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 \( h_{i,m} = \sum_k (W_k)_i, (\bar{a}_{l-1})_{k,m} \),

\[
\frac{\partial L_n}{\partial (W_l)_{a,b}} = \sum_{m,t} \frac{\partial L_n}{\partial (h_l)_{i,m}} \frac{\partial (h_l)_{i,m}}{\partial (W_l)_{a,b}} = \sum_{m} \frac{\partial L_n}{\partial (h_l)_{a,m}} (\bar{a}_{l-1})_{b,m}.
\]

Then,

\[
\frac{\partial^2 L_n}{\partial (W_l')_{c,d} \partial (W_l)_{a,b}} = \sum_{m} \left( \frac{\partial}{\partial (W_l')_{c,d}} \left( \frac{\partial L_n}{\partial (h_l)_{a,m}} (\bar{a}_{l-1})_{b,m} + \frac{\partial L_n}{\partial (h_l)_{a,m}} \frac{\partial (a_{l-1})_{b,m}}{\partial (W_l')_{c,d}} \right) \right).
\]

Using the chain rule,

\[
\frac{\partial}{\partial (W_l')_{c,d}} \left( \frac{\partial L_n}{\partial (h_l)_{a,m}} \right) = \sum_{m',t} \frac{\partial^2 L_n}{\partial (h_l')_{i,m'} \partial (h_l)_{a,m}} \frac{\partial (h_l')_{i,m'}}{\partial (W_l')_{c,d}} \]

\[
= \sum_{m'} \frac{\partial^2 L_n}{\partial (h_l')_{c,m'} \partial (h_l)_{a,m}} (\bar{a}_{l-1})_{d,m'}.
\]

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

\[
\frac{\partial (\bar{a}_{l-1})_{b,m}}{\partial (W_l')_{c,d}} = 0,
\]

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,

\[
\frac{\partial^2 L_n}{\partial (W_l)_{a,b} \partial (W_l')_{c,d}} = \sum_{m,m'} (\bar{a}_{l-1})_{b,m} (\bar{a}_{l-1})_{d,m'} \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_l')_{c,m'}}.
\]

A.3. Extended K-FAC

We divide the summands of

\[
E_{(x,t)} \left[ E_n \left[ \sum_{m,m'} (\bar{a}_{l-1})_{b,m} (\bar{a}_{l-1})_{d,m'} \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_l')_{c,m'}} \right] \right]
\]

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

\[
(\{H''\})_{l,l'} a, c = E_{(x,t)} \left[ E_n \left[ \sum_m \frac{\partial^2 L_n}{\partial (h_l)_{a,c} \partial (h_l')_{c,m}} \right] \right].
\]

Note that \( H'' \) is not symmetric in general unlike the others. In addition, let \( \pi \) and \( \pi' \) be permutations of \( \{1, 2, \cdots, N\} \) such that \( n \neq \pi(n) \neq \pi'(n) \neq n \) for all \( n \).
(i) $G_1$: $m = m' = n$

$$E_{(x,t)} \left[ E_n \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'}^{-1})_{d,n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right]$$

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

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

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

$$= (N - 1)E_{(x,t)} \left[ E_n \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'}^{-1})_{d,n} \right] \right] \cdot E_{(x,t)} \left[ E_n \left[ \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right]$$

$$= E_x \left[ E_n \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'}^{-1})_{d,n} \right] \right] \cdot E_{(x,t)} \left[ E_n \left[ \sum_{m \neq n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m}} \right] \right]$$

since for any $m \neq n$,

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

Though $E_{(x,t)} \left[ E_n \left[ (N - 1) \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right]$ and $E_{(x,t)} \left[ E_n \left[ \sum_{m \neq n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m}} \right] \right]$ 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.

$$E_{(x,t)} \left[ E_n \left[ (N - 1) \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right] = E_{(x,t)} \left[ E_n \left[ \sum_{m \neq n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m}} \right] \right]$$

$$= E_{(x,t)} \left[ E_n \left[ \sum_{m} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m}} \right] \right] = (\mathcal{H}_l' - \mathcal{H}_l)_{a,c}$$

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

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

$$= (N - 1)E_{(x,t)} \left[ E_n \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'}^{-1})_{d,n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right]$$

$$= E_x \left[ E_n \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{l'}^{-1})_{d,n} \right] \right] \cdot E_{(x,t)} \left[ E_n \left[ (N - 1) \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \right]$$

$$= E_x \left[ E_n \left[ \sum_{m \neq n} \right] \right] \cdot E_{(x,t)} \left[ E_n \left[ \sum_{m \neq n} \frac{\partial^2 \mathcal{L}_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,m}} \right] \right]$$
\[
\begin{align*}
E_x \left[ E_n \left[ E_{m(\neq n)} \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{\nu-1})_{d,m} \right] \right] \right] \\
= E_x \left[ E_n \left[ \frac{1}{N-1} \left( N E_m \left[ (\bar{a}_{l-1})_{b,n}(\bar{a}_{\nu-1})_{d,m} \right] - (\bar{a}_{l-1})_{b,n}(\bar{a}_{\nu-1})_{d,n} \right) \right] \right] \\
= \frac{1}{N-1} \left( N E_x \left[ E_n \left[ (\bar{a}_{l-1})_{b,n} \right] E_m \left[ (\bar{a}_{\nu-1})_{d,m} \right] \right] - (\{\bar{A}\}_{l,\nu})_{b,d} \right) \\
= \frac{1}{N-1} (\{N\bar{A}' - \bar{A}\}_{l,\nu})_{b,d}
\end{align*}
\]
\[
E_{(x,t)} \left[ E_n \left[ \sum_{m(\neq n)} \frac{\partial^2 L_n}{\partial (h_t)_{a,n} \partial (h_{\nu'})_{c,m}} \right] \right] \\
= E_{(x,t)} \left[ E_n \left[ \sum_{m(\neq n)} \frac{\partial^2 L_n}{\partial (h_t)_{a,n} \partial (h_{\nu'})_{c,m}} \right] \right] = (\{H''' - \mathcal{H}\}_{l,\nu})_{a,c}
\]
\[
\therefore G_3 = \frac{1}{N-1} (N \bar{A}' - \bar{A}) * (H''' - \mathcal{H})
\]
\[
E_{(x,t)} \left[ E_n \left[ \sum_{m,m'} (\bar{a}_{l-1})_{b,m}(\bar{a}_{\nu-1})_{d,m'} \frac{\partial^2 L_n}{\partial (h_t)_{a,m} \partial (h_{\nu'})_{c,m'}} \right] \right] \\
= (N-1)(N-2) E_{(x,t)} \left[ E_n \left[ (\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{\nu-1})_{d,\pi'(n)} \frac{\partial^2 L_n}{\partial (h_t)_{a,\pi(n)} \partial (h_{\nu'})_{c,\pi'(n)}} \right] \right]
\]
\[
= E_x \left[ E_n \left[ (\bar{a}_{l-1})_{b,\pi(n)}(\bar{a}_{\nu-1})_{d,\pi'(n)} \right] E_{(x,t)} \left[ E_n \left[ (N-1)(N-2) \frac{\partial^2 L_n}{\partial (h_t)_{a,\pi(n)} \partial (h_{\nu'})_{c,\pi'(n)}} \right] \right] \right]
\]
\[
= E_x \left[ E_n \left[ \sum_{m,m'} \frac{\partial^2 L_n}{\partial (h_t)_{a,n} \partial (h_{\nu'})_{c,m'}} \right] \right] E_{(x,t)} \left[ E_n \left[ \sum_{m,m'} \frac{\partial^2 L_n}{\partial (h_t)_{a,n} \partial (h_{\nu'})_{c,m'}} \right] \right]
\]
\[
\begin{align*}
E_x \left[ E_n \left[ \sum_{m,m'} \frac{\partial^2 L_n}{\partial (h_t)_{a,n} \partial (h_{\nu'})_{c,m'}} \right] \right] \\
= E_x \left[ E_n \left[ \frac{1}{(N-1)(N-2)} \left( N^2 E_m[(\bar{a}_{l-1})_{b,m}(\bar{a}_{\nu-1})_{d,m}] \\
- N E_m[(\bar{a}_{l-1})_{b,n}(\bar{a}_{\nu-1})_{d,m}] - N E_m[(\bar{a}_{l-1})_{b,m}(\bar{a}_{\nu-1})_{d,n}] + (\bar{a}_{l-1})_{b,n}(\bar{a}_{\nu-1})_{d,n} \right) \right] \right]
\end{align*}
\]
\[
= \frac{1}{(N-1)(N-2)} \left((N^2 - 2N) \bar{A}' - (N - 2) \bar{A}_{l,\nu})_{b,d} = \frac{1}{N-1} (\{N\bar{A}' - \bar{A}\}_{l,\nu})_{b,d}
\]
\[
\begin{align*}
\mathbb{E}(x,t) \left[ \mathbb{E}_n \left[ \sum_{m,m'} \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m'}} \right] \right] \\
= \mathbb{E}(x,t) \left[ \mathbb{E}_n \left[ \sum_{m,m'} \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m'}} ight] \\
\quad - \sum_m \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,m}} - \sum_m \frac{\partial^2 L_n}{\partial (h_l)_{a,m} \partial (h_{l'})_{c,n}} + \frac{\partial^2 L_n}{\partial (h_l)_{a,n} \partial (h_{l'})_{c,n}} \right] \\
= \left( \{H'' - H'' - H'' + 2H\}_{l',l'} \right)_{a,c} \\
\therefore G_5 = \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' - H'' - H'' - \bar{H}' + 2\bar{H}) \\
\end{align*}
\]

Therefore, for \( N > 2 \),

\[
H = G_1 + G_2 + G_3 + G_4 + G_5 \\
= \bar{A} \ast H + \bar{A} \ast (H' - H) + \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' + H'' - 2\bar{H}) \\
\quad + \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' - H'' - H'' - \bar{H}' + 2\bar{H}) \\
= \bar{A} \ast H' + \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' - H') \tag{59}
\]

For \( N = 2 \),

\[
H = G_1 + G_2 + G_3 + G_4 = \bar{A} \ast H' + \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' + H'' - 2\bar{H}) \\
= \bar{A} \ast H' + \frac{1}{N-1} (N \bar{A}' - \bar{A}) \ast (H'' - H') \tag{60}
\]

since

\[
\left( \{H'' - H'\}_{l',l'} \right)_{a,c} = \left( \{H'' + H'' - 2\bar{H}\}_{l',l'} \right)_{a,c} \\
= \frac{1}{2} \mathbb{E}(x,t) \left[ \frac{\partial^2 L_1}{\partial (h_l)_{a,1} \partial (h_{l'})_{c,1}} + \frac{\partial^2 L_1}{\partial (h_l)_{a,2} \partial (h_{l'})_{c,1}} + \frac{\partial^2 L_2}{\partial (h_l)_{a,1} \partial (h_{l'})_{c,2}} + \frac{\partial^2 L_2}{\partial (h_l)_{a,2} \partial (h_{l'})_{c,1}} \right]. \tag{61}
\]

For \( N = 1 \),

\[
H = G_1 = A \ast H = \bar{A} \ast H' + (N \bar{A}' - \bar{A}) \ast (H'' - H') \tag{62}
\]

since \( H = H' = H'' \).

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

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

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 \( (\bar{a}_{0:L-1})_{:,n} = [(\bar{a}_0)_{:,n}^\top (\bar{a}_1)_{:,n}^\top \cdots (\bar{a}_{L-1})_{:,n}^\top]^\top \). 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,
\[
\bar{A} - \bar{A}' = \mathbb{E}_x [\mathbb{E}_n ([\bar{a}_{0,L-1},n] (\bar{a}_{0,L-1})^\top_n)] - \mathbb{E}_n ([\bar{a}_{0,L-1},n] (\bar{a}_{0,L-1})^\top_n) \geq 0 \tag{66}
\]

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, \(\bar{\mathcal{H}}' \succeq 0, \bar{\mathcal{H}}'' \succeq 0, \) and \(N\bar{\mathcal{H}}' - \bar{\mathcal{H}}'' \succeq 0,\) because
\[
\bar{\mathcal{H}}' = \mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ -N \mathbb{E}_m \left[ \begin{bmatrix} \partial L_n & \partial L_n \\ \partial(h_{1:L},m) & \partial(h_{1:L},m) \end{bmatrix} \right] \right] \right] \succeq 0, \tag{68}
\]
\[
\bar{\mathcal{H}}'' = \mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ N^2 \mathbb{E}_m \left[ \begin{bmatrix} \partial L_n & \partial L_n \\ \partial(h_{1:L},m) & \partial(h_{1:L},m) \end{bmatrix} \right] \right] \right] \succeq 0, \tag{69}
\]
\[
N\bar{\mathcal{H}}' - \bar{\mathcal{H}}'' = \mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ \begin{bmatrix} \partial L_n & \partial L_n \\ \partial(h_{1:L},m) & \partial(h_{1:L},m) \end{bmatrix} \right] \right] \succeq 0. \tag{70}
\]

Therefore, for \(N \geq 2,\)
\[
\bar{H} = \bar{A} \ast \bar{\mathcal{H}}' + \frac{1}{N-1} (N\bar{A}' - \bar{A} \ast \bar{\mathcal{H}}'' + \bar{A}' \ast \bar{\mathcal{H}}'') \succeq 0 \tag{71}
\]
since the Khatri–Rao product of two symmetrically partitioned positive semi-definite matrices is positive semi-definite [19, 36]. For \(N = 1, \bar{H} = \bar{A} \ast \bar{\mathcal{H}}' \succeq 0.\)

Also, this proof reveals how to implement XK-FAC efficiently. Computing each block matrices in XK-FAC, \(\bar{A}, \bar{A}', \bar{\mathcal{H}}',\) and \(\bar{\mathcal{H}}'',\) 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 \texttt{im2col}). Let \(a_{l-1}\) of size \((C_{l-1}K_l) \times (S_lN)\) be the unrolled input of \(a_{l-1}\) (denoted by \([^]\) in [10]) and \(a_{l-1}\) of size \((C_{l-1}K_l + 1) \times (S_lN)\) be the unrolled input with homogeneous dimension appended (denoted by \([^]\) in [10]). Then, the output \(h_l\) of size \(C_l \times (S_lN)\) is
\[
h_l = W_l \bar{a}_{l-1}. \tag{72}
\]

Thus, if the Hessian is approximated by the Fisher information matrix, then Equation 7 becomes
\[
\mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ \sum_{s, s', m, m'} (\bar{a}_{l-1})_{b,(s,m)} (\bar{a}_{l-1})_{d,(s',m')} \frac{\partial L_n}{\partial(h_l)_{a,(s,m)}} \frac{\partial L_n}{\partial(h_l)_{c,(s',m')}} \right] \right], \tag{73}
\]

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

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

\[
\begin{align*}
\{\bar{A}\}_{l,l}^{b,d} &= \sum_s \mathbb{E}_x \left[ \mathbb{E}_n \left[ (\bar{a}_{l-1})_{b,(s,n)} (\bar{a}_{l-1})_{d,(s,n)} \right] \right], \\
\{\bar{A}'\}_{l,l}^{b,d} &= \sum_s \mathbb{E}_x \left[ \mathbb{E}_n \left[ (\bar{a}_{l-1})_{b,(s,n)} \right] \mathbb{E}_n \left[ (\bar{a}_{l-1})_{d,(s,n)} \right] \right], \\
\{\bar{H}'\}_{l,l}^{a,c} &= \frac{1}{S_l} \sum_s \mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ \sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}} \frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s,m)}} \right] \right], \\
\{\bar{H}''\}_{l,l}^{a,c} &= \frac{1}{S_l} \sum_s \mathbb{E}_{(x,y)} \left[ \mathbb{E}_n \left[ \left( \sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{a,(s,m)}} \right) \left( \sum_m \frac{\partial \mathcal{L}_n}{\partial (h_l)_{c,(s,m)}} \right) \right] \right],
\end{align*}
\]

and \(\otimes\) is the Kronecker product.

### A.6. Effect of \(\alpha_s\) and \(\alpha_t\)

![Figure 2](image)

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