Supplementary Material to
Do We Really Need Gold Samples for Sample Weighting under Label Noise?

7. Convergence of Robust-Meta-Noisy-Weight-Network

We will detail the proof of convergence in Theorem 2 for our proposed method. Recall when we have clean meta samples, the meta loss is computed as

$$\mathcal{L}^{\text{meta}}(\mathbf{w}^*(\Theta)) = \frac{1}{M}\sum_{j=1}^{M} \ell^{\cdot,\text{meta}}(\mathbf{w}^*(\Theta)), \tag{7}$$

whereas for corrupted meta samples, the meta loss is computed as,

$$\mathcal{L}^{\text{noisy-meta}}(\mathbf{w}^*(\Theta)) = \frac{1}{M}\sum_{j=1}^{M} \ell^{\cdot,\text{noisy-meta}}(\mathbf{w}^*(\Theta)), \tag{8}$$

where $\mathbf{w}^*$ is the optimal classifier network, and Θ is the parameter of the weighting network. The classifier network is trained on the following objective,

$$\mathcal{L}^{\text{train}}(\mathbf{w};\Theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{W}(\ell_{\text{CE}}^{\cdot,\text{train}}(\mathbf{w});\mathbf{\Theta})\ell_{\text{CE}}^{\cdot,train}(\mathbf{w}). \tag{9}$$

We use the following lemma from [54] for proving convergence results.

Lemma 3. *Suppose the meta loss function is Lipschitz smooth with constant L, and $\mathcal{W}(\cdot)$ is differential with a δ-bounded gradient and twice differential with its Hessian bounded by $\mathcal{B}$, and the loss function $\ell_{CE}^{\cdot,train}$ have ρ-bounded gradients with respect to training/meta data. Then the gradient of Θ with respect to meta loss is Lipschitz continuous.*

Proof. Detailed proof can be found in [54]. □

In Theorem 1, we showed that the expectation of the meta-gradient remains same for corrupted meta samples under uniform noise model. Next, we bound the variance of meta-gradient under uniformly corrupted meta samples.

Lemma 4. *Suppose the meta loss function $\ell^{\cdot,meta}$ ($\ell^{\cdot,noisy\text{-}meta}$), satisfying symmetric condition in Eq. 6, have ρ-bounded gradients with respect to meta data. Let the variance of drawing a minibatch (of m samples) randomly is σ^2. Then the variance of the meta-gradients under uniformly corrupted meta samples (with rate η) is bounded by $\hat{\sigma}^2 = \sigma^2 + \frac{2\eta\rho^2}{m}$.*

Proof. Under clean meta samples, we have,

$$\begin{aligned}\xi^t &= \nabla\mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t} - \nabla\mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\ &= \nabla\mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t} - \mathcal{K}(\hat{\mathbf{w}}(\Theta^t))\end{aligned}$$

where the mini-batch of size m, ζ_t is drawn uniformly from the entire clean meta data set and $\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))$ is the unbiased meta-gradient. We also have $\mathbb{E}[||\xi^t||^2] = \sigma^2$ for clean meta dataset.

Under corrupted meta dataset, we have,

$$\begin{aligned}\xi^t &= \nabla\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t} - (1-\eta)\nabla\mathcal{L}^{meta}(\hat{\mathbf{w}}^t(\Theta^t)) \\ &= \nabla\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t} - (1-\eta)\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))\end{aligned}$$

since, we have shown in Theorem 1, meta-gradients of the corrupted meta dataset is upto a constant of the unbiased ones when the meta loss function satisfies symmetric condition. We note that for noisy meta dataset, the meta-gradient is,

$$\frac{1}{m}\sum_{j=1}^{m}\frac{\partial\ell^{j,\text{noisy-meta}}(\hat{\mathbf{w}}(\Theta^t))}{\partial\hat{\mathbf{w}}(\Theta^t)}\Big|_{\hat{\mathbf{w}}^t}$$

We compute variance for a single meta sample and then use the variance of sum of independent random variable rule to compute the final variance. Note, for a single sample, $\mathbb{E}[||\nabla\mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t} - \mathcal{K}(\hat{\mathbf{w}}(\Theta^t))||^2] = m\sigma^2$. Now we can compute the variance of corrupted meta-gradient when mini batch size is 1.

$$\begin{aligned}&\mathbb{E}_{\zeta_t,\eta_t}[||\xi^t||^2] = \mathbb{E}_{\zeta_t,\eta_t}[||\nabla\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t} \\ &\quad - (1-\eta)\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))||^2] \\ &= \mathbb{E}_{\zeta_t,\eta_t}\Big[\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t}^{\intercal}\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t} \\ &\quad + (1-\eta)^2\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))^{\intercal}\mathcal{K}(\hat{\mathbf{w}}(\Theta^t)) \\ &\quad - 2(1-\eta)\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))^{\intercal}\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t}\Big] \\ &= \mathbb{E}_{\zeta_t}\Big[(1-\eta)||\frac{\partial\ell^{\cdot,\text{meta}}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}}||^2 + (1-\eta)^2||\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))||^2 \\ &\quad - 2(1-\eta)^2\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))^{\intercal}\frac{\partial\ell^{\cdot,\text{meta}}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}} + \\ &\quad + \frac{2\eta}{K}\sum_c||\frac{\partial\ell^{\text{meta}}(c,f(\mathbf{x},\hat{\mathbf{w}}))}{\partial\hat{\mathbf{w}}}||^2 \\ &\quad - 2\frac{\eta(1-\eta)}{K}\mathcal{K}(\hat{\mathbf{w}}(\Theta^t))^{\intercal}\sum_c\frac{\partial\ell^{\text{meta}}(c,f(\mathbf{x},\hat{\mathbf{w}}))}{\partial\hat{\mathbf{w}}}\Big] \\ &\leq \mathbb{E}_{\zeta_t}\Big[||\nabla\mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t} - \mathcal{K}(\hat{\mathbf{w}}(\Theta^t))||^2\Big] \\ &\quad + \frac{2\eta}{K}\sum_c||\frac{\partial\ell^{\text{meta}}(c,f(\mathbf{x},\hat{\mathbf{w}}))}{\partial\hat{\mathbf{w}}}||^2 \\ &= m\sigma^2 + 2\eta\rho^2\end{aligned}$$

For a minibatch of size of m, the variance decreases as,

$$\mathbb{E}_{\zeta_t,\eta_t}[||\xi^t||^2] \leq \frac{m}{m^2}(m\sigma^2 + \eta\rho^2) = \sigma^2 + \frac{2\eta\rho^2}{m} \tag{10}$$

□

With Lemma 3 and 4, we can now prove Theorem 2. We will prove only for corrupted meta datasets. Proving convergence of clean meta dataset is easy and can be obtained by simply putting $\eta = 0$ in the noisy result.

Theorem 5. *Suppose the meta loss function ℓ and the classifier network loss $\ell_{CE}^{\cdot,train}$ is Lipschitz smooth with constant L, and have ρ-bounded gradients. The weighting function $\mathcal{W}(\cdot)$ has bounded gradient and twice differential with bounded Hessian. Let the classifier network learning rate α_t satisfies $\alpha_t = \min\{1, \frac{k}{T}\}$, for some $k > 0, k < T$. The learning rate of the weighting network satisfies $\beta_t = \min\{\frac{1}{L}, \frac{b}{\hat{\sigma}\sqrt{T}}\}$ for some $b > 0$, such that $\frac{\hat{\sigma}\sqrt{T}}{b} \geq L$ where $\hat{\sigma}^2$ is the variance of drawing a minibatch corrupted with noise. Then Robust-Meta-Noisy-Weight-Net can achieve $\mathbb{E}[\|\nabla \mathcal{L}^{noisy\text{-}meta}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2] \leq \epsilon$ in $\mathcal{O}(\frac{\hat{\sigma}^2}{(1-\eta)^2\epsilon^2})$ steps when meta loss function ℓ satisfies symmetric condition in Eq. 6. In particular,*

$$\min_{0\leq t\leq T} \mathbb{E}[\|\nabla \mathcal{L}^{noisy\text{-}meta}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2] \leq \mathcal{O}(\frac{\hat{\sigma}}{(1-\eta)\sqrt{T}}). \tag{11}$$

Proof. The update of Θ in each iteration is:

$$\Theta^{t+1} = \Theta^t - \beta_t \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t}.$$

We can rewrite the update equation as:

$$\Theta^{t+1} = \Theta^t - \beta_t[(1-\eta)\nabla \mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t)) + \xi^t],$$

where $\xi^t = \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\big|_{\zeta_t,\eta_t} - (1-\eta)\nabla \mathcal{L}^{\text{meta}}(\hat{\mathbf{w}}^t(\Theta^t))$. We have, $\mathbb{E}[\xi^t] = 0$, as we have shown in Theorem 1. We have,

$$\begin{aligned}
&\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\
&= \left\{\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1}))\right\} \\
&+ \left\{\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\right\}.
\end{aligned} \tag{12}$$

Since meta loss function is Lipschitz smooth, we have

$$\begin{aligned}
&\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1})) \\
&\leq \langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1})), \hat{\mathbf{w}}^{t+1}(\Theta^{t+1}) - \hat{\mathbf{w}}^t(\Theta^{t+1})\rangle \\
&\quad + \frac{L}{2}\|\hat{\mathbf{w}}^{t+1}(\Theta^{t+1}) - \hat{\mathbf{w}}^t(\Theta^{t+1})\|_2^2
\end{aligned}$$

Further, using the SGD update equation, we can write, $\hat{\mathbf{w}}^{t+1}(\Theta^{t+1}) - \hat{\mathbf{w}}^t(\Theta^{t+1}) = -\alpha_t \frac{1}{n}\sum_{i=1}^n \mathcal{W}(\ell_{\text{CE}}^{i,\text{train}}(\mathbf{w}^{t+1}); \Theta^{t+1}) \nabla_{\mathbf{w}} \ell_{\text{CE}}^{i,train}(\mathbf{w})\Big|_{\mathbf{w}^{t+1}}$. Thus, using the fact $\left\|\frac{\partial \ell_{\text{CE}}^{i,\text{train}}(\mathbf{w})}{\partial \mathbf{w}}\Big|_{\mathbf{w}^t}\right\| \leq \rho$, $\left\|\frac{\partial \ell^{j,\text{noisy-meta}}(\hat{\mathbf{w}})}{\partial \hat{\mathbf{w}}}\Big|_{\hat{\mathbf{w}}^t}^T\right\| \leq \rho$, we can bound,

$$\begin{aligned}
&\|\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1}))\| \\
&\leq \alpha_t\rho^2 + \frac{L\alpha_t^2}{2}\rho^2 = \alpha_t\rho^2(1 + \frac{\alpha_t L}{2})
\end{aligned}$$

Since meta loss function is Lipschitz continuous from Lemma 3, we get the following,

$$\begin{aligned}
&\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\
&\leq \langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \Theta^{t+1} - \Theta^t\rangle + \frac{L}{2}\|\Theta^{t+1} - \Theta^t\|_2^2 \\
&= \langle \nabla \mathcal{L}^{meta}(\hat{\mathbf{w}}^t(\Theta^t)), -\beta_t[(1-\eta)\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\
&\qquad + \xi^t]\rangle + \frac{L\beta_t^2}{2}\|(1-\eta)\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) + \xi^t\|_2^2 \\
&= -((1-\eta)\beta_t - (1-\eta)^2\frac{L\beta_t^2}{2})\|\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2 + \\
&\quad \frac{L\beta_t^2}{2}\|\xi^t\|_2^2 - (\beta_t - L(1-\eta)\beta_t^2)\langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \xi^t\rangle.
\end{aligned}$$

Using the above inequality, we bound Eq.(12) as,

$$\begin{aligned}
&\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\
&\leq \alpha_t\rho^2(1 + \frac{\alpha_t L}{2}) - ((1-\eta)\beta_t - (1-\eta)^2\frac{L\beta_t^2}{2}) \\
&\qquad \times \|\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2 + \frac{L\beta_t^2}{2}\|\xi^t\|_2^2 \\
&\qquad - (\beta_t - L(1-\eta)\beta_t^2)\langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \xi^t\rangle.
\end{aligned} \tag{13}$$

We can simplifiy as,

$$\begin{aligned}
&((1-\eta)\beta_t - (1-\eta)^2\frac{L\beta_t^2}{2})\|\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2 \\
&\leq \alpha_t\rho^2(1 + \frac{\alpha_t L}{2}) + \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)) \\
&\qquad - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) + \frac{L\beta_t^2}{2}\|\xi^t\|_2^2 \\
&\qquad - (\beta_t - L(1-\eta)\beta_t^2)\langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \xi^t\rangle.
\end{aligned}$$

We can simplify as,

$$\begin{aligned}
&\sum_{t=1}^T((1-\eta)\beta_t - (1-\eta)^2\frac{L\beta_t^2}{2})\|\nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2 \\
&\leq \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1)(\Theta^1) - \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^{t+1}(\Theta^{t+1})) \\
&\quad + \sum_{t=1}^T \alpha_t\rho^2(1 + \frac{\alpha_t L}{2}) - \sum_{t=1}^T(\beta_t - L(1-\eta)\beta_t^2) \\
&\quad \langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \xi^t\rangle + \frac{L}{2}\sum_{t=1}^T\beta_t^2\|\xi^t\|_2^2 \\
&\leq \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1)) + \sum_{t=1}^T \alpha_t\rho^2(1 + \frac{\alpha_t L}{2}) \\
&\quad - \sum_{t=1}^T(\beta_t - L(1-\eta)\beta_t^2)\langle \nabla \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t)), \xi^t\rangle \\
&\quad + \frac{L}{2}\sum_{t=1}^T\beta_t^2\|\xi^t\|_2^2,
\end{aligned} \tag{14}$$

Taking expectations with respect to ξ^N on both sides of Eq.

14, we get,

$$\sum_{t=1}^{T}((1-\eta)\beta_t-(1-\eta)^2\frac{L\beta_t^2}{2})\mathbb{E}_{\xi^N}\|\nabla\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2$$
$$\leq \mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))+\sum_{t=1}^{T}\alpha_t\rho^2(1+\frac{\alpha_t L}{2})+\frac{L\hat{\sigma}^2}{2}\sum_{t=1}^{T}\beta_t^2,$$

since $\mathbb{E}_{\xi^N}\langle\nabla\mathcal{L}^{\text{noisy-meta}}(\Theta^t),\xi^t\rangle=0$ and $\mathbb{E}[\|\xi^t\|_2^2]\leq\hat{\sigma}^2$ (using Lemma 4,) where $\hat{\sigma}^2$ is the variance with respect to ξ^t. Finally, we can obtain the the bound as,

$$\begin{aligned}
&\min_t \mathbb{E}[\|\nabla\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2]\\
&\leq \frac{\sum_{t=1}^{T}((1-\eta)\beta_t-(1-\eta)^2\frac{L\beta_t^2}{2})\mathbb{E}_{\xi^N}\|\nabla\mathcal{L}^{\eta}(\hat{\mathbf{w}}^t(\Theta^t))\|_2^2}{\sum_{t=1}^{T}((1-\eta)\beta_t-(1-\eta)^2\frac{L\beta_t^2}{2})}\\
&\leq \frac{1}{\sum_{t=1}^{T}(2(1-\eta)\beta_t-(1-\eta)^2L\beta_t^2)}\Big[2\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))\\
&\quad+\sum_{t=1}^{T}\alpha_t\rho^2(2+\alpha_t L)+L\hat{\sigma}^2\sum_{t=1}^{T}\beta_t^2\Big]\\
&\leq \frac{1}{\sum_{t=1}^{T}2(1-\eta)\beta_t}\Big[2\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))+\sum_{t=1}^{T}\alpha_t\rho^2(2+\alpha_t L)\\
&\quad+L\hat{\sigma}^2\sum_{t=1}^{T}\beta_t^2\Big]\\
&\leq \frac{1}{2T(1-\eta)\beta_t}\Big[2\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))+\alpha_1\rho^2T(2+L)\\
&\quad+L\hat{\sigma}^2\sum_{t=1}^{T}\beta_t^2\Big]\\
&=\frac{\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))}{(1-\eta)T}\frac{1}{\beta_t}+\frac{2\alpha_1\rho^2(2+L)}{2(1-\eta)\beta_t}+\frac{L\hat{\sigma}^2}{2(1-\eta)T}\sum_{t=1}^{T}\beta_t\\
&\leq\frac{\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))}{(1-\eta)T}\frac{1}{\beta_t}+\frac{2\alpha_1\rho^2(2+L)}{2(1-\eta)\beta_t}+L\hat{\sigma}^2\beta_t\frac{1}{2(1-\eta)}\\
&=\frac{\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1))}{(1-\eta)T}\max\{L,\frac{\hat{\sigma}\sqrt{T}}{b}\}\\
&\quad+\min\{1,\frac{k}{T}\}\max\{L,\frac{\hat{\sigma}\sqrt{T}}{b}\}\frac{\rho^2(2+L)}{2(1-\eta)}\\
&\quad+L\hat{\sigma}^2\min\{\frac{1}{L},\frac{b}{\hat{\sigma}\sqrt{T}}\}\frac{1}{2(1-\eta)}\\
&\leq\frac{\hat{\sigma}\mathcal{L}^{\text{noisy-meta}}(\hat{\mathbf{w}}^1(\Theta^1)}{(1-\eta)b\sqrt{T}}+\frac{k\hat{\sigma}\rho^2(2+L)}{b(1-\eta)\sqrt{T}}+\frac{L\hat{\sigma}b}{(1-\eta)\sqrt{T}}\\
&=\mathcal{O}(\frac{\hat{\sigma}}{(1-\eta)\sqrt{T}}).
\end{aligned}\tag{15}$$

The third inequality holds for $\sum_{t=1}^{T}(2\beta_t-L\beta_t^2)\geq\sum_{t=1}^{T}\beta_t$. Therefore, Robust-Meta-Noisy-Weight-Network can achieve $\min_{0\leq t\leq T}\mathbb{E}[\|\nabla\mathcal{L}^{\text{noisy-meta}}(\Theta^t)\|_2^2]\leq\mathcal{O}(\frac{\hat{\sigma}}{(1-\eta)\sqrt{T}})$ in T steps. □

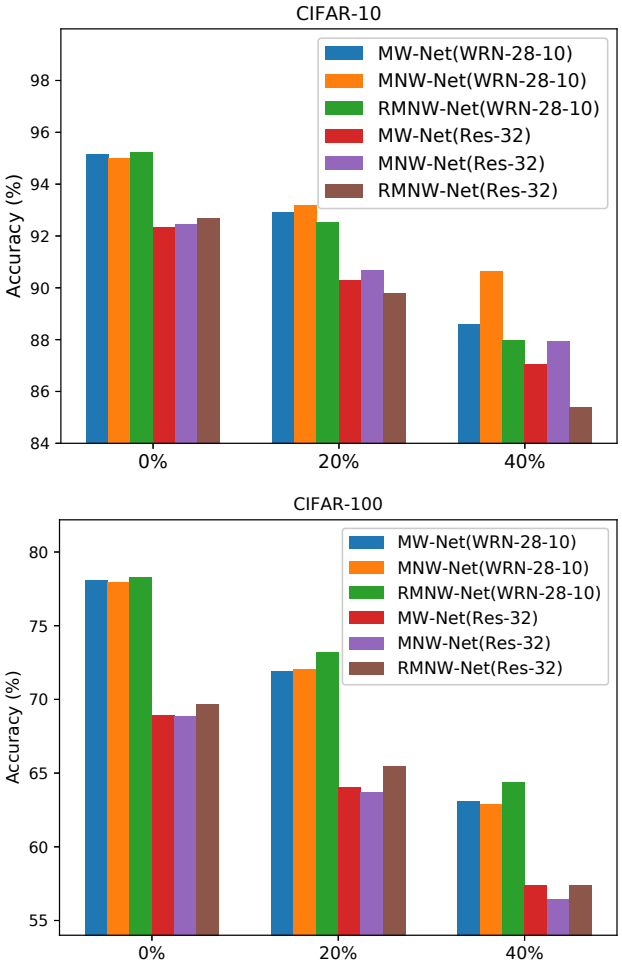

Figure 3. Performance comparison for different classifier architecture (WRN-28-10 and ResNet32) for CIFAR flip2 noise [54]

8. Additional Results

Indifference to network architecture on flip2 noise model In Table 1, we experimented with ResNet-32 architecture for flip2 noise model following [54]. For parity with uniform noise, we also experiment with WRN-28-10 architecture for flip2 noise. Figure 3 shows performances of MW-Net*, MNW-Net, and RMNW-Net on these two architectures for flip2 noise model. We observe similar trends on both architectures for the flip2 noise model. On CIFAR-10/100, MNW-Net/RMNW-Net performs marginally better than the other on both architectures.