限してしまう.

・qを扱いやすい分布にできるが、代わりに表現力が制限されてしまうため、近似精度には限界がある。

・制限の仕方。

1. qをパラメトリックな分布 $q(z|\xi)$ に制限する。

・パラメータξを変分パラメータ(variational parameter)という

→このときELBOはξの函数 $\mathcal{L}(\xi)$ になる。最適なパラメータξ^*を求めるには、

$$\xi^* = \underset{\xi}{\operatorname{argmax}}\, \mathcal{L}(\xi)$$

を適当な最適化手法を用いて解けばよい。

2. 平均場近似(mean field approximation)。

$z = (z_1, \dots, z_M)$ とMコに分割して、独立性を仮定;

$$q(z) = \prod_{i=1}^{M} q_i(z_i)$$

→各近似分布$q_1(z_1), \dots, q_M(z_M)$を順にそれぞれ最適化する

・$q_j(z_j)$についての最適化をする 他の$q_i(z_i)$ $(i \neq j)$は固定する。

$$
\begin{aligned}
&\mathcal{L}[q] \\
&= \mathbb{E}_{q(z)}\left[\log \frac{p(x,z)}{q(z)}\right] \\
&= \int q_j(z_j) \prod_{i\neq j} q_i(z_i) \left(\log p(x,z) - \sum_{i\neq j} \log q_i(z_i) - \log q_j(z_j)\right) dz \\
&= \int q_j(z_j)\left(\int \prod_{i\neq j} q_i(z_i) \log p(x,z)\, d\underbrace{z_{\setminus j}}_{j\text{以外}}\right) dz_j \\
&\quad - \sum_{i\neq j} \int q_i(z_i) \log q_i(z_i)\, dz_i - \int q_j(z_j) \log q_j(z_j)\, dz_j.
\end{aligned}
$$

ここで，$\int \prod_{i \neq j} q_i(z_i) \log p(x, z) dz_{\setminus j} =: \mathbb{E}_{i \neq j}[\log p(x, z)]$ と書き，
$\sum_{i \neq j} \int q_i(z_i) \log q_i(z_i) dz_i = \text{const.}$ に注意すると，

$$\mathcal{L}[q] = \int q_j(z_j) \left(\mathbb{E}_{i \neq j}[\log p(x, z)] - \log q_j(z_j) \right) dz_j + \text{const.}$$

さらに，$\log \tilde{p}(x, z_j) = \mathbb{E}_{i \neq j}[\log p(x, z)] + \text{const.}$ となる分布 $\tilde{p}(x, z_j)$

を考える．すると，

$$\mathcal{L}[q] = -\int q_j(z_j) \log \frac{q_j(z_j)}{\tilde{p}(x, z_j)} dz_j + \text{const.}$$
$$= -D_{KL}[q_j(z_j) \| \tilde{p}(x, z_j)] + \text{const.}$$

よって，

$$q_j^*(z_j)$$
$$= \underset{q_j}{\operatorname{argmax}}\ \mathcal{L}[q]$$
$$= \underset{q_j}{\operatorname{argmin}}\ D_{KL}[q_j(z_j) \| \tilde{p}(x, z_j)]$$

← 一般にはここで変分法を使うことで最適解を求める．
ここではKL divergenceの最小化なので簡単．

$$= \tilde{p}(x, z_j)$$
$$\propto \exp\left(\mathbb{E}_{i \neq j}[\log p(x, z)] \right).$$

$q_j^*(z_j)$ が分布になるように，正規化定数を計算すると，

$$q_j^*(z_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\log p(x, z)] \right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\log p(x, z)] \right) dz_j}.$$

座標降下法．

→全ての因子 $q_i(z_i)$ を適当に初期化して，各 $q_i(z_i)$ を上式で逐次最適化する．

$\mathcal{L}[q]$ は各 q_i に関して凸なので，このスキームで q^* に収束する．

Remark $\mathcal{L}[q]$

$$= \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ$$

$$= \int q(Z) \left(\log p(X|Z) + \log \frac{p(Z)}{q(Z)} \right) dZ$$

$$= \mathbb{E}_{q(Z)}[\log p(X|Z)] - D_{KL}[q(Z) \| p(Z)].$$

$\mathcal{L}[q] \to \max$ とするとき

第1項：観測データの対数尤度の q による期待値

$\to$ なるべく大きくなる.

$\log p(X|Z)$ を最大にする Z を $\hat{Z}$ として, $q(Z) = \delta(Z - \hat{Z})$ (デルタ分布) のとき最大

第2項：q の p に対する KL-divergence.

$\to$ なるべく小さくなる.

q が p から大きく離れた分布になるのを防ぐ. (正則化の効果)

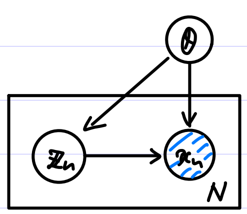
変分推論により得られる q^* は, 2項のバランスのとれた解になっている.

◦ 変分EMアルゴリズム (variational expectation maximization algo.)

・潜在変数 $Z = \{z_1, \dots, z_N\}$ の他に, パラメータ θ が存在する状況を考える.

・Z：データ数が増えるにつれ次元が増える (外延的)

・θ：固定次元 (内延的)



・モデル：$p(X, Z, \theta) = p(X|Z, \theta) p(Z|\theta) p(\theta)$

・変分EMアルゴリズムでは、

$$p(z,\theta)=p(z|\theta)p(\theta)\approx q(z,\theta)=q(z)q(\theta)$$

と平均場近似をして変分推論を行う.

。変分E-step : パラメータの近似分布 $q(\theta)$ を固定し、

潜在変数の近似分布 $q(z)$ を更新する.

$$q^*(z)=\frac{\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z,\theta)]\right)}{\int\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z,\theta)]\right)dz}.$$

ここで、

$$\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z,\theta)]\right)$$

$$=\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)+\log p(\theta)]\right)$$

$$=\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)]\right)\underbrace{\exp\left(\mathbb{E}_{q(\theta)}[\log p(\theta)]\right)}_{z\text{に依らない}}$$

なので、

$$\int\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)]\right)dz$$

$$=\exp\left(\mathbb{E}_{q(\theta)}[\log p(\theta)]\right)\int\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)]\right)dz.$$

結局、

$$q^*(z)=\frac{\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)]\right)}{\int\exp\left(\mathbb{E}_{q(\theta)}[\log p(x,z|\theta)]\right)dz}.$$

と更新する.

◦変分M-step：潜在変数の近似分布 $q(z)$ を固定し，

パラメータの近似分布 $q(\theta)$ を更新する．

$$q^*(\theta) = \frac{\exp\left(\mathbb{E}_{q(z)}[\log p(x, z, \theta)]\right)}{\int \exp\left(\mathbb{E}_{q(z)}[\log p(x, z, \theta)]\right) d\theta}$$

先と同様に，

$$\exp\left(\mathbb{E}_{q(z)}[\log p(x, z, \theta)]\right)$$

$$= \exp\left(\mathbb{E}_{q(z)}[\log p(x, z | \theta)]\right) \exp\left(\mathbb{E}_{q(z)}[\underbrace{\log p(\theta)}_{z\text{に依らない}}]\right)$$

$$= \exp\left(\mathbb{E}_{q(z)}[\log p(x, z | \theta)]\right) \exp\left(\log p(\theta)\right)$$

$$= p(\theta) \exp\left(\mathbb{E}_{q(z)}[\log p(x, z | \theta)]\right).$$

よって

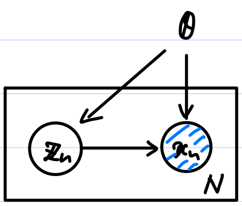
$$q^*(\theta) = \frac{p(\theta) \exp\left(\mathbb{E}_{q(z)}[\log p(x, z | \theta)]\right)}{\int p(\theta) \exp\left(\mathbb{E}_{q(z)}[\log p(x, z | \theta)]\right) d\theta}$$

と更新する．

cf.)（通常の）EMアルゴリズム．

モデル：$p(x, z; \theta) = p(x | z; \theta) p(z; \theta)$.

θにはは事前分布を設定しない



θ はパラメータ．

θ について，対数尤度

$$\log p(x; \theta) = \log \int p(x, z; \theta) dz$$

を最大化する（θ の最尤推定）．

z についての適当な分布 $q(z)$ を考えると、

$$\log p(x;\theta) = \log p(x;\theta) \int q(z)dz$$
$$= \int q(z) \log p(x;\theta)\, dz$$
$$= \int q(z) \log \frac{p(x,z;\theta)}{p(z|x;\theta)}\, dz$$
$$= \int q(z) \log \frac{p(x,z;\theta)}{q(z)}\, dz + \int q(z) \log \frac{q(z)}{p(z|x;\theta)}\, dz$$
$$= \int q(z) \log \frac{p(x,z;\theta)}{q(z)}\, dz + D_{KL}[q(z) \| p(z|x;\theta)].$$

$f=g$
$\Rightarrow \int_A f(x)dx = \int_A g(x)dx$

ここで $\mathcal{L}_\theta[q] := \int q(z) \log \frac{p(x,z;\theta)}{q(z)}\, dz$ とおけば、

$$\log p(x;\theta) = \mathcal{L}_\theta[q] + \underbrace{D_{KL}[q(z) \| p(z|x;\theta)]}_{\geq 0}$$
$$\geq \mathcal{L}_\theta[q].$$

下界 $\mathcal{L}_\theta[q]$ を θ に関して最大化することで尤度を最大にする θ を求める。

このために、q と θ について交互に最適化していく。

- E-step：θ を固定して $q(z)$ を更新する。

$$\mathcal{L}_\theta[q] = \underbrace{\log p(x;\theta)}_{q\text{に依らない}} - D_{KL}[q(z) \| p(z|x;\theta)]$$ なので、

$$q^*(z) = \underset{q(z)}{\operatorname{argmax}}\ \mathcal{L}_\theta[q]$$
$$= \underset{q(z)}{\operatorname{argmin}}\ D_{KL}[q(z) \| p(z|x;\theta)]$$
$$= p(z|x;\theta). \quad (z\text{の事後分布})$$

このとき、

この期待値をとるためのステップなので E-step.

$$\mathcal{L}[q^*] = \int q^*(z) \log p(x,z;\theta)dz - \int q^*(z) \log q^*(z) dz = \mathbb{E}_{q^*(z)}[\log p(x,z;\theta)] + \text{const}$$

・M-step：$q(z)$を固定して θ を更新する.

$$\theta^* = \underset{\theta}{\mathrm{argmax}}\ \mathcal{L}_\theta[q]$$
$$= \underset{\theta}{\mathrm{argmax}} \int q(z)\log p(x, z;\theta)\,dz$$
$$= \underset{\theta}{\mathrm{argmax}}\ \mathbb{E}_{q(z)}[\log p(x, z;\theta)].$$

← θについて最大化するステップなのでM-step.

最大化してるので.

・以上のE-step, M-stepでは $\mathcal{L}_\theta[q]$ が減少することはない.

E-stepを行う前のパラメータを θ^{old} とする.

E-step終了後の q を q^{new} とすると, $\mathcal{L}_{\theta^{old}}[q^{new}] = \log p(x;\theta^{old})$ が成立.

M-step終了後のパラメータを θ^{new} とすると, $q^{new}(z) = p(z|x;\theta^{old})$ なので

$$\log p(x;\theta^{new}) = \mathcal{L}_{\theta^{new}}[q^{new}] + D_{KL}[q^{new}(z)\,\|\,p(z|x;\theta^{new})]$$
$$\geq \mathcal{L}_{\theta^{old}}[q^{new}] + D_{KL}[q^{new}(z)\,\|\,p(z|x;\theta^{new})]$$
$$= \log p(x;\theta^{old}) + D_{KL}[q^{new}(z)\,\|\,p(z|x;\theta^{new})].$$
$$= \log p(x;\theta^{old}) + D_{KL}[p(z|x;\theta^{old})\,\|\,p(z|x;\theta^{new})].$$
$$\geq \log p(x;\theta^{old}).$$

M-stepで θ について最大化しているので $\mathcal{L}_{\theta^{new}}[q] \geq \mathcal{L}_{\theta^{old}}[q]$ が $\forall q$ について成立

よってEM algo. 1ステップで対数尤度は増加する.

$\theta^{new} = \theta^{old}$ のとき, またそのときに限り等号成立.

対数尤度が収束したときのパラメータ θ が最大推定量になる.

4.2.2 例：平均場近似による潜在変数モデルの学習.

- 潜在変数モデル (latent variable model) による次元削減を扱う.
  - $\mathcal{X} = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$ ：観測データ

    $\mathcal{Z} = \{z_1, \dots, z_N\} \subseteq \mathbb{R}^\ell$ ：潜在変数 $(\ell \ll d)$.

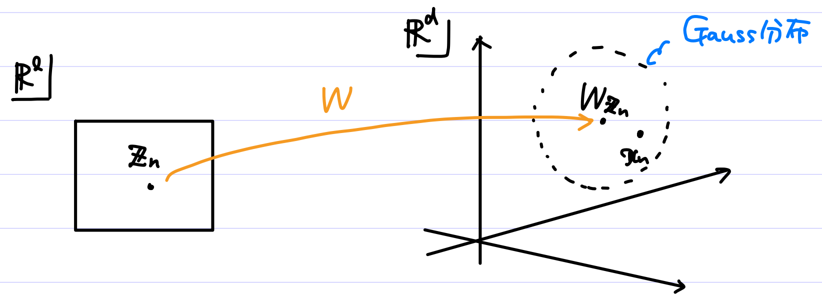
    x_i を z_i で表現して，データの圧縮，可視化，特徴量抽出などを行う.
  - PCA, ICA, 行列分解，k-means などがこのモデルで表現できる.

4.2.2.1 線形次元削減への適用.

- 仮定.

  $W \in M_{d,\ell}(\mathbb{R})$ をパラメータ，$\sigma_x^2 > 0$ を定数として

  $$p(\mathcal{X} \mid \mathcal{Z}, W) = \prod_{n=1}^{N} p(x_n \mid z_n, W) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid W z_n, \sigma_x^2 I_d)$$



  $\mathcal{Z}$ の事前分布： $p(\mathcal{Z}) = \prod_{n=1}^{N} \mathcal{N}(z_n \mid 0, I_\ell)$

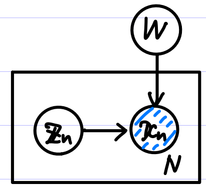
  W の事前分布：

  $$W = \begin{pmatrix} w_1^T \\ \vdots \\ w_d^T \end{pmatrix} \quad (w_j \in \mathbb{R}^\ell), \quad \sigma_w^2 > 0 \text{：定数} \quad \text{として.}$$

  $$p(W) = \prod_{j=1}^{d} \mathcal{N}(w_j \mid 0, \sigma_w^2 I_\ell).$$

・モデル：

$$p(\mathcal{X}, \mathcal{Z}, W) = p(\mathcal{X}|\mathcal{Z}, W)\,p(\mathcal{Z}, W)$$
$$= p(\mathcal{X}|\mathcal{Z}, W)\,p(\mathcal{Z})\,p(W).$$



・事後分布 $p(\mathcal{Z}, W|\mathcal{X})$ を次で近似：

$$p(\mathcal{Z}, W|\mathcal{X}) \approx q(\mathcal{Z})\,q(W).$$

変分推論（変分EM algo.）を行う.

i 回目の更新後の近似分布を $q_i(\mathcal{Z})$, $q_i(W)$ と書く.

◦ 変分 E-step. $\mathcal{Z}$ の近似分布を更新.

$$q_{i+1}(\mathcal{Z}) \propto \exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}, \mathcal{Z}|W)]\right).$$

Wに依存しない

$$= \exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}|\mathcal{Z}, W)] + \mathbb{E}_{q_i(W)}[\underline{\log p(\mathcal{Z})}]\right)$$
$$= \exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}|\mathcal{Z}, W)]\right)\exp\left(\log p(\mathcal{Z})\right)$$
$$= p(\mathcal{Z})\exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}|\mathcal{Z}, W)]\right).$$

◦ 変分 M-step. W の近似分布を更新.

$$q_{i+1}(W) \propto p(W)\exp\left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\log p(\mathcal{X}, \mathcal{Z}|W)]\right)$$
$$\propto p(W)\exp\left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\log p(\mathcal{X}|\mathcal{Z}, W)]\right).$$

・各 step の計算をもう少し進めてみる.

$$\log p(\mathcal{X}|\mathcal{Z}, W)$$
$$= \log\left(\prod_{n=1}^{N} \mathcal{N}(x_n | W z_n, \sigma_x^2 I_d)\right)$$

$$= \sum_{n=1}^{N} \log \mathcal{N}(x_n \mid W z_n, \sigma_x^2 I_d)$$

z_n, W に依存しない項は const. にまとめる.

$$= \sum_{n=1}^{N} \left(-\frac{1}{2\sigma_x^2} (x_n - W z_n)^T (x_n - W z_n) \right) + \text{const.}$$

スカラーなので転置をとっても同じ.

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\underbrace{x_n^T x_n}_{\text{const.}} - x_n^T W z_n - \underline{z_n^T W^T x_n} + z_n^T W^T W z_n \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(z_n^T W^T W z_n - 2 \underbrace{x_n^T W z_n}_{= z_n^T W^T x_n} \right) + \text{const.}$$

スカラーなので転置をとっても同じ.

よって,

$$\mathbb{E}_{q_i(W)}[\log p(\mathcal{X} \mid \mathcal{Z}, W)]$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(z_n^T \mathbb{E}_{q_i(W)}[W^T W] z_n - 2 z_n^T \mathbb{E}_{q_i(W)}[W]^T x_n \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(z_n^T \left(\sum_{j=1}^{d} \mathbb{E}_{q_i(W)}[w_j w_j^T] \right) z_n - 2 z_n^T \left(\sum_{j=1}^{d} x_{n,j} \mathbb{E}_{q_i(W)}[w_j] \right) \right) + \text{const.}$$

$$\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\log p(\mathcal{X} \mid \mathcal{Z}, W)]$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\underline{z_n^T W^T W z_n}] - 2 x_n^T W \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n] \right) + \text{const.}$$

スカラーは tr をとっても同じ

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\underline{\mathrm{tr}(z_n^T W^T W z_n)}] - 2 x_n^T W \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n] \right) + \text{const.}$$

$\mathrm{tr}(AB) = \mathrm{tr}(BA)$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\underline{\mathrm{tr}(W z_n z_n^T W^T)}] - 2 x_n^T W \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n] \right) + \text{const.}$$

tr の計算

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\mathbb{E}_{q_{i+1}(\mathcal{Z})}\left[\sum_{j=1}^{d} w_j^T z_n z_n^T w_j \right] - 2 x_n^T W \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n] \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\sum_{j=1}^{d} w_j^T \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n z_n^T] w_j - 2 \underline{x_n^T W \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n]} \right) + \text{const.}$$

スカラーの転置

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\sum_{j=1}^{d} w_j^T \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n z_n^T] w_j - 2 \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n]^T W^T x_n \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{n=1}^{N} \left(\sum_{j=1}^{d} w_j^T \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n z_n^T] w_j - 2 \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n]^T \left(\sum_{j=1}^{d} x_{n,j} w_j \right) \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{j=1}^{d} \sum_{n=1}^{N} \left(w_j^T \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n z_n^T] w_j - 2 x_{n,j} \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n]^T w_j \right) + \text{const.}$$

$$= -\frac{1}{2\sigma_x^2} \sum_{j=1}^{d} \left(w_j^T \left(\sum_{n=1}^{N} \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n z_n^T] \right) w_j - 2 w_j^T \left(\sum_{n=1}^{N} x_{n,j} \mathbb{E}_{q_{i+1}(\mathcal{Z})}[z_n] \right) \right) + \text{const.}$$

変分E-step での更新は,

$$\log q_{i+1}(\mathcal{Z})$$

$$= \log p(\mathcal{Z}) + \mathbb{E}_{q_i(W)}[\log p(\mathcal{X}|\mathcal{Z}, W)] + \text{const.}$$

$$= -\frac{1}{2}\sum_{n=1}^{N} \mathbf{z}_n^T \mathbf{z}_n - \frac{1}{2\sigma_x^2}\sum_{n=1}^{N}\left(\mathbf{z}_n^T\left(\sum_{j=1}^{d}\mathbb{E}_{q_i(W)}[\mathbf{w}_j \mathbf{w}_j^T]\right)\mathbf{z}_n - 2\mathbf{z}_n^T\left(\sum_{j=1}^{d} x_{n,j}\mathbb{E}_{q_i(W)}[\mathbf{w}_j]\right)\right)$$

$$+ \text{const.}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{z}_n^T \underbrace{\left(I_\ell + \frac{1}{\sigma_x^2}\sum_{j=1}^{d}\mathbb{E}_{q_i(W)}[\mathbf{w}_j \mathbf{w}_j^T]\right)}_{=:\hat{\Sigma}_{z,i+1}^{-1}} \mathbf{z}_n - 2\mathbf{z}_n^T \underbrace{\left(\frac{1}{\sigma_x^2}\sum_{j=1}^{d} x_{n,j}\mathbb{E}_{q_i(W)}[\mathbf{w}_j]\right)}_{=:\hat{\Sigma}_{z,i+1}^{-1}\hat{\mu}_{z_n,i+1}}\right) + \text{const}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{z}_n^T \hat{\Sigma}_{z,i+1}^{-1}\mathbf{z}_n - 2\mathbf{z}_n^T \hat{\Sigma}_{z,i+1}^{-1}\hat{\mu}_{z_n,i+1}\right) + \text{const}$$

$$= \sum_{n=1}^{N}\log \mathcal{N}(\mathbf{z}_n | \hat{\mu}_{z_n,i+1}, \hat{\Sigma}_{z,i+1})$$

つまり, $q_{i+1}(\mathcal{Z}) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{z}_n | \hat{\mu}_{z_n,i+1}, \hat{\Sigma}_{z,i+1})$.

右辺は各 $\mathbf{z}_n$ の分布の積の形なので, $q_{i+1}(\mathcal{Z}) = \prod_{n=1}^{N} q_{i+1}(\mathbf{z}_n)$ と独立な形に書けて, 各 $n = 1, \dots, N$ に対して

$$q_{i+1}(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n | \hat{\mu}_{z_n,i+1}, \hat{\Sigma}_{z,i+1}),$$

$$\hat{\mu}_{z_n,i+1} = \frac{1}{\sigma_x^2}\hat{\Sigma}_{z,i+1}\sum_{j=1}^{d} x_{n,j}\mathbb{E}_{q_i(W)}[\mathbf{w}_j],$$

$$\hat{\Sigma}_{z,i+1} = \left(I_\ell + \frac{1}{\sigma_x^2}\sum_{j=1}^{d}\mathbb{E}_{q_i(W)}[\mathbf{w}_j \mathbf{w}_j^T]\right)^{-1}$$

と更新すればよい.

$\mathbb{E}_{q_i(W)}[\mathbf{w}_j]$ と $\mathbb{E}_{q_i(W)}[\mathbf{w}_j \mathbf{w}_j^T]$ については, 次の変分 M-step の更新の結果から計算する.

変分M-stepでの更新は

$$\log q_{i+1}(W)$$

$$= \log p(W) + \mathbb{E}_{q_{i+1}(Z)}[\log p(\mathcal{X}|Z, W)] + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{d}\frac{1}{\sigma_w^2} w_j^T w_j - \frac{1}{2}\sum_{j=1}^{d}\left(w_j^T\left(\frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(Z)}[z_n z_n^T]\right)w_j - 2w_j^T\left(\frac{1}{\sigma_x^2}\sum_{n=1}^{N}x_{n,j}\mathbb{E}_{q_{i+1}(Z)}[z_n]\right)\right) + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{d}\left(w_j^T\underbrace{\left(\frac{1}{\sigma_w^2}I_d + \frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(Z)}[z_n z_n^T]\right)}_{=:\hat{\Sigma}_{w,i+1}^{-1}}w_j - 2w_j^T\underbrace{\left(\frac{1}{\sigma_x^2}\sum_{n=1}^{N}x_{n,j}\mathbb{E}_{q_{i+1}(Z)}[z_n]\right)}_{=:\hat{\Sigma}_{w,i+1}^{-1}\hat{\mu}_{w_j,i+1}}\right) + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{d}\left(w_j^T\hat{\Sigma}_{w,i+1}^{-1}w_j - 2w_j^T\hat{\Sigma}_{w,i+1}^{-1}\mu_{w_j,i+1}\right) + \text{const.}$$

$$= \sum_{j=1}^{d}\log\mathcal{N}(w_j | \mu_{w_j,i+1}, \hat{\Sigma}_{w,i+1})$$

つまり，$q_{i+1}(W) = \prod_{n=1}^{d}\mathcal{N}(w_j | \mu_{w_j,i+1}, \hat{\Sigma}_{w,i+1})$

右辺は各 w_j の分布の積の形になっているので，$q_{i+1}(W) = \prod_{n=1}^{d} q_{i+1}(w_j)$ と独立な形に

書けて，各 $j = 1, \dots, d$ に対して

$$q_{i+1}(w_j) = \mathcal{N}(w_j | \mu_{w_j,i+1}, \hat{\Sigma}_{w,i+1})$$

$$\mu_{w_j,i+1} = \frac{1}{\sigma_x^2}\hat{\Sigma}_{w,i+1}\sum_{n=1}^{N}x_{n,j}\mathbb{E}_{q_{i+1}(Z)}[z_n]$$

$$\hat{\Sigma}_{w,i+1} = \left(\frac{1}{\sigma_w^2}I_d + \frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(Z)}[z_n z_n^T]\right)^{-1}$$

と更新すればよい．

結局，$q_i(z_n)$ も $q_i(w_j)$ も Gauss分布なので，期待値が計算でき，

$$\mathbb{E}_{q_i(W)}[w_j] = \mathbb{E}_{q_i(w_j)}[w_j] = \hat{\mu}_{w_j,i},$$

$$\mathbb{E}_{q_i(W)}[w_j w_j^T] = \mathbb{E}_{q_i(w_j)}[w_j w_j^T] = \hat{\Sigma}_{w,i} + \hat{\mu}_{w_j,i}\hat{\mu}_{w_j,i}^T,$$

$$\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\mathbf{z}_n] = \mathbb{E}_{q_{i+1}(\mathbf{z}_n)}[\mathbf{z}_n] = \hat{\mu}_{z_n, i+1},$$

$$\mathbb{E}_{q_{i+1}(\mathcal{Z})}[\mathbf{z}_n \mathbf{z}_n^T] = \mathbb{E}_{q_{i+1}(\mathbf{z}_n)}[\mathbf{z}_n \mathbf{z}_n^T] = \hat{\Sigma}_{z, i+1} + \hat{\mu}_{z_n, i+1} \hat{\mu}_{z_n, i+1}^T.$$

最終的に、以下のように Gauss分布のパラメータを更新すればよいことが分かった：

[変分 E-step]

$$\hat{\Sigma}_{z, i+1} = \left(I_l + \frac{d}{\sigma_x^2} \hat{\Sigma}_{w, i} + \frac{1}{\sigma_x^2} \sum_{j=1}^{d} \hat{\mu}_{w_j, i} \hat{\mu}_{w_j, i}^T \right)^{-1},$$

$$\hat{\mu}_{z_n, i+1} = \frac{1}{\sigma_x^2} \hat{\Sigma}_{z, i+1} \sum_{j=1}^{d} x_{n,j} \hat{\mu}_{w_j, i}.$$

[変分 M-step]

$$\hat{\Sigma}_{w, i+1} = \left(\frac{1}{\sigma_w^2} I_l + \frac{N}{\sigma_x^2} \hat{\Sigma}_{z, i+1} + \frac{1}{\sigma_x^2} \sum_{n=1}^{N} \hat{\mu}_{z_n, i+1} \hat{\mu}_{z_n, i+1}^T \right)^{-1},$$

$$\hat{\mu}_{w_j, i+1} = \frac{1}{\sigma_x^2} \hat{\Sigma}_{w, i+1} \sum_{n=1}^{N} x_{n,j} \hat{\mu}_{z_n, i+1}.$$

4.2.2.2 混合Gauss分布への適用.

・連続な潜在変数 $\mathcal{Z}$ の代わりに、離散の潜在変数

$$\mathcal{S} = \{ s_1, \dots, s_N \} \subseteq \{0, 1\}^K, \quad \sum_{k=1}^{K} s_{n,k} = 1 \quad (n = 1, \dots, N)$$

を割り当てることで、クラスタリングができる.

・仮定.

$W = (w_1, \dots, w_K) \in M_{d,K}(\mathbb{R})$ $(w_j \in \mathbb{R}^d)$ をパラメータ, $\sigma_x^2 > 0$ を定数として,

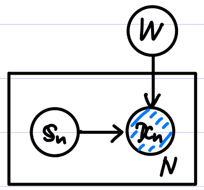
$$p(\mathcal{X} | \mathcal{S}, W) = \prod_{n=1}^{N} p(x_n | s_n, W) = \prod_{n=1}^{N} \mathcal{N}(x_n | W s_n, \sigma_x^2 I_d).$$

$\mathcal{S}$ の事前分布 : $p(\mathcal{S}) = \prod_{n=1}^{N} \mathrm{Cat}(s_n | \pi)$ （π : 定ベクトル）

W の事前分布 : $\sigma_w^2 > 0$ として $p(W) = \prod_{k=1}^{K} \mathcal{N}(w_k | 0, \sigma_w^2 I_d)$.

・モデル:

$$p(\mathcal{X}, \mathcal{S}, W) = p(\mathcal{X}|\mathcal{S}, W)p(\mathcal{S}, W)$$
$$= p(\mathcal{X}|\mathcal{S}, W)p(\mathcal{S})p(W)$$



・事後分布 $p(\mathcal{S}, W|\mathcal{X})$ を次で近似:

$$p(\mathcal{S}, W|\mathcal{X}) \approx q(\mathcal{S})q(W).$$

変分推論（変分EM algo.）を行う.

i回目の更新後の近似分布を $q_i(\mathcal{S})$, $q_i(W)$ と書く.

・変分E-step. $\mathcal{S}$の近似分布を更新.

$$q_{i+1}(\mathcal{S}) \propto \exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}, \mathcal{S}|W)]\right)$$
$$= p(\mathcal{S})\exp\left(\mathbb{E}_{q_i(W)}[\log p(\mathcal{X}|\mathcal{S}, W)]\right)$$

・変分M-step. Wの近似分布を更新.

$$q_{i+1}(W) \propto p(W)\exp\left(\mathbb{E}_{q_{i+1}(\mathcal{S})}[\log p(\mathcal{X}, \mathcal{S}|W)]\right)$$
$$\propto p(W)\exp\left(\mathbb{E}_{q_{i+1}(\mathcal{S})}[\log p(\mathcal{X}|\mathcal{S}, W)]\right)$$

・各stepの計算をもう少し進める.

$$\log p(\mathcal{X}|\mathcal{S}, W)$$
$$= -\frac{1}{2}\sum_{n=1}^{N}\left(\frac{1}{\sigma_x^2}s_n^T W^T W s_n - \frac{2}{\sigma_x^2}x_n^T W s_n\right) + \text{const.}$$
$$= -\frac{1}{2}\sum_{n=1}^{N}\left(\frac{1}{\sigma_x^2}\underline{\left(\sum_{k=1}^{K}s_{n,k}w_k\right)^T\left(\sum_{k=1}^{K}s_{n,k}w_k\right)} - \frac{2}{\sigma_x^2}x_n^T\left(\sum_{k=1}^{K}s_{n,k}w_k\right)\right) + \text{const.}$$
$$= -\frac{1}{2}\sum_{n=1}^{N}\left(\frac{1}{\sigma_x^2}\sum_{k=1}^{K}s_{n,k}w_k^T w_k - \frac{2}{\sigma_x^2}x_n^T\left(\sum_{k=1}^{K}s_{n,k}w_k\right)\right) + \text{const}$$

$$s_{n,k}s_{n,k'} = \begin{cases} 0 & (k \neq k') \\ s_{n,k} & (k = k') \end{cases}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{\sigma_x^2} s_{n,k} w_k^T w_k - \frac{2}{\sigma_x^2} s_{n,k} x_n^T w_k\right) + \text{const}$$

よって，

$$\mathbb{E}_{q_i(w)}[\log p(\mathcal{X}|\mathcal{S}, W)]$$
$$= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{\sigma_x^2} s_{n,k} \mathbb{E}_{q_i(w)}[w_k^T w_k] - \frac{2}{\sigma_x^2} s_{n,k} x_n^T \mathbb{E}_{q_i(w)}[w_k]\right) + \text{const}.$$

$$\mathbb{E}_{q_{i+1}(\mathcal{S})}[\log p(\mathcal{X}|\mathcal{S}, W)]$$
$$= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{\sigma_x^2} \mathbb{E}_{q_{i+1}(\mathcal{S})}[s_{n,k}] w_k^T w_k - \frac{2}{\sigma_x^2} \mathbb{E}_{q_{i+1}(\mathcal{S})}[s_{n,k}] x_n^T w_k\right) + \text{const}$$

変分E-stepでは，

$$\log q_{i+1}(\mathcal{S})$$
$$= \log p(\mathcal{S}) + \mathbb{E}_{q_i(w)}[\log p(\mathcal{X}|\mathcal{S}, W)] + \text{const}.$$
$$= \sum_{n=1}^{N}\log \mathrm{Cat}(s_n|\pi) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{\sigma_x^2} s_{n,k} \mathbb{E}_{q_i(w)}[w_k^T w_k] - \frac{2}{\sigma_x^2} s_{n,k} x_n^T \mathbb{E}_{q_i(w)}[w_k]\right)$$
$$+ \text{const}.$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} s_{n,k}\underbrace{\left(\log \pi_k - \frac{1}{2\sigma_x^2}\mathbb{E}_{q_i(w)}[w_k^T w_k] + \frac{1}{\sigma_x^2} x_n^T \mathbb{E}_{q_i(w)}[w_k]\right)}_{=:\ \log \alpha_{n,k,i+1}} + \text{const}.$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} s_{n,k}\log \alpha_{n,k,i+1} + \text{const}$$

$$\hat{\pi}_{n,k,i+1} := \frac{\alpha_{n,k,i+1}}{\sum_{k'=1}^{K}\alpha_{n,k',i+1}}, \quad \hat{\pi}_{n,i+1} := \begin{pmatrix}\hat{\pi}_{n,1,i+1}\\ \vdots \\ \hat{\pi}_{n,K,i+1}\end{pmatrix}$$ とおくと，

$$\log q_{i+1}(\mathcal{S}) = \sum_{n=1}^{N}\log \mathrm{Cat}(s_n|\hat{\pi}_{n,i+1}) \Leftrightarrow q_{i+1}(\mathcal{S}) = \prod_{n=1}^{N}\mathrm{Cat}(s_n|\hat{\pi}_{n,i+1}).$$

右辺は各s_nの分布の積の形なので，$q_{i+1}(\mathcal{S}) = \prod_{n=1}^{N} q_{i+1}(s_n)$ と独立な形にかける．

各s_n $(n=1,\dots,N)$について，

$$\alpha_{n,k,i+1} = \exp\left(\log \pi_k - \frac{1}{2\sigma_x^2}\mathbb{E}_{q_i(w)}\left[w_k^T w_k\right] + \frac{1}{\sigma_x^2} x_n^T \mathbb{E}_{q_i(w)}\left[w_k\right]\right)$$

$$= \pi_k \exp\left(-\frac{1}{2\sigma_x^2}\mathbb{E}_{q_i(w)}\left[w_k^T w_k\right] + \frac{1}{\sigma_x^2} x_n^T \mathbb{E}_{q_i(w)}\left[w_k\right]\right) \quad (k=1,\dots,K)$$

$$\hat{\pi}_{n,k,i+1} = \frac{\alpha_{n,k,i+1}}{\sum_{k'=1}^{K}\alpha_{n,k',i+1}} \quad (k=1,\dots,K), \quad \hat{\pi}_{n,i+1} = \begin{pmatrix} \hat{\pi}_{n,1,i+1} \\ \vdots \\ \hat{\pi}_{n,K,i+1} \end{pmatrix},$$

$$q_{i+1}(s_n) = \mathrm{Cat}(s_n \mid \hat{\pi}_{n,i+1})$$

と更新すればよい.

変分M-stepでは,

$$\log q_{i+1}(W)$$

$$= \log p(W) + \mathbb{E}_{q_{i+1}(s)}\left[\log p(x \mid s, W)\right] + \text{const.}$$

$$= -\frac{1}{2\sigma_w^2}\sum_{k=1}^{K} w_k^T w_k - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{\sigma_x^2}\mathbb{E}_{q_{i+1}(s)}\left[s_{n,k}\right] w_k^T w_k - \frac{2}{\sigma_x^2}\mathbb{E}_{q_{i+1}(s)}\left[s_{n,k}\right] x_n^T w_k\right)$$

$$+ \text{const}$$

$$= -\frac{1}{2}\sum_{k=1}^{K}\left(\underbrace{\left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(s)}\left[s_{n,k}\right]\right)}_{=:\ \hat{\sigma}_{k,i+1}^{-2}} w_k^T w_k - 2 w_k^T \underbrace{\left(\frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(s)}\left[s_{n,k}\right] x_n\right)}_{=:\ \hat{\sigma}_{k,i+1}^{-2}\hat{\mu}_{k,i+1}}\right)$$

$$+ \text{const.}$$

$$= -\frac{1}{2}\sum_{k=1}^{K}\left(w_k^T\left(\hat{\sigma}_{k,i+1}^2 I_d\right)^{-1} w_k - 2 w_k^T\left(\hat{\sigma}_{k,i+1}^2 I_d\right)^{-1}\hat{\mu}_{k,i+1}\right) + \text{const.}$$

$$= \sum_{k=1}^{K}\mathcal{N}\left(w_k \mid \hat{\mu}_{k,i+1}, \hat{\sigma}_{k,i+1}^2 I_d\right).$$

$$\Leftrightarrow q_{i+1}(W) = \prod_{k=1}^{K}\mathcal{N}\left(w_k \mid \hat{\mu}_{k,i+1}, \hat{\sigma}_{k,i+1}^2 I_d\right).$$

右辺は各 w_k の分布の積の形なので, $q_{i+1}(W) = \prod_{k=1}^{K} q_{i+1}(w_k)$ と独立な形にかけて,

各 w_k $(k=1,\dots,K)$ に対して,

$$q_{i+1}(w_k) = \mathcal{N}(w_k \mid \hat{\mu}_{k,i+1}, \hat{\sigma}_{k,i+1}^2 I_d),$$

$$\hat{\sigma}_{k,i+1}^2 = \left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(s)}[s_{n,k}]\right)^{-1}$$

$$\hat{\mu}_{k,i+1} = \frac{\hat{\sigma}_{k,i+1}^2}{\sigma_x^2}\sum_{n=1}^{N}\mathbb{E}_{q_{i+1}(s)}[s_{n,k}]x_n$$

と更新すればよい.

分布の形が分かったので期待値を計算する.

$\mathcal{N}(w_k \mid \hat{\mu}_{k,i}, \hat{\sigma}_{k,i}^2 I_d)$ のとき
w_k の各成分 $w_{k,j}$ はそれぞれ
$\mathcal{N}(\hat{\mu}_{k,i,j}, \hat{\sigma}_{k,i}^2)$ に従う.

$$\mathbb{E}_{q_i(w)}[w_k] = \mathbb{E}_{q_i(w_k)}[w_k] = \hat{\mu}_{k,i}$$

$$\mathbb{E}_{q_i(w)}[w_k^T w_k] = \mathbb{E}_{q_i(w_k)}[w_k^T w_k] = \sum_{j=1}^{d}\mathbb{E}_{q_i(w_k)}[w_{k,j}^2]$$

$$= \sum_{j=1}^{d}(\hat{\sigma}_{k,i}^2 + \hat{\mu}_{k,i,j}^2) = d\hat{\sigma}_{k,i}^2 + \hat{\mu}_{k,i}^T\hat{\mu}_{k,i}.$$

$$\mathbb{E}_{q_{i+1}(s)}[s_{n,k}] = \mathbb{E}_{q_{i+1}(s_n)}[s_{n,k}] = \hat{\pi}_{n,k,i+1}.$$

以上より, パラメータの更新式は次のとおり.

[変分E-step]

$$\alpha_{n,k,i+1} = \pi_k \exp\left(-\frac{1}{2\sigma_x^2}\left(d\hat{\sigma}_{k,i}^2 + \hat{\mu}_{k,i}^T\hat{\mu}_{k,i}\right) + \frac{1}{\sigma_x^2}x_n^T\hat{\mu}_{k,i}\right)$$

$$\hat{\pi}_{n,k,i+1} = \frac{\alpha_{n,k,i+1}}{\sum_{k'=1}^{K}\alpha_{n,k',i+1}}, \quad \hat{\pi}_{n,i+1} = \begin{pmatrix}\hat{\pi}_{n,1,i+1}\\ \vdots \\ \hat{\pi}_{n,K,i+1}\end{pmatrix}.$$

[変分M-step]

$$\hat{\sigma}_{k,i+1}^2 = \left(\frac{1}{\sigma_w^2} + \frac{1}{\sigma_x^2}\sum_{n=1}^{N}\hat{\pi}_{n,k,i+1}\right)^{-1},$$

$$\hat{\mu}_{k,i+1} = \frac{\hat{\sigma}_{k,i+1}^2}{\sigma_x^2}\sum_{n=1}^{N}\hat{\pi}_{n,k,i+1}x_n.$$

4.2.3 Laplace近似.

- Laplace近似 (Laplace approximation):

事後分布 $p(\mathbf{z}|\mathbf{x})$など.
複雑な分布をより簡単なGauss分布を使って近似的に表現する手法.

- 分布 $p(\mathbf{z})$のモード (mode): $p(\mathbf{z})$の極大値を与える点.

- $\mathbf{z}_0$: 近似したい分布 $p(\mathbf{z})$の任意のモード.

Gauss分布では,

$$\log \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \Sigma) = -\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \underset{\mathbf{z}\text{について}}{\mathrm{const}}.$$

の形になっているので, $\log p(\mathbf{z})$を$\mathbf{z}_0$まわりでTaylor展開してみる.

$$\begin{aligned}
&\log p(\mathbf{z}) \\
&= \underbrace{\log p(\mathbf{z}_0)}_{\mathrm{const.}} + \left(\nabla_{\mathbf{z}} \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}\right)^T(\mathbf{z}-\mathbf{z}_0) + \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\left(\nabla_{\mathbf{z}}^2 \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}\right)(\mathbf{z}-\mathbf{z}_0) \\
&\quad + O(\|\mathbf{z}-\mathbf{z}_0\|^3) \\
&\approx \left(\nabla_{\mathbf{z}} \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}\right)^T(\mathbf{z}-\mathbf{z}_0) + \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\left(\nabla_{\mathbf{z}}^2 \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}\right)(\mathbf{z}-\mathbf{z}_0) + \mathrm{const.}
\end{aligned}$$

$\mathbf{z}_0$は$p(\mathbf{z})$のモードであり, $\log p(\mathbf{z})$の極大値でもあるので,

$$\nabla_{\mathbf{z}} \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0} = \mathbf{0}.$$

ゆえに,

$$\log p(\mathbf{z}) \approx \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\left(\nabla_{\mathbf{z}}^2 \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}\right)(\mathbf{z}-\mathbf{z}_0) + \mathrm{const.}$$

← $\mathbf{z}_0$がモードなら, $\Lambda(\mathbf{z}_0)$は正定値行列.

よって, $\Lambda(\mathbf{z}_0) := -\nabla_{\mathbf{z}}^2 \log p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}$ とおけば,

$$\log p(\mathbf{z}) = -\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\Lambda(\mathbf{z}_0)(\mathbf{z}-\mathbf{z}_0) + \mathrm{const.}$$

$$\Leftrightarrow \quad p(\mathbf{z}) \approx C\exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\Lambda(\mathbf{z}_0)(\mathbf{z}-\mathbf{z}_0)\right) = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \Lambda(\mathbf{z}_0)^{-1})$$ と近似できる.

・適当な最適化手法で $p(\boldsymbol{z})$ のモード $\boldsymbol{z}_0$ を求め，$\Lambda(\boldsymbol{z}_0)$ を計算することで

近似 $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}_0, \Lambda(\boldsymbol{z}_0)^{-1})$ を得る．

Remark ・多峰的 (multimodal) な分布では，モード $\boldsymbol{z}_0$ の選び方により得られる近似が異なる．

・モード $\boldsymbol{z}_0$ 近傍のみを考慮した近似なので，分布全体を近似できてはいない．

・Hesse行列の計算をする必要がある．

・$p(\boldsymbol{z})$ が実変数の分布でないと，直接適用できない．

4.2.4 モーメントマッチングによる近似．

4.2.4.1 モーメントマッチング．

・分布 $p(\boldsymbol{z})$ を，より簡単な分布 $q(\boldsymbol{z})$ で近似する．

[仮定] $q(\boldsymbol{z})$ は指数型分布族，i.e.,

$$q(\boldsymbol{z};\eta) = h(\boldsymbol{z})\exp\left(\eta^T t(\boldsymbol{z}) - a(\eta)\right).$$

・p を q で近似するため，p の q に対する KL-divergence $D_{KL}[p(\boldsymbol{z}) \| q(\boldsymbol{z};\eta)]$ (forward KL-divergence) を最小化する

Remark ・変分推論では reverse KL-divergence $D_{KL}[q(\boldsymbol{z};\eta) \| p(\boldsymbol{z})]$ を最小化していた．

今回は，「p を真の分布とみたときの，q の p からの乖離度」を最小化する．

・解く問題は $\eta^* = \underset{\eta \in \mathcal{H}}{\operatorname{argmin}}\, D_{KL}[p(\boldsymbol{z}) \| q(\boldsymbol{z};\eta)]$.

q が分布となるような範囲．

但し，$\mathcal{H} = \left\{\eta \,\middle|\, \exp(a(\eta)) = \int \exp(\eta^T t(\boldsymbol{z}))h(\boldsymbol{z})d\boldsymbol{z} < \infty\right\}$

$D_{KL}[p(\mathbf{z})\|q(\mathbf{z};\eta)]$ を整理してみる.

$$D_{KL}[p(\mathbf{z})\|q(\mathbf{z};\eta)] = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z};\eta)} d\mathbf{z}$$

$$= \int p(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z} - \int p(\mathbf{z}) \log q(\mathbf{z};\eta) d\mathbf{z}$$

$$= -\mathbb{E}_p[\log q(\mathbf{z};\eta)] + \underbrace{\mathbb{E}_p[\log p(\mathbf{z})]}_{\text{const.}}$$

$$= -\mathbb{E}_p[\log h(\mathbf{z}) + \eta^T \mathbf{t}(\mathbf{z}) - a(\eta)] + \text{const.}$$

$$= \underbrace{-\mathbb{E}_p[\log h(\mathbf{z})]}_{\text{const.}} - \eta^T \mathbb{E}_p[\mathbf{t}(\mathbf{z})] + a(\eta) + \text{const.}$$

$$= -\eta^T \mathbb{E}_p[\mathbf{t}(\mathbf{z})] + a(\eta) + \text{const.}$$

よって. $\eta^* = \underset{\eta}{\operatorname{argmin}} \left(-\eta^T \mathbb{E}_p[\mathbf{t}(\mathbf{z})] + a(\eta)\right)$.

[Claim] $-\eta^T \mathbb{E}_p[\mathbf{t}(\mathbf{z})] + a(\eta)$ は η について凸関数.

pf. $-\eta^T \mathbb{E}_p[\mathbf{t}(\mathbf{z})]$ は η について線形なので凸関数.

$\nabla_\eta^2 a(\eta) = \mathbb{V}_p[\mathbf{t}(\mathbf{z})] \succ \underset{\text{正定値.}}{0}$ なので凸関数. 凸関数の和は, やはり凸関数. ▨

[Claim] 自然パラメータ η の動ける空間

$$\mathcal{H} = \left\{\eta \,\middle|\, \exp(a(\eta)) = \int \exp(\eta^T \mathbf{t}(\mathbf{z})) h(\mathbf{z}) d\mathbf{z} < \infty \right\}$$

は凸集合.

pf. $^\forall \eta_1, \eta_2 \in \mathcal{H}$, $^\forall \alpha \in (0.1)$ として. $\alpha\eta_1 + (1-\alpha)\eta_2 \in \mathcal{H}$ を示す.

$$\int \exp\left((\alpha\eta_1 + (1-\alpha)\eta_2)^T \mathbf{t}(\mathbf{z})\right) h(\mathbf{z}) d\mathbf{z}$$

$$= \int \exp(\alpha\eta_1^T \mathbf{t}(\mathbf{z})) \exp((1-\alpha)\eta_2^T \mathbf{t}(\mathbf{z})) h(\mathbf{z}) d\mathbf{z}$$

$$= \int \left(\exp(\eta_1^T \mathbf{t}(\mathbf{z}))\right)^\alpha \left(\exp(\eta_2^T \mathbf{t}(\mathbf{z}))\right)^{1-\alpha} h(\mathbf{z}) d\mathbf{z}$$

$$\leq \left(\int \exp(\boldsymbol{\eta}_1^T \mathbf{t}(\mathbf{z})) h(\mathbf{z}) d\mathbf{z}\right)^{\alpha} \left(\int \exp(\boldsymbol{\eta}_2^T \mathbf{t}(\mathbf{z})) h(\mathbf{z}) d\mathbf{z}\right)^{1-\alpha} \quad (\because \text{Hölderの不等式})$$

$$< \infty.$$

∎

以上より，この問題は凸最適化問題なので，極値が最小値を与える．

$$\therefore -\mathbb{E}_{p(\mathbf{z})}[\mathbf{t}(\mathbf{z})] + \nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta})\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} = 0.$$

$\nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta}) = \mathbb{E}_q[\mathbf{t}(\mathbf{z})]$

$$\Leftrightarrow \mathbb{E}_{q(\mathbf{z};\boldsymbol{\eta}^*)}[\mathbf{t}(\mathbf{z})] = \mathbb{E}_{p(\mathbf{z})}[\mathbf{t}(\mathbf{z})]$$

結局，q を指数型分布族に制限した際の最適な q は，十分統計量 $\mathbf{t}(\mathbf{z})$ の p による平均と q による平均が一致するようなパラメータ $\boldsymbol{\eta}^*$ を選んだときの $q(\mathbf{z};\boldsymbol{\eta}^*)$ になる．

このようにして近似する手法をモーメントマッチング (moment matching) という．

Remark ・$p(\mathbf{z})$ のモーメントとは $\mathbf{z} \sim p(\mathbf{z})$ のときの $\mathbb{E}_p[\mathbf{z}]$, $\mathbb{E}_p[\mathbf{z}\mathbf{z}^T]$ などの量をいう

$\mathbb{E}_{p(\mathbf{z})}[\mathbf{t}(\mathbf{z})]$ は，$p(\mathbf{z})$ のモーメントではないのだが，指数型分布族 $q(\mathbf{z};\boldsymbol{\eta})$ に対し，

$\boldsymbol{\mu} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\eta}^*)}[\mathbf{t}(\mathbf{z})]$ のことを「モーメントパラメータ」と呼ぶ．実際，$\mathbf{t} = \mathbf{t}(\mathbf{z})$ の分布は

cf.) 吉田『数理統計学』

$q(\mathbf{t};\boldsymbol{\eta}) = \tilde{h}(\mathbf{t}) \exp(\boldsymbol{\eta}^T \mathbf{t} - a(\boldsymbol{\eta}))$ であり，$\mathbb{E}_{q(\mathbf{t};\boldsymbol{\eta})}[\mathbf{t}] = \nabla_{\boldsymbol{\eta}} a(\boldsymbol{\eta}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\eta})}[\mathbf{t}(\mathbf{z})]$ なので

$\boldsymbol{\mu}$ は $q(\mathbf{t};\boldsymbol{\eta})$ の1次のモーメント．「モーメントマッチング」とはモーメントパラメータをあわせること，と考えられそう．

4.2.4.2 仮定密度フィルタリング．

制御分野での呼び方だとか．
"online Bayesian learning", "weak marginalization" とも．

・仮定密度フィルタリング (assumed density filtering)

データ $\mathcal{D}_1, \mathcal{D}_2, \ldots$ を逐次的に学習する．

パラメータ θ の事前分布を $p(\theta)$ とすると，データ $\mathcal{D}_1$ を観測後の事後分布は，

$p(\theta|\mathcal{D}_1) \propto p(\mathcal{D}_1|\theta)p(\theta)$ となる．尤度関数 $p(\mathcal{D}_1|\theta)$ と $p(\theta)$ が共役なら，

$p(\theta|\mathcal{D}_1),\ p(\theta|\mathcal{D}_1,\mathcal{D}_2),\ p(\theta|\mathcal{D}_1,\mathcal{D}_2,\mathcal{D}_3),\ \ldots$

は全て $p(\theta)$ と同じ形の分布で解析的に計算できる.

・共役でない状況を考える. $p(\theta|\mathcal{D}_1)$ に対する近似分布として, $p(\theta)$ と同じ分布になるよう $q_1(\theta)$ を設定し,

$$q_1(\theta) \approx r_1(\theta) = \frac{1}{Z_1} p(\mathcal{D}_1|\theta)p(\theta), \quad Z_1 = \int p(\mathcal{D}_1|\theta)p(\theta)d\theta.$$

と近似する. q_1 は $q_1(\theta) = \operatorname*{argmin}_{q_1} D_{KL}[r_1(\theta) \| q_1(\theta)]$ により決定する.

p が指数型分布族のとき, これはモーメントマッチングで できる.

以降も,

$$q_i(\theta) \approx r_i(\theta) = \frac{1}{Z_i} p(\mathcal{D}_i|\theta)q_{i-1}(\theta), \quad Z_i = \int p(\mathcal{D}_i|\theta)q_{i-1}(\theta)d\theta,$$

$$q_i(\theta) = \operatorname*{argmin}_{q_i} D_{KL}[r_i(\theta) \| q_i(\theta)]$$

と近似していく. こうすることで, $q_i(\theta)$ たちは同じ分布を保ったまま逐次的に事後分布を更新していける.

以降 $f_i(\theta) = p(\mathcal{D}_i|\theta)$ と書く.

4.2.4.3 1次元Gauss分布の例. $\mathcal{N}(\theta|\mu_i,v_i) = \exp\left(-\frac{1}{2v_i}\theta^2 + \frac{\mu_i}{v_i}\theta - \frac{\mu_i^2}{2v_i} - \frac{1}{2}\log 2\pi v_i\right).$

・$q_i(\theta) = \mathcal{N}(\theta|\mu_i, v_i)$ の場合. これは指数型分布族なのでモーメントマッチングでできる.

十分統計量は $u(\theta) = \begin{pmatrix}\theta \\ \theta^2\end{pmatrix}$. $r_{i+1}(\theta)$ の正規化定数は,

$$Z_{i+1} = \int f_{i+1}(\theta) q_i(\theta)\, d\theta.$$

$$= \int f_{i+1}(\theta) \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(\theta-\mu_i)^2\right) d\theta.$$

まず,

$$\frac{\partial}{\partial \mu_i} \log Z_{i+1}$$

$$= \underbrace{\frac{1}{Z_{i+1}} \int f_{i+1}(\theta) \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(\theta-\mu_i)^2\right)}_{= r_{i+1}(\theta)} \left(\frac{\theta-\mu_i}{v_i}\right) d\theta$$

$$= \frac{1}{v_i} \int r_{i+1}(\theta)(\theta-\mu_i)\, d\theta$$

$$= \frac{1}{v_i}\left(\mathbb{E}_{r_{i+1}(\theta)}[\theta] - \mu_i\right).$$

$$\therefore \mathbb{E}_{r_{i+1}(\theta)}[\theta] = \mu_i + v_i \frac{\partial}{\partial \mu_i} \log Z_{i+1}.$$

続いて,

$$\frac{\partial}{\partial v_i} \log Z_{i+1}$$

$$= \frac{1}{Z_{i+1}} \int f_{i+1}(\theta) \frac{1}{\sqrt{2\pi}} \left(-\frac{1}{2}\frac{1}{v_i\sqrt{v_i}} + \frac{1}{\sqrt{v_i}}\frac{(\theta-\mu_i)^2}{2v_i^2}\right) \exp\left(-\frac{1}{2v_i}(\theta-\mu_i)^2\right) d\theta$$

$$= \int r_{i+1}(\theta)\left(-\frac{1}{2v_i} + \frac{(\theta-\mu_i)^2}{2v_i^2}\right) d\theta$$

$$= \frac{1}{2v_i^2} \int r_{i+1}(\theta)\left(-v_i + \theta^2 - 2\mu_i\theta + \mu_i^2\right) d\theta$$

$$= -\frac{1}{2v_i} + \frac{1}{2v_i^2}\left(\mathbb{E}_{r_{i+1}(\theta)}[\theta^2] - 2\mu_i \mathbb{E}_{r_{i+1}(\theta)}[\theta] + \mu_i^2\right).$$

$$\therefore \mathbb{E}_{r_{i+1}(\theta)}[\theta^2] = v_i - \mu_i^2 + 2\mu_i \mathbb{E}_{r_{i+1}(\theta)}[\theta] + 2v_i^2 \frac{\partial}{\partial v_i} \log Z_{i+1}.$$

モーメントマッチングにより,

$$\mathbb{E}_{q_{i+1}(\theta)}[\theta] = \mathbb{E}_{r_{i+1}(\theta)}[\theta], \quad \mathbb{E}_{q_{i+1}(\theta)}[\theta^2] = \mathbb{E}_{r_{i+1}(\theta)}[\theta^2]$$

とすればよい.

$\mathbb{E}_{q_{i+1}(\theta)}[\theta] = \mu_{i+1}$ なので, $\mu_{i+1} = \mu_i + v_i \frac{\partial}{\partial \mu_i} \log Z_{i+1}.$

$\mathbb{E}_{q_{i+1}(\theta)}[\theta^2] = v_{i+1} + \mu_{i+1}^2$ なので,

$$v_{i+1} = v_i - (\mu_{i+1} - \mu_i)^2 + 2v_i^2 \frac{\partial}{\partial v_i} \log Z_{i+1}$$

$$= v_i - v_i^2 \left(\left(\frac{\partial}{\partial \mu_i} \log Z_{i+1} \right)^2 - 2 \frac{\partial}{\partial v_i} \log Z_{i+1} \right).$$

各正規化定数 Z_i が計算できれば、このように逐次更新ができる.

4.2.4.4 ガンマ分布の例.

$$q_i(\theta) = \mathrm{Gam}(\theta | a_i, b_i) = \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i - 1} e^{-b_i \theta}$$

$$= \exp\left((a_i - 1) \log \theta - b_i \theta + a_i \log b_i - \log \Gamma(a_i) \right)$$

の場合. 指数型分布族. 十分統計量は $t(\theta) = \begin{pmatrix} \log \theta \\ \theta \end{pmatrix}$.

$r_{i+1}(\theta)$ の正規化定数は,

$$Z_{i+1}(a_i, b_i) = \int f_{i+1}(\theta) \mathrm{Gam}(\theta | a_i, b_i) d\theta.$$

$\mathbb{E}_{r_{i+1}(\theta)}[\theta]$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int \theta f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i - 1} e^{-b_i \theta} d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i} e^{-b_i \theta} d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \frac{\Gamma(a_i + 1)}{b_i \Gamma(a_i)} Z_{i+1}(a_i + 1, b_i)$$

$\Gamma(a_i + 1) = a_i \Gamma(a_i)$

$$= \frac{a_i Z_{i+1}(a_i + 1, b_i)}{b_i Z_{i+1}(a_i, b_i)}.$$

$\mathbb{E}_{r_{i+1}(\theta)}[\log \theta]$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int \log \theta \, f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i - 1} e^{-b_i \theta} d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \left(\frac{\partial}{\partial a_i} \theta^{a_i - 1} \right) e^{-b_i \theta} d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \frac{b_i^{a_i}}{\Gamma(a_i)} \frac{\partial}{\partial a_i} \int f_{i+1}(\theta) \theta^{a_i - 1} e^{-b_i \theta} d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \frac{b_i^{a_i}}{\Gamma(a_i)} \frac{\partial}{\partial a_i} \left(Z_{i+1}(a_i, b_i) \frac{\Gamma(a_i)}{b_i^{a_i}} \right)$$

$$= \frac{\partial}{\partial a_i} \log \left(Z_{i+1}(a_i, b_i) \frac{\Gamma(a_i)}{b_i^{a_i}} \right)$$

$$= \frac{\partial}{\partial a_i} \log Z_{i+1}(a_i, b_i) + \frac{d}{d a_i} \log \Gamma(a_i) - \log b_i .$$

ここで，ディガンマ函数 $\psi(a) = \frac{d}{da} \log \Gamma(a)$ を使うと，

$$\mathbb{E}_{r_{i+1}(\theta)}[\log \theta] = \frac{\partial}{\partial a_i} \log Z_{i+1}(a_i, b_i) + \psi(a_i) - \log b_i .$$

モーメントマッチングをすると，

$$\mathbb{E}_{q_{i+1}(\theta)}[\theta] = \frac{a_{i+1}}{b_{i+1}} = \frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)}$$

$$\mathbb{E}_{q_{i+1}(\theta)}[\log \theta] = \psi(a_{i+1}) - \log b_{i+1} = \frac{\partial}{\partial a_i} \log Z_{i+1}(a_i, b_i) + \psi(a_i) - \log b_i .$$

となる．この更新は数値的に行うしかない．

この妥当性は不明．
もちろんKL-divergenceは最小にならない．

・解析的な更新式を得るため，$\log \theta$ の代わりに θ^2 の平均をあわせることにする．

$$\mathbb{E}_{r_{i+1}(\theta)}[\theta^2]$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int \theta^2 f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i - 1} e^{-b_i \theta} \, d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \int f_{i+1}(\theta) \frac{b_i^{a_i}}{\Gamma(a_i)} \theta^{a_i + 1} e^{-b_i \theta} \, d\theta$$

$$= \frac{1}{Z_{i+1}(a_i, b_i)} \frac{\Gamma(a_i+2)}{b_i^2 \Gamma(a_i)} Z_{i+1}(a_i+2, b_i)$$

$$= \frac{(a_i+1) a_i Z_{i+1}(a_i+2, b_i)}{b_i^2 Z_{i+1}(a_i, b_i)} .$$

更新のためには，以下のようにする．

$$\mathbb{E}_{q_{i+1}(\theta)}[\theta] = \frac{a_{i+1}}{b_{i+1}} = \frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)} ,$$

$$\mathbb{E}_{q_{i+1}(\theta)}[\theta^2] = \frac{a_{i+1}}{b_{i+1}^2} = \frac{(a_i+1) a_i Z_{i+1}(a_i+2, b_i)}{b_i^2 Z_{i+1}(a_i, b_i)} - \left(\frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)} \right)^2 .$$

$$a_{i+1} = \frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)} b_{i+1}$$ を第2式に代入し、

$$\frac{1}{b_{i+1}} \frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)} = \frac{(a_i+1) a_i Z_{i+1}(a_i+2, b_i)}{b_i^2 Z_{i+1}(a_i, b_i)} - \left(\frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)} \right)^2$$

$$\frac{1}{b_{i+1}} a_i Z_{i+1}(a_i+1, b_i) = \frac{(a_i+1) a_i Z_{i+1}(a_i+2, b_i)}{b_i} - \frac{a_i^2 Z_{i+1}(a_i+1, b_i)^2}{b_i Z_{i+1}(a_i, b_i)}$$

$$\frac{1}{b_{i+1}} = \frac{(a_i+1) Z_{i+1}(a_i+2, b_i)}{b_i Z_{i+1}(a_i+1, b_i)} - \frac{a_i Z_{i+1}(a_i+1, b_i)}{b_i Z_{i+1}(a_i, b_i)}$$

$$= \frac{1}{b_i} \left(\frac{(a_i+1) Z_{i+1}(a_i+2, b_i)}{Z_{i+1}(a_i+1, b_i)} - \frac{a_i Z_{i+1}(a_i+1, b_i)}{Z_{i+1}(a_i, b_i)} \right).$$

$$\therefore b_{i+1} = b_i \left(\frac{(a_i+1) Z_{i+1}(a_i+2, b_i)}{Z_{i+1}(a_i+1, b_i)} - \frac{a_i Z_{i+1}(a_i+1, b_i)}{Z_{i+1}(a_i, b_i)} \right)^{-1}.$$

$$a_{i+1} = a_i \frac{Z_{i+1}(a_i+1, b_i)}{Z_{i+1}(a_i, b_i)} \left(\frac{(a_i+1) Z_{i+1}(a_i+2, b_i)}{Z_{i+1}(a_i+1, b_i)} - \frac{a_i Z_{i+1}(a_i+1, b_i)}{Z_{i+1}(a_i, b_i)} \right)^{-1}$$

$$= a_i \left(\frac{(a_i+1) Z_{i+1}(a_i+2, b_i) Z_{i+1}(a_i, b_i)}{Z_{i+1}(a_i+1, b_i)^2} - a_i \right)^{-1}.$$

各正規化定数 $Z(a_i, b_i)$ が計算できれば、このような更新ができる。

4.2.5 例：モーメントマッチングによるプロビット回帰モデルの学習。

probability unit からきている

・プロビット回帰 (probit regression) モデル：

2値分類のための一般化線形モデルの一つ。モデルによる出力値は分類確率。

・リンク関数はプロビット関数 (probit func.)：

$$\mathrm{probit}(p) = \Phi^{-1}(p), \quad p \in (0,1)$$ ← Φ は標準正規分布の cdf.

cf.) ロジスティック回帰も同様のモデル。こちらはリンク関数がロジット関数 (logit func.)

ロジスティック関数の逆関数 「対数オッズ比」ともいう。

$$\mathrm{logit}(p) = \log \frac{p}{1-p}, \quad p \in (0,1)$$

の一般化線形モデル。

ロジスティック回帰よりもプロビット回帰の方が外れ値に敏感らしい。(PRML Ch 4.3)

- 入力 : $\mathcal{X} = \{x_1, \dots, x_N\} \subseteq \mathbb{R}$,　← 多次元でもできる

クラス : $\mathcal{Y} = \{y_1, \dots, y_N\}$

プロビットモデル : $p(y|x, w) = \Phi(ywx)$.　($w \in \mathbb{R}$ はパラメータ)

尤度函数 : $p(\mathcal{Y}|\mathcal{X}, w) = \prod_{n=1}^{N} p(y_n|x_n, w) = \prod_{n=1}^{N} \Phi(y_n w x_n)$.

w の事前分布 $p(w) = \mathcal{N}(w|0, v_0)$　($v_0 > 0$ はハイパーパラメータ)

- モデルの周辺尤度は

$$p(\mathcal{Y}|\mathcal{X}) = \int p(\mathcal{Y}|\mathcal{X}, w) p(w) dw.$$

これは解析的に計算できないので, データを一つずつ追加しながら

仮定密度フィルタリングをする.

- $f_i(w) = p(y_i|x_i, w)$ とし, パラメータ w の事後分布を

$$q_i(w) = \mathcal{N}(w|\mu_i, v_i)$$

で近似する. $\mu_0 = 0$, $q_0(w) = p(w) = \mathcal{N}(w|\mu_0, v_0)$ とする.

i ステップ目の更新は

$$q_{i+1}(w) \approx \frac{1}{Z_{i+1}} f_{i+1}(w) q_i(w)$$

をモーメントマッチングで行えばよい.

- 正規化定数

$$Z_{i+1} = \int p(y_{i+1}|x_{i+1}, w) \mathcal{N}(w|\mu_i, v_i) dw$$

は解析的に計算することができる.

- $x_{i+1} = 0$ のとき, $p(y_{i+1} | x_{i+1}, w) = \Phi(0) = \frac{1}{2}$.

  このとき $Z_{i+1} = \int \Phi(0) \mathcal{N}(w | \mu_i, v_i) dw = \Phi(0) = \frac{1}{2}$.

- $x_{i+1} \neq 0$ のとき, $y_{i+1} \in \{\pm 1\}$ なので $x_{i+1} y_{i+1} \neq 0$.

$$Z_{i+1}$$

$$= \int_{-\infty}^{\infty} \Phi(y_{i+1} w x_{i+1}) \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(w-\mu_i)^2\right) dw$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(w-\mu_i)^2\right) \left(\int_{-\infty}^{y_{i+1} w x_{i+1}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right) dt \right) dw$$

$s = \frac{t}{x_{i+1} y_{i+1}}$ とおく.

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(w-\mu_i)^2\right) \left(\int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_{i+1}^2 y_{i+1}^2 s^2\right) x_{i+1} y_{i+1} \, ds \right) dw$$

$u = w - \mu_i$ とおく.

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i} u^2\right) \left(\int_{-\infty}^{u+\mu_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_{i+1}^2 y_{i+1}^2 s^2\right) x_{i+1} y_{i+1} \, ds \right) du$$

$v = s - u$ とおく.

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i} u^2\right) \left(\int_{-\infty}^{\mu_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_{i+1}^2 y_{i+1}^2 (u+v)^2\right) x_{i+1} y_{i+1} \, dv \right) du$$

積分順序の変更

$$= \int_{-\infty}^{\mu_i} \frac{1}{\sqrt{2\pi}} x_{i+1} y_{i+1} \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i} u^2\right) \exp\left(-\frac{1}{2} x_{i+1}^2 y_{i+1}^2 \underline{(u+v)^2}\right) du \right) dv$$

展開

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1} y_{i+1}}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{u^2}{2v_i}\right) \exp\left(-\frac{1}{2} x_{i+1}^2 y_{i+1}^2 (\underline{u^2 + 2uv + v^2})\right) du \right) dv$$

展開

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1} y_{i+1}}{\sqrt{2\pi}} \exp\left(-\frac{x_{i+1}^2 y_{i+1}^2 v^2}{2}\right) \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}\left((1+\underbrace{v_i x_{i+1}^2 y_{i+1}^2}_{=: C_{i+1}}) u^2 + 2 v_i x_{i+1}^2 y_{i+1}^2 uv\right)\right) du \right) dv$$

平方完成

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1} y_{i+1}}{\sqrt{2\pi}} \exp\left(-\frac{x_{i+1}^2 y_{i+1}^2 v^2}{2}\right) \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1+C_{i+1}}{2v_i}\left(u + \frac{C_{i+1} v}{1+C_{i+1}}\right)^2 + \frac{C_{i+1}^2 v^2}{2v_i(1+C_{i+1})}\right) du \right) dv$$

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1} y_{i+1}}{\sqrt{2\pi}} \exp\left(-\frac{x_{i+1}^2 y_{i+1}^2 v^2}{2} + \frac{C_{i+1}^2 v^2}{2v_i(1+C_{i+1})}\right) \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1+C_{i+1}}{2v_i}\left(u + \frac{C_{i+1} v}{1+C_{i+1}}\right)^2\right) du \right) dv.$$

ここで,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1+C_{i+1}}{2v_i}\left(u + \frac{C_{i+1} v}{1+C_{i+1}}\right)^2\right) du$$

$$= \frac{1}{\sqrt{1+C_{i+1}}} \int_{-\infty}^{\infty} \sqrt{\frac{1+C_{i+1}}{2\pi v_i}} \exp\left(-\frac{1+C_{i+1}}{2v_i}\left(u + \frac{C_{i+1} v}{1+C_{i+1}}\right)^2\right) du$$

$$= \frac{1}{\sqrt{1+C_{i+1}}} \int_{-\infty}^{\infty} \mathcal{N}\left(u \,\middle|\, -\frac{C_{i+1} v}{1+C_{i+1}}, \frac{v_i}{1+C_{i+1}}\right) du = \frac{1}{\sqrt{1+C_{i+1}}}.$$

よって，

$$Z_{i+1}$$

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1}y_{i+1}}{\sqrt{2\pi}} \exp\left(-\frac{x_{i+1}^2 y_{i+1}^2 v^2}{2} + \frac{c_{i+1}^2 v^2}{2v_i(1+c_{i+1})}\right) \frac{1}{\sqrt{1+c_{i+1}}}\, dv$$

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1}y_{i+1}}{\sqrt{2\pi(1+c_{i+1})}} \exp\left(-\frac{1}{2}\cdot\frac{1}{v_i}\left(c_{i+1} - \frac{c_{i+1}^2}{1+c_{i+1}}\right)v^2\right) dv$$

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1}y_{i+1}}{\sqrt{2\pi(1+c_{i+1})}} \exp\left(-\frac{1}{2}\frac{c_{i+1}}{v_i(1+c_{i+1})}v^2\right) dv.$$

$$= \int_{-\infty}^{\mu_i} \frac{x_{i+1}y_{i+1}}{\sqrt{2\pi(1+c_{i+1})}} \exp\left(-\frac{1}{2}\frac{x_{i+1}^2 y_{i+1}^2}{1+c_{i+1}}v^2\right) dv$$

$z = \frac{x_{i+1}y_{i+1}}{\sqrt{1+c_{i+1}}} v = d_{i+1} v$ とおくと，

$$Z_{i+1}$$

$$= \int_{-\infty}^{d_{i+1}\mu_i} \frac{\cancel{d_{i+1}}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)\cdot\frac{1}{\cancel{d_{i+1}}}\, dz$$

$$= \Phi(d_{i+1}\mu_i)$$

$$= \Phi\left(\frac{x_{i+1}y_{i+1}\mu_i}{\sqrt{1+v_i x_{i+1}^2 y_{i+1}^2}}\right).$$

これは $x_{i+1}=0$ のときも定義され値が一致するので，任意の $x_{i+1}\in\mathbb{R}$ で成立．

以上より，$a_{i+1} := \frac{x_{i+1}y_{i+1}\mu_i}{\sqrt{1+v_i x_{i+1}^2 y_{i+1}^2}}$ とすると $Z_{i+1} = \Phi(a_{i+1})$.

モーメントマッチングを行う．

$$\frac{\partial}{\partial\mu_i}\log Z_{i+1}$$

$$= Z_{i+1}^{-1}\frac{\partial}{\partial\mu_i}Z_{i+1}$$

$$= \Phi(a_{i+1})^{-1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}a_{i+1}^2\right)\frac{\partial}{\partial\mu_i}a_{i+1}$$

$$= \Phi(a_{i+1})^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} a_{i+1}^2\right) \frac{a_{i+1}}{\mu_i}.$$

$$\frac{\partial}{\partial v_i} \log Z_{i+1}$$

$$= Z_{i+1}^{-1} \frac{\partial}{\partial v_i} Z_{i+1}$$

$$= \Phi(a_{i+1})^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} a_{i+1}^2\right) \frac{\partial}{\partial v_i} a_{i+1}$$

$$= -\frac{1}{2} \Phi(a_{i+1})^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} a_{i+1}^2\right) \frac{x_{i+1}^3 y_{i+1}^3 \mu_i}{\left(\sqrt{1+v_i x_{i+1}^2 y_{i+1}^2}\right)^3}$$

$$= -\frac{1}{2} \Phi(a_{i+1})^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} a_{i+1}^2\right) \frac{a_{i+1}^3}{\mu_i^2}.$$

よって更新式は,

$$\mu_{i+1} = \mu_i + v_i \frac{\partial}{\partial \mu_i} \log Z_{i+1} = \mu_i + \frac{1}{\sqrt{2\pi}} \frac{v_i a_{i+1}}{\mu_i} \Phi(a_{i+1})^{-1} \exp\left(-\frac{1}{2} a_{i+1}^2\right).$$

$$v_{i+1} = v_i - v_i^2 \left(\left(\frac{\partial}{\partial \mu_i} \log Z_{i+1}\right)^2 - 2 \frac{\partial}{\partial v_i} \log Z_{i+1}\right)$$

$$= v_i - v_i^2 \left(\frac{1}{2\pi} \Phi(a_{i+1})^{-2} \exp\left(-a_{i+1}^2\right) \frac{a_{i+1}^2}{\mu_i^2} + \Phi(a_{i+1})^{-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} a_{i+1}^2\right) \frac{a_{i+1}^3}{\mu_i^2}\right)$$

$$= v_i - \frac{v_i^2 a_{i+1}^2}{\mu_i^2} \Phi(a_{i+1})^{-2} \left(\frac{1}{2\pi} \exp\left(-a_{i+1}^2\right) + \frac{a_{i+1}}{\sqrt{2\pi}} \Phi(a_{i+1}) \exp\left(-\frac{1}{2} a_{i+1}^2\right)\right)$$

4.2.6 期待値伝播法.

- 仮定密度フィルタリング: 一度学習したデータは捨ててしまい, 再び学習に使わない.

  ✓ メモリ効率は良い

  ✗ 逐次的に入ってくるデータの順序に近似結果が強く依存.

- 期待値伝播法 (expectation propagation)

  仮定密度フィルタリングをバッチ学習もできるように一般化した手法.

  最適化の過程で同じデータを何度も用いることで, 精度の高い近似ができる.

- パラメータ θ に事前分布 $p(\theta)$ を設定し，各 x_n が $p(x_n|\theta)$ に従うとする．

考えるモデルは，

$$p(\mathcal{X}, \theta) = p(\theta) \prod_{n=1}^{N} p(x_n|\theta).$$

ここで，

$$f_n(\theta) := \begin{cases} p(\theta) & (n=0) \\ p(x_n|\theta) & (n=1, \ldots, N) \end{cases}$$

とすると，$p(\mathcal{X}, \theta) = \prod_{n=0}^{N} f_n(\theta)$ と表せる．f_n を因子 (factor) という．

- 事後分布は，

$$p(\theta|\mathcal{X}) = \frac{1}{p(\mathcal{X})} p(\mathcal{X}, \theta) = \frac{1}{p(\mathcal{X})} \prod_{n=0}^{N} f_n(\theta).$$

この事後分布の近似を，次のように近似因子 $\tilde{f}_n$ の積で表す：

$$p(\theta|\mathcal{X}) \approx q(\theta) = \frac{1}{Z} \prod_{n=0}^{N} \tilde{f}_n(\theta). \qquad (Z：正規化定数)$$

$\tilde{f}_n$ を指数型分布族の分布にしておけば，q も指数型分布族の分布になる．

- $q(\theta)$ が $p(\theta|\mathcal{X})$ をよく近似するよう，各 $\tilde{f}_n$ を逐次的に更新する．

$q(\theta)$ のパラメータを適当に初期化しておく．

現在の $q(\theta)$ を $q_{old}(\theta)$ と書く．$\tilde{f}_n$ の更新を行う．

まず，$q_{old}(\theta)$ から n 番目の因子を取り除く：

$$q_{\backslash n}(\theta) = \frac{q_{old}(\theta)}{\tilde{f}_n(\theta)}$$

$\tilde{f}_n$ の代わりに f_n を用いる分布を $r(\theta) = \frac{1}{Z_n} f_n(\theta) q_{\backslash n}(\theta)$ とし，

新たな近似分布 $q_{new}(\theta)$ を次をみたすように更新:

$$q_{new}(\theta) = \underset{q}{\mathrm{argmin}}\, D_{KL}[r(\theta) \| q(\theta)].$$

これは、q が指数型分布族なのでモーメントマッチングでできる。

最後に、$\tilde{f}_n$ を次のように更新する。

$$\tilde{f}_n(\theta) \leftarrow Z_n \frac{q_{new}(\theta)}{q_{\backslash n}(\theta)}.$$

以上の更新を $n = 0, \dots, N$ に対して繰り返し行う。

- 以上の計算は、各 Z_n が計算できるならば実行可能
- 収束性については理論的に保証されてはいないが、実験的には良い性能を示す。