A. Proofs

A.1. Proof of Proposition 1

Proof. The student's update rule parameterized by α is given by:

$$w^{(t+1)}(\alpha) = w^{(t)} - \eta_w \nabla_w \widetilde{\mathcal{L}}(w, \alpha)\Big|_{w=w^{(t)}} \tag{13}$$

Differentiating eq. (13) w.r.t α yields:

$$\frac{dw^{(t+1)}}{d\alpha}\Big|_{\alpha=\alpha^{(t)}} = -\eta_w \nabla_{w\alpha} \widetilde{\mathcal{L}}(w, \alpha)\Big|_{(w^{(t)},\alpha^{(t)})} \tag{14}$$

We now can compute the meta-gradient explicitly using the chain rule as follows:

$$\frac{d\mathcal{L}}{d\alpha}\Big|_{\alpha=\alpha^{(t)}} = \frac{d\mathcal{L}}{dw^{(t+1)}}\Big|_{w^{(t+1)}=w^{(t+1)}} \cdot \frac{dw^{(t+1)}}{d\alpha}\Big|_{\alpha=\alpha^{(t)}} \tag{15}$$

Where $w^{(t+1)}$ is obtained from the updated student. Substituting eq. (14) we obtain the desired result. □

A.2. Proof of Proposition 2

Proof. In the case $k > 1$, we have:

$$w^{(t-k+2)}(\alpha) = w^{(t-k+1)} - \eta_w \nabla_w \widetilde{\mathcal{L}}(w^{(t-k+1)}, \alpha) \tag{16}$$

Recall that we neglect the dependency of $w^{(t-k+1)}$ on α. In addition, for all $t-k+2 \leq \tau \leq t$:

$$w^{(\tau+1)}(\alpha) = w^{(\tau)}(\alpha) - \eta_w \nabla_w \widetilde{\mathcal{L}}(w^{(\tau)}(\alpha), \alpha) \tag{17}$$

Recall also that:

$$\forall t-k+1 \leq \tau \leq t : \alpha^{(\tau)} = \alpha^{(t)} \tag{18}$$

We denote:

$$H^{(\tau)}_{w\alpha} := \nabla_{w\alpha} \widetilde{\mathcal{L}}(w^{(\tau)}, \alpha^{(\tau)}) \tag{19}$$

$$H^{(\tau)}_{ww} := \nabla_{ww} \widetilde{\mathcal{L}}(w^{(\tau)}, \alpha^{(\tau)}) \tag{20}$$

deriving eq. (17) w.r.t α (considering the derivative at $\alpha = \alpha^{(t)}$), using the chain rule again and substituting eq. (18) yields:

$$\frac{dw^{(\tau+1)}}{d\alpha} = \frac{dw^{(\tau)}}{d\alpha} - \eta_w [H^{(\tau)}_{ww} \cdot \frac{dw^{(\tau)}}{d\alpha} + H^{(\tau)}_{w\alpha}] \tag{21}$$

We rewrite the above equation as:

$$\frac{dw^{(\tau+1)}}{d\alpha} = [I - \eta_w H^{(\tau)}_{ww}] \cdot \frac{dw^{(\tau)}}{d\alpha} - \eta_w H^{(t)}_{w\alpha} \tag{22}$$

If we now approximate $H^{(\tau)}_{ww} \approx I$ we get:

$$\frac{dw^{(\tau+1)}}{d\alpha} = (1-\eta_w)\frac{dw^{(\tau)}}{d\alpha} - \eta_w H^{(\tau)}_{w\alpha} \tag{23}$$

By setting $\gamma_w = 1 - \eta_w$, opening up the recursion formula and using eq. (16) at the end of the recursion, we get the desired result.

□

A.3. Proof of Proposition 3

Proof. This follows from the definition of cross-entropy and simple differentiation rules. Indeed,

$$\widetilde{\ell}_i(w, \alpha) = CE(q_\alpha(y|x_i, \widetilde{y}_i), p_w(y|x_i)) = \tag{24}$$

$$-\sum_{c=1}^{C}(q_\alpha(y=c|x_i, \widetilde{y}_i) \cdot \log(p_w(y=c|x_i)))) = \tag{25}$$

$$-\langle q_\alpha(y|x_i, \widetilde{y}_i), \log p_w(y|x_i))\rangle \tag{26}$$

Now, for two general differential functions $f(w) : \mathbb{R}^W \to \mathbb{R}^C$, and $g(\alpha) : \mathbb{R}^A \to \mathbb{R}^C$ consider $h(w, \alpha) = \langle f(w), g(\alpha)\rangle$. Then:

$$\nabla_w h(w, \alpha) = \nabla_w \langle f(w), g(\alpha)\rangle = (g(\alpha))^T J_w(f) \tag{27}$$

Differentiating both sides of the equation w.r.t α yields:

$$\nabla_{w\alpha} h(w, \alpha) = \frac{d\nabla_w h}{dg} \cdot \frac{dg}{d\alpha} = [J_w(f)]^T [J_\alpha(g)] \tag{28}$$

And hence $\nabla_{w\alpha} h(w, \alpha)$ exists and equals to $[J_w(f)]^T[J_\alpha(g)]$. Substituting f, g from eq. (26) in the above equation yields the desired result. □

B. Comparison of the Meta Gradient Computation

We compare the differences of our FPMG algorithm and MLC [41] in terms of the quality and efficiency of the meta-gradient computation in table 3. Prominently, FPMG avoids computing second order derivatives which yield a large memory and computation overhead.

C. Additional Experiments

C.1. Ablation studies

We perform an ablation study on the different components of our method. Each setting is validated on the Clothing1M dataset. The results are summarized in table 4.

Criterion	FPMG	MLC
Exact GD recursion	✓	✗
Exact mixed Hessian $H_{w\alpha}$	✓	✓
Avoids second-order derivative	✓	✗
Approximation of H_{ww}	$H_{ww} \approx I$	$H_{ww} \approx I$

Table 3: Comparison of FPMG and MLC [41] regarding the meta-gradient quality and efficiency of the computation.

Meta Regularization	Corruption Strategy	Strong Augmentations	Accuracy
✗	✗	✗	74.35
✓	✗	✗	77.52
✓	Rand.	✗	78.51
✓	Adv.	✗	78.60
✓	Adv.	✓	**79.35**

Table 4: Ablation study on the distinct components of EMLC. We validate the effectiveness of each combination by considering test accuracy (%) on the Clothing1M dataset. We verify the effectiveness of the meta-learning regularization, the effectiveness of the proposed proactive noise injection and the benefit of applying AutoAugment to the labeled data. The corruption strategy values might be either none (✗– in which case, the BCE loss is omitted), random (rand.) or adversarial (adv.).

As can be observed from table 4, our approach possesses a very strong baseline. Each of our proposed components incrementally improves the effectiveness, as expected. Notably, it can be observed that the meta-regularization and the artificial corruption were crucial for the success of our method.

Regarding the number of look-ahead steps, we perform two ablative experiments. Following the finding that the multi-step strategy outclassed the single-step strategy in some of the CIFAR experiments, we perform an ablation study on the number of steps in the multi-step strategy on the CIFAR-100 dataset with 50% symmetric noise and a fixed seed. As can be observed from the results in fig. 6, $k = 5$ arbitrated to be the best. In addition, due to the importance of the Clothing1M dataset, we perform an additional ablation study to demonstrate the robustness of our method to varying number of look-ahead steps on a real-world dataset, as presented in fig. 7.

C.2. Teacher's Label Recovery

To further verify the teacher's ability to cleanse the training labels, we compare the teacher's label recovery rate

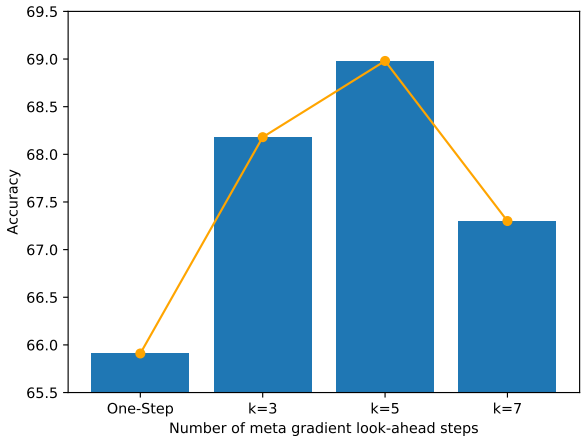

Figure 6: Ablation study on the number of look-ahead steps, measuring the effectiveness on the CIFAR-100 dataset with 50% symmetric noise.

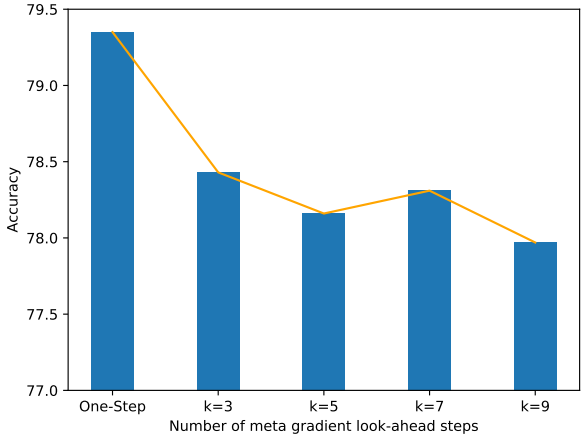

Figure 7: Ablation study on the number of look-ahead steps, measuring the effectiveness on the Clothing1M dataset.

(total and wrongly annotated samples) on the CIFAR-10 dataset with different noise levels of EMLC against MLC [41] and MLC with self-supervised pretraining in fig. 8.

D. A probabilistic interpretation of the teacher

The goal of the teacher is to model the conditional distribution of the true label y given the sample x and its noisy label $\tilde{y}$, namely $p(y|x, \tilde{y})$.

The above conditional distribution can be decomposed as follows:

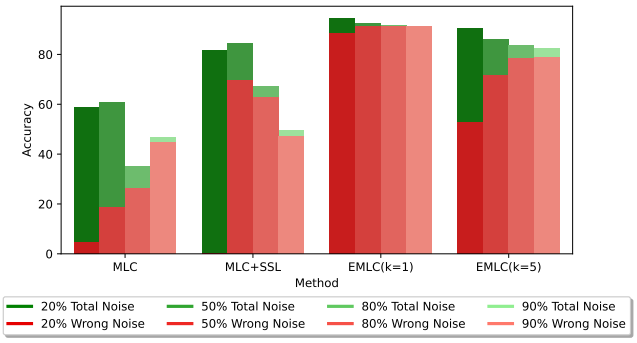

Figure 8: Comparison of the teacher's label recovery rate (total and wrongly annotated samples) on the CIFAR-10 dataset with different noise levels.

$$p(y|x,\tilde{y}) = \tag{29}$$
$$w \cdot p(y|x,\tilde{y},\mathcal{A}) + (1-w) \cdot p(y|x,\tilde{y},\mathcal{A}^c) = \tag{30}$$
$$w \cdot \delta_{\tilde{y}y} + (1-w) \cdot p(y|x,\tilde{y},\mathcal{A}^c) \tag{31}$$

where $\mathcal{A}$ is the event of $y = \tilde{y}$ and $w := p(\mathcal{A}|x,\tilde{y})$.

In our teacher architecture, we model w directly. However, we approximate $p(y|x,\tilde{y},\mathcal{A}^c) \approx p(y|x,\mathcal{A}^c) \approx p(y|x)$ and model $p(y|x)$ instead. In the first approximation, we throw away the conditioning on $\tilde{y}$, intuitively ignoring the cases in which $y|x$ is dependent of $\tilde{y}$ whenever the label is corrupted. While the second approximation is technically unnecessary (as $p(y|x,\mathcal{A}^c)$ can be easily modeled by masking the *SoftMax* layer), we found out that the latter modeling was better in practice since it effectively enhances the weight of hard clean samples.

E. Further Details of the Experimental Setting

E.1. Symmetric and Asymmetric Artificial Noise

In the *Symmetric noise* setting, a fraction of ϵ samples are chosen randomly. The samples are then assigned with a random label selected uniformly over the classes. The expected fraction of *wrong samples* is effectively smaller than ϵ. This noise definition is convenient, as $\epsilon = 1$ would mean that the label assignment is entirely random. In the *Asymmetric noise* setting proposed by Partini *et al*. [23], a fraction of ϵ samples are chosen randomly. Then, their label is replaced by a label of a visually similar category, to better model real-world noise. In the CIFAR-10 dataset, this is done by applying the following transitions: TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, CAT $\leftrightarrow$ DOG. In the CIFAR-100 dataset, each label is changed to its successor(circularly) in its super-class.

E.2. Implementation Details and Hyperparameters

Further experimental details are presented in table 5.

Dataset	CIFAR-10	CIFAR-100	Clothing1M
Architecture	ResNet-34	ResNet-50	ResNet-50
Pretraining	SimCLR	SimCLR	ImageNet
MLP hidden layers	1	1	1
MLP hidden units	128	128	128
Noisy batch size	128	128	1024
Clean batch size	32	32	1024
Epochs	15	15	3
Noisy augmentations	Horizontal Flip Random Crop	Horizontal Flip Random Crop	None
Clean augmentations	Horizontal Flip Random Crop CIFAR AutoAugment	Horizontal Flip Random Crop CIFAR AutoAugment	Horizontal Flip ImageNet AutoAugment
Optimizer	SGD	SGD	SGD
Scheduler	None	None	LR-step
Momentum	0.9	0.9	0.9
Weight decay	0	0	0
Initial LR	0.02	0.02	0.1
Number of GPUs	1	1	8

Table 5: Experimental setup and hyper-parameters.