# Supplementary of Semi-Supervised Learning with Meta-Gradient

## A Table of Notations

The notations in this work are summarized in Table 1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^l_k$</td>
<td>The $k^{th}$ labeled training example</td>
</tr>
<tr>
<td>$x^u_i$</td>
<td>The $i^{th}$ unlabeled training example</td>
</tr>
<tr>
<td>$y_k$</td>
<td>Actual label of $x^l_k$</td>
</tr>
<tr>
<td>$\hat{y}_i$</td>
<td>Updated proximal label of $x^u_i$</td>
</tr>
<tr>
<td>$\tilde{y}_{i,j}$</td>
<td>The $j^{th}$ entry of $\tilde{y}<em>i$ (Same for $\hat{y}</em>{i,j}$)</td>
</tr>
<tr>
<td>$\tilde{y}$</td>
<td>Proximal labels of the unlabeled mini-batch ${\tilde{y}_i : i = 1, \cdots, B^u}$ (Seen as a vector)</td>
</tr>
<tr>
<td>$y_{i,t}$</td>
<td>Proximal labels of the unlabeled mini-batch at the $t^{th}$ step$^1$</td>
</tr>
<tr>
<td>$\hat{y}_{i,t}$</td>
<td>The $i^{th}$ proximal label of $\tilde{y}_{t}$</td>
</tr>
<tr>
<td>$\theta_t$</td>
<td>Model parameters at the $t^{th}$ step</td>
</tr>
<tr>
<td>$\hat{\theta}_t$</td>
<td>Model parameters after the pseudo-update at the $t^{th}$ step (i.e., Eq. (4) in the main text)</td>
</tr>
<tr>
<td>$\theta_{t,l}$</td>
<td>The $l^{th}$ entry of $\theta_t$ (Same for $\hat{\theta}_{t,l}$)</td>
</tr>
<tr>
<td>$\nabla \theta_t$</td>
<td>Gradients of the pseudo-update at the $t^{th}$ step</td>
</tr>
<tr>
<td>$\nabla \hat{\theta}_t$</td>
<td>Gradients of the actual update at the $t^{th}$ step</td>
</tr>
<tr>
<td>$\nabla \theta_{t,l}$</td>
<td>Gradients of labeled mini-batch at the $t^{th}$ step</td>
</tr>
<tr>
<td>$\nabla y_i$</td>
<td>Gradients of the proximal label $\tilde{y}_i$</td>
</tr>
<tr>
<td>$\nabla \tilde{y}$</td>
<td>Gradients of proximal labels ${\nabla \tilde{y}_i : i = 1, \cdots, B^u}$ (Seen as a vector)</td>
</tr>
<tr>
<td>$\alpha_t, \beta_t$</td>
<td>Regular learning rate and meta learning rate at the $t^{th}$ step</td>
</tr>
<tr>
<td>$N^l, N^u$</td>
<td>Numbers of labeled examples and unlabeled examples</td>
</tr>
<tr>
<td>$B^l, B^u$</td>
<td>batch sizes for labeled examples and unlabeled examples</td>
</tr>
</tbody>
</table>

$^1$For a bit abuse of notations, the subscript $t$ or $\tau$ of $\tilde{y}$ 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 ommitted when there is no ambiguity.
B Convergence Analysis of Semi-Supervised Learning with Meta-Gradient

Lemma 1. Let

\[ G(\theta; D^l) = \frac{1}{N^l} \sum_{k=1}^{N^l} \mathcal{L}(x_k^l, y_k; \theta) \]

be the loss function of the labeled examples. Assume

(i) the gradient function \( \nabla \theta G \) is Lipschitz-continuous with a Lipschitz constant \( L_0 \); and

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

\[ \| \nabla \theta f(x_i^u; \theta) \| \leq M, \quad \forall i \in \{1, \ldots, N^u\}. \]

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

Proof. Recall the SGD update formula

\[ \tilde{\theta}_{t+1} = \theta_t - \frac{\alpha_t}{B^u} \sum_{i=1}^{B^u} \nabla \theta \mathcal{L}(x_i^l, \tilde{y}_i; \theta_t), \]

and we have

\[ \frac{\partial \tilde{\theta}_{t+1,i}}{\partial y_{i,j}} = -\frac{\alpha_t}{B^u} \frac{\partial^2 \mathcal{L}}{\partial \tilde{y}_{i,j} \partial \theta_t}(x_i^l, \tilde{y}_i; \theta_t). \]

Then, we expand the partial derivative of each entry \( \tilde{y}_{i,j} \):

\[
\frac{\partial H}{\partial y_{i,j}} = \frac{1}{N^l} \sum_{k=1}^{N^l} \sum_{l} \frac{\partial \mathcal{L}(x_k^l, y_k; \tilde{\theta}_{t+1})}{\partial \theta_t} \frac{\partial \tilde{\theta}_{t+1,i}}{\partial y_{i,j}} = -\frac{\alpha_t}{B^u N^l} \sum_{k=1}^{N^l} \sum_{l} \frac{\partial \mathcal{L}(x_k^l, y_k; \tilde{\theta}_{t+1})}{\partial \theta_t} \frac{\partial^2 \mathcal{L}}{\partial \tilde{y}_{i,j} \partial \theta_t}(x_i^l, \tilde{y}_i; \theta_t)
\]

\[ = -\frac{\alpha_t}{B^u N^l} \sum_{k=1}^{N^l} \nabla \theta \mathcal{L}(x_k^l, y_k; \tilde{\theta}_{t+1}) \cdot \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^l, \tilde{y}_i; \theta_t) \]

Then, for arbitrary \( \tilde{y}_1 \) and \( \tilde{y}_2 \),

\[
\frac{\partial H}{\partial y_{i,j}} \bigg|_{\tilde{y}_1 = \tilde{y}_2} = \frac{\partial H}{\partial y_{i,j}} \bigg|_{\tilde{y}_1 = \tilde{y}_2} = \frac{\alpha_t}{B^u} \left( \nabla \theta G(\tilde{\theta}_{t+1}) \cdot \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^u, \tilde{y}_i^2; \theta_t) - \nabla \theta G(\tilde{\theta}_{t+1}) \cdot \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^u, \tilde{y}_i^1; \theta_t) \right)
\]

\[ = \frac{\alpha_t}{B^u} \left( \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^u, \tilde{y}_i^2; \theta_t) \cdot \left( \nabla \theta G(\tilde{\theta}_{t+1}) - \nabla \theta G(\tilde{\theta}_{t+1}) \right) + \nabla \theta G(\tilde{\theta}_{t+1}) \cdot \left( \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^u, \tilde{y}_i^2; \theta_t) - \nabla \theta \frac{\partial \mathcal{L}}{\partial \tilde{y}_{i,j}}(x_i^u, \tilde{y}_i^1; \theta_t) \right) \right), \]

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

\[
\frac{\partial H}{\partial y_{i,j}} \bigg|_{\tilde{y}_1 = \tilde{y}_2} = \frac{\partial H}{\partial y_{i,j}} \bigg|_{\tilde{y}_1 = \tilde{y}_2} = \frac{2\alpha_t}{B^u} \nabla \theta f_j(x_i^u; \theta_t) \cdot \left( \nabla \theta G(\tilde{\theta}_{t+1}^2) - \nabla \theta G(\tilde{\theta}_{t+1}^1) \right),
\]

\[ \nabla \tilde{y}_i H(\tilde{y}^1) - \nabla \tilde{y}_i H(\tilde{y}^2) = \frac{2\alpha_t}{B^u} \nabla \theta f_j(x_i^u; \theta_t) \cdot \left( \nabla \theta G(\tilde{\theta}_{t+1}^2) - \nabla \theta G(\tilde{\theta}_{t+1}^1) \right). \]
By taking the norm, we have
\[
\left\| \nabla_{\hat{y}} H(\hat{y}^1) - \nabla_{\hat{y}} H(\hat{y}^2) \right\| \leq \frac{2\alpha_t}{B_u} \left\| J_\theta f(x_i^n; \theta_t) \right\| \left\| \nabla_\theta G(\theta^2_{t+1}) - \nabla_\theta G(\theta^1_{t+1}) \right\| .
\]
(8)

By assumptions, we have
\[
\left\| J_\theta f(x_i^n; \theta_t) \right\| \leq M,
\]
\[
\left\| \nabla_\theta G(\theta^2_{t+1}) - \nabla_\theta G(\theta^1_{t+1}) \right\| \leq L_0 \left\| \theta^2_{t+1} - \theta^1_{t+1} \right\| .
\]
(9)

Considering
\[
\left\| \theta^2_{t+1} - \theta^1_{t+1} \right\| = \frac{\alpha_t}{B_u} \sum_{i=1}^{B_u} \left\| \nabla_\theta L(x_i^n; \tilde{y}_i; \theta_t) - \nabla_\theta L(x_i^n; y_i; \theta_t) \right\|
\]
\[
\leq \frac{\alpha_t}{B_u} \sum_{i=1}^{B_u} \left\| \nabla_\theta L(x_i^n; \tilde{y}_i; \theta_t) - \nabla_\theta L(x_i^n; y_i; \theta_t) \right\|
\]
\[
= \frac{2\alpha_t}{B_u} \sum_{i=1}^{B_u} \left\| J_\theta f(x_i^n; \theta_t) \cdot (\tilde{y}_i - y_i) \right\|
\]
\[
\leq \frac{2\alpha_t}{B_u} \sum_{i=1}^{B_u} \left\| J_\theta f(x_i^n; \theta_t) \right\| \left\| \tilde{y}_i - y_i \right\|
\]
\[
\leq 2\alpha_t M \left\| \tilde{y}^1 - \tilde{y}^2 \right\| .
\]
(10)

thus we have
\[
\left\| \nabla_{\hat{y}} H(\hat{y}^1) - \nabla_{\hat{y}} H(\hat{y}^2) \right\| \leq \frac{4\alpha_t^2 M^2 L_0}{B_u} \left\| \tilde{y}^1 - \tilde{y}^2 \right\| ,
\]
\[
\left\| \nabla_{\hat{y}} H(\hat{y}^1) - \nabla_{\hat{y}} H(\hat{y}^2) \right\| \leq \sum_{i=1}^{B_u} \left\| \nabla_{\hat{y}} H(\hat{y}^1) - \nabla_{\hat{y}} H(\hat{y}^2) \right\| \leq 4\alpha_t^2 M^2 L_0 \left\| \tilde{y}^1 - \tilde{y}^2 \right\| .
\]
(11)

Therefore, \( \nabla_{\hat{y}} H \) 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 \( \alpha_t \) and meta learning rate \( \beta_t \) satisfy \( \alpha_t^2 \beta_t < (4M^2L_0)^{-1} \), then each SGD step of Alg. 1 will lead to the decrease of the validation loss \( \hat{G}(\theta) \), regardless of the selected unlabeled examples, i.e.,
\[
\hat{G}(\theta_{t+1}) \leq \hat{G}(\theta_t), \quad \text{for each} \ t.
\]
(12)

Moreover, the equality holds if and only if \( \nabla \hat{y} = 0 \) 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
\[
H(\tilde{y}) = H(\hat{y}) + \nabla_\hat{y}^\top H(\hat{y} + \xi(\tilde{y} - \hat{y})) \cdot (\tilde{y} - \hat{y}).
\]
(13)

Recall the update formula of the pseudo-targets, i.e., \( \tilde{y} = \hat{y} - \beta_t \nabla \hat{y} \). Then, by the Lipschitz-continuity of \( \nabla_\hat{y} H \), we have
\[
H(\tilde{y}) = H(\hat{y}) - \beta_t \nabla_\hat{y} H(\hat{y} - \xi \beta_t \nabla_\hat{y}) \cdot \nabla_\hat{y}
\]
\[
= H(\hat{y}) - \beta_t \nabla_\hat{y} H(\hat{y}) \cdot \nabla_\hat{y} - \beta_t (\nabla_\hat{y} H(\hat{y} - \xi \beta_t \nabla_\hat{y}) - \nabla_\hat{y} H(\hat{y})) \cdot \nabla_\hat{y}
\]
\[
\leq H(\hat{y}) - \beta_t \nabla_\hat{y} H(\hat{y}) \cdot \nabla_\hat{y} + \beta_t^2 L_t \| \nabla_\hat{y} \|^2_2 \quad \text{(By (11))}
\]
\[
\leq H(\hat{y}) - (\beta_t - \beta_t^2 L_t) \| \nabla_\hat{y} \|^2 \quad \text{(Since} \ \nabla_\hat{y} = \nabla_\hat{y} H(\hat{y}) \text{).}
\]
\[
\leq H(\hat{y}).
\]
(14)

Therefore, \( \hat{G}(\theta_{t+1}) = H(\tilde{y}) \leq H(\hat{y}) = \hat{G}(\theta_t) \).

Moreover, as long as \( \alpha_t^2 \beta_t < (4M^2L_0)^{-1} \) is satisfied, the equality holds if and only if \( \nabla \tilde{y} = 0 \).
Theorem 2. Assume the same conditions as in Lemma 1, and
\begin{equation}
\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.
\end{equation}

We further assume that the unlabeled dataset contains the labeled dataset, i.e., \( \mathcal{D}^l \subseteq \mathcal{D}^u \). Then, Alg. 1 achieves \( \mathbb{E} \left[ \| \nabla_{\theta} G(\theta_{t}) \|^{2} \right] \leq c \) in \( O(1/c^{2}) \) steps, i.e.,
\begin{equation}
\min_{1 \leq t \leq T} \mathbb{E} \left[ \| \nabla_{\theta} G(\theta_{t}) \|^{2} \right] \leq \frac{C}{\sqrt{T}},
\end{equation}
where \( C \) is a constant independent of the training process.

Proof. According to (13) in the proof of Theorem 1, we have
\begin{equation}
G(\theta_{t+1}) \leq G(\theta_{t}) - (\beta_{t} - \beta_{t}^{2}L_{t}) \| \nabla \bar{y}_{t} \|^{2} \leq G(\theta_{t}) - (\beta_{t} - 4\alpha_{t}^{2}\beta_{t}^{2}M^{2}L_{0}) \| \nabla \bar{y}_{t} \|^{2}.
\end{equation}
Therefore,
\begin{equation}
G(\theta_{t}) - G(\theta_{t+1}) \geq (\beta_{t} - 4\alpha_{t}^{2}\beta_{t}^{2}M^{2}L_{0}) \| \nabla \bar{y}_{t} \|^{2} \geq D_{1} \| \nabla \bar{y}_{t} \|^{2}.
\end{equation}

By taking the expectation, we have
\begin{equation}
\mathbb{E}_{1 \sim t} \left[ G(\theta_{t}) \right] - \mathbb{E}_{1 \sim t} \left[ G(\theta_{t+1}) \right] \geq D_{1} \mathbb{E}_{1 \sim t} \left[ \| \nabla \bar{y}_{t} \|^{2} \right].
\end{equation}

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

Here, the first equality comes from the law of total expectation, while the second comes from the fact that \( G(\theta_{t}) \) is deterministic given the selected batches of the first \( t-1 \) steps. Besides, when \( t = 1 \), (19) is adapted to
\begin{equation}
G(\theta_{1}) - \mathbb{E}_{1} \left[ G(\theta_{2}) \right] \geq D_{1} \mathbb{E}_{1} \left[ \| \nabla \bar{y}_{1} \|^{2} \right],
\end{equation}
where \( G(\theta_{1}) \) 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
\begin{equation}
D_{1} \sum_{t=1}^{T} \mathbb{E}_{1 \sim t} \left[ \| \nabla \bar{y}_{t} \|^{2} \right] \leq G(\theta_{1}) - \mathbb{E}_{1 \sim T} \left[ G(\theta_{T+1}) \right] \leq G(\theta_{1}).
\end{equation}

Therefore, there exists \( \tau \in \{1, \cdots, T\} \), s.t.
\begin{equation}
\mathbb{E}_{1 \sim \tau} \left[ \| \nabla \bar{y}_{\tau} \|^{2} \right] \leq \frac{G(\theta_{1})}{D_{1}T}.
\end{equation}

Then, we attempt to build a relationship between \( \nabla \bar{y}_{\tau} \) and \( \nabla_{\theta} G(\theta_{\tau}) \). Similar to Eq. (5), we have
\begin{equation}
\nabla \bar{y}_{\tau} = -\frac{\alpha_{\tau}}{B^{u}} \nabla_{u,\theta}^{T} L(\bar{x}^{u}_{\tau}; \bar{y}_{\tau}; \theta_{\tau}) \cdot \nabla_{\theta} G(\theta_{\tau}) = \frac{2\alpha_{\tau}}{B^{u}} J_{\theta} f(\bar{x}^{u}_{\tau}; \theta_{\tau}) \cdot \nabla_{\theta} G(\theta_{\tau}).
\end{equation}

Therefore,
\begin{equation}
\| \nabla \bar{y}_{\tau} \|^{2} = \sum_{i=1}^{B^{u}} \nabla_{i} \bar{y}_{\tau} \cdot \nabla_{i} \bar{y}_{\tau} = \frac{4\alpha_{\tau}^{2}}{(B^{u})^{2}} \nabla_{\theta} G(\theta_{\tau}) \cdot \left( \sum_{i=1}^{B^{u}} J_{\theta} f(\bar{x}^{u}_{\tau}; \theta_{\tau}) \cdot J_{\theta} f(\bar{x}^{u}_{\tau}; \theta_{\tau}) \right) \cdot \nabla_{\theta} G(\theta_{\tau}).
\end{equation}

Now consider the potential unlabeled batches \( \{ B_{k} : k = 1, \cdots, N^{l} \} \) of the \( \tau \)th step. Since, \( \mathcal{D}^l \subseteq \mathcal{D}^u \), we can assume \( x^{l}_{k} \in B_{k}, k = 1, \cdots, N^{l} \) 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

$$E_{1 \sim \tau} \left[ \| \nabla \bar{y}_r \|^2 \right] = E_{1 \sim \tau-1} \left[ E_r \left[ \| \nabla \bar{y}_r \|^2 \right] \right]$$

$$= E_{1 \sim \tau-1} \left[ \frac{4\alpha^2}{(B^u)^2} \nabla \theta^T G(\theta_r) \cdot E_r \left[ \sum_{i=1}^{B^n} J_{\theta}^T f(x_i^u; \theta_r) \cdot J_{\theta} f(x_i^u; \theta_r) \right] \cdot \nabla \theta G(\theta_r) \right]$$

$$\geq E_{1 \sim \tau-1} \left[ \frac{4\alpha^2}{(B^u)^2} \nabla \theta^T G(\theta_r) \cdot \left( \sum_{k=1}^{N^l} p_{k} J_{\theta}^T f(x_k^l; \theta_r) \cdot J_{\theta} f(x_k^l; \theta_r) \right) \cdot \nabla \theta G(\theta_r) \right]$$

$$\geq \frac{4pD_2^2}{(B^u)^2} E_{1 \sim \tau-1} \left[ \nabla \theta^T G(\theta_r) \cdot \left( \sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta_r) \cdot J_{\theta} f(x_k^l; \theta_r) \right) \cdot \nabla \theta G(\theta_r) \right].$$

(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 $\theta_r$.

By applying the chain rule, we have

$$\nabla \theta G(\theta) = \frac{2}{N^l} \sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta) \cdot (f(x_k^l; \theta) - y_k).$$

(27)

Since both $f(x_k^l; \theta)$ and $y_k$ are distributions on the category space, there exists a constant $R > 0$, s.t. $\|f(x_k^l; \theta) - y_k\| \leq R$. Therefore,

$$\sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta_r) \cdot J_{\theta} f(x_k^l; \theta_r)$$

$$\geq \frac{1}{R^2} \sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta_r) \cdot (f(x_k^l; \theta_r) - y_k) \cdot (f(x_k^l; \theta_r) - y_k)^T \cdot J_{\theta} f(x_k^l; \theta_r)$$

$$\geq \frac{1}{N^lR^2} \left( \sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta_r) \cdot (f(x_k^l; \theta_r) - y_k) \right) \cdot \left( \sum_{k=1}^{N^l} J_{\theta}^T f(x_k^l; \theta_r) \cdot (f(x_k^l; \theta_r) - y_k) \right)^T$$

$$= \frac{N^l}{4R^2} \nabla \theta^T G(\theta_r) \cdot \nabla \theta^T G(\theta_r).$$

(28)

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

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

$$u^T A^T v v^T A u = \|v\|^2 \|Au\|^2 \leq \|v\|^2 \|Au\|^2 \leq R^2 \|Au\|^2 = R^2 u^T A^T A u.$$ 

(29)


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

$$E_{1 \sim \tau} \left[ \| \nabla \bar{y}_r \|^2 \right] \geq \frac{pD_2^2 N^l}{(B^u)^2 R^2} E_{1 \sim \tau-1} \left[ \| \nabla \theta G(\theta_r) \|^2 \right] \geq \frac{pD_2^2 N^l}{(B^u)^2 R^2} \left( E_{1 \sim \tau-1} \left[ \| \nabla \theta G(\theta_r) \|^2 \right] \right)^2.$$

(30)

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

$$E_{1 \sim \tau-1} \left[ \| \nabla \theta G(\theta_r) \|^2 \right] \leq \frac{C}{\sqrt{T}}, \quad \text{where } C = \frac{B^u R}{D_2} \sqrt{\frac{G(\theta_1)}{pN^lD_1}}.$$

(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 $\beta_t$ is always set equal to the regular learning rate $\alpha_t$. 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 $5 \times 10^{-5}$ for SVHN. The batch size is 128 for both labeled and unlabeled data. The shape parameter $\gamma$ 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 $\gamma$ is set to 1.0.

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

<table>
<thead>
<tr>
<th>Layer</th>
<th>#Filters</th>
<th>Kernel Size</th>
<th>Stride</th>
<th>Paddings</th>
<th>Output Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution</td>
<td>128</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>32 × 32</td>
</tr>
<tr>
<td>Convolution</td>
<td>128</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>32 × 32</td>
</tr>
<tr>
<td>Convolution</td>
<td>128</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>32 × 32</td>
</tr>
<tr>
<td>MaxPooling</td>
<td>128</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>16 × 16</td>
</tr>
<tr>
<td>Dropout</td>
<td>Drop probability = 0.5</td>
<td>16 × 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convolution</td>
<td>256</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16 × 16</td>
</tr>
<tr>
<td>Convolution</td>
<td>256</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16 × 16</td>
</tr>
<tr>
<td>Convolution</td>
<td>256</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16 × 16</td>
</tr>
<tr>
<td>MaxPooling</td>
<td>128</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>8 × 8</td>
</tr>
<tr>
<td>Dropout</td>
<td>Drop probability = 0.5</td>
<td>8 × 8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convolution</td>
<td>512</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>6 × 6</td>
</tr>
<tr>
<td>Convolution</td>
<td>256</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6 × 6</td>
</tr>
<tr>
<td>Convolution</td>
<td>128</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6 × 6</td>
</tr>
<tr>
<td>AvgPooling</td>
<td>128</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1 × 1</td>
</tr>
<tr>
<td>Linear</td>
<td>128 → 10</td>
<td></td>
<td></td>
<td></td>
<td>1 × 1</td>
</tr>
</tbody>
</table>
References


