A. Proof

Proof of Lemma 2: Based on the implicit function theorem (Lorraine et al., 2020), or we simply set $\frac{\mathcal{L}_1(w^*,\alpha)}{\partial w} = 0$ since the model weights w achieved the local optimal in the training set with α, we have:

$$\frac{\partial \mathcal{L}_1(w^*(\alpha),\alpha)}{\partial w} = 0, \tag{10}$$

and we have

$$\begin{aligned} \frac{\partial}{\partial \alpha}\left(\frac{\partial \mathcal{L}_1(w^*(\alpha),\alpha)}{\partial w}\right) &= 0, \\ \frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w} + \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\frac{\partial (w^*(\alpha))}{\partial \alpha} &= 0, \\ \frac{\partial (w^*(\alpha))}{\partial \alpha} &= -\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1}\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}. \end{aligned} \tag{11}$$

In this way, the hypergradient could be formulated as

$$\nabla_\alpha \mathcal{L}_2 = \frac{\partial \mathcal{L}_2}{\partial \alpha} - \frac{\partial \mathcal{L}_2}{\partial w}\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1}\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}. \tag{12}$$

□

Proof of Corollary 1: The key in this proposition is to use the Neumann series to approximate the $\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1}$.

Based on the Neumann series approximation, for $\|I - A\| < 1$, we have:

$$A^{-1} = \sum_{k=0}^{\infty}(I - A)^k. \tag{13}$$

Based Assumption 2.1, we have $\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} < L_1^{\nabla_w}$. With $\gamma < \frac{1}{L_1^{\nabla_w}}$, we have $\left\|I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right\| < 1$ (Shaban et al., 2019; Lorraine et al., 2020). When we conduct the Neumann series approximation for $\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1}$ in the optimal point, we have:

$$\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1} = \gamma(I - I + \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w})^{-1} = \gamma \sum_{j=0}^{\infty}\left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j. \tag{14}$$

So that:

$$\nabla_\alpha \mathcal{L}_2 = \frac{\partial \mathcal{L}_2}{\partial \alpha} - \gamma \frac{\partial \mathcal{L}_2}{\partial w}\sum_{j=0}^{\infty}\left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j \frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}. \tag{15}$$

□

Proof of Theorem 1 Based on the Eq. (8) and (7), we have

$$\nabla_\alpha \mathcal{L}_2 - \nabla_\alpha \tilde{\mathcal{L}}_2 = \gamma \frac{\partial \mathcal{L}_2}{\partial w}\sum_{j=K+1}^{\infty}\left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j \frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}. \tag{16}$$

Since the $\mathcal{L}_1$ is μ-strongly convex, and $\gamma\mu I \preceq \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \preceq I$, we have

$$\sum_{j=K+1}^{\infty}\left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j \leq \sum_{j=K+1}^{\infty}\left[I - \gamma\mu\right]^j. \tag{17}$$

Based on the sum of geometric sequence, we have

$$\sum_{j=K+1}^{\infty}\left[I-\gamma\frac{\partial^2\mathcal{L}_1}{\partial w\partial w}\right]^j \leq \frac{1}{\gamma\mu}(1-\gamma\mu)^{K+1}. \tag{18}$$

Since $\frac{\partial\mathcal{L}_2}{\partial w}$ and $\frac{\partial^2\mathcal{L}_1}{\partial\alpha\partial w}$ are bounded, we have

$$\left\|\nabla_\alpha\mathcal{L}_2-\nabla_\alpha\tilde{\mathcal{L}}_2\right\| \leqslant C_{\mathcal{L}_1^{w\alpha}}\, C_{\mathcal{L}_2^{w}}\,\frac{1}{\mu}(1-\gamma\mu)^{K+1}. \tag{19}$$

□

Proof: Corollary 2 Based on the definitions, the hypergradient of truncated back-propagation and the proposed Neumann approximation based hypergradient are defined in Eq.(4) and Eq.(8). When we assume that w_t has converged to a stationary point w^* in the last K steps, we have

$$\begin{aligned} w_i(\alpha) = w_j(\alpha) = w^*(\alpha), &\qquad for\ all\ i,j\in[T-K+1,T];\\ \frac{\partial\Phi(w_i,\alpha)}{\partial w_i} = \frac{\partial\Phi(w_j,\alpha)}{\partial w_j} = \frac{\partial\Phi(w^*(\alpha),\alpha)}{\partial w^*(\alpha)} = A_T, &\qquad for\ all\ i,j\in[T-K+1,T];\\ \frac{\partial\Phi(w_i,\alpha)}{\partial\alpha} = \frac{\partial\Phi(w_j,\alpha)}{\partial\alpha} = \frac{\partial\Phi(w^*(\alpha),\alpha)}{\partial\alpha} = B_T, &\qquad for\ all\ i,j\in[T-K+1,T]. \end{aligned} \tag{20}$$

Now the truncated back-propagation could be formulated as:

$$\begin{aligned} h_{T-K} &= \frac{\partial\mathcal{L}_2}{\partial\alpha}+\frac{\partial\mathcal{L}_2}{\partial w_T}\Big(\sum_{t=T-K+1}^{T}B_tA_{t+1}...A_T\Big)\\ &= \frac{\partial\mathcal{L}_2}{\partial\alpha}+\frac{\partial\mathcal{L}_2}{\partial w_T}\Big(\sum_{t=0}^{K}B_TA_T^t\Big). \end{aligned} \tag{21}$$

We have

$$\begin{aligned} A_T &= \frac{\partial\Phi(w^*(\alpha),\alpha)}{\partial w^*(\alpha)} = \frac{\partial(w^*-\eta\frac{\partial\mathcal{L}_1}{\partial w})}{\partial w^*} = I-\gamma\frac{\partial^2\mathcal{L}_1(w^*)}{\partial w\partial w},\\ B_T &= \frac{\partial\Phi(w^*(\alpha),\alpha)}{\partial\alpha} = \frac{\partial(w^*-\eta\frac{\partial\mathcal{L}_1}{\partial w})}{\partial\alpha} = -\gamma\frac{\partial^2\mathcal{L}_1(w^*)}{\partial\alpha\partial w}. \end{aligned} \tag{22}$$

From the above, we have

$$\begin{aligned} h_{T-K} &= \frac{\partial\mathcal{L}_2}{\partial\alpha}+\frac{\partial\mathcal{L}_2}{\partial w_T}\Big(\sum_{t=0}^{K}B_TA_T^t\Big)\\ &= \frac{\partial\mathcal{L}_2}{\partial\alpha}-\gamma\frac{\partial\mathcal{L}_2}{\partial w}\sum_{j=0}^{K}\left[I-\gamma\frac{\partial^2\mathcal{L}_1}{\partial w\partial w}\right]^j\frac{\partial^2\mathcal{L}_1}{\partial\alpha\partial w}\\ &= \nabla_\alpha\tilde{\mathcal{L}}_2. \end{aligned} \tag{23}$$

□

Proof of Lemma 4: First, for $\forall(\alpha,\alpha')$, we have

$$\begin{aligned} &\|\nabla_\alpha\mathcal{L}_2(w,\alpha)-\nabla_\alpha\mathcal{L}_2(w,\alpha')\| = \|\nabla_\alpha\mathcal{L}_2(\cdot,\alpha)-\nabla_\alpha\mathcal{L}_2(\cdot,\alpha')+\nabla_\alpha\mathcal{L}_2(w(\alpha),\cdot)-\nabla_\alpha\mathcal{L}_2(w(\alpha'),\cdot)\|\\ =&\|\nabla_\alpha\mathcal{L}_2(\cdot,\alpha)-\nabla_\alpha\mathcal{L}_2(\cdot,\alpha')+\nabla_w\mathcal{L}_2(w(\alpha),\cdot)\nabla_\alpha w(\alpha)-\nabla_w\mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha')\|\\ \leq&\|\nabla_\alpha\mathcal{L}_2(\cdot,\alpha)-\nabla_\alpha\mathcal{L}_2(\cdot,\alpha')\|+\|\nabla_w\mathcal{L}_2(w(\alpha),\cdot)\nabla_\alpha w(\alpha)-\nabla_w\mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha')\|. \end{aligned} \tag{24}$$

Then we divide Eq.(24) to two parts. For the first part, based on the Assumption 1.2, we have:

$$\|\nabla_\alpha \mathcal{L}_2(\cdot,\alpha) - \nabla_\alpha \mathcal{L}_2(\cdot,\alpha')\| \leq L_2^{\nabla_\alpha}(\alpha - \alpha'). \tag{25}$$

And for the second part of Eq.(24), we have

$$\begin{aligned}
&\|\nabla_w \mathcal{L}_2(w(\alpha),\cdot)\nabla_\alpha w(\alpha) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha')\| \\
=&\|\nabla_w \mathcal{L}_2(w(\alpha),\cdot)\nabla_\alpha w(\alpha) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha') + \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha)\| \\
\leq&\|\nabla_w \mathcal{L}_2(w(\alpha'),\cdot) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\| \|\nabla_\alpha w(\alpha)\| + \|\nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\| \|\nabla_\alpha w(\alpha) - \nabla_\alpha w(\alpha')\|.
\end{aligned} \tag{26}$$

Based Assumption 1.3, we have

$$\|\nabla_w \mathcal{L}_2(w(\alpha'),\cdot) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\| \leq L_2^{\nabla_w} \|w(\alpha) - w(\alpha')\|, \tag{27}$$

and based Assumption 2.2 that we have

$$\|w(\alpha) - w(\alpha')\| \leq L_w \|\alpha - \alpha'\|, \quad \text{and} \quad \|\nabla_\alpha w(\alpha) - \nabla_\alpha w(\alpha')\| \leq L_{\nabla_\alpha w} \|\alpha - \alpha'\|. \tag{28}$$

Based on Assumption 1.3, we know $\nabla_w \mathcal{L}_2(w(\alpha'),\cdot)$ is bounded that $\nabla_w \mathcal{L}_2(w(\alpha'),\cdot) \leq L_2^w$. $\nabla_\alpha w(\alpha)$ is also bounded by $\|\nabla_\alpha w(\alpha)\| \leq L_w$. In this way, Eq.(26) could be rephrased as:

$$\|\nabla_w \mathcal{L}_2(w(\alpha),\cdot)\nabla_\alpha w(\alpha) - \nabla_w \mathcal{L}_2(w(\alpha'),\cdot)\nabla_\alpha w(\alpha')\| \leq L_2^{\nabla_w} L_w^2 \|\alpha - \alpha'\| + L_2^w L_{\nabla_\alpha w} \|\alpha - \alpha'\|. \tag{29}$$

Based on Eq. (24), Eq. (25) and (29) we have

$$\|\nabla_\alpha \mathcal{L}_2(w,\alpha) - \nabla_\alpha \mathcal{L}_2(w,\alpha')\| \leq (L_2^{\nabla_\alpha} + L_2^{\nabla_w} L_w^2 + L_2^w L_{\nabla_\alpha w}) \|\alpha - \alpha'\|. \tag{30}$$

Therefore, Lemma 4 is proved.

□

Proof of Theorem 2: We first define the noise term between the stochastic estimate $\nabla_\alpha \mathcal{L}_2^i$ and the true gradient $\nabla_\alpha \mathcal{L}_2$ as:

$$\varepsilon_i = \nabla_\alpha \mathcal{L}_2 - \nabla_\alpha \mathcal{L}_2^i, \tag{31}$$

and the error between the approximated hypergradient $\nabla_\alpha \tilde{\mathcal{L}}_2$ and the exact hypergradient $\nabla_\alpha \mathcal{L}_2$ as:

$$e_m = \nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m) - \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m). \tag{32}$$

We then prove that $\nabla_\alpha \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)$ is an unbiased estimate of $\nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m)$ that:

$$E[\nabla_\alpha \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m) \mid \alpha_m] = \nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m). \tag{33}$$

Based on IFT in Eq.(7), we have

$$\nabla_\alpha \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m) = \frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial \alpha} - \frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial w}\left[\frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial w \partial w}\right]^{-1} \frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial \alpha \partial w}. \tag{34}$$

So that

$$\begin{aligned}
&E\left[\nabla_\alpha \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right] \\
=&E\left[\frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial \alpha} - \frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial w}\left[\frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial w \partial w}\right]^{-1} \frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial \alpha \partial w} \mid \alpha_m\right].
\end{aligned} \tag{35}$$

Based on the linear assumption for $\mathcal{L}_1^j$ in the condition 4 of the Theorem 2, we have $\frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial w \partial w} = \frac{\partial^2 \mathcal{L}_1(w^*(\alpha_m),\alpha_m)}{\partial w \partial w}$, and

$$\begin{aligned}
&E\left[\nabla_\alpha \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right] \\
&= \frac{1}{R}\sum_{i=1}^{R} \frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial \alpha} - \frac{1}{R}\sum_{i=1}^{R} \frac{\partial \mathcal{L}_2^i(w^*(\alpha_m),\alpha_m)}{\partial w} \left[\frac{\partial^2 \mathcal{L}_1(w^*(\alpha_m),\alpha_m)}{\partial w \partial w}\right]^{-1} \frac{1}{J}\sum_{j=1}^{J} \frac{\partial^2 \mathcal{L}_1^j(w^*(\alpha_m),\alpha_m)}{\partial \alpha \partial w} \\
&= \frac{\partial \mathcal{L}_2(w^*(\alpha_m),\alpha_m)}{\partial \alpha} - \frac{\partial \mathcal{L}_2(w^*(\alpha_m),\alpha_m)}{\partial w} \left[\frac{\partial^2 \mathcal{L}_1(w^*(\alpha_m),\alpha_m)}{\partial w \partial w}\right]^{-1} \frac{\partial^2 \mathcal{L}_1(w^*(\alpha_m),\alpha_m)}{\partial \alpha \partial w} \\
&= \nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m).
\end{aligned} \tag{36}$$

Based on the Lemma 4, we know that $\nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m)$ is Lipschitz continuous with $L_{\nabla_\alpha \mathcal{L}_2} = L_2^{\nabla_\alpha} + L_2^{\nabla_w} L_w^2 + L_2^w L_{\nabla_\alpha w}$. Based on Lipschitz condition, we have

$$\begin{aligned}
&E\left[\mathcal{L}_2(w^*(\alpha_{m+1}),\alpha_{m+1}) \mid \alpha_m\right] \leq E\left[\mathcal{L}_2(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right] \\
&+ E\left[\langle \nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m), \alpha_{m+1} - \alpha_m\rangle \mid \alpha_m\right] + \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} E\left[\|\alpha_{m+1} - \alpha_m\|^2\right] \\
&= \mathcal{L}_2(w^*(\alpha_m),\alpha_m) + \left\langle E\left[\nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m)\right], -\gamma_{\alpha_m} E\left[\nabla_\alpha \mathcal{L}_2^{i'}(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right]\right\rangle \\
&+ \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}^2 E\left[\left\|\nabla_\alpha \mathcal{L}_2^{i'}(w^*(\alpha_m),\alpha_m)\right\|^2\right].
\end{aligned} \tag{37}$$

From our definitions, we have

$$\begin{aligned}
E\left[\nabla_\alpha \mathcal{L}_2(w^*(\alpha_m),\alpha_m)\right] &= E\left[\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m) + e_m\right] = E\left[\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right] + E\left[e_m\right], \\
E\left[\nabla_\alpha \hat{\mathcal{L}}_2^i(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right] &= E\left[\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m) - \varepsilon_m \mid \alpha_m\right] = E\left[\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right], \\
E\left[\left\|\nabla_\alpha \hat{\mathcal{L}}_2^i(w^j(\alpha_m),\alpha_m)\right\|^2 \mid \alpha_m\right] &= E\left[\left\|\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m) - \varepsilon_m\right\|^2\right] = E\left[\left\|\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right\|^2\right] + E\left[\|\varepsilon_m\|^2\right],
\end{aligned} \tag{38}$$

since $E(\varepsilon_m) = 0$. In this way, we have

$$\begin{aligned}
E\left[\mathcal{L}_2(w^*(\alpha_{m+1}),\alpha_{m+1}) \mid \alpha_m\right] \leq\ & E\left[\mathcal{L}_2(w^*(\alpha_m),\alpha_m) \mid \alpha_m\right] - \gamma_{\alpha_m} E\left[\left\|\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right\|^2\right] \\
&- \gamma_{\alpha_m} E\left\langle e_m, \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right\rangle + \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}^2 E\left[\left\|\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right\|^2\right] \\
&+ \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}^2 E\left[\|\varepsilon_m\|^2\right].
\end{aligned} \tag{39}$$

Based on Theorem 1, we have $\|e_m\| \leqslant C_{\mathcal{L}_1^{w\alpha}}\ C_{\mathcal{L}_2^w}\ \frac{1}{\mu}(1-\gamma\mu)^{K+1}$. In this way, for all $\nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)$, we have

$$\begin{aligned}
\left\langle e_m, \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m),\alpha_m)\right\rangle &\geq -C_{\mathcal{L}_1^{w\alpha}}\ C_{\mathcal{L}_2^w}\ \frac{1}{\mu}(1-\gamma\mu)^{K+1} \left\|\nabla_\alpha \tilde{\mathcal{L}}_2\right\| \\
&= -\frac{C_{\mathcal{L}_1^{w\alpha}}\ C_{\mathcal{L}_2^w}\ (1-\gamma\mu)^{K+1}}{\mu \left\|\nabla_\alpha \tilde{\mathcal{L}}_2\right\|} \left\|\nabla_\alpha \tilde{\mathcal{L}}_2\right\|^2 \\
&= -P\left\|\nabla_\alpha \tilde{\mathcal{L}}_2\right\|^2,
\end{aligned} \tag{40}$$

where $P = \frac{C_{\mathcal{L}_1^{w\alpha}} C_{\mathcal{L}_2^w} (1-\gamma\mu)^{K+1}}{\mu \left\| \nabla_\alpha \tilde{\mathcal{L}}_2 \right\|}$. In this way, we have:

$$\begin{aligned} E\left[\mathcal{L}_2(w^*(\alpha_{m+1}), \alpha_{m+1})\right] &\leq E\left[\mathcal{L}_2(w^*(\alpha_m), \alpha_m)\right] - \gamma_{\alpha_m}(1-P) E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2 \right\|^2\right] \\ &+ \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}^2 (1+D) E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2 \right\|^2\right] \\ &\leq E\left[\mathcal{L}_2(w^*(\alpha_m), \alpha_m)\right] - \gamma_{\alpha_m}[(1-P) - \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}(1+D)] E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2 \right\|^2\right]. \end{aligned} \tag{41}$$

If we choose γ_{α_m} to make $(1-P) - \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}(1+D) > 0$, we have $\gamma_{\alpha_m} < \frac{(1-P)}{\frac{L_{\nabla_\alpha \mathcal{L}_2}}{2}(1+D)}$. In addition, since the learning rate should be positive, we should make that $1 - P > 0$, which could be reached by choose appropriate γ and K that $\frac{C_{\mathcal{L}_1^{w\alpha}} C_{\mathcal{L}_2^w} (1-\gamma\mu)^{K+1}}{\mu \left\| \nabla_\alpha \tilde{\mathcal{L}}_2 \right\|} < 1$, where $0 < 1 - \gamma\mu \leq 1$. In this way, we could find that $\mathcal{L}_2$ is decreasing with α_m, and we know that with sufficiently large m, $\mathcal{L}_2$ will decrease and converge since $\mathcal{L}_2$ is bounded.

Furthermore, we have:

$$\begin{aligned} &E\left[\mathcal{L}_2(w^*(\alpha_m), \alpha_m)\right] - E\left[\mathcal{L}_2(w^*(\alpha_{m+1}), \alpha_{m+1})\right] \\ \geq &\gamma_{\alpha_m}[(1-P) - \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}(1+D)] E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m), \alpha_m) \right\|^2\right]. \end{aligned} \tag{42}$$

By telescoping sum, we can show that

$$E\left[\mathcal{L}_2(w^*(\alpha_0), \alpha_0)\right] - E\left[\mathcal{L}_2(w^*(\alpha_m), \alpha_m)\right] \geq \sum_{k=0}^{K} q_t E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m), \alpha_m) \right\|^2\right], \tag{43}$$

where $q_t = \gamma_{\alpha_m}[(1-P) - \frac{L_{\nabla_\alpha \mathcal{L}_2}}{2} \gamma_{\alpha_m}(1+D)] > 0$. Since $\mathcal{L}_2$ is bounded, we have $\sum_{k=0}^{K}_{K\to\infty} q_t E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m), \alpha_m) \right\|^2\right] < \infty$. In addition, based on condition 3, we have $\sum_{k=0}^{K}_{K\to\infty} q_t = \infty$, which imply that $\lim_{k\to\infty} E\left[\left\| \nabla_\alpha \tilde{\mathcal{L}}_2(w^*(\alpha_m), \alpha_m) \right\|\right] = 0$, so as $\lim_{m\to\infty} E\left[\left\| \nabla_\alpha \hat{\mathcal{L}}_2^i(w^j(\alpha_m), \alpha_m) \right\|\right] = 0$.

□

B. Practical implementation of hypergradient

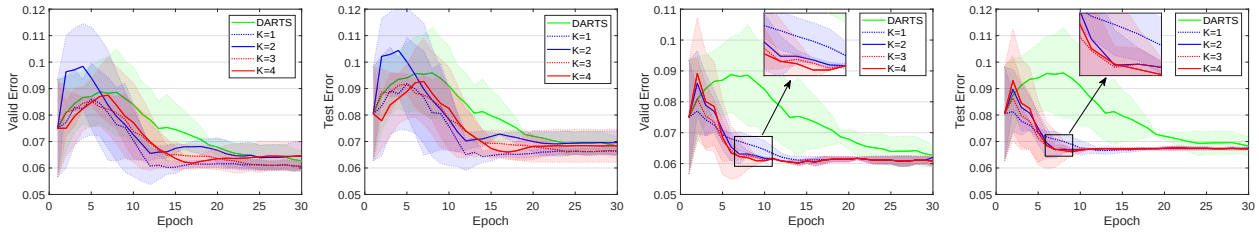
As described, our iDARTS is built based on the DARTS framework with reformulation the hypergradient calculation as:

$$\nabla_\alpha \tilde{\mathcal{L}}_2 = \frac{\partial \mathcal{L}_2}{\partial \alpha} - \gamma \frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^k \frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}. \tag{44}$$

where the different part is the $\left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^k$. As known, it is costly to calculate the Hessian matrix $\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}$ and $\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$ for a large neural network, and we propose two approximations to reduce the computational cost for pracyical implementation, as described below.

Approximation 1: Although it is hard to directly calculate the Hessian matrix $\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}$, we could consider Hessian-vector product technique with autograd to calculate $\frac{\partial \mathcal{L}_2}{\partial w} \cdot \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}$. In this way, we can calculate $\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^k$ step by step:

$$\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^k = \sum_{k=0}^{K} \frac{\partial \mathcal{L}_2}{\partial w} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^k = \sum_{k=0}^{K} V_0 \left[I - \gamma H\right]^k = V_0 + V_1 + V_2 + ... + V_k. \tag{45}$$



(a) Validation error with T=1 (b) Test errors with T=1 (c) Validation error with T=5 (d) Test errors with T=5

Figure 4. Ablation study on K for iDARTS with $T = 1$ and $T = 5$ on NAS-Bench-1Shot1.

where we define $V_0 = \frac{\partial \mathcal{L}_2}{\partial w}$, $H = \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}$, and $V_1 = V_0(I - H), V_2 = V_1(I - H), ..., V_K = V_{K-1}(I - H)$. We can find that, $\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \right]^k$ could be calculated with K steps of Hessian-vector product.

Approximation 2: Apart from the Hessian matrix $\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}$, it is also costly to calculate $\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$ for large neural networks, and we follow DARTS to use the Taylor expansion to approximate $\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$. After calculating $\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \right]^k$, considering the function $\frac{\partial \mathcal{L}_1(w,\alpha)}{\partial \alpha}$ with Taylor expansion, we have

$$\begin{aligned} \frac{\partial \mathcal{L}_1(w + \epsilon A, \alpha)}{\partial \alpha} &= \frac{\partial \mathcal{L}_1(w, \alpha)}{\partial \alpha} + \frac{\partial^2 \mathcal{L}_1(, \alpha)}{\partial \alpha \partial w} \epsilon A + ..., \\ \frac{\partial \mathcal{L}_1(w - \epsilon A, \alpha)}{\partial \alpha} &= \frac{\partial \mathcal{L}_1(w, \alpha)}{\partial \alpha} - \frac{\partial^2 \mathcal{L}_1(w, \alpha)}{\partial \alpha \partial w} \epsilon A + ..., \end{aligned} \tag{46}$$

where ϵ is a very small scalar. When we replace A with $\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \right]^k$, we have

$$\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \right]^k \frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w} = \frac{\frac{\partial \mathcal{L}_1(w + \epsilon A, \alpha)}{\partial \alpha} - \frac{\partial \mathcal{L}_1(w - \epsilon A, \alpha)}{\partial \alpha}}{2\epsilon}. \tag{47}$$

As described, the proposed approximated hypergradient $\nabla_\alpha \tilde{\mathcal{L}}_2$ is easy to implement based on the DARTS framework with only replacing $\frac{\partial \mathcal{L}_2}{\partial w}$ to $\frac{\partial \mathcal{L}_2}{\partial w} \sum_{k=0}^{K} \left[I - \gamma \frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} \right]^k$, which could be computed using the the Hessian-vector product technique.

Therefor, we can practically implement our approximated hypergradient $\nabla_\alpha \tilde{\mathcal{L}}_2$, so as the stochastic approximated hypergradient $\nabla_\alpha \hat{\mathcal{L}}_2^i(w^j(\alpha), \alpha)$ with minibatches based on the DARTS framework [2].

C. Ablation study on the number of approximation terms K

As we described before, there are two additional hyperparameters in our practical iDARTS, the inner optimization steps T and the number of terms for the approximation in Eq.(8). We have analyzed T in previous experiments. In this section, we analyze another hyperparameter K on the NAS-Bench-1Shot1 benchmark dataset. In the first experiment, we set a default hyperparameter $T = 1$ the same as DARTS for the inner supernet training to remove the bias from T. From Eq.(8) and (3), we could further find that the hypergradient calculation in our iDARTS with $T = 1$ and $K = 0$ is the same as DARTS. Figure 4 (a) (b) plots the performance of iDARTS with different K on the NAS-Bench-1Shot1. As shown, our iDARTS is very robust to K with limited training steps $T = 1$, where iDARTS with different K all outperform the DARTS baseline with the same inner training steps $T = 1$, showing the superiority of the proposed approximation over DARTS. Another interesting finding is that, our iDARTS with $K = 1$ and $T = 1$ even achieve slightly more competitive results than $K > 1$. An underlying reason is that, when the inner training step is too small, it is hard to achieve the local optimal w^* and the corresponding hypergradient is not accurate.

[2]The codes and training log files could be found in the supplementary material. The best trained models on CIFAR-10, CIFR-100, and ImageNet could be found https://github.com/MiaoZhang0525/iDARTS.

Table 3. Ablation study on K for iDARTS with on NAS-Bench-201.

Method	CIFAR-10		CIFAR-100		ImageNet-16-120	
	Valid(%)	Test(%)	Valid(%)	Test(%)	Valid(%)	Test(%)
DARTS($T = 1$, $K = 0$)	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
iDARTS($T = 1$, $K = 1$)	86.85±0.93	89.67±1.31	64.09±2.92	64.17±3.26	36.26±5.71	36.11± 5.77
iDARTS($T = 4$, $K = 0$)	87.31±1.33	90.36±1.79	64.76±2.54	64.43±2.47	32.53±1.31	32.42±1.54
iDARTS($T = 4$, $K = 1$)	89.30±1.47	92.44±1.14	67.88±1.86	68.17±2.81	37.11±7.79	36.61±7.47
iDARTS($T = 4$, $K = 2$)	89.86±0.60	93.58±0.32	70.57±0.24	70.83±0.48	40.38±0.59	40.89±0.68
iDARTS($T = 4$, $K = 3$)	89.35±0.03	92.29±0.26	68.51±0.77	68.58±1.18	42.37±0.48	42.26±0.41

To further investigate the effectiveness of the proposed approximation, we consider setting enough inner training steps with $T = 5$, and Figure 4 (c) (d) plots the performance of iDARTS with different K on the NAS-Bench-1Shot1 under $T = 5$. The first impression from Figure 4 is that increasing inner training steps could significantly improve the performance, where all cases with $T = 5$ generally outperform $T = 1$. Another interesting finding is that, with enough inner training steps, the number of approximation terms K has a positive impact on the performance of iDARTS. As shown in Figure 4 (c) (d), increasing K also helps iDARTS converge to excellent solutions faster, verifying that the proposed $\nabla_\alpha \hat{\mathcal{L}}_2^i$ could asymptotically approach to the exact hypergradient $\nabla_\alpha \mathcal{L}_2^i$ with the increase of approximation term K. Besides, we can find that, $K = 2$ is large enough to result in competitive performance for our iDARTS on NAS-Bench-1shot1, which results in similar performance as $K \geq 3$.

We also conduct an ablation study on NAS-Bench-201 dataset to analyse the hyperparameter K, and Table 3 summarizes the performance of iDARTS on NAS-Bench-201 with a different number of approximation term K. The results in Table 3 are similar to those on the NAS-Bench-1Shot1 dataset, also showing that K has a positive impact on the performance of iDARTS. Firstly, we can find that, with the same inner training steps $T = 1$ as DARTS baseline, our iDARTS ($T = 1$, $K = 1$) with one approximation term outperform DARTS by large margins in this case, verifying the superiority of the proposed approximation over DARTS. Secondly, the results in Table 3 also demonstrate that considering more approximation terms does indeed help improve our iDARTS to a certain degree. With enough inner training steps, the performance of iDARTS increases with K from 0 to 2. Another interesting finding is that the performance of iDARTS does not always increase with the K, and there is a decrease for $K \geq 3$. One underlying reason may be that, the iDARTS with smaller K brings more noises into the hypergradient, which in turn enhances the exploration. Several recent works (Chen & Hsieh, 2020; Zhang et al., 2020a) show the importance of the exploration in the differentiable NAS, where adding more noises into the hypergradient could improve the performance. Our experimental results suggest that a $K = 2$ achieves an excellent trade-off between the accuracy of hypergradient and the exploration, thus achieving the competitive performance on the NAS-Bench-201 dataset.

D. Experimental settings in all experiments

In the first experimental set, we choose the third search space of NAS-Bench-1Shot1 (Zela et al., 2020b) to analyze iDARTS, since it is much more complicated than the remaining two search spaces and is a better case to identify the advantages of iDARTS. In Section 5.1, we analyzed the hyperparameter T for our iDARTS and compared it with baseline on the NAS-Bench-1Shot1, and we set another hyperparameter $K = 3$ in all cases. In Appendix C, we further conduct the ablation study to investigate another important hyperparameter K, where we consider two cases with $T = 1$ and $T = 5$, and the remaining experimental settings are the same as the default settings.

In the second experimental set, we choose the NAS-Bench-201 dataset (Dong & Yang, 2020) to analyze differentiable NAS methods. In Section 5.2, we first conduct a comparison experiment with several NAS baselines, and the hyperparameters for our iDARTS in this experiments are $T = 4$, $K = 2$, and γ=0.01. Then we conduct a series ablation studies to investigate three important hyperparameters, inner supernet training steps T, supernet learning rate γ, and architecture learning rate γ_α. In the experiment for the investigation of T, see Figure 2 (a), we set $K = 1$ and other hyperparameters are default settings. In the Figure 2 (b) and (c), we set $T = 4$ and $K = 1$ to investigate both the supernet learning rate γ and architecture learning rate γ_α. In Appendix C, we also analyze the impact of K in iDARTS on NAS-Bench-201 dataset, where we set $T = 4$ and γ=0.01, and the remaining settings are the default.

In the common DARTS search space, we follow the experimental settings in (Liu et al., 2019) to compare with the state-of-the-art NAS methods. We search for micro-cell structures on CIFAR-10 to stack more cells to form the final structure for

Table 4. An overview of different hypergradient approximations.

Method	Steps	Memory Cost	Hypergradient Calculation
Exact IFT hypergradient	∞	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \frac{\partial \mathcal{L}_2}{\partial w}\left[\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^{-1}\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$
DARTS (Liu et al., 2019)	1	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \gamma\frac{\partial \mathcal{L}_2}{\partial w}\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$
$T_1 - T_2$ (Luketina et al., 2016)	1	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \frac{\partial \mathcal{L}_2}{\partial w}[I]^{-1}\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$
Reverse-mode (Franceschi et al., 2017)	T	$\mathcal{O}((P+H)T)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} + \frac{\partial \mathcal{L}_2}{\partial w_T}(\sum_{t=0}^{T} B_t A_{t+1}...A_T)$
Truncated Reverse-mode (Shaban et al., 2019)	K	$\mathcal{O}((P+H)K)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} + \frac{\partial \mathcal{L}_2}{\partial w_T}(\sum_{t=T-K}^{T} B_t A_{t+1}...A_T)$
Neumann Series (Bengio, 2000)	∞	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \gamma\frac{\partial \mathcal{L}_2}{\partial w}\sum_{j=0}^{\infty}\left[I - \gamma\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$
Conjugate Gradient (Rajeswaran et al., 2019)	S	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \left(\text{argmin}_x \left\|x\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w} - \frac{\partial \mathcal{L}_2}{\partial w}\right\|\right)\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$
Our Neumann approximation $\nabla_\alpha \hat{\mathcal{L}}_2^i$	K	$\mathcal{O}(P+H)$	$\frac{\partial \mathcal{L}_2}{\partial \alpha} - \gamma\frac{\partial \mathcal{L}_2}{\partial w}\sum_{j=0}^{K}\left[I - \gamma\frac{\partial^2 \mathcal{L}_1}{\partial w \partial w}\right]^j\frac{\partial^2 \mathcal{L}_1}{\partial \alpha \partial w}$

architecture evaluation. There are two types of cells with the unified search space: a normal cell α_{normal} and a reduction cell α_{reduce}. Cell structures are repeatedly stacked to form the final CNN structure. There are only two reduction cells in the final CNN structure, located in the 1/3 and 2/3 depths of the network. The best architecture searched by our iDARTS on the DARTS search space is obtained with $T = 4$ and $K = 2$. In CIFAR-10, we stack 20 cells to form the final structure for training. The batch size is set as 96, and the number of initial filters is 36. We then transfer the best-searched cells to CIFAR-100 and ImageNet to evaluate the transferability. The experiment setting for the evaluation in CIFAR-100 is the same as CIFAR-10. In the ImageNet dataset, the experiment setting is slightly different from CIFAR-10 in that only 14 cells are stacked, and the number of initial channels is changed to 48, and the batch size is set as 128. We use a linear learning rate scheduler and also following PDART (Chen et al., 2019a) and PCDARTS (Xu et al., 2020) to use a smaller slope in the last five epochs for the architecture evaluation on the ImageNet.

E. Comparison of methods to approximate the hypergradient

We compare different hypergradient approximations in Table 4, which summarizes the computational complexity and memory cost for each method. Under the assumption that Hessian vector products are computed with the *autograd*, we know that the compute time and memory cost for computing a Hessian vector product are with a constant factor of the compute time and memory used for computing a single derivative $\frac{\partial \mathcal{L}_2}{\partial w}$ (Rajeswaran et al., 2019; Griewank, 1993; Griewank & Walther, 2008). We denote that the memory cost for computing the gradient of supernet weight w and architecture parameters α are P and H, respectively. We consider each step in the **Steps** means the computational time of computing a Hessian vector product. The Conjugate Gradient considers iterative solver (e.g., CG) to calculate the inverse of Hessian, where S is the CG solver optimization steps, and each step contains the computation of Hessian vector product.