On Training Implicit Models (Supplementary material)

Anonymous Author(s)
Affiliation
Address
email

A Algorithm of Phantom Gradients

The following PyTorch-style [1] pseudo code describes the implementation of phantom gradients in both the unrolling form (see Alg. 1) and the Neumann form (see Alg. 2). To implement phantom gradients with TensorFlow [2], replace the no_grad context manager with the $stop\_gradient$ operator.

The unrolling-based phantom gradient is computed by the automatic differentiation engine, while the Neumann-series-based phantom gradient is given by Alg. 3. A special reminder is that, for a trained model, removing the unrolling steps in the test stage will not cause the performance decay but accelerate the inference instead. After training, even increasing the unrolling steps k to 20 in the test stage can not further improve the performance. This implies that the final results are fully obtained by the implicit model rather than the unrolling steps.

Algorithm 1 Unrolling-based phantom gradient, PyTorch-style

```
# solver: the solver to find h*, e.g., the Broyden solver in MDEQ.
# func: the explicit function F that defines the implicit model.
# z: the input variables z to solve h* = F(h*, z)
# h: the solution h* of the implicit models.
# training: a bool variable that indicates training or inference.
# k: the unrolling step k.
# lambda_: the damping factor λ.

# a plain forward pass using Pytorch
# calculate the phantom gradient by automatic differentiation
# input: z & output: h
def forward(z):
    with torch.no_grad():
        h = solver(func, z)

    # define the computational graph for the backward pass.
    # only used in the training stage
    if training:
        for _ in range(k):
            h = (1 - lambda_) * h + lambda_ * func(h, z)

    return h
```

Algorithm 2 Neumann-series-based Phantom Gradient, Pytorch-style

```
# solver: the solver to find h*, e.g., the Broyden solver in MDEQ.
# func: the explicit function F that defines the implicit model.
# grad(a, b, c): the function to compute the Jacobian-vector product (∂a/∂b) c
# z: the input variables z to solve h* = F(h*, z)
# h: the output h* of the implicit model.
# k: the unrolling step k.
# lambda_: the damping factor λ.

# a plain forward pass using Pytorch
# input: z & output: h
def forward(z):
    with torch.no_grad():
        h = solver(func, z)

    return h

# phantom gradient for the backward pass
# input: dl/ dh & output: dl / dz
def phantom_grad(g):
    # forward pass for automatic differentiation
    f = (1 - lambda_) * h + lambda_ * func(h, z)

    g_hat = g
    for _ in range(k-1):
        # compute Jacobian-vector product with automatic differentiation
        g_hat = g + grad(f, h, g_hat)

    # compute Jacobian-vector product to obtain dl / dz
    g_hat = grad(f, z, g_hat)
    return g_hat
```

Algorithm 3 Neumann-series-based phantom gradient with $\mathcal{O}(1)$ memory

Input $\partial\mathcal{L}/\partial\boldsymbol{h}$, $\mathcal{F}$, $\boldsymbol{h}^*$, k, λ.
Initialize $\widehat{\boldsymbol{g}} = \boldsymbol{g} = \partial\mathcal{L}/\partial\boldsymbol{h}$;
$\boldsymbol{f} \leftarrow (1-\lambda)\boldsymbol{h}^* + \lambda\mathcal{F}(\boldsymbol{h}^*, \boldsymbol{z})$
for $i = 1, 2, \cdots, k-1$ **do**
    $\widehat{\boldsymbol{g}} \leftarrow \boldsymbol{g} + (\partial\boldsymbol{f}/\partial\boldsymbol{h})\,\widehat{\boldsymbol{g}}$; ▷ Compute Jacobian-vector product with automatic differentiation
end for
$\widehat{\boldsymbol{g}} \leftarrow (\partial\boldsymbol{f}/\partial\boldsymbol{z})\,\widehat{\boldsymbol{g}}$ ▷ Compute Jacobian-vector product to obtain the phantom gradient *w.r.t.* $\boldsymbol{z}$
return $\widehat{\boldsymbol{g}}$.

B Proof of Theorems

Theorem 1. *Suppose the exact gradient and the phantom gradient are given by Eq. (4) and Eq. (5), respectively. Let σ_{max} and σ_{min} be the maximal and minimal singular value of $\partial\mathcal{F}/\partial\boldsymbol{\theta}$. If*

$$\left\| \boldsymbol{A}\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right) - \frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}} \right\| \le \frac{\sigma_{min}^2}{\sigma_{max}}, \tag{A-1}$$

then the phantom gradient provides a descent direction of the function $\mathcal{L}$, i.e.,

$$\left\langle \widehat{\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}}, \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} \right\rangle \ge 0. \tag{A-2}$$

Proof. Denote $\boldsymbol{J} = \partial\mathcal{F}/\partial\boldsymbol{\theta}$, $\boldsymbol{v} = \partial\mathcal{L}/\partial\boldsymbol{h}$, and $\boldsymbol{u} = (\boldsymbol{I} - \partial\mathcal{F}/\partial\boldsymbol{h})^{-1}\boldsymbol{v}$. Let

$$\boldsymbol{E} = \boldsymbol{A}\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right) - \frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}}, \tag{A-3}$$

and we have $\|\boldsymbol{E}\| \le \sigma_{\text{min}}^2/\sigma_{\text{max}}$. Then,

$$\begin{aligned}\left\langle \widehat{\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}}, \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\rangle &= \boldsymbol{v}^\top\boldsymbol{A}^\top\boldsymbol{J}\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right)^{-1}\boldsymbol{v} = \boldsymbol{u}^\top\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right)^\top\boldsymbol{A}^\top\boldsymbol{J}\boldsymbol{u} = \boldsymbol{u}^\top(\boldsymbol{J}+\boldsymbol{E})^\top\boldsymbol{J}\boldsymbol{u}\\ &\ge \|\boldsymbol{J}\boldsymbol{u}\|^2 - \|\boldsymbol{E}\|\|\boldsymbol{J}\|\|\boldsymbol{u}\|^2 \ge \left(\sigma_{\text{min}}^2 - \sigma_{\text{max}}\|\boldsymbol{E}\|\right)\|\boldsymbol{u}\|^2 \ge 0,\end{aligned} \tag{A-4}$$

which concludes the proof. □

Proof of Remerk 1. Suppose $\boldsymbol{A} = (\partial\mathcal{F}/\partial\boldsymbol{\theta})\boldsymbol{D}$ and the condition in Eq. (8). Then,

$$\left\|\boldsymbol{A}\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right) - \frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}}\right\| \le \left\|\frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}}\right\|\left\|\boldsymbol{D}\left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right) - \boldsymbol{I}\right\| \le \sigma_{\text{max}} \cdot \frac{1}{\kappa^2} = \frac{\sigma_{\text{min}}^2}{\sigma_{\text{max}}}, \tag{A-5}$$

indicating the condition in Eq. (6) is satisfied. □

Theorem 2. *Suppose the matrix $\partial\mathcal{F}/\partial\boldsymbol{h}$ is a contractive mapping. Then,*

(i) the Neumann series in (15) converges to the Jacobian-inverse $(\boldsymbol{I} - \partial\mathcal{F}/\partial\boldsymbol{h})^{-1}$; and

(ii) if the function $\mathcal{F}$ is continuously differentiable w.r.t. both $\boldsymbol{h}$ and $\boldsymbol{\theta}$, the sequence in Eq. (14) converges to the exact Jacobian $\partial\boldsymbol{h}^*/\partial\boldsymbol{\theta}$ as $T \to \infty$, i.e.,

$$\lim_{T\to\infty}\frac{\partial\boldsymbol{h}_T}{\partial\boldsymbol{\theta}} = \left.\frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{h}^*}\left(\boldsymbol{I} - \left.\frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right|_{\boldsymbol{h}^*}\right)^{-1}. \tag{A-6}$$

Proof. (i) Since $\|\partial\mathcal{F}/\partial\boldsymbol{h}\| < 1$,

$$\|\boldsymbol{B}\| \le \lambda\left\|\frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right\| + (1-\lambda)\|\boldsymbol{I}\| < 1. \tag{A-7}$$

Let $\boldsymbol{B}_k = \sum_{t=0}^{k-1}\boldsymbol{B}^t$, and for each $p \in \mathbb{N}_+$, we have

$$\|\boldsymbol{B}_{k+p} - \boldsymbol{B}_k\| = \left\|\sum_{t=k}^{k+p-1}\boldsymbol{B}^t\right\| \le \|\boldsymbol{B}\|^k\left\|\sum_{t=0}^{p-1}\boldsymbol{B}^t\right\| \le \|\boldsymbol{B}\|^k\sum_{t=0}^{p-1}\|\boldsymbol{B}\|^t < \frac{\|\boldsymbol{B}\|^k}{1-\|\boldsymbol{B}\|}. \tag{A-8}$$

By the Cauchy's convergence test, the sequence $\{\boldsymbol{B}_k\}$ is convergent. Since

$$(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{B}_k = \boldsymbol{I} - \boldsymbol{B}^k \to \boldsymbol{I}, \quad \text{as } k \to \infty, \tag{A-9}$$

it follows that $\boldsymbol{B}_k \to (\boldsymbol{I} - \boldsymbol{B})^{-1}$, as $k \to \infty$. Therefore,

$$\lambda\sum_{t=0}^{\infty}\boldsymbol{B}^t = \lambda(\boldsymbol{I} - \boldsymbol{B})^{-1} = \left(\boldsymbol{I} - \frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right)^{-1}. \tag{A-10}$$

(ii) Let $\mathcal{H}(\boldsymbol{h}, \boldsymbol{z}) = \lambda\mathcal{F}(\boldsymbol{h}, \boldsymbol{z}) + (1-\lambda)\boldsymbol{h}$, and

$$\frac{\partial\mathcal{H}}{\partial\boldsymbol{h}} = \lambda\frac{\partial\mathcal{F}}{\partial\boldsymbol{h}} + (1-\lambda)\boldsymbol{I}. \tag{A-11}$$

Similar to Eq. (A-7), $\partial\mathcal{H}/\partial\boldsymbol{h}$ is also a contractive mapping. By the Banach Fixed Point Theorem [3], the sequence $\{\boldsymbol{h}_t\}$ converges to an exact fixed point $\boldsymbol{h}^*$ of $\mathcal{H}$, which is also a fixed point of $\mathcal{F}$.

Denote

$$\boldsymbol{U}_t = \left.\frac{\partial\mathcal{F}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{h}_t}, \quad \boldsymbol{V}_t = \lambda\left.\frac{\partial\mathcal{F}}{\partial\boldsymbol{h}}\right|_{\boldsymbol{h}_t} + (1-\lambda)\boldsymbol{I}. \tag{A-12}$$

Since the function $\mathcal{F}$ is continuously differentiable *w.r.t.* both $\boldsymbol{h}$ and $\boldsymbol{\theta}$, we have

$$\lim_{t\to\infty} \boldsymbol{U}_t = \left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{h}^*} = \boldsymbol{U}_\infty, \quad \lim_{t\to\infty} \boldsymbol{V}_t = \lambda \left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{h}}\right|_{\boldsymbol{h}^*} + (1-\lambda)\,\boldsymbol{I} = \boldsymbol{V}_\infty. \tag{A-13}$$

According to the conclusion in (i), we have

$$\left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{h}^*} \left(\boldsymbol{I} - \left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{h}}\right|_{\boldsymbol{h}^*}\right)^{-1} = \lambda\, \boldsymbol{U}_\infty \sum_{t=0}^{\infty} \boldsymbol{V}_\infty^t. \tag{A-14}$$

Comparing Eq. (14) with Eq. (A-14), we have

$$\begin{aligned}
&\left\| \frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{\theta}} - \left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{h}^*} \left(\boldsymbol{I} - \left.\frac{\partial \mathcal{F}}{\partial \boldsymbol{h}}\right|_{\boldsymbol{h}^*}\right)^{-1} \right\| = \lambda \left\| \sum_{t=0}^{T-1} \boldsymbol{U}_t \prod_{s=t+1}^{T-1} \boldsymbol{V}_s - \boldsymbol{U}_\infty \sum_{t=0}^{\infty} \boldsymbol{V}_\infty^t \right\| \\
&\le \lambda \left(\underbrace{\left\| \sum_{t=0}^{T-2} \boldsymbol{U}_t \left(\prod_{s=t+1}^{T-1} \boldsymbol{V}_s - \boldsymbol{V}_\infty^{T-t-1} \right) \right\|}_{\boldsymbol{\Delta}_1} + \underbrace{\left\| \sum_{t=0}^{T-1} (\boldsymbol{U}_t - \boldsymbol{U}_\infty)\, \boldsymbol{V}_\infty^{T-t-1} \right\|}_{\boldsymbol{\Delta}_2} + \underbrace{\left\| \boldsymbol{U}_\infty \sum_{t=T}^{\infty} \boldsymbol{V}_\infty^t \right\|}_{\boldsymbol{\Delta}_3} \right).
\end{aligned} \tag{A-15}$$

In the following context, we prove Eq. (A-6) by showing that $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3$ can be arbitrarily small when T is sufficiently large.

Preparations. For any $\epsilon > 0$, since $\boldsymbol{U}_t \to \boldsymbol{U}_\infty$ and $\boldsymbol{V}_t \to \boldsymbol{V}_\infty$ as $t \to \infty$, there exists $N \in \mathbb{N}_+$ s.t.

$$\|\boldsymbol{U}_t - \boldsymbol{U}_\infty\| < \epsilon, \quad \|\boldsymbol{V}_t - \boldsymbol{V}_\infty\| < \epsilon, \quad \forall t > N. \tag{A-16}$$

Since $\partial \mathcal{H}/\partial \boldsymbol{h}$ is a contractive mapping, there exists $\gamma \in (0,1)$ s.t.

$$\|\boldsymbol{V}_t\| \le \gamma, \quad \|\boldsymbol{V}_\infty\| \le \gamma. \tag{A-17}$$

Besides, since $\partial \mathcal{F}/\partial \boldsymbol{\theta}$ is a continuous function and $\{\boldsymbol{h}_t\}$ is a convergent sequence, it follows that $\{\boldsymbol{h}_t\}$ forms a compact set and that $\partial \mathcal{F}/\partial \boldsymbol{\theta}$ is bounded on $\{\boldsymbol{h}_t\}$. Therefore, there exists $M > 0$, s.t.

$$\|\boldsymbol{U}_t\| \le M, \quad t = 0, 1, 2, \cdots. \tag{A-18}$$

Taking $t \to \infty$, we have $\|\boldsymbol{U}_\infty\| \le M$.

For $\boldsymbol{\Delta}_1$. For $t > N$, consider

$$\begin{aligned}
\left\| \boldsymbol{U}_t \left(\prod_{s=t+1}^{T-1} \boldsymbol{V}_s - \boldsymbol{V}_\infty^{T-t-1} \right) \right\| &\le \|\boldsymbol{U}_t\| \sum_{s=t+1}^{T-1} \left\| \boldsymbol{V}_{t+1} \boldsymbol{V}_{t+2} \cdots \boldsymbol{V}_s \boldsymbol{V}_\infty^{T-s-1} - \boldsymbol{V}_{t+1} \boldsymbol{V}_{t+2} \cdots \boldsymbol{V}_{s-1} \boldsymbol{V}_\infty^{T-s} \right\| \\
&\le \|\boldsymbol{U}_t\| \sum_{s=t+1}^{T-1} \|\boldsymbol{V}_{t+1}\| \, \|\boldsymbol{V}_{t+2}\| \cdots \|\boldsymbol{V}_{s-1}\| \, \|\boldsymbol{V}_s - \boldsymbol{V}_\infty\| \, \|\boldsymbol{V}_\infty\|^{T-s-1} \\
&\le M(T-t-1)\gamma^{T-t-2}\epsilon,
\end{aligned} \tag{A-19}$$

and for $t \le N$, we simply have

$$\left\| \boldsymbol{U}_t \left(\prod_{s=t+1}^{T-1} \boldsymbol{V}_s - \boldsymbol{V}_\infty^{T-t-1} \right) \right\| \le \|\boldsymbol{U}_t\| \left(\prod_{s=t+1}^{T-1} \|\boldsymbol{V}_s\| + \|\boldsymbol{V}_\infty\|^{T-t-1} \right) \le 2M\gamma^{T-t-1}. \tag{A-20}$$

Therefore, when $T > N + 2$, $\boldsymbol{\Delta}_1$ can be bounded as follows:

$$\begin{aligned}
\boldsymbol{\Delta}_1 &\le \left(\sum_{t=0}^{N} + \sum_{t=N+1}^{T-2} \right) \left\| \boldsymbol{U}_t \left(\prod_{s=t+1}^{T-1} \boldsymbol{V}_s - \boldsymbol{V}_\infty^{T-t-1} \right) \right\| \\
&\le 2M \sum_{t=0}^{N} \gamma^{T-t-1} + M\epsilon \sum_{t=N+1}^{T-2} (T-t-1)\gamma^{T-t-2} \\
&\le 2M\gamma^{T-N-1} \frac{1-\gamma^{N+1}}{1-\gamma} + \left(\frac{1-\gamma^{T-N-2}}{(1-\gamma)^2} - \frac{(T-N-2)\gamma^{T-N-2}}{1-\gamma} \right) M\epsilon \\
&\le \frac{2M}{1-\gamma}\gamma^{T-N-1} + \frac{M}{(1-\gamma)^2}\epsilon.
\end{aligned} \tag{A-21}$$

Since $M/(1-\gamma)^2$ is a constant and $\gamma^{T-N-1} \to 0$ as $T \to \infty$, $\boldsymbol{\Delta}_1$ can be arbitrarily small for a sufficiently large T.

For $\boldsymbol{\Delta}_2$. Consider

$$\left\|\left(\boldsymbol{U}_t - \boldsymbol{U}_\infty\right) \boldsymbol{V}_\infty^{T-t-1}\right\| \le \left\|\boldsymbol{U}_t - \boldsymbol{U}_\infty\right\| \left\|\boldsymbol{V}_\infty\right\|^{T-t-1} \le \begin{cases} \gamma^{T-t-1}\epsilon, & \text{when } t \ge N; \\ 2M\gamma^{T-t-1} & \text{when } t < N. \end{cases} \tag{A-22}$$

Therefore, when $T > N+2$, $\boldsymbol{\Delta}_2$ can be bounded as follows:

$$\begin{aligned} \boldsymbol{\Delta}_2 &\le \left(\sum_{t=0}^{N} + \sum_{t=N+1}^{T-1}\right) \left\|\left(\boldsymbol{U}_t - \boldsymbol{U}_\infty\right) \boldsymbol{V}_\infty^{T-t-1}\right\| \\ &\le 2M \sum_{t=0}^{N} \gamma^{T-t-1} + \epsilon \sum_{t=N+1}^{T-1} \gamma^{T-t-1} \\ &\le \frac{2M}{1-\gamma}\gamma^{T-N-1} + \frac{\epsilon}{1-\gamma}. \end{aligned} \tag{A-23}$$

Since $1/(1-\gamma)$ is a constant and $\gamma^{T-N-1} \to 0$ as $T \to \infty$, $\boldsymbol{\Delta}_2$ can be arbitrarily small for a sufficiently large T.

For $\boldsymbol{\Delta}_3$. As $t \to \infty$, we have

$$\left\|\boldsymbol{U}_\infty \sum_{t=T}^{\infty} \boldsymbol{V}_\infty^t\right\| \le \left\|\boldsymbol{U}_\infty\right\| \left\|\boldsymbol{V}_\infty\right\|^T \left\|\left(\boldsymbol{I} - \boldsymbol{V}_\infty\right)^{-1}\right\| \le M \cdot \gamma^T \cdot \frac{1}{1-\gamma} \to 0. \tag{A-24}$$

As a result, the conclusion in Eq. (A-6) is proved. □

Theorem 3. *Suppose the loss function $\mathcal{R}$ in Eq. (3) is ℓ-smooth, lower-bounded, and has bounded gradient almost surely in the training process. Besides, assume the gradient in Eq. (4) is an unbiased estimator of $\nabla\mathcal{R}(\boldsymbol{\theta})$ with a bounded covariance. If the phantom gradient in Eq. (5) is an ϵ-approximation to the gradient in Eq. (4), i.e.,*

$$\left\|\widehat{\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}} - \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}\right\| \le \epsilon, \quad \textit{almost surely}, \tag{A-25}$$

then using Eq. (5) as a stochastic first-order oracle with a step size of $\eta_\tau = O(1/\sqrt{\tau})$ to update $\boldsymbol{\theta}$ with gradient descent, it follows after T iterations that

$$\mathbb{E}\left[\frac{\sum_{\tau=1}^{T} \eta_\tau \|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\|^2}{\sum_{\tau=1}^{T} \eta_\tau}\right] \le O\left(\epsilon + \frac{\log T}{\sqrt{T}}\right). \tag{A-26}$$

Proof. Let $\widehat{\frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\theta}}}$ be the phantom gradient at the τ^{th} iteration. By ℓ-smoothness of $\mathcal{R}$, we have

$$\begin{aligned} \mathcal{R}(\boldsymbol{\theta}_{\tau+1}) &\le \mathcal{R}(\boldsymbol{\theta}_\tau) + \langle \nabla\mathcal{R}(\boldsymbol{\theta}_\tau), \boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau \rangle + \frac{\ell}{2} \left\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\right\|^2 \\ &= \mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau \left\langle \nabla\mathcal{R}(\boldsymbol{\theta}_\tau), \widehat{\frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\theta}}} \right\rangle + \frac{\ell\eta_\tau^2}{2} \left\|\widehat{\frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\theta}}}\right\|^2. \end{aligned} \tag{A-27}$$

Let

$$\boldsymbol{e}_\tau = \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau} - \widehat{\frac{\partial \mathcal{L}_\tau}{\partial \boldsymbol{\theta}}} \tag{A-28}$$

be the approximation error at the τ^{th} iteration. Taking expectation *w.r.t.* the first τ iterations, we have

$$\mathbb{E}_{1\sim\tau}\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1})\right] = \mathbb{E}_{1\sim\tau-1}\left[\mathbb{E}_\tau\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1}) \mid 1 \sim \tau-1\right]\right] = \mathbb{E}_{1\sim\tau-1}\left[\mathbb{E}_\tau\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1}) \mid \boldsymbol{\theta}_\tau\right]\right], \tag{A-29}$$

where the first equality comes from the *law of total expectation*, while the second from the fact that the stochasticity of the first $\tau-1$ steps is totally captured by the value $\boldsymbol{\theta}_\tau$. Consider the inner expectation in Eq. (A-29), and we omit the condition on $\boldsymbol{\theta}_\tau$ when no ambiguity is made. Note that in the following derivation, all expectations and variances are conditioned on $\boldsymbol{\theta}_\tau$:

$$\begin{aligned}\mathbb{E}_\tau\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1})\right] &\leq \mathbb{E}_\tau\left[\mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau\left\langle \nabla\mathcal{R}(\boldsymbol{\theta}_\tau), \frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right\rangle + \frac{\ell\eta_\tau^2}{2}\left\|\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right\|^2\right] \\ &= \mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau\left\langle \nabla\mathcal{R}(\boldsymbol{\theta}_\tau), \mathbb{E}_\tau\left[\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right]\right\rangle + \frac{\ell\eta_\tau^2}{2}\mathbb{E}_\tau\left[\left\|\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right\|^2\right],\end{aligned} \tag{A-30}$$

where

$$\mathbb{E}_\tau\left[\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right] = \mathbb{E}_\tau\left[\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau} - \boldsymbol{e}_\tau\right] = \nabla\mathcal{R}(\boldsymbol{\theta}_\tau) - \mathbb{E}_\tau\left[\boldsymbol{e}_\tau\right], \tag{A-31}$$

and

$$\mathbb{E}_\tau\left[\left\|\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right\|^2\right] = \left\|\mathbb{E}_\tau\left[\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right]\right\|^2 + \operatorname{tr}\left(\operatorname{Cov}_\tau\left(\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right)\right). \tag{A-32}$$

Suppose $\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\| \leq G$ almost surely, and then we have

$$\left\|\mathbb{E}_\tau\left[\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right]\right\|^2 = \left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau) - \mathbb{E}_\tau\left[\boldsymbol{e}_\tau\right]\right\|^2 \leq (G+\epsilon)^2. \tag{A-33}$$

Moreover, by the properties of covariance,

$$\begin{aligned}&\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\frac{\partial\widehat{\mathcal{L}_\tau}}{\partial\boldsymbol{\theta}}\right)\right) = \operatorname{tr}\left(\operatorname{Cov}_\tau\left(\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau} - \boldsymbol{e}_\tau\right)\right) \\ &= \operatorname{tr}\left(\operatorname{Cov}_\tau\left(\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau}\right)\right) + \operatorname{tr}\left(\operatorname{Cov}_\tau\left(\boldsymbol{e}_\tau\right)\right) - 2\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau}, \boldsymbol{e}_\tau\right)\right) \\ &\leq 2\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau}\right)\right) + 2\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\boldsymbol{e}_\tau\right)\right),\end{aligned} \tag{A-34}$$

where the last inequility comes from

$$\begin{aligned}\left|\operatorname{tr}\left(\operatorname{Cov}\left(\boldsymbol{a},\boldsymbol{b}\right)\right)\right| &\leq \sum_i \left|\operatorname{Cov}\left(a_i,b_i\right)\right| \leq \sum_i \sqrt{\operatorname{Var}\left(a_i\right)\operatorname{Var}\left(b_i\right)} \leq \sum_i \frac{\operatorname{Var}\left(a_i\right)+\operatorname{Var}\left(b_i\right)}{2} \\ &= \frac{1}{2}\left(\operatorname{tr}\left(\operatorname{Cov}\left(\boldsymbol{a}\right)\right) + \operatorname{tr}\left(\operatorname{Cov}\left(\boldsymbol{b}\right)\right)\right).\end{aligned} \tag{A-35}$$

By the Popoviciu's inequality on variances [4], the second term in (A-34) can be bounded by $d_{\boldsymbol{\theta}}\,\epsilon^2$, *i.e.,*

$$\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\boldsymbol{e}_\tau\right)\right) \leq d_{\boldsymbol{\theta}}\,\epsilon^2, \tag{A-36}$$

where $d_{\boldsymbol{\theta}}$ denotes the dimension of $\boldsymbol{\theta}$. Finally, since the gradient estimator $\partial\mathcal{L}/\partial\boldsymbol{\theta}$ has a bounded covariance, there exists $M > 0$, s.t.

$$\operatorname{tr}\left(\operatorname{Cov}_\tau\left(\left.\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\tau}\right)\right) \leq M, \quad \text{almost surely.} \tag{A-37}$$

Combining (A-30), (A-31), (A-32), (A-33), (A-34), (A-37), we have

$$\begin{aligned}\mathbb{E}_\tau\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1})\right] &\leq \mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|^2 + \eta_\tau\left\langle\nabla\mathcal{R}(\boldsymbol{\theta}_\tau), \mathbb{E}_\tau\left[\boldsymbol{e}_\tau\right]\right\rangle + K\eta_\tau^2, \\ &\leq \mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|^2 + \eta_\tau\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|\left\|\mathbb{E}_\tau\left[\boldsymbol{e}_\tau\right]\right\| + K\eta_\tau^2 \\ &\leq \mathcal{R}(\boldsymbol{\theta}_\tau) - \eta_\tau\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|^2 + \eta_\tau G\epsilon + K\eta_\tau^2,\end{aligned} \tag{A-38}$$

where $K = \ell\left((G+\epsilon)^2 + 2M + 2d_{\boldsymbol{\theta}}\,\epsilon^2\right)/2$ is a constant. Substitute (A-38) into Eq. (A-29), and it becomes

$$\mathbb{E}_{1\sim\tau}\left[\mathcal{R}(\boldsymbol{\theta}_{\tau+1})\right] \le \mathbb{E}_{1\sim\tau-1}\left[\mathcal{R}(\boldsymbol{\theta}_{\tau})\right] - \eta_\tau \mathbb{E}_{1\sim\tau-1}\left[\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|^2\right] + \eta_\tau G\epsilon + K\eta_\tau^2. \tag{A-39}$$

By taking a summation over the first T steps,

$$\begin{aligned}\mathbb{E}_{1\sim T}\left[\sum_{\tau=1}^{T}\eta_\tau\left\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\right\|^2\right] &\le \mathcal{R}(\boldsymbol{\theta}_1) - \mathbb{E}_{1\sim T}\left[\mathcal{R}(\boldsymbol{\theta}_{T+1})\right] + G\epsilon\sum_{\tau=1}^{T}\eta_\tau + K\sum_{\tau=1}^{T}\eta_\tau^2\\ &\le \mathcal{R}(\boldsymbol{\theta}_1) - m + G\epsilon\sum_{\tau=1}^{T}\eta_\tau + K\sum_{\tau=1}^{T}\eta_\tau^2,\end{aligned} \tag{A-40}$$

where $m = \inf_{\boldsymbol{\theta}}\mathcal{R}(\boldsymbol{\theta})$ since $\mathcal{R}$ is lower-bounded. Dividing a factor of $\sum_{\tau=1}^{T}\eta_\tau$, we have

$$\mathbb{E}_{1\sim T}\left[\frac{\sum_{\tau=1}^{T}\eta_\tau\|\nabla\mathcal{R}(\boldsymbol{\theta}_\tau)\|^2}{\sum_{\tau=1}^{T}\eta_\tau}\right] \le G\epsilon + \frac{\mathcal{R}(\boldsymbol{\theta}_1) - m}{\sum_{\tau=1}^{T}\eta_\tau} + K\frac{\sum_{\tau=1}^{T}\eta_\tau^2}{\sum_{\tau=1}^{T}\eta_\tau}. \tag{A-41}$$

Since $\eta_\tau = O(1/\sqrt{T})$, it follows that

$$\sum_{\tau=1}^{T}\eta_\tau = O\left(\sqrt{T}\right), \quad \frac{\sum_{\tau=1}^{T}\eta_\tau^2}{\sum_{\tau=1}^{T}\eta_\tau} = O\left(\frac{\log T}{\sqrt{T}}\right). \tag{A-42}$$

Combining (A-41) and Eq. (A-42) concludes the proof. □

Remark 4. The assumption that $\mathcal{R}$ has almost-surely bounded gradient at $\{\boldsymbol{\theta}_\tau\}_{\tau=0}^{T}$ is reasonable. Because of the norm-based regularization, *e.g.,* weight decay, we can assume $\boldsymbol{\theta}$ is almost surely optimized within a compact set in the parameter space. If we further assume $\mathcal{R}$ is continuously differentiable, the almost-sure boundedness of $\|\nabla\mathcal{R}\|$ within the compact set follows its continuity.

Remark 5. We justify the assumption that the gradient in Eq. (4) has a bounded covariance. For the SGD algorithm, the stochasticity of Eq. (4) comes from the random sampling of the training example (or the training mini-batch) from the dataset. Since there are finite samples in the training set, the covariance of Eq. (4) remains finite. Moreover, as Theorem 2 only considers a finite training schedule, *i.e.,* T steps, the possible combination of the selected sample at each step is still finite (even though its number grows exponentially). Therefore, it is reasonable to assume the gradient in Eq. (4) has a bounded covariance.

C Implementation Details

C.1 Synthetic Settings

For the synthetic setting, the following model is used:

$$\boldsymbol{h}^* = \mathcal{F}\left(\boldsymbol{h}^* + \boldsymbol{u}\right) \tag{A-43}$$

where $\mathcal{F}$ is an 1-layer convolutional network with spectral normalization [5], and $\boldsymbol{u}, \boldsymbol{h}^* \in \mathbb{R}^{B\times C\times N}$. The loss $\mathcal{L}$ is given by the mean squared error between $\boldsymbol{h}^*$ and $\boldsymbol{y}$. We choose $C = 128$, $N = 32$, and randomly sample 50000 data pairs $(\boldsymbol{u}, \boldsymbol{y})$ to compute the gradient $\partial\mathcal{L}/\partial\boldsymbol{u}$.

We generate a symmetric weight matrix for the network and constrain the Lipschitz constant $L_{\boldsymbol{h}}$ to a given level with spectral normalization. For the visualization in the main paper, we adopt $L_{\boldsymbol{h}} = 0.9$. For the additional visualization on the stability of the solver in Fig. 1, we choose $L_{\boldsymbol{h}}$ from $\{0.9, 0.99, 0.999, 0.9999\}$.

To solve $\boldsymbol{h}^*$, we employ the fixed-point iteration as the solver. For the synthetic setting shown in the main paper, we use 100 fixed-point iterations to obtain $\boldsymbol{h}^*$ that satisfies the relative error $\|\boldsymbol{h}^* - \mathcal{F}(\boldsymbol{h}^*, \boldsymbol{u})\|/\|\boldsymbol{h}^*\| \le 10^{-5}$. For the visualization in Fig. 1, we also apply 100 fixed-point iterations for each $L_{\boldsymbol{h}}$.

C.2 Ablation Settings

For the ablation settings, we use the original MDEQ-Tiny [6] model (170K parameters) on the CIFAR-10 [7] classification benchmark without any architecture modification. Therefore, the performance gain upon state-of-the-art methods is due to the improved training efficiency thanks to the proposed phantom gradient.

The experiments are conducted without data augmentation as in [6]. The training schedule, batch size, cosine learning rate annealing strategy, and other hyperparameters are kept unchanged for all ablation experiments. We also follow the official training protocol of MDEQ[1] to reproduce its results.

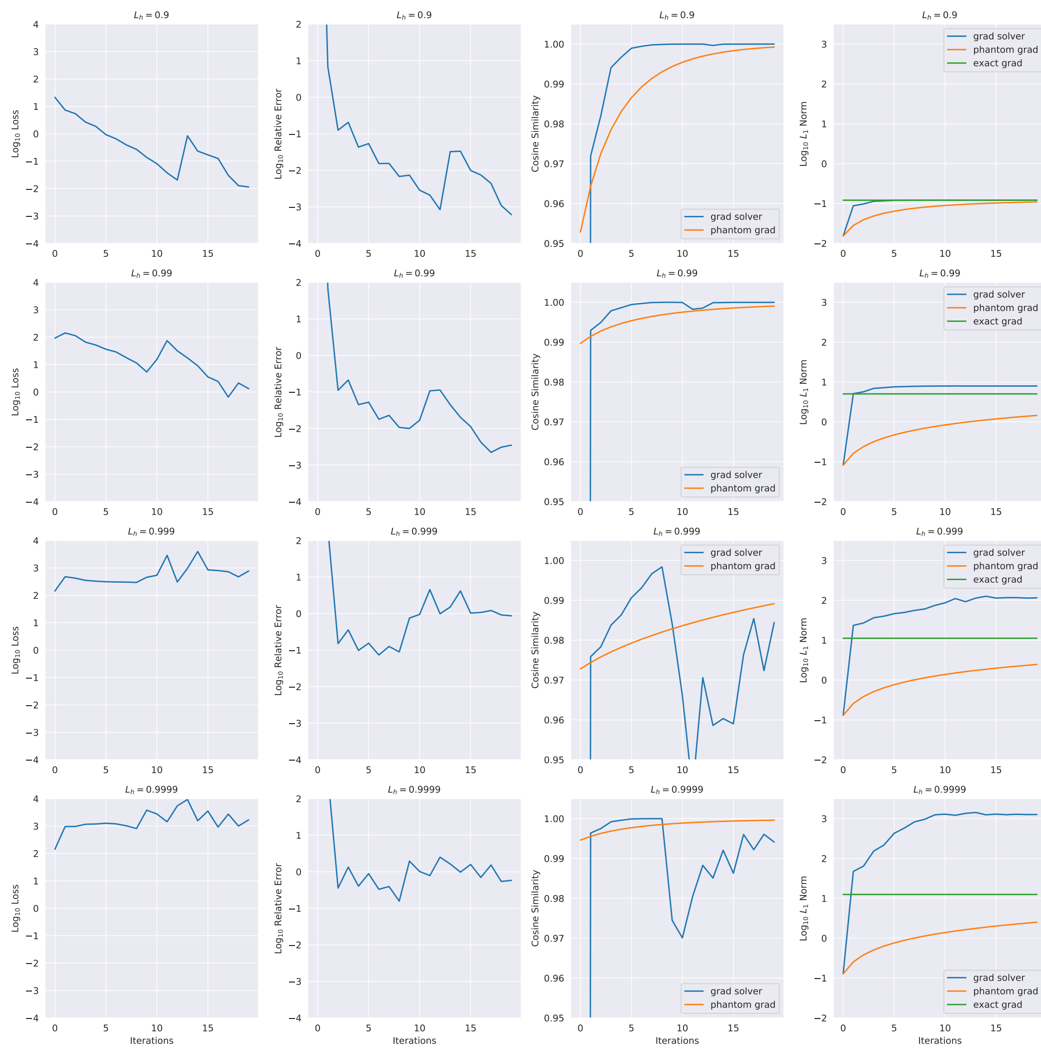

Figure 1: Visualization of gradient solver under different $L_{\boldsymbol{h}}$.

For the training protocol without pretraining, we substitute the unrolled pretraining stage by implicit differentiation. For the training protocol without Dropout, we remove the variational Dropout from the model. We also try the SGD optimizer with a regular hyperparameter setting of learning rate 0.1 and weight decay 0.0001.

We train the MDEQ model with two types of phantom gradient using the SGD optimizer with a learning rate of 0.1, a weight decay of 0.0001, and other hyperparameters unchanged from the original setting. The model is trained without shallow-layer pretraining, suggesting an $\mathcal{O}(k)$ or $\mathcal{O}(1)$ peak

[1]Code available at https://github.com/locuslab/mdeq.

memory usage in the unrolling form or the Neumann form, as presented in the main paper. In both cases, the damped fixed-point iteration starts at the solution $\widehat{\boldsymbol{h}}^*$ obtained by the Broyden's method.

C.3 Large-Scale Experiments

For large-scale experiments, we adopt the MDEQ model (10M parameters) and MDEQ-Small model (18M parameters) on the CIFAR-10 [7] and ImageNet [8] benchmarks, respectively. To train MDEQ on CIFAR-10, we employ the unrolling-based phantom gradient with $\lambda = 0.5$ and $k = 5$, *i.e.,* $\boldsymbol{A}_{5,0.5}$. Besides, we use the SGD optimizer with a learning rate of 0.1 and a weight decay 0.0001, and keep other experimental setting unchanged, including the number of training epochs, batch size, the learning rate annealing strategy, and *etc*. On the ImageNet dataset, we follow the practice of [6] to pretrain the model for the same number of epochs. Afterwards, the unrolling-based phantom gradient (*i.e.,* $\boldsymbol{A}_{5,0.5}$) is used to train the model for the rest training schedule.

D Additional Analysis on the Gradient Solver

To illustrate the vulnerability the gradient solver for implicit differentiation in the ill-conditioned cases, we provide the optimization dynamics in Fig. 1 and its comparison with the phantom gradient in the synthetic setting. We plot the optimization objective $\|(I - \partial\mathcal{F}/\partial\boldsymbol{h})\boldsymbol{g} - \partial\mathcal{L}/\partial\boldsymbol{h}\|$, the relative error $\|(I - \partial\mathcal{F}/\partial\boldsymbol{h})\boldsymbol{g} - \partial\mathcal{L}/\partial\boldsymbol{h}\|/\|\boldsymbol{g}\|$, the cosine similarity between the solved gradient $\boldsymbol{g}$ and the exact gradient, and the L_1 norm of the gradient. Here, in the optimization-based context, $\boldsymbol{g}$ is the solution of the backward linear system solved by the Broyden solver.

Fig. 1 shows that the gradient solver may diverge in ill-conditioned situations. It can be seen that the phantom gradient demonstrates much better stability especially in the extremely ill-conditioned cases. On the contrary, the optimization step does not necessarily lead to the solved gradient being more aligned to the exact gradient, as indicated by the cosine similarity, and its angle may oscillate in the great range. Besides, the norm of the solved gradient also tends to explode in the optimization process, while the phantom gradient maintains a moderate norm throughout.

References

[1] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 1

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, page 265–283, 2016. 1

[3] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. math*, 3(1):133–181, 1922. 3

[4] Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145, 1935. 6

[5] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2018. 7

[6] Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Multiscale Deep Equilibrium Models. In *Neural Information Processing Systems (NeurIPS)*, pages 5238–5250, 2020. 8, 9

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009. 8, 9

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 9