FedAdamom: Adaptive Momentum for Improved Generalization in Federated Optimization

Supplementary Material

A. Experiment Details

A.1. Real Data

We evaluate our models on three benchmark datasets: CIFAR-10, CIFAR-100 and TinyImagenet. For image classification tasks, CIFAR-10, and CIFAR-100, TinyImagenet are used, with image dimensions of $3 \times 32 \times 32$ for both CIFAR-10 and CIFAR-100, $3 \times 64 \times 64$ for TinyImagenet. These datasets contain 10, 100 and 200 classes, respectively.

To generate Non-IID data partitions for CIFAR-10, CIFAR-100 and TinyImagenet, we allocate training samples to clients based on class labels. Specifically, following prior works [19, 40], we use the Dirichlet distribution [13] to create federated heterogeneous datasets by sampling a class-probability vector for each client, where each vector is drawn from a Dirichlet distribution with a concentration parameter that controls the degree of data heterogeneity. For each client, labels are sampled according to this probability vector, and corresponding images are drawn without replacement. This process is repeated until all data points are allocated. As a result, each client's label distribution follows the Dirichlet distribution, with the concentration parameter governing the level of statistical heterogeneity across devices. For instance, when the Dirichlet factor is set to 0.3, approximately 80% of each client's data is concentrated in 3–4 dominant classes. In the IID setting, data is randomly shuffled and evenly distributed across clients. For the FEMNIST, CelebA and Shakespeare, we adopt the LEAF framework [3] to generate Non-IID data.

A.2. Hyper-parameters

Implementation details. In our experiments, all algorithms are implemented using PyTorch 2.0.0 with CUDA 11.8 on a GEFORCE RTX 4090 GPU. We consider different hyperparameter configurations for various setups and datasets. In the moderate-scale experiments, we fix the batch size to 50. In the large-scale experiments, we set the batch size to 20. In addition, the batch size is set to 100 for Shakespeare, and 50 for both CelebA and FEMNIST. Hyperparameters are searched for all algorithms in all IID and Dirichlet settings for a fixed 100 communication rounds. The learning rate decay is selected from the range of $[0.99, 0.998, 0.9995, 1.0]$. The weight decay is selected from the range of $[0.01, 0.001, 0.0001, 0.00001]$. The learning rate is selected from the range of $[0.001, 0.01, 0.1, 1.0]$. The global learning rate is set to 1, except for FedAdam, FedYogi, FedAdagrad, FAFED and FADAS which is set to 0.01.

Hyperparameter selection. To reproduce the competing methods, we mainly follow the configurations provided in their original papers and adjust certain parameters only when it leads to clear performance improvements. Specifically, β is is chosen from $[0.8, 0.85, 0.9, 0.95]$ in FedAvgM and FedAdamom, $[0.99, 0.95, 0.9]$ in FedAdam, FedYogi, FedAdagrad, FedCAda, FAFED and FADAS. For β_0, in FedAdam, FedYogi, FedAdagrad, FedCAda, FAFED, FADAS, choices are from $[0.01, 0.05, 0.1]$.

A.3. Evaluation on Various Data Heterogeneity

To evaluate the performance of FedAdamom under different levels of data heterogeneity, we conduct experiments on CIFAR-10 and CIFAR-100 using a mild heterogeneity setting with a Dirichlet distribution parameter of 0.6. Tables B show that FedAdamom also matches or outperforms the performance of competitive methods. As shown in Table 4, FedAdamom still shows strong robustness to data heterogeneity.

A.4. Evaluation on Large-scale

To evaluate the performance of FedAdamom under different client scales and extremely low participation rates, we conducted experiments with 500 clients and a 2% participation rate. As shown in Table 5, the results for the large-scale setting show a lower overall performance than those in the moderate-scale experiments. This is because each client has fewer training samples and the data distribution becomes more heterogeneous. However, FedAdamom still shows strong robustness to data heterogeneity and low participation rates.

Table 4. $\alpha = 0.6$: 100 clients, 5% participation

Method	CIFAR-10				CIFAR-100			
	Rounds (↓)		Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)	
	75%	81%	500R	1000R	45%	50%	500R	1000R
FedAvg [34]	195	303	83.23	86.70	342	470	50.23	56.89
FedAvgM [13]	178	279	83.56	89.09	309	432	51.15	58.51
FedAdam [40]	179	294	83.31	86.54	285	476	50.31	57.24
FedAdagrad [40]	183	332	83.01	86.28	301	484	50.16	57.15
FedYogi [40]	170	261	86.55	89.25	270	412	52.14	58.72
FAFED [51]	184	297	83.22	86.16	386	490	47.48	56.55
FedCAda [61]	173	264	85.57	87.65	410	1000	48.09	53.58
FADAS [49]	179	264	86.81	88.33	278	457	50.70	57.55
FedAdamom (ours)	**176**	**248**	**87.83**	**89.83**	**234**	**389**	**53.76**	**60.04**

Table 5. Large-scale: 500 clients, 2% participation

Method	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	Rounds (↓)		Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)	
	75%	81%	500R	1000R	40%	42%	500R	1000R	25%	30%	500R	1000R
FedAvg [34]	610	896	74.53	81.77	828	895	31.34	45.34	412	626	28.97	36.86
FedAvgM [13]	577	882	74.64	81.79	663	782	33.10	47.11	396	564	29.30	37.16
FedAdam [40]	534	908	74.50	81.71	714	845	31.70	45.54	343	545	29.21	36.87
FedAdagrad [40]	548	943	73.76	81.01	798	861	31.31	44.87	384	560	28.69	36.24
FedYogi [40]	**485**	877	74.68	82.09	**630**	775	33.18	47.16	328	503	29.87	37.38
FAFED [51]	636	1000+	73.17	80.97	935	1000+	29.19	40.30	401	592	28.35	36.05
FedCAda [61]	513	865	74.01	81.93	1000+	1000+	30.46	34.18	**322**	651	27.76	33.93
FADAS [49]	501	864	74.52	81.99	746	856	32.22	46.25	335	515	29.97	37.41
FedAdamom (ours)	544	**831**	**74.03**	**82.83**	635	**737**	**36.83**	**48.15**	327	**483**	**30.86**	**37.98**

A.5. Evaluation on CNN

To evaluate the performance of FedAdamom under different network architectures, we conduct experiments on the CIFAR-10 and CIFAR-100 datasets using a CNN model. As shown in Table 6, FedAdamom achieves comparable or superior performance to competitive methods under the CNN architecture.

A.6. Communication and Computation Cost

Let d denote the model dimension and s the number of participating clients. Both FedAdam and our method have a computational complexity of $\Theta(sd)$, arising from gradient aggregation across participating clients. The memory complexity of both methods is $\Theta(d)$, due to storing the global update gradient Δ, second moment v_t, and momentum m_t (where $\beta_{1,t}$ and v_t sharing memory). We evaluate FedAdam and our method on ResNet-18 with CIFAR-10 and measure the average server-side computation and memory overhead per round. FedAdam requires $0.042s$ of computation and $70.15M$ of memory, while our method incurs $0.052s$ and $70.60M$, respectively.

A.7. More Sensitivity Analysis

By default, we set $\epsilon = 10^{-3}$, $\beta_2 = 0.05$, $\eta_l = 0.1$, and $\eta = 1$. We further conduct additional hyperparameter sensitivity experiments, as shown in Tables 7 and 8. From Table 7, the performance is relatively stable when η lies in the range $[0.7, 1]$, achieving the best accuracy at $\eta = 1$. From Table 8, the performance remains largely stable across a wide range of ϵ values. These results indicate that the proposed method is relatively robust to hyperparameter variations and does not require careful tuning to achieve strong performance.

Table 6. CNN: 100 clients, 5% participation

Method	CIFAR-10				CIFAR-100			
	Rounds (↓)		Acc. (%, ↑)		Rounds (↓)		Acc. (%, ↑)	
	71%	78%	500R	1000R	38%	42%	500R	1000R
FedAvg [34]	207	906	73.80	78.42	417	703	40.09	44.91
FedAvgM [13]	199	769	76.81	78.89	398	566	40.27	45.38
FedAdam [40]	180	947	73.30	78.20	402	624	40.04	45.21
FedAdagrad [40]	196	961	73.14	78.01	308	607	40.81	45.13
FedYogi [40]	165	875	75.99	79.25	**285**	424	42.59	45.64
FAFED [51]	330	965	72.78	78.24	539	793	37.69	44.30
FedCAda [61]	**121**	722	77.32	78.93	387	1000+	38.83	41.18
FADAS [49]	140	669	77.23	79.09	352	541	40.55	45.32
FedAdamom (ours)	153	**543**	**77.81**	**80.22**	303	**418**	**42.92**	**46.47**

Table 7. Performance of different η with $\epsilon = 10^{-3}$ and $\beta_2 = 0.05$.

η	3	2	1	0.9	0.7	0.5	0.1
Acc.	45.9	88.73	88.93	87.94	87.33	84.89	69.55

Table 8. Performance of different ϵ with $\eta = 1$ and $\beta_2 = 0.05$.

ϵ	0.5	0.1	0.01	10^{-3}	10^{-4}	10^{-5}	10^{-8}
Acc.	88.28	88.33	88.80	88.93	88.91	88.64	88.58

A.8. Visualizing 3D Loss Landscapes

We use the visualization technique proposed by [24] to analyze the loss landscape, and we adapt their code to support our specific datasets and network architectures. The visualization is performed by evaluating the loss along random directions in the parameter space, which is achieved by perturbing the model parameters within a predefined range. In our setup, the perturbations are restricted to the range $[1, 1]$ for both the x and y directions. To ensure consistent comparison across different models, we use the same set of random directions for all visualizations.

A.9. The Mean Escaple Time Analysis

The mean escape time is the expected time for particle gonverned by Eq. (2) to escape from Sharp Valley a to Flat Valley d. We follow the experiment setup of [54]. Specifically, we generate 50000 Gaussian samples as the training data set, where $z \sim \mathcal{N}(0, 4I)$. The model is two-layer fully-connected networks with one hidden layer and 10 neurons per hidden layer. The batch size is set 10. No weight decay. The test function is Styblinski-Tang Function:

$$h(x) = \frac{1}{2}\sum_{i=1}^{d}(x_i^4 - 16x_i^2 + 5x).$$

The one-dimensional Styblinski-Tang Function has one global minimum located at $a = -2.903534$ and one saddle point $b = 0.155731$ as the boundary. For a n-dimension Styblinski-Tang Function, we initialize parameter $x_0 = \frac{1}{\sqrt{k}}(-2.903534, ..., -2.903534)$, and set the valley's boundary as $x_i < \frac{1}{\sqrt{k}}0.156731$, where i is the dimension index. We record the number of iterations required to escape from the valley to the outside of valley. We repeat experiments 100 times to estimate the escape rate Γ and the mean escape time τ. As the escape time is approximately a random variable obeying an exponential distribution, $t \sim Exp(\Gamma)$, estimated escape rate can be written as

$$\Gamma = \frac{100 - 2}{\sum_{j=1}^{100} t_j}.$$

The 95% confidence interval of this estimator is

$$\Gamma(1 - \frac{1.96}{\sqrt{100}}) \leq \Gamma \leq \Gamma(1 + \frac{1.96}{\sqrt{100}}).$$

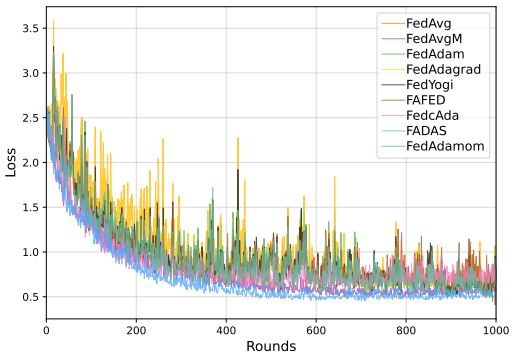
A.10. Variance measurements of top accuracy with different seeds

Table 9. The variance measurement of top validation accuracy that can be achieved, with 4 random seeds, 100 clients, 5% participation.

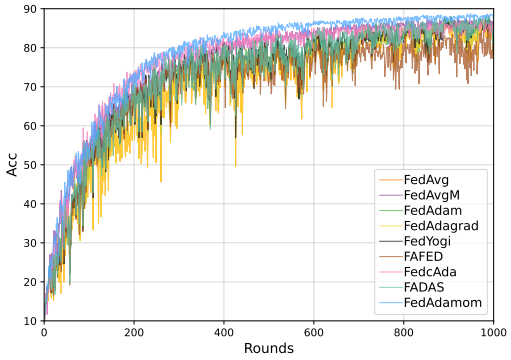
Dataset	FedAvg	FedAvgM	FedAdam	FedAdagrad	FedYogi
CIFAR-10	87.20 ± 0.33	87.54 ± 0.26	87.31 ± 0.36	87.03 ± 0.35	87.91 ± 0.40
CIFAR-100	53.39 ± 0.38	54.10 ± 0.22	53.67 ± 0.34	53.13 ± 0.41	54.05 ± 0.37
Dataset	FAFED	FedCAda	FADAS	FedAdamom	
CIFAR-10	86.11 ± 0.44	86.90 ± 0.58	88.14 ± 0.41	**88.93** ± 0.47	
CIFAR-100	52.65 ± 0.47	47.51 ± 0.62	54.67 ± 0.39	**57.58** ± 0.41	

A.11. Convergence and Top Validation ACCURACY

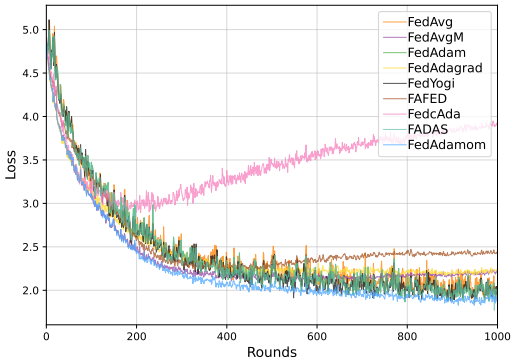
Fig. 4 shows the accuracy and loss curves of all algorithms. The exnerimental results demonstrate that our algorithm consistently outnerforms the maiority of existing adapt FL optimization algorithms.



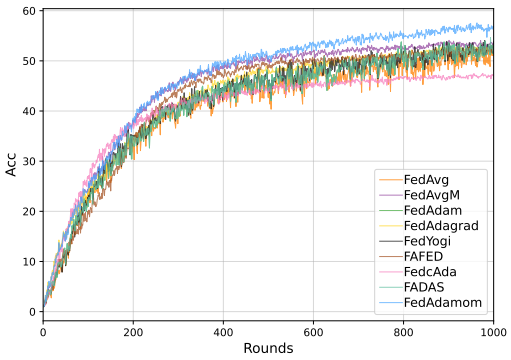
(a) Training loss on CIFAR-10.



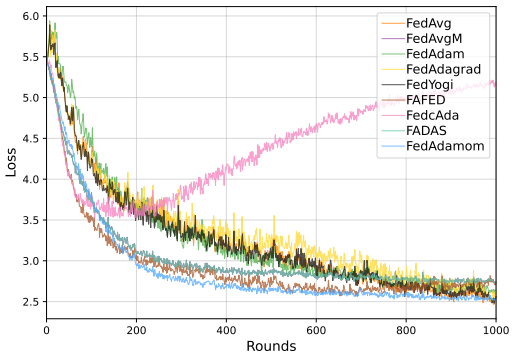
(b) Accuracy on CIFAR-10.



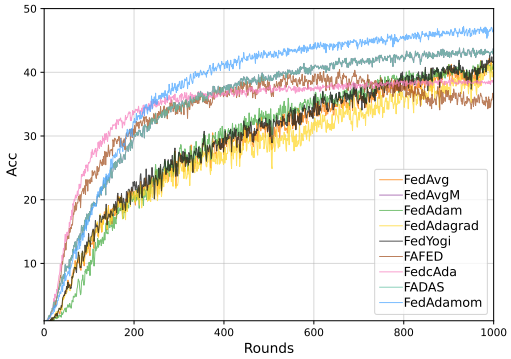
(c) Training loss on CIFAR-100.



(d) Accuracy on CIFAR-100.



(e) Training loss on TinyImagenet.



(f) Accuracy on TinyImagenet.

Figure 4. Training loss and accuracy versus communication rounds on different datasets. The setting is Dirichlet (0.3), 100 clients, 5% participation. Each method updates for 1000 communication rounds. The FedAdamom achieves the best and stable performance in the training.

B. Diffusion Dynamics

B.1. Saddle point escaping

Without of generality, we set the number of the local update $K = 1$ to make it convenient to analyze the dynamic of different methods. The global model update formula for FedAvg as follows:

$$x_{t+1} = x_t + \eta \Delta_t, \tag{12}$$

where $\Delta_t = \frac{1}{n}\sum_{i=1}^{n}(x_{t,K}^i - x_t)$. Since $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, we have $\mathbb{E}[\Delta_t] = -\eta_l \nabla f(x_t)$. Then we write the Langevin Equation as

$$dx = -\nabla f(x)dt + [\eta C(x)]^{\frac{1}{2}} dW_t, \tag{13}$$

where $dW_t \sim \mathcal{N}(0, Idt)$, I is the identity matrix, and $C(x)$ is the gradient noise convariance matrix. In practice, the learning rate η incorporates the local learning rate η_l. For notational simplicity, we denote it as η'.The Fokker-Planck Equation is

$$\frac{\partial P(x,t)}{\partial t} = \nabla \cdot [P(x,t)\nabla f(x)] + \nabla \cdot \nabla D(x) P(x,t), \tag{14}$$

where $D(x) = \frac{\eta' C(x)}{2}$ is the diffusion matrix [53]. Near a critical point c, we have

$$\begin{aligned} D(x) &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{\eta'}{2B}\left[\frac{1}{N}\sum_{j=1}^{N} \nabla f_i(x;\xi_i^j)\nabla f_i(x;\xi_i^j)^\top\right] \\ &\approx \frac{\eta'}{2B}[\frac{1}{n}\sum_{i=1}^{n} H_i(x)]^+ = \frac{\eta'}{2B}[H(x)]^+, \end{aligned}$$

where N is the number of training samples, $H(x)$ is the Hessian matrix of the global loss function at x. Given $H = Udiag(H_1, ..., H_{n-1}, H_n)U^\top$, we use $[\cdot]^+$ to denote the transformation that $[H]^+ = Udiag(|H_1|, ..., |H_{n-1}|, |H_n|)U^\top$. The i-th column vector of U is the eigenvector corresponding to H_i. Now we present the proof of the saddle points escape time.

For FedAvg, we can validate the probability density function

$$P(x,t) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma(t))}} exp(-\frac{1}{2}(x-c)^\top \Sigma(t)(x-c)) \tag{15}$$

is the solution of the Fokker-Planck Equation (14). Without losing generality, we only validate one-dimensional solution. The left-hand side of the Equation (14) can be written as

$$\begin{aligned} \frac{\partial P(x,t)}{\partial t} &= -\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^2}}\frac{1}{\theta^2} exp(-\frac{x^2}{2\theta^2})\frac{\partial \theta^2}{\partial t} \\ &\quad + \frac{1}{\sqrt{2\pi\theta^2}} exp(-\frac{x^2}{2\theta^2})\frac{x^2}{2\theta^4}\frac{\partial \theta^2}{\partial t} \\ &= \frac{1}{2}\left(\frac{x^2}{\theta^4} - \frac{1}{\theta^2}\right) P(x,t)\frac{\partial \theta^2}{\partial t}. \end{aligned} \tag{16}$$

The right-hand of the Equation (14) can be written as

$$\begin{aligned} \nabla \cdot [P(x,t)\nabla f(x)] &= P(x)H + Hx\frac{1}{\sqrt{2\pi\theta^2}} exp(-\frac{x^2}{2\theta^2})(-\frac{x}{\theta^2}) \\ &= H(1 - \frac{x^2}{\theta^2})P(x,t), \end{aligned} \tag{17}$$

and

$$\begin{aligned} D\nabla^2 P(x,t) &= -D\frac{\theta^2 - x^2}{\theta^5\sqrt{2\pi}} exp(-\frac{x^2}{2\theta^2}) \\ &= D(\frac{x^2}{\theta^4} - \frac{1}{\theta^2})P(x,t). \end{aligned} \tag{18}$$

Thus, we have

$$\begin{aligned} \frac{1}{2}(x^2 - \theta^2)\frac{\partial \theta^2}{\partial t} &= H\theta^2(\theta^2 - x^2) + D(x^2 - \theta^2) \\ \frac{\partial \theta^2}{\partial t} &= 2D - 2H\theta^2. \end{aligned} \tag{19}$$

The initial condition of θ^2 is given by $\theta^2(0) = 0$. We can validate that θ^2 satisfies

$$\theta_i^2(t) = \frac{D_i}{H_i}[1 - exp(-2H_i t]. \tag{20}$$

It is true along all eigenvectors' directions. By $D = \frac{\eta}{2B}H$, we can get the results of FedAvg diffusion:

$$\begin{aligned}\theta_i^2(T) &= sign(H_i)\frac{\eta'}{2B}[1 - exp(-2H_i\eta' T] \\ &= \frac{|H_i|\eta^2\eta_l^2 T}{B} + \mathcal{O}(B^{-1}H_i^2\eta^3\eta_l^3 T^2),\end{aligned} \tag{21}$$

where the second equality follows by $|H_i|\eta T \ll 1$ near the saddle points.

For FedAvgM, according to Eq. (6), we can write the deformed motion equation as follows:

$$\begin{cases} m_t = (1 - \gamma\delta t)m_{t-1} + \frac{F}{M}\delta t, \\ x_{t+1} = x_t + m_t\delta t, \end{cases} \tag{22}$$

where $F = -\Delta_t$, $\delta t = \eta$, $1 - \gamma\delta t = \beta$, and $\frac{\delta t}{M} = \eta_l$. Thus, we obtain the differential form of the motion equation as

$$Md\dot{x} = -\gamma M d\dot{x} - \frac{\partial f(x)}{\partial x}\delta t + [2D]^{\frac{1}{2}}dW_t. \tag{23}$$

Its Fokker-Planck Equation in the phase space (the $x - \dot{x}$ space) is well known as

$$\frac{\partial P(x,r,t)}{\partial t} - \nabla_x \cdot [rP(x,r,t)] + \nabla_r \cdot [\gamma r + M^{-1}\nabla_x f(x)] + \nabla_r \cdot M^{-2} D \cdot \nabla_r P(x,r,t), \tag{24}$$

where $r = \dot{x}$. As we disucssed in FedAvg dynamic, the position-space distribution in equilibirum is time-dependent near saddle points. Following the the form of the solution to the position-space Fokker-Planck Equation Eq. (15), the ansatz solution of $P(x,t)$ is given by

$$\begin{cases} P(x,t) = \frac{1}{(\sqrt{2\pi})^d \det(\Sigma(t))}\exp(-\frac{1}{2}(x - c(t))^\top \Sigma(t)(x - c(t))), \\ \Sigma(t) = U diag(\theta_1^2(t), ..., \theta_d^2(t))U^\top, \end{cases}$$

The time-dependent components in $P(x,t)$ are caused by momentum drift effect and the diffusion effect. The former decides $c(t)$, the center position of the probability density, the latter decides the covariance of the probability density, $\Sigma(t)$.

We can write the dynamics of the momentum drift effect as

$$M\ddot{c}(t) = -\gamma M\dot{c}(t). \tag{25}$$

Following [54], we ignore the conservative force near saddle points and focus on the behaviors near ill-conditioned saddle points, where Hessian eigenvalues are small along the escape directions. Then we obtain the solution $c(t)$ as

$$c_i(t) = c_i + \frac{r_{i,eq}}{\gamma}[1 - \exp(-\gamma t)], \tag{26}$$

where $\dot{c}(0) = r_{eq}$ and $\ddot{c}(0) = -\gamma r_{eq}$ are initial conditions, $r_{eq,i}^2 = \frac{D_i}{\gamma M^2}$.

The result of diffusion effect is equivalent to FedAvg with $\hat{\eta} = \frac{\eta'}{\gamma M}$. The expression of $P(x,t)$ and $\theta_i^2(t)$ is

$$\theta^2(t) = \frac{D_i}{\gamma M H_i}[1 - \exp(-\frac{2H_i t}{\gamma M})]. \tag{27}$$

We combine the momentum drift effect and the diffusion effect together, and then obtain the mean squared displacement of x as

$$\begin{aligned}\langle \Delta x_i^2(t)\rangle &= (c_i(t) - c_i)^2 + \theta_i^2(t) \\ &= \frac{D_i}{\gamma^3 M^2}[1 - \exp(-\gamma t)]^2 + \frac{D_i}{\gamma M H_i}[1 - \exp(-\frac{2H_i t}{\gamma M})].\end{aligned} \tag{28}$$

Then we apply the second order Taylor expansion in case of small $-\frac{2H_i t}{\gamma M}$. Then we obtain

$$\langle \Delta x_i^2 \rangle = \frac{|H_i|\eta^2\eta_l^2}{2(1-\beta)^3 B}[1-\exp(-(1-\beta)T)]^2 + \frac{|H_i|\eta^2\eta_l^2 T}{(1-\beta)^2 B} + \mathcal{O}(B^{-1}H_i^2\eta^3\eta_l^3 T^2). \tag{29}$$

We can directly obtained the conclusion for FedAdam with $\hat{\eta} = \eta C^{-\frac{1}{2}}$. The mean squared displacement is written as

$$\langle \Delta x_i^2(t) = \frac{D_i}{\gamma^2 M}[1-\exp(-\gamma t)]^2 + \frac{D_i}{\gamma M H_i}[1-\exp(\frac{2H_i t}{\gamma M})].$$

Then we can get

$$\langle \Delta x_i^2 \rangle = \frac{\eta^2\eta_l}{2(1-\beta_1)}[1-\exp(-(1-\beta_1)T)]^2 + \eta^2\eta_l^2 T + \mathcal{O}(\sqrt{B|H_i|}\eta^3\eta_l^3 T^2). \tag{30}$$

Similarly, for FedAdamom, as $M = \frac{\eta}{(I-\beta_1)\eta_l}$, $\gamma = \frac{I-\beta_1}{\eta}$, and $I-\beta_1 = \frac{v}{\bar{v}}$, we can derive

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

B.2. Proof of Theorem 1

Proof. We decompose the proof into two steps : 1) compute the probalility of locating in valley a, $P(x \in V_a)$, 2) compute the probability flux $j = \int_{S_a} J \cdot dS$. We first analyze the one-dimensional case. Step1: Under Assumption 3, we may only consider the second order Taylor approximation of the density function around critical points.

$$\begin{aligned}
P(x \in V_a) &= \int_{x\in V_a} P(x)dV \\
&= \int_{x\in V_a} P(a)\exp\left[-\frac{f(x)-f(a)}{T_a}\right]dV \\
&= P(a)\int_{x\in V_a} \exp\left[-\frac{\frac{1}{2}(x-a)^{a(x-a)+\mathcal{O}(\Delta x^3)}}{T_a}\right]dx \\
&= P(a)\frac{(2\pi T_a)^{\frac{1}{2}}}{H_a^{\frac{1}{2}}}.
\end{aligned}$$

Step 2: In case of SGD [53], we can obtain Smoluchowski Equation in position space:

$$J = D(x)\exp\left(-\frac{f(x)}{T}\right)\nabla\left[\exp\left(\frac{f(x)}{T}\right)P(x)\right],$$

where $T = D$. We assume the point ρ is the midpoint on the most possible path between a and b, where $f(s) = (1-s)f(a) + sL(b)$. The temperature T_a dominates the path $a \to s$, while temperature T_b dominates the path $s \to b$. Rearrange the above inequality we have

$$\nabla\left[\exp\left(\frac{f(x)-f(s)}{T}\right)P(x)\right] = JD^{-1}\exp\left(\frac{f(x)-f(s)}{T}\right).$$

Under Assumption 2, we integrate the equation from Valley a to the outside of Valley a along the most possible escape path

$$\begin{aligned}
Left &= \int_a^c \frac{\partial}{\partial x}\left[\exp\left(\frac{f(x)-f(s)}{T}\right)P(x)\right]dx \\
&= \int_a^s \frac{\partial}{\partial x}\left[\exp\left(\frac{f(x)-f(s)}{T}\right)P(x)\right]dx \\
&\quad + \int_s^c \frac{\partial}{\partial x}\left[\exp\left(\frac{f(x)-f(s)}{T}\right)P(x)\right]dx \\
&= [P(s) - \exp\left(\frac{f(a)-f(s)}{T_a}\right)P(a)] + [0 - P(s)] \\
&= -\exp\left(\frac{f(a)-f(s)}{T_a}\right)P(a) \\
Right &= -J\int_a^c D^{-1}\exp\left(\frac{f(x)-f(s)}{T}\right)dx.
\end{aligned}$$

Since J is fixed on an escape path from sharp minimum **a** to flat minimum **d** through saddle point **b**, we obtain

$$J = \frac{\exp\left(\frac{f(a)-f(s)}{T_a}\right)P(a)}{\int_a^c D^{-1}\exp\left(\frac{f(x)-f(s)}{T}\right)dx}.$$

Under Assumption 3, we have

$$\begin{aligned}
&\int_a^c D^{-1}\exp\left(\frac{f(x)-f(s)}{T}\right)dx \\
&\approx \int_a^c D^{-1}\exp\left[\frac{f(b)-f(s)+\frac{1}{2}(x-b)^{\mathrm{b}}(x-b)}{T_b}\right]dx \\
&\approx D_b^{-1}\int_{-\infty}^{+\infty}\exp\left[\frac{f(b)-f(s)+\frac{1}{2}(x-b)^{\mathrm{b}}(x-b)}{T_b}\right]dx \\
&= D_b^{-1}\exp\left(\frac{f(b)-f(s)}{T_b}\right)\sqrt{\frac{2\pi T_b}{|H_b|}}.
\end{aligned}$$

The last equality follows from the properties of the Gaussian distribution. Based on the results of the above two steps, we have

$$\begin{aligned}
\tau &= \frac{P(x\in V_a)}{\int_{S_a} J\cdot dS} \\
&= P(a)\sqrt{\frac{2\pi T_a}{H_a}}\frac{D_b^{-1}\exp\left(\frac{f(b)-f(s)}{T_b}\right)\sqrt{\frac{2\pi T_b}{|H_b|}}}{P(a)\exp\left(\frac{f(a)-f(s)}{T_a}\right)} \\
&= 2\pi\frac{1}{|H_b|}\exp\left[\frac{2B\Delta L}{\eta}\left(\frac{s}{H_a}+\frac{1-s}{|H_b|}\right)\right].
\end{aligned}$$

We then generalize the proof of one-dimensional diffusion to high-dimensional diffusion. Firstly, we can deduce that $P(x\in V_a) = P(a)\frac{(2\pi)^{\frac{n}{2}}}{\det(D_a{}^{-1}H_a)^{\frac{1}{2}}}$. Based on the formula of the one-dimensional probability current and flux, we obtain

$$\begin{aligned}
\int_{S_b} J\cdot dS &= J_b\int_{S_b}\exp\left[-\frac{\frac{1}{2}(x-b)^\top H_b^+(x-b)}{T}\right]dS \\
&= J_b\frac{(2\pi T)^{\frac{n-1}{2}}}{(\Pi_{i=1}^{n-1}H_{bi})^{\frac{1}{2}}}.
\end{aligned}$$

So we have

$$\begin{aligned}\tau &= \frac{P(x \in V_a)}{\int_{S_b} J \cdot dS} \\ &= J_{1d} \int_{S_b} \exp\left[-\frac{1}{2}(x-b)^\top [D_b^{-\frac{1}{2}} H_b D_b^{-\frac{1}{2}}]^{\perp e}(x-b)\right] dS \\ &= J_{1d} \frac{(2\pi)^{\frac{n-1}{2}}}{(\Pi_{i\neq e}(D_{bi}^{-1}H_{bi}))^{\frac{1}{2}}},\end{aligned}$$

where $[\cdot]^{\perp e}$ indicates the directions perpendicular to the escape direction e. So we have

$$\tau = 2\pi \sqrt{\frac{-\det(H_b D_b^{-1})}{\det(H_a D_a^{-1})}} \frac{1}{|H_{be}|} \exp\left(\frac{s\Delta L}{T_a} + \frac{(1-s)\Delta L}{T_b}\right).$$

T_a and T_b are the eigenvalues of $H_a^{-1}D_a$ and $H_b^{-1}D_b$ corresponding to the escape direction. We know $D_a = \frac{\eta'}{2B}H_a$ and $D_b = \frac{\eta'}{2B}[H_b]^+$. Thus, we have

$$\tau = 2\pi \frac{1}{|H_{be}|} \exp\left[\frac{2B\Delta L}{\eta'}\left(\frac{s}{H_{ae}} + \frac{1-s}{|H_{be}|}\right)\right].$$

□

B.3. Proof of Theorem 2

Proof. The proof closely similar to the proof of Theorem 1. According to Gauss Divergence Theorem, we rewrite the Fokker-Planck Equation Eq. (24) as

$$\begin{aligned}\frac{\partial P(x,r,t)}{\partial t} &= -r \cdot \nabla_x P(x,r,t) + \nabla_x f(x) \cdot M^{-1} \nabla_r P(x,r,t) + \nabla_r \cdot M^{-2} D(x) \cdot P_{eq}(r) \nabla_r [P_{eq}(r)^{-1} P(x,r,t)] \\ &= -\nabla \cdot J(x,t).\end{aligned}$$

According to [16], in case of finite inertia, we can transform the phase-space equation into the position-space Smoluchowski-like form with the effective diffusion correction:

$$J = \hat{D}(x) \exp(\frac{-f(x)}{T}) \nabla \left[\exp(\frac{f(x)}{T}) P(x)\right], \tag{31}$$

where $T = \frac{D}{\gamma M}$, and

$$\hat{D}_i(x) = D_i(x) \left(1 - \sqrt{1 - \frac{4H_i(x)}{\gamma^2 M}}\right) \left(\frac{2H_i(x)}{\gamma^2 M}\right)^{-1}. \tag{32}$$

Since the τ is computed in the same way as in Theorem 1, we can obtain

$$\tau = \pi \left[\sqrt{1 + \frac{4|H_{be}|}{\gamma^2 M}} + 1\right] \frac{1}{|H_{be}|} \exp\left[\frac{2\gamma M B \Delta f}{\eta'}\left(\frac{s}{H_{ae}} + \frac{(1-s)}{|H_{be}|}\right)\right] \tag{33}$$

So we have

$$\log(\tau) = \mathcal{O}\left(\frac{2(1-\beta)B\Delta f}{\eta \eta_l H_{ae}}\right) \tag{34}$$

□

B.4. Proof of Theorem 3

Proof. The proof closely related to the proof of Theorem 1. We only need replace the standard learning rate by the adaptive learning rate $\hat{\eta} = \eta C^{-\frac{1}{2}}$, and set $\gamma M = \frac{1}{\eta_l}$. Then we can obtain

$$\begin{aligned}
\tau &= \frac{P(x \in V_a)}{\int_{S_b} J \cdot dS} \\
&= \left[\frac{|\det(D_b^{-1} V_b^{\frac{1}{2}} H_b)|}{\det(D_a^{-1} V_a^{\frac{1}{2}} H_a)}\right]^{\frac{1}{2}} \pi \left[\sqrt{1 + \frac{4|H_{be}|}{\gamma^2 M}} + 1\right] \frac{1}{|H_{be}|} \exp\left[\frac{2\gamma M B \Delta f}{\eta}\left(\frac{s}{V_{ae}^{-\frac{1}{2}} H_{ae}} + \frac{1-s}{V_{be}^{-\frac{1}{2}} |H_{be}|}\right)\right] \\
&= \pi \left[\sqrt{1 + \frac{4\eta\eta_l \sqrt{B|H_{be}|}}{1-\beta_1}} + 1\right] \frac{|\det(H_a^{-1} H_b)|^{\frac{1}{4}}}{|H_{be}|} \exp\left[\frac{2\sqrt{B}\Delta f}{\eta\eta_l}\left(\frac{s}{\sqrt{H_{ae}}} + \frac{1-s}{\sqrt{|H_{be}|}}\right)\right].
\end{aligned} \tag{35}$$

So we have

$$\log(\tau) = \mathcal{O}\left(\frac{2\sqrt{B}\Delta f}{\eta\eta_l \sqrt{H_{ae}}}\right). \tag{36}$$

□

B.5. Proof of Theorem 4

Proof. The proof closely related to the proof of Theorem 1. We only need introduce the mass matrix $\hat{M}$ and the dampening matrix $\hat{\gamma}$. We have $\hat{M} = \frac{\eta}{\eta_l(I-\beta_1)}$, $\hat{\gamma} = \frac{I-\beta_1}{\eta}$, $I - \beta_1 = \frac{v}{\bar{v}}$ and $\hat{\gamma}\hat{M} = \frac{1}{\eta_l} I$. Then we can obtain

$$\begin{aligned}
\tau &= \frac{P(x \in V_a)}{\int_{S_b} J \cdot dS} \\
&= \pi \left[\sqrt{1 + \frac{4|H_{be}|}{\hat{\gamma}^2 \hat{M}}} + 1\right] \frac{1}{|H_{be}|} \exp\left[\frac{2\hat{\gamma}\hat{M} B \Delta f}{\eta\eta_l}\left(\frac{s}{H_{ae}} + \frac{1-s}{|H_{be}|}\right)\right] \\
&= \pi \left[\sqrt{1 + \frac{4\eta\eta_l \sum_{i=1}^{d} |H_{bi}|}{n}} + 1\right] \frac{1}{|H_{be}|} \exp\left[\frac{2B\Delta f}{\eta\eta_l}\left(\frac{s}{H_{ae}} + \frac{1-s}{|H_{be}|}\right)\right].
\end{aligned} \tag{37}$$

So we have

$$\log(\tau) = \mathcal{O}\left(\frac{2B\Delta f}{\eta\eta_l H_{ae}}\right). \tag{38}$$

□

C. Convergence Analysis

Lemma 1. *For any $t \geq 0$, we have*

$$\begin{aligned}
x_{t+1} - x_t &= -\eta \sum_{k=0}^{t} q_{k,t} \Delta_k \\
1 - \beta_{1,max}^{t+1} &\leq \sum_{k=0}^{t} q_{k,t} \leq 1.
\end{aligned}$$

where $q_{k,t} = (1 - \beta_{1,k}) \prod_{i=k+1}^{t} \beta_{1,i}$.

Proof. Recall that

$$x_{t+1} = x_t + \eta(1-\beta_{1,t})\Delta_t + \beta_{1,t}(x_t - x_{t-1})$$
$$x_{t+1} - x_t = \beta_{1,t}(x_t - x_{t-1}) + \eta(1-\beta_{1,t})\Delta_t.$$

So we can get

$$x_{t+1} - x_t = \eta \sum_{k=0}^{t} (1-\beta_{1,k})\Delta_t \prod_{i=k+1}^{t} \beta_{1,i}.$$

Let $q_{k,t} = (1-\beta_{1,k})\prod_{i=k+1}^{t}\beta_{1,i}$, then we calculated the derivatives with respect to $\beta_{1,k}$ for any $0 \le k \le t$:

$$\frac{\partial \sum_{k=0}^{t} q_{k,t}}{\partial \beta_{1,0}} = -\prod_{i=k+1}^{t} \beta_{1,i} \le 0.$$

Since $0 \le \beta_{1,k} \le \beta_{1,max}$, we have

$$\sum_{k=0}^{t} q_{k,t}|_{\beta_{1,0}=\beta_{1,max}} \le \sum_{k=0}^{t} q_{k,t} \le \sum_{k=0}^{t} q_{k,t}|_{\beta_{1,0}=0}.$$

Recursively, we can calculate the derivatives with respect to $\beta_{1,k}, k > 1$. Then we obtain

$$1 - \beta_{1,max}^{t+1} \le \sum_{k=0}^{t} q_{k,t} \le 1.$$

. □

Lemma 2. *client drift satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|x_{t,k}^i - x_t\|^2 \le 40K^2\eta_l^2\mathbb{E}\|\nabla f(x_t)\|^2 + 5K\eta_l^2\sigma_l^2 + 40K^2\eta_l^2\sigma_g^2.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_t\|^2 \\
&= \mathbb{E}\|x_{t,k-1}^i - x_t - \eta_l g_{t,k-1}^i\|^2 \\
&= \mathbb{E}\|x_{t,k-1}^i - x_t - \eta_l(g_{t,k-1}^i - \nabla F_i(x_{t,k-1}^i) + \nabla F_i(x_{t,k-1}^i) - \nabla F_i(x_t) + \nabla F_i(x_t))\|^2 \\
&\le \mathbb{E}\|x_{t,k-1}^i - x_t - \eta_l(\nabla F_i(x_{t,k-1}^i) - \nabla F_i(x_t) + \nabla F_i(x_t))\|^2 + \eta_l^2\sigma_l^2 \\
&\le (1 + \frac{1}{2K-1})\mathbb{E}\|x_{t,k-1}^i - x_t\|^2 + 4K\mathbb{E}\|\eta_l(\nabla F_i(x_{t,k-1}^i) - \nabla F_i(x_t))\|^2 + 4K\mathbb{E}\|\eta_l\nabla F_i(x_t)\|^2 + \eta_l^2\sigma_l^2 \\
&\le (1 + \frac{1}{2K-1} + 4K\eta_l^2L^2)\mathbb{E}\|x_{t,k-1}^i - x_t\|^2 + 8K\eta_l^2\mathbb{E}\|\nabla f(x_t)\|^2 + \eta_l^2\sigma_l^2 + 8K\eta_l^2\sigma_g^2 \\
&\le (1 + \frac{1}{K-1})\mathbb{E}\|x_{t,k-1}^i - x_t\|^2 + 8K\eta_l^2\mathbb{E}\|\nabla f(x_t)\|^2 + \eta_l^2\sigma_l^2 + 8K\eta_l^2\sigma_g^2,
\end{aligned}$$

where the second inequality is due to that $\|x_1+x_2\|^2 \le (1+a)\|x_1\|^2 + (1+\frac{1}{a})\|x_2\|^2$ with $a = \frac{1}{2K-1}$ and the last inequality is because of $\eta_l \le \frac{1}{6LK}$. Unrolling it we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|x_{t,k}^i - x_t\|^2 \le 40K^2\eta_l^2\mathbb{E}\|\nabla f(x_t)\|^2 + 5K\eta_l^2\sigma_l^2 + 40K^2\eta_l^2\sigma_g^2.$$

□

Lemma 3. *The global update satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|g_{t,k}^{i}\|^2 \leq (80L^2K^3\eta_l^2+4K)\mathbb{E}\|\nabla f(x_t)\|^2+(10L^2K^2\eta_l^2+K)\sigma_l^2+(80L^2K^3\eta^2+4K)\sigma_g^2.$$

Proof.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|g_{t,k}^{i}\|^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}(\mathbb{E}\|\nabla F_i(x_{t,k}^{i})\|^2+\sigma_l^2)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1}(\mathbb{E}\|\nabla F_i(x_{t,k}^{i})-\nabla F_i(x_t)+\nabla F_i(x_t)\|^2+\sigma_l^2)\\
&\leq \frac{2L^2}{m}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\mathbb{E}\|x_{t,k}^{i}-x_t\|^2+4K\mathbb{E}\|\nabla f(x_t)\|^2+K\sigma_l^2+4K\sigma_g^2\\
&\leq (80L^2K^3\eta_l^2+4K)\mathbb{E}\|\nabla f(x_t)\|^2+(10L^2K^2\eta_l^2+K)\sigma_l^2+(80L^2K^3\eta^2+4K)\sigma_g^2.
\end{aligned}$$

□

Theorem 6. *Suppose that local functions $\{F_i\}_{i=1}^m$ are non-convex. Let the local learning rate satisfy $\eta_l=\mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$, the global satisfy $\eta=\mathcal{O}\left(\sqrt{sK}\right)$. Then, the convergence rate for FedAdamom satisfies:*

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

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

Proof. According to the L-smooth of f, we have

$$\begin{aligned}
&f(x_{t+1})-f(x_t)\\
&\leq\langle\nabla f(x_t),x_{t+1}-x_t\rangle+\frac{L}{2}\|x_{t+1}-x_t\|^2\\
&=\langle\nabla f(x_t),\eta\sum_{k=0}^{t}q_{k,t}\Delta_k\rangle+\frac{L\eta^2}{2}\|\sum_{k=0}^{t}q_{k,t}\Delta_k\|^2\\
&=\eta\sum_{k=0}^{t}q_{k,t}\sum_{j=0}^{t-1}\langle\nabla f(t_{j+1})-\nabla f(t_j),\Delta_k\rangle+\eta\sum_{k=0}^{t}q_{k,t}\langle\nabla f(x_k),\Delta_k\rangle+\frac{L\eta^2}{2}\|\sum_{k=0}^{t}q_{k,t}\Delta_k\|^2\\
&\leq\frac{1}{2}\sum_{k=0}^{t}q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\|\nabla f(x_{j+1})-\nabla f(x_j)\|^2]+\frac{\eta^2}{2}\sum_{k=0}^{t}q_{k,t}\mathbb{E}\|\Delta_k\|^2+\eta\sum_{k=0}^{t}q_{k,t}\langle\nabla f(x_k),\Delta_k\rangle+\frac{L\eta^2}{2}\mathbb{E}\|\sum_{k=0}^{t}q_{k,t}\Delta_k\|^2.
\end{aligned}$$

For the first term, we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\|\nabla f(x_{j+1})-\nabla f(x_j)\|^2] \\
&\leq \frac{L^2}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\|x_{t+1}-x_t\|^2] \\
&= \frac{L^2}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\|-\eta\sum_{\tau=0}^{j} q_{\tau,j}\Delta_\tau\|^2] \\
&\leq \frac{L^2\eta^2}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\sum_{\tau=0}^{j} q_{\tau,j}\|\Delta_\tau\|^2] \\
&\leq \frac{L^2\eta^2\eta_l^2(n-s)}{2ms(n-1)}\sum_{k=0}^{t} q_{k,t}\mathbb{E}[\sum_{j=k}^{t-1}\sum_{\tau=0}^{j} q_{\tau,j}\sum_{i=1}^{n}\|\sum_{k=0}^{K-1} g_{t,k}^i\|^2] \\
&\leq \frac{L^2\eta^2\eta_l^2(n-s)}{2s(n-1)}\sum_{k=0}^{t} q_{k,t}\sum_{j=k}^{t-1}\sum_{\tau=0}^{j} q_{\tau,j}[(80L^2K^3\eta_l^2+4K)\mathbb{E}\|\nabla f(x_t)\|^2+(10L^2K^2\eta_l^2+K)\sigma_l^2 \\
&\quad +(80L^2K^3\eta_l^2+4K)\sigma_g^2] \\
&\leq \frac{L^2\eta^2\eta_l^2(n-s)}{2s(n-1)}[\frac{\beta_{1,max}}{1-\beta_{1,max}}((80L^2K^3\eta^2+4K)\sigma_g^2+(10L^2K^2\eta_l^2+K)\sigma_l^2) \\
&\quad +(80L^2K^3\eta_l^2+4K)\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\nabla f(x_k)\|^2].
\end{aligned}
$$

For the third term, we have

$$
\begin{aligned}
&\eta\sum_{k=0}^{t} q_{k,t}\mathbb{E}\langle\nabla f(x_k),\Delta_k\rangle \\
&= -\eta\eta_l K\sum_{k=0}^{t} q_{k,t}\langle\nabla f(x_k),\frac{1}{sK}\sum_{i\in\mathcal{S}_k}\sum_{\tau=0}^{K-1}(\nabla F_i(x_{k,\tau}^i)-\nabla F_i(x_k))+\nabla f(x_k)\rangle \\
&= -\eta\eta_l K\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\nabla f(x_k)\|^2+\eta\eta_l K\sum_{k=0}^{t} q_{k,t}\langle\nabla f(x_k),\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\sum_{\tau=0}^{K-1}(\nabla F_i(x_{k,\tau}^i)-\nabla f(x_k))\rangle \\
&\leq \frac{-\eta\eta_l K}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\nabla f(x_k)\|^2+\frac{L^2K\eta\eta_l}{2}\sum_{k=0}^{t} q_{k,t}\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\sum_{\tau=0}^{K-1}\mathbb{E}\|x_{k,\tau}^i-x_k\|^2 \\
&\leq \frac{-\eta\eta_l K}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\nabla f(x_k)\|^2+\frac{L^2K\eta\eta_l}{2}\sum_{k=0}^{t} q_{k,t}(40K^2\eta_l^2\mathbb{E}\|\nabla f(x_t)\|^2+5K\eta_l^2\sigma_l^2+40K^2\eta_l^2\sigma_g^2).
\end{aligned}
$$

For the second and last term, we have

$$
\begin{aligned}
&\frac{\eta^2}{2}\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\Delta_k\|^2+\frac{L\eta^2}{2}\mathbb{E}\|\sum_{k=0}^{t} q_{k,t}\Delta_k\|^2 \\
&\leq (\frac{\eta^2}{2}+\frac{L\eta^2}{2})\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\Delta_k\|^2 \\
&\leq \frac{\eta^2\eta_l^2(L+1)(n-s)}{2s(n-1)}\sum_{k=0}^{t} q_{k,t}[(80L^2K^3\eta_l^2+4K)\mathbb{E}\|\nabla f(x_t)\|^2+(10L^2K^2\eta_l^2+K)\sigma_l^2+(80L^2K^3\eta^2+4K)\sigma_g^2].
\end{aligned}
$$

So we have

$$
\begin{aligned}
&f(x_{t+1}) - f(x_t) \\
&\leq -\eta\eta_l K[\frac{1}{2} - 20L^2K^2\eta_l^2 - \frac{\eta\eta_l(L^2+L+1)(n-s)}{2s(n-1)}(80L^2K^2\eta_l^2+4)]\mathbb{E}\sum_{k=0}^{t} q_{k,t}\|\nabla f(x_k)\|^2 \\
&\quad + (\frac{L^2\eta^2\eta_l^2(n-s)}{2s(n-1)}\frac{\beta_{1,max}}{1-\beta_{1,max}} + \frac{\eta^2\eta_l^2(L+1)(n-s)}{2s(n-1)}\sum_{k=0}^{t} q_{k,t})[(80L^2K^3\eta_l^2+4K)\sigma_g^2 + (10L^2K^2\eta_l^2+K)\sigma_l^2] \\
&\quad + \frac{L^2K\eta\eta_l}{2}\sum_{k=0}^{t} q_{k,t}(5K\eta_l^2\sigma_l^2 + 40K^2\eta_l^2\sigma_g^2).
\end{aligned}
$$

If $\eta_l \leq \frac{1}{12LK}$, there exists a constant $c_1 > 0$ such that $\frac{1}{2} - 20L^2K^2\eta_l^2 - \frac{\eta\eta_l(L^2+L+1)(n-s)}{2s(n-1)}(90L^2K^2\eta_l^2+3) \leq c_1 < \frac{1}{2}$. Rearranging the above inequality and summing from $t = 0, ..., T-1$, we have

$$
\begin{aligned}
&c_1\sum_{t=0}^{T-1}\sum_{k=0}^{t} q_{k,t}\mathbb{E}\|\nabla f(x_k)\|^2 \\
&\leq \frac{\mathbb{E}[f(x_0) - f(x_T)]}{\eta\eta_l K} + \frac{L^2}{2}\sum_{t=0}^{T-1}\sum_{k=0}^{t} q_{k,t}(5K\eta_l^2\sigma_l^2 + 40K^2\eta_l^2\sigma_g^2) \\
&\quad + \sum_{t=0}^{T-1}(\frac{L^2\eta\eta_l(n-s)}{2s(n-1)}\frac{\beta_{1,max}}{1-\beta_{1,max}} + \frac{\eta\eta_l(L+1)(n-s)}{2s(n-1)}\sum_{k=0}^{t} q_{k,t})[(80L^2K^2\eta_l^2+4)\sigma_g^2 + (10L^2K\eta_l^2+1)\sigma_l^2].
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \\
&\leq \frac{\mathbb{E}[f(x_0) - f(x_T)]}{c_1\eta\eta_l KT} + \frac{L^2}{2}\frac{\beta_{1,max}}{1-\beta_{1,max}}(5K\eta_l^2\sigma_l^2 + 40K^2\eta_l^2\sigma_g^2) \\
&\quad + \frac{(L^2+L+1)\eta\eta_l(n-s)}{2s(n-1)}\frac{\beta_{1,max}}{1-\beta_{1,max}}[(80L^2K^2\eta_l^2+4)\sigma_g^2 + (10L^2K\eta_l^2+1)\sigma_l^2].
\end{aligned}
$$

By setting $\Psi = \frac{n-s}{n-1}\frac{1}{\sqrt{sKT}}\left[\left(\frac{80}{T}+4\right)\sigma_g^2 + \left(\frac{10}{KT}+1\right)\sigma_l^2\right]$ and $\Theta_0 = \mathbb{E}[f(x_0) - f(x_T)]$, we obtain the conclusion. □