Flash: Concept Drift Adaptation in Federated Learning

Kunjal Panchal [1] **Sunav Choudhary** [2] **Subrata Mitra** [2] **Koyel Mukherjee** [2] **Somdeb Sarkhel** [3] **Saayan Mitra** [3] **Hui Guan** [1]

Abstract

In Federated Learning (FL), adaptive optimization is an effective approach to addressing the statistical heterogeneity issue but cannot adapt quickly to concept drifts. In this work, we propose a novel adaptive optimizer called FLASH that simultaneously addresses both statistical heterogeneity and the concept drift issues. The fundamental insight is that a concept drift can be detected based on the magnitude of parameter updates that are required to fit the global model to each participating client's local data distribution. FLASH uses a two-pronged approach that synergizes client-side early-stopping training to facilitate detection of concept drifts and the server-side drift-aware adaptive optimization to effectively adjust effective learning rate. We theoretically prove that FLASH matches the convergence rate of state-of-the-art adaptive optimizers and further empirically evaluate the efficacy of FLASH on a variety of FL benchmarks using different concept drift settings.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017) is a distributed machine learning paradigm where edge devices (called "clients") jointly train a machine learning (ML) model while keeping the training data on their devices. FL has cross-device and cross-silo settings (Wu et al., 2022). In cross-device FL, which is our primary focus, a server picks a small fraction of clients from the set of available clients to train an ML model (called "global model") on their local data at each round. The server then aggregates the trained models from the participating clients to update the global model. As the server does not have access and control over each client's data, FL is inherently privacy-preserving and has been applied in a number of industries such as healthcare (Loftus et al., 2022; du Terrail et al., 2022) and IoT (Dara et al., 2022; Li et al., 2022) where data privacy is of paramount importance.

Adaptive optimization is an effective approach to addressing the *statistical heterogeneity issue* in FL, where one client's data distribution could be different from another client. As participating clients are different from one round to another, statistical heterogeneity causes differences in data distributions across rounds, leading to convergence problems (Karimireddy et al., 2020). Adaptive optimizers such as FEDYOGI and FEDADAM (Reddi et al., 2021) use adaptive learning rates which incorporate knowledge of past rounds to perform more informed model updates. They successfully smooth the updates applied to the global model and improve the convergence in FL.

Problem. Existing adaptive optimization approaches, however, cannot adapt quickly to *concept drift*, another practical issue faced by deploying FL in real-world applications yet largely overlooked by the literature. A *concept drift* happens when the participating clients change their class conditional distributions due to phenomena like seasonality effects, geographic biases, diurnal patterns, and change in users' habits (Kairouz et al., 2021; Canonaco et al., 2021). For example, the use of the term "corona" would generate different outputs pre- and post-pandemic. Formally, the class conditional distribution before concept drift, $\mathcal{P}(y \mid x) = \sum_{c \in \mathcal{C}} q_c \mathcal{P}_c(y \mid x)$, is different from the distribution after concept drift, $\mathcal{P}'(y \mid x)$, where q_c is weight of c^{th} client, $\mathcal{C}$ is a set of available client for a particular round in FL, and (x, y) are data samples.

Adapting quickly to a concept drift requires a *large adaptive learning rate* to adjust global model parameters so that updated parameters can fit new conditional distribution $\mathcal{P}'(y \mid x)$. But existing adaptive optimizers often result in a relatively small effective learning rate towards an optimum despite a concept drift. Without loss of generality, existing adaptive optimizers update model parameters using $w_g^{(r)} = w_g^{(r-1)} + \eta_g \frac{m^{(r)}}{\sqrt{v^{(r)}}+\tau}$, where $w_g^{(r-1)}$ are global model parameters at round $r-1$, and $m^{(r)}$ and $v^{(r)}$ are the estimates of first and second moments of the gradients at round r. We refer to $\frac{\eta_g}{\sqrt{v^{(r)}}+\tau}$ as the *effective learning rate* of a parameter, which is the learning rate η_g scaled by the

[1]University of Massachusetts, Amherst, USA [2]Adobe Research, Bangalore, India [3]Adobe Research, San Jose, USA. Correspondence to: Kunjal Panchal <kpanchal@umass.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

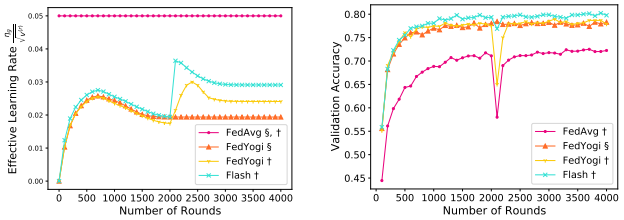


Figure 1. (Left) Effective learning rate and (Right) Accuracy across rounds on CIFAR10 dataset, where drift occurs at round 2000. † = with drift, § = without drift.

inverse of the second moments of the gradients. The second moments of the gradients would be relatively high when clients train the global model, which has captured $\mathcal{P}$, on the new distribution $\mathcal{P}'$, resulting in an effective learning rate at the onset of a concept drift which is not large enough to adapt to the new distribution quickly.

To illustrate the problem, Figure 1 shows the effective learning rate and the validation accuracy of the state-of-the-art adaptive optimizer FedYogi (Reddi et al., 2021). We calculate the learning rate by taking average of per-parameter $\eta_g/\sqrt{v^{(r)}}$ values. Before the drift, the effective learning rate of FEDYOGI is decreasing as the training approaches convergence. But after a drift at round 2000, the effective learning rate increases slowly for FEDYOGI. This slowly increased effective learning rate delays the drift adaptation, while also causing a large accuracy dip during the drift.

Proposed Approach. In this paper, we design a novel adaptive optimization algorithm called FLASH that can quickly adapt to concept drift while simultaneously addressing the statistical heterogeneity issue. The core intuition in FLASH is to use a larger effective learning rate at the onset of concept drifts while preserving the effectiveness of adaptivity to address statistical heterogeneity as shown in Figure 1.

Our fundamental insight is that a concept drift can be detected based on *how large client parameter updates are required to fit the received global model to each client's local data distribution* in a round. Specifically, when the global model needs larger updates to fit a client's data in the current round, it implies that the client's data is unlikely to be sampled from the global distribution captured by the global model. Furthermore, when larger updates are required by all participating clients in the current round, it indicates a concept drift and thus the needs of a larger effective learning rate to update the global model.

To this end, FLASH uses a two-pronged approach that synergizes client-side training to facilitate concept drift detection and server-side drift-aware adaptive optimization to dynamically adjust effective learning rate. On the client side, FLASH allows clients to train the global model until it reaches a steady state, where the trained model fits well to their local data, in each round. This results in a varying number of epochs per client. In contrast, existing adaptive optimizers typically have participating clients train the global model for a fixed number of epochs.

On the server side, the server orchestrates the update of the global model and thus can adapt effective learning rate based on *effective gradients*, which results from aggregating local parameter updates from participating clients. The principle of the adaptivity is to retain the similar effective learning rate as existing adaptive optimizers when no concept drift happens but significantly increase it at the onset of the concept drift. We achieve the adaptivity by tracking $||(\Delta^{(r)})^2 - v^{(r)}||$ at each round r–that is, the difference between the current squared gradients $(\Delta^{(r)})^2$ and the second moment of the gradients $v^{(r)}$ and found that this metric leads to a boost in the effective learning rate in case of a drift, while still getting a stabilized performance when no distribution change.

We evaluate the effectiveness of FLASH both theoretically and empirically. Theoretically, we prove that FLASH can match the convergence rate of FEDYOGI, a state-of-the-art adaptive optimizer, while it can also adapt quicker to a drift by decreasing the lower bound of the change in second order effective gradient.

Empirically, we compare FLASH against the best of the adaptive optimizers, personalization, drift correction, and drift detection methods using CIFAR10/100, and EMNIST datasets. In no concept drift settings, FLASH gives a comparable performance than the best performing baselines. In concept drift settings, FLASH has the highest accuracy at the onset of concept drift, needs the least number of FL rounds to recover accuracy, and achieves the highest recovered accuracy after concept drifts. The fast concept drift adaptation from FLASH saves 11.18% (CIFAR10), 10.42% (CIFAR100) , and 11.79% (EMNIST) of local epochs done by clients, as the global model adapts to the new distribution. In concept drift settings, FLASH only underperforms against ORACLE (an algorithm which has knowledge of when a drift occurs and access to both *pre-* and *post-* drift distributions) with the lowest accuracy dip difference of 1.48%-2.99%, while the best performing baselines exhibit 3.15%-9.22% more accuracy dip compared to the ORACLE.

We summarize the contributions of this work as follows:

- We propose a drift-aware adaptive optimization strategy that can quickly adapt to various concept drift patterns (sudden, incremental, and recurrent).
- Empirical evaluation demonstrates FLASH can achieve comparable performance as the most performing baselines in no concept drift settings while outperforms all baselines in various concept drift settings.
- We give theoretical analysis on convergence of FLASH, an epoch bound for early-stopping SGD for client-side

training, and draw a relation between the change in second order effective gradients with concept drifts.

2. Related Work

FLASH is related to techniques that address statistical heterogeneity and adaptive optimizations in particular, and techniques that address concept drift.

Techniques to Address Statistical Heterogeneity. *Adaptive optimization* to address the challenges of statistical heterogeneity has been studied in FEDYOGI/FEDADAM (Reddi et al., 2021). They are slow to adapt to concept drifts because of their relatively small effective learning rate when a synchronized concept drift happens. FLASH proposes a new adaptive update rule where the effective learning rate is higher at the onset of a drift for faster adaptation.

Personalization also addresses statistical heterogeneity HYPCLUSTER (Mansour et al., 2020), MAML (Jiang et al., 2019), PERFEDAVG (Fallah et al., 2020), APFL (Deng et al., 2020), DITTO (Li et al., 2021), FEDROD (Chen and Chao, 2022). It usually creates a separate personalized model, along with a local model. FLASH uses early-stopping training to get a well-trained local model for a client. And the same model, when sent to the server, is used to detect and adapt to (if any) concept drifts. We do not need to store any model states on client, as the model states can be rendered useless in case of drifts. *Drift correction* strategies address statistical heterogeneity by reducing gradient mismatch caused by solving two separate objectives (local and global). SCAFFOLD (Karimireddy et al., 2020) introduces a gradient correction term named "control variate" to help reduce the variance across client updates. FEDDYN (Acar et al., 2021) further reduces the communication cost related to drift-correction. These methods consider the concept drift as client-level noise, resulting in a local model update correction towards the pre-drift distribution. FLASH instead adapts the global model to the drift, resulting in less efforts from each client in local training and gradient correction.

Techniques to Address Concept Drift. Concept drifts in FL fall into two categories, distributed drift, and synchronized drift. Distributed drift, where clients can drift to multiple distributions in the same round, is explored in FEDDRIFT (Jothimurugesan et al., 2022). They propose a multi-model solution where each new distribution drift by any client spawns a new global model. Maintaining multiple global models to track all the past and current distributions poses a scalability challenge. Besides, their experiments do not consider heterogeneous data setting. FLASH focuses on synchronized drift, where all the clients face a drift towards one new distribution. This setting is explored in centralized learning in (Gama et al., 2014; Tahmasbi et al., 2020). FL poses an additional challenge of telling apart client-level heterogeneity from a concept drift. ADAPTIVEFEDAVG (Canonaco et al., 2021) attempts to addresses the synchronized drift issue through a preliminary study on adaptive optimizers, but it is incomplete and it falls short in improving the global accuracy dip *during* the drift. FLASH has a rigorous study with various patterns of synchronized drift, on various baselines based on adaptive optimization, drift correction, and drift adaptation. We also provide theoretical analysis for FLASH.

3. Methodology

This section describes our proposed adaptive optimization algorithm FLASH. Compared to existing adaptive optimizers, FLASH has two distinct features: (1) Clients train local model via *early-stopping training* to facilitate the concept drift detection and, at the same time, address statistical heterogeneity, and (2) the server updates the global model via *drift-aware adaptive optimization*. The synergy of the two features enables FLASH to quickly adapt to concept drift. Algorithm 1 provides the pseudo-code for FLASH.

3.1. Client-side Early-stopping Training

In one round of FL, each selected client receives the global model from the server, trains the global model locally via *early-stopping training*, and sends the locally-updated global model (called "local model") back to the server. Early-stopping training trains the global model until an early stop criteria based on model fitness is met. This is in contrast to the prevalent way of client-side training (Canonaco et al., 2021; Deng et al., 2020; Li et al., 2021; Mansour et al., 2020), i.e, each client trains the the global model for a fixed number of epochs. Lines 6 to 15 in Algorithm 1 describe the training procedure of each of the participating clients.

As a stopping criterion, we use decrement in the validation loss value (see Line 8) to indicate when a local model $w_c^{(r)}$ reaches its steady state. If the validation loss stops to decrease by a threshold γ/e, where γ is a threshold hyperparameter and e is the current epoch count, we stop training for the client. The threshold decreases as epoch count grows since the validation loss is also expected to decrease. Otherwise a client can train the local model until a set number of maximum epochs E.

The rationale behind the early-stopping training are twofolds. First, training for longer epochs in an FL round is demonstrate to be an effective approach for addressing statistical heterogeneity because it creates local models which can better fit the clients' data distributions than training for only a few epochs (Wu et al., 2022). Early-stopping further avoids potential over-fitting of the local model. Second, early-stopping training produces parameter updates that are necessary for the global model to fit a client's local data

Algorithm 1: FLASH

Input: R: total number of rounds, r: round index, p: participation rate, $\mathcal{C}^{(r)}$: a set of available clients for r^{th} round, $\mathcal{C}$: a set of currently participating clients, c: client index, η_ℓ: local learning rate, η_g: global learning rate, $\mathcal{D}_c^{(r)}$: local dataset of client c in r^{th} round, $e_c^{(r)}$: number of epochs for client c for round r, E: maximum number of epochs a client can train for, $\mathcal{L}_c$: local objective for client c, $\beta_{\{1,2\}}$: exponential decay rates, τ: adaptivity rate, γ: loss decrement threshold

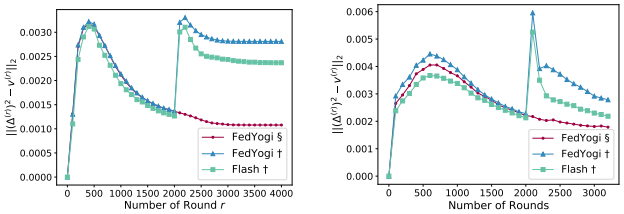
Output: $w_g^{(R)}$: global model

1 server randomly initializes $w_g^{(0)}$
2 **for** $r \in [R]$ *round* **do**
3 sample $\mathcal{C}$ from $\mathcal{C}^{(r)}$ with the rate of p
4 send $w_g^{(r-1)}$ to all the clients in $\mathcal{C}$
5 **for** $c \in \mathcal{C}$ *in parallel* **do**
6 $w_{c,0}^{(r)} \leftarrow w_g^{(r-1)}$
7 **for** $e \in [E]$ *epochs* **do**
8 **if** $\ell_{c,e-1}^{(r)} - \ell_{c,e}^{(r)} \geq \gamma/e$ **then**
9 $w_{c,e}^{(r)} \leftarrow w_{c,e-1}^{(r)} - \eta_\ell \nabla \mathcal{L}_c(w_{c,e-1}^{(r)}; \mathcal{D}_{c,train}^{(r)})$
10 $\ell_{c,e}^{(r)} \leftarrow \sum_{(x,y)\in\mathcal{D}_{c,valid}^{(r)}} f_c(w_{c,e}^{(r)}, x, y)$
11 **else**
12 **break**
13 **end**
14 **end**
15 **return** $w_c^{(r)}$ to the server
16 **end**
17 $\Delta^{(r)} \leftarrow \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(w_c^{(r)} - w_g^{(r-1)})$
18 $m^{(r)} \leftarrow \beta_1 m^{(r-1)} + (1-\beta_1)\Delta^{(r)}$
19 $v^{(r)} \leftarrow \beta_2 v^{(r-1)} + (1-\beta_2)(\Delta^{(r)})^2$
20 $\beta_{3j} \leftarrow \frac{\|v_j^{(r-1)}\|_2}{\|(\Delta_j^{(r)})^2 - v_j^{(r)}\|_2 + \|v_j^{(r-1)}\|_2} \; \forall j \in [n]$
21 $d_j^{(r)} \leftarrow \beta_{3j} d_j^{(r-1)} + (1-\beta_{3j})\left((\Delta_j^{(r)})^2 - v_j^{(r)}\right) \; \forall j \in [n]$
22 $w_g^{(r)} \leftarrow w_g^{(r-1)} + \eta_g \frac{m^{(r)}}{\sqrt{v^{(r)}} - d^{(r)} + \tau}$
23 **end**

distribution. The magnitude of the parameter updates could indicate if concept drifts. We empirically verify the two rationales in Section 5.4 and demonstrated that early-stopping training is able to improve validation accuracy by 1.41%-3.87% across datasets in no concept drift settings. And in concept drift settings, early-stopping training is more effective in detecting concept drifts than fixed-epoch training.

3.2. Server-side Drift-aware Adaptive Optimization

After receiving the local model from each participating client, the server aggregates the parameter updates into *effective gradients* (Line 17) and updates the global model via *drift-aware adaptive optimization* (Lines 18-22). Here the goal is to automatically increase effective learning rate



(a) CIFAR 10 (b) CIFAR 100

Figure 2. The gradient disparity $||(\Delta^{(r)})^2 - v^{(r)}||$ v.s. round r. Sudden concept drift occurs at $r = 2000$ for FLASH and FEDYOGI. † = with drift, § = without drift.

when there is a concept drift while preserving the convergence properties of existing adaptive optimizers. Although the magnitude of effective gradients is a good indicator of concept drift, the challenge lies in how much the effective learning rate needs to be adjusted accordingly.

To address the challenge, we propose a metric called *gradient disparity* that calculates the difference between the squared gradients $(\Delta^{(r)})^2$ and the second moment of the gradients $v^{(r)}$ at round r, i.e., $||(\Delta^{(r)})^2 - v^{(r)}||$. We find that this metric not only correlates well with the inside of a concept drift but also lies in a value range that is comparable to $\sqrt{v^{(r)}}$ and thus can be used to adjust the effective learning rate. Figure 2 shows how the concept drift impacts gradient disparity. Concept drift at round $r = 2000$ results in a larger magnitude of the effective gradient $\Delta^{(r)}$ and thus a larger gradient disparity. This sudden increase in gradient disparity indicates that the global model needs to adapt to the new distribution.

Since an absolute value of $\left((\Delta^{(r)})^2 - v^{(r)}\right)$ doesn't indicate whether the value is larger or nearly equal to the previous ones , we track a moving average of $\left((\Delta^{(r)})^2 - v^{(r)}\right)$ as shown in Line 21 in Algorithm 1, symbolized by $d^{(r)}$. Instead of a fixed weight parameter β_3, we use an adaptive β_3 for weighing the two terms of $d^{(r)}$. Our aim is to weigh $d^{(r-1)}$ term more in case there is no "large" change in $\left((\Delta^{(r)})^2 - v^{(r)}\right)$ compared to its previous values. Inversely, we want a larger weighted $\left((\Delta^{(r)})^2 - v^{(r)}\right)$ if the value widely differs from the previous ones. We assign β_3 as shown in Line 20 in Algorithm 1. We use the equation in Line 22 to update the global model.

4. Analysis of FLASH

In this section, we theoretically analyze the convergence of early-stopping training on the client side and the convergence of FLASH. The goal is to show that, despite the drift-aware design, FLASH retains the convergence property of existing adaptive optimizers. Specifically, given a client with a local variance of σ_c, (defined in Assumption C.3), we

derive the upper bound of the number of epochs e_c needed to train a global model to meet our stopping criteria. We prove that FLASH has the same convergence rate as FEDYOGI. We also derive a relation between FEDYOGI and FLASH in terms of drift. The proofs are in Appendix C.

Theorem 4.1 (Convergence of Early-stopping SGD)**.** *Assume functions $\{F_c\}$ satisfy Assumptions C.1, C.2, C.3, and C.4. The output of the early-stopping SGD with the early stopping criteria, $\sum_{x,y} f_c(w_{c,e-1}^{(r)}; x, y) - \sum_{x,y} f_c(w_{c,e}^{(r)}; x, y) \geq \gamma/e, \ \forall\ e \in [E]$ and $\forall\ (x, y) \in \mathcal{D}_{c,valid}^{(r)}$, has an expected error smaller than ϵ for $\gamma \geq \frac{(F-\epsilon)}{\ln E + \frac{1}{E}}$ and η_ℓ, e_c satisfying*

- *Strongly convex, $\frac{1}{\mu E} \leq \eta_\ell \leq \frac{\log(\max(1, \mu^2 ED/c))}{\mu E}$, and $e_c = \mathcal{O}\left(\min\left(\frac{\mu D^2}{\epsilon} + \frac{G^2}{\mu\epsilon} + \frac{\sigma_c^2}{2\mu\epsilon}, \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E})\right)\right)$*
- *General convex, $\frac{1}{E} \leq \eta_\ell$, and $e_c = \mathcal{O}\left(\min\left(D^2 + \frac{G^2D^2}{\epsilon^2} + \frac{\sigma_c^2 D^2}{2\epsilon^2}, \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E})\right)\right)$*
- *Non-convex, $\frac{1}{E} \leq \eta_\ell$, and $e_c = \mathcal{O}\left(\min\left(F + \frac{G^2F}{\epsilon^2} + \frac{\sigma_c^2 LF^2}{2\epsilon^2}, \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E})\right)\right)$*

where $c := G^2 + \frac{\mu_c^2}{2}$, $D := \mathbb{E}\left\|w_{c,0}^{(r)} - w_c^\star\right\|$, and $F = f_c(w_{c,0}^{(r)}) - f_c(w_c^\star)$.

Discussion. Here we discuss the relationship between e_c, epochs taken to achieve an optimization error of $\epsilon = \mathbb{E}_r\left[f_c(w_{c,e_c^{(r)}}^{(r)})\right] - f_c(w_c^\star)$, and the concept drift $D = \left\|w_{c,0}^{(r)} - w_c^\star\right\|$. In case of FL, $w_{c,0}^{(r)} := w_g^{(r-1)}$. We focus on the general convex case, but the same analysis can be applied for non-convex case as well. Dropping all the terms related to L, G, and lower powers of ϵ, we get $e_c = O\left(D^2(1 + \frac{1}{\epsilon^2} + \frac{\sigma_c^2}{\epsilon^2})\right)$. This indicates that number of epochs a client would run the local training for depends on the drift between $w_g^{(r-1)}$ and $w_c^\star$. If $\mathcal{P}^{(r-1)} \neq \mathcal{P}_c^{(r)}$, then according to Lemma C.8, D would be larger, implying a large e_c. Besides, e_c also depends on the early stopping parameter γ. A small γ mean large e_c, as a lower limit to the error difference $f_c(w_{c,e-1}^{(r)}) - f_c(w_{c,e}^{(r)})$ allows for more epochs of local SGD, and vice versa.

Theorem 4.2 (Convergence of FLASH)**.** *Let assumptions C.1 to C.4 hold. Suppose the server and client learning rates satisfy $\eta_\ell \leq \min\left[\left(\frac{|\mathcal{C}|}{30L^2E}\right)^{\frac{1}{2}}, \left(\frac{\tau}{6(B^2-1)\left[G(\beta_2+\sqrt{\beta_2})+L\eta_g\right]}\right)\right]$. Then the iterates of Algorithm 1 for $\eta_\ell = \Theta(1/L\sqrt{E})$, $\eta_g = \Theta(1/\sqrt{R})$, and $\tau = G/L$ for FLASH satisfy*

$$\min_{0\leq r\leq R} \mathbb{E}\left\|\nabla f(w_g^{(r)})\right\|^2 \leq \mathcal{O}\Bigg(\frac{f(w_g^{(0)}) - \mathbb{E}_r[f(w_g^{(R)})]}{\sqrt{ER}} + \frac{G}{\sqrt{ER|\mathcal{C}|}}(\sigma_\ell^2 + 6E\sigma_g^2) + \frac{6L\sigma_\ell^2}{RG^2|\mathcal{C}|} + \frac{6L}{R}\Bigg)$$

Discussion FLASH obtains convergence at the rate of $\mathcal{O}(1/\sqrt{R})$, which matches FEDYOGI's convergence rate. As $R \to \infty$, the error tends to 0. The local and global variance terms are also weighed by $1/|\mathcal{C}|$. Meaning that with more participating clients, the impact of local and global variance decreases. In Section 5, we empirically show that FLASH converges faster than FEDYOGI.

Theorem 4.3 (Lower Bounding the Change in the Second Moment of Effective Gradients)**.** *Let $\{F_c\}$ satisfy Assumptions C.1, C.3, and C.5. In round r, the updates in* FLASH *and* FEDYOGI *satisfy,*

$$\begin{aligned}
&\text{FEDYOGI}\left(\mathbb{E}_r\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) - \text{FLASH}\left(\mathbb{E}_r\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) \\
&\geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} + \tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \\
&- \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} - \eta_\ell EG(1-\beta_3^r) + \tau\right)}\right\|\right. \\
&\quad \left. - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right|
\end{aligned}$$

Discussion. Lower bounding the change in the second moment of effective gradients gives us an intuition of what would be the behavior of an adaptive optimizer during, and after a concept drift. We see that as $\beta_3 \to 1$ (steady state), FLASH behaves similar to FEDYOGI and there's no major change in the lower bound, confirming with the results shown in Figure 2. But at the onset of a drift, $\beta_3 \to 0$ and the lower bound for FLASH decreases due to the term $\eta_\ell EG(1-\beta_3^r)$. This is the reason why we see a lower value of $\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|$ for FLASH during and after a concept drift. Besides, assuming that $w_g^{(r-1)}$ and $w_g^{(r)}$ are capturing *pre-* and *post*-drift global distributions $\mathcal{P}$ and $\mathcal{P}'$ respectively, the lower bounds would be high for both.

5. Experiments

We evaluate FLASH on both convex and non-convex tasks using 6 non-iid federated datasets. We test the efficacy of FLASH on three concept drift setups: sudden, incremental, and recurrent, and compare our results against state-of-the-art adaptive optimization, personalization, drift correction, and drift adaptation approaches.

5.1. Experiment Settings

Datasets, Tasks, and Models. We have a convex task: Classification of Synthetic data (Li et al., 2020) with a 2 layer fully connected model (Li et al., 2020) (**Synthetic**). For non-convex tasks, we used Stackoverflow Next Word Prediction (Stackoverflow, 2023) with a 1 layer LSTM model (Reddi et al., 2021) (**Stackoverflow**), EMNIST (Cohen et al., 2017) Image Classification with a 2 layer CNN model (Reddi et al., 2021) (**EMNIST**), Shakespeare (McMahan et al., 2017) Next Character Prediction with a 2 layer LSTM model (Reddi et al., 2021) (**Shakespeare**). The above three datasets have natural non-IID partitions based on the authors. Each author is considered a client in the federated setup, the images/texts generated by that author are samples for that client. We also use CIFAR10/100 (Krizhevsky et al., 2009) for image classification tasks based on ResNet18 (He et al., 2016) (**CIFAR10/100**). Both datasets are generated artificially in a federated non-iid fashion based on Dirichlet distribution as described in (Reddi et al., 2021). More details in Appendix A.

Baselines. Our baselines fall in four groups: (a) Server- and client-side optimization (FEDAVG, FEDPROX, FEDYOGI, FEDNOVA), (b) Personalization (APFL, DITTO, HYPCLUSTER), (c) Drift correction (SCAFFOLD, FEDDYN, FEDDC), and (d) Drift adaptation (FEDDRIFT, ADAPTIVEFEDAVG shortened to ADAPFA). Note that in case of FEDDRIFT, we are using FEDDRIFTEAGER since our use case demands a synchronized concept drift. We also use an ORACLE algorithm which has knowledge about *at what round which client faces a concept drift*. ORACLE uses FEDYOGI as its base algorithm. ORACLE simultaneously trains two global models, each based on the pre- and post- drift distributions. With its knowledge about when the drift occurs, ORACLE simply switches the current global model to reflect the drift.

Metrics. Here we define four metrics used to compare performance of FLASH and its baselines. **Generalized Accuracy** for a round r is the average accuracy of the global model $w_g^{(r)}$ on the test data $\mathcal{D}_c^{(r)}$ of clients $c \in \mathcal{C}^{(r)}$, where $\mathcal{C}^{(r)}$ is a set of available clients for the round r. **Personalized Accuracy** for a round r is the average accuracy of the local model $w_{c,e_c}^{(r)}$ on the test data $\mathcal{D}_c^{(r)}$ of clients $c \in \mathcal{C}$, where $\mathcal{C}$ is a set of participating clients for the round r. **Accuracy Dip** is the lowest generalized accuracy during a concept drift starting from round r_1 to round r_2. It is defined as $dip^{(r_1,r_2)} = \min(\{acc_g^{(r)}\ \forall r \in [r_1, r_2]\})$. **Rounds till Recovery** is the number of rounds taken for the global model $w_g^{(r)}$ to reach a steady state after a concept drift has occurred.

Concept Drift Setups. Recall that concept drift occurs when the conditional distribution $\mathcal{P}(y \mid x)$ changes. Following the practice in (Tahmasbi et al., 2020; Jothimurugesan et al., 2022), we use *label swapping* to simulate concept drifts, given the same input features. Since it is not feasible to swap labels for next word prediction tasks, we use the classification tasks (EMNIST, CIFAR10, CIFAR100, and Synthetic) for concept drift experiments. For a task with n labels, we swap i^{th} label with $i + 1^{st}$ label $\forall\ i \in [0, 2, \ldots, n]$. Specifically, we simulate three types of concept drifts. (1) **Sudden Drift**. All the clients face the distribution change abruptly, at the same round. For EMNIST, CIFAR10, CIFAR100, and Synthetic datasets, the concept drift occurs at 600^{th}, 2000^{th}, 2000^{th}, and 500^{th} rounds respectively. (2) **Incremental Drift**. Starting on the same rounds as described in sudden drift, for every 100 rounds, 20% more clients change their distributions to the new one. (3) **Recurrent Drift**. First drift occurs abruptly for EMNIST, CIFAR10, CIFAR100, and Synthetic datasets at 600^{th}, 2000^{th}, 2000^{th}, and 500^{th} rounds respectively. Next drift to the old (initial) distribution occurs at 1000^{th}, 2500^{th}, 2500^{th}, and 800^{th} rounds.

We use Flower (Beutel et al., 2020) library to implement FLASH and all its baselines. The implementation is available on [1]. We use an NVidia 2080ti GPU to run all the experiments with 3 runs for each. The random seeds used are 0, 44, and 56. For all the tasks and datasets, we have uniformly randomly sampled 10 clients per round. Hyperparameter details are given in Appendix A.

Table 1. Generalized accuracy (the higher, the better) for FLASH and baselines **with no concept drift**. EM = EMNIST, SO = Stackoverflow, SH = Shakespeare, C10 = CIFAR10, C100 = CIFAR100. Full results in Appendix Tables 3 - 6.

Tasks	EM	SO	SH	C10	C100	C10	C100
"Non-iid"ness	Writers	Authors	Authors	$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.6$	$\alpha = 0.6$
FEDAVG	84.12%	28.24%	57.29%	56.91%	36.00%	75.28%	41.52%
FEDPROX	87.52%	27.78%	57.43%	56.74%	35.83%	76.23%	42.63%
FEDYOGI	87.71%	28.96%	57.82%	67.57%	39.92%	78.30%	44.33%
FEDNOVA	88.53%	28.56%	58.23%	66.59%	37.45%	77.89%	43.82%
SCAFFOLD	88.10%	27.68%	57.59%	65.86%	32.48%	75.43%	44.14%
FEDDYN	88.18%	26.08%	57.43%	65.08%	35.20%	77.33%	45.33%
FEDDC	**90.64%**	27.50%	56.46%	65.16%	39.89%	79.28%	44.89%
APFL	89.78%	24.32%	**59.46%**	68.90%	36.44%	78.05%	**45.78%**
DITTO	87.17%	27.97%	59.16%	69.41%	35.41%	78.34%	45.73%
HYPCLUSTER	89.66%	27.96%	58.50%	68.23%	39.24%	78.66%	45.49%
ADAPFA	88.46%	28.76%	58.42%	66.91%	39.48%	78.31%	44.36%
FEDDRIFT	89.12%	27.65%	58.28%	69.18%	38.09%	78.46%	44.49%
FLASH	88.17%	**29.13%**	58.27%	**69.45%**	**40.65%**	**79.58%**	45.65%

5.2. Performance Comparison without Concept Drift

This set of experiments aim to demonstrate that FLASH retains the benefits of adaptivity to address statistical heterogeneity issues in FL. We ran FLASH and its baselines in a no concept drift setup to get the base generalized and personalized accuracies. For a fair comparison, we used the same early-stopping criteria for client-side training for all the baselines.

[1] Source Code

Table 1 reports the generalized accuracy. Personalized accuracy shows similar trends and is reported in Appendix B, tables 5 and 6. *Overall,* FLASH *achieves comparable generalized accuracy as state-of-the-art techniques in addressing statistical heterogeneity.* Specifically, FLASH outperforms adaptive optimization approaches FEDYOGI and drift correction methods SCAFFOLD and FEDDYN (except with EMNIST) across all tasks. It demonstrates that the new model updating rules in FLASH, designed for concept drift adaptation, retains the benefits of adaptive learning rates. Moreover, FLASH gives comparable performance against personalization methods APFL, DITTO, and HYPCLUSTER. The results echo the findings in (Wu et al., 2022) that multiple epochs of local training (or finetuning) works better than other personalization methods. FEDDRIFT in no drift setting shows no advantage over another clustering algorithm, HYPCLUSTER. ADAPFA, also being an optimizer for drift adaptation, shows comparable performance to FEDYOGI.

5.3. Performance Comparison with Concept Drift

This set of experiments aims to demonstrate the better performance of FLASH than baselines in adapting to concept drifts. Here, we report the results from one type of concept drift, *incremental drift*, as described in Section 5.1. Results for the other two types, sudden and recurrent concept drifts, are similar and reported in Appendix B.

Figure 3 shows validation generalized accuracy curves with an incremental drift. Table 2 further reports the accuracy dip and rounds till recovery to the steady state. We observe that although ADAPFA competes with FLASH in terms of round till recovery, it does it with the sacrifice in its *during*-drift accuracy degradation of up to 49.14% (C100) compared to FLASH. While other methods like FEDDRIFT come close to FLASH in the accuracy dip criteria, they still struggle to recover quickly due to new model creation and training at the onset of a drift. ORACLE does not face any performance degradation due to possessing global models based on both the *pre-* and *post-* drift distribution, FLASH still outperforms it through a more stable effective learning rate adaptivity.

Given the same number of rounds *pre*-drift and *post*-drift, FEDYOGI is unable to recover to the same steady-state accuracies in case of EMNIST, CIFAR100, and CIFAR10 ($\alpha = 0.1$) tasks. For all the datasets, the accuracy dip (see Table 2) is also larger for FEDYOGI compared to FLASH. The faster recovery of FLASH when the drift is injected for 20% more clients every 100 rounds indicates that in the first 100 rounds interval, FLASH can still detect the drift and adapt accordingly.

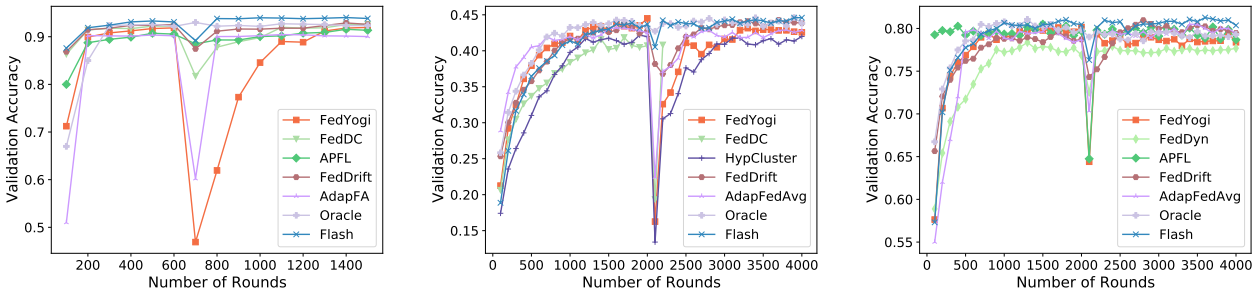
We also make observations for the cases of *adaptation* and *starting from scratch* in the concept drift setup. The main difference between FLASH and FEDDRIFT is that while FLASH lets the same global model adapt (on server side) to a new distribution, FEDDRIFT starts a new global model when a client encounters a new distribution. As described in Section 5.1, for this specific case of concept drift, the feature distribution $\mathcal{P}(x)$ remains the same after the labels have been swapped. Hence, the global model based on $\mathcal{P}_1(x, y)$ would still offer utility when learning a new distribution $\mathcal{P}_2(x, y)$ in terms of having similar feature extraction parameters before and after the concept drift. Hence we see a lower accuracy dip and faster adaptation in case of FLASH.

Table 2. The lowest accuracy (the higher, the better) during the concept drift (D) and rounds till recovery (the lower, the better) to the steady state (R) in an **incremental concept drift setting**. EM = EMNIST, C10 = CIFAR10, C100 = CIFAR100.

Tasks	EM (D) (R)	C10 (D) (R)	C100 (D) (R)	C10 (D) (R)	C100 (D) (R)
Non-iid	Writers	$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.6$	$\alpha = 0.6$
FEDAVG	14.64% 510	43.63% 290	18.39% 610	62.75% 420	14.88% 670
FEDPROX	64.94% 460	44.40% 270	17.67% 570	63.10% 400	19.48% 540
FEDYOGI	46.90% 360	45.35% 220	19.46% 510	64.40% 180	16.26% 510
FEDNOVA	58.16% 430	44.13% 240	20.64% 530	62.76% 270	17.49% 550
SCAFFOLD	64.35% 620	47.30% 230	20.89% 500	56.30% 250	15.40% 380
FEDDYN	64.24% 380	52.60% 220	19.83% 430	72.55% 230	18.64% 490
FEDDC	81.73% 300	53.15% 230	20.19% 400	71.64% 240	19.33% 480
APFL	88.50% 320	59.44% 180	25.43% 350	64.75% 190	20.07% 560
DITTO	59.71% 540	56.18% 160	28.49% 430	64.28% 170	18.99% 600
HYPCLUST	85.35% 400	58.65% 140	26.20% 580	64.00% 490	13.41% 620
ADAPFA	59.46% 320	52.25% 70	20.41% 260	70.31% 110	23.17% 270
FEDDRIFT	83.63% 350	63.63% 130	36.10% 240	74.35% 160	38.08% 430
ORACLE	91.65% 0	72.85% 0	40.13% 0	79.29% 0	42.72% 0
FLASH	**89.18% 260**	**70.45% 60**	**38.65% 210**	**76.30% 80**	**40.52% 220**

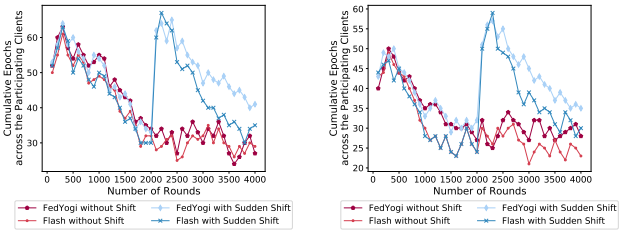
Drift correction methods are rigid to a single global distribution. They update local models based on the assumption of a single global distribution, rather than adapt global model to the distributions learned by the local models. Personalization helps in case of the aforementioned issue of rigidity since the personalized models of each client do not need to "correct" themselves according to any global distribution. Yet the global models generated by each client remain suboptimal because of the slower adaptability of the server-side aggregation methods of these personalization approaches.

Computation Savings after Concept Drifts. We further report the computation savings resulting from FLASH's faster adaptation to concept drifts compared to FEDYOGI. After concept drifts, FLASH requires less number of federated rounds to recover accuracy, which directly translates to the less number of epochs on the client side to train the global model. Figure 4 shows the total number of epochs to train the global model to reach the same early-stopping criteria for FLASH and FEDYOGI in each round. The total number of epochs is the sum of the number of epochs per participating client in that round. At the onset of both the training (round 0) and the drift (round 2000), a client has to train the global model for many epochs (4-6 epochs in average per client) as the global model has not yet reached a steady state. As the training approaches the steady state (starting from around $r = 1000$), each client uses less number of epochs (less than 3 epochs in average). We observe that, *after concept drift occurs,* FLASH *has 11.18-11.79% lower*



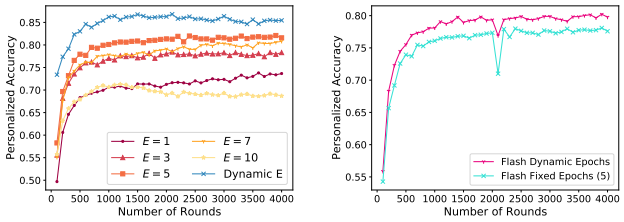
(a) EMNIST (b) CIFAR 100 ($\alpha = 0.6$) (c) CIFAR 10 ($\alpha = 0.6$)

Figure 3. Accuracy curves with **incremental** drift after steady state. The validation accuracies have been averaged across 100 rounds of interval. Plotted baselines are best of each categories from server-side optimization (FEDYOGI), personalization, and drift correction (varies depending on the dataset). All drift adaptation baselines along with FLASH and ORACLE are shown. Rest of the baselines depicted in Figure 8 in Appendix B.



(a) CIFAR 10 (b) CIFAR 100

Figure 4. Total number of epochs ran across participating clients per round r for FLASH vs FEDYOGI. Global model adaptation in FLASH leads to lower amount of local training done by the clients.



(a) No drift (b) With sudden drift

Figure 5. Personalized accuracy of FLASH with dynamic number of epochs (through early-stopping) versus fixed number of epochs $E \in \{1, 3, 5, 7, 10\}$ on CIFAR10.

number of epochs per round compared to FedYogi because FLASH *adapts to the new distribution faster, resulting in clients having to do less work.*

Impact of the Ratio of Drifted Clients. We observe the impact of a ratio p_{drift} of "clients facing a concept drift" to total clients in a round for FLASH. As the ratio decreases, the global impact of only a few clients ($p_{drift} < 0.25$) getting injected with a drift gets absorbed by rest of the clients who do not face a drift. Results on the drift ratio and its impact on adaptation are in Appendix B.

5.4. Fixed Epochs versus Early Stopping

This set of experiments aims to verify the importance of early-stopping training by comparing it with fixed-epoch training. We empirically verify the two benefits of early-stopping training (see Section 3.1): (1) It improves the personalized accuracy of the local model on a client and (2) It produces a more reliable signal in detecting concept drifts. Figure 5a reports the personalized accuracy curves from fixed-epoch training and early-stopping training in a no concept drift setting. We observe that *early-stopping training outperforms the best of the fixed-epoch training (with $E = 5$) in personalized accuracy.* With early stopping, each client's epoch count depends on the local gradient variance of the client and its rate of loss reduction (see Theorem 4.1), which in turn is based on its heterogeneity. Hence, a *less* heterogeneous client could spend less epochs on local training, avoiding over-fitting. The same trend is also seen in generalized accuracy curves.

Figure 5b shows the accuracy curves with sudden concept drifts at round 2000 using CIFAR10 dataset. The lower accuracy dip from early-stopping training at the onset of the concept drift demonstrates that early-stopping training is better than fixed-epoch training in adapting concept drift.

6. Conclusion

In this work, we studied concept drift adaptation in federated learning. We proposed FLASH that leverages client-side early-stopping training to facilitate the concept drift detection and server-side drift-aware adaptive optimization to address both statistical heterogeneity and concept drift adaptation. We gave convergence rate for FLASH and its

client-side early stopping training. We also empirically evaluated the effectiveness of FLASH in improving generalized and personalized accuracy and reducing accuracy dips at the onset of concept drifts and the federated rounds taken to recover from concept drifts using a set of tasks and different drift patterns. For future work, it would be interesting to see how this approach translates to real-world drifts where the drifts can occur in an ad-hoc manner and changes in joint distribution instead of class conditional distribution.

References

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Shanshan Wu, Tian Li, Zachary Charles, Yu Xiao, Ziyu Liu, Zheng Xu, and Virginia Smith. Motley: Benchmarking heterogeneity and personalization in federated learning. *arXiv preprint 2206.09262*, 2022.

Tyler J Loftus, Matthew M Ruppert, Benjamin Shickel, Tezcan Ozrazgat-Baslanti, Jeremy A Balch, Philip A Efron, Jr. Gilbert R Upchurch, Parisa Rashidi, Christopher Tignanelli, Jiang Bian, and Azra Bihorac. Federated learning for preserving data privacy in collaborative healthcare research. *Digital Health*, 2022.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, Boris Muzellec, Constantin Philippenko, Santiago Silva, Maria Teleńczuk, Shadi Albarqouni, Salman Avestimehr, Aurélien Bellet, Aymeric Dieuleveut, Martin Jaggi, Sai Praneeth Karimireddy, Marco Lorenzi, Giovanni Neglia, Marc Tommasi, and Mathieu Andreux. FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Suresh Dara, Ambedkar Kanapala, A. Ramesh Babu, Swetha Dhamercherala, Ankit Vidyarthi, and Ruchi Agarwal. Scalable federated-learning and internet-of-things enabled architecture for chest computer tomography image classification. *Computers and Electrical Engineering*, 2022.

Li Li, Xi Yu, Xuliang Cai, Xin He, and Yanhong Liu. Contract theory based incentive mechanism for federated learning in health crowdsensing. *IEEE Internet of Things Journal*, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

Giuseppe Canonaco, Alex Bergamasco, Alessio Mongelluzzo, and Manuel Roveri. Adaptive federated learning in presence of concept drift. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arxiv preprint 2002.10619*, 2020.

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arxiv preprint 1909.12488*, 2019.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint 2003.13461*, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. *International Conference on Machine Learning*, 2021.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip Gibbons. Federated learning under distributed concept drift. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014.

Ashraf Tahmasbi, Ellango Jothimurugesan, Srikanta Tirthapura, and Phillip B. Gibbons. Driftsurf: A risk-competitive learning algorithm under concept drift. *arxiv preprint arXiv:2003.06508*, 2020.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 2020.

Stackoverflow. TensorFlow Federated Datasets. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets, 2023. [Online; accessed 09-Jan-2023].

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint 2007.14390*, 2020.

Stackoverflow Dataset FedML. FedML Federated Stackoverflow Heterogenity . https://github.com/FedML-AI/FedML/blob/ecd2d81222301d315ca3a84be5a5ce4f33d6181c/doc/en/simulation/benchmark/BENCHMARK_MPI.md, 2022. [Online; accessed 14-Sep-2022].

Shakespeare. TensorFlow Federated Datasets. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/shakespeare, 2023. [Online; accessed 09-Jan-2023].

CIFAR100. TensorFlow Federated Datasets. https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/cifar100/load_data, 2023. [Online; accessed 18-Jan-2023].

A. Datasets and Hyperparameters

EMNIST EMNIST dataset (Cohen et al., 2017) has 3400 unique clients where each client has train, validation, and test datasets. The train datasets of all the clients combined have in total 671,585 instances. The validation and test datasets of all the clients combined have a total of 77,483 instances each. The heterogeneity of EMNIST clients stems from the individual writing style of each client (where one client is one person), as discussed in Appendix C.2 of (Reddi et al., 2021).

We have trained each of our baselines and FLASH for $R = 1500$ rounds, with the batch size of $N = 20$ instances, 10 clients per round. All the experiments have ran for $E = 10$ with the early-stopping criteria discussed in 3.1. The default learning rates for all the experiments is $\eta_\ell = 0.05$ and $\eta_g = 1.00$. Although SCAFFOLD and FEDDYN required $\eta_\ell = 0.03$. For both FEDPROX and FEDDYN, λ was assigned 0.001. APFL has $\alpha = 0.25$. And DITTO has $\lambda = 0.1$ and client learning rate of $\eta_\ell = 0.01$. α in FEDDC has been assigned to 0.5. While ρ in FEDNOVA has been assigned to 0.8. FLASH has $\gamma = 0.04$.

Stackoverflow Stackoverflow dataset (Stackoverflow, 2023) has separate clients for training, validation, and testing. There are 342,477 train clients, having a combined sample count of 135,818,730. Similarly, there are 38,758 validation and 204,088 test clients having a combined sample count of 16,491,230 and 16,586,035 respectively. Since we need validation set for early-stopping training for each client, we divide a client's train dataset in a 7:3 split to get the training and the validation sets. This is a naturally heterogeneous dataset (FedML, 2022). Each user of Stackoverflow is a client and their posts form a dataset for the client. The dataset is heterogeneous in two ways: First, users have different writing styles and thus clients' datasets are not i.i.d. Second, the total number of posts from each user is also different, leading to different sizes of datasets per client.

We have trained each of our baselines and FLASH for $R = 2000$ rounds, with the batch size of $N = 16$ instances, 10 clients per round. The vocabulary consists is of 10,000 tokens. Similar to EMNIST, all the experiments have ran for $E = 10$ with the validation loss based early-stopping criteria. The default learning rates for all the experiments is $\eta_\ell = 0.3$ and $\eta_g = 1.00$. FEDPROX has λ set to 0.01. While λ for FEDDYN and DITTO have been set to 0.001 and 0.1 respectively. For APFL, the interpolation factor α is 0.25. The value of ρ for FEDNOVA is 0.9. FLASH has γ set to 0.05.

Shakespeare Shakespeare dataset (Shakespeare, 2023) has 715 unique clients where each client has train, validation, and test datasets. The train datasets of all the clients combined have in total 12,854 instances. The validation and test datasets of all the clients combined have a total of 3,214 and 2,356 instances respectively. Shakespeare dataset is heterogeneous because of each client is one play written by William Shakespeare, and all the plays have different setting and characters.

Each of our baselines and FLASH are trained for $R = 1500$ rounds, with the batch size of $N = 4$, 10 clients per round. The vocabulary size is 90 as each token is related to a character in the English language. The maximum number of epochs E is set to 10 for all the algorithms, with early-stopping. Default learning rates for all the experiments are $\eta_\ell = 0.1$ and $\eta_g = 1.00$. FEDPROX and FEDDYN has their $\lambda = 0.001$. While DITTO has $\lambda = 0.1$. APFL and FEDDC have their $\alpha = 0.5$. FEDNOVA has $\rho = 0.85$. Early stopping threshold γ of FLASH is set to 0.05.

CIFAR10 CIFAR10 dataset is created from the centralized version of CIFAR10 dataset (Krizhevsky et al., 2009) having 50,000 images. Federated CIFAR10 dataset has 500 unique clients, each having 100 samples for training, and 20 samples for testing. A client would receive the training and testing samples according to the Dirichlet distribution (Reddi et al., 2021). A Dirichlet distribution with the parameter $\alpha \in [0, 1]$ determines the heterogeneity of a client, with a client being more heterogeneous as $\alpha \to 0$. Here, heterogeneity means how dissimilar the dataset instances sampled from a distribution are. We have experimented on $\alpha = 0.1$ and $\alpha = 0.6$.

FLASH and its baselines are trained for $R = 4000$ rounds, with batch size of $N = 20$, with 10 clients per round. Maximum number of epochs is set to $E = 15$ for the early stopping criteria. Default learning rates for all the experiments are $\eta_\ell = 0.05$ and $\eta_g = 0.5$. SCAFFOLD, HYPCLUSTER, FEDDYN and APFL are set for $\eta_\ell = 0.01$, 0.01 and 0.09 respectively. FEDPROX, DITTO and FEDDYN have their λ set to 0.08, 0.01, and 0.05. APFL and FEDDC have their $\alpha = 0.25$ and 0.01. FEDNOVA has $\rho = 0.9$. FLASH has $\gamma = 0.04$.

CIFAR100 Similar to CIFAR10, CIFAR100 dataset (CIFAR100, 2023) is also created from CIFAR100 dataset (Krizhevsky et al., 2009) having 50,000 images. The client count and train-test image count are same as that of CIFAR10. Here too, we have experimented with the Dirichlet parameter $\alpha = 0.1$ and $\alpha = 0.6$.

FLASH and its baselines are trained for $R = 4000$ rounds, with batch size of $N = 20$, with 10 clients per round. Maximum number of epochs is set to $E = 15$ for the early stopping criteria. Default learning rates for all the experiments are $\eta_\ell = 0.01$ and $\eta_g = 0.5$. Although SCAFFOLD works best at $\eta_\ell = 0.03$. FEDPROX, DITTO and FEDDYN have their λ set to 0.01. APFL and FEDDC have their $\alpha = 0.25$ and 0.01. FEDNOVA has $\rho = 0.8$. FLASH has $\gamma = 0.05$.

Synthetic Synthetic dataset is generated with the same procedure described in (Li et al., 2020). The total number of clients is 30. The dimension of input features is 60, and there are 10 output classes. Each client has sample count randomly generated from the log-normal distribution with its $\mu = 4$ and $\sigma = 2$, and size = number of clients. The parameter which controls how much local models differ from each other is set to $\alpha = 0.5$. And the parameter which controls how much the local data at each client differs from that of other clients is set to $\beta = 0.5$.

We have trained all the baselines and FLASH for $R = 1000$ rounds, with batch size of $N = 10$, with 10 clients per round. Maximum number of epochs is set to $E = 8$ for the early stopping criteria. Default learning rates for all the experiments are $\eta_\ell = 0.01$ and

Table 3. Generalized accuracy (P), standard deviation of individual client's accuracy (I), and standard deviation of individual experiment run (E) for FLASH vs baselines, in a **no distribution change setting**. EM = EMNIST, SO = Stackoverflow, C10 = CIFAR10, C100 = CIFAR100.

Tasks "Non-iid"ness	EM (P) Writers	(I)	(E)	SO (P) Authors	(I)	(E)	C10 (P) $\alpha = 0.1$	(I)	(E)	C100 (P) $\alpha = 0.1$	(I)	(E)
LOCAL ONLY	-	-	-	-	-	-	-	-	-	-	-	-
FEDAVG	84.12%	7.92%	1.35%	28.24%	4.44%	0.22%	56.91%	17.69%	1.42%	36.00%	10.21%	0.54%
FEDPROX	87.52%	4.92%	1.10%	27.78%	4.26%	0.09%	56.74%	18.34%	1.23%	35.83%	11.80%	0.30%
FEDYOGI	87.71%	7.14%	1.02%	28.96%	4.36%	0.19%	67.57%	15.08%	1.57%	39.92%	10.39%	0.34%
FEDNOVA	88.53%	7.15%	0.96%	28.56%	4.03%	0.22%	66.59%	8.23%	1.77%	37.45%	10.49%	0.46%
SCAFFOLD	88.10%	7.47%	1.12%	27.68%	4.32%	0.20%	65.86%	13.17%	1.65%	32.48%	10.46%	0.40%
FEDDYN	88.18%	7.84%	0.86%	26.08%	4.52%	0.12%	65.08%	12.90%	1.43%	35.20%	11.19%	0.62%
FEDDC	90.64%	6.81%	0.79%	27.50%	4.76%	0.26%	65.16%	9.30%	1.14%	39.89%	10.44%	0.48%
APFL	89.78%	7.48%	0.90%	24.32%	4.49%	0.31%	68.90%	15.82%	0.95%	36.44%	11.13%	0.44%
DITTO	87.17%	5.01%	0.95%	27.97%	4.10%	0.20%	69.41%	13.73%	1.08%	35.41%	10.77%	0.35%
HYPCLUSTER	89.66%	7.31%	1.25%	27.96%	4.51%	0.24%	68.23%	11.08%	1.76%	39.24%	10.85%	0.55%
ADAPFA	88.46%	6.26%	1.06%	28.76%	4.68%	0.19%	66.91%	10.20%	1.27%	39.48%	10.60%	0.37%
FEDDRIFT	89.12%	6.36%	1.07%	27.65%	4.75%	0.23%	69.18%	11.27%	1.18%	38.09%	11.06%	0.39%
FLASH	88.17%	5.27%	1.20%	29.13%	4.22%	0.12%	69.45%	9.10%	1.24%	40.65%	11.24%	0.42%

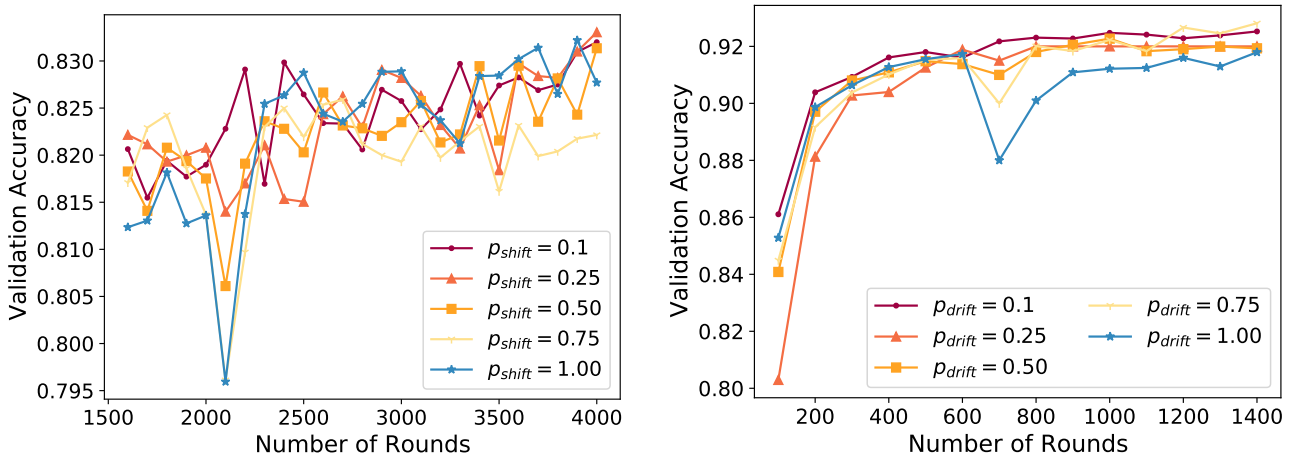
$\eta_g = 0.05$. Although SCAFFOLD, FEDDYN, FEDDC, DITTO prefer $\eta_\ell = 0.005$, $\eta_\ell = 0.005$, $\eta_\ell = 0.05$, $\eta_\ell = 0.005$. FEDPROX, DITTO and FEDDYN have their λ set to 0.25, 0.01, and 0.01. APFL and FEDDC have their $\alpha = 0.5$ and 0.1. FEDNOVA has $\rho = 0.95$. FLASH has $\gamma = 0.03$.

B. Additional Results

B.1. Few clients with sudden drift: FLASH balances local training and global adaptation

We recall that the goal of FLASH is to let all heterogeneous clients have well-trained local models, and if the majority of those local models exhibit a drift towards a new data distribution, we let the global model adapt to that new distribution. To this end, we find it interesting to see what ratio of clients have to face a concept drift for the global adaptation to get triggered. We experiment with different ratios of clients which face a sudden concept drift over all the participating clients of a certain round. As shown in Figure 6, $p_{drift} = 1.00$ is the most extreme case where all the clients face the drift. We see the highest accuracy dip there.

We also observe that the accuracy achieved after global drift adaptation is lower than what a global model following the same distribution throughout the training can achieve.



(a) CIFAR 10

(b) EMNIST

Figure 6. Behavior of FLASH with changing p_{drift}, the ratio of participating clients facing the sudden concept drift over total participating clients for a round r. The sudden concept drift occurs at 2000^{th} and 600^{th} rounds for CIFAR10 and EMNIST respectively.

Table 4. Generalized accuracy (P), standard deviation of individual client's accuracy (I), and standard deviation of individual experiment run (E) for FLASH vs baselines, in a **no distribution change setting**. SH = Shakespeare, SY = Synthetic, C10 = CIFAR10, C100 = CIFAR100.

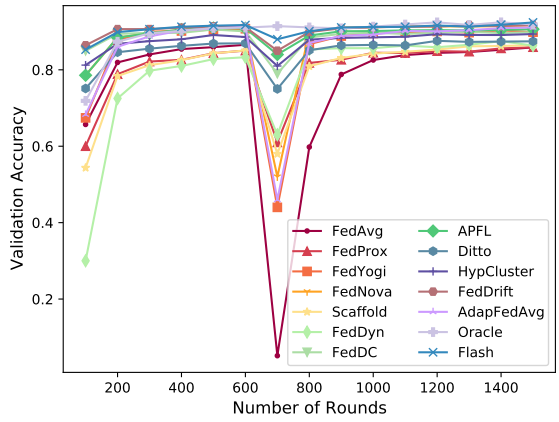
Tasks	SH (P)	(I)	(E)	SY (P)	(I)	(E)	C10 (P)	(I)	(E)	C100 (P)	(I)	(E)
"Non-iid"ness	Authors			$\alpha = 0.5$,	$\beta = 0.5$		$\alpha = 0.6$			$\alpha = 0.6$		
LOCAL ONLY	-	-	-	-	-	-	-	-	-	-	-	-
FEDAVG	57.29%	9.72%	0.35%	90.46%	24.07%	0.09%	75.28%	9.54%	0.78%	41.52%	10.52%	0.19%
FEDPROX	57.43%	10.48%	0.41%	92.76%	24.58%	0.09%	76.23%	9.00%	0.69%	42.63%	9.89%	0.36%
FEDYOGI	57.82%	10.59%	0.44%	93.20%	24.65%	0.16%	78.30%	9.02%	0.94%	44.33%	10.42%	0.24%
FEDNOVA	58.23%	9.43%	0.53%	93.57%	22.18%	0.23%	77.89%	8.41%	1.15%	43.82%	8.52%	0.53%
SCAFFOLD	57.59%	10.81%	0.53%	92.92%	28.64%	0.20%	75.43%	8.74%	0.86%	44.14%	9.80%	0.35%
FEDDYN	57.43%	11.57%	0.39%	91.89%	24.68%	0.14%	77.33%	9.35%	0.71%	45.33%	10.08%	0.14%
FEDDC	56.46%	11.55%	0.41%	93.10%	26.07%	0.15%	79.28%	9.83%	1.02%	44.89%	10.44%	0.35%
APFL	59.46%	7.04%	0.51%	91.78%	25.72%	0.08%	78.05%	8.72%	0.75%	45.78%	10.63%	0.29%
DITTO	59.16%	11.23%	0.43%	92.34%	15.79%	0.10%	78.34%	10.79%	0.83%	45.73%	10.58%	0.26%
HYPCLUSTER	58.50%	10.99%	0.60%	92.36%	25.15%	0.19%	78.66%	8.74%	0.74%	45.49%	9.91%	0.19%
ADAPFA	58.42%	11.40%	0.58%	92.88%	27.77%	0.17%	78.31%	9.33%	0.67%	44.36%	10.02%	0.14%
FEDDRIFT	58.28%	9.30%	0.51%	92.38%	28.36%	0.24%	78.46%	9.38%	0.83%	44.49%	9.27%	0.13%
FLASH	58.27%	6.98%	0.56%	93.92%	25.42%	0.11%	79.58%	8.14%	0.85%	45.65%	10.23%	0.26%

Table 5. Personalized accuracy (P), standard deviation of individual client's accuracy (I), and standard deviation of individual experiment run (E) for FLASH vs baselines, in a **no distribution change setting**. EM = EMNIST, SO = Stackoverflow, C10 = CIFAR10, C100 = CIFAR100.

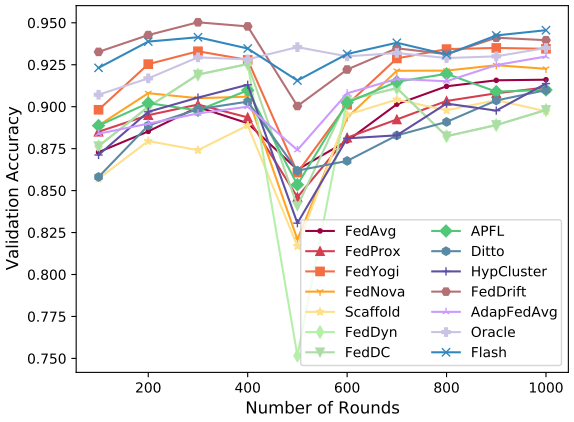
Tasks	EM (P)	(I)	(E)	SO (P)	(I)	(E)	C10 (P)	(I)	(E)	C100 (P)	(I)	(E)
"Non-iid"ness	Writers			Authors			$\alpha = 0.1$			$\alpha = 0.1$		
LOCAL ONLY	28.18%	18.56%	1.14%	15.93%	5.31%	0.25%	49.78%	16.65%	1.56%	36.19%	11.91%	0.43%
FEDAVG	83.06%	7.80%	0.95%	28.12 %	4.45%	0.12%	55.81%	17.77%	1.03/%	35.69%	10.41%	0.53%
FEDPROX	86.67%	4.75%	1.02%	28.28%	4.51%	0.16%	55.56%	17.99%	1.25%	35.39%	10.60%	0.47%
FEDYOGI	87.31%	6.68%	1.25%	29.42%	4.85%	0.21%	68.35%	15.44%	1.63%	39.84%	10.42%	0.36%
FEDNOVA	89.26%	4.52%	0.89%	29.06%	4.67%	0.15%	67.56%	14.67%	1.27%	37.26%	9.76%	0.34%
SCAFFOLD	87.51%	7.83%	1.08%	27.82%	4.31%	0.11%	65.84%	13.03%	1.78%	38.44%	10.53%	0.54%
FEDDYN	87.22%	7.92%	1.12%	26.00%	4.61%	0.18%	65.90%	12.85%	0.96%	39.90%	10.55%	0.49%
FEDDC	89.98%	6.91%	1.32%	27.14%	4.75%	0.22%	66.47%	9.58%	1.12%	41.72%	10.38%	0.61%
APFL	90.22%	7.19%	1.27%	27.34%	4.32%	0.13%	69.11%	15.84%	1.43%	40.66%	9.81%	0.41%
DITTO	88.11%	5.01%	0.93%	28.92%	4.55%	0.20%	69.28%	13.09%	1.11%	39.39%	10.56%	0.56%
HYPCLUSTER	89.04%	7.47%	1.18%	28.65%	4.50%	0.19%	70.17%	11.12%	1.52%	39.10%	10.49%	0.59%
ADAPFA	88.86%	6.22%	0.83%	29.01%	4.66%	0.24%	67.29%	10.34%	1.53%	39.25%	9.76%	0.53%
FEDDRIFT	89.68%	7.42%	0.92%	28.34%	4.19%	0.28%	70.45%	12.78%	1.67%	39.22%	10.62%	0.47%
FLASH	89.44 %	7.90%	1.04%	30.24%	4.36%	0.14%	70.07%	8.37%	1.21%	41.23%	11.57%	0.46%

Table 6. Personalized accuracy (P), standard deviation of individual client's accuracy (I), and standard deviation of individual experiment run (E) for FLASH vs baselines, in a **no distribution change setting**. SH = Shakespeare, SY = Synethtic, C10 = CIFAR10, C100 = CIFAR100.

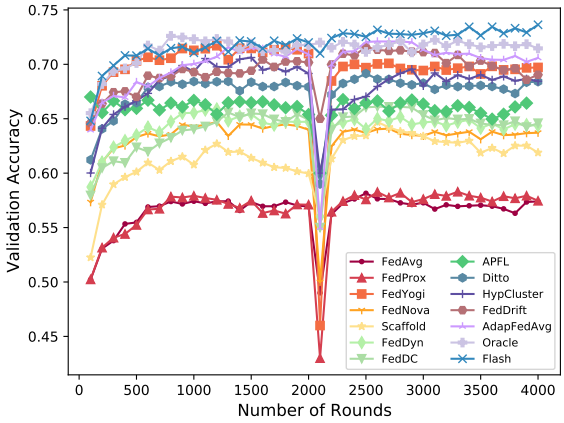
Tasks	SH (P)	(I)	(E)	SY (P)	(I)	(E)	C10 (P)	(I)	(E)	C100 (P)	(I)	(E)
"Non-iid"ness	Authors			$\alpha = 0.5$,	$\beta = 0.5$		$\alpha = 0.6$			$\alpha = 0.6$		
LOCAL ONLY	-	-	-	-	-	-	-	-	-	-	-	-
FEDAVG	57.35%	9.72%	0.35%	90.46%	24.07%	0.09%	75.53%	9.54%	0.78%	41.81%	10.52%	0.19%
FEDPROX	57.27%	10.48%	0.41%	92.76%	24.58%	0.09%	76.35%	9.00%	0.69%	42.76%	9.89%	0.36%
FEDYOGI	57.75%	10.59%	0.44%	93.20%	24.65%	0.16%	78.47%	9.02%	0.94%	41.71%	10.42%	0.24%
FEDNOVA	58.86%	9.43%	0.53%	93.57%	22.18%	0.23%	75.93%	8.41%	1.15%	46.12%	8.52%	0.53%
SCAFFOLD	57.65%	10.81%	0.53%	92.92%	28.64%	0.20%	75.84%	8.74%	0.86%	44.40%	9.80%	0.35%
FEDDYN	56.54%	11.57%	0.39%	91.89%	24.68%	0.14%	77.49%	9.35%	0.71%	45.57%	10.08%	0.14%
FEDDC	56.62%	11.55%	0.41%	93.10%	26.07%	0.15%	79.43%	9.83%	1.02%	45.06%	10.44%	0.35%
APFL	59.06%	7.04%	0.51%	91.78%	25.72%	0.08%	80.56%	8.72%	0.75%	46.87%	10.63%	0.29%
DITTO	60.05%	11.23%	0.43%	92.34%	15.79%	0.10%	79.86%	10.79%	0.83%	46.77%	10.58%	0.26%
HYPCLUSTER	59.84%	10.99%	0.60%	92.36%	25.15%	0.19%	79.76%	8.74%	0.74%	46.92%	9.91%	0.19%
ADAPFA	59.07%	11.40%	0.58%	92.88%	27.77%	0.17%	78.56%	9.33%	0.67%	43.86%	10.02%	0.14%
FEDDRIFT	59.18%	9.30%	0.51%	92.38%	28.36%	0.24%	78.82%	9.38%	0.83%	46.04%	9.27%	0.13%
FLASH	59.19%	6.98%	0.56%	93.92%	25.42%	0.11%	79.95%	8.14%	0.85%	45.90%	10.23%	0.26%



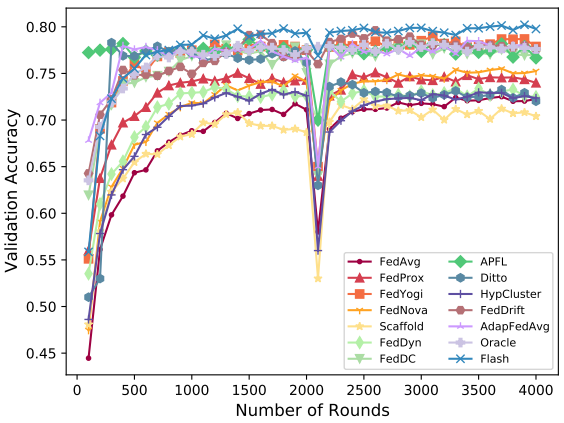
(a) EMNIST



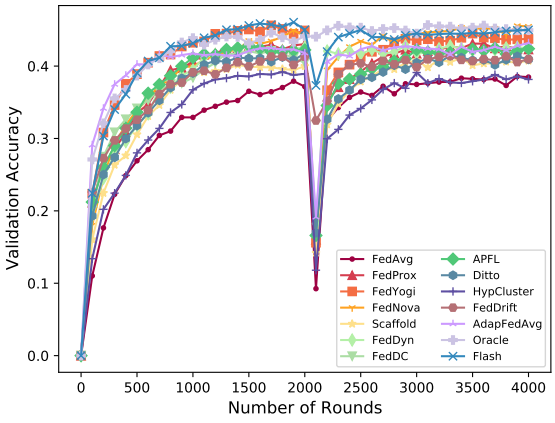
(b) Synthetic



(c) CIFAR 10 ($\alpha = 0.1$)

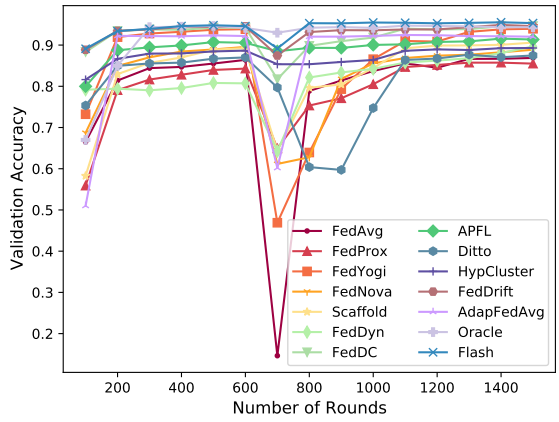


(d) CIFAR 10 ($\alpha = 0.6$)

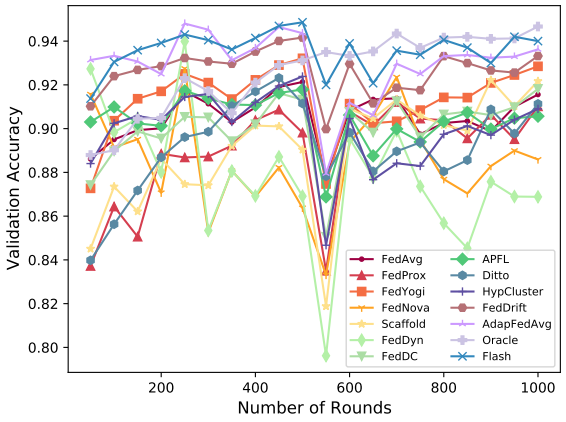


(e) CIFAR 100 ($\alpha = 0.6$)

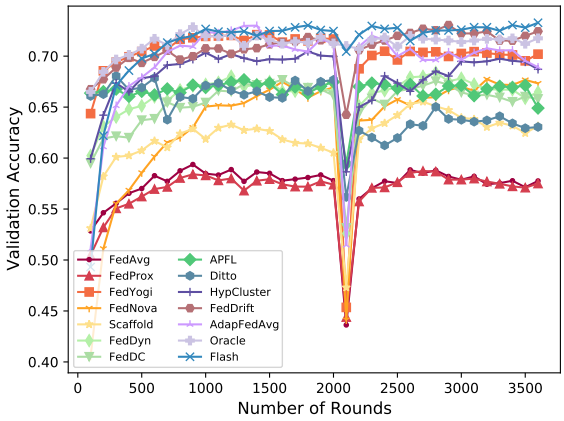
Figure 7. Accuracy curves for EMNIST, Synthetic, and CIFAR 10 / 100 datasets with sudden concept drift after steady state



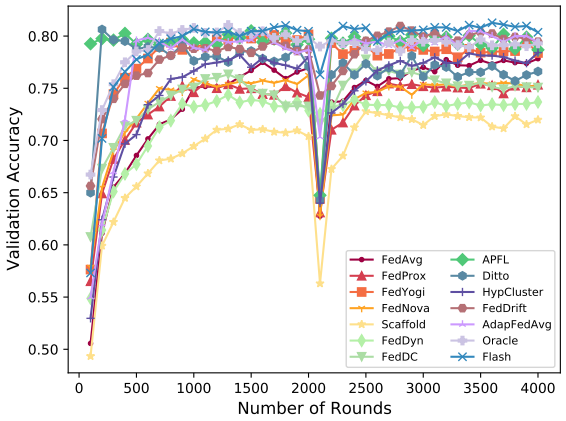
(a) EMNIST



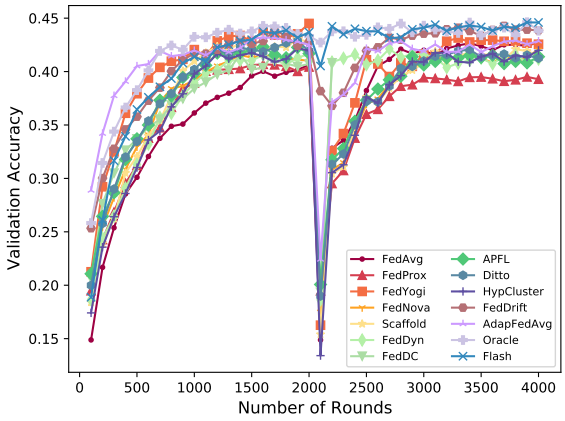
(b) Synthetic



(c) CIFAR 10 ($\alpha = 0.1$)

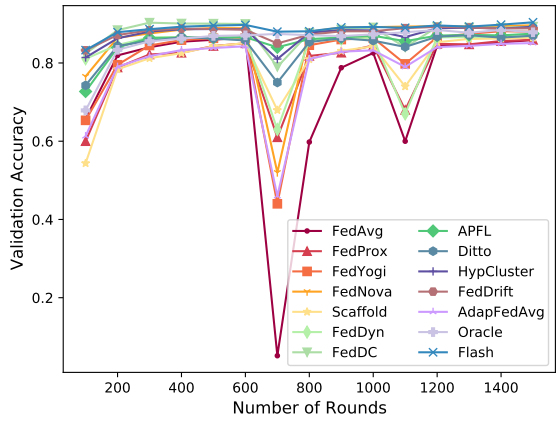


(d) CIFAR 10 ($\alpha = 0.6$)

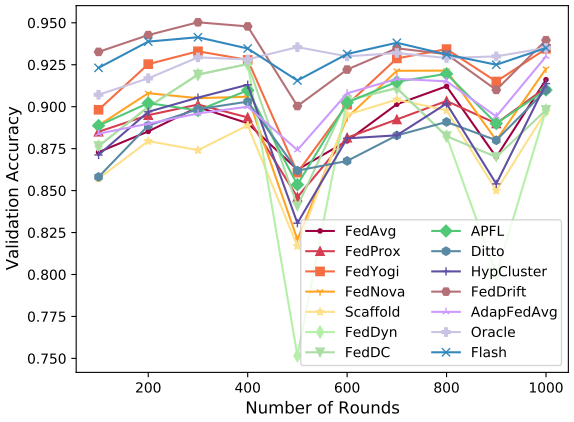


(e) CIFAR 100 ($\alpha = 0.6$)

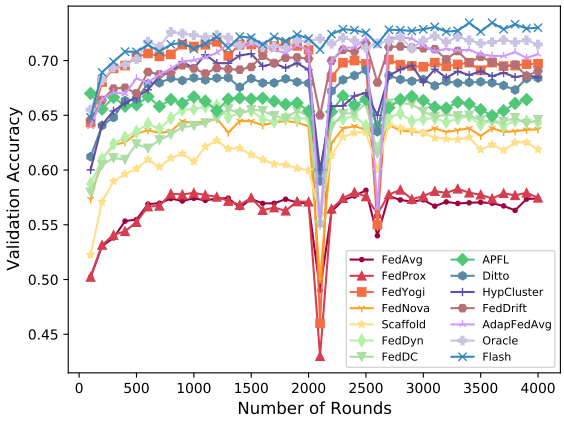
Figure 8. Accuracy curves for EMNIST, Synthetic, and CIFAR 10 / 100 datasets with incremental concept drift after steady state



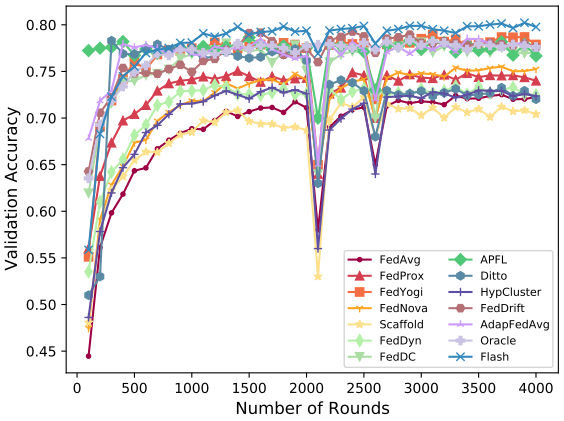
(a) EMNIST



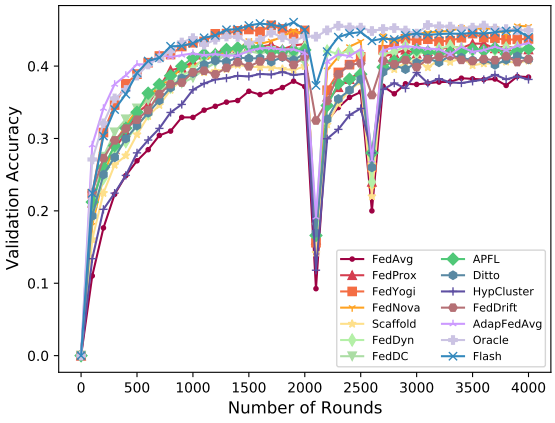
(b) Synthetic



(c) CIFAR 10 ($\alpha = 0.1$)



(d) CIFAR 10 ($\alpha = 0.6$)



(e) CIFAR 100 ($\alpha = 0.6$)

Figure 9. Accuracy curves for EMNIST, Synthetic, and CIFAR 10 / 100 datasets with recurrent concept drift after steady state

Table 7. Lowest accuracy during the concept drift (P), rounds till recovery to the steady state (R) for FLASH vs baselines for all the tasks, in a **sudden concept drift setting**. EM = EMNIST, C10 = CIFAR10, C100 = CIFAR100, SY = Synthetic.

Tasks	EM (P)	(R)	C10 (P)	(R)	C100 (P)	(R)	C10 (P)	(R)	C100 (P)	(R)	SY (P)	(R)
"Non-iid"ness	Writers		$\alpha = 0.1$		$\alpha = 0.1$		$\alpha = 0.6$		$\alpha = 0.6$		$\alpha = 0.5$,	$\beta = 0.5$
FEDAVG	5.11%	490	49.05%	240	13.05%	570	57.85%	380	9.25%	610	86.25%	160
FEDPROX	61.50%	390	43.10%	230	12.65%	530	63.60%	310	15.99%	510	84.61%	80
FEDYOGI	43.82%	370	46.42%	190	17.50%	480	65.40%	150	15.60%	550	86.07%	150
FEDNOVA	52.48%	370	50.12%	210	16.48%	500	64.18%	220	16.74%	520	82.09%	120
SCAFFOLD	58.51%	600	56.50%	240	18.73%	460	53.45%	200	14.20%	500	81.69%	80
FEDDYN	63.18%	360	56.22%	210	17.95%	380	69.65%	190	18.25%	430	75.16%	70
FEDDC	80.43%	260	55.38%	210	19.30%	360	64.30%	190	18.75%	410	84.07%	90
APFL	84.33%	270	59.70%	140	21.00%	320	69.95%	190	16.60%	520	85.35%	110
DITTO	75.05%	510	58.86%	150	15.21%	410	62.83%	220	16.96%	470	86.19%	80
HYPCLUSTER	81.32%	350	60.50%	150	14.16%	560	56.00%	450	11.80%	600	83.06%	50
ADAPFA	44.08%	270	54.08%	410	19.00%	230	65.95%	90	18.48%	90	87.04%	60
FEDDRIFT	82.31%	310	65.67%	100	34.22%	300	76.32%	130	31.07%	380	90.30%	90
ORACLE	90.28%	0	72.15%	0	35.75%	0	78.25%	0	44.06%	0	93.55%	0
FLASH	88.00%	210	70.99%	50	35.81%	180	76.90%	60	37.33%	160	91.56%	40

C. Analysis of FLASH

In Federated Learning, we solve an optimization problem of the form:

$$\min_{w_g \in \mathbb{R}^d} f(w_g) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F_c(w_g),$$

where $F_c(w_g) = \mathbb{E}_{(x,y)\sim\mathcal{D}_c}\left[f_c(w_g; x, y)\right]$ is the loss function of the c^{th} client, and $\mathcal{D}_c$ is the data for the c^{th} client. The functions F_c and therefore f may be non-convex. For each c and w_g, we assume access to an unbiased stochastic gradient $\nabla f_c(w_g)$ of the client's true gradient $\nabla F_c(w_g)$. In addition, we make the following assumptions.

Assumption C.1 (Lipschitz Gradient)**.** The function F_c is L-smooth for all $c \in \mathcal{C}$ i.e.,

$$||\nabla F_c(w) - \nabla F_c(z)|| \leq L||w - z|| \qquad \forall w, z \in \mathbb{R}^d.$$

Assumption C.2 (μ-convexity)**.** The function F_c is μ-convex for $\mu \geq 0$ and satisfies:

$$\langle \nabla F_c(w), z - w \rangle \leq -\left(F_c(w) - F_c(z) + \frac{\mu}{2}\|w - z\|^2\right) \qquad \forall\, c, w, z.$$

Assumption C.3 (Bounded Variance)**.** (Assumption 2 in (Reddi et al., 2021)) The function F_c for any $c \in \mathcal{C}$ has σ_ℓ-bounded local variance i.e., $\mathbb{E}\left[||\nabla[f_c(w; x, y)]_j - [\nabla F_c(w)]_j||^2\right] = \sigma_{c,j}^2 \leq \sigma_{\ell,j}^2$ for all $w \in \mathbb{R}^d, j \in [d]$, and $c \in \mathcal{C}$. Furthermore, we assume the global variance is bounded, $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} ||\nabla[F_c(w)]_j - [\nabla f(w)]_j||^2 \leq \sigma_{g,j}^2$ for all $w \in \mathbb{R}^d$ and $j \in [d]$.

Assumption C.4 (Bounded Local Gradients)**.** The function $f_c(w; x, y)$ have G-bounded gradients i.e., for any $c \in \mathcal{C}, w \in \mathbb{R}^d$, and $(x, y) \in \mathcal{D}_c^{(r)}$ we have $|[\nabla f_c(w; x, y)]_j| \leq G$ for all $j \in [d]$.

Assumption C.5 (Bounded Gradient Dissimilarity or (G, B)-BGD)**.** (Assumption 1 in (Karimireddy et al., 2020)) There exists constants such as $G \geq 0$ and $B \geq 1$ such that $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\|\nabla f_c(w_g)\|^2 \leq G^2 + B^2\|\nabla f(w_g)\|^2, \forall\ w_g$. If $\{F_c\}$ are convex, we can relax the assumption to $\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\|\nabla f_c(w_g)\|^2 \leq G^2 + 2LB^2(f(w_g) - f(w_g^\star)), \forall\ w_g$.

Lemma C.6 (One epoch progress of client-side SGD)**.** *Suppose $\{F_c\}$ satisfies Assumptions C.1, C.2, C.3, and C.4. For a local step-size η_ℓ, the updates of SGD satisfy*

$$\mathbb{E}_e\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2 \leq \left(1 - \frac{\mu\eta_\ell}{2}\right)\mathbb{E}_{e-1}\left\|w_{c,e-1}^{(r)} - w_c^\star\right\|^2 + 2\eta_\ell\left(f_c(w_c^\star) - \mathbb{E}_{e-1}[f_c(w_{c,e-1}^{(r)})]\right) + 2L\eta_\ell^3 G^2 + 2\eta_\ell^4 L^2 G^2 + 2\eta_\ell^2 G^2 + \eta_\ell^2\sigma_c^2$$

Proof. We start with restating the SGD update

$$w_{c,e}^{(r)} \leftarrow w_{c,e-1}^{(r)} - \eta_\ell \nabla f_c(w_{c,e-1}^{(r)})$$

which gives

$$\Delta w_{c,e}^{(r)} = -\eta_\ell \nabla f_c(w_{c,e-1}^{(r)}) \implies \mathbb{E}_{e-1}\left[\Delta w_{c,e}^{(r)}\right] = -\eta_\ell \mathbb{E}_{e-1}\left[\nabla f_c(w_{c,e-1}^{(r)})\right]$$

Using the above update, we proceed as,

$$\mathbb{E}_{e-1}\left\|w_{c,e}^{(r)} + \Delta w_{c,e}^{(r)} - w_c^\star\right\|^2 = \mathbb{E}_{e-1}\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2 - 2\eta_\ell\langle\nabla f_c(w_{c,e-1}^{(r)}), w_{c,e}^{(r)} - w_c^\star\rangle + \eta_\ell^2\mathbb{E}_{e-1}\left\|\nabla f_c(w_{c,e-1}^{(r)})\right\|^2 \tag{1}$$

$$\leq \mathbb{E}_{e-1}\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2 \underbrace{-2\eta_\ell\langle\nabla f_c(w_{c,e-1}^{(r)}), w_{c,e}^{(r)} - w_c^\star\rangle}_{T_1} + \underbrace{\eta_\ell^2\mathbb{E}_{e-1}\left\|\nabla F_c(w_{c,e-1}^{(r)})\right\|^2}_{T_2} + \eta_\ell^2\sigma_c^2 \tag{2}$$

(Separating variance and mean according to Lemma 4 in (Karimireddy et al., 2020))

Bounding T_1

$$T_1 = 2\eta_\ell\langle\nabla f_c(w_{c,e-1}^{(r)}), w_c^\star - w_{c,e}^{(r)}\rangle \tag{3}$$

$$\leq 2\eta_\ell\left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) + L\left\|w_{c,e}^{(r)} - w_{c,e-1}^{(r)}\right\|^2 - \frac{\mu}{4}\left\|w_c^\star - w_{c,e}^{(r)}\right\|^2\right) \text{ from Lemma 5 of (Karimireddy et al., 2020)} \tag{4}$$

$$= 2\eta_\ell\left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) - \frac{\mu}{4}\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2\right) + 2L\eta_\ell\left\|-\eta_\ell\nabla f_c(w_{c,e-1}^{(r)})\right\|^2 \tag{5}$$

$$= 2\eta_\ell\left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) - \frac{\mu}{4}\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2\right) + 2L\eta_\ell^3\left\|\nabla f_c(w_{c,e-1}^{(r)})\right\|^2 \tag{6}$$

$$\leq 2\eta_\ell\left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) - \frac{\mu}{4}\left\|w_{c,e}^{(r)} - w_c^\star\right\|^2\right) + 2L\eta_\ell^3 G^2 \tag{7}$$

Bounding T_2

$$T_2 = \eta_\ell^2 \mathbb{E}_{e-1} \left\| \nabla F_c(w_{c,e-1}^{(r)}) - \nabla F_c(w_{c,e}^{(r)}) + \nabla F_c(w_{c,e}^{(r)}) \right\|^2 \tag{8}$$

$$\leq 2\eta_\ell^2 \mathbb{E}_{e-1} \left\| \nabla F_c(w_{c,e-1}^{(r)}) - \nabla F_c(w_{c,e}^{(r)}) \right\|^2 + 2\eta_\ell^2 \mathbb{E}_{e-1} \left\| \nabla F_c(w_{c,e}^{(r)}) \right\|^2 \tag{9}$$

$$\leq 2\eta_\ell^2 L^2 \mathbb{E}_{e-1} \left\| w_{c,e-1}^{(r)} - w_{c,e}^{(r)} \right\|^2 + 2\eta_\ell^2 \left\| \nabla F_c(w_{c,e}^{(r)}) \right\|^2 \tag{10}$$

$$\leq 2\eta_\ell^2 L^2 \left(\mathbb{E}_{e-1} \left\| \eta_\ell \nabla f_c(w_{c,e-1}^{(r)}) \right\|^2 \right) + 2\eta_\ell^2 G^2 \tag{11}$$

$$\leq 2\eta_\ell^4 L^2 G^2 + 2\eta_\ell^2 G^2 = 2\eta_\ell^2 G^2 (\eta_\ell^2 L^2 + 1) \tag{12}$$

The inequality (9) comes from the relaxed triangular inequality (Lemma 3 in (Karimireddy et al., 2020)). Inequalities (10) and (12) are respectively implied by Assumptions C.1 and C.4.

Plugging in the bounds for T_1 and T_2 in Equation 2,

$$\mathbb{E}_{e-1} \left\| w_{c,e}^{(r)} + \Delta w_{c,e}^{(r)} - w_c^\star \right\|^2 \leq \mathbb{E}_{e-1} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 \underbrace{- 2\eta_\ell \langle \nabla f_c(w_{c,e-1}^{(r)}), w_{c,e}^{(r)} - w_c^\star \rangle}_{T_1} + \underbrace{\eta_\ell^2 \mathbb{E} \left\| \nabla F_c(w_{c,e-1}^{(r)}) \right\|}_{T_2} + \eta_\ell^2 \sigma_c^2 \tag{13}$$

$$\leq \mathbb{E}_{e-1} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 + 2\eta_\ell \left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) - \frac{\mu}{4} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 \right) + 2L\eta_\ell^3 G^2 + 2\eta_\ell^4 L^2 G^2 + 2\eta_\ell^2 G^2 + \eta_\ell^2 \sigma_c^2 \tag{14}$$

$$= \left(1 - \frac{\mu \eta_\ell}{2} \right) \mathbb{E}_{e-1} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 + 2\eta_\ell \left(f_c(w_c^\star) - f_c(w_{c,e}^{(r)}) \right) + 2L\eta_\ell^3 G^2 + 2\eta_\ell^4 L^2 G^2 + 2\eta_\ell^2 G^2 + \eta_\ell^2 \sigma_c^2 \tag{15}$$

□

Theorem C.7 (Convergence of Early-stopping SGD)**.** *For functions $\{F_c\}$ which satisfy Assumptions C.1, C.2, C.3, and C.4, the output of the early-stopping SGD with early stopping criteria $\sum_{x,y} f_c(w_{c,e-1}^{(r)}; x, y) - \sum_{x,y} f_c(w_{c,e}^{(r)}; x, y) \geq \gamma/e, \ \forall\ e \in [E]$ and $\forall\ (x, y) \in \mathcal{D}_{c,valid}^{(r)}$ has expected error smaller than ϵ for $\gamma \geq \frac{(F-\epsilon)}{\ln E + \frac{1}{E}}$ and some values of η_ℓ, e_c satisfying*

- *Strongly convex, $\frac{1}{\mu E} \leq \eta_\ell \leq \frac{\log(\max(1, \mu^2 E D/c))}{\mu E}$, and $e_c = \mathcal{O}\left(\min\left(\frac{\mu D^2}{\epsilon} + \frac{G^2}{\mu\epsilon} + \frac{\sigma_c^2}{2\mu\epsilon} + \frac{LG^2}{\mu^2\epsilon} + \frac{L^2G^2}{\mu^3\epsilon} \right), \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E}) \right)$*
- *General convex, $\frac{1}{E} \leq \eta_\ell$, and $e_c = \mathcal{O}\left(\min\left(D^2 + \frac{G^2D^2}{\epsilon^2} + \frac{\sigma_c^2 D^2}{2\epsilon^2} + \frac{\sqrt{L}GD^2}{\epsilon^{3/2}} + \frac{(L^2G^2)^{1/3}D^2}{\epsilon^{4/3}} \right), \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E}) \right)$*
- *Non-convex, $\frac{1}{E} \leq \eta_\ell$, and $e_c = \mathcal{O}\left(\min\left(F + \frac{G^2F}{\epsilon^2} + \frac{\sigma_c^2 LF^2}{2\epsilon^2} + \frac{\sqrt{L}GF}{\epsilon^{3/2}} + \frac{(L^2G)^{1/3}F}{\epsilon^{4/3}} \right), \exp(\frac{F-\epsilon}{\gamma} - \frac{1}{E}) \right)$*

where $c := G^2 + \frac{\mu_c^2}{2}$, $D := \mathbb{E} \left\| w_{c,0}^{(r)} - w_c^\star \right\|$, and $F = f_c(w_{c,0}^{(r)}) - f_c(w_c^\star)$.

Proof. Using the Lemma C.6 statement, and moving $f_c(w_c^\star) - f_c(w_{c,e}^{(r)})$ to LHS and rearranging the terms on RHS,

$$\mathbb{E}\left[f_c(w_{c,e-1}^{(r)}) \right] - f_c(w_c^\star) \leq \frac{1}{2\eta_\ell} \left(1 - \frac{\mu\eta_\ell}{2} \right) \mathbb{E} \left\| w_{c,e-1}^{(r)} - w_c^\star \right\|^2 - \frac{1}{2\eta_\ell} \mathbb{E} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 + \frac{\eta_\ell^2}{2\eta_\ell} (2L\eta_\ell G^2 + 2\eta_\ell^2 L^2 G^2 + 2G^2 + \sigma_c^2) \tag{16}$$

$$= \frac{1}{2\eta_\ell} \left(1 - \frac{\mu\eta_\ell}{2} \right) \mathbb{E} \left\| w_{c,e-1}^{(r)} - w_c^\star \right\|^2 - \frac{1}{2\eta_\ell} \mathbb{E} \left\| w_{c,e}^{(r)} - w_c^\star \right\|^2 + \eta_\ell (G^2 + \frac{\sigma_c^2}{2}) + \eta_\ell^2 (LG^2) + \eta_\ell^3 (L^2 G^2) \tag{17}$$

Note that the indices are off by 1 with respect to Lemma C.6 to be consistent with the theorem statement.

In case of $\mu = 0$ (general convex case), we apply the sublinear convergence rate on a non-negative sequence (refer to Lemma 2 in (Karimireddy et al., 2020)) and the condition $\frac{1}{E} \leq \eta_\ell$,

$$\mathbb{E}\left[f_c(w_{c,e}^{(r)}) \right] - f_c(w_c^\star) \leq \frac{D^2}{2\eta_\ell E} + \left(G^2 + \frac{\sigma_c^2}{2} \right) \eta_\ell + (LG^2)\eta_\ell^2 + (LG)\eta_\ell^3 \tag{18}$$

$$\leq \frac{D^2}{2} + \left(G^2 + \frac{\sigma_c^2}{2} \right)^{\frac{1}{2}} \left(\frac{D^2}{E} \right)^{\frac{1}{2}} + (LG^2)^{\frac{1}{3}} \left(\frac{D^2}{E} \right)^{\frac{2}{3}} + (LG)^{\frac{1}{2}} \left(\frac{D^2}{E} \right)^{\frac{3}{4}} \tag{19}$$

For the μ-convex case, similar to the usage of Lemma 1 in (Karimireddy et al., 2020), we apply the linear convergence rate with the condition $\eta_\ell \geq \frac{1}{\mu E}$ and get,

$$\mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - f_c(w_c^\star) \leq \frac{3}{2}D^2\mu \exp\left(-\frac{\eta_\ell \mu E}{2}\right) + \left(G^2 + \frac{\sigma_c^2}{2}\right)\eta_\ell + (LG^2)\eta_\ell^2 + (L^2G^2)\eta_\ell^3 \tag{20}$$

$$\leq \frac{3D^2\mu}{2} + \left(G^2 + \frac{\sigma_c^2}{2}\right)\frac{1}{\mu E} + (LG^2)\frac{1}{\mu^2E^2} + (L^2G^2)\frac{1}{\mu^3E^3} \tag{21}$$

Rearranging the terms and assigning the error as $\mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - f_c(w_c^\star) = \epsilon$ get us the bounds shown in the theorem statement.

For the bound on early stopping parameter γ, we get the lower bound of $\mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - f_c(w_c^\star)$ as follows,

$$\mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - f_c(w_c^\star) = \mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - \mathbb{E}\left[f_c(w_{c,e-1}^{(r)})\right] + \mathbb{E}\left[f_c(w_{c,e-1}^{(r)})\right] - \mathbb{E}\left[f_c(w_c^{(E-2)})\right] + \mathbb{E}\left[f_c(w_c^{(E-2)})\right]$$

$$\cdots - \mathbb{E}\left[f_c(w_c^{(0)})\right] + \mathbb{E}\left[f_c(w_c^{(0)})\right] - f_c(w_c^\star) \tag{22}$$

$$\therefore \epsilon \geq -\gamma\sum_{i=1}^{E}\frac{1}{i} + \mathbb{E}\left[f_c(w_c^{(0)})\right] - f_c(w_c^\star) \tag{23}$$

$$\geq -\gamma\left(\ln E + \frac{1}{E}\right) + \mathbb{E}\left[f_c(w_c^{(0)})\right] - f_c(w_c^\star) \tag{24}$$

$$\therefore \gamma \geq \frac{(\mathbb{E}f_c(w_c^{(0)}) - \mathbb{E}f_c(w_{c,e}^{(r)}))}{\ln E + \frac{1}{E}} = \frac{(F-\epsilon)}{\ln E + \frac{1}{E}} \tag{25}$$

where $F = \mathbb{E}\left[f_c(w_c^{(0)})\right] - f_c(w_c^\star)$ and $\epsilon = \mathbb{E}\left[f_c(w_{c,e}^{(r)})\right] - f_c(w_c^\star)$. □

Lemma C.8 (Bounding the Client Drift wrt Current Round)**.** *Let $\{F_c\}$ satisfy Assumptions C.1, C.3, and C.5. For any step size satisfying $\eta_\ell \leq \sqrt{\frac{1}{L^2E(E-1)}}$, we can bound the drift for $E = \max(\{E_c \mid c \in \mathcal{C}\})$ where E_c is the number of epochs client c trains $w_{c,0}^{(r)} := w_g^{(r-1)}$ for, as*

$$\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mathbb{E}\left\|w_{c,e}^{(r)} - w_g^{(r)}\right\|^2 \leq 5E\eta_\ell^2\mathbb{E}[\sigma_\ell^2 + 6E\sigma_g^2] + 30E^2\eta_\ell^2\mathbb{E}\left\|\nabla f(w_g^{(r)})\right\|^2$$

Proof. We have followed the same proof technique as in Lemma 3 of (Reddi et al., 2021). □

Lemma C.9 (Upper Bounding the Effective Gradients)**.** *Let $\{F_c\}$ satisfy Assumptions C.1, C.3, and C.5. In round r, the updates in* FLASH *and* FEDYOGI *satisfy,*

$$\mathbb{E}_r\left\|(\Delta^{(r)})^2\right\| \leq \frac{6\eta_\ell^2E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2EG^2 + \frac{30\eta_\ell^4L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2 \tag{26}$$

Proof.

$$\mathbb{E}_r\left\|(\Delta^{(r)})^2\right\| \leq \mathbb{E}_r\left\|\Delta^{(r)} + \eta_\ell E\nabla f(w_g^{(r)}) - \eta_\ell E\nabla f(w_g^{(r)})\right\|^2 \tag{27}$$

$$\leq 2\mathbb{E}_r\left\|\Delta^{(r)} + \eta_\ell E\nabla f(w_g^{(r)})\right\|^2 + 2\mathbb{E}_r\left\|\eta_\ell E\nabla f(w_g^{(r)})\right\|^2 \tag{28}$$

$$= 2\mathbb{E}_r\left\|-\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{e=0}^{E-1}\eta_\ell\nabla f_c(w_{c,e}^{(r)}) + \eta_\ell E\nabla f(w_g^{(r)})\right\|^2 + 2\mathbb{E}_r\left\|\eta_\ell E\nabla f(w_g^{(r)})\right\|^2 \tag{29}$$

$$= 2\mathbb{E}_r\left\|-\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{e=0}^{E-1}(\eta_\ell\nabla f_c(w_{c,e}^{(r)}) - \eta_\ell\nabla F_c(w_{c,e}^{(r)}) + \eta_\ell\nabla F_c(w_{c,e}^{(r)}) - \eta_\ell\nabla F_c(w_g^{(r)}) + \eta_\ell\nabla F_c(w_g^{(r)})) + \eta_\ell E\nabla f(w_g^{(r)})\right\|^2$$

$$+ 2\mathbb{E}_r\left\|\eta_\ell E\nabla f(w_g^{(r)})\right\|^2 \tag{30}$$

$$= 2\eta_\ell^2 \mathbb{E}_r \left\| -\frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E-1} (\nabla f_c(w_{c,e}^{(r)}) - \nabla F_c(w_{c,e}^{(r)}) + \nabla F_c(w_{c,e}^{(r)}) - \nabla F_c(w_g^{(r)})) - \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E_c} \nabla F_c(w_g^{(r)}) + E\nabla f(w_g^{(r)}) \right\|^2$$
$$+ 2E^2\eta_\ell^2 \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \tag{31}$$

$$\leq 6\eta_\ell^2 \mathbb{E}_r \left\| \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E_c} (\nabla f_c(w_{c,e}^{(r)}) - \nabla F_c(w_{c,e}^{(r)})) \right\|^2 + 6\eta_\ell^2 \mathbb{E}_r \left\| \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E_c} (\nabla F_c(w_{c,e}^{(r)}) - \nabla F_c(w_g^{(r)})) \right\|^2$$
$$+ 6\eta^2 E \left(G^2 + (B^2-1)\mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \right) + 2E^2\eta_\ell^2 \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \tag{32}$$

$$\leq \frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + \frac{6\eta_\ell^2}{|\mathcal{C}|^2} \mathbb{E}_r \left\| \sum_{c\in\mathcal{C}} \sum_{e=0}^{E_c} L(w_{c,e}^{(r)} - w_g^{(r)}) \right\|^2 + (2E^2\eta_\ell^2 + 6\eta_\ell^2 E(B^2-1)) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 + 6\eta_\ell^2 EG^2 \tag{33}$$

$$\leq \frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + \frac{6\eta_\ell^2 L^2 E^2}{|\mathcal{C}|} \left(5E\eta_\ell^2(\sigma_\ell^2 + 6E\sigma_g^2) + 30E^2\eta_\ell^2 \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \right)$$
$$+ (2E^2\eta_\ell^2 + 6\eta_\ell^2 E(B^2-1)) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 + 6\eta_\ell^2 EG^2 \tag{34}$$

$$\leq \frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4 L^2 E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2 E(B^2-1) \right) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \tag{35}$$

The second to last inequality follows from Lemma C.8. □

Lemma C.10 (Upper Bounding the Rolling Average of Effective Gradients)**.** *Let $\{F_c\}$ satisfy Assumptions C.1, C.3, and C.5. In round r, the updates in* FLASH *and* FEDYOGI *satisfy,*

$$\mathbb{E}_r \left[\left\| v^{(r)} \right\| \right] \leq \eta_\ell^2 E^2 G^2 (1-\beta_2^{r-1}) + (2-\beta_2) \left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \left(\sigma_\ell^2 + 6E\sigma_g^2\right) \right)$$
$$+ (2-\beta_2) \left(\left(\frac{180\eta_\ell^4 L^2 E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2 E(B^2-1) \right) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \right) \tag{36}$$

Proof. First we recall that $v^{(r)} = \beta_2 v^{(r-1)} + (1-\beta_2)(\Delta^{(r)})^2$, similar to FEDADAM/FEDYOGI updates in (Reddi et al., 2021). Unrolling the recursion, we get $v^{(r)} = (1-\beta_2)\sum_{i=1}^{r} \beta_2^{r-i} (\Delta^{(i)})^2$.

Replacing $v^{(r)}$ with its unrolled version,

$$\mathbb{E}_r \left[\left\| v^{(r)} \right\| \right] = (1-\beta_2) \mathbb{E}_r \left[\left\| \sum_{i=1}^{r} \beta_2^{r-i} (\Delta^{(i)})^2 \right\| \right] \tag{37}$$
$$\leq (1-\beta_2) \left[\mathbb{E}_r \left\| \sum_{i=1}^{r-1} \beta_2^{r-1-i} (\Delta^{(i)})^2 \right\| + \mathbb{E}_r \left\| \beta_2^0 (\Delta^{(r)})^2 \right\| \right] \tag{38}$$
$$= (1-\beta_2) \left[\left\| \sum_{i=1}^{r-1} \beta_2^{r-1-i} (\Delta^{(i)})^2 \right\| + \mathbb{E}_r \left\| (\Delta^{(r)})^2 \right\| \right] \tag{39}$$
$$\leq (1-\beta_2) \sum_{i=1}^{r-1} \beta_2^{r-1-i} \left\| (\Delta^{(i)})^2 \right\| + (1-\beta_2) \underbrace{\mathbb{E}_r \left\| (\Delta^{(r)})^2 \right\|}_{T_1} \tag{40}$$

Plugging in the bounds for T_1, from Lemma C.9,

$$\mathbb{E}_r\left[\left\|v^{(r)}\right\|\right] \leq (1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(\Delta^{(i)})^2\right\| + (2-\beta_2)\mathbb{E}_r\left\|(\Delta^{(r)})^2\right\| \tag{41}$$

$$\leq (1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(\Delta^{(i)})^2\right\| + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{42}$$

$$\leq (1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|\left(-\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{e=0}^{E}\eta_\ell\nabla f_c(w_{c,e}^{(i)})\right)^2\right\| + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{43}$$

$$\leq \frac{(1-\beta_2)\eta_\ell^2}{|\mathcal{C}|^2}\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|\left(\sum_{c\in\mathcal{C}}\sum_{e=0}^{E}\nabla f_c(w_{c,e}^{(i)})\right)^2\right\| + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{44}$$

$$\leq \frac{(1-\beta_2)\eta_\ell^2}{|\mathcal{C}|^2}\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(|\mathcal{C}|EG)^2\right\| + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{45}$$

$$\leq \frac{(1-\beta_2)\eta_\ell^2|\mathcal{C}|^2E^2G^2}{|\mathcal{C}|^2}\cdot\frac{(1-\beta_2^{r-1})}{(1-\beta_2)} + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{46}$$

$$\leq \eta_\ell^2E^2G^2(1-\beta_2^{r-1}) + (2-\beta_2)\left(\frac{6\eta_\ell^2 E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2 EG^2 + \frac{30\eta_\ell^4 L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2 + 6E\sigma_g^2\right) + \left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right) \tag{47}$$

□

Theorem C.11 (Lower Bounding the Change in the Second Moment of Effective Gradients)**.** *Let $\{F_c\}$ satisfy Assumptions C.1, C.3, and C.5. In round r, the updates in* FLASH *and* FEDYOGI *satisfy,*

$$\begin{aligned}\textsc{FedYogi}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) - \textsc{Flash}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) \\ \geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r}+\tau\right)}\right\| - \eta_\ell^2E^2G^2(1-\beta_2^{r-1})\right| \\ - \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} - \eta_\ell EG(1-\beta_3^r)+\tau\right)}\right\| - \eta_\ell^2E^2G^2(1-\beta_2^{r-1})\right|\end{aligned} \tag{48}$$

Proof. Using Jensen's Inequality, we have

$$\mathbb{E}\left[||a-b||\right] \geq |\mathbb{E}||a|| - \mathbb{E}||b|||$$

Hence we get,

$$\mathbb{E}_r\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right] \geq \left|\mathbb{E}_r||(\Delta^{(r)})^2|| - \mathbb{E}_r||v^{(r)}||\right| \tag{49}$$

$$\geq \left|\mathbb{E}_r||(\Delta^{(r)})^2|| - \left((1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(\Delta^{(i)})^2\right\| + (1-\beta_2)\mathbb{E}_r\left\|(\Delta^{(r)})^2\right\|\right)\right| \tag{50}$$

$$\geq \left|(1-1+\beta_2)\mathbb{E}_r||(\Delta^{(r)})^2|| - \left((1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(\Delta^{(i)})^2\right\|\right)\right| \tag{51}$$

$$\geq \left|(1-1+\beta_2)\mathbb{E}_r||(\Delta^{(r)})^2|| - \left((1-\beta_2)\sum_{i=1}^{r-1}\beta_2^{r-1-i}\left\|(\Delta^{(i)})^2\right\|\right)\right| \tag{52}$$

$$\geq \left|\beta_2\mathbb{E}_r||(\Delta^{(r)})^2|| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \tag{53}$$

The second and the last inequalities both are based on the derivation shown in Lemma C.10.

For FEDYOGI, we know that $w_g^{(r)} = w_g^{(r-1)} + \eta_g \frac{\Delta^{(r)}}{\sqrt{v^{(r)}}+\tau}$, hence we get lower bound of $\mathbb{E}_r\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]$ as,

$$\textsc{FedYogi}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) \geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r}+\tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \tag{54}$$

and For FLASH, we know that $w_g^{(r)} = w_g^{(r-1)} + \eta_g \frac{\Delta^{(r)}}{\sqrt{v^{(r)}-d^{(r)}}+\tau}$, hence we get lower bound of $\mathbb{E}_r\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]$ as,

$$\textsc{Flash}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) \geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} - d^{(r)}+\tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \tag{55}$$

$$\geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} - \eta_\ell EG(1-\beta_3^r)+\tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \tag{56}$$

Second inequality follows from the fact that $-d^{(r)} \geq -\eta_\ell EG(1-\beta_3^r)$. Note that $\eta_\ell EG(1-\beta_3^r)$ would always be positive.

Therefore, we get

$$\textsc{FedYogi}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right) - \textsc{Flash}\left(\mathbb{E}\left[\left\|(\Delta^{(r)})^2 - v^{(r)}\right\|\right]\right)$$

$$\geq \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r}+\tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right|$$

$$- \left|\frac{\beta_2}{\sqrt{\eta_g}}\mathbb{E}_r\left\|\sqrt{\left(w_g^{(r)} - w_g^{(r-1)}\right)\left(\sqrt{v^r} - \eta_\ell EG(1-\beta_3^r)+\tau\right)}\right\| - \eta_\ell^2 E^2 G^2(1-\beta_2^{r-1})\right| \tag{57}$$

□

Theorem C.12 (Convergence of FLASH)**.** *Let assumptions C.1 to C.4 hold. Suppose the server and client learning rates satisfy*

$$\eta_\ell \leq \min\left[\left(\frac{|\mathcal{C}|}{30L^2E}\right)^{\frac{1}{2}}, \left(\frac{\tau}{6(B^2-1)\left[G(\beta_2+\sqrt{\beta_2})+L\eta_g\right]}\right)\right].$$

Then the iterates of Algorithm 1 for $\eta_\ell = \Theta(1/L\sqrt{E})$, $\eta_g = \Theta(1/\sqrt{R})$, and $\tau = G/L$ for FLASH *satisfy*

$$\min_{0\leq r\leq R}\mathbb{E}\left\|\nabla f(w_g^{(r)})\right\|^2 \leq \mathcal{O}\left(\frac{f(w_g^{(0)}) - \mathbb{E}_r[f(w_g^{(R)})]}{\sqrt{ER}} + \frac{G}{\sqrt{ER}|\mathcal{C}|}(\sigma_\ell^2 + 6E\sigma_g^2) + \frac{6L\sigma_\ell^2}{RG^2|\mathcal{C}|} + \frac{6L}{R}\right)$$

Proof. The proof strategy is similar to that of FEDADAM as described in (Reddi et al., 2021), except now we also have to manage the addition of moving average of the difference between $(\Delta^{(r)})^2$ and $v^{(r)}$.

We note that FLASH has the following update rule (see Algorithm 1 Line 22)

$$w_g^{(r+1)} \leftarrow w_g^{(r)} + \eta_g \frac{\Delta^{(r)}}{\sqrt{v^{(r)}} - d^{(r)} + \tau}.$$

Similar to the analysis of FEDADAM, we are assuming $\beta_1 = 0$. Hence, $m^{(r)} = \Delta^{(r)}$.
Note that $d_j^{(r)} \leftarrow \beta_{3j} d_j^{(r-1)} + (1-\beta_{3j})((\Delta_j^{(r)})^2 - v_j^{(r)})$ where $\beta_{3j} = \frac{||v_j^{(r-1)}||_2}{||(\Delta_j^{(r)})^2 - v_j^{(r)}||_2 + ||v_j^{(r-1)}||_2}$ for all $j \in [d]$.

Using the L-smooth nature of the function f and the above update rule, we have the following,

$$f(w_g^{(r+1)}) \leq f(w_g^{(r)}) + \left\langle \nabla f(w_g^{(r)}), w_g^{(r+1)} - w_g^{(r)} \right\rangle + \frac{L}{2} \left\| w_g^{(r+1)} - w_g^{(r)} \right\|^2 \tag{58}$$

$$= f(w_g^{(r)}) + \eta_g \sum_{j=1}^{d} \left([\nabla f(w_g^{(r)})]_j \times \frac{\Delta_j^{(r)}}{\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau} \right) + \frac{L\eta_g^2}{2} \sum_{j=1}^{d} \frac{(\Delta_j^{(r)})^2}{\left(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau\right)^2} \tag{59}$$

The second step follows from the update rule of FLASH stated initially.

Now we take expectation of $f(w_g^{r+1})$ (over randomness at round r) and rewrite the above inequality as,

$$\mathbb{E}_r[f(w_g^{(r+1)})] \leq f(w_g^{(r)}) + \eta_g \left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r \left[\frac{\Delta^{(r)}}{\sqrt{v^{(r)}} - d^{(r)} + \tau} - \frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} + \frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} \right] \right\rangle$$
$$+ \frac{L\eta_g^2}{2} \mathbb{E}_r \left[\frac{(\Delta^{(r)})^2}{\left(\sqrt{v^{(r)}} - d^{(r)} + \tau\right)^2} \right] \tag{60}$$

$$= f(w_g^{(r)}) + \eta_g \underbrace{\left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r \left[\frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} \right] \right\rangle}_{T_1}$$
$$+ \eta_g \underbrace{\left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r \left[\frac{\Delta^{(r)}}{\sqrt{v^{(r)}} - d^{(r)} + \tau} - \frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} \right] \right\rangle}_{T_2} + \frac{L\eta_g^2}{2} \mathbb{E}_r \left[\frac{(\Delta^{(r)})^2}{\left(\sqrt{v^{(r)}} - d^{(r)} + \tau\right)^2} \right] \tag{61}$$

Bounding T_1

$$T_1 = \left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r \left[\frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} \right] \right\rangle \tag{62}$$

$$= \left\langle \frac{\nabla f(w_g^{(r)})}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau}, \mathbb{E}_r \left[\Delta^{(r)} - \eta_\ell E \nabla f(w_g^{(r)}) + \eta_\ell E \nabla f(w_g^{(r)}) \right] \right\rangle \tag{63}$$

$$= \frac{-\eta_\ell E [\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} + \underbrace{\left\langle \frac{\nabla f(w_g^{(r)})}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau}, \mathbb{E}_r \left[\Delta^{(r)} + \eta_\ell E \nabla f(w_g^{(r)}) \right] \right\rangle}_{T_3} \tag{64}$$

Bounding T_3

$$T_3 = \left\langle \frac{\nabla f(w_g^{(r)})}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}}, \mathbb{E}_r \left[-\frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E} \eta_\ell \nabla f_c(w_{c,e}^{(r)}) + \eta_\ell E \nabla f(w_g^{(r)}) \right] \right\rangle \tag{65}$$

$$\leq \frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{\eta_\ell}{2} \mathbb{E}_r \left[\left\| \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E} \frac{\nabla F_c(w_{c,e}^{(r)})}{\sqrt{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau}} - \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \sum_{e=0}^{E} \frac{\nabla F_c(w_g^{(r)})}{\sqrt{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau}} \right\|^2 \right] \tag{66}$$

$$\leq \frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{\eta_\ell}{2|\mathcal{C}|} \mathbb{E}_r \left[\sum_{c\in\mathcal{C}} \sum_{e=0}^{E} \left\| \frac{\nabla F_c(w_{c,e}^{(r)}) - \nabla F_c(w_g^{(r)})}{\sqrt{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau}} \right\|^2 \right] \tag{67}$$

$$\leq \frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{\eta_\ell L^2}{2|\mathcal{C}|\tau} \mathbb{E}_r \left[\sum_{c\in\mathcal{C}} \sum_{e=0}^{E} \left\| w_{c,e}^{(r)} - w_g^{(r)} \right\|^2 \right] \tag{68}$$

$$\leq \frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{\eta_\ell L^2 E}{2\tau} \left(5E\eta_\ell^2 \left(\sigma_\ell^2 + 6E\sigma_g^2 \right) + 30E^2 \eta_\ell^2 \mathbb{E}_r \left[\left\| \nabla f(w_g^{(r)}) \right\|^2 \right] \right) \tag{69}$$

Plugging the bound of T_3 in T_1,

$$T_1 \leq -\frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{\eta_\ell L^2 E}{2\tau} \left(5E\eta_\ell^2 \left(\sigma_\ell^2 + 6E\sigma_g^2 \right) + 30E^2 \eta_\ell^2 \mathbb{E}_r \left[\left\| \nabla f(w_g^{(r)}) \right\|^2 \right] \right) \tag{70}$$

$$= -\frac{\eta_\ell E}{2} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)} - \beta_3 d^{(r-1)} + \tau}} + \frac{5\eta_\ell^3 L^2 E^2}{2\tau} \left(\sigma_\ell^2 + 6E\sigma_g^2 \right) + \frac{15\eta_\ell^3 L^2 E^3}{\tau} \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 \tag{71}$$

Bounding T_2

$$T_2 = \left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r \left[\frac{\Delta^{(r)}}{\sqrt{v^{(r)}} - d^{(r)} + \tau} - \frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} \right] \right\rangle \tag{72}$$

$$= \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} - \sqrt{v_j^{(r)}} + d_j^{(r)}}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \tag{73}$$

$$= \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\left(\sqrt{\beta_2 v_j^{(r-1)}} - \sqrt{v_j^{(r)}} \right) + \left(d_j^{(r)} - \beta_{3j} d_j^{(r-1)} \right)}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \tag{74}$$

$$= \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\frac{\beta_2 v_j^{(r-1)} - v_j^{(r)}}{\sqrt{\beta_2 v_j^{(r-1)}} + \sqrt{v_j^{(r)}}} + (1-\beta_{3j})((\Delta_j^{(r)})^2 - v_j^{(r)})}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \tag{75}$$

$$\leq \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\beta_2 v_j^{(r-1)} - v_j^{(r)} + (1-\beta_{3j})((\Delta_j^{(r)})^2 - v_j^{(r)})}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \tag{76}$$

Now we rearrange the nominator to convert it into terms with only $\Delta^{(r)}$ and $v^{(r)}$.

$$= \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\beta_2 (\Delta_j^{(r)})^2 - v_j^{(r)} + \beta_{3j}(v_j^{(r)} - (\Delta_j^{(r)})^2)}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \tag{77}$$

$$\begin{aligned} &= \beta_2 \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{(\Delta_j^{(r)})^2}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \\ &\quad + \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{-v_j^{(r)}}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \\ &\quad + \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\beta_{3j}(v_j^{(r)} - (\Delta_j^{(r)})^2)}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \end{aligned} \tag{78}$$

$$\begin{aligned} &\leq \frac{\beta_2}{\tau} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{(\Delta_j^{(r)})^2}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)} \\ &\quad - \frac{1}{\tau} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\beta_2 v_j^{(r-1)}}{(\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} + \tau)} \\ &\quad + \frac{1}{\tau} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \frac{\beta_{3j}(1-\beta_2) \sum_{i=0}^{r-1} \beta_2^{r-i} (\Delta_j^{(i)})^2}{(\sqrt{v_j^{(r)}} - d_j^{(r)} + \tau)} \end{aligned} \tag{79}$$

Here, the first term inequality follows from the fact that $\sqrt{\beta_2 v_j^{(r-1)}} \geq \beta_3 d_j^{(r-1)}$, since $\sqrt{\beta_2 v_j^{(r-1)}} - \beta_3 d_j^{(r-1)} \leq \eta_\ell E G \left(\sqrt{\beta_2 (1-\beta_2^r)} + \beta_3 (1-\beta_3^r) \right)$. Similarly for the second term, we can remove $\sqrt{v_j^{(r)}} - d_j^{(r)}$ from the denominator to get an upper bound since the term is positive ($\sqrt{v_j^{(r)}} - d_j^{(r)} \leq \eta_\ell E G \left(\sqrt{(1-\beta_2^r)} + (1-\beta_3^r) \right)$). And $-v_j^{(r)} \leq -\beta_2 v_j^{(r-1)}$, since $v_j^{(r)}$ is always positive. For the third term, $v_j^{(r)} - (\Delta_j^{(r)})^2 = (1-\beta_2) \sum_{i=0}^{r} \beta_2^{r-i} (\Delta_j^{(i)})^2 - (\Delta_j^{(r)})^2 = (1-\beta_2) \sum_{i=0}^{r-1} \beta_2^{r-i} (\Delta_j^{(i)})^2$. For the same of simplicity, we have assumed $\beta_3 \in [0,1]$ to be a constant, but a similar anaysis can be derived for a dynamic β_3 as well.

Bounding the above term further,

$$\begin{aligned} T_2 \leq \frac{\beta_2}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times (\Delta_j^{(r)})^2 + \frac{\sqrt{\beta_2}}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \Delta_j^{(r)} \times \sqrt{v_j^{(r-1)}} \\ + \frac{1}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \beta_{3j}(1-\beta_2) \sum_{i=0}^{r-1} \beta_2^{r-i} (\Delta_j^{(i)})^2 \end{aligned} \tag{80}$$

The first term inequality follows from $\sqrt{v_j^{(r)}} - d_j^{(r)} \geq \Delta_j^{(r)}$. Second term inequality follows from $-\sqrt{\beta_2 v_j^{(r-1)}} - \beta_{3j} d_j^{(r-1)} \geq -\sqrt{\beta_2 v_j^{(r-1)}}$.

$$\begin{aligned} T_2 \leq \frac{\beta_2}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times (\Delta_j^{(r)})^2 + \frac{\sqrt{\beta_2}}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times (\Delta_j^{(r)})^2 \\ + \frac{1-\beta_2}{\tau^2} \mathbb{E}_r \sum_{j=1}^{d} [\nabla f(w_g^{(r)})]_j \times \beta_{3j} \sum_{i=0}^{r-1} \beta_2^{r-i} (\Delta_j^{(i)})^2 \end{aligned} \tag{81}$$

$$\therefore T_2 \leq \frac{(\beta_2+\sqrt{\beta_2})G}{\tau^2}\mathbb{E}_r\sum_{j=1}^{d}(\Delta_j^{(r)})^2 + \frac{(1-\beta_2)G}{\tau^2}\mathbb{E}_r\sum_{i=0}^{r-1}\beta_2^{r-i}(\Delta_j^{(i)})^2 \tag{82}$$

$$\leq \frac{(\beta_2+\sqrt{\beta_2})G}{\tau^2}\mathbb{E}_r\sum_{j=1}^{d}(\Delta_j^{(r)})^2 + \frac{(1-\beta_2^{r-1})G^3|\mathcal{C}|^2E^2}{\tau^2} \tag{83}$$

$$\leq \frac{(\beta_2+\sqrt{\beta_2})G}{\tau^2}\mathbb{E}_r\sum_{j=1}^{d}(\Delta_j^{(r)})^2 + \frac{(1-\beta_2^{r-1})G^3|\mathcal{C}|^2E^2}{\tau^2} \tag{84}$$

Plugging in the bounds of T_1 and T_2 in Equation 61,

$$\mathbb{E}_r[f(w_g^{(r+1)})] = f(w_g^{(r)}) + \eta_g \underbrace{\left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r\left[\frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau}\right]\right\rangle}_{T_1}$$
$$+ \eta_g \underbrace{\left\langle \nabla f(w_g^{(r)}), \mathbb{E}_r\left[\frac{\Delta^{(r)}}{\sqrt{v^{(r)}}-d^{(r)}+\tau} - \frac{\Delta^{(r)}}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau}\right]\right\rangle}_{T_2} + \frac{L\eta_g^2}{2}\mathbb{E}_r\left[\frac{(\Delta^{(r)})^2}{\left(\sqrt{v^{(r)}}-d^{(r)}+\tau\right)^2}\right] \tag{85}$$

$$\leq f(w_g^{(r)}) + \eta_g\left(-\frac{\eta_\ell E}{2}\frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau} + \frac{5\eta_\ell^3 L^2E^2}{2\tau}\left(\sigma_\ell^2+6E\sigma_g^2\right) + \frac{15\eta_\ell^3L^2E^3}{\tau}\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2\right)$$
$$+ \eta_g\left(\frac{G(\beta_2+\sqrt{\beta_2})}{\tau^2}\mathbb{E}_r\sum_{j=1}^{d}(\Delta_j^{(r)})^2 + \frac{G^3|\mathcal{C}|^2E^2(1-\beta_2^{r-1})}{\tau^2}\right) + \frac{L\eta_g^2}{2\tau^2}\mathbb{E}_r(\Delta^{(r)})^2 \tag{86}$$

Using the bounds derived on $\mathbb{E}_r(\Delta^{(r)})^2$ from Lemma C.9,

$$\therefore \mathbb{E}_r[f(w_g^{(r+1)})] \leq f(w_g^{(r)}) - \frac{\eta_g\eta_\ell E}{2\tau}\frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau} + \frac{5\eta_g\eta_\ell^3L^2E^2}{2\tau}(\sigma_\ell^2+6E\sigma_g^2)$$
$$+ \frac{15\eta_g\eta_\ell^3L^2E^3}{\tau}\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2 + \left(\frac{L\eta_g^2}{2\tau^2}+\frac{\eta_g G(\beta_2+\sqrt{\beta_2})}{\tau^2}\right)\mathbb{E}_r(\Delta^{(r)})^2 + \frac{G^3|\mathcal{C}|^2E^2(1-\beta_2^{r-1})}{\tau^2} \tag{87}$$

$$\leq f(w_g^{(r)}) - \frac{\eta_g\eta_\ell E}{2\tau}\frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau} + \frac{5\eta_g\eta_\ell^3L^2E^2}{2|\mathcal{C}|\tau}(\sigma_\ell^2+6E\sigma_g^2) + \frac{15\eta_g\eta_\ell^3L^2E^3}{|\mathcal{C}|\tau}\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2$$
$$+ \left(\frac{L\eta_g^2}{2\tau^2}+\frac{\eta_g G(\beta_2+\sqrt{\beta_2})}{\tau^2}\right)\left(\frac{6\eta_\ell^2E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2EG^2 + \frac{30\eta_\ell^4L^2E^3}{|\mathcal{C}|}\left(\sigma_\ell^2+6E\sigma_g^2\right)\right)$$
$$+ \left(\frac{L\eta_g^2}{2\tau^2}+\frac{\eta_g G(\beta_2+\sqrt{\beta_2})}{\tau^2}\right)\left(\frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2$$
$$+ \frac{G^3|\mathcal{C}|^2E^2(1-\beta_2^{r-1})}{\tau^2} \tag{88}$$

$$\leq f(w_g^{(r)}) - \frac{\eta_g\eta_\ell E}{2\tau}\frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}}-\beta_3 d^{(r-1)}+\tau} + \left(\frac{5\eta_g\eta_\ell^3L^2E^2}{2|\mathcal{C}|\tau} + \frac{30q_1\eta_\ell^4L^2E^3}{|\mathcal{C}|}\right)(\sigma_\ell^2+6E\sigma_g^2)$$
$$+ \left(\frac{15\eta_g\eta_\ell^3L^2E^3}{|\mathcal{C}|\tau}+q_1q_2\right)\mathbb{E}_r\left\|\nabla f(w_g^{(r)})\right\|^2 + q_1q_3 + \frac{G^3|\mathcal{C}|^2E^2(1-\beta_2^{r-1})}{\tau^2} \tag{89}$$

where $q_1 = \frac{L\eta_g^2}{2\tau^2} + \frac{\eta_g G(\beta_2+\sqrt{\beta_2})}{\tau^2}$, $q_2 = \frac{180\eta_\ell^4L^2E^4}{|\mathcal{C}|} + 2E^2\eta_\ell^2 + 6\eta_\ell^2E(B^2-1)$, $q_3 = \frac{6\eta_\ell^2E\sigma_\ell^2}{|\mathcal{C}|} + 6\eta_\ell^2EG^2$.

Summing over $r = 1$ to R gives us,

$$\mathbb{E}_r[f(w_g^{(R)})] \leq f(w_g^{(0)}) - \sum_{r=1}^{R} \frac{\eta_g \eta_\ell E}{2\tau} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} + \sum_{r=1}^{R} \left(\frac{5\eta_g \eta_\ell^3 L^2 E^2}{2|\mathcal{C}|\tau} + \frac{30 q_1 \eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \right) (\sigma_\ell^2 + 6E\sigma_g^2)$$
$$+ \sum_{r=1}^{R} \left(\frac{15\eta_g \eta_\ell^3 L^2 E^3}{|\mathcal{C}|\tau} + q_1 q_2 \right) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 + \sum_{r=1}^{R} \left(q_1 q_3 + \frac{G^3 |\mathcal{C}|^2 E^2 (1 - \beta_2^{r-1})}{\tau^2} \right) \tag{90}$$

$$\leq f(w_g^{(0)}) - \sum_{r=1}^{R} \frac{\eta_g \eta_\ell E}{2\tau} \frac{[\nabla f(w_g^{(r)})]^2}{\sqrt{\beta_2 v^{(r-1)}} - \beta_3 d^{(r-1)} + \tau} + R \left(\frac{5\eta_g \eta_\ell^3 L^2 E^2}{2|\mathcal{C}|\tau} + \frac{30 q_1 \eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \right) (\sigma_\ell^2 + 6E\sigma_g^2)$$
$$+ \sum_{r=1}^{R} \left(\frac{15\eta_g \eta_\ell^3 L^2 E^3}{|\mathcal{C}|\tau} + q_1 q_2 \right) \mathbb{E}_r \left\| \nabla f(w_g^{(r)}) \right\|^2 + R q_1 q_3 + \frac{G^3 |\mathcal{C}|^2 E^2}{\tau^2} \left(R - \frac{1 - \beta_2^R}{(1 - \beta_2)} \right) \tag{91}$$

Using the fact $\frac{15\eta_g \eta_\ell^3 L^2 E^3}{|\mathcal{C}|\tau} + q_1 q_2 \leq \frac{\eta_g \eta_\ell E}{2\tau}$ derived from the bounds on the local learning rate $\eta_\ell \leq \min \left[\left(\frac{|\mathcal{C}|}{30 L^2 E} \right)^{\frac{1}{2}}, \left(\frac{\tau}{6(B^2-1)\left[G(\beta_2 + \sqrt{\beta_2}) + L\eta_g\right]} \right) \right]$, we get

$$\mathbb{E}_r[f(w_g^{(R)})] \leq f(w_g^{(0)}) - \frac{\eta_g \eta_\ell E}{2\tau} \sum_{r=1}^{R} \sum_{j=1}^{d} \frac{[\nabla f(w_g^{(r)})]_j^2}{\sqrt{\beta_2 v_j^{(r-1)}} - \beta_3 d_j^{(r-1)} + \tau} + R \left(\frac{5\eta_g \eta_\ell^3 L^2 E^2}{2|\mathcal{C}|\tau} + \frac{30 q_1 \eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \right) (\sigma_\ell^2 + 6E\sigma_g^2)$$
$$+ R q_1 q_3 + \frac{G^3 |\mathcal{C}|^2 E^2}{\tau^2} \left(R - \frac{1 - \beta_2^R}{(1 - \beta_2)} \right) \tag{92}$$

Using the following lower bound,

$$\sum_{r=1}^{R} \sum_{j=1}^{d} \frac{[\nabla f(w_g^{(r)})]_j^2}{\sqrt{\beta_2 v_j^{(r-1)}} - \beta_3 d_j^{(r-1)} + \tau} \geq \sum_{r=1}^{R} \sum_{j=1}^{d} \frac{[\nabla f(w_g^{(r)})]_j^2}{\eta_\ell E G(\sqrt{\beta_2(1 - \beta_2^r)} + \beta_3(1 - \beta_3^r)) + \tau} \tag{93}$$

$$\geq \frac{R}{\eta_\ell E G(\sqrt{\beta_2(1 - \beta_2^r)} + \beta_3(1 - \beta_3^r)) + \tau} \min_{1 \leq r \leq R} \left\| \nabla f(w_g^{(r)}) \right\|^2 \tag{94}$$

we derive the convergence bound as follows,

$$\min_{0 \leq r \leq R} \mathbb{E} \left\| \nabla f(w_g^{(r)}) \right\|^2 \leq \frac{2(\eta_\ell E G(\sqrt{\beta_2(1 - \beta_2^r)} + \beta_3(1 - \beta_3^r)) + \tau)}{\eta_\ell \eta_g E R} \left[f(w_g^{(0)}) - \mathbb{E}_r[f(w_g^{(R)})] \right]$$
$$+ \frac{2(\eta_\ell E G(\sqrt{\beta_2(1 - \beta_2^r)} + \beta_3(1 - \beta_3^r)) + \tau)}{\eta_\ell \eta_g E} \left[\left(\frac{5\eta_g \eta_\ell^3 L^2 E^2}{2|\mathcal{C}|\tau} + \frac{30 q_1 \eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \right) (\sigma_\ell^2 + 6E\sigma_g^2) + q_1 q_3 \right]$$
$$+ \frac{2(\eta_\ell E G(\sqrt{\beta_2(1 - \beta_2^r)} + \beta_3(1 - \beta_3^r)) + \tau)}{\eta_\ell \eta_g E R} \left[\frac{G^3 |\mathcal{C}|^2 E^2}{\tau^2} \left(R - \frac{1 - \beta_2^R}{1 - \beta_2} \right) \right] \tag{95}$$

$$\therefore \min_{0 \leq r \leq R} \mathbb{E} \left\| \nabla f(w_g^{(r)}) \right\|^2 \leq \mathcal{O} \Bigg(\frac{\eta_\ell E(\beta_3(1 - \beta_3^r)) + \tau}{\eta_\ell \eta_g E R} \left[f(w_g^{(0)}) - \mathbb{E}_r[f(w_g^{(R)})] \right]$$
$$+ \frac{\eta_\ell E(\beta_3(1 - \beta_3^r)) + \tau}{\eta_\ell \eta_g E} \left[\left(\frac{5\eta_g \eta_\ell^3 L^2 E^2}{2|\mathcal{C}|\tau} + \frac{30 q_1 \eta_\ell^4 L^2 E^3}{|\mathcal{C}|} \right) (\sigma_\ell^2 + 6E\sigma_g^2) + q_1 q_3 \right]$$
$$+ \frac{\eta_\ell E(\beta_3(1 - \beta_3^r)) + \tau}{\eta_\ell \eta_g E R} \left[\frac{G^3 |\mathcal{C}|^2 E^2}{\tau^2} \left(R - \frac{1 - \beta_2^R}{1 - \beta_2} \right) \right] \Bigg) \tag{96}$$

Assuming $\eta_\ell = \Theta(1/L\sqrt{E})$, $\eta_g = \Theta(1/\sqrt{R})$, and $\tau = G/L$, we get the following asymptotic bound for the convergence of FLASH,

$$\min_{0 \leq r \leq R} \mathbb{E} \left\| \nabla f(w_g^{(r)}) \right\|^2 \leq \mathcal{O} \left(\frac{f(w_g^{(0)}) - \mathbb{E}_r[f(w_g^{(R)})]}{\sqrt{ER}} + \frac{G}{\sqrt{ER}|\mathcal{C}|} (\sigma_\ell^2 + 6E\sigma_g^2) + \frac{6L\sigma_\ell^2}{RG^2|\mathcal{C}|} + \frac{6L}{R} \right) \tag{97}$$

□