# L2B: Learning to Bootstrap Robust Models for Combating Label Noise

## Supplementary Material

## 6. Appendix

### 6.1. Normalization function comparision.

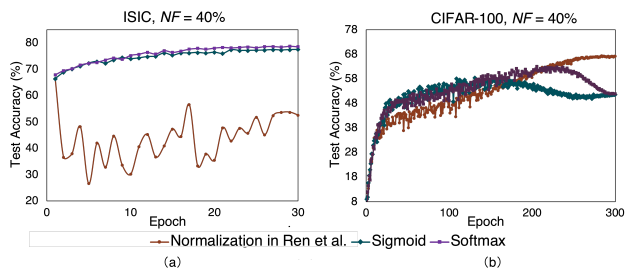


Figure 3. Comparison among different normalization functions (i.e., Eq. 9, Sigmoid function and Softmax function). Testing accuracy curve: (a) with different normalization functions under 40% symmetric noise label on the ISIC dataset. (b) with different normalization under 40% symmetric label noise on CIFAR-100.

### 6.2. Alleviate potential overfitting to noisy examples.

We also plot the testing accuracy curve under different noise fractions in Figure 4, which shows that our proposed L2B would help preventing potential overfitting to noisy samples compared with standard training. Meanwhile, compared to simply sample reweighting (L2RW), our L2B introduces pseudo-labels for bootstrapping the learner and is able to converge to a better optimum.

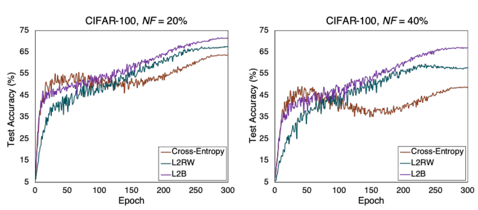


Figure 4. Test accuracy v.s. number of epochs on CIFAR-100 under the noise fraction of 20% and 40%.

## 7. Theoretical Analysis

### 7.1. Equivalence of the two learning objectives

We show that Eq. 3 is equivalent with Eq. 2 when $\forall i\ \alpha_i + \beta_i = 1$. For convenience, we denote $y_i^{\text{real}}, y_i^{\text{pseudo}}, \mathcal{F}(x_i, \theta)$

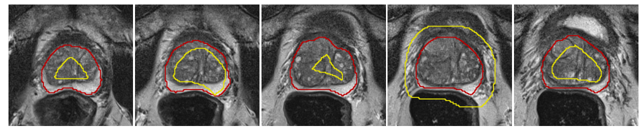

Figure 5. Visual comparison of prostate MRI images with noisy (contoured in yellow) and accurate (contoured in red) segmentation masks to demonstrate the discrepancy in segmentation quality between the two.

using $y_i^r, y_i^p, p_i$ respectively.

$$\alpha_i \mathcal{L}(p_i, y_i^r) + \beta_i \mathcal{L}(p_i, y_i^p) = \sum_{l=1}^{L} \alpha_i y_{i,l}^r \log p_{i,l} \tag{11}$$

$$+ \beta_i y_{i,l}^p \log p_{i,l} = \sum_{l=1}^{L} (\alpha_i y_{i,l}^r + \beta_i y_{i,l}^p) \log p_{i,l} \tag{12}$$

Due to that $\mathcal{L}(\cdot)$ is the cross-entropy loss, we have $\sum_{l=1}^{L} y_{i,l}^r = \sum_{l=1}^{L} y_{i,l}^p = 1$. Then $\sum_{l=1}^{L} \alpha_i y_{i,l}^r + \beta_i y_{i,l}^p = \alpha_i + \beta_i$. So if $\alpha_i + \beta_i = 1$, we have

$$\sum_{l=1}^{L} (\alpha_i y_{i,l}^r + \beta_i y_{i,l}^p) \log p_{i,l} = \mathcal{L}(p_i, \alpha_i y_i^r + \beta_i y_i^p) \tag{13}$$

$$= \mathcal{L}(p_i, (1 - \beta_i) y_i^r + \beta_i y_i^p) \tag{14}$$

### 7.2. Gradient used for updating $\theta$

We derivative the update rule for $\boldsymbol{\alpha}, \boldsymbol{\beta}$ in Eq. 10.

$$\alpha_{t,i} = -\eta \frac{\partial}{\partial \alpha_i} \Big(\sum_{j=1}^{m} f_j^v(\hat{\theta}_{t+1})\Big)\Big|_{\alpha_i=0} \tag{15}$$

$$= -\eta \sum_{j=1}^{m} \nabla f_j^v(\hat{\theta}_{t+1})^T \frac{\partial \hat{\theta}_{t+1}}{\partial \alpha_i}\Big|_{\alpha_i=0} \tag{16}$$

$$= -\eta \sum_{j=1}^{m} \nabla f_j^v(\hat{\theta}_{t+1})^T \tag{17}$$

$$\frac{\partial(\theta_t - \lambda \nabla(\sum_k \alpha_k\, f_k(\theta) + \beta_k\, g_k(\theta))\big|_{\theta=\theta_t})}{\partial \alpha_i}\Big|_{\alpha_i=0} \tag{18}$$

$$= \eta\lambda \sum_{j=1}^{m} \nabla f_j^v(\theta_t)^T \nabla f_i(\theta_t) \tag{19}$$

$$\beta_{t,i} = -\eta \frac{\partial}{\partial \beta_i}(\sum_{j=1}^{m} f_j^v(\hat{\theta}_{t+1}))\Big|_{\beta_i=0} \tag{20}$$

$$= -\eta \sum_{j=1}^{m} \nabla f_j^v(\hat{\theta}_{t+1})^T \frac{\partial \hat{\theta}_{t+1}}{\partial \beta_i}\Big|_{\beta_i=0} \tag{21}$$

$$= -\eta \sum_{j=1}^{m} \nabla f_j^v(\hat{\theta}_{t+1})^T \tag{22}$$

$$\frac{\partial(\theta_t - \lambda\nabla(\sum_k \alpha_k\, g_k(\theta) + \beta_k\, g_k(\theta))\Big|_{\theta=\theta_t})}{\partial \beta_i}\Big|_{\beta_i=0} \tag{23}$$

$$= \eta\lambda \sum_{j=1}^{m} \nabla f_j^v(\theta_t)^T \nabla g_i(\theta_t) \tag{24}$$

Then $\theta_{t+1}$ can be calculated by Eq. 10 using the updated $\alpha_{t,i}, \beta_{t,i}$.

### 7.3. Convergence

This section provides the proof for covergence (Section 3.3).

**Theorem.** *Suppose that the training loss function $f, g$ have $\sigma$-bounded gradients and the validation loss $f^v$ is Lipschitz smooth with constant $L$. With a small enough learning rate $\lambda$, the validation loss monotonically decreases for any training batch $B$, namely,*

$$G(\theta_{t+1}) \leq G(\theta_t), \tag{25}$$

*where $\theta_{t+1}$ is obtained using Eq. 10 and $G$ is the validation loss*

$$G(\theta) = \frac{1}{M}\sum_{i=1}^{M} f_i^v(\theta), \tag{26}$$

*Furthermore, Eq. 25 holds for all possible training batches only when the gradient of validation loss function becomes 0 at some step $t$, namely, $G(\theta_{t+1}) = G(\theta_t)\ \forall B \Leftrightarrow \nabla G(\theta_t) = 0$*

*Proof.* At each training step $t$, we pick a mini-batch $B$ from the union of training and validation data with $|B| = n$. From section B we can derivative $\theta_{t+1}$ as follows:

$$\theta_{t+1} = \theta_t - \lambda \sum_{i=1}^{n} (\alpha_{t,i}\nabla f_i(\theta_t) + \beta_{t,i}\nabla g_i(\theta_t)) \tag{27}$$

$$= \theta_t - \eta\lambda^2 M \sum_{i=1}^{n} (\nabla G^T \nabla f_i \nabla f_i + \nabla G^T \nabla g_i \nabla g_i) \tag{28}$$

We omit $\theta_t$ after every function for briefness and set $m$ in section B equals to $M$. Since $G(\theta)$ is Lipschitz-smooth, we have

$$G(\theta_{t+1}) \leq G(\theta_t) + \nabla G^T \Delta\theta + \frac{L}{2}||\Delta\theta||^2. \tag{29}$$

Then we show $\nabla G^T \Delta\theta + \frac{L}{2}||\Delta\theta||^2 \leq 0$ with a small enough $\lambda$. Specifically,

$$\nabla G^T \Delta\theta = -\eta\lambda^2 M \sum_i (\nabla G^T \nabla f_i)^2 + (\nabla G^T \nabla g_i)^2. \tag{30}$$

Then since $f_i, g_i$ have $\sigma$-bounded gradients, we have

$$\frac{L}{2}||\Delta\theta||^2 \leq \frac{L\eta^2\lambda^4 M^2}{2} \sum_i (\nabla G^T \nabla f_i)^2 ||\nabla f_i||^2 \tag{31}$$

$$+ (\nabla G^T \nabla g_i)^2 ||\nabla g_i||^2 \tag{32}$$

$$\leq \frac{L\eta^2\lambda^4 M^2\sigma^2}{2} \sum_i (\nabla G^T \nabla f_i)^2 + (\nabla G^T \nabla g_i)^2 \tag{33}$$

Then if $\lambda^2 < \frac{2}{\eta\sigma^2 ML}$,

$$\nabla G^T \Delta\theta + \frac{L}{2}||\Delta\theta||^2 \leq (\frac{L\eta^2\lambda^4 M^2\sigma^2}{2} - \eta\lambda^2 M) \tag{34}$$

$$\sum_i (\nabla G^T \nabla f_i)^2 + (\nabla G^T \nabla g_i)^2 \leq 0. \tag{35}$$

Finally we prove $G(\theta_{t+1}) = G(\theta_t)\ \forall B \Leftrightarrow \nabla G(\theta_t) = 0$: If $\nabla G(\theta_t) = 0$, from section B we have $\alpha_{t,i} = \beta_{t,i} = 0$, then $\theta_{t+1} = \theta_t$ and thus $G(\theta_{t+1}) = G(\theta_t)\ \forall B$. Otherwise, if $\nabla G(\theta_t) \neq 0$, we have

$$0 < ||\nabla G||^2 = \nabla G^T \nabla G = \frac{1}{M}\sum_{i=1}^{M} \nabla G^T \nabla f_i^v, \tag{36}$$

which means there exists a $k$ such that $\nabla G^T \nabla f_k^v > 0$. So for the mini-batch $B_k$ that contains this example, we have

$$G(\theta_{t+1}) - G(\theta_t) \leq \nabla G^T \Delta\theta + \frac{L}{2}||\Delta\theta||^2 \tag{37}$$

$$\leq (\frac{L\eta^2\lambda^4 M^2\sigma^2}{2} - \eta\lambda^2 M) \tag{38}$$

$$\sum_{i\in B} (\nabla G^T \nabla f_i)^2 + (\nabla G^T \nabla g_i)^2 \tag{39}$$

$$\leq (\frac{L\eta^2\lambda^4 M^2\sigma^2}{2} - \eta\lambda^2 M)\nabla G^T \nabla f_k^v \tag{40}$$

$$< 0. \tag{41}$$