Appendices

A Proofs

A.1 Proof of Theorem 3.1

Proof. The integral form of Taylor's theorem gives

$$\begin{aligned}\boldsymbol{\theta}(k\eta+\eta)-\boldsymbol{\theta}(k\eta) =&\eta\dot{\boldsymbol{\theta}}(k\eta)+\eta^2\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)\\ =&-\eta\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\eta^2\boldsymbol{\xi}(\boldsymbol{\theta}(k\eta))+\eta^2\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)\,. \end{aligned} \tag{18}$$

Remember the definition of the discrete gradient descent:

$$\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k=-\eta\boldsymbol{g}(\boldsymbol{\theta}_k)\,. \tag{19}$$

Subtracting Equation (19) from Equation (18), we have

$$\boldsymbol{e}_{k+1}-\boldsymbol{e}_k=-\eta(\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\boldsymbol{g}(\boldsymbol{\theta}_k))-\eta^2\boldsymbol{\xi}(\boldsymbol{\theta}(k\eta))+\eta^2\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s) \tag{20}$$

$$=-\eta(\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\boldsymbol{g}(\boldsymbol{\theta}(k\eta)-\boldsymbol{e}_k))-\eta^2\boldsymbol{\xi}(\boldsymbol{\theta}(k\eta))+\eta^2\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)\,. \tag{21}$$

□

A.2 Proof of Theorem 3.2

Proof. The proof is by induction. For $k=0$, $\boldsymbol{e}_0=O(\eta^\gamma)$ by assumption. If $\boldsymbol{e}_k=O(\eta^\gamma)$ for $k\geq 1$, Theorem 3.1 gives

$$\boldsymbol{e}_{k+1}=\boldsymbol{e}_k-\eta(\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\boldsymbol{g}(\boldsymbol{\theta}(k\eta)-\boldsymbol{e}_k))+\boldsymbol{\Lambda}(\boldsymbol{\theta}(k\eta))=O(\eta^\gamma)+O(\eta^{\gamma+1})+O(\eta^\gamma)=O(\eta^\gamma)\,. \tag{22}$$

$\eta(\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\boldsymbol{g}(\boldsymbol{\theta}(k\eta)-\boldsymbol{e}_k))=O(\eta^{\gamma+1})$ follows from Taylor's expansion of $\boldsymbol{g}(\boldsymbol{\theta}(k\eta)-\boldsymbol{e}_k)$ around $\boldsymbol{\theta}(k\eta)$ and from assumption $\boldsymbol{e}_k=O(\eta^\gamma)$:

$$-\eta(\boldsymbol{g}(\boldsymbol{\theta}(k\eta))-\boldsymbol{g}(\boldsymbol{\theta}(k\eta)-\boldsymbol{e}_k))=\eta(\boldsymbol{e}_k\cdot\nabla\boldsymbol{g}(\boldsymbol{\theta}(k\eta))+O(||\boldsymbol{e}_k||^2))=O(\eta^{\gamma+1})\,. \tag{23}$$

□

A.3 Proof of Theorem 3.3

Proof. The proof of Theorem 3.3 consists of the following three Lemmas, all of which are proved in the following sections.

Lemma A.1.

$$\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)=\sum_{n=0}^{\infty}\frac{\eta^n}{(n+2)!}\frac{d^{n+2}}{dt^{n+2}}\boldsymbol{\theta}(k\eta) \tag{24}$$

Lemma A.2. *For $n\geq 1$,*

$$\frac{d^n}{dt^n}\boldsymbol{\theta}(t)=(-1)^n\sum_{k_1,\cdots,k_n=0}^{\infty}\eta^{k_1+\cdots k_n}\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{n-1}}\Xi_{k_n}\,, \tag{25}$$

where $\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{n-1}}:=1$ for $n=1$.

Lemma A.3.

$$\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)=\sum_{j=0}^{\infty}\sum_{i=2}^{j+2}\sum_{k_1+\cdots+k_i=j-i+2}\frac{(-1)^i}{i!}\eta^j D_{k_1}\cdots D_{k_{i-1}}\Xi_{k_i} \tag{26}$$

Theorem 3.3 follows by comparing both sides of Equation (6) order-by-order with using Equation (26) and the expansion of $\boldsymbol{\xi}$ (7). □

A.3.1 Proof of Lemma A.1

Proof.

$$\begin{aligned}
&\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)\\
=&\frac{1}{\eta^2}\int_{k\eta}^{k\eta+\eta} ds\ddot{\boldsymbol{\theta}}(s)(k\eta+\eta-s) && (27)\\
=&\frac{1}{\eta^2}\int_0^{\eta} ds'[\ddot{\boldsymbol{\theta}}(k\eta)(\eta-s')+\dddot{\boldsymbol{\theta}}(k\eta)(\eta-s')s'+\frac{1}{2!}\ddddot{\boldsymbol{\theta}}(k\eta)(\eta-s')s'^2+\cdots] && (28)\\
=&\sum_{n=0}^{\infty}\frac{\eta^n}{(n+2)!}\frac{d^{n+2}}{dt^{n+2}}\boldsymbol{\theta}(k\eta) && (29)
\end{aligned}$$

From Line (27) to (28), we used $s' := s - k\eta$ and the Taylor expansion of $\ddot{\boldsymbol{\theta}}(k\eta + s')$ around $k\eta$. From Line (28) to (29), we used $\int_0^\eta ds'(\eta - s')s'^n = \frac{\eta^{n+2}}{(n+1)(n+2)}$ for $n \geq 0$. □

A.3.2 Proof of Lemma A.2

Proof. Note that given $\dot{\boldsymbol{\theta}}(t) = -\boldsymbol{g}(\boldsymbol{\theta}(t)) - \eta\boldsymbol{\xi}(\boldsymbol{\theta}(t))$, we have

$$\frac{d}{dt}\left(\frac{d^{n-1}}{dt^{n-1}}\boldsymbol{\theta}(t)\right) = -\mathcal{D}\left(\frac{d^{n-1}}{dt^{n-1}}\boldsymbol{\theta}(t)\right) \quad (n \geq 1)\,, \tag{30}$$

where $d^0\boldsymbol{\theta}/dt^0 := \boldsymbol{\theta}$. Therefore,

$$\frac{d^n}{dt^n}\boldsymbol{\theta}(t) = (-1)^{n-1}\mathcal{D}^{n-1}(-\boldsymbol{g} - \eta\boldsymbol{\xi}) = (-1)^n\mathcal{D}^{n-1}\Xi \quad (n \geq 1)\,. \tag{31}$$

Thus, by definition of $\mathcal{D}$, $\mathcal{D}_\alpha$, and Ξ_α (Theorem 3.3 in Section 3.3), we have

$$\begin{aligned}
\frac{d^n}{dt^n}\boldsymbol{\theta}(t) =&(-1)^n(\sum_{k_1=0}^{\infty}\eta^{k_1}\mathcal{D}_{k_1})\cdots(\sum_{k_{n-1}=0}^{\infty}\eta^{k_{n-1}}\mathcal{D}_{k_{n-1}})\Xi && (32)\\
=&(-1)^n\sum_{k_1,\cdots,k_n=0}^{\infty}\eta^{k_1+\cdots k_n}\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{n-1}}\Xi_{k_n}\,. && (33)
\end{aligned}$$

□

A.3.3 Proof of Lemma A.3

Proof. From Lemma A.1 and A.2, we have

$$\begin{aligned}
&\int_0^1 ds\ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s)\\
=&\sum_{n=0}^{\infty}\frac{\eta^n}{(n+2)!}\frac{d^{n+2}}{dt^{n+2}}\boldsymbol{\theta}(k\eta) && (34)\\
=&\sum_{n=0}^{\infty}\frac{\eta^n}{(n+2)!}(-1)^{n+2}\sum_{k_1,\cdots,k_{n+2}=0}^{\infty}\eta^{k_1+\cdots+k_{n+2}}\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{n+1}}\Xi_{k_{n+2}} && (35)\\
=&\sum_{n=0}^{\infty}\sum_{k_1,\cdots,k_{n+2}=0}^{\infty}\frac{(-1)^n}{(n+2)!}\eta^{n+k_1+\cdots+k_{n+2}}\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{n+1}}\Xi_{k_{n+2}} && (36)\\
=&\sum_{j=0}^{\infty}\sum_{i=2}^{j+2}\sum_{k_1+\cdots+k_i=j-i+2}\frac{(-1)^{i-2}}{i!}\eta^j\mathcal{D}_{k_1}\cdots\mathcal{D}_{k_{i-1}}\Xi_{k_i}\,. && (37)
\end{aligned}$$

On the last line, we replaced $n + 2$ and $n + k_1 + \cdots + k_{n+2}$ with i and j, respectively. □

A.4 Proof of Corollary 4.1

Proof. By assumption, we use

$$\boldsymbol{\xi}(\boldsymbol{\theta}) = \eta^2 \sum_{\alpha=0}^{\gamma-1} \eta^\alpha \tilde{\boldsymbol{\xi}}_\alpha \,. \tag{38}$$

From Theorem 3.3, we have

$$\begin{aligned}
\boldsymbol{\Lambda}(\boldsymbol{\theta}) &= \eta^2 \int_0^1 ds \ddot{\boldsymbol{\theta}}(\eta(k+s))(1-s) - \eta^2 \boldsymbol{\xi}(\boldsymbol{\theta}(k\eta)) & (39)\\
&= \eta^2 \sum_{\alpha=0}^{\infty} \eta^\alpha \tilde{\boldsymbol{\xi}}_\alpha - \eta^2 \sum_{\alpha=0}^{\gamma-1} \eta^\alpha \tilde{\boldsymbol{\xi}}_\alpha & (40)\\
&= \eta^2 \sum_{\alpha=\gamma}^{\infty} \eta^\alpha \tilde{\boldsymbol{\xi}}_\alpha & (41)\\
&= \eta^{\gamma+2} \tilde{\boldsymbol{\xi}}_\gamma + O(\eta^{\gamma+3}) \,. & (42)
\end{aligned}$$

Therefore, Theorem 3.2 gives

$$\begin{aligned}
\boldsymbol{e}_{k+1} =& \boldsymbol{e}_k + \boldsymbol{\Lambda}(\boldsymbol{\theta}(k\eta)) + O(\eta^{\gamma+3}) & (43)\\
=& \boldsymbol{e}_k + \eta^{\gamma+2} \tilde{\boldsymbol{\xi}}_\gamma + O(\eta^{\gamma+3}) + O(\eta^{\gamma+3}) & (44)\\
=& \boldsymbol{e}_k + \eta^{\gamma+2} \tilde{\boldsymbol{\xi}}_\gamma + O(\eta^{\gamma+3}) \,. & (45)
\end{aligned}$$

□

A.5 Proof of Corollary 4.2

Proof. From Equation (12), we have

$$\boldsymbol{e}_k = \boldsymbol{e}_0 + \sum_{s=0}^{k-1} \frac{\eta^2}{2} (H(\boldsymbol{\theta}(s\eta) + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + O(\eta^3) \,. \tag{46}$$

Because $\boldsymbol{e}_0 = O(\eta^3)$ by assumption, we have

$$\begin{aligned}
\boldsymbol{e}_k &= \sum_{s=0}^{k-1} \frac{\eta^2}{2} (H(\boldsymbol{\theta}(s\eta) + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + O(\eta^3) & (47)\\
\therefore \ ||\boldsymbol{e}_k|| &\leq \frac{\eta^2}{2} \sum_{s=0}^{k-1} ||(H(\boldsymbol{\theta}(s\eta) + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(s\eta))|| + O(\eta^3) & (48)\\
&\leq \frac{\eta^2 k}{2} \max_{0 \leq s \leq k-1} \{||(H(\boldsymbol{\theta}(s\eta) + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(s\eta))||\} + O(\eta^3) \,. & (49)
\end{aligned}$$

Let $t > 0$ be a given arbitrary number. Then, for $k \in \{1, 2, ..., \lfloor \frac{t}{\eta} \rfloor\}$,

$$||\boldsymbol{e}_k|| \leq \frac{\eta^2 k}{2} \max_{0 \leq t' \leq t} \{||(H(\boldsymbol{\theta}(t') + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(t'))||\} + O(\eta^3) \,. \tag{50}$$

Therefore, if $\eta < \sqrt{\epsilon/k} \sqrt{2/\max_{0 \leq t' \leq t}\{||(H(\boldsymbol{\theta}(t') + \lambda I) \boldsymbol{g}(\boldsymbol{\theta}(t'))||\}}$, then

$$||\boldsymbol{e}_k|| < \epsilon + O(\epsilon^{3/2}) \,. \tag{51}$$

□

A.6 Proof of Corollary A.1

Corollary A.1 (Learning rate bound when $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$)**.** *Let* $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$ *and assume that* $\boldsymbol{e}_0 = O(\eta^4)$*. Let* ϵ *and* t *be arbitrary positive numbers. If the step size satisfies*

$$\eta < \sqrt[3]{\frac{\epsilon}{k}} \sqrt[3]{\frac{12}{\max\limits_{0 \le t' \le t} \{||4(H(\boldsymbol{\theta}(t')) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(t')) + \boldsymbol{g}(\boldsymbol{\theta}(t'))^\top \nabla H(\boldsymbol{\theta}(t')) \boldsymbol{g}(t')||\}}}\,, \tag{52}$$

for some $k \in \{1, 2, ..., \lfloor \frac{t}{\eta} \rfloor\}$*, then the discretization error can be arbitrarily small:*

$$||\boldsymbol{e}_k|| < \epsilon + O(\epsilon^{\frac{4}{3}})\,. \tag{53}$$

Proof. From Equation (10) and Corollary 4.1 and by assumption, we have

$$\boldsymbol{e}_k = \boldsymbol{e}_0 + \eta^3 \sum_{s=0}^{k-1} \{\frac{1}{2}(\tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta}(s\eta) \cdot \nabla))\boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + \frac{1}{6}(\boldsymbol{g}(\boldsymbol{\theta}(s\eta)) \cdot \nabla)\tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta}(s\eta))\} + O(\eta^4)\,. \tag{54}$$

Because $\boldsymbol{e}_0 = O(\eta^4)$ by assumption, we have

$$\boldsymbol{e}_k = \eta^3 \sum_{s=0}^{k-1} \{\frac{1}{2}(\tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta}(s\eta) \cdot \nabla))\boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + \frac{1}{6}(\boldsymbol{g}(\boldsymbol{\theta}(s\eta)) \cdot \nabla)\tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta}(s\eta))\} + O(\eta^4) \tag{55}$$

$$= \eta^3 \sum_{s=0}^{k-1} \{\frac{1}{3}(H(\boldsymbol{\theta}(s\eta)) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + \frac{1}{12}\boldsymbol{g}^\top(\boldsymbol{\theta}(s\eta)) \nabla H(\boldsymbol{\theta}(s\eta)) \boldsymbol{g}(\boldsymbol{\theta}(s\eta))\} + O(\eta^4)\,. \tag{56}$$

Therefore,

$$||\boldsymbol{e}_k|| \le \eta^3 \sum_{s=0}^{k-1} ||\frac{1}{3}(H(\boldsymbol{\theta}(s\eta)) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + \frac{1}{12}\boldsymbol{g}^\top(\boldsymbol{\theta}(s\eta)) \nabla H(\boldsymbol{\theta}(s\eta)) \boldsymbol{g}(\boldsymbol{\theta}(s\eta))|| + O(\eta^4) \tag{57}$$

$$\le \frac{\eta^3 k}{12} \max_{0 \le s \le k-1} \{||4(H(\boldsymbol{\theta}(s\eta)) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(s\eta)) + \boldsymbol{g}^\top(\boldsymbol{\theta}(s\eta)) \nabla H(\boldsymbol{\theta}(s\eta)) \boldsymbol{g}(\boldsymbol{\theta}(s\eta))||\} + O(\eta^4)\,. \tag{58}$$

Let $t > 0$ be a given arbitrary number. Then, for $k \in \{1, 2, ..., \lfloor \frac{t}{\eta} \rfloor\}$,

$$||\boldsymbol{e}_k|| \le \frac{\eta^3 k}{12} \max_{0 \le t' \le t} \{||4(H(\boldsymbol{\theta}(t')) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(t')) + \boldsymbol{g}^\top(\boldsymbol{\theta}(t')) \nabla H(\boldsymbol{\theta}(t')) \boldsymbol{g}(\boldsymbol{\theta}(t'))||\} + O(\eta^4)\,. \tag{59}$$

Therefore, if

$$\eta < \sqrt[3]{\frac{\epsilon}{k}} \sqrt[3]{\frac{12}{\max\limits_{0 \le t' \le t} \{||4(H(\boldsymbol{\theta}(t')) + \lambda I)^2 \boldsymbol{g}(\boldsymbol{\theta}(t')) + \boldsymbol{g}(\boldsymbol{\theta}(t'))^\top \nabla H(\boldsymbol{\theta}(t')) \boldsymbol{g}(t')||\}}}\,, \tag{60}$$

then $||\boldsymbol{e}_k|| < \epsilon + O(\epsilon^{4/3})$. ☐

A.7 Proof of Theorem 5.1

We use the following Lemmas.

Lemma A.4. *For scale-invariant layers* $\mathcal{A}$*, the following equations hold:*

$$\boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{61}$$

$$H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{62}$$

$$\nabla_{\mathcal{A}^c} \nabla_{\mathcal{A}}^\top f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} = 0\,, \tag{63}$$

where $H_{\mathcal{A}}(\boldsymbol{\theta}) := (\mathbb{1}_{\mathcal{A}} \odot \nabla)(\mathbb{1}_{\mathcal{A}} \odot \nabla)^\top f(\boldsymbol{\theta})$*.*

Proof. Differentiating both sides of $f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) = f(\boldsymbol{\theta})$ with respect to α, we have

$$\boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla_{\mathcal{A}} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) = 0\,, \tag{64}$$

where $\nabla_{\mathcal{A}} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})$ means $(\nabla_{\mathcal{A}} f(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\alpha_{\mathcal{A}}\odot\boldsymbol{\theta}}$. For $\alpha = 1$, we have

$$\boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0\,. \tag{65}$$

Applying ∇, we have

$$(\boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla_{\mathcal{A}})\nabla f(\boldsymbol{\theta}) + \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{66}$$

$$\Longleftrightarrow (\boldsymbol{\theta}_{\mathcal{A}} \cdot \nabla_{\mathcal{A}})(\nabla_{\mathcal{A}} + \nabla_{\mathcal{A}^{\mathsf{c}}}) f(\boldsymbol{\theta}) + \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{67}$$

$$\Longleftrightarrow H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \nabla_{\mathcal{A}^{\mathsf{c}}}\nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0\,. \tag{68}$$

Multiplying by $\mathbb{1}_{\mathcal{A}^{\mathsf{c}}}\odot$, we have

$$\nabla_{\mathcal{A}^{\mathsf{c}}}\nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} = 0\,. \tag{69}$$

Therefore,

$$H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0\,. \tag{70}$$

□

Lemma A.5. *For scale-invariant layers $\mathcal{A}$, the following equations hold:*

$$\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = \frac{1}{r_{\mathcal{A}}} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})\,, \tag{71}$$

where $\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) := (\nabla_{\mathcal{A}} f(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathcal{A}}+\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}}$.

Proof. Note that $f(\boldsymbol{\theta}) = f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) = f(\alpha\boldsymbol{\theta}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})$. Differentiating both sides with respect to $\boldsymbol{\theta}$, we have

$$\nabla f(\boldsymbol{\theta}) \tag{72}$$

$$= \nabla(f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) \tag{73}$$

$$= (\nabla_{\mathcal{A}} + \nabla_{\mathcal{A}^{\mathsf{c}}})(f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) \tag{74}$$

$$= \alpha\nabla_{\mathcal{A}} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \nabla_{\mathcal{A}^{\mathsf{c}}} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\,. \tag{75}$$

For $\alpha = 1/r_{\mathcal{A}}$, we have

$$\nabla f(\boldsymbol{\theta}) = \frac{1}{r_{\mathcal{A}}} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \nabla_{\mathcal{A}^{\mathsf{c}}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})\,. \tag{76}$$

Therefore,

$$\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{A}} \odot \nabla f(\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{A}} \odot (\frac{1}{r_{\mathcal{A}}} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \nabla_{\mathcal{A}^{\mathsf{c}}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})) \tag{77}$$

$$= \frac{1}{r_{\mathcal{A}}} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})\,. \tag{78}$$

□

Lemma A.6. *For scale-invariant layers $\mathcal{A}$, the following equations hold for all $\alpha > 0$:*

$$H(\boldsymbol{\theta}) = \alpha^2 H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \alpha(\nabla_{\mathcal{A}^{\mathsf{c}}}\nabla_{\mathcal{A}}^{\top} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) + H_{\mathcal{A}^{\mathsf{c}}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) \tag{79}$$

$$H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} = \alpha^2 H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} \tag{80}$$

$$H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} = H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}}\,, \tag{81}$$

where $H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) := ((\mathbb{1}_{\mathcal{A}} \odot \nabla)(\mathbb{1}_{\mathcal{A}} \odot \nabla)^{\top} f(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\alpha_{\mathcal{A}}\odot\boldsymbol{\theta}}$.

Proof. Because $\nabla f(\boldsymbol{\theta}) = \alpha \nabla_{\mathcal{A}} f(\alpha \boldsymbol{\theta}_{\mathcal{A}}) + \nabla_{\mathcal{A}^c} f(\boldsymbol{\theta}_{\mathcal{A}^c})$ (Equation 75),

$$\begin{aligned}
H(\boldsymbol{\theta}) &= \nabla \nabla^\top f(\boldsymbol{\theta}) \qquad (82)\\
&= \nabla(\alpha \nabla_{\mathcal{A}}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \nabla_{\mathcal{A}^c}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) \qquad (83)\\
&= (\nabla_{\mathcal{A}} + \nabla_{\mathcal{A}^c})(\alpha \nabla_{\mathcal{A}}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \nabla_{\mathcal{A}^c}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) \qquad (84)\\
&= \alpha^2 H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \alpha(\nabla_{\mathcal{A}^c} \nabla_{\mathcal{A}}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta}) + \nabla_{\mathcal{A}} \nabla_{\mathcal{A}^c}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})) + H_{\mathcal{A}^c}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\,. \qquad (85)
\end{aligned}$$

Therefore,

$$\begin{aligned}
H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} &= \alpha^2 H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \alpha \nabla_{\mathcal{A}^c} \nabla_{\mathcal{A}}^\top f(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} \qquad (86)\\
&= \alpha^2 H_{\mathcal{A}}(\alpha_{\mathcal{A}} \odot \boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}}\,. \qquad (87)
\end{aligned}$$

For $\alpha = 1$, we have

$$H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} = H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}}\,. \qquad (88)$$

□

We now prove Theorem 5.1.

Proof. We use Lemmas A.4, A.5, and A.6.

$$\begin{aligned}
\dot{r_{\mathcal{A}}^2}(t) =& 2\boldsymbol{\theta}_{\mathcal{A}}(t) \cdot \dot{\boldsymbol{\theta}}_{\mathcal{A}}(t) \qquad (89)\\
=& 2\boldsymbol{\theta}_{\mathcal{A}}(t) \cdot (-\nabla_{\mathcal{A}} f(\boldsymbol{\theta}(t)) - \lambda \boldsymbol{\theta}_{\mathcal{A}}(t) - \eta \boldsymbol{\xi}(\boldsymbol{\theta}(t))) \qquad (90)\\
=& -2\lambda r_{\mathcal{A}}^2(t) - 2\eta \boldsymbol{\theta}_{\mathcal{A}}(t) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}(t))\,. \qquad (91)
\end{aligned}$$

For $\boldsymbol{\xi} = \mathbf{0}$,

$$\dot{r_{\mathcal{A}}^2}(t) = -2\lambda r_{\mathcal{A}}^2(t)\,. \qquad (92)$$

For $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$,

$$\begin{aligned}
\dot{r_{\mathcal{A}}^2}(t) =& -2\lambda r_{\mathcal{A}}^2(t) - 2\eta \boldsymbol{\theta}_{\mathcal{A}}(t) \cdot \tilde{\boldsymbol{\xi}}_0(\boldsymbol{\theta}(t)) \qquad (93)\\
=& -2\lambda r_{\mathcal{A}}^2(t) - \eta(\lambda^2 r_{\mathcal{A}}^2(t) - ||\nabla_{\mathcal{A}} f(\boldsymbol{\theta}(t))||^2) \qquad (94)\\
=& -2\lambda(1 + \frac{\eta\lambda}{2}) r_{\mathcal{A}}^2(t) + \frac{\eta}{r_{\mathcal{A}}^2(t)} ||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t))||^2\,. \qquad (95)
\end{aligned}$$

We used

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{A}} \cdot \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} &= \frac{1}{2}\boldsymbol{\theta}_{\mathcal{A}} \cdot (H(\boldsymbol{\theta}) + \lambda I)(\nabla f(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}) \qquad (96)\\
&= \frac{1}{2}\boldsymbol{\theta}_{\mathcal{A}} \cdot (H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda H(\boldsymbol{\theta})\boldsymbol{\theta} + \lambda \nabla f(\boldsymbol{\theta}) + \lambda^2 \boldsymbol{\theta}) \qquad (97)\\
&= \frac{1}{2}(\boldsymbol{\theta}_{\mathcal{A}}^\top H_{\mathcal{A}}(\boldsymbol{\theta}) \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}_{\mathcal{A}}^\top H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \lambda^2 r_{\mathcal{A}}^2) \qquad (98)\\
&= \frac{1}{2}(-||\nabla_{\mathcal{A}} f(\boldsymbol{\theta})||^2 + \lambda^2 r_{\mathcal{A}}^2)\,. \qquad (99)
\end{aligned}$$

Using $\dot{\boldsymbol{x}}(t) = -a\boldsymbol{x} + \boldsymbol{y}(t) \Leftrightarrow \boldsymbol{x}(t) = \boldsymbol{x}(0)e^{-at} + \int_0^t d\tau e^{-a(t-\tau)}\boldsymbol{y}(\tau)$, we can show the remaining equations.

□

A.8 Proof of Corollary 5.1

Proof. When $\boldsymbol{\xi} = \mathbf{0}$, $r_{\mathcal{A}} \xrightarrow{t\to\infty} 0$ is obvious from the EoM for $r_{\mathcal{A}}$ (Theorem 5.1). When $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$, EoM is given by

$$\dot{r^2}_{\mathcal{A}}(t) = -2\lambda(1 + \frac{\eta\lambda}{2}) r_{\mathcal{A}}^2(t) + \frac{\eta}{r^2(t)} ||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) + \boldsymbol{\theta}_{\mathcal{A}^c}(t))||^2\,. \qquad (100)$$

At equilibrium, $\dot{r}_{\mathcal{A}} = 0$ and $||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^c})|| = c_*$ by assumption; thus, we have

$$
\begin{aligned}
0 &= -2\lambda(1 + \frac{\eta\lambda}{2}) r_{\mathcal{A}*}^2 + \frac{\eta}{r_{\mathcal{A}*}^2} c_*^2 \qquad (101)\\
\iff & r_{\mathcal{A}*}^2 = \sqrt{\frac{\eta}{2\lambda + \eta\lambda^2}} c_* \,. \qquad (102)
\end{aligned}
$$

□

A.9 Proof of Theorem C.1

Proof. We use Lemmas A.4, A.5, and A.6:

$$
\begin{aligned}
\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}} &= \frac{d}{dt} \frac{\boldsymbol{\theta}_{\mathcal{A}}}{r_{\mathcal{A}}} \qquad (103)\\
&= -\frac{\dot{r}_{\mathcal{A}}}{r_{\mathcal{A}}^2} \boldsymbol{\theta}_{\mathcal{A}} + \frac{1}{r_{\mathcal{A}}} \dot{\boldsymbol{\theta}}_{\mathcal{A}} \qquad (104)\\
&= \frac{\boldsymbol{\theta}_{\mathcal{A}}}{r_{\mathcal{A}}^2} (\lambda r_{\mathcal{A}} + \eta \hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}(\boldsymbol{\theta})) + \frac{1}{r_{\mathcal{A}}} (-\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}} - \eta \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta})) \qquad (105)\\
&= \frac{\eta}{r_{\mathcal{A}}} \hat{\boldsymbol{\theta}}_{\mathcal{A}} (\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta})) - \frac{1}{r_{\mathcal{A}}} \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \frac{\eta}{r_{\mathcal{A}}} \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}) \qquad (106)\\
&= -\frac{1}{r_{\mathcal{A}}^2} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^c}) + \frac{\eta}{r_{\mathcal{A}}} ((\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta})) \hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta})) \,, \qquad (107)
\end{aligned}
$$

where $\boldsymbol{\xi}_{\mathcal{A}} := \mathbb{1}_{\mathcal{A}} \odot \boldsymbol{\xi}$. We used $\dot{r}_{\mathcal{A}} = -\lambda r_{\mathcal{A}} - \eta \hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}(\boldsymbol{\theta})$ (Theorem 5.1). Note that $\frac{\eta}{r_{\mathcal{A}}}((\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}))\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}))$ has no $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ component; i.e., it is orthogonal to $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$. When $\boldsymbol{\xi} = \mathbf{0}$, Equation (107) is equivalent to $\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}} = -\frac{1}{r_{\mathcal{A}}^2} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}})$. When $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$, note that from Equation (99),

$$
\boldsymbol{\theta}_{\mathcal{A}} \cdot \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} = \frac{1}{2}(-||\nabla_{\mathcal{A}} f(\boldsymbol{\theta})||^2 + \lambda^2 r_{\mathcal{A}}^2) \,. \qquad (108)
$$

Therefore,

$$
(\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \tilde{\boldsymbol{\xi}}_{0\mathcal{A}}) \hat{\boldsymbol{\theta}}_{\mathcal{A}} = -\frac{1}{2} \frac{1}{r_{\mathcal{A}}^3} ||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^c})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}} + \frac{\lambda^2}{2} \boldsymbol{\theta}_{\mathcal{A}} \,. \qquad (109)
$$

Also,

$$
\begin{aligned}
\tilde{\boldsymbol{\xi}}_{0\mathcal{A}} &= \frac{1}{2} \mathbb{1}_{\mathcal{A}} \odot (H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \lambda H(\boldsymbol{\theta}) \boldsymbol{\theta} + \lambda \nabla f(\boldsymbol{\theta}) + \lambda^2 \boldsymbol{\theta}) \qquad (110)\\
&= \frac{1}{2} (\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \lambda \mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})(\boldsymbol{\theta}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^c}) + \lambda \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda^2 \boldsymbol{\theta}_{\mathcal{A}}) \qquad (111)\\
&= \frac{1}{2} (\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \lambda H_{\mathcal{A}}(\boldsymbol{\theta}) \boldsymbol{\theta}_{\mathcal{A}} + \lambda \nabla_{\mathcal{A}} \nabla_{\mathcal{A}^c}^\top f(\boldsymbol{\theta}) \boldsymbol{\theta}_{\mathcal{A}^c} + \lambda \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda^2 \boldsymbol{\theta}_{\mathcal{A}}) \qquad (112)\\
&= \frac{1}{2} (\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \lambda \nabla_{\mathcal{A}} \nabla_{\mathcal{A}^c}^\top f(\boldsymbol{\theta}) \boldsymbol{\theta}_{\mathcal{A}^c} + \lambda^2 \boldsymbol{\theta}_{\mathcal{A}}) \,. \qquad (113)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&(\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \tilde{\boldsymbol{\xi}}_{0\mathcal{A}}(\boldsymbol{\theta}))\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} && (114)\\
=& -\frac{1}{2}\frac{1}{r_{\mathcal{A}}^3}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}} + \frac{\lambda^2}{2}\boldsymbol{\theta}_{\mathcal{A}} - \frac{1}{2}(\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda \nabla_{\mathcal{A}} \nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} + \lambda^2 \boldsymbol{\theta}_{\mathcal{A}}) && (115)\\
=& -\frac{1}{2}\frac{1}{r_{\mathcal{A}}^3}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}} - \frac{1}{2}\nabla_{\mathcal{A}}\nabla^{\top} f(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) - \frac{\lambda}{2}\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} && (116)\\
=& -\frac{1}{2}\frac{1}{r_{\mathcal{A}}^3}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}} - \frac{1}{2}H_{\mathcal{A}}(\boldsymbol{\theta})\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \frac{1}{2}\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\boldsymbol{\theta})\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}) \\
&- \frac{1}{2}\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\boldsymbol{\theta})\lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} && (117)\\
=& -\frac{1}{2}\frac{1}{r_{\mathcal{A}}^3}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}} - \frac{1}{2}\frac{1}{r_{\mathcal{A}}}H_{\mathcal{A}}\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \\
&- \frac{1}{2}\frac{1}{r_{\mathcal{A}}}\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})(\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})\,. && (118)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}} =& -\frac{1}{r_{\mathcal{A}}^2}\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) - \frac{\eta}{2r_{\mathcal{A}}^2}(H_{\mathcal{A}}(\boldsymbol{\theta})\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \\
&+ \nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})(\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta})\lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \frac{1}{r_{\mathcal{A}}^2}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})||^2 \hat{\boldsymbol{\theta}}_{\mathcal{A}}) && (119)\\
=& -\frac{1}{r_{\mathcal{A}}^2}(I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}) \\
&+ \frac{\eta}{2}(\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \cdot \nabla_{\mathcal{A}^{\mathsf{c}}} + \frac{\eta}{2}\frac{1}{r_{\mathcal{A}}^2}\hat{\boldsymbol{\theta}}_{\mathcal{A}}\nabla_{\mathcal{A}}^{\top} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}))\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) && (120)\\
=& -\frac{1}{r_{\mathcal{A}}^2}(I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}) + \frac{\eta}{2}(\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \cdot \nabla_{\mathcal{A}^{\mathsf{c}}} \\
&+ \frac{\eta}{2}\hat{\boldsymbol{\theta}}_{\mathcal{A}}\nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta}))\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})\,. && (121)
\end{aligned}
$$

□

A.10 Proof of Corollary 5.2

Proof. We use Lemmas A.4 and A.5. The angular update is defined as

$$
\cos\Delta(t) = \frac{\boldsymbol{\theta}_{\mathcal{A}}(t)}{r_{\mathcal{A}}(t)} \cdot \frac{\boldsymbol{\theta}_{\mathcal{A}}(t+\eta)}{r_{\mathcal{A}}(t+\eta)}\,. \tag{122}
$$

We evaluate the higher order terms in $\boldsymbol{\theta}_{\mathcal{A}}(t+\eta)$ and $r_{\mathcal{A}}(t+\eta)$. First,

$$
\begin{aligned}
\boldsymbol{\theta}_{\mathcal{A}}(t+\eta) =&\boldsymbol{\theta}_{\mathcal{A}}(t) + \eta\dot{\boldsymbol{\theta}}_{\mathcal{A}}(t) + \frac{\eta^2}{2}\ddot{\boldsymbol{\theta}}_{\mathcal{A}}(t) + O(\eta^3) \\
=&\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta\nabla f(\boldsymbol{\theta}_{\mathcal{A}}(t)) - \eta\lambda\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta^2\boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}(t)) + \frac{\eta^2}{2}\ddot{\boldsymbol{\theta}}_{\mathcal{A}}(t) + O(\eta^3)\,. && (123)
\end{aligned}
$$

The second derivative $\ddot{\boldsymbol{\theta}}(t)$ is given by

$$
\begin{aligned}
\ddot{\boldsymbol{\theta}}_{\mathcal{A}} &= \frac{d}{dt}\dot{\boldsymbol{\theta}}_{\mathcal{A}} && (124)\\
&= \frac{d}{dt}(-\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda\boldsymbol{\theta}_{\mathcal{A}}) + O(\eta) && (125)\\
&= -(\dot{\boldsymbol{\theta}} \cdot \nabla)\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda\dot{\boldsymbol{\theta}}_{\mathcal{A}} + O(\eta) && (126)\\
&= \nabla_{\mathcal{A}}\nabla^{\top} f(\boldsymbol{\theta})(\nabla f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}) + \lambda(\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}_{\mathcal{A}}) + O(\eta) && (127)\\
&= \mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} + \lambda\nabla_{\mathcal{A}}\nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} + \lambda\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda^2\boldsymbol{\theta}_{\mathcal{A}} + O(\eta) && (128)\\
&= \mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^{\mathsf{c}}}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} + \lambda^2\boldsymbol{\theta}_{\mathcal{A}} + O(\eta)\,. && (129)
\end{aligned}
$$

Therefore,

$$\begin{aligned}\boldsymbol{\theta}_{\mathcal{A}}(t+\eta) = \boldsymbol{\theta}_{\mathcal{A}}(t) - \eta\nabla_{\mathcal{A}} f(\boldsymbol{\theta}(t)) - \eta\lambda\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta^2\boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}(t)) \\ + \frac{\eta^2}{2}(\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}(t))\nabla f(\boldsymbol{\theta}(t)) + \lambda\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^c}^{\top} f(\boldsymbol{\theta}(t))\boldsymbol{\theta}_{\mathcal{A}^c}(t) + \lambda^2\boldsymbol{\theta}_{\mathcal{A}}(t)) + O(\eta^3)\,.\end{aligned} \tag{130}$$

Next,

$$r_{\mathcal{A}}(t+\eta) = r_{\mathcal{A}}(t) + \dot{r}_{\mathcal{A}}(t)\eta + \frac{\eta^2}{2}\ddot{r}_{\mathcal{A}}(t) + O(\eta^3)\,. \tag{131}$$

Because $\dot{r}_{\mathcal{A}} = -\lambda r_{\mathcal{A}} - \eta\hat{\boldsymbol{\theta}}_{\mathcal{A}} \cdot \boldsymbol{\xi}$ (use Equation (91) and $\dot{r^2}_{\mathcal{A}} = 2r_{\mathcal{A}}\dot{r}_{\mathcal{A}}$),

$$r_{\mathcal{A}}(t+\eta) = r_{\mathcal{A}}(t) - \eta\lambda r_{\mathcal{A}}(t) - \eta^2\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) \cdot \boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}(t)) + \frac{\eta^2}{2}\ddot{r}_{\mathcal{A}}(t) + O(\eta^3)\,. \tag{132}$$

In addition, because $\ddot{r}_{\mathcal{A}} = -\lambda\dot{r}_{\mathcal{A}} + O(\eta) = \lambda^2 r_{\mathcal{A}} + O(\eta)$,

$$r_{\mathcal{A}}(t+\eta) = r_{\mathcal{A}}(t) - \eta\lambda r_{\mathcal{A}}(t) - \eta^2\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}(t)) + \frac{\eta^2}{2}\lambda^2 r_{\mathcal{A}}(t) + O(\eta^3)\,. \tag{133}$$

Therefore,

$$\cos\Delta(t) = \frac{\boldsymbol{\theta}_{\mathcal{A}}(t)}{r_{\mathcal{A}}(t)} \cdot \frac{\boldsymbol{\theta}_{\mathcal{A}}(t+\eta)}{r_{\mathcal{A}}(t+\eta)} \tag{134}$$

$$\begin{aligned}=&\frac{\boldsymbol{\theta}_{\mathcal{A}}(t)}{r_{\mathcal{A}}(t)} \cdot \Big(\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta\nabla_{\mathcal{A}} f(\boldsymbol{\theta}(t)) - \eta\lambda\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta^2\boldsymbol{\xi}_{\mathcal{A}}(\boldsymbol{\theta}(t)) \\ &+ \frac{\eta^2}{2}(\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}(t))\nabla f(\boldsymbol{\theta}(t)) + \lambda\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^c}^{\top} f(\boldsymbol{\theta}(t))\boldsymbol{\theta}_{\mathcal{A}^c}(t) + \lambda^2\boldsymbol{\theta}_{\mathcal{A}}(t))\Big)\Big/\Big(r_{\mathcal{A}}(t) - \eta\lambda r_{\mathcal{A}}(t) \\ &- \eta^2\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}(t)) + \frac{\eta^2}{2}\lambda^2 r_{\mathcal{A}}(t)\Big) \\ &+ O(\eta^3)\,.\end{aligned} \tag{135}$$

Substituting $\boldsymbol{\xi}_{\mathcal{A}} = \tilde{\boldsymbol{\xi}}_{0\mathcal{A}}$, and using

$$\tilde{\boldsymbol{\xi}}_{0\mathcal{A}} = \frac{1}{2}(\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda\nabla_{\mathcal{A}}\nabla_{\mathcal{A}^c}^{\top} f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^c} + \lambda^2\boldsymbol{\theta}_{\mathcal{A}}) \quad \text{(Equation 113)} \tag{136}$$

$$\boldsymbol{\theta}_{\mathcal{A}} \cdot \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} = \frac{1}{2}(-||\nabla_{\mathcal{A}} f(\boldsymbol{\theta})||^2 + \lambda^2 r_{\mathcal{A}}^2) \quad \text{(Equation 108)}\,, \tag{137}$$

we have

$$\cos\Delta(t) = \frac{\boldsymbol{\theta}_{\mathcal{A}}(t)}{r_{\mathcal{A}}(t)} \cdot \frac{\boldsymbol{\theta}_{\mathcal{A}}(t) - \eta\nabla_{\mathcal{A}} f(\boldsymbol{\theta}(t)) - \eta\lambda\boldsymbol{\theta}_{\mathcal{A}}(t)}{r_{\mathcal{A}}(t) - \eta\lambda r_{\mathcal{A}}(t) - \frac{\eta^2}{2r_{\mathcal{A}}(t)}(-||\nabla_{\mathcal{A}} f(\boldsymbol{\theta})||^2 + \lambda^2 r_{\mathcal{A}}^2(t)) + \frac{\eta^2}{2}\lambda^2 r_{\mathcal{A}}(t)} + O(\eta^3) \tag{138}$$

$$= \frac{(1-\eta\lambda)r_{\mathcal{A}}^2(t)}{(1-\eta\lambda)r_{\mathcal{A}}^2(t) + \frac{\eta^2}{2r_{\mathcal{A}}^2(t)}||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^c})||^2} + O(\eta^3)\,. \tag{139}$$

At equilibrium, we have $r_{\mathcal{A}}^2 \xrightarrow{t\to\infty} r_{\mathcal{A}*}^2 = \sqrt{\frac{\eta}{2\lambda+\eta\lambda^2}c_*}$ and $||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) + \boldsymbol{\theta}_{\mathcal{A}^c}(t))|| \xrightarrow{t\to\infty} c_*$ because of Corollary 5.1. Thus,

$$\cos\Delta_* = \frac{(1-\eta\lambda)r_{\mathcal{A}*}^2}{(1-\eta\lambda)r_{\mathcal{A}*}^2 + \frac{\eta^2}{2r_{\mathcal{A}*}^2}c_*^2} + O(\eta^3) \tag{140}$$

$$= \frac{1-\eta\lambda}{1-\eta^2\lambda^2/2} + O(\eta^3)\,, \tag{141}$$

and we have shown the first statement of the theorem.

The second statement follows from Equation (141). By definition of cosine and tangent, we have

$$\tan\Delta_* = \frac{\sqrt{(1-\eta^2\lambda^2/2)^2-(1-\eta\lambda)^2}}{1-\eta\lambda} + O(\eta^3) = \frac{\sqrt{2\eta\lambda - 2\eta^2\lambda^2 + \eta^4\lambda^4/4}}{1-\eta} + O(\eta^3)\,. \tag{142}$$

Therefore, using Taylor's series of the tangent function, we have

$$\Delta_* = \tan\Delta_* - \frac{1}{3}\Delta_*^3 - \frac{2}{15}\Delta_*^5 - ... = \sqrt{2\eta\lambda} + O((\eta\lambda)^{3/2})\,. \tag{143}$$

This concludes the proof. □

A.11 Proof of Theorem 5.2

We use the following Lemma:

Lemma A.7. *For translation-invariant layers $\mathcal{A}$, the following equations hold:*

$$\boldsymbol{\theta}_{\mathcal{A}\perp}\cdot\boldsymbol{\theta}_{\mathcal{A}\|} = 0 \tag{144}$$

$$\mathbb{1}_{\mathcal{A}}\cdot\nabla f(\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{A}}\cdot\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{145}$$

$$P\nabla f(\boldsymbol{\theta}) = P\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{146}$$

$$\boldsymbol{\theta}_{\mathcal{A}\perp}\cdot\nabla f(\boldsymbol{\theta}) = \boldsymbol{\theta}_{\mathcal{A}\perp}\cdot\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0 \tag{147}$$

$$H(\boldsymbol{\theta})\mathbb{1}_{\mathcal{A}} = 0 \tag{148}$$

$$PH(\boldsymbol{\theta}) = 0 \tag{149}$$

$$H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}\perp} = 0 \tag{150}$$

$$\nabla f(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^c}) \tag{151}$$

$$H(\boldsymbol{\theta}) = H(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^c})\,. \tag{152}$$

Proof. Note that $P^\top = P$, $P^2 = P$, and thus, $P^\top(I-P) = P(I-P) = P - P = 0$. Therefore,

$$\boldsymbol{\theta}_{\mathcal{A}\perp}\cdot\boldsymbol{\theta}_{\mathcal{A}\|} = \boldsymbol{\theta}_{\mathcal{A}}^\top P^\top (I-P)\boldsymbol{\theta}_{\mathcal{A}} = 0\,. \tag{153}$$

Next, differentiating $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta} + \alpha\mathbb{1}_{\mathcal{A}})$ with respect to α, we have

$$\mathbb{1}_{\mathcal{A}}\cdot\nabla f(\boldsymbol{\theta}+\alpha\mathbb{1}_{\mathcal{A}}) = 0\,. \tag{154}$$

For $\alpha = 0$, we have

$$\mathbb{1}_{\mathcal{A}}\cdot\nabla f(\boldsymbol{\theta}) = \mathbb{1}_{\mathcal{A}}\cdot\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0\,. \tag{155}$$

Therefore,

$$P\nabla f(\boldsymbol{\theta}) = P\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = (\mathbb{1}_{\mathcal{A}}\cdot\nabla f(\boldsymbol{\theta}))\frac{1}{d_{\mathcal{A}}}\mathbb{1}_{\mathcal{A}} = 0 \tag{156}$$

and

$$\boldsymbol{\theta}_{\mathcal{A}\perp}\cdot\nabla f(\boldsymbol{\theta}) = \frac{\mathbb{1}_{\mathcal{A}}\cdot\boldsymbol{\theta}_{\mathcal{A}}}{d_{\mathcal{A}}}\mathbb{1}_{\mathcal{A}}\cdot\nabla f(\boldsymbol{\theta}) = \frac{\mathbb{1}_{\mathcal{A}}\cdot\boldsymbol{\theta}_{\mathcal{A}}}{d_{\mathcal{A}}}\mathbb{1}_{\mathcal{A}}\cdot\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = 0\,. \tag{157}$$

Next, differentiating Equation 155 with respect to $\boldsymbol{\theta}$, we have

$$H(\boldsymbol{\theta})\mathbb{1}_{\mathcal{A}} = 0\,. \tag{158}$$

Therefore,

$$PH(\boldsymbol{\theta}) = \frac{\mathbb{1}_{\mathcal{A}}}{d_{\mathcal{A}}}\mathbb{1}_{\mathcal{A}}^\top H(\boldsymbol{\theta}) = 0 \tag{159}$$

and

$$H(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}\perp} = \frac{\mathbb{1}_{\mathcal{A}}\cdot\boldsymbol{\theta}_{\mathcal{A}}}{d_{\mathcal{A}}}H(\boldsymbol{\theta})\mathbb{1}_{\mathcal{A}} = 0\,. \tag{160}$$

Next, differentiating $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta} + \alpha \mathbb{1}_{\mathcal{A}})$ with respect to $\boldsymbol{\theta}$, we have

$$\nabla f(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta} + \alpha \mathbb{1}_{\mathcal{A}}) \tag{161}$$

and

$$H(\boldsymbol{\theta}) = H(\boldsymbol{\theta} + \alpha \mathbb{1}_{\mathcal{A}}) \,. \tag{162}$$

For $\alpha = -\frac{\mathbb{1}_{\mathcal{A}} \cdot \boldsymbol{\theta}_{\mathcal{A}}}{d_{\mathcal{A}}}$, we have

$$\nabla f(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta} - P\boldsymbol{\theta}_{\mathcal{A}}) = \nabla f(\boldsymbol{\theta}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}} - P\boldsymbol{\theta}_{\mathcal{A}}) = \nabla f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \tag{163}$$

and

$$H(\boldsymbol{\theta}) = H(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \,. \tag{164}$$

□

We begin the proof of Theorem 5.2.

Proof. We use Lemma A.7.

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\perp} = P\dot{\boldsymbol{\theta}}_{\mathcal{A}} = P(-\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}} - \eta \boldsymbol{\xi}_{\mathcal{A}}) = -\lambda \boldsymbol{\theta}_{\mathcal{A}\perp} - \eta P \boldsymbol{\xi}_{\mathcal{A}} \,. \tag{165}$$

When $\boldsymbol{\theta} = \mathbf{0}$, EoM is

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\perp}(t) = -\lambda \boldsymbol{\theta}_{\mathcal{A}}(t) \,. \tag{166}$$

When $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_{0\mathcal{A}}$, note that

$$\tilde{\boldsymbol{\xi}}_0 = \frac{1}{2}(H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) + \lambda \nabla f(\boldsymbol{\theta}) + \lambda H(\boldsymbol{\theta})\boldsymbol{\theta} + \lambda^2 \boldsymbol{\theta}) \tag{167}$$

and

$$\tilde{\boldsymbol{\xi}}_0 \cdot \mathbb{1}_{\mathcal{A}} = \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} \cdot \mathbb{1}_{\mathcal{A}} = \frac{\lambda^2}{2} \mathbb{1}_{\mathcal{A}} \cdot \boldsymbol{\theta}_{\mathcal{A}} \,. \tag{168}$$

Thus,

$$P\tilde{\boldsymbol{\xi}}_0 = \frac{\lambda^2}{2} \boldsymbol{\theta}_{\mathcal{A}\perp} \,. \tag{169}$$

Therefore,

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\perp} = -\lambda \boldsymbol{\theta}_{\mathcal{A}\perp} - \eta P \tilde{\boldsymbol{\xi}}_{0\mathcal{A}} = -\lambda \boldsymbol{\theta}_{\mathcal{A}\perp} - \eta \frac{\lambda^2}{2} \boldsymbol{\theta}_{\mathcal{A}\perp} = -(\lambda + \frac{\eta\lambda^2}{2}) \boldsymbol{\theta}_{\mathcal{A}\perp} \,. \tag{170}$$

Using $\dot{\boldsymbol{v}}(t) = -a\boldsymbol{v}(t) \Leftrightarrow \boldsymbol{v}(t) = \boldsymbol{v}(0)e^{-at}$, we can show the remaining equations. □

A.12 Proof of Theorem D.1

Proof. We use Lemma A.7. First, note that

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\|} = \dot{\boldsymbol{\theta}}_{\mathcal{A}} - \dot{\boldsymbol{\theta}}_{\mathcal{A}\perp} \,. \tag{171}$$

Because

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}} = -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}} - \eta \boldsymbol{\xi}_{\mathcal{A}} \tag{172}$$

and

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\perp} = -\lambda \boldsymbol{\theta}_{\mathcal{A}\perp} - \eta P \boldsymbol{\xi}_{\mathcal{A}} \,, \tag{173}$$

we have

$$\begin{aligned} \dot{\boldsymbol{\theta}}_{\mathcal{A}\|} = \dot{\boldsymbol{\theta}}_{\mathcal{A}} - \dot{\boldsymbol{\theta}}_{\mathcal{A}\perp} &= -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} - \eta (I - P) \boldsymbol{\xi}_{\mathcal{A}} & (174) \\ &= -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} - \eta (I - P) \boldsymbol{\xi}_{\mathcal{A}} \,. & (175) \end{aligned}$$

Note that $\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}$ is orthogonal to $\boldsymbol{\theta}_{\mathcal{A}\|}$ because $\boldsymbol{\theta}_{\mathcal{A}\perp} \cdot \dot{\boldsymbol{\theta}}_{\mathcal{A}\|} = -\boldsymbol{\theta}_{\mathcal{A}\perp} \cdot \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\perp} \cdot \boldsymbol{\theta}_{\mathcal{A}\|} - \eta \boldsymbol{\theta}_{\mathcal{A}\perp}^\top (I - P)\tilde{\boldsymbol{\xi}}_{0\mathcal{A}} = 0 - 0 - 0 = 0$ (we used $\boldsymbol{\theta}_{\mathcal{A}\perp}^\top (I-P) = \boldsymbol{\theta}_{\mathcal{A}}^\top P^\top (I-P) = \boldsymbol{\theta}_{\mathcal{A}}^\top (P-P) = 0$).

When $\boldsymbol{\xi} = \mathbf{0}$, we have

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\|} = -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} = -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} \,. \tag{176}$$

When $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$, we have

$$\begin{aligned}
\dot{\boldsymbol{\theta}}_{\mathcal{A}\|} &= -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} \\
&\quad - \eta\Big(\frac{1}{2}(\mathbb{1}_{\mathcal{A}} H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \lambda \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda \mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\boldsymbol{\theta} + \lambda^2 \boldsymbol{\theta}_{\mathcal{A}}) - \frac{\lambda^2}{2}\boldsymbol{\theta}_{\mathcal{A}\perp}\Big) && (177)\\
&= -\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \lambda \boldsymbol{\theta}_{\mathcal{A}\|} - \eta\Big(\frac{1}{2}\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta}) \nabla f(\boldsymbol{\theta}) + \frac{\lambda}{2}\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \frac{\lambda}{2}\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\boldsymbol{\theta} + \lambda\lambda^2 2\boldsymbol{\theta}_{\mathcal{A}\|}\Big) && (178)\\
&= -\lambda \boldsymbol{\theta}_{\mathcal{A}\|} - \frac{\eta\lambda^2}{2}\boldsymbol{\theta}_{\mathcal{A}\|} - \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \frac{\eta\lambda}{2}\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \frac{\eta}{2}\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\nabla f(\boldsymbol{\theta}) - \frac{\eta\lambda}{2}\mathbb{1}_{\mathcal{A}} \odot H(\boldsymbol{\theta})\boldsymbol{\theta} && (179)\\
&= -(1 + \frac{\eta\lambda}{2})(\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}_{\mathcal{A}\|}) \\
&\quad - \frac{\eta}{2} H_{\mathcal{A}}(\boldsymbol{\theta}) \nabla_{\mathcal{A}} f(\boldsymbol{\theta}) - \frac{\eta}{2}\nabla_{\mathcal{A}} \nabla_{\mathcal{A}^{\mathsf{c}}}^\top f(\boldsymbol{\theta}) \nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}) - \frac{\eta\lambda}{2} H_{\mathcal{A}}(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}} - \frac{\eta\lambda}{2}\nabla_{\mathcal{A}} \nabla_{\mathcal{A}^{\mathsf{c}}}^\top f(\boldsymbol{\theta})\boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}} && (180)\\
&= -(I + \frac{\eta\lambda}{2}I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}))(\nabla_{\mathcal{A}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \lambda \boldsymbol{\theta}_{\mathcal{A}\|}) \\
&\quad - \frac{\eta}{2}\nabla_{\mathcal{A}} \nabla_{\mathcal{A}^{\mathsf{c}}}^\top f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}})(\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \lambda \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) && (181)\\
&= -\lambda(I + \frac{\eta\lambda}{2}I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}))\boldsymbol{\theta}_{\mathcal{A}\|} - (I + \frac{\eta\lambda}{2}I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \\
&\quad + \frac{\eta}{2}I((\nabla_{\mathcal{A}^{\mathsf{c}}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) + \lambda \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \cdot \nabla_{\mathcal{A}^{\mathsf{c}}}))\nabla_{\mathcal{A}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathsf{c}}}) \,. && (182)
\end{aligned}$$

□

A.13 Proof of Theorem B.1

Proof. First, note that

$$\nabla f(\boldsymbol{\theta}) \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) = 0 \,, \tag{183}$$

which can be shown by differentiating $f(\boldsymbol{\theta}) = f(\boldsymbol{G}(\boldsymbol{\theta}, \alpha))$ with respect to α. Thus, assuming $\boldsymbol{\theta} \cdot ((\nabla f(\boldsymbol{\theta}) \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}, \alpha))$ and using $\dot{\boldsymbol{\theta}}(t) = -\nabla f(\boldsymbol{\theta}(t)) - \lambda \boldsymbol{\theta}(t) - \eta \boldsymbol{\xi}(\boldsymbol{\theta}(t))$, we have

$$\begin{aligned}
&\frac{d}{dt}(\boldsymbol{\theta}(t) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha)) && (184)\\
=&\dot{\boldsymbol{\theta}} \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) + \boldsymbol{\theta} \cdot (\dot{\boldsymbol{\theta}} \cdot \nabla \boldsymbol{G}(\boldsymbol{\theta}, \alpha)) && (185)\\
=& - \nabla f(\boldsymbol{\theta}) \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) - \lambda \boldsymbol{\theta} \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) - \eta \boldsymbol{\xi}(\boldsymbol{\theta}) \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) + \boldsymbol{\theta} \cdot (-(\nabla f(\boldsymbol{\theta}) \cdot \nabla) - \lambda(\boldsymbol{\theta} \cdot \nabla) \\
&- \eta \boldsymbol{\xi}(\boldsymbol{\theta}) \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}, \alpha) && (186)\\
=& - \lambda(\boldsymbol{\theta} \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) + \boldsymbol{\theta} \cdot ((\boldsymbol{\theta} \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}, \alpha))) - \eta \boldsymbol{\xi}(\boldsymbol{\theta}) \cdot \boldsymbol{G}(\boldsymbol{\theta}, \alpha) - \boldsymbol{\theta} \cdot ((\eta \boldsymbol{\xi}(\boldsymbol{\theta}) \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}, \alpha)) \,. && (187)
\end{aligned}$$

Using $\dot{\boldsymbol{v}}(t) = -a\boldsymbol{v}(t) + \boldsymbol{u}t \Leftrightarrow \boldsymbol{v}(t) = \boldsymbol{v}(0)e^{-at} + \int_0^t d\tau e^{-a(t-\tau)}\boldsymbol{u}(\tau)$, we have

$$\begin{aligned}
&\boldsymbol{\theta}(t) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha) && (188)\\
=& \boldsymbol{\theta}(0) \cdot \boldsymbol{G}(\boldsymbol{\theta}(0), \alpha) && (189)\\
&- \lambda \int_0^t d\tau e^{-\lambda(t-\tau)} \boldsymbol{\theta}(\tau) \cdot ((\boldsymbol{\theta}(\tau) \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}(\tau), \alpha)) && (190)\\
&- \eta \int_0^t d\tau e^{-\lambda(t-\tau)} (\boldsymbol{\xi}(\boldsymbol{\theta}(\tau)) \cdot \boldsymbol{G}(\boldsymbol{\theta}(\tau), \alpha) + \boldsymbol{\theta}(\tau) \cdot ((\boldsymbol{\xi}(\boldsymbol{\theta}(\tau)) \cdot \nabla)\boldsymbol{G}(\boldsymbol{\theta}(\tau), \alpha))) \,. && (191)
\end{aligned}$$

□

B Learning Dynamics Induced by Symmetry Breaking: Neural Mechanics

To show the benefits of the counter term, we apply it to broken conservation laws [31]. In [31], the authors build relationships between the symmetries of weights and conserved quantities (i.e., Noether's theorem [57, 58] for DNNs), and they also investigate the dynamics of DNNs under symmetry breaking. We address three shortcomings of their analysis: 1) it includes a counter term only up to order one, 2) a discretization error analysis is missing, and 3) their experiment makes too optimistic an assumption on gradients.

First, we generalize broken conservation laws (Equations (18–20) in [31]) by adding all orders of the counter term. Let $\boldsymbol{G}(\boldsymbol{\theta}, \alpha) := \partial_\alpha \boldsymbol{\psi}(\boldsymbol{\theta}, \alpha)$, which is called the generator of symmetry transformation $\boldsymbol{\psi}$.

Theorem B.1 (Generalized broken conservation law)**.** *Let f be symmetric under transformation $\boldsymbol{\psi}$. Assume that $\boldsymbol{G}$ satisfies $\boldsymbol{\theta}(t) \cdot \{(\nabla f(\boldsymbol{\theta}(t)) \cdot \nabla) \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha)\} = 0$. Then,*

$$\begin{aligned}&\frac{d}{dt}(\boldsymbol{\theta}(t) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha)) = \\ &- \lambda \boldsymbol{\theta}(t) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha) - \lambda \boldsymbol{\theta}(t) \cdot \{(\boldsymbol{\theta}(t) \cdot \nabla) \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha)\} - \eta (\boldsymbol{\xi}(\boldsymbol{\theta}(t)) \cdot \nabla) \cdot (\boldsymbol{\xi}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha)) \,.\end{aligned} \tag{192}$$

Note that the assumption holds for translation, scale, and rescale transformation [31]. Furthermore, Equation (192) can be formally solved:

$$\begin{aligned}&\boldsymbol{\theta}(t) \cdot \boldsymbol{G}(\boldsymbol{\theta}(t), \alpha) = \boldsymbol{\theta}(0) \cdot \boldsymbol{G}(\boldsymbol{\theta}(0), \alpha) e^{-\lambda t} \\ &- \lambda \int_0^t e^{-\lambda(t-\tau)} \boldsymbol{\theta}(\tau) \cdot \{(\boldsymbol{\theta}(\tau) \cdot \nabla) \boldsymbol{G}(\boldsymbol{\theta}(\tau), \alpha)\} d\tau \\ &- \eta \int_0^t e^{-\lambda(t-\tau)} (\boldsymbol{\xi}(\boldsymbol{\theta}(\tau)) \cdot \nabla)(\boldsymbol{\xi}(\boldsymbol{\theta}(\tau)) \cdot \boldsymbol{G}(\boldsymbol{\theta}(\tau), \alpha)) d\tau \,.\end{aligned} \tag{193}$$

The proof is given in Appendix A.13. Now, Equation (193) includes all orders of the counter term $\boldsymbol{\xi} = \sum_{\alpha=0}^{\infty} \tilde{\boldsymbol{\xi}}_\alpha$. We can reproduce [31] by setting $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$. In addition, we already know the discretization error (Corollary 4.1), which is lacking in [31]. We also provide empirical results on Equation (193) in the following sections.

B.1 Scale-invariant Layers

For scale transformation, $\boldsymbol{G}(\boldsymbol{\theta}, \alpha) = \alpha_{\mathcal{A}} \boldsymbol{\theta}$, and thus, the left hand side of Equation (193) becomes $||\boldsymbol{\theta}_{\mathcal{A}}||^2$. Therefore, Equation 193 describes the temporal evolution of the weight norm of scale-invariant layers. Figure 7 shows the temporal evolution of $||\boldsymbol{\theta}_{\mathcal{A}}||^2$ for the network explained in Section 6. Figure 8 shows the gap of $||\boldsymbol{\theta}_{\mathcal{A}}||^2$ between GD and its theoretical predictions (GF and EoM) (Equation 193). We see that the counter term reduces the gap. There is an improvement in the experimental settings compared with [31]. As described in [31], they substitute the gradients computed in GD for the gradients used for GF's simulation instead of using small learning rates to simulate continuous trajectories of GF. This approximation reduces computational costs, but it causes an additional gap between the surrogate gradients and the true gradients of GF along the continuous trajectories. Therefore, we avoid this approximation; we use a small learning rate ($\eta = 10^{-5}$) to simulate GF and EoM, as explained in Section 6.

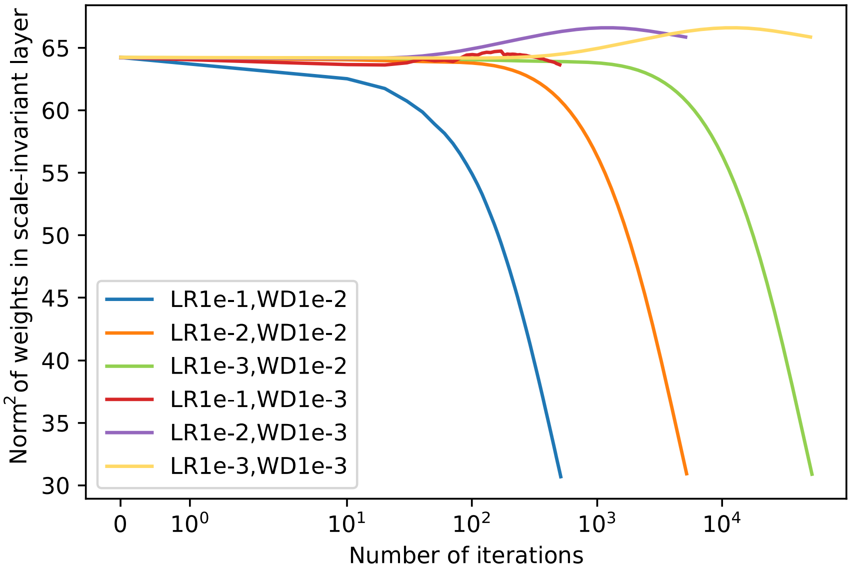


Figure 7: **Dynamics of squared weight norm of scale-invariant layer.** LR and WD mean learning rate and weight decay, respectively. See Section 6 for experimental settings.

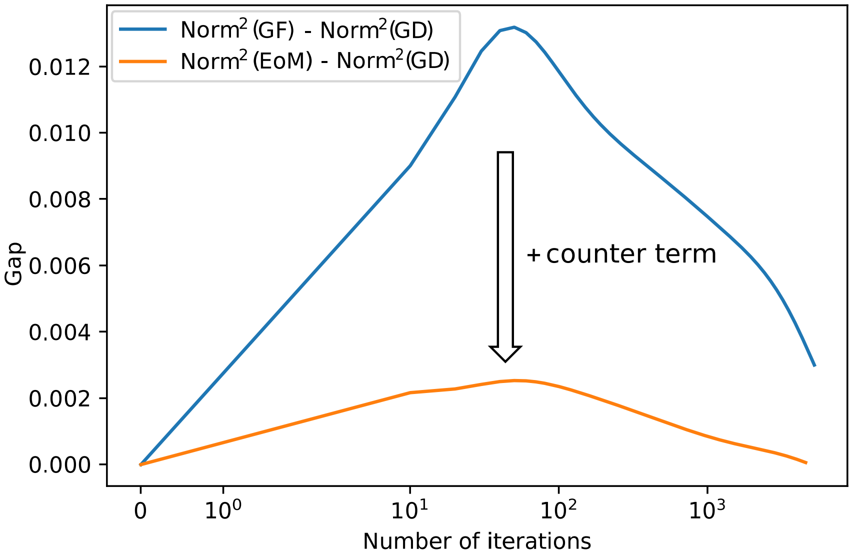


Figure 8: **Discrepancy between actual dynamics of GD and its theoretical prediction (GF and EoM) of squared weight norm of scale-invariant layer.** We see that our counter term reduces the gap between the actual dynamics of GD and its theoretical prediction. See Section 6 for experimental settings.

B.2 Translation-invariant Layers

We also provide an empirical result for translation-invariant layers. For translation transformation, $\boldsymbol{G}(\boldsymbol{\theta}, \alpha) = \alpha \mathbb{1}_{\mathcal{A}}$ and thus the left hand side of Equation (193) becomes $\mathbb{1}_{\mathcal{A}} \cdot \boldsymbol{\theta}_{\mathcal{A}}$ (sum of weights). Therefore, Equation (193) describes the temporal evolution of the sum of weights of translation-invariant layers. Figure 9 shows the temporal evolution of $\mathbb{1}_{\mathcal{A}} \cdot \boldsymbol{\theta}_{\mathcal{A}}$ for the network described in Section 6. Figure 10 shows the gap of $\mathbb{1}_{\mathcal{A}} \cdot \boldsymbol{\theta}_{\mathcal{A}}$ between GD and its theoretical predictions (GF and EoM) (Equation 193). We see that the counter term reduces the gap.

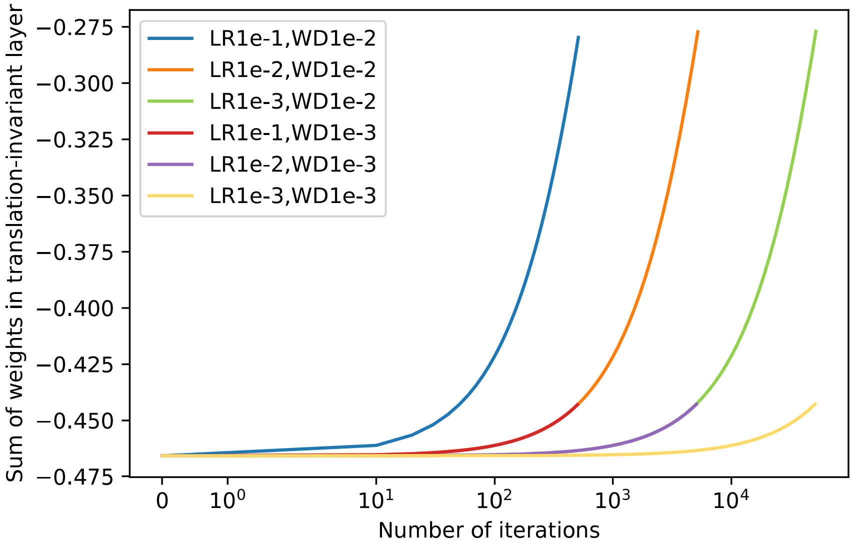


Figure 9: **Sum of weights of translation-invariant layer.** LR and WD mean learning rate and weight decay, respectively. See Section 6 for experimental settings.

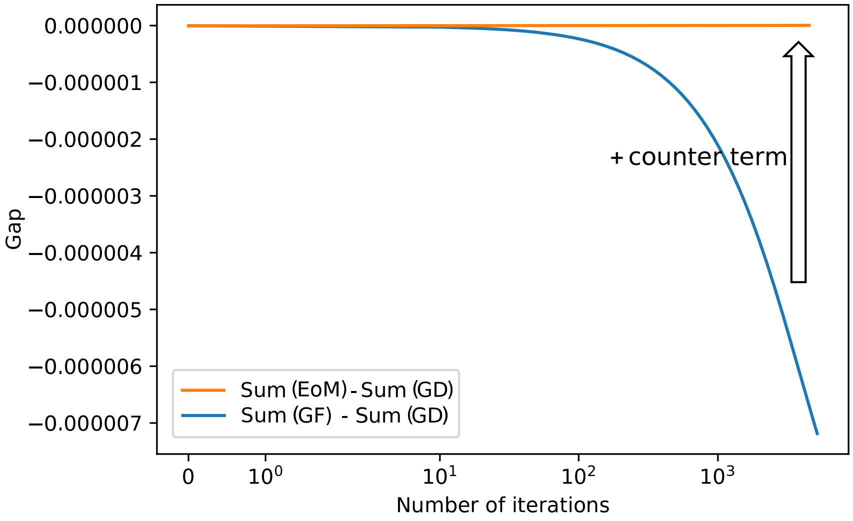


Figure 10: **Discrepancy between actual dynamics of GD and its theoretical prediction (GF and EoM) of sum of weights of translation-invariant layer.** We see that our counter term reduces the gap. See Section 6 for experimental settings.

C Equation of Motion for $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$

For completeness, we construct the EoM for $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ for scale-invariant layers $\mathcal{A}$. See Section 5.1 for the EoM for $r_{\mathcal{A}}$.

Theorem C.1 (EoM for $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$)**.** *EoM (1) gives* $\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}}(t) = -\frac{1}{r_{\mathcal{A}}^2(t)} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t)) + \frac{\eta}{r_{\mathcal{A}}(t)}((\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) \cdot \boldsymbol{\xi}(\boldsymbol{\theta}(t)))\, \hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) - \boldsymbol{\xi}(\boldsymbol{\theta}(t)))$*. Specifically, this is equivalent to:*

$$\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}}(t) = -\frac{1}{r_{\mathcal{A}}^2(t)} \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t)) \tag{194}$$

for $\boldsymbol{\xi} = \mathbf{0}$ *(GF) and*

$$\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}} = -\frac{1}{r_{\mathcal{A}}^2} \left(I + \frac{\eta}{2} H_{\mathcal{A}}(\boldsymbol{\theta}) + \frac{\eta}{2} I((\nabla_{\mathcal{A}^{\mathrm{c}}} f(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \cdot \nabla_{\mathcal{A}^{\mathrm{c}}}) + \frac{\eta}{2} \hat{\boldsymbol{\theta}}_{\mathcal{A}} \nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta}) \right) \nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \tag{195}$$

for $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$ *(EoM), where* $H_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) := (\mathbb{1}_{\mathcal{A}} \odot \nabla)(\mathbb{1}_{\mathcal{A}} \odot \nabla)^{\top} f(\boldsymbol{\theta})|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathcal{A}}}$*.*

The proof is given in Appendix A.9.

Effective learning rate. This result highlights the differences between GD and GF on scale-invariant layers. The factor $\frac{1}{r_{\mathcal{A}}^2}$ (Equation (194)), which is $\frac{\eta}{r_{\mathcal{A}}^2}$ at discretization, is called the *effective learning rate* [29, 42, 30, 43, 44, 33, 45, 34, 46, 47]. The dynamics of $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ is induced by $\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}})$ with the effective learning rate $\frac{\eta}{r_{\mathcal{A}}^2}$, not η. We find that the counter term corrects the effective learning rate to a matrix operator form (Equation (195)). Let us see the meaning of each correction in order. First, I (identity matrix) corresponds to the original effective learning rate. Second, $\frac{\eta}{2} H_{\mathcal{A}}$ directs the gradient $\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}})$ toward the maximum eigenvector of $H_{\mathcal{A}}$, i.e., a flat direction. Therefore, GD tends to go through flatter regions than GF. Third, $\frac{\eta}{2} I((\nabla_{\mathcal{A}^{\mathrm{c}}} f(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \cdot \nabla_{\mathcal{A}^{\mathrm{c}}})$ involves $\nabla_{\mathcal{A}^{\mathrm{c}}} f$ into the learning dynamics of $\mathcal{A}$; therefore, $\mathcal{A}$ is explicitly affected by $\mathcal{A}^{\mathrm{c}}$ in GD, unlike in GF. This point is often missing in the literature on scale-invariant networks because it is often assumed that the whole network is scale-invariant. Fourth, $\frac{\eta}{2} \hat{\boldsymbol{\theta}}_{\mathcal{A}} \nabla_{\mathcal{A}}^{\top} f(\boldsymbol{\theta})$ cancels the $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ component of the right hand side of Equation (195), which may not seem obvious but can be seen from the proof of Theorem C.1 (see Appendix A.9), and thus, $\dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}}$ is orthogonal to $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$, which should be satisfied anyway because $||\hat{\boldsymbol{\theta}}_{\mathcal{A}}||^2 \equiv 1 \Longrightarrow 2 \dot{\hat{\boldsymbol{\theta}}}_{\mathcal{A}} \cdot \hat{\boldsymbol{\theta}}_{\mathcal{A}} = 0$.

D Equation of Motion for $\theta_{\mathcal{A}\|}$

For completeness, we provide the EoM for $\boldsymbol{\theta}_{\mathcal{A}\|}$. The proof is given in Appendix A.12.

Theorem D.1 (EoM for $\boldsymbol{\theta}_{\mathcal{A}\|}$)**.** *EoM (1) gives*

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}(t) = -\lambda \boldsymbol{\theta}_{\mathcal{A}\|}(t) - \nabla f(\boldsymbol{\theta}_{\mathcal{A}\|}(t)) - \eta(I - P)\boldsymbol{\xi}(\boldsymbol{\theta}(t))\,. \tag{196}$$

Specifically, this is equivalent to:

$$\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}(t) = -\lambda \boldsymbol{\theta}_{\mathcal{A}\|}(t) - \nabla f(\boldsymbol{\theta}_{\mathcal{A}\|}(t) + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \tag{197}$$

for $\boldsymbol{\xi} = \mathbf{0}$ (GF) and

$$\begin{aligned}
&\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}(t) = -\lambda\Big(I + \frac{\eta\lambda}{2}I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}})\Big)\boldsymbol{\theta}_{\mathcal{A}\|} \\
&- \Big(I + \frac{\eta\lambda}{2}I + \frac{\eta}{2}H_{\mathcal{A}}(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) + \frac{\eta}{2}I((\nabla_{\mathcal{A}^{\mathrm{c}}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) + \lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \cdot \nabla_{\mathcal{A}^{\mathrm{c}}})\Big)\nabla_{\mathcal{A}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}})
\end{aligned} \tag{198}$$

for $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0$ (EoM).

This result highlights the differences between the dynamics of GD and GF. The two factors $\frac{\eta\lambda}{2}I$ in Equation (198) mean that the existence of weight decay increases the learning rate (increases the velocity $\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}$). The factor $\frac{\eta}{2}H$ means that, as mentioned in Appendix C, GD tends to go along sharper paths than GF. Note that velocity $\dot{\boldsymbol{\theta}}_{\mathcal{A}\|}$ is orthogonal to $\boldsymbol{\theta}_{\mathcal{A}\perp}$ because ∇f, $\boldsymbol{\theta}_{\mathcal{A}\|}$, and $H(\nabla f + \lambda\boldsymbol{\theta}_{\mathcal{A}\|})$ are orthogonal to $\boldsymbol{\theta}_{\mathcal{A}\perp}$. $H(\nabla f + \lambda\boldsymbol{\theta}_{\mathcal{A}\|}) \perp \boldsymbol{\theta}_{\mathcal{A}\perp}$ follows because $H\boldsymbol{v} \perp \boldsymbol{\theta}_{\mathcal{A}\perp}$ for arbitrary non-zero vector $\boldsymbol{v} \in \mathbb{R}^d$ ($\because H\mathbb{1}_{\mathcal{A}} = H\boldsymbol{\theta}_{\mathcal{A}\perp} = 0$) (see Lemma A.7). $\frac{\eta}{2}I((\nabla_{\mathcal{A}^{\mathrm{c}}} f(\boldsymbol{\theta}_{\mathcal{A}\|} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) + \lambda\boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}) \cdot \nabla_{\mathcal{A}^{\mathrm{c}}})$ involves $\nabla_{\mathcal{A}^{\mathrm{c}}} f$ into the learning dynamics of $\mathcal{A}$. We see that the dynamics of $\boldsymbol{\theta}_{\mathcal{A}\|}$ is also independent of that of $\boldsymbol{\theta}_{\mathcal{A}\perp}$, and thus, they are completely separable. A summary of Theorems 5.2 and D.1 is given in Figure 5.

E Details of Experiment

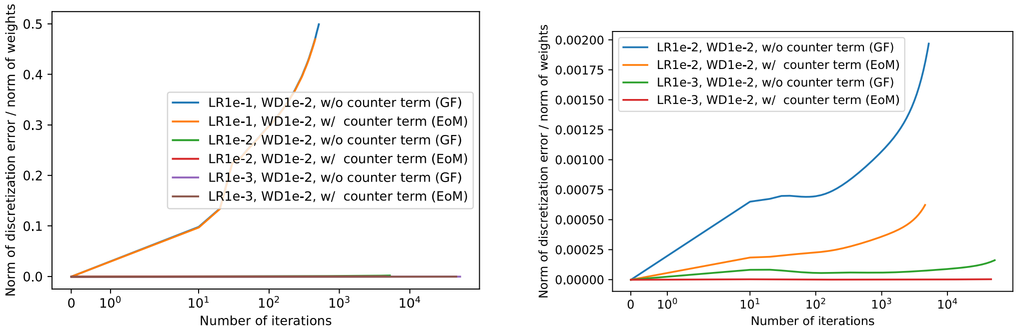
We provide detailed experimental settings (see also Section 6). Our computational infrastructure is a DGX-1 server. The fundamental libraries used in the experiment are TensorFlow 2.3 [59], Numpy 1.18 [60], and Python 3.6.8 [61]. The random seeds used for TensorFlow and Numpy are both 7. The input image is first divided by 127.5 and subtracted by 1. The maximum total number of iterations is 5 million steps for GF and EoM. The total runtime is approximately a month. We use least square fitting (`np.polyfit`) to calculate the decay rates in Table 1. More information and detailed experimental results can be found in our code.

In Figures 2 and 12, the theoretical prediction of discretization error is defined as $||\mathbf{e}_k|| = \frac{\eta^2}{2}||\sum_{s=0}^{k-1}(H(\theta(s\eta)) + \lambda I)\mathbf{g}(\theta(s\eta))||$ (Equation (12)). To reduce computational costs, we approximate the r.h.s.: $(H(\theta(t)) + \lambda I)\mathbf{g}(\boldsymbol{\theta}(t)) \sim \frac{\mathbf{g}(\theta(t)+\epsilon\mathbf{g}(\theta(t)))-\mathbf{g}(\theta(t)-\epsilon\mathbf{g}(\theta(t)))}{2\epsilon}$, where ϵ is set to 10^{-7}. The green curve in Figure 2 is defined as $\boldsymbol{e}_k = \tilde{\boldsymbol{e}}_{100} + \frac{\eta^2}{2}\sum_{s=100}^{k-1}(H(\theta(s\eta)) + \lambda I)\mathbf{g}(\theta(s\eta))$ (compare this with Equation (12)), where $\tilde{\boldsymbol{e}}_{100}$ is the actual discretization error at the 100th step that is obtained from GD. Therefore, the green curve represents the theoretical prediction of discretization error after the 100th step, given $\tilde{\boldsymbol{e}}_{100}$.

F Supplementary Experiment

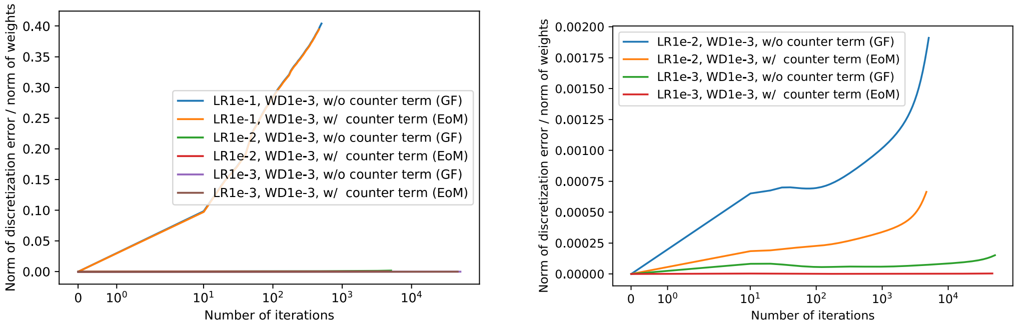
F.1 Relative Discretization Error

We provide the relative discretization error, which is defined as $||\boldsymbol{e}_k||/||\boldsymbol{\theta}_k||$ ($k \in \mathbb{Z}_{\geq 0}$). See Figure 11. We can see that a large learning rate ($\eta = 10^{-1}$) leads to a large discretization error (Figure 11 (a) and (c)). We also see that the counter term reduces the discretization error as expected (Figure 11 (b) and (d)).



(a) Weight decay = 10^{-2}.

(b) Weight decay = 10^{-2}. Magnified.



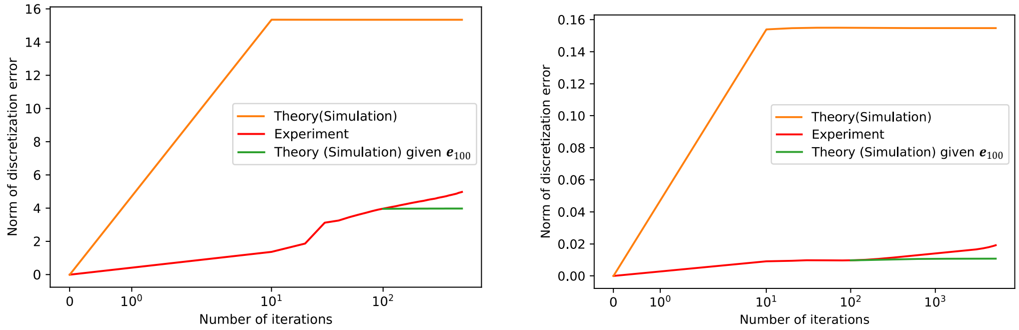
(c) Weight decay = 10^{-3}.

(d) Weight decay = 10^{-3}. Magnified.

Figure 11: **Relative discretization error.** In (a) and (c), the LR1e-1 curves overlap each other, and the LR1e-2 and LR1e-3 curves collapse in the lower region of the figure. The LR1e-2 and LR1e-3 are magnified and shown in (c) and (d). See Section 6 and Appendix E for experimental settings.

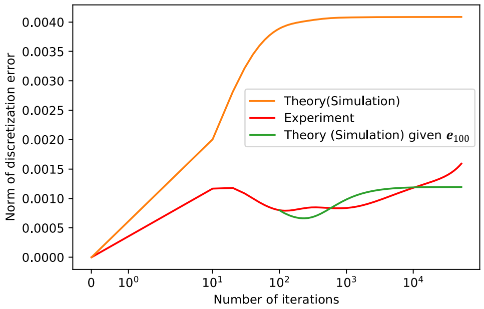
F.2 Theoretical Prediction Vs. Experimental Result of Discretization Error

We compare the theoretical prediction of discretization error between GF and GD (Equation (12)) with the actual discretization error obtained in the experiment. The green curve is defined as $\boldsymbol{e}_k = \tilde{\boldsymbol{e}}_{100} + \frac{\eta^2}{2}\sum_{s=100}^{k-1}(H(\theta(s\eta)) + \lambda I)\mathbf{g}(\theta(s\eta)) + O(\eta^3)$ (compare this with Equation (12)), where $\tilde{\boldsymbol{e}}_{100}$ is the actual discretization error at the 100th step. Therefore, the green curve represents the theoretical prediction of discretization error after the 100th step given $\tilde{\boldsymbol{e}}_{100}$.



(a) Learning rate = 10^{-1}.

(b) Learning rate = 10^{-2}.



(c) Learning rate = 10^{-3}.

Figure 12: **Theoretical prediction (Equation (12)) vs. experimental result of discretization error between GF and GD.** The weight decay is 10^{-2}. See Section 6 and Appendix E for experimental settings.

G Supplementary Discussion

Supplementary related work (Section 2). To show the benefits of EoM, we focus on scale-invariant layers [29, 42, 30, 43, 44, 33, 45, 34, 46, 47] and translation-invariant layers [31, 32] in Section 5. To carry over the stability of a continuous optimization algorithm to a discretized system, the authors of [19] add a feedback term to the optimization, and after that, they apply a discretization method to it. The authors' primary motivation is to keep the orthogonality of the weight parameters of DNNs, which is different from ours.

Convergence of $\boldsymbol{\xi}$ (Section 3.3). Note that the expansion of $\boldsymbol{\xi}$ in terms of η is not necessarily convergent, as is also pointed out in [35]. Thus, we have to truncate the expansion at a suitable order. The discretization error at the truncation is given in Theorem 4.1.

Beyond leading order of discretization error (Theorem 3.2 and Section 4.1). In this work, we analyze the leading order of discretization error. However, higher-order terms cannot always be negligible. We discuss in Section 4.1 that the higher-order terms are important at the beginning of training.

Existence of $\mathcal{A}^{\mathrm{c}}$ (Section 5). In our theoretical analysis of scale- and translation-invariant layers, the network contains both invariant ($\mathcal{A}$) and non-invariant layers ($\mathcal{A}^{\mathrm{c}}$), while previous works assume the whole network is invariant for simplicity [29, 42, 30, 43, 44, 33, 45, 34, 46, 47]. We avoid this assumption and show that such mixed networks require appropriate modifications to analyses of invariant networks. For example, $\nabla f(\boldsymbol{\theta}) = \frac{1}{||\boldsymbol{\theta}||}\nabla f(\hat{\boldsymbol{\theta}})$ for invariant networks, while $\nabla_{\mathcal{A}} f(\boldsymbol{\theta}) = \frac{1}{||\boldsymbol{\theta}_{\mathcal{A}}||}\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}} + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}})$ for mixed networks (Lemma A.5), not $\frac{1}{||\boldsymbol{\theta}_{\mathcal{A}}||}\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}})$. Such a naive replacement is not allowed.

Higher-order corrections to decay rate of $r_{\mathcal{A}}$ (Section 5.1). We can compute more corrections to the decay rate of $r_{\mathcal{A}}$ ($\mathcal{A}$ is a scale-invariant layer), using more counter terms. For example, a long algebra gives decay rate $\eta\lambda(1 + \frac{\eta\lambda}{2} + \frac{\eta^2\lambda^2}{3})$ for $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_0 + \eta\tilde{\boldsymbol{\xi}}_1$. The proof is similar to Appendix A.7.

On equilibrium assumptions in Corollaries 5.1 and 5.2 (Section 5.1). We make assumptions in Corollaries 5.1 and 5.2; there exist two constants $r_{\mathcal{A}*} \geq 0$ and $c_* \geq 0$ such that $r_{\mathcal{A}}(t) \xrightarrow{t\to\infty} r_{\mathcal{A}*}$ and $||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}(t))|| \xrightarrow{t\to\infty} c_*$. These assumptions are similar to those given in previous studies [29, 34]. However, whether the assumptions are valid in the actual learning dynamics of DNNs is of independent interest. In fact, the equilibrium assumption ($r_{\mathcal{A}*}(t)$ and $||\nabla_{\mathcal{A}} f(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(t) + \boldsymbol{\theta}_{\mathcal{A}^{\mathrm{c}}}(t))|| \xrightarrow{t\to\infty}$ constant) could not be satisfied even at one million steps of GD, and potentially because of it, $r_{\mathcal{A}*}$ and Δ_* have a large discrepancy between the empirical results and theoretical predictions. Deeper analyses on this point are needed. Under what conditions are the equilibrium assumptions valid? Can we relax the equilibrium assumptions and obtain realistic limiting dynamics of scale-invariant layers? This is exciting future work.

In contrast to our empirical result mentioned above, in [34], their experiments dramatically match their theoretical prediction. This is potentially because of differences in experimental settings; in [34], SGD is used (ours is GD) and variance is induced, ResNet-50 [62, 63] is used (ours is a fully-connected network with three layers), ImageNet [64, 65] and MSCOCO [66] are used (ours is MNIST [50]), and large learning rates ($\sim 10^{-1}$) and small weight decays ($\sim 10^{-4}$) are used (ours are given in Appendix E).

Extension of EoM to general settings (Section 7). While we focus on GD and GF for simplicity, our counter-term-based approach and discretization error analysis can be extended to more general settings, such as SGD, acceleration methods (e.g., momentum SGD), and adaptive optimizers (e.g., Adam [53]). First, to extend our analysis to SGD, discretization error analysis of the Euler-Maruyama method, e.g., [67], can be used. SDE's error analysis [23, 24] is also relevant. Second, we can extend our counter-term-based approach and discretization error analysis to acceleration methods by modifying the analysis for different differential equations from GF and different discretization schemes from the Euler method, as is discussed in [7, 14, 12]. Third, [56] is the first work that provides a continuous approximation of Adam. However, its counter term and discretization error are open questions.