FedCVC: Federated Primal-Dual Learning with Client-Driven Virtual Compensation for Mitigating Dual Drift

Supplementary Material

8. Details on Experiments

We conduct experiments on two 4090 GPUs with 24GB RAM each, using PyTorch 3.9. All experiments were conducted with five different random seeds: [0, 20, 23, 40, 47].

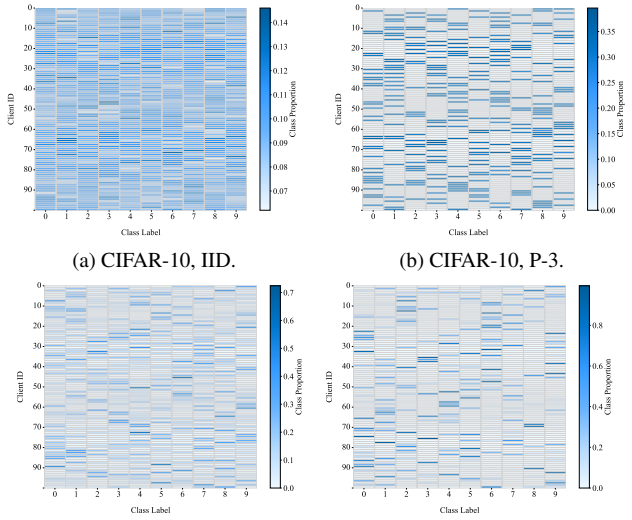
We evaluate FedCVC on five benchmark datasets: MNIST, CIFAR-10, CIFAR-100, Tiny-ImageNet, and AG-News. Experiments are conducted under two FL scenarios with 100 clients (10% participation rate) and 500 clients (2% participation rate). MNIST consists of 60,000 training and 10,000 test grayscale images (28x28) across 10 classes. CIFAR-10 and CIFAR-100 each have 50,000 training and 10,000 test RGB images (32x32), with 10 and 100 classes, respectively. Tiny-ImageNet includes 100,000 training and 10,000 test RGB images (64x64) spanning 200 classes. AG-News comprises 40,000 training and 10,000 test samples from 4 news categories. For MNIST, we employ a LeNet-based architecture with three fully connected layers (MNIST-2NN). On CIFAR-10, CIFAR-100, and Tiny-ImageNet, we implement ResNet-18 with group normalization instead of batch normalization, adhering to FL best practices. Additionally, we evaluate CIFAR-10 and CIFAR-100 using a modified LeNet architecture that features two convolutional layers, max-pooling, and three fully connected layers for configuration alignment. The AG-News model consists of an embedding layer followed by a two-layer classifier (AG-NN).

8.1. Data Partitions

To assess data heterogeneity, we partition client data using Dirichlet and Pathological distributions. For the Dirichlet distribution, we set concentration parameters to 0.6 and 0.1 to simulate mild, moderate, and severe heterogeneity levels, respectively. Pathological partitioning employs parameter pairs (6,3) for MNIST and CIFAR-10, (20,10) for CIFAR-100, (40,20) for Tiny-ImageNet, and (3,1) for AG-News. We also include standard IID partitioning as a baseline, resulting in six distinct data distribution scenarios for each dataset. Fig. 10 illustrates these partitioning schemes across 100 clients for CIFAR-10.

8.2. Data Augmentation

We adopt established normalization schemes with dataset-specific parameters. For MNIST, images are normalized using a mean of 0.1307 and a standard deviation of 0.3081. CIFAR-10 employs per-channel means of [0.491, 0.482, 0.447] and standard deviations of [0.247, 0.243, 0.262]. Similarly, CIFAR-100 uses means [0.5071, 0.4867, 0.4408] and standard deviations [0.2675, 0.2565, 0.2761]. Tiny-ImageNet is uniformly normalized with mean and standard deviation both set to [0.5, 0.5, 0.5] across all channels. Furthermore, we apply standard techniques, including random horizontal flipping and cropping. Input images for MNIST and CIFAR datasets are padded with 4 pixels before cropping, whereas Tiny-ImageNet uses 8-pixel padding due to its higher resolution. This consistent padding approach ensures uniform augmentation effects across varying input sizes. All experiments adhere to the same preprocessing pipeline to maintain fair comparisons.



(a) CIFAR-10, IID. (b) CIFAR-10, P-3.

(c) CIFAR-10, Dir-0.6. (d) CIFAR-10, Dir-0.1.

Figure 10. The data distribution (four data partitions) of CIFAR-10 dataset across 100 clients.

8.3. Baselines

- **FedAvg** serves as the foundational framework for FL, providing a standard benchmark for task-specific evaluations.
- **FedProx** uses a proximal term to constrain the distance between the global model and local models in Euclidean space, mitigating client drift.
- **Scaffold** introduces auxiliary variables for variance reduction to mitigate client drift.
- **FedDyn** introduces dual variables and employs primal-dual updates to alternately optimize model parameters and their dual variables, dynamically achieving consistency between local updates and global consensus.
- **A-FedPD** introduces a server-side virtual dual update

mechanism based on FedDyn to simulate the dual updates of non-participating clients, effectively mitigating the DD.

- **FedSpeed** introduces the SAM optimizer based on FedDyn and combines SGD and SAM optimizers to form smoother updates, enhancing the generalization capability of the global model.
- **FedSMOO** applies primal-dual updates separately to parameters and sharpness measures to enforce global and local consistency.
- **FedLESAM-D** is a federated sharpness-aware optimization method based on FedDyn that uses client-side local estimation of the global model update direction as the perturbation direction, aiming to improve model generalization and reduce computational overhead.
- **A-FedPDSAM** replaces the SGD optimizer in A-FedPD with the SAM local optimizer.
- **FedGLOSS**, based on FedDyn, uses SAM on the server-side and leverages the pseudo-gradient from the previous round to approximate sharpness, directly optimizing global flatness to improve the model's generalization ability and robustness while reducing communication overhead.

For fairness, we compare FedCVC with methods based on the SGD optimizer and FedCVCSAM with methods based on the SAM optimizer, with the algorithm flow of FedCVCSAM shown in Algorithm 2 .

8.4. Hyperparameters Selection

We conduct comprehensive grid searches to identify optimal hyperparameters for all methods. The global learning rate is selected from [0.1, 0.5, 1.0], with 1.0 proving optimal. For local learning rates, we evaluate values in [0.01, 0.05, 0.1, 0.5, 1.0], and 0.1 yields the best performance. Weight decay is tuned from [1e-4, 5e-4, 1e-3, 0.01], with 1e-3 achieving the best results. Learning rate decay is explored across [0.995, 0.998, 0.9998, 1], where 0.998 is optimal for 100 clients and 0.9998 for 500 clients. Batch size is searched over [10, 20, 50, 100]; for 100 clients, 50 performs best, while for 500 clients, 20 is optimal for FedDyn, FedSpeed, FedSMOO, and FedLESAM-D, and 50 for the other methods.

Specifically, for methods that require a penalized weight α, we perform grid search over [0.001, 0.01, 0.05, 0.1, 0.2, 0.5]. For the discount factor β of FedCVC and FedCVCSAM, we conduct grid search over [0.01, 0.05, 0.1, 0.2, 0.5]. For SAM-based methods, we search for the perturbation coefficient ρ in [0.01, 0.05, 0.1, 0.2, 0.5, 1.0] and for SAM eps in [1e-2, 1e-5, 1e-8].

8.5. Settings of the m $\sim$ n dual update operation

Fig. 11 depicts the procedure for the $m \sim n$ dual update operations. Under full participation of 100 clients, each client reuses stale dual variables from the previous $m-1$ rounds (with $m \geq 1$) to execute local primal-dual updates and stores all local updates from these rounds. In the m-th round, the clients first perform primal updates using the stale dual variables. Subsequently, they utilize the stored local updates from the most recent n rounds to compensate for the staleness, deriving new dual variables. These new dual variables are then combined with the local updates from the m-th round to finalize the dual update.

Algorithm 2 FedCVCSAM Algorithm

Input: Global model parameter $\mathbf{w}_0$
Parameter: Penalized weight α, discount factor β, local datasets D_i, number of clients N, number of communication rounds T, number of local epochs K, learning rate η, local drifts variable $\boldsymbol{\lambda}_i^0 = 0$, silence table $\mathbf{R}^0 = \mathbf{1}$
Output: The final model parameter $\mathbf{w}_0^T$

for $t = 0$ to $T-1$ **do**
  Randomly select client-set C^t at round t.
  for client $i \in C^t$ **do**
    Broadcast $\mathbf{w}_0^t$ to client i and set $\mathbf{w}_i^t = \mathbf{w}_0^t$.
    for $k = 0$ to $K-1$ **do**
      Randomly sample a mini-batch ξ_i and compute the stochastic gradients:
      $\mathbf{g}_i(\mathbf{w}_i^{t,k}) = \nabla f_i(\mathbf{w}_i^{t,k}; \xi_i)$.
      Perform gradient ascent: $\breve{\mathbf{w}}_{i,k}^t = \mathbf{w}_{i,k}^t + \rho \mathbf{g}_i(\mathbf{w}_i^{t,k})$
      Compute the stochastic gradients at $\breve{\mathbf{w}}_{i,k}^t$: $\breve{\mathbf{g}}_i(\mathbf{w}_i^{t,k}) = \nabla f_i(\breve{\mathbf{w}}_{i,k}^t; \xi_i)$
      Update the local model parameter:
      $\mathbf{w}_i^{t,k+1} = \mathbf{w}_i^{t,k} - \eta_i(\breve{\mathbf{g}}_i(\mathbf{w}_i^{t,k}) + \alpha(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t)$.
    end for
    Set $\mathbf{w}_i^{t+1} = \mathbf{w}_i^{t,K}$.
    Update dual variable usingEq. (12) and Eq. (13).
    Set R_i^t = 1.
    Communicate $\mathbf{w}_i^{t+1} - \mathbf{w}_0^t$ to the server.
  end for
  for client $i \notin C^t$ **do**
    $\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t$.
    $R_i^t = R_i^t + 1$.
  end for
  Update global drifts variable using Eq. (14).
  Update global model using Eq. (15).
end for

9. More Experiments Results

9.1. Performance on MNIST and AG-News

As shown in Table Tab. 3, FedCVC consistently outperforms SOTA methods across most data heterogeneity scenarios. Notably, FedCVCSAM demonstrates superior

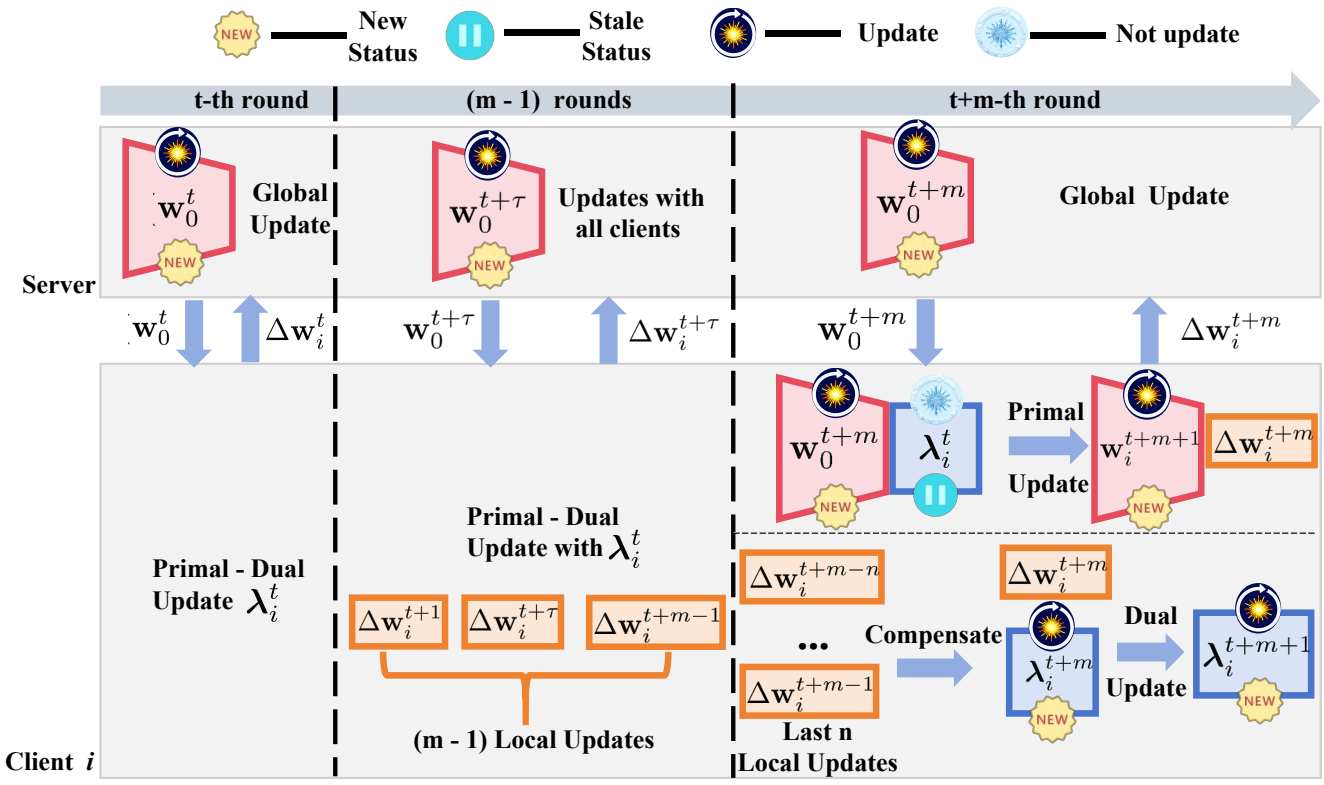


Figure 11. The framework of $m \sim n$ dual update operation, where τ ranges from 0 to m-1, representing the rounds from 0 to m-1.

Methods	MNIST			AG-News		
	IID	Dir-0.6	Dir-0.1	IID	Dir-0.6	Dir-0.1
FedAvg	98.15±0.01 (0.00)	98.09±0.02(0.00)	97.09±0.06 (0.00)	89.66±0.05 (0.00)	88.38±0.06 (0.00)	84.13±0.09 (0.00)
FedProx	98.18±0.01 (↑0.03)	98.14±0.02 (↑0.05)	97.77±0.05 (↑0.68)	89.79±0.03 (↑0.13)	89.08±0.06 (↑0.70)	84.39±0.07 (↑0.26)
Scaffold	98.45±0.01 (↑0.30)	98.51±0.02 (↑0.42)	98.27±0.05 (↑1.18)	86.71±0.22 (↓ 2.95)	86.42±0.36 (↓1.96)	84.80±1.21 (↑0.67)
FedDyn	98.52±0.01 (↑0.37)	98.39±0.03 (↑0.30)	98.19±0.05 (↑1.10)	51.17±0.21 (↓38.49)	40.51±0.37 (↓47.87)	26.67±0.85 (↓57.96)
A-FedPD	98.04±0.03 (↓0.11)	97.80±0.08 (↓0.29)	96.93±0.13 (↓0.16)	48.91±0.28 (↓40.75)	30.32±0.47 (↓58.06)	25.00±0.00 (↓59.13)
FedCVC	98.54±0.01 (↑**0.39**)	98.49±0.02 (↑**0.40**)	98.39±0.07 (↑**1.30**)	90.51±0.03 (↑**0.85**)	89.70±0.04 (↑**1.32**)	87.09±0.07 (↑**2.96**)
FedSpeed	98.74±0.01 (↑0.59)	98.64±0.02 (↑0.55)	98.28±0.03 (↑1.19)	90.33±0.02 (↑0.67)	89.47±0.05 (↑1.09)	86.24±0.06 (↑2.11)
FedSMOO	98.62±0.01 (↑0.47)	98.54±0.02 (↑0.45)	98.29±0.04 (↑1.20)	90.32±0.02 (↑0.66)	89.50±0.04 (↑1.12)	86.21±0.07 (↑2.08)
FedLESAM-D	98.54±0.01 (↑0.39)	98.43±0.06 (↑0.34)	98.17±0.03 (↑1.08)	90.33±0.02 (↑0.67)	89.47±0.05 (↑1.09)	86.32±0.07 (↑2.19)
A-FedPDSAM	98.46±0.01 (↑0.31)	97.85±0.06 (↓0.24)	97.67±0.07 (↓0.58)	49.55±0.32 (↓40.11)	31.67±0.63 (↓56.71)	25.00±0.00 (↓59.13)
FedGLOSS	98.48±0.01 (↑0.33)	98.46±0.02 (↑0.37)	98.04±0.06 (↑0.95)	90.38±0.03 (↑0.72)	89.63±0.03 (↑1.25)	86.76±0.09 (↑2.63)
FedCVCSAM	98.78±0.01 (↑**0.63**)	98.69±0.01 (↑**0.60**)	98.41±0.02 (↑**1.32**)	90.63±0.02 (↑**0.97**)	89.91±0.03 (↑**1.53**)	87.39±0.06 (↑**3.26**)

Table 3. A comparison of Test Top-1 accuracy across different algorithms under varying data heterogeneity distributions on MNIST and AG-News. The bold text represents the optimal values.

performance across all heterogeneity levels by effectively leveraging the SAM optimizer to reduce inter-client model variance, thereby enhancing the precision of client-driven virtual compensation mechanism.

For the classification benchmark, Tab. 3 reveals that FedCVC obtains SOTA performance among SGD-based FL approaches, while FedCVCSAM achieves the best results in the SAM-based methods. The experimental results demonstrate that FedCVC significantly outperforms both FedDyn and A-FedPD, primarily due to client-driven virtual compensation mechanism which guarantees global model convergence while simultaneously achieving superior accuracy. This validates the mechanism's effectiveness in resolving convergence instability and performance deterioration. Comparative analysis shows that A-FedPD with SGD optimization suffers from optimization errors arising from misalignment between server-side virtual dual updates and actual client updates, resulting in unstable convergence be-

Methods	CIFAR-10			CIFAR-100		
	IID	Dir-0.6	Dir-0.1	IID	Dir-0.6	Dir-0.1
FedAvg	81.85±0.02 (0.00)	79.74±0.13 (0.00)	75.94±0.16 (0.00)	39.54±0.03 (0.00)	41.82±0.04 (0.00)	39.89±0.08 (0.00)
FedProx	81.74±0.03 (↓0.11)	79.62±0.05 (↓0.12)	76.39±0.11 (↑0.45)	41.42±0.02 (↑1.88)	41.78±0.03 (↓0.04)	40.55±0.10 (↑0.66)
Scaffold	84.51±0.02 (↑2.66)	83.57±0.05 (↑3.83)	79.42±0.06 (↑3.48)	49.32±0.06 (↑9.78)	49.31±0.05 (↑7.49)	45.92±0.08 (↑6.03)
FedDyn	84.77±0.01 (↑2.92)	82.06±0.06 (↑2.32)	78.49±0.11 (↑2.55)	50.12±0.04 (↑10.58)	49.83±0.07 (↑10.58)	46.28±0.08 (↑8.01)
A-FedPD	84.55±0.07 (↑2.70)	83.36±0.15 (↑3.62)	77.46±0.26 (↑1.52)	48.22±0.10 (↑8.68)	47.17±0.10 (↑5.35)	45.16±0.18 (↑5.27)
FedCVC	85.15±0.02 (↑**3.30**)	84.05±0.03 (↑**4.31**)	79.69±0.07 (↑**3.75**)	53.21±0.03 (↑**13.67**)	53.03±0.03 (↑**11.21**)	47.76±0.04 (↑**7.87**)
FedSpeed	85.99±0.01 (↑4.14)	85.91±0.05 (↑6.17)	82.15±0.16 (↑6.21)	55.93±0.01 (↑16.39)	54.95±0.03 (↑13.13)	52.22±0.05 (↑12.33)
FedSMOO	86.36±0.01 (↑4.51)	85.69±0.03 (↑5.95)	81.75±0.09 (↑5.81)	54.92±0.02 (↑15.38)	54.93±0.03 (↑13.11)	51.75±0.05 (↑11.86)
FedLESAM-D	85.23±0.01 (↑3.38)	84.42±0.06 (↑4.68)	79.06±0.14 (↑3.12)	54.23±0.01 (↑14.69)	54.02±0.02 (↑12.20)	52.13±0.06 (↑12.24)
A-FedPDSAM	86.36±0.01 (↑4.51)	85.79±0.10 (↑6.05)	81.11±0.47 (↑5.17)	54.95±0.02 (↑15.41)	53.77±0.04 (↑11.95)	51.54±0.12 (↑14.62)
FedGLOSS	86.29±0.01 (↑4.44)	85.42±0.02 (↑5.68)	81.62±0.11 (↑11.95)	53.91±0.02 (↑13.37)	54.20±0.02 (↑12.38)	51.45±0.13 (↑14.56)
FedCVCSAM	86.68±0.01 (↑**4.83**)	85.98±0.05 (↑**6.24**)	82.69±0.15 (↑**6.75**)	56.80±0.02 (↑**17.26**)	57.04±0.03 (↑**15.22**)	53.60±0.04 (↑**13.71**)

Table 4. A comparison of Test Top-1 accuracy across different algorithms under varying data heterogeneity distributions on CIFAR-10 and CIFAR-100(The Backbone is LeNet). The bold text represents the optimal values.

havior. In contrast, FedCVCSAM leverages SAM's smooth optimization landscape to enhance model performance by reducing estimation errors in client-driven virtual compensation. While SAM optimization improves FedSpeed's resilience against data stratification and stabilizes its convergence, our results indicate it cannot fully compensate for the estimation bias introduced by A-FedPDSAM's virtual dual updates.

9.2. Performance on CIFAR-10/100 with Different Backbones.

We assessed the performance of FedCVC and FedCVCSAM across diverse model architectures by conducting experiments with a lightweight LeNet backbone on the CIFAR-10 and CIFAR-100 datasets. As indicated in Tab. 4, even with this simpler architecture, both FedCVC and FedCVCSAM attained SOTA convergence accuracy, highlighting their robustness to model complexity variations. Notably, A-FedPD exhibits strong performance on simple models, attributed to reduced estimation errors in virtual updates under such conditions.

9.3. Communication Overhead in More Settings

We provide communication overhead and computational time for CIFAR-10 and CIFAR-100 across various FL scenarios, as detailed in Tabs. 5 to 7. Results demonstrate that FedCVC and FedCVCSAM typically achieve target accuracy with minimal communication and computational costs. This efficiency arises from a novel client-driven virtual compensation mechanism that preserves the freshness of stale dual variables, reducing optimization errors and accelerating convergence.

A key benefit is communication efficiency: the mechanism enhances performance while matching FedAvg's per-round communication overhead. In contrast, Scaffold, A-FedPD, FedSMOO, and A-FedPDSAM incur double FedAvg's communication costs. Although they may require fewer rounds to reach target accuracy, their total communication overhead exceeds that of FedCVC. Moreover, FedGLOSS, despite fewer communication rounds in some cases, involves additional server-side computation, resulting in longer overall computation times than FedCVCSAM. Thus, FedCVC and FedCVCSAM offer substantial advantages in communication and computation efficiency, making them well-suited for resource-constrained federated learning environments.

Methods	Comm.Round ↓		Comm.Cost (MiB) ↓		Compu.Time (min) ↓	
T.Acc (SGD)	30%	40%	30%	40%	30%	40%
FedAvg	–	–	–	–	–	–
FedProx	–	–	–	–	–	–
Scaffold	**360**	1393	612	2368	**34**	132
FedDyn	–	–	–	–	–	–
A-FedPD	449	–	763	–	48	–
FedCVC	457	**920**	**388**	**782**	36	**72**
T.Acc (SAM)	38%	43%	38%	43%	38%	43%
FedSpeed	840	1434	714	1218	100	170
FedSMOO	686	1017	1166	1728	133	197
FedLESAM-D	1323	–	1809	–	197	–
A-FedPDSAM	449	708	763	1203	57	90
FedGLOSS	1028	–	1747	–	162	–
T FedCVCSAM	**370**	**506**	**314**	**430**	**45**	**61**

Table 5. Communication rounds and communication costs (in gigabytes) when reaching the target accuracy on CIFAR-100 dataset (Dir-0.6). The number of clients is 500, with a participation rate of 2%. "-" indicates failure to achieve target accuracy.

9.4. Data heterogeneity on Tiny-ImageNet

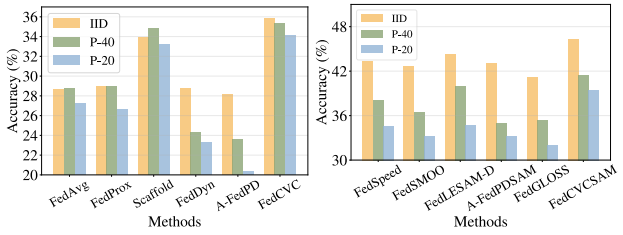
To evaluate the effectiveness of FedCVC on complex datasets and data heterogeneity, we report the accuracy of the Tiny-ImageNet dataset under various heterogeneity distributions, as shown in Fig. 12. The results demonstrate that FedCVC and FedCVCSAM outperform SGD-based and SAM-based baseline methods, respectively, across dif-

Methods	Comm.Round ↓		Comm.Cost (MiB) ↓		Compu.Time (min) ↓	
T.Acc (SGD)	81%	83%	81%	83%	81%	83%
FedAvg	713	–	606	–	46	–
FedProx	–	–	–	–	–	–
Scaffold	287	591	487	1007	27	56
FedDyn	230	1546	195	1314	23	156
A-FedPD	379	–	644	–	40	–
FedCVC	**210**	**327**	**179**	**278**	**16**	**26**
T.Acc (SAM)	84%	86%	84%	86%	84%	86%
FedSpeed	459	768	390	652	55	91
FedSMOO	335	711	570	1209	65	138
FedLESAM-D	964	–	820	–	144	–
A-FedPDSAM	420	785	714	1335	53	99
FedGLOSS	**287**	603	488	1022	48	95
T FedCVCSAM	396	**602**	**337**	**512**	**45**	**73**

Table 6. The number of communication rounds, communication cost, and computation time required to achieve the target accuracy on CIFAR-10. The number of clients is 100, with a participation rate of 10%. "–" indicates that the target accuracy was not achieved.

Methods	Comm.Round ↓		Comm.Cost (MiB) ↓		Compu.Time (min) ↓	
T.Acc (SGD)	42%	48%	42%	48%	42%	48%
FedAvg	931	–	791	–	–	–
FedProx	–	–	–	–	–	–
Scaffold	144	473	244	802	14	44
FedDyn	–	–	–	–	–	–
A-FedPD	–	–	–	–		–
FedCVC	**129**	**469**	**110**	**399**	**9**	**35**
T.Acc (SAM)	48%	52%	48%	52%	48%	52%
FedSpeed	238	511	202	434	28	61
FedSMOO	230	–	392	–	44	–
FedLESAM-D	324	441	275	375	48	66
A-FedPDSAM	**187**	930	318	1581	23	117
FedGLOSS	262	–	446	–	43	–
T FedCVCSAM	197	**394**	**167**	**335**	**22**	**68**

Table 7. Communication rounds and communication costs (in gigabytes) when reaching the target accuracy on CIFAR-100 dataset (Dir-0.6). The number of clients is 100, with a participation rate of 10%. "–" indicates failure to achieve target accuracy.



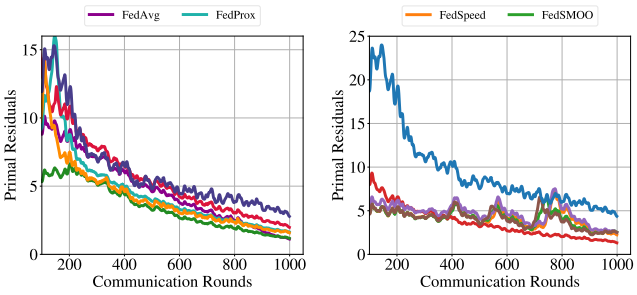
(a) Tiny-Image, SGD. (b) Tiny-Image, SAM.

Figure 12. Top-1 test accuracy under different data heterogeneity.

ferent heterogeneity levels, confirming their robustness.

9.5. Primal Residuals

We present the primal residuals for various methods, defined as $\frac{1}{N}\sum_i \|\mathbf{w}_0^t - \mathbf{w}_i^t\|$, which serve as a consistency measure. As shown in Fig. 13, FedCVC and FedCVCSAM

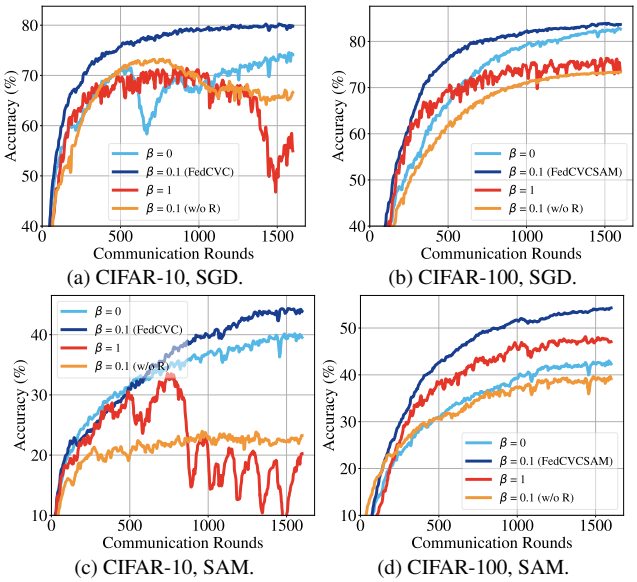


(a) Primal. (b) Primal.

Figure 13. Dual Residuals and Ration of Primal and Dual.

exhibit the highest primal residuals, demonstrating their effectiveness in maintaining consistency between local models and the global consensus. Additionally, Fig. 6 indicates that these methods achieve the highest dual residuals, corresponding to global updates, and preserve consistency while converging rapidly. The low and stable primal/dual residual ratio further implies a balanced progression of both residuals, resulting in a smooth convergence process.

9.6. Training Curve of Ablation Study



(a) CIFAR-10, SGD. (b) CIFAR-100, SGD.
(c) CIFAR-10, SAM. (d) CIFAR-100, SAM.

Figure 14. Training curve of ablation study of FedCVC and FedCVCSAM.

We conduct ablation studies to evaluate the training accuracy of FedCVC under extreme client participation rates, as illustrated in Fig. 14. Results indicate that disabling the client-driven virtual compensation mechanism ($\beta = 0$) slows convergence and reduces accuracy. Similarly, excluding the discount factor ($\beta = 1$) also impairs convergence and accuracy, as compensating for highly stale dual variables introduces additional errors, consistent with the $m \sim$

observation. Furthermore, replacing the client silence table with a constant R_i^t induces convergence oscillations due to inadequate adaptive virtual compensation, leading to over- or under-compensation.

10. Proofs

10.1. Preliminaries

We consider a federated learning system with N clients, where each client i maintains a local dataset D_i drawn from distribution $\mathcal{D}_i$ with D samples. The global data distribution is given by $\mathcal{D} = \bigcup_{i=1}^{N} \mathcal{D}_i$. During each communication round t, the server selects a subset C^t of M active clients for training. To properly account for partial client participation, we introduce the virtual sequences $\tilde{\mathbf{w}}_i^t$ and $\tilde{\boldsymbol{\lambda}}_i^t$ to assume that all clients participate the training at t-th round. Due to Eq. (3), we have the following update rules:

$$\tilde{\mathbf{w}}_i^{t,k+1} = \tilde{\mathbf{w}}_i^{t,k} - \eta(\mathbf{g}_i(\tilde{w}_i^{t,k}) - \boldsymbol{\lambda}_i^t + \alpha(\tilde{\mathbf{w}}_i^{t,k} - \mathbf{w}_0^t)), \quad \forall i \in N. \tag{18}$$

$$\tilde{\boldsymbol{\lambda}}_i^{t+1} = \boldsymbol{\lambda}_i^t + \alpha(1 + \beta R_i^t)(\mathbf{w}_0^t - \tilde{\mathbf{w}}_i^{t+1}), \quad \forall i \in N. \tag{19}$$

In addition, we have $\tilde{\mathbf{w}}_i^t = \mathbf{w}_i^t$, $\tilde{\boldsymbol{\lambda}}_i^t = \boldsymbol{\lambda}_i^t$ if $i \in C^t$. For ease of subsequent proofs, we define some commonly used notations as follows:

$$\begin{aligned} C_i^t &= (1 + \beta R_i^t) \\ Q_i^t &= (1 + C_i^t)\alpha\eta r \\ P_i^t &= C_i^t \alpha\eta r, \end{aligned} \tag{20}$$

where $r = \sum_{k=0}^{K-1} r_k$ and $r_k = (1 - \alpha\eta)^{K-1-k}$.

10.2. Client-Driven Virtual Compensation

Based on Definition 1, according to Eq. (5), the dual variable update at the $(t + \tau)$-th round is given by:

$$\begin{aligned} &\boldsymbol{\lambda}_i^{t+\tau+1} - \boldsymbol{\lambda}_i^{t+\tau} = \alpha(\mathbf{w}_0^{t+\tau} - \mathbf{w}_i^{t+\tau+1}) \\ &= -\alpha\Delta\mathbf{w}_i^{t+\tau} = -\alpha b\Delta\mathbf{w}_0^{t+\tau} - \alpha\boldsymbol{\epsilon}_1^{\tau}. \end{aligned} \tag{21}$$

Considering the accumulated changes across rounds from $t + 1$ to $t + \tau_0 - 1$, we have:

$$\boldsymbol{\lambda}_i^{t+\tau_0} - \boldsymbol{\lambda}_i^{t+1} = -\alpha b \sum_{\tau=1}^{\tau_0-1} \Delta\mathbf{w}_0^{t+\tau} - \alpha \sum_{\tau=1}^{\tau_0-1} \boldsymbol{\epsilon}_1^{\tau}. \tag{22}$$

Since the primal-dual update ensures global consistency, we make the following approximation:

$$\sum_{\tau=1}^{\tau_0-1} \Delta\mathbf{w}_0^{t+\tau} = (\tau_0 - 1)\Delta\mathbf{w}_0^{t+\tau_0} \tag{23}$$

Using Definition 1 again, we can obtain:

$$\boldsymbol{\lambda}_i^{t+\tau_0} - \boldsymbol{\lambda}_i^{t+1} \approx -\alpha(\tau_0 - 1)\Delta\mathbf{w}_i^{t+\tau_0} + \alpha(\tau_0 - 1)\boldsymbol{\epsilon}_1^{\tau_0} - \alpha \sum_{\tau=1}^{\tau_0-1} \boldsymbol{\epsilon}_1^{\tau}. \tag{24}$$

In consecutive communication rounds, the global model undergoes minimal changes, and the client drift for client i can be approximated as constant. Thus, the discrepancy between $\boldsymbol{\epsilon}_1^{\tau_0}$ and $\boldsymbol{\epsilon}_1^{\tau_0}$ is negligible and we obtain:

$$\boldsymbol{\lambda}_i^{t+\tau_0} - \boldsymbol{\lambda}_i^{t+1} \approx -\alpha(\tau_0 - 1)\Delta\mathbf{w}_i^{t+\tau_0} \tag{25}$$

10.3. Client Slience Table

Lemma 1. *Consider a sequence $\mathbf{R}^t \in (\mathbb{Z}^+)^N$ initialized to $\mathbf{1}$, updated at round t as follows for each client i:*

$$R_i^t = \begin{cases} 1, & \text{if client } i \text{ is selected at round } t, \\ R_i^{t-1} + 1, & \text{otherwise}. \end{cases}$$

At each round, M clients are sampled uniformly at random without replacement, where $1 \leq M \leq N$. Then, for any client i, the steady-state expectation satisfies:

$$\mathbb{E}ss[R_i^t] = \lim_{t\to\infty} \mathbb{E}[R_i^t] = \frac{N}{M}. \tag{26}$$

Proof. For any client i, the selection probability per round is $p = M/N$. The process $\{R_i^t\}_{t\geq 0}$ forms a Markov chain on $\mathbb{Z}^+$ with transitions:

$$\Pr(R_i^t = 1 \mid R_i^{t-1} = m) = p,$$
$$\Pr(R_i^t = m+1 \mid R_i^{t-1} = m) = 1-p.$$

This chain is irreducible, aperiodic, and positive recurrent (as the return time to state 1 has finite expectation), guaranteeing a unique stationary distribution $\{\pi_k\}_{m=1}^{\infty}$. The stationary distribution satisfies:

$$\pi_1 = p\sum_{j=1}^{\infty}\pi_j = p,$$
$$\pi_m = (1-p)\pi_{m-1}, \quad m \geq 2.$$

Solving it recursively yields:

$$\pi_m = p(1-p)^{m-1}, \quad m \in \mathbb{Z}^+.$$

The steady-state expectation is:

$$\mathbb{E}\text{ss}[R_i^t] = \sum_{m=1}^{\infty} m\pi_m = \sum_{m=1}^{\infty} mp(1-p)^{m-1} = \frac{1}{p} = \frac{N}{M},$$

where the above equation uses the formula for the expectation of a geometric distribution (support $m \geq 1$). By ergodicity, $\lim_{t\to\infty}\mathbb{E}[R_i^t] = \mathbb{E}_{\text{ss}}[R_i^t]$. This intuitively means that each client is selected once every $\frac{N}{M}$ rounds on average, thus the mean value of R_i^t (the recorded number of unselected rounds) is $\frac{N}{M}$. □

In the experiments, to avoid overly large client silence values, we adopt the following strategy: for client i with $R_i^t > \frac{N}{M}$, the step size used in their virtual compensation mechanism was set to $\frac{N}{M}$. The update rule of R_i^t is as detailed below:

$$R_i^t = \begin{cases} 1, & \text{client } i \text{ is selected and } R_i^t < \frac{N}{M}, \\ R_i^t - \frac{N}{M}, & \text{client } i \text{ is selected and } R_i^t > \frac{N}{M}, \\ R_i^{t-1} + 1, & \text{otherwise.} \end{cases}$$

10.4. Optimization

This section first introduces several key lemmas and then provides the convergence rate of FedCVC based on these lemmas. Finally, a comparative analysis of the convergence rates between primal and dual update methods is presented.

Lemma 2. *For client i, the local primal and dual updates of the original model are as follows:*

$$\begin{aligned} \mathbf{w}_i^{t+1} - \mathbf{w}_0^t &= -\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \boldsymbol{\lambda}_i^t) \\ \boldsymbol{\lambda}_i^{t+1} &= (1 - C_i^t\alpha\eta r)\boldsymbol{\lambda}_i^t + C_i^t\alpha\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbf{g}_i(\mathbf{w}_i^{t,k}), \end{aligned} \tag{27}$$

where the coefficient C_i^t serves as an adaptive factor for controlling the dual update magnitude. Importantly, C_i^t increases when client i remains inactive for prolonged periods, thereby amplifying the influence of current model updates on the dual variable.

Proof. According to Eq. (18) and applying $\mathbf{w}_i^{t+1} = \mathbf{w}_i^{t,K}$ and $\mathbf{w}_i^{t,0} = \mathbf{w}_0^t$, we obtain the following variation:

$$\begin{aligned}
&\mathbf{w}_i^{t+1} - \mathbf{w}_0^t = \mathbf{w}_i^{t,K} - \mathbf{w}_i^{t,0} \\
&= (1-\alpha\eta)(\mathbf{w}_i^{t,K-1} - \mathbf{w}_0^t) - \eta(\mathbf{g}_i(\mathbf{w}_i^{t,K-1}) - \boldsymbol{\lambda}_i^t) \\
&= -\eta \sum_{k=0}^{K-1} (1-\alpha\eta)^{K-1-k}(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \boldsymbol{\lambda}_i^t) \\
&= -\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbf{g}_i(\mathbf{w}_i^{t,k}) + \eta r \boldsymbol{\lambda}_i^t \\
&= -\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \boldsymbol{\lambda}_i^t).
\end{aligned}$$

The dual update in Eq. (19) can be reformulated as:

$$\begin{aligned}
\boldsymbol{\lambda}_i^{t+1} &= \boldsymbol{\lambda}_i^t + \alpha C_i^t(\mathbf{w}_0^t - \mathbf{w}_i^{t+1}) \\
&= \boldsymbol{\lambda}_i^t + C_i^t \alpha\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbf{g}_i(\mathbf{w}_i^{t,k}) - C_i^t\alpha\eta r\boldsymbol{\lambda}_i^t \\
&= (1 - C_i^t\alpha\eta r\boldsymbol{\lambda}_i^t) + C_i^t\alpha\eta r \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbf{g}_i(\mathbf{w}_i^{t,k})
\end{aligned}$$

□

Lemma 3. *In the non-convex case, we can bound* $\mathbb{E}[f(\mathbf{w}_0^{t+1})] - \mathbb{E}[f(\mathbf{w}_0^t)]$ *as follows:*

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{w}_0^{t+1})] - \mathbb{E}[f(\mathbf{w}_0^t)] \leq \\
&\frac{1}{\alpha N}\sum_{i=1}^{N}(1 - \frac{M}{2N})Q_i^t\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 \\
&+ \frac{5\alpha\sigma^2 M}{4SN^2}\sum_{i=1}^{N} Q_i^t - (\frac{3\alpha}{4} - \frac{L}{2})\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2].
\end{aligned} \tag{28}$$

Proof. Expanding f according to its definition, we obtain:

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{w}_0^{t+1})] - \mathbb{E}[f(\mathbf{w}_0^t)] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[f_i(\mathbf{w}_0^{t+1}) - f_i(\mathbf{w}_0^t)] \\
&\overset{(a)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\langle\nabla f_i(\mathbf{w}_0^t), \mathbf{w}_0^{t+1} - \mathbf{w}_0^t\rangle] + \frac{L}{2N}\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] \\
&\overset{(b)}{=} \mathbb{E}[\langle\nabla f(\mathbf{w}_0^t), \mathbf{w}_0^{t+1} - \mathbf{w}_0^t\rangle] + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] \\
&\overset{(c)}{=} \mathbb{E}[\langle\nabla f_i(\mathbf{w}_0^t) + \alpha(\mathbf{w}_0^{t+1} - \mathbf{w}_0^t), \mathbf{w}_0^{t+1} - \mathbf{w}_0^t\rangle] - (\alpha - \frac{L}{2})\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] \\
&\overset{(d)}{\leq} \frac{1}{\alpha}\underbrace{\mathbb{E}[\|\nabla f(\mathbf{w}_0^t) + \alpha(\mathbf{w}_0^{t+1} - \mathbf{w}_0^t)\|^2]}_{A.1} - (\frac{3\alpha}{4} - \frac{L}{2})\mathbb{E}\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2],
\end{aligned} \tag{29}$$

where (a) follows from the L-smoothness property of f_i. (b) results from the linearity of expectation and the uniform distribution of clients, where $\frac{L}{2N}\sum_{i=1}^{N}\mathbb{E}[\|\cdot\|^2] = \frac{L}{2}\mathbb{E}[\|\cdot\|^2]$. (c) is obtained by perturbing the objective with $\alpha\|w_0^{t+1} - w_0^t\|^2$ and reordering inner product terms. d) holds by applying the Cauchy-Schwarz inequality followed by Young's inequality:

$\langle \mathbf{a}, \mathbf{b} \rangle \leq \alpha \|\mathbf{a}\|^2 + \frac{1}{4\alpha}\|\mathbf{b}\|^2$, where we set $\mathbf{a} = \nabla f(\mathbf{w}_0^t) + \alpha(\mathbf{w}_0^{t+1} - \mathbf{w}_0^t)$ and $\mathbf{b} = \mathbf{w}_0^{t+1} - \mathbf{w}_0^t$. Expanding $A.1$ based on the update rule of $\mathbf{w}_0^t$, we obtain:

$$
\begin{aligned}
A.1 &= \mathbb{E}[\| \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{w}_0^t) + \frac{1}{N}\sum_{i=1}^{N} \alpha(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\lambda}_i^{t+1} \|^2] \\
&\overset{(a)}{\leq} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) + \alpha(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] \\
&\overset{(b)}{=} \frac{1}{N}\sum_{i=1}^{N} (\frac{M}{N} \underbrace{\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) + \alpha(\tilde{\mathbf{w}}_i^{t+1} - \mathbf{w}_0^t) - \tilde{\boldsymbol{\lambda}}_i^{t+1}\|^2]}_{B.1} + (1 - \frac{M}{N})\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2]),
\end{aligned}
\tag{30}
$$

where (a) holds due to Jensen's inequality: $\mathbb{E}\left[\|\sum X_i\|^2\right] \leq \sum \mathbb{E}\left[\|X_i\|^2\right]$. (b) The expression is decomposed into participating and non-participating clients: For participating clients, the updated $\tilde{\mathbf{w}}_i^{t+1}$ and $\tilde{\boldsymbol{\lambda}}_i^{t+1}$ are used, while for non-participating clients, $\boldsymbol{\lambda}_i^t$ remains unchanged. For term $B.1$, an upper bound can be obtained by applying Lemma 2, which yields:

$$
\begin{aligned}
B.1 &= \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t - Q_i^t \sum_{k=0}^{K-1} \frac{r_k}{r}(\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k}) - \boldsymbol{\lambda}_i^t)\|^2] \\
&= \mathbb{E}[\|(1 - Q_i^t)(\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t) + Q_i^t \sum_{k=0}^{K-1} \frac{r_k}{r}(\nabla f_i(\mathbf{w}_0^t) - \mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k}))\|^2] \\
&\leq (1 - Q_i^t)\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + Q_i^t \underbrace{\sum_{k=0}^{K-1} \frac{r_k}{r}\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k})\|^2]}_{C.1}.
\end{aligned}
\tag{31}
$$

For the term $C.1$, we add a zero term $\nabla f_i(\tilde{\mathbf{w}}_i^{t,k})$ and utilize the property of $\mathbb{E}$ to obtain:

$$
\begin{aligned}
C.1 &\leq \sum_{k=0}^{K-1} \sum_{r_k}^{r} \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \nabla f_i(\tilde{\mathbf{w}}_i^{t,k})\|^2] + \frac{\sigma^2}{S} \\
&\leq L^2 \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbb{E}[\|\mathbf{w}_0^t - \tilde{\mathbf{w}}_i^{t,k}\|^2] + \frac{\sigma^2}{S} \\
&\overset{(a)}{\leq} L^2 \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbb{E}[\eta \sum_{j=0}^{k-1} r_j(\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,j}) - \boldsymbol{\lambda}_i^t)\|^2] + \frac{\sigma^2}{S} \\
&= L^2\eta^2 \sum_{k=0}^{K-1} \frac{r_k}{r}\mathbb{E}[\| \sum_{j=0}^{k-1} r_j(\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,j}) - \boldsymbol{\lambda}_i^t)\|^2] + \frac{\sigma^2}{S} \\
&\overset{(b)}{\leq} L^2\eta^2 \sum_{k=0}^{K-1} \frac{r_k}{r}(\sum_{j=0}^{k-1} r_j)(\sum_{j=0}^{k-1} r_j \mathbb{E}[\|g_i(\tilde{\mathbf{w}}_i^{t,j}) - \boldsymbol{\lambda}_i^t\|^2]) + \frac{\sigma^2}{S} \\
&\overset{(c)}{\leq} L^2\eta^2 r^2 \underbrace{\sum_{k=0}^{K-1} \frac{r_k}{r}\mathbb{E}[\|g_i(\tilde{\mathbf{w}}_i^{t,k}) - \boldsymbol{\lambda}_i^t\|^2]}_{D.1} + \frac{\sigma^2}{S},
\end{aligned}
\tag{32}
$$

where (a) holds due to the local update $\tilde{\mathbf{w}}_i^{t,k} - \mathbf{w}_0^t = -\eta \sum_{j=0}^{k-1} r_j(\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,j}) - \boldsymbol{\lambda}_i^t)$. (b) holds due to Jensen's inequality: $\mathbb{E}\left[\|\sum X_i\|^2\right] \leq \sum \mathbb{E}\left[\|X_i\|^2\right]$. (c) holds due to $\sum_{k=0}^{K-1} \frac{r_k}{r}(\sum_{j=0}^{k-1} r_j) \leq \sum_{k=0}^{K-1} \frac{r_k}{r}(\sum_{j=0}^{K-2} r_j) \leq r - r_{K-1} \leq r$. For term

$D.1$, add zero term $\nabla f_i(\tilde{\mathbf{w}}_i^{t,k})$ again, we have:

$$\begin{aligned} D.1 &\leq \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbb{E}[\|\nabla f_i(\tilde{\mathbf{w}}_i^{t,k}) - \boldsymbol{\lambda}_i^t] + \frac{\sigma^2}{S} \\ &= \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbb{E}[\|\nabla f_i(\tilde{\mathbf{w}}_i^{t,k}) - \nabla f_i(\mathbf{w}_0^t) + \nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + \frac{\sigma^2}{S} \\ &\overset{(a)}{\leq} 2L^2 \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbb{E}[\|\tilde{\mathbf{w}}_i^{t,k} - \mathbf{w}_0^t\|^2] + 2\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + \frac{\sigma^2}{S} \\ &\overset{(b)}{\leq} 2\eta^2 r^2 L^2 \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbb{E}[\|\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k}) - \boldsymbol{\lambda}_i^t\|^2] + 2\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + \frac{\sigma^2}{S}, \end{aligned} \tag{33}$$

where (a) utilizes the quadratic expansion property that $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ along with the L-smoothness assumption (b) holds due to $\|\tilde{\mathbf{w}}_i^{t,k} - \mathbf{w}_0^t\|^2 \leq \eta^2 r^2 \mathbb{E}\left[\|\mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k}) - \boldsymbol{\lambda}_i^t\|^2\right]$. Simplifying the relationship yields:

$$\begin{aligned} D.1 &\leq \frac{1}{1 - 2\eta^2 r^2 L^2} (2\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + \frac{\sigma^2}{S}) \\ &\leq \frac{1}{4\eta^2 r^2 L^2} (2\mathbb{E}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{\sigma^2}{S}), \end{aligned} \tag{34}$$

where (a) holds because we enforce the condition $\sqrt{6}\eta r L \leq 1$. Substituting the expression for $D.1$ into $C.1$ yields:

$$C.1 \leq \frac{1}{2}\mathbb{E}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\sigma^2}{4S}. \tag{35}$$

We obtain the following results by substituting the bound from $C.1$ back into $B.1$:

$$B.1 \leq (1 - \frac{Q_i^t}{2})\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\sigma^2}{4S} Q_i^t. \tag{36}$$

Substituting the expression for $B.1$ into $A.1$ yields:

$$A.1 \leq \frac{1}{N} \sum_{i=1}^{N} (1 - \frac{M}{2N} Q_i^t) \|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\sigma^2 Q_i^t M}{4SN}. \tag{37}$$

Simplifying the relationship yields:

$$\mathbb{E}[f(\mathbf{w}_0^{t+1})] - \mathbb{E}[f(\mathbf{w}_0^t)] \leq \frac{1}{\alpha N} \sum_{i=1}^{N} (1 - \frac{M}{2N}) Q_i^t \|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\alpha\sigma^2 M}{4SN^2} \sum_{i=1}^{N} Q_i^t - (\frac{3\alpha}{4} - \frac{L}{2}) \mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] \tag{38}$$

□

Lemma 4. *The expectation $\mathbb{E}[\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t]$ is bounded as follows:*

$$\begin{aligned} \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \boldsymbol{\lambda}_i^{t+1}\|^2] \leq{}& \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + (1 + \frac{4N}{MP_i^t}) L^2 \mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] \\ &- \frac{MP_i^t(\frac{M}{N}P_i^t + 2)}{8N} \|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\sigma^2 MP_i^t}{4SN}(1 + \frac{MP_i^t}{4N}). \end{aligned} \tag{39}$$

Proof. We analyze the variation in gradient estimation error from t to $t+1$. First, we decompose the error by adding and subtracting $\nabla f_i(\mathbf{w}_0^t)$:

$$\begin{aligned} &\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \boldsymbol{\lambda}_i^t\|^2] = \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \nabla f_i(\mathbf{w}_0^t) + \nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] \\ &\overset{(a)}{\leq} (1 + \frac{1}{d}) \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \nabla f_i(\mathbf{w}_0^t)\|^2] + (1 + d) \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^{t+1}\|^2] \\ &\overset{(b)}{=} (1 + \frac{1}{d}) L^2 \mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] + (1 + d)(1 - \frac{M}{N}) \mathbb{E}[\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + (1 + d)\frac{M}{N} \underbrace{\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \tilde{\boldsymbol{\lambda}}_i^{t+1}\|^2]}_{E.1}, \end{aligned} \tag{40}$$

where (a) applies Young's inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq \left(1+\frac{1}{d}\right)\|\mathbf{a}\|^2 + (1+d)\|\mathbf{b}\|^2$ (where $d>0$). (b) decomposes $\boldsymbol{\lambda}_i^{t+1}$ into $\tilde{\boldsymbol{\lambda}}_i^{t+1}$ for participating clients and retain λ_i^t for non-participating clients. According to Eq. (27), the term $E.1$ can be bounded as follows:

$$\begin{aligned}
E.1 &= \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - (1-P_i^t)\boldsymbol{\lambda}_i^t - P_i^t \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k})\|^2] \\
&= \mathbb{E}[\|(1-P_i^t)(\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t) + P_i^t \sum_{k=0}^{K-1} (\nabla f_i(\mathbf{w}_0^t) - \mathbf{g}_i(\tilde{\mathbf{w}}_i^{t,k}))\|^2] \\
&\leq (1-P_i^t)\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] + P_i^t \sum_{k=0}^{K-1} \frac{r_k}{r} \mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - g_i(\tilde{\mathbf{w}}_i^{t,k})\|^2] \\
&\overset{(a)}{\leq} (1-\frac{P_i^t}{2})\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|] + \frac{5\sigma^2}{4S^2} P_i^t,
\end{aligned} \tag{41}$$

where (a) follows Eq. (35). Simplifying the relationship yields:

$$\begin{aligned}
&\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \boldsymbol{\lambda}_i^{t+1}\|^2] - \mathbb{E}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] \\
&\leq (1+\frac{1}{d})\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] + \frac{5\sigma^2}{4S^2}(1+d)\frac{M}{N}P_i^t + ((1+d)(1-\frac{MP_i^t}{2N}) - 1)\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2].
\end{aligned} \tag{42}$$

Let $d = \frac{MP_i^t}{4N}$, and $(1+d)(1-\frac{MP_i^t}{2N}) - 1 = -\frac{MP_i^t}{8N}(\frac{M}{N}P_i^t + 2)$,we have:

$$\begin{aligned}
&\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \boldsymbol{\lambda}_i^{t+1}\|^2] - \mathbb{E}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] \\
&\leq (1+\frac{4N}{MP_i^t})L^2\mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] - \frac{MP_i^t(\frac{M}{N}P_i^t + 2)}{8N}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2 + \frac{5\sigma^2 MP_i^t}{4SN}(1+\frac{MP_i^t}{4N}),
\end{aligned} \tag{43}$$

which shows that when the global update $\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|$ is small, a tighter upper bound can be achieved by increasing β to amplify the weight P_i^t. □

Theorem 3. *Let f_i be L-smooth, and $\mathbb{E}_{ss}[R_i^t] = \frac{N}{M}$ for all clients. Set the $\alpha > \frac{L}{2} + \frac{13L}{2\beta}$. Then, after T iterations, with initial global model $\mathbf{w}_0^0$, optimal value $f^* = \min_{\mathbf{w}_0} f(\mathbf{w}_0)$ and $E = f(\mathbf{w}_0^0) - f^*$, the average expected squared gradient norm satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\mathbf{w}_0^t)\|^2] = \mathcal{O}\left(\frac{E}{T} + \frac{M\sigma^2}{NS}\right),$$

where the expectation is over all randomness in the algorithm.

Proof. Define the Lyapunov function $G^t = \mathbb{E}[f(\mathbf{w}_0^t)] + \frac{4}{M\alpha}\sum_{i=1}^N \frac{\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2]}{P_i^t}$. Assume $\alpha > \frac{L}{2} + \frac{13LN}{2MP_i^t}$ for all i and t, ensuring $B_i^t > \alpha/4$ later in the proof. To derive the main inequality, multiply Eq. (39) by $\frac{4N}{MP_i^t\alpha}$ and add it to Eq. (28), yielding:

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{w}_0^{t+1})] - \mathbb{E}[f(\mathbf{w}_0^t)] + \frac{4}{M\alpha}\sum_{i=1}^N (\frac{\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1}) - \boldsymbol{\lambda}_i^{t+1}\|^2]}{P_i^{t+1}} - \frac{\mathbb{E}\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2}{P_i^t}) \\
&\leq -\frac{1}{2\alpha N}\sum_{i=1}^N \frac{M}{N}(Q_i^t + P_i^t)\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^t) - \boldsymbol{\lambda}_i^t\|^2] - \frac{1}{N}\sum_{i=1}^N B_i^t \mathbb{E}[\|\mathbf{w}_0^{t+1} - \mathbf{w}_0^t\|^2] + \frac{5\sigma^2}{S\alpha N}\sum_{i=1}^N (1 + \frac{M(P_i^t + Q_i^t)}{4N}),
\end{aligned} \tag{44}$$

where $B_i^t = \frac{3\alpha}{4} - \frac{4M}{\alpha N P_i^t}(1+\frac{M}{NP_i^t})L^2 - \frac{L}{2}$. The condition $\alpha > \frac{L}{2} + \frac{13LN}{2MP_i^t}$ implies $D_i^t > \frac{\alpha}{4}$. Consequently, rearranging

terms gives:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{M}{N}(P_i^t+Q_i^t)\|\nabla f_i(\mathbf{w}_0^t)-\boldsymbol{\lambda}_i^t\|^2+\frac{\alpha^2}{2}\mathbb{E}[\|\mathbf{w}_0^{t+1}-\mathbf{w}_0^t\|^2]\\
&\leq 2\alpha(\mathbb{E}[f(\mathbf{w}_0^{t+1})]-\mathbb{E}[f(\mathbf{w}_0^t)])+\frac{4}{M\alpha}\sum_{i=1}^{N}(\frac{\mathbb{E}[\|\nabla f_i(\mathbf{w}_0^{t+1})-\boldsymbol{\lambda}_i^{t+1}\|^2]}{P_i^{t+1}}-\frac{\mathbb{E}\|\nabla f_i(\mathbf{w}_0^t)-\boldsymbol{\lambda}_i^t\|^2}{P_i^t})+\\
&\frac{10\sigma^2}{SN}\sum_{i=1}^{N}(1+\frac{M(P_i^t+Q_i^t)}{4N})
\end{aligned}\tag{45}$$

Furthermore, bounding the squared gradient norm:

$$\begin{aligned}
&\mathbb{E}[\|\nabla f(\mathbf{w}_0^t)\|^2]=\alpha^2\mathbb{E}[\|\mathbf{w}_0^t-\mathbf{w}_0^{t+1}+\mathbf{w}_0^{t+1}-\mathbf{w}_0^t+\frac{1}{\alpha}\nabla f(\mathbf{w}_0^t)\|^2]\\
&\leq\alpha^2(1+c)\mathbb{E}[\|\mathbf{w}_0^t-\mathbf{w}_0^{t+1}\|^2]+\alpha^2(1+\frac{1}{c})\mathbb{E}[\|\mathbf{w}_0^{t+1}-\mathbf{w}_0^t+\frac{1}{\alpha}\nabla f(\mathbf{w}_0^t)\|^2]\\
&\leq(1+c)\alpha^2\mathbb{E}[\|\mathbf{w}_0^t-\mathbf{w}_0^{t+1}\|^2]+\frac{1+\frac{1}{c}}{2N}\sum_{i=1}^{N}(2-\frac{MQ_i^t}{N})\|\nabla f(\mathbf{w}_0^t)-\boldsymbol{\lambda}_i^t\|^2+\frac{5\sigma^2MP_i^t}{2SN}\\
&=(1+c)\alpha^2\mathbb{E}[\|\mathbf{w}_0^t-\mathbf{w}_0^{t+1}\|^2]+\frac{1+\frac{1}{c}}{2N}\sum_{i=1}^{N}(2-\frac{MQ_i^t}{N})\|\nabla f_i(\mathbf{w}_0^t)-\boldsymbol{\lambda}_i^t\|^2+\frac{5(1+\frac{1}{c}\sigma^2)}{4SN}\sum_{i=1}^{N}\frac{M}{N}Q_i^t\\
&\overset{(a)}{\leq}2\alpha B_0(G^t-G^{t+1})+\frac{10B_0\sigma^2}{S}(1+\frac{M(Q+P)}{4N})+\frac{5(1+\frac{1}{c})\sigma^2M}{4SN}Q,
\end{aligned}\tag{46}$$

where (a) holds because $\mathbb{E}_{ss}[R_i^t]=\frac{N}{M}$, the expectations satisfy $\mathbb{E}[Q_i^t]=\left(\frac{\beta N}{M}+2\right)\alpha\eta r=Q$ and $\mathbb{E}[P_i^t]=\left(\frac{\beta N}{M}+1\right)\alpha\eta r=P$.

$$B_0=2(1+c)+\frac{1}{2}(1+\frac{1}{c})\frac{2-\frac{M}{N}Q}{\frac{M}{N}(Q+P)}=(1+\frac{1}{c})\frac{2N}{M(Q+P)}+2(1+c)-(1+\frac{1}{c})\frac{Q}{2(Q+P)}$$

Setting $c=\frac{1}{8}$ and we have $2(1+c)-(1+\frac{1}{c})\frac{Q}{2(Q+P)}<0$. Substituting Q and P into B_0 yields $B_0\leq\frac{18N}{M(P+Q)}$.

Summing over $t=0$ to $T-1$ and averaging, the telescoping sum for G^t-G^{t+1} simplifies to G^0-G^T. Since $G^T\geq\mathbb{E}[f(\mathbf{w}_0^T)]\geq f^*$:

$$\begin{aligned}
\frac{1}{T}\mathbb{E}[\|\nabla f(\mathbf{w}_0^t)\|^2]&\leq\frac{2B_0\alpha(G^0-f^*)}{T}+\frac{10B_0\sigma^2}{STN}\sum_{t=1}^{T}\sum_{i=1}^{N}(1+\frac{M(Q_i^t+P_i^t)}{4N})+\frac{45\sigma^2}{4ST}\sum_{t=1}^{T}\frac{M}{N}Q_i^t\\
&\leq\frac{36\alpha(G^0-f^*)}{\beta T}+\frac{\sigma^2}{S}(\frac{90}{\beta}+\frac{45\beta}{4}+\frac{45M}{2N}+45)\\
&\leq\frac{36\alpha(G^0-f^*)}{\beta T}+\frac{\sigma^2}{S}(\frac{90}{\beta}+\frac{45M}{2N}+\frac{225}{4}).
\end{aligned}\tag{47}$$

Furthermore, let $E=(f(\mathbf{w}_0^0)-f^*)$:

$$\frac{1}{T}\mathbb{E}[\|\nabla f(w_0^t)\|^2]=\mathcal{O}(\frac{E}{T}+\frac{M\sigma^2}{NS})\tag{48}$$

□

10.4.1. Discussion of Optimization Error

Adopting the same proof framework, we compare the convergence rates of various algorithms in Tab. 8. The results demonstrate that while FedPD [46] necessitates full client participation, FedCVC achieves a comparable convergence rate through

	Assumption	Convergence rate	Mitigate DD?
FedPD	smoothness, bounded SGD variance, ϵ-inexact solution	$\mathcal{O}\left(\frac{E}{T}+\frac{\sigma^2}{S}+\epsilon\right)$	$\times$
FedDyn	smoothness, bounded SGD variance	$\mathcal{O}\left(\frac{NE}{MT}+\frac{N\sigma^2}{MS}\right)$	$\times$
FedADMM	smoothness, bounded SGD variance, $\epsilon_{i,t}$-inexact solution	$\mathcal{O}\left(\frac{NE}{MT}+\frac{1}{NT}\sum_{i,t}\epsilon_{i,t}\right)$	$\times$
A-FedPD	smoothness, bounded SGD variance, ϵ-inexact solution	$\mathcal{O}\left(\frac{E}{T}+\frac{\sigma^2}{S}+\epsilon\right)$	✓
FedCVC	smoothness, bounded SGD variance	$\mathcal{O}\left(\frac{E}{T}+\frac{M\sigma^2}{NS}\right)$	✓

Table 8. Comparison of optimization errors of different algorithms.

its client-driven virtual compensation mechanism. Relative to FedDyn [1], a primal-dual method accommodating partial client participation, FedCVC eliminates the $\frac{N}{M}$ term, thereby mitigating DD. A-FedPD [29] also removes the $\frac{N}{M}$ term by simulating client dual updates on the server side. However, this approach introduces approximation errors owing to local bias and lacks the capability to dynamically adjust update confidence based on data distribution and client participation patterns. In contrast, FedCVC controls approximation errors via adaptive step sizes and hyperparameter β, enhancing robustness to virtual compensation of stale dual variables.

10.5. Generalization

This section discusses the generalization capability of FedCVC on two adjacent datasets, D and $\hat{D}$. We first define key notations and present relevant lemmas, followed by a bound on the generalization error. Notably, the notation $\hat{\cdot}$ denotes results derived from the dataset $\hat{D}$.

Following with [29], we define $u^t = \mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\|$ to track the difference of global model parameter. Define $v^t = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\mathbf{w}_i^t - \hat{\mathbf{w}}_i^t\|$ to track the discrete difference of local parameters. Define $z^t = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|$ to track the discrete difference of the dual variables. Finally, define $h^t = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\|(\mathbf{w}_i^t - \mathbf{w}_0^{t-1}) - (\hat{\mathbf{w}}_i^t - \hat{\mathbf{w}}_0^{t-1})\|$.

Lemma 5.

$$\epsilon^G \leq \sup_{D,\hat{D},\xi}[f(\mathbf{w}^T,\xi) - f(\hat{\mathbf{w}}^T,\xi)] \leq L_G\mathbb{E}\|\mathbf{w}^t - \hat{\mathbf{w}}^T\| + \frac{HM\tau_0}{ND} \tag{49}$$

Proof. Let ζ represents the event where models $\mathbf{w}^T$ and $\hat{\mathbf{w}}^T$ maintain equivalence (i.e., $\mathbf{w}^T = \hat{\mathbf{w}}^T$) prior to the τ_0-th training round, indicating undisturbed states before data discrepancy propagation. The complement event ζ^D corresponds to the case where model divergence has already occurred. The expectation difference is decomposed into two parts using the event ζ, as follows:

$$\begin{aligned}\mathbb{E}[|f(\mathbf{w}^T,\xi) - f(\hat{\mathbf{w}}^T,\xi)|] &\leq P(\zeta)\mathbb{E}[|f(\mathbf{w}^T,\xi) - f(\hat{\mathbf{w}}^T,\xi)||\zeta] + P(\zeta^D)\mathbb{E}[|f(\mathbf{w}^T,\xi) - f(\hat{\mathbf{w}}^T,\xi)||\zeta^D] \\ &\leq L_G\mathbb{E}[\|\mathbf{w}^T - \hat{\mathbf{w}}^T\||\zeta] + HP(\zeta^D)\end{aligned}$$

where $H = \sup_{\mathbf{w},\xi} f(\mathbf{w},\xi) < +\infty$. The first term corresponds to ζ (no discrepancy has propagated), and the model difference is controlled by gradient boundedness. The second term represents ζ^D where the event fails (discrepancy has propagated), for which the function value boundedness is directly applied. Assume the different pairs are selected in s-th round, the probability of selecting differing samples before round s_0 is as follows:

$$P(\zeta^D) = P(s \leq s_0) \leq \sum_{t=0}^{s_0} P(s=t) = \sum_{t=0}^{s_0} P(i^* \in C^t)P(j^*) \leq \frac{M}{N}\frac{1}{D}s_0 = \frac{Ms_0}{ND},$$

where i^* denotes the client index possessing distinct data pairs, while j^* represents the index of differing samples in the client dataset. By bounding $\mathbb{E}\|\mathbf{w}^T - \hat{\mathbf{w}}^T\|$, the algorithmic stability is transformed into generalization error. □

Lemma 6. *Given a client $i(i \neq i^*)$, we bound the local updates as follows:*

$$\mathbb{E}[\|(\mathbf{w}_i^{t,k+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k+1} - \hat{\mathbf{w}}_0^t)\|] \leq \eta KL\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta K\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|. \tag{50}$$

Proof. According to the update rule, we have:

$$\mathbf{w}_i^{t,k+1} - \mathbf{w}_0^t = (1-\alpha\eta)(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - \eta(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \boldsymbol{\lambda}_i^t),$$

$$\begin{aligned}
&\mathbb{E}[\|(\mathbf{w}_i^{t,k+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k+1} - \hat{\mathbf{w}}_0^t)\|] \\
&= \mathbb{E}[\|(1-\alpha\eta)[(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)] - \eta(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \mathbf{g}_i(\hat{\mathbf{w}}_i^{t,k})) + \eta(\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t)\|] \\
&\leq (1-\alpha\eta)\mathbb{E}\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\| + \eta L\|\mathbf{w}_i^{t,k} - \hat{\mathbf{w}}_i^{t,k}\| + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| \\
&\leq (1-\alpha\eta)\mathbb{E}\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\| + \eta L\|\mathbf{w}_i^{t,k} - \hat{\mathbf{w}}_i^{t,k} + \mathbf{w}_0^t + \hat{\mathbf{w}}_0^t - \mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| \\
&\leq (1-\alpha_L\eta)\mathbb{E}\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\| + \eta L\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|,
\end{aligned}$$

where $\alpha_L = \alpha - L$ denote a positive constant. Unrolling the recursion from step 0 to $K-1$ and substituting the boundary conditions $\mathbf{w}_i^{t+1} = \mathbf{w}_i^{t,K}$ and $\mathbf{w}_0^t = \mathbf{w}_i^{t,0}$, we derive the expected model discrepancy as:

$$\begin{aligned}
&\mathbb{E}[\|(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t+1} - \hat{\mathbf{w}}_0^t)\|] = \mathbb{E}[\|(\mathbf{w}_i^{t,K} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,K} - \hat{\mathbf{w}}_0^t)] \\
&\leq [\prod_k^{K-1}(1-\alpha_L\eta)]\mathbb{E}[\|(\mathbf{w}_i^{t,0} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,0} - \hat{\mathbf{w}}_0^t)\|] + \sum_{k=0}^{K-1}\eta[\prod_{m=k+1}^{K-1}(1-\alpha_L\eta)](L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|).
\end{aligned}$$

Noting that the initial discrepancy $\mathbb{E}[\|(\mathbf{w}_i^{t,0} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,0} - \hat{\mathbf{w}}_0^t)\|]$ is zero due to identical initialization, we obtain the tighter bound:

$$\begin{aligned}
\mathbb{E}[\|(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t+1} - \hat{\mathbf{w}}_0^t)\|] &\leq \sum_{k=0}^{K-1}\eta[\prod_{m=k+1}^{K-1}(1-\alpha_L\eta)](L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|) \\
&= \frac{1-(1-\alpha_L\eta)^K}{\alpha_L}(L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|) \\
&\leq \eta KL\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta K\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\|,
\end{aligned}$$

where the final inequality leverages the binomial approximation $(1+a)^n \geq 1 + ax$ for $a > -1$ and integer $n \geq 1$. □

Lemma 7. *For client $i = i^*$ possessing distinct data pairs, the difference between local updates is bounded by:*

$$\mathbb{E}[\|(\mathbf{w}_i^{t,k+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k+1} - \hat{\mathbf{w}}_0^t)\|] \leq \eta KL\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta K\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2\eta KG}{D} \tag{51}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|(\mathbf{w}_i^{t,k+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k+1} - \hat{\mathbf{w}}_0^t)\|] \\
&= \mathbb{E}[\|(1-\alpha\eta)[(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)] - \eta(\mathbf{g}_i(\mathbf{w}_i^{t,k}) - \mathbf{g}_i(\hat{\mathbf{w}}_i^{t,k})) + \eta(\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t)\|] \\
&\leq (1-\frac{1}{D})[(1-\alpha_L\eta)\mathbb{E}[\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\|] + \eta L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| \\
&+ \frac{1}{D}(1-\alpha_L\eta)\mathbb{E}[\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\|] + 2\eta L_G + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| \\
&\leq (1-\alpha_L\eta)\mathbb{E}[\|(\mathbf{w}_i^{t,k} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,k} - \hat{\mathbf{w}}_0^t)\|] + \eta L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2\eta G}{D}
\end{aligned}$$

For the case where the data size D is sufficiently large, we observe that $1 - \frac{1}{D}$ asymptotically approaches 1. By unrolling the recursive relationship from step 0 to $K-1$, we obtain the following key expression:

$$\begin{aligned}
&\mathbb{E}[\|(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t+1} - \hat{\mathbf{w}}_0^t)\|] = \mathbb{E}[\|(\mathbf{w}_i^{t,K} - \mathbf{w}_0^t) - (\hat{\mathbf{w}}_i^{t,K} - \hat{\mathbf{w}}_0^t)] \\
&\leq \sum_{k=0}^{K-1}\eta[\prod_{m=k+1}^{K-1}(1-\alpha_L\eta)](L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2L_G}{N}) \\
&= \frac{1-(1-\alpha_L\eta)^K}{\alpha_L}(L\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2L_G}{D}) \\
&\leq \eta KL\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta K\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2\eta KL_G}{D}
\end{aligned}$$

□

By introducing Lemmas 5 to 7, we establish the connections between local stability, global stability, and generalization error, laying the foundation for subsequent analysis.

Theorem 4. *Under Assumptions 1 and 3, when the learning rate η satisfies $\eta \leq KL$ and the hyperparameters α, β are properly chosen, the generalization error of the global model is upper bounded by: For constant learning rates:*

$$\epsilon_G \leq \frac{2L_G(1+l)^T}{LND},$$

For diminishing learning rates $\eta^t = \frac{\eta_0}{t+1}$:

$$\epsilon_G \leq \frac{2}{ND}\left(\frac{2L_G^2}{L}\right)^{\frac{1}{1+\mu L}}(HMT)^{\frac{\mu L}{1+\mu L}},$$

where l, μ are constants, L is the Lipschitz smoothness, K is the number of local steps, and D is client's data size.

Proof. Following the update rule of global model parameters ($\mathbf{w}_0$), we derive:

$$\mathbf{w}_0^{t+1} = \frac{1}{M}\sum_{i\in C^t}\mathbf{w}_i^{t+1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i^{t+1}. \tag{52}$$

Moreover, we establish the following bound on the model difference term u^{t+1}:

$$\begin{aligned} u^{t+1} &= \mathbb{E}\|\mathbf{w}_0^{t+1} - \hat{\mathbf{w}}_0^{t+1}\| = \mathbb{E}\|\frac{1}{M}\sum_{i\in C^t}(\mathbf{w}_i^{t+1} - \hat{\mathbf{w}}_i^{t+1}) - \frac{1}{N\alpha}\sum_{i=1}^{N}(\boldsymbol{\lambda}_i^{t+1} - \hat{\boldsymbol{\lambda}}_i^{t+1})\| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\mathbf{w}_i^{t+1} - \hat{\mathbf{w}}_i^{t+1}\| + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\alpha}\mathbb{E}\|\boldsymbol{\lambda}_i^{t+1} - \hat{\boldsymbol{\lambda}}_i^{t+1}\| \leq v^{t+1} + \frac{1}{\alpha}z^t \end{aligned}$$

Following the dual variable update rule, we derive:

$$\begin{aligned} \boldsymbol{\lambda}^{t+1} = \sum_{i=1}^{N}\boldsymbol{\lambda}_i^{t+1} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i^t - \frac{\alpha(1+\beta(R_i^t-1))}{N}\sum_{i\in C^t}(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t) \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\lambda}_i^t - \alpha\beta\frac{1}{M}\sum_{i\in C^t}(\mathbf{w}_i^{t+1} - \mathbf{w}_0^t). \end{aligned}$$

We obtain the final inequality by noting that $\mathbb{E}_{ss}[R_i^t] = \frac{N}{M}$. Analyzing the gap between $\boldsymbol{\lambda}^{t+1}$ and $\hat{\boldsymbol{\lambda}}^{t+1}$ under expectation yields:

$$z^{t+1} \leq z^t + \alpha\beta h^{t+1}.$$

Based on Lemma 7, we derive the following key bound:

$$h^{t+1} \leq \eta KL\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + \eta K\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + \frac{2\eta KL_G}{D}$$

By introducing the constants $e_1 = \eta KL$, $e_2 = \eta K$, and $e_3 = \frac{2\eta KL_G}{D}$, we simplify the bound as follows:

$$h^{t+1} \leq e_1\mathbb{E}\|\mathbf{w}_0^t - \hat{\mathbf{w}}_0^t\| + e_2\mathbb{E}\|\boldsymbol{\lambda}_i^t - \hat{\boldsymbol{\lambda}}_i^t\| + e_3$$

By combining the inequalities derived, we obtain the following key theoretical relationship:

$$\begin{cases} u^{t+1} \leq v^{t+1} + \frac{1}{\alpha}z^{t+1} \\ z^{t+1} \leq z^t + \alpha\beta h^{t+1} \\ v^{t+1} \leq h^{t+1} + u^{t+1} \\ h^{t+1} \leq e_1u^t + e_2z^t + e_3 \end{cases} \tag{53}$$

To consolidate the inequalities, we introduce positive coefficients x, y, l:

$$xz^{t+1} + y\delta^{t+1} + lh^{t+1} \leq x(z^t + \alpha\beta h^{t+1}) + y(h^{t+1} + u^t) + l(e_1 u^t + e_2 z^t + e_3).$$

Rearranging the right-hand side:

$$xz^{t+1} + yv^{t+1} + dh^{t+1} \leq (x + le_2)z^t + (y + le_1)u^t + (x\alpha\beta + y)h^{t+1} + le_3.$$

Transposing terms yields:

$$xz^{t+1} + yv^{t+1} + [l - (x\alpha\beta + y)]h^{t+1} \leq (x + le_2)z^t + (y + le_1)u^t + le_3.$$

By setting $y = 1$ and $l = x\alpha\beta + y = x\alpha\beta + 1$, we ensure: $l - (x\alpha\beta + y) = 0$. The inequality then simplifies to:

$$xz^{t+1} + y\delta^{t+1} \leq (x + le_2)z^t + (y + le_1)u^t + le_3.$$

Substituting $y = 1$:

$$xz^{t+1} + v^{t+1} \leq (x + le_2)z^t + (1 + le_1)u^t + le_3,$$

where $l = x\alpha\beta + 1$. From Inequality: $u^{t+1} \leq v^{t+1} + \frac{1}{\alpha}z^{t+1}$, we derive $v^{t+1} \geq u^{t+1} - \frac{1}{\alpha}z^{t+1}$. Substituting this into the above:

$$xz^{t+1} + \left(u^{t+1} - \frac{1}{\alpha}z^{t+1}\right) \leq (x + le_2)z^t + (1 + le_1)u^t + le_3.$$

Rearranging the left-hand side:

$$u^{t+1} + \left(x - \frac{1}{\alpha}\right)z^{t+1} \leq (1 + le_1)u^t + (x + le_2)z^t + le_3.$$

Letting $x_\alpha = x - \frac{1}{\alpha}$, we obtain:

$$u^{t+1} + x_\alpha z^{t+1} \leq (1 + le_1)u^t + (x + le_2)z^t + le_3.$$

Define the Lyapunov function $V^t = u^t + x_\alpha z^t$. Make the following condition holds:

$$x + le_2 \leq x_\alpha(1 + le_1),$$

then the inequality can be reduced to:

$$V^{t+1} \leq (1 + le_1)V^t + le_3.$$

Substituting $e_1 = \eta KL$, $e_2 = \eta K$, $e_3 = \frac{2\eta^t KL_G}{ND}$, and $l = x\alpha\beta + 1$:

$$V^{t+1} \leq (1 + (x\alpha\beta + 1)\eta KL)\, V^t + (x\alpha\beta + 1)\frac{2\eta KL_G}{ND}.$$

where the hyperparameter β controls the influence of local update differences h^{t+1} on the dual variable differences z^{t+1}. When $0 < \beta < 1$, it is robust to high-variance local updates through this contraction mapping. Unfolding the recursion from $t = s_0$ to $T - 1$ and leveraging the initial conditions $u^{s_0} = 0$ and $\boldsymbol{\lambda}^{s_0} = 0$ (i.e., $V^{s_0} = 0$). When the diminishing learning rate is given by $\eta = \eta^t = \frac{\eta_0}{t+1}$, and $\eta_0 \leq \frac{\mu}{dK}$($\mu > 0$, is a constant), we obtain:

$$V^T \leq \sum_{t=s_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \left(1 + \frac{l\eta_0 KL}{j+1}\right) \right) \frac{2l\eta_0 KL_G}{ND(t+1)}.$$

By leveraging the inequality $1+a \leq e^a$ and the properties of harmonic series, we simplify the product as follows:

$$\prod_{j=t+1}^{T-1}\left(1+\frac{l\eta_0 KL}{j+1}\right) \leq \exp\left(l\eta_0 KL \sum_{j=t+1}^{T-1}\frac{1}{j+1}\right) \leq \left(\frac{T}{t+1}\right)^{l\eta_0 KL},$$

and calculate the sum:

$$V^T \leq \frac{2l\eta_0 KL_G}{ND} T^{l\eta_0 KL} \sum_{t=s_0}^{T-1}(t+1)^{-1-l\eta_0 KL}$$

By applying the condition $l\eta_0 KL \leq \mu L$, we perform integral relaxation as follows:

$$\sum_{t=s_0}^{T-1}(t+1)^{-1-l\eta_0 KL} \leq \int_{s_0}^{T} t^{-1-l\eta_0 KL}dt \leq \frac{s_0^{-l\eta_0 KL}}{l\eta_0 KL} \leq \frac{s_0^{-\mu L}}{\mu L}$$

Simplying the relationship:

$$V^T \leq \frac{2l\eta_0 KL_G}{ND} T^{l\eta_0 KL} \cdot \frac{s_0^{-\mu L}}{\mu L} \leq \frac{2L_G}{LND}\left(\frac{T}{s_0}\right)^{\mu L}. \tag{54}$$

According to Lemma 5:

$$\epsilon_G \leq L_G \Delta^T + \frac{HM\tau_0}{ND}.$$

Substituting the bounds under diminishing learning rates, we obtain:

$$\epsilon_G \leq L_G \cdot \frac{2L_G}{LND}\left(\frac{T}{s_0}\right)^{\mu L} + \frac{HMs_0}{ND} = \frac{2L_G^2}{LND}\left(\frac{T}{s_0}\right)^{\mu L} + \frac{HMs_0}{ND}.$$

To minimize the upper bound, we set the two terms equal:

$$\frac{2L_G^2}{L}\left(\frac{T}{s_0}\right)^{\mu L} = HMs_0 \implies s_0 = \left(\frac{2L_G^2}{HML}\right)^{\frac{1}{1+\mu L}} T^{\frac{\mu L}{1+\mu L}}.$$

Simplying the ralationship:

$$\epsilon_G \leq 2\frac{HMs_0}{ND} = \frac{2}{ND}\left(\frac{2L_G^2}{L}\right)^{\frac{1}{1+\mu L}} (HMT)^{\frac{\mu L}{1+\mu L}}. \tag{55}$$

□

10.6. Discuss of generalization

We compare the generalization error bounds of various methods in Tab. 9. Our results indicate that FedCVC attains the same generalization error bound as [29], while introducing a tunable parameter β. This parameter directly regulates the amplification of the local update difference h^{t+1} on the dual variable difference z^{t+1}, leading to a stable low ratio of primal to dual residual. Consequently, it compensates for DD's optimization error and reduces error propagation, thereby improving generalization performance in practical optimization.

	Assumption	Generalization
[8]	Lipschitz	$\mathcal{O}\left(\frac{(KT)^{\frac{\mu L}{1+\mu L}}}{ND}\right)$
[23]	Lipschitz, VC	$\mathcal{O}\left(\frac{TK}{\sqrt{TD}}\right)$
[11]	Lipschitz, Bernstein	$\mathcal{O}\left(\frac{TK}{\sqrt{ND}}\right)$
[35]	Lipschitz, Stochastic	$\mathcal{O}\left(\frac{\sqrt{T}}{N\sqrt{ND}}\right)$
[27]	Lipschitz	$\mathcal{O}\left(\frac{(MKT)^{\frac{\mu L}{1+\mu L}}}{ND}\right)$
[29]	Lipschitz	$\mathcal{O}\left(\frac{(MT)^{\frac{\mu L}{1+\mu L}}}{ND}\right)$
our	Lipschitz	$\mathcal{O}\left(\frac{(MT)^{\frac{\mu L}{1+\mu L}}}{ND}\right)$

Table 9. Comparison of generalization error of different algorithms.