A. Appendix

A.1. Related Work

Continual learning. Many penalty-based approaches have been proposed to overcome catastrophic forgetting. [15] protects the source task performance by a quadratic penalty loss where the importance of each weight is measured by the diagonal of Fisher. [20] proposes a network reparameterization technique that approximately diagonalizes the Fisher Information Matrix of the network parameters. In [31], the block diagonal K-FAC is used for a quadratic penalty loss to take interaction between parameters in each single layer into account. [1] proposes to measure the importance of a parameter by the magnitude of the gradient. [35] also defines a quadratic penalty loss designed with the change in loss over an entire trajectory of parameters. [27] approximates a true loss function using an asymmetric quadratic function with one of its sides overestimated.

A.2. The Hessian of a linear layer

As $(h_l)_{i,m} = \sum_k (W_l)_{i,k}(\bar{a}_{l-1})_{k,m}$,

$$\frac{\partial \mathcal{L}_n}{\partial (W_l)_{a,b}} = \sum_{m,i} \frac{\partial \mathcal{L}_n}{\partial (h_l)_{i,m}} \frac{\partial (h_l)_{i,m}}{\partial (W_l)_{a,b}} = \sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,m}} (\bar{a}_{l-1})_{b,m}. \tag{36}$$

Then,

$$\frac{\partial^2 \mathcal{L}_n}{\partial (W_{l'})_{c,d} \partial (W_l)_{a,b}} = \sum_m \left(\frac{\partial}{\partial (W_{l'})_{c,d}} \left(\frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,m}} \right) (\bar{a}_{l-1})_{b,m} + \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,m}} \frac{\partial (\bar{a}_{l-1})_{b,m}}{\partial (W_{l'})_{c,d}} \right). \tag{37}$$

Using the chain rule,

$$\begin{aligned} \frac{\partial}{\partial (W_{l'})_{c,d}} \left(\frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,m}} \right) &= \sum_{m',i} \frac{\partial^2 \mathcal{L}_n}{\partial (h_{l'})_{i,m'} \partial (h_l)_{a,m}} \frac{\partial (h_{l'})_{i,m'}}{\partial (W_{l'})_{c,d}} \\ &= \sum_{m'} \frac{\partial^2 \mathcal{L}_n}{\partial (h_{l'})_{c,m'} \partial (h_l)_{a,m}} (\bar{a}_{l'-1})_{d,m'}. \end{aligned} \tag{38}$$

Here, as in [3, 30], we can assume $l \leq l'$ by the symmetry of Hessian, so

$$\frac{\partial (\bar{a}_{l-1})_{b,m}}{\partial (W_{l'})_{c,d}} = 0, \tag{39}$$

since $\bar{a}_{l-1}$ is a function of $W_1, W_2, \cdots, W_{l-1}$, but does not depend on $W_l, W_{l+1}, \cdots, W_L$. Therefore,

$$\frac{\partial^2 \mathcal{L}_n}{\partial (W_l)_{a,b} \partial (W_{l'})_{c,d}} = \sum_{m,m'} (\bar{a}_{l-1})_{b,m} (\bar{a}_{l'-1})_{d,m'} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m'}}. \tag{40}$$

A.3. Extended K-FAC

We divide the summands of

$$\mathbb{E}_{(x,t)} \left[\mathbb{E}_n \left[\sum_{m,m'} (\bar{a}_{l-1})_{b,m} (\bar{a}_{l'-1})_{d,m'} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m'}} \right] \right] \tag{41}$$

into the five groups so that it can be expressed as $G_1 + G_2 + G_3 + G_4 + G_5$. For the derivation, we define

$$(\{\mathcal{H}'''\}_{l,l'})_{a,c} = \mathbb{E}_{(x,t)} \left[\mathbb{E}_n \left[\sum_m \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,m}} \right] \right]. \tag{42}$$

Note that $\mathcal{H}'''$ is not symmetric in general unlike the others. In addition, let π and π' be permutations of $\{1, 2, \cdots, N\}$ such that $n \neq \pi(n) \neq \pi'(n) \neq n$ for all n.

(i) G_1: $m = m' = n$

$$\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,n}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\right]\right]$$
$$= \mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,n}\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\right]\right] = (\{\bar{A}\}_{l,l'})_{b,d}(\{\mathcal{H}\}_{l,l'})_{a,c} \quad (43)$$

$$\therefore G_1 = \bar{A} * \mathcal{H} \quad (44)$$

(ii) G_2: $m = m' \neq n$

$$\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}\right]\right]\\
&= (N-1)\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{l'-1})_{d,\pi(n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi(n)}}\right]\right]\\
&= \mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{l'-1})_{d,\pi(n)}\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(N-1)\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi(n)}}\right]\right]\\
&= (\{\bar{A}\}_{l,l'})_{b,d}\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}\right]\right]
\end{aligned} \quad (45)$$

since for any $m \neq n$,

$$\mathbb{E}_{(x,t)}\left[\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi(n)}}\right] = \mathbb{E}_{(x,t)}\left[\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}\right]. \quad (46)$$

Though $\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(N-1)\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi(n)}}\right]\right]$ and $\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}\right]\right]$ are equivalent, the latter is more efficient in estimating the expectation as all possible combinations in a single backward pass are considered. We use this type of efficient reformulation for other groups as well.

$$\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(N-1)\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi(n)}}\right]\right] = \mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}\right]\right]\\
&= \mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_m\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}} - \frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\right]\right] = (\{\mathcal{H}' - \mathcal{H}\}_{l,l'})_{a,c}
\end{aligned} \quad (47)$$

$$\therefore G_2 = \bar{A} * (\mathcal{H}' - \mathcal{H}) \quad (48)$$

(iii) G_3: $m = n \neq m'$

$$\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,m}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,m}}\right]\right]\\
&= (N-1)\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,\pi(n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,\pi(n)}}\right]\right]\\
&= \mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,\pi(n)}\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(N-1)\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,\pi(n)}}\right]\right]\\
&= \mathbb{E}_x\left[\mathbb{E}_n\left[\mathbb{E}_{m(\neq n)}\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,m}\right]\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,m}}\right]\right]
\end{aligned} \quad (49)$$

$$
\begin{aligned}
&\mathbb{E}_x\left[\mathbb{E}_n\left[\mathbb{E}_{m(\neq n)}\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,m}\right]\right]\right] \\
&\quad= \mathbb{E}_x\left[\mathbb{E}_n\left[\frac{1}{N-1}\left(N\mathbb{E}_m\left[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,m}\right] - (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,n}\right)\right]\right] \\
&\quad= \frac{1}{N-1}\left(N\mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,n}\right]\mathbb{E}_m\left[(\bar{a}_{l'-1})_{d,m}\right]\right] - (\{\bar{A}\}_{l,l'})_{b,d}\right) \\
&\quad= \frac{1}{N-1}(\{N\bar{A}' - \bar{A}\}_{l,l'})_{b,d}
\end{aligned}
\tag{50}
$$

$$
\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,m}}\right]\right] \\
&\quad= \mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,m}} - \frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\right]\right] = (\{\mathcal{H}''' - \mathcal{H}\}_{l,l'})_{a,c}
\end{aligned}
\tag{51}
$$

$$
\therefore G_3 = \frac{1}{N-1}(N\bar{A}' - \bar{A}) * (\mathcal{H}''' - \mathcal{H}) \tag{52}
$$

(iv) G_4: $m \neq n = m'$

$$
\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,n}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,n}}\right]\right] \\
&\qquad\qquad= \mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m(\neq n)}(\bar{a}_{l'-1})_{d,n}(\bar{a}_{l-1})_{b,m}\frac{\partial^2\mathcal{L}_n}{\partial(h_{l'})_{c,n}\partial(h_l)_{a,m}}\right]\right]
\end{aligned}
\tag{53}
$$

$$
\therefore G_4 = G_3^\top = \frac{1}{N-1}(N\bar{A}' - \bar{A}) * (\mathcal{H}'''^\top - \mathcal{H}) \tag{54}
$$

(v) G_5: $n \neq m \neq m' \neq n$

$$
\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{\substack{m,m'\\(n\neq m\neq m'\neq n)}}(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m'}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m'}}\right]\right] \\
&\quad= (N-1)(N-2)\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{l'-1})_{d,\pi'(n)}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi'(n)}}\right]\right] \\
&\quad= \mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{l'-1})_{d,\pi'(n)}\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[(N-1)(N-2)\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,\pi(n)}\partial(h_{l'})_{c,\pi'(n)}}\right]\right] \\
&\quad= \mathbb{E}_x\left[\mathbb{E}_n\left[\mathbb{E}_{\substack{m,m'\\(n\neq m\neq m'\neq n)}}\left[(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m'}\right]\right]\right]\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{\substack{m,m'\\(n\neq m\neq m'\neq n)}}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m'}}\right]\right]
\end{aligned}
\tag{55}
$$

$$
\begin{aligned}
&\mathbb{E}_x\left[\mathbb{E}_n\left[\mathbb{E}_{\substack{m,m'\\(n\neq m\neq m'\neq n)}}\left[(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m'}\right]\right]\right] \\
&\quad= \mathbb{E}_x\Bigg[\mathbb{E}_n\Bigg[\frac{1}{(N-1)(N-2)}\Big(N^2\mathbb{E}_{m,m'}[(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m'}] \\
&\qquad - N\mathbb{E}_m[(\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,m}] - N\mathbb{E}_m[(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,n}] + (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,n} \\
&\qquad - N\mathbb{E}_m[(\bar{a}_{l-1})_{b,m}(\bar{a}_{l'-1})_{d,m}] + (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'-1})_{d,n}\Big)\Bigg]\Bigg] \\
&\quad= \frac{1}{(N-1)(N-2)}(\{(N^2-2N)\bar{A}' - (N-2)\bar{A}\}_{l,l'})_{b,d} = \frac{1}{N-1}(\{N\bar{A}' - \bar{A}\}_{l,l'})_{b,d}
\end{aligned}
\tag{56}
$$

$$
\begin{aligned}
&\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{\substack{m,m'\\(n\neq m\neq m'\neq n)}}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m'}}\right]\right]\\
&=\mathbb{E}_{(x,t)}\left[\mathbb{E}_n\left[\sum_{m,m'}\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m'}}\right.\right.\\
&\quad-\sum_m\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,m}}-\sum_m\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,n}}+\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\\
&\quad\left.\left.-\sum_m\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,m}\partial(h_{l'})_{c,m}}+\frac{\partial^2\mathcal{L}_n}{\partial(h_l)_{a,n}\partial(h_{l'})_{c,n}}\right]\right]\\
&=(\{\mathcal{H}''-\mathcal{H}'''-\mathcal{H}'''^\top-\mathcal{H}'+2\mathcal{H}\}_{l,l'})_{a,c}
\end{aligned}
\tag{57}
$$

$$
\therefore G_5=\frac{1}{N-1}(N\bar{A}'-\bar{A})*(\mathcal{H}''-\mathcal{H}'''-\mathcal{H}'''^\top-\mathcal{H}'+2\mathcal{H})
\tag{58}
$$

Therefore, for $N>2$,

$$
\begin{aligned}
H&=G_1+G_2+G_3+G_4+G_5\\
&=\bar{A}*\mathcal{H}+\bar{A}*(\mathcal{H}'-\mathcal{H})+\frac{1}{N-1}(N\bar{A}'-\bar{A})*(\mathcal{H}'''+\mathcal{H}'''^\top-2\mathcal{H})\\
&\quad+\frac{1}{N-1}(N\bar{A}'-\bar{A})*(\mathcal{H}''-\mathcal{H}'''-\mathcal{H}'''^\top-\mathcal{H}'+2\mathcal{H})\\
&=\bar{A}*\mathcal{H}'+\frac{1}{N-1}\left(N\bar{A}'-\bar{A}\right)*(\mathcal{H}''-\mathcal{H}')
\end{aligned}
\tag{59}
$$

For $N=2$,

$$
\begin{aligned}
H=G_1+G_2+G_3+G_4&=\bar{A}*\mathcal{H}'+\frac{1}{N-1}(N\bar{A}'-\bar{A})*(\mathcal{H}'''+\mathcal{H}'''^\top-2\mathcal{H})\\
&=\bar{A}*\mathcal{H}'+\frac{1}{N-1}(N\bar{A}'-\bar{A})*(\mathcal{H}''-\mathcal{H}')
\end{aligned}
\tag{60}
$$

since

$$
\begin{aligned}
&(\{\mathcal{H}''-\mathcal{H}'\}_{l,l'})_{a,c}=(\{\mathcal{H}'''+\mathcal{H}'''^\top-2\mathcal{H}\}_{l,l'})_{a,c}\\
&\quad=\frac{1}{2}\mathbb{E}_{(x,t)}\left[\frac{\partial^2\mathcal{L}_1}{\partial(h_l)_{a,1}\partial(h_{l'})_{c,2}}+\frac{\partial^2\mathcal{L}_1}{\partial(h_l)_{a,2}\partial(h_{l'})_{c,1}}+\frac{\partial^2\mathcal{L}_2}{\partial(h_l)_{a,1}\partial(h_{l'})_{c,2}}+\frac{\partial^2\mathcal{L}_2}{\partial(h_l)_{a,2}\partial(h_{l'})_{c,1}}\right].
\end{aligned}
\tag{61}
$$

For $N=1$,

$$
H=G_1=\bar{A}*\mathcal{H}=\bar{A}*\mathcal{H}'+\left(N\bar{A}'-\bar{A}\right)*(\mathcal{H}''-\mathcal{H}')
\tag{62}
$$

since $\mathcal{H}=\mathcal{H}'=\mathcal{H}''$.

A.4. Positive semi-definiteness of XK-FAC

If we denote the n-th column vector of $\bar{a}_{l-1}$ by $(\bar{a}_{l-1})_{:,n}$, then

$$
\{\bar{A}\}_{l,l'}=\mathbb{E}_x[\mathbb{E}_n[(\bar{a}_{l-1})_{:,n}(\bar{a}_{l'-1})_{:,n}^\top]],
\tag{63}
$$

so

$$
\bar{A}=\mathbb{E}_x[\mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}(\bar{a}_{0:L-1})_{:,n}^\top]]\succeq 0
\tag{64}
$$

where $(\bar{a}_{0:L-1})_{:,n}=\begin{bmatrix}(\bar{a}_0)_{:,n}^\top & (\bar{a}_1)_{:,n}^\top & \cdots & (\bar{a}_{L-1})_{:,n}^\top\end{bmatrix}^\top$. For $\bar{A}'$,

$$
\bar{A}'=\mathbb{E}_x[\mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}]\mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}]^\top]\succeq 0.
\tag{65}
$$

Also,

$$\bar{A} - \bar{A}' = \mathbb{E}_x[\mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}(\bar{a}_{0:L-1})_{:,n}^\top] - \mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}]\mathbb{E}_n[(\bar{a}_{0:L-1})_{:,n}]^\top] \succeq 0 \tag{66}$$

since it is an expectation of covariance matrices. Thus,

$$\bar{A} \succeq 0, \quad \bar{A}' \succeq 0, \quad \bar{A} - \bar{A}' \succeq 0. \tag{67}$$

Similarly, $\hat{\mathcal{H}}' \succeq 0$, $\hat{\mathcal{H}}'' \succeq 0$, and $N\hat{\mathcal{H}}' - \hat{\mathcal{H}}'' \succeq 0$, because

$$\hat{\mathcal{H}}' = \mathbb{E}_{(x,y)}\left[\mathbb{E}_n\left[N\mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}^\top\right]\right]\right] \succeq 0, \tag{68}$$

$$\hat{\mathcal{H}}'' = \mathbb{E}_{(x,y)}\left[\mathbb{E}_n\left[N^2\mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\right]\mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\right]^\top\right]\right] \succeq 0, \tag{69}$$

$$N\hat{\mathcal{H}}' - \hat{\mathcal{H}}'' = N^2\mathbb{E}_{(x,y)}\left[\mathbb{E}_n\left[\mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}^\top\right] - \mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\right]\mathbb{E}_m\left[\frac{\partial \mathcal{L}_n}{\partial (h_{1:L})_{:,m}}\right]^\top\right]\right] \succeq 0. \tag{70}$$

Therefore, for $N \geq 2$,

$$H = \bar{A} * \hat{\mathcal{H}}' + \frac{1}{N-1}\left(N\bar{A}' - \bar{A}\right) * (\hat{\mathcal{H}}'' - \hat{\mathcal{H}}') = \frac{1}{N-1}(\bar{A} - \bar{A}') * (N\hat{\mathcal{H}}' - \hat{\mathcal{H}}'') + \bar{A}' * \hat{\mathcal{H}}'' \succeq 0 \tag{71}$$

since the Khatri–Rao product of two symmetrically partitioned positive semi-definite matrices is positive semi-definite [19, 36]. For $N = 1$, $H = \bar{A} * \hat{\mathcal{H}}' \succeq 0$.

Also, this proof reveals how to implement XK-FAC efficiently. Computing each block matrices in XK-FAC, $\bar{A}$, $\bar{A}'$, $\hat{\mathcal{H}}'$, and $\hat{\mathcal{H}}''$, just requires some averages and matrix-matrix multiplications (Equations 64, 65, 68, 69). Thus, XK-FAC can be implemented very easily using any basic linear algebra subprograms (BLAS).

A.5. Combining XK-FAC and KFC

For the l-th convolutional layer, let a_{l-1} of size $C_{l-1} \times (S_{l-1}N)$ be an input to the layer and W_l of size $C_l \times (C_{l-1}K_l + 1)$ be the weight, where S and K represent the flattened spatial dimension and kernel dimension, respectively. A convolution operation can be converted to a matrix-matrix multiplication by unrolling the input [5] (this unrolling function is often called `im2col`). Let $\boldsymbol{a}_{l-1}$ of size $(C_{l-1}K_l) \times (S_lN)$ be the unrolled input of a_{l-1} (denoted by $[\![\cdot]\!]$ in [10]) and $\bar{\boldsymbol{a}}_{l-1}$ of size $(C_{l-1}K_l + 1) \times (S_lN)$ be the unrolled input with homogeneous dimension appended (denoted by $[\![\cdot]\!]_H$ in [10]). Then, the output h_l of size $C_l \times (S_lN)$ is

$$h_l = W_l\bar{\boldsymbol{a}}_{l-1}. \tag{72}$$

Thus, if the Hessian is approximated by the Fisher information matrix, then Equation 7 becomes

$$\mathbb{E}_{(x,y)}\left[\mathbb{E}_n\left[\sum_{s,s',m,m'}(\bar{\boldsymbol{a}}_{l-1})_{b,(s,m)}(\bar{\boldsymbol{a}}_{l-1})_{d,(s',m')}\frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}}\frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s',m')}}\right]\right], \tag{73}$$

where s and s' index the spatial location.

KFC assumes three conditions: IAD, SH, and SUD, and these conditions can be straightforwardly extended for a different mini-batches case. If we apply KFC for each (m, m'), we get

$$\sum_{m,m'}\left(\sum_s \mathbb{E}_x\left[\mathbb{E}_n\left[(\bar{\boldsymbol{a}}_{l-1})_{b,(s,m)}(\bar{\boldsymbol{a}}_{l-1})_{d,(s,m')}\right]\right]\right)\left(\frac{1}{S_l}\sum_s \mathbb{E}_{(x,y)}\left[\mathbb{E}_n\left[\frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}}\frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s,m')}}\right]\right]\right). \tag{74}$$

Now, the N^2 summands here can also be divided into the five groups, and the remaining processes are exactly the same as Appendix A.3. Therefore,

$$\{H\}_{l,l} = \{\bar{A}\}_{l,l} \otimes \{\hat{\mathcal{H}}'\}_{l,l} + \frac{1}{\max(N-1, 1)}\left(N\{\bar{A}'\}_{l,l} - \{\bar{A}\}_{l,l}\right) \otimes (\{\hat{\mathcal{H}}''\}_{l,l} - \{\hat{\mathcal{H}}'\}_{l,l}), \tag{75}$$

where

$$(\{\bar{A}\}_{l,l})_{b,d} = \sum_s \mathbb{E}_x \left[\mathbb{E}_n \left[(\bar{\boldsymbol{a}}_{l-1})_{b,(s,n)} (\bar{\boldsymbol{a}}_{l-1})_{d,(s,n)} \right] \right], \tag{76}$$

$$(\{\bar{A}'\}_{l,l})_{b,d} = \sum_s \mathbb{E}_x \left[\mathbb{E}_n \left[(\bar{\boldsymbol{a}}_{l-1})_{b,(s,n)} \right] \mathbb{E}_n \left[(\bar{\boldsymbol{a}}_{l-1})_{d,(s,n)} \right] \right], \tag{77}$$

$$(\{\hat{\mathcal{H}}'\}_{l,l})_{a,c} = \frac{1}{S_l} \sum_s \mathbb{E}_{(x,y)} \left[\mathbb{E}_n \left[\sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}} \frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s,m)}} \right] \right], \tag{78}$$

$$(\{\hat{\mathcal{H}}''\}_{l,l})_{a,c} = \frac{1}{S_l} \sum_s \mathbb{E}_{(x,y)} \left[\mathbb{E}_n \left[\left(\sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}} \right) \left(\sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s,m)}} \right) \right] \right], \tag{79}$$

and $\otimes$ is the Kronecker product.

A.6. Effect of α_s and α_t

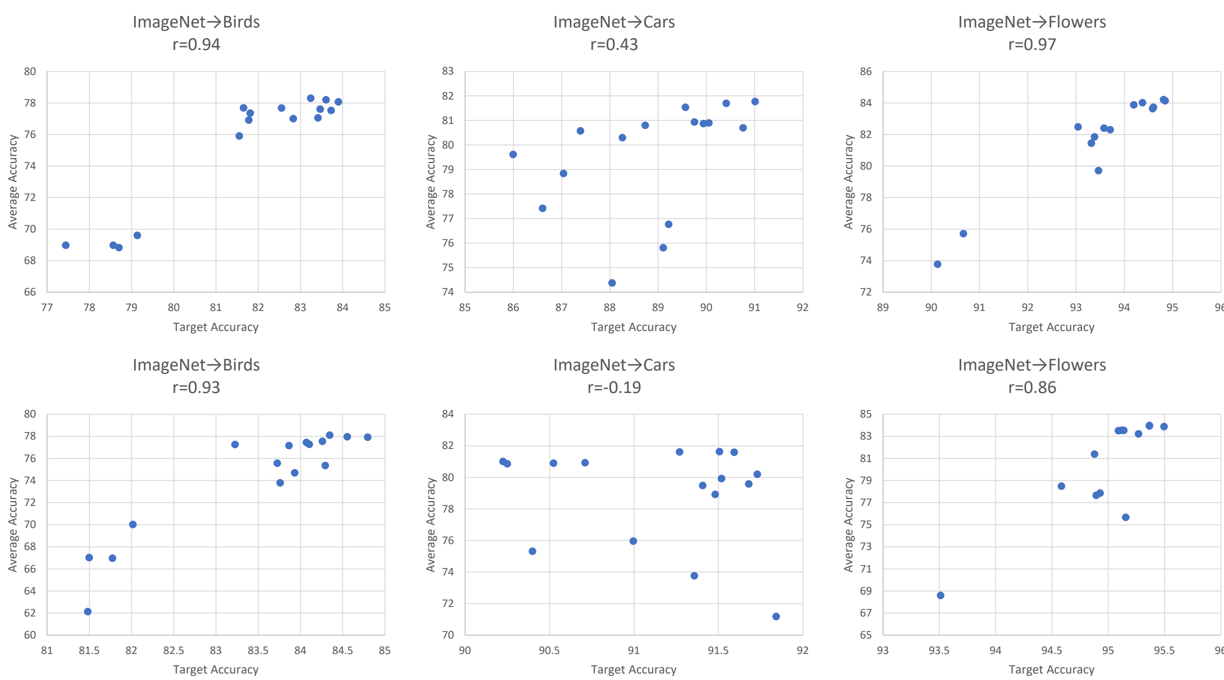

Figure 2: Each point represents the result of a specific hyperparameters (learning rate, damping) setting. α_s and α_t are used in the top graphs, and they are not used in the bottom graphs. With α_s and α_t, the target accuracy and average accuracy have a more positive correlation.

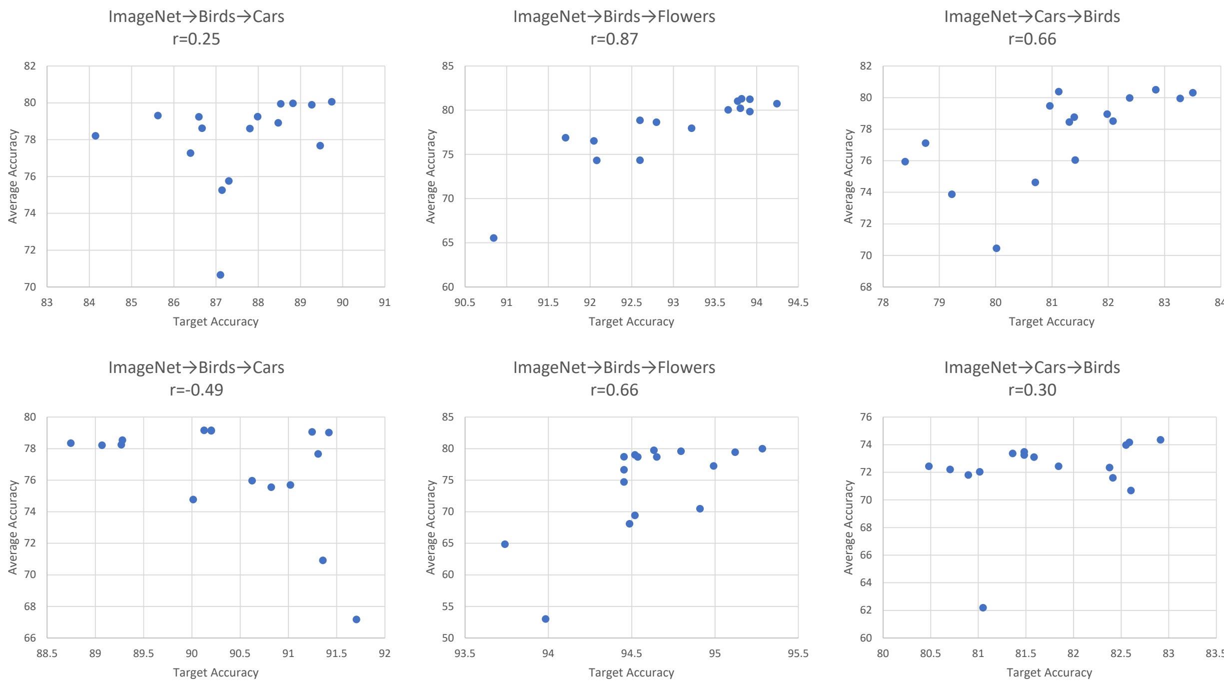

Figure 3: Each point represents the result of a specific hyperparameters (learning rate, damping) setting. α_s and α_t are used in the top graphs, and they are not used in the bottom graphs. With α_s and α_t, the target accuracy and average accuracy have a more positive correlation.

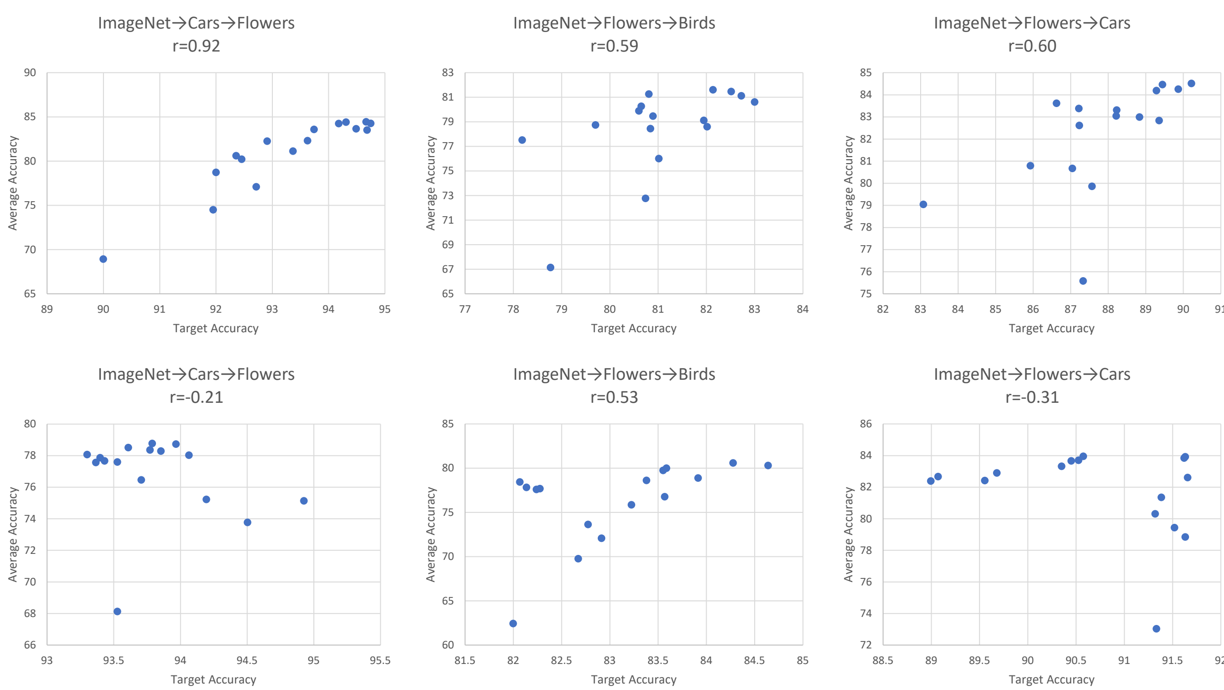

Figure 4: Each point represents the result of a specific hyperparameters (learning rate, damping) setting. α_s and α_t are used in the top graphs, and they are not used in the bottom graphs. With α_s and α_t, the target accuracy and average accuracy have a more positive correlation.

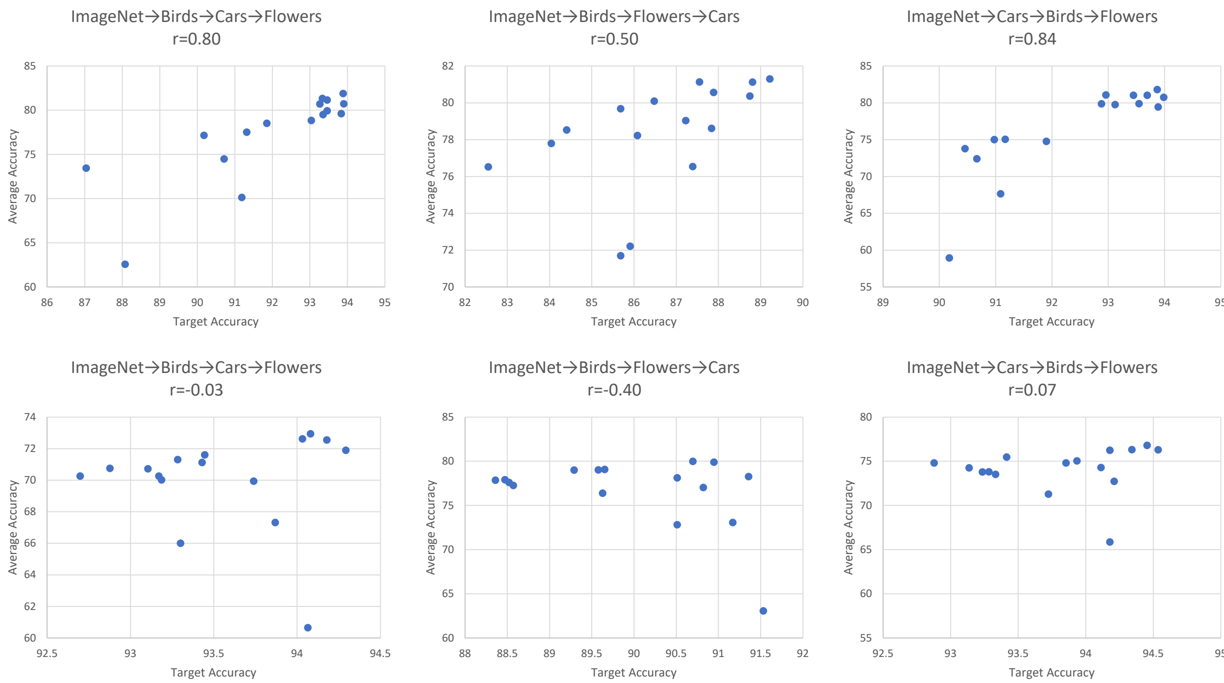

Figure 5: Each point represents the result of a specific hyperparameters (learning rate, damping) setting. α_s and α_t are used in the top graphs, and they are not used in the bottom graphs. With α_s and α_t, the target accuracy and average accuracy have a more positive correlation.

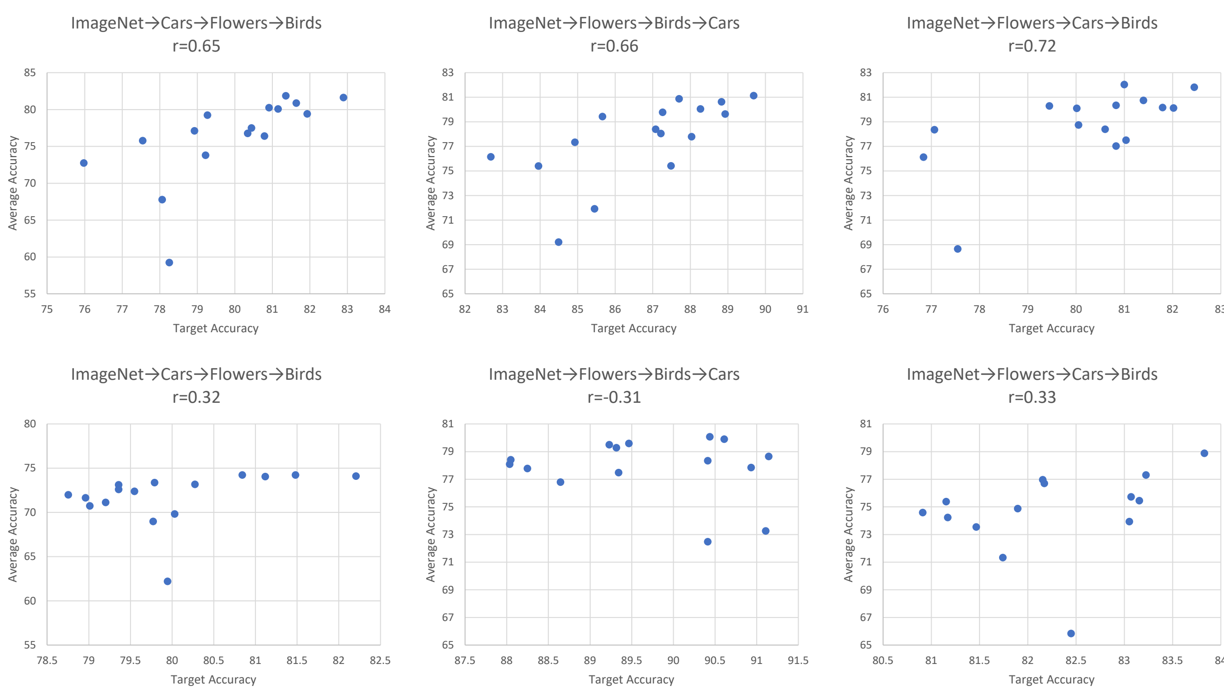

Figure 6: Each point represents the result of a specific hyperparameters (learning rate, damping) setting. α_s and α_t are used in the top graphs, and they are not used in the bottom graphs. With α_s and α_t, the target accuracy and average accuracy have a more positive correlation.