Appendix

A Theoretical Analysis 12
A.1. Proof of Proposition 1 12
A.2. Proposition 1 with Relaxed Assumption . . . 12

B Dataset Description 12

C Baseline Algorithms 13

D Additional Ablations and Analyses 13
D.1. Multi-Scale Multi-Grid Function 13
D.2. Learnable Multi-Grid Function 13
D.3. Impact of Model and Dataset Size 14
D.4. Impact of the Number of Snapshots 14
D.5. Compatibility with Image Compression . . . 14
D.6. Scalability under Extreme Heterogeneity . . 14

A. Theoretical Analysis

A.1. Proof of Proposition 1

Definition 1. *A function $D : \mathcal{D} \times \mathcal{D} \to [0, \infty)$ is a gradient L2 distance measure if, for any two datasets $\forall X, X' \in \mathcal{D}$ it is defined as the Euclidean (L2) distance between their corresponding loss gradients:*

$$D(X, X') = \left\| \nabla \mathcal{L}(X, \mathbf{w}) - \nabla \mathcal{L}(X', \mathbf{w}) \right\|_2$$

where $\| \cdot \|_2$ denotes the L2 norm. This measure satisfies the following properties inherited from the L2 norm:

1. $D(X, X) = 0$ *and* $D(X, X') = D(X', X)$.
2. $\forall d \in \mathbb{R}^m$ *s.t.* d *is closer to* X' *than* d_i, $D(X \setminus \{d_i\} \cup \{d\}, X') \leq D(X, X')$.
3. $D(X, X' \cup \{d_i\}) \leq D(X, X')$.

Proposition 1. *If $\mathcal{N}^{n'} \subseteq \mathcal{M}$, then for the above-defined gradient L2 distance measure D,*

$$\min_{\boldsymbol{S} \in \mathbb{R}^{1 \times m}} D(\mathcal{F}(\boldsymbol{S}), \mathcal{T}) \leq \min_{\boldsymbol{S} \in \mathbb{R}^{1 \times m}} D(\boldsymbol{S}, \mathcal{T}).$$

Proof. For simplicity, we denote $[1, \ldots, n]$ as $[n]$. Let us denote $\mathcal{T} = \{t_i\}_{i=1}^{n_t}$ and $\boldsymbol{S} = \{s_j\}_{j=1}^{n}$, where $t_i \in \mathcal{N} \subset \mathbb{R}^m$ and $s_j \in \mathbb{R}^m$, $\forall i \in [n_t]$ and $\forall j \in [n]$. Under the assumption that $\mathcal{N}$ is a subspace of $\mathbb{R}^m$, there exists the projection of s_j onto $\mathcal{N}$, $\bar{s}_j \in \mathcal{N}$. Because $t_i \in \mathcal{N}$ for $i = 1, \ldots, n_t$, $\|\bar{s}_j - t_i\| \leq \|s_j - t_i\|, \forall j \in [n]$ and $\forall i \in [n_t]$. This means the projection $\bar{s}_j$ is closer to $\mathcal{T}$ than s_j, $\forall j \in [n]$. Let us define a partially projected dataset $\bar{\boldsymbol{S}}_k = \{\bar{s}_j\}_{j=1}^{k} \cup \{s_j\}_{j=k+1}^{n}$. Then by the second axiom of Definition 1,

$$D(\bar{\boldsymbol{S}}_n, \mathcal{T}) \leq D(\bar{\boldsymbol{S}}_{n-1}, \mathcal{T}) \leq \ldots \leq D(\boldsymbol{S}, \mathcal{T}).$$

This result means that the optimum $\boldsymbol{S}^* = \arg\min D(\boldsymbol{S}, \mathcal{T})$ satisfies $\boldsymbol{S}^* \in \mathcal{N}^n$. Note our multi-grid function $\mathcal{F}$ augments the number of data from n to n' where $n = 1$. Let us denote $k' = n' - 1$ and $S^*_{add} = \boldsymbol{S}^* \cup \{t_i\}_{i=1}^{k'}$. By the third axiom of Definition 1,

$$D(S^*_{add}, \mathcal{T}) \leq D(\boldsymbol{S}^*, \mathcal{T}).$$

The elements of S^*_{add} lie in $\mathcal{N}$ and $S^*_{add} \in \mathcal{N}^{n'}$. From the assumption $\mathcal{N}^{n'} \subseteq \mathcal{M}$, $\exists \boldsymbol{S} \in \mathbb{R}^{1 \times m}$ *s.t.* $\mathcal{F}(\boldsymbol{S}) = S^*_{add}$. Thus,

$$\begin{aligned} \min_{\boldsymbol{S} \in \mathbb{R}^{1 \times m}} D(\mathcal{F}(\boldsymbol{S}), \mathcal{T}) &\leq D(S^*_{add}, \mathcal{T}) \\ &\leq D(\boldsymbol{S}^*, \mathcal{T}) = \min_{\boldsymbol{S} \in \mathbb{R}^{1 \times m}} D(\boldsymbol{S}, \mathcal{T}). \end{aligned}$$

A.2. Proposition 1 with Relaxed Assumption

In Proposition 1, we assume $\mathcal{N}^{n'} \subseteq \mathcal{M}$) that the synthetic samples space by f is sufficiently large to contain all data points in $\mathcal{N}$. Relaxing the assumption, we consider when $\mathcal{M}$ approximately covers $\mathcal{N}^{n'}$. With the following notion of ϵ-cover, we describe the trade-off between the effects from the increase in the number of data and the decrease in representability of the synthetic samples.

Definition 2. *Given a gradient distance measure D, $\mathcal{M}$ is a ϵ-cover of $\mathcal{N}^{n'}$ on D if $\forall X' \in \mathcal{N}^{n'}$, $\exists \boldsymbol{S} \in \mathbb{R}^{n \times m}$ s.t. $D(\mathcal{F}(\boldsymbol{S}), X') \leq \epsilon$.*

Here, we assume a gradient distance measure D satisfies the triangular inequality. From the proof above in Proposition 1, $\exists S^*_{add} \in \mathcal{N}^{n'}$ *s.t.* $D(S^*_{add}, \mathcal{T}) \leq \min_{\boldsymbol{S} \in \mathbb{R}^{n \times m}} D(\boldsymbol{S}, \mathcal{T})$. Let us denote the gain $G = \min_{\boldsymbol{S}} D(\boldsymbol{S}, \mathcal{T}) - D(S^*_{add}, \mathcal{T})$. If $\mathcal{M}$ is a ϵ-cover of $\mathcal{N}^{n'}$ on D, then $\exists \boldsymbol{S} \in \mathbb{R}^{n \times m}$ *s.t.*

$$\begin{aligned} D(\mathcal{F}(\boldsymbol{S}), \mathcal{T}) &\leq D(S^*_{add}, \mathcal{T}) + D(\mathcal{F}(\boldsymbol{S}), S^*_{add}) \\ &\leq D(S^*_{add}, \mathcal{T}) + \epsilon. \end{aligned}$$

Note, we use the triangular inequality in the first inequality above and use the definition of ϵ-cover in the second inequality. We can conclude that

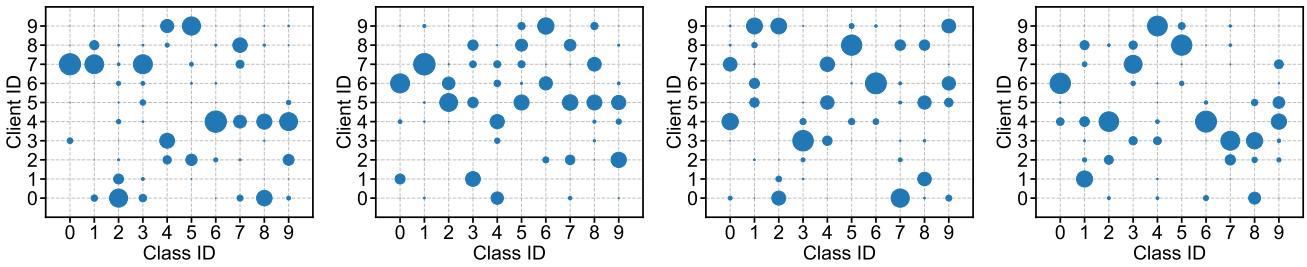
$$\begin{aligned} \min_{\boldsymbol{S} \in \mathbb{R}^{n \times m}} D(\mathcal{F}(\boldsymbol{S}), \mathcal{T}) &\leq D(S^*_{add}, \mathcal{T}) + \epsilon \\ &= \min_{\boldsymbol{S} \in \mathbb{R}^{n \times m}} D(\boldsymbol{S}, \mathcal{T}) - G + \epsilon. \end{aligned}$$

To summarize, the optimization with multi-grid function $\mathcal{F}$ can generate a synthetic gradient that has at least $G - \epsilon$ smaller distance to the target gradient compared to when not using $\mathcal{F}$. We can interpret G as a possible gain by the increase in the number of data, *i.e.*, from n to n', and ϵ as the representability loss in parameterization by $\mathcal{F}$.

B. Dataset Description

In our experiments, we evaluate OS-FED on a diverse range of benchmarks, including large-scale image classification and text classification tasks.

ImageNet-10. Following common practice in validating algorithms on large-scale data with reasonable computation [42], we use a subset of the ImageNet dataset. We adopt



(a) MNIST (b) FMNIST (c) CIFAR-10 (d) ImageNet-10

Figure 8. **Data distribution for 10 clients on (a) MNIST, (b) FMNIST, (c) CIFAR-10, and (d) ImageNet-10, generated using a Dirichlet distribution (Dir($\alpha = 0.5$)).** The size of the dot corresponds to the number of samples for a given class on a given client.

a subclass list from [42] and select the first 10 classes for our ImageNet-10 experiments. All images are preprocessed to a fixed size of 224x224 using resize and center crop functions. To simulate the non-IID data distributions inherent in federated learning, we partition the dataset among 10 clients using a Dirichlet distribution (Dir($\alpha = 0.5$)). The resulting data distribution for each client is visualized in Figure 8.

AGNews. The AGNews [56] dataset is a widely-used standard benchmark for English text classification, consisting of news articles clearly categorized into four classes: World, Sports, Business, and Sci/Tech. It contains 120,000 training samples and 7,600 testing samples.

SogouNews. The SogouNews [56] dataset is another text classification benchmark, containing Chinese news articles. It is larger and more complex than AGNews, comprising 450,000 training samples and 60,000 testing samples, distributed across five categories: Sports, Finance, Entertainment, Automobile, and Technology.

C. Baseline Algorithms

We compare OS-FED against several competitive baselines that facilitate efficient federated learning, including FedAvg [31], FedMPQ [8], RandTopk [59], LoRA-FAIR [6], FedSD2C [55] and FedAF [48].

- **FedAvg:** this is the standard federated learning algorithm that serves as our performance and communication upper bound. Clients transmit the full, uncompressed model gradients or parameters to the server for aggregation.
- **FedMPQ:** a mixed-precision quantization method where different layers of the model are assigned varying bit-widths to compress the gradient updates.
- **RandTopk:** a sparsification technique where clients transmit the top-k gradients by magnitude. It introduces randomness by allowing non-top-k values to be selected with a small probability.
- **LORA-FAIR:** an approach that applies Low-Rank Adaptation (LoRA) to federated learning. Clients only train and transmit the parameters of small, low-rank adapter modules, significantly reducing communication.
- **FedSD2C:** a sophisticated multi-stage framework that first selects an informative core-set from local data, applies privacy-enhancing Fourier perturbations, and then uses a server-provided autoencoder to distill it into a highly compact synthetic distillate for transmission.
- **FedAF:** a dataset distillation method that uses distribution matching to condense local data. It notably incorporates a regularization term based on Sliced Wasserstein Distance to better align the local knowledge distribution with that of other clients.

D. Additional Ablations and Analyses

D.1. Multi-Scale Multi-Grid Function

The standard uniform multi-grid function typically divides the snapshot into a grid of same-sized cells, each requiring the same upsampling factor. In this section, we explore a multi-scale variant where cells can have heterogeneous resolutions. For instance, a 2x2 grid can have some cells further subdivided, creating a mix of cells that require different upsampling factors (e.g., 4x and 8x) to be restored to the target size. As shown in Table 8, under a tight budget (1 snapshot per client), the multi-scale function is more effective as it can allocate its information capacity more flexibly. When the budget is ample (8 snapshots), the simpler uniform function is already sufficient.

D.2. Learnable Multi-Grid Function

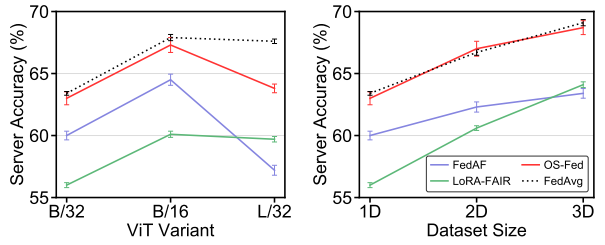
We further study the potential of exploiting a learnable multi-grid function, aiming to synthesize more diverse and representative snapshots, albeit at the cost of additional computation and storage. In this experiment, we replace the fixed bilinear upsampling with a learnable function using a lightweight Fast Super-Resolution Convolutional Neural Network (FSRCNN). Table 9 summarizes the resulting performance. While the extra learnable module offers little benefit at a lower complexity (Factor 4), its advantage becomes much more significant at a higher complexity (Factor 8). We conjecture that standard upsampling is sufficient for lower factors but suffers from a lack of representational

Table 8. **Performance comparison of the uniform and multi-scale multi-grid functions on ImageNet-10.** #Shape denotes the size of snapshots sent by each client per round. The multi-scale function outperforms the uniform one under tight budgets.

#Shape	Test Model	Uniform (default)	Multi-Scale
1×224×244 (0.1%)	ResNet-18	63.5	**65.2**
	RegNet-X	68.0	**69.3**
8×224×224 (0.8%)	ResNet-18	**64.6**	63.9
	RegNet-X	**68.8**	68.1

Table 9. **Performance comparison of a standard bilinear upsampling versus a learnable FSRCNN for the multi-grid function on ImageNet-10.** Using a learnable function consistently improves performance, especially for higher complexity factors.

Test Model	Factor 4		Factor 8	
	Upsample	FSRCNN	Upsample	FSRCNN
ResNet-18	63.5	64.0	64.7	**66.9**
RegNet-X	68.0	68.1	68.8	**70.7**



(a) Impact of Model Size (b) Impact of Dataset Size

Figure 9. **Impact of (a) model size and (b) dataset size on performance.** OS-FED consistently outperforms other baselines across all configurations. Here, (a) is evaluated on ImageNet-10. (b) is evaluated with the ViT-B/32 model.

power for higher factors. In such scenarios, the learnable function therefore shows promising results.

D.3. Impact of Model and Dataset Size

We evaluate the robustness of OS-FED under varying model and dataset scales to demonstrate its broad applicability. As shown in Figure 9 (a), OS-FED consistently maintains a significant performance advantage across different Vision Transformer (ViT) model sizes. It is noteworthy, however, that for extremely large models like ViT-L/32 (304M parameters), a single snapshot may struggle to fully capture the vast update information. Figure 9 (b) demonstrates that OS-FED 's superiority also holds as the dataset size is scaled up (1D, 2D, and 3D correspond to 1x, 2x, and 3x the original dataset size, respectively).

D.4. Impact of the Number of Snapshots

As identified in D.3, large models pose a challenge for single-snapshot compression. A natural and flexible solution is to simply increase the number of snapshots transmitted by each client per round. We investigate this strategy using the large ViT-L/32 model, with results shown in Figure 10. The plot clearly indicates that increasing the number of snapshots steadily improves server accuracy. This effectively mitigates the information bottleneck identified earlier and progressively closes the performance gap with the uncompressed FedAvg upper bound.Crucially, even when increasing the number of snapshots from one to eight, the total communication cost remains over 260x lower than the FedAvg baseline (approx. 45.6MB vs. 12160MB), showcasing a practical and still highly efficient path to scale OS-FED for extremely large models.

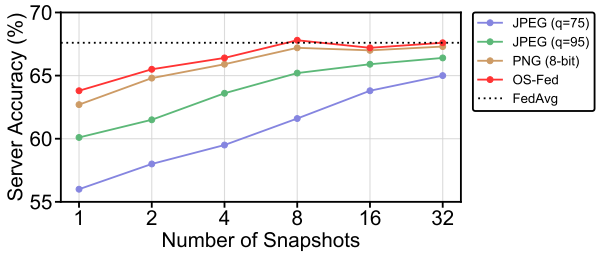


Figure 10. **Impact of the number of snapshots per client on server accuracy, evaluated on ImageNet-10 with ViT-L/32.** FedAvg represents the uncompressed performance upper bound.

Resolution	Top-1 Accuracy (FedAvg / **OS-FED**)			Overhead (per-client)	
	20 clients	40 clients	60 clients	Memory Cost	Wall-time
224^2	37.3% / 37.0%	36.5% / 36.6%	35.8% / 35.1%	4.55GB	3.6s
336^2	38.5% / 38.7%	38.1% / 37.7%	37.7% / 38.0%	8.8GB	5.3s
448^2	41.3% / 41.5%	40.6% / 40.6%	39.1% / 38.7%	14.9GB	7.5s

Table 10. **Scalability on ImageNet-1K in challenging non-IID (α=0.01).** Comparing OS-FED with FedAvg across different client scales and resolutions, while reporting overheads.

D.5. Compatibility with Image Compression

Our snapshot-based approach is also compatible with standard lossy image compression techniques. In Figure 10, we plot the performance when snapshots are compressed using JPEG and PNG. The results demonstrate a compelling trade-off: for a communication budget equivalent to one 32-bit snapshot (0.57MB), transmitting four PNG-compressed snapshots achieves higher server accuracy. This shows that multiple, structurally-rich but lossy snapshots can be more informative than a single, high-precision one, highlighting the synergistic potential of combining OS-FED with traditional compression.

D.6. Scalability under Extreme Heterogeneity

We evaluate OS-FED on the ImageNet-1K benchmark (1000 classes) with up to 60 clients. To simulate extreme statistical heterogeneity, the dataset is partitioned using a Dirichlet distribution with α=0.01. As shown in Table 10, we compare OS-FED against the uncompressed FedAvg baseline across varying client scales (20, 40, 60) and input resolutions (224^2, 336^2, 448^2). The results indicate that OS-FED maintains competitive accuracy under these scaled and highly non-IID conditions.