

---

# Appendix for “Continual Optimization with Symmetry Teleportation for Multi-Task Learning”

---

Anonymous Author(s)

Affiliation

Address

email

1	<b>Contents</b>	
2	<b>A Additional Results</b>	<b>2</b>
3	A.1 Conflict and Gradient Examinations . . . . .	2
4	A.2 Plug-and-Play Verification . . . . .	2
5	A.3 Motivation . . . . .	3
6	A.4 Analysis on Trigger Condition . . . . .	3
7	A.5 Analysis on Alternative of PEFT . . . . .	4
8	A.6 Time Cost . . . . .	4
9	<b>B Implementation Details</b>	<b>5</b>
10	B.1 Baseline Implementation . . . . .	5
11	B.2 PEFT Implementation . . . . .	5
12	<b>C Algorithm</b>	<b>6</b>
13	<b>D Discussion &amp; Limitation</b>	<b>6</b>
14	<b>E Pareto Concept</b>	<b>6</b>
15	<b>F Convergence Analysis</b>	<b>6</b>

## 16 A Additional Results

### 17 A.1 Conflict and Gradient Examinations

18 Although our method achieves competitive performance, it remains unclear whether it effectively  
 19 resolves the targeted issues, i.e., conflict mitigation and greater gradient norm discovery. To investigate  
 20 this, we analyze the training process by recording the results before and after teleportation, as shown  
 21 in Figure 1. The findings indicate that conflict is significantly alleviated, with task gradients becoming  
 22 positively correlated in most cases after teleportation. Besides, teleportation consistently yields  
 greater gradient norms, confirming the effectiveness of COST’s design.

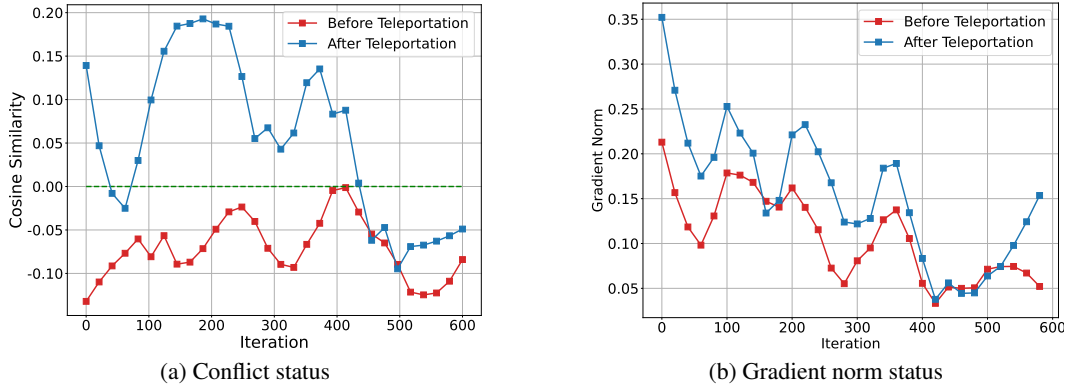


Figure 1: Examinations before/after teleportation.

23

### 24 A.2 Plug-and-Play Verification

25 Intuitively, our method is orthogonal to existing MTL approaches and is therefore plug-and-play,  
 26 enabling augmentation when integrated. Here, we take three baselines (i.e., CAGrad, Nash-MTL, and  
 27 FairGrad) to demonstrate the effectiveness of COST, and present the results in Table 1. As anticipated,  
 28 our method successfully brings considerable augmentation to its baselines, with improvements ranging  
 29 from 0.88 to 3.21 according to  $\Delta m\%$ . Specifically, CAGrad and FairGrad receive improvements on  
 30 almost each individual metric.

31 We also conduct an additional verification experiment using CAGrad on the NYUv2 dataset. The  
 32 results, presented in Table 2, show that COST consistently enhances CAGrad’s performance across all  
 metrics.

Table 1: Plug-and-play verification on *CityScapes* (2 tasks) dataset. We adopt FAMO’s implemen-  
 tation for Nash-MTL (denoted as Nash-R) and augment it with COST, since Nash-MTL does not  
 provide the official implementation on *CityScapes*.

Method	Segmentation $\uparrow$		Depth $\downarrow$		$\Delta m\% \downarrow$
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.	
CAGrad	75.16	93.48	0.0141	37.60	11.58
CAGrad + COST	75.46	93.57	0.0134	35.68	8.37
Nash-R	75.87	93.57	0.0135	37.29	9.89
Nash-R + COST	75.70	93.56	0.0134	34.34	7.15
FairGrad	75.72	93.68	0.0134	32.25	5.18
FairGrad + COST	75.73	93.53	0.0133	31.53	4.30

33

Table 2: Plug-and-play verification on *NYUv2* (3 tasks).

Method	Segmentation		Depth		Surface Normal					$\Delta m\%$ ↓
	(Higher Better)		(Lower Better)		Angle Distance		Within $t^\circ$			
					(Lower Better)		(Higher Better)			
	mIoU	Pix. Acc.	Abs Err	Rel Err	Mean	Median	11.25	22.5	30	
CAGrad	39.79	65.49	0.55	0.23	26.31	21.58	25.61	52.36	65.58	0.29
CAGrad + COST	40.76	66.42	0.53	0.22	26.00	21.13	26.44	53.25	66.30	-1.61
FairGrad	39.74	66.01	0.54	0.22	24.84	19.60	29.26	56.58	69.16	-4.66
FairGrad + COST	38.06	64.71	0.54	0.23	24.47	18.80	30.84	58.25	0.30	-5.39

### 34 A.3 Motivation

35 To further elucidate our motivation, we additionally carry out a verification experiment to observe  
 36 the dominant conflict status within Nash-MTL. Based on the results shown in Figure 2, Nash-MTL  
 37 exhibits fewer dominated conflicts in comparison to CAGrad and FairGrad, yet still encounters a  
 38 significant number. Moreover, it also possesses symmetry points that have the same loss level but  
 with different conflict status.

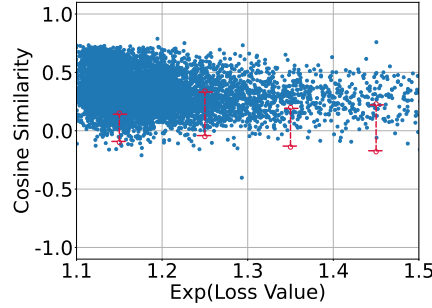


Figure 2: Dominated conflict vs. loss examination of Nash-MTL.

39

### 40 A.4 Analysis on Trigger Condition

41 As stated in the main text, the trigger condition is of importance for our symmetry teleportation  
 42 process. To illustrate this, we carry out a comparison between our system when triggered by  
 43 dominated conflict and when triggered by weak conflict, and the results are presented in Table 3. It  
 44 can be observed that although COST-weak surpasses COST-dominated in terms of most individual metrics,  
 45 it accomplishes this by sacrificing a significant portion of the performance on the Rel. Err. metric.  
 46 This is an unexpected outcome in the context of MTL. Consequently, it does not perform satisfactorily  
 47 on the overall metric ( $\Delta m\%$ ). These results further emphasize that addressing both imbalance and  
 48 conflict issues is crucial for MTL. This is because the scenario of weak conflict only focuses on the  
 conflict issue, overlooking the importance of handling imbalance as well.

Table 3: Trigger condition comparison on *CityScapes* (2 tasks) dataset.

Method	Segmentation		Depth		$\Delta m\%$ ↓
	(Higher Better)		(Lower Better)		
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.	
COST-dominated	75.73	93.53	0.0133	31.53	4.30
COST-weak	75.92	93.64	0.0127	35.94	6.94

49

50 We also conduct an additional experiment on the trigger condition in the presence of numerous  
 51 tasks, following Eqn. 4 in the main text. As previously stated, this condition is designed to balance

effectiveness and efficiency—relaxing it would improve performance but at the cost of increased inefficiency. The results, presented in Figure 3, confirm this trade-off. As shown,  $\Delta m\%$  gradually decreases as the trigger condition is relaxed, with improvements driven by more frequent teleportation. However, this comes at the cost of an almost linear increase in time complexity.

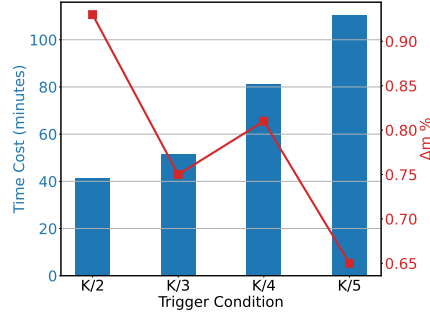


Figure 3: Trigger condition under the massive tasks scenario. Time cost is measured for one epoch on CelebA.

Table 4: PEFT alternatives comparison on *CityScapes* (2 tasks) dataset.

Method	Segmentation		Depth		$\Delta m\%$ ↓
	(Higher Better)		(Lower Better)		
	mIoU	Pix. Acc.	Abs. Err.	Rel. Err.	
FairGrad	75.72	93.68	0.0134	32.25	5.18
COST-LoRA	75.73	93.53	0.0133	31.53	4.30
COST-LoHa Hyeon-Woo et al. [2021]	75.52	93.43	0.0130	35.07	7.10
COST-OFT Qiu et al. [2023]	68.17	91.40	0.0151	45.25	23.39

## A.5 Analysis on Alternative of PEFT

Currently, numerous PEFT alternatives are available for teleportation purposes. To further examine the effect of PEFT on our method, we employ additional PEFT options to assess their impact on MTL performance. Specifically, we evaluate a LoRA variant (LoHa) and another PEFT alternative, OFT. The results, presented in Table 4, show that neither PEFT option improves upon their baselines. This suggests that while advanced PEFT methods may enable more efficient tuning, their complex designs can limit generalizability across various scenarios, aligning with some recent observations Pu et al. [2023]. Identifying a suitable PEFT approach remains a future direction for our framework.

The obtained results demonstrate that our framework can indeed benefit from certain other PEFT alternatives, e.g., LoHa. However, it also encounters setbacks when using alternatives like OFT. This implies that the selection of the PEFT method warrants further investigation.

## A.6 Time Cost

Applying teleportation at every instance of a conflict would significantly increase the computational burden and training time. To mitigate this, we introduce two strategies: delayed start and frequency control. The delayed start strategy postpones the application of teleportation until after  $E$  epochs. Meanwhile, frequency control limits the number of teleportation operations within each epoch, reducing overhead without much compromising the optimization process.

Here, we measure the running time of a single epoch on CelebA, comparing the scenarios with and without the COST augmentation, and present the results in Figure 4. As can be observed, our applied strategies introduce only an additional 30% of the training time compared to its baselines. Nonetheless, we acknowledge this still constitutes one of our limitations.

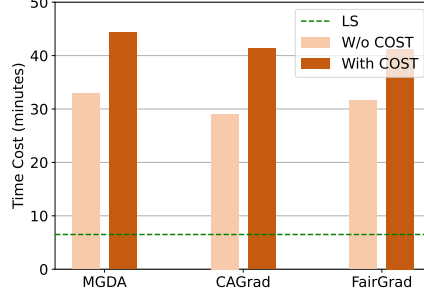


Figure 4: Time cost comparison.

## B Implementation Details

### B.1 Baseline Implementation

**CAGrad:** CAGrad strikes a balance between Pareto optimality and globe convergence by regulating the combined gradients in proximity to the average gradient:

$$\max_{d \in \mathbb{R}^m} \min_{\omega \in \mathcal{W}} g_{\omega}^{\top} d \quad \text{s.t.} \quad \|d - g_0\| \leq c \|g_0\| \quad (1)$$

In our experiments, we extend the official implementation to include COST. And all our training settings follow the same manner with its original one. For more comprehensive information, please consult the official implementation.

**Nash-MTL:** Nash-MTL provides the individual progress guarantee via the following objective:

$$\min_{\omega} \sum_i \beta_i(\omega) + \varphi_i(\omega) \quad (2)$$

$$\text{s.t.} \quad \forall i, -\varphi_i(\omega) \leq 0, \quad \omega_i > 0. \quad (3)$$

where  $\varphi_i(\omega) = \log(\omega_i) + \log(g_i^{\top} G \omega)$ ,  $G = [g_1, g_2, \dots, g_K]$ . As demonstrated, the individual progress is ensured through the projection, subject to the constraint  $\beta_i = g_i^{\top} G \omega \geq \frac{1}{\omega_i}$ .

We utilize the FAMO’s implementation of Nash-MTL, since Nash-MTL’s official code does not provide the implementation on CityScapes (2 tasks). In the Table 4 of the main text, we honestly report the results of this implementation, and show improvements brought by COST.

**FairGrad:** FairGrad is a pioneer MTL algorithm that proposes fairness measurements to promote maximal loss decrease, and formulate the following objective to derive the combination of individual gradients:

$$G^{\top} G \omega = \omega^{-1/\alpha} \quad (4)$$

where  $\alpha$  is the hyper-parameter, which is set to 1. We regard FairGrad as the advanced version of Nash-MTL that is able to balance task progresses in a finer grained. We adopt its official implementation for all our implementation through the paper.

### B.2 PEFT Implementation

**LoHa:** As a variant of LoRA, it approximates large weight matrices with low-rank ones through the Hadamard product. This approach has the potential to be more parameter-efficient than LoRA itself.

**OFT:** OFT draws inspiration from continual learning. It operates by re-parameterizing the pre-trained weight matrices using its orthogonal matrix, thereby preserving the information within the pre-trained model. To decrease the number of parameters, OFT incorporates a block-diagonal structure into the orthogonal matrix.

We utilize the implementations of LoRA, LoHa, and OFT provided by Hugging Face’s PEFT. **The rank for each of these is set to 5, while other configurations are left at their default values.** We apply LoRA to the backbone network across three datasets (CityScapes, NYUv2, and CelebA). For the QM9 benchmark, since it employs a graph network which is currently not supported by PEFT, we only apply LoRA to input and output linear layers instead.

---

**Algorithm 1** COST for MTL

---

Model parameters  $\theta^0$ Initialize Initialize  $\theta^0$  randomly $t = 1$  to  $T$  Compute task gradients  $G = [g_i]_{i=1}^K$ Dominated conflict detected Freeze  $\theta^t$  and train LoRA ( $\Delta\theta^t$ ) according Eqn.7, 9, 10, and 11 in the main text;Merge  $\theta^t$  and  $\Delta\theta^t$ :  $\theta^t = \theta^t + \Delta\theta^t$ , and unfreeze  $\theta^t$ 

Apply HTR on the optimizer according to Eqn.12, and 13 in the main text; Other MTL optimization

Have applied HTR Reset  $\sigma$  to 1;

---

**C Algorithm**

We conclude the learning paradigm of COST in Algorithm 1. It should be noted that since COST is a scalable framework, thus the other MTL optimization in Algorithm 1 could be mainstream MTL approaches (e.g., CAGrad, Nash-MTL, and FairGrad, etc).

**D Discussion & Limitation**

We offer a new framework to address the challenges of MTL, which is highly scalable and can be further improved by integrating more advanced components. For instance, LoRA could be substituted with an alternative PEFT method, and sharpness estimation can be substituted with some efficient gradient estimation methods Liu et al. [2024]. However, there are still some limitations. The current training paradigm requires additional training costs, and its performance on regression tasks is less competitive compared to the others. We have discussed some of these limitations in the Appendix, while others are left for future work.

**E Pareto Concept**

Formally, let us assume the weighted loss as  $\mathcal{L}_\omega = \sum_{i=1}^K \omega_i \mathcal{L}_i(\theta)$ , where  $\omega \in \mathcal{W}$  and  $\mathcal{W}$  represents the probability simplex on  $[K]$ . A point  $\theta'$  is said to Pareto dominate  $\theta$  if and only if  $\forall i, \mathcal{L}_i(\theta') \leq \mathcal{L}_i(\theta)$ . Consequently, the Pareto optimal situation arises when no  $\theta'$  can be found that satisfies  $\forall i, \mathcal{L}_i(\theta') \leq \mathcal{L}_i(\theta)$  for the given point  $\theta$ . All points that meet these conditions are referred to as Pareto sets, and their solutions are known as Pareto fronts. Another concept, known as Pareto stationary, requires  $\min_{\omega \in \mathcal{W}} \|g_\omega\| = 0$ , where  $g_\omega$  represents the weighted gradient  $\omega^\top G$ , and  $G$  is the gradients matrix whose each row is an individual gradient. We also provide the definition of gradient similarity for ease of description.

**F Convergence Analysis**

**Lemma 1.** Let  $\mathcal{L}(\theta, \xi)$  be a  $\Lambda$ -smooth function, where  $\xi$  is the i.i.d sampled mini-batch data. It follows that:

$$\mathbb{E} \left[ \|\nabla \mathcal{L}(\theta, \xi)\|^2 \right] \leq 2\Lambda (\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) + 2\Lambda \left( \mathcal{L}(\theta^*) - \mathbb{E} \left[ \inf_{\theta} \mathcal{L}(\theta, \xi) \right] \right)$$

*Proof.* We have the following inequality according to the  $\Lambda$ -smooth property of  $\mathcal{L}(\theta, \xi)$ :

$$\begin{aligned} \mathcal{L}(\theta', \xi) - \mathcal{L}(\theta, \xi) &\leq \\ \langle \nabla \mathcal{L}(\theta, \xi), \theta' - \theta \rangle + \frac{\Lambda}{2} \|\theta' - \theta\|^2, \forall \theta', \theta \in \mathbb{R}^d \end{aligned} \quad (5)$$

And  $\theta' = \theta - \frac{1}{\Lambda} \nabla \mathcal{L}(\theta, \xi)$ , thus we have:

$$\mathcal{L}(\theta - (1/\Lambda) \nabla \mathcal{L}(\theta, \xi), \xi) \leq \mathcal{L}(\theta, \xi) - \frac{1}{2\Lambda} \|\nabla \mathcal{L}(\theta, \xi)\|^2 \quad (6)$$

134 Assume  $\theta^* = \arg \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$ , then we can re-arranging the above inequality to have:

$$\begin{aligned}
& \mathcal{L}(\theta^*, \xi) - \mathcal{L}(\theta, \xi) = \\
& \mathcal{L}(\theta^*, \xi) - \inf_{\theta} \mathcal{L}(\theta, \xi) + \inf_{\theta} \mathcal{L}(\theta, \xi) - \mathcal{L}(\theta, \xi) \\
& \leq \mathcal{L}(\theta^*, \xi) - \inf_{\theta} \mathcal{L}(\theta, \xi) + \mathcal{L}(\theta - \frac{1}{\Lambda} \nabla \mathcal{L}(\theta, \xi), \xi) - \mathcal{L}(\theta, \xi) \\
& \leq \mathcal{L}(\theta^*, \xi) - \inf_{\theta} \mathcal{L}(\theta, \xi) - \frac{1}{2\Lambda} \|\nabla \mathcal{L}(\theta, \xi)\|^2
\end{aligned} \tag{7}$$

135 where the first inequality holds because  $\inf_{\theta} \mathcal{L}(\theta, \xi) \leq \mathcal{L}(\theta, \xi), \forall \theta$ . Taking expectation on above  
136 gives:

$$\begin{aligned}
& \mathbb{E} \left[ \|\nabla \mathcal{L}(\theta, \xi)\|^2 \right] \\
& \leq 2\mathbb{E} \left[ \Lambda \left( \mathcal{L}(\theta^*, \xi) - \inf_{\theta} \mathcal{L}(\theta, \xi) + \mathcal{L}(\theta, \xi) - \mathcal{L}(\theta^*, \xi) \right) \right] \\
& \leq 2\Lambda \mathbb{E} \left[ \mathcal{L}(\theta^*, \xi) - \inf_{\theta} \mathcal{L}(\theta, \xi) + \mathcal{L}(\theta, \xi) - \mathcal{L}(\theta^*, \xi) \right] \\
& \leq 2\Lambda (\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) + 2\Lambda \left( \mathcal{L}(\theta^*) - \mathbb{E} \left[ \inf_{\theta} \mathcal{L}(\theta, \xi) \right] \right)
\end{aligned}$$

137 □

138 **Theorem 2.** Assume task loss functions  $\mathcal{L}_1, \dots, \mathcal{L}_K$  are differentiable and  $\Lambda$ -smooth ( $\Lambda > 0$ ) such that  
139  $\|\nabla \mathcal{L}_i(\theta_1) - \nabla \mathcal{L}_i(\theta_2)\| \leq \Lambda \|\theta_1 - \theta_2\|$  for any two points  $\theta_1, \theta_2$ , and our symmetry teleportation  
140 property holds. Set the step size as  $\eta = \frac{1}{\Lambda\sqrt{T-1}}$ ,  $T$  is the training iteration. Then, there exists a  
141 subsequence  $\{\theta^{t_j}\}$  of the output sequence  $\{\theta^t\}$  that converges to a Pareto stationary point  $\theta^*$ .

142 *Proof.* We have the following inequality according to the  $\Lambda$ -smooth property of  $\mathcal{L}(\theta)$ :

$$\mathcal{L}(\theta') - \mathcal{L}(\theta) \leq \langle \nabla \mathcal{L}(\theta), \theta' - \theta \rangle + \frac{\Lambda}{2} \|\theta' - \theta\|^2 \tag{8}$$

143 Let  $\theta' = \theta^{t+1}$ ,  $\theta^{t'} = \theta^t + \Delta \theta^t$  (Gradient maximization:  $\Delta \theta^t = \arg \max \nabla \mathcal{L}(\theta^t + \Delta \theta^t)$ ), and  
144  $\mathcal{L}(\theta^t) = \mathcal{L}(\theta^{t'})$  (Loss invariance), we have:

$$\mathcal{L}(\theta^{t+1}) \leq \mathcal{L}(\theta^{t'}) + \langle \nabla \mathcal{L}(\theta^{t'}), \theta^{t+1} - \theta^{t'} \rangle \tag{9}$$

$$+ \frac{\Lambda}{2} \|\theta^{t+1} - \theta^{t'}\|^2 \tag{10}$$

$$= \mathcal{L}(\theta^t) - \eta_t \langle \nabla \mathcal{L}(\theta^{t'}), \nabla \mathcal{L}(\theta^{t'}, \xi^t) \rangle + \frac{\Lambda \eta_t^2}{2} \|\nabla \mathcal{L}(\theta^{t'}, \xi^t)\|^2 \tag{11}$$

145 Taking expectation conditioned on  $\theta^t$ , we have:

$$\begin{aligned}
\mathbb{E}_t [\mathcal{L}(\theta^{t+1})] & \leq \mathcal{L}(\theta^t) - \eta_t \|\nabla \mathcal{L}(\theta^{t'})\|^2 \\
& + \frac{\Lambda \eta_t^2}{2} \mathbb{E}_t \left[ \|\nabla \mathcal{L}(\theta^{t'}, \xi^t)\|^2 \right]
\end{aligned} \tag{12}$$

146 According to Lemma 1, we have:

$$\begin{aligned}
& \mathbb{E} \left[ \|\nabla \mathcal{L}(\theta, \xi)\|^2 \right] \leq \\
& 2\Lambda (\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) + 2\Lambda \left( \mathcal{L}(\theta^*) - \mathbb{E} \left[ \inf_{\theta} \mathcal{L}(\theta, \xi) \right] \right)
\end{aligned} \tag{13}$$

147 Inserting Eqn. 13 into Eqn. 12, we have:

$$\begin{aligned}
& \mathbb{E}_t [\mathcal{L}(\theta^{t+1})] \leq \mathcal{L}(\theta^t) - \eta_t \|\nabla \mathcal{L}(\theta^{t'})\|^2 \\
& + \Lambda^2 \eta_t^2 \left( \mathcal{L}(\theta^{t'}) - \mathcal{L}(\theta^*) + \mathcal{L}(\theta^*) - \mathbb{E} \left[ \inf_{\theta} \mathcal{L}(\theta, \xi) \right] \right)
\end{aligned} \tag{14}$$

148 By taking full expectation and re-arranging terms, we have:

$$\begin{aligned} \eta_t \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] &\leq (1 + \Lambda^2 \eta_t^2) \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^t) - \mathcal{L}^*] \\ &\quad - \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^{t+1}) - \mathcal{L}^*] + \Lambda^2 \eta_t^2 \sigma^2 \end{aligned} \quad (15)$$

149 where  $\sigma^2 = \mathcal{L}(\boldsymbol{\theta}^*) - \mathbb{E} [\inf_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \xi)]$ . Then we consider to introduce the re-weighting trick in Stich  
150 [2019]. Let  $\gamma_t$  ( $\gamma_t > 0$ ) be a sequence such that  $\gamma_t(1 + \Lambda^2 \eta_t^2) = \gamma_{t-1}$ . Assume  $\gamma_{-1} = 1$ , then  
151  $\gamma_t = 1 + \Lambda^2 \eta_t^{2-(t+1)}$ . By multiplying  $\gamma_t$  on both sides of Eqn. 15, we have:

$$\begin{aligned} \gamma_t \eta_t \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] &\leq \gamma_{t-1} \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^t) - \mathcal{L}^*] \\ &\quad - \gamma_t \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^{t+1}) - \mathcal{L}^*] + \gamma_t \Lambda^2 \eta_t^2 \sigma^2 \end{aligned} \quad (16)$$

152 Summing up the above equation from  $t = 0, \dots, T-1$ , we have:

$$\begin{aligned} \sum_{t=0}^{T-1} \gamma_t \eta_t \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] &\leq \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}^*] \\ &\quad + \Lambda^2 \sigma^2 \sum_{t=0}^{T-1} \gamma_t \eta_t^2 \end{aligned} \quad (17)$$

153 Dividing both sides by  $\sum_{t=0}^{T-1} \gamma_t \eta_t^2$ , we have:

$$\begin{aligned} \min_{t=0, \dots, T-1} \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] &\leq \frac{1}{\sum_{t=0}^{T-1} \gamma_t \eta_t} \sum_{t=0}^{T-1} \gamma_t \eta_t \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \\ &\leq \frac{\mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}^*] + \Lambda^2 \sigma^2 \sum_{t=0}^{T-1} \gamma_t \eta_t^2}{\sum_{t=0}^{T-1} \gamma_t \eta_t} \end{aligned} \quad (18)$$

154 Assume  $\eta_t \equiv \eta$ , then we have:

$$\begin{aligned} \sum_{t=0}^{T-1} \gamma_t \eta_t &= \eta \sum_{t=0}^{T-1} (1 + \Lambda^2 \eta^2)^{-(t+1)} \\ &= \frac{\gamma}{1 + \Lambda^2 \eta^2} \frac{1 - (1 + \Lambda^2 \eta^2)^{-T}}{1 - (1 + \Lambda^2 \eta^2)^{-1}} \\ &= \frac{1 - (1 + \Lambda^2 \eta^2)^{-T}}{\Lambda^2 \eta} \end{aligned} \quad (19)$$

155 Note that  $(1 + \Lambda^2 \eta^2)^{-T} \leq \frac{1}{2}$  and  $\frac{x}{1+x} \leq \log(1+x)$ , thus we have

$$\frac{\log(2)}{\log(1 + \Lambda^2 \eta^2)} \leq \frac{\log(2)(1 + \Lambda^2 \eta^2)}{\Lambda^2 \eta^2} \leq T \quad (20)$$

156 From this, we can obtain:

$$\sum_{t=0}^{T-1} \gamma_t \eta_t \geq \frac{1}{2\Lambda^2 \eta}, \text{ for } T \geq \frac{\log(2)(1 + \Lambda^2 \eta^2)}{\Lambda^2 \eta^2} \quad (21)$$

157 Inserting the above equation into the Eqn. 18, we have:

$$\min_{t=0, \dots, T-1} \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] \quad (22)$$

$$\leq 2\Lambda^2 \eta \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}^*] + \eta \Lambda^2 \sigma^2, \text{ for } T \geq \frac{\log(2)(1 + \Lambda^2 \eta^2)}{\Lambda^2 \eta^2} \quad (23)$$



158 Setting  $\eta = \frac{1}{\Lambda\sqrt{T-1}}$ , we finally have:

$$\min_{t=0,\dots,T-1} \mathbb{E} \left[ \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^{t'}) \right\|^2 \right] \quad (24)$$

$$\leq \frac{2\Lambda}{\sqrt{T-1}} \mathbb{E} [\mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}^*] + \frac{\Lambda\sigma^2}{\sqrt{T-1}} \quad (25)$$

159 Thus, our method can readily reach to the Pareto Stationary point.  $\square$

## 160 References

- 161 Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for  
162 communication-efficient federated learning. *arXiv preprint arXiv:2108.06098*, 2021.
- 163 Xinyue Liu, Hualin Zhang, Bin Gu, and Hong Chen. General stability analysis for zeroth-order  
164 optimization algorithms. In *The Twelfth International Conference on Learning Representations*,  
165 2024.
- 166 George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. Empirical analysis of the strengths and  
167 weaknesses of peft techniques for llms. *arXiv preprint arXiv:2304.14999*, 2023.
- 168 Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller,  
169 and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances*  
170 *in Neural Information Processing Systems*, 36:79320–79362, 2023.
- 171 Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint*  
172 *arXiv:1907.04232*, 2019.