Supplementary Material

1 Online Posterior Update for the Inverse Noise Variance

To update $q_{\text{cur}}(\tau)$, we consider the blending distribution only in terms of the NN output f_o and τ,

$$\tilde{p}(f_o, \tau) \propto q_{\text{cur}}(f_o) q_{\text{cur}}(\tau) \mathcal{N}(y_{\mathbf{i}_n} | f_o, \tau^{-1}) = \mathcal{N}(f_o | \alpha_n, \beta_n) \text{Gamma}(\tau | a, b) \mathcal{N}(y_{\mathbf{i}_n} | f_o, \tau^{-1}). \quad (1)$$

Following the conditional expectation propagation (CEP) framework proposed by Wang and Zhe (2019), we first derive the conditional moments of τ given f_o and then approximate the expectation of the conditional moments to obtain the moments and update the posterior of τ. Specifically, from (1), we can easily derive the conditional blending distribution,

$$\tilde{p}(\tau | f_o) = \text{Gamma}(\tau | \hat{a}, \hat{b}) \quad (2)$$

where $\hat{a} = a + \frac{1}{2}$ and $\hat{b} = b + \frac{1}{2}(y_{\mathbf{i}_n}^2 - 2f_o + f_o^2)$. We can obtain the conditional moments of τ,

$$\mathbb{E}_{\tilde{p}(\tau|f_o)}[\tau] = \frac{\hat{a}}{\hat{b}}, \qquad \mathbb{E}_{\tilde{p}(\tau|f_o)}[\log(\tau)] = \Psi(\hat{a}) - \log(\hat{b}).$$

where $\Psi(\cdot)$ is the digamma function. Note that these moments are based on the sufficient statistics of Gamma distribution, which are standard for moment matching in ADF and EP framework. The true moments can therefore be calculated by taking the expectation of the conditional moments,

$$\mathbb{E}_{\tilde{p}}[\tau] = \mathbb{E}_{\tilde{p}(f_o)} \mathbb{E}_{\tilde{p}(\tau|f_o)}[\tau] = \mathbb{E}_{\tilde{p}(f_o)}[\frac{\hat{a}}{\hat{b}}],$$
$$\mathbb{E}_{\tilde{p}}[\log(\tau)] = \mathbb{E}_{\tilde{p}(f_o)} \mathbb{E}_{\tilde{p}(\tau|f_o)}[\log(\tau)] = \mathbb{E}_{\tilde{p}(f_o)}[\Psi(\hat{a}) - \log(\hat{b})].$$

However, the normalization constant for (1) is intractable and it is difficult to compute the marginal blending distribution $\tilde{p}(f_o)$. To overcome this problem, we approximate $\tilde{p}(f_o)$ with the current posterior of f_o, namely $q_{\text{cur}}(f_o)$. This is reasonable, because $\tilde{p}(f_o)$ is an integration of $q(f_o)$ and one new data point; when we have processed many data points, adding one more data point is unlikely to significantly change the posterior. In other words, we can assume $q(f_o)$ and $\tilde{p}(f_o)$ are close in high density regions. Hence, we can approximate

$$\mathbb{E}_{\tilde{p}}[\tau] \approx \mathbb{E}_{q_{\text{cur}}(f_o)}[\frac{\hat{a}}{\hat{b}}],$$
$$\mathbb{E}_{\tilde{p}}[\log(\tau)] \approx \mathbb{E}_{q_{\text{cur}}(f_o)}[\Psi(\hat{a}) - \log(\hat{b})].$$

A second problem is that due to the nonlinearity of the conditional moments, even with $q_{\text{cur}}(f_o)$ (which has a nice Gaussian form), we still cannot analytically compute the expectation. To address this issue, we further observe that the conditional moments are functions of f_o and f_o^2,

$$h_1(f_o, f_o^2) = \frac{\hat{a}}{\hat{b}} = \frac{a + \frac{1}{2}}{b + \frac{1}{2}(y_{\mathbf{i}_n}^2 - 2f_o + f_o^2)},$$
$$h_2(f_o, f_o^2) = \Psi(\hat{a}) - \log(\hat{b}) = \Psi(a + \frac{1}{2}) - \log\big(b + \frac{1}{2}(y_{\mathbf{i}_n}^2 - 2f_o + f_o^2)\big).$$

Define $\mathbf{f} = [f_o, f_o^2]^\top$. We use a Taylor expansion at the mean of f_o and f_o^2 to approximate the conditional moments,

$$h_1(f_o, f_o^2) \approx h_1(\mathbb{E}_{q_{\text{cur}}}[f_o], \mathbb{E}_{q_{\text{cur}}}[f_o^2]) + (\mathbf{f} - \mathbb{E}_{q_{\text{cur}}}[\mathbf{f}])^\top \nabla h_1|_{\mathbf{f}=\mathbb{E}_{q_{\text{cur}}}[\mathbf{f}]},$$
$$h_2(f_o, f_o^2) \approx h_2(\mathbb{E}_{q_{\text{cur}}}[f_o], \mathbb{E}_{q_{\text{cur}}}[f_o^2]) + (\mathbf{f} - \mathbb{E}_{q_{\text{cur}}}[\mathbf{f}])^\top \nabla h_2|_{\mathbf{f}=\mathbb{E}_{q_{\text{cur}}}[\mathbf{f}]}. \quad (3)$$

We take expectation over the Taylor expansion, and obtain a closed-form result,

$$\mathbb{E}_{\tilde{p}}[\tau] = \mathbb{E}_{\tilde{p}}[h_1] \approx \frac{a^*}{b^*}, \quad \mathbb{E}_{\tilde{p}}[\log \tau] = \mathbb{E}_{\tilde{p}}[h_2] \approx \Psi(a^*) - \log(b^*) \quad (4)$$

where

$$a^* = a + \frac{1}{2}, \quad b^* = b + \frac{1}{2}((y_{\mathbf{i}_n} - \alpha_n)^2 + \beta_n).$$

Finally, from these moments, we can obtain the updated posterior, $q(\tau) = \text{Gamma}(\tau | a^*, b^*)$.

2 The Updates for Spike-and-Slab Prior Approximation

In our streaming posterior inference, after we execute ADF to process all the entries in the newly received batch, we use standard EP to update the spike-and-slab prior approximation. In this way, we can refine the approximation quality so as to effectively sparsify and condense the neural network to prevent overfitting. Specifically, for each weight w_{mjt}, we first divide the posterior by the prior approximation to obtain the calibrated (or context) distribution,

$$q^{\setminus}(w_{mjt}, s_{mjt}) \propto \frac{q_{\text{cur}}(w_{mjt}, s_{mjt})}{A(w_{mjt}, s_{mjt})} = \text{Bern}(s_{mjt}|\rho_0)\mathcal{N}(w_{mjt}|\mu^{\setminus}_{mjt}, v^{\setminus}_{mjt})$$

where $A(w_{mjt}, s_{mjt}) = \text{Bern}\big(s_{mjt}|c(\rho_{mjt})\big)\mathcal{N}(w_{mjt}|\mu^0_{mjt}, v^0_{mjt})$ (see (13) in the main paper). Because both q_{cur} and A belong to the exponential family, this can be easily done by subtracting the natural parameters. Note that $\text{Bern}(s_{mjt}|\rho_0)$ is the prior of s_{mjt} (see (4) and (5) in the main paper) — this comes from the fact that the (approximate) posterior of s_{mjt} is proportional to the product of its prior and the approximation term in A.

Next, we combine the calibrated distribution and the exact prior to obtain a tilted distribution (which is similar to the blending distribution in the streaming case),

$$\tilde{p}(w_{mjt}, s_{mjt}) \propto q^{\setminus}(w_{mjt}, s_{mjt})\big(s_{mjt}\mathcal{N}(w_{mjt}|0, \sigma_0^2) + (1 - s_{mjt})\delta(w_{mjt})\big). \tag{5}$$

We then project $\tilde{p}$ to the exponential family to obtain the updated posterior,

$$q^*(w_{mjt}, s_{mjt}) = \text{Bern}(s_{mjt}|c(\rho^*_{mjt}))\mathcal{N}(w_{mjt}|\mu^*_{mjt}, v^*_{mjt}),$$

where $c(\cdot)$ is the sigmoid function,

$$\rho^*_{mjt} = \log\big(\frac{\mathcal{N}(\mu^{\setminus}_{mjt}|0, \sigma_0^2 + v^{\setminus}_{mjt})}{\mathcal{N}(\mu^{\setminus}_{mjt}|0, v^{\setminus}_{mjt})}\big), \tag{6}$$

$$\mu^*_{mjt} = c(\widehat{\rho}_{mjt})\widehat{\mu}_{mjt}, \tag{7}$$

$$v^*_{mjt} = c(\widehat{\rho}_{mjt})\big(\widehat{v}_{mjt} + (1 - c(\widehat{\rho}_{mjt}))\widehat{\mu}^2_{mjt}\big), \tag{8}$$

$$\tag{9}$$

and

$$\widehat{\rho}_{mjt} = \rho^*_{mjt} + c^{-1}(\rho_0),$$
$$\widehat{v}_{mjt} = \big(\big(v^{\setminus}_{mjt}\big)^{-1} + \sigma_0^{-2}\big)^{-1},$$
$$\widehat{\mu}_{mjt} = \widehat{v}_{mjt}\frac{\mu^{\setminus}_{mjt}}{v^{\setminus}_{mjt}}.$$

Finally, we can update the prior approximation term via dividing the updated posterior by the calibrated distribution, $A^*(w_{mjt}, s_{mjt}) \propto q^*(w_{mjt}, s_{mjt})/q^{\setminus}(w_{mjt}, s_{mjt})$. Now, we replace the current prior approximation by A^* and set $q_{\text{cur}} = q^*$, to prepare the steaming inference for the next batch. Therefore, the learned posterior weights and selection probabilities are consistent, and they effectively deactivate many weights to adjust the complexity of the network.

References

Wang, Z. and Zhe, S. (2019). Conditional expectation propagation. In *UAI*, page 6.