Supplementary for "FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction"

A. Appendix: More Experiment Results

We run experiments on the true world datasets of image classification tasks including CIFAR10, CIFAR100, MNIST, fashion MNIST, Tiny ImageNet and EMNIST-L datasets. We also evaluate on a Synthetic dataset. We explored multiple types of models: the FCN for MNIST and EMNIST-L, the CNN architecture network for CIFAR10 and CIFAR100, and a Multi-class logistic network for Synthetic dataset, ResNet18 for Tiny ImageNet. We conduct comprehensive investigations for the impact of client heterogeneity by designing iid and non-iid data scenarios, balance and unbalance data, full client participating and part client participating scenarios. We also justify the robustness of FedDC in flexible devices and large-scale setup by assigning the data to different amounts of clients. For comparison, we utilize the FedAvg, FedProx, Scaffold and FedDyn algorithms as baselines. We detailed describe the experiment settings, the models and datasets, the comparison methods in the following.

A.1. Synthetic Dataset

We conduct experiments on a Synthetic dataset which adopts the same setting as [1]. We generate the samples for each clients $(x, y) \in D_i$, where the samples and labels follow the rule of $y = argmax(\theta_i x + b_i)$, the shape of x is 30×1, y contains 5 categories, θ_i (the shape is 5×30) and b_i (the shape is 5×1) are the best parameter to fitting the data distribution in i-th client. We use γ_1 to control the value of (θ_i, b_i) which sampled from $N(\mu_i, 1)$ where $\mu_i \sim N(0, \gamma_1)$, and γ_2 to control the data distribution in each client. In the experiments on Synthetic dataset, we set only one of γ_1, γ_2 as 1 to allow one type heterogeneity for one set of experiments, and we set all of them as 0 to simulate the homogeneous settings. Thus, the settings for (γ_1, γ_2) include $(0, 0)$, $(0, 1)$, $(1, 0)$, that represent a homogeneous setting and two heterogeneous settings. In all experiments on Synthetic dataset, the amount of clients is 20, the average amount of samples for each client is 200.

A.2. Real World Dataset

Datasets and models. We adopt the real-world datasets for the image classification task, including MNIST, EMNIST-L, CIFAR10, fashion MNIST, Tiny ImageNet and CIFAR100 datasets. The EMNIST-L is used for characters classification, which is a subset of the EMNIST dataset that only contains the first 10 categories. The MNIST is the dataset for the classification of handwritten digits, which contains 10 categories. FashionMNIST is an image dataset that replaces the MNIST handwritten digit set. Different from the MNIST handwriting data set, the Fashion-MNIST data set contains 10 categories of images, namely: t-shirt (T-shirt), trouser (jeans), pullover (pullover), dress (skirt), coat (coat) , Sandal (sandals), shirt (shirt), sneaker (sports shoes), bag (bag), ankle boot (short boots). The sample size of the EMNIST-L, fashion MNIST and MNIST is $(1 \times 28 \times 28)$. For MNIST and fashion MNIST, the sample amount in the training set is 60000, and the sample number in the test set is 10000. For EMNIST-L, the sample amount in the training set is 48000, and the sample amount in the test set is 8000. Both the CIFAR10 and CIFAR100 datasets contain 60000 of $3 \times 32 \times 32$ images. For CIFAR10, there are 10 categories, and there are 100 categories on CIFAR100. For both the CIFAR10 and CIFAR100, the sample amount in the training set is 50000, and the sample amount in the test set is 10000.

A fully-connected network (FCN) as [5] is adopted for the classification of MNIST and EMNIST-L. The FCN includes an input layer, two fully connected hidden layers and an output layer. The two hidden layers both contain 200 neurons. A network with CNN-based structure is employed to classify samples on CIFAR10 and CIFAR100. The CNN follows similar setting as [5], which consists of the basic modules of CNN, including two conventional layers with 64 of 5×5 convolution kernels, each conventional layer followed a down—pooling larger, after that are two fully connected layers with 394 and 192 neurons and a softmax layer for prediction. The classic ResNet18 network is adopted in the Tiny ImageNet dataset.

iid and non-iid setting. The experiments mainly contain three types of balanced data settings, including an iid setting and two non-iid settings. For the iid data distribution, all clients get the same number of samples that are independently identically distributed on the training dataset. For the non-iid settings, we obey the Dirichlet distribution to sample data. In non-iid settings, the label ratio of each client follows the Dirichlet distribution. For each client, its samples are sampled without replacement from the full training dataset according to the label ratio that obeys the Dirichlet distribution. A hyper-parameter of the Dirichlet distribution controls the data heterogeneity degree, and we set two types of Dirichlet distributions where the hyper-parameter is 0.3 and 0.6, respectively. Dirichlet-0.3 distribution is stronger non-iid than Dirichlet -0.6 distribution . In most experiments, we set 100 clients in the experiments. Each of them contains 1% samples of full training data in the balanced settings.

Unbalanced setting. In unbalanced data settings, the sample amount of clients are different from each other. To produce the unbalanced dataset, each client owns data points in which the amount follow a lognormal distribution. The hyper-parameter in the lognormal distribution is the variance of the distribution. In the balanced setting, the variance is 0, and we set the variance as 0.3 in the unbalanced setting.

Hyper-parameter setting. We give the hyper-parameter settings in different datasets. For all the true world datasets including MNIST, EMNIST-L, CIFAR10 and CIFAR100, we set the batch size as 50, the number of local epochs in one communication round as 5, the initial learning rate as 0.1 and the learning rate decay per round as 0.998, the weight decay as 0.001. We search the α of FedDC in $[0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 1]$. In experiments of CIFAR10 and CIFAR100, when the number of client is 100 we set $\alpha = 0.01$ for FedDC, and when the number of clients is 500 we set $\alpha = 0.05$ for FedDC. In experiments with 100 clients, we set the hyper-parameter $\alpha = 0.1$ for MNIST, fashion MNIST and $\alpha = 0.2$ for EMNIST-L in FedDC, respectively. In experiments of MNIST with 500 clients, we set the hyper-parameter $\alpha = 0.2$ for FedDC. In experiments on Tiny ImageNet dataset, we set the client number as 20, besides, the pretrained ResNet18 is adopted. For the Synthetic dataset, we use a multi-class logistic classification model to classify samples, and we set the batch size as 10, the epoch amount in each local communication round as 10, the learning rate as 0.1, the hyper-parameter in FedDC as $\alpha = 0.005$. For the hyper-parameters in baselines, $\alpha = 0.01$ in FedDyn, $\mu = 0.0001$ and weight decay as $1e-5$ in FedProx.

We adopt the same hyper-parameter for a specific dataset for all iid or non-iid data, full client participating or 15% client participating settings. In the experiments, we use ”MNIST iid 100-clients-100%” to represent the result on iid distributed MNIST dataset with 100 clients and 100% client participating, and so on.

A.3. Results

Sensitive of hyper-parameter in FedDC. In FedDC, there is only one manually controlled parameter α. It controls the weight of the penalized term in the local objective function. The α is related to the dissimilarity between local parameters and the global parameter. To analyze the impact of α, we run experiments with different α. The hyper-parameter α is explored in $[0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1]$. The convergence plots of experiments use different α with 100 clients 15% client participating on MNIST and CIFAR10 datasets are shown in Figure 1. A large value of α increases the weight of the penalized term and that leads to less attention to the local experience loss, which will cause the model difficult to converge. As the figure shows, $\alpha = 1$ and $\alpha = 0.5$ get the worst performance. The value of α needs to trade off the empirical loss term and the penalized term, a more reasonable value of α for CIFAR10 is in the range $[0.1, 0.01]$. All experiments of different α converge to a stationary point, but a carefully selected α better improve the model performance.

Ablation study In order to show the effects of the gradient correction term and the penalized term in FedDC, respectively, we conduct the ablation study. We denote the standard local empirical loss as le, the gradient correction term as lg, and the penalized term as lp. The training process of FedDC is different with FedAvg because we use the drift variable to decouple the local models and the global model, in which the global parameters are the sum of local parameters and drift variables. We compare the results of using the process of FedDC with different combinations of local objective functions. $FedDC(le)$ represents the method that adopts the training process of FedDC and uses the standard local empirical loss as the local objective function. With FedDC’s training process, $FedDC(lelg)$ and $FedDC(lelp)$ represent adding a gradient correction term and a penalized term to the local objective function, respectively. $FedDC(lelglp)$ is the proposed method in which the local objective function includes the empirical loss, the penalized term and the gradient correction term. Figure 2 shows the convergence plots of the ablation study on iid distributed and non-iid distributed CIFAR10 dataset with 100 clients 15% client participating. From the figure, we can observe that the performance of the $FedDC(le)$ and $FedDC(lelg)$ methods are the worst, which even significantly lower than the standard FedAvg. The local objective functions of both $FedDC(le)$ and $FedDC(lelg)$ do not contain the penalized term to limit the drift variable, so that the global parameters can not be treated as the sum of the local parameters and the drift variables. Thus, the drift variable does not take effect; instead, it leads to worse performance. Compared with FedAvg, $FedDC(lelp)$ method greatly improves the performance, which indicates that decoupling the local parameters and global parameters with effective drift variables improves the model performance. The $FedDC(lelplg)$ method achieves the best performance compared with all the ablation methods and baselines, which indicates that the gradient correction term can reduce the risk of randomness and make the tracking of parameter drift more accurate.

Convergence plots. We display a lot of convergence plots that trained in different settings and different datasets to verify the robustness and effectiveness of FedDC. Figure 3 shows the convergence plots on the Synthetic dataset with 20 clients 15% and 100% client participating in three types data settings. The results indicate that FedDC is consistently the best compared with the baselines on the Synthetic dataset. Figure 5 shows the convergence curves with the massive devices which contains 500 clients on MNIST, CIFAR10 and CIFAR100 datasets. The results show that FedDC is robust to the setting with large-scale clients. Figure 4 shows the convergence plots of models which are trained on the unbalanced datasets. The results indicate that FedDC is robust on unbalanced settings, FedDC gets the best performance on the unbalanced CIFAR10 and CIFAR100 datasets in both settings of 15% client participating and 100% client participating. Figure 6, 7 and 8 show the convergence plots that trained with different client participating partitions and different data distributions on fashion MNIST, CIFAR10 and CIFAR100. FedDC's accuracy outperforms the baselines, and FedDC also converges faster than baselines. Figure 9 and 10 are the convergence plots on MNIST and EMNIST datasets. In the experiments on MNIST and EMNIST, FedDC achieves competitive performance.

The comparison of convergence speed. We compare the convergence speed of FedDC with baselines on CIFAR10, CIFAR100, MNIST and EMNIST-L datasets, the results are given in Table 2, 3, 4 and 5. We report the communication speedup to achieve the target accuracy in which the benchmark is the number of communication rounds consumed by FedAvg. All results show that FedDC reaches the target accuracy using fewer communication rounds than other methods. As a case, Table 2 reports the communication rounds of different methods (FedAvg, FedProx, Scaffold, FedDyn and FedDC) to achieve three accuracy degrees on CIFAR10 and CIFAR100 datasets with 100 clients and 100% client participating. In the experiments which train models in non-iid distributed data, we compare the FedDC's communication speedup of FedDC with FedAvg method. In the table, the number of communication round (>1000) means the method can not achieve the target accuracy in 1000 rounds, and the $SpeedUp$ with the symbol ($>$) means it is calculated with the benchmark of FedAvg where the communication rounds (>1000). The results indicate that FedDC outperforms all the comparison methods in both iid and non-iid (0.6-Dirichlet and 0.3-Dirichlet distribution) settings. Tacking the settings with 100 clients and 100% client sampling ratio as an example, we can draw the following conclusions. In the experiments on CIFAR10 dataset, when FedAvg costs 149 communication rounds to achieve the 78% accuracy, FedDC only spends 35 rounds in the same setting, where FedDC is $4.25\times$ faster than FedAvg. Besides, FedDyn and Scaffold are $3.47\times$ and $1.67\times$ faster than FedAvg, respectively. In the non-iid setting, we increase the data heterogeneity over clients. In the setting of 0.6-Dirichlet distributed data, FedDC is $4.77\times$ faster than FedAVG to reach 78% accuracy on CIFAR10. When the data is more non-iid distributed (0.3-Dirichlet), the FedDC is over $18.86\times$ faster than FedAvg to reach 80% accuracy on CIFAR10. The convergence curves of the FedAvg method is more stable, which indicates that the local optimization of the FedAvg is slower and causes less fluctuation in parameters. All the above results demonstrate that FedDC outperforms the baselines in both convergence speed and model accuracy.

Comparison with other recent methods. We added experiments to compare with the following three recent methods including FedAdam, FedYogi and FedAdagrad [6]. Table 1 shows the evaluation results. We can find that FedDC performs the best in different settings. They solve the difficulty of tune and exhibiting unfavorable convergence behavior with adaptive optimization methods. Specifically, they propose federated versions of adaptive optimizers, including ADAGRAD, ADAM, and YOGI to improve the convergence in the presence of heterogeneous data. These methods take effect in the parameter aggregation stage with gradient momentum update method, which reduces the negative effect of client drift and accelerates the convergence speed in a certain extent. However, they have no way to solve the parameter drift in the local training phase. FedDC decouples and learns client drift in the client training phase, and uses it to correct local parameters, which has better adaptability. FedDC takes effect in local training phase that is orthogonal to these improved aggregation methods (FedAdam, FedYogi, FedAdagrad etc.), and they can be used in combination.

A.4. Discussion

Impact of data heterogeneity. We set various types of data heterogeneity settings including Dirichlet-0.3 distributed datasets, Dirichlet-0.6 distributed datasets and unbalanced datasets. From the results, we observe that the non-iid degree significantly impacts the model performance for federated learning. With a higher non-iid setting in Dirichlet-0.3 distributed datasets, both the accuracy and the convergence speed of the global model is lower than iid settings. That indicates that data

Table 1. Comparison of FedDC with FedAdam, FedYogi and FedAdagrad. There are 100 clients and 15% of them randomly participate in training per round. The table shows the test accuracy on one iid and two non-iid (D1, D2) settings of CIFAR100 dataset.

Method	iid	D1	D2
FedAdam	40.9%	41.5%	41.6%
FedYogi	42.3%	42.6%	42.7%
FedAdagrad	42.5%	42.5%	42.1%
FedDC	55.4%	54.7%	53.9%

heterogeneity makes the model convergence in federated learning more unstable and challenging. Fortunately, the proposed FedDC has an advantage over baselines in all non-iid settings. FedDC achieves the fastest convergence and the best accuracy compared with baselines, indicating FedDC is much robust to data heterogeneity.

Impact of clients size and client sampling. We first discuss the impact of different client settings for model convergence. Because part of the data cannot accurately describe the global data distribution in each round, part client participating introduce more randomness to the model than full client participating. The total number of data points is fixed so that more clients means fewer samples in each client. The reduction in the amount of client local data would trigger more randomness in local optimization that makes it more challenging to track the parameter drift. Figure 5 shows the results with massive clients. In the massive clients setting, all methods are slower to reach a reasonable performance because each client owns fewer data. The FedDC spend 43 rounds to reach 80% accuracy in 100 clients 100% participating on iid CIFAR10, but that is 143 rounds in the massive setting with 500 clients. The results demonstrate that FedDC has a stronger ability to integrate information from massive clients to save communication and enhance model performance compared with baselines. FedDC brings communication-saving, which results in faster convergence than the baselines. In the convergence plots of CIFAR10, CIFAR100, MNIST and EMNIST-L, we compare the model performance of full participating and part participating. All methods spend more communication rounds to achieve acceptable performance in the experiments with 15% client participating than full participation. FedDC outperforms FedAvg and FedProx a lot in test accuracy, and there are also satisfactory improvements of FedDC over Scaffold and FedDyn. Nevertheless, compared with other methods, FedDC improves the convergence speed and model accuracy significantly. By the way, in the experiments with the different number of clients, we find it is beneficial to model convergence if the hyper-parameter α appropriately increased when the number of clients increases. That indicates that FedDC requires stronger constraints on the penalized terms in heterogeneous settings with bigger randomness.

In summary, FedDC can better handle data heterogeneity, so that FedDC converges faster and obtains better model performance in the experiments. In addition, from the results with different numbers of customers, different data distributions, and different levels of client participation, we conclude that FedDC is strong robustness in various heterogeneous scenarios.

Table 2. The number of communication round in different methods to achieve a target accuracy on CIFAR10 and CIFAR100 while containing with 100 clients which 100% participating each round. The *SpeedUp* denotes the communication-saving relative to FedAvg.

Method	Accuracy	Non-iid (0.6-Dirichlet)		Non-iid (0.3-Dirichlet)		iid	
		Round	SpeedUp	Round	SpeedUp	Round	SpeedUp
CIFAR10 100 clients 100% participating							
FedAvg	0.78	205	-	346	-	149	-
	0.8	>1000	-	>1000	-	286	-
	0.82	>1000	-	>1000	-	803	-
FedProx	0.78	195	1.05×	350	0.99×	142	1.05×
	0.8	474	>2.11×	>1000	1×	277	1.03×
	0.82	>1000	1×	>1000	1×	>1000	1×
Scaffold	0.78×	123	1.67×	148	2.34×	89	1.67×
	0.8	165	>6.06×	218	>4.59×	120	2.38×
	0.82	283	>1.71×	387	>2.58×	194	4.14×
FedDyn	0.78	44	4.66×	57	6.07×	43	3.47×
	0.8	60	>16.67×	75	>17.54×	55	5.2×
	0.82	84	>11.90×	114	>8.77×	75	10.7×
FedDC	0.78	43	4.77×	53	6.53×	35	4.25×
	0.8	53	>18.86×	70	>14.28×	43	6.65×
	0.82	70	>14.28×	114	>8.77×	56	14.34×
CIFAR100 100 clients 100% participating							
FedAvg	0.35	142	-	112	-	201	-
	0.4	476	-	847	-	>1000	-
	0.5	>1000	-	>1000	-	>1000	-
FedProx	0.35	190	0.75×	124	0.9×	145	1.39×
	0.4	502	0.95×	507	1.67×	273	>3.66×
	0.5	>1000	1×	>1000	1×	>1000	1×
Scaffold	0.35	64	2.22×	67	1.67×	58	3.47×
	0.4	91	5.23×	94	9.01×	84	>11.9×
	0.5	424	>2.35×	501	>2×	305	>3.28×
FedDyn	0.35	38	3.74×	38	2.95×	45	4.47×
	0.4	51	9.33×	53	15.98×	56	>17.85×
	0.5	154	>6.49×	182	>5.95×	169	>5.92×
FedDC	0.35	30	4.73×	33	3.39×	29	6.93×
	0.4	39	12.2×	41	20.65×	37	>27.03×
	0.5	70	>14.28×	81	>12.35×	70	>14.28×

Table 3. The number of communication round in different methods to achieve a target accuracy while containing with 100 clients which 15% participating each round. The $SpeedUp$ denotes the communication-saving relative to FedAvg.

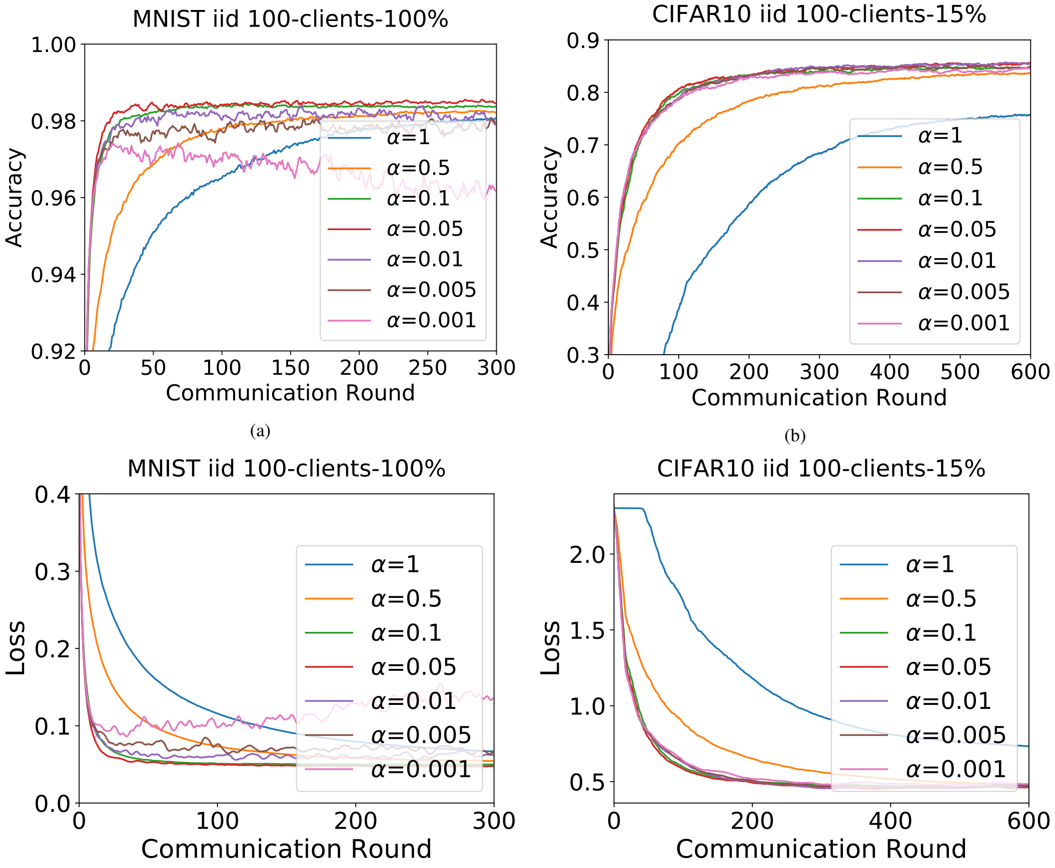
Method	Accuracy	Non-iid (0.6-Dirichlet)		Non-iid (0.3-Dirichlet)		iid	
		Round	SpeedUp	Round	SpeedUp	Round	SpeedUp
CIFAR10 100 clients 15% participating							
FedAvg	0.78	259	-	491	-	177	-
	0.8	616	-	>1000	-	>1000	-
	0.82	>1000	-	>1000	-	>1000	-
FedProx	0.78	228	1.13×	485	1.1×	153	1.15×
	0.8	459	1.34×	>1000	1×	307	>3.28
	0.82	>1000	1×	>1000	1×	>1000	1×
Scaffold	0.78	132	1.96×	169	2.91×	94	1.88×
	0.8	200	3.08×	263	>3.80×	126	>7.93×
	0.82	332	>3.01×	600	>1.67×	204	>4.9×
FedDyn	0.78	118	2.19×	146	3.39×	110	1.61×
	0.8	193	3.19×	195	>5.12×	145	>6.9×
	0.82	254	>3.93×	512	>1.95×	231	>4.33×
FedDC	0.78	101	2.56×	105	4.68×	88	2.01×
	0.8	141	4.37×	143	>6.99×	108	>9.26×
	0.82	211	>4.74×	242	>4.13×	162	>6.17×
CIFAR100 100 clients 15% participating							
FedAvg	0.35	170	-	144	-	260	-
	0.4	615	-	520	-	724	-
	0.5	>1000	-	>1000	-	>1000	-
FedProx	0.35	227	0.75×	148	0.97×	187	1.39×
	0.4	980	0.63×	503	1.03×	650	1.11×
	0.5	>1000	1×	>1000	1×	>1000	1×
Scaffold	0.35	68	2.5×	72	2×	68	3.82×
	0.4	106	5.8×	114	3.56×	113	6.41×
	0.5	>1000	1×	>1000	1×	>1000	1×
FedDyn	0.35	98	1.73×	78	1.46×	106	2.45×
	0.4	149	4.42×	148	3.51×	143	5.06×
	0.5	574	>1.74×	710	>1.41×	619	>1.62×
FedDC	0.35	78	2.18×	74	1.54×	74	3.51×
	0.4	102	6.03×	103	5.05×	100	7.04×
	0.5	249	>4.02×	278	>3.6×	206	>4.85×

Table 4. The number of communication round in different methods to achieve a target accuracy on MNIST and EMNIST-L while containing with 100 clients which 100% participating each round. The $SpeedUp$ denotes the communication-saving relative to FedAvg.

Method	Accuracy	Non-iid (0.6-Dirichlet)		Non-iid (0.3-Dirichlet)		iid	
		Round	SpeedUp	Round	SpeedUp	Round	SpeedUp
MNIST 100 clients 100% participating							
FedAvg	0.96	25	-	28	-	16	-
	0.98	258	-	492	-	142	-
FedProx	0.96	24	1.04×	27	1.04×	16	1×
	0.98	263	0.98×	480	1.03×	136	1.04×
Scaffold	0.96	11	2.27×	14	2×	9	1.78×
	0.98	58	4.45×	58	8.48×	53	2.68×
FedDyn	0.96	8	3.13×	9	3.11×	7	2.29×
	0.98	46	5.61×	51	9.65×	27	5.26×
FedDC	0.96	8	3.13×	10	2.8×	7	2.29×
	0.98	35	7.37×	37	13.3×	26	5.46×
EMNIST-L 100 clients 100% participating							
FedAvg	0.94	142	-	192	-	107	-
	0.95	>300	-	>300	-	>300	-
FedProx	0.94	135	1.05×	198	0.97×	92	1.16×
	0.95	>300	1×	>300	1×	>300	1×
Scaffold	0.94	43	3.30×	52	3.69×	30	3.57×
	0.95	75	>4×	150	>2×	66	> 4.55×
FedDyn	0.94	30	4.73×	52	3.69×	27	3.96×
	0.95	137	>2.19×	160	>1.88×	69	>4.35×
FedDC	0.94	43	3.3×	60	3.2×	21	5.1×
	0.95	78	>3.85×	134	>2.24×	50	>6×

Table 5. The number of communication round in different methods to achieve a target accuracy on MNIST and EMNIST-L while containing with 100 clients which 15% participating each round. The $SpeedUp$ denotes the communication-saving relative to FedAvg.

Method	Accuracy	Non-iid (0.6-Dirichlet)		Non-iid (0.3-Dirichlet)		iid	
		Round	SpeedUp	Round	SpeedUp	Round	SpeedUp
MNIST 100 clients 15% participating							
FedAvg	0.96	32	-	35	-	23	-
	0.98	361	-	>600	-	158	-
FedProx	0.96	31	1.03×	34	1.03×	23	1×
	0.98	383	0.94×	418	>1.44×	149	1.06×
Scaffold	0.96	20	1.6×	23	1.52×	16	1.44×
	0.98	62	5.82×	72	> 8.33×	50	3.16×
FedDyn	0.96	21	1.52×	23	1.52×	18	1.28×
	0.98	122	2.96×	153	>3.92×	71	2.23×
FedDC	0.96	18	1.78×	22	1.59×	16	1.44×
	0.98	60	6.02×	62	> 9.68×	46	3.43×
EMNIST-L 100 clients 15% participating							
FedAvg	0.94	153	-	245	-	108	-
	0.95	>300	-	>300	-	>300	-
FedProx	0.94	145	1.06×	240	1.02×	105	1.03×
	0.95	>300	1×	>300	1×	>300	1×
Scaffold	0.94	44	3.48×	68	3.6×	42	2.57×
	0.95	95	>4.21×	>300	1×	87	>3.45×
FedDyn	0.94	73	2.1×	81	3.06×	61	1.61×
	0.95	127	>2.36×	>300	1×	255	>1.18×
FedDC	0.94	48	3.19×	74	3.31×	47	2.3×
	0.95	92	>3.26×	>300	1×	81	>3.7×



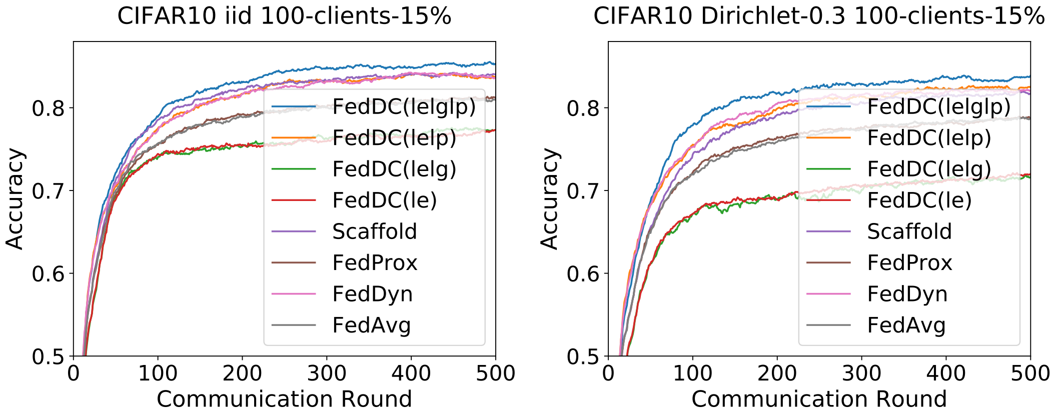
(a)

(b)

(c)

(d)

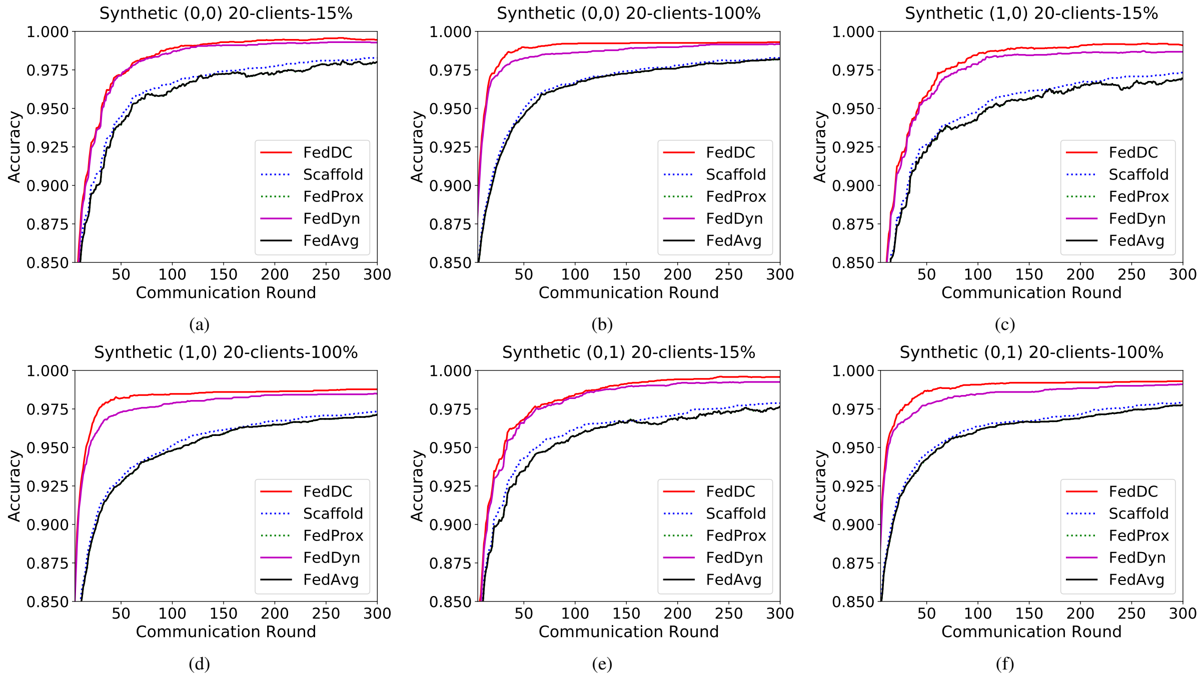
Figure 1. Convergence plots of FedDC for different hyper-parameter-α settings with 100 clients adopting 100% and 15% client participating settings on iid MNIST and CIFAR10.



(a)

(b)

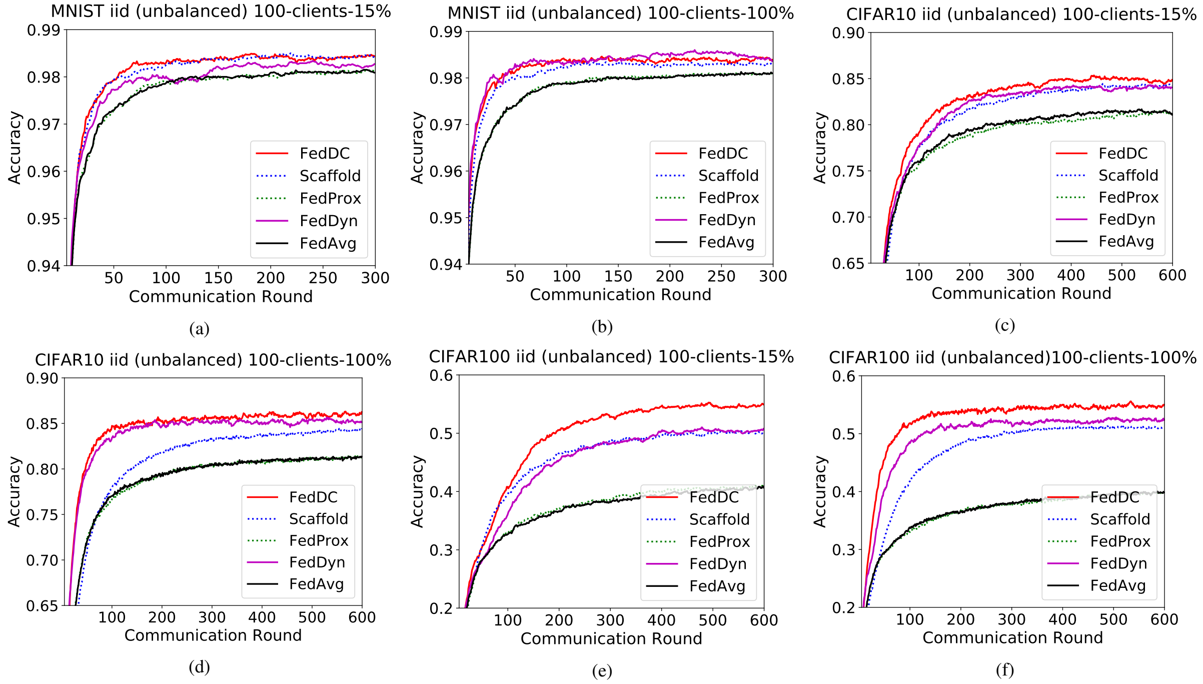
Figure 2. Ablation study for FedDC on CIFAR10. FedDC(le) adopts the proposed training process and uses the empirical loss as local objection function. FedDC(lelp) adopts the proposed training process and uses the sum of the empirical loss and the penalized term as local objection function. FedDC(lelg) adopts the proposed training process and uses the sum of the empirical loss and the gradient correction term as local objection function. FedDC(lelglp) adopts the proposed training process and uses the sum of the empirical loss, the gradient correction term and the penalized term as local objection function.



(a) (b) (c)

(d) (e) (f)

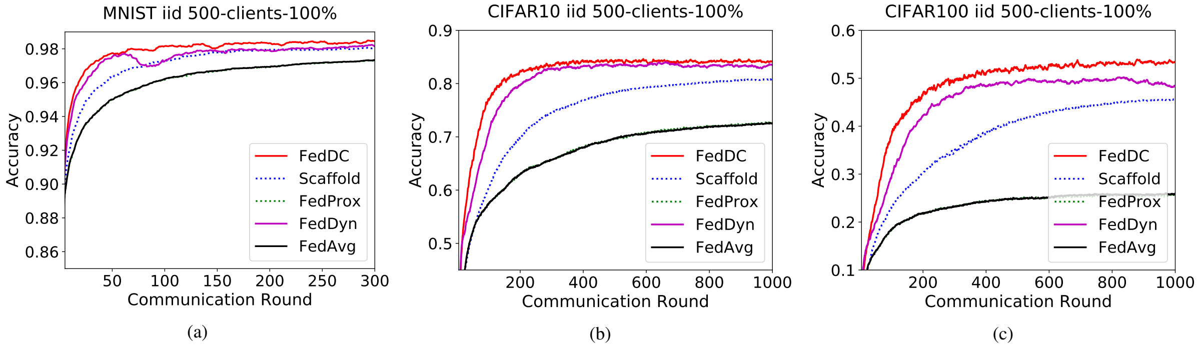
Figure 3. Convergence plots on Synthetic dataset. There are three types of settings, including the homogeneous setting where (γ_1, γ_2) equal $(0, 0)$, and two heterogeneous settings where (γ_1, γ_2) equal $(1, 0)$ and $(0, 1)$, respectively.



(a) (b) (c)

(d) (e) (f)

Figure 4. Convergence plots with 100 clients adopting 100% and 15% client participating settings on unbalanced data of MNIST, CIFAR10 and CIFAR100.

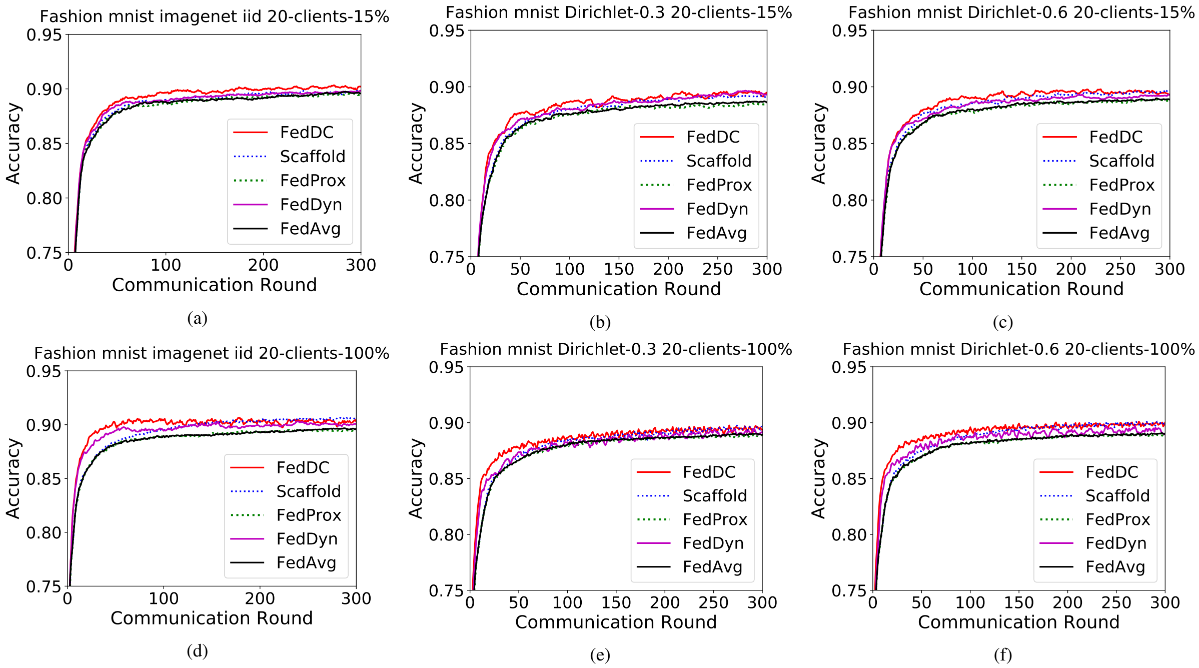


(a)

(b)

(c)

Figure 5. Convergence plots for massive clients (500) with 100% client participating settings in the iid datasets of MNIST, CIFAR10 and CIFAR100.



(a)

(b)

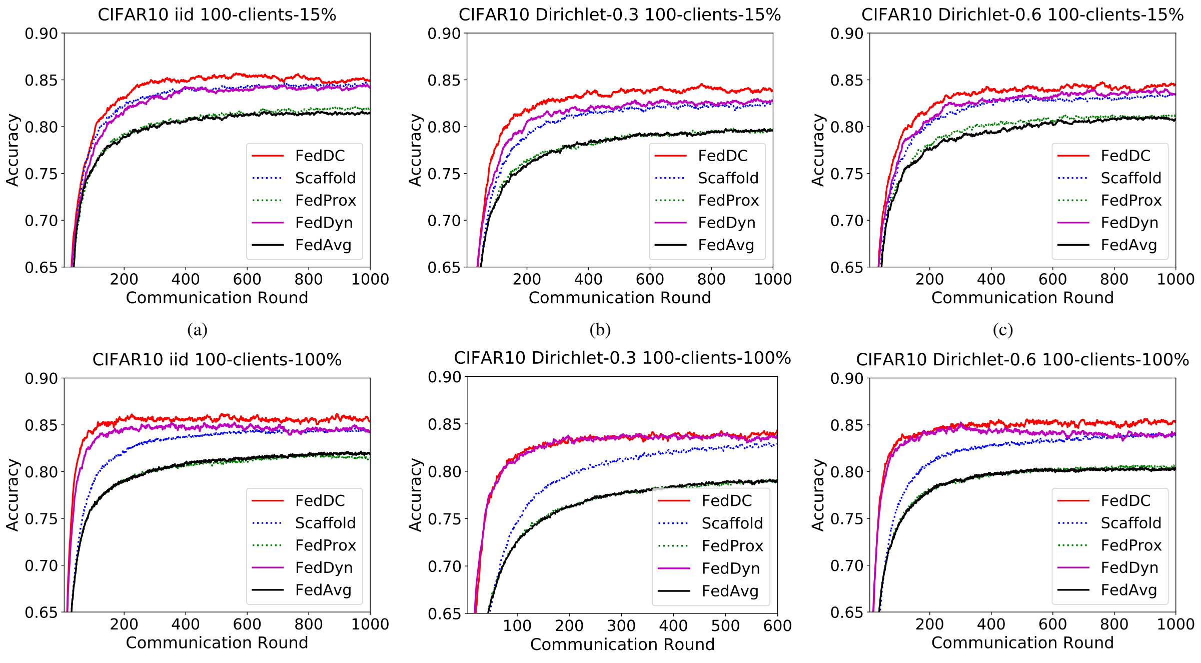
(c)

(d)

(e)

(f)

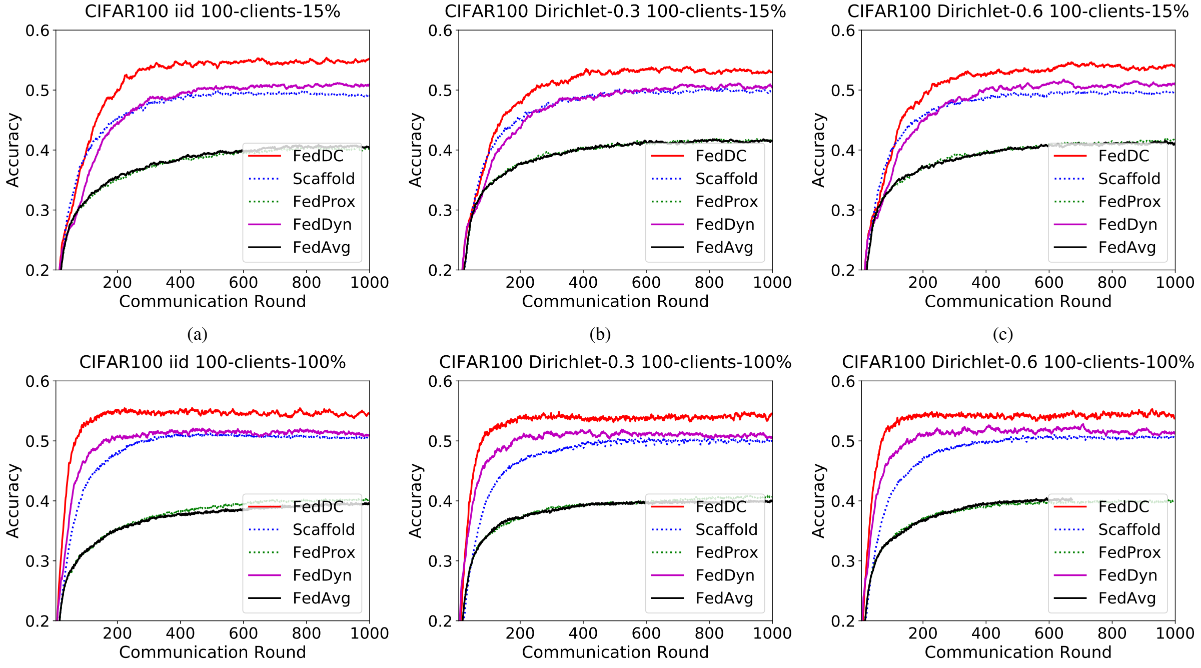
Figure 6. Convergence plots for iid and non-iid data with 100 clients adopting 100% and 15% client participating settings on fashion MNIST.



(a) (b) (c)

(d) (e) (f)

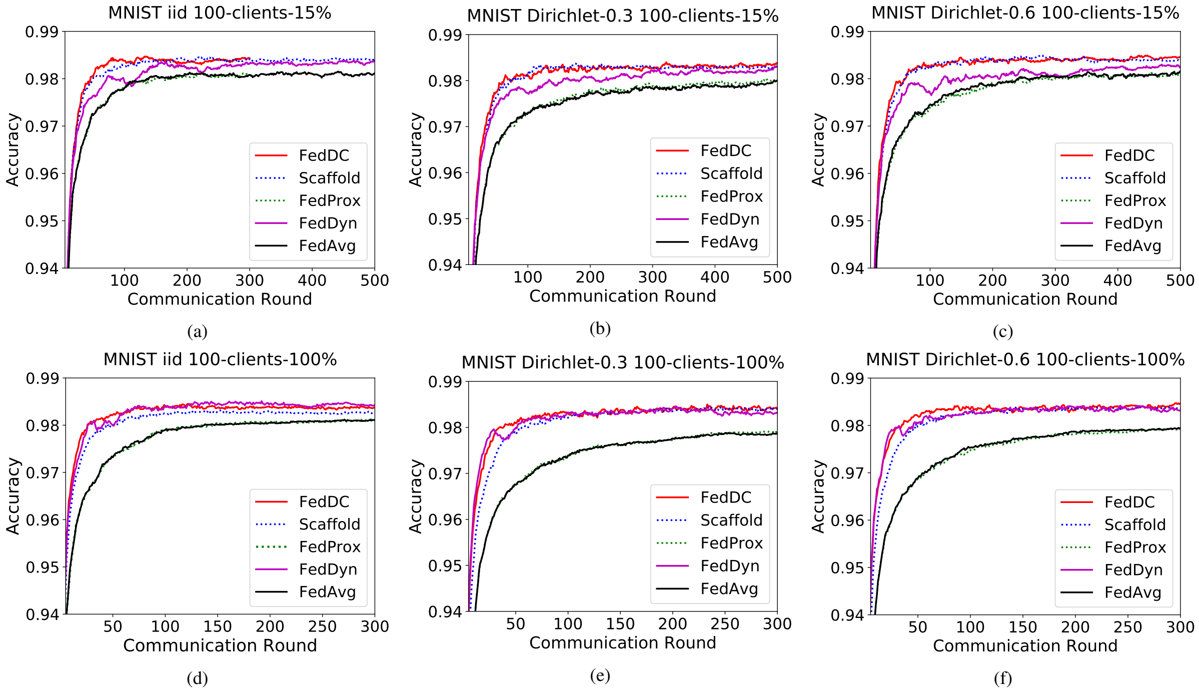
Figure 7. Convergence plots for iid and non-iid data with 100 clients adopting 100% and 15% client participating settings on CIFAR10.



(a) (b) (c)

(d) (e) (f)

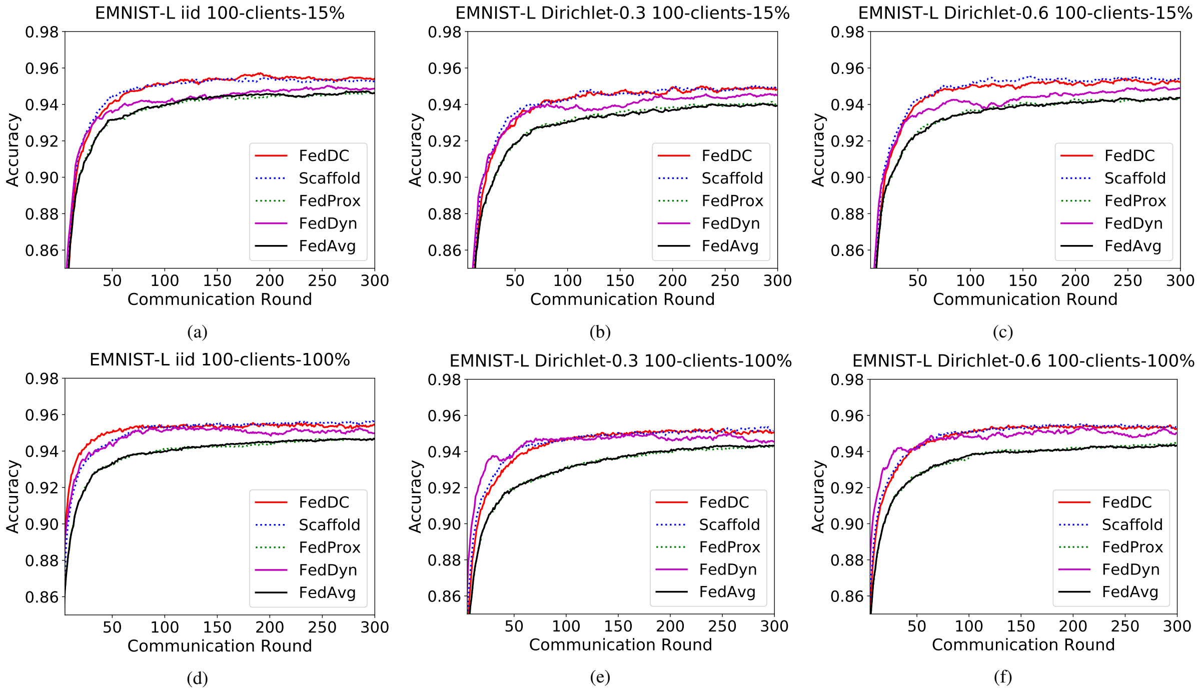
Figure 8. Convergence plots for iid and non-iid data with 100 clients adopting 100% and 15% client participating settings on CIFAR100.



(a) (b) (c)

(d) (e) (f)

Figure 9. Convergence plots for iid and non-iid data with 100 clients adopting 100% and 15% client participating settings on MNIST.



(a) (b) (c)

(d) (e) (f)

Figure 10. Convergence plots for iid and non-iid data with 100 clients adopting 100% and 15% client participating settings on EMNIST-L.

B. Appendix: Convergence Proof of FedDC

First we present algorithm of the proposed FedDC in Algorithm 1. In each training round t, the server first selects the active client set C_t (where $C_t \subseteq [N]$) and boardcasts the global model parameters to them. Then each active client updates the local model parameter and the corresponding local drift variables on its own local datasets. Finally, the server aggregates the sum of the local models and local drift variables to update the global model. In the algorithm, F is the objective loss function, $g_i = \Delta\theta_i$ and $g = \Delta\theta$ are auxiliary variables used in the third term (the gradient correction item $G_i(\theta_i; g_i, g)$) of the objective function F.

Algorithm 1: Algorithm of FedDC

```
Input: Random initial global model parameter $w$, set training round $T$, the number of clients $N$, initial local drift variables as all
       zero matrix, the learning rate $\eta$, the number of local training batches $K$.
Output: The trained global model $w$.
for $t = 1, 2, ..., T$ do
    Sample the active client set $C_t \subseteq [N]$.
    for each client $i \in C_t$ in parallel do
        Set the local model parameter $\theta_i = w$,
        for $k = 1, 2, ..., K$ do
            Update the local model parameter:
                $\theta_i = \theta_i + \eta \frac{\partial F(\theta_i; h_i, D_i, w)}{\partial \theta_i}$,
        Set local gradient drift $\Delta\theta_i = \theta_i - w$,
        Update the local drift: $h_i = h_i + \Delta\theta_i$,
    Update the global model: $w = \frac{1}{|C_t|} \sum_{i \in C_t} (\theta_i + h_i)$,
    Set global gradient drift $\Delta\theta = \frac{1}{|C_t|} \sum_{i \in C_t} \Delta\theta_i$,
Return $w$.
```

We newly define some symbols to facilitate the convergence analysis of FedDC step by step. Specifically, we use the superscript t to represent the communication round, and the subscript i to represent the client index. For example, θ_i^t represents the local model parameter of client i in the t-th round. Considering only the variables to be optimized, the objective function F can be rewritten as:

$$F_i(\theta_i) = \mathbb{E}_{(x,y)\in D_i} l(\theta_i, (x, y)) + \frac{\alpha}{2} ||\theta_i - (w^{t-1} - h_i^{t-1})||^2 \\ + \frac{1}{K\eta} \langle \theta_i, \Delta\theta_i^{t-1} - \Delta\theta^{t-1} \rangle, \tag{1}$$

where $\Delta\theta_i^{t-1} = \Delta g_i^{t-1}$ and $\Delta\theta^{t-1} = \mathbb{E}\Delta g_i^{t-1}$, In the k-th local training iteration of the t-th communication round, the client first optimizes local model with the gradient of local objective function which is represented as follows:

$$\theta_i^{t,k} = \theta_i^{t,k-1} - \eta \nabla_\theta F_i(\theta_i^{t,k-1}). \tag{2}$$

The update value of the corresponding local drift variable is represented as $\Delta h_i^t = -\eta \sum_{k=1}^{K} \nabla_\theta F_i(\theta_i^{t,k}, h_i^{t,k})$. After the local training process completes, the server updates the global parameters based on the updated θ_i^t and h_i^t with the following approach:

$$w^t = \mathbb{E}_{i\in[C_t]}(\theta_i^t + h_i^t) = \mathbb{E}_{i\in[C_t]}(\theta_i^{t-1} + h_i^{t-1} + \Delta\theta_i^t + \Delta h_i^t), \tag{3}$$

where θ_i^t and h_i^t are abbreviations for $\theta_i^{t,K}$ and $h_i^{t,K}$, respectively.

B.1. Discussion of FedDC.

We first present the intuition and results of FedDC convergence analysis. The parameters of clients' local model indirectly align with global parameters by adding the local drift variables. In federated learning, suppose the clients' local optimal points $\theta_i^*, \forall i \in [N]$ are arbitrarily different from each other due to the heterogeneous training data located on various clients. We show that in FedDC, when each client reaches the local optimal points, and the global model also converges to a stationary point at the same time. Based on the Eq. (3), if the local model converges to local optima, then $\Delta\theta_i^t \to 0, \forall i \in [N]$, that implies $w^t = w^{t-1} + \Delta w^{t+1} = w^{t-1} + 2\Delta\theta^t = w^{t-1}$. That indicates the global model also converges when clients' local models all converge to their local stationary points.

Penalized Term. The penalized term is mainly used to help the local drift variables to track the parameter gap. Due to the data heterogeneity among clients, it is impractical to assume that all local models converge to a consistent stationary point. In FedDC, the parameter drift specifically tracks the gap between local models and global models. The parameter deviation from local models to the global model is caused by the following two factors: 1) the update drift in the current round, and 2) the residual parameter deviation. FedProx [3] and Scaffold [2] have proved that reducing the update drift is effective for speeding up the convergence time. However, the residual parameter drift has a cumulative effect between communication rounds, making it more critical to convergence and performance of the training process. We use auxiliary local drift variables to denote the parameter deviation of the client's local model in federated learning from the unbiased global model. In this way, we decouple the training of the global model from the clients' local models. Each client updates its local drift variables under the limitation of the penalized term $\frac{\alpha}{2}||\theta_i - (w - h_i)||^2$ which ensure the effectiveness of the local drift variables.

Gradient correction. To briefly and clearly illustrate the effectiveness of the gradient correction term, we first disregard the drift variable and the penalized term. We assume that all clients are active. Under this condition, we suppose $L_i(\theta_i^t) = \mathbb{E}_{(x,y)\in D_i} l(\theta_i, (x,y))$ and $L(\theta) = \mathbb{E}_{(x,y)\in D} l(\theta, (x,y))$. Client i's corrected gradient which uses the gradient correction term satisfies $g_i^{t,k} = \nabla_\theta L_i(\theta_i^{t,k}) + \frac{1}{K\eta}(\Delta\theta_i^{t-1} - \Delta\theta^{t-1}) \approx \nabla_\theta L_i(\theta_i^{t,k}) + (\nabla_\theta L(\theta_i^{t-1,k}) - \nabla_\theta L_i(\theta_i^{t-1,k}))$ to optimize its model instead of $\nabla_\theta L_i(\theta_i^{t,k})$. The gradient variance is $\frac{1}{N}\sum_{i=1}^N ||g_i^{t,k} - g^{t,k}||^2$, where $g^{t,k} = \frac{1}{N}\sum_{i=1}^N g_i^{t,k} \approx \nabla F(\theta_i^t)$. Thus, the different degrees of local gradient can be expressed as

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}||g_i^{t,k} - g^{t,k}||^2 &\approx \frac{1}{N}\sum_{i=1}^{N}||\nabla_\theta L_i(\theta_i^{t,k}) + (\nabla_\theta L(\theta_i^{t-1,k}) - \nabla_\theta L_i(\theta_i^{t-1,k})) - \nabla_\theta L(\theta_i^{t,k})||^2 \\ &\leq \frac{2}{N}\sum_{i=1}^{N}[||\nabla_\theta L_i(\theta_i^{t,k}) - \nabla_\theta L_i(\theta_i^{t-1,k})||^2 + ||\nabla_\theta L(\theta_i^{t-1,k}) - \nabla_\theta L(\theta_i^{t,k})||^2].\end{aligned} \tag{4}$$

The variance of the local gradients is bounded by the above inequality, which is independent of the dissimilarity of their local objective functions. With the smoothness assumption of L and $L_i, (\forall i \in [N])$, their gradients would not change a lot. We deduce that the gradient of each client is strictly bounded. Thus, the gradient correction term is effective to reduce the gradient drift.

Convergence results. We show the convergence theoretical analysis for FedDC in convex and non-convex functions. With the local drift variable to correct the local parameter, we denote $L(\cdot)$ as the global empirical loss objective and we have $L(w) = F(w)$. We suppose the objective function is β-Lipschitz continuous gradient and B-local dissimilarity bounded under the γ-inexact solution assumption [4], that implies the following expected objective decent in each round:

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - (\frac{2 - 2\gamma B}{\alpha} - \frac{2\beta B(1+\gamma)}{\alpha\bar{\alpha}} - 2\beta\frac{B^2(1+\gamma)^2}{\hat{\alpha}^2})||\nabla L(w^{t-1})||^2, \tag{5}$$

where $\bar{\alpha} = \alpha - \beta_d > 0$ is a constant, C_t is the selected active client set in round t. We can use the objective function decrease to note the convergence on convex and non-convex $L(\cdot)$. In non-convex case, assuming $\Gamma = L(w^0) - L(w^*)$, $p = (\frac{1}{\alpha} - (\frac{\gamma}{\alpha} - \frac{(1+\gamma)\sqrt{2}}{\bar{\alpha}\sqrt{C}} - \frac{\beta(1+\gamma)}{\alpha\bar{\alpha}})\sqrt{1 + \frac{\sigma^2}{\epsilon}} - (\frac{\beta(1+\gamma)^2}{2\bar{\alpha}^2} - \frac{\beta(1+\gamma)^2(2*\sqrt{2}+2)}{\bar{\alpha}^2 C}))(1 + \frac{\sigma^2}{\epsilon}) > 0$, the relation in Eq. (5) holds for FedDC, we get $\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - 2p||\nabla L(w^{t-1})||^2$. Giving a $\epsilon > 0$, we prove that $\sum_{t=1}^T ||\nabla L(w^t)||^2 \leq \epsilon$ when the number of communication round satisfies $T = O(\frac{\Gamma}{p\epsilon})$. If the objective functions are convex, setting $\beta_d = 0, \bar{\alpha} = \alpha$, if $\gamma = 0$, $B \leq \sqrt{C}$ and $1 << B \leq 0.5\sqrt{C}$, we have

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - \frac{2}{\alpha^2}[\alpha(1 - \frac{\sqrt{2}B}{\sqrt{C}}) - (B + (\frac{(2\sqrt{2}+2)}{C} + \frac{1}{2})\beta B^2)]||\nabla L(w^{t-1})||^2. \tag{6}$$

Let $\alpha = 6\beta B^2$, to achieve the convergence state where $\sum_{t=1}^T ||\nabla L(w^t)||^2 \leq \epsilon$, FedDC costs $T = O(\frac{\beta B^2\Gamma}{\epsilon})$ communication rounds. The detailed proof is given in following.

B.2. Detailed convergence Proof of FedDC.

A1: β-smoothness function. f is β-smoothness that satisfies

$$||\nabla f(\theta_1) - \nabla f(\theta_2)|| \leq \beta||\theta_1 - \theta_2||, \quad for \quad \forall\theta_1, \theta_2, \tag{7}$$

that also implies a quadratic upper bound for f,

$$f(\theta_2) \leq f(\theta_1) + \langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle + \frac{\beta}{2} ||\theta_2 - \theta_1||^2. \tag{8}$$

A2: μ-convex function. f is a μ-convex function for $\mu > 0$ that satisfies

$$\langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle \leq f(\theta_2) - f(\theta_1) - \frac{\mu}{2} ||\theta_1 - \theta_2||^2, \quad for \quad \forall \theta_1, \theta_2. \tag{9}$$

D1: B-local dissimilarity bounded. If the local empirical loss L_i is B-local dissimilarity where $\mathbb{E}||\nabla L_i(\theta)||^2 \leq ||\nabla L(w)||^2 B^2$, and we define $B(\theta) = \sqrt{\frac{\mathbb{E}||\nabla L_i(\theta)||^2}{||\nabla L(w)||^2}}$.

A3: γ-inexact solution. We define function $F_i(\theta_i, \hat{\theta}_i)$ as $F_i(\theta_i, \hat{\theta}_i) = L_i(\theta_i) + \frac{\alpha}{2}||\theta_i - \hat{\theta}_i||^2, \quad where \quad \alpha \in [0,1] \quad and \quad \hat{\theta}_i = w - h_i$, we get the gradient of F_i: $\nabla F_i(\theta_i, \hat{\theta}_i) = \nabla L_i(\theta_i) + \alpha(\theta_i - \hat{\theta}_i)$. If θ_i^* is a γ-inexact point of $\min F_i(\theta_i, \hat{\theta}_i)$, it satisfies $||\nabla F_i(\theta_i^*, \hat{\theta}_i)|| \leq \gamma ||\nabla F_i(\hat{\theta}_i, \hat{\theta}_i)||$.

A4. Bounded dissimilarity assumption for L. There exists a B_ϵ while $\epsilon > 0$, for any w, that satisfies $||\nabla L(w)||^2 > \epsilon$, and $B(w) > B_\epsilon$.

Our convergence proof for FedDC use a similar method as that in FedProx [3]. In FedDC, the parameter of global model consists of the average of local model parameters and the average of local drift variables.

$$w = \mathbb{E}_i(\theta_i + h_i) = \mathbb{E}_i \theta_i + \mathbb{E}_i h_i \tag{10}$$

We define a virtual variable $\hat{\theta}_i^t$ as the corrected local parameter in t-th round, that satisfies

$$\hat{\theta}_i^t = w^t - h_i^t. \tag{11}$$

We get $\mathbb{E}\hat{\theta}_i^t = w^t - \mathbb{E}h_i^t = \mathbb{E}_i \theta_i^t$ from the definition of 10 and 11, where $\hat{\theta}_i^t$ is independent with the active client set C_t. In FedDC, we define $L_i = \mathbb{E}_{(x,y) \in D_i} l(\theta_i, (x,y)) + \frac{1}{K\eta} \langle \theta_i, \Delta\theta_i^{t-1} - \Delta\theta^{t-1} \rangle$, thus, the local objective function of i-th client is

$$F_i(\theta_i) = L_i(\theta_i) + \frac{\alpha}{2} ||\theta_i - (w^{t-1} - h_i^{t-1})||^2 = L_i(\theta_i) + \frac{\alpha}{2} ||\theta_i - \hat{\theta}_i^{t-1}||^2. \tag{12}$$

The gradient of F_i in round $t+1$ is

$$\nabla F_i(\theta_i^{t+1}) = \nabla L_i(\theta_i^{t+1}) + \alpha(\theta_i^{t+1} - \hat{\theta}_i^t). \tag{13}$$

In addition, let $L(\theta) = \mathbb{E}_{(x,y) \in D} l(\theta_i, (x,y))$, from the denifition of L_i, we have $L(\theta) = \mathbb{E}_{(x,y) \in D} l(\theta_i, (x,y)) = \mathbb{E}L_i(\theta)$. We define $\bar{\theta}^t = \mathbb{E}\theta_i^t = \mathbb{E}\hat{\theta}_i^t$, the expectation of Eq. 13 satisfies

$$\mathbb{E}_i \nabla F_i(\theta_i^{t+1}) = \mathbb{E}_i \nabla L_i(\theta_i^{t+1}) + \alpha \mathbb{E}_i(\theta_i^{t+1} - \hat{\theta}_i^t) = \mathbb{E}_i \nabla L_i(\theta_i^{t+1}) + \alpha \mathbb{E}(\bar{\theta}^{t+1} - \bar{\theta}^t), \tag{14}$$

then we get

$$\bar{\theta}^{t+1} - \bar{\theta}^t = \frac{1}{\alpha}(\mathbb{E}_i \nabla F_i(\theta_i^{t+1}) - \mathbb{E}_i \nabla L_i(\theta_i^{t+1})). \tag{15}$$

In addition, form the process of FedDC, we get

$$h_i^{t+1} = h_i^t + \theta_i^{t+1} - \theta_i^t \quad \rightarrow \quad h_i^{t+1} - h_i^t = \theta_i^{t+1} - \theta_i^t, \tag{16}$$

so that the difference of the global parameters in $t+1$-th round and t-th round is

$$w^{t+1} - w^t = \mathbb{E}[(h_i^{t+1} - h_i^t) + (\theta_i^{t+1} - \theta_i^t)] = 2(\bar{\theta}^{t+1} - \bar{\theta}^t). \tag{17}$$

Theorem 1: Convergence of FedDC in non-convex case. For non-convex and β-Lipschitz smooth function $L_i, \forall i \in [N]$, there exists a $\beta_d > 0$, where $\bar{\alpha} = \alpha - \beta_d > 0$ and $\nabla^2 L_i \geq -\beta_d I$. We assume the local empirical loss L_i is non-convex and B-dissimilarity, in which $B(\theta^t) \leq B$. The global objective of FedDC decreases as follows:

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - 2p||\nabla L(w^{t-1})||^2, \tag{18}$$

where $p = (\frac{\gamma}{\alpha} - \frac{B(1+\gamma)\sqrt{2}}{\bar{\alpha}\sqrt{N}} - \frac{\beta B(1+\gamma)}{\alpha\bar{\alpha}} - \frac{\beta(1+\gamma)^2 B^2}{2\bar{\alpha}^2} - \frac{\beta B^2(1+\gamma)^2(2\sqrt{2C}+2)}{\bar{\alpha}^2 N}) > 0$, C_t is the active client set in round t which contains C clients.

Proof for Theorem 1. In the proof, we follow the techniques of [3], assume the local empirical loss L_i is γ-inexactness solver. We define e_i^t as

$$\nabla L_i(\theta_i^t) + \alpha(\theta_i^t - \hat{\theta}_i^{t-1}) - e_i^t = 0. \tag{19}$$

In addition, we have $\nabla F_i(\hat{\theta}_i^{t-1}, \hat{\theta}_i^{t-1}) = \nabla L_i(\hat{\theta}_i^{t-1})$, so with the B-local dissimilarity bounded assumption we can get: $||\nabla F_i(\theta_i^t, \hat{\theta}_i^{t-1})|| \leq ||\nabla F_i(\theta_i^*, \hat{\theta}_i^{t-1})|| \leq \gamma||\nabla F_i(\hat{\theta}_i^{t-1}, \hat{\theta}_i^{t-1})||$, that implies

$$||e_i^t|| \leq \gamma||\nabla L_i(\hat{\theta}_i^{t-1})||. \tag{20}$$

As $\bar{\theta}^t = \mathbb{E}_i\theta_i^t = \mathbb{E}_i\hat{\theta}_i^t$, so that we get the following equation

$$\bar{\theta}^t - \bar{\theta}^{t-1} = \mathbb{E}_i[\theta_i^t - \hat{\theta}_i^{t-1}] = \frac{1}{\alpha}\mathbb{E}_i(-\nabla L_i(\theta_i^t) + e_i^t). \tag{21}$$

Let $\bar{\alpha} = \alpha - L_d > 0$ and $\ddot{\theta}_i^t = \arg\min_\theta F_i(\theta, \hat{\theta}_i^{t-1})$. Due to that F_i is $\bar{\alpha}$ strong convex function, we get

$$||\ddot{\theta}_i^t - \theta_i^t|| \leq \frac{\gamma}{\bar{\alpha}}||\nabla L_i(\hat{\theta}_i^{t-1})||. \tag{22}$$

With the strong convex nature of F_i again, we get

$$||\ddot{\theta}_i^t - \hat{\theta}_i^{t-1}|| \leq \frac{1}{\bar{\alpha}}||\nabla L_i(\hat{\theta}_i^{t-1})||. \tag{23}$$

Using triangle inequality for 22 and 23, we get:

$$||\theta_i^t - \hat{\theta}_i^{t-1}|| \leq \frac{1+\gamma}{\bar{\alpha}}||\nabla L_i(\hat{\theta}_i^{t-1})||. \tag{24}$$

With the bounded dissimilarity assumption and $\mathbb{E}\hat{\theta}_i^{t-1} = \mathbb{E}\hat{\theta}^{t-1} = \bar{\theta}_i^{t-1}$, we get

$$\begin{aligned} ||\bar{\theta}^t - \bar{\theta}^{t-1}|| &\leq \mathbb{E}_i||\theta_i^t - \hat{\theta}_i^{t-1}|| \leq \frac{1+\gamma}{\bar{\alpha}}\mathbb{E}_i||\nabla L_i(\hat{\theta}_i^{t-1})|| \\ &\leq \frac{1+\gamma}{\bar{\alpha}}\sqrt{\mathbb{E}_i||\nabla L_i(\hat{\theta}_i^{t-1})||^2} \leq B\frac{1+\gamma}{\bar{\alpha}}||\nabla L(\bar{w}^{t-1})||, \end{aligned} \tag{25}$$

where the last inequality is due to the bounded dissimilarity assumption and $F(w) = L(w)$, $\nabla\nabla_w F(w) = \nabla_\theta L_i(\theta_i)$.

We define M_t as $\bar{\theta}^t - \hat{\theta}^{t-1} = -\frac{1}{\alpha}(\nabla L(w^{t-1}) + M_t))$. Taking Eq. 21 into it, we get $M_t = \mathbb{E}_i[(\nabla L_i(\theta_i^t) - \nabla L_i(\hat{\theta}_i^{t-1}) - e_i^t]$. M_t is bounded with

$$\begin{aligned} ||M_t|| &\leq \mathbb{E}_i[(\beta||\theta^t - \hat{\theta}^{t-1}|| + ||e_i^t||] \leq (\frac{\beta(1+\gamma)}{\bar{\alpha}} + \gamma)\mathbb{E}_i||\nabla L_i(\hat{\theta}^{t-1})|| \\ &\leq (\frac{\beta(1+\gamma)}{\bar{\alpha}} + \gamma)B||\nabla L(w^{t-1})||, \end{aligned} \tag{26}$$

The last is due to $\nabla_w L(w) = \nabla_w F(w) = \nabla_\theta L_i(\theta)$ and the bounded dissimilarity assumption.

Because $h_i^t = h_i^{t-1} + \Delta\theta_i^t$, we get $\mathbb{E}(\theta_i^t - \theta_i^{t-1}) = \mathbb{E}(h_i^t - h_i^{t-1}) = \mathbb{E}(\bar{\theta}^t - \ddot{\theta}^{t-1})$, and $w^t - w^{t-1} = \mathbb{E}(\theta_i^t + h_i^t) - \mathbb{E}(\theta_i^{t-1} + h_i^{t-1}) = 2\mathbb{E}(\theta_i^t - \theta_i^{t-1})$.

With β-Lipschitz smoothness assumption of L and Taylor expansion, we get

$$\begin{aligned} L(w^t) &\leq L(w^{t-1}) + < \nabla L(w^{t-1}), w^t - w^{t-1} > + \frac{\beta}{2}||w^t - w^{t-1}||^2 \\ &\leq_1 L(w^{t-1}) + < \nabla L(w^{t-1}), 2\mathbb{E}_i(\theta_i^t - \theta_i^{t-1}) > + \frac{\beta}{2}||2\mathbb{E}(\theta_i^t - \theta_i^{t-1})||^2 \\ &\leq_2 L(w^{t-1}) - \frac{2}{\alpha}||\nabla L(w^{t-1})||^2 - \frac{2}{\alpha} < \nabla L(w^{t-1}), M_t > + \frac{2\beta B^2(1+\gamma)^2}{\hat{\alpha}^2}||\nabla L(w^{t-1})||^2 \\ &\leq L(w^{t-1}) - (\frac{2-2\gamma B}{\alpha} - \frac{2\beta B(1+\gamma)}{\alpha\bar{\alpha}} - 2\beta\frac{B^2(1+\gamma)^2}{\hat{\alpha}^2})||\nabla L(w^{t-1})||^2, \end{aligned} \tag{27}$$

where $(\leq_1)$ is due to $w^t - w^{t-1} = 2\mathbb{E}(\theta_i^t - \theta_i^{t-1})$, $(\leq_2)$ is due to the definition of M. Set a proper α for the above inequality, $L(w^t) - L(w^{t-1})$ is decrease proportional to $||\nabla L(w^{t-1})||^2$. The above inequality demonstrates that if the hyper-parameter α of the penalized term is large enough, the works would be decreased.

Proof for partial client participation settings. In practice, FedDC runs on sampled active clients each round. We assume there are C clients are chosen randomly to the active set C_t in round t. With a local Lipschitz continuity assumption for L, if β_l is the continuity constant, we get

$$L(w_2) \leq L(w_1) + \beta_l ||w_1 - w_2||, \quad for \quad \forall w_1, w_2, \tag{28}$$

besides, we assume $\theta^t = \mathbb{E}_{C_t}\theta_i^t$ and $\bar{\theta}^t = \mathbb{E}_i\theta_i^t$, the following satisfies that

$$\begin{aligned} \beta_l &\leq ||\nabla L(w^{t-1})|| + \beta \max(||\bar{w}^t - w^{t-1}||, ||w^t - w^{t-1}||) \\ &\leq ||\nabla L(w^{t-1})|| + \beta(||\bar{w}^t - w^{t-1}|| + ||w^t - w^{t-1}||). \end{aligned} \tag{29}$$

So that in the partial client participating settings we need to bound

$$\mathbb{E}_{C_t} L(w^t) \leq L(\bar{w}^t) + \mathbb{E}_{C_t}\beta_l ||w^t - \bar{w}^t|| \leq L(\bar{w}^t) + \mathbb{E}_{C_t}\beta_l ||w^t - \bar{w}^t||, \tag{30}$$

where the expectation is calculated on the active client set C_t.

$$\begin{aligned} &\mathbb{E}_{C_t}\beta_l ||w^t - \bar{w}^t|| \leq \mathbb{E}_{C_t}[||\nabla L(w^{t-1})|| + \beta(||\bar{w}^t - w^{t-1}|| + ||w^t - w^{t-1}||)] * ||w^t - \bar{w}^t|| \\ &\leq [||\nabla L(w^{t-1})|| + \beta||\bar{w}^t - w^{t-1}||] * \mathbb{E}_{C_t}||w^t - \bar{w}^t|| + \beta\mathbb{E}_{C_t}||w^t - \bar{w}^t|| * ||w^t - w^{t-1}|| \\ &\leq (||\nabla L(w^{t-1})|| + 2\beta(||\bar{w}^t - w^{t-1}||)\mathbb{E}_{C_t}||w^t - \bar{w}^t|| + \mathbb{E}_{C_t}||w^t - \bar{w}^t||^2. \end{aligned} \tag{31}$$

Taking $\mathbb{E}(\theta_i^t - \theta_i^{t-1}) \leq B\frac{1+\gamma}{\bar{\alpha}}||\nabla L(\hat{\theta}^{t-1})||$ from 26, we have

$$\mathbb{E}_{C_t}||\theta^t - \bar{\theta}^t|| \leq \sqrt{\mathbb{E}_{C_t}||\theta^t - \bar{\theta}^t||^2}, \tag{32}$$

and

$$\begin{aligned} \mathbb{E}_{C_t}||\theta^t - \bar{\theta}^t||^2 &\leq \frac{1}{C}\mathbb{E}_{i \in C_t}(||\theta_i^t - \bar{\theta}^t||^2) \\ &\leq \frac{2}{C}\mathbb{E}_{i \in C_t}||\theta_i^t - \theta^{t-1}||^2 \\ &\leq \frac{2}{C}\frac{(1+\gamma)^2}{\bar{\alpha}^2}\mathbb{E}_{i \in C_t}||\nabla L_i(\theta^{t-1})||^2, \end{aligned} \tag{33}$$

where the last inequality is due to bounded dissimilarity assumption. Further, with h_i fixed, we get $\mathbb{E}_{C_t}||w^t - \bar{w}^t||^2 = \mathbb{E}_{C_t}||\theta^t - \bar{\theta}^t||^2 \leq \frac{2}{C}\frac{(1+\gamma)^2}{\bar{\alpha}^2}\mathbb{E}_{i \in C_t}||\nabla L_i(\theta^{t-1})||^2 \leq \frac{2B^2(1+\gamma)^2}{C\bar{\alpha}^2}||\nabla L(w^{t-1})||^2$. Replace the bound in 31, the inequality becomes

$$\mathbb{E}_{C_t}\beta_l ||w^t - \bar{w}^t|| \leq (\frac{B\sqrt{2}(1+\gamma)}{\bar{\alpha}\sqrt{C}} + \frac{\beta B^2(1+\gamma)^2}{\bar{\alpha}^2 C}(2 * \sqrt{2} + 2))||\nabla L(w^{t-1})||^2. \tag{34}$$

We combine 27,30,34 to get

$$\begin{aligned} &\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - 2(\frac{1-\gamma B}{\alpha} - \frac{B(1+\gamma)\sqrt{2}}{\bar{\alpha}\sqrt{C}} - \\ &\frac{\beta B(1+\gamma)}{\alpha\bar{\alpha}} - \frac{\beta(1+\gamma)^2 B^2}{2\bar{\alpha}^2} - \frac{\beta B^2(1+\gamma)^2(2 * \sqrt{2} + 2)}{\bar{\alpha}^2 C})||\nabla L(w^{t-1})||^2. \end{aligned} \tag{35}$$

B.3. Bounded Gradients

We get prove the following corollary. We hold the bounded dissimilarity assumption for any L_i. The bounded variance of gradients is

$$\mathbb{E}||\nabla L_i(\theta) - \nabla L(w)||^2 \leq \sigma^2, \quad \forall\epsilon > 0, \tag{36}$$

Then we get $B_\epsilon \leq \sqrt{1+\frac{\sigma^2}{\epsilon}}$. We can restate the convergence result in Theorem 1 based on this Corollary and the the bounded variance assumption.

Proof for bounded Gradients. We get the following inequalities,

$$\begin{aligned}
&\mathbb{E}||\nabla L_i(\theta) - \nabla L(w)||^2 = \mathbb{E}||\nabla L_i(\theta)|| - ||\nabla L(w)||^2 \leq \mathbb{E}||\nabla L_i(\theta) - \nabla L(w)||^2 \leq \sigma^2 \\
&\mathbb{E}||\nabla L_i(\theta)||^2 \leq \sigma^2 + ||\nabla L(w)||^2 \\
&B_\epsilon = (\frac{(\mathbb{E}_i||\nabla L_i(\theta)||^2)}{||\nabla L(w)||^2})^{\frac{1}{2}} \leq (1+\frac{\sigma^2}{\epsilon})^{\frac{1}{2}}.
\end{aligned} \tag{37}$$

B.4. Convergence of FedDC in non-convex case

Assumption:B-local dissimilarity If L_i is non-convex, β-Lipschitz smooth function, and B-local dissimilarity bounded. There existing β_d makes $\nabla^2 L_i \geq -\beta_d - I$ and $\bar{\alpha} = \alpha - \beta_d > 0$. $B(\theta) \leq B$. We can select α, C, γ which satisfies that:

$$p = (\frac{1}{\alpha} - (\frac{\gamma}{\alpha} - \frac{(1+\gamma)\sqrt{2}}{\bar{\alpha}\sqrt{C}} - \frac{\beta(1+\gamma)}{\alpha\bar{\alpha}})\sqrt{1+\frac{\sigma^2}{\epsilon}} - (\frac{\beta(1+\gamma)^2}{2\bar{\alpha}^2} - \frac{\beta(1+\gamma)^2(2*\sqrt{2}+2)}{\bar{\alpha}^2 C}))(1+\frac{\sigma^2}{\epsilon}) > 0. \tag{38}$$

In each round of FedDC, the global objective decreases as

$$\begin{aligned}
\mathbb{E}_{C_t} L(w^t) \leq\ & L(w^{t-1}) - 2(\frac{1}{\alpha} - (\frac{\gamma}{\alpha} - \frac{(1+\gamma)\sqrt{2}}{\bar{\alpha}\sqrt{C}} - \frac{\beta(1+\gamma)}{\alpha\bar{\alpha}})\sqrt{1+\frac{\sigma^2}{\epsilon}} \\
& - (\frac{\beta(1+\gamma)^2}{2\bar{\alpha}^2} - \frac{\beta(1+\gamma)^2(2*\sqrt{2}+2)}{\bar{\alpha}^2 C}))(1+\frac{\sigma^2}{\epsilon}) * ||\nabla L(w^{t-1})||^2 \\
\leq\ & L(w^{t-1}) - 2p||\nabla L(w^{t-1})||^2.
\end{aligned} \tag{39}$$

B.5. Convergence of FedDC in convex case

We suppose $\beta_d = 0, \bar{\alpha} = \alpha$ in the convex case, if $\gamma = 0$, $B \leq \sqrt{C}$, we can find that $||\nabla_L(w^t)||$ is proportional decreased. Assuming $1 << B \leq 0.5\sqrt{C}$, we get

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - 2(\frac{1-\frac{\sqrt{2}B}{\sqrt{C}}}{\alpha} - \frac{B + (\frac{(2\sqrt{2}+2)}{C} + \frac{1}{2})\beta B^2}{\alpha^2})||\nabla L(w^{t-1})||^2, \tag{40}$$

and

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - (\frac{1}{\alpha} - \frac{3\beta B^2}{\alpha^2})||\nabla L(w^{t-1})||^2. \tag{41}$$

Setting $\alpha = 6\beta B^2$, we get

$$\mathbb{E}_{C_t} L(w^t) \leq L(w^{t-1}) - \frac{1}{12\beta B^2}||\nabla L(w^{t-1})||^2. \tag{42}$$

We can use the decrease in the global objective according to above inequality to characterize the FedDC's convergence rate. To achieve a threshold ϵ where $\sum_{t=1}^{T} ||\nabla L(w^t)||^2 \leq \epsilon$, if the model achieve the optimal point at T round, we denoted $\Gamma = L(w^0) - L(w^*)$. From 42 and the above definition, we get:

$$\begin{aligned}
&\mathbb{E}_{S_T} L(w^T) - L(w^0) = \mathbb{E}_{C_t} L(w^*) - L(w^0) \leq -\sum_{t=1}^{T} \frac{1}{12\beta B^2}||\nabla L(w^{t-1})||^2 \\
&\rightarrow \sum_{t=1}^{T} ||\nabla L(w^{t-1})||^2 \leq 12\beta B^2(L(w^0) - L(w^*)).
\end{aligned} \tag{43}$$

Thus, FedDC spend $O(\frac{\beta B^2 \Gamma}{\epsilon})$ to achieve convergence state where $\sum_{t=1}^{T} ||\nabla L(w^{t-1})||^2 \leq \epsilon$.

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. 1

[2] Sai Praneeth Reddy Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Jakkam Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning (ICML)*, pages 5132–5143, 2020. 14

[3] 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. 14, 15, 16

[4] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020. 14

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 1

[6] Sashank J. Reddi, Zachary Charles, et al. Adaptive federated optimization. In *ICLR*, 2021. 3