Supplementary Material – An Internal Covariate Shift Bounding Algorithm for Deep Neural Networks by Unitizing Layers' Outputs

You Huang Yuanlong Yu

In this document, the proofs of all the theorems are given, with an example of the unbounded Earth Mover (EM) distance. Then, we propose an algorithm for estimating the EM distances that measure Internal Covariate Shift (ICS) in practice, based on the Wasserstein Generative Adversarial Network (WGAN) [1]. The estimated distances for the unitization in network training are reported. Besides, for micro-batches, the proposed unitization is compared against Batch Normalization (BN) [3] and Group Normalization (GN) [5]. Code is available at `https://github.com/unknown9567/unitization.git`.

1. Proofs of the Theorems

1.1. Notations

- $W(p_l^{(t+\Delta t)}, p_l^{(t)})$: the EM distance between the distributions $p_l^{(t+\Delta t)}$ and $p_l^{(t)}$ at the $(t+\Delta t)$-th and tth iterations, respectively, for the lth layer's outputs
- $\mu_i^{(t)}$: the expectation of the ith element x_i of a random vector $\boldsymbol{x} \sim p_l^{(t)}$
- $(\sigma_i^{(t)})^2$: the variance of x_i
- $g(\boldsymbol{x})$: the vanilla unitization defined as

$$g(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{||\boldsymbol{x}||_2} & , ||\boldsymbol{x}||_2 \neq 0 \\ \boldsymbol{c} & , ||\boldsymbol{x}||_2 = 0 \end{cases}$$

- $g(\boldsymbol{x}; \alpha)$: the modified unitization with a parameter $\alpha \in [0, 1]$, defined as

$$g(\boldsymbol{x}; \alpha) = \begin{cases} \boldsymbol{c} & , ||\boldsymbol{x}||_2 = 0, \alpha = 1 \\ \dfrac{\boldsymbol{x}}{\alpha||\boldsymbol{x}||_2 + (1-\alpha)} & , other \end{cases}$$

- $g(\boldsymbol{x}; \boldsymbol{\alpha})$: the modified unitization with a parameter vector $\boldsymbol{\alpha} \in [0, 1]^d$, defined as

$$g(\boldsymbol{x}; \boldsymbol{\alpha}) = \begin{cases} \boldsymbol{0} & , ||\boldsymbol{x}||_2 = 0 \\ \big((||\boldsymbol{x}||_2 - 1) \cdot \mathrm{diag}(\boldsymbol{\alpha}) + E\big)^{-1} \boldsymbol{x} & , ||\boldsymbol{x}||_2 > 0 \end{cases}$$

  where $\mathrm{diag}(\boldsymbol{\alpha})$ is a diagonal matrix of $\boldsymbol{\alpha}$ and E is an identity matrix

- $g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)$: the practical unitization with a parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $\epsilon > 0$, defined as

$$g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon) = \left[\frac{1}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} \boldsymbol{\alpha} + (\boldsymbol{1} - \boldsymbol{\alpha})\right] \odot \boldsymbol{x},$$

  where $\odot$ represents element-wise production

1.2. Useful facts

There are two facts that can be derived directly from the EM distance, and they'll be used in the proof of the theorems.

Fact 1.

$$\begin{aligned} W(p_l^{(t+\Delta t)}, p_l^{(t)}) &= \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \\ &= \sup_{||f||_L \leq 1} \left| \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \right|, \end{aligned}$$

since f is a 1*-Lipschitz function iff $-f$ is a* 1*-Lipschitz function, and*

$$\left| \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \right| = \max \left\{ \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})], \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[-f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[-f(\boldsymbol{y})] \right\}.$$

Fact 2.

$$\begin{aligned} W(p_l^{(t+\Delta t)}, p_l^{(t)}) &= \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \\ &= \sup_{f: ||f||_L \leq 1, f(\mathbf{0})=0} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})], \end{aligned}$$

since for any 1*-Lipschitz function f,*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(\boldsymbol{y})\big] =& \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(\boldsymbol{x})\big] - f(\mathbf{0}) - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(\boldsymbol{y})\big] + f(\mathbf{0}) \\ =& \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(\boldsymbol{x}) - f(\mathbf{0})\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(\boldsymbol{y}) - f(\mathbf{0})\big] \\ =& \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[\tilde{f}(\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[\tilde{f}(\boldsymbol{y})\big], \end{aligned}$$

where $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\mathbf{0})$ is still a 1*-Lipschitz function, and satisfies $\tilde{f}(\mathbf{0}) = 0$.*

1.3. Proofs

Theorem 1. *Suppose that* $|\mu_i^{(t)}| < \infty, |\mu_i^{(t+\Delta t)}| < \infty, 1 \leq i \leq d$. *Then,*

$$W(p_l^{(t+\Delta t)}, p_l^{(t)}) \leq \sum_{i=1}^{d} (\sigma_i^{(t+\Delta t)})^2 + \sum_{i=1}^{d} (\sigma_i^{(t)})^2 + \Big(\sum_{i=1}^{d} (\mu_i^{(t+\Delta t)} - \mu_i^{(t)})^2 \Big)^{\frac{1}{2}} + 2.$$

Proof. Denote by $\mathbb{I}_A(x)$ an indicator function defined by $\mathbb{I}_A(x) = 1$ if $x \in A$, otherwise $\mathbb{I}_A(x) = 0$. Then, according to **Fact 1**,

$$\begin{aligned}
W(p_l^{(t+\Delta t)}, p_l^{(t)}) =& \sup_{||f||_L \leq 1} \big| \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \big| \\
\leq& \sup_{||f||_L \leq 1} \big| \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x}) - f(\boldsymbol{\mu}^{(t+\Delta t)})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y}) - f(\boldsymbol{\mu}^{(t)})] \big| + \big| f(\boldsymbol{\mu}^{(t+\Delta t)}) - f(\boldsymbol{\mu}^{(t)}) \big| \\
\leq& \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\big| f(\boldsymbol{x}) - f(\boldsymbol{\mu}^{(t+\Delta t)}) \big| \big] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\big| f(\boldsymbol{y}) - f(\boldsymbol{\mu}^{(t)}) \big| \big] + \big| f(\boldsymbol{\mu}^{(t+\Delta t)}) - f(\boldsymbol{\mu}^{(t)}) \big| \\
\leq& \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\big|\big| \boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)} \big|\big|_2 \big] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\big|\big| \boldsymbol{y} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \big] + \big|\big| \boldsymbol{\mu}^{(t+\Delta t)} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \\
=& \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\mathbb{I}_{||\boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)}||_2 \leq 1}(\boldsymbol{x}) \big|\big| \boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)} \big|\big|_2 \big] + \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\mathbb{I}_{||\boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)}||_2 > 1}(\boldsymbol{x}) \big|\big| \boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)} \big|\big|_2 \big] \\
&+ \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\mathbb{I}_{||\boldsymbol{y} - \boldsymbol{\mu}^{(t)}||_2 \leq 1}(\boldsymbol{y}) \big|\big| \boldsymbol{y} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \big] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\mathbb{I}_{||\boldsymbol{y} - \boldsymbol{\mu}^{(t)}||_2 > 1}(\boldsymbol{y}) \big|\big| \boldsymbol{y} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \big] + \big|\big| \boldsymbol{\mu}^{(t+\Delta t)} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \\
\leq& \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\mathbb{I}_{||\boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)}||_2 \leq 1}(\boldsymbol{x}) \cdot 1 \big] + \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \big[\mathbb{I}_{||\boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)}||_2 > 1}(\boldsymbol{x}) \big|\big| \boldsymbol{x} - \boldsymbol{\mu}^{(t+\Delta t)} \big|\big|_2^2 \big] \\
&+ \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\mathbb{I}_{||\boldsymbol{y} - \boldsymbol{\mu}^{(t)}||_2 \leq 1}(\boldsymbol{y}) \cdot 1 \big] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \big[\mathbb{I}_{||\boldsymbol{y} - \boldsymbol{\mu}^{(t)}||_2 > 1}(\boldsymbol{y}) \big|\big| \boldsymbol{y} - \boldsymbol{\mu}^{(t)} \big|\big|_2^2 \big] + \big|\big| \boldsymbol{\mu}^{(t+\Delta t)} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \\
\leq& 1 + \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}} \left[\sum_{i=1}^{d} (x_i - \mu_i^{(t+\Delta)})^2 \right] + 1 + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \left[\sum_{i=1}^{d} (y_i - \mu_i^{(t)})^2 \right] + \big|\big| \boldsymbol{\mu}^{(t+\Delta t)} - \boldsymbol{\mu}^{(t)} \big|\big|_2 \\
=& \sum_{i=1}^{d} (\sigma_i^{(t+\Delta)})^2 + \sum_{i=1}^{d} (\sigma_i^{(t)})^2 + \Big(\sum_{i=1}^{d} (\mu_i^{(t+\Delta t)} - \mu_i^{(t)})^2 \Big)^{\frac{1}{2}} + 2.
\end{aligned}$$

□

Theorem 2. *Suppose that $C > 0$ is a real number, and $n \geq 2$ is an integer. Then,*

$$\begin{aligned} W(p_l^{(t+\Delta t)}, p_l^{(t)}) &= \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})] \\ &\geq \big|\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f_{n,C}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f_{n,C}(\boldsymbol{y})]\big|, \end{aligned}$$

where $f_{n,C}$ is the 1-Lipschitz function defined as

$$f_{n,C}(\boldsymbol{x}) = \frac{1}{nC^{n-1}d^{\frac{1}{2}}}\bigg(\sum_{|x_i| \leq C} x_i^n + \sum_{x_i < -C} (-C)^n + \sum_{x_i > C} C^n\bigg).$$

Proof. According to **Fact 1**, it's obvious that the inequality holds if $f_{n,C}$ is a 1-Lipschitz function. Thus, only the proof of Lipschitz continuity of $f_{n,C}$ is required. For convenience, let $f_{n,C} = f_b/(nC^{n-1}d^{\frac{1}{2}})$, where

$$f_b(\boldsymbol{x}) = \sum_{|x_i| \leq C} x_i^n + \sum_{x_i < -C} (-C)^n + \sum_{x_i > C} C^n.$$

Then, we'll prove the Lipschitz continuity of f_b. In fact, for any $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$,

$$\begin{aligned} |f_b(\boldsymbol{v}) - f_b(\boldsymbol{w})| =& \bigg| \sum_{|v_i| \leq C} v_i^n + \sum_{v_i < -C} (-C)^n + \sum_{v_i > C} C^n - \sum_{|w_i| \leq C} w_i^n - \sum_{w_i < -C} (-C)^n - \sum_{w_i > C} C^n \bigg| \\ =& \bigg| \sum_{i:|v_i| \leq C, |w_i| \leq C} (v_i^n - w_i^n) + \sum_{i:|v_i| \leq C, w_i < -C} [v_i^n - (-C)^n] + \sum_{i:|v_i| \leq C, w_i > C} (v_i^n - C^n) \\ &+ \sum_{i:v_i < -C, |w_i| \leq C} [(-C)^n - w_i^n] + \sum_{i:v_i < -C, w_i < -C} [(-C)^n - (-C)^n] + \sum_{i:v_i < -C, w_i > C} [(-C)^n - C^n] \\ &+ \sum_{i:v_i > C, |w_i| \leq C} (C^n - w_i^n) + \sum_{i:v_i > C, w_i < -C} [C^n - (-C)^n] + \sum_{i:v_i > C, w_i > C} (C^n - C^n) \bigg| \\ =& \bigg| \sum_{i:|v_i| \leq C, |w_i| \leq C} (v_i - w_i) \sum_{m=0}^{n-1} v_i^{n-1-m} w_i^m + \sum_{i:|v_i| \leq C, w_i < -C} [v_i - (-C)] \sum_{m=0}^{n-1} v_i^{n-1-m} (-C)^m \\ &+ \sum_{i:|v_i| \leq C, w_i > C} (v_i - C) \sum_{m=0}^{n-1} v_i^{n-1-m} C^m + \sum_{i:v_i < -C, |w_i| \leq C} (-C - w_i) \sum_{m=0}^{n-1} (-C)^{n-1-m} w_i^m \\ &+ \sum_{i:v_i < -C, w_i > C} (-C - C) \sum_{m=0}^{n-1} (-C)^{n-1-m} C^m + \sum_{i:v_i > C, |w_i| \leq C} (C - w_i) \sum_{m=0}^{n-1} C^{n-1-m} w_i^m \\ &+ \sum_{i:v_i > C, w_i < -C} [C - (-C)] \sum_{m=0}^{n-1} C^{n-1-m} (-C)^m \bigg| \\ \leq& nC^{n-1} \bigg(\sum_{i:|v_i| \leq C, |w_i| \leq C} |v_i - w_i| + \sum_{i:|v_i| \leq C, w_i < -C} |v_i - (-C)| + \sum_{i:|v_i| \leq C, w_i > C} |v_i - C| \\ &+ \sum_{i:v_i < -C, |w_i| \leq C} |-C - w_i| + \sum_{i:v_i < -C, w_i > C} |-C - C| + \sum_{i:v_i > C, |w_i| \leq C} |C - w_i| \\ &+ \sum_{i:v_i > C, w_i < -C} |C - (-C)| \bigg) \end{aligned}$$

$$
\begin{aligned}
\leq & nC^{n-1}\Bigg(\sum_{i:|v_i|\leq C,|w_i|\leq C}|v_i-w_i|+\sum_{i:|v_i|\leq C,w_i<-C}|v_i-w_i|+\sum_{i:|v_i|\leq C,w_i>C}|v_i-w_i| \\
& +\sum_{i:v_i<-C,|w_i|\leq C}|v_i-w_i|+\sum_{i:v_i<-C,w_i>C}|v_i-w_i|+\sum_{i:v_i>C,|w_i|\leq C}|v_i-w_i|+\sum_{i:v_i>C,w_i<-C}|v_i-w_i|\Bigg) \\
\leq & nC^{n-1}\sum_{i=1}^{d}|v_i-w_i| \\
= & nC^{n-1}||\boldsymbol{v}-\boldsymbol{w}||_1.
\end{aligned}
$$

Note that Hölder's inequality implies that

$$
||\boldsymbol{x}||_1=\sum_{i=1}^{d}|x_i|\leq\left(\sum_{j=1}^{d}|x_j|^2\right)^{1/2}\left(\sum_{k=1}^{d}1\right)^{1/2}=d^{\frac{1}{2}}||\boldsymbol{x}||_2
$$

for any $\boldsymbol{x}\in\mathbb{R}^d$. Thus,

$$
|f_b(\boldsymbol{v})-f_b(\boldsymbol{w})|\leq nC^{n-1}d^{\frac{1}{2}}||\boldsymbol{v}-\boldsymbol{w}||_2
$$

Therefore,

$$
|f_{n,C}(\boldsymbol{v})-f_{n,C}(\boldsymbol{w})|=\frac{1}{nC^{n-1}d^{\frac{1}{2}}}|f_b(\boldsymbol{v})-f_b(\boldsymbol{w})|\leq||\boldsymbol{v}-\boldsymbol{w}||_2
$$

implying $f_{n,C}$ is a 1-Lipschitz function. □

Example of the Unbounded EM Distance: Given $C'>0$, consider the two distributions as follows:

$$
p_l^{(t+\Delta t)}(\boldsymbol{x})=\begin{cases}\dfrac{1}{\int_{[C'/2,C']^d}1d\boldsymbol{x}} & ,\boldsymbol{x}\in[C'/2,C']^d\\ 0 & ,other\end{cases}
$$

and

$$
p_l^{(t)}(\boldsymbol{y})=\begin{cases}\dfrac{1}{\int_{[0,C'/4]^d}1d\boldsymbol{y}} & ,\boldsymbol{y}\in[0,C'/4]^d\\ 0 & ,other\end{cases}
$$

Then, the supports of the distributions are subsets of $[0,C']^d$ and for any $i\in\{1,2,\ldots,d\}$,

$$
\mathbb{E}_{\boldsymbol{x}\sim p_l^{(t+\Delta t)}}[x_i^n]>\frac{C'^n}{2^n},\mathbb{E}_{\boldsymbol{y}\sim p_l^{(t)}}[y_i^n]<\frac{C'^n}{4^n}
$$

Thus, according to **Theorem 2**, the lower bound on the EM distance between $p_l^{(t+\Delta)}$ and $p_l^{(t)}$ is

$$
\begin{aligned}
W(p_l^{(t+\Delta t)},p_l^{(t)})\geq & \frac{1}{nC'^{n-1}d^{\frac{1}{2}}}\left|\mathbb{E}_{\boldsymbol{x}\sim p_l^{(t+\Delta t)}}\left[\sum_{i=1}^{d}x_i^n\right]-\mathbb{E}_{\boldsymbol{y}\sim p_l^{(t)}}\left[\sum_{i=1}^{d}y_i^n\right]\right| \\
= & \frac{1}{nC'^{n-1}d^{\frac{1}{2}}}\left|\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{x}\sim p_l^{(t+\Delta t)}}[x_i^n]-\mathbb{E}_{\boldsymbol{y}\sim p_l^{(t)}}[y_i^n]\right| \\
> & \frac{(2^{-n}-4^{-n})d^{\frac{1}{2}}}{n}C'
\end{aligned}
$$

Note that n and d are fixed, and then the lower bound is dominated by C'. Thus, the distance is unbounded and would go to infinity as $C'\to\infty$. Intuitively, the samples from $p_l^{(t+\Delta t)}$ can be regarded as the scaled and shifted samples from $p_l^{(t)}$. The unstable scale and center of the distributions lead to the unbounded lower bound, implying the unbounded distance and upper bound. The distributions of deep layers' outputs might perform in the same way if they are not normalized. In contrast, the upper bound on the distance for the outputs processed by BN is relatively stable (though it depends on the moments with noise), and can constrain the distance to a reasonable range in some cases discussed in the paper.

Theorem 3.1. *Suppose that for $\boldsymbol{x} \sim p_l^{(t)}$, $g(\boldsymbol{x}) \sim p_U^{(t)}$. Then,*

$$W(p_U^{(t+\Delta t)}, p_U^{(t)}) \leq 2.$$

Proof. Note that for any 1-Lipschitz function f such that $f(\mathbf{0}) = 0$,

$$\big|f(g(\boldsymbol{x}))\big| = \big|f(g(\boldsymbol{x})) - f(\mathbf{0})\big| \leq \big|\big|g(\boldsymbol{x}) - \mathbf{0}\big|\big|_2 = 1, \forall \boldsymbol{x}$$

which yields

$$-1 \leq f(g(\boldsymbol{x})) \leq 1, \forall \boldsymbol{x}$$

Therefore, according to **Fact 2**,

$$\begin{aligned}
W(p_U^{(t+\Delta t)}, p_U^{(t)}) &= \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x}))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y}))\big] \\
&= \sup_{f:||f||_L \leq 1, f(\mathbf{0})=0} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x}))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y}))\big] \\
&\leq \sup_{f:||f||_L \leq 1, f(\mathbf{0})=0} \mathbb{E}_{\boldsymbol{x} \sim p^{(t+\Delta)}}[1] - \mathbb{E}_{\boldsymbol{y} \sim p^{(t)}}[-1] \\
&= 2.
\end{aligned}$$

□

Theorem 3.2. *Suppose that for $\alpha \in [0, 1]$ and $\boldsymbol{x} \sim p_l^{(t)}$, $g(\boldsymbol{x}; \alpha) \sim p_U^{(t)}$. Then,*

$$W(p_U^{(t+\Delta t)}, p_U^{(t)}) \leq \mathbb{I}_{\alpha=0}(\alpha) \cdot (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2]) + \mathbb{I}_{\alpha>0}(\alpha) \cdot \frac{2}{\alpha}.$$

Proof. Note that given a 1-Lipschitz function f that satisfies $f(\mathbf{0}) = 0$, for any $\boldsymbol{x}$ and $\alpha \in [0, 1]$,

$$\begin{aligned}
\big|f(g(\boldsymbol{x}; \alpha))\big| &= \big|f(g(\boldsymbol{x}; \alpha)) - f(\mathbf{0})\big| \\
&\leq \big|\big|g(\boldsymbol{x}; \alpha) - \mathbf{0}\big|\big|_2 \\
&= \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{x}) \cdot \frac{||\boldsymbol{x}||_2}{\alpha||\boldsymbol{x}||_2 + (1-\alpha) \times 1} + \mathbb{I}_{||\boldsymbol{x}||_2=0, \alpha=1}(\boldsymbol{x}, \alpha) \cdot 1 + \mathbb{I}_{||\boldsymbol{x}||_2=0, \alpha<1}(\boldsymbol{x}, \alpha) \cdot 0 \\
&= \mathbb{I}_{||\boldsymbol{x}||_2>0, \alpha=0}(\boldsymbol{x}, \alpha) \cdot \frac{||\boldsymbol{x}||_2}{\alpha||\boldsymbol{x}||_2 + (1-\alpha) \times 1} + \mathbb{I}_{||\boldsymbol{x}||_2>0, \alpha>0}(\boldsymbol{x}, \alpha) \cdot \frac{1}{\alpha + (1-\alpha)/||\boldsymbol{x}||_2} + \mathbb{I}_{||\boldsymbol{x}||_2=0, \alpha=1}(\boldsymbol{x}, \alpha) \cdot 1 \\
&\leq \mathbb{I}_{||\boldsymbol{x}||_2>0, \alpha=0}(\boldsymbol{x}, \alpha) \cdot ||\boldsymbol{x}||_2 + \mathbb{I}_{||\boldsymbol{x}||_2>0, \alpha>0}(\boldsymbol{x}, \alpha) \cdot \frac{1}{\alpha} + \mathbb{I}_{||\boldsymbol{x}||_2=0, \alpha>0}(\boldsymbol{x}, \alpha) \cdot 1 \\
&\leq \mathbb{I}_{\alpha=0}(\alpha) \cdot ||\boldsymbol{x}||_2 + \mathbb{I}_{\alpha>0}(\alpha) \cdot \frac{1}{\alpha}
\end{aligned}$$

which yields

$$-\mathbb{I}_{\alpha=0}(\alpha) \cdot ||\boldsymbol{x}||_2 - \mathbb{I}_{\alpha>0}(\alpha) \cdot \frac{1}{\alpha} \leq f(g(\boldsymbol{x}; \alpha)) \leq \mathbb{I}_{\alpha=0}(\alpha) \cdot ||\boldsymbol{x}||_2 + \mathbb{I}_{\alpha>0}(\alpha) \cdot \frac{1}{\alpha}.$$

Therefore, according to **Fact 2**,

$$\begin{aligned}
W(p_U^{(t+\Delta t)}, p_U^{(t)}) &= \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x}; \alpha))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y}; \alpha))\big] \\
&= \sup_{f:||f||_L \leq 1, f(\mathbf{0})=0} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x}; \alpha))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y}; \alpha))\big] \\
&\leq \mathbb{I}_{\alpha=0}(\alpha) \cdot (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2]) + \mathbb{I}_{\alpha>0}(\alpha) \cdot \frac{2}{\alpha}.
\end{aligned}$$

□

Theorem 3.3. *Suppose that for $\boldsymbol{\alpha} \in [0,1]^d$ and $\boldsymbol{x} \sim p_l^{(t)}$, $g(\boldsymbol{x};\boldsymbol{\alpha}) \sim p_U^{(t)}$. Then,*

$$W(p_U^{(t+\Delta t)}, p_U^{(t)}) \leq \mathbb{I}_{\min_j \alpha_j > 0}(\boldsymbol{\alpha}) \cdot \frac{2}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j = 0}(\boldsymbol{\alpha}) \cdot (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2] + 2).$$

Proof. Given a 1-Lipschitz function f satisfying $f(\mathbf{0}) = 0$, for any $\boldsymbol{x}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in [0,1]^d$,

$$\begin{aligned}
\big|f(g(\boldsymbol{x};\boldsymbol{\alpha}))\big| \leq & \big|\big|g(\boldsymbol{x};\boldsymbol{\alpha})\big|\big|_2 = \mathbb{I}_{||\boldsymbol{x}||_2=0}(\boldsymbol{x}) \cdot 0 + \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{y}) \cdot \left[\sum_{i=1}^{d} \left(\frac{x_i}{\alpha_i ||\boldsymbol{x}||_2 + (1-\alpha_i) \cdot 1}\right)^2\right]^{\frac{1}{2}} \\
\leq & \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{x}) \cdot \left[\sum_{i=1}^{d} \left(\frac{x_i}{\min_j \alpha_j ||\boldsymbol{x}||_2 + (1-\alpha_j)}\right)^2\right]^{\frac{1}{2}} \\
\leq & \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{x}) \cdot \frac{||\boldsymbol{x}||_2}{\min_j \alpha_j ||\boldsymbol{x}||_2 + (1-\alpha_j)} \\
= & \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{x}) \cdot \left(\mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{1}{\min_j \alpha_j + (1-\alpha_j)/||\boldsymbol{x}||_2} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \frac{||\boldsymbol{x}||_2}{\min_j \alpha_j ||\boldsymbol{x}||_2 + (1-\alpha_j)}\right) \\
\leq & \mathbb{I}_{||\boldsymbol{x}||_2>0}(\boldsymbol{x}) \cdot \left(\mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{1}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \max\big\{||\boldsymbol{x}||_2, 1\big\}\right) \\
\leq & \mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{1}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \max\big\{||\boldsymbol{x}||_2, 1\big\}.
\end{aligned}$$

which yields

$$\begin{aligned}
f(g(\boldsymbol{x};\boldsymbol{\alpha})) \geq & -\mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{1}{\min_j \alpha_j} - \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \max\big\{||\boldsymbol{x}||_2, 1\big\} \\
f(g(\boldsymbol{x};\boldsymbol{\alpha})) \leq & \mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{1}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \max\big\{||\boldsymbol{x}||_2, 1\big\}.
\end{aligned}$$

Therefore, according to **Fact 2**,

$$\begin{aligned}
W(p_U^{(t+\Delta t)}, p_U^{(t)}) = & \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x};\boldsymbol{\alpha}))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y};\boldsymbol{\alpha}))\big] \\
= & \sup_{f:||f||_L \leq 1, f(\mathbf{0})=0} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\big[f(g(\boldsymbol{x};\boldsymbol{\alpha}))\big] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\big[f(g(\boldsymbol{y};\boldsymbol{\alpha}))\big] \\
\leq & \mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{2}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \left(\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}\Big[\max\big\{||\boldsymbol{x}||_2, 1\big\}\Big] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}\Big[\max\big\{||\boldsymbol{y}||_2, 1\big\}\Big]\right) \\
\leq & \mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{2}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot \left(\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[1] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[1]\right) \\
\leq & \mathbb{I}_{\min_j \alpha_j>0}(\boldsymbol{\alpha}) \cdot \frac{2}{\min_j \alpha_j} + \mathbb{I}_{\min_j \alpha_j=0}(\boldsymbol{\alpha}) \cdot (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2] + 2).
\end{aligned}$$

□

Theorem 3.4. *Suppose that for $\boldsymbol{\alpha} \in \mathbb{R}^d, \epsilon > 0$ and $\boldsymbol{x} \sim p_l^{(t)}$, $g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon) \sim p_g^{(t)}$. Then,*

$$W(p_g^{(t+\Delta t)}, p_g^{(t)}) \leq 2||\boldsymbol{\alpha}||_\infty + ||\mathbf{1} - \boldsymbol{\alpha}||_\infty (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2]).$$

Proof. Given a 1-Lipschitz function f satisfying $f(\mathbf{0}) = 0$, for any $\boldsymbol{x}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{R}^d$,

$$\begin{aligned}
|f(g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon))| \leq & ||g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)||_2 \\
= & \left|\left| \frac{\boldsymbol{x}}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} \odot \boldsymbol{\alpha} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \boldsymbol{x} \right|\right|_2 \\
\leq & \left|\left| \frac{\boldsymbol{x}}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} \odot \boldsymbol{\alpha} \right|\right|_2 + \left|\left| (\mathbf{1} - \boldsymbol{\alpha}) \odot \boldsymbol{x} \right|\right|_2 \\
= & \frac{1}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} \left(\sum_{i=1}^d \alpha_i^2 x_i^2 \right)^{1/2} + \left(\sum_{i=1}^d (1 - \alpha_i)^2 x_i^2 \right)^{1/2} \\
\leq & \frac{1}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} \left(\max_{1 \leq i \leq d} \alpha_i^2 \sum_{i=1}^d x_i^2 \right)^{1/2} + \left(\max_{1 \leq i \leq d} (1 - \alpha_i)^2 \sum_{i=1}^d x_i^2 \right)^{1/2} \\
= & ||\boldsymbol{\alpha}||_\infty \cdot \frac{||\boldsymbol{x}||_2}{\sqrt{||\boldsymbol{x}||_2^2 + \epsilon}} + ||\mathbf{1} - \boldsymbol{\alpha}||_\infty ||\boldsymbol{x}||_2 \\
\leq & ||\boldsymbol{\alpha}||_\infty + ||\mathbf{1} - \boldsymbol{\alpha}||_\infty ||\boldsymbol{x}||_2
\end{aligned}$$

which yields

$$-||\boldsymbol{\alpha}||_\infty - ||\mathbf{1} - \boldsymbol{\alpha}||_\infty ||\boldsymbol{x}||_2 \leq f(g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)) \leq ||\boldsymbol{\alpha}||_\infty + ||\mathbf{1} - \boldsymbol{\alpha}||_\infty ||\boldsymbol{x}||_2.$$

Therefore, according to **Fact 2**,

$$\begin{aligned}
W(p_g^{(t+\Delta t)}, p_g^{(t)}) = & \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}} \left[f(g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)) \right] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \left[f(g(\boldsymbol{y}; \boldsymbol{\alpha}, \epsilon)) \right] \\
= & \sup_{f: ||f||_L \leq 1, f(\mathbf{0}) = 0} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta)}} \left[f(g(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)) \right] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}} \left[f(g(\boldsymbol{y}; \boldsymbol{\alpha}, \epsilon)) \right] \\
\leq & 2||\boldsymbol{\alpha}||_\infty + ||\mathbf{1} - \boldsymbol{\alpha}||_\infty (\mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[||\boldsymbol{x}||_2] + \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[||\boldsymbol{y}||_2]).
\end{aligned}$$

□

2. Earth Mover Distance Estimation

In the theoretical analysis, the EM distance has been used to measure ICS. However, whether there is a correlation between the performance of networks and ICS in practice is unclear. Thus, an algorithm is proposed to estimate the EM distance in network training, and the results of the following experiments yield evidence that ICS is related to the networks' performance.

2.1. Algorithm

By definition, the accurate EM distance is obtained by optimizing a function f over the 1-Lipschitz function space:

$$W(p_l^{(t+\Delta t)}, p_l^{(t)}) = \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{y} \sim p_l^{(t)}}[f(\boldsymbol{y})].$$

Thus, the proposed algorithm aims to get the optimal 1-Lipschitz function f^* and use f^* to compute the difference between the expectations. In WGAN [1], the 1-Lipschitz function f is parameterized by a neural network f_w with a scalar output in the algorithm. Then, the parameters of f_w are bounded to satisfy the constraint $||f_w||_L \leq 1$. Furthermore, the constraint can be relaxed such that w is required to satisfy $||f_w||_L \leq K$ for some $K > 0$, and then the EM distance is $K \cdot W(p_l^{(t+\Delta t)}, p_l^{(t)})$. For convenience, the parameter space is denoted by $\mathcal{W} = \big\{w \big| ||f_w||_L \leq K\big\}$. Then, a suboptimal K-Lipschitz function f_w^* is obtained by training f_w on the dataset consisting of the real and generated images. The EM distance is computed by f_w^*'s outputs with the real and generated images as the inputs.

In a convolutional neural network $f_{1:L}$ with L layers, the EM distance for measuring ICS is estimated in the same way. For the lth ($l < L$) convolutional layer of $f_{1:L}$, the multi-channel outputs can be treated as images, though there are more than 3 channels. Thus, analogously, the proposed algorithm trains f_w to maximize the same objective function of WGANs, with the training samples generated by the local networks $f_{1:l}^{(t)}$ and $f_{1:l}^{(t+\Delta t)}$ formed by the previous l layers of $f_{1:L}$ at the tth and $t + \Delta t$th iterations, respectively. Then, the algorithm estimates the EM distance by f_w^* similarly.

To formulate the process, the approximated EM distance, which replaces the 1-Lipschitz function space with $\mathcal{W}$, is defined as

$$\widetilde{W}(p_l^{(t+\Delta t)}, p_l^{(t)}) = \sup_{w \in \mathcal{W}} E_{\boldsymbol{x} \sim p_l^{(t+\Delta t)}}[f_w(\boldsymbol{x})] - E_{\boldsymbol{y} \sim p_l^{(t)}}[f_w(\boldsymbol{y})].$$

Then, in practice, $\widetilde{W}(p_l^{(t+\Delta t)}, p_l^{(t)})$ is further approximated by the empirical estimation:

$$\widetilde{W}(p_l^{(t+\Delta t)}, p_l^{(t)}) \approx \sup_{w \in \mathcal{W}} \frac{1}{N} \sum_{i=1}^{N} f_w(f_{1:l}^{(t+\Delta t)}(\boldsymbol{x}_i)) - \frac{1}{N} \sum_{i=1}^{N} f_w(f_{1:l}^{(t)}(\boldsymbol{x}_i)),$$

where N is the number of samples and $\boldsymbol{x}_i$ is a sample from the dataset for $f_{1:L}$. The algorithm is presented in Algorithm 3.

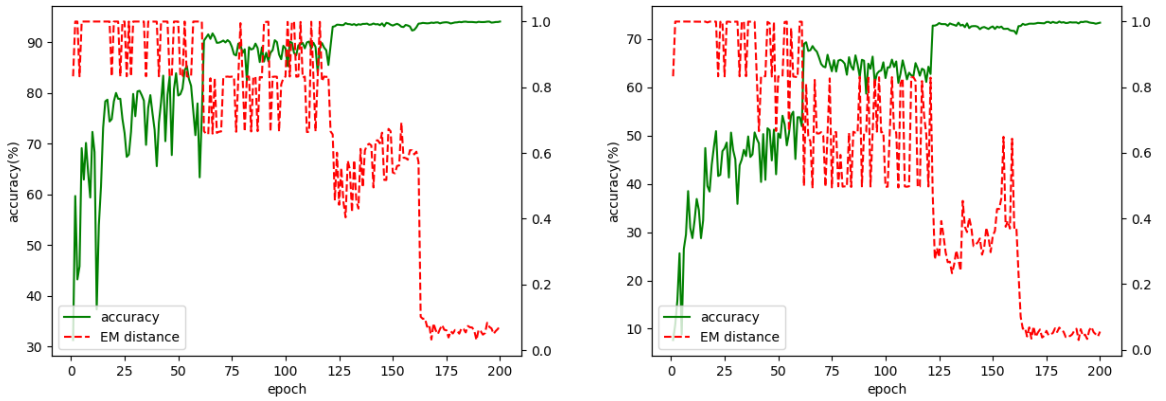
Algorithm 3 EM Distance Estimation Algorithm

Input: datasets $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$, local networks $f_{1:l}^{(t+\Delta t)}$ and $f_{1:l}^{(t)}$, initialized f_w, parameters T, n and c
Output: estimated EM distance d

1: **for** $i \leftarrow 1$ to T **do**
2: Sample $\{\boldsymbol{x}_j\}_{j=1}^{n}$ randomly from $\mathcal{D}_{train}$
3: $g_w \leftarrow \nabla_w \Big[\frac{1}{n} \sum_{j=1}^{n} \big(f_w(f_{1:l}^{(t+\Delta t)}(\boldsymbol{x}_j)) - f_w(f_{1:l}^{(t)}(\boldsymbol{x}_j))\big)\Big]$
4: Update w by g_w
5: $w \leftarrow clip(w, -c, c)$
6: **end for**
7: $d \leftarrow \frac{1}{|\mathcal{D}_{test}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{test}} \big(f_w(f_{1:l}^{(t+\Delta t)}(\boldsymbol{x})) - f_w(f_{1:l}^{(t)}(\boldsymbol{x}))\big)$

2.2. Experiments

The EM distances for some specific unitization layers of ResNet-110s are estimated. The ResNet-110s, *i.e.*, $f_{1:L}$s, are trained on CIFAR-10 and CIFAR-100 datasets with the data augmentation method [4], while the f_ws corresponding to the specific layers are trained on the same dataset without data augmentation.



(a) CIFAR-10 (b) CIFAR-100

Figure 1: Illustration of test accuracy of ResNet-110s on the CIFAR datasets and the average EM distances. On both the CIFAR-10 (a) and CIFAR-100 (b) datasets, the accuracy increases significantly in about 60th, 120th and 160th epochs, and the average EM distance sharply drops concurrently.

Network Architectures: Each f_w is parameterized as a vanilla CNN. The inputs of f_w are the outputs of the corresponding layer, followed by four 3×3 convolutions on the feature maps of sizes $\{64, 64, 128, 128\}$ with strides of $\{2, 1, 2, 1\}$, respectively. The convolutional outputs are activated using ReLU but not normalized/unitized. A fully-connected layer with a sigmoid activation is applied to generate the network's outputs.

Implementation Details: The experiments of training ResNet-110s on the CIFAR datasets have been described in the paper, and the experiments of estimating the distances are described as follows. The EM distances are computed per epoch. At the end of each epoch, the weights of the current local network $f_{1:l}^{(t+\Delta t)}$ are first saved for the next epoch. Then another local network $f_{1:l}^{(t)}$ is initialized by the weights saved in the previous epoch. f_w is trained by Algorithm 3 with these two local networks as the algorithm's inputs. For each estimated EM distance, the same hyperparameters are used: the iteration number T is $1.5k$; the mini-batch size n is 128; the bound c is 0.01. f_w is initialized by the same weights that are generated using the method [2]. Finally, the average EM distance for the deep layers including the 60th, 69th, 78th, 87th, 96th and 105th layers, along with the test accuracy of the ResNet-110s, are reported.

Results: As is shown in Figure 1, the accuracy of each experiment significantly increases in about 60th, 120th and 160th epoch as the learning rate decreases. Meanwhile, the average EM distance dramatically drops in these epochs. The results demonstrate that the EM distance is a suitable ICS measure since it drops immediately as the learning rate decreases, in which ICS obviously decreases. Then, based on the effectiveness of the EM distance in measuring ICS, the results further substantiate that ICS is related to the performance of networks.

3. Experiments of the Unitization for Micro-Batches

The unitization is evaluated in the case of micro-batches to further demonstrate the performance. For comparison, BN and GN [5] techniques are also evaluated in this case. Only the CIFAR-10 dataset with the same data augmentation in the previous experiments are used.

Network Architectures: The unitization, BN and GN are evaluated with the same ResNet-110s in the previous experiments. For GN, there are different ways of group division, according to the experiments in [5]. For convenience, denote by G the number of the groups in GN, with the value in $\{1, 2, 4, 8, 16\}$. Furthermore, for a network, there are two types of settings for G in each convolutional layer: fixing the numbers of (1) the groups or (2) the channels per group. Thus, there are 10 ways of group division in total.

Implementation Details: The ResNet-110s are trained with the same settings as the previous experiments, except for the

batch size, denoted by s. Unlike [5], we experiment on **only one GPU**, and report the median accuracy of 3 runs for each experiment of $s = 2$. In fact, for the cases of $s \in \{4, 8, 16\}$, BN still outperforms GN. Thus, only the case of $s = 2$, in which BN degrades, is considered.

Table 1: Classification accuracy ($\%$) on the CIFAR-10 dataset

Groups (G)					Channels per group					BatchNorm	Unitization
1	2	4	8	16	1	2	4	8	16		
10.00	10.00	10.00	71.63	73.76	79.54	73.58	70.68	72.04	10.00	71.50	**81.88**

Results: As is shown in Table 1, the unitization still outperforms the other techniques for micro-batches. Some networks with GN cannot even converge, where the accuracy (of only 10%) degrades, and BN suffers from highly noisy estimations in this case, with the almost lowest accuracy among all the comparable results ($> 10\%$). The results further demonstrate the effectiveness of the unitization in controlling ICS.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[4] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[5] Yuxin Wu and Kaiming He. Group normalization. *arXiv preprint arXiv:1803.08494*, 2018.