Appendix - Fed-ADE: Adaptive Learning Rate for Federated Post-adaptation under Distribution Shift

Contents

A. Notation Table

Table 5. Notation Table

Notation	Description	Notation	Description
$\mathbf{x}_G^0$	Pre-training data at server	$\mathbf{y}_G^0$	Pre-training labels at server
$\mathbf{Q}_{\mathcal{G},\mathbf{x}}^0$	Feature distribution of $\mathbf{x}_G^0$	$\mathbf{Q}_{\mathcal{G},\mathbf{y}}^0$	Label distribution of $\mathbf{y}_G^0$
$\mathcal{I}$	Set of class indices	i	Index for class
t	Timestep	T	Total number of timesteps
$\mathcal{C}$	Set of FL clients	c	Client index
θ	Global model parameters	θ_c	Model parameters of client c
$\mathbf{Q}_c^t$	Overall data distribution at time t	$\omega(t)$	Weighting function controlling the distribution at time t
$\mathcal{H}(\cdot)$	Softmax prediction of the model	$\mathcal{L}(\cdot)$	Loss function
$\mathcal{F}_c^{t,i}(\theta_c)$	Class-wise risk for class i	$\widehat{\mathcal{F}}_c^{0,i}(\theta_c)$	Empirical risk
$\mathbf{M}$	Confusion matrix computed using pre-training data	$\mathbf{Q}_{c,\hat{\mathbf{y}}}^t$	Predicted distribution over $\mathbf{x}_c^t$
R	Total number of communication rounds	r	Communication round
$\mathcal{C}^{(r)}$	Selected client set in round r	N_c^t	Number of samples of client c at t
ψ_c	Shared layers of client c	ϕ_c	Personalized layers of client c
$\bar{\psi}$	Aggregated global shared layers	$\{\psi_c, \phi_c\}$	Model split: shared/personalized
$[\eta_{\min}, \eta_{\max}]$	Min/max learning rate bounds	$\mathcal{S}_c^t$	Distribution dynamics signal
$\mathcal{S}_{\text{unc},c}^t$	Uncertainty dynamic signal	$\mathcal{S}_{\text{rep},c}^t$	Representation dynamic signal
$\mathbf{q}_c^t$	Aggregated predictive distribution of current data batch	$\mathbf{z}_c^t$	Batch-level latent feature vector
$h_{\psi_c}(\cdot)$	Feature extractor	$K_{\cos}, K_h, K_\psi, K_\phi$	Lipschitz constant
B	Loss value upper bound	σ	Min. singular value of $\mathbf{M}$
Γ	Projection constant	$\lvert\mathcal{I}\rvert$	Number of classes
Reg_T	Dynamic regret over T rounds	$\Delta\mathcal{L}^{(0)}$	Initial loss gap for convergence

B. Proof of Unbiased Risk Estimation

Lemma 5 (Unbiased Risk Estimator)**.** *Given model parameters θ_c independent of the test-time data $(\mathbf{x}_c^t, \mathbf{y}_c^t)$, the estimator $\widehat{\mathcal{F}}_c^t$ in (4) satisfies*

$$\mathbb{E}_{\mathbf{y}_c^t \sim \mathbf{Q}_{c,\mathbf{y}}^t}[\widehat{\mathcal{F}}_c^t(\theta_c)] = \mathcal{F}_c^t(\theta_c),$$

provided that the confusion matrix $\mathbf{M}$, estimated at the server using sufficient pre-training data $(\mathbf{x}_G^0, \mathbf{y}_G^0)$ ensuring $\widehat{\mathbf{M}} = \mathbf{M}$, and that each client has sufficient local initial data $(\mathbf{x}_c^0, \mathbf{y}_c^0)$ such that $\widehat{\mathcal{F}}_c^0(\theta_c) = \mathcal{F}_c^0(\theta_c)$ for all $i \in \mathcal{I}$.

Proof. The expected risk decomposes into class-specific components:

$$\mathcal{F}_c^t(\theta_c) = \sum_{i \in \mathcal{I}} [\mathbf{Q}_{\mathbf{y}_c^t}]_i \cdot \mathcal{F}_c^{0,i}(\theta_c)$$

where $\mathcal{I}$ denotes the class index set. Through BBSE [40], we approximate $\mathbf{Q}_{\mathbf{y}_c^t} \approx \mathbf{M}^{-1}\mathbf{Q}_{\hat{\mathbf{y}}_c^t}$ using the confusion matrix $\mathbf{M}$ from pre-training and empirical predictions $\mathbf{Q}_{\hat{\mathbf{y}}_c^t}$.

Expanding the estimator's expectation:

$$\mathbb{E}[\widehat{\mathcal{F}}_c^t] = \sum_{i \in \mathcal{I}} [\widehat{\mathbf{M}}^{-1}\mathbb{E}[\mathbf{Q}_{\hat{\mathbf{y}}_c^t}]]_i \cdot \widehat{\mathcal{F}}_c^{0,i}(\theta_c)$$

Under condition $\mathbb{E}[\mathbf{Q}_{\hat{\mathbf{y}}_c^t}] = \mathbf{M}\mathbf{Q}_{\mathbf{y}_c^t}$, yielding:

$$\mathbb{E}[\widehat{\mathcal{F}}_c^t] = \sum_{i \in \mathcal{I}} [\mathbf{Q}_{\mathbf{y}_c^t}]_i \cdot \mathcal{F}_c^{0,i}(\theta_c) = \mathcal{F}_c^t(\theta_c)$$

The estimation error $|\mathbb{E}[\widehat{\mathcal{F}}_c^t] - \mathcal{F}_c^t|$ is bounded by $\mathcal{O}(1/\sqrt{n_0})$ via concentration inequalities, where a number of pre-training data $n_0 = |(\mathbf{x}_G^0, \mathbf{y}_G^0)|$. This becomes negligible when $n_0 \geq \Omega(|\mathcal{I}|^2/\epsilon^2)$ for error tolerance ϵ. Practical implementations may employ singular value thresholding for numerical stability when inverting $\mathbf{M}$. □

Complete derivation expanding the expectation operator:

$$\mathbb{E}_{\mathbf{x}_c^t}[\widehat{\mathcal{F}}_c^t] = \sum_{i \in \mathcal{I}} \mathbf{M}^{-1}\mathbb{E}[\mathbf{Q}_{\hat{\mathbf{y}}_c^t}] \cdot \mathcal{F}_c^{0,i}(\theta_c) = \sum_{i \in \mathcal{I}} [\mathbf{Q}_{\mathbf{y}_c^t}]_i \cdot \mathcal{F}_c^{0,i}(\theta_c)$$

confirming the unbiasedness through the relationship $\mathbf{Q}_{\mathbf{y}_c^t} = \mathbf{M}^{-1}\mathbb{E}[\mathbf{Q}_{\hat{\mathbf{y}}_c^t}]$ established via BBSE methodology.

Remark 6. The complexity function $\Omega(\cdot)$ in the sample complexity $n_0 \geq \Omega(|\mathcal{I}|^2/\epsilon^2)$ denotes an asymptotic lower bound, meaning the pre-training dataset size must grow at least quadratically with the number of classes to guarantee estimation accuracy. This aligns with information-theoretic limits for distribution estimation.

C. Proof of Theorem 1

We restate the theorem for completeness. Let $\mathbf{Q}^t_{c,\mathbf{y}}$ denote the true predictive distribution marginal at timestep t, and $\mathbf{q}^t_c$ the empirical mean softmax vector defined in (11). Under ϵ-calibration with respect to $\mathbf{Q}^t_{c,\mathbf{y}}$, we have $\|\mathbf{q}^t_c - \mathbf{Q}^t_{c,\mathbf{y}}\|_2 \leq \epsilon_t$ in expectation. The cumulative surrogate of predictive dynamics is

$$\bar{\mathcal{S}}_{\text{unc}} = \sum_{t=1}^{T} \mathcal{S}^t_{\text{unc}} = \sum_{t=1}^{T} \big(1 - \cos(\mathbf{q}^{t-1}_c, \mathbf{q}^t_c)\big).$$

Our goal is to show that $\bar{\mathcal{S}}_{\text{unc}}$ approximates the cumulative true deviation $\sum_{t=1}^{T} \|\mathbf{Q}^t_{c,\mathbf{y}} - \mathbf{Q}^{t-1}_{c,\mathbf{y}}\|_1$ up to an additive error of order $\mathcal{O}\big(\sum_t \epsilon_t\big)$.

Step 1: Lipschitz continuity of cosine distance. The cosine similarity function $\cos(\cdot,\cdot)$ is $K_{\cos}$-Lipschitz continuous in each of its arguments on the unit sphere, i.e.,

$$\big|\, \cos(\mathbf{a}, \mathbf{b}) - \cos(\mathbf{a}', \mathbf{b}') \,\big| \leq K_{\cos}\big(\|\mathbf{a} - \mathbf{a}'\|_2 + \|\mathbf{b} - \mathbf{b}'\|_2\big).$$

Since both $\mathbf{q}^t_c$ and $\mathbf{Q}^t_{c,\mathbf{y}}$ are probability vectors normalized to unit norm, this Lipschitz property holds directly.

Step 2: Bounding the surrogate deviation. By the definition of $\mathcal{S}^t_{\text{unc}}$, we have

$$\begin{aligned}&\big|\mathcal{S}^t_{\text{unc}} - \big(1 - \cos(\mathbf{Q}^{t-1}_{c,\mathbf{y}}, \mathbf{Q}^t_{c,\mathbf{y}})\big)\big| \\ &\leq K_{\cos}\big(\|\mathbf{q}^{t-1}_c - \mathbf{Q}^{t-1}_{c,\mathbf{y}}\|_2 + \|\mathbf{q}^t_c - \mathbf{Q}^t_{c,\mathbf{y}}\|_2\big) \leq K_{\cos}(\epsilon_{t-1} + \epsilon_t).\end{aligned}$$

Step 3: Summing over timesteps. Summing over $t = 2, \ldots, T$ yields

$$\Big|\sum_{t=1}^{T} \mathcal{S}^t_{\text{unc}} - \sum_{t=1}^{T} \big(1 - \cos(\mathbf{Q}^{t-1}_{c,\mathbf{y}}, \mathbf{Q}^t_{c,\mathbf{y}})\big)\Big| \leq K_{\cos} \sum_{t=2}^{T} (\epsilon_t + \epsilon_{t-1}).$$

Step 4: Relating cosine distance to total variation. For normalized probability vectors, the cosine distance upper bounds the ℓ_1 deviation:

$$1 - \cos(\mathbf{Q}^{t-1}_{c,\mathbf{y}}, \mathbf{Q}^t_{c,\mathbf{y}}) \leq \|\mathbf{Q}^{t-1}_{c,\mathbf{y}} - \mathbf{Q}^t_{c,\mathbf{y}}\|_1.$$

Combining this with the bound above yields

$$\Big|\bar{\mathcal{S}}_{\text{unc}} - \sum_{t=1}^{T} \|\mathbf{Q}^{t-1}_{c,\mathbf{y}} - \mathbf{Q}^t_{c,\mathbf{y}}\|_1\Big| \leq K_{\cos} \sum_{t=2}^{T} (\epsilon_t + \epsilon_{t-1}).$$

□

D. Proof of Theorem 2

Let $\mathbf{z}_c^t$ denote the empirical ℓ_2-normalized batch-mean feature vector defined in (13), and $\bar{\mathbf{z}}_c^t = \mathbb{E}[h_{\psi_c}(x)/\|h_{\psi_c}(x)\|_2]$ be its expectation under the local distribution at time t. We assume $\|\mathbf{z}_c^t - \bar{\mathbf{z}}_c^t\|_2 \le \epsilon_t'$ in expectation, and that the shared representation extractor h_{ψ_c} is K_h-Lipschitz.

The cumulative surrogate of representation dynamics is defined as

$$\bar{\mathcal{S}}_{\text{rep}} = \sum_{t=1}^{T} \mathcal{S}_{\text{rep}}^t = \tfrac{1}{2} \sum_{t=1}^{T} \big(1 - \cos(\mathbf{z}_c^{t-1}, \mathbf{z}_c^t)\big).$$

Since all feature vectors are ℓ_2-normalized, the cosine distance $\frac{1}{2}(1 - \cos(\mathbf{a}, \mathbf{b}))$ is smooth on the unit sphere, and its deviation with respect to small perturbations of $\mathbf{a}$ and $\mathbf{b}$ is bounded linearly by their ℓ_2 distance. Using the Lipschitz continuity of h_{ψ_c}, the deviation between the empirical mean $\mathbf{z}_c^t$ and its population counterpart $\bar{\mathbf{z}}_c^t$ thus induces a bounded perturbation in the cosine term.

For each timestep t, we have

$$\Big| \mathcal{S}_{\text{rep}}^t - \tfrac{1}{2}\big(1 - \cos(\bar{\mathbf{z}}_c^{t-1}, \bar{\mathbf{z}}_c^t)\big) \Big| \le K_h \big(\epsilon_{t-1}' + \epsilon_t'\big).$$

Summing over all $t = 2, \ldots, T$ gives

$$\Big| \sum_{t=1}^{T} \mathcal{S}_{\text{rep}}^t - \sum_{t=1}^{T} \tfrac{1}{2}\big(1 - \cos(\bar{\mathbf{z}}_c^{t-1}, \bar{\mathbf{z}}_c^t)\big) \Big| \le K_h \sum_{t=2}^{T} (\epsilon_t' + \epsilon_{t-1}').$$

Therefore, the cumulative surrogate $\bar{\mathcal{S}}_{\text{rep}}$ satisfies

$$\Big| \bar{\mathcal{S}}_{\text{rep}} - \sum_{t=1}^{T} \tfrac{1}{2} \left(1 - \frac{\langle \bar{\mathbf{z}}_c^{t-1}, \bar{\mathbf{z}}_c^t \rangle}{\|\bar{\mathbf{z}}_c^{t-1}\|_2 \, \|\bar{\mathbf{z}}_c^t\|_2} \right) \Big| \le K_h \sum_{t=2}^{T} (\epsilon_t' + \epsilon_{t-1}'),$$

which shows that $\bar{\mathcal{S}}_{\text{rep}}$ accurately approximates the temporal trajectory of the expected local feature representations, with error bounded by the feature-level estimation noise scaled by the Lipschitz constant K_h. □

E. Proof of Theorem 3

E.1. Regret Bound

We present the proof of Theorem 3 by decomposing the dynamic regret into two terms. Our approach introduces a reference sequence that evolves in a piecewise-constant manner across predefined intervals. The first term evaluates the algorithm's performance relative to this sequence, while the second term quantifies how well the reference sequence itself approximates the optimal comparator. For the following proof, we use $\{\psi_c^t, \phi_c^t\}$ for model parameters to differentiate the model per timestep.

Proof. The regret bound can be split into two terms by introducing a reference sequence $\{\psi_c^t, \phi_c^t\}$ that only changes every τ steps. Specifically, let $\mathcal{J}_m = [(m-1)\tau + 1, m\tau]$ denote the m-th interval. For each interval, the comparator $\{\psi_c^t, \phi_c^t\}$ is chosen as the best fixed decision in that interval, i.e., $\{\psi_c^{\mathcal{J}_m}, \phi_c^{\mathcal{J}_m}\} = \arg\min_{\{\psi_c^t, \phi_c^t\}} \sum_{t\in\mathcal{J}_m} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\})$ for $t \in \mathcal{J}_m$. Then,

$$\begin{aligned}
&\mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) - \sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^*, \phi_c^*\})\right] \\
&\quad= \mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) - \sum_{m=1}^{M}\sum_{t\in\mathcal{J}_m} \mathcal{F}_c^t(\{\psi_c^{\mathcal{J}_m}, \phi_c^{\mathcal{J}_m}\})\right] \qquad \text{(a)} \\
&\quad+ \mathbb{E}_{1:T}\left[\sum_{m=1}^{M}\sum_{t\in\mathcal{J}_m} \mathcal{F}_c^t(\{\psi_c^{\mathcal{J}_m}, \phi_c^{\mathcal{J}_m}\}) - \sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^*, \phi_c^*\})\right] \qquad \text{(b)}
\end{aligned}$$

where $M = \lceil \frac{T}{\tau} \rceil \le T/\tau + 1$ is the number of intervals. Next, we analyze term (a) and term (b) separately.

Analysis of term (a) This term represents the regret of the algorithm compared to the piecewise-stationary reference sequence. The regret with respect to the expected risk $\mathcal{F}_c^t(\cdot)$ can be related to the unbiased empirical risk estimator $\widehat{\mathcal{F}}_c^t(\cdot)$:

$$\begin{aligned}
\text{term (a)} &= \mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) - \sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\})\right] \\
&\le \mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \langle \nabla \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}), (\{\psi_c^t, \phi_c^t\}) - (\{\psi_c^t, \phi_c^t\})\rangle\right] \\
&= \mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \langle \nabla \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) - \nabla \widehat{\mathcal{F}}_c^t(\{\psi_c^t, \phi_c^t\}), (\{\psi_c^t, \phi_c^t\}) - (\{\psi_c^t, \phi_c^t\})\rangle\right] \\
&\quad+ \mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \langle \nabla \widehat{\mathcal{F}}_c^t(\{\psi_c^t, \phi_c^t\}), (\{\psi_c^t, \phi_c^t\}) - (\{\psi_c^t, \phi_c^t\})\rangle\right],
\end{aligned}$$

where the first inequality uses the convexity of $\mathcal{F}_c^t(\cdot)$. The first term above is zero because $\widehat{\mathcal{F}}_c^t$ is an unbiased estimator, i.e., $\nabla \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) = \mathbb{E}_t[\nabla \widehat{\mathcal{F}}_c^t(\{\psi_c^t, \phi_c^t\}) \mid 1 : t-1]$. For the model sequence $\{\{\psi_c^t, \phi_c^t\}\}_{t=1}^T$ generated by Fed-ADE, the following lemma holds:

Lemma 7. *Under the assumptions of Theorem 3, Fed-ADE with learning rate $\eta > 0$ in equation (6) satisfies*

$$\sum_{t=1}^{T} \left\langle \nabla \widehat{\mathcal{F}}_c^t(\{\psi_c^t, \phi_c^t\}), (\{\psi_c^t, \phi_c^t\}) - (\{\psi_c^t, \phi_c^t\}) \right\rangle \le \frac{2\eta|\mathcal{I}|G^2T}{\sigma^2} + \frac{2\Gamma P_T + \Gamma^2}{2\eta},$$

where $P_T = \sum_{t=2}^{T} \|(\{\psi_c^t, \phi_c^t\}) - (\{\psi_c^{t-1}, \phi_c^{t-1}\})\|_2$ is the total variation of the comparator sequence and $G \triangleq \sup_{(\mathbf{x}_c^t, \mathbf{y}_c^t)} \left\|\nabla_{\{\psi_c, \phi_c\}} \mathcal{L}\big(\mathcal{H}\big(\{\psi_c, \phi_c\}, \mathbf{x}_c^t\big), \mathbf{y}_c^t\big)\right\|_2$ is an upper bound on the gradient norm.

Since the comparator sequence for term (a) changes only $M-1$ times, $P_T \le \Gamma(M-1) \le (\Gamma T)/\tau$. Taking expectations gives

$$\text{term (a)} \le \frac{2\eta|\mathcal{I}|G^2T}{\sigma^2} + \frac{2\Gamma^2 T/\tau + \Gamma^2}{2\eta}.$$

Analysis of term (b) This term accounts for the error from changing the reference sequence. Following [5], we get

$$\text{term (b)} \leq 2\tau \sum_{t=2}^{T} \sup_{\{\psi_c^t, \phi_c^t\}} |\mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\}) - \mathcal{F}_c^{t-1}(\{\psi_c^t, \phi_c^t\})| \triangleq 2\tau \bar{\mathcal{S}}_c.$$

Combining the two terms, we obtain

$$\begin{aligned}\mathbb{E}_{1:T}\left[\sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^t, \phi_c^t\})\right] - \sum_{t=1}^{T} \mathcal{F}_c^t(\{\psi_c^*, \phi_c^*\}) &\leq \frac{2\eta|\mathcal{I}|G^2T}{\sigma^2} + \frac{2\Gamma^2 T/\tau + \Gamma^2}{2\eta} + 2B\tau\bar{\mathcal{S}}_c \\ &\leq \left(\frac{2|\mathcal{I}|G^2}{\sigma^2} + 2B^2\right)\eta T + \frac{\Gamma^2}{\eta} + 4(\Gamma+1)\sqrt{\frac{BT\bar{\mathcal{S}}_c}{\eta}},\end{aligned}$$

where we set $\tau = \left\lceil \sqrt{\Gamma^2 T/(\eta B \bar{\mathcal{S}}_c)} \right\rceil$ for optimal balance and $B \triangleq \sup_{(\mathbf{x}_c^t, \mathbf{y}_c^t)} \left|\mathcal{L}\big(\mathcal{H}(\{\psi_c, \phi_c\}, \mathbf{x}_c^t), \mathbf{y}_c^t\big)\right|$ is an upper bound on the loss value, and G represents gradient norm upper bound.

Under the distribution shift assumption, the variation $\bar{\mathcal{S}}_c$ can be further bounded by the change in class and representation priors. Thus,

$$\mathbb{E}\left[\text{Reg}_T\right] \leq \left(\frac{2|\mathcal{I}|G^2}{\sigma^2} + 2B^2\right)\eta T + \frac{\Gamma^2}{\eta} + 4(\Gamma+1)\sqrt{\frac{BT\bar{\mathcal{S}}_c}{\eta}}.$$

□

E.2. Min-max Optimality

For online convex optimization with general convex losses, [5] showed that the dynamic regret has a lower bound of $\Omega\left(\bar{\mathcal{S}}_c^{1/3} T^{2/3}\right)$ when only noisy feedback is observed. Here, $\bar{\mathcal{S}}_c$ measures the total variation of the loss functions. The upper bound in Theorem 3 matches this rate, showing that Fed-ADE achieves min-max optimality up to constants.

Lemma 8. *Under the same assumptions as Theorem 3, Fed-ADE with learning rate η satisfies*

$$\mathbb{E}[\text{Reg}_T] \leq 2\left(\frac{|\mathcal{I}|G^2}{\sigma^2} + B^2\right)\eta T + \frac{\Gamma^2}{\eta} + 4(\Gamma+1)\sqrt{\frac{B\bar{\mathcal{S}}_c T}{\eta}} = \mathcal{O}\left(\eta T + \frac{1}{\eta} + \frac{\sqrt{\bar{\mathcal{S}}_c T}}{\eta}\right),$$

where $\sigma > 0$ is the minimum singular value of the confusion matrix $\mathbf{M}$ and $\bar{\mathcal{S}}_c$ is the temporal variation of the loss.

This result follows by setting $\eta = \Theta\left(T^{-1/3}\bar{\mathcal{S}}_c^{1/3}\right)$, which yields $\mathcal{O}\left(\bar{\mathcal{S}}_c^{1/3} T^{2/3}\right)$ regret, matching the lower bound. Similar reasoning shows that the algorithm also achieves $\mathcal{O}\left(\max\{\bar{\mathcal{S}}_c^{1/3} T^{2/3}, \sqrt{T}\}\right)$ dynamic regret.

F. Convergence Analyses

We next show that our Fed-ADE provably converges when each per-timestep learning rate $\eta_c^t \in [\eta_{min}, \eta_{max}]$ obeys the same upper bound used in Theorem 12. For the proof, we use three commonly used assumptions [6, 51] as follows.

Assumption 9. (Lipschitz parameter) For every client c, the loss function $\mathcal{L}$ is continuously differentiable and there exist *Lipschitz* constants $K_\psi, K_\phi, K_{\psi\phi}, K_{\phi\psi}$ holds that

- $\nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})$ is K_ψ *Lipschitz* with respect to ψ and $K_{\psi\phi}$ *Lipschitz* with respect to $\phi_c^{(r)}$.
- $\nabla_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})$ is K_ϕ *Lipschitz* with respect to ϕ and $K_{\phi\psi}$ *Lipschitz* with respect to $\bar{\psi}^{(r)}$.

When the average loss function of total clients is defined as $\mathcal{L}(\bar{\psi}^{(r)}, \phi^{(r)}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathbb{E}\big[\mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\big]$, it has *Lipschitz* constant K_ψ with respect to $\bar{\psi}^{(r)}$, $K_{\phi\psi}/\sqrt{|\mathcal{C}|}$ with respect to $\phi^{(r)}$, and $K_{\psi\phi}/|\mathcal{C}|$ with respect to any $\phi_c^{(r)}$. Further, there exist χ to measure the relative cross-sensibility of $\nabla_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})$ and $\nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})$ as follows.

$$\chi = \frac{\max\{K_{\psi\phi}, K_{\phi\psi}\}}{\sqrt{K_\psi K_\phi}} \tag{19}$$

Assumption 10. (Bounded variance) The stochastic gradients in the client-side update algorithm are unbiased and have bounded variance as follows.

$$\mathbb{E}\big[\tilde{\nabla}_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\big] = \nabla_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)}) \tag{20}$$

$$\mathbb{E}\big[\tilde{\nabla}_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\big] = \nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)}) \tag{21}$$

Furthermore, there exist constants σ_ϕ and σ_ψ that meet the following inequation.

$$\mathbb{E}\big[\|\tilde{\nabla}_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)}) - \nabla_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\|^2\big] \le \sigma_\phi^2 \tag{22}$$

$$\mathbb{E}\big[\|\tilde{\nabla}_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)}) - \nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\|^2\big] \le \sigma_\psi^2 \tag{23}$$

We can consider $\nabla_\psi \mathcal{L}_c(\bar{\psi}^{(r)}, \phi_c^{(r)})$ as a stochastic partial gradient of average loss function $\mathcal{L}(\bar{\psi}^{(r)}, \phi^{(r)})$ with respect to $\bar{\psi}^{(r)}$.

Assumption 11. (Partial gradient diversity) There exist constants $\delta \ge 0$ and $\rho \ge 0$ such that for $\bar{\psi}^{(r)}$ and all personalized layers $\phi^{(r)} = [\phi_1^{(r)}, \phi_2^{(r)}, ...]$, which meet the following equations.

$$\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \|\nabla_\psi \mathcal{L}_(\bar{\psi}^{(r)}, \phi_c^{(r)}) - \nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi^{(r)})\|^2 \le \delta^2 + \rho^2 \|\nabla_\psi \mathcal{L}(\bar{\psi}^{(r)}, \phi^{(r)})\|^2 \tag{24}$$

Additionally, $\mathcal{L}(\bar{\psi}, \phi^{(r)})$ is bounded below by $\hat{\mathcal{L}}$, where $\hat{\mathcal{L}}$ meets the equation $\Delta\mathcal{L}^{(0)} = \mathcal{L}(\bar{\psi}^{(0)}, \phi^{(0)}) - \hat{\mathcal{L}}$ at initial communication round.

Theorem 12. *The convergence of the Fed-ADE is bounded as follows, where $\Delta\mathcal{L}^{(0)}$ denotes the difference between a bound and initial value of loss $\mathcal{L}(\bar{\psi}, \phi_c)$.*[2]

$$\frac{1}{R}\sum_{r=0}^{R-1}\left(\frac{1}{K_\psi}\mathbb{E}\big[\Delta_\psi^{(r)}\big] + \frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|K_\phi}\mathbb{E}\big[\Delta_\phi^{(r)}\big]\right) \le \frac{\Delta\mathcal{L}^{(0)}}{\eta R} + \eta\sigma_1^2 + \eta^2\sigma_2^2 \tag{25}$$

[2]To simplify the inequality in the following assumptions and theorems, we use following shorthands: $\phi^{(r)} = \{\phi_c^{(r)} | c \in \mathcal{C}^{(r)}\}$, $\mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)}) = \mathcal{L}(\{\bar{\psi}^{(r)}, \phi_c^{(r)}\}; \mathbf{x}_c^t)$, $\mathcal{L}(\bar{\psi}^{(r)}, \phi^{(r)}) = \sum_{c\in\mathcal{C}} \frac{N_c}{N}\mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})$, $\Delta_\psi^{(r)} = \big\|\nabla_\psi \sum_{c\in\mathcal{C}} \frac{N_c}{N}\mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\big\|^2$, $\Delta_\phi^{(r)} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \big\|\nabla_\phi \mathcal{L}(\bar{\psi}^{(r)}, \phi_c^{(r)})\big\|^2$.

Here, the constants σ_1^2 and σ_2^2 are defined as follows.

$$\sigma_1^2 = \frac{\delta^2}{K_\psi}\Big(1 - \frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big) + \frac{\sigma_\psi^2}{K_\psi} + \frac{\sigma_\phi^2\big(|\mathcal{C}^{(r)}| + \chi^2(|\mathcal{C}| - |\mathcal{C}^{(r)}|)\big)}{K_\phi|\mathcal{C}|} \tag{26}$$

$$\sigma_2^2 = \big(\frac{\sigma_\psi^2 + \delta^2}{K_\psi} + \frac{\sigma_\phi^2|\mathcal{C}^{(r)}|}{K_\phi|\mathcal{C}|}\big)\big(1 - E^{-1}\big) + \frac{\chi^2\sigma_\phi^2}{K_\phi} \tag{27}$$

This convergence is bounded when the learning rates are chosen as $\eta_\psi = \eta/(K_\psi E)$ and $\eta_\phi = \eta/(K_\phi E)$, where η satisfies the following inequality.

$$\eta \leq \min\left\{\frac{1}{24(1+\rho^2)}, \frac{|\mathcal{C}^{(r)}|}{128\chi^2(|\mathcal{C}| - |\mathcal{C}^{(r)}|)}, \sqrt{\frac{|\mathcal{C}^{(r)}|}{\chi^2|\mathcal{C}|}}\right\} \tag{28}$$

Proof. The proof of Theorem 12 begins with the following equation describing the update from round r to round $r+1$.

$$\mathcal{L}\Big(\psi^{(r+1)}, \phi^{(r+1)}\Big) - \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r)}\Big) = \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big) - \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r)}\Big) \tag{29}$$

$$+ \mathcal{L}\Big(\psi^{(r+1)}, \phi^{(r+1)}\Big) - \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big) \tag{30}$$

In the case of the update for the personalized layers in (29), since each client is training with their own data, it can be demonstrated similarly to the gradient update in conventional deep learning approaches. However, in the case of the update for the shared layers in (30), as gradient descent is performed for different client models, we aim to establish its convergence.

The smoothness bound for the updating of the shared layers gives an equation as follows.

$$\mathcal{L}\Big(\psi^{(r+1)}, \phi^{(r+1)}\Big) - \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big) \leq \Big\langle \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big), \psi^{(r+1)} - \bar{\psi}^{(r)} \Big\rangle + \frac{K_\psi}{2}\|\psi^{(r+1)} - \bar{\psi}^{(r)}\|^2 \tag{31}$$

According to Lipschitz continuity, we can express the gradient as shown in the (31), when $\langle A, B\rangle$ means $A^T B$. The proof proceeds by demonstrating that for all communication rounds $0 < r < R$, the total sum of this gradient does not diverge but rather remains bounded, as is well known. However, a key distinction between the conventional deep learning approach and the personalized federated learning method is that only a subset of clients among the total clients participate in training during a single round. In these conditions, the virtual full participants $\tilde{\phi}^{(r+1)}$ are used to move all the dependencies of the model update on the participants $\mathcal{C}^{(r)} \in \mathcal{C}$ [51]. When applying the virtual full participants to the (31), we can get an equation as follows.

$$\Big\langle \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big), \psi^{(r+1)} - \bar{\psi}^{(r)} \Big\rangle = \Big\langle \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \tilde{\phi}^{(r+1)}\Big), \psi^{(r+1)} - \psi^{(r)} \Big\rangle + \Big\langle \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \tilde{\phi}^{(r+1)}\Big) - \nabla_\psi \mathcal{L}\Big(\psi^{(r)}, \phi^{(r+1)}\Big), \psi^{(r+1)} - \bar{\psi}^{(r)} \Big\rangle \tag{32}$$

Applying Young's inequality and Lipschitz inequality to the (32), the (31) can be expressed as follows.

$$\mathcal{L}\Big(\psi^{(r+1)}, \phi^{(r+1)}\Big) - \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big) \leq \Big\langle \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big), \psi^{(r+1)} - \bar{\psi}^{(r)} \Big\rangle + \frac{1}{2K_\psi}\|\nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \tilde{\phi}^{(r+1)}\Big) - \nabla_\psi \mathcal{L}\Big(\bar{\psi}^{(r)}, \phi^{(r+1)}\Big)\|^2 + K_\psi\|\psi^{(r+1)} - \bar{\psi}^{(r)}\|^2 \tag{33}$$

This allows us to take an expectation. Therefore, the final bounded form can be described as follows with Assumption 11, where the $\tilde{\Delta}_\psi^{(r)}$ is the analog of $\Delta_\psi^{(r)}$ with the virtual variable $\tilde{\phi}^{(r+1)}$.

$$\frac{1}{R}\sum_{r=0}^{R-1}\left(\frac{\eta_\psi E}{8}\mathbb{E}[\tilde{\Delta}_\psi^{(r)}] + \frac{\eta_\phi E c}{16|\mathcal{C}|}\mathbb{E}[\Delta_\phi^{(r)}]\right) \leq \frac{\Delta\mathcal{L}^{(0)}}{R} + \mathcal{O}(\eta_\psi^2 + \eta_\phi^2). \tag{34}$$

Bound of shared layers update When using (34), the update of the shared layers can be bounded as follows.

$$\begin{aligned}
&\mathcal{L}\Big(\psi^{(r+1)},\phi^{(r+1)}\Big)-\mathcal{L}\Big(\psi^{(r)},\phi^{(r+1)}\Big)\\
\leq&\Big\langle\nabla_p\mathcal{L}\Big(\psi^{(r)},\tilde{\phi}^{(r+1)}\Big),\psi^{(r+1)}-\psi^{(r)}\Big\rangle K_\psi\|\psi^{(r+1)}-\psi^{(r)}\|^2+\frac{\chi^2K_\phi}{2|\mathcal{C}|}\sum_{c\in\mathcal{C}}\|\tilde{\phi}_c^{(r+1)}-\phi_c^{(r+1)}\|^2
\end{aligned}\tag{35}$$

Thus, the dependence of the second line term on $\phi^{(r+1)}$ is successfully eliminated. To bind the third line term, expectation can be used as follows.

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{L}\Big(\psi^{(r+1)},\phi^{(r+1)}\Big)-\mathcal{L}\Big(\psi^{(r)},\phi^{(r+1)},\Big)\Big]\\
&\leq-\frac{\eta_\psi E}{4}\mathbb{E}\big[\tilde{\Delta}_\psi^{(r)}\big]+\frac{2\eta_\psi K_\psi^2}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{e=0}^{E}\mathbb{E}\big\|\tilde{\psi}_{c,e}^{(r)}-\psi^{(r)}\big\|^2+4\eta_\phi^2E^2K_\phi\sigma_\phi^2\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\\
&\quad+\frac{K_\psi\eta_\psi^2E^2}{|\mathcal{C}^{(r)}|}\Big(\sigma_\psi^2+3\delta^2\big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\big)\Big)+8\eta_\psi^2E^2K_\phi\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\Delta_\phi^{(r)}
\end{aligned}\tag{36}$$

Note that $24K_\psi\eta_\psi E(1+\rho^2)\leq 1$ is used to simplify the coefficients of some of the terms above. The term in the third line is referred to as client shift in the literature. This term can be described using the virtual variable $\tilde{\phi}^{(r+1)}$ as follows.

$$\begin{aligned}
&\frac{2\eta_\psi K_\psi^2}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\sum_{e=0}^{E}\mathbb{E}\|\tilde{\psi}_{c,e}^{(r)}-\psi^{(r)}\|^2\\
&\leq\frac{16\eta_\psi^3K_\psi^2E(E-1)}{|\mathcal{C}|}\Big(\delta^2+\rho^2\mathbb{E}\|\nabla_\psi\mathcal{L}\big(\psi^{(r)},\tilde{\phi}^{(r+1)}\big)\|^2\Big)+8\eta_\psi^3K_\psi^2E^2(E-1)\sigma_\psi^2
\end{aligned}\tag{37}$$

By inputting (37) into the previous inequality of expectation (36), the inequality can be described as follows.

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\psi^{(r+1)},\phi^{(r+1)})-\mathcal{L}(\psi^{(r)},\phi^{(r+1)})]\\
&\leq-\frac{\eta_\psi E}{8}\mathbb{E}[\tilde{\Delta}_\psi^{(r)}]+\frac{K_\psi\eta_\psi^2E^2}{|\mathcal{C}^{(r)}|}\Big(\sigma_\psi^2+2\delta^2(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|})\Big)+4\eta_\phi^2E^2K_\phi\sigma_\phi^2\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\\
&+8\eta_\phi^2E^2K_\phi\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\Delta_\phi^{(r)}+8\eta_\psi^2K_\psi^3E^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)(\sigma_\psi^2+2\delta_\psi)
\end{aligned}\tag{38}$$

Bound of personalized layers The above analysis for updating the shared layers can be applied to updating the personalized layers. This analysis provides the following, which displays the bound of the update of the personalized layers. It can be simplified some coefficients using $128\eta_\phi EK_\phi\chi^2(\frac{|\mathcal{C}|}{|\mathcal{C}^{(r)}|}-1)\leq 1$.

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{L}\big(\phi^{(r+1)},\psi^{(r+1)}\big)-\mathcal{L}\big(\psi^{(r)},\phi^{(r)}\big)\Big]\\
&\leq-\frac{\eta_\psi E}{8}\mathbb{E}\big[\tilde{\Delta}_\psi^{(r)}\big]-\frac{\eta_\phi E|\mathcal{C}^{(r)}|}{16|\mathcal{C}|}\mathbb{E}\big[\Delta_\phi^{(r)}\big]+4\eta_\phi^2K_\phi E^2\sigma_\phi^2\Big(\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}+\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\Big)\\
&+\frac{\eta_\psi^2K_\psi E^2}{|\mathcal{C}^{(r)}|}\Big(\sigma_\psi^2+2\delta^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\Big)+8\eta_\psi^2K_\psi^2E^2(E-1)(\sigma_\psi^2+2\delta^2)+\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}4\eta_\phi^3K_\phi^2E^2(E-1)\sigma_\phi^2
\end{aligned}\tag{39}$$

The summation of inequality (39) for considering the communication round is given below.

$$
\begin{aligned}
&\frac{1}{R}\sum_{r=0}^{R-1}\left(\frac{\eta_\psi E}{8}\mathbb{E}\big[\tilde{\Delta}_\psi^{(r)}\big]+\frac{\eta_\phi E|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\mathbb{E}\big[\Delta_\phi^{(r)}\big]\right)\\
\leq&\frac{\Delta\mathcal{L}^{(0)}}{R}+4\eta_\phi^2K_\phi E^2\sigma_\phi^2\left(\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}+\chi^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\right)+\frac{\eta_\psi^2K_\psi E^2}{|\mathcal{C}^{(r)}|}\left(\sigma_\psi^2+2\delta^2\Big(1-\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}\Big)\right)\\
&+8\eta_\psi^3K_\psi^2E^2(E-1)(\sigma_\psi^2+2\delta^2)+\frac{|\mathcal{C}^{(r)}|}{|\mathcal{C}|}4\eta_\phi^3K_\phi^2E^2(E-1)\sigma_\phi^2
\end{aligned}
\tag{40}
$$

This is the bound about virtual variable $\tilde{\phi}^{(r+1)}$. To describe this bound with the participants $\phi^{(r+1)}$, the assumption of the Lipschitz parameter is used as follows.

$$
\begin{aligned}
&\mathbb{E}\big\|\nabla_\psi\mathcal{L}(\psi^{(r)},\phi^{(r)})-\nabla_\psi\mathcal{L}(\psi^{(r)},\tilde{\phi}^{(r+1)})\big\|^2\\
&\leq\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mathbb{E}\big\|\nabla_\psi\mathcal{L}(\psi^{(r)},\phi_c^{(r)})-\nabla_\psi\mathcal{L}(\psi^{(r)},\tilde{\phi}_c^{(r+1)})\big\|^2\\
&\leq\frac{\chi^2K_\phi K_\psi}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mathbb{E}\big\|\tilde{\phi}_c^{(r+1)}-\phi_c^{(r)}\big\|^2\\
&\leq\frac{\chi^2K_\phi K_\psi}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\Big(16\eta_\phi^2E^2\big\|\nabla_p\mathcal{L}(\psi^{(r)},\phi_c^{(r)})\big\|^2+8\eta_\phi^2E^2\sigma_\phi^2\Big)=8\eta_\phi^2E^2\chi^2K_\phi K_\psi\Big(\sigma_\phi^2+2\mathbb{E}\big[\Delta_\phi^{(r)}\big]\Big)
\end{aligned}
\tag{41}
$$

Using the Cauchy–Schwarz inequality described as follows.

$$
\|\nabla_\psi\mathcal{L}(\psi^{(r)},\phi^{(r)})\|^2\leq2\|\nabla_\psi\mathcal{L}(\psi^{(r)},\phi^{(r)})-\nabla_\psi\mathcal{L}(\psi^{(r)},\tilde{\phi}^{(r+1)})\|^2+2\|\nabla_\psi\mathcal{L}(\psi^{(r)},\tilde{\phi}^{(r+1)})\|^2 \tag{42}
$$

By utilizing inequality (42), we obtain the following result.

$$
\mathbb{E}\big[\Delta_\phi^r\big]\leq\mathbb{E}[\tilde{\Delta}_\psi^{(r)}]+16\eta_\phi^2E^2\chi^2K_\phi K_\psi(\sigma_\phi^2+2\mathbb{E}[\Delta_\phi^{(r)}]) \tag{43}
$$

Through inequality (43), we can get the following inequality.

$$
\frac{\eta_\psi E}{16}\mathbb{E}[\Delta_\psi^{(r)}]+\frac{\eta_\phi E|\mathcal{C}^{(r)}|}{32|\mathcal{C}|}\mathbb{E}[\Delta_\phi^{(r)}]\leq\frac{\eta_\psi E}{8}\mathbb{E}[\tilde{\Delta}_\psi^{(r)}]+\frac{\eta_\phi E|\mathcal{C}^{(r)}|}{16|\mathcal{C}|}\mathbb{E}[\Delta_\phi^{(r)}]+\eta_\psi\eta_\phi^2E^3\sigma_\phi^2\chi^2K_\phi K_\psi \tag{44}
$$

By summing inequality (44) over the total number of communication rounds and applying the parameter settings $\eta_\phi=\eta/(K_\phi E)$ and $\eta_\psi=\eta/(K_\psi E)$, the proof is completed. □

To ensure that the convergence analysis above remains valid when using an adaptive learning rate in (9), we require that the value of the adaptive learning rate η_c across all clients and rounds is bounded by the same upper bound used in Theorem 12. Since our proposed label shift adaptive learning rate can not have value outside of the bound $[\eta_{min},\eta_{max}]$ (where $\eta_c^t\in[\eta_{min},\eta_{max}]$), by appropriately choosing $[\eta_{min},\eta_{max}]\leq\min\left\{\frac{1}{24(1+\rho^2)},\frac{|\mathcal{C}^{(r)}|}{128\chi^2(|\mathcal{C}|-|\mathcal{C}^{(r)}|)},\sqrt{\frac{|\mathcal{C}^{(r)}|}{\chi^2|\mathcal{C}|}}\right\}$, we guarantee that all learning rates used in the adaptive scheme satisfy the conditions required for Theorem 12.

G. Simulation Settings

G.1. Datasets

We evaluate on various benchmarks including five image-based benchmarks- Tiny ImageNet [32], CIFAR-10 [29], CIFAR-10-C [19], CIFAR-100 [29], and CIFAR-100-C [19], and a text-based benchmark LAMA [50]. The following is a detailed description of the datasets used in this work.

- **Tiny ImageNet [32]:** Tiny ImageNet is a subset of the ImageNet dataset, containing 200 object classes, each with 500 training images, 50 validation images, and 50 test images. All images are RGB and resized to 64×64 pixels.
- **CIFAR-10 [29]:** The CIFAR-10 dataset consists of 60,000 RGB images of size 32×32 across 10 classes, with 50,000 images for training and 10,000 for testing. We applied standard preprocessing including normalization with dataset-wide mean and standard deviation.
- **CIFAR-10-C [19]:** CIFAR-10-C is an extension of the CIFAR-10 dataset designed to evaluate model robustness against common corruptions. It contains 950,000 images created by applying 19 different types of corruption—such as noise, blur, weather effects, and digital distortions—at 5 severity levels to the original CIFAR-10 test set of 10,000 images.
- **CIFAR-100 [29]:** CIFAR-100 is a fine-grained variant of CIFAR-10 containing 100 object classes. It consists of 60,000 color images at 32×32 pixels, split into 50,000 training and 10,000 test images. The 100 classes are grouped into 20 superclasses for hierarchical classification research.
- **CIFAR-100-C [19]:** CIFAR-100C applies the same 19 corruption types with 5 severity levels from CIFAR-10-C onto the CIFAR-100 dataset. It consists of 950,000 corrupted images used as a benchmark for analyzing model robustness on fine-grained image classification tasks.
- **LAMA [50]:** LAnguage Model Analysis is a benchmark dataset designed to evaluate the factual and commonsense knowledge acquired by language models. In our experiments, we focus on the T-REx subset of LAMA, which consists of (subject, relation, object) triples derived from Wikidata. This subset consists of 41 relational tasks, with 100 samples randomly sampled per task. To simulate heterogeneous client environments, we assign disjoint subsets of tasks to different clients.

G.2. Online Distribution Shift Modeling

To emulate online distribution shift in (1), each client's data distribution moves from a common initial uniform vector $\mathbf{Q}_c^0$ toward a client-specific target $\mathbf{Q}_c^T$ according to a time-dependent weight $\omega(t)$. We evaluate under four prototypical distribution shift schedules—linear (Lin.), sine (Sin.), square (Squ.), and Bernoulli (Ber.), each driving a client's distribution from a uniform start toward a client-specific target [10, 45, 56, 65, 69]. Following is a detailed description of four prototypical label-shift schedules according to different $\omega(t)$ functions, and Figure 3 shows how the parameter $\omega(t)$ changes over time for various shift types.

- Linear shift (Lin.): $\omega(t) = \frac{t}{T}$, where t is the current timestep and T is the total number of timesteps. This provides a smooth and gradual transition from $\mathbf{Q}_c^0$ to $\mathbf{Q}_c^T$ over time.
- Sine shift (Sin.): $\omega(t) = \sin\left(\frac{\pi t}{\sqrt{T}}\right)$. This introduces cyclical variations, capturing periodic shifts in the data distribution.
- Square Shift (Squ.): $\omega(t)$ alternates between 0 and 1 every $\frac{\sqrt{T}}{2}$ timestep, resulting in abrupt, periodic changes in the class distribution.
- Bernoulli shift (Bern.): $\omega(t)$ retains its previous value $\omega(t-1)$ with a probability of $\frac{1}{\sqrt{T}}$, or flips to $1 - \omega(t-1)$. This configuration captures stochastic variations, ensuring that class priors change randomly while maintaining a scale proportional to $\sqrt{T}$.

For $\mathbf{Q}_c^T$, we used the Dirichlet distribution to consider heterogeneous data distributions among clients. The distribution factor α of the Dirichlet distribution controls the level of heterogeneity among clients. Smaller α leads more heterogeneous distribution, and Larger α leads to Independent and identically distributed data system. In this study, we set $Dir(\alpha = 0.1)$. Figure 4 visualizes the heterogeneity among clients according to Dirichlet distribution factor α.

G.3. Distribution Shift Scenarios

In all experiments, each target distribution of client $\mathbf{Q}_{\mathbf{y}_c^T}$ in (1) is defined according to the specific type of distribution shift being simulated. We summarize below the construction and datasets used for each shift scenario.

- **Label Shift.** Label shift simulations are conducted on Tiny ImageNet [32], CIFAR-10 [29], and LAMA [50], where each client's class prior evolves over time according to one of four temporal schedules (Lin., Sin, Squ., Ber.) which follows (1).
- **Covariate Shift.** The target feature distribution $\mathbf{Q}_{c,\mathbf{x}}^T$ reflects corruption-induced changes to the input domain. Each

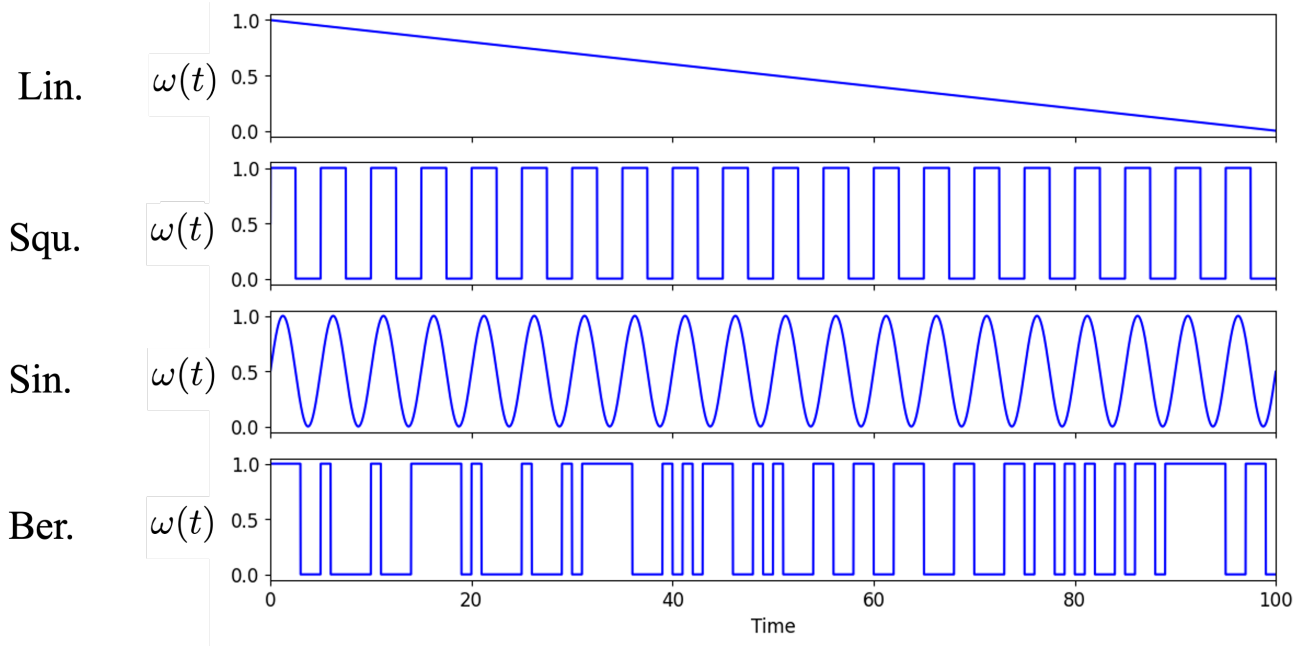
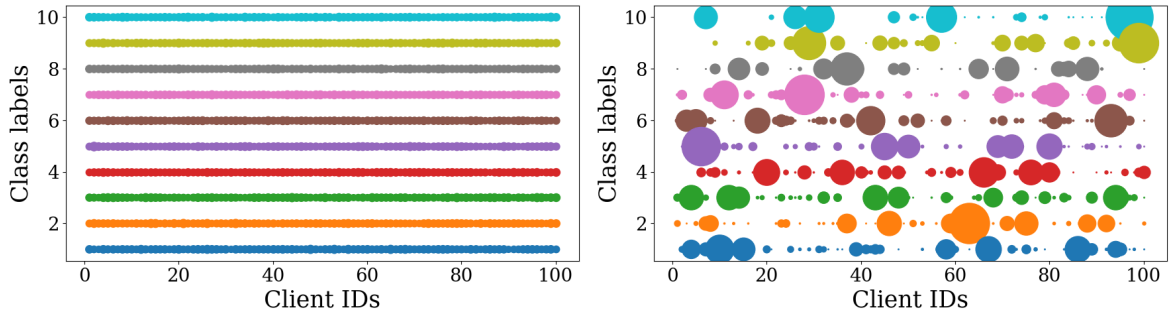


Figure 3. Visualization of how the parameter $\omega(t)$ changes over time for various shift types. The Lin. presents a steady and continuous increase or decrease, reflecting a gradual and predictable change. Both the Squ. and Sin. display periodic patterns. The Squ. alternates sharply between two values at regular intervals, while the Sin. oscillates smoothly in a wave-like manner. The Ber. introduces randomness at each timestep, resulting in a stochastic and less predictable trajectory for $\omega(t)$.



(a) Dir($\alpha = 100$) (b) Dir($\alpha = 0.1$)

Figure 4. Visualization of heterogeneity among clients according to Dirichlet distribution factor α in CIFAR-10 dataset. The x-axis represents the ID of the client, while the y-axis represents the class of the data. The color of the circles varies with each class type, and the size of the circles indicates the size of the data. This level of heterogeneity is not dependent on the data set but on α values.

client is assigned a distinct and fixed corruption type (e.g., Gaussian noise, blur), while the corruption severity level shifts dynamically over time according to one of four temporal schedules. We use the CIFAR-10 [29] $\rightarrow$ CIFAR-10-C [19] and CIFAR-100 [29] $\rightarrow$ CIFAR-100-C [19] benchmarks to simulate these covariate-shift environments.

G.4. Compared Methods

We evaluate our proposed Fed-ADE against six existing unsupervised online adaptation baselines:

- **FTH** [58]: FTH maintains an average of estimated label distributions across all previous timesteps. By aggregating the entire history of predictions, FTH assumes that past distributional information is informative for current adaptation, making it robust to gradual, monotonic shifts but potentially less responsive to abrupt changes.
- **ATLAS** [2]: ATLAS maintains an ensemble of base learners, each optimized with a distinct learning rate. Using an exponential weighting scheme, ATLAS dynamically aggregates these learners according to their recent performance,

enabling robust adaptation to various patterns and magnitudes of label shift.

- **UDA** [18]: UDA is a technique where a model is trained on labeled data from a source domain and adapted to work well on a target domain that has no labels, even though the data distributions are different. The main idea is to reduce the gap between the source and target domains by aligning their feature representations or adjusting the model so it can generalize to the target data, all without using any target domain labels.
- **UNIDA** [63]: UNiDA is a method designed to handle domain adaptation scenarios. This assumes the same class set across domains, which aims to adapt models by distinguishing between common, source-private, and target-private classes, often using prototype-based or confidence-based techniques to avoid misclassifying unknown target classes as known ones.
- **Fed-POE** [43]: Fed-POE is a federated online adaptation framework in which each client combines local fine-tuning with an ensemble of periodically aggregated global models. While this approach supports client personalization and leverages federated knowledge, it employs a fixed learning rate and incurs computational overhead from managing multiple model instances.
- **FedCCFA** [9]: FedCCFA is a federated learning framework for concept shift, where clients may experience different distributions. It performs class-level classifier clustering to share aggregated class-wise classifiers among similar clients, and aligns feature spaces using clustered feature anchors with an entropy-based adaptive weighting.

G.5. Hyperparameters

In this section, we summarize hyperparameters used in the simulations.

Table 6. Hyperparameters for Tiny ImageNet, CIFAR-10, and CIFAR-100.

Hyperparameter	**Tiny ImageNet**	**CIFAR-10**	**CIFAR-100**
Number of clients	100	100	100
Communication rounds (R)	10	10	10
Local epochs	4	4	4
Optimizer	SGD	SGD	SGD
Batch size	128	32	32
Participant rate	10%	10%	10%
Model architecture	ResNet18	CNN w/ 3 residual blocks	CNN w/ 3 residual blocks
Shared layers (ψ_c)	Representation layers	Representation layers	Representation layers
Lowest learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}
Highest learning rate	10^{-4}	10^{-4}	10^{-4}

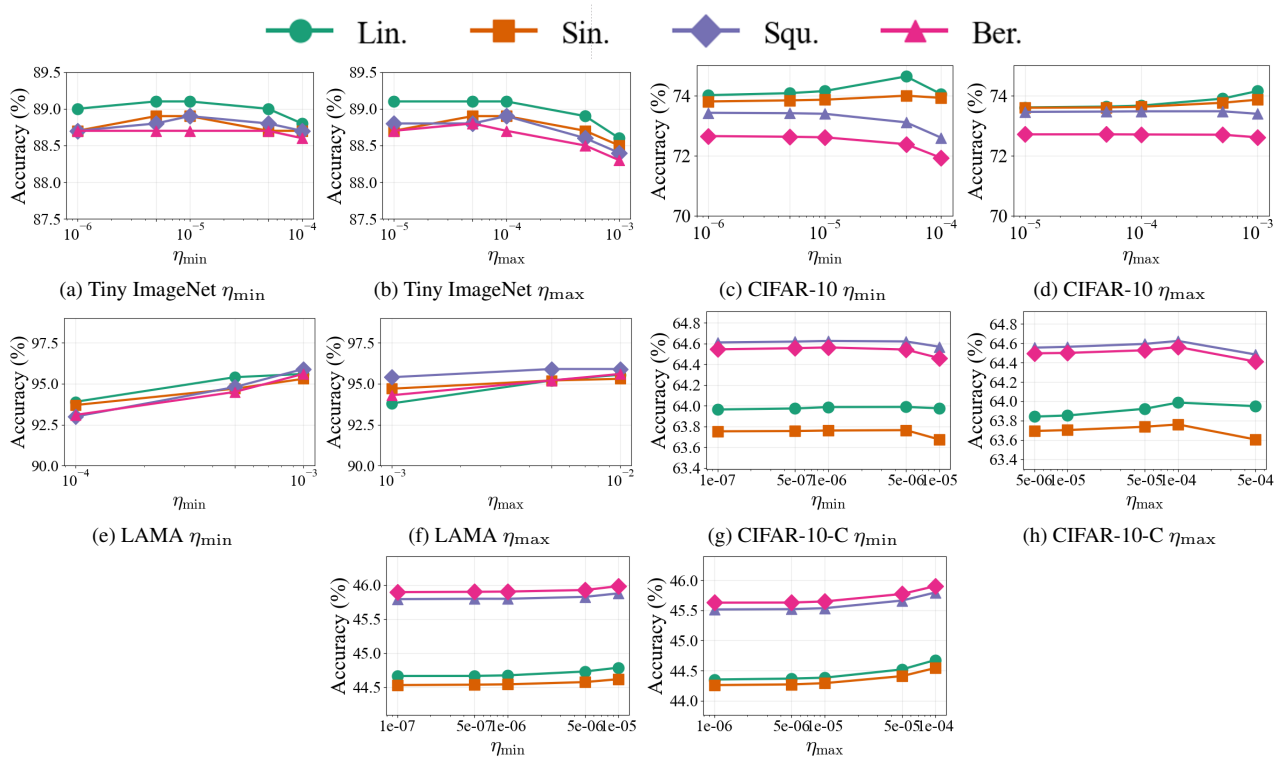
Table 7. Hyperparameters for CIFAR-10-C, CIFAR-100-C, and LAMA

Hyperparameter	**CIFAR-10-C**	**CIFAR-100-C**	**LAMA**
Number of clients	100	100	5
Communication rounds (R)	10	10	5
Local epochs	4	4	1
Optimizer	SGD	SGD	AdamW
Batch size	32	32	4
Participant rate	10%	10%	100%
Model architecture	CNN w/ 3 residual blocks	CNN w/ 3 residual blocks	Llama-3.2-3B
Shared layers (ψ_c)	Representation layers	Representation layers	LoRA on q/k/v (layers 0–13)
Lowest learning rate	5×10^{-6}	5×10^{-6}	1×10^{-3}
Highest learning rate	1×10^{-4}	1×10^{-4}	1×10^{-2}

G.6. Hardware Specifications

Table 8. Hardware specifications used for all simulations.

Component	**Specification**
CPU	AMD Ryzen 9 7950X, 16-Core, 32 Threads
GPU	NVIDIA GeForce RTX 3090X2
RAM	256 GB DDR5
Storage	1.8 TB NVMe SSD



(a) Tiny ImageNet $\eta_{\min}$ (b) Tiny ImageNet $\eta_{\max}$ (c) CIFAR-10 $\eta_{\min}$ (d) CIFAR-10 $\eta_{\max}$

(e) LAMA $\eta_{\min}$ (f) LAMA $\eta_{\max}$ (g) CIFAR-10-C $\eta_{\min}$ (h) CIFAR-10-C $\eta_{\max}$

(i) CIFAR-100-C $\eta_{\min}$ (j) CIFAR-100-C $\eta_{\max}$

Figure 5. Impact of learning rate bounds on Fed-ADE under label shift scenarios.

H. Extra Simulation Results

H.1. Impact of Learning Rate Selection

Figure 5 examines the sensitivity of Fed-ADE to the choice of learning rate bounds under all scenarios and benchmarks. For image benchmarks, we vary $\eta_{\min}$ while fixing $\eta_{\max} = 10^{-4}$, and vary $\eta_{\max}$ while fixing $\eta_{\min} = 5 \times 10^{-6}$. For text benchmark LAMA, we vary $\eta_{\min}$ while fixing $\eta_{\max} = 10^{-2}$, and vary $\eta_{\max}$ while fixing $\eta_{\min} = 10^{-4}$. Across all cases, Fed-ADE consistently maintains high accuracy within a broad range of hyperparameter choices. While extreme values lead to slight performance degradation, the overall trend shows that our dynamics-driven adaptation is far less sensitive to the exact learning rate bounds compared to fixed-rate baselines. This robustness highlights that Fed-ADE reduces the need for extensive hyperparameter tuning, an essential property in federated and online learning where search budgets are limited.

Table 9. Performance comparison of different online distribution shift adaptation methods across datasets and shift types with pretrained model from Gaussian distribution (average accuracy (%) and average wall time (sec.))

		Localized Learning				Federated Learning					
Dataset	**Shift**	FTH	ATLAS	UNIDA	UDA	Fed-POE	FedCCFA	FixLR(Low)	FixLR(Mid)	FixLR(High)	**Fed-ADE**
(i) Label Shift Scenarios											
Tiny ImageNet	Lin.	74.1±2.3	70.7±2.1	75.1±2.1	63.1±2.3	81.5±1.2	81.7±1.8	81.5±1.1	81.2±1.3	78.1±1.5	**84.5±1.1**
	Sin.	73.7±2.4	71.1±1.9	74.8±2.5	61.3±2.1	82.1±1.2	80.9±1.1	81.9±1.7	81.0±2.1	77.5±2.1	**84.7±0.7**
	Squ.	72.6±2.4	70.9±1.8	74.6±2.2	61.7±2.1	81.5±1.3	78.8±1.9	82.4±1.2	80.9±1.6	77.7±1.3	**83.9±0.9**
	Ber.	72.1±2.3	71.0±2.0	74.7±2.1	61.1±1.7	80.9±1.6	79.1±1.1	82.1±0.9	80.8±1.7	77.8±1.5	**84.2±0.4**
CIFAR-10	Lin.	28.4±1.7	22.6±3.3	17.2±4.1	27.4±4.2	66.1±1.2	53.6±2.1	64.8±2.2	65.7±2.1	51.3±3.1	**68.0±1.7**
	Sin.	28.9±2.1	23.1±3.1	18.3±4.5	27.1±3.7	66.7±1.2	53.3±2.1	64.7±1.9	65.1±1.8	51.6±4.1	**67.7±1.7**
	Squ.	25.9±1.7	21.6±2.6	15.8±3.7	25.1±2.2	67.1±1.5	54.3±1.3	64.8±1.2	65.7±1.6	52.3±3.5	**66.4±1.6**
	Ber.	26.1±2.1	21.7±2.9	16.1±4.1	24.5±2.5	67.0±1.3	54.1±1.3	64.3±1.7	65.4±1.4	52.6±2.8	**67.1±1.3**
(ii) Covariate Shift Scenarios											
CIFAR-10 CIFAR-10-C	Lin.	15.4±0.4	10.1±0.8	39.1±0.5	37.8±1.1	41.6±0.4	35.2±0.5	58.0±0.6	58.1±0.6	36.7±2.4	**59.5±0.5**
	Sin.	15.2±0.5	11.9±0.7	39.2±0.4	37.2±0.9	41.7±0.6	33.1±0.4	57.1±0.5	57.8±0.7	35.3±1.7	**59.3±0.5**
	Squ.	13.5±0.5	13.1±0.6	38.3±0.4	36.8±1.2	43.1±0.5	32.1±0.7	56.9±0.9	57.3±1.1	35.6±2.1	**60.4±0.5**
	Ber.	14.8±0.5	12.2±1.1	38.7±0.4	36.6±1.1	43.2±0.5	33.0±0.5	56.7±0.7	67.1±1.0	46.1±2.7	**60.3±0.4**
CIFAR-100 CIFAR-100-C	Lin.	4.2±1.7	2.4±0.2	29.1±0.1	31.3±0.3	20.2±0.4	20.7±0.6	42.0±0.4	40.8±0.2	37.1±1.1	**43.1±0.5**
	Sin.	5.6 ±2.2	2.5±0.1	29.5±0.3	30.7±0.8	20.3±0.5	20.7±0.6	41.9±0.4	40.9±0.3	36.6±1.3	**44.4±0.4**
	Squ.	5.7±2.1	3.2±0.3	27.1±0.4	29.1±0.7	22.9±0.4	18.6±0.8	41.6±0.3	41.2±0.5	37.1±1.5	**45.1±0.5**
	Ber.	6.1±1.8	3.1±0.4	27.8±0.4	29.6±0.8	21.8±0.8	19.3±0.8	40.3±0.5	41.4±0.5	34.1±1.4	**44.8±0.6**

H.2. Impact of Distribution of Pre-training Data

In practical federated deployments, the exact distribution of the pre-training data on the server is often unknown to clients. To examine the robustness of Fed-ADE under such realistic conditions, we evaluate post-adaptation performance on image benchmarks when the pre-training data follows two non-uniform distributions: *Gaussian* and *Exponential Decay*. As shown in Table 9, and Table 10, Fed-ADE maintains consistently strong performance across all distribution settings and shift types, with only a minor variation compared to the uniform case (cf. Table 1). These results demonstrate that our dynamics-based adaptation effectively generalizes even when the pre-training distribution differs from the assumed prior, validating its robustness under practical deployment scenarios.

H.3. Why Cosine Similarity? Alternatives and Ablations

In this paper, our Fed-ADE must be (i) online & per-client (no extra communication), (ii) label-free (only pseudo-labels/representations), (iii) robust to pseudo-label noise and transient shifts, (iv) lightweight in compute and memory, and (v) bounded and interpretable so that we can stably map it to a learning rate in $[\eta_{\min}, \eta_{\max}]$ across heterogeneous clients.

Cosine similarity meets these requirements

- **Bounded**: S_c^t is bounded in [0,1], so it leads to a stable mapping to learning across clients and timestep
- **Efficient**: Linear-time, no optimization/probabilistic inference
- **Robust**: Scale-invariant and less sensitive to sparsity/outliers in pseudo-label histograms; temperature rescaling or mild miscalibration does not change directions.

Alternatives considered We implemented three commonly suggested alternatives and found consistent drawbacks in our setting.

(1) KL divergence on batch histograms: *Pros:* classical information-theoretic distance. *Cons:* asymmetric and *unbounded*; becomes unstable/infinite under zeros/sparsity—common with noisy pseudo-labels on small batches. Requires smoothing/regularization and hyperparameters (e.g., ϵ-floor), which are highly dataset/client dependent and harm cross-client consistency. Mapping an unbounded statistic to a stable learning rate is delicate.

(2) Wasserstein distance: Requires a ground metric on classes or high-dim features; even with Sinkhorn regularization, it needs iterative optimization (per batch pair) and careful tuning. *Pros:* captures geometric discrepancies. *Cons:* higher compute

Table 10. Performance comparison of different online distribution shift adaptation methods across datasets and shift types with pretrained model from exponential decay distribution(average accuracy (%) and average wall time (sec.))

		Localized Learning				**Federated Learning**					
Dataset	**Shift**	FTH	ATLAS	UNIDA	UDA	Fed-POE	FedCCFA	FixLR(Low)	FixLR(Mid)	FixLR(High)	**Fed-ADE**
(i) Label Shift Scenarios											
Tiny ImageNet	Lin.	78.2±1.0	70.2±2.1	77.3±0.8	69.0±0.6	79.9±0.3	82.1±0.4	81.8±0.5	82.5±0.3	81.4±0.2	**83.1±0.4**
	Sin.	77.9±0.8	71.4±1.6	77.1±0.6	69.4±0.5	80.4±0.6	81.5±0.4	82.1±0.2	82.7±0.5	81.3±0.3	**82.7±0.5**
	Squ.	77.2±0.8	71.1±1.8	76.9±0.6	69.4±0.5	80.9±0.5	81.2±0.3	82.3±0.4	82.1±0.4	80.1±0.3	**82.6±0.3**
	Ber.	78.2±1.1	70.7±1.6	76.6±0.7	69.7±0.4	80.4±0.5	80.4±0.3	81.1±0.4	82.1±0.2	80.4±0.3	**81.8±0.4**
CIFAR-10	Lin.	27.7±1.4	29.0±2.0	20.2±1.3	27.6±1.8	65.1±1.0	63.8±0.9	64.1±1.4	64.8±1.3	60.1±1.3	**66.2±1.7**
	Sin.	28.4±1.2	28.3±2.0	20.3±1.1	28.1±1.6	64.7±0.6	63.1±1.3	63.8±1.2	64.3±1.4	60.3±0.8	**65.9±1.7**
	Squ.	27.1±1.2	28.5±1.8	21.1±0.9	24.8±1.6	64.6±0.9	62.8±1.1	64.5±1.1	64.1±1.6	61.1±0.9	**66.4±1.5**
	Ber.	26.6±1.5	27.1±1.9	20.7±1.1	25.1±2.1	64.3±1.1	62.3±0.7	64.7±1.4	63.8±1.5	60.8±1.2	**67.1±1.3**
(ii) Covariate Shift Scenarios											
CIFAR-10 CIFAR-10-C	Lin.	23.7±0.2	13.9±0.2	43.3±0.6	45.7±0.8	44.5±0.8	43.1±0.5	63.4±0.2	63.9±0.3	40.6±2.1	**58.3±0.5**
	Sin.	22.9±0.2	13.2±0.2	44.8±0.1	43.5±0.3	44.7±1.2	43.3±0.4	63.5±0.3	63.7±0.3	38.1±3.1	**58.2±0.5**
	Squ.	23.8±0.1	14.1±0.4	42.2±0.1	43.3±0.2	48.5±1.2	41.6±0.7	62.7±2.1	64.5±2.1	39.6±2.8	**59.0±0.5**
	Ber.	23.6±0.2	14.2±0.3	42.7±0.2	42.3±0.1	48.7±0.9	42.0±0.5	64.1±2.1	64.4±2.1	40.8±2.2	**59.0±0.4**
CIFAR-100 CIFAR-100-C	Lin.	5.1±1.8	6.4±0.4	24.4±0.5	27.6±0.5	19.1±2.1	17.9±1.1	40.8±0.5	39.1±0.6	37.7±1.2	**42.7±0.5**
	Sin.	5.3 ±1.7	5.1±0.4	24.3±0.6	28.1±0.5	20.1±1.9	18.3±1.1	40.7±0.6	38.6±0.7	37.6±1.2	**42.4±0.6**
	Squ.	4.8±1.5	4.7±0.5	24.1±0.4	28.3±0.7	19.8±2.1	18.1±0.9	39.4±0.6	38.2±0.7	38.1±1.1	**42.1±0.4**
	Ber.	4.3±1.3	5.1±0.3	23.5±0.4	27.8±0.6	19.7±2.3	16.9±1.0	40.1±0.4	37.7±0.6	37.8±1.2	**42.0±0.5**

and memory; sensitive to transient fluctuations; scale depends on ground metric ⇒ learning rate scaling across clients becomes ill-conditioned. In high-dim feature space, pairwise costs/transport exacerbate overhead.

(3) Bayesian change-point detection: *Pros:* probabilistic treatment of regime switches. *Cons:* assumes i.i.d. within segments and requires prior modeling; posterior updates/inference add computation; fragile to batch-level noise (false alarms/delays). Typically designed for detection (binary), not for per-step *continuous* shift magnitudes needed to drive learning rate.

Performance comparison of cosine similarity vs. alternatives We compared cosine similarity with KL, Wasserstein, and Bayesian CPD under four dynamic shift patterns (Lin./Sin./Squ./Ber.) on CIFAR-10. Cosine consistently achieved the best accuracy:

Measure	**Lin.**	**Sin.**	**Squ.**	**Ber.**
Fed-ADE (Cosine similarity)	**73.8 ± 0.6**	**73.6 ± 0.5**	**72.2± 1.6**	**72.9 ± 2.2**
KL divergence	63.2 ± 3.8	62.4 ± 2.3	61.1 ± 3.4	61.4 ± 3.6
Wasserstein	70.8 ± 0.6	69.9 ± 1.1	68.8 ± 1.3	66.5 ± 2.1
Bayesian CPD	71.5 ± 0.5	70.8 ± 0.8	69.7 ± 2.4	69.0 ± 1.1

Gains of the cosine similarity method are largest under label noise and transient shifts, where bounded, direction-only comparison stabilizes learning rate scaling across clients.