A. Appendix

In Sec. 3.2.1, we describe how the WiMA update is equivalent to updating the first model comprised in the window frame \( w^\tau \) with various SGD steps, using a learning rate decay dependent on the position in the queue, given by \( \frac{t + \tau - \tau}{W} \) (Eq. 6,7). We describe here the steps to reach this conclusion.

We recall that

\[
\frac{w^{t+W-1}}{W_{\text{FedAvg}}} = \frac{1}{W} \sum_{\tau=t'}^{t+W-1} w^{\tau}_{\text{FedAvg}} \tag{Eq. 4}
\]

\[
= \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \sum_{S' \in S^\tau} \frac{N_i}{N} w^\tau_i \tag{FedAvg in Eq. 3}
\]

\[
= \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \left( w^\tau - \eta_i \sum_{l \in S'} \frac{N_l}{N} (w^\tau - w^\tau_l) \right) \tag{FedOpt in Eq. 3}
\]

where \( w^{t+1}_{\text{FedAvg}} \) is the new global model built with FedAvg at the end of round \( \tau \), \( W \) the window size, \( t' \) the first round comprised in window frame, \( w_i \) the local update of client \( i \), \( S' \) the subset of clients selected at round \( t \), \( \eta_i \), the server learning rate.

For simplicity, we first assume all clients have access to the same number of images, i.e. \( \frac{N_i}{N} = \frac{1}{|S'|} \). Since the same number of clients is selected at each round, \( \frac{1}{|S'|} = \frac{1}{|S|^{\tau-\tau}} \).

First, we recursively rewrite \( w^\tau \) following Eq. 3 as

\[
\frac{w^{t+W}}{W_{\text{WiMA}}} = \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \left( w^\tau - \frac{1}{|S'|} \sum_{S' \in S^\tau} (w^\tau - w^\tau_0) \right) \tag{8}
\]

\[
= \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \left( w^\tau - \frac{1}{|S'|} \sum_{S' \in S^\tau} (w^{\tau-1} - \frac{1}{|S'|} \sum_{j \in S'^{\tau-1}} (w^{\tau-1} - w^{\tau-1}_j) - w^\tau_j) \right) \tag{9}
\]

\[
= \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \left( w^\tau - \frac{1}{|S'|} \sum_{S' \in S^\tau} (w^{\tau-2} - \frac{1}{|S'|} \sum_{j \in S'^{\tau-1}} (w^{\tau-2} - w^{\tau-2}_j) - w^{\tau-2}_j) \right) \tag{10}
\]

\[
= \ldots \frac{1}{W} \sum_{\tau=t'}^{t+W-1} \left( w^\tau - \frac{1}{|S'|} \sum_{S' \in S^\tau} (w^0 - \frac{1}{|S'|} \sum_{m \in S'^0} (w^0 - w^0_m) \right) \tag{11}
\]

As in standard SGD, each model implicitly contains information on the previous updates. By unraveling the summation over
\[ w^{t+W}_{\text{WMA}} = w^0 - \frac{1}{|S^t|} \left( \sum_{i \in S^t} (w^{0} - w^{0}_{i}) + \sum_{i \in S^{t'}} (w^{t'} - w^{t'}_{i}) + \right. \]
\[ \left. + \frac{W-1}{W} \sum_{i \in S^{t'}+1} (w^{t+1} - w^{t+1}_{i}) + \ldots + \frac{1}{W} \sum_{i \in S^{t+W-1}} (w^{t+W-1} - w^{t+W-1}_{i}) \right) = \]
\[ w^{t'} - \frac{1}{|S^t|} \left( \frac{W-1}{W} \sum_{i \in S^{t'}+1} (w^{t'+1} - w^{t'+1}_{i}) + \ldots + \frac{1}{W} \sum_{i \in S^{t+W-1}} (w^{t'+W-1} - w^{t'+W-1}_{i}) \right) = \]
\[ = w^{t'} - \frac{1}{|S^t|} \sum_{i=1}^{t'} \frac{t'+W-1}{W} \sum_{i \in S^{t'}} (w^{t'} - w^{t'}_{i}). \] (17)

If we drop the constraint \( \frac{N_i}{N} = \frac{1}{|S^t|} \) and insert the server learning rate \( \eta_s \), we can summarize the results as
\[ w^{t+W}_{\text{WMA}} = w^{t'} - \eta_s \sum_{i=1}^{t'} \frac{t'+W-1}{W} \sum_{i \in S^{t'}} \frac{N_i}{N} (w^{t'} - w^{t'}_{i}), \] (18)

obtaining Eq. 6.