Weight Uncertainty in Neural Networks (arXiv:1505.05424v2)

Thakur

Weight Uncertainty in Neural Networks (arXiv:1505.05424v2)

3.2 Gaussian Variational Posterior:

Suppose that the variational posterior is a diagonal Gaussian distribution, then a sample of the weights $\mathbf{w}$ can be obtained by sampling a unit Gaussian, shifting it by a mean μ and scaling by a standard deviation σ. We parameterise the standard deviation pointwise as $\sigma = \log(1 + \exp(\rho))$ and so σ is always non-negative. The variational posterior parameters are $\theta = (\mu, \rho)$. Thus the transform from a sample of parameter-free noise and the variational posterior parameters that yields a posterior sample of the weights $\mathbf{w}$ is: $\mathbf{w} = t(\theta, \epsilon) = \mu + \log(1 + \exp(\rho)) \circ \epsilon$ where $\circ$ is pointwise multiplication. Each step of optimisation proceeds as follows:

1. Sample $\epsilon \sim \mathcal{N}(0, I)$.
2. Let $\mathbf{w} = \mu + \log(1 + \exp(\rho)) \circ \epsilon$.
3. Let $\theta = (\mu, \rho)$.
4. Let $f(\mathbf{w}, \theta) = \log q(\mathbf{w}|\theta) - \log P(\mathbf{w})P(\mathcal{D}|\mathbf{w})$.
5. Calculate the gradient with respect to the mean

$$\Delta_\mu = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \mu}. \tag{3}$$

6. Calculate the gradient with respect to the standard deviation parameter ρ

$$\Delta_\rho = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} \frac{\epsilon}{1 + \exp(-\rho)} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \rho}. \tag{4}$$

7. Update the variational parameters:

$$\mu \leftarrow \mu - \alpha \Delta_\mu \tag{5}$$

$$\rho \leftarrow \rho - \alpha \Delta_\rho. \tag{6}$$

Derivation of weights and biases gradient for the objective function (1/n)

<u>Objective Function to Minimize Equ. 8 in the Paper</u>:

$$\mathcal{F}_i^{\mathrm{EQ}}(\mathcal{D}_i, \theta) = \frac{1}{M}\mathrm{KL}\left[q(\mathbf{w}|\theta) \;||\; P(\mathbf{w})\right] - \mathbb{E}_{q(\mathbf{w}|\theta)}\left[\log P(\mathcal{D}_i|\mathbf{w})\right]. \quad (8)$$

$$F(D, \theta) = \frac{1}{M}\{\log(q(w|\,\theta)\text{-log(p(w))-log(p(D|w))}\}$$

Where M is number of batches in a Epoch

$\log(q(w|\,\theta))$ → logarithmic of Postrior

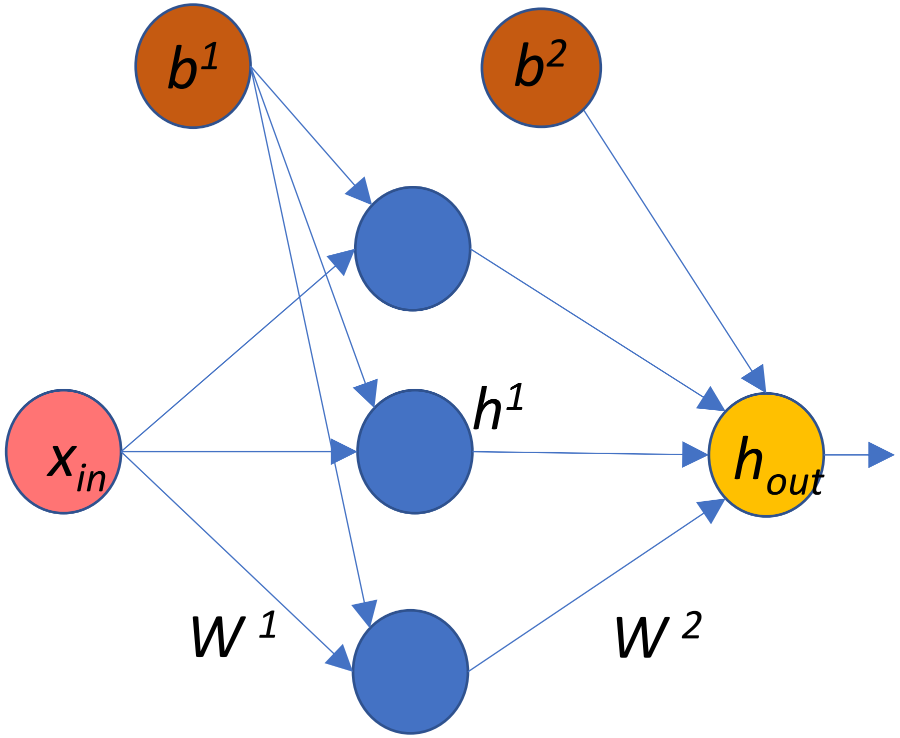
log(p(w)) → logarithmic of Prior

log(p(D|w)) → logarithmic of Likelihood

Derivation of weights and biases gradient for the objective function (2/n)

- Using a Simple example I am going to explain gradient calculation of Objective function.

Consider a 2-layer neural network for Sine Curve implementation



$h^1 = sigmoid(x_{in} * W^1 + b^1)$

$h_{out} = (h^1 * W^2 + b^2)$

x_{in} → Input value of Data point (Input)

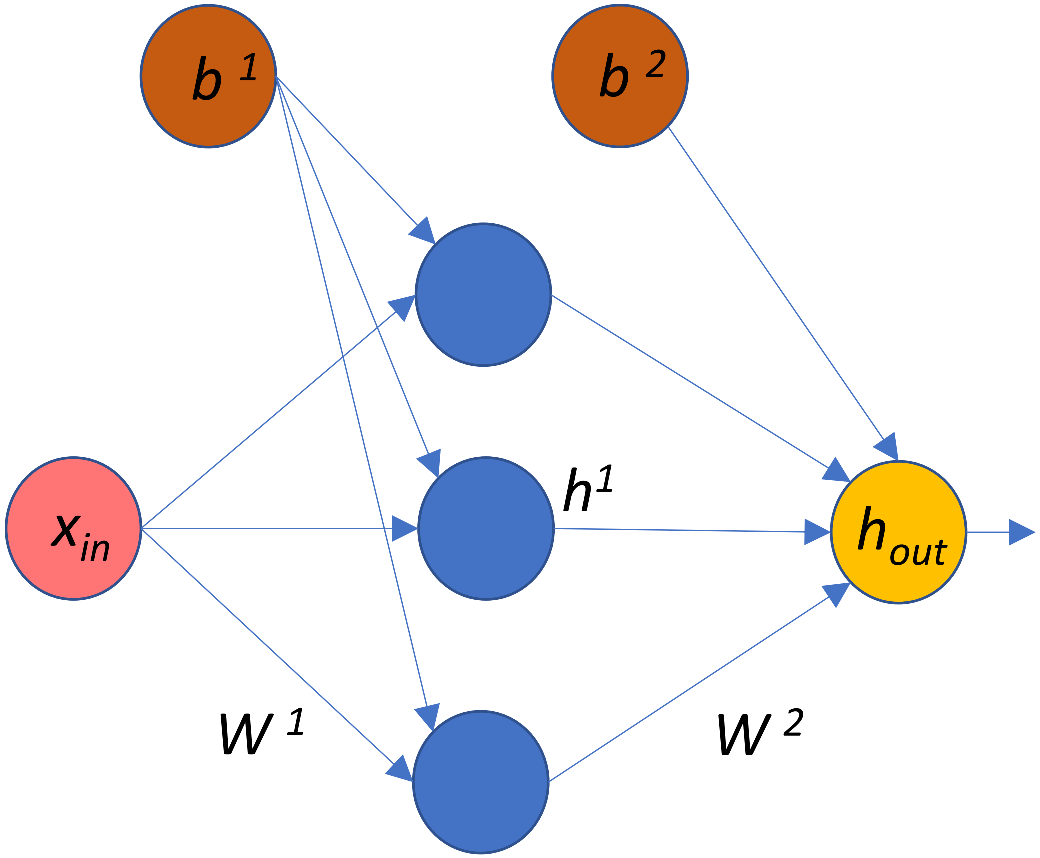
W^1, b^1 → Weights and Biases of Layer-1

W^2, b^2 → Weights and Biases of Layer-2

h_{out} → Output of neural network

In Matrix form: Shapes of weights and biases

x_{in} → [1 x 1]
W^1 → [1 x 3]
b^1 → [1 x 3]
W^2 → [3 x 1]
b^2 → [1 x 1]



$$z^1 = (x_{in} * W^1 + b^1)$$
$$h^1 = sigmoid(z^1)$$
$$h_{out} = (h^1 * W^2 + b^2)$$

Each Neural Network weights are combination of (μ, ρ)

$$\mathrm{W} = \mu + \log(1 + \exp(\rho)) \,.{*}\, \epsilon$$

$$\mathrm{b} = \mu + \log(1 + \exp(\rho)) \,.{*}\, \epsilon$$

In our case:

$$W\,1 = \mu_w{}^1 + log(1 + exp(\rho_w{}^1)) \,.{*}\, \epsilon_w{}^1$$

$$b\,1 = \mu_b{}^1 + log(1 + exp(\rho_b{}^1)) \,.{*}\, \epsilon_b{}^1$$

$$W\,2 = \mu_w{}^2 + log(1 + exp(\rho_w{}^2)) \,.{*}\, \epsilon_w{}^2$$

$$b\,2 = \mu_b{}^2 + log(1 + exp(\rho_b{}^2)) \,.{*}\, \epsilon_b{}^2$$

Here, .* is a Element-wise multiplication
ϵ = is a random variable (same shape as ρ)

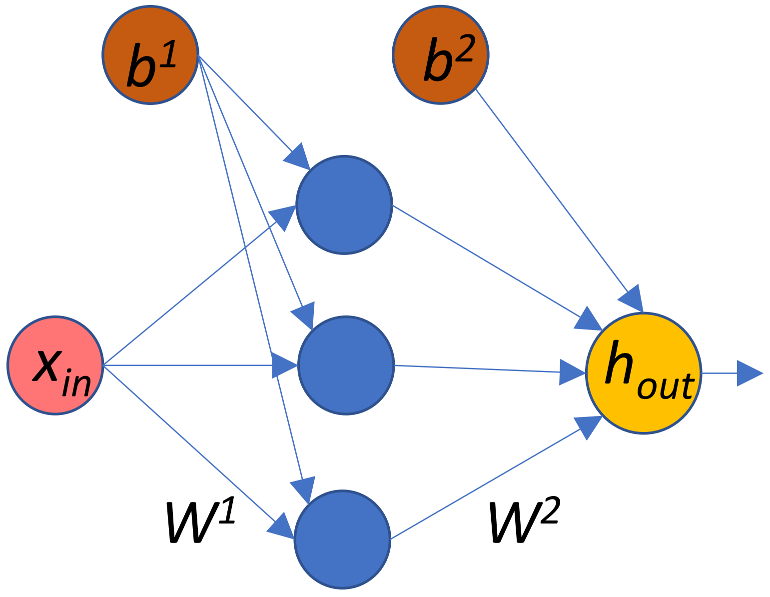
$$W1 = \mu_w{}^1 + log(1 + exp(\rho_w{}^1)) .* \epsilon_w{}^1$$

$$b1 = \mu_b{}^1 + log(1 + exp(\rho_b{}^1)) .* \epsilon_b{}^1$$

$$W2 = \mu_w{}^2 + log(1 + exp(\rho_w{}^2)) .* \epsilon_w{}^2$$

$$b2 = \mu_b{}^2 + log(1 + exp(\rho_b{}^2)) .* \epsilon_b{}^2$$

Here, .* is a Element-wise multiplication
ϵ = is a random variable



- At first time (epoch one, batch size), Sample μ, ρ, ϵ from a Gaussian distribution (mean=0.0, std=0.05)
- Each weight and bias has (μ, ρ).
- In the Bayesian, learning parameters are (μ, ρ) of each weight. i.e. we update (μ, ρ) during neural network training.
- Remember every time we sample ϵ a new value randomly.

Objective function:

$F(D, \theta) = \frac{1}{M}$ {log($q(w|\ \theta)$-log(p(w))}-log(p(D|w))

Negative log-likelihood {-log(p(D|w)} gradient calculation:

First, log-likelihood function:

Likelihood:

$$p(D|w) = \frac{1}{\sqrt{2\pi * \sigma}} * e^{-\frac{(y-h_{out})^2}{2\sigma^2}}$$

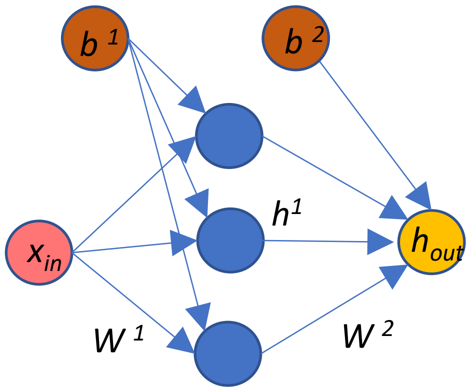
Log-likelihood:

$$logp(D|w) = -0.5 * \log(2\pi) - \log(\sigma) - \frac{(y - h_{out})^2}{2\sigma^2}$$

here, y → true label

$$\frac{\partial \log(p(D|w))}{\partial \mu_w^2}, \frac{\partial \log(p(D|w))}{\partial \rho_w^2}, \frac{\partial \log(p(D|w))}{\partial \mu_b^2}, \frac{\partial \log(p(D|w))}{\partial \rho_b^2} = ?$$

$$\frac{\partial \log(p(D|w))}{\partial \mu_w^1}, \frac{\partial \log(p(D|w))}{\partial \rho_w^1}, \frac{\partial \log(p(D|w))}{\partial \mu_b^1}, \frac{\partial \log(p(D|w))}{\partial \rho_b^1} = ?$$



Log-likelihood gradient with respect to μ, ρ

$$\frac{\partial \log(p(D|w))}{\partial \mu_w^2}, \frac{\partial \log(p(D|w))}{\partial \rho_w^2}, \frac{\partial \log(p(D|w))}{\partial \mu_b^2}, \frac{\partial \log(p(D|w))}{\partial \rho_b^2} = ?$$

$$logp(D|w) = -0.5 * \log(2\pi) - \log(\sigma) - \frac{(y - h_{out})^2}{2\sigma^2}$$

$$W\,1 = \mu_w{}^1 + log(1 + exp(\rho_w{}^1)) .* \epsilon_w{}^1$$
$$b\,1 = \mu_b{}^1 + log(1 + exp(\rho_b{}^1)) .* \epsilon_b{}^1$$
$$W\,2 = \mu_w{}^2 + log(1 + exp(\rho_w{}^2)) .* \epsilon_w{}^2$$
$$b\,2 = \mu_b{}^2 + log(1 + exp(\rho_b{}^2)) .* \epsilon_b{}^2$$

Here, $.*$ is a Element-wise multiplication

$\epsilon =$ is a random variable

$$z^1 = (x_{in} * W^1 + b^1)$$
$$h^1 = sigmoid(z^1)$$
$$h_{out} = (h^1 * W^2 + b^2)$$

$\dfrac{\partial \log(p(D|w))}{\partial \mu_w^2}$:

$$\frac{\partial \log(p(D|w))}{\partial \mu_w^2} = \frac{\partial \log(p(D|w))}{\partial h_{out}} * \frac{\partial h_{out}}{\partial W^2} * \frac{\partial W^2}{\partial \mu_w^2}$$

$$\frac{\partial \log(p(D|w))}{\partial h_{out}} = \frac{(y-hout)}{\sigma^2}, \quad \frac{\partial h_{out}}{\partial W^2} = h^1, \quad \frac{\partial W^2}{\partial \mu_w^2} = 1.0$$

$$\frac{\partial \log(p(D|w))}{\partial \mu_w^2} = \frac{(y-hout)}{\sigma^2} * h^1 \qquad \text{---}$$

(1)

$\dfrac{\partial \log(p(D|w))}{\partial \rho_w^2}$:

$$\frac{\partial \log(p(D|w))}{\partial \rho_w^2} = \frac{\partial \log(p(D|w))}{\partial h_{out}} * \frac{\partial h_{out}}{\partial W^2} * \frac{\partial W^2}{\partial \rho_w^2}$$

$$\frac{\partial \log(p(D|w))}{\partial h_{out}} = \frac{(y-hout)}{\sigma^2}$$

$$\frac{\partial h_{out}}{\partial W^2} = h^1$$

$$\frac{\partial W^2}{\partial \rho_w^2} = \epsilon_w^2 * \left(\frac{1}{1+e^{-\rho_w^2}}\right)$$

$$\frac{\partial \log(p(D|w))}{\partial \rho_w^2} = \frac{(y-hout)}{\sigma^2} * h^1 * \epsilon_w^2 * \left(\frac{1}{1+e^{-\rho_w^2}}\right) \qquad \text{--- (2)}$$

$\dfrac{\partial \log(p(D|w))}{\partial \mu_b^2}$:

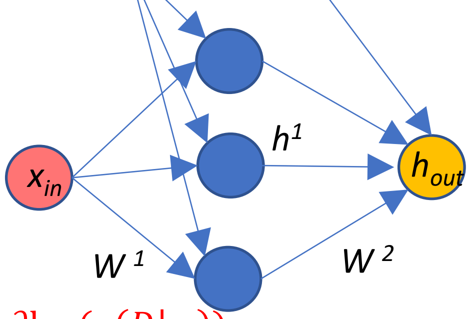
$$\frac{\partial \log(p(D|w))}{\partial \mu_b^2} = \frac{\partial \log(p(D|w))}{\partial h_{out}} * \frac{\partial h_{out}}{\partial b^2} * \frac{\partial b^2}{\partial \mu_b^2}$$

$$\frac{\partial \log(p(D|w))}{\partial \mu_b^2} = \frac{(y-hout)}{\sigma^2} * 1.0 * 1.0 \qquad \text{---- (3)}$$

$\dfrac{\partial \log(p(D|w))}{\partial \rho_b^2}$:

$$\frac{\partial \log(p(D|w))}{\partial \rho_b^2} = \frac{\partial \log(p(D|w))}{\partial h_{out}} * \frac{\partial h_{out}}{\partial b^2} * \frac{\partial b^2}{\partial \rho_b^2}$$

$$\frac{\partial \log(p(D|w))}{\partial \rho_b^2} = \frac{(y-hout)}{\sigma^2} * 1.0 * \epsilon_b^2 * \left(\frac{1}{1+e^{-\rho_b^2}}\right) \qquad \text{---- (4)}$$



Log-likelihood gradient with respect to μ, ρ

$$\frac{\partial\log(p(D|w))}{\partial\mu_w^1}, \frac{\partial\log(p(D|w))}{\partial\rho_w^1}, \frac{\partial\log(p(D|w))}{\partial\mu_b^1}, \frac{\partial\log(p(D|w))}{\partial\rho_b^1} = ?$$

$$W\,1 = \mu_w{}^1 + log(1 + exp(\rho_w{}^1))\,.*\,\epsilon_w{}^1$$
$$b\,1 = \mu_b{}^1 + log(1 + exp(\rho_b{}^1))\,.*\,\epsilon_b{}^1$$
$$W\,2 = \mu_w{}^2 + log(1 + exp(\rho_w{}^2))\,.*\,\epsilon_w{}^2$$
$$b\,2 = \mu_b{}^2 + log(1 + exp(\rho_b{}^2))\,.*\,\epsilon_b{}^2$$

Here, .* is a Element-wise multiplication

ϵ = is a random variable

$$z^1 = (x_{in} * W^1 + b^1)$$
$$h^1 = sigmoid(z^1)$$
$$h_{out} = (h^1 * W^2 + b^2)$$

$\frac{\partial\log(p(D|w))}{\partial\mu_w^1}$:

$$\frac{\partial\log(p(D|w))}{\partial\mu_w^1} = \frac{\partial\log(p(D|w))}{\partial h^1} * \frac{\partial h^1}{\partial z^1} * \frac{\partial z^1}{\partial W_w^1} * \frac{\partial W_w^1}{\partial \mu_w^1}$$

$$\frac{\partial\log(p(D|w))}{\partial h^1} = \frac{(y-hout)}{\sigma^2} * W^2$$

$$\frac{\partial h^1}{\partial z^1} = h^1 * (1 - h^1)$$ ➔ " derivation of sigmoid unit"

$$\frac{\partial z^1}{\partial W_w^1} = x_{in}\ , \quad \frac{\partial W_w^1}{\partial \mu_w^1} = 1.0$$

$$\frac{\partial\log(p(D|w))}{\partial\mu_w^1} = \frac{(y-hout)}{\sigma^2} * W^2 * h^1 * (1 - h^1) * x_{in} \qquad \text{-----(5)}$$

$\frac{\partial\log(p(D|w))}{\partial\rho_w^1}$:

$$\frac{\partial\log(p(D|w))}{\partial\rho_w^1} = \frac{\partial\log(p(D|w))}{\partial h^1} * \frac{\partial h^1}{\partial z^1} * \frac{\partial z^1}{\partial W_w^1} * \frac{\partial W_w^1}{\partial \rho_w^1}$$

$$\frac{\partial W_w^1}{\partial \rho_w^1} = \epsilon_w^1 * \frac{1.0}{(1+e^{-(\rho_w^1)})}$$

$$\frac{\partial\log(p(D|w))}{\partial\rho_w^1} = \frac{(y-hout)}{\sigma^2} * W^2 * h^1 * (1 - h^1) * x_{in} * \epsilon_w^1 * \frac{1.0}{(1+e^{-(\rho_w^1)})} \qquad \text{----- (6)}$$

$\frac{\partial\log(p(D|w))}{\partial\mu_b^1}$:

$$\frac{\partial\log(p(D|w))}{\partial\mu_b^2} = \frac{\partial\log(p(D|w))}{\partial h^1} * \frac{\partial h^1}{\partial z^1} * \frac{\partial z^1}{\partial b^1} * \frac{\partial b^1}{\partial \mu_b^1}$$

$$\frac{\partial\log(p(D|w))}{\partial\mu_b^2} = \frac{(y-hout)}{\sigma^2} * W^2 * h^1 * (1 - h^1) \qquad \text{---- (7)}$$

$\frac{\partial\log(p(D|w))}{\partial\rho_b^1}$:

$$\frac{\partial\log(p(D|w))}{\partial\rho_b^2} = \frac{\partial\log(p(D|w))}{\partial h^1} * \frac{\partial h^1}{\partial z^1} * \frac{\partial z^1}{\partial b^1} * \frac{\partial b^1}{\partial \mu_b^1} * \frac{\partial b^1}{\partial \rho_b^1}$$

$$\frac{\partial\log(p(D|w))}{\partial\rho_b^2} = \frac{(y-hout)}{\sigma^2} * W^2 * h^1 * (1 - h^1) * \epsilon_b^2 * \left(\frac{1}{1+e^{-\rho_b^2}}\right) \qquad \text{---- (8)}$$

Log-Prior gradient with respect to μ, ρ

$\underline{logp(w):}$

In the case of prior, mu=0.0, σ_P =0.05

$$Gaussian\ Prior:\ p(W) = \frac{1}{\sqrt{2\pi} * \sigma_P} * e^{-\frac{(W-mu)^2}{2\sigma_P^2}}$$

$$logp(w) = -0.5 * \log(2\pi) - \log(\sigma_P) - \frac{(W)^2}{2\sigma_P^2}$$

$$W\,1 = \mu_w{}^1 + log(1 + exp(\rho_w{}^1))\ .*\ \epsilon_w{}^1$$
$$b\,1 = \mu_b{}^1 + log(1 + exp(\rho_b{}^1))\ .*\ \epsilon_b{}^1$$
$$W\,2 = \mu_w{}^2 + log(1 + exp(\rho_w{}^2))\ .*\ \epsilon_w{}^2$$
$$b\,2 = \mu_b{}^2 + log(1 + exp(\rho_b{}^2))\ .*\ \epsilon_b{}^2$$

Here, .* is a Element-wise multiplication

$\epsilon =$ is a random variable

$$\frac{\partial \log(p(W^2))}{\partial \mu_w^2}, \frac{\partial \log(p(W^2))}{\partial \rho_w^2}, \frac{\partial \log(p(b^2))}{\partial \mu_b^2}, \frac{\partial \log(p(b^2))}{\partial \rho_b^2} = ?$$

$\frac{\partial \log(p(W^2))}{\partial \mu_w^2}$:

$$logp(W^{(2)}) = -0.5 * log(2\pi) - log(\sigma_P) - \frac{(W^{(2)})^2}{2\sigma_P^2}$$

$$\frac{\partial log(p(W^2))}{\partial \mu_w^2} = 0 - 0 - \frac{W^{(2)}}{\sigma_p^2} * \frac{\partial(W^2)}{\partial \mu_w^2} \quad (since\ \frac{\partial(W^2)}{\partial \mu_w^2} = 1.0)$$

$$\frac{\partial log(p(W^2))}{\partial \mu_w^2} = -\frac{W^{(2)}}{\sigma_p^2}$$

$\frac{\partial \log(p(W^2))}{\partial \rho_w^2}$:

$$\frac{\partial log(p(W^2))}{\partial \rho_w^2} = 0 - 0 - \frac{W^{(2)}}{\sigma_p^2} * \frac{\partial(W^2)}{\partial \rho_w^2} \quad (since\ \frac{\partial(W^2)}{\partial \rho_w^2} = \epsilon_w^2 * \frac{1.0}{(1+e^{-\rho_w^2})})$$

$$\frac{\partial log(p(W^2))}{\partial \rho_w^2} = -\frac{W^{(2)}}{\sigma_p^2} * \epsilon_w^2 * \frac{1.0}{(1+e^{-\rho_w^2})}$$

Similarly ➔

$$\frac{\partial \log(p(W^1))}{\partial \mu_w^1}, \frac{\partial \log(p(W^1))}{\partial \rho_w^1}, \frac{\partial \log(p(b^1))}{\partial \mu_b^1}, \frac{\partial \log(p(b^1))}{\partial \rho_b^1} = ?$$

$\frac{\partial \log(p(b^2))}{\partial \mu_b^2}$:

$$\frac{\partial \log(p(b^2))}{\partial \mu_b^2} = -\frac{b^{(2)}}{\sigma_p^2}$$

$\frac{\partial \log(p(b^2))}{\partial \rho_b^2}$:

$$\frac{\partial \log(p(b^2))}{\partial \rho_b^2} = -\frac{b^{(2)}}{\sigma_p^2} * \epsilon_b^2 * \frac{1.0}{(1+e^{-\rho_b^2})}$$

$\frac{\partial \log(p(W^1))}{\partial \mu_w^1}$:

$$\frac{\partial \log(p(W^1))}{\partial \mu_w^1} = -\frac{W^{(1)}}{\sigma_p^2}$$

$\frac{\partial \log(p(W^1))}{\partial \rho_w^1}$:

$$\frac{\partial \log(p(W^1))}{\partial \rho_w^1} = -\frac{W^{(1)}}{\sigma_p^2} * \epsilon_w^1 * \frac{1.0}{(1+e^{-\rho_w^1})}$$

$\frac{\partial \log(p(b^1))}{\partial \mu_b^1}$:

$$\frac{\partial \log(p(b^1))}{\partial \mu_b^1} = -\frac{b^{(1)}}{\sigma_p^2}$$

$\frac{\partial \log(p(b^1))}{\partial \rho_b^2}$:

$$\frac{\partial \log(p(b^1))}{\partial \rho_b^2} = -\frac{b^{(1)}}{\sigma_p^2} * \epsilon_b^1 * \frac{1.0}{(1+e^{-\rho_b^1})}$$

Log-Variational Posterior gradient with respect to μ, ρ

$\log(q(w|\ \theta))$:

- It is similar to Gaussian Prior, but there is trick in calculating the log_variational_posterior
- In the case of log-likelihood, we wrote $logp(D|w) \rightarrow -0.5 * \log(2\pi) - \log(\sigma) - \frac{(y-h_{out})^2}{2\sigma^2}$

  In likelihood case, the $\sigma = 0.05$ costant, since we sampled from random normal (0, 0.05).

  In the Mean-Squared-Error loss $\sigma = 1.0$ constant.
- $logq(w|\ \theta) = -0.5 * \log(2\pi) - \log(\sigma) - \frac{(w-\mu)^2}{2\sigma^2}$ here $w = \mu + log(1 + e^{\rho}) * \epsilon$