Supplementary of Semi-Supervised Learning with Meta-Gradient

A Table of Notations

The notations in this work are summarized in Table 1.

Table 1: Table of notations in this work.

Symbol	Description
	Data
$\boldsymbol{x}_k^l$	The k^{th} labeled training example
$\boldsymbol{x}_i^u$	The i^{th} unlabeled training example
	Labels
$\boldsymbol{y}_k$	Actual label of $\boldsymbol{x}_k^l$
$\widetilde{\boldsymbol{y}}_i$	Initialized proximal label of $\boldsymbol{x}_i^u$
$\widehat{\boldsymbol{y}}_i$	Updated proximal label of $\boldsymbol{x}_i^u$
$\widetilde{y}_{i,j}$	The j^{th} entry of $\widetilde{\boldsymbol{y}}_i$ (Same for $\widehat{y}_{i,j}$)
$\widetilde{\boldsymbol{y}}$	Proximal labels of the unlabeled mini-batch $\widetilde{\boldsymbol{y}} = \{\widetilde{\boldsymbol{y}}_i : i = 1, \cdots, B^u\}$ (Seen as a vector)
$\widetilde{\boldsymbol{y}}_t$	Proximal labels of the unlabeled mini-batch at the t^{th} step[1]
$\widetilde{\boldsymbol{y}}_{i,t}$	The i^{th} proximal label of $\widetilde{\boldsymbol{y}}_t$
	Functions
$f(\cdot;\boldsymbol{\theta})$	Convolutional Neural Network parameterized by $\boldsymbol{\theta}$
$\Phi(\cdot,\cdot)$	Non-negative function that measures discrepancy of distributions
$\mathcal{L}(\cdot,\cdot;\boldsymbol{\theta})$	Loss function of the data-label pair when the model parameter is $\boldsymbol{\theta}$
$G(\boldsymbol{\theta};\mathcal{D})$	Validation loss on the dataset $\mathcal{D}$ when the model parameter is $\boldsymbol{\theta}$
$\nabla_{\boldsymbol{x}} g, J_{\boldsymbol{x}} g$	The gradient or Jacobian function of a generic function g *w.r.t.* $\boldsymbol{x}$
	Parameters
$\boldsymbol{\theta}_t$	Model parameters at the t^{th} step
$\widetilde{\boldsymbol{\theta}}_t$	Model parameters after the pseudo-update at the t^{th} step (*i.e.,* Eq. (4) in the main text)
$\theta_{t,l}$	The l^{th} entry of $\boldsymbol{\theta}_t$ (Same for $\widetilde{\theta}_{t,l}$)
	Gradients
$\nabla\boldsymbol{\theta}_t$	Gradients of the pseudo-update at the t^{th} step
$\nabla\widehat{\boldsymbol{\theta}}_t$	Gradients of the actual update at the t^{th} step
$\nabla\boldsymbol{\theta}_t^l$	Gradients of labeled mini-batch at the t^{th} step
$\nabla\widetilde{\boldsymbol{y}}_i$	Gradients of the proximal label $\widetilde{\boldsymbol{y}}_i$
$\nabla\widetilde{\boldsymbol{y}}$	Gradients of proximal labels $\nabla\widetilde{\boldsymbol{y}} = \{\nabla\widetilde{\boldsymbol{y}}_i : i = 1, \cdots, B^u\}$ (Seen as a vector)
	Configurations
$\alpha_t,\ \beta_t$	Regular learning rate and meta learning rate at the t^{th} step
$N^l,\ N^u$	Numbers of labeled examples and unlabeled examples
$B^l,\ B^u$	batch sizes for labeled examples and unlabeled examples

[1] For a bit abuse of notations, the subscript t or τ of $\widetilde{\boldsymbol{y}}$ specify the current step number, while subscript (i,j) of indicates the j^{th} entry of the i^{th} proximal label. The step subscript is ommited when there is no ambiguity.

B Convergence Analysis of Semi-Supervised Learning with Meta-Gradient

Lemma 1. *Let*

$$G(\boldsymbol{\theta};\mathcal{D}^l)=\frac{1}{N^l}\sum_{k=1}^{N^l}\mathcal{L}(\boldsymbol{x}_k^l,\boldsymbol{y}_k;\boldsymbol{\theta}_t) \tag{1}$$

be the loss function of the labeled examples. Assume

(i) the gradient function $\nabla_{\boldsymbol{\theta}}G$ is Lipschitz-continuous with a Lipschitz constant L_0; and

(ii) the norm of the Jacobian matrix of f w.r.t. $\boldsymbol{\theta}$ is upper-bounded by a constant M, i.e.,

$$\|J_{\boldsymbol{\theta}}f(\boldsymbol{x}_i^u;\boldsymbol{\theta})\|\le M,\quad \forall\, i\in\{1,\cdots,N^u\}\,. \tag{2}$$

If the labeled data loss is considered as a function of the pseudo-targets $\widetilde{\boldsymbol{y}}=\{\widetilde{\boldsymbol{y}}_i:i=1,\cdots,B^u\}$, i.e., $H(\widetilde{\boldsymbol{y}})=G(\widetilde{\boldsymbol{\theta}}_{t+1}(\widetilde{\boldsymbol{y}}))$, then the gradient function $\nabla_{\widetilde{\boldsymbol{y}}}H$ is also Lipschitz-continuous and its Lipschitz constant is upper-bounded by $4\alpha_t^2M^2L_0$.

Proof. Recall the SGD update formula

$$\widetilde{\boldsymbol{\theta}}_{t+1}=\boldsymbol{\theta}_t-\frac{\alpha_t}{B^u}\sum_{i=1}^{B^u}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t), \tag{3}$$

and we have

$$\frac{\partial\widetilde{\theta}_{t+1,l}}{\partial\widetilde{y}_{i,j}}=-\frac{\alpha_t}{B^u}\frac{\partial^2\mathcal{L}}{\partial\widetilde{y}_{i,j}\partial\theta_l}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t). \tag{4}$$

Then, we expand the partial derivative of each entry $\widetilde{y}_{i,j}$:

$$\begin{aligned}\frac{\partial H}{\partial\widetilde{y}_{i,j}}=&\frac{1}{N^l}\sum_{k=1}^{N^l}\sum_l\frac{\partial\mathcal{L}}{\partial\theta_l}(\boldsymbol{x}_k^l,\boldsymbol{y}_k;\widetilde{\boldsymbol{\theta}}_{t+1})\frac{\partial\widetilde{\theta}_{t+1,l}}{\partial\widetilde{y}_{i,j}}\\=&-\frac{\alpha_t}{B^uN^l}\sum_{k=1}^{N^l}\sum_l\frac{\partial\mathcal{L}}{\partial\theta_l}(\boldsymbol{x}_k^l,\boldsymbol{y}_k;\widetilde{\boldsymbol{\theta}}_{t+1})\frac{\partial^2\mathcal{L}}{\partial\widetilde{y}_{i,j}\partial\theta_l}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t)\\=&-\frac{\alpha_t}{B^uN^l}\sum_{k=1}^{N^l}\nabla_{\boldsymbol{\theta}}^\top\mathcal{L}(\boldsymbol{x}_k^l,\boldsymbol{y}_k;\widetilde{\boldsymbol{\theta}}_{t+1})\cdot\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t)\\=&-\frac{\alpha_t}{B^u}\nabla_{\boldsymbol{\theta}}^\top G(\widetilde{\boldsymbol{\theta}}_{t+1})\cdot\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t).\end{aligned} \tag{5}$$

Then, for arbitrary $\widetilde{\boldsymbol{y}}^1$ and $\widetilde{\boldsymbol{y}}^2$,

$$\begin{aligned}&\left.\frac{\partial H}{\partial\widetilde{y}_{i,j}}\right|_{\widetilde{\boldsymbol{y}}=\widetilde{\boldsymbol{y}}^1}-\left.\frac{\partial H}{\partial\widetilde{y}_{i,j}}\right|_{\widetilde{\boldsymbol{y}}=\widetilde{\boldsymbol{y}}^2}\\=&\frac{\alpha_t}{B^u}\left(\nabla_{\boldsymbol{\theta}}^\top G(\widetilde{\boldsymbol{\theta}}_{t+1}^2)\cdot\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i^2;\boldsymbol{\theta}_t)-\nabla_{\boldsymbol{\theta}}^\top G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\cdot\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i^1;\boldsymbol{\theta}_t)\right)\\=&\frac{\alpha_t}{B^u}\left(\nabla_{\boldsymbol{\theta}}^\top\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i^2;\boldsymbol{\theta}_t)\cdot\left(\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^2)-\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\right)+\right.\\&\left.\nabla_{\boldsymbol{\theta}}^\top G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\cdot\left(\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i^2;\boldsymbol{\theta}_t)-\nabla_{\boldsymbol{\theta}}\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i^1;\boldsymbol{\theta}_t)\right)\right),\end{aligned} \tag{6}$$

where $\widetilde{\boldsymbol{\theta}}_{t+1}^r=\widetilde{\boldsymbol{\theta}}_{t+1}(\widetilde{\boldsymbol{y}}^r)$, $r=1,2$. As the MSE loss is used for unlabeled data, we have $\frac{\partial\mathcal{L}}{\partial\widetilde{y}_{i,j}}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_t)=-2(f_j(\boldsymbol{x}_i^u;\boldsymbol{\theta}_t)-\widetilde{y}_{i,j})$. Here, f_j denotes the j^{th} entry of f. Therefore,

$$\begin{aligned}\left.\frac{\partial H}{\partial\widetilde{y}_{i,j}}\right|_{\widetilde{\boldsymbol{y}}=\widetilde{\boldsymbol{y}}^1}-\left.\frac{\partial H}{\partial\widetilde{y}_{i,j}}\right|_{\widetilde{\boldsymbol{y}}=\widetilde{\boldsymbol{y}}^2}&=-\frac{2\alpha_t}{B^u}\nabla_{\boldsymbol{\theta}}^\top f_j(\boldsymbol{x}_i^u;\boldsymbol{\theta}_t)\cdot\left(\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^2)-\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\right),\\\nabla_{\widetilde{\boldsymbol{y}}_i}H(\widetilde{\boldsymbol{y}}^1)-\nabla_{\widetilde{\boldsymbol{y}}_i}H(\widetilde{\boldsymbol{y}}^2)&=-\frac{2\alpha_t}{B^u}J_{\boldsymbol{\theta}}f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_t)\cdot\left(\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^2)-\nabla_{\boldsymbol{\theta}}G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\right).\end{aligned} \tag{7}$$

By taking the norm, we have

$$\left\|\nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^1) - \nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^2)\right\| \leq \frac{2\alpha_t}{B^u} \left\|J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u; \boldsymbol{\theta}_t)\right\| \left\|\nabla_{\boldsymbol{\theta}} G(\widetilde{\boldsymbol{\theta}}_{t+1}^2) - \nabla_{\boldsymbol{\theta}} G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\right\|. \tag{8}$$

By assumptions, we have

$$\begin{aligned} \left\|J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u; \boldsymbol{\theta}_t)\right\| &\leq M, \\ \left\|\nabla_{\boldsymbol{\theta}} G(\widetilde{\boldsymbol{\theta}}_{t+1}^2) - \nabla_{\boldsymbol{\theta}} G(\widetilde{\boldsymbol{\theta}}_{t+1}^1)\right\| &\leq L_0 \left\|\widetilde{\boldsymbol{\theta}}_{t+1}^2 - \widetilde{\boldsymbol{\theta}}_{t+1}^1\right\|. \end{aligned} \tag{9}$$

Considering

$$\begin{aligned} \left\|\widetilde{\boldsymbol{\theta}}_{t+1}^2 - \widetilde{\boldsymbol{\theta}}_{t+1}^1\right\| &= \frac{\alpha_t}{B^u} \left\|\sum_{i=1}^{B^u} \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}_i^u, \widetilde{\boldsymbol{y}}_i^2; \boldsymbol{\theta}_t) - \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}_i^u, \widetilde{\boldsymbol{y}}_i^1; \boldsymbol{\theta}_t)\right)\right\| \\ &\leq \frac{\alpha_t}{B^u} \sum_{i=1}^{B^u} \left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}_i^u, \widetilde{\boldsymbol{y}}_i^2; \boldsymbol{\theta}_t) - \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}_i^u, \widetilde{\boldsymbol{y}}_i^1; \boldsymbol{\theta}_t)\right\| \\ &= \frac{2\alpha_t}{B^u} \sum_{i=1}^{B^u} \left\|J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u; \boldsymbol{\theta}_t) \cdot (\widetilde{\boldsymbol{y}}_i^1 - \widetilde{\boldsymbol{y}}_i^2)\right\| \\ &\leq \frac{2\alpha_t}{B^u} \sum_{i=1}^{B^u} \left\|J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u; \boldsymbol{\theta}_t)\right\| \left\|\widetilde{\boldsymbol{y}}_i^1 - \widetilde{\boldsymbol{y}}_i^2\right\| \\ &\leq 2\alpha_t M \left\|\widetilde{\boldsymbol{y}}^1 - \widetilde{\boldsymbol{y}}^2\right\|, \end{aligned} \tag{10}$$

thus we have

$$\begin{aligned} \left\|\nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^1) - \nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^2)\right\| &\leq \frac{4\alpha_t^2 M^2 L_0}{B^u} \left\|\widetilde{\boldsymbol{y}}^1 - \widetilde{\boldsymbol{y}}^2\right\|, \\ \left\|\nabla_{\widetilde{\boldsymbol{y}}} H(\widetilde{\boldsymbol{y}}^1) - \nabla_{\widetilde{\boldsymbol{y}}} H(\widetilde{\boldsymbol{y}}^2)\right\| &\leq \sum_{i=1}^{B^u} \left\|\nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^1) - \nabla_{\widetilde{\boldsymbol{y}}_i} H(\widetilde{\boldsymbol{y}}^2)\right\| \leq 4\alpha_t^2 M^2 L_0 \left\|\widetilde{\boldsymbol{y}}^1 - \widetilde{\boldsymbol{y}}^2\right\|. \end{aligned} \tag{11}$$

Therefore, $\nabla_{\widetilde{\boldsymbol{y}}} H$ is Lipschitz-continuous with a Lipschitz constant $L_t \leq 4\alpha_t^2 M^2 L_0$. □

Theorem 1. *Assume the same conditions as in Lemma 1. If the regular learning rate α_t and meta learning rate β_t satisfy $\alpha_t^2 \beta_t < (4M^2 L_0)^{-1}$, then each SGD step of Alg. 1 will lead to the decrease of the validation loss $G(\boldsymbol{\theta})$, regardless of the selected unlabeled examples, i.e.,*

$$G(\boldsymbol{\theta}_{t+1}) \leq G(\boldsymbol{\theta}_t), \quad \textit{for each } t. \tag{12}$$

Moreover, the equality holds if and only if $\nabla \widetilde{\boldsymbol{y}} = \mathbf{0}$ for the selected unlabeled batch at the t^{th} step.

Proof. According to the Lagrange Mean Value Theorem, there exists $\xi \in (0, 1)$, such that

$$H(\widehat{\boldsymbol{y}}) = H(\widetilde{\boldsymbol{y}}) + \nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}} + \xi(\widehat{\boldsymbol{y}} - \widetilde{\boldsymbol{y}})) \cdot (\widehat{\boldsymbol{y}} - \widetilde{\boldsymbol{y}}). \tag{13}$$

Recall the update formula of the pseudo-targets, *i.e.,* $\widehat{\boldsymbol{y}} = \widetilde{\boldsymbol{y}} - \beta_t \nabla \widetilde{\boldsymbol{y}}$. Then, by the Lipschitz-continuity of $\nabla_{\widetilde{\boldsymbol{y}}} H$, we have

$$\begin{aligned} H(\widehat{\boldsymbol{y}}) &= H(\widetilde{\boldsymbol{y}}) - \beta_t \nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}} - \xi\beta_t \nabla \widetilde{\boldsymbol{y}}) \cdot \nabla \widetilde{\boldsymbol{y}} \\ &= H(\widetilde{\boldsymbol{y}}) - \beta_t \nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}}) \cdot \nabla \widetilde{\boldsymbol{y}} - \beta_t \left(\nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}} - \xi\beta_t \nabla \widetilde{\boldsymbol{y}}) - \nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}})\right) \cdot \nabla \widetilde{\boldsymbol{y}} \\ &\leq H(\widetilde{\boldsymbol{y}}) - \beta_t \nabla_{\widetilde{\boldsymbol{y}}}^\top H(\widetilde{\boldsymbol{y}}) \cdot \nabla \widetilde{\boldsymbol{y}} + \beta_t^2 L_t \|\nabla \widetilde{\boldsymbol{y}}\|_2^2 && \text{(By (11))} \\ &= H(\widetilde{\boldsymbol{y}}) - (\beta_t - \beta_t^2 L_t) ||\nabla \widetilde{\boldsymbol{y}}||^2 && \text{(Since } \nabla \widetilde{\boldsymbol{y}} = \nabla_{\widetilde{\boldsymbol{y}}} H(\widetilde{\boldsymbol{y}})) \\ &\leq H(\widetilde{\boldsymbol{y}}). && \text{(Since } \beta_t < L_t^{-1}) \end{aligned} \tag{14}$$

Therefore, $G(\boldsymbol{\theta}_{t+1}) = H(\widehat{\boldsymbol{y}}) \leq H(\widetilde{\boldsymbol{y}}) = G(\boldsymbol{\theta}_t)$.

Moreover, as long as $\alpha_t^2 \beta_t < (4M^2 L_0)^{-1}$ is satisfied, the equality holds if and only if $\nabla \widetilde{\boldsymbol{y}} = \mathbf{0}$. □

Theorem 2. *Assume the same conditions as in Lemma 1, and*

$$\inf_t \left(\beta_t - 4\alpha_t^2\beta_t^2 M^2 L_0\right) = D_1 > 0, \quad \inf_t \alpha_t = D_2 > 0. \tag{15}$$

We further assume that the unlabeled dataset contains the labeled dataset, i.e., $\mathcal{D}^l \subseteq \mathcal{D}^u$. Then, Alg. 1 achieves $\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_t)\|^2\right] \leq \epsilon$ in $O(1/\epsilon^2)$ steps, i.e.,

$$\min_{1\leq t\leq T} \mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_t)\|^2\right] \leq \frac{C}{\sqrt{T}}, \tag{16}$$

where C is a constant independent of the training process.

Proof. According to (13) in the proof of Theorem 1, we have

$$G(\boldsymbol{\theta}_{t+1}) \leq G(\boldsymbol{\theta}_t) - (\beta_t - \beta_t^2 L_t)\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2 \leq G(\boldsymbol{\theta}_t) - (\beta_t - 4\alpha_t^2\beta_t^2 M^2 L_0)\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2. \tag{17}$$

Therefore,

$$G(\boldsymbol{\theta}_t) - G(\boldsymbol{\theta}_{t+1}) \geq (\beta_t - 4\alpha_t^2\beta_t^2 M^2 L_0)\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2 \geq D_1\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2. \tag{18}$$

By taking the expectation, we have

$$\mathbb{E}_{1\sim t}\left[G(\boldsymbol{\theta}_t)\right] - \mathbb{E}_{1\sim t}\left[G(\boldsymbol{\theta}_{t+1})\right] \geq D_1\mathbb{E}_{1\sim t}\left[\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2\right]. \tag{19}$$

Here, $\mathbb{E}_{1\sim t}$ indicates the expectation is taken over the selected mini-batches of the first t steps. Next, we show $\mathbb{E}_{1\sim t}\left[G(\boldsymbol{\theta}_t)\right] = \mathbb{E}_{1\sim t-1}\left[G(\boldsymbol{\theta}_t)\right]$, which is intuitive as the value of $\boldsymbol{\theta}_t$ only relies on the selected batches of the first $t-1$ steps. We rigorously prove it with conditional expectation:

$$\mathbb{E}_{1\sim t}\left[G(\boldsymbol{\theta}_t)\right] = \mathbb{E}_{1\sim t-1}\left[\mathbb{E}_t\left[G(\boldsymbol{\theta}_t)|1\sim t-1\right]\right] = \mathbb{E}_{1\sim t-1}\left[G(\boldsymbol{\theta}_t)\right]. \tag{20}$$

Here, the first equality comes from the *law of total expectation*, while the second comes from the fact that $G(\boldsymbol{\theta}_t)$ is deterministic given the selected batches of the first $t-1$ steps. Besides, when $t=1$, (19) is adapted to

$$G(\boldsymbol{\theta}_1) - \mathbb{E}_1\left[G(\boldsymbol{\theta}_2)\right] \geq D_1\mathbb{E}_1\left[\left\|\nabla\widetilde{\boldsymbol{y}}_1\right\|^2\right], \tag{21}$$

where $G(\boldsymbol{\theta}_1)$ is the loss of the initialized model parameters so the expectation is omitted. Then, by taking a summation over the first T steps, we have

$$D_1\sum_{t=1}^{T}\mathbb{E}_{1\sim t}\left[\left\|\nabla\widetilde{\boldsymbol{y}}_t\right\|^2\right] \leq G(\boldsymbol{\theta}_1) - \mathbb{E}_{1\sim T}\left[G(\boldsymbol{\theta}_{T+1})\right] \leq G(\boldsymbol{\theta}_1). \tag{22}$$

Therefore, there exists $\tau \in \{1,\cdots,T\}$, s.t.

$$\mathbb{E}_{1\sim\tau}\left[\left\|\nabla\widetilde{\boldsymbol{y}}_\tau\right\|^2\right] \leq \frac{G(\boldsymbol{\theta}_1)}{D_1 T}. \tag{23}$$

Then, we attempt to build a relationship between $\nabla\widetilde{\boldsymbol{y}}_\tau$ and $\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)$. Similar to Eq. (5), we have

$$\nabla\widetilde{\boldsymbol{y}}_{i,\tau} = -\frac{\alpha_\tau}{B^u}\nabla^2_{\widetilde{\boldsymbol{y}}_i,\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{x}_i^u,\widetilde{\boldsymbol{y}}_i;\boldsymbol{\theta}_\tau)\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau) = \frac{2\alpha_\tau}{B^u}J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_\tau)\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau). \tag{24}$$

Therefore,

$$\left\|\nabla\widetilde{\boldsymbol{y}}_\tau\right\|^2 = \sum_{i=1}^{B^u}\nabla^\top\widetilde{\boldsymbol{y}}_{i,\tau}\cdot\nabla\widetilde{\boldsymbol{y}}_{i,\tau} = \frac{4\alpha_\tau^2}{(B^u)^2}\nabla_{\boldsymbol{\theta}}^\top G(\boldsymbol{\theta}_\tau)\cdot\left(\sum_{i=1}^{B^u}J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_\tau)\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_\tau)\right)\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau). \tag{25}$$

Now consider the potential unlabeled batches $\{\boldsymbol{B}_k : k=1,\cdots,N^l\}$ of the τ^{th} step. Since, $\mathcal{D}^l \subseteq \mathcal{D}^u$, we can assume $\boldsymbol{x}_k^l \in \boldsymbol{B}_k,\ k=1,\cdots,N^l$ and these batches are sampled with non-zero probabilities $\{p_k : k=1,\cdots,N^l\}$.

Let $p = \min_k p_k > 0$, and we have

$$\begin{aligned}
\mathbb{E}_{1\sim\tau}\left[\left\|\nabla\widetilde{\boldsymbol{y}}_\tau\right\|^2\right] &= \mathbb{E}_{1\sim\tau-1}\left[\mathbb{E}_\tau\left[\left\|\nabla\widetilde{\boldsymbol{y}}_\tau\right\|^2\right]\right] \\
&= \mathbb{E}_{1\sim\tau-1}\left[\frac{4\alpha_\tau^2}{(B^u)^2}\nabla_{\boldsymbol{\theta}}^\top G(\boldsymbol{\theta}_\tau)\cdot\mathbb{E}_\tau\left[\sum_{i=1}^{B^u} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_\tau)\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_i^u;\boldsymbol{\theta}_\tau)\right]\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right] \\
&\geq \mathbb{E}_{1\sim\tau-1}\left[\frac{4\alpha_\tau^2}{(B^u)^2}\nabla_{\boldsymbol{\theta}}^\top G(\boldsymbol{\theta}_\tau)\cdot\left(\sum_{k=1}^{N^l} p_k\, J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\right)\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right] \\
&\geq \frac{4pD_2^2}{(B^u)^2}\mathbb{E}_{1\sim\tau-1}\left[\nabla_{\boldsymbol{\theta}}^\top G(\boldsymbol{\theta}_\tau)\cdot\left(\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\right)\cdot\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right].
\end{aligned} \tag{26}$$

Note that similar to Eq. (20), the inner expectation is also conditioned on the selected batches of the first $\tau-1$ steps, which is equivalent to that conditioned on $\boldsymbol{\theta}_t$.

By applying the chain rule, we have

$$\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) = \frac{2}{N^l}\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta})\cdot\left(f(\boldsymbol{x}_k^l;\boldsymbol{\theta})-\boldsymbol{y}_k\right). \tag{27}$$

Since both $f(\boldsymbol{x}_k^l;\boldsymbol{\theta})$ and $\boldsymbol{y}_k$ are distributions on the category space, there exists a constant $R>0$, s.t. $\left\|f(\boldsymbol{x}_k^l;\boldsymbol{\theta})-\boldsymbol{y}_k\right\|\leq R$. Therefore,

$$\begin{aligned}
&\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau) \\
\succeq\ &\frac{1}{R^2}\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot\left(f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)-\boldsymbol{y}_k\right)\cdot\left(f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)-\boldsymbol{y}_k\right)^\top\cdot J_{\boldsymbol{\theta}} f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau) \\
\succeq\ &\frac{1}{N^l R^2}\left(\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot\left(f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)-\boldsymbol{y}_k\right)\right)\cdot\left(\sum_{k=1}^{N^l} J_{\boldsymbol{\theta}}^\top f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)\cdot\left(f(\boldsymbol{x}_k^l;\boldsymbol{\theta}_\tau)-\boldsymbol{y}_k\right)\right)^\top \\
=\ &\frac{N^l}{4R^2}\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\cdot\nabla_{\boldsymbol{\theta}}^\top G(\boldsymbol{\theta}_\tau).
\end{aligned} \tag{28}$$

Here, the symbol $\succeq$ indicates certain matrix relationship where $\boldsymbol{A}\succeq\boldsymbol{B}$ means $\boldsymbol{A}-\boldsymbol{B}$ is a positive semidefinite matrix.

We prove the first inequality in (28) with simplified notations. Suppose $\boldsymbol{v}$ is a vector and $\boldsymbol{A}$ is a matrix of proper dimension. Then, we show that if $\|\boldsymbol{v}\|\leq R$, then $R^2\boldsymbol{A}^\top\boldsymbol{A}\succeq\boldsymbol{A}^\top\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{A}$. For an arbitrary vector $\boldsymbol{u}$ of proper dimension, we have

$$\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{u} = \left\|\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{u}\right\|^2 \leq \|\boldsymbol{v}\|^2\left\|\boldsymbol{A}\boldsymbol{u}\right\|^2 \leq R^2\left\|\boldsymbol{A}\boldsymbol{u}\right\|^2 = R^2\boldsymbol{u}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{u}. \tag{29}$$

By definition, $R^2\boldsymbol{A}^\top\boldsymbol{A}-\boldsymbol{A}^\top\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{A}$ is positive semidefinite. The second inequality in (28) comes from the Cauchy-Schwartz inequality that $\mathbb{E}\left[\boldsymbol{A}^\top\boldsymbol{A}\right]\succeq\mathbb{E}\left[\boldsymbol{A}^\top\right]\mathbb{E}\left[\boldsymbol{A}\right]$ for any random matrix $\boldsymbol{A}$.

With (26) and (28), it is easy to show that

$$\mathbb{E}_{1\sim\tau}\left[\left\|\nabla\widetilde{\boldsymbol{y}}_\tau\right\|^2\right] \geq \frac{pD_2^2N^l}{(B^u)^2R^2}\mathbb{E}_{1\sim\tau-1}\left[\left\|\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right\|^4\right] \geq \frac{pD_2^2N^l}{(B^u)^2R^2}\left(\mathbb{E}_{1\sim\tau-1}\left[\left\|\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right\|^2\right]\right)^2. \tag{30}$$

Again, the second inequality comes from the Cauchy-Schwartz inequality. Incorporating with (23), we have

$$\mathbb{E}_{1\sim\tau-1}\left[\left\|\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}_\tau)\right\|^2\right] \leq \frac{C}{\sqrt{T}}, \quad \text{where } C = \frac{B^u R}{D_2}\sqrt{\frac{G(\boldsymbol{\theta}_1)}{pN^lD_1}}. \tag{31}$$

which concludes this proof. □

C Implementation Details

Our implementation is based on the PyTorch (Paszke et al., 2019) library and the proposed algorithm is evaluated on the SVHN (Netzer et al., 2011), CIFAR (Krizhevsky et al., 2009), and ImageNet (Russakovsky et al., 2015) datasets.

Evaluation on the SVHN and CIFAR datasets. As the standard evaluation protocol, 1k category-balanced labels are used for supervision out of the 73,257 training examples of the SVHN dataset. For the CIFAR-10 (*resp.* CIFAR-100) dataset, the number of labeled examples is 4k (*resp.* 10k) out the 50k training examples. For the backbone architectures, the Conv-Large architecture is the same as the one in previous work (Laine and Aila, 2017; Miyato et al., 2018; Tarvainen and Valpola, 2017; Athiwaratkun et al., 2019; Wang et al., 2019). The detailed configurations are summarized in Table 2. For the ResNet (He et al., 2016) architecture, we adopt the ResNet-26-2x96d Shake-Shake regularized architecture with 12 residual blocks as in Gastaldi (2017). The same architecture is used in prior SSL methods (Tarvainen and Valpola, 2017; Athiwaratkun et al., 2019). We follow a common practice of data augmentation, *i.e.,* zero-padding of 4 pixels on each side of the image, random crop of a 32×32 patch, and random horizontal flip, for the CIFAR datasets, and omit the random horizontal flip for SVHN. The meta learning rate β_t is always set equal to the regular learning rate α_t. We train from scratch for 400k iterations with an initial learning rate of 0.1, and decay the learning rate by a factor of 10 at the end of 300k and 350k iterations. We use the SGD optimizer with a momentum of 0.9, and the weight decay is set to 10^{-4} for the CIFAR datasets, and 5×10^{-5} for SVHN. The batch size is 128 for both labeled and unlabeled data. The shape parameter γ of the Beta distribution is set to 1.0 for the CIFAR datasets, and 0.1 for SVHN, as suggested by Wang et al. (2019).

Evaluation on the ImageNet dataset. The large-scale ImageNet benchmark contains 1.28M training images of 1k fine-grained classes. We evaluate on the ResNet-18 (He et al., 2016) backbone with 10% labels. The standard data augmentation strategy (Simonyan and Zisserman, 2015; He et al., 2016; Xie et al., 2017) is adopted: image resize such that the shortest edge is of 256 pixels, random crop of a 224×224 patch, and random horizontal flip. The overall batch size is 512, and the same optimizer as the aforementioned one is employed with a weight decay of 10^{-4}. We train for 600 epochs in total, and decay the learning rate from 0.1 according to the cosine annealing strategy (Loshchilov and Hutter, 2017). The shape parameter γ is set to 1.0.

Table 2: Conv-Large (Tarvainen and Valpola, 2017) Architecture.

Layer	**Configurations**				**Output Size**
	#Filters	Kernel Size	Stride	#Paddings	
Convolution	128	3	1	1	32×32
Convolution	128	3	1	1	32×32
Convolution	128	3	1	1	32×32
MaxPooling	128	2	2	0	16×16
Dropout	Drop probability = 0.5				16×16
Convolution	256	3	1	1	16×16
Convolution	256	3	1	1	16×16
Convolution	256	3	1	1	16×16
MaxPooling	128	2	2	0	8×8
Dropout	Drop probability = 0.5				8×8
Convolution	512	3	1	0	6×6
Convolution	256	1	1	0	6×6
Convolution	128	1	1	0	6×6
AvgPooling	128	6	1	0	1×1
Linear	$128 \rightarrow 10$				1×1

References

Athiwaratkun, B., Finzi, M., Izmailov, P., and Wilson, A. G. (2019). There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *International Conference on Learning Representations (ICLR)*. 6

Gastaldi, X. (2017). Shake-Shake Regularization. *arXiv preprint arXiv:1705.07485*. 6

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. 6

Krizhevsky, A., Hinton, G., et al. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer. 6

Laine, S. and Aila, T. (2017). Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*. 6

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*. 6

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 41(8):1979–1993. 6

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading Digits in Natural Images with Unsupervised Feature Learning. In *Neural Information Processing Systems Workshops*. 6

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-performance Deep Learning Library. In *Neural Information Processing Systems (NeurIPS)*, pages 8026–8037. 6

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal on Computer Vision (IJCV)*, 115(3):211–252. 6

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-scale Image Recognition. In *International Conference on Learning Representations (ICLR)*. 6

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Neural Information Processing Systems (NeurIPS)*, pages 1195–1204. 6

Wang, Q., Li, W., and Gool, L. V. (2019). Semi-Supervised Learning by Augmented Distribution Alignment. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1466–1475. 6

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500. 6