To Reviewer #1:

Q1.1: MLP architecture of Meta-Weight-Net. We actually have tried different MLP architecture settings in experiments. The right table depicts some representative results under 6 different structures, with different depths and widths. It can be seen that varying MLP settings have unsubstantial effects to the final result. We thus prefer to use the simple and shallow one.

Table 1: Test accuracy on CIFAR-10 and CIFAR-100 of different MW-Nets.

architcture	Imbalance (factor 100)		Uniform noise (40%)		Flip noise (40%)	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
1-50-1	73.50	41.87	89.01	67.63	87.38	57.83
1-100-1	75.21	42.09	89.27	67.73	87.54	58.64
1-200-1	74.70	41.72	89.58	67.84	87.74	58.41
1-100-100-1	75.01	41.97	89.09	66.48	87.28	57.39
1-10-10-1	74.71	41.94	89.10	66.53	87.58	57.11
1-10-10-10-1	74.96	42.31	88.82	66.67	87.36	57.29

To Reviewer #2:

Q2.1: How the choices of weight function influence the results? Succeeded from the understanding of conventional sample reweighting approaches, this explicit weighting function is set as mapping from loss to weight, and thus MLP is suitable. Instead, since LSTM is functioned on temporal feature input, it is not proper to be used here. As introduced in Q1.1, we have also tested different structures for MW-Net (with different depths and widths), which have only unsubstantial influence to the final result.

Q2.2: Experiment results with more training epochs. In our experiments, we have tried to specifically set the epoch number for each compared method to guarantee possibly the optimal performance. Actually, we have shown in Fig. 1(a) of SM the performance tendency of our method with more than 100 epochs. It is easy to see the convergence of our method after about 40 epochs. Similar phenomena have been observed from all our experiments. Comparatively, most of other methods could get the best performance before 100 epochs, while the state-of-the-art L2RW needs more than 100 epochs, as shown in Fig. 1(a) of SM as well as Fig. 6 of the paper. This supports us to say that our method converges relatively faster. We'll add more results in revision for more clarification.

To Reviewer #3:

Q3.1: Definition of $\mathcal{L}^{meta}(\Theta)$ and the proof of the inequality after line 47 (supp). We sincerely thank the reviewer for pointing this out. The function $\mathcal{L}^{(meta)}$ does depend on $\mathbf{w}$, and thus it should be inappropriate to neglect the symbol $\mathbf{w}$. Specifically, in our algorithm (Algorithm 1), we use one step gradient descend result $\hat{\mathbf{w}}^{(t)}(\Theta)$ as the variable in $\mathcal{L}^{meta}$ function, and $\Theta^{(t+1)}$ and $\Theta^{(t)}$ appearing between Line 47-48 of SM do be evaluated with different $\mathbf{w}$s, which should be under $\hat{\mathbf{w}}^{(t+1)}$ and $\hat{\mathbf{w}}^{(t)}$, respectively, just as the reviewer properly indicates. The deduction under line 47 thus should be rectified as follows:

$$\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t+1)}(\Theta^{(t+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t)}))=\{\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t+1)}(\Theta^{(t+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t+1)}))\}+\{\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t)}))\}.$$

The deduction below line 47 in SM actually deduces the upper bound of the above second term (difference of $\mathcal{L}^{(meta)}$ under the same $w^{(t)}$). That is, let $Re=-(\beta_t-\frac{L\beta_t^2}{2})\|\nabla\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t)}))\|_2^2+\frac{L\beta_t^2}{2}\|\xi^{(t)}\|_2^2-(\beta_t-L\beta_t^2)\langle\nabla\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t)})),\xi^{(t)}\rangle$ we then have

$$\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t+1)}(\Theta^{(t+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t)}))\leq\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t+1)}(\Theta^{(t+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta^{(t+1)}))+Re.$$

Summing up the above inequalities and rearranging the terms, we can obtain

$$\textstyle\sum_{t=1}^{T}(\beta_t-\frac{L\beta_t^2}{2})\|\nabla\mathcal{L}^{meta}(\Theta^{(t)})\|_2^2\leq\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(T+1)}(\Theta^{(T+1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(2)}(\Theta^{(1)}))+\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(1)})(\Theta^{(1)})-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(T+1)}(\Theta^{(T+1)}))$$
$$\textstyle-\sum_{t=1}^{T}(\beta_t-L\beta_t^2)\langle\nabla\mathcal{L}^{meta}(\Theta^{(t)}),\xi^{(t)}\rangle+\frac{L}{2}\sum_{t=1}^{T}\beta_t^2\|\xi^{(t)}\|_2^2=\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(1)}(\Theta^{(1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(2)}(\Theta^{(1)}))-\sum_{t=1}^{T}(\beta_t-L\beta_t^2)\langle\nabla\mathcal{L}^{meta}(\Theta^{(t)}),\xi^{(t)}\rangle+\frac{L}{2}\sum_{t=1}^{T}\beta_t^2\|\xi^{(t)}\|_2^2,$$

This is almost similar to (14), with only $\mathcal{L}^{meta}(\Theta^{(1)})-\mathcal{L}^{meta}(\Theta^{*})$ replaced by $\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(1)}(\Theta^{(1)}))-\mathcal{L}^{meta}(\hat{\mathbf{w}}^{(2)}(\Theta^{(1)}))$. So do the following inequalities (15)(16). We'll revise the proof accordingly to avoid possible confusions of readers.

Q3.2: Prove that $\mathcal{L}^{(meta)}$ is Lipshitz smooth in lemma 1. Many thanks to the reviewer for carefully checking our proof. To guarantee that Eq.(9) holds based on the proof already being presented, we do need to additionally prove $\nabla^2_{\Theta^2}L_i^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta))\big|_{\Theta^{(t)}}$ is bounded, which needs another two mild conditions: the meta loss function is Lipschitz smooth with constant L, and $\mathcal{V}(\cdot)$ is a twice differential with its Hessian bounded by $\mathcal{B}$. The proof is then presented as follows: Let $\mathcal{V}_j(\Theta)=\mathcal{V}(L_j^{train}(\mathbf{w}^{(t)});\Theta)$ and G_{ij} being defined in line 18 of SM, taking gradient of Θ in both sides of (6), we have

$$\nabla^2_{\Theta^2}L_i^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta))\big|_{\Theta^{(t)}}=\frac{-\alpha}{n}\textstyle\sum_{j=1}^{n}\left[\frac{\partial}{\partial\Theta}(G_{ij})\big|_{\Theta^{(t)}}\frac{\partial\mathcal{V}_j(\Theta)}{\partial\Theta}\big|_{\Theta^{(t)}}+(G_{ij})\frac{\partial^2\mathcal{V}_j(\Theta)}{\partial\Theta^2}\big|_{\Theta^{(t)}}\right].$$

For the first term in the right hand side, we have that $\left\|\frac{\partial}{\partial\Theta}(G_{ij})\big|_{\Theta^{(t)}}\frac{\partial\mathcal{V}_j(\Theta)}{\partial\Theta}\big|_{\Theta^{(t)}}\right\|\leq\rho\left\|\frac{\partial}{\partial\hat{\mathbf{w}}}\left(\frac{\partial L_i^{meta}(\hat{\mathbf{w}})}{\partial\Theta}\big|_{\Theta^{(t)}}\right)\big|_{\hat{\mathbf{w}}^{(t)}}^{T}\frac{\partial L_j^{train}(\mathbf{w})}{\partial\mathbf{w}}\big|_{\mathbf{w}^{(t)}}\right\|$ $=\rho\left\|\frac{\partial}{\partial\hat{\mathbf{w}}}\left(\frac{\partial L_i^{meta}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}}\big|_{\hat{\mathbf{w}}^{(t)}}\frac{-\alpha}{n}\sum_{j=1}^{n}\frac{\partial L_j^{train}(\mathbf{w})}{\partial\mathbf{w}}\big|_{\mathbf{w}^{(t)}}\frac{\partial\mathcal{V}_j(\Theta)}{\partial\Theta}\big|_{\Theta^{(t)}}\right)\right\|=\rho\left\|\frac{\partial^2 L_i^{meta}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}^2}\big|_{\hat{\mathbf{w}}^{(t)}}\frac{-\alpha}{n}\sum_{j=1}^{n}\frac{\partial L_j^{train}(\mathbf{w})}{\partial\mathbf{w}}\big|_{\mathbf{w}^{(t)}}\frac{\partial\mathcal{V}_j(\Theta)}{\partial\Theta}\big|_{\Theta^{(t)}}\right\|\leq\alpha L\rho^3,$ since $\left\|\frac{\partial^2 L_i^{meta}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}^2}\big|_{\hat{\mathbf{w}}^{(t)}}\right\|\leq L,\left\|\frac{\partial L_j^{train}(\mathbf{w})}{\partial\mathbf{w}}\big|_{\mathbf{w}^{(t)}}\right\|\leq\rho,\left\|\frac{\partial\mathcal{V}_j(\Theta)}{\partial\Theta}\big|_{\Theta^{(t)}}\right\|\leq\rho$. And for the second term $\left\|(G_{ij})\frac{\partial^2\mathcal{V}_j(\Theta)}{\partial\Theta^2}\big|_{\Theta^{(t)}}\right\|=\left\|\frac{\partial L_i^{meta}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}}\big|_{\hat{\mathbf{w}}^{(t)}}^{T}\frac{\partial L_j^{train}(\mathbf{w})}{\partial\mathbf{w}}\big|_{\mathbf{w}^{(t)}}\frac{\partial^2\mathcal{V}_j(\Theta)}{\partial\Theta^2}\big|_{\Theta^{(t)}}\right\|\leq\mathcal{B}\rho^2$, since $\left\|\frac{\partial L_i^{meta}(\hat{\mathbf{w}})}{\partial\hat{\mathbf{w}}}\big|_{\hat{\mathbf{w}}^{(t)}}^{T}\right\|\leq\rho,\left\|\frac{\partial^2\mathcal{V}_j(\Theta)}{\partial\Theta^2}\big|_{\Theta^{(t)}}\right\|\leq\mathcal{B}$. Combining the above two inequalities, we have $\left\|\nabla^2_{\Theta^2}L_i^{meta}(\hat{\mathbf{w}}^{(t)}(\Theta))\big|_{\Theta^{(t)}}\right\|\leq\alpha(\alpha L\rho^3+\mathcal{B}\rho^2)$. Therefore, using Lagrange's mean value theorem, Eq. (9) holds.

Q3.3: The proof of theorem 2 at line 71 in the SM. We have skipped several steps, and the detailed proof is as follows: Taking expectation of both sides of (21) and since $\mathbb{E}[\psi^{(t)}]=0$ (line 66-68 of SM), we have

$$\mathbb{E}[\mathcal{L}^{train}(\mathbf{w}^{(t+1)};\Theta^{(t+1)})]-\mathbb{E}[\mathcal{L}^{train}(\mathbf{w}^{(t)};\Theta^{(t+1)})]\leq-\alpha_t\mathbb{E}[\|\nabla\mathcal{L}^{train}(\mathbf{w}^{(t)};\Theta^{(t+1)})\|_2^2]+\frac{L\alpha_t^2}{2}\{\mathbb{E}[\|\nabla\mathcal{L}^{train}(\mathbf{w}^{(t)};\Theta^{(t+1)})\|_2^2]+\mathbb{E}[\|\psi^{(t)}\|_2^2]\}$$

Summing up the above inequalities over $t=1,...,\infty$ in both sides, we obtain (There exists a typo at line 71, $\|\cdot\|$ should be $\|\cdot\|_2^2$)

$$\sum_{t=1}^{\infty}\alpha_t\mathbb{E}[\|\nabla\mathcal{L}^{tr}(\mathbf{w}^{(t)};\Theta^{(t+1)})\|_2^2]\leq\sum_{t=1}^{\infty}\frac{L\alpha_t^2}{2}\{\mathbb{E}[\|\nabla\mathcal{L}^{tr}(\mathbf{w}^{(t)};\Theta^{(t+1)})\|_2^2]+\mathbb{E}[\|\psi^{(t)}\|_2^2]\}+\mathbb{E}[\mathcal{L}^{tr}(\mathbf{w}^{(1)};\Theta^{(2)})]-\lim_{T\to\infty}\mathbb{E}[\mathcal{L}^{tr}(\mathbf{w}^{(t+1)};\Theta^{(t+1)})]\leq\sum_{t=1}^{\infty}\frac{L\alpha_t^2}{2}\{\rho^2+\sigma^2\}+\mathbb{E}[\mathcal{L}^{tr}(\mathbf{w}^{(1)};\Theta^{(2)})]\leq\infty,$$

where tr is short for $train$ to save space, since $\mathbb{E}[\|\nabla\mathcal{L}^{train}(\mathbf{w}^{(t)};\Theta^{(t+1)})\|_2^2\leq\rho^2,\mathbb{E}[\|\psi^{(t)}\|_2^2]\leq\sigma^2$.

Q3.4: Some typos and a different bound for $\mathcal{L}^{(meta)}$ instead of ρ. Yes, there is a L_2 norm in the sixth line of Eq. (24), and we'll modify this and other typos in revision. We assume the gradients with respect to training/meta data are both ρ-bounded for avoiding the abuse of symbols.