6. Variational Dropout

1) Gaussian Dropout: $B = WA, \; W \sim \mathcal{N}(\hat{W}, \alpha \odot \hat{W}^2)$

$$W \sim \hat{W} \odot \mathcal{N}(1, \alpha)$$

2) Variational Dropout:

ELBO: $\hat{\mathcal{L}}_V(\varphi) = \mathbb{E}_{q(W|\varphi)} \log p(y|x, W) - D_{KL}(q(W|\varphi) \,\|\, p(W)) \to \max_{\varphi}$

posterior: $w_{ij} = \hat{w}_{ij}(1 + \sqrt{\alpha_{ij}}\, \varepsilon_{ij}), \; \varepsilon_{ij} \sim \mathcal{N}(0,1)$

$$\Downarrow$$

$$q(w_{ij} | \varphi_{ij}) = \mathcal{N}(w_{ij} \,|\, \hat{w}_{ij}, \alpha_{ij} \hat{w}_{ij}^2)$$

prior: $p(w_{ij}) \propto \frac{1}{|w_{ij}|}$ - log-uniform

KL-divergence: $-D_{KL}(q(w_{ij} | \hat{w}_{ij}, \alpha_{ij}) \,\|\, p(w_{ij})) =$

$$= 0.5 \log \alpha_{ij} - \mathbb{E}_{\varepsilon \sim \mathcal{N}(1, \alpha_{ij})} \log|\varepsilon| + C$$

Doesn't depend on w_{ij}. $\Rightarrow$ can fix α and optimize only $\hat{w}_{ij}$

$\Downarrow$

gaussian dropout is Bayesian technique

3) Local reparametriz. to Bayesian Dropout:

$p(w_i) = \mathcal{N}(w_i | \mu_i, \sigma_i^2), \; b = w^T x = \sum_i (\mu_i + \varepsilon_i \sigma_i) \cdot x_i = \sum_i \mu_i x_i + \varepsilon \sqrt{\sum_i \sigma_i^2 x_i^2}$

$q(w_{ij} | \varphi_{ij}) = \mathcal{N}(w_{ij} | \hat{w}_{ij}, \alpha_{ij} \hat{w}_{ij}^2)$

output $= Wx$

output $=$ Relu(output)

$\longrightarrow$

Output $\sim \mathcal{N}(Wx, [W \odot W][X \odot X])$

Output $=$ Relu(Output)

4) Variance Reduction.

$$w_{ij} = \hat{w}_{ij}(1 + \sqrt{\alpha_{ij}} \cdot \varepsilon_{ij}) \Rightarrow \frac{\partial w_{ij}}{\partial \hat{w}_{ij}} = 1 + \sqrt{\alpha_{ij}}\, \varepsilon_{ij}$$

very noisy for $\alpha > 1$.

Solution: Additive Noise Parametrization:

$$w_{ij} = \hat{w}_{ij} + \sigma_{ij} \varepsilon_{ij}, \quad \sigma_{ij}^2 = \alpha_{ij} \hat{w}_{ij}^2 \qquad \frac{\partial w_{ij}}{\partial \hat{w}_{ij}} = 1 \leftarrow \text{no noise}$$

optimize ~~the~~ ELBO w.r.t. $(\hat{w}, \sigma)$

KL-div approximated analytically.

5) ~~Warm-up~~ Pruning problem: ~~Sol~~

Too many weights are pruned at the beginning

Use annealing of the KL coefficient with τ weight.

warmup solution.

One sigma per layer works just as well.

For large models - use pretrained model.

6) Bayesian Sparsification of RNNs.

$$q(w_{ij}^x) = \mathcal{N}(w_{ij}^x \mid \hat{w}_{ij}^x, \sigma_{ij}^{x\,2}), \quad q(w_{ij}^h) = \mathcal{N}(w_{ij}^h \mid \hat{w}_{ij}^h, \sigma_{ij}^{h\,2})$$

- log-uniform prior
- same sample W ~~fo~~ for all timestamps
- cannot use local reparameteriz. trick
- up to 200 x compression on ~~language modelling~~ sentiment analysis.

7) Group Sparsity: Drop neurons instead of weights.

Cannot use one α per layer.

Log-Normal Dropout: $p(\theta_i) = \text{Log}\, U_{[a,b]}(\theta_i) \quad q(\theta_i \mid \mu_i, \sigma_i) = \text{Log}\,\mathcal{N}_{[a,b]}(\theta_i \mid \mu_i, \sigma_i^2)$.

(structured BP)