E Theoretical Analysis

This appendix provides theoretical analysis and derivation details to supplement the methodology section. Please note that the meanings of some symbols differ from those in the main text.

E.1 Component I - Data Volume Estimation Module

We provide a rigorous theoretical analysis of our module from the perspectives of both privacy and utility.

E.1.1 Privacy Guarantee Analysis. We begin with the formal definition of Local Differential Privacy (LDP).

Definition 1 (ϵ-Local Differential Privacy). *A randomized algorithm $\mathcal{A}$ satisfies ϵ-LDP if for any two inputs x, x' in its domain, and for any subset of outputs O in its range, the following inequality holds:*

$$\Pr[\mathcal{A}(x) \in O] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(x') \in O] \tag{28}$$

Based on this definition, we can prove that our proposed module satisfies LDP in every round.

Theorem 1 (Privacy Guarantee of the Module). *Our proposed data size estimation mechanism ensures that the estimate n'_j reported by client j in any round t satisfies ϵ_d-LDP.*

Proof. Let the entire client-side procedure be represented by a randomized algorithm $\mathcal{A}$, which takes the true data size n_j as input and produces the reported estimate n'_j as output. We analyze the privacy properties of this algorithm.

Sensitivity Analysis: The first step of the algorithm is to clip the input n_j (Eq. (??)). For any two possible inputs n_j and n^*_j, the L1 distance between their clipped counterparts, $n_{j,\text{clip}}$ and $n^*_{j,\text{clip}}$, defines the global sensitivity Δf:

$$\Delta f = \sup_{n_j, n^*_j} |n_{j,\text{clip}} - n^*_{j,\text{clip}}| \leq (c_d - 1) - 0 = c_d - 1 \tag{29}$$

We use its upper bound, c_d, as the global sensitivity.

Laplace Mechanism Analysis: The core privacy-preserving step is the addition of Laplace noise (Eq. (??)). By the standard definition of the Laplace mechanism, when the noise scale parameter is set to $b_d = \Delta f/\epsilon_d = c_d/\epsilon_d$, the output $noisy\_d^t_j$ satisfies ϵ_d-differential privacy with respect to its input $n_{j,\text{clip}}$. For any two inputs $n_{j,\text{clip}}$ and $n^*_{j,\text{clip}}$, and any possible output value v, the ratio of their probability density functions is:

$$\begin{aligned} \frac{p(v|n_{j,\text{clip}})}{p(v|n^*_{j,\text{clip}})} &= \frac{\frac{1}{2b_d}\exp(-\frac{|v-n_{j,\text{clip}}|}{b_d})}{\frac{1}{2b_d}\exp(-\frac{|v-n^*_{j,\text{clip}}|}{b_d})} \\ &= \exp\left(\frac{|v-n^*_{j,\text{clip}}| - |v-n_{j,\text{clip}}|}{b_d}\right) \end{aligned} \tag{30}$$

By the reverse triangle inequality, $|v-n^*_{j,\text{clip}}| - |v-n_{j,\text{clip}}| \leq |n_{j,\text{clip}} - n^*_{j,\text{clip}}| \leq c_d$. Therefore:

$$\frac{p(v|n_{j,\text{clip}})}{p(v|n^*_{j,\text{clip}})} \leq \exp\left(\frac{|n_{j,\text{clip}} - n^*_{j,\text{clip}}|}{b_d}\right) \leq \exp\left(\frac{c_d}{c_d/\epsilon_d}\right) = e^{\epsilon_d} \tag{31}$$

This demonstrates that the generation of $noisy\_d^t_j$ satisfies ϵ_d-DP.

Immunity to Post-processing: A fundamental property of differential privacy is its immunity to post-processing. In our algorithm, the EMA smoothing step (Eq. (??)) and the assignment of the result to n'_j are computations performed on the value $noisy\_d^t_j$, which already satisfies ϵ_d-DP. Since these subsequent operations do not depend on the original private data n_j, they do not compromise the existing privacy guarantee.

□

In summary, the entire client-side reporting procedure satisfies ϵ_d-LDP.

Remark 1 (Cumulative Privacy Budget). *It is important to note that ϵ_d represents the privacy budget for a single round. If a client participates in T consecutive rounds, its total privacy cost, according to the principle of Sequential Composition in differential privacy, is bounded by $T \cdot \epsilon_d$. In practice, a sufficiently small ϵ_d can be chosen to ensure that the cumulative privacy leakage over the entire FL process remains within an acceptable range.*

E.1.2 Utility Analysis: Unbiasedness and Order-Preserving Property. We demonstrate the utility of our estimator by proving that its expectation is unbiased.

Theorem 2 (Asymptotic Unbiasedness of the Estimator). *Assuming the smoothing factor $\gamma_1 \in (0, 1)$ and that the client's true clipped data volume $n_{j,\text{clip}}$ remains stable, the mathematical expectation of the smoothed estimate reported by client j, $\mathbb{E}[s^t_j]$, asymptotically converges to its true clipped data volume $n_{j,\text{clip}}$:*

$$\lim_{t\to\infty} \mathbb{E}[s^t_j] = n_{j,\text{clip}} \tag{32}$$

Proof. We begin by analyzing the expectation of the single-round noisy observation $noisy\_d^t_j$. The expectation of the Laplace distribution $\text{Laplace}(0, b_d)$ is zero. Therefore, by the linearity of expectation:

$$\begin{aligned} \mathbb{E}[noisy\_d^t_j] &= \mathbb{E}[n_{j,\text{clip}} + \text{noise}^t_j] \\ &= \mathbb{E}[n_{j,\text{clip}}] + \mathbb{E}[\text{noise}^t_j] \\ &= n_{j,\text{clip}} + 0 \\ &= n_{j,\text{clip}} \end{aligned} \tag{33}$$

This shows that the noisy observation in each round is, by itself, an unbiased estimator of the true clipped value.

Next, we analyze the expectation of the smoothed value s^t_j. Applying the linearity of expectation to the equation:

$$\begin{aligned} \mathbb{E}[s^t_j] &= \mathbb{E}[\gamma_1 \cdot s^{t-1}_j + (1-\gamma_1) \cdot noisy\_d^t_j] \\ &= \gamma_1 \cdot \mathbb{E}[s^{t-1}_j] + (1-\gamma_1) \cdot \mathbb{E}[noisy\_d^t_j] \\ &= \gamma_1 \cdot \mathbb{E}[s^{t-1}_j] + (1-\gamma_1) \cdot n_{j,\text{clip}} \end{aligned} \tag{34}$$

This is a first-order linear recurrence relation for $\mathbb{E}[s^t_j]$. To find its general solution, we unroll the recurrence, using the initial

condition $E[s_j^0] = E[0] = 0$:

$$\begin{aligned}\mathbb{E}[s_j^t] &= \gamma_1 \mathbb{E}[s_j^{t-1}] + (1-\gamma_1) n_{j,\text{clip}} \\ &= \gamma_1 \left(\gamma_1 \mathbb{E}[s_j^{t-2}] + (1-\gamma_1) n_{j,\text{clip}}\right) + (1-\gamma_1) n_{j,\text{clip}} \\ &= \gamma_1^2 \mathbb{E}[s_j^{t-2}] + \gamma_1(1-\gamma_1) n_{j,\text{clip}} + (1-\gamma_1) n_{j,\text{clip}} \\ &= \dots \\ &= \gamma_1^t \mathbb{E}[s_j^0] + \left(\sum_{i=0}^{t-1} \gamma_1^i\right)(1-\gamma_1) n_{j,\text{clip}} \\ &= \gamma_1^t \cdot 0 + \left(\frac{1-\gamma_1^t}{1-\gamma_1}\right)(1-\gamma_1) n_{j,\text{clip}} \\ &= (1-\gamma_1^t) n_{j,\text{clip}}\end{aligned} \tag{35}$$

We now take the limit of the above expression as $t \to \infty$. Since the smoothing factor $\gamma_1 \in (0,1)$, it follows that $\lim_{t\to\infty} \gamma_1^t = 0$. Therefore:

$$\lim_{t\to\infty} \mathbb{E}[s_j^t] = \lim_{t\to\infty} (1-\gamma_1^t) n_{j,\text{clip}} = (1-0)\cdot n_{j,\text{clip}} = n_{j,\text{clip}} \tag{36}$$

This result proves that the expectation of the smoothed estimate asymptotically converges to the true clipped data volume.

□

Remark 2 (Statistical Order-Preserving Property). *The Theorem above serves as the theoretical cornerstone for the utility of our module. It reveals that although random noise is introduced in every round, the system effectively filters out the mean effect of this noise through temporal smoothing. Consequently, the set of estimates reported by the clients, $\{n'_j\}_{j\in\mathcal{M}}$, will, in statistical expectation, faithfully reflect the true ranking and relative magnitudes of each client's clipped data volume. This ensures that the final score vector $\mathbf{S}^t_{\text{data}}$ computed by the server is a meaningful and robust input, providing a solid foundation for subsequent FL algorithms that rely on client contribution assessment.*

E.2 Theoretical Analysis of Component I - Contribution Estimation Module

E.2.1 Privacy Guarantee Analysis. We first present the relevant theorem on the composition of differential privacy, and then leverage it to prove the privacy guarantee of our module.

Definition 2 (Basic Composition Theorem). *Let there be a series of privacy mechanisms $\mathcal{B}_1, \dots, \mathcal{B}_k$, where each $\mathcal{B}_i$ provides ϵ_i-DP. The sequence of their outputs, $(\mathcal{B}_1(D), \dots, \mathcal{B}_k(D))$, when applied to the same input database D, satisfies $(\sum_{i=1}^k \epsilon_i)$-DP.*

Theorem 3 (Privacy Guarantee of the Contribution Module). *The contribution estimate v_j reported by client j in any round t satisfies $\epsilon_{\text{contrib}}$-LDP.*

Proof. We consider the entire client-side reporting procedure as a randomized algorithm $\mathcal{B}$, with the client's private dataset D_j as input and the public value v_j as output. This algorithm $\mathcal{B}$ can be decomposed into two independent sanitization sub-mechanisms and a post-processing step.

(1) **Privacy of Sub-mechanism $\mathcal{B}_l$**: This sub-mechanism processes the initial loss. Its input is $l_j^t = \mathcal{L}(w^t_{\text{start},j}; D_j)$. The clipping operation in Eq. (5) bounds its L1 sensitivity to c_l. Subsequently, Eq. (6) adds Laplace noise with a scale of $b_l = c_l/\epsilon_l$. By definition of the Laplace mechanism, the output of sub-mechanism $\mathcal{B}_l$, $noisy\_l_j$, satisfies ϵ_l-LDP.

(2) **Privacy of Sub-mechanism $\mathcal{B}_g$**: This sub-mechanism processes the gradient norm. Its input is $g_j^t = \nabla\mathcal{L}(w^t_{\text{start},j}; D_j)$. The norm clipping in Eq. (5) bounds its L2 sensitivity to c_g. The Laplace noise added in Eq. (6) has a scale of $b_g = c_g/\epsilon_g$. This operation ensures that the output of sub-mechanism $\mathcal{B}_g$, $noisy\_g_j$, satisfies ϵ_g-LDP.

(3) **Applying Composition**: The client effectively reports a tuple $(noisy\_l_j, noisy\_g_j)$ to the server, as the final value v_j is computed from these two components. Since both $noisy\_l_j$ and $noisy\_g_j$ are derived from the same private dataset D_j, we must apply the Basic Composition Theorem from **Definition 2**. Viewing $\mathcal{B}_l$ and $\mathcal{B}_g$ as two privacy mechanisms acting on the same data D_j, the composite mechanism that releases both of their results satisfies $(\epsilon_l + \epsilon_g)$-LDP.

(4) **Post-processing Invariance**: The final reported value v_j is the result of a deterministic linear transformation (Eq. (7)) on the private tuple $(noisy\_l_j, noisy\_g_j)$. A fundamental property of differential privacy is its immunity to post-processing, which states that performing any computation on the output of a DP mechanism, without accessing the original private data, does not weaken the privacy guarantee. Therefore, the computation of v_j inherits the privacy level of its input tuple.

In conclusion, by combining steps 3 and 4, the entire algorithm $\mathcal{B}$ satisfies $(\epsilon_l + \epsilon_g)$-LDP. As per our hyperparameter setting $\epsilon_{\text{contrib}} = \epsilon_l + \epsilon_g$, the output v_j of our module strictly satisfies $\epsilon_{\text{contrib}}$-LDP. □

E.2.2 Utility and Rationale of Contribution Proxies. The utility of this module is founded on the rationale that the two chosen proxy metrics can reasonably represent a client's potential contribution. We articulate this rationale through the following propositions.

Proposition 1 (Gradient Norm as a Proxy for Potential Contribution). *The gradient norm, $\|g_j^t\|_2 = \|\nabla\mathcal{L}(w^t_{\text{start},j}; D_j)\|_2$, quantifies the local rate of change of the loss function, defined by the client's data D_j, at the current model parameter point. A large gradient norm implies that the loss surface is steep at that point, which intuitively suggests a significant misalignment between the current model $w^t_{\text{start},j}$ and the optimal parameters for client j's data. Therefore, the client's data holds substantial information for correcting the global model, indicating a high potential contribution.*

Proposition 2 (Initial Loss as a Proxy for Improvement Headroom). *The initial loss, $l_j^t = \mathcal{L}(w^t_{\text{start},j}; D_j)$, directly quantifies the performance of the current model on the client's local data. A high initial loss value indicates poor performance, which is synonymous with a large headroom for improvement. Consequently, such a client has the potential to achieve a more significant loss reduction during its local training, thereby making a greater tangible contribution to the global model's performance enhancement.*

Remark 3 (Holistic Contribution Metric). *Relying on either proxy alone could lead to sub-optimal selections. For instance, anomalous data could produce an extremely large gradient norm,*

while low-quality data might also result in a very high initial loss. Our module, by taking a weighted sum of the two noisy proxies, $v_j = \eta_l \cdot noisy\_l_j + \eta_g \cdot noisy\_g_j$, aims to strike a balance. It seeks to identify clients that not only present a discrepancy with the current model (proxied by the gradient norm) but also on which the model can effectively learn from this discrepancy to achieve significant performance gains (proxied by the initial loss). The weighting hyperparameters η_l and η_g provide the necessary flexibility to tune this trade-off between the two dimensions.

E.3 Component II - Hedonic Game-Based Stable Coalition Formation

The core goal of this section is to prove that our proposed online estimators $\widehat{\mu_e}$ and $\widehat{\sigma^2}$ are effective approximations of the prior parameters μ_e and σ^2 in Donahue K. theory, thereby proving the rationality of the dynamic game threshold $\hat{T}$.

E.3.1 The validity of $\widehat{\mu_e}$.

Assumption 1 (Theoretical Basis and Error Decomposition). *According to the work of Donahue K. et al., there is a set of $\mathcal{M}$. The real data of each client j is determined by an unknown D-dimensional parameter vector θ_j^* and a noise variance $\sigma_{\epsilon,j}^2$. These parameters are drawn independently and identically (IID) from a higher-level prior distribution. We define the following two core theoretical prior parameters, the Mean Irreducible Error and Inherent Heterogeneity:*

$$\mu_e := \mathbb{E}_j[\sigma_{\epsilon,j}^2], \quad \sigma^2 := \mathrm{Var}_j(\theta_j^*). \tag{37}$$

Assumption 2 (Local Model Convergence). *We assume that in an effective FL process, for any client j, its local model θ_j^t obtained after t rounds of training will converge probabilistically or mean-squared to the optimal model θ_j^* on its local data distribution D_j. This assumption means that as training progresses, the model will fully learn generalizable patterns in the local data. We assume that $\lim_{t\to\infty} \theta_j^t = \theta_j^*$ is a reasonable simplification under ideal convergence conditions, which captures the core dynamics of the local model constantly approaching the optimal representation of its data distribution.*

Theorem 4 (Asymptotically Unbiased Estimation). *Under the condition that Assumption 2 holds, our estimator $\widehat{\mu_e}^t$ is an asymptotically unbiased estimator of the theoretical parameter μ_e, that is:*

$$\lim_{t\to\infty} \mathbb{E}[\widehat{\mu_e}^t] = \mu_e \tag{38}$$

Proof. According to the setting of Donahue K. et al., μ_e is the expected irreducible error of the system average. For any client j, the expectation of its loss function $\mathbb{E}[L_j]$ can be decomposed. Let f_j^* be the best possible function (i.e., Bayesian optimal classifier) on the data distribution D_j of client j, θ_j^t be the model parameters after t rounds of training, and its corresponding function is $f(\theta_j^t, \cdot)$. Then the expected loss of t rounds is:

$$\begin{aligned}
\mathbb{E}[L_j^t] &= \mathbb{E}_{(x,y)\sim D_j}\left[\left(f(\theta_j^t; x) - y\right)^2\right] \\
&= \mathbb{E}_{(x,y)\sim D_j}\left[\left(f(\theta_j^t; x) - f_j^*(x) + f_j^*(x) - y\right)^2\right] \\
&= \mathbb{E}_x\left[\left(f(\theta_j^t; x) - f_j^*(x)\right)^2\right] + \mathbb{E}_{(x,y)}\left[\left(f_j^*(x) - y\right)^2\right] \\
&= \mathrm{Error}_{\mathrm{reducible}}(\theta_j^t) + \mathrm{Error}_{\mathrm{irreducible},j},
\end{aligned} \tag{39}$$

where the cross term $\mathbb{E}[(f(\theta_j^t, x) - f_j^*(x))(f_j^*(x) - y)]$ is zero since $f_j^*(x)$ is the conditional expectation $\mathbb{E}[y|x]$ given x. $\mathrm{Error}_{\mathrm{irreducible},j} = \mathbb{E}[(f_j^*(x) - y)^2] = \sigma_{\epsilon,j}^2$ is the noise variance inherent to client j, and the theoretical parameter $\mu_e = \mathbb{E}_j[\sigma_{\epsilon,j}^2]$.

Assumption 3 (Convergence of Estimators - 1). *We assume that in an effective FL process, the model parameters θ_j^t of client j converge to their local optimal solution θ_j^*, so that the reducible error converges to a minimum value, ideally zero:*

$$\lim_{t\to\infty} \mathrm{Error}_{\mathrm{reducible}}(\theta_j^t) = \lim_{t\to\infty} \mathbb{E}[(f_j(\theta_j^t, x) - f_j^*(x))^2] = 0. \tag{40}$$

This means that the model has been fully learned in non-convex optimization, and the error caused by model inaccuracy cannot be significantly reduced through further training.

Based on this assumption, we can have:

Corollary 1. *Substituting this limit into the loss decomposition, we obtain that the expectation of the local loss will converge to the local irreducible error in the later stages of training:*

$$\lim_{t\to\infty} \mathbb{E}[L_j^t] = \lim_{t\to\infty} \left(\mathrm{Error}_{\mathrm{reducible}}(\theta_j^t) + \sigma_{\epsilon,j}^2\right) = \sigma_{\epsilon,j}^2. \tag{41}$$

The instantaneous observation value $\mu_{e,\mathrm{obs}}^t$ we calculated in round t is the mean of L_j^{t-1} on the current participating client set $\mathcal{N}^{t-1}$:

$$\mu_{e,\mathrm{obs}}^t = \frac{1}{|\mathcal{N}^{t-1}|} \sum_{j\in\mathcal{N}^{t-1}} L_j^{t-1}. \tag{42}$$

If we assume that the client sampling process is unbiased, that is, $E_{\mathcal{N}^{t-1}}[\cdot] = \mathbb{E}_j[\cdot]$, then its expectation is:

$$\mathbb{E}[u_{e,\mathrm{obs}}^t] = \mathbb{E}\left[\frac{1}{|\mathcal{N}^{t-1}|} \sum_{j\in\mathcal{N}^{t-1}} L_j^{t-1}\right] = \mathbb{E}_j\left[L_j^{t-1}\right]. \tag{43}$$

Therefore, when $t \to \infty$:

$$\lim_{t\to\infty} \mathbb{E}[u_{e,\mathrm{obs}}^t] = \mathbb{E}_j\left[\lim_{t-1\to\infty} \mathbb{E}[L_j^{t-1}]\right] = \mathbb{E}_j[\sigma_{\epsilon,j}^2] = \mu_e. \tag{44}$$

The EMA update mechanism $\widehat{\mu_e}^t = \gamma_2 \cdot \widehat{\mu_e}^{t-1} + (1-\gamma_2) \cdot \mu_{e,\mathrm{obs}}^t$ can be expanded into a weighted sum of all historical observations:

$$\hat{\mu}_e^t = (1-\gamma_2) \sum_{i=0}^{t-1} \gamma_2^i \mu_{e,\mathrm{obs}}^{t-i} + \gamma_2^t \hat{\mu}_e^0. \tag{45}$$

When $t \to \infty$, $\widehat{\mu_e}^t$ will converge to the expected value of $\mu_{e,\mathrm{obs}}^t$. In the steady state, $\mathbb{E}[\widehat{\mu_e}^t] \approx \mathbb{E}[\mu_{e,\mathrm{obs}}^t]$. Therefore, our final estimator $\widehat{\mu_e}$ is an asymptotically unbiased and consistent estimate of the theoretical parameter μ_e, and its validity is guaranteed. □

E.3.2 The validity of $\widehat{\sigma^2}$.

Theorem 5 (Theoretical Basis and Estimator). *Under the condition that Assumption 2 holds, estimator $\widehat{\sigma^2}^t$ is an asymptotically unbiased estimator of the theoretical parameter σ^2, that is, $\lim_{t\to\infty}\mathbb{E}[\widehat{\sigma^2}^t]=\sigma^2$.*

Proof. The theoretical parameter σ^2 is the variance between the optimal local models θ_j^*:

$$\sigma^2=\mathrm{Var}_j(\theta_j^*)=\mathbb{E}_j[\|\theta_j^*-\mathbb{E}_j[\theta_j^*]\|^2]. \tag{46}$$

Instantaneous observation $\sigma^2_{\mathrm{obs},t}$ is the unbiased sample variance of the model θ_j^{t-1} in round $t-1$, and its expectation is:

$$\mathbb{E}[\sigma^2_{\mathrm{obs},t}]=\mathbb{E}\left[\frac{1}{|\mathcal{N}^{t-1}|-1}\sum_{j\in\mathcal{N}^{t-1}}\left\|\theta_j^{t-1}-\bar{\theta}^{t-1}_{\mathrm{obs}}\right\|^2\right]=\mathrm{Var}_j(\theta_j^{t-1}) \tag{47}$$

Assumption 4 (Convergence of Estimators - 2). *According to Assumption 2, each local model θ_j^t converges to its own ideal model θ_j^*:*

$$\lim_{t\to\infty}\theta_j^t=\theta_j^*. \tag{48}$$

Corollary 2 (Convergent Transitivity). *Variance $\mathrm{Var}(\cdot)$ is a continuous functional. According to the Continuous Mapping Theorem, the convergence of the variable can be transferred to its continuous function. According to Assumption 1, the definition of $\mathrm{Var}_j(\theta_j^*)$ is σ^2. Therefore:*

$$\lim_{t\to\infty}\mathrm{Var}_j(\theta_j^{t-1})=\mathrm{Var}_j\left(\lim_{t\to\infty}\theta_j^{t-1}\right)=\mathrm{Var}_j(\theta_j^*)=\sigma^2 \tag{49}$$

In Sec 4.1, our instantaneous heterogeneity observation $\sigma^2_{\mathrm{obs},t}$ is:

$$\sigma^2_{\mathrm{obs},t}=\frac{1}{|\mathcal{N}^{t-1}|-1}\sum_{j\in\mathcal{N}^{t-1}}\left\|\theta_j^{t-1}-\bar{\theta}^{t-1}_{\mathrm{obs}}\right\|^2. \tag{50}$$

This is a standard unbiased sample variance estimator, whose expectation is equal to the population variance:

$$\mathbb{E}[\sigma^2_{\mathrm{obs},t}]=\mathrm{Var}_j(\theta_j^{t-1}), \tag{51}$$

assuming client sampling is unbiased.

Therefore, when t is large enough:

$$\lim_{t\to\infty}\mathbb{E}[\sigma^2_{\mathrm{obs},t}]=\lim_{t-1\to\infty}\mathrm{Var}_j(\theta_j^{t-1})=\sigma^2. \tag{52}$$

Similarly, EMA estimator

$$\widehat{\sigma^2}^t=\gamma_2\cdot\widehat{\sigma^2}^{t-1}+(1-\gamma_2)\cdot\sigma^2_{\mathrm{obs},t} \tag{53}$$

will converge to the expected value of $\sigma^2_{\mathrm{obs},t}$.

Therefore:

$$\lim_{t\to\infty}\widehat{\sigma^2}^t=\lim_{t\to\infty}\mathbb{E}[\sigma^2_{\mathrm{obs},t}]=\sigma^2, \tag{54}$$

and this proves that $\widehat{\sigma^2}$ is an asymptotically unbiased estimator of the theoretical parameter σ^2.

□

Remark 4 (Warm-up and EMA Conditions). *The above proofs all rely on the limit condition of $t\to\infty$. In the early stages of limited training, the convergence assumption does not actually hold, and L_j^t and θ_j^t contain huge noise and deviations.*

The above proves the asymptotic properties of the estimator. In a finite number of training rounds, the initial estimates $\widehat{\mu_e}^0$ and $\widehat{\sigma^2}^0$ obtained in the warm-up phase by batch averaging T_{warmup} rounds are, according to the law of large numbers, closer to their expected initial estimates than any single-round instantaneous values. This provides a robust "anchor point" for EMA, preventing it from being "biased" by extreme noise values in the early stages, thereby accelerating the convergence of the estimator to its true value. In EMA smoothing, μ_e and σ^2 are static properties that describe the data distribution, while $\mu^t_{e,\mathrm{obs}}$ and $\sigma^2_{\mathrm{obs},t}$ are dynamic observations affected by client sampling and training randomness. Through the smoothing factor γ_2, the system controls the trade-off between the estimated bias and variance, ensuring that throughout the dynamic process, $\hat{T}^t=\widehat{\mu_e}^t/\widehat{\sigma^2}^t$ is a stable proxy that gradually approaches the true value.

The mathematical logic of Sec 4.2 is rigorous. It transforms the two theoretical priors (μ_e,σ^2) that are not directly observable into empirical quantities $(\widehat{\mu_e}^t,\widehat{\sigma^2}^t)$ that can be dynamically estimated through L_j^t and θ_j^t in the FL process. Its effectiveness is guaranteed by the basic convergence assumptions of machine learning and statistical estimation theory.

E.4 Component III - Multi-Granularity Aggregation with Soft Update

This section aims to prove that the aggregation strategy we designed for the coalitions is reasonable and effective, and the soft update mechanism is a necessary supplement to ensure the long-term convergence of the system. We omit the Uniform and Coarse-grained here, as related work has provided sufficient proof already.

E.4.1 Theoretical Basis of Fine-grained Aggregation Strategy. The optimal weight v_{ji} of its Fine-grained federation is derived by minimizing the expected MSE of player j. The solution (as shown in Lemma 7.1 of Donahue K.) is of a complex form and depends on μ_e, σ^2, and n_i of all clients. The underlying logic is that client j should give higher weights to clients whose models θ_i are "similar" to its own θ_j, in order to minimize the introduced bias. We demonstrate that our Fine-grained aggregation strategy is more adaptable to dynamic systems in the following proof.

Corollary 3 (Gradient Direction as Proxy for Data Distribution). *The direction of the model update vector $\Delta\theta_j$ is a valid approximation of the average gradient direction $\mathbb{E}_{x\sim D_j}[\nabla_\theta L_j(\theta)]$ of the local data distribution D_j of client j at the current model point $w_{\mathrm{start},j}$. Therefore, the cosine similarity between update vectors is an effective measure of local task similarity.*

Proof. For E-step local updates (learning rate is η), $\Delta\theta_j^t$ can be written as:

$$\Delta\theta_j^t=\theta_j^t-w^t_{start,j}=\sum_{e=1}^{E}\Delta\theta^t_{j,e}, \tag{55}$$

where $\Delta\theta^t_{j,e}$ is the update at step e. For SGD,

$$\Delta\theta^t_{j,e}=-\eta\cdot g_j(\theta^t_{j,e-1}), \tag{56}$$

where g_j is the gradient over a mini-batch. When η is small enough and E is not too large, θ does not change much during the local update, and $\theta^t_{j,e-1}\approx w^t_{\mathrm{start},j}$.

Therefore, in FL, the model update $\Delta\theta_j = \theta_j^t - w_{\text{start},j}^t$ can be viewed as an approximation of the average gradient direction driven by client j's local data D_j in the model parameter space. Therefore, $\text{sim}(\Delta\theta_i, \Delta\theta_j)$ directly measures the consistency of the goals of the two client local learning tasks:

$$\Delta\theta_j^t \approx \sum_{e=1}^{E} -\eta \cdot \mathbb{E}_{batch_e \sim D_j}[g_j(w_{start,j}^t)] = -E \cdot \eta \cdot \mathbb{E}_{x \sim D_j}[g_j(w_{start,j}^t)]. \tag{57}$$

This approximation shows that the direction of $\Delta\theta_j^t$ is approximately collinear with the local expected gradient direction $\nabla_\theta L_j(w_{\text{start},j}^t)$. Since the expected gradient is uniquely determined by the data distribution D_j, $\text{sim}(\Delta\theta_i, \Delta\theta_j)$ is an approximation of $\text{sim}(\nabla L_i, \nabla L_j)$, and is therefore a reasonable measure of the similarity of the tasks of D_i and D_j at the current point.

Because our weight calculation formula is:

$$\text{vec}_{j,k} = (1 + \Gamma_{k,j}^t) / \sum_{i \in C_f^t} (1 + \Gamma_{i,j}^t), \tag{58}$$

where $\Gamma_{k,j}^t = \text{sim}(\Delta_k^t, \Delta_j^t)$.

If the update direction of client k is highly consistent with that of j ($\Gamma_{k,j}^t \to 1$), $vec_{j,k}$ gets a high weight. This is completely consistent with the spirit of "cooperating with similar parties" in Donahue K.'s theory. If the update direction of client k is orthogonal or opposite to that of j ($\Gamma_{k,j}^t \le 0$), $vec_{j,k}$ gets a low weight. This also achieves the goal of "avoiding cooperation with those who conflict with tasks". □

Based on the above, our Fine-grained aggregation strategy reshapes Donahue K.'s theory in terms of implementation by using the cosine similarity of model updates as a similarity metric. It is not only consistent with the original work in terms of logical goals, but also has more advantages in mathematics: 1. Adaptability: No need to rely on global, static prior parameters μ_e, σ^2, but adaptively adjust according to the dynamic learning behavior $\Delta\theta$ of each round. 2. Robustness: Directly operating "knowledge increment" rather than "model state" can better handle the situation where model parameters vary greatly in non-IID scenarios, avoiding catastrophic forgetting or parameter conflicts.

Remark 5 (Advantages of Weighting Property). *Donahue K.'s theory is based on the similarity of model parameters θ_i, which can be misleading in non-IID settings (e.g., two models with similar functions may have parameters far apart). Our method is based on gradient direction and focuses on the consistency of learning objectives, which is a more robust similarity measure in the non-convex optimization landscape.*

E.4.2 Effectiveness of Soft Update Mechanism. Since the theoretical basis of Donahue K. is limited and based on a single-round static game, it does not involve the evolution path of the model in multiple rounds of iterations. Therefore, we introduce a soft update mechanism after multi-granularity aggregation.

Theorem 6 (Bounded Divergence via Soft Update). *In fully personalized FL, introducing a global reference model w_{ref} as the starting point for training some clients can provide an upper bound for the variance $\text{Var}_j(\theta_j^t)$ between client models, thereby preventing the system from diverging due to knowledge drift.*

Proof. **Divergence risk of unconstrained systems**: In a fully personalized FL framework, without any global constraints, the model evolution of each client j: $\theta_j^{t+1} = \text{LocalTrain}_j(\theta_j^t)$ is an independent Markov chain. Under non-IID, the steady-state distributions of these chains may be different, resulting in $\lim_{t\to\infty} \text{Var}_j(\theta_j^t)$ being non-zero or even divergent, causing the entire system to fail to converge to a good performance level. This can be formalized as:

$$\lim_{t\to\infty} \text{Var}_j(\|\theta_j^t - \bar{\theta}^t\|_2^2) \to \infty \text{ or const} > 0, \tag{59}$$

where $\bar{\theta}^t$ is form of global average model.

Centroid property of w_{ref}^t: According to the definition of w_{ref}^t, $w_{\text{ref}}^t = E_{j \in \mathcal{N}^t, n_j'}[\theta_j^t]$ (before aggregation). This shows that w_{ref}^t is the weighted centroid of the current active client group in the parameter space. It is defined as Eq. (23) where:

$$p_i^t = (\sum_{j \in C_i^t} n_j') / (\sum_{k \in \mathcal{N}^t} n_k'), \tag{60}$$

$$\theta_{i,\text{rep}}^t = \frac{\sum_{j \in C_i^t} n_j' \theta_j^t}{\sum_{j \in C_i^t} n_j'}. \tag{61}$$

Regularization Effect of Soft Update: The global reference model w_{ref}^t we introduced is designed to solve this problem. The soft update mechanism sets $w_{\text{start},k}^{t+1} = w_{\text{ref}}^t$ for a portion of clients k (especially those that have just joined or have not participated for a long time). This is equivalent to "pulling back" the models of these clients to the knowledge center of round t before the start of round $t+1$.

We can regard the evolution of the entire system as a controlled random process, as a variant of a conditional reset or proximal update. At each time step t, the soft update operation effectively reduces the model variance. Let $V^t = \text{Var}_j(\theta_j^t)$. At the beginning of $t+1$ round, due to soft update, $\mathbb{E}[V_{\text{start}}^{t+1}] < V^t$. Although local training will increase the variance again, the periodic pullback operation ensures that the variance does not grow unbounded.

More formally, we can model this process as $V^{t+1} \le (1-\alpha)V^t + C$, where $\alpha > 0$ is the variance reduction rate brought by soft update and C is the variance increase introduced by local training. As long as C is bounded (which holds under the bounded gradient assumption), this iterative process ensures that V^t will eventually converge to a bounded steady-state value, thus preventing the system from diverging. □

Based on the above, the multi-granularity aggregation strategy proposed by our method is a rigorous inheritance and optimization of Donahue K.'s theory, especially the similarity measure in the Fine-grained strategy is more robust mathematically. The soft update mechanism fundamentally solves the convergence shortcomings of static theory in dynamic applications, providing a solid theoretical guarantee for the stable operation of our entire framework.