Using Node Biases with Conservative Learning

John Mastroberti

April 25, 2019

As a modification to the conservative learning algorithm outlined here [1], we want to add a bias term to the nodes:

$$y_j^k = \sum_{i \to j} x_i^k w_{i \to j} - b_j \tag{1}$$

We can accomplish this by introducing a special input node which is always set to 1. This node is then connected to all of the hidden and output nodes, and if we give this special node the index s, we can define

$$w_{s \to j} = -b_j \tag{2}$$

so that we can retain the formula

$$y_j^k = \sum_{i \to j} x_i^k w_{i \to j} \tag{3}$$

Using this formalism leaves the update algorithm largely unchanged. The important differences are as follows.

While equation (1.15),

$$\forall (i \in H, k \in K) : \ \gamma_i^k = df_i^k \sum_{i \to j} w_{i \to j} \gamma_j^k \tag{4}$$

remains unchanged, since $s \notin H$, equation (1.16),

$$\begin{aligned} \forall (i \to j \in E) : \ \ \Delta w_{i \to j} + \delta w_{i \to j} &= \sum_{k \in K} \tilde{x}_i^k \gamma_j^k & (5) \\ &= \tilde{x}_i \cdot \gamma_j & (6) \end{aligned}$$

now encompasses the special case

$$\forall (j \in H \cup O) : \ \ -\Delta b_j + -\delta b_j = \tilde{x}_s \cdot \gamma_j \tag{7}$$

which means that equation (1.18) now reads

$$\forall (j \in H_-, k \in K) : \ \ \delta y_j^k = \sum_{i \to j} \tilde{x}_i^k (\tilde{x}_i \cdot \gamma_j - \Delta w_{i \to j}) + \tilde{x}_s^k (\tilde{x}_s \cdot \gamma_j + \Delta b_j) \tag{8}$$

$$= \sum_{i\to j} \tilde{x}_i^k(\tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j}) + \sum_{\kappa\in K} \gamma_j^\kappa + \Delta b_j \tag{9}$$

since $\tilde{x}_s^k = 1$. Furthermore, the forward propagation equation (1.19) gets the slight modification

$$\forall(j \notin H_-, k \in K): \ \ \delta y_j^k = \sum_{i\to j}(df_i^k \delta y_i^k w_{i\to j} + \tilde{x}_i^k(\tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j})) + \tilde{x}_s^k(\tilde{x}_s \cdot \gamma_j + \Delta b_j) \tag{10}$$

$$= \sum_{i\to j}(df_i^k \delta y_i^k w_{i\to j} + \tilde{x}_i^k(\tilde{x}_i \cdot \gamma_j - \Delta w_{i\to j})) + \sum_{\kappa\in K} \gamma_j^\kappa + \Delta b_j \tag{11}$$

since there are no δy_j^k's or df_i^k's for $i \in I$, including our special input node s.

Finally, the initialization formulas need to be modified slightly; most notably, an extra row is added to $\tilde{X}$ for our special node s, and δW and ΔW get an extra row for the biases:

$$\forall(j \in O, k \in K): \ \ \epsilon_j^k = \sum_{i\to j}(df_i^k \delta y_i^k w_{i\to j} + \tilde{x}_i^k \delta w_{i\to j}) - \delta b_j \tag{12}$$

$$\mathcal{E}(n)_{kj} = \epsilon_j^k - \sum_{i\to j} df_i^k w_{i\to j} \delta y(n)_i^k \tag{13}$$

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1^1 & \tilde{x}_1^2 & \cdots & \tilde{x}_1^k & \cdots \\ \tilde{x}_2^1 & \tilde{x}_2^2 & \cdots & \tilde{x}_2^k & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ 1 & 1 & \cdots & 1 & \cdots \end{pmatrix} \tag{14}$$

$$\delta W(n) = \begin{pmatrix} \delta w_{1\to 1} & \delta w_{1\to 2} & \cdots & \delta w_{1\to j} & \cdots \\ \delta w_{2\to 1} & \delta w_{2\to 2} & \cdots & \delta w_{2\to j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ -\delta b_1 & -\delta b_2 & \cdots & -\delta b_j & \cdots \end{pmatrix} \tag{15}$$

$$\Delta W(n) = \begin{pmatrix} \Delta w_{1\to 1} & \Delta w_{1\to 2} & \cdots & \Delta w_{1\to j} & \cdots \\ \Delta w_{2\to 1} & \Delta w_{2\to 2} & \cdots & \Delta w_{2\to j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ -\Delta b_1 & -\Delta b_2 & \cdots & -\Delta b_j & \cdots \end{pmatrix} \tag{16}$$

With these slight modifications, and maintaining $\Gamma(n)_{kj} = \gamma(n)_j^k$ from equation (1.29), the form of equation (1.30) remains unchanged:

$$\Gamma(n) = (\tilde{X}^\top \tilde{X})^{-1}(\tilde{X}^\top \Delta W + \mathcal{E}(n)) \tag{17}$$

Thus, the steps for an inner loop now read

1. Initialize γ's at the output nodes:

$$\Gamma(n) = (\tilde{X}^\top \tilde{X})^{-1}(\tilde{X}^\top \Delta W + \mathcal{E}(n)) \tag{17}$$

2. Back-propagate γ's down to all the hidden nodes:

$$\forall(i \in H, k \in K): \ \gamma_i^k = df_i^k \sum_{i\to j} w_{i\to j} \gamma_j^k \tag{4}$$

3. Initialize δy's at nodes adjacent to the input nodes:

$$\forall(j \in H_-, k \in K): \;\; \delta y_j^k = \sum_{i \to j} \tilde{x}_i^k (\tilde{x}_i \cdot \gamma_j - \Delta w_{i \to j}) + \sum_{\kappa \in K} \gamma_j^\kappa + \Delta b_j \tag{9}$$

4. Forward propagate δy's up to the nodes adjacent to the output nodes:

$$\forall(j \notin H_-, k \in K): \;\; \delta y_j^k = \sum_{i \to j} (df_i^k \delta y_i^k w_{i \to j} + \tilde{x}_i^k (\tilde{x}_i \cdot \gamma_j - \Delta w_{i \to j})) + \sum_{\kappa \in K} \gamma_j^\kappa + \Delta b_j \tag{11}$$

We still obtain the weight changes from (1.33),

$$\forall(i \to j \in E): \;\; \delta w_{i \to j} = \tilde{x}_i \cdot \gamma_j - \Delta w_{i \to j}$$

and we obtain the bias changes from

$$\forall(j \in H \cup O): \delta b_j = -\sum_{\kappa \in K} \gamma_j^\kappa + \Delta b_j$$

References

[1] Veit Elser. Conservative updates by backward-forward cycles. 2019.