AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

A APPENDIX

A.1 KALMAN FILTER

For neural networks, the model of interest

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} = \boldsymbol{w}, \\ y_t = h(\boldsymbol{\theta}_t, x_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, R_t), \end{cases} \tag{1}$$

is formulated in stochastic language as an EKF problem targeting on $\hat{\boldsymbol{\theta}}_t$, where $\boldsymbol{w}$ is the vector of all trainable parameters in the network $h(\cdot,\cdot)$, $\{(x_t, y_t)\}_{t\in\mathbb{N}}$ are pairs of feature and label, $\{y_t\}_{t\in\mathbb{N}}$ can also be seen as measurements of EKF, $\{\eta_t\}_{t\in\mathbb{N}}$ are noise terms subject to normal distribution with mean 0 and variances $\{R_t\}_{t\in\mathbb{N}}$ correspondingly, and $\forall t \in \mathbb{N}, \hat{\boldsymbol{\theta}}_{t|t-1} := \mathbb{E}[\boldsymbol{\theta}_t | y_{t-1}, y_{k-2}, \ldots, y_1]$. With fixed $\boldsymbol{\theta}_t$ and bounded x_t, $h(\boldsymbol{\theta}_t, x_t)$ will be approximated well by its linearization at $\hat{\boldsymbol{\theta}}_{t|t-1}$, just omitting a term $\mathcal{O}((\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t|t-1})^2)$.

$$y_t \approx h(\hat{\boldsymbol{\theta}}_{t|t-1}, x_t) + \mathbf{H}_t^\mathsf{T}(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_{t|t-1}) + \eta_t \tag{2}$$

$$\mathbf{H}_t = \left.\frac{\partial h(\boldsymbol{\theta}, x_t)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{t|t-1}}$$

If set $m_t = y_t - h(\hat{\boldsymbol{\theta}}_{t|t-1}, x_t) + \mathbf{H}_t^\mathsf{T}\hat{\boldsymbol{\theta}}_{t|t-1}$ and rewrite equation 2 the following KF problem

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} = \boldsymbol{w}, & (3) \\ m_t \approx \mathbf{H}_t^\mathsf{T}\boldsymbol{\theta}_t + \eta_t. & (4) \end{cases}$$

At the beginning of training, the estimator $\hat{\boldsymbol{\theta}}_{t|t-1}$ is far away from $\boldsymbol{w}$, so less attention should be paid to those data fed to the network at an earlier stage of training than those at later stage. Through timing a factor $\alpha_t := \Pi_{i=1}^t \lambda_i^{-1/2}$ and $\alpha_0 := 1$, where $0 < \lambda_i \le 1$ and $\lambda_i \to 1$, the last problem enjoys the better variant as below

$$\begin{cases} \boldsymbol{\theta}_t = \lambda_t^{-1/2}\boldsymbol{\theta}_{t-1}, \ \boldsymbol{\theta}_1 = \boldsymbol{w} \\ \tilde{m}_t = \alpha_t m_t \approx \mathbf{H}_t^\mathsf{T}\boldsymbol{\theta}_t + \alpha_t\eta_t = \mathbf{H}_t^\mathsf{T}\boldsymbol{\theta}_t + \tilde{\eta}_t, \end{cases} \tag{5}$$

where λ_t is called memory factor. The greater λ_t is, the more weight, or say attention, is paid to previous data. According to basic KF theory Haykın & Haykın (2001) , we obtain

$$\begin{aligned} \mathbf{a}_t &= \lambda_t^{-1}\mathbf{H}_t^\mathsf{T}\mathbf{P}_{t-1}\mathbf{H}_t + \alpha_t^2 R_t = \mathbb{E}[\tilde{\epsilon}_t^2], \\ \mathbf{K}_t &= \lambda_t^{-1}\mathbf{P}_{t-1}\mathbf{H}_t^\mathsf{T}\mathbf{a}_t^{-1}, \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\,\lambda_t^{-1}\mathbf{P}_{t-1}, \\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\tilde{\epsilon}_t, \\ \tilde{\epsilon}_t &= \tilde{m}_t - \mathbf{H}_t^\mathsf{T}\hat{\boldsymbol{\theta}}_{t|t-1} = \alpha_t(y_t - h(\alpha_t^{-1}\hat{\boldsymbol{\theta}}_{t|t-1}, x_t)). \end{aligned}$$

Finally, we recover the estimator of w via that of θ divided by the factor α_t, define $\forall t \in \mathbb{N}, \hat{\boldsymbol{w}}_{t|t-1} := \alpha_t^{-1}\hat{\boldsymbol{\theta}}_{t|t-1}, \boldsymbol{w}_t := \alpha_t^{-1}\hat{\boldsymbol{\theta}}_t$, find $\hat{\boldsymbol{w}}_{t|t-1} = \boldsymbol{w}_{t-1}$, and then get our weights updating strategy

$$\begin{aligned} \epsilon_t &= y_t - h(\boldsymbol{w}_{t-1}, x_t), \\ \boldsymbol{w}_t &= \boldsymbol{w}_{t-1} + \mathbf{K}_t\epsilon_t. \end{aligned}$$