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 $w^{t'}$ with various SGD steps, using a learning rate decay dependent on the position in the queue, given by ${t'+w-\tau}/{W}$ (Eq. 6,7). We describe here the steps to reach this conclusion.

We recall that

$$\begin{aligned} w_{\text{WIMA}}^{t'+W} &= \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} w_{\text{FEDAVG}}^{\tau+1} && \text{(Eq. 4)} \\ &= \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \sum_{i\in\mathcal{S}^{\tau}} \frac{N_i}{N} w_i^{\tau} && \text{(FedAvg in Eq. 3)} \\ &= \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \Big(w^{\tau} - \eta_s \sum_{i\in\mathcal{S}^{\tau}} \frac{N_i}{N}(w^{\tau} - w_i^{\tau})\Big), && \text{(FedOpt in Eq. 3)} \end{aligned}$$

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

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

First, we recursively rewrite w^{τ} following Eq. 3 as

$$w_{\text{WIMA}}^{t'+W} = \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \Big(w^{\tau} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{i\in\mathcal{S}^{\tau}}(w^{\tau} - w_i^{\tau})\Big) \tag{8}$$

$$= \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \Big(w^{\tau} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{i\in\mathcal{S}^{\tau}} \underbrace{\big(w^{\tau-1} - \frac{1}{|\mathcal{S}^{\tau-1}|}\sum_{j\in\mathcal{S}^{\tau-1}}(w^{\tau-1} - w_j^{\tau-1})}_{w^{\tau}} - w_i^{\tau})\Big) \tag{9}$$

$$\overset{|\mathcal{S}^{\tau-1}|=|\mathcal{S}^{\tau}|}{=} \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \Big(w^{\tau} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{i\in\mathcal{S}^{\tau}}(w^{\tau-1} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{j\in\mathcal{S}^{\tau-1}} \underbrace{(w^{\tau-2} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{l\in\mathcal{S}^{\tau-2}}(w^{\tau-2} - w_l^{\tau-2}) +}_{w^{\tau-1}} \tag{10}$$

$$- w_j^{\tau-1}) - w_i^{\tau})\Big) \tag{11}$$

$$= \ldots = \frac{1}{W}\sum_{\tau=t'}^{t'+W-1} \Big(w^{\tau} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{i\in\mathcal{S}^{\tau}}(w^{\tau-1} - \ldots - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{m\in\mathcal{S}^{1}}(w^{0} - \frac{1}{|\mathcal{S}^{\tau}|}\sum_{l\in\mathcal{S}^{0}}(w^{0} - w_l^{0}) + \tag{12}$$

$$- w_m^{1}) - \ldots - w_j^{\tau-1}) - w_i^{\tau})\Big) \tag{13}$$

As in standard SGD, each model implicitly contains information on the previous updates. By unraveling the summation over

τ, we get

$$w_{\text{WIMA}}^{t'+W} = w^0 - \frac{1}{|\mathcal{S}^0|}\Big(\underbrace{\sum_{i\in\mathcal{S}^0}(w^0 - w_i^0) + ... + \sum_{i\in\mathcal{S}^{t'}}(w^{t'} - w_i^{t'}) +}_{\tau\leq t'} \tag{14}$$

$$\underbrace{+ \underbrace{\frac{W-1}{W}}_{t'+W-(t'+1)=W-1} \sum_{i\in\mathcal{S}^{t'+1}}(w^{t'+1} - w_i^{t'+1}) + ... + \frac{1}{W}\sum_{i\in\mathcal{S}^{t'+W-1}}(w^{t'+W-1} - w_i^{t'+W-1})\Big)}_{t'<\tau<t'+W} = \tag{15}$$

$$= w^{t'} - \frac{1}{|\mathcal{S}^0|}\Big(\frac{W-1}{W}\sum_{i\in\mathcal{S}^{t'+1}}(w^{t'+1} - w_i^{t'+1}) + ... + \frac{1}{W}\sum_{i\in\mathcal{S}^{t'+W-1}}(w^{t'+W-1} - w_i^{t'+W-1})\Big) = \tag{16}$$

$$= w^{t'} - \frac{1}{|\mathcal{S}^0|}\sum_{\tau=t'+1}^{t'+W-1}\frac{t'+W-\tau}{W}\sum_{i\in\mathcal{S}^\tau}(w^\tau - w_i^\tau). \tag{17}$$

If we drop the constraint $\frac{N_i}{N} = \frac{1}{|\mathcal{S}^t|}$ and insert the server learning rate η_s, we can summarize the results as

$$w_{\text{WIMA}}^{t'+W} = w^{t'} - \eta_s \sum_{\tau=t'}^{t'+W-1}\frac{t'+W-\tau}{W}\sum_{i\in\mathcal{S}^\tau}\frac{N_i}{N}(w^\tau - w_i^\tau), \tag{18}$$

obtaining Eq. 6.