A. Experimental Details

A.1. Data Sets and Network Architectures

The MNIST, CIFAR-10, CIFAR-100 data sets are obtained from PyTorch's torchvision package.[1] A summary is shown in Table 3. The networks (MLP on MNIST, and CNN on CIFAR-10, CIFAR-100) used are shown in Table 4. Models 1 and 2 have been used in (Yu et al., 2019) on CIFAR-10 and CIFAR-100, respectively. Model 3 has been used in (Han et al., 2018).

Table 3. Data sets with artificial label noise.

	#tra	#val	#test	#classes
MNIST	60,000	5,000	5,000	10
CIFAR-10	50,000	5,000	5,000	10
CIFAR-100	50,000	5,000	5,000	100

A.2. Details for Figure 1

A.2.1. Figure 1(a)

We use the CIFAR-10 dataset (Table 3), and model 1 in Table 4. The number of training epochs is 200. We use the Adam optimizer (Kingma & Ba, 2014) with momentum 0.9 and batch size 128. The initial learning rate is 0.001, and is linearly decayed to zero from the 80th epoch. The 5 random $R(T)$s (denoted "Random R(T)" 1-5) are generated by uniform sampling the corresponding hyperparameter $\boldsymbol{x} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$.

Besides the test accuracies shown in Figure 1(a), we also show in Figure 8 the random $R(T)$s, the original $R(T)$ used in Co-teaching (Han et al., 2018), the $R(T)$ obtained by *S2E* (denoted "Searched"), and the implicit $R(T)$ corresponding to training on the whole noisy dataset (denoted "Baseline").

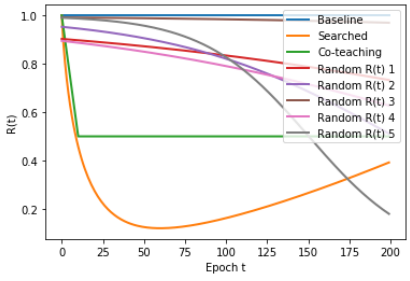


Figure 8. $R(t)$ used in Figure 1(a).

A.2.2. Figures 1(b)-1(c)

Experiment is performed on the MNIST/CIFAR-10/CIFAR-100 datasets (Table 3). The number of training epochs, batch size, and learning rate schedule are the same as that in Figure 1(a).

A.2.3. Figure 1(d)

We use the CNN models 1-3 in Table 4. As CIFAR-100 has 100 outputs, we also change the number of outputs of model 1 to 100. The number of training epochs, batch size, and learning rate schedule are the same as that in Figure 1(a).

A.2.4. Figure 1(e)

We use model 1 in Table 4. For Adam, the learning rate schedule is the same as that in Figure 1(a). For SGD, the initial learning rate is 0.1, and decayed to 0.01 and 0.001 at the 500th and 750th epoch, respectively. Moreover, the number of training epochs is 1000 instead of 200. For RMSProp, the learning rate is fixed at 0.01.

A.2.5. Figure 1(f)

The number of training epochs, batch size, and learning rate schedule are the same as that in Figure 1(a). We only change the batch size and initial learning rate as shown in the figure of Figure 1(f). Moreover, to better demonstrate the memorization effect for small learning rates, the number of training epochs is set to 1000 instead of 200.

A.3. Additional Plots for Section 4.1.2

Figure 9 compares the label precisions of the various methods on CIFAR-10 and CIFAR-100.

B. Additional Experiments

B.1. Approximation to $R(\cdot)$ in Co-teaching

Recall that $R(t)$ in Co-teaching is generated from (1). As all basis functions in Table 1 are smooth, it is not possible for (4) to exactly subsume (1). However, $R(t)$ in (4) can well approximate (1). to illustrate this, we randomly generate three $R(t)$'s in Co-teaching's search space by uniform sampling the corresponding hyperparameters $\tau \in (0, 1)$, $c \in \{0.5, 1, 2\}$ and $t_k \in (0, 200)$. Figure 10 shows the function in (4) that best approximates each of these $R(t)$'s with the least squared error.

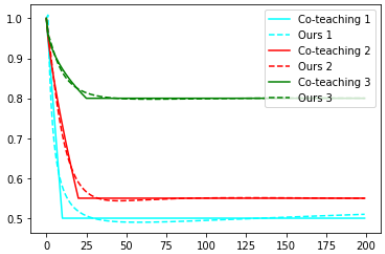


Figure 10. $R(t)$ in Co-teaching and the best approximation from the proposed search space.

[1] https://pytorch.org/docs/stable/torchvision/datasets.html

Table 4. MLP and CNN models used in the experiments.

MLP on MNIST	CNN on CIFAR-10 (Model 1)	CNN on CIFAR-100 (Model 2)	CNN (Model 3)
28×28 gray image	32×32 RGB image	32×32 RGB image	32×32 RGB image
Dense 28×28→256 ReLU	5×5 Conv, 6 ReLU 2×2 Max-pool	3×3 Conv, 64 BN, ReLU 3×3 Conv, 64 BN, ReLU 2×2 Max-pool	3×3 Conv, 128 BN, LReLU 3×3 Conv, 128 BN, LReLU 3×3 Conv, 128 BN, LReLU 2×2 Max-pool, stride 2 Dropout, p=0.25
	5×5 Conv, 16 ReLU 2×2 Max-pool	3×3 Conv, 128 BN, ReLU 3×3 Conv, 128 BN, ReLU 2×2 Max-pool	3×3 Conv, 256 BN, LReLU 3×3 Conv, 256 BN, LReLU 3×3 Conv, 256 BN, LReLU 2×2 Max-pool, stride 2 Dropout, p=0.25
	Dense 16×5×5→120 ReLU Dense 120→84 ReLU	3×3 Conv, 196 BN, ReLU 3×3 Conv, 196 BN, ReLU 2×2 Max-pool	3×3 Conv, 512 BN, LReLU 3×3 Conv, 256 BN, LReLU 3×3 Conv, 128 BN, LReLU Avg-pool
Dense 256→10	Dense 84→10	Dense 256→100	Dense 128→ 10

B.2. Comparison with Weight Sharing

Weight sharing (Pham et al., 2018; Liu et al., 2019) is a popular method to speed up the search in NAS. In this experiment, we study if weight-sharing is also beneficial to the search of $R(\cdot)$. We compare *S2E* with *ASNG* (Akimoto et al., 2019), which is a weight-sharing version of NG. Specifically, *ASNG* optimizes

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{G}(\boldsymbol{\theta},\boldsymbol{w}) \equiv \int_{\boldsymbol{x}\in\mathcal{F}} \mathcal{L}_{\text{val}}(f(\boldsymbol{w};R(\boldsymbol{x})),\mathcal{D}_{val}) p_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{x},$$

by alternating the updates of $\boldsymbol{w}$ (using gradient descent) and $\boldsymbol{\theta}$ (using natural gradient descent). Unlike *S2E* in (6), in which each $\boldsymbol{\theta}$ has its own optimal $\boldsymbol{w}^*$, *ASNG* only uses one $\boldsymbol{w}$ that is shared by all $\boldsymbol{\theta}$.

Table 5 compare the test accuracies of *S2E* and *ASNG*. As can be seen, the $R(\cdot)$ obtained by *ASNG* is much worse than that from *S2E*, indicating weight-sharing is not a good choice here. Recently, the problem of weight sharing is also discussed in (Sciuto et al., 2020), which shows that it is not useful in NAS for convolutional and recurrent neural works.

Table 5. Testing accuracies (%) obtained on CIFAR-10 by ASNG and S2E.

	sym-20%	sym-50%	pair-45%
ASNG	57.82	47.34	41.46
S2E	58.73	50.82	47.58

C. Proofs

C.1. Proposition 1

Proof. By definition,

$$\begin{aligned}\nabla^2 \mathcal{J}(\boldsymbol{\theta}) &= \int \bar{f}(\boldsymbol{x}) \nabla^2 p_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \mathbb{E}_{p_{\boldsymbol{\theta}}}\left[\bar{f}(\boldsymbol{x}) \frac{\nabla^2 p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right]. \end{aligned} \tag{10}$$

Now,

$$\begin{aligned}\nabla^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \nabla\left(\frac{\nabla p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right) \\ &= \frac{\nabla^2 p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} - \frac{\nabla p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla p_{\boldsymbol{\theta}}(\boldsymbol{x})^\top}{p_{\boldsymbol{\theta}}^2(\boldsymbol{x})}.\end{aligned}$$

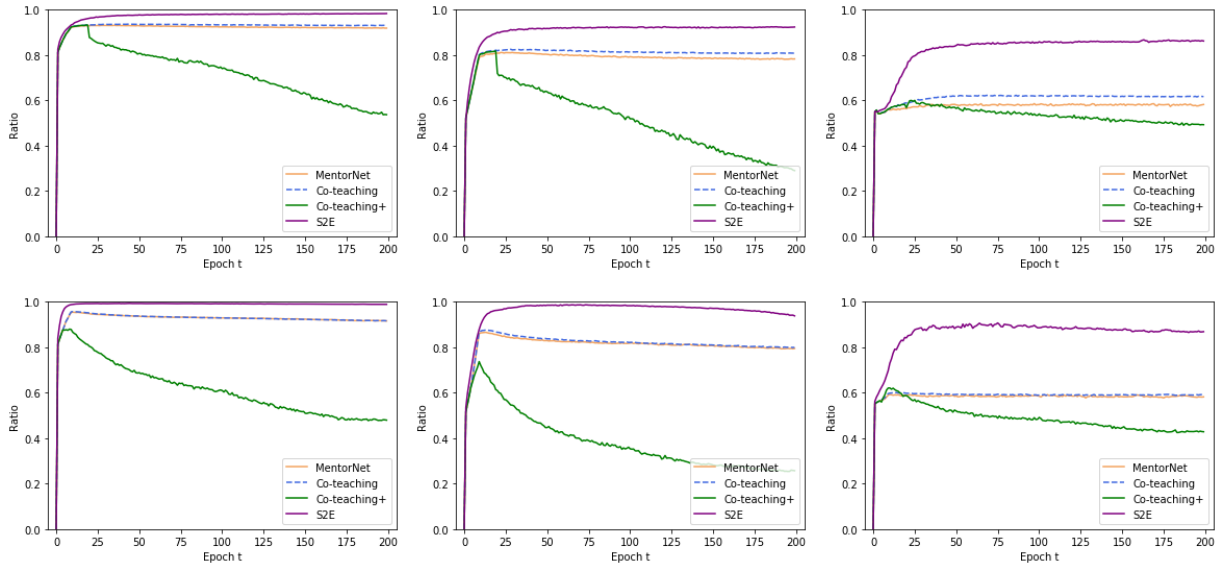
Thus,

$$\begin{aligned}\frac{\nabla^2 p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} &= \nabla^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{\nabla p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla p_{\boldsymbol{\theta}}(\boldsymbol{x})^\top}{p_{\boldsymbol{\theta}}^2(\boldsymbol{x})} \\ &= \nabla^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \left(\frac{\nabla p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)\left(\frac{\nabla p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)^\top \\ &= \nabla^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \bar{\boldsymbol{p}}_{\boldsymbol{\theta}} \bar{\boldsymbol{p}}_{\boldsymbol{\theta}}^\top,\end{aligned}$$

Result follows on substituting this into (10). □

C.2. Proposition 2

Proof. First, we introduce the following Lemma 1 which results from Assumption 2.



(a) symmetry flipping (20%). (b) symmetry flipping (50%). (c) pair flipping (45%).

Figure 9. Label precisions of *MentorNet*, *Co-teaching*, *Co-teaching+* and *S2E* on CIFAR-10 (top) and CIFAR-100 (bottom).

Lemma 1 ((Rockafellar, 1970)). *Since $\mathcal{J}$ is L-Lipschitz smooth, we have $\mathcal{J}(\boldsymbol{y}) \leq \mathcal{J}(\boldsymbol{x}) + \langle \nabla \mathcal{J}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2$ for any $\boldsymbol{x}$ and $\boldsymbol{y}$.*

Define a function g as

$$g(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{H}) = (\boldsymbol{\theta} - \boldsymbol{y})^\top \boldsymbol{z} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{y})^\top \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{y}).$$

Due to (9), we can express $\boldsymbol{\theta}^{m+1}$ as

$$\boldsymbol{\theta}^{m+1} = \arg\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{H}), \tag{11}$$

where

$$\boldsymbol{y} = \boldsymbol{\theta}^m, \boldsymbol{z} = \nabla \mathcal{J}(\boldsymbol{\theta}^m) - \boldsymbol{e}^m \text{ and } \boldsymbol{H} = \boldsymbol{\Delta}^m. \tag{12}$$

Note that $\boldsymbol{\Delta}^m$ is a positive definite matrix, thus g is a convex function on $\boldsymbol{\theta}$. Consider the directional derivative of g w.r.t. $\boldsymbol{\theta}$ at the optimal point $\boldsymbol{\theta} = \boldsymbol{\theta}^{m+1}$, and using the fact that g is a convex function, we have

$$\left\langle \boldsymbol{z} + \boldsymbol{H}(\boldsymbol{\theta}^{m+1} - \boldsymbol{y}), \boldsymbol{w}^m \right\rangle \geq 0 \tag{13}$$

for any direction $\boldsymbol{w}$.

Let $\boldsymbol{w} = \boldsymbol{\theta}^m - \boldsymbol{\theta}^{m+1}$. Combining (12) and (13), we have

$$\langle \nabla \mathcal{J}(\boldsymbol{\theta}^m) - \boldsymbol{e}^m, \boldsymbol{\gamma}^m \rangle \leq -(\boldsymbol{\gamma}^m)^\top \boldsymbol{\Delta}^m \boldsymbol{\gamma}^m. \tag{14}$$

Next, using Lemma 1, we have

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^{m+1}) \leq & \mathcal{J}(\boldsymbol{\theta}^m) + \left\langle \nabla \mathcal{J}(\boldsymbol{\theta}^m), \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\rangle \\ & + \frac{L}{2} \left\| \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\|^2. \end{aligned} \tag{15}$$

Now, we add the error term $\boldsymbol{e}^m$ in (15), i.e.,

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^{m+1}) \leq & \mathcal{J}(\boldsymbol{\theta}^m) + \left\langle \nabla \mathcal{J}(\boldsymbol{\theta}^m) - \boldsymbol{e}^m, \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\rangle \\ & + \frac{L}{2} \left\| \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\|^2 + \left\langle \boldsymbol{e}^m, \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\rangle, \\ \leq & \mathcal{J}(\boldsymbol{\theta}^m) - (\boldsymbol{\gamma}^m)^\top \boldsymbol{\Delta}^m \boldsymbol{\gamma}^m + \frac{L}{2} \left\| \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\|^2 \\ & + \left\langle \boldsymbol{e}^m, \boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m \right\rangle, \end{aligned} \tag{16}$$

$$= \mathcal{J}(\boldsymbol{\theta}^m) - (\boldsymbol{\gamma}^m)^\top \boldsymbol{\Delta}^m \boldsymbol{\gamma}^m + \frac{L}{2} \left\| \boldsymbol{\gamma}^m \right\|^2 + \left\langle \boldsymbol{e}^m, \boldsymbol{\gamma}^m \right\rangle,$$

$$\begin{aligned}\leq & \mathcal{J}(\boldsymbol{\theta}^m) - (\boldsymbol{\gamma}^m)^\top \boldsymbol{\Delta}^m \boldsymbol{\gamma}^m + \frac{L}{2} \left\| \boldsymbol{\gamma}^m \right\|^2 \\ & + \left\| \boldsymbol{e}^m \right\|_2 \left\| \boldsymbol{\gamma}^m \right\|_2, \end{aligned} \tag{17}$$

$$\begin{aligned}\leq & \mathcal{J}(\boldsymbol{\theta}^m) - \frac{1}{\eta} \left\| \boldsymbol{\gamma}^m \right\|^2 + \frac{L}{2} \left\| \boldsymbol{\gamma}^m \right\|^2 \\ & + \left\| \boldsymbol{e}^m \right\|_2 \left\| \boldsymbol{\gamma}^m \right\|_2, \end{aligned} \tag{18}$$

$$= \mathcal{J}(\boldsymbol{\theta}^m) + \frac{\eta L - 2\eta}{2\eta} \left\| \boldsymbol{\gamma}^m \right\|^2 + \left\| \boldsymbol{e}^m \right\|_2 \left\| \boldsymbol{\gamma}^m \right\|_2, \tag{19}$$

where (16) is from (14), (17) is from inequality $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle \leq \|\boldsymbol{\alpha}\| \|\boldsymbol{\beta}\|$, (18) results from Assumption 3, i.e., the smallest eigen value of $\boldsymbol{\Delta}^m$ is not smaller than η. Finally, rearranging terms in (19), we will obtain the Proposition. □

C.3. Theorem 1

Before proving this Theorem 1, we first introduce the following Lemma 1.

Lemma 2. *Define*

$$\varepsilon^m = \left\| (\boldsymbol{\Delta}^m)^{-1} \boldsymbol{e}^m \right\| \text{ and } \rho^m = (\boldsymbol{\Delta}^m)^{-1} \mathcal{J}(\boldsymbol{\theta}^m).$$

We have

$$c^m \leq \|\boldsymbol{\gamma}^m\| \leq \|\rho^m\| + \varepsilon^m,$$

where $c^m = \max\left(\|\rho^m\| - \varepsilon^m, \varepsilon^m - \|\rho^m\|\right)$.

Proof. Since $\boldsymbol{\theta}^{m+1}$ is generated by (9), thus

$$\begin{aligned}\|\boldsymbol{\gamma}^m\| &= \left\|\left(\boldsymbol{\Delta}^m\right)^{-1}\left(\nabla\mathcal{J}(\boldsymbol{\theta}^m) - \boldsymbol{e}^m\right)\right\|, \\ &= \left\|\left(\boldsymbol{\Delta}^m\right)^{-1}\boldsymbol{e}^m + \rho^m\right\|.\end{aligned}$$

Then, the Lemma follows from Cauchy inequality. □

Now, we start to prove Theorem 1.

Proof of Theorem 1. Since the eigen values of $\boldsymbol{\Delta}^m$ are in $[\eta, L]$ (by Assumption 3), we have

$$\frac{1}{L}\|\boldsymbol{e}^m\| \leq \left\|\left(\boldsymbol{\Delta}^m\right)^{-1}\boldsymbol{e}^m\right\| \leq \frac{1}{\eta}\|\boldsymbol{e}^m\|. \tag{20}$$

Combining (20) and Proposition 2, we obtain

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^m) - \mathcal{J}(\boldsymbol{\theta}^{m+1}) &\geq \frac{2-L\eta}{2\eta}\|\boldsymbol{\gamma}^m\|^2 - \|\boldsymbol{e}^m\|\,\|\boldsymbol{\gamma}^m\|, \\ &\geq \frac{2-L\eta}{2\eta}\|\boldsymbol{\gamma}^m\|^2 - \eta(\varepsilon^m)^2\|\boldsymbol{\gamma}^m\|. \end{aligned}\tag{21}$$

Next, using Lemma 2 in (21), we have

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^m) - \mathcal{J}(\boldsymbol{\theta}^{m+1}) \geq \frac{2-L\eta}{2\eta}\left(\|\rho^m\| - \varepsilon^m\right)^2 \\ - \eta(\varepsilon^m)^2\left(\|\rho^m\| + \varepsilon^m\right).\end{aligned}$$

Rearranging teams in the above inequality, we have where

$$\begin{aligned} b_1 &= \frac{2-L\eta}{2\eta}, \\ b_2 &= \frac{2-L\eta+\eta^2}{\eta}, \\ \text{and}\quad b_3 &= \frac{2\eta^2+L\eta-2}{2\eta}.\end{aligned}$$

First Assertion. Define the following auxiliary function

$$\psi(\boldsymbol{\theta}^m) = b_1\|\rho^m\|^2 - b_2\|\rho^m\|\,\varepsilon - b_3(\varepsilon^m)^2.$$

With this definition, we have

$$\mathcal{J}(\boldsymbol{\theta}^m) - \mathcal{J}(\boldsymbol{\theta}^{m+1}) \geq \psi(\boldsymbol{\theta}^m). \tag{22}$$

It is easy to see that since $\|\rho^m\|$ and ε are continuous, b_1, b_2 and b_3 are non-negative, then $\psi(\boldsymbol{\theta}^m)$ is lower semi-continuous. Let the sub-level set of ψ be

$$\mathcal{L}(\psi, a) \equiv \{\boldsymbol{\theta} \mid \psi(\boldsymbol{\theta}) \leq a\}, \quad a \geq 0.$$

Note that the sub-level set of $\mathcal{L}(\psi, a)$ is closed for any $a \geq 0$ (see Theorem 7.1 in (Rockafellar, 1970)). Denote $u = \|\rho^m\|$, and resolving the quadratic inequality in u:

$$b_1u^2 - b_2\varepsilon^m u - b_3(\varepsilon^m)^2 - t \leq 0,$$

we conclude

$$u \leq \frac{b_2\varepsilon^m}{2b_1} + \frac{1}{2b_1}\sqrt{(b_2^2+4b_1b_3)(\varepsilon^m)^2 + 4b_1a}.$$

Thus,

$$\begin{aligned}\mathcal{L}(\psi, a) = \Big\{\boldsymbol{\theta} \mid \|\rho^m\| \leq \frac{b_2\varepsilon^m}{2b_1} \\ + \frac{1}{2b_1}\sqrt{(b_2^2+4b_1b_3)(\varepsilon^m)^2 + 4b_1a}\Big\},\end{aligned}$$

In particular,

$$\mathcal{L}(\psi, 0) = \left\{\boldsymbol{\theta} \mid \|\rho^m\| \leq \frac{b_2 + \sqrt{(b_2^2+4b_1b_3)}}{2b_1}\varepsilon\right\}.$$

Define

$$d_1 = \frac{b_2 + \sqrt{b_2^2+4b_1b_3}}{2b_1} \quad \text{and} \quad d_2 = b_1^{-1}.$$

We conclude

$$\mathcal{L}(\psi, a) \subseteq \left\{\boldsymbol{\theta} \mid \|\rho^m\| \leq d_1\varepsilon^m + d_2a^{\frac{1}{}}\right\},$$

and

$$\mathcal{L}(\psi, 0) \subseteq \{\boldsymbol{\theta} \mid \|\rho^m\| \leq d_1\varepsilon^m\}. \tag{23}$$

Next, we prove that there exists a limit point $\bar{\boldsymbol{\theta}}$ of $\{\boldsymbol{\theta}^m\}$ such that $\bar{\boldsymbol{\theta}} \in \mathcal{L}(\psi, 0)$. Suppose the opposite holds. By (22), we have

$$\mathcal{J}(\boldsymbol{\theta}^m) - \mathcal{J}(\boldsymbol{\theta}^{m+1}) \geq \psi(\boldsymbol{\theta}^m), \quad \forall m \geq m_1. \tag{24}$$

By Assumption $\lim_{m\to\infty}\{\boldsymbol{\theta}^m\} \cap \mathcal{L}(\psi, 0) = \emptyset$. Then, since ψ is lower semi-continuous, we have

$$\psi(\boldsymbol{\theta}^m) \geq c > 0,$$

when $m \geq m_2$ for some sufficiently large m_2 and a positive constant c.

Denote $k = \max\{m_1, m_2\}$, for any $m \geq k$, we have

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}^k) - \mathcal{J}(\boldsymbol{\theta}^m) &= \sum_{j=m}^{k}\left(\mathcal{J}(\boldsymbol{\theta}^j) - \mathcal{J}(\boldsymbol{\theta}^{j+1})\right), \\ &\geq (k-m)c.\end{aligned}$$

Let $k \to \infty$, we have $\lim_{m\to\infty}\mathcal{J}(\boldsymbol{\theta}^m) = -\infty$, which contradicts with Assumption 2, i.e., $\inf\mathcal{J} > -\infty$. Thus, from (23), for every limit point $\bar{\boldsymbol{\theta}}$ of $\{\boldsymbol{\theta}^m\}$, we must have

$$\|\rho^m\| \leq d_1\varepsilon^m.$$

Recall the definition of ρ^m in Lemma 2, and by Assumption 3 that the error ε^m on gradient is upper-bounded by $\bar{\varepsilon}$, we obtain the first assertion.

Second Assertion. If the sequence $\{\boldsymbol{\theta}^m\}$ converges, then for every sub-sequence $\{\boldsymbol{\theta}^{m_i}\}$ of $\{\boldsymbol{\theta}^m\}$ it follows

$$\lim_{i\to\infty} \sup \mathcal{J}(\boldsymbol{\theta}^{m_i}) = \lim_{m\to\infty} \inf \mathcal{J}(\boldsymbol{\theta}^m),$$

Thus,

$$\lim_{m\to\infty} \boldsymbol{\theta}^m \subseteq \mathcal{L}(\psi, 0), \tag{25}$$

where $\mathcal{L}(\psi, 0)$ is in (23). Combing (25) with the first assertion, we then have

$$\lim_{m\to\infty} \|\varepsilon^m\| \leq c_1\bar{\varepsilon}.$$

Finally, by (i) in Assumption 2 and definition of ε^m in Lemma 2, we have

$$\lim_{m\to\infty} \|\boldsymbol{e}^m\| \leq c_2\bar{\varepsilon},$$

for a positive constant c_2, which proves the second assertion. □