The derivations here are very close to those for the simple EMA and EMV without batches. These can be found in [3].

Definitions

Data $\{x_{k,j}\}$ first index is for batch second index is for element in batch. The first index starts from zero the second starts from 1. I'm sorry it just ended up working out nicer that way. Batches contain b elements. Each batch will be given a certain weight in the average $w_{m,k}$ where m denotes the total number of batches minus 1 and k denotes the batches index, starting from zero.

Property 1

$$\frac{\sum_{k=0}^{m-1} w_{m,k} \sum_{j=1}^{b} f(x_{k,j})}{\sum_{k=0}^{m-1} w_{m,k}} = \frac{\sum_{k=0}^{m-1} w_{m-1,k} \sum_{j=1}^{b} f(x_{k,j})}{\sum_{k=0}^{m-1} w_{m-1,k}}$$

Property 2

$$\sum_{k=0}^{m} w_{m,k} = 1$$

Definition 1: Expectation

$$\mathbb{E}_{bm}\left[f(x)\right] := \sum_{k=0}^{m} w_{m,k} \left[\frac{\sum_{j=1}^{b} f(x_{k,j})}{b}\right]$$

Definition 2: Mean

$$\mu_{bm} := \mathbb{E}_{bm}\left[x\right]$$

Definition 3: Variance

$$\sigma_{bm}^2 := \mathbb{E}_{bm}[(x - \mu_{bm})^2]$$

Algorithms

Lemma 2: Incremental Expectation

$$\mathbb{E}_{bm}[f(x)] = \mathbb{E}_{b(m-1)}(f(x)) + w_{m,m}\left(\frac{\sum_{j=1}^{b} f(x_{m,j})}{b} - \mathbb{E}_{b(m-1)}(f(x))\right)$$

Proof:

$$
\begin{aligned}
\mathbb{E}_{bm}[f(x)] &= \sum_{k=0}^{m} w_{m,k} \frac{\sum_{j=1}^{b} f(x_{k,j})}{b} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + \sum_{k=0}^{m-1} w_{m,k} \frac{\sum_{j=1}^{b} f(x_{k,j})}{b} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \frac{\sum_{k=0}^{m-1} w_{m,k} \sum_{j=1}^{b} f(x_{k,j})}{b(1 - w_{m,m})} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \frac{\sum_{k=0}^{m-1} w_{m,k} \sum_{j=1}^{b} f(x_{k,j})}{b(1 - w_{m,m})} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \frac{\sum_{k=0}^{m-1} w_{m,k} \sum_{j=1}^{b} f(x_{k,j})}{b \left[\sum_{k=0}^{m} w_{m,k} - w_{m,m}\right]} \text{ by property 2} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \frac{\sum_{k=0}^{m-1} w_{m-1,k} \sum_{j=1}^{b} f(x_{k,j})}{b \sum_{k=0}^{m-1} w_{m-1,k}} \text{ by property 1} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \frac{\sum_{k=0}^{m-1} w_{m-1,k} \sum_{j=1}^{b} f(x_{k,j})}{b} \text{ by property 2} \\
&= w_{m,m} \frac{\sum_{j=1}^{b} f(x_{m,j})}{b} + (1 - w_{m,m}) \mathbb{E}_{b(m-1)}\left[f(x)\right] \\
&= \mathbb{E}_{b(m-1)}(f(x)) + w_{m,m} \left(\frac{\sum_{j=1}^{b} f(x_{m,j})}{b} - \mathbb{E}_{b(m-1)}(f(x)) \right)
\end{aligned}
$$

Now observe that the classic formula as a difference the first and second moments still holds in our setting.

$$
\begin{aligned}
\sigma_{bm}^2 &= \mathbb{E}_{bm}[(x - \mu_{bm})^2] \\
&= \mathbb{E}_{bm}[x^2] - 2\mathbb{E}_{bm}[x\mu_{bm}] + \mathbb{E}_{bm}[\mu_{bm}^2] \\
&= \mathbb{E}_{bm}[x^2] - 2\mathbb{E}_{bm}[x]\mu_{bm} + \mu_{bm}^2 \text{ since expectation is linear} \\
&= \mathbb{E}_{bm}[x^2] - 2\mu_{bm}^2 + \mu_{bm}^2 \\
&= \mathbb{E}_{bm}[x^2] - \mu_{bm}^2
\end{aligned}
$$

This suggests the following naive one pass algorithm which is correct by lemma 2,

Algorithm: ***Algorithm 1***

$\mu_{b0} \leftarrow \sum_{j=1}^{b} \frac{x_{0,j}}{b}$

$\mathbb{E}_{b0}[x^2] \leftarrow \sum_{j=1}^{b} \frac{x_{0,j}^2}{b}$

$\sigma_{b0}^2 \leftarrow \mathbb{E}_{b0}[x^2] - \mu_{b0}^2$

for *i from* 1 *to* m **do**

  $\mu_{bi} \leftarrow \mu_{b(i-1)} + w_{i,i} \left(\frac{\sum_{j=1}^{b} x_{m,j}}{b} - \mu_{b(i-1)} \right)$

  $\mathbb{E}_{bi}[x^2] \leftarrow \mathbb{E}_{b(m-1)}(x^2) + w_{i,i} \left(\frac{\sum_{j=1}^{b} x_{i,j}^2}{b} - \mathbb{E}_{b(k-1)}(x^2) \right)$

  $\sigma_{bi}^2 \leftarrow \mathbb{E}_{bi}[x^2] - \mu_{bi}^2$

end

Sadly I don't think this approach will work due to catastrophic cancellation [2, 1].
If we continue trying to follow the derivations from [3] section 8 in the batch setting, we can arrive at the following result.
This gives us the following algorithm,

Lemma 3

$$\sigma_{bm}^2 = (1 - w_{m,m})\sigma_{b(m-1)}^2 + w_{m,m}\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - (\mu_{bm} - \mu_{b(m-1)})^2$$

Proof:

$$\begin{aligned}
\sigma_{bm}^2 &= \mathbb{E}_{bm}[(x - \mu_{bm})^2] \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)} + \mu_{b(m-1)} - \mu_{bm})^2] \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)})^2 - 2(x - \mu_{b(m-1)})(\mu_{bm} - \mu_{b(m-1)}) + (\mu_{bm} - \mu_{b(m-1)})^2] \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)})^2] + -2\mathbb{E}_{bm}[(x - \mu_{b(m-1)})(\mu_{bm} - \mu_{b(m-1)})] + \mathbb{E}_{bm}[(\mu_{bm} - \mu_{b(m-1)})^2] \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)})^2] + -2(\mu_{bm} - \mu_{b(m-1)})\mathbb{E}_{bm}[(x - \mu_{b(m-1)})] + (\mu_{bm} - \mu_{b(m-1)})^2 \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)})^2] + -2(\mu_{bm} - \mu_{b(m-1)})^2 + (\mu_{bm} - \mu_{b(m-1)})^2 \\
&= \mathbb{E}_{bm}[(x - \mu_{b(m-1)})^2] - (\mu_{bm} - \mu_{b(m-1)})^2 \\
&= (1 - w_{m,m})\mathbb{E}_{b(m-1)}[(x - \mu_{b(m-1)})^2] + w_{m,m}\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - (\mu_{bm} - \mu_{b(m-1)})^2 \text{ by lemma 2} \\
&= (1 - w_{m,m})\sigma_{b(m-1)}^2 + w_{m,m}\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - (\mu_{bm} - \mu_{b(m-1)})^2 \text{ by lemma 2}
\end{aligned}$$

The correctness of algorithm 2 follows directly from lemmas 2 and 3.

Algorithm: ***Algorithm 2***

$\mu_{b0} \leftarrow \sum_{j=1}^{b}\frac{x_{0,j}}{b}$
$\sigma_{b0}^2 \leftarrow \mathbb{E}_{b0}[x^2] - \mu_{b0}^2$
for *i from* 1 *to* m **do**
  $\Delta \leftarrow w_{m,m}\left(\frac{\sum_{j=1}^{b} x_{m,j}}{b} - \mu_{b(i-1)}\right)$
  $S \leftarrow \frac{\sum_{j=1}^{b}\left[x_{i,j} - \mu_{b(i-1)}\right]^2}{b}$
  $\mu_{bi} \leftarrow \mu_{b(i-1)} + \Delta$
  $\sigma_{bi}^2 \leftarrow (1 - w_{m,m})\sigma_{b(i-1)}^2 + w_{m,m}S - \Delta^2$
end

Lemma 4

Let $\hat{\mu}_k := \frac{\sum_{j=1}^{b} x_{k,j}}{b}$
Let $\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{b}[x_{k,j} - \hat{\mu}_k]^2}{b}$ i.e. the sample variance of batch k using the MLE estimator. Lets call this the intarbatch variance.
Let $\delta_{bm}^2 = \left(\mu_{b(m-1)} - \hat{\mu}_m\right)^2$. I'm going to call this term the interbatch variance.

$$\sigma_{bm}^2 = \sigma_{b(m-1)}^2 + w_{m,m}\left[\hat{\sigma}_m^2 + (1 - w_{m,m})\delta_{bm}^2 - \sigma_{b(m-1)}^2\right]$$

Proof:

$$\sigma^2_{bm} = \gamma\sigma^2_{b(m-1)} + w_{m,m}\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - (\mu_{bm} - \mu_{b(m-1)})^2 \text{ by lemma 3}$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - w^2_{m,m}\delta^2_{bm} \text{ by lemma 2}$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - w_{m,m}\delta^2_{bm}\right]$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]^2}{b} - \left(\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)}\right]}{b}\right)^2 + (1 - w_{m,m})\delta^2_{bm}\right]$$

Recall that $\frac{\sum_i y_i^2}{b} - \left(\frac{\sum_i y_i}{b}\right)^2 = \frac{\sum_i \left(y_i - \frac{\sum_i y_i}{b}\right)^2}{b}$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\frac{\sum_{j=1}^{b}\left[x_{m,j} - \mu_{b(m-1)} - \frac{\sum_{j=1}^{b} x_{m,j} - \mu_{b(m-1)}}{b}\right]^2}{b} + (1 - w_{m,m})\delta^2_{bm}\right]$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\frac{\sum_{j=1}^{b}\left[x_{m,j} - \frac{\sum_{j=1}^{b} x_{m,j}}{b}\right]^2}{b} + (1 - w_{m,m})\delta^2_{bm}\right]$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\frac{\sum_{j=1}^{b}\left[x_{m,j} - \hat{\mu}_k\right]^2}{b} + (1 - w_{m,m})\delta^2_{bm}\right]$$

$$= (1 - w_{m,m})\sigma^2_{b(m-1)} + w_{m,m}\left[\hat{\sigma}^2_m + (1 - w_{m,m})\delta^2_{bm}\right]$$

$$= \sigma^2_{b(m-1)} + w_{m,m}\left[\hat{\sigma}^2_m + (1 - w_{m,m})\delta^2_{bm} - \sigma^2_{b(m-1)}\right]$$

Again the correctness of algorithm 3 follows directly from lemma 4 and 2.

Algorithm: ***Algorithm 3***

$\mu_{b0} \leftarrow \sum_{j=1}^{b} \frac{x_{0,j}}{b}$
$\sigma^2_{b0} \leftarrow \mathbb{E}_{b0}[x^2] - \mu^2_{b0}$
for k *from* 1 *to* m **do**
  $\delta_{bk} \leftarrow \left(\hat{\mu}_k - \mu_{b(k-1)}\right)$
  $\mu_{bk} \leftarrow \mu_{b(k-1)} + w_{m,m}\delta_{bk}$
  $\sigma^2_{bk} \leftarrow \sigma^2_{b(k-1)} + w_{m,m}\left[\hat{\sigma}^2_k + (1 - w_{m,m})\delta^2_{bk} - \sigma^2_{b(k-1)}\right]$
end

Weightings

Let $0 < \alpha < 1$ and let $\gamma = 1 - \alpha$

Standard EMA Weighting

$$w_{m,k} := \begin{cases} \gamma^{m-k}\alpha, 1 \leq k \leq m \\ \gamma^m, k = 0 \end{cases}$$

The advantages of this weighting is that the learning rate $w_{m,m} = \alpha$ which does not depend in any way on m. However, this weighting disproportionately emphasizes the first batch especially when α is close to zero.

Adjusting EMA Weighting

$w_{m,k} := \frac{\gamma^{m-k}}{\sum_{k=0}^{m} \gamma^k}$
To address the shortcomings of the first weighting scheme, we can use the above weighting scheme. This fixes the problem of having a large weight on the initial batch. However the trade off is that the learning rate $w_{m,m} = \frac{1}{\sum_{k=0}^{m} \gamma^k} = \frac{\alpha}{1-\gamma^{m+1}}$ is no longer constant. Although the learning rate is no longer constant it is still the case that $w_{m,m} \to \alpha$ as $m \to \infty$.

References

[1] Catastrophic cancellation. *Wikipedia*, May 2022.

[2] Tony F Chan, Gene H Golub, and Randall J LeVeque. Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247, 1983.

[3] Tony Finch. Incremental calculation of weighted mean and variance. [Online; accessed 9-August-2022] `https://fanf2.user.srcf.net/hermes/doc/antiforgery/stats.pdf`.