# Measuring Asymmetric Gradient Discrepancy in Parallel Continual Learning (Appendix)

Fan Lyu, Qing Sun, Fanhua Shang, Liang Wan, Wei Feng*
College of Intelligence and Computing, Tianjin University
{fanlyu, sssunqing, fhshang, lwan}@tju.edu.cn, wfeng@ieee.org
Source Code: https://github.com/fanlyu/maxdo

## A. Dataset Construction

For effective transformation, several requirements are needed: (1) Random label set for each task, in which the data stream length of each task can be different; (2) Random timeline for each label set, in which the debut of each task can be any time between the first access of the former and latter tasks. For simplicity, we omit all blank time that all data streams are unavailable.

- *Parallel Split EMNIST* (PS-EMNIST): We split EMNIST (62 classes) into 5 tasks and randomly generate 3 label sets for each task and 3 timelines for each label set (say 9 different situations). The size of the label set for each task is set to $\{12, 12, 12, 13, 13\}$.
- *Parallel Split CIFAR-100* (PS-CIFAR-100): We split CIFAR-100 into 10 tasks and randomly generate 3 label sets for each task and 3 timelines for each label set. The size of the label set for each task is set to 10.
- *Parallel Split ImageNet-TINY* (PS-ImageNet-TINY): We split it into 10 tasks w.r.t. random 3 label sets, and each label set has 3 randomly generated timelines. The size of the label set for each task is set to 20.

## B. Proof of Lemma 1 on AGD

As an asymmetric metric, the proposed Asymmetric Gradient Discrepancy (AGD) measure needs to satisfy the two features in Lemma 1.

**Proof:** Given three arbitrary gradients $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, and assume that at least one gradient is non-zero, we have

(1) If $\mathbf{x} = \mathbf{y}$, $D(\mathbf{x}, \mathbf{y}) = 0$.

(2) **Positivity**: If $\mathbf{x} \neq \mathbf{y}$, then $\|\mathbf{x} - \mathbf{y}\| \neq 0$, and we have $D(\mathbf{x}, \mathbf{y}) = \dfrac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{y}\| + \|\mathbf{x} - \mathbf{y}\|} > 0$.

(3) **The triangle inequality**:

$$\begin{aligned} \frac{\|\mathbf{x} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{z}\|} &= \frac{\|\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{y} + \mathbf{y} - \mathbf{z}\|} \\ &\leq \frac{\|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|} \\ &= \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|} + \frac{\|\mathbf{y} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|} \\ &\leq \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{z}\| + \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|} + \frac{\|\mathbf{y} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\|} \\ &\leq \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{y}\| + \|\mathbf{x} - \mathbf{y}\|} + \frac{\|\mathbf{y} - \mathbf{z}\|}{\|\mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\|}. \end{aligned} \tag{1}$$

(4) **Asymmetric**: $D(\mathbf{x}, \mathbf{y}) = \dfrac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{y}\| + \|\mathbf{x} - \mathbf{y}\|}$, and $D(\mathbf{y}, \mathbf{x}) = \dfrac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{x}\| + \|\mathbf{x} - \mathbf{y}\|}$. Thus, it is obvious that $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$ is not always satisfied when $\mathbf{x} \neq \mathbf{y}$ and depends on the magnitude $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ .

Therefore, the proposed AGD is an asymmetric metric. ■

## C. Proof of Corollary 1

Let us review the definition of AGD:

$$\widehat{D}(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{y}\| + \|\mathbf{x} - \mathbf{y}\|}. \tag{2}$$

$\widehat{D}(\mathbf{x}, \mathbf{y})$ represents the gradient influence from $\mathbf{x}$ to $\mathbf{y}$. The nature of this asymmetric measure is the norm effect should *only* be from gradient difference $\|\mathbf{x} - \mathbf{y}\|$ to $\|\mathbf{y}\|$ rather than to both $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$. That is, the discrepancy should only depend on the ratio $\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}$, which can be further reduced to

$$\begin{aligned} \frac{\|\mathbf{x} - \mathbf{y}\|}{\|\mathbf{y}\|} &= \frac{\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos \angle \mathbf{x}, \mathbf{y}}}{\|\mathbf{y}\|} \\ &= \sqrt{\left(\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\right)^2 - 2\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} \cos \angle \mathbf{x}, \mathbf{y} + 1}. \end{aligned} \tag{3}$$

*Corresponding author.

It is easy to know that

$$\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|+\|\mathbf{x}-\mathbf{y}\|}=1-\frac{1}{1+\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}}. \tag{4}$$

Because $\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}\geq 0$, $\widehat{D}(\mathbf{x},\mathbf{y})\in[0,1]$.

In the paper, we illustrate the proposed AGD is an asymmetric measure of gradient discrepancy because $\widehat{D}(\mathbf{x},\mathbf{y})$ brings a tolerance when $\|\mathbf{x}\|\ll\|\mathbf{y}\|$ instead of the absolute difference between them. To analyze the values of gradient discrepancy measure $D$ regarding $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}$, we consider the following asymmetric limits with $\|y\|\neq 0$:

- $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty} D$: When $\|\mathbf{x}\|\gg\|\mathbf{y}\|$, the conflict should be large from $\mathbf{x}$ to $\mathbf{y}$;
- $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0} D$: When $\|\mathbf{x}\|\ll\|\mathbf{y}\|$, the conflict is acceptable to some extend and should approach a tolerance value that less than $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty} D$.

We show the two limits for different discrepancy measures including Cosine Similarity, Euclidean Distance, Normalized Euclidean Distance, and AGD.

**Cosine Similarity**: Using the Cosine Similarity to measure the discrepancy has no relevance to the magnitude difference.

$$\begin{aligned}&\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}1-\frac{\mathbf{x}^\top\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\\ =&\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}1-\frac{\mathbf{x}^\top\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\\ =&\ 1-\cos\angle\mathbf{x},\mathbf{y}.\end{aligned} \tag{5}$$

**Euclidean Distance**: When $\|\mathbf{y}\|\neq 0$, we have

$$\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|}=\frac{1}{1+\|\mathbf{y}\|\cdot\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}}. \tag{6}$$

Thus, we have

$$\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}1-\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|}=\frac{\|\mathbf{y}\|}{1+\|\mathbf{y}\|}, \tag{7}$$

$$\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}1-\frac{1}{1+\|\mathbf{x}-\mathbf{y}\|}=1. \tag{8}$$

When $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0$, by using the Euclidean Distance highly depends on $\|\mathbf{y}\|$, which makes it unpredictable.

**Normalized Euclidean Distance**: When $\|\mathbf{y}\|\neq 0$, we have

$$\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{x}\|+\|\mathbf{y}\|}=\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}\frac{\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}}{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}+1}=1, \tag{9}$$

$$\begin{aligned}&\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{x}\|+\|\mathbf{y}\|}\\ =&\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}\frac{\sqrt{\left(\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\right)^2-2\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\cos\angle\mathbf{x},\mathbf{y}+1}}{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}+1}\\ =&\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}\sqrt{\frac{2\cos\angle\mathbf{x},\mathbf{y}+2}{\left(\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}+1\right)^2}-\frac{2\cos\angle\mathbf{x},\mathbf{y}+2}{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}+1}+1}\\ =&1.\end{aligned} \tag{10}$$

The discrepancy using Normalized EuDist has the same value when $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}$ and $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}$, which means no tolerance.

**AGD** and **Proof of Corollary 1**: According to Eq. (4), we have

$$\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}\widehat{D}(\mathbf{x},\mathbf{y})=\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}1-\frac{1}{1+\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}}=\frac{1}{2}, \tag{11}$$

$$\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}\widehat{D}(\mathbf{x},\mathbf{y})=\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}1-\frac{1}{1+\frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|}}=1. \tag{12}$$

The two equations denote that when $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0$, AGD has the tolerance value $\frac{1}{2}<\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to\infty}=1$, which means that $\|\mathbf{x}\|<<\|\mathbf{y}\|$ is acceptable as the half of perfect equal.

## D. Contour of AGD

We show more function contour comparisons with existing measurement methods in Fig. 1, where the axes are the angle $\angle\mathbf{x},\mathbf{y}$, the ratio $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}$ and the metric contour value $z$ for better visualization. As we can see, the CosDist (Fig. 1(a)) has no relation to the ratio. The tolerance for $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}$ of EuDist depends on the norm of $\mathbf{y}$ (Fig. 1(b)). The proposed AGD has fixed tolerance for $\lim_{\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}\to 0}$ as shown in Fig. 1(c).

## E. Introduction of MGDA

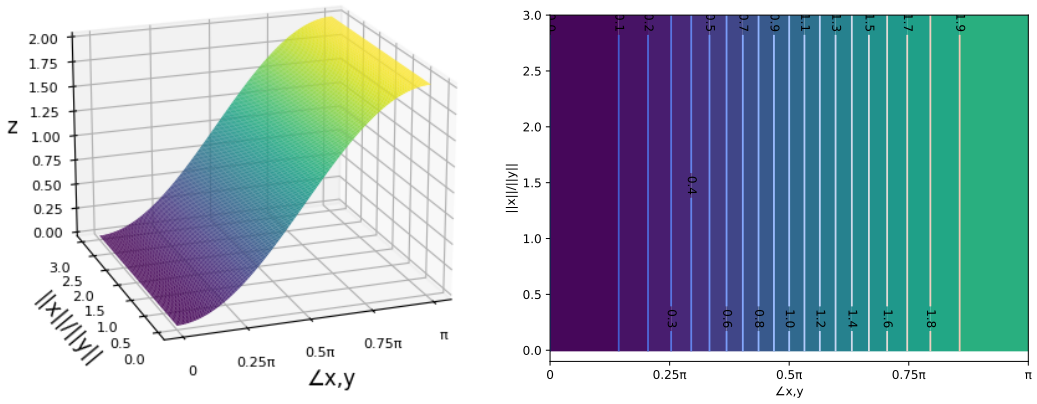
At any time, PCL training yields the following dynamic multi-objective empirical risk minimization formulation:

$$\min_{\boldsymbol{\theta},\{\boldsymbol{\theta}_i|i\in\mathcal{T}\}}\quad\{\ell_i\left(\mathcal{D}_i\right)\mid\forall i\in\mathcal{T}_t\}, \tag{13}$$

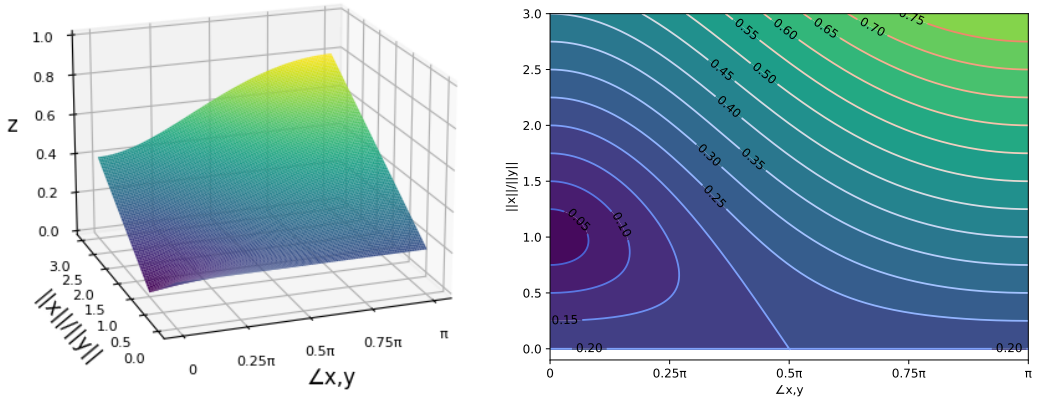
where $\mathcal{T}$ is the task index set with activated data streams at time $t$.

An elegant solution to the MOO for Pareto optimality [1] is the Steepest Descent Method (SDM) [3], which aims to obtain an optimal descent direction $d^*$ that satisfies

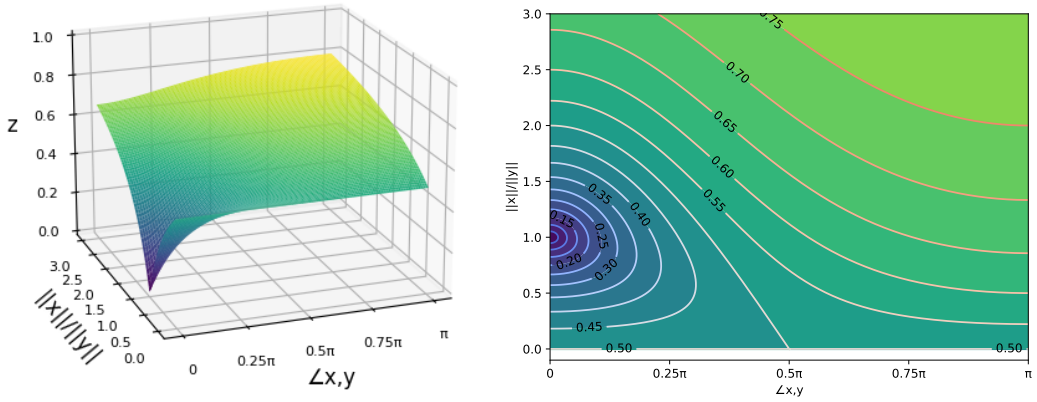
$$\begin{aligned}\mathbf{d}^*,\alpha^*=\arg\min_{\mathbf{d},\alpha}&\quad\alpha+\frac{1}{2}\|d\|^2,\\ \text{s.t.}&\quad\mathbf{g}_i^\top\mathbf{d}\leq\alpha,\quad\forall i\in\mathcal{T},\end{aligned} \tag{14}$$




(a) $z = 1 - \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$




(b) $z = \|\mathbf{x} - \mathbf{y}\|,\ \|\mathbf{y}\| = 0.2$




(c) $z = \frac{\|\mathbf{x}-\mathbf{y}\|}{\|\mathbf{y}\|+\|\mathbf{x}-\mathbf{y}\|}$

Figure 1. Contours of different measures. Note that the $x$- and $y$-axes are the angle (i.e., $\angle\mathbf{x}, \mathbf{y}$) between $\mathbf{x}$ and $\mathbf{y}$, and the magnitude ratio $\frac{\|\mathbf{x}\|}{\|\mathbf{y}\|}$, respectively. (a) Cosine distance; (b) Euclidean distance where $\|\mathbf{y}\| = 0.2$; (c) Asymmetric gradient distance.

where the constraints let each task have non-conflict with gradient $d$. Considering the Lagrange multipliers and Karush–Kuhn–Tucker (KKT) condition, the dual problem solved by the Multi-Gradient Descent Algorithm (MGDA) [2] is

$$\begin{aligned} \mathbf{w}^* = \arg\min_{\mathbf{w}} \quad & \left\| \sum_i \mathbf{w}_i \mathbf{g}_i \right\|^2, \\ \text{s.t.} \quad & \sum_i \mathbf{w}_i = 1 \ \text{ and } \ \mathbf{w}_i \geq 0, \forall i. \end{aligned} \tag{15}$$

The objective of MGDA is $0$ and the resulting point satisfies the KKT conditions, or the solution gives a Pareto descent direction that improves all tasks.

## References


[1] James M Buchanan. The relevance of pareto optimality. *Journal of conflict resolution*, 1962. 2

[2] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 2012. 3

[3] Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research*, 2000. 2