

$\{X: \text{data}, Z: \text{latent variable}, \theta: \text{model parameter}\}$

大前提として、事象確率 $p(x|z)$ を求めることとする。①これが data の説明尺度を説明するものにあたる。

$$p(z|x) \propto p(x|z)p(z)$$

これが成り立つ。予測方程式である。

(左側は π と θ の joint model parameter である。
② deterministic な π が確率分布を持つ場合と呼ばれる。
③ fully Bayesian の π は全ての θ が含まれる。)

$$\text{新しいデータの予測: } p(x|x) = \int dz p(x|z)p(z|x)$$

$$\text{model の比較: } p(x|\theta) = \int dz p(x|z, \theta)p(z)$$

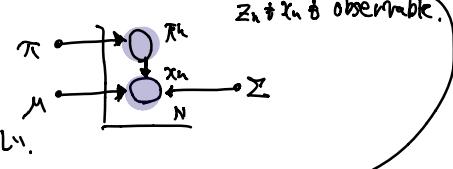
(θ は deterministic parameter と呼ばれる)

$p(z|x)$
左側は π が x によって決まる。↑ つまり π は x によって決まるが、序盤でも既に示した。

chap. 9. EM 法.

$$\frac{\partial}{\partial \theta} p(x, z, \theta) \ln p(z|x, \theta)$$

この「 π が決める場合。
事象分布から被除能分布の場合」



π

z_n

x_n is observable.

x

N

→ 現実的で実行可能な方法が必要。

① stochastic approach (MCMC, chapter 11)

② deterministic approach (variational inference, chap 10)

chap 10.1

後方法: 決定論的の極値問題。 $S[\theta(t)] := \int dt L[\theta(t), \dot{\theta}(t), t]$, $0 = \delta S[\theta(t)] = \int [S(t) + \delta S(t)] - S[t]$

解説

→ lead to Euler-Lagrange Eq.

後方法: 関数を引数として後方法で最小化する。 (10.3) $L[\theta] \leq \text{minimum}$ は。 (ただし複雑な事象分布を simple)
アントラジカルなものがいる)

$$\Delta p(\theta) = L(\theta) + KL(p||p) = \int dz \Delta(\theta) \ln \left\{ \frac{p(z, \theta)}{p(z)} \right\} - \int dz p(z) \ln \left\{ \frac{p(z|\theta)}{p(z)} \right\}$$

$$= L(\theta)$$

$L(\theta)$ is evidence & to be optimized.
 $L(\theta) \geq$ Evidence (ELBO).
Lower Bound.
とも呼ぶ。

この問題を解決する一般的な関数形、これを tractable (= 10.3 で述べた) 良い形 (tractable) に制限して $p(\theta)$ を定めた。

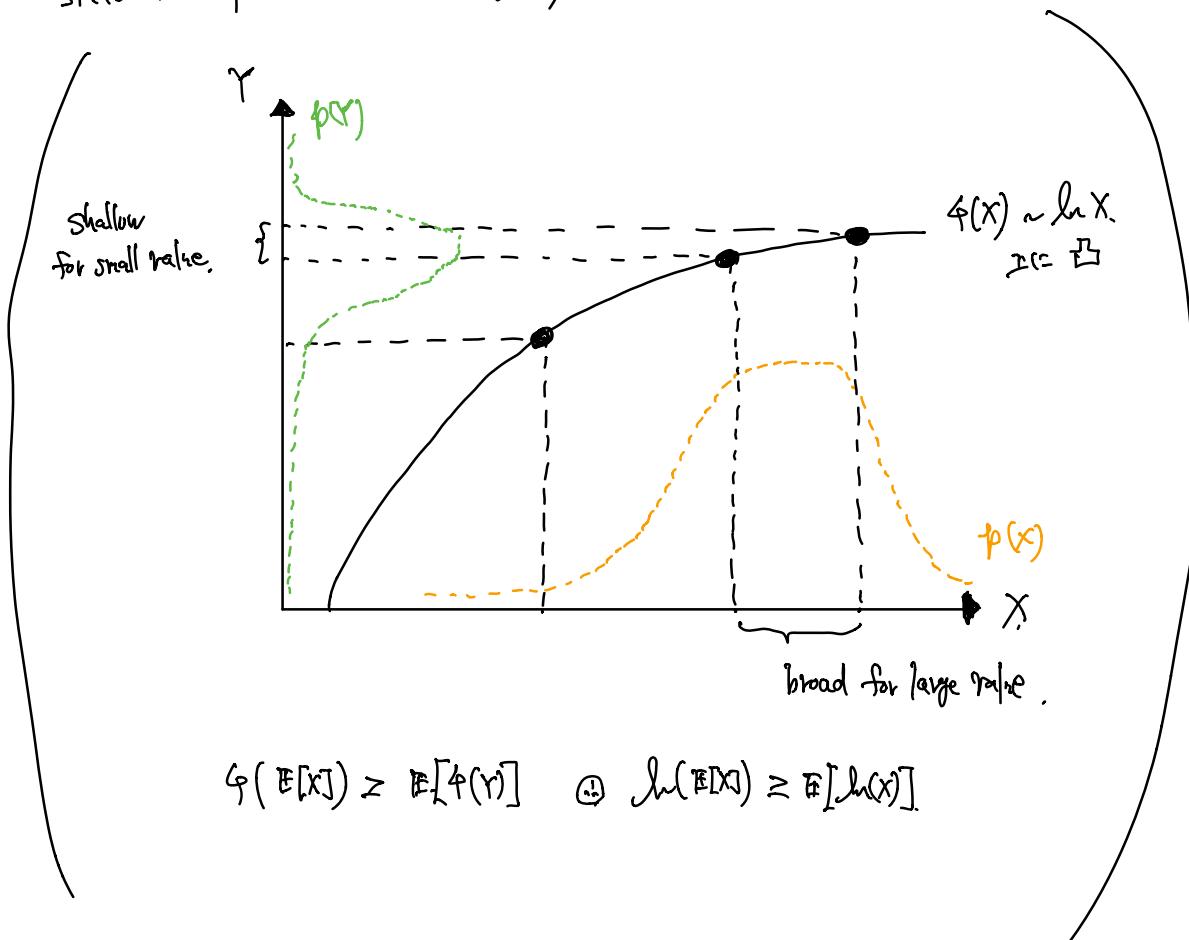
① $p(\theta)$ は parameter を関数を假定して、それを θ とする。(gaussian を假定した結果が Fig. 10.1)

② $p(\theta)$ が各々に個別に積を取って $p(\theta)$ から θ の factorization が可能。

ELBO is obtained by using Jensen's Inequality.

$$\begin{aligned}
 \ln p(x) &= \ln \int dz p(x, z) \\
 &= \ln \int dz p(x, z) \frac{\delta(z)}{\delta(z)} \\
 &= \ln \left(\mathbb{E}_z \left[\frac{p(x, z)}{\delta(z)} \right] \right) \\
 &\geq \mathbb{E}_z \left[\ln \left(\frac{p(x, z)}{\delta(z)} \right) \right] = \underbrace{\int dz \delta(z) \ln \left(\frac{p(x, z)}{\delta(z)} \right)}_{\text{Jensen}}
 \end{aligned}$$

Sketch of the proof of Jensen's inequality.



$$g(z) = \prod_{i=1}^n g_i(z_i) \quad \longleftrightarrow \text{mean field theory と対応}.$$

$$\begin{aligned} L(g) &= \int d\mathbf{z}_S g_S \underbrace{\mathbb{E}_{i \neq S} [\ln p(x_i | z_i)]}_{\ln \tilde{p}(x_i | z_i)} - \int d\mathbf{z}_S g_S h_S + \text{const.} \rightarrow \text{等式を最大化する} \\ &\rightarrow -KL(g || \tilde{p}) \rightarrow \text{最大化. つまり 0 に近づける} \\ &\rightarrow g_S^*(z_S) = \tilde{p}(x_i | z_S) \\ &\rightarrow \ln g_S^*(z_S) = \ln \tilde{p}(x_i | z_S) = \mathbb{E}_{i \neq S} [\ln p(x_i | z_i)] + (\text{const.}) \\ &\rightarrow \ln g_S^*(z_S) = \mathbb{E}_{i \neq S} [\ln p(x_i | z_i)] + (\text{const.}) \quad \text{ただし (※)} \end{aligned}$$

以上は RNN の計算過程で closed form で表せる。
初期段階で、各字を update して 時系列を更新。

10.1.2

2つの潜在変数 $z = \{z_1, z_2\}$ の場合で適用。exact で計算できるか。確認。

$$g(z) = g(z_1)g(z_2)$$

$$(※) \text{ も解く。 } g_S^*(z_S) = N(z_S | \mu_S, \Lambda_{SS}^{-1}) \text{ が手に入る。}$$

$$\begin{aligned} p(z) &= N(z | \mu, \Lambda^{-1}) \\ z &= (z_1, z_2) \\ \mu &= (\mu_1, \mu_2) \\ \Lambda &= \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \end{aligned}$$

（※） g_S^* が求めば $p(z)$ を近似できるので、 $p(z) p(z)$ を解く。

ここで $KL(g || p)$ の値、reverse で計算 $KL(p || g)$ の値。特に、真の分布に対する二つの特徴 (Fig. 10.2)
と divergence の話。

10.1.3

univariate Gaussian の場合。 $\{x_i = x_i\}$ Goal は $\{\text{mean } \frac{\mu}{\tau}\}$ の事後分布を推測する。

$$\begin{cases} p(D | \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^N (x_i - \mu)^2\right\} \\ p(\mu | \tau) = N(\mu | \mu_0, (\lambda_0 \tau)^{-1}) \\ p(\tau) = \text{Gam}(\tau | a_0, b_0) \end{cases}$$

$$g(\mu, \tau) = g(\mu)g(\tau) \propto \tau^{-1/2} e^{-\frac{1}{2\tau} \sum (x_i - \mu)^2} \quad (\text{解析的解: } \frac{1}{\tau} = \frac{1}{N} \sum (x_i - \bar{x})^2)$$

$$\text{期待値: } \left\{ \mathbb{E}[x_i] = \bar{x}, \mathbb{E}[\tau] = \frac{1}{N} \sum (x_i - \bar{x})^2 \right\} \text{ が得られる。}$$

10.1.4

Model comparison が図示。Model が重力下で 2 次元が重力下で比較が難しい。Model が index で比較されるべきである。

$$g(z, m) = g(z|m)g(m)$$

$$\text{Jointly: } p(h|x) \propto p(x|h)p(h).$$

$$\begin{aligned} \text{dilemma: } & p(x) = P - \frac{1}{N} \sum_i g(z_i|m) g(m) \ln \frac{g(z_i|m)}{\mathbb{E}[g(z_i|m)]} \\ & \left. \begin{aligned} & \text{optimization: } g(m) \propto p(m) e^{-\lambda m} \\ & \text{optimization: } g(z|m) \end{aligned} \right\} \text{ ここで } g(z|m) \leftrightarrow g(m|x) \end{aligned}$$

Chap. 10.2.

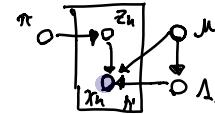
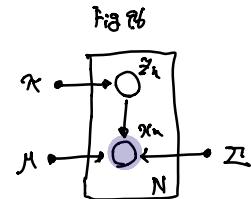
Gaussian Mixture Model \approx 考虑 (右圖)

基本元氣, 2.2 版的複雜度在本節適用 LT+IT.

$$X = \{x_1, \dots, x_N\}$$

$$Z = \{z_1, \dots, z_N\}, \quad z_i = \{z_{i1}, \dots, z_{iK}\}, \quad \mu = \{\mu_k\}, \quad \Lambda = \{\Lambda_k\}.$$

$$\left\{ \begin{array}{l} p(z|\pi) = \prod_{n=1}^N \prod_{k=1}^K \pi_n^{z_{nk}} \\ p(x|z, \mu, \Lambda) = \prod_{n=1}^N \prod_{k=1}^K N(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}} \\ p(\pi) = D_{\text{Dir}}(\pi | \alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_{0k}-1} \\ p(\mu, \Lambda) = p(\mu|\Lambda) p(\Lambda) = \prod_{k=1}^K N(\mu_k | \mu_0, (\Lambda_0 \Lambda_k)^{-1}) W(\Lambda_k | \Lambda_0, V_0) \end{array} \right. \quad \begin{array}{l} \text{conjugate distrib.} \\ \text{未知参数为 unknown \& conjugate prior} \\ \text{prior \& 导入此 modeling 时结果为 prior} \end{array}$$



10.2.1

$$p(x, z, \pi, \mu, \Lambda) = p(x|z, \pi, \mu, \Lambda) p(z|\pi, \mu, \Lambda) = p(x|z, \pi, \mu, \Lambda) p(z|\pi) p(\pi) p(\mu|\Lambda) p(\Lambda)$$

$$\star \text{ Factorization } f^*(z, \pi, \mu, \Lambda) = f(z) f(\pi, \mu, \Lambda)$$

等价 exact 解
(graphical model 基于)

$f^*(z)$ Variational method 一般解法. $\ln f^*(z) = \mathbb{E}_{\pi, \mu, \Lambda} [\ln p(x, z, \pi, \mu, \Lambda)] + (\text{const.})$

$$= \mathbb{E}_\pi [\ln p(z|\pi)] + \mathbb{E}_{\mu, \Lambda} [\ln p(x|z, \mu, \Lambda)] + (\text{const.})$$

$$\downarrow \quad \quad \quad \downarrow$$

$$\sum_{n=1}^N \sum_{k=1}^K z_{nk} \mathbb{E}_\pi [\ln \pi_n] \quad \sum_{n=1}^N \sum_{k=1}^K z_{nk} \mathbb{E}_\mu [\ln N(x_n | \mu_k, \Lambda_k^{-1})]$$

$$\Rightarrow \ln f^*(z) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \ln p_{nk}.$$

$$\Rightarrow f^*(z) \propto \prod_{n=1}^N \prod_{k=1}^K p_{nk}^{z_{nk}}.$$

$$\Rightarrow f^*(z) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}} \quad \text{where } r_{nk} = \frac{p_{nk}}{\sum_{k'=1}^K p_{nk'}}$$

$\sum_{n=1}^N \sum_{k=1}^K z_{nk} \frac{p_{nk}}{\sum_{k'=1}^K p_{nk'}} = \sum_{n=1}^N r_{nk} \sum_{k=1}^K z_{nk} = \sum_{n=1}^N r_{nk}$

Using this $f^*(z)$,

$$\mathbb{E}[z_{nk}] = \sum_{n=1}^N \prod_{k'=1}^{K-1} r_{nk'} \cdot r_{nk} = \underline{r_{nk}}. \quad \text{④ } z_{nk} = 1 \text{ 为混合成分的强度.}$$

responsibility

z 的 k 分量 贡献计算.

$$N_k = \sum_{n=1}^N r_{nk}. \quad \text{scalar}$$

$$\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} x_n. \quad \text{vector}$$

$$S_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T. \quad \text{matrix.}$$

$f^*(\pi, \mu, \Lambda)$

$$\ln f^*(\pi, \mu, \Lambda) = \ln p(\pi) + \sum_{k=1}^K \ln p(\mu_k, \Lambda_k) + \mathbb{E}_z [\ln p(z|\pi)] + \sum_{k=1}^K \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln N(x_n | \mu_k, \Lambda_k^{-1}) + (\text{const.})$$

★ Factorization $f(\pi, \mu, \lambda) = f(\pi) \prod_{k=1}^K f(\mu_k, \lambda_k)$ ↑ の式が $\pi \in (0, 1)$ のため不適である。(prior が conjugate でない)

$$\ln f^*(\pi) \xrightarrow{\text{Factorization}} f^*(\pi) = D \ln(\pi) \quad \alpha_k = d_0 + N_k, \quad d = \{d_1, d_2, \dots, d_K\}$$

$$\sum_{k=1}^K \ln f^*(\mu_k, \lambda_k) \longrightarrow f^*(\mu_k, \lambda_k) = N(\mu_k | m_k, (\beta_k \lambda_k)^{-1}) \mathcal{N}(\lambda_k | w_k, v_k)$$

以上の式は EM 法の M-step (parameter 更新) を示す。この update は $E[\ln \pi_k] = h_k$ が基準。
この E-step は $\hat{\pi}_k$ を計算するには下記の量を計算する。

$$E_{\mu_k, \lambda_k}[(x_n - \mu_k)^T \lambda_k (x_n - \mu_k)] = D p_k^{-1} + V_k (x_n - m_k)^T W_k (x_n - m_k)$$

$$\ln \hat{\pi}_k := E[\ln \pi_k] = \frac{D}{2} \psi\left(\frac{D+1-k}{2}\right) + D \ln 2 + \ln[V_k].$$

$$\ln \tilde{\pi}_k = E[\ln \pi_k] = \psi(d_k) - \psi(D)$$

$$\Rightarrow \ln \hat{\pi}_k, V_{nk} = \frac{p_k^{-1}}{\sum_{j=1}^K p_j^{-1}} \Rightarrow V_{nk} \propto \tilde{\pi}_k \lambda_k^{-1} \exp\left\{-\frac{D}{2p_k} - \frac{V_k}{2} (x_n - m_k)^T W_k (x_n - m_k)\right\}$$

EM 法の結果
同じ形。

EM 法と同様に 特殊場合で E-step が M-step と等しい。

① $m_k \approx 0, N_k \approx 0$ のときの component 等子から導くことができる。自動的にクラス数が決まる。

(解説的には、多少手間であるが、prior が broad になると各成分は独立的で negative k)

② $E[\pi_k] = \frac{d_0 + N_k}{K d_0 + N}$ prior が broad ($d_0 \rightarrow 0$) $E[\pi_k] \rightarrow 0$. 等分布。
prior が tight ($d_0 \rightarrow \infty$) $E[\pi_k] \rightarrow \frac{1}{K}$ 独立性。

Maximum Likelihood EM と似た方針

- 1つの gaussian component が 0 のときに導入されてしまう singularity が 固定される。
- 大きな K では 過度に over-fit してしまう。
- 最適な components の数を自動的に決定できる。

10.2.2 Variational Lower Bound と、re-estimating 式を再導出する。 $\tilde{\pi}_k$ と $\tilde{\lambda}_k$ は一般的な parameter と見なして f を maximize することで導出。

10.2.3 predictive density $f(x|x) = \prod_{k=1}^K \int d\mu_k d\lambda_k \underbrace{f(\frac{1}{2} \lambda_k, 1)}_{\text{Factorisation}} \underbrace{f(\frac{1}{2} \pi_k | \lambda_k)}_{\text{Factorisation}} f(\pi_k, \mu_k, \lambda_k | x)$

$$\approx \frac{1}{2} \frac{K}{\pi_k} \alpha_k S_t(\frac{1}{2} \pi_k, \mu_k, \lambda_k, v_k + 1 - D)$$

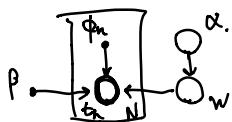
これが t 時点で得たもので
実際には算出する。
Student's t 分布が角溝。

10.2.4: parameter space で同じ極値点構の解が複数現れる。
Bayesian では全ての parameter 空間で確率的で各解の導出が行われる。
→ その分考慮して ELBO は Lkl とつなげたい。



10.2.5:
要請について Factorization は factorization ができる場合がある。 $\theta^*(\lambda, \alpha) \rightarrow \prod_{k=1}^K \theta^*(\lambda_k, \alpha_k)$
この数値計算で結構よく要素の都合に適している。
check では graphical model で条件独立性 (d-separation) を実現。
基本的には conjugate prior の設定 (Covar がかかる)。

Chap. 10.3



In graphical models like Variational Inference for predictive distribution となる。

fit prediction parameter θ using time value t for $t = 1, \dots, T$.
→ まず 分布の扱い一般化が straightforward.

Chap. 10.4

指數型分布族の話。 PRML で扱う complete-data likelihood の多く場合もこれで記述できる。

参数の種類 → observed.
→ hidden → expensive (data 計算量大) $\theta := \mu$
→ infinite. (data 計算量無限) $\theta := \mu, \lambda, \pi_E$

General Mixture Model の場合,

$$p(x, z | \theta) = \prod_{n=1}^N f(x_n, z_n) g(z) \exp\{z^T u(x_n, z_n)\}, \text{ parameter } \theta \text{ の prior } p(z | \nu_0, \gamma_0) = f(z, \nu_0) g(z)^{\nu_0} \exp\{\gamma_0 z^T x_0\}$$

★ factorization $g(z, \theta) = g(z) g(\theta)$.

$$\begin{aligned} -\text{解法} & \quad \ln g^*(z) = \mathbb{E}_\theta [\ln p(x, z | \theta)] + (\text{const.}) = \sum_{n=1}^N \left[\ln f(x_n, z_n) + \mathbb{E}[z^T u(x_n, z_n)] + (\text{const.}) \right] \quad g^*(z) = (\text{指數型分布}) \\ & \quad \ln g^*(\theta) = \ln p(z | \nu_0, \gamma_0) + \mathbb{E}_z [\ln p(x, z | \theta)] + (\text{const.}) = \dots \quad g^*(\theta) = f(z, \nu_0) g(z)^{\nu_0} \exp\{\gamma_0 z^T x_0\} \end{aligned}$$

$$\begin{aligned} \nu_n &= \nu_0 + N \\ \nu_n x_n &= \nu_0 x_0 + \sum_{n=1}^N \mathbb{E}_z [z^T x_n] \end{aligned}$$

E step. → M step → E ...

$\mathbb{E}[u(x_n, z_n)]$ $\mathbb{E}[z^T]$ Σ 見付けるため
current $g(z)$ を使おう。 $\frac{1}{N} \sum_{n=1}^N g(z_n)$ を使おう。
Z(z) $g(z)$ を計算。 $\frac{1}{N} \sum_{n=1}^N g(z_n)$ θ の更新

10.4.1

具体的な計算のための用語をみてみたが、一般的な directed graph の場合で説明する。

$$p(x) = \prod_i p(x_i | p_{a_i}) \quad \text{where } p_{a_i} \text{ is node } i \text{ の父 node の集合.}$$



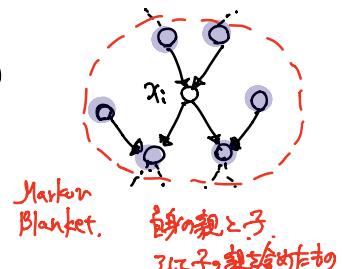
★ Factorization $f(x) = \prod_i f_i(x_i)$

- 複雑な場合.

$$\ln f_i^*(x_i) = \mathbb{E}_{x \in \Omega} \left[\sum_i \ln p(x_i | p_{a_i}) \right] + (\text{const.})$$

この条件確率計算は必要なのは graph local to factor \$f_i\$. (Markov blanket)

⇒ Variational Inference の実装に際して model の構造を理解する必要がある。



全ての条件分布が conjugate 在相数型/分布族で表される場合.

→ Variational to update & local to message passing で表現可能.

→ parent \rightarrow children \leftarrow children が親から message を受け取る過程で update される.

→ Lower bound の計算に必要な大部の message passing は得られる。この分散の仕組みは大规模ネットワークに適している。

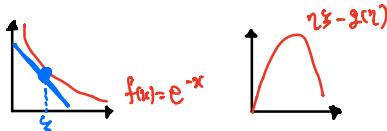
Chap 10.5

これまで posterior / 全体（全局）を近似していた。これを “global” と呼ぶこと。

local では局所的な一部を近似して説明するもの。

(例.) $f(x) = e^{-x}$.

ここで $f(x)$ を線形関数で近似してみる。



Taylor 展開 1 次.

$$g(x) = e^{-3} - e^{-3}(x-3) + O((x-3)^2)$$

凸関数の性質により、 $g(x) \leq f(x)$ $x=3$ のとき等号

係数 $f'(3) = -e^{-3} = -1 < 0$ すなはち $f(x)$ の値が減少傾向。

$$g(x, \eta) = \eta x - \eta + \eta \ln(-\eta)$$

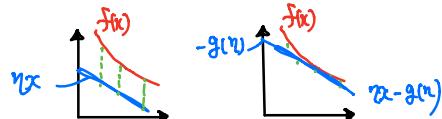
$g(x, \eta) \leq f(x)$ が示す。

$$f(x) = \max_x \{ \eta x - \eta + \eta \ln(-\eta) \}$$

η の範囲は $\eta > 0$ である $f(x) \in \text{interval}$ を表す。

(一般形) $f(x)$ が凸関数である場合

$g(x)$ は線形関数で近似した。



Taylor 展開 1 次.

$$g(x) = f(3) + f'(3)(x-3) + O((x-3)^2)$$

$$g(x) \leq f(x)$$

すなはち $f(x) \geq g(x)$ が成り立つ。これは $f'(3) = -1$ である。

$$g(x) = -\max_x \{ f(x) - g(x) \}$$

$$= \max_x \{ \eta x - f(x) \}$$

これが適切な η は $g(x) \geq -f(x)$ となる。

(一般的な導論と適用)

$$g(x) = \max_{\eta} \{ \eta x - C^{\eta} \}.$$

$$\frac{\partial}{\partial x} \{ \cdot \} = \eta + e^{-\eta} = 0 \quad \text{④} \quad x = -\ln(-\eta).$$

$$g(x) = \eta - \eta \ln(-\eta)$$

係数 η を $f(x)$ や $g(x)$ の関数の場合を表す。

$$f(x) = \max_{\eta} \{ \eta x - g(\eta) \}$$

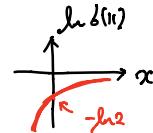
\Rightarrow Concave function $f(x)$ が $\min_{\eta} \{ \eta x - g(\eta) \}$ で表される。

$$\begin{cases} f(x) = \max_{\eta} \{ \eta x - g(\eta) \} \\ g(\eta) = \min_{x} \{ \eta x - f(x) \} \end{cases}$$

\Rightarrow 対数的・凸関数の場合で、 $g(x)$ を直接つかうと便利である。

[重要な例として sigmoid function の場合] $\delta(x) = \frac{1}{1+e^{-x}}$

$$f(x) = \ln \delta(x) = -\ln(1+e^{-x}) \text{ は凸 (=+)。}$$



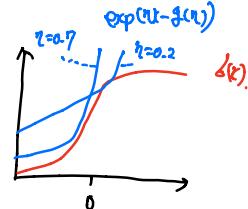
$$\begin{aligned} g(x) &= \max_{\eta} \{ \eta x - \delta(\eta) \} = \max_{\eta} \{ \eta x - \ln(1+e^{-\eta}) \} \\ &= -\eta \ln(\frac{\eta}{1-\eta}) - \ln(1+\frac{\eta}{1-\eta}) \\ &= -\eta \ln \eta - (1-\eta) \ln(1-\eta). \end{aligned}$$

binary entropy function.

$$\begin{aligned} \frac{\partial}{\partial x} \{ \cdot \} &= \eta - \frac{1}{1+e^{-\eta}} \{-e^{-\eta}\} = 0. \\ \rightarrow \eta(He^{-\eta}) &= e^{-\eta}. \\ \rightarrow e^{-\eta}(1-\eta) &= \eta. \\ \rightarrow x &= -\ln(\frac{\eta}{1-\eta}). \end{aligned}$$

$$f(x) = \max_{\eta} \{ \eta x - g(\eta) \} \rightarrow \ln \delta(x) \leq \eta x - g(\eta)$$

$$\rightarrow \delta(x) \leq \exp(\eta x - g(\eta))$$



\Rightarrow sigmoid function の下限は Gaussian で近似できる。 (convex function は近似可能)

$$\ln \delta(x) = -\ln(1+e^{-x}) = -\ln \{ e^{-\frac{x}{2}} (e^{\frac{x}{2}} + e^{-\frac{x}{2}}) \} = \frac{x}{2} - \ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}})$$

$f(x) = -\ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}})$ は convex である。

$$g(x) = \max_{\eta} \{ \eta x^2 - f(\eta) \}$$

$$\frac{\partial}{\partial x^2} \{ \cdot \} = \eta - \frac{d\eta}{dx^2} \frac{d}{dx} f(x) = \eta + \frac{1}{4x^2} \tanh(\frac{x}{2}) = 0. \quad \text{⑤} \quad \eta = -\frac{1}{4x^2} \tanh(\frac{x}{2})$$

2 値 $x \in [0, 1]$ の 2 値 y の確率 $p(y|x)$ contact point の値で $y \in [0, 1]$ 。
 $y = -\frac{1}{4x^2} \tanh(\frac{x}{2}) = -\frac{1}{2x^2} [\tanh(\frac{x}{2}) - \frac{1}{2}]$

$$g(x) = \eta x^2 - f(\eta) = \eta x^2 + \ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}})$$

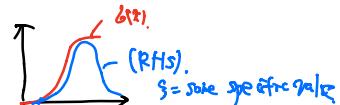
$$f(x) = \max_{\eta} \{ \eta x - g(\eta) \} \text{ は凸。}$$

$$f(x) \geq \eta x^2 - g(\eta) = \eta x^2 - \eta x^2 - \ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = -\ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = \frac{x^2 - x^2}{2} + \ln \delta(\frac{x}{2}) - \frac{x}{2}.$$

$$\text{f(x)} = \frac{x}{2} + \ln \delta(\frac{x}{2})$$

$$\ln \delta(x) \geq \frac{(x-\xi)}{2} + \ln \delta(\xi) + \frac{\eta(x^2-\xi^2)}{2}$$

$$\delta(x) \geq \delta(\xi) \exp \left\{ \frac{x-\xi}{2} + \eta \frac{x^2-\xi^2}{2} \right\}, \text{ Gaussian の形。}$$



chap. 10.6

§ 4.5 で Laplace approximation と logistic regression と local variational method と直結する。(精度向上が3)

$$p(t) = \int dw p(t|w)p(w) = \int dw \left[\prod_{n=1}^N p(t_n|w) \right] p(w)$$

$$\downarrow p(t|w) = \prod_{n=1}^N \frac{t_n^{t_n}}{(1-t_n)^{1-t_n}} \stackrel{\text{2.16}}{=} \prod_{n=1}^N p(c_i|w) (1-p(c_i|w))^{1-t_n} \text{ where } t_i = \delta_i \beta_i$$

(local variational method と直結)

(10.6.1) ではこれを近似する、 $p(w) = N(w|\mu_N, S_N)$ とする

(10.6.2) では予測分布を求める際の variational parameter $\{t_n\}$ を定め、① w の潜在変数分布を求める ② w の解説的分布を求める。
 $p(t|t) = \int dw p(t|w) p(w|t) \propto \int dw p(t|w) p(t|w) p(w)$
 ここで近似法 (local variational inference) とする。
 $t_i(\mu, \Sigma)$ を与えると t_i を求めること。

(10.6.3) Hyperparameters も考慮する分布を求める inference.

chap. 10.7

Variational Inference は 10.6.3 Inference method. Expectation Propagation,

Expectation Propagation で 10.6.3

- ① $KL(p||q)$ reverse で計算する。② $E_{\tilde{p}(q)}[u(q)] = E_{p(q)}[u(q)]$ で moment matching で計算する。
- ③ $\tilde{p}(q) = T_q f_q(q)$ とし形を決定 (この構造は model がどのよう分布するか)
- ④ $f_q(q) = \frac{1}{2} T_q f_q(q)$ と指數型分布で近似する。 $\tilde{f}_q(q) \in \mathcal{F}_q(q) = \frac{f_q(q)}{f_q(q)}$ で moment matching update。
- ⑤ 最終的に $\tilde{p}(q) \approx \int T_q f_q(q) dq$ で分布を近似する。

factor graph $\begin{cases} \text{clutter problem} \\ \text{factor graph} \end{cases} \rightarrow \tilde{f}_q(q) \approx f_q(q) \sim$ 低層加重が良い精度を保つ範囲をコントロール。
 belief propagation の message を使う。