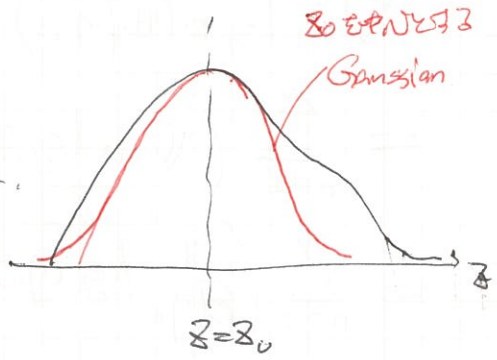
4.4 Laplace近似

○目的

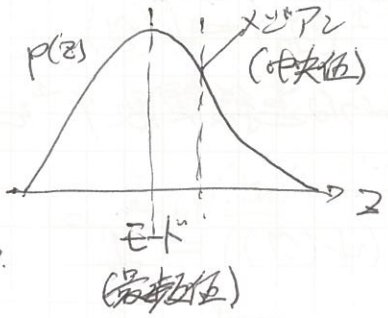
確率密度分布 $p(z)$ を Gaussian で近似する.



○ 1変数の case

仮定: $p(z) = \frac{1}{Z} f(z)$ (4.125)

正規化係数 $Z = \int dz\, f(z)$



・モード … 最頻値

$p(z)$ は $z=z_0$ をモードにもつ

$\Longleftrightarrow$ $p(z)$ は $z=z_0$ で最大値をもつ.

$\Longrightarrow \left.\frac{dp}{dz}\right|_{z=z_0} = 0$ (? ; z は定数.)

$\Longleftrightarrow \left.\frac{df(z)}{dz}\right|_{z=z_0} = 0$

・Gaussian の対数は2次関数.

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad (2.42)$$

$$\log \mathcal{N}(x|\mu,\sigma^2) = 定 - \frac{1}{2\sigma^2}(x-\mu)^2 + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)$$

$$= \left(-\frac{1}{2\sigma^2}\right)x^2 + \frac{\mu}{\sigma^2}x + \left(-\frac{\mu^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right)$$

$$= (x\text{の2次関数})$$

つまり、$\log f(z)$ が $z=z_0$ 近傍で z の2次関数であればよい。

Taylor展開

$$\log f(z) = \log f(z_0) + \underbrace{\frac{d}{dz}\log f(z)\Big|_{z=z_0}}_{\frac{1}{f(z)}\frac{df(z)}{dz}\Big|_{z=z_0}=0}(z-z_0) + \frac{1}{2}\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0}(z-z_0)^2 + O((z-z_0)^3)$$

$$A = -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0} \quad (4.128)$$

$$\simeq \log f(z_0) - \frac{1}{2}A(z-z_0)^2$$

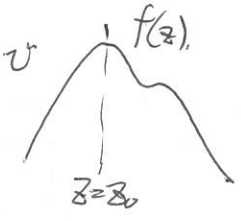
↓ 両辺の対数を比較

$$f(z) \simeq f(z_0)\exp\left(-\frac{A}{2}(z-z_0)^2\right) \quad (4.129)$$

・$f(z_0)$ を求める。(正規化)

仮定: $A \overset{(4.128)}{=} -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0} > 0$

$f(z)$ が $z=z_0$ で局所極大(上凸) ~~(4.128)~~



$$1 = \int_{-\infty}^{\infty} dz\, f(z) = f(z_0)\underbrace{\int_{-\infty}^{\infty} dz \exp\left(-\frac{A}{2}(z-z_0)^2\right)}_{\sqrt{\frac{A}{2}}(z-z_0)=x \text{ とおく}}$$

$$= \sqrt{\frac{2\pi}{A}} f(z_0) \underbrace{\sqrt{\frac{2}{A}}\int_{-\infty}^{\infty} dx e^{-x^2}}_{\sqrt{\pi}}$$

$$\therefore f(z_0) = \sqrt{\frac{A}{2\pi}} \qquad \sqrt{\frac{2\pi}{A}}$$

よって、近似した分布 $q(z)$ は

$$q(z) = \sqrt{\frac{A}{2\pi}}\exp\left(-\frac{A}{2}(z-z_0)^2\right) \quad (4.130)$$

$$= \mathcal{N}(z|z_0, A^{-1})$$

$\sqrt{2\pi A^{-1}}^{-1}$ $\quad$ $\frac{1}{2A^{-1}}$

○M次元

分布： $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$

仮定：$p(\boldsymbol{z})$は $\boldsymbol{z} = \boldsymbol{z}_0$ でモードをもつ

$$\nabla f(\boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{z}_0} = 0$$

$\log f(\boldsymbol{z})$を $\boldsymbol{z} = \boldsymbol{z}_0$ まわりで Taylor展開する.

$$\log f(\boldsymbol{z}) \overset{(2次まで)}{\simeq} \log f(\boldsymbol{z}_0) + \underbrace{\nabla \log f(\boldsymbol{z}_0)}_{=0} (\boldsymbol{z}-\boldsymbol{z}_0) + \frac{1}{2} (\boldsymbol{z}-\boldsymbol{z}_0)^T \nabla\nabla \log f(\boldsymbol{z}_0) (\boldsymbol{z}-\boldsymbol{z}_0)$$

$$A = -\nabla\nabla \log f(\boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{z}_0} \quad (4.132)$$

ヘッセ行列

$$= \log f(\boldsymbol{z}_0) - \frac{1}{2} (\boldsymbol{z}-\boldsymbol{z}_0)^T A (\boldsymbol{z}-\boldsymbol{z}_0)$$

↓ 両辺の指数をとる.

$$f(\boldsymbol{z}) \simeq f(\boldsymbol{z}_0) \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)^T A (\boldsymbol{z}-\boldsymbol{z}_0)\right) \quad (4.133)$$

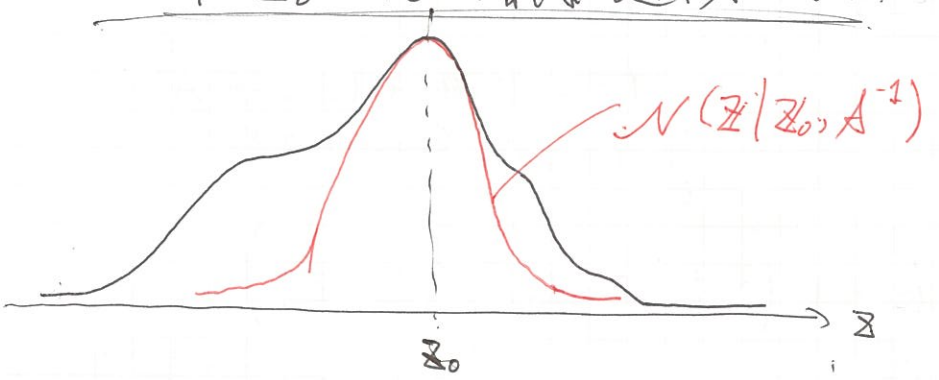
↓ 正規化 $\int d\boldsymbol{z}\, f(\boldsymbol{z}) = 1 \Rightarrow f(\boldsymbol{z}_0) = \frac{\sqrt{|A|}}{\sqrt{2\pi}^M}$

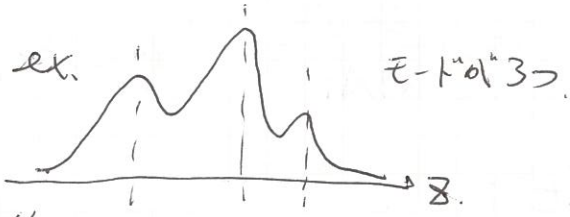
$$q(\boldsymbol{z}) = \frac{\sqrt{|A|}}{\sqrt{2\pi}^M} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}_0)^T A (\boldsymbol{z}-\boldsymbol{z}_0)\right)$$

$$= \mathcal{N}(\boldsymbol{z}|\boldsymbol{z}_0, A^{-1}) \quad (4.134)$$

つまり...

モード z_0 を見つければ近似できる.





○注意など

- 一般的に分布は多峰的
  → どのモードを使うので Laplace 近似が異なる.
- Laplace 近似を使う際、真の分布の正規化係数又は不要.
  ~~知る必要は~~
- 中心極限定理より、観測データが増えるほど
  モデルの事後確率分布は Gaussian に近づく.
  → Laplace 近似が有効
- Gaussian 及び 実数変数にしか適用できない.
- 真の分布のある1点（モード付近）の特性しか捉えられない. ← 局所的
  cf. 10章 ... 全体的な特性を捉える方法.

4.4.1 モデルの比較とBIC

○ 正規化係数Zの近似

$$Z = \int dz\, f(z)$$

$f(z)$ を平均値：モード z_0，分散 A^{-1}（←正定値ヘッセ行列）の Gaussian で近似 (4.133)

$$\simeq f(z_0) \int dz \exp\left(-\frac{1}{2}(z-z_0)^T A (z-z_0)\right)$$

$$= f(z_0)\frac{(2\pi)^{\frac{M}{2}}}{\sqrt{|A|}} \quad (4.135)$$

$z \in \mathbb{R}^M$ ← M：空間の次元の数

○ モデルエビデンス

L 個のモデル $\mathcal{M} = \{M_i\}_{i=1,\cdots,L}$ のなから
訓練データ $\mathcal{D}$ を最もよく説明できるモデルを選ぶ．

訓練データ $\mathcal{D}$ が与えられたときのモデルの事後確率

$$p(M_i|\mathcal{D}) \propto p(\mathcal{D}|M_i)\underline{p(M_i)}$$

どのモデルが選ばれるか？
モデルの不確かさ．

仮定：
$p(M_i) = \text{const.}$（離散型一様分布 ex. サイコロの出目の確率分布）

↓

~~モデルの事後確率は~~ $p(\mathcal{D}|M_i)$ で決まる． ← どのモデルが選ばれるかは $p(\mathcal{D}|M_i)$ で決まる．

〈model evidence〉

“モデル M_i の下で $\mathcal{D}$ がどれくらいよく説明されるか？”

モデル M_i がパラメータ $\{\theta_i\}$ をもつとき，モデルエビデンスは

$$p(\mathcal{D}|M_i) = \int d\theta_i\, p(\mathcal{D}|\theta_i, M_i)\, p(\theta_i|M_i) \quad (3.68)$$

~~(4.150)~~

で与えられる．

$$p(\mathcal{D}|M_i) = \int d\theta_i\, p(\mathcal{D}|\theta_i, M_i)\, p(\theta_i|M_i) \qquad (4.136)$$

$p(\mathcal{D}|M_i) = Z$

$p(\mathcal{D}|\theta_i, M_i)\, p(\theta_i|M_i) = f(\theta_i)$

(4.135)を適用後logをとる.

(演習4.22)

$$Z \simeq f(\theta_0)\frac{(2\pi)^{\frac{M}{2}}}{|A|} \qquad (4.135)$$

$\downarrow \log$

$$\log Z \simeq \log f(\theta_0) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A|$$

$$\log p(\mathcal{D}|M_i) \simeq \log p(\mathcal{D}|\theta_{MAP}, M_i) + \log p(\theta_{MAP}|M_i) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A| \qquad (4.137)$$

θ_{MAP}: 事後確率分布のモードでのθの値

最適なパラメータを使って評価した対数尤度

Occam係数

モデルの複雑さにペナルティーを科す.

where

$$A \overset{(4.132)}{=} -\nabla\nabla \log f(\theta_0)$$

$$= -\nabla\nabla \log p(\mathcal{D}|\theta_{MAP}, M_i)\, p(\theta_{MAP}|M_i)$$

$$= -\nabla\nabla \log p(\mathcal{D}|\theta_{MAP}, M_i) \qquad (4.138)$$

$p(\theta_{MAP}|M_i)$ 本の誤植

〈Hessian matrix〉

typo

$\nabla\nabla p(\theta_{MAP}|M_i) = 0$

?

$\nabla\nabla p(\theta_{MAP}|M_i) \neq 0$

$$= \nabla\nabla\left(-\log p(\mathcal{D}|\theta_{MAP}, M_i)\right) - \nabla\nabla \log p(\theta_{MAP}|M_i) \qquad (4.138)$$

$= H$ (演習4.23)

「正規化係数Zを近似することで
モデルエビデンスを計算できる.」

⑦ 式(4.139)の導出（演習(4.23)）

~~パラメータ上での事前分布 $p(\theta|\cdot)$~~

仮定：
$p(\theta|M_i)$ が Gaussian $\mathcal{N}(\theta|m, V_0)$.

$$\mathcal{N}(\theta|m, V_0) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\sqrt{|V_0|}} \exp\left(-\frac{1}{2}(\theta-m)^T V_0^{-1}(\theta-m)\right) \quad (2.43)$$

$$\log \mathcal{N}(\theta|m, V_0) = -\frac{1}{2}(\theta-m)^T V_0^{-1}(\theta-m) + \log\left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\sqrt{|V_0|}}\right)$$

$$\begin{aligned}(\langle\theta| - \langle m|)\hat{V}_0(|\theta\rangle - |m\rangle) &= \langle\theta|\hat{V}_0|\theta\rangle + \langle m|\hat{V}_0|m\rangle - \langle\theta|\hat{V}_0|m\rangle - \langle m|\hat{V}_0|\theta\rangle \\ \hat{V}_0^t = \hat{V}_0 \quad &= \langle\theta|\hat{V}_0|\theta\rangle - 2\langle\theta|\hat{V}_0|m\rangle + \langle m|\hat{V}_0|m\rangle \\ &\to \theta^T V_0^{-1}\theta - 2\theta^T V_0^{-1} m + m^T V_0^{-1} m\end{aligned}$$

$$= \underbrace{-\frac{1}{2}\theta^T V_0^{-1}\theta}_{\theta\text{の二次項}} + \underbrace{\theta^T V_0^{-1} m}_{\theta\text{の一次項}} \underbrace{- \frac{1}{2} m^T V_0^{-1} m + \log\left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\sqrt{|V_0|}}\right)}_{\text{定数項}}$$

$$\nabla \log \mathcal{N}(\theta|m, V_0) = -V_0^{-1}\theta + V_0^{-1} m.$$

$$\nabla\nabla \log \mathcal{N}(\theta|m, V_0) = -V_0^{-1}.$$

より、(4.138)は

$$\begin{aligned}A &= H - \nabla\nabla \log p(\theta_{MAP}|M_i) \\ &= H + V_0^{-1}.\end{aligned}$$

したがって、(4.137)は

$$\log p(\mathcal{D}|M_i) \quad \text{(4.137)}$$

$$\simeq \log p(\mathcal{D}|\theta_{MAP}, M_i) + \log p(\theta_{MAP}|M_i) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A|$$

① ② ③ ④

② $\mathcal{N}(\theta_{MAP}|m, V_0)$

$$= \log p(\mathcal{D}|\theta_{MAP}, M_i) \quad ①$$

$$-\frac{1}{2}(\theta_{MAP}-m)^T V_0^{-1}(\theta_{MAP}-m) + \underbrace{\log\left(\frac{1}{(2\pi)^{\frac{M}{2}}}\frac{1}{\sqrt{|V_0|}}\right)}_{\text{Const.}} \quad ②$$

$$+\underbrace{\frac{M}{2}\log(2\pi)}_{\text{Const.}} \quad ③$$

$$-\frac{1}{2}\log|H+V_0^{-1}| \quad ④$$

$$= |H(1+H^{-1}V_0^{-1})|$$

$$= |H||1+H^{-1}V_0^{-1}| \quad \text{(仮定)}$$

$$\simeq |H| \quad \text{「Gauss分布が広がっている」}$$

→ V_0が大きい

→ V_0^{-1}は微小

$$\simeq \log p(\mathcal{D}|\theta_{MAP}, M_i) - \frac{1}{2}(\theta_{MAP}-m)^T V_0^{-1}(\theta_{MAP}-m) - \frac{1}{2}\log|H| + \text{const.}$$

$$H = -\nabla\nabla \log p(\mathcal{D}|\theta_{MAP})$$

独立同分布
仮定: i.i.d data.

$$= \sum_{n=1}^{N} H_n$$

$$p(\mathcal{D}|\theta_{MAP}) = \prod_{n=1}^{N} p_n$$

$$= N\left(\frac{1}{N}\sum_{n=1}^{N} H_n\right)$$

n番目のデータ点からの寄与

$$= N\hat{H}$$

N: データ数

より,

$$\log|H| = \log|N\hat{H}|$$

$\hat{H}$の次元数 … M × θのパラメータ数

$$= \log(N^M|\hat{H}|)$$

$$= M\log N + \log|\hat{H}|$$

$O(N)$ vs $O(1)$

$N \gg 1$

$$\simeq M\log N$$

つまり,

$$\log p(\mathcal{D}|M_i)$$

$$\simeq \log p(\mathcal{D}|\theta_{MAP}, M_i) - \frac{1}{2}(\theta_{MAP} - m)V_0^{-1}(\theta_{MAP} - m) - \frac{1}{2}\log|H| + \text{const.}$$

V_0^{-1}は微小パラメータ

$$\simeq \log p(\mathcal{D}|\theta_{MAP}, M_i) - \frac{1}{2}M\log N \quad (4.139)$$

□.

〈ベイズ情報量規準 (BIC)
Bayesian Information Criterion〉
or
〈Schwarz criterion〉

cf. 赤池情報量規準 (AIC)

$$\log p(\mathcal{D}|w_{ML}) - M \quad (4.73)$$

BICの方が重い.
→ BICのモデルが選ばれやすい