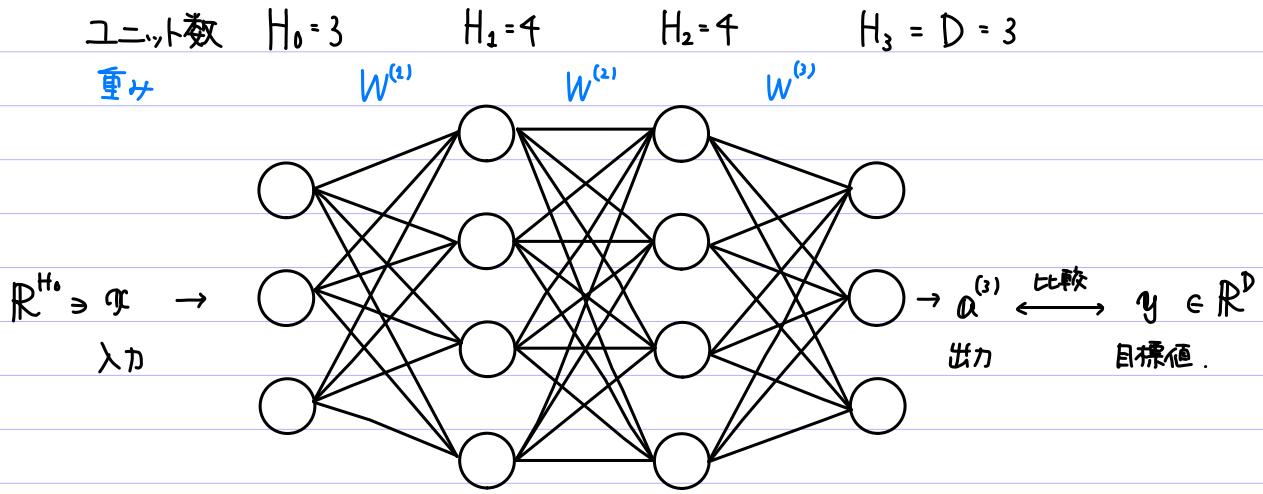
Ch.5 ニューラルネットワークのBayes推論.

- Feedforward NN 等の深層学習モデルを Bayes的に扱って,

  Ch.4 でやった近似推論手法をNNモデルに適用していく.

用語・記法の復習 (Ch.2より)

順伝播型ニューラルネットワーク (feedforward NN)



入力層	隠れ層		出力層	層数
layer 0	layer 1	layer 2	layer 3	$L=3$

各層の出力: $z^{(0)} = x$, $z^{(l)} = \phi(a^{(l)})$ $(l=1,\dots,L-2)$

(ϕ: 活性化函数(非線形), 要素ごとの適用.)

活性: $a^{(l)} = W^{(l)} z^{(l-1)}$ $(l=1,\dots,L)$

重みの各要素を全てまとめて W と書く. ベクトルとして扱う.

重み W をパラメータとするNNを $f(x;W)$ と表す. $f(x;W) = a^{(L)}$.

§5.1 Bayesニューラルネットワークモデルの近似推論法.

- NNのバッチ学習(全データを一度に学習)について考える.
- 回帰問題を扱う.(分類でも本筋は同じ)

5.1.1 Bayesニューラルネットワークモデル

- ネットワークの挙動を支配するパラメータ W に事前分布 $p(W)$ を設定することで Bayes的に扱えるようにする.

入力 $\mathcal{X} = \{x_1, \dots, x_N\}$ に対する観測データ $\mathcal{Y} = \{y_1, \dots, y_N\}$: given.

- モデルを次のように設定:

$$p(\mathcal{Y}, W \mid \mathcal{X}) = p(W)p(\mathcal{Y} \mid \mathcal{X}, W) = p(W)\prod_{n=1}^{N} p(y_n \mid x_n, W)$$

観測モデル: $p(y_n \mid x_n, W) = \mathcal{N}(y_n \mid f(x_n; W), \sigma_y^2 I_D)$

($\sigma_y^2 > 0$ は固定のノイズパラメータ)

事前分布: $p(W) = \prod_{w \in W} p(w) = \prod_{w \in W} \mathcal{N}(w \mid 0, \sigma_w^2)$

($\sigma_w^2 > 0$ は固定のノイズパラメータ)

以下では,簡単のため出力次元を $D = 1$ で考える.

5.1.2 Laplace近似による学習.

5.1.2.1 事後分布の近似

1. 事後分布 $p(W \mid \mathcal{Y}, \mathcal{X})$ の最大化問題の局所最適解 $\tilde{W}$ を求める.

$\tilde{W}$ が $p(W \mid \mathcal{Y}, \mathcal{X})$ の局所最適解であるなら,

$\tilde{\mathcal{W}}$ は $\log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$ の局所最適解でもある.

$-\log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$ に対して勾配降下法で最小化すれば,

$-\log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$ の極小解 $\Leftrightarrow$ $\log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$ の極大解 が得られる.

反復スキーム: 学習率を $\alpha > 0$ として,

$$\mathcal{W}_{new} \leftarrow \mathcal{W}_{old} + \alpha \nabla_{\mathcal{W}} \log p(\mathcal{W}|\mathcal{Y},\mathcal{X})\Big|_{\mathcal{W}=\mathcal{W}_{old}}.$$

- $\nabla_{\mathcal{W}} \log p(\mathcal{W}|\mathcal{Y},\mathcal{X}) = \left(\frac{\partial}{\partial w} \log p(\mathcal{W}|\mathcal{Y},\mathcal{X})\right)_{w\in\mathcal{W}}$ について.

$p(\mathcal{W}|\mathcal{Y},\mathcal{X}) \propto p(\mathcal{W})p(\mathcal{Y}|\mathcal{X},\mathcal{W})$ ← $\mathcal{W}$ に依存する項のみ考える なので,

$$\log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$$
$$= \log p(\mathcal{Y}|\mathcal{X},\mathcal{W}) + \log p(\mathcal{W}) + \text{const.}$$
$$= \sum_{n=1}^{N} \log \mathcal{N}(y_n|f(x_n;\mathcal{W}),\sigma_y^2) + \sum_{w\in\mathcal{W}} \log \mathcal{N}(w|0,\sigma_w^2) + \text{const.}$$

- 各項は,

$$\log \mathcal{N}(y_n|f(x_n;\mathcal{W}),\sigma_y^2) = -\frac{1}{2\sigma_y^2}\left(y_n - f(x_n;\mathcal{W})\right)^2 + \text{const.}$$
$$\log \mathcal{N}(w|0,\sigma_w^2) = -\frac{w^2}{2\sigma_w^2} + \text{const.}$$

なので, パラメータ $w\in\mathcal{W}$ について微分すると,

$$\frac{\partial}{\partial w} \log p(\mathcal{W}|\mathcal{Y},\mathcal{X})$$
$$= -\frac{1}{\sigma_y^2}\frac{\partial}{\partial w}\underbrace{\left(\frac{1}{2}\sum_{n=1}^{N}\left(y_n - f(x_n;\mathcal{W})\right)^2\right)}_{=:E(\mathcal{W})} - \frac{1}{\sigma_w^2}\frac{\partial}{\partial w}\underbrace{\left(\frac{1}{2}\sum_{w\in\mathcal{W}} w^2\right)}_{=:\Omega_{L2}(\mathcal{W})}$$
$$= -\left(\frac{1}{\sigma_y^2}\frac{\partial}{\partial w}E(\mathcal{W}) + \frac{1}{\sigma_w^2}\underbrace{\frac{\partial}{\partial w}\Omega_{L2}(\mathcal{W})}_{=w}\right).$$

$$\nabla_w \log p(w|y,x) = -\left(\frac{1}{\sigma_y^2}\nabla_w E(w) + \frac{1}{\sigma_w^2} w\right).$$

- 第1項は backpropagation で計算できる.

→ スキーム $w_{new} \leftarrow w_{old} + \alpha \nabla_w \log p(w|y,x)\big|_{w=w_{old}}$ はきちんと計算できる.

2. $p(w|y,x) \approx q(w) = \mathcal{N}(w|\tilde{w}, (\Lambda(\tilde{w}))^{-1})$ と近似.

精度行列は,

$$\Lambda(\tilde{w}) = -\nabla_w^2 \log p(w|y,x) = \frac{1}{\sigma_y^2}\underbrace{\nabla_w^2 E(w)}_{\text{Hesse行列.}} + \frac{1}{\sigma_w^2} I$$

5.1.2.2 予測分布の近似

- 次は予測分布 $p(y_*|x_*,y,x) = \int p(y_*|x_*,w)p(w|y,x)dw$

の近似をする. 上の事後分布の近似 $p(w|y,x) \approx q(w)$ を使う.

- $p(y_*|x_*,w) = \mathcal{N}(y_*|f(x_*;w), \sigma_y^2)$ と, NN $f(x_*;w)$ が含まれるので

事後分布の近似としても依然として解析的に計算できない.

→ パラメータの事後分布の密度が $\tilde{w}$ まわりに集中していて, (サンプル数が多くなれば ある程度は妥当か?)

その近傍で $f(x_*;w)$ を w に関して線形近似できると仮定.

Taylor展開として1次近似:

$$f(x_*;w) \approx f(x_*;\tilde{w}) + \underbrace{\left(\nabla_w f(x_*;w)\big|_{w=\tilde{w}}\right)}_{=:g}^{\top}(w-\tilde{w}).$$

- 結局, 予測分布は次で近似する.

$$p(y_*|x_*,y,x)$$
$$\approx \int \mathcal{N}(y_*|f(x_*;\tilde{w}) + g^{\top}(w-\tilde{w}), \sigma_y^2)\mathcal{N}(w|\tilde{w}, (\Lambda(\tilde{w}))^{-1})dw.$$

これを計算しよう．$f = f(x_*, \tilde{w})$ と書く．

$$\log \mathcal{N}(y_* \mid f + g^\top (w - \tilde{w}), \sigma_y^2)$$

$$= -\frac{1}{2\sigma_y^2}\left(y_* - (f + g^\top (w - \tilde{w}))\right)^2 + \text{const.}$$

$$= -\frac{1}{2\sigma_y^2}\left(y_*^2 - 2(f + g^\top (w - \tilde{w}))y_* + (f + g^\top (w - \tilde{w}))^2\right) + \text{const.}$$

$$= -\frac{1}{2\sigma_y^2}\left(y_*^2 - 2(f - g^\top \tilde{w})y_* + (f - g^\top \tilde{w})^2 - 2g^\top w y_* + 2g^\top w (f - g^\top \tilde{w}) + (g^\top w)^2\right) + \text{const.}$$

$$= -\frac{1}{2\sigma_y^2}\left(y_* - (f - g^\top \tilde{w})\right)^2 - \frac{1}{2\sigma_y^2}\left(w^\top g g^\top w - 2w^\top g\left(y_* - (f - g^\top \tilde{w})\right)\right) + \text{const.}$$

また，

$$\log \mathcal{N}(w \mid \tilde{w}, (\Lambda(\tilde{w}))^{-1})$$

$$= -\frac{1}{2}(w - \tilde{w})^\top \Lambda(\tilde{w})(w - \tilde{w}) + \text{const.}$$

$$= -\frac{1}{2}\left(w^\top \Lambda(\tilde{w}) w - 2w^\top \Lambda(\tilde{w})\tilde{w}\right) + \text{const.}$$

これより．

$$\log \mathcal{N}(y_* \mid f + g^\top (w - \tilde{w}), \sigma_y^2)\mathcal{N}(w \mid \tilde{w}, (\Lambda(\tilde{w}))^{-1})$$

$$= -\frac{1}{2\sigma_y^2}\left(y_* - (f - g^\top \tilde{w})\right)^2$$

$$-\frac{1}{2}\Big(w^\top \underbrace{\left(\Lambda(\tilde{w}) + \frac{1}{\sigma_y^2} g g^\top\right)}_{=: \Lambda_*} w - 2w^\top \underbrace{\left(\Lambda(\tilde{w})\tilde{w} + \frac{1}{\sigma_y^2} g\left(y_* - (f - g^\top \tilde{w})\right)\right)}_{=: \Lambda_* \mathcal{M}(y_*)}\Big) + \text{const.}$$

$$= \log \mathcal{N}(y_* \mid f - g^\top \tilde{w}, \sigma_y^2) - \frac{1}{2}\left(w^\top \Lambda_* w - 2w^\top \Lambda_* \mathcal{M}(y_*)\right) + \text{const.}$$

$$= \log \mathcal{N}(y_* \mid f - g^\top \tilde{w}, \sigma_y^2) - \frac{1}{2}(w - \mathcal{M}(y_*))^\top \Lambda_* (w - \mathcal{M}(y_*)) + \frac{1}{2}\mathcal{M}(y_*)^\top \Lambda_* \mathcal{M}(y_*) + \text{const.}$$

$$= \log \mathcal{N}(y_* \mid f - g^\top \tilde{w}, \sigma_y^2)\mathcal{N}(w \mid \mathcal{M}(y_*), \Lambda_*^{-1}) + \frac{1}{2}\mathcal{M}(y_*)^\top \Lambda_* \mathcal{M}(y_*) + \text{const.}$$

となるので，

$$\int \mathcal{N}(y_* \mid f + g^\top (w - \tilde{w}), \sigma_y^2)\, \mathcal{N}(w \mid \tilde{w}, (\Lambda(\tilde{w}))^{-1})\, dw$$

$$\propto \mathcal{N}(y_* \mid f - g^\top \tilde{w}, \sigma_y^2) \exp\left(\frac{1}{2} m(y_*)^\top \Lambda_* m(y_*)\right) \underbrace{\int \mathcal{N}(w \mid m(y_*), \Lambda_*^{-1})\, dw}_{=1}$$

$$= \mathcal{N}(y_* \mid f - g^\top \tilde{w}, \sigma_y^2) \exp\left(\frac{1}{2} m(y_*)^\top \Lambda_* m(y_*)\right).$$

ここで,

$$m(y_*)^\top \Lambda_* m(y_*)$$

$$= \left(\Lambda(\tilde{w})\tilde{w} + \frac{1}{\sigma_y^2} g\,(y_* - (f - g^\top \tilde{w}))\right)^\top \left(\Lambda(\tilde{w}) + \frac{1}{\sigma_y^2} g g^\top\right)^{-1} \left(\Lambda(\tilde{w})\tilde{w} + \frac{1}{\sigma_y^2} g\,(y_* - (f - g^\top \tilde{w}))\right).$$

であるが, Sherman-Morrison-Woodbury の公式より

$$\left(\Lambda(\tilde{w}) + \frac{1}{\sigma_y^2} g g^\top\right)^{-1}$$

スカラーなのでこれはただの逆数

$$= \Lambda(\tilde{w})^{-1} - \frac{1}{\sigma_y^2} \Lambda(\tilde{w})^{-1} g \underbrace{\left(1 + \frac{1}{\sigma_y^2} g^\top \Lambda(\tilde{w})^{-1} g\right)^{-1}}_{\text{ℓ}} g^\top \Lambda(\tilde{w})^{-1}$$

$$= \Lambda(\tilde{w})^{-1} - \underbrace{\left(\sigma_y^2 + g^\top \Lambda(\tilde{w})^{-1} g\right)^{-1}}_{=: \sigma(x_*)^2} \Lambda(\tilde{w})^{-1} g g^\top \Lambda(\tilde{w})^{-1}$$

$$= \Lambda(w)^{-1} \left(\Lambda(\tilde{w}) - \sigma(x_*)^{-2} g g^\top\right) \Lambda(\tilde{w})^{-1}$$

なので,

$$m(y_*)^\top \Lambda_* m(y_*)$$

$$= \left(\tilde{w} + \sigma_y^{-2} \Lambda(\tilde{w})^{-1} g\,(y_* - (f - g^\top \tilde{w}))\right)^\top \left(\Lambda(\tilde{w}) - \sigma(x_*)^{-2} g g^\top\right) \left(\tilde{w} + \sigma_y^{-2} \Lambda(\tilde{w})^{-1} g\,(y_* - (f - g^\top \tilde{w}))\right)$$

$$= \underbrace{\tilde{w}^\top \left(\Lambda(\tilde{w}) - \sigma(x_*)^{-2} g g^\top\right) \tilde{w}}_{\text{const.}} + 2\sigma_y^{-2} \tilde{w}^\top \left(\Lambda(\tilde{w}) - \sigma(x_*)^{-2} g g^\top\right) \Lambda(\tilde{w})^{-1} g\,(y_* - (f - g^\top \tilde{w}))$$

$$+ \sigma_y^{-4} (y_* - (f - g^\top \tilde{w}))^2 g^\top \Lambda(\tilde{w})^{-1} \left(\Lambda(\tilde{w}) - \sigma(x_*)^{-2} g g^\top\right) \Lambda(\tilde{w})^{-1} g$$

$$= 2\sigma_y^{-2} g^\top \tilde{w} \underbrace{\left(1 - \sigma(x_*)^{-2} g^\top \Lambda(\tilde{w})^{-1} g\right)}_{= \sigma_y^2 \sigma(x_*)^{-2}} (y_* - (f - g^\top \tilde{w}))$$

$$+ \sigma_y^{-4} (y_* - (f - g^\top \tilde{w}))^2 \underbrace{g^\top \Lambda(\tilde{w})^{-1} g}_{\sigma(x_*)^2 - \sigma_y^2} \underbrace{\left(1 - \sigma(x_*)^{-2} g^\top \Lambda(\tilde{w})^{-1} g\right)}_{\sigma_y^2 \sigma(x_*)^{-2}} + \text{const.}$$

$$= 2\sigma(x_*)^{-2} g^\top \tilde{W} \left(y_* - (f - g^\top \tilde{W})\right) + \left(\sigma_y^{-2} - \sigma(x_*)^{-2}\right)\left(y_* - (f - g^\top \tilde{W})\right)^2 + \text{const.}$$

結局,

$$\log \mathcal{N}(y_* \mid f - g^\top \tilde{W}, \sigma_y^2) \exp\left(\tfrac{1}{2} M(y_*)^\top \Lambda_* M(y_*)\right)$$

$$= -\tfrac{1}{2}\sigma_y^{-2}\left(y_* - (f - g^\top \tilde{W})\right)^2$$

$$+ \tfrac{1}{2}\left(2\sigma(x_*)^{-2} g^\top \tilde{W}\left(y_* - (f - g^\top \tilde{W})\right) + \left(\sigma_y^{-2} - \sigma(x_*)^{-2}\right)\left(y_* - (f - g^\top \tilde{W})\right)^2\right) + \text{const.}$$

$$= -\tfrac{1}{2}\sigma(x_*)^{-2}\left(\left(y_* - (f - g^\top \tilde{W})\right)^2 - 2 g^\top \tilde{W}\left(y_* - (f - g^\top \tilde{W})\right)\right) + \text{const.}$$

$$= -\tfrac{1}{2}\sigma(x_*)^{-2}\left(y_* - (f - g^\top \tilde{W})\right)\left(y_* - (f + g^\top \tilde{W})\right) + \text{const.}$$

$$= -\tfrac{1}{2}\sigma(x_*)^{-2}\left(y_*^2 - 2 f y_*\right) + \text{const.}$$

$$= \log \mathcal{N}(y_* \mid f(x_*; \tilde{W}), \sigma(x_*)) + \text{const.}$$

$$\sigma_y^2 + g^\top \Lambda(\tilde{W})^{-1} g$$

以上より,

$$p(y_* \mid x_*, y, X) \approx \mathcal{N}(y_* \mid f(x_*; \tilde{W}), \sigma_y^2 + g^\top \Lambda(\tilde{W})^{-1} g)$$

5.1.3 ハミルトニアンモンテカルロ法による学習.

- HMCで $p(w|y,x)$ からサンプリングをする.

→ サンプリングした w を使って $y_* = f(x_*;w)$ と予測できる.

複数回サンプリングすれば, 予測の不確実性も分かる.

- 離散変数をパラメータに含まないなら, $\nabla_w \log p(w|y,x)$ の計算は

backpropagationでできる.

→ ポテンシャルの微分ができるからHMCが使える

- 計算時間は比較的長いものの, 理論的には真の事後分布からのサンプルが得られる.

5.1.3.1 重みパラメータの推論.

$\log p(w|y,x) = \log p(y|x,w) + \log p(w) + \text{const.}$ なので,

ポテンシャルエネルギーとして $\mathcal{U}(w) = \log p(y|x,w) + \log p(w)$

を利用すればよい.

リープフロッグ法で, この微分が必要になる (backpropagationでできる)

- NNに対するHMCの適用に関する問題点.

✗ 複雑な分布からのサンプリングになるので, 少ないサンプリング数では分布の特徴を

とらえきれているかどうか不明. 適切なサンプリング数も不明.

✗ ステップサイズ ε やステップ数 L の調整が難しい.

✗ 学習が低速.

5.1.3.2 ハイパーパラメータの推論.

- 上では σ_w^2, σ_y^2 をハイパーパラメータとして扱ってきた.

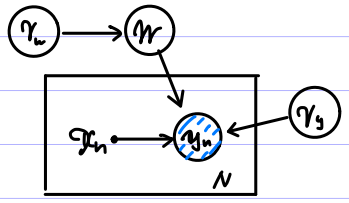
これにも事前分布を与えることで, 推論できるようにする.

- 以下, 精度パラメータ $\gamma_w = \sigma_w^{-2}$, $\gamma_y = \sigma_y^{-2}$ を利用する.

事前分布: $p(\gamma_w) = \mathrm{Gam}(\gamma_w | a_w, b_w)$ ← ガンマ分布は $\gamma_w > 0$ で確率が非ゼロ. 正の実数をとる値の分布として使える.

$p(\gamma_y) = \mathrm{Gam}(\gamma_y | a_y, b_y)$

$(a_w, b_w, a_y, b_y > 0$: 固定値$)$



- モデル.

$$p(\mathcal{Y}, w, \gamma_w, \gamma_y | \mathcal{X}) = p(\gamma_w) p(\gamma_y) p(w|\gamma_w) \prod_{n=1}^{N} p(y_n | x_n, w, \gamma_y)$$

- 事後分布は, $p(w, \gamma_w, \gamma_y | \mathcal{Y}, \mathcal{X})$.

- Gibbsサンプリングをする.

- γ_w, γ_y : given として, w の条件付き分布は $p(w | \mathcal{Y}, \mathcal{X}, \gamma_w, \gamma_y)$.

これは w の事後分布なので, HMCで w をサンプリングできる

- w, γ_y : given として. γ_w の条件付き分布は,

$$p(\gamma_w | \mathcal{Y}, \mathcal{X}, w, \gamma_y)$$

$$\propto p(\mathcal{Y}, w, \gamma_y, \gamma_w | \mathcal{X})$$

$$\propto p(w|\gamma_w) p(\gamma_w).$$

$p(w|\gamma_w)$ は Gauss分布, $p(\gamma_w)$ はその共役事前分布のガンマ分布なので,

条件付き分布 $p(\gamma_w | \gamma_y, \mathcal{X}, W, \gamma_y)$ はガンマ分布. $K_w :=$ (Wのパラメータ数) として,

Ch.3 での結果より

$$\gamma_w \sim \mathrm{Gam}(\hat{a}_w, \hat{b}_w), \quad \hat{a}_w = a_w + \frac{1}{2}K_w, \quad \hat{b}_w = b_w + \frac{1}{2}\sum_{w \in W} w^2$$

とサンプリングすればよい.

- W, γ_w : given として. γ_y の条件付き分布は,

$$p(\gamma_y | \mathcal{Y}, \mathcal{X}, W, \gamma_w)$$

$$\propto p(\mathcal{Y}, W, \gamma_y, \gamma_w | \mathcal{X})$$

$$\propto p(\gamma_y)\, p(\mathcal{Y} | \mathcal{X}, W, \gamma_y).$$

$p(\mathcal{Y} | \mathcal{X}, W, \gamma_y)$ は Gauss 分布で, $p(\gamma_y)$ はガンマ分布なので,

$p(\gamma_y | \mathcal{Y}, \mathcal{X}, W, \gamma_w)$ もガンマ分布.

$$\therefore \quad \gamma_y \sim \mathrm{Gam}(\hat{a}_y, \hat{b}_y),$$

$$\hat{a}_y = a_y + \frac{1}{2}N,$$

$$\hat{b}_y = b_y + \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n; W))^2 = b_y + E(W).$$

現在の W で表現できない NN の誤差の情報を ハイパーパラメータに加味している.

- ガンマ分布の平均 : $\mathbb{E}[\gamma_y] = \frac{\hat{a}_w}{\hat{b}_w}$

→ $\hat{b}_w$ が大きい ⇔ 誤差 $E(W)$ が大きい.

このとき $\mathbb{E}[\gamma_y]$ は小さくなるので, y_n に対する分散は大きくなる (γ_y の逆数)

- 重みのハイパーパラメータ γ_w は共通としていたが, 層ごとにパラメータを変える, などの改変もできる.

§5.2 近似Bayes推論の効率化

- Bayes NN:

✗ パラメータの周辺化に伴う計算量が膨大.

学習に必要なデータ数が多い

→ ✗ バッチ学習は効率が良くない.

- 巨大なネットワークや膨大なデータ量に対しても, 近似的なBayes推論ができるようにする.

Q ミニバッチを利用した効率的な学習.

5.2.1 確率的勾配Langevin動力学法による学習.

- SGD

✓ 大規模なNNモデルの効率的な学習.

✗ 正則化項を追加した最適化やMAP推定(点推定)などでは,

パラメータの不確実性を取り扱えない.

→ 過適合をおこす.

不確実性に基づく予測やモデルの評価ができない.

- HMC ← 勾配を利用したサンプリング方法

  - よく利用されている

✗ (MCMC全般的に)大規模データに対し, 計算効率が悪い.

・確率的勾配Langevin動力学法

(Stochastic gradient Langevin dynamics method)

・SGD + Langevin dynamics method

✓ 計算効率が高い.

✓ 不確実性の推定が可能.

・確率的MCMCの一種.

・SGDのアルゴリズムを用いて，NNの正則化項付きのコスト関数を最適化.

ミニバッチを指定する添字集合をSと書く. $|S| = M$.

ミニバッチ: $\mathcal{D}_S = \{(x_n, y_n)\}_{n \in S}$.

$$\mathcal{X}_S := \{x_n\}_{n \in S}, \quad \mathcal{Y}_S := \{y_n\}_{n \in S}$$

MAP推定

$$W_{MAP} = \underset{W}{\operatorname{argmax}} \log p(W | \mathcal{Y}_S, \mathcal{X}_S)$$

が正則化項付きのコスト関数の最小化と等価だった

$$\begin{aligned} &\nabla_W \log p(W | \mathcal{Y}_S, \mathcal{X}_S) \\ &= \nabla_W \log p(\mathcal{Y}_S | W, \mathcal{X}_S) + \nabla_W \log p(W) - \nabla_W \underline{\log p(\mathcal{Y}_S | \mathcal{X}_S)} \quad \text{（}W\text{に依らない}\text{）} \\ &= \sum_{n \in S} \nabla_W \log p(y_n | x_n, W) + \nabla_W \log p(W). \end{aligned}$$

であり，ミニバッチに対するコスト関数が，期待値としては全データに対するコスト関数と等価になるよう係数を調整すると，

t 回目のパラメータの更新は

$$W \leftarrow W + \frac{\alpha_t}{2}\left(\frac{N}{M}\sum_{n\in S}\nabla_W \log p(y_n|x_n,W) + \nabla_W \log p(W)\right)$$

係数は便宜上こう設定する

とすればよい．学習率 α_t は最適解への収束のため

$$\sum_{i=1}^{\infty}\alpha_i = \infty \ , \quad \sum_{i=1}^{\infty}\alpha_i^2 < \infty$$

をみたすようにとる．

- Langevin動力学法で事後分布 $p(W|Y,X)$ からサンプル候補 W^* を得る．

ポテンシャルエネルギーを $\mathcal{U} = -\log p(W|Y,X)$，

運動量ベクトルを $\mathbf{p} \sim \mathcal{N}(\mathbf{p}|0,I)$ とし，

ステップサイズを $\varepsilon = \sqrt{\alpha_t}$ とすると，

$$W^* = W - \frac{\varepsilon^2}{2}\nabla_W \mathcal{U} + \varepsilon \mathbf{p}$$

$$= W + \frac{\alpha_t}{2}\left(\sum_{n=1}^{N}\nabla_W \log p(y_n|x_n,W) + \nabla_W \log p(W)\right) + \sqrt{\alpha_t}\,\mathbf{p}.$$

- W^* の受容率は，ε を小さく（i.e., α_t を小さく）していくと 1 に近づく．

シミュレーション誤差が小さくなるから．

- 上の二つのアナロジーで，Langevin動力学法のミニバッチ版を

$$W \leftarrow W + \frac{\alpha_t}{2}\left(\frac{N}{M}\sum_{n\in S}\nabla_W \log p(y_n|x_n,W) + \nabla_W \log p(W)\right) + \sqrt{\alpha_t}\,\mathbf{p}$$

$$\mathbf{p} \sim \mathcal{N}(\mathbf{p}|0,I)$$

とする．

✓ $\nabla_W \log p(W|Y,X)$ の不偏推定量が得られる．

✓ $t \to \infty$ でサンプル W の受容率が漸近的に 1．

⇒ はじめは SGD の利点を生かして W を効率的に探索.

t が大きくなるにつれて Langevin 動力学法による真の事後分布からの

近似的なサンプルを得ることができるようになる.

5.2.2 確率的変分推論法による学習

- **確率的変分推論法** (stochastic variational inference method)

SGD+変分推論 → 効率的に学習

- $p(W|Y,X) \approx q(W;\xi)$ と近似.

ここでは, $\xi = \{\mu_{ij}^{(l)}, \sigma_{ij}^{(l)2}\}_{i,j,l}$ として $q(W;\xi) = \prod_{i,j,l} \mathcal{N}(w_{ij}^{(l)} | \mu_{ij}^{(l)}, \sigma_{ij}^{(l)2})$ とする.

ELBO は,

$$\begin{aligned}
\mathcal{L}(\xi) &= \int q(W;\xi) \log \frac{p(W,Y|X)}{q(W;\xi)} dW \\
&= \int q(W;\xi) \left(\log \frac{p(Y|W,X)}{q(W;\xi)} + \log p(W) \right) dW \\
&= \int q(W;\xi) \log p(Y|W,X) dW - D_{KL}[q(W;\xi) \| p(W)] \\
&= \sum_{n=1}^{N} \int q(W;\xi) \log p(y_n | f(x_n;W)) dW - D_{KL}[q(W;\xi) \| p(W)].
\end{aligned}$$

- $\mathcal{L}(\xi)$ を勾配降下法で ξ について最大化しようとすると (最大化なので「上昇法」), 各ステップでの

$\nabla_\xi \mathcal{L}(\xi)$ の評価のために全データを読みこむ必要がある.

→ 効率的でない.

- ミニバッチ $\mathcal{D}_s$ に対する部分的な ELBO を以下のように設定:

$$\mathcal{L}_s(\xi) := \frac{N}{M} \sum_{n \in s} \int q(W;\xi) \log p(y_n | f(x_n;W)) dW - D_{KL}[q(W;\xi) \| p(W)].$$

→ $\mathbb{E}_\delta[\mathcal{L}_\delta(\xi)] = \mathcal{L}(\xi)$ (不偏推定量) となる.

(pf. は Ch.2 でやったのと同様.)

$\mathcal{L}(\xi)$ を最大化する代わりに, $\mathcal{L}_\delta(\xi)$ を最大化するようにすればよい.

5.2.3 勾配のモンテカルロ近似.

- $\mathcal{L}_\delta(\xi)$ に対して勾配降下法を行うには, $\nabla_\xi \mathcal{L}_\delta(\xi)$ を計算する必要がある
- $\nabla_\xi D_{KL}[q_\phi(w;\xi)\|p(w)]$ は, どちらも Gauss 分布なので解析的に計算できる
- $\nabla_\xi \int q_\phi(w;\xi)\log p(y_n|f(x_n;w))\,dw$ は解析的に計算できない.

→ モンテカルロ法で積分を近似し, 勾配の推定を得る.

- 簡易的に表記する　パラメータ $w\in\mathbb{R}$ に対して, ある関数 $f(w)$ と

分布 $q_\phi(w;\xi)$ を考えて, 勾配

$$I(\xi) = \nabla_\xi \int f(w) q_\phi(w;\xi)\,dw$$

を評価したい.

5.2.3.1 スコア関数推定

- 分布 $q_\phi(w;\xi)$ に対し, $\nabla_\xi \log q_\phi(w;\xi)$ を**スコア関数**という.

$$
\begin{aligned}
&I(\xi) \\
&= \int f(w)\nabla_\xi q_\phi(w;\xi)\,dw \qquad \text{(微分と積分の交換)} \\
&= \int f(w) q_\phi(w;\xi)\nabla_\xi \log q_\phi(w;\xi)\,dw \qquad \text{(スコア関数)} \\
&= \mathbb{E}_{q_\phi(w;\xi)}[f(w)\nabla_\xi \log q_\phi(w;\xi)].
\end{aligned}
$$

→ 分布 $q(w;\xi)$ から w を複数サンプリングし，$f(w)\nabla_\xi \log q(w;\xi)$ のサンプル平均をとることで $I(\xi)$ を推定できる．

✓ $\nabla_\xi \log q(w;\xi)$ が計算できれば利用できる．

✗ 分散が大きい． → 制御変量法（control variates method）と併用．

5.2.3.2 再パラメータ化勾配．

・再パラメータ化勾配（reparameterization gradient）

$w \sim q(w;\xi)$ とする代わりに，変分パラメータ ξ のない分布 $p(\varepsilon)$ から $\varepsilon \sim p(\varepsilon)$ とサンプリングし，$w = g(\xi;\varepsilon)$ と変換することで w を得る．

（↑ w が q に従うように p と g を設定する．）

→ $I(\xi)$

（$q(w;\xi)dw = p(\varepsilon)d\varepsilon$）

$$= \nabla_\xi \int f(g(\xi;\varepsilon))p(\varepsilon)d\varepsilon$$

$$= \int f'(g(\xi;\varepsilon))\left(\nabla_\xi g(\xi;\varepsilon)\right)p(\varepsilon)d\varepsilon$$

$$= \mathbb{E}_{p(\varepsilon)}\left[f'(g(\xi;\varepsilon))\nabla_\xi g(\xi;\varepsilon)\right].$$

分布 $p(\varepsilon)$ から ε を複数サンプリングし，$f'(g(\xi;\varepsilon))\nabla_\xi g(\xi;\varepsilon)$ のサンプル平均をとることで $I(\xi)$ を推定できる．

ex) $\xi = \{\hat{\mu}, \hat{\sigma}\}$ として，$q(w;\xi) = \mathcal{N}(w|\hat{\mu}, \hat{\sigma}^2)$ の場合を考える．

$\tilde{\varepsilon} \sim \mathcal{N}(0,1)$ とし，$\tilde{w} = g(\xi;\varepsilon) = \hat{\mu} + \hat{\sigma}\tilde{\varepsilon}$ とすれば

$\tilde{w} \sim \mathcal{N}(w|\hat{\mu}, \hat{\sigma}^2)$ となる．

このとき，$\xi = \{\hat{\mu}, \hat{\sigma}\}$ に関する勾配は，

$$\frac{\partial}{\partial \hat{\mu}} \int f(w) q_\phi(w;\xi)\, dw = \int f'(w) q_\phi(w;\xi)\, dw = \mathbb{E}_{q_\phi(w;\xi)}[f'(w)].$$

$$\frac{\partial}{\partial \hat{\sigma}} \int f(w) q_\phi(w;\xi)\, dw = \int f'(w)\, \tilde{\varepsilon}\, q_\phi(w;\xi)\, dw = \int f'(w) \frac{w-\hat{\mu}}{\hat{\sigma}} q_\phi(w;\xi)\, dw$$

$$= \mathbb{E}_{q_\phi(w;\xi)}\left[f'(w) \frac{w-\hat{\mu}}{\hat{\sigma}} \right].$$

5.2.3.3 再パラメータ化勾配の一般化.

- 再パラメータ化勾配

✓ 勾配の分散を小さく抑えられる傾向がある.

✗ ξ に依存しない $p(\varepsilon)$ をつくれないことが多い.

・ ガンマ分布, ベータ分布などで, 再パラメータ化勾配による勾配推定ができない.

- 一般化再パラメータ化勾配 (generalized reparametrization gradient):

g に関する制約を緩めて, ε の分布に対して ξ の依存性が残るのを許容する ($\varepsilon \sim p(\varepsilon;\xi)$).

- 陰関数微分 (implicit differentiation) を用いる方法.

・ g を求めるのは困難だが, g^{-1} は容易に得られるケース.

ex) ガンマ分布, Dirichlet分布, von Mises分布, etc.

cf.) von Mises分布: 角度 $\theta \in [0, 2\pi)$ の分布

pdf: $p(\theta) = \frac{1}{2\pi I_0(\beta)} \exp(\beta \cos(\theta-\mu))$

$\mu \in [0, 2\pi)$, $\beta \geq 0$: パラメータ

$I_j(\beta) = \left(\frac{\beta}{2}\right)^j \sum_{i=0}^{\infty} \frac{1}{i!\,\Gamma(j+i+1)} \left(\frac{\beta^2}{4}\right)^i$ (j次第一種変形Bessel関数)

・逆変換 $\varepsilon = g^{-1}(\xi, w)$ を ξ に関して微分

→ g を介さずに期待値の勾配を得られる.

・w のサンプルを得るのには（g が計算できないので）棄却サンプリングなどを使う必要がある.

・離散の確率変数に対する再パラメータ化勾配を用いる方法もある.

・Gumbel ソフトマックス分布（連続分布）で,

分布の温度パラメータを 0 にすることでカテゴリ分布に一致させることができる.

→ カテゴリ分布の Gumbel ソフトマックス分布による連続緩和ができる.

5.2.4 勾配近似による変分推論法.

再パラメータ化勾配を利用して, ミニバッチ $\mathcal{D}_s = \{(x_n, y_n)\}_{n \in s}$ に対する

ELBO

$$\mathcal{L}_s(\xi)$$

$$:= \frac{N}{M} \sum_{n \in s} \int q(w;\xi) \log p(y_n | f(x_n; w)) dw - D_{KL}[q(w;\xi) \| p(w)]$$

を最大化する

→ 事後分布の近似精度を上げる.

$\varepsilon \sim \mathcal{N}(0, I)$, $w = g(\xi; \varepsilon)$ と変数変換.

$q(w;\xi)\, dw = p(\varepsilon)\, d\varepsilon$ に注意して,

$$\mathcal{L}_s(\xi) = \frac{N}{M} \sum_{n \in s} \int p(\varepsilon) \log p(y_n | f(x_n; g(\xi;\varepsilon)) d\varepsilon - D_{KL}[q(w;\xi) \| p(w)].$$

すごい雑だけど計算効率的に
↙ 大変なので仕方ない. 複数サンプリング してももちろんよい.

積分を, 1回のサンプル値 $\tilde{\varepsilon} \sim \mathcal{N}(\varepsilon|0, I)$ で近似する.

↑ データ x_n, y_n ごとにサンプリングする

$$\mathcal{L}_{\mathcal{B}}(\xi) \approx \mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)$$

$$= \frac{N}{M} \sum_{n \in \mathcal{B}} \log p(y_n | f(x_n; g(\xi; \tilde{\varepsilon})) - D_{KL}[q(w;\xi) \| p(w)]$$

$\mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)$ は $\mathcal{L}(\xi)$ の不偏推定量 : $\mathbb{E}_{\mathcal{B},\tilde{\varepsilon}}[\mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)] = \mathcal{L}(\xi)$.

・計算すれば分かる.

・勾配を近似すると,

$$\nabla_{\xi} \mathcal{L}_{\mathcal{B}}(\xi) \approx \nabla_{\xi} \mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)$$

$$= \frac{N}{M} \sum_{n \in \mathcal{B}} \underbrace{\nabla_{\xi} \log p(y_n | f(x_n; g(\xi; \tilde{\varepsilon}))}_{\text{↑ backpropagation で計算できる.}} - \underbrace{\nabla_{\xi} D_{KL}[q(w;\xi) \| p(w)]}_{\text{解析的に求まる.}}$$

まとめると, 以下のように ξ を最適化する.

1. ミニバッチ $\mathcal{D}_{\mathcal{B}}$ を $\mathcal{D}$ からつくる.
2. M 個のノイズ $\tilde{\varepsilon}_i \sim \mathcal{N}(0, I)$ を取得.
3. $\nabla_{\xi} \mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)$ を計算.
4. $\xi \leftarrow \xi + \alpha \nabla_{\xi} \mathcal{L}_{\mathcal{B},\tilde{\varepsilon}}(\xi)$ と更新. 1へ.

5.2.5 期待値伝播法による学習

・確率的逆伝播法 (probabilistic backpropagation method)

・順伝播 : ネットワークを通して確率を伝播, 周辺尤度の評価.

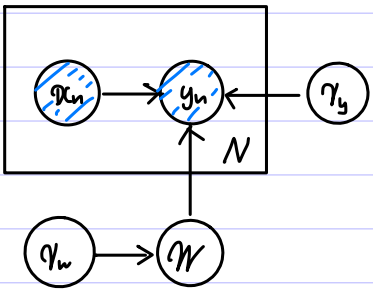
・逆伝播 : パラメータ学習のための周辺尤度の勾配計算.

✓ データを逐次的に処理でき，大量データでの学習ができる．

✓ ハイパーパラメータの近似推論．

5.2.5.1 モデル

$y_n \in \mathbb{R}$ ：1次元のラベルで考える．



モデル：

$$p(y, W, \gamma_y, \gamma_w \mid X) = p(y \mid X, W, \gamma_y) p(W \mid \gamma_w) p(\gamma_y) p(\gamma_w).$$

$$p(y \mid X, W, \gamma_y) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1})$$

$$p(\gamma_y) = \mathrm{Gam}(\gamma_y \mid \alpha_{\gamma_{y0}}, \beta_{\gamma_{y0}}) \quad (\alpha_{\gamma_{y0}}, \beta_{\gamma_{y0}} : \text{ハイパーパラメータ})$$

$$p(W \mid \gamma_w) = \prod_{l=1}^{L} \prod_{i=1}^{H_l} \prod_{j=1}^{H_{l-1}} \mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1}) \quad \leftarrow \text{全て独立}$$

$$p(\gamma_w) = \mathrm{Gam}(\gamma_w \mid \alpha_{\gamma_{w0}}, \beta_{\gamma_{w0}}) \quad (\alpha_{\gamma_{w0}}, \beta_{\gamma_{w0}} : \text{ハイパーパラメータ})$$

ここでは，NN $f(x_n; W)$ を次のように設定している：

$$z_n^{(0)} = x_n, \quad z_n^{(l)} = \mathrm{ReLU}(a_n^{(l)}) \quad (l = 1, \dots, L-1)$$

$$a_n^{(l)} = \frac{1}{\sqrt{H_{l-1}}} W^{(l)} z_n^{(l-1)} \quad (l = 1, \dots, L) \quad \leftarrow \text{係数を調整}$$

$$f(x_n; W) = z_n^{(L)} = a_n^{(L)}$$

- 事後分布

$$p(W, \gamma_y, \gamma_w \mid y, X)$$

$$\propto \frac{1}{p(y \mid X)} p(y, W, \gamma_y, \gamma_w \mid X)$$

$$\propto p(y \mid X, W, \gamma_y) p(W \mid \gamma_w) p(\gamma_y) p(\gamma_w)$$

5.2.5.2 近似分布.

- 近似分布を次のとおりに設定（事前分布と同じ種類の分布を選択）：

$$p(W, \gamma_y, \gamma_w | Y, X) \approx q_f(W, \gamma_y, \gamma_w),$$

$$q_f(W, \gamma_y, \gamma_w) = \frac{\mathrm{Gam}(\gamma_y | \alpha_{\gamma_y}, \beta_{\gamma_y})}{q_f(\gamma_y)} \frac{\mathrm{Gam}(\gamma_w | \alpha_{\gamma_w}, \beta_{\gamma_w})}{q_f(\gamma_w)} \underbrace{\prod_{l=1}^{L} \prod_{i=1}^{H_l} \prod_{j=1}^{H_{l-1}} \mathcal{N}(w_{i,j}^{(l)} | m_{i,j}^{(l)}, v_{i,j}^{(l)})}_{q_f(W) = \prod_l \prod_i \prod_j q_f(w_{i,j}^{(l)}),\ q_f(w_{i,j}^{(l)}) = \mathcal{N}(\cdots)}$$

$$= q_f(W) q_f(\gamma_y) q_f(\gamma_w).$$

5.2.5.3 初期化と事前分布因子の導入.

$q_f(W, \gamma_y, \gamma_w)$ が"無情報"になるように，パラメータを設定：（"一様"ぐらいの意味.）

$$m_{i,j}^{(l)} = 0, \quad v_{i,j}^{(l)} = \infty$$

$$\alpha_{\gamma_y} = 1, \quad \beta_{\gamma_y} = 0$$

$$\alpha_{\gamma_w} = 1, \quad \beta_{\gamma_w} = 0$$

この辺とすぐに更新されるのであまり気にしなくていいかも.

cf.) 無情報事前分布（noninformative prior），非正則事前分布（improper prior）

- やること：逐次学習

$$q_{i+1}(\theta) \approx \frac{1}{Z_{i+1}} f_{i+1}(\theta) q_i(\theta), \quad \theta = \{W, \gamma_y, \gamma_w\}.$$

$f_{i+1}(\theta)$ として

$$p(W, \gamma_y, \gamma_w | Y, X)$$

$$\propto p(Y | X, W, \gamma_y) p(W | \gamma_w) p(\gamma_y) p(\gamma_w)$$

$$= \underbrace{\left(\prod_{n=1}^{N} \mathcal{N}(y_n | f(x_n; W), \gamma_y^{-1}) \right)}_{\text{尤度因子}} \underbrace{\left(\prod_{l=1}^{L} \prod_{i=1}^{H_l} \prod_{j=1}^{H_{l-1}} \mathcal{N}(w_{i,j}^{(l)} | 0, \gamma_w^{-1}) \right) p(\gamma_y) p(\gamma_w)}_{\text{事前分布因子}}$$

の各因子を次々と設定することで $q(\theta)$ を更新していく.

・まずは事前分布因子の情報を追加していく：

f として $p(\gamma_w)$, $p(\gamma_y)$, $\mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1})$ を設定して，

$q(\gamma_w)$, $q(\gamma_y)$, $q(W)$ のパラメータを更新．

・続いて，尤度因子の情報を追加していく：

f として $\mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1})$ を設定して

$q(\gamma_y)$, $q(W)$ のパラメータを更新．

1. $p(\gamma_y)p(\gamma_w)$ の導入．

$q(\gamma_y)q(\gamma_w)$ と同じ形にしているので，q のパラメータを単純に

$$\alpha_{\gamma_y} \leftarrow \alpha_{\gamma_{y0}},\quad \beta_{\gamma_y} \leftarrow \beta_{\gamma_{y0}},\quad \alpha_{\gamma_w} \leftarrow \alpha_{\gamma_{w0}},\quad \beta_{\gamma_w} \leftarrow \beta_{\gamma_{w0}}$$

と更新すればよい．

2. $\mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1})$ の導入．

$$\begin{aligned}
&q_{new}(w_{i,j}^{(l)})\, q_{new}(\gamma_w) \\
&\approx \frac{1}{Z_0}\mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1})\, q(w_{i,j}^{(l)})\, q(\gamma_w) \\
&= \frac{1}{Z_0}\mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1})\, \mathcal{N}(w_{i,j}^{(l)} \mid m_{i,j}^{(l)}, v_{i,j}^{(l)})\, \mathrm{Gam}(\gamma_w \mid \alpha_{\gamma_w}, \beta_{\gamma_w}),
\end{aligned}$$

$$Z_0 = Z(\alpha_{\gamma_w}, \beta_{\gamma_w})$$

$$= \int \mathcal{N}(w_{i,j}^{(l)} \mid 0, \gamma_w^{-1})\, \mathcal{N}(w_{i,j}^{(l)} \mid m_{i,j}^{(l)}, v_{i,j}^{(l)})\, \mathrm{Gam}(\gamma_w \mid \alpha_{\gamma_w}, \beta_{\gamma_w})\, dw_{i,j}^{(l)}\, d\gamma_w$$

- $m_{ij}^{(l)}$, $v_{ij}^{(l)}$ の更新. §4.2.4.3 の結果から,

$$m_{ij}^{(l)} \leftarrow m_{ij}^{(l)} + v_{ij}^{(l)} \frac{\partial}{\partial m_{ij}^{(l)}} \log Z_0$$

$$v_{ij}^{(l)} \leftarrow v_{ij}^{(l)} - v_{ij}^{(l)2} \left(\left(\frac{\partial}{\partial m_{ij}^{(l)}} \log Z_0 \right)^2 - 2 \frac{\partial}{\partial v_{ij}^{(l)}} \log Z_0 \right).$$

- α_{γ_w}, β_{γ_w} の更新. §4.2.4.4 の結果から,

$$\alpha_{\gamma_w} \leftarrow \left(Z_0 Z_2 Z_1^{-1} \frac{\alpha_{\gamma_w}+1}{\alpha_{\gamma_w}} - 1 \right)^{-1} \qquad Z_1 = Z(\alpha_{\gamma_w}+1, \beta_{\gamma_w}),$$

$$\beta_{\gamma_w} \leftarrow \left(Z_2 Z_1^{-1} \frac{\alpha_{\gamma_w}+1}{\beta_{\gamma_w}} - Z_1 Z_0^{-1} \frac{\alpha_{\gamma_w}}{\beta_{\gamma_w}} \right)^{-1} \qquad Z_2 = Z(\alpha_{\gamma_w}+2, \beta_{\gamma_w}).$$

$Z(\alpha_{\gamma_w}, \beta_{\gamma_w})$ を求める必要があるが,

厳密解は得られないので近似する.

$$Z(\alpha_{\gamma_w}, \beta_{\gamma_w})$$

$$= \int \mathcal{N}(w_{ij}^{(l)} | 0, \gamma_w^{-1}) \mathcal{N}(w_{ij}^{(l)} | m_{ij}^{(l)}, v_{ij}^{(l)}) \mathrm{Gam}(\gamma_w | \alpha_{\gamma_w}, \beta_{\gamma_w}) dw_{ij}^{(l)} d\gamma_w$$

$$= \int \left(\int \mathcal{N}(w_{ij}^{(l)} | 0, \gamma_w^{-1}) \mathrm{Gam}(\gamma_w | \alpha_{\gamma_w}, \beta_{\gamma_w}) d\gamma_w \right) \mathcal{N}(w_{ij}^{(l)} | m_{ij}^{(l)}, v_{ij}^{(l)}) dw_{ij}^{(l)}$$

↓ §3.2.4.4 の予測分布

$$= \int \mathrm{St}(w_{ij}^{(l)} | 0, \frac{\alpha_{\gamma_w}}{\beta_{\gamma_w}}, 2\alpha_{\gamma_w}) \mathcal{N}(w_{ij}^{(l)} | m_{ij}^{(l)}, v_{ij}^{(l)}) dw_{ij}^{(l)}$$

↓ Student の t 分布を平均と分散の等しい Gauss 分布で近似

$$\approx \int \mathcal{N}(w_{ij}^{(l)} | 0, \frac{\beta_{\gamma_w}}{\alpha_{\gamma_w}-1}) \mathcal{N}(w_{ij}^{(l)} | m_{ij}^{(l)}, v_{ij}^{(l)}) dw_{ij}^{(l)}.$$

ここで,

$$\int \mathcal{N}(w | 0, \lambda_1^{-1}) \mathcal{N}(w | m, \lambda_2^{-1}) dw$$

$$= \int \frac{\sqrt{\lambda_1}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda_1}{2} w^2\right) \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda_2}{2}(w-m)^2\right) dw$$

$$= \frac{\sqrt{\lambda_1}}{\sqrt{2\pi}} \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}\left((\lambda_1+\lambda_2)w^2 - 2\lambda_2 m w + \lambda_2 m^2\right)\right) dw$$

↓ 平方完成

$$= \frac{\sqrt{\lambda_1}}{\sqrt{2\pi}} \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} \int \exp\left(-\frac{\lambda_1+\lambda_2}{2}\left(w - \frac{\lambda_2}{\lambda_1+\lambda_2} m\right)^2\right) \exp\left(-\frac{1}{2} \frac{\lambda_2\lambda_2}{\lambda_1+\lambda_2} m^2\right) dw$$

$$= \frac{\sqrt{\lambda_1}}{\sqrt{2\pi}} \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} m^2\right) \underbrace{\int \exp\left(-\frac{\lambda_1+\lambda_2}{2}\left(w - \frac{\lambda_2}{\lambda_1+\lambda_2} m\right)^2\right) dw}_{= \sqrt{\frac{2\pi}{\lambda_1+\lambda_2}}}$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}} \exp\left(-\frac{1}{2}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} m^2\right)$$

$$= \mathcal{N}\left(m \mid 0, \left(\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\right)^{-1}\right)$$

なので，

$$Z(\alpha_{\gamma_w}, \beta_{\gamma_w}) \approx \mathcal{N}\left(m_{ij}^{(l)} \mid 0, \frac{\beta_{\gamma_w}}{\alpha_{\gamma_w}-1} + V_{ij}^{(l)}\right).$$

$\frac{\partial}{\partial m_{ij}^{(l)}} \log Z_0$, $\frac{\partial}{\partial v_{ij}^{(l)}} \log Z_0$, Z_1, Z_2 はこれで計算できる．

5.2.5.4 尤度因子の導入．

3. $\mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1})$ の導入．

$$q_{new}(W) q_{new}(\gamma_y)$$

$$\approx \frac{1}{Z} \mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1}) q_f(W) q_f(\gamma_y)$$

$$= \frac{1}{Z} \mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1}) q_f(W) \mathrm{Gam}(\gamma_y \mid \alpha_{\gamma_y}, \beta_{\gamma_y}).$$

$$= \frac{1}{Z} \mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1}) \left(\prod_{l=1}^{L} \prod_{i=1}^{H_l} \prod_{j=1}^{H_{l-1}} q(w_{ij}^{(l)})\right) \mathrm{Gam}(\gamma_y \mid \alpha_{\gamma_y}, \beta_{\gamma_y}).$$

$$Z = Z(\alpha_{\gamma_y}, \beta_{\gamma_y})$$

$$= \int \mathcal{N}(y_n \mid f(x_n; W), \gamma_y^{-1}) q_f(W) \mathrm{Gam}(\gamma_y \mid \alpha_{\gamma_y}, \beta_{\gamma_y}) dW d\gamma_y$$

- $m_{ij}^{(l)}$, $v_{ij}^{(l)}$ の更新．§4.2.4.3 の結果から，

$$m_{ij}^{(l)} \leftarrow m_{ij}^{(l)} + v_{ij}^{(l)} \frac{\partial}{\partial m_{ij}^{(l)}} \log Z$$

$$v_{ij}^{(l)} \leftarrow v_{ij}^{(l)} - v_{ij}^{(l)2} \left(\left(\frac{\partial}{\partial m_{ij}^{(l)}} \log Z\right)^2 - 2 \frac{\partial}{\partial v_{ij}^{(l)}} \log Z\right).$$

・ α_{γ_y}, β_{γ_y} の更新. §4.2.4.4 の結果から,

$$\alpha_{\gamma_y} \leftarrow \left(Z Z_2 Z_1^{-1} \frac{\alpha_{\gamma_y}+1}{\alpha_{\gamma_y}} - 1 \right)^{-1} \qquad Z_1 = Z(\alpha_{\gamma_y}+1, \beta_{\gamma_y}),$$

$$\beta_{\gamma_y} \leftarrow \left(Z_2 Z_1^{-1} \frac{\alpha_{\gamma_y}+1}{\beta_{\gamma_y}} - Z_1 Z^{-1} \frac{\alpha_{\gamma_y}}{\beta_{\gamma_y}} \right)^{-1} \qquad Z_2 = Z(\alpha_{\gamma_y}+2, \beta_{\gamma_y}).$$

→ $Z = Z(\alpha_{\gamma_y}, \beta_{\gamma_y})$ が求まれば"パラメータの更新"が可能.

$W \sim q(W)$ のとき, $z^{(l)} = f(x_n; W)$ の分布は複雑になるので",

中心極限定理.

$z^{(l)} \sim \mathcal{N}(z^{(l)} | m_{z^{(l)}}, \sigma^2_{z^{(l)}})$ で"近似してしまう.

つまり, $q(W)dW = p(z^{(l)})dz^{(l)} \approx \mathcal{N}(z^{(l)} | m_{z^{(l)}}, \sigma^2_{z^{(l)}})dz^{(l)}$ とする.

Remark 式 $q(W)dW = p(z^{(l)})dz^{(l)}$ は, 変数変換の公式そのものではない.

(∵ 一般には W と $z^{(l)}$ の次元が違うので, $z^{(l)} = f(x_n; W)$ は全単射にならない)