

FedAdamom: Adaptive Momentum for Improved Generalization in Federated Optimization

Wenjie Hou^{*1} Tianxiang Chen^{*1,2} Feng Wang^{1,2} Tiantong Wu² Zhiming Zheng¹
Shaoting Tang¹ Wei Yang Bryan Lim²

¹ Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing,
School of Artificial Intelligence, Beihang University

² College of Computing and Data Science, Nanyang Technological University

{houwj, txchen, fengwang}@buaa.edu.cn

Abstract

Federated learning (FL) has emerged as a widely adopted training paradigm for privacy-preserving machine learning. Despite the past success of SGD-based methods, they still suffer from severe data heterogeneity and the lack of adaptivity in practical applications. While several adaptive federated optimization methods (such as FedAdam) have been proposed and demonstrated to achieve faster convergence, they fail to show significant improvements in generalization performance under highly heterogeneous data distributions, and their optimization and generalization mechanisms remain insufficiently understood. To fill this gap, we introduce diffusion theory into the adaptive federated optimization framework and analyze the distinct effects of adaptive learning rate and global momentum from the perspectives of saddle-point escaping and flat-minima selection. Theoretical results show that although FedAdam outperforms FedAvg/FedAvgM in escaping saddle points, the latter escapes sharp minima more efficiently. The root cause lies in the fact that adaptive learning rate, while enhancing saddle-point escape, weakens the preference for flat minima. Motivated by these insights, we propose FedAdamom, a new adaptive federated optimization algorithm that adapts the momentum hyperparameter rather than the learning rate. FedAdamom maintains strong saddle-point escaping capability while enhancing flat-minima selection. We further establish its convergence guarantees under non-convex objectives. Extensive experiments demonstrate that FedAdamom significantly outperforms existing adaptive federated optimization methods in terms of convergence speed, generalization performance, and preference for flat minima. Our code is available at <https://github.com/Tenshawn/FedAdamom>.

^{*} indicates equal contribution. Corresponding author: Shaoting Tang.

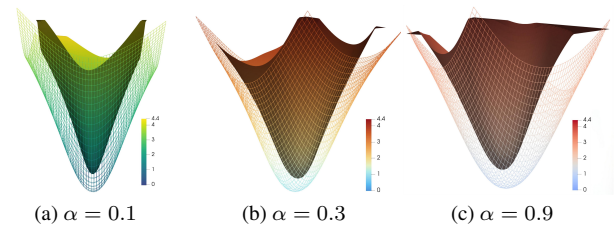


Figure 1. Comparison of FedAvg(net) and FedAdam(solid) loss landscapes with varying degrees of data heterogeneity (α) on the CIFAR-10 datasets. The effectiveness of FedAdam and FedAvg in converging to global flat minima is highly influenced by data heterogeneity, with higher heterogeneity ($\alpha \downarrow$) leading to convergence toward sharper minima. This highlights the importance of optimizing global sharpness. Model: CNN.

1. Introduction

Federated Learning (FL) is a large-scale machine learning framework in which multiple clients collaboratively train a shared model under the coordination of a central server [34]. The key benefit of this paradigm is that each client’s private data remains local and is not shared with the server or other clients, thereby providing a degree of data privacy protection. The baseline algorithm for FL, FedAvg [34], proceeds as follows: each of the selected clients updates its local model using stochastic gradient descent based on their local data. After local updates, the selected clients upload their local model weights to the server, which then computes the new global model by averaging the model weights.

Although FedAvg has achieved remarkable success in its early applications, it essentially shares the same characteristics as Stochastic Gradient Descent (SGD) and may perform poorly when facing heavy-tailed stochastic gradient noise distributions—a phenomenon frequently encountered during model training [36]. These limitations primarily arise from two factors: i) the high heterogeneity of client

data in FL settings [17]; and ii) the lack of adaptivity such as learning rate adjustments or momentum terms [40]. In this context, employing adaptive learning rate methods that leverage information from past iterations can lead to more efficient optimization. For example, FedAdam [40] incorporates adaptive learning rates with momentum, and has become one of the most popular optimization algorithms for accelerating FL training. However, although FedAdam achieves faster convergence compared to FedAvg, its generalization performance does not always show significant improvement. Studies [50] have shown that in common convolutional neural network (CNN) architectures, using Adam optimizer tends to converge to sharper local minima, which are widely believed to be closely associated with poorer generalization performance [4, 5, 12, 14, 37]. Consequently, while FedAdam improves optimization efficiency, it may introduce new challenges in achieving robust generalization.

Diffusion theory has been widely employed to analyze the dynamical behavior of SGD-based methods in the process of minima selection [29, 54]. Xie et al. [53] used the diffusion theory to quantitatively demonstrate that, due to the Hessian-dependent covariance structure of stochastic gradient noise, SGD tends to converge to flatter minima with better generalization ability among numerous minima. Subsequent studies [8, 14, 54, 62] have further shown that injecting or amplifying gradient noise correlated with curvature sharpness, similar to the inherent noise in SGD, can effectively facilitate convergence toward flatter minima.

Although FedAdam and FedAvg both use SGD on the clients and can converge to flatter minima locally, prior work [2] shows that the local and global loss landscapes are often inconsistent due to data heterogeneity in FL. Accordingly, we incorporate diffusion theory into the FL framework to investigate how FedAvg selects minima during the optimization process. Our analysis reveals that FedAvg typically escapes from sharp global minima faster than FedAdam (Fig. 1), thereby providing a theoretical explanation for its better generalization performance. In addition, saddle point escape dynamics have been recognized as a critical factor influencing training efficiency in deep learning, and previous studies have analyzed convergence properties from this perspective [45]. Similarly, the mismatch between the local and global loss landscapes means that the local preference for flat minima does not necessarily guarantee good global generalization.

To further investigate local and global dynamical behaviors of FedAvg and FedAdam, we systematically examine the distinct roles of the adaptive global learning rate and the momentum mechanism in FL, characterizing their behaviors in global saddle point escape and flat minima selection. Our theoretical analysis explains why FedAdam converges faster yet generalizes worse than FedAvg. Specifically, we

demonstrate that in the FL context, momentum facilitates efficient escape from saddle points but has a limited effect on escaping sharp minima. In contrast, adaptive learning rate schemes accelerate the escape from the saddle point while weakening the algorithm’s preference for flatter minima. The main contributions of this paper are summarized as follows:

- We introduce diffusion theory into the FL context to systematically analyze the distinct roles of adaptive learning rates and momentum mechanisms in federated adaptive optimization algorithms such as FedAdam. By examining from the perspectives of both saddle point escape and flat minimum selection, our theoretical analysis provides insight into why FedAdam often achieves faster convergence yet generalizes worse than the FedAvg algorithm.
- Building on our theoretical foundation, we propose FedAdamom, a novel adaptive federated optimization algorithm that adapts the momentum hyperparameter rather than the global learning rate in conventional adaptive methods. This design preserves the advantage of efficient saddle point escaping characteristics while avoiding the drawback of impairing flat minimum selection.
- We provide a convergence upper bound for FedAdamom in the general non-convex setting with partial participation. The theoretical result matches the convergence rates of existing FL algorithms.
- We conduct extensive experiments comparing FedAdamom against several established adaptive federated optimization baselines. The results empirically validate our theoretical findings, demonstrating that FedAdamom consistently outperforms existing methods in terms of convergence speed, generalization performance and flat minimum selection.

2. Related Work

Federated Learning Since the introduction of the FedAvg framework [34], FL has attracted significant attention from both machine learning and computer vision communities to address the key challenges of the communication bottleneck in conventional centralized learning. In machine learning, the focus is primarily on theoretical analysis from the perspective of optimization methods [19, 28, 30, 44, 46, 56], whereas computer vision promotes exploring its real-world applications, such as autonomous driving and healthcare [7, 35, 42, 55, 60]. Following FedAvg, numerous research efforts have aimed to improve both theoretical and empirical performance of FL algorithms. FedProx [27] introduces a proximal term to restrict local model drift under data heterogeneity. FedNova [47] utilizes different local steps in asynchronous aggregation settings. Notably, Karimireddy et al. [17] propose SCAFFOLD, which employs variance reduction via control variates to mitigate client drift. FedDyn [1] dynamically adjusts local objectives to closely align

with the global objective. MOON [26] leverages model-level contrastive learning to align local and global models, thereby mitigating optimization discrepancies. Mime [18] combines control variates with server-side optimizer state to mitigate the adverse effects of non-i.i.d. data. FedSSD [10] and FedCAD [11] adaptively adjust knowledge distillation for local models based on the confidence of the global model to alleviate data heterogeneity. FedDisco [57] incorporates both dataset size and distribution discrepancy into the aggregation weights, while FedAWA [43] adjusts the aggregation weights based on client vectors during training, enhancing the stability and generalization of the global model. FedSAM [38] applies SAM as a client-side optimizer to reduce the difference between global and local losses. Despite the remarkable progress, FL still suffers from fundamental challenges, including substantial communication overhead, inefficient local updates, and statistical heterogeneity across client datasets. Given this context, our proposed algorithm aims to learn a global model that generalizes better without increasing communication cost or local computational complexity.

Adaptive Optimization In centralized learning, a wide range of adaptive optimization algorithms such as AdaGrad [6], Adam [20], Amsgrad [39] and their powerful variants have been developed [31, 32, 58]. These methods automatically adapt to different learning tasks and training challenges, and have demonstrated superior empirical performance compared to SGD. Afterward, Adan [52] further improves performance via an adaptive Nesterov momentum estimator, while MIAAdam [15] integrates a multiple integral term into Adam, outperforming Adam and its variants. In FL settings, Reddi et al. [40] first propose adaptive FL methods, which deploy adaptive optimizers on the global server to handle long-tailed stochastic gradient noise and accelerate convergence. FedAMS [48] incorporates gradient compression to achieve a communication-efficient adaptive method. FAFED [51] employs a momentum-based variance reduction technique, while FADAS [49] integrates asynchronous updates into adaptive federated optimization and extends it with a delay-adaptive learning adjustment strategy. FedCada [61] adjusts the adaptive correction process of the Adam algorithm, achieving favorable results in both adaptivity and performance.

However, most of these methods require additional communication costs or extra client computation, which are problematic in realistic FL scenarios.

3. Preliminaries

3.1. Problem setting

In FL, the objective is to learn a global model $f(x) : \mathcal{X} \rightarrow \mathcal{Y}$ parametrized by $x \in \mathbb{R}^d$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the input and the output spaces, respectively. In image classification,

$\mathcal{X}$ contains the images and $\mathcal{Y}$ their corresponding labels. Let $F_i(x) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [F_i(x; \xi_i)]$ be the empirical loss function of client $i \in \{1, \dots, n\}$ with a local dataset denoted by $\mathcal{D}_i$, and ξ_i is the data sample randomly drawn from the local data distribution $\mathbb{D}_i$. The global objective is:

$$\min_x \left\{ \frac{1}{n} \sum_{i=1}^n F_i(x) \right\}. \quad (1)$$

We focus on the heterogeneous data distributions setting, i.e., $\mathbb{D}_i \neq \mathbb{D}_j, \forall i \neq j$. The training proceeds in three stages: (i) Due to communication constraints and other practical considerations, the central server selects a subset of clients and broadcasts the global model to them; (ii) Each selected client trains its local model using its local datasets and a local optimizer. After training, the client sends its local updates back to the server; And (iii) the server updates the global model using the received local updates with the global optimizer, GLOBALOPT (e.g., SGD, Adam). FEDOPT [40] solves Eq (1) as

$$\begin{cases} \Delta_t = \frac{1}{s} \sum_{i \in \mathcal{S}_t} (x_t - x_{t,K}^i), \\ x_{t+1} = x_t - \text{GLOBALOPT}(x_t, \Delta_t, \eta), \end{cases}$$

where Δ_t denotes the global pseudo-gradient at round t and K is the number of local updates. In the FedAvg algorithm [34], the server updates the global model as $x_{t+1} = x_t + \eta \Delta_t$, which is equivalent to performing a single SGD step on the pseudo-gradient.

3.2. Diffusion theory

The diffusion theory is an important theoretical tool for understanding how deep learning dynamics work. It characterizes the diffusion process of probability densities of parameters instead of model parameters themselves. The SGD diffusion theory [53] analyzes how SGD escapes from minima. Noting that the gradient vanishes near critical points, the behavior of SGD escaping from a critical point is essentially a noise-dominated diffusion process, which can be rigorously modeled by a Langevin equation [25, 33]:

$$dx = -\nabla f(x)dt + [\eta C(x)]^{\frac{1}{2}} dW_t, \quad (2)$$

where $dW_t \sim \mathcal{N}(0, Idt)$, I is the identity matrix, and $C(x)$ is the gradient noise covariance matrix. Then, the probability density governed by the Langevin Equation is described by the Fokker-Planck Equation [41]:

$$\frac{\partial P(x, t)}{\partial t} = \nabla \cdot [P(x, t) \nabla f(x)] + \nabla \cdot \nabla D(x) P(x, t), \quad (3)$$

where $\nabla \cdot$ is the divergence operator and $D(x) = \frac{\eta C(x)}{2}$ is the diffusion matrix [53]. Before conducting the analysis,

we introduce three fundamental assumptions that are crucial for applying the diffusion theory framework [54].

Assumption 1. *The loss around the critical point c is*

$$f(x) \approx f(c) + \frac{1}{2}(x - c)^\top H(c)(x - c). \quad (4)$$

Assumption 2. (Quasi-equilibrium approximation). *The system is in a quasi-equilibrium state near the minima.*

Assumption 3. (Low-temperature approximation). *The system's gradient noise magnitude is small (i.e., low temperature).*

These assumptions are used to model the process of SGD escaping minima as a Kramers Escape Problem [16] in deep learning. Recent works [15, 54, 62] also leveraged Assumption 2 and Assumption 3 implicitly or explicitly.

4. Diffusion Dynamics of FL

In this section, we analyze how FedAvg, FedAvgM, and FedAdam escape from saddle points and sharp minima.

4.1. FedAvg Diffusion Dynamics

Reddi et al. [40] show that FedAvg is equivalent to SGD on the server side. So we can deduce that the time evolution of the probability distribution of the parameter displacement is governed by the Fokker–Planck equation (Eq. (3)). Under the Assumption 1, the dynamic is governed by Eq. (2), then we can get the displacement of x from a saddle point c follows a Gaussian distribution, i.e., $x \sim \mathcal{N}(c, \Sigma(t))$, where $\Sigma(t) = U \text{diag}(\theta_1^2, \theta_2^2, \dots, \theta_n^2) U^\top$ and

$$\theta_i^2(T) = \frac{|H_i| \eta^2 \eta_l^2 T}{B} + \mathcal{O}(B^{-1} H_i^2 \eta^3 \eta_l^3 T^2),$$

where H_i is the i -th eigenvalue of the Hessian matrix H at c . The column vectors of U are exactly the eigenvectors of H . For additional details, please refer to the Appendix B.1. Since $x \sim \mathcal{N}(c, \Sigma(t))$, the mean squared displacement is determined by the variance, i.e., $\langle \Delta x_i^2(t) \rangle = \theta_i^2(t)$. In the following, the diffusion theory framework is utilized to measure the generalization capabilities of the FedAvg by computing the mean escape time, represented as τ . Following the theoretical analysis in [53, 54], we derive Theorem 1. The complete proof can be found in Appendix B.2.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 hold while saddle point $\mathbf{b}$ is the exit from sharp minimum $\mathbf{a}$. Then the mean escape time of FedAvg from sharp minimum $\mathbf{a}$ to flat minimum $\mathbf{d}$ through saddle point $\mathbf{b}$ is given by*

$$\log(\tau_{\text{FedAvg}}) = \mathcal{O}\left(\frac{2B\Delta f}{\eta \eta_l H_{ae}}\right), \quad (5)$$

where the subscript e indicates the escape direction; B denotes the batch size; $\Delta f = f(\mathbf{b}) - f(\mathbf{a})$; H is the Hessian matrix.

Note that the escape direction is aligned with some eigenvector in the diffusion theory [53].

4.2. FedAvgM Diffusion Dynamics

FedAvgM [13] employs momentum-based SGD updates on the server side:

$$\begin{cases} m_t = \beta m_{t-1} + \Delta_t, \\ x_{t+1} = x_t + \eta m_t, \end{cases} \quad (6)$$

where m_t denotes the global momentum at round t . Compared with FedAvg, the dynamics of FedAvgM are additionally influenced by the momentum drift term. The mean-squared displacement near saddle points is

$$\begin{aligned} \langle \Delta x_i^2 \rangle &= \frac{|H_i| \eta^2 \eta_l^2}{2(1-\beta)^3 B} [1 - \exp(-(1-\beta)T)]^2 \\ &\quad + \frac{|H_i| \eta^2 \eta_l^2 T}{(1-\beta)^2 B} + \mathcal{O}(B^{-1} H_i^2 \eta^3 \eta_l^3 T^2). \end{aligned}$$

Compared with FedAvg, we observe that FedAvgM escaping saddle points is influenced not only by the diffusion effect but also by an additional momentum drift effect. We then formulate how FedAvgM escapes the minimum, as stated in Theorem 2.

Theorem 2. *Suppose that Assumption 1, 2, and 3 hold while saddle point $\mathbf{b}$ is the exit from sharp minimum $\mathbf{a}$. Then the mean escape time of FedAvgM from sharp minimum $\mathbf{a}$ to flat minimum $\mathbf{d}$ through saddle point $\mathbf{b}$ is given by*

$$\log(\tau_{\text{FedAvgM}}) = \mathcal{O}\left(\frac{2(1-\beta)B\Delta f}{\eta \eta_l H_{ae}}\right). \quad (7)$$

We note that $\log(\tau_{\text{FedAvg}}), \log(\tau_{\text{FedAvgM}}) = \mathcal{O}(H_{ae}^{-1})$, which implies that momentum has no impact on escaping from the minimum.

4.3. FedAdam Diffusion Dynamics

The global update formula of FedAdam is shown as

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1) \Delta_t, \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) \Delta_t^2, \\ x_{t+1} = x_t + \eta \frac{m_t}{\sqrt{v_t} + \epsilon}, \end{cases} \quad (8)$$

where ϵ is a small value to prevent division by zero in the denominator. In contrast to FedAvg and FedAvgM, the dynamics of FedAdam are further affected by the adaptive learning rate. The mean squared displacement is

$$\begin{aligned} \langle \Delta x_i^2 \rangle &= \frac{\eta^2 \eta_l}{2(1-\beta_1)} [1 - \exp(-(1-\beta_1)T)]^2 \\ &\quad + \eta^2 \eta_l^2 T + \mathcal{O}(\sqrt{B|H_i|} \eta^3 \eta_l^3 T^2). \end{aligned} \quad (9)$$

The first term represents the momentum drift effect, and the second term is the diffusion effect. In comparison, FedAdam escapes saddle points faster because both effects are approximately Hessian-independent and isotropic [45]. This also explains why FedAdam typically converges faster. We next formulate how FedAdam escapes the minimum as stated in Theorem 3.

Theorem 3. *Suppose that Assumption 1, 2, and 3 hold while saddle point $\mathbf{b}$ is the exit from sharp minimum $\mathbf{a}$. Then the mean escape time of FedAdam from sharp minimum $\mathbf{a}$ to flat minimum $\mathbf{d}$ through saddle point $\mathbf{b}$ is given by*

$$\log(\tau_{\text{FedAdam}}) = \mathcal{O}\left(\frac{2\sqrt{B}\Delta f}{\eta\eta_l\sqrt{H_{ae}}}\right). \quad (10)$$

Since the above three algorithms all perform SGD locally, their local updates satisfy $\log(\tau_{\text{local}}) = \mathcal{O}(H_{ae}^{-1})$ [53]. In diffusion theory, the sharpness of a minimum is reflected by the eigenvalue of the Hessian corresponding to the escape direction along an eigenvector [54]. FedAvg and FedAvgM satisfy $\log(\tau) = \mathcal{O}(H_{ae}^{-1})$ while FedAdam satisfies $\log(\tau) = \mathcal{O}(H_{ae}^{-\frac{1}{2}})$, which implies that the global update of FedAdam is less dependent on the Hessian and is inconsistent with the local update in terms of escaping sharp minima. As H_{ae} increases, i.e., the escape direction becomes steeper, the escaping time $\log(\tau)$ of FedAvg and FedAvgM decreases more significantly than that of FedAdam. Therefore, compared with FedAvg, FedAdam is less efficient at selecting flat minima, thereby degrading the generalization.

5. Adaptive FL Momentum Optimization

To simultaneously improve the generalization of FedAdam while preserving the fast convergence of adaptive methods, we propose FedAdamom, an algorithm designed to efficiently escape saddle points without damaging the consistency of flat minima selection. Algorithm 1 presents the details of FedAdamom. According to Eq.(9), the adaptive learning rate utilizes second-order gradient information, rendering the momentum drift and the diffusion effect of FedAdam approximately Hessian-independent near saddle points. Although this property accelerates the escape from saddle points, it leads to an inconsistency between the global and local abilities to escape from sharp minima, thereby impairing the generalization performance of the global model. Inspired by this observation, we apply the adaptive update to the momentum parameter.

Adaptive Momentum Parameter To transfer the adaptive update into the momentum parameter, we set $\beta_{1,t} = (1 - \frac{v_t}{\bar{v}_t}).\text{Clip}(0, 1 - \epsilon)$, where $\bar{v}_t$ is the mean of all elements in the estimated v_t (Lines 15 and 16, Algorithm 1).

Algorithm 1 FedAdamom Algorithm Framework

```

1: Input: model parameters  $x_0$ ; total communication
   rounds  $T$ ; number of local iterations  $K$ ; momentum pa-
   rameter  $\beta_2$ ; model dimension  $d$ ; and local learning rate
    $\eta_l$ .
2: Output: model parameters  $x_T$ .
3: for  $t = 0, 1, 2, \dots, T - 1$  do
4:   Select active clients-sets  $\mathcal{S}_t$  at round  $t$ 
5:   for client  $i \in \mathcal{S}_t$  in parallel do
6:     Transmit  $x_t$  and  $v_t$  to local client  $i$  and initialize
        $x_{t,0}^i = x_t$ 
7:     for  $k = 0, 1, 2, \dots, K - 1$  do
8:        $g_{t,k}^i = \nabla F_i(x_{t,k}^i, \xi_i)$   $\triangleright$  compute unbiased
       stochastic gradient
9:        $x_{t,k+1}^i = x_{t,k}^i - \eta_l g_{t,k}^i$   $\triangleright$  update gradient
       step
10:    end for
11:     $\Delta x_t^i = x_{t,K}^i - x_{t,0}^i$  and transmit  $\Delta x_t^i$  to server
12:    end for
13:     $\Delta_t = \frac{1}{s} \sum_{i \in \mathcal{S}_t} \Delta x_t^i$ 
14:     $v_t = \beta_2 v_{t-1} + (1 - \beta_2) \Delta_t^2$ 
15:     $\bar{v}_t = \frac{1}{d} \sum_{i=1}^d v_{t,i}$ 
16:     $\beta_{1,t} = (1 - \frac{v_t}{\bar{v}_t}).\text{Clip}(0, 1 - \epsilon)$ 
17:     $m_t = \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) \Delta_t$ 
18:     $x_{t+1} = x_t + \eta m_t$   $\triangleright$  update global model
19:  end for

```

We generalize the momentum parameter $\beta_{1,t}$ from a scalar to a vector, and we use $\beta_{1,t}$ to update the momentum m_t (Line 17, Algorithm 1). Note that this momentum update is parameter-wisely adaptive. Then we utilized the obtained momentum m_t to update the global model (Line 18, Algorithm 1). We will show that this design enables efficient escape from saddle points without compromising the ability to escape from sharp minima, thereby enhancing the generalization performance of the global model.

5.1. Analysis of FedAdamom Dynamics

According to the Algorithm 1, the mean-squared displacement near saddle points is

$$\begin{aligned} \langle \Delta x_i^2 \rangle &= \frac{\sum_{i=1}^d |H_i| \eta^2 \eta_l^2}{nB} + \frac{|H_i| \eta^2 \eta_l^2 T}{B} \\ &\quad + \mathcal{O}(B^{-1} H_i^2 \eta^3 \eta_l^3 T^2). \end{aligned}$$

We can see that the momentum drift term is also isotropic and approximately Hessian-independent. Thus, FedAdamom can efficiently escape saddle points. Then we formulate how FedAdamom escapes minima in Theorem 4.

Theorem 4. *Suppose that Assumptions 1, 2, 3 hold while saddle point $\mathbf{b}$ is the exit from sharp minimum $\mathbf{a}$. Then the*

mean escape time of FedAdamom from sharp minimum $\mathbf{a}$ to flat minimum $\mathbf{d}$ through saddle point $\mathbf{b}$ is given by

$$\log(\tau_{\text{FedAdamom}}) = \mathcal{O}\left(\frac{2B\Delta f}{\eta\eta_l H_{ae}}\right). \quad (11)$$

We can see that FedAdamom also achieves $\log(\tau) = \mathcal{O}(H_{ae}^{-1})$. Therefore, escaping from sharp minima remains consistent between the global and local updates. Compared with FedAdam, FedAdamom tends to converge to flatter minima, thereby improving the generalization performance.

5.2. Convergence analysis of FedAdamom

We now discuss the convergence of the proposed algorithm in general non-convex FL scenarios. To establish the convergence analysis, we make three assumptions; (i) the local loss function $F_i(\cdot)$ is L -smooth, (ii) its stochastic gradient g_t^i is an unbiased estimator of ∇F_i with bounded variance, i.e., $\mathbb{E}[g_t^i] = \nabla F_i(x_t^i)$ and $\mathbb{E}[\|g_t^i - \nabla F_i(x_t^i)\|^2] \leq \sigma_l^2$, (iii) The dissimilarity between local clients is bounded on the gradients, i.e., $\mathbb{E}[\|\nabla F_i(x) - \nabla f(x)\|^2] \leq \sigma_g^2, \forall i$. These assumptions are mild and widely used in characterizing the convergence rate of FL [40, 59]. Note that our convergence proof is free from the bounded gradient assumption of the global or local loss, while it is used in [40]. We now state the convergence result of FedAdamom; the detailed proof is provided in Appendix C in the supplementary document.

Theorem 5. Suppose that local functions $\{F_i\}_{i=1}^m$ are non-convex. Let the local and global learning rate satisfy $\eta_l = \mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$ and $\eta = \mathcal{O}\left(\sqrt{sK}\right)$, respectively. Then, the convergence rate for FedAdamom satisfies:

$$\begin{aligned} & \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(x_t)\|^2 \\ & \leq \mathcal{O} \left(\frac{L\Theta_0}{\sqrt{sKT}} + \frac{\beta_{1,max}}{1 - \beta_{1,max}} \left(\frac{\sigma_l^2}{KT} + \frac{\sigma_g^2}{T} + \Psi \right) \right), \end{aligned}$$

where $\Psi = \frac{n-s}{n-1} \frac{1}{\sqrt{sKT}} \left[\left(\frac{80}{T} + 4\right) \sigma_g^2 + \left(\frac{10}{KT} + 1\right) \sigma_l^2 \right]$ and $\Theta_0 = \mathbb{E}[f(x_0) - f(x_T)]$, $\beta_{1,max} := \max_{t \leq T, i \in [d]} \beta_{1,t}^i$.

The result in Theorem 5 matches the best convergence rate of existing FL methods [18, 40, 59].

6. Experimental Analysis

6.1. Experimental Setup

Datasets. We conduct experiments on three datasets, CIFAR-10 [21], CIFAR-100 [21], and TinyImageNet [22], each chosen to reflect different levels of data heterogeneity and participation rates. CIFAR-10 and CIFAR-100 are real-life image classification datasets with a total of 10 and 100 classes, respectively. We partition the training and test

sets among clients according to a Dirichlet distribution (e.g., $\alpha = 0.3$). A smaller α yields the setting with greater heterogeneity. We extend our experiments to the widely adopted FL benchmark, LEAF [3], known for its realistic settings. LEAF introduces heterogeneity in class distribution, data quantity, and feature alignment. For FEMNIST, CelebA, and ShakeSpeare, we use the non-i.i.d. data splits provided by LEAF.

Baselines. We compare our method, FedAdamom, with state-of-the-art adaptive FL optimization methods, which include FedAvg [34], FedAvgM [13], FedAdam [40], FedYogi [40], FedAdagrad [40], FAFED [51], FedCAda [61] and FADAS [49].

Models. We adopt the standard ResNet-18 [9] architecture for experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet. To further demonstrate the effectiveness of FedAdamom across different network architectures, we also conduct experiments using a CNN similar to LeNet-5 [23]. Additional experimental results are provided in the supplementary material.

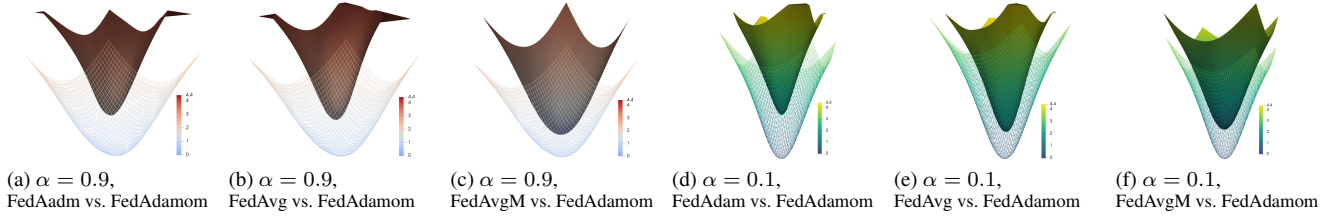
6.2. Generalization and Convergence Speed

We first present the convergence and generalization performance of the proposed FedAdamom compared with other adaptive federated optimizers on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. To examine the effects of the number of clients, data heterogeneity, and participation rate, we conduct experiments under two different settings: (i) Moderate scale, 100 clients, with 5% of clients randomly selected in each communication round; (ii) Large scale, 500 clients, with 2% of clients participating per round. To maintain the same overall dataset size, each client in the medium-scale setting is assigned $\frac{1}{5}$ of the data used in the large-scale setting. The results for the large-scale setting are provided in Appendix A.

The results in Tab. 1 show that FedAdamom achieves clear and consistent improvements in both accuracy and convergence speed over other adaptive federated optimization methods in most cases. These improvements result from the use of adaptive momentum parameters, which strengthen the selection of flat minima while preserving efficient saddle point escaping ability. This allows FedAdamom to maintain the fast convergence of adaptive methods while improving generalization. It is worth noting that FedCAda requires 3 communication overhead per round because it transmits both the current model and gradient-related information, whereas the other methods only transmit model parameters. In addition, FAFED is designed for full client participation, so its performance degrades significantly under partial participation. FedCAda is a client-side adaptive momentum algorithm that achieves fast early-stage convergence but is more sensitive to high heterogeneity. For the large-scale setting, the results show lower overall performance than in the moderate-scale ex-

Table 1. Moderate-scale: 100 clients, 5% participation

Method	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	Rounds ($\downarrow$)		Acc. (% , $\uparrow$)		Rounds ($\downarrow$)		Acc. (% , $\uparrow$)		Rounds ($\downarrow$)		Acc. (% , $\uparrow$)	
	75%	81%	500R	1000R	45%	50%	500R	1000R	35%	40%	500R	1000R
FedAvg [34]	308	460	82.08	87.20	403	693	46.74	53.39	602	860	32.93	42.41
FedAvgM [13]	260	383	83.48	87.54	291	435	50.87	54.10	317	537	37.60	43.13
FedAdam [40]	244	403	82.12	87.31	390	670	47.04	53.67	562	805	33.34	41.75
FedAdagrad [40]	301	507	80.88	87.03	385	633	48.98	53.13	655	879	31.03	41.57
FedYogi [40]	241	397	82.54	87.91	390	597	48.38	54.05	549	801	33.82	42.47
FAFED [51]	289	482	80.45	86.11	327	527	49.11	52.65	292	690	38.94	40.45
FedCada [61]	209	325	84.48	86.90	487	1000+	45.53	47.51	232	1000+	37.85	39.06
FADAS [49]	241	324	82.94	88.14	387	623	48.27	54.67	290	517	39.92	43.83
FedAdamom (ours)	213	307	85.62	88.93	276	392	52.32	57.58	244	353	43.43	47.38

Figure 2. **Loss landscapes** of models trained with FedAdamom (net) vs. FedAdam and FedAvgM (solid) on CIFAR10. FedAdamom achieves flatter minima and lower loss values w.r.t. FedAdam and FedAvgM.Table 2. Sensitivity of FedAdamom with respect to β_2 .

β_2	0.01	0.05	0.1	0.2	0.3
<i>Dir</i> (0.3)	88.75	88.93	88.73	88.42	88.13
<i>i.i.d.</i>	91.73	91.83	91.47	91.41	91.32

periments. This is because each client has fewer training samples and the data distribution becomes more heterogeneous. However, FedAdamom still shows strong robustness to data heterogeneity and low participation rates. More detailed convergence results for FedAdamom are provided in the Appendix A.

Hyperparameters sensitivity. We evaluate the performance of the algorithm under Dirichlet (0.3) and i.i.d. data splits by varying the values of β_2 . The experiments are conducted on the CIFAR-10 dataset using 1000 communication rounds and a 5% client participation rate among 100 clients. As shown in Tab. 2, the performance of FedAdamom remains stable when β_2 ranges from 0.01 to 0.3. Although there are slight fluctuations, the accuracy stays at a high level and reaches its peak at $\beta_2 = 0.05$.

6.3. The Loss Landscape and Escape Time

Global loss landscape. We visualize the global loss landscapes of different algorithms using the technique de-

scribed in [24], and conduct the experiments under the *Dirichlet*(0.3) data partition. The experiments are performed on the CIFAR-10 dataset with a CNN model. As shown in Fig. 2, the global model obtained by FedAdamom converges to a significantly flatter minimum in the global loss landscape compared with those obtained by FedAdam. This confirms that FedAdamom is capable of learning flatter minima.

The Escape Time Analysis. We empirically investigate the escape rates Γ of different algorithms, whose values are equal to the inverse of the mean escape time and depend on the sharpness of the minima. We use the Styblinski-Tang Function [54] as the test function, which has clear boundaries between loss valleys. We repeat experiments 100 times to estimate the escape rate Γ and the mean escape time τ . Since the escape time t is approximately a random variable obeying an exponential distribution, i.e., $t \sim \text{Exp}(\Gamma)$, the estimated escape rate can be written as $\Gamma = (100 - 2) / \sum_{i=1}^{100} t_i$. We use a scaling factor k to adjust the sharpness of minima, and the details are provided in the Appendix A. Fig. 3 shows that FedAdamom and FedAvgM satisfies $\log(\tau) = \mathcal{O}(H_{ae}^{-1})$, while FedAdam satisfies $\log(\tau) = \mathcal{O}(H_{ae}^{-\frac{1}{2}})$. This further confirms that FedAdam’s escape from sharp minima is less dependent on the Hessian compared to FedAvgM and FedAdamom.

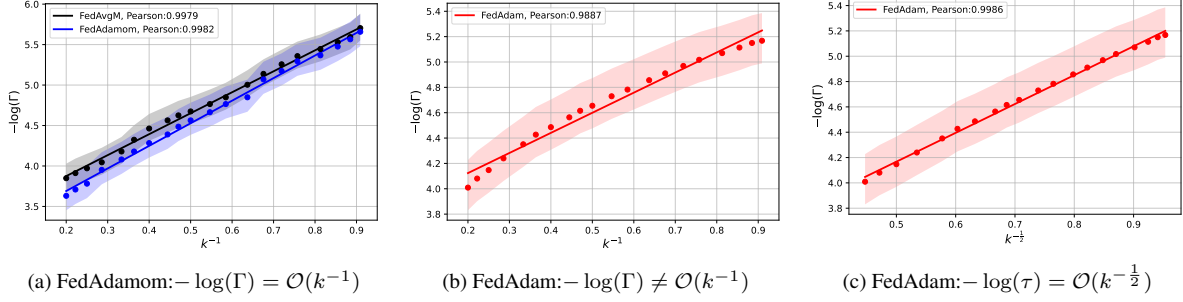


Figure 3. Flat Minima Selection: The log-scale mean escape time $-\log(\Gamma)$ with the 95% confidence interval is displayed. We empirically verify that FedAdamom and FedAvgM satisfy $-\log(\Gamma) = \mathcal{O}(k^{-1})$ but FedAdam does not. Instead, we observe that $-\log(\tau) = \mathcal{O}(k^{-\frac{1}{2}})$ holds better for Adam.

Table 3. Results on the realistic datasets involving feature skewness and data imbalance between clients.

Method	FEMNIST				CelebA				Shakespeare			
	Acc. (% , $\uparrow$)		Rounds ($\downarrow$)		Acc. (% , $\uparrow$)		Rounds ($\downarrow$)		Acc. (% , $\uparrow$)		Rounds ($\downarrow$)	
	250R	500R	75%	80%	250R	500R	80%	88%	500R	1000R	42%	45%
FedAvg [34]	80.79	82.24	72	241	88.16	89.15	106	174	45.13	46.58	98	499
FedAvgM [13]	81.23	82.32	69	216	88.23	89.41	83	153	45.32	46.71	83	472
FedAdam [40]	78.99	81.51	64	320	87.26	87.65	77	500+	45.21	46.37	67	785
FedAdagrad [40]	78.61	80.62	68	332	86.73	87.22	81	500+	44.51	45.88	73	897
FedYogi [40]	79.59	82.11	58	318	87.54	88.17	72	456	46.37	47.10	62	432
FAFED [51]	79.14	81.66	66	347	86.72	87.39	74	500+	44.31	45.95	87	841
FedCAda [61]	73.62	79.25	303	500+	85.58	86.31	62	500+	45.01	46.42	58	480
FADAS [49]	79.34	82.06	56	314	87.77	88.58	69	345	46.73	47.40	65	364
FedAdamom (ours)	81.41	82.85	62	177	88.76	89.95	68	133	46.94	48.02	64	289

6.4. Experiments on Realistic Datasets

We further evaluate our method on two additional real-world datasets: FEMNIST and CelebA, both provided by the LEAF framework [3]. These datasets exhibit various types of non-i.i.d. characteristics, including feature skewness and data imbalance across clients. In this set of experiments, we use 2000 clients, follow the data partition protocol in [3], and randomly sample 5 clients per round for participation. For model architectures, we adopt a two-layer CNN for FEMNIST and a four-layer CNN for CelebA. As shown in Tab. 3, FedAdamom consistently outperforms other baseline methods on both datasets in most cases, highlighting its robustness to heterogeneous data distributions, including differences in data volume and feature alignment across clients.

We further evaluate FedAdamom in another task domain, namely next-word prediction, using the Shakespeare dataset from the LEAF framework. This task also exhibits substantial data imbalance across clients. In this experiment, we employ an LSTM as the backbone network and set the client participation rate to 5% in each communication round. The results in Tab. 3 show that FedAdamom also performs well

on the language modeling task.

7. Conclusion

This paper investigates the distinct roles of the adaptive global learning rate and global momentum in saddle-point escape and flat-minimum selection within FL. We introduce diffusion theory to systematically characterize their dynamical behaviors during global saddle-point escape and flat-minimum selection. Under reasonable assumptions, our theoretical analysis explains why FedAdam excels at escaping saddle points but performs poorly in selecting flatter minima. We further propose a novel adaptive federated optimization framework, FedAdamom, which adaptively adjusts the global momentum hyperparameter at the parameter level. This design enables Hessian-independent saddle-point escape acceleration while preserving favorable flat-minimum selection properties. Furthermore, we provide a formal mathematical proof for the convergence rate of FedAdamom. Our empirical analysis and extensive experimental evaluations demonstrate that FedAdamom generalizes significantly better than popular FedAdam variants and FedAvg.

Acknowledgements

This research is supported by the National Science and Technology Major Project (2022ZD0116800), the Program of the National Natural Science Foundation of China (12425114, 62141605, 12201026, 12301305, 62441617, 12501702), and the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing; by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award MOE-T2EP20125-0005); and by the China Scholarship Council (Grant 202406020279).

References

- [1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *International Conference on Learning Representations*, 2021. 2
- [2] Debora Caldarola, Pietro Cagnasso, Barbara Caputo, and Marco Ciccone. Beyond local sharpness: Communication-efficient global sharpness-aware minimization for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25187–25197, 2025. 2
- [3] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018. 6, 8, 1
- [4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019. 2
- [5] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *Advances in Neural Information Processing Systems*, 35: 23439–23451, 2022. 2
- [6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive sub-gradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 3
- [7] Lidia Fantauzzo, Eros Fanì, Debora Caldarola, Antonio Tavera, Fabio Cermelli, Marco Ciccone, and Barbara Caputo. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11504–11511. IEEE, 2022. 2
- [8] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021. 2
- [9] 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. 6
- [10] Yuting He, Yiqiang Chen, Xiaodong Yang, Hanchao Yu, Yihua Huang, and Yang Gu. Learning critically: Selective self-distillation in federated learning on non-iid data. *IEEE Transactions on Big Data*, 10(6):789–800, 2022. 3
- [11] Yuting He, Yiqiang Chen, Xiaodong Yang, Yingwei Zhang, and Bixiao Zeng. Class-wise adaptive self distillation for heterogeneous federated learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, Virtual*, 2022. 3
- [12] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 2
- [13] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 4, 6, 7, 8, 1, 2, 3
- [14] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. 2
- [15] Long Jin, Han Nong, Liangming Chen, and Zhenming Su. A method for enhancing generalization of adam by multiple integrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4147–4155, 2025. 3, 4
- [16] Pavol Kalinay and Jerome K Percus. Phase space reduction of the one-dimensional fokker-planck (kramers) equation. *Journal of Statistical Physics*, 148(6):1135–1155, 2012. 4, 10
- [17] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020. 2
- [18] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021. 3, 6
- [19] Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12385–12394, 2024. 2, 1
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 3
- [21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [22] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6
- [23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 6
- [24] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 7, 3
- [25] Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathemat-

- ical foundations. *Journal of Machine Learning Research*, 20 (40):1–47, 2019. 3
- [26] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 3
- [27] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 2
- [28] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *Proceedings of ICLR*, 2020. 2
- [29] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021. 2
- [30] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don’t use large mini-batches, use local sgd. *Proceedings of ICLR*, 2020. 2
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3
- [32] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019. 3
- [33] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. 3
- [34] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 3, 6, 7, 8
- [35] Jiaxu Miao, Zongxin Yang, Leilei Fan, and Yi Yang. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8042–8052, 2023. 2
- [36] Zhang Michael, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems*, 2019. 1
- [37] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021. 2
- [38] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pages 18250–18280. PMLR, 2022. 3
- [39] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. 3
- [40] 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. 2, 3, 4, 6, 7, 8, 1
- [41] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990. PMLR, 2014. 3
- [42] Donald Shenaj, Eros Fani, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro Zanuttigh, and Barbara Caputo. Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 444–454, 2023. 2
- [43] Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30651–30660, 2025. 3
- [44] Yifan Shi, Kang Wei, Li Shen, Yingqi Liu, Xueqian Wang, Bo Yuan, and Dacheng Tao. Toward the flatter landscape and better generalization in federated learning under client-level differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11632–11643, 2025. 2
- [45] Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra. Escaping saddle points with adaptive gradient methods. In *International Conference on Machine Learning*, pages 5956–5965. PMLR, 2019. 2, 5
- [46] Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [47] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020. 2
- [48] Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *International conference on machine learning*, pages 22802–22838. PMLR, 2022. 3
- [49] Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. FADAS: Towards federated adaptive asynchronous optimization. In *International Conference on Machine Learning*, 2024. 3, 6, 7, 8, 2
- [50] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017. 2
- [51] Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster adaptive federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10379–10387, 2023. 3, 6, 7, 8, 2
- [52] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions*

- on *Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024. [3](#)
- [53] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021. [2](#), [3](#), [4](#), [5](#), [6](#), [8](#)
 - [54] Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International conference on machine learning*, pages 24430–24459. PMLR, 2022. [2](#), [4](#), [5](#), [7](#), [3](#)
 - [55] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20866–20875, 2022. [2](#)
 - [56] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *Proceedings of ICLR*, 2021. [2](#)
 - [57] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pages 39879–39902. PMLR, 2023. [3](#)
 - [58] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018. [3](#)
 - [59] Hao Zhang, Chenglin Li, Wenrui Dai, Ziyang Zheng, Junni Zou, and Hongkai Xiong. Stabilizing and accelerating federated learning on heterogeneous data with partial client participation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [6](#)
 - [60] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023. [2](#)
 - [61] Liuzhi Zhou, Yu He, Kun Zhai, Xiang Liu, Sen Liu, Xingjun Ma, Guangnan Ye, and Hongfeng Chai. Fedcada: Adaptive client-side optimization for accelerated and stable federated learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. [3](#), [6](#), [7](#), [8](#), [2](#)
 - [62] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020. [2](#), [4](#)