PROGRESS-GUIDED CLIENT PARTICIPATION IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication is a primary bottleneck in federated learning (FL), yet standard pipelines typically fix the number of participating clients per round. A static participation budget may incur unnecessary communication when additional participation offers limited optimization benefit and, conversely, fail to capture sufficient client diversity under strong data heterogeneity. We propose *Intelligent Selection of Participants* (ISP), an adaptive mechanism that treats the per-round client count as a controllable optimization variable and adjusts it online to balance communication and optimization progress. ISP uses a lightweight client-count search and complements methods that select *which* clients participate. Across public vision and NLP benchmarks and a large-scale real-world ECG dataset, ISP reduces overall communication by up to 30% while maintaining the downstream performance of fixed-budget baselines. To assess generality, we evaluate ISP across a wide range of heterogeneity levels and system settings, demonstrating robust adaptation under noisy per-round estimates. We also provide a convergence analysis of the adaptive client-count rule underlying ISP.

1 INTRODUCTION

Federated learning (FL) allows decentralized data owners to jointly train a model without sharing raw data (Kairouz et al., 2021; Konečný et al., 2016; Li et al., 2020a). Communication efficiency is often its primary bottleneck: clients must upload model updates over bandwidth-constrained channels. Under uniform sampling, using more clients reduces the variance of the aggregate gradient but increases per-round communication overhead. Conversely, selecting too few clients can worsen convergence and harm generalization.

While client-selection strategies have been extensively studied, the number of clients per training round remains underexplored. In widely used baselines such as FEDAVG (McMahan et al., 2017) and FEDPROX (Li et al., 2020b), the per-round participation budget is fixed, leaving an important degree of freedom unexploited. Our insight is that the participant count should be treated as a dynamic control variable that adapts to the training process, balancing communication cost against convergence speed. We propose *Intelligent Selection of Participants* (ISP), a lightweight mechanism that adjusts this count online.

MAIN CONTRIBUTIONS.

We summarize the key contributions of this work as follows:

- **Adaptive participant-count selection.** We formalize the round-wise choice of the number of participating clients as a full-pool constrained optimization problem. Based on a tractable finite-sample approximation of this formulation, we propose *Intelligent Selection of Participants* (ISP), which adapts the participation budget as training evolves. Since ISP controls how many clients participate, it complements methods for client selection, local optimization, or update compression.
- **Empirical verification.** We conduct extensive experiments under heterogeneous settings across three domains: image classification, text modeling, and large-scale real-world ECG time-series classification. The evaluation covers established client selections and federated optimization algorithms, system-constrained settings, and gradient compression. ISP achieves up to a 30% reduc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion in client-update exchanges for most baselines while maintaining downstream performance and reducing overall training time, especially in large-scale regimes. We further provide comprehensive ablations of the main ISP components, confirming the robustness of its design choices.
- **Theoretical analysis.** We bound the response of the count returned by the finite-sample search, jointly accounting for auxiliary-pool and Monte Carlo errors, and quantify how this bound changes while the count is reused. Additional stationarity results and exact constructions identify regimes with sustained progress and controlled auxiliary-update cost. The analysis also separates the mean distortion and response variance introduced by general client selectors.

2 Related Work

Federated optimization. Federated optimization alternates local training with server aggregation; subsequent developments address heterogeneity by constraining or correcting local updates, normalizing their contributions, or adapting server updates (McMahan et al., 2017; Li et al., 2020b; Karimireddy et al., 2020; Wang et al., 2020; Reddi et al., 2020). These mechanisms improve how participating clients contribute to learning and can be combined with adaptive participant counts.

Client selection. Client-selection strategies use local losses, gradient information, or label distributions to identify informative participants (Cho et al., 2020; Tang et al., 2022; Wang et al., 2023; Zhang et al., 2023). System-aware approaches additionally account for client availability and execution speed (Lai et al., 2021; Li et al., 2022). These strategies primarily determine which clients participate for a given budget, leaving the complementary question of how many clients to select. Appendix B reviews these approaches and describes the selectors used in our experiments.

Client-Count Selection. There is, indeed, a small number of papers that explored dynamic control over the number of clients participating in a communication round. The authors of (de Souza et al., 2024) proposed an exponential decay schedule to gradually reduce the number of clients. This rule varies participation over time but does not react to the observed optimization state, which may be particularly limiting under statistical heterogeneity. Other studies, such as (Chen et al., 2022), examined how the number of clients influences fine-tuning of large transformer models in federated settings. Recent works also study adaptive participation in specialized settings, including bandwidth-constrained LoRA-based LLM adaptation (Solat & Lee, 2025) and vehicular FL (Lin & Zhou, 2026). Most relevant is Li et al. (2024), who directly addressed round-wise client count selection through AdaFL. However, AdaFL uses a predefined linear schedule rather than responding to the observed training dynamics.

Taken together, these approaches leave the client count governed by externally specified rules instead of an explicit progress-based criterion. We instead consider a progress-aware formulation that selects the client count using an estimated next-round loss response.

3 Method

3.1 Problem Setup

We define the reference global objective as

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \; f(\boldsymbol{x}) := \frac{1}{M}\sum_{i=1}^{M} f_i(\boldsymbol{x}), \tag{1}$$

where $f_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_i\sim\mathcal{D}_i}[f_i(\boldsymbol{x}, \boldsymbol{z}_i)]$ is the expected loss of client i under distribution $\mathcal{D}_i$, $\boldsymbol{x}$ are the model parameters. To analyze FL procedure with varying client set, we introduce the concept of a client sampling strategy, formalized below.

Definition 3.1 (Client Sampling Strategy (CSS))**.** A CSS $\mathcal{S}_\tau$ at round τ is a (possibly randomized) procedure that returns a subset of chosen size $m = |C|$ for a given pool of clients A:

$$C \sim \mathcal{S}_\tau(A, m), \quad C = \{i_1, \ldots, i_m\} \subseteq A.$$

Algorithm 1 presents the standard federated training procedure with uniform aggregation and a controller-specified number of participating clients, which serves as the basis for our later developments. A central server coordinates training across M clients, each with its own dataset. At round τ, it samples a subset $C_{\tau+1} \sim \mathcal{S}_\tau([M], m_{\tau+1})$ of size $m_{\tau+1}$ from a whole client pool $[M] := \{1, 2, \ldots, M\}$. The selected clients compute gradient $\nabla f_i(\boldsymbol{x}^\tau)$ of their local objective using the entire dataset. This local procedure defines the reference formulation used in the analysis; the experimental implementation instead employs standard multi-step local updates, as detailed in Appendix A. The server then aggregates these gradients into $g^\tau(C_{\tau+1}) := \frac{1}{|C_{\tau+1}|} \sum_{i \in C_{\tau+1}} \nabla f_i(\boldsymbol{x}^\tau)$ and makes a descent step $\boldsymbol{x}^{\tau+1} = \boldsymbol{x}^\tau - \gamma g^\tau(C_{\tau+1})$. After T rounds, the final model $\boldsymbol{x}^T$ is obtained.

Algorithm 1 Communication with Varying Client Set

Require: initial model $\boldsymbol{x}^0$, initial count $m_0 \in [M]$, sampling strategies $\{\mathcal{S}_\tau\}$.
1: **for** $\tau = 0$ to $T-1$ **do**
2: $m_{\tau+1} \leftarrow ISP(\boldsymbol{x}^\tau, m_\tau, \tau)$ ▷ Algorithm 2
3: Sample $C_{\tau+1} \sim \mathcal{S}_\tau([M], m_{\tau+1})$
4: $g_i^\tau \leftarrow \nabla f_i(\boldsymbol{x}^\tau)$ for all $i \in C_{\tau+1}$ ▷ in parallel
5: $\boldsymbol{x}^{\tau+1} \leftarrow \boldsymbol{x}^\tau - \gamma \frac{1}{m_{\tau+1}} \sum_{i \in C_{\tau+1}} g_i^\tau$
6: **end for**
7: **Output:** $\boldsymbol{x}^T$

In standard FL procedure, the number of participants $m_{\tau+1}$ is constant across rounds. While this simplifies the analysis, it overlooks an important degree of freedom: varying $m_{\tau+1}$ can trade off communication cost against convergence speed. We study this trade-off through the *next-round loss change* $\delta f_\tau(m_{\tau+1}, C) := f\left(\boldsymbol{x}^\tau - \gamma g^\tau(C)\right) - f(\boldsymbol{x}^\tau)$.

Under uniform sampling, the count controls aggregate-gradient variance, while its expectation remains $\nabla f(\boldsymbol{x}^\tau)$. With biased selectors, the expected update may also depend on the count. We address this trade-off by *minimizing* the client count as a proxy for per-round communication. Specifically, we seek the smallest $m_{\tau+1}$ that still ensures average model improvement:

$$\begin{aligned} &m_{\tau+1} = \arg\min_{m \in [M]} \{m\}, \\ &\text{s.t.} \quad \Phi_\tau([M], m) := \mathbb{E}_{C \sim \mathcal{S}_\tau([M], m)} \delta f_\tau(m, C) < 0. \end{aligned} \tag{2}$$

Equation (2) serves as a full-pool reference rule for adaptive client-count selection. Its strict feasible set may be empty, in which case the argmin is undefined; Algorithm 2 specifies a separate fallback for its finite search. We also consider an alternative optimization formulation (Appendix E.7). However, several challenges currently hinder (2) direct solution. We next examine these *Research Questions (RQ)* and propose resolution strategies.

3.2 Our methodology

We address the requirements of the full-pool reference rule through a sequence of research questions. We begin with the computation of $\Phi_\tau([M], m)$.

RQ 1 (Full-objective estimation)**.** *How can the expected next-round loss change $\Phi_\tau([M], m)$ be evaluated before committing to the next-round cohort size?*

Evaluating $\Phi_\tau([M], m)$ requires candidate models $\boldsymbol{x}^\tau - \gamma g^\tau(C)$ and therefore access to the corresponding client updates. We introduce an intermediate round $\tau + 1/2$, in which updates are collected from $b_{\tau+1/2}$ clients before selecting $m_{\tau+1}$. Setting $b_{\tau+1/2} = M$ provides the updates required to evaluate candidate subsets drawn from $\mathcal{S}_\tau([M], m)$, but requires full intermediate participation whenever the client count is updated. To avoid this cost, we impose an auxiliary budget $b_{\tau+1/2} < M$, sample $I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})$, and replace the full-pool problem with the following intermediate-pool surrogate:

$$m_{\tau+1} = \arg\min_{m \in [b_{\tau+1/2}]} \{m : \Phi_\tau(I_{\tau+1/2}, m) < 0\}. \tag{3}$$

We provide a structural bound on the approximation gap of (3) in Appendix F.5 and empirically study the corresponding inclusion-probability diagnostics in Appendix G.

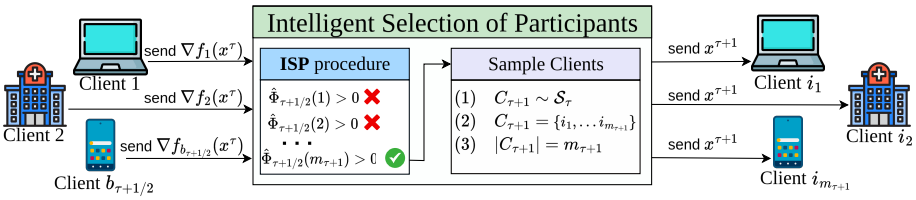


Figure 1: ISP selects a client count using an auxiliary pool, then reuses it for Δ rounds with a fresh training cohort each round.

The auxiliary budget $b_{\tau+1/2}$ controls the trade-off between intermediate communication and the fidelity of the surrogate problem. Since m_τ is the most recently selected client count, we use it as a local anchor and set $b_{\tau+1/2} = \min\{M, \lceil (1+p)m_\tau \rceil\}$ where $p \geqslant 0$ is the *expansion ratio*. Larger values of p give a wider intermediate search pool and can improve the fidelity of the surrogate, while smaller values reduce intermediate communication. This choice uses the previously selected client count as a local anchor for the next budget search, while allowing controlled expansion as the optimization state changes. We ablate this expansion ratio in Appendix E.1.

Because $m_{\tau+1} \leq b_{\tau+1/2}$, each adaptation event collects at least as many auxiliary client updates as in the subsequent training round. This motivates the following question:

RQ 2 (Communication overhead)**.** *How can the overhead of intermediate communication be amortized rather than incurred at every training round?*

In practice, the client count acts as a coarse communication-policy variable rather than an instantaneous model parameter. We therefore treat client-count selection as a *strategy-level* decision: search it once every Δ rounds reuse the selected $m_{\tau+1}$ until the next adaptation event. The effect of Δ is ablated in Appendix E.2.

We call the resulting periodic client-count controller **Intelligent Selection of Participants** (ISP). The workflow is illustrated in Figure 1, and the pseudocode is given in Algorithm 2. After intermediate communication, ISP tests candidate counts in increasing order and returns the first with a negative estimated response; if none of the tested counts is accepted, it returns the auxiliary budget $b_{\tau+1/2}$. We now turn to the internal computation required by this enumeration.

Algorithm 2 ISP (Intelligent Selection of Participants)

Require: Previous state $\{\tau, m_\tau, \boldsymbol{x}^\tau\}$. Ratio p, window Δ, depth N.

if $\tau \bmod \Delta \neq 0$ **then return** m_τ ▷ Run ISP each Δ rounds
end if
$b_{\tau+1/2} \leftarrow \min\{M, \lceil (1+p)m_\tau \rceil\}$ ▷ Auxiliary budget
$I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})$
$g_i^\tau \leftarrow \nabla f_i(\boldsymbol{x}^\tau)$ for all $i \in I_{\tau+1/2}$ ▷ Intermediate round
$m \leftarrow 1$
while $m < b_{\tau+1/2}$ **do**
  $\hat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}^\tau - \gamma g^\tau(C^n)) - f(\boldsymbol{x}^\tau)$ ▷ $C^n \sim \mathcal{S}_\tau(I_{\tau+1/2}, m)$
  if $\hat{\Phi}_\tau(I_{\tau+1/2}, m) < 0$ **then Break** ▷ Early exit
  end if
  $m \leftarrow m + 1$
end while
return m

RQ 3 (Restricted-objective estimation)**.** *How can the restricted-pool response $\Phi_\tau(I_{\tau+1/2}, m)$ be estimated across candidate counts $m \in [b_{\tau+1/2}]$?*

Conditioned on $I_{\tau+1/2}$, the quantity $\Phi_\tau(I_{\tau+1/2}, m)$ remains an expectation over subsets drawn from the restricted strategy $\mathcal{S}_\tau(I_{\tau+1/2}, m)$. We estimate it by Monte Carlo sampling:

$$\hat{\Phi}_\tau(I_{\tau+1/2}, m) = \frac{1}{N}\sum_{n=1}^{N} \left[f(\boldsymbol{x}^\tau - \gamma g^\tau(C^n)) - f(\boldsymbol{x}^\tau) \right], \tag{4}$$

where $C^n \sim \mathcal{S}_\tau(I_{\tau+1/2}, m)$. With the pre-adaptation history fixed and exact evaluations of f, the estimator is conditionally unbiased:

$$\mathbb{E}\Big[\hat{\Phi}_\tau(I_{\tau+1/2}, m) \mid I_{\tau+1/2}\Big] = \Phi_\tau(I_{\tau+1/2}, m).$$

Finite N introduces subset-sampling error, while auxiliary restriction changes the response being estimated. Section 4 and Appendix F.2 control these errors jointly for the count selected by the uniform-sampling reference procedure. Algorithm 5 in Appendix A gives the implementation, and Appendix E.5 ablates the depth N.

The required depth depends on the variability of $\mathcal{S}_\tau(I_{\tau+1/2}, m)$. For deterministic CSS, including ranking-based methods such as FEDCOR (Tang et al., 2022) and POWER-OF-CHOICE (Cho et al., 2020) when the candidate set is fixed, $N = 1$ exactly evaluates the restricted-pool response. For stochastic selectors, structured sampling can reduce variability: FEDCBS (Zhang et al., 2023) promotes class-balanced subsets, while DELTA (Wang et al., 2023) uses a variance-optimal unbiased scheme. For a fixed count, the response variability depends on the underlying CSS; selecting among several estimated counts additionally requires simultaneous error control. Appendix G analyzes the induced sampling mass under auxiliary restriction.

Evaluating $f(\boldsymbol{x}^\tau - \gamma g^\tau(C^n))$ requires client-side loss evaluation at the candidate model. Clients return scalar losses, while delivering each candidate model to its evaluators adds downlink traffic and device-side inference cost. Alternatively, a server-side surrogate loss $\hat{f}$ can approximate the same validation signal, as in proxy-objective approaches for Byzantine-tolerant FL (Xie et al., 2019; Molodtsov et al., 2026). This optional assumption is useful when public data are available; we use it for ECG in Section 5.3, analyze its computational impact in Appendix C.4, and relax it in Appendix E.8.

When validation remains strictly client-held, we suggest an optional Gaussian process ($\mathcal{GP}$) add-on that approximates the restricted-pool response $\hat{\Phi}_\tau(I_{\tau+1/2}, m)$, drawing on prior use of Gaussian processes to predict loss changes in FL Tang et al. (2022). The $\mathcal{GP}$ fits the queried pairs $(m, \hat{\Phi}_\tau(I_{\tau+1/2}, m))$ and uses its predictions at unqueried candidate counts to obtain $\hat{m}$ according to (2), thereby reducing downlinks of candidate models $\boldsymbol{x}^\tau - \gamma g^\tau(C^n)$ and client-side evaluation burden. Implementation details are provided in Appendix A, with the complete $\mathcal{GP}$ analysis deferred to Appendix D.2.

4 THEORETICAL ANALYSIS

We study two questions behind ISP: when the count selected by Algorithm 2 gives expected progress, and how reusing that count changes the outcome. We analyze the reference updates of Section 3 with exact gradients and losses, stepsize $\gamma > 0$, and $M \geq 2$ clients. All subsets are sampled uniformly without replacement; conditional on the auxiliary pool, the $N \geq 1$ Monte Carlo samples for each candidate are independent, and each training round uses a fresh uniform subset conditional on all preceding draws. Practical local updates and stabilization are discussed in Appendix F.6.

Assumption 4.1 (Smoothness)**.** The function f is differentiable and L-smooth for some $L > 0$, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$. All norms are Euclidean.

Assumption 4.2 (Gradient heterogeneity)**.** The differentiable client objectives satisfy, for constants $\sigma_c, \rho \geq 0$ and every $x \in \mathbb{R}^d$,

$$\frac{1}{M}\sum_{i=1}^{M} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma_c^2 + \rho\|\nabla f(x)\|^2.$$

The additive term permits client gradients to disagree even at a stationary point of the global objective. No convexity assumption is needed.

Participation and expected progress. First, the variance of a fresh update gives a sufficient progress condition at any count, including the count returned by the search.

Theorem 4.3 (Expected response at the selected count)**.** *Under Assumptions 4.1 and 4.2, the count returned by Algorithm 2 satisfies*

$$\begin{aligned}\Phi_\tau([M], m_{\tau+1}) \le & -\gamma(1 - L\gamma/2)\|\nabla f(\boldsymbol{x}^\tau)\|^2 \\ & + \frac{L\gamma^2}{2}\frac{M - m_{\tau+1}}{m_{\tau+1}(M-1)}\big(\sigma_c^2 + \rho\|\nabla f(\boldsymbol{x}^\tau)\|^2\big).\end{aligned} \tag{5}$$

Here the response averages over the fresh training subset, conditional on the completed search.

The factor $(M - m)/(m(M - 1))$ quantifies the reduction in gradient variance with participation. For $\sigma_c = 0$ and $\gamma = 1/L$, the bound gives strict expected descent at a nonstationary model whenever $m_{\tau+1} > M\rho/(M - 1 + \rho)$: a sufficient count scales with ρ when $1 \ll \rho \ll M$. The search need not satisfy this sufficient threshold; its own guarantee must account for estimation error. The proof is in Appendix F.1.

What the finite search guarantees. Fix an adaptation round τ with $b_{\tau+1/2} \ge 2$ and a failure probability $\delta \in (0, 1)$. Probabilities below condition on all rounds preceding this adaptation, so $\boldsymbol{x}^\tau$ and $b_{\tau+1/2}$ are fixed (Appendix F.2 formalizes this convention). Define the common error radius

$$\begin{aligned}\epsilon_\tau := \gamma\Big(\|\nabla f(\boldsymbol{x}^\tau)\| + L\gamma \max_i \|\nabla f_i(\boldsymbol{x}^\tau)\|\Big) \max_{i,k} \|\nabla f_i(\boldsymbol{x}^\tau) - \nabla f_k(\boldsymbol{x}^\tau)\| \\ \times \sqrt{\frac{1}{2}\log\frac{4b_{\tau+1/2}}{\delta}}\left(\frac{\sqrt{\min\{b_{\tau+1/2}, M - b_{\tau+1/2}\}}}{b_{\tau+1/2}} + \frac{1}{\sqrt{N}}\right).\end{aligned} \tag{6}$$

The two terms in parentheses measure auxiliary-pool restriction and Monte Carlo error. The first vanishes for a full pool; increasing N reduces only the second.

Theorem 4.4 (Accuracy of the selected response)**.** *Under Assumption 4.1, with probability at least* $1 - \delta$ *over Algorithm 2,*

$$\big|\widehat{\Phi}_\tau(I_{\tau+1/2}, m) - \Phi_\tau([M], m)\big| \le \epsilon_\tau \quad \textit{for every tested } m < b_{\tau+1/2}. \tag{7}$$

Simultaneous accuracy remains valid after selecting the count using the same estimates. On this event, an accepted count satisfies $\Phi_\tau([M], m_{\tau+1}) \le \widehat{\Phi}_\tau(I_{\tau+1/2}, m_{\tau+1}) + \epsilon_\tau$, whereas every rejected count $m < m_{\tau+1}$ satisfies $\Phi_\tau([M], m) \ge -\epsilon_\tau$. Thus an accepted estimate below $-\epsilon_\tau$ certifies expected descent; the untested fallback has no acceptance certificate. The search recovers the first count with negative full-pool response whenever that response is below $-\epsilon_\tau$, all earlier responses exceed ϵ_τ, and the count is below $b_{\tau+1/2}$. Appendix F.2 proves these conclusions and gives sharper radii. Evaluating the radius requires bounds on the population gradients in (6).

Progress while reusing the count.

Theorem 4.5 (Expected progress over a reuse block)**.** *Under Assumptions 4.1 and 4.2, assume additionally that every f_i has an L-Lipschitz gradient on $\mathbb{R}^d$. On the event of Theorem 4.4, reuse an accepted count $m_{\tau+1} < b_{\tau+1/2}$ for an integer $\Delta \ge 1$ rounds. The expectation below conditions on all preceding rounds and the completed search, and averages over the subsequent fresh training subsets:*

$$\begin{aligned}\frac{\mathbb{E}[f(\boldsymbol{x}^{\tau+\Delta}) - f(\boldsymbol{x}^\tau)]}{\Delta} \le & \ \widehat{\Phi}_\tau(I_{\tau+1/2}, m_{\tau+1}) + \epsilon_\tau \\ & + L\gamma^2(\Delta - 1)(1 + L\gamma)^{2(\Delta-1)} \\ & \times \left[\|\nabla f(\boldsymbol{x}^\tau)\|^2 + \frac{M - m_{\tau+1}}{m_{\tau+1}(M-1)}\big(\sigma_c^2 + \rho\|\nabla f(\boldsymbol{x}^\tau)\|^2\big)\right].\end{aligned} \tag{8}$$

The block makes expected progress when the magnitude of the accepted estimate exceeds the error radius plus the drift term, which vanishes at $\Delta = 1$. This controls average progress over the block. Example F.4 shows how accepting a count in error can turn mean-square contraction into growth when it is reused longer; increasing N can restore contraction. Appendix F.3 proves the bound and gives additional conditions for stationarity; Appendix F.4 develops exact examples and a partial-pool convergence and cost construction. General selectors and inclusion diagnostics are treated in Appendices F.5 and G.

5 EXPERIMENTS AND RESULTS

5.1 EXPERIMENT SETTINGS

To evaluate the impact of the proposed technique, we conducted extensive experiments on three domains. For the image domain, we use CIFAR-10 (Krizhevsky et al., 2009) with RESNET-18 (He et al., 2016) for client-selection experiments. For the text domain, we use the SHAKESPEARE dataset from the LEAF benchmark (Caldas et al., 2018) with a two-layer LSTM for next-character prediction. For the ECG domain, we use a proprietary real-world dataset of approximately $800{,}000$ 12-lead ECG segments and train RESNET1D-18 for multi-label time-series classification. We induce client heterogeneity with Dirichlet partitioning (Hsu et al., 2019): $\alpha = 0.1$ for the main CIFAR-10 client-selection experiments and $\alpha = 5.0$ for ECG; SHAKESPEARE follows the LEAF client split. Additional experiments in Appendix D evaluate ISP across various federated optimization algorithms (Appendix D.1) on TINY-IMAGENET (Appendix D.3), with the gradient-compression (Appendix D.4) and against client-count schedules (Appendix D.5).

Table 1: Hyperparameter setup across domains.

Parameters	**Image**	**Text**	**ECG**
Clients	$M = 100$	$M = 1129$	$M = 2000$
ISP-Ratio	$p = 4$	$p = 4$	$p = 4$
ISP-Window	$\Delta = 20$	$\Delta = 20$	$\Delta = 20$
ISP-Depth	$N = 10$	$N = 10$	$N = 10$
ISP-Surrogate	No	No	Yes

To match practical FL, all experiments use local client updates rather than single-gradient exchanges. Unless explicitly ablated, ISP uses the default EMA and count momentum specified in Appendix C.1. We count communication as client update calls, including intermediate ISP calls, and report average training time to capture the computation discussed in RQ 3. Appendices A and C provide the full implementation details; Table 1 summarizes the main ISP hyperparameters.

5.2 DYNAMIC PARTICIPANTS WITH CLIENT SELECTION

We evaluate ISP with UNIFORM, POW-D ($D = 40$), FEDCOR, FEDCBS ($\lambda = 10$), DELTA ($\alpha_1 = 0.8$, $\alpha_2 = 0.2$), and OORT ($\alpha = 1$, $\varepsilon = 0.05$, $c = 0.6$); Appendix B.2 describes these selectors and their parameters. Fixed-count baselines use $m = 20$, motivated by the comparison in Appendix D.7. Table 2 reports results over five runs; communication counts include intermediate ISP rounds performed every $\Delta = 20$ rounds.

ISP reduces communication by up to 25% for FEDCOR and DELTA, while preserving or slightly improving accuracy, and lowers training time for five of the six CSS baselines. For POW-D, a small communication increase yields substantially better accuracy (Figure 2a), highlighting the benefit of adapting participation to training dynamics. With OORT, we additionally simulate slow clients and intermittent availability ($q_{\text{slow}} = q_{\text{drop}} = 0.1$). The intermediate pool in (3) uses available clients, taking the full available set when it falls below the requested budget. Even under these constraints, ISP saves more than 20% of communication while maintaining accuracy, demonstrating its compatibility with system-aware selection. Appendix C.3 details this setup; Appendix G analyzes CSS-specific sampling distortions.

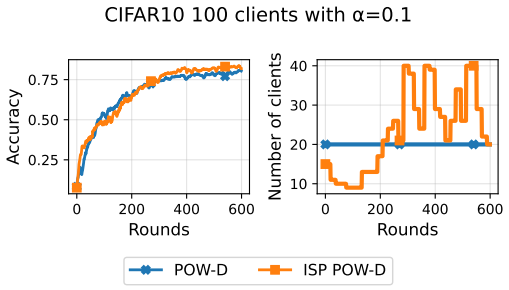
5.3 ADDITIONAL DOMAINS EXPERIMENT

ECG experiment For ECG classification, we evaluate ISP with $2{,}000$ clients as a large-scale stress test for adaptive participation. Unlike the image and text experiments, this setting uses a server-side surrogate loss $\hat{f}$ derived from the public PTB-XL dataset (Wagner et al., 2020). The task is imbalanced multi-label classification of six clinically relevant ECG abnormalities, so we report pathology-weighted Sensitivity and Specificity (Strodthoff et al., 2020). Additional setup details are provided in Appendices C.1 and C.2.

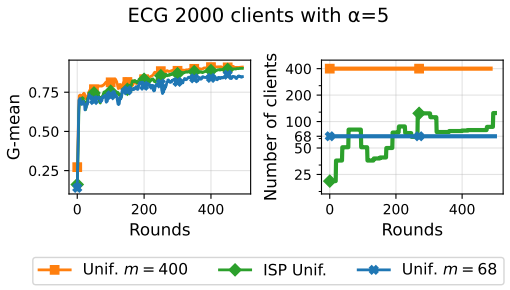
Following the participation ratio from Section 5.2, we set the fixed ECG baseline to $m = 400$. Table 3 and Figure 2b show that ISP reduces communication and training time by more than 67% while preserving downstream metrics. During training, ISP selects 68 clients per round on average; however, the fixed $m = 68$ baseline degrades performance despite using slightly fewer updates.

Table 2: Impact of ISP on CIFAR-10 across CSS baselines. Metrics average 5 runs. Communications count client update exchanges, including intermediate ISP communications.

Method	**Comm.** (↓)	**Acc.** (↑)	**Avg. Time** (↓)
Uniform	17,767 ± 1,937	0.842 ± 0.004	08h 54m 49s
ISP-Uniform	**14,343 ± 556**	0.836 ± 0.013	**06h 14m 21s**
POW-D	**10,347 ± 493**	0.812 ± 0.009	**03h 10m 33s**
ISP-POW-D	10,864 ± 546	0.830 ± 0.011	03h 20m 14s
FedCor	19,360 ± 557	0.848 ± 0.006	05h 44m 05s
ISP-FedCor	**15,037 ± 468**	0.848 ± 0.010	**04h 34m 57s**
FedCBS	19,207 ± 837	0.830 ± 0.007	11h 18m 09s
ISP-FedCBS	**16,611 ± 816**	0.836 ± 0.001	**09h 28m 14s**
DELTA	19,700 ± 191	0.721 ± 0.007	32h 05m 21s
ISP-DELTA	**14,791 ± 214**	0.722 ± 0.010	**21h 45m 26s**
OORT	19500 ± 724	0.834 ± 0.007	08h 27m 30s
ISP-OORT	**15448 ± 611**	0.835 ± 0.011	**08h 02m 46s**



(a) Client count dynamics and test accuracy for POW-D (2nd run).



(b) Client count dynamics and test G-mean for UNIFORM on ECG.

Figure 2: Dynamics of adaptive participation across CIFAR-10 and ECG experiments.

ISP varies the client count over training rather than using a fixed low budget, which yields a better communication–performance trade-off in the large-scale ECG setting.

Text domain For the text experiment, we compare ISP with UNIFORM sampling using a fixed baseline budget $m = 225$, matching the CIFAR-10 participation ratio. Experiment and dataset details are provided in Appendices C.1, C.2. We report TOP-1 accuracy, and perplexity in Table 4 and Figure 7 (top) (Appendix D.6). ISP reduces communication by 27.7%, selecting 81 clients per round on average, while maintaining comparable down-

Table 4: SHAKESPEARE next-character prediction under UNIFORM CSS.

Method	**Comm.** (↓)	**Acc.** (↑)	**PPL** (↓)
UNIFORM	18000	**0.546**	5.35
ISP-UNIFORM	**13015**	0.541	**5.07**

Table 3: Multilabel ECG classification with the fixed budget $m = 400$, ISP, and ISP-matched fixed budget $m = 68$. Specificity (Spec.) and sensitivity (Sens.) are pathology-weighted.

Method	**Comm.** (↓)	**Spec.** (↑)	**Sens.** (↑)	**Time** (↓)
Uniform ($m = 400$)	106400	**0.9129**	0.8511	22h 25m 50s
ISP Uniform	**34975**	0.9049	**0.8596**	**04h 41m 51s**
Uniform ($m = 68$)	**28560**	0.8913	0.8364	**04h 13m 54s**
ISP Uniform	34975	**0.9049**	**0.8596**	04h 41m 51s

stream performance. Although TOP-1 accuracy slightly decreases, perplexity improves. The client-count trajectory in Figure 7 (top) shows that ISP adapts the client count over training rather than relying on a fixed reduced cohort.

5.4 ABLATION STUDY

Ablation (p, Δ). To assess the sensitivity of ISP, we ablate its two key hyperparameters (p, Δ). We also report the communication overhead factor introduced by the intermediate ISP rounds.

Figure 3 shows that ISP remains stable across a broad range of these parameters. The default setting $(p = 4, \Delta = 20)$ reduces communication while preserving downstream performance, and smaller auxiliary expansions remain competitive, indicating that ISP does not rely on a finely tuned intermediate pool. Increasing Δ further reduces communication and overhead by running the intermediate search less frequently, at the cost of a moderate accuracy decrease for large windows. Overall, (p, Δ) provide interpretable knobs for controlling the communication–performance trade-off without changing the core method. Full sweeps and client-count dynamics are reported in Appendices E.1 and E.2; additional ablations are provided in Appendix E.

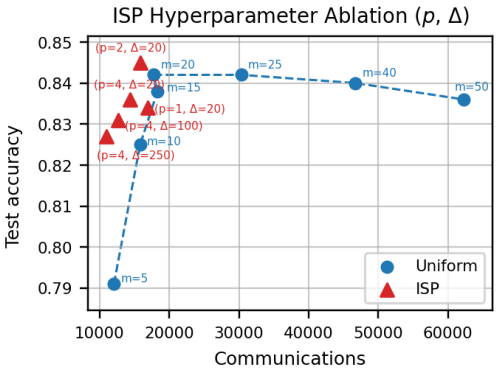


Figure 3: (p, Δ) ablation Pareto front.

$\mathcal{GP}$-guided candidate evaluation. To address settings without a centralized trust dataset, we evaluate the optional $\mathcal{GP}$ add-on using client-held validation. With an adaptation interval $\Delta = 100$, it queries up to $Q = 6$ distinct candidate counts per adaptation event, with adaptive Monte Carlo depth up to $N = 10$. A fixed evaluator set E_τ, selected using label-distribution proxies, evaluates the current and candidate models through paired loss differences. Appendix C.6 details the query procedure and evaluator construction.

Table 5 shows that ISP-GP extends adaptive participation to client-held validation, reducing client-update uploads and training time while retaining competitive predictive performance on image and text tasks. Budgeted candidate evaluation keeps the additional model-downlink cost moderate without requiring a centralized trust dataset. Appendix D.2 extends the evaluation to FEDCBS and FEDNOVA and examines the effect of evaluator coverage. It also checks the GP-predicted decision boundary against full-client evaluations and compares evaluator-selection policies at matched evaluator-set sizes.

Table 5: GP evaluation under Uniform sampling. Comm. counts client-update uploads. Relative downlink includes training and candidate-evaluation model transmissions, normalized to the fixed-budget baseline.

Dataset	Method	Comm.	Acc.	Rel. downlink	Time
CIFAR-10	Uniform	17,767	0.842	×1.00	08h 54m 49s
	ISP-GP	13,620	0.832	×1.29	06h 01m 36s
Shakespeare	Uniform	18,000	0.546	×1.00	15h 10m 25s
	ISP-GP	12,261	0.532	×1.29	13h 05m 44s

6 CONCLUSION

We introduced ISP, a progress-guided controller that adapts the number of participating clients in federated learning. Across image, text, and ECG settings, ISP achieves favorable communication–performance trade-offs, while our analysis characterizes finite-sample count selection and the effect of reusing selected counts between adaptations.

AI USE STATEMENT

Generative AI tools were used solely for formatting purposes and did not impact the core methodology, scientific rigor, or originality of the research. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of the work.

REPRODUCIBILITY STATEMENT

The anonymized source code is available at https://anonymous.4open.science/r/ISPFL-ICLR.
Implementation and experimental details are provided in the appendices.

REFERENCES

Mohammad Mohammadi Amiri, Deniz Gündüz, Sanjeev R Kulkarni, and H Vincent Poor. Convergence of update aware device scheduling for federated learning at the wireless edge. *IEEE Transactions on Wireless Communications*, 20(6):3643–3658, 2021.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Advances in Neural Information Processing Systems*, 37:65525–65561, 2024.

Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous sgd, 2017. URL https://arxiv.org/abs/1604.00981.

Jinyu Chen, Wenchao Xu, Song Guo, Junxiao Wang, Jie Zhang, and Haozhao Wang. Fedtune: A deep dive into efficient federated fine-tuning with pre-trained transformers. *arXiv preprint arXiv:2211.08025*, 2022.

Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

Allan M de Souza, Filipe Maciel, Joahannes BD da Costa, Luiz F Bittencourt, Eduardo Cerqueira, Antonio AF Loureiro, and Leandro A Villas. Adaptive client selection with personalization for communication efficient federated learning. *Ad Hoc Networks*, 157:103462, 2024.

Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. Active federated learning. *arXiv preprint arXiv:1909.12641*, 2019.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification, 2019. URL https://arxiv.org/abs/1909.06335.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pp. 19–35, 2021.

Chenning Li, Xiao Zeng, Mi Zhang, and Zhichao Cao. Pyramidfl: A fine-grained client selection framework for efficient federated learning. In *Proceedings of the 28th annual international conference on mobile computing and networking*, pp. 158–171, 2022.

Qingming Li, Xiaohang Li, Li Zhou, and Xiaoran Yan. Adafl: Adaptive client selection and dynamic contribution evaluation for efficient federated learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649. IEEE, 2024.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Wenhao Lin and Yang Zhou. Adaptive client participation mechanism for federated learning in heterogeneous vehicular networks. *Scientific Reports*, 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Gleb Molodtsov, Daniil Medyakov, Sergey Skorik, Nikolas Khachaturov, Shahane Tigranyan, Vladimir Aletov, Aram Avetisyan, Martin Takáč, and Aleksandr Beznosikov. Bant: Byzantine antidote via trial function and trust scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24431–24440, 2026.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pp. 1–7. IEEE, 2019.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

Mónica Ribero, Haris Vikalo, and Gustavo De Veciana. Federated learning under intermittent client availability and time-varying communication constraints. *IEEE Journal of Selected Topics in Signal Processing*, 17(1):98–111, 2022.

Elsa Rizk, Stefan Vlaski, and Ali H Sayed. Federated learning under importance sampling. *IEEE Transactions on Signal Processing*, 70:5381–5396, 2022.

Atal Sahu, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. Rethinking gradient sparsification as total error minimization. *Advances in Neural Information Processing Systems*, 34:8133–8146, 2021.

Faranaksadat Solat and Joohyung Lee. Optimizing client participation in communication-constrained federated llm adaptation with lora. *Sensors*, 25(21):6538, 2025.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, and Yiran Chen. Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10102–10111, 2022.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Lin Wang, YongXin Guo, Tao Lin, and Xiaoying Tang. Delta: Diverse client sampling for fasting federated learning. *Advances in Neural Information Processing Systems*, 36:47626–47668, 2023.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International conference on machine learning*, pp. 6893–6901. PMLR, 2019.

Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2): 1188–1200, 2020.

Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, and Hai Li. Fed-cbs: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In *International Conference on Machine Learning*, pp. 41354–41381. PMLR, 2023.

Appendix

A Additional Pseudocode

Algorithm 3 Federated Optimization with Varying Client Set

Require: Initial model $\boldsymbol{x}^0$, client count m_0, sampling strategies $\{\mathcal{S}_\tau\}_{\tau=0}^{T-1}$, aggregation rule $\mathcal{A}$, compressor $\mathcal{C}$.

1: **for** $\tau = 0$ to $T-1$ **do**
2: $m_{\tau+1} \leftarrow$ Ext-ISP$(\tau, m_\tau, \boldsymbol{x}^\tau)$ ▷ Algorithm 4
3: Sample $C_{\tau+1} \sim \mathcal{S}_\tau([M], m_{\tau+1})$
4: $\{\mathbf{u}_i^\tau\}_{i\in C_{\tau+1}} \leftarrow$ ClientsUpdate$(\boldsymbol{x}^\tau, C_{\tau+1}, \mathcal{C})$ ▷ Algorithm 6
5: $\boldsymbol{x}^{\tau+1} \leftarrow \mathcal{A}\big(\boldsymbol{x}^\tau, \{\mathbf{u}_i^\tau\}_{i\in C_{\tau+1}}\big)$ ▷ Server aggregation
6: **end for**
7: **Output:** $\boldsymbol{x}^T$

Algorithms 1 and 2 define the reference client-count formulation used in the main text and its theoretical analysis. The algorithms below provide its implementation-level realization: Algorithms 3 and 4 specify the federated training and adaptive client-count procedures, Algorithms 5 and 6 define response estimation and the underlying client routines, and Algorithms 7 and 8 describe the optional $\mathcal{GP}$ add-on.

Relation to the reference formulation. Algorithms 1 and 2 define the reference transition used to formulate the client-count problem in Equations (2)–(4) and to support the analysis in Section 4. The implementation-level procedures above implement the same progress-based approach, but replace the single-gradient transition $\boldsymbol{x}^\tau - \gamma g^\tau(C)$ with the federated optimization pipeline used in the experiments. Specifically, Algorithm 3 first obtains the next client count from Ext-ISP, samples the corresponding training cohort, computes its local updates, and applies the server aggregation rule (lines 2–5). At an adaptation round, Ext-ISP constructs the auxiliary pool and obtains the corresponding intermediate client states through the same local-update and compression pipeline (Algorithm 4, lines 4, 6). In particular, candidate models are produced through multi-step local training, optional compression $\mathcal{C}$, and the aggregation rule $\mathcal{A}$ associated with the underlying FL method. Consequently, ISP adapts the client count to the complete update pipeline rather than to a separate proxy update. This extension establishes compatibility with different client sampling strategies, federated optimizers, and compression regimes, while the theoretical guarantees concern uniform reference sampling with exact single-gradient updates, including finite auxiliary pools in Section 4.

Operational response estimation. Algorithm 5 implements the restricted-pool response in Equation (4) using the objective signal available in the corresponding experimental regime. For each candidate count m, subsets $C^n \sim \mathcal{S}_\tau(I_{\tau+1/2}, m)$ produce candidate models through the same aggregation rule used for training. The exact objective values in the reference formulation are then replaced by empirical estimates $\ell_n(\boldsymbol{x}^{(\tau+1/2)_n})$ (Algorithm 5, lines 4, 5) and $\ell_\tau(\boldsymbol{x}^\tau)$ obtained through ObjectiveEval (Algorithm 6). In the client-side regime, this objective is the mean of the local losses over the available clients; when a server-side surrogate $\hat{f}$ is available (Algorithm 6, line 16), the same interface evaluates the corresponding surrogate objective. Using the same objective-evaluation rule for the current and candidate models preserves the interpretation of $\widehat{\Phi}_\tau(I_{\tau+1/2}, m)$ as an estimated next-round loss change. EMA smoothing, enabled in the default configuration, stabilizes this empirical signal before the zero-threshold test (Algorithm 5, line 10). Since smoothing can change its sign, the theoretical implications require a bound on its discrepancy from the current response, as discussed in Appendix F.6.

GP-guided response approximation. Algorithm 4 also permits the candidate-response relation to be approximated by the optional $\mathcal{GP}$ add-on, which is selected and invoked in lines 8–9. The GP uses the same auxiliary-state construction and the ExpectEstim interface; in particular, each queried candidate is evaluated through Algorithm 5 at line 10. It retains the progress-based count-selection

criterion and predicts the queried loss-change signal across candidate counts using observations and round-level context. Restricting evaluation to E_τ introduces an evaluator surrogate for the full-client objective; its discrepancy and the GP prediction error are additional approximations discussed in Appendix F.6. Algorithm 7 constructs this context from the auxiliary-update statistics, their alignment with the previous server update, and the loss history (lines 4–7); Algorithm 8 then alternates posterior updates, candidate selection, and response collection in lines 8–11, before returning the final posterior-selected count in lines 17–18. The default implementation evaluates the response directly through the search in Algorithm 4, lines 13–20, so that the contribution of the client-count controller can be studied independently of an additional learned approximation layer across the sampling, optimization, and compression experiments. The GP variant is considered separately for regimes in which repeated objective evaluation constitutes a relevant cost and its additional context construction, posterior updates, query policy, and accounting can be assessed explicitly. In both modes, the resulting $m^\star$ is processed by the same momentum rule before determining $m_{\tau+1}$ (lines 10 and 22); further GP details and analysis are provided in Appendices C.1,D.2.

Algorithm 4 EXT-ISP (Extended Intelligent Selection of Participants)

Require: Round τ, previous count m_τ, current model $\boldsymbol{x}^\tau$.
Context: strategy $\mathcal{S}_\tau$; total clients M; ratio p; window Δ; depth N; resolution w; momentum β; $\mathbb{I}\{\text{Use } \mathcal{GP}\}$.

1: **if** $\tau \bmod \Delta \neq 0$ **then**
2: **return** m_τ ▷ Run ISP each Δ rounds
3: **end if**
4: $b_{\tau+1/2} \leftarrow \min\{M, \lceil (1+p)m_\tau \rceil\}$ ▷ Auxiliary budget
5: $I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})$ ▷ Auxiliary pool
6: $\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}} \leftarrow \textsc{ClientsUpdate}(\boldsymbol{x}^\tau, I_{\tau+1/2}, \mathcal{C})$
7: $\ell(\boldsymbol{x}^\tau) \leftarrow \frac{1}{M}\sum_{i=1}^M \ell_i(\boldsymbol{x}^\tau), \quad \ell_i(\boldsymbol{x}^\tau) \leftarrow \textsc{ObjectiveEval}(\boldsymbol{x}^\tau, \{i\}),\ i \in [M]$ ▷ Algorithm 6
8: **if** $\mathbb{I}\{\text{Use } \mathcal{GP}\}$ **then**
9: $m^\star \leftarrow \text{GP-}\textsc{Estimation}\big(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}, I_{\tau+1/2}, w, N, \{\ell_i(\boldsymbol{x}^\tau)\}_{i\in[M]}\big)$ ▷ Optional $\mathcal{GP}$ estimator
10: $m_{\tau+1} \leftarrow \min\big\{b_{\tau+1/2}, \max\{1, \lfloor \beta m^\star + (1-\beta)m_\tau \rfloor\}\big\}$ ▷ Apply β-momentum
11: **return** $m_{\tau+1}$
12: **end if**
13: $m \leftarrow 1, \quad m^\star \leftarrow b_{\tau+1/2}$
14: **while** $m < b_{\tau+1/2}$ **do**
15: $\widehat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow \textsc{ExpectEstim}\big(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}, \boldsymbol{x}^\tau, N, I_{\tau+1/2}, [M], m, \ell(\boldsymbol{x}^\tau)\big)$
16: **if** $\widehat{\Phi}_\tau(I_{\tau+1/2}, m) < 0$ **then**
17: $m^\star \leftarrow m$
18: **break** ▷ Early exit
19: **end if**
20: $m \leftarrow m + w$ ▷ Search with resolution w
21: **end while**
22: $m_{\tau+1} \leftarrow \min\big\{b_{\tau+1/2}, \max\{1, \lfloor \beta m^\star + (1-\beta)m_\tau \rfloor\}\big\}$ ▷ Apply β-momentum
23: **return** $m_{\tau+1}$

Algorithm 5 EXPECTESTIM

Require: States $(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}, x^\tau)$; depth N; auxiliary pool $I_{\tau+1/2}$; infer pool I; candidate m; estimated loss ℓ_τ.
Context: strategy $\mathcal{S}_\tau$; aggregation $\mathcal{A}$; $\mathbb{I}\{\text{Use EMA}\}$; estimated loss history $\mathcal{L}$.

$\widehat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow 0$
for $n = 1$ to N **do**
  $C^n \sim \mathcal{S}_\tau(I_{\tau+1/2}, m)$ ▷ Restricted-pool sampling
  $\boldsymbol{x}^{(\tau+1/2)_n} \leftarrow \mathcal{A}\Big(\boldsymbol{x}^\tau, \{\mathbf{u}_i^{\tau+1/2}\}_{i\in C^n}\Big)$
  $\ell_n\big(\boldsymbol{x}^{(\tau+1/2)_n}\big) \leftarrow \textsc{ObjectiveEval}\big(\boldsymbol{x}^{(\tau+1/2)_n}, I\big)$
  $\widehat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow \widehat{\Phi}_\tau(I_{\tau+1/2}, m) + \frac{1}{N}\ell_n\big(\boldsymbol{x}^{(\tau+1/2)_n}\big)$
end for

if $\mathbb{I}\{\text{Use EMA}\}$ **then**
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\tau, \ell_\tau)\}$ ▷ Update once per round
  $\widehat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow \text{EMA}\Big(\widehat{\Phi}_\tau(I_{\tau+1/2}, m), \mathcal{L}\Big)$
end if

$\widehat{\Phi}_\tau(I_{\tau+1/2}, m) \leftarrow \widehat{\Phi}_\tau(I_{\tau+1/2}, m) - \ell_\tau$
return $\widehat{\Phi}_\tau(I_{\tau+1/2}, m)$

Algorithm 6 CLIENTSUPDATE and OBJECTIVEEVAL

Context: local datasets $\{\mathcal{D}_i\}_{i=1}^M$ and objectives $\{f_i\}_{i=1}^M$; local optimizer OPT; local epochs E; batch size B; $\mathbb{I}\{\text{Use } \hat{f}\}$.

function CLIENTSUPDATE($\boldsymbol{x}^\tau, C, \mathcal{C}$)
  for all clients $i \in C$ **in parallel do**
    $\boldsymbol{x}_i \leftarrow \boldsymbol{x}^\tau$
    for $e = 1$ to E **do**
      for mini-batches $\mathcal{B} \subseteq \mathcal{D}_i$ **do**
        $g_i \leftarrow \nabla f_i(\boldsymbol{x}_i; \mathcal{B})$
        $\boldsymbol{x}_i \leftarrow \textsc{OptStep}(\boldsymbol{x}_i, g_i)$
      end for
    end for
    $\mathbf{u}_i^\tau \leftarrow \mathcal{C}(\boldsymbol{x}_i - \boldsymbol{x}^\tau)$ ▷ Compressed client update
  end for
  return $\{\mathbf{u}_i^\tau\}_{i\in C}$
end function

function OBJECTIVEEVAL($\boldsymbol{x}, C$)
  if $\mathbb{I}\{\text{Use } \hat{f}\}$ **then**
    $\ell(\boldsymbol{x}; C) \leftarrow \hat{f}(\boldsymbol{x})$ ▷ Server-side evaluation
  else
    for all clients $i \in C$ **in parallel do**
      $\ell_i(\boldsymbol{x}) \leftarrow 0$
      for mini-batches $\mathcal{B} \subseteq \mathcal{D}_i$ **do**
        $\ell_i(\boldsymbol{x}) \leftarrow \ell_i(\boldsymbol{x}) + \frac{|\mathcal{B}|}{|\mathcal{D}_i|} f_i(\boldsymbol{x}; \mathcal{B})$
      end for
    end for
    $\ell(\boldsymbol{x}; C) \leftarrow \frac{1}{|C|}\sum_{i\in C} \ell_i(\boldsymbol{x})$ ▷ Uniform client-average loss
  end if
  return $\ell(\boldsymbol{x}; C)$
end function

Algorithm 7 CREATECONTEXT and PROXYCOHORT

Context: counts $(m_\tau, b_{\tau+1/2}, M)$; stale states $\{\mathbf{u}_i\}_{i \notin I_{\tau+1/2}}$; dataset sizes $\{n_i\}_{i=1}^M$; temperature T; coverage k; history $\mathcal{L}$.

function CREATECONTEXT($\{\mathbf{u}_i^{\tau+1/2}\}_{i \in I_{\tau+1/2}}, \boldsymbol{x}^\tau, \boldsymbol{x}^{\tau-1}$)
  $\mathbf{u}^{\tau-1} \leftarrow \boldsymbol{x}^\tau - \boldsymbol{x}^{\tau-1}$ ▷ Previous server update
  $\overline{\mathbf{u}}_{\tau+1/2} \leftarrow \frac{1}{b_{\tau+1/2}} \sum_{i \in I_{\tau+1/2}} \mathbf{u}_i^{\tau+1/2}$ ▷ Mean auxiliary-client update
  $\overline{r}_{\tau+1/2} \leftarrow \frac{1}{b_{\tau+1/2}} \sum_{i \in I_{\tau+1/2}} \|\mathbf{u}_i^{\tau+1/2}\|_2$ ▷ Mean auxiliary-update norm
  $s_{\tau+1/2} \leftarrow \sqrt{\frac{1}{b_{\tau+1/2}} \sum_{i \in I_{\tau+1/2}} \left(\|\mathbf{u}_i^{\tau+1/2}\|_2 - \overline{r}_{\tau+1/2}\right)^2}$ ▷ Auxiliary-update norm deviation
  $\rho_{\tau+1/2} \leftarrow \frac{\langle \overline{\mathbf{u}}_{\tau+1/2}, \mathbf{u}^{\tau-1} \rangle}{\|\overline{\mathbf{u}}_{\tau+1/2}\|_2 \|\mathbf{u}^{\tau-1}\|_2}$ ▷ Alignment with the previous server update
  return $\left[\frac{m_\tau}{M}, \frac{b_{\tau+1/2}}{M}, \overline{r}_{\tau+1/2}, s_{\tau+1/2}, \rho_{\tau+1/2}, \mathcal{L}\right]$ ▷ Round-level $\mathcal{GP}$ context
end function

function PROXYCOHORT($\{\mathbf{u}_i^{\tau+1/2}\}_{i \in I_{\tau+1/2}}$)
  $\{\mathbf{u}_i\}_{i \in [M]} \leftarrow \{\mathbf{u}_i^{\tau+1/2}\}_{i \in I_{\tau+1/2}} \cup \{\mathbf{u}_i\}_{i \notin I_{\tau+1/2}}$ ▷ Collect full states
  $\mathbf{p}_i \leftarrow \operatorname{softmax}\left([\mathbf{u}_i]_{\text{bias}}/T\right),\ i \in [M]$ ▷ $\mathbf{p}_i \in \mathbb{R}^C$ proxy class distribution (Chen & Vikalo, 2024)
  $E_\tau \leftarrow \emptyset,\ \mathbf{d} \leftarrow k \cdot \mathbf{1}_C,\ \mathbf{d} \in \mathbb{R}^C$
  while $\sum_{c=1}^C \mathbb{I}\{d_c > 0\} > 0$ **do**
    $i^\star \leftarrow \arg\max_{i \in [M] \setminus E_\tau} \sum_{c=1}^C \min\{n_i p_{i,c}, d_c\}$ ▷ Maximum marginal coverage
    $E_\tau \leftarrow E_\tau \cup \{i^\star\}$
    $\mathbf{d} \leftarrow [\mathbf{d} - n_{i^\star} \mathbf{p}_{i^\star}]_+$ ▷ Element-wise
  end while
  return E_τ
end function

Algorithm 8 GP-ESTIMATION

Require: States $\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}$; auxiliary pool $I_{\tau+1/2}$; resolution w; depth N; estimated losses $\{\ell_i(\boldsymbol{x}^\tau)\}_{i\in[M]}$.
Context: strategy $\mathcal{S}_\tau$; response history $\mathcal{D}_{\mathrm{GP}}$; global states $(\boldsymbol{x}^\tau, \boldsymbol{x}^{\tau-1})$; previous count m_τ; budget Q; confidence β_σ.

$g \sim \mathcal{GP}(c_\theta, k_\theta)$ ▷ Use fitted $\mathcal{GP}$ with $\theta = \theta^{\tau-1}$
$C_\tau \leftarrow \textsc{CreateContext}\Big(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}, \boldsymbol{x}^\tau, \boldsymbol{x}^{\tau-1}\Big)$, $\boldsymbol{z}_\tau(m) := [C_\tau, m]$ ▷ Algorithm 7
$E_\tau \leftarrow \textsc{ProxyCohort}\Big(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}\Big)$, $\ell_{E_\tau}(\boldsymbol{x}^\tau) \leftarrow \frac{1}{|E_\tau|}\sum_{i\in E_\tau}\ell_i(\boldsymbol{x}^\tau)$ ▷ Algorithm 7
$\mathcal{M}_\tau \leftarrow [1, 1+w, \ldots, b_{\tau+1/2}]$ ▷ Candidate counts at resolution w
$\mathbf{Z}_\tau \leftarrow \big[\boldsymbol{z}_\tau(m)\big]_{m\in\mathcal{M}_\tau}$, $\mathbf{g}_\tau \leftarrow g(\mathbf{Z}_\tau) = \big[g(z_\tau(m))\big]_{m\in\mathcal{M}_\tau}$ ▷ Context and prior initalization
$\mathcal{D}^\tau \leftarrow \varnothing$, $q \leftarrow 0$

while $|\mathcal{D}^\tau| < Q$ **do**
  $\mathbf{g}_\tau \mid \mathcal{D}_{\mathrm{GP}}, \mathcal{D}^\tau \sim \mathcal{N}\Big(\boldsymbol{\mu}_\tau^{(q)}, \boldsymbol{\Sigma}_\tau^{(q)}\Big)$ ▷ Posterior update based on actual $\mathcal{D}^\tau$
  $m_q \leftarrow \min\{m \in \mathcal{M}_\tau : \mu_\tau^{(q)}(m) + \beta_\sigma\sigma_\tau^{(q)}(m) < 0\}$ ▷ Fallback: Appendix C.6
  $\widehat{\Phi}_\tau(I_{\tau+1/2}, m_q) \leftarrow \textsc{ExpectEstim}\Big(\{\mathbf{u}_i^{\tau+1/2}\}_{i\in I_{\tau+1/2}}, \boldsymbol{x}^\tau, N, I_{\tau+1/2}, E_\tau, m_q, \ell_{E_\tau}(\boldsymbol{x}^\tau)\Big)$
  $\mathcal{D}^\tau \leftarrow \mathcal{D}^\tau \cup \big(z_\tau(m_q), \widehat{\Phi}_\tau(I_{\tau+1/2}, m_q)\big)$ ▷ Update posterior-path observation
  $q \leftarrow q + 1$
end while

$\mathcal{D}_{\mathrm{GP}} \leftarrow \mathcal{D}_{\mathrm{GP}} \cup \mathcal{D}^\tau$
$\theta^\tau \leftarrow \arg\max_\theta \log p(\mathbf{y}_{\mathrm{GP}}^\tau \mid \mathbf{Z}_{\mathrm{GP}}^\tau, \theta)$ ▷ $\mathcal{GP}$ fitting
$\mathbf{g}_\tau \mid \mathcal{D}_{\mathrm{GP}} \sim \mathcal{N}(\boldsymbol{\mu}_\tau^\star, \boldsymbol{\Sigma}_\tau^\star)$ ▷ Posterior update based on fitted θ^τ
$m^\star \leftarrow \min\{m \in \mathcal{M}_\tau : \mu_\tau^\star(m) + \beta_\sigma\sigma_\tau^\star(m) < 0\}$ ▷ Same fallback rule
return $m^\star$

B Additional Related Work

B.1 Federated Optimization Methods

Federated optimization builds on distributed synchronous SGD (Chen et al., 2017), with Federated Averaging (FedAvg) (McMahan et al., 2017) alternating local client training and global model averaging to trade off computation and communication. Allowing multiple local steps reduces the frequency of communication, but heterogeneous client objectives can cause local updates to diverge. This has motivated methods that modify either local optimization or the way client updates are combined.

To mitigate client drift, FedProx (Li et al., 2020b) constrains local optimization through a proximal term, while SCAFFOLD (Karimireddy et al., 2020) uses control variates to correct the discrepancy between local and global update directions. A complementary source of inconsistency arises when clients perform different numbers of local steps: FedNova (Wang et al., 2020) addresses this through update normalization. Adaptive server optimization (Reddi et al., 2020) further develops the aggregation stage by applying Adam-like updates to the information received from clients. Together, these approaches improve how local computation contributes to the global model; broader surveys discuss their role among the statistical and system challenges of federated learning (Li et al., 2020a).

B.2 Client Selection

Alongside changes to local training and aggregation, communication efficiency can be improved by choosing which clients participate. Uniform sampling provides a simple reference that does not use differences in client data or execution conditions. Under heterogeneous data, however, equally sized participant sets can provide different learning contributions, motivating selection rules based on statistical utility. System-aware approaches additionally account for resource availability and execution speed (Nishio & Yonetani, 2019; Xu & Wang, 2020; Ribero et al., 2022).

Local loss offers a direct signal of how well the current model fits a client's data. Early work used this signal to assign sampling probabilities (Goetz et al., 2019), but obtaining fresh losses introduces communication of its own. Power-of-Choice (POW-D) (Cho et al., 2020) limits this requirement by evaluating a randomly sampled candidate pool and selecting its highest-loss clients; the pool size D controls the scope of this comparison ($D = 40$ in our experiments). Loss-based selection can also account for interactions between clients rather than ranking their losses independently. FedCor (Tang et al., 2022) models correlated client loss changes with a Gaussian process and selects participants according to their predicted effect on the global loss. Periodic observations update the model, reducing the need for repeated loss collection from all clients.

Gradient information provides another view of a client's optimization contribution. Selection rules have used gradient norms (Amiri et al., 2021), local-update norms (Chen et al., 2020), and importance sampling to reduce sampling variance (Rizk et al., 2022). Large update magnitudes alone, however, do not distinguish complementary contributions from similar updates. DELTA (Wang et al., 2023) addresses this issue by incorporating gradient diversity into its sampling criterion. Its practical rule combines between-client gradient differences with within-client stochastic variability, weighted by α_1 and α_2, respectively ($\alpha_1 = 0.8$, $\alpha_2 = 0.2$ in our experiments).

A related concern is whether the selected clients collectively represent the classes needed for training. FED-CBS (Zhang et al., 2023) uses label-distribution information to reduce class imbalance in the combined data of the selected clients. Thus, the criterion concerns the composition of the participant set rather than the loss or update magnitude of each client in isolation. Its exploration–exploitation mechanism is controlled by λ ($\lambda = 10$ in our experiments).

Statistical utility alone does not capture the time required to obtain a client's contribution. Oort (Lai et al., 2021) combines loss-based utility with a runtime penalty and exploration, bringing learning contribution and execution speed into the same selection rule. The runtime penalty is controlled by α, exploration by ε, and the confidence width by c ($\alpha = 1$, $\varepsilon = 0.05$, $c = 0.6$ in our experiments). PyramidFL (Li et al., 2022) further develops this direction through local utility optimization to better exploit the contributions of selected clients.

These approaches provide different criteria for composing a participant set, typically for a specified participation budget. The number of participants introduces an additional degree of freedom: it determines how many contributions are collected, while the selection rule determines their composition.

C TECHNICAL DETAILS

This appendix provides implementation and experimental details that are omitted from the main text. We first describe the practical ISP components used in the experiments, then provide dataset-specific setup details, the system-aware OORT implementation, and the accounting rules for communication and computation.

C.1 PRACTICAL ISP DETAILS

The implementation-level training and client-count procedures are specified in Algorithms 3–6. At each adaptation round, the clients in the auxiliary pool initialize from the current server model and perform multi-step local optimization through CLIENTSUPDATE (Algorithm 6, lines 1–12). Unless stated otherwise, each client trains for one local epoch. Since the batch size is fixed within each experiment while local dataset sizes may differ, heterogeneous partitioning can induce different numbers of local optimization steps across clients. The resulting updates are processed by the same compressor $\mathcal{C}$ and aggregation rule $\mathcal{A}$ as the subsequent training update: EXT-ISP first constructs the intermediate client updates (Algorithm 4, line 6), and EXPECTESTIM aggregates their sampled subsets to form candidate models (Algorithm 5, lines 3–4). Thus, the estimated response reflects the local-training, compression, and aggregation regime actually used in the experiment rather than the single-gradient reference transition from the main analysis.

The practical controller additionally uses three controls for search granularity and stability. The resolution parameter w evaluates candidate counts on the grid $1, 1+w, \ldots, b_{\tau+1/2}$ instead of checking every feasible count (Algorithm 4, line 20), reducing candidate-evaluation cost in large-client settings; its effect is studied in Appendix E.6. After obtaining the first candidate $m^\star$ satisfying the estimated descent condition, the controller applies

$$m_{\tau+1} = \min\Big\{b_{\tau+1/2}, \max\{1, \lfloor \beta m^\star + (1-\beta) m_\tau \rfloor\}\Big\}.$$

This momentum step limits abrupt round-to-round changes in the selected cohort size: $\beta = 1$ recovers the unsmoothed selection, while smaller values retain more of the previous count. The same post-processing is applied after either direct or $\mathcal{GP}$-guided response estimation (Algorithm 4, lines 10 and 22); the corresponding ablation is reported in Appendix E.3. Finally, EMA smoothing in the default configuration combines the current candidate-objective estimate with the stored loss history before forming the final response (Algorithm 5, lines 9–10). The same zero threshold is applied to the smoothed signal, whose sign can differ from the current unsmoothed response; its empirical effect is studied in Appendix E.4.

Across experiments, we use $\beta = 0.5$. For the large-client text and ECG settings, we additionally use $w = 20$ to avoid unnecessarily fine-grained enumeration.

C.2 DATASET AND TASK DETAILS

ECG details. The proprietary ECG recordings are represented as 10-second 12-lead segments sampled at 500Hz from patients older than 18 years. The independent test set contains approximately 400,000 ECG segments obtained under the same data-generation protocol. We consider six clinically relevant abnormalities commonly used in ECG classification studies (Ribeiro et al., 2020; Hannun et al., 2019): Atrial Fibrillation (AFIB), Premature Ventricular Complex (PVC), Complete Left Bundle Branch Block (CLBBB), Complete Right Bundle Branch Block (CRBBB), Left Anterior Fascicular Block (LAFB), and first-degree AV Block (1AVB). To handle the strong imbalance toward normal recordings, we use positive-label weights in the local binary cross-entropy losses.

SHAKESPEARE. For the SHAKESPEARE experiment, each speaking role from the LEAF benchmark is treated as a client, resulting in $M = 1129$ naturally non-IID clients (Caldas et al., 2018). The task is next-character prediction: each input is a sequence of 80 consecutive characters, and the target is the next character. For each client, we preserve the natural temporal order of the text stream and split it chronologically into train and test parts in a $9/1$ ratio, preventing future-character leakage into training. The model is a two-layer LSTM with 256 hidden units per layer and 8-dimensional character embeddings. We use ADAM with learning rate 0.01.

Image datasets. For CIFAR-10, we use RESNET-18 and Dirichlet partitioning with $\alpha = 0.1$ in the main client-selection experiments. For the gradient-compression experiment, we use a nearly uniform partition with $\alpha = 1000$ to isolate the interaction between ISP and compressed updates. For TINY-IMAGENET, we use SWIN-T with Dirichlet heterogeneity $\alpha = 0.5$. The main image-domain hyperparameters are summarized in Table 9.

C.3 OORT

We include OORT (Lai et al., 2021) to evaluate ISP on top of a system-aware client selection strategy. Unlike the other baselines in Section 5.2, OORT combines statistical and system utility in a single score. For client i, the utility is

$$\mathrm{Util}(i) = \underbrace{|B_i|\sqrt{\frac{1}{|B_i|}\sum_{k\in B_i}\mathrm{Loss}(k)^2}}_{\text{statistical utility } U(i)} \cdot \underbrace{\left(\frac{T}{t_i}\right)^{\mathbf{1}_{\{T<t_i\}}\cdot\alpha}}_{\text{system utility}},$$

where B_i is the set of local mini-batches used to estimate the client contribution, $\mathrm{Loss}(k)$ is the training loss on mini-batch k, t_i is the profiled completion time of the client, T is the target round duration, and α controls the penalty for slow clients. The statistical term favors clients with large and informative local data, while the system term penalizes clients whose completion time exceeds T.

We use exploration factor $\varepsilon = 0.05$ and confidence width $c = 0.6$. At each round, OORT selects the top $(1-\varepsilon)m$ clients by utility and fills the remaining part through exploration among clients whose utilities fall within the confidence interval. ISP is applied on top of this selection rule by adapting only the requested number of clients m.

Our simulation runs on homogeneous hardware, so we inject artificial runtime variability to emulate system heterogeneity. We first mark a random subset of clients as computation-limited with probability $q_{\mathrm{slow}} = 0.1$. Whenever such a client participates, we draw $\xi \sim \mathrm{Geom}(q_{\mathrm{geom}})$ with $q_{\mathrm{geom}} = 0.1$ and add $0.5 \cdot \xi$ seconds of delay to its local training, where 0.5 seconds corresponds to the measured duration of one local training step in our setup. We also simulate intermittent availability by dropping each client independently with probability $q_{\mathrm{drop}} = 0.1$ before selection. If fewer clients are available than requested during an intermediate ISP step, the full available set is used as both the auxiliary and inference pools.

We set $\alpha = 1$ in the OORT system penalty. The target duration T is chosen from profiling the CIFAR-10 training loop. With batch size 64, one local training step takes approximately 0.5 seconds. Treating $1{,}500$ local examples as a conservative upper bound for a client gives roughly 24 local steps; including communication and bookkeeping overhead, we set $T \approx 14$ seconds. This makes the system penalty active for clients that are substantially slower than the intended round duration.

C.4 COMMUNICATION AND COMPUTATION: ACCOUNTING AND OVERHEAD

For a distributed routine $\mathcal{R}$, let $\mathrm{size}_{\uparrow}(\mathcal{R})$ and $\mathrm{size}_{\downarrow}(\mathcal{R})$ denote the total sizes of the messages transmitted from clients to the server and from the server to clients, respectively. For CLIENTSUPDATE, each client $i \in C$ receives the current state $\boldsymbol{x}$ and returns its compressed update $\mathbf{u}_i$ (Algorithm 6, lines 1–12). Hence,

$$\mathrm{size}_{\uparrow}(\textsc{ClientsUpdate}(\boldsymbol{x}, C, \mathcal{C})) = \sum_{i\in C}\mathrm{size}(\mathbf{u}_i) = |C|\,\kappa_{\mathcal{C}}\,\mathrm{size}(\boldsymbol{x}),$$

where $\kappa_{\mathcal{C}} = 1$ for dense updates and $\kappa_{\mathcal{C}} < 1$ for a compressed representation. As $\mathrm{size}(\mathbf{u}_i) \equiv \kappa_{\mathcal{C}}\mathrm{size}(\mathbf{x})$ in our horizontal FL procedure, we get

$$\mathrm{size}_{\downarrow}(\textsc{ClientsUpdate}(\boldsymbol{x}, C, \mathcal{C})) = \sum_{i\in C}\left[\mathrm{size}(\boldsymbol{x}) + \mathrm{size}(\mathrm{ctrl}_i^{\mathrm{upd}})\right] \approx |C|\,\mathrm{size}(\boldsymbol{x}),$$

where $\mathrm{ctrl}_i^{\mathrm{upd}}$ denotes scalar control information required for local training. Its size is negligible relative to the model state. Throughout the paper, *Comm.* reports the number of client-update

exchanges rather than byte volume. In the compression experiments, $\kappa_{\mathcal{C}}$ is held fixed within each baseline–ISP comparison, so the reported exchange count isolates the additional effect of adaptive participation.

For client-side ObjectiveEval, each client $i \in C$ receives the evaluated state $\boldsymbol{x}$ and returns one scalar local-objective value $\ell_i(\boldsymbol{x})$ (Algorithm 6, lines 18–24). Therefore,

$$\mathrm{size}_{\uparrow}(\textsc{ObjectiveEval}(\boldsymbol{x}, C)) = \sum_{i \in C} \mathrm{size}(\ell_i(\boldsymbol{x}))\,,$$

and

$$\mathrm{size}_{\downarrow}(\textsc{ObjectiveEval}(\boldsymbol{x}, C)) = \sum_{i \in C} \Big[\mathrm{size}(\boldsymbol{x}) + \mathrm{size}(\mathrm{ctrl}_i^{\mathrm{eval}})\Big] \approx |C|\,\mathrm{size}(\boldsymbol{x}).$$

When the server-side surrogate $\hat{f}$ is used, both directed sizes of ObjectiveEval are zero. As above, the scalar control messages and returned scalar losses are negligible compared with a model-sized transmission.

For each training round τ, let C_τ denote the multiset union of the client cohorts passed to ClientsUpdate. Thus, $C_\tau = C_{\tau+1}$ at a regular round and $C_\tau = C_{\tau+1} \uplus I_{\tau+1/2}$ when an intermediate ISP update is performed. Similarly, let $\widetilde{C}_\tau$ denote the multiset union of the client cohorts used by all client-side ObjectiveEval calls during round τ. The multiset cardinalities $|C_\tau|$ and $|\widetilde{C}_\tau|$ therefore count repeated transmissions arising from separate calls within the same round.

Definition C.1. We define the total uplink communication volume over a training run as

$$\mathrm{Comm} = \sum_{\tau=0}^{T-1} \left[\sum_{i \in C_\tau} \mathrm{size}(\mathbf{u}_i^\tau) + \sum_{i \in \widetilde{C}_\tau} \mathrm{size}(\ell_i^\tau) \right].$$

Because ObjectiveEval returns only one scalar per queried client,

$$\mathrm{Comm} \approx \sum_{\tau=0}^{T-1} \sum_{i \in C_\tau} \mathrm{size}(\mathbf{u}_i^\tau) = \mathrm{size}(\boldsymbol{x}) \sum_{\tau=0}^{T-1} |C_\tau|.$$

Accordingly, we report *Comm.* as the number of client updates returned to the server:

$$\mathrm{Comm.} = \sum_{\tau=0}^{T-1} |C_\tau|.$$

This includes both the standard training updates in Algorithm 3 and the intermediate updates collected by Ext-ISP at adaptation rounds in Algorithm 4. The primary communication bottleneck in our setting is the client-to-server transfer captured by $\mathrm{size}_{\uparrow}(\cdot)$, since clients upload model-sized updates, whereas objective evaluations return only scalars. Accordingly, *Comm.* is used as the main measure of communication efficiency throughout the paper. Different CSS and federated optimization methods may alter local client behavior and the resulting client–server interaction pattern; evaluating each method under its native interaction structure is standard in FL experiments. For compression experiments, communication is compared within the same compressor configuration.

Definition C.2. We define the total downlink communication volume over a training run as

$$\mathrm{Downlink} = \sum_{\tau=0}^{T-1} \Big[\mathrm{size}_{\downarrow}(\textsc{ClientsUpdate}(\boldsymbol{x}^\tau, C_\tau, \mathcal{C})) + \mathrm{size}_{\downarrow}\Big(\textsc{ObjectiveEval}(\widetilde{\boldsymbol{x}}^{\,\tau}, \widetilde{C}_\tau)\Big) \Big].$$

Up to negligible scalar control messages, this gives

$$\mathrm{Downlink} \approx \mathrm{size}(\boldsymbol{x}) \sum_{\tau=0}^{T-1} \Big(|C_\tau| + |\widetilde{C}_\tau| \Big).$$

The first term corresponds to distributing server states before local training, while the second accounts for transmitting states required for client-side objective evaluation. When the server-side surrogate $\hat{f}$ is available, the second term vanishes and, under the fixed-size approximation above, $\mathrm{Downlink} \approx \mathrm{Comm}$. These client–server interactions also induce method-dependent local computation and orchestration costs, which are conventionally assessed through runtime and are reported in our tables as *Avg. Time*. The optional $\mathcal{GP}$ add-on is intended to mitigate the second downlink component and the associated client burden caused by repeated objective inferences.

Computation Time. We report computational cost through measured end-to-end training time rather than an additive per-client expression $\mathrm{Comp}(\tau)$. The measured time includes local optimization in CLIENTSUPDATE, local-data traversal during client-side OBJECTIVEEVAL, server aggregation, intermediate ISP evaluation, and orchestration overhead. For the $\mathcal{GP}$ configuration, it additionally includes context construction, posterior updates, and GP fitting. Wall-clock time is more representative of the implemented simulator because clients are executed concurrently in fixed-size batches, so elapsed time is not equal to the sum of individual client execution times. Consequently, the measured duration of a communication round exhibits an explicit dependence on the number of clients involved.

The measurement of time was performed up to the best training round, which was identified based on the loss observed on the clients' validation datasets. Upon receiving the global model, each client first conducted validation before commencing local training.

Table 6 reports the computational overhead introduced by the ISP technique as a result of intermediate communication (see Research Question 1). We present the average overhead of full-client intermediate communication and quantify its relative contribution in comparison to a standard training round, given that ISP is positioned as a strategy with communication triggered every $\Delta = 20$ rounds. We see that overall overhead is no more than 10% which is neglect due to communication savings.

Table 7 reports the computational overhead introduced by the ISP technique under a different relaxing strategies: increasing step size, adding surrogate loss, inreasing Δ. With this strategies we can almost neglect computational overhead in overall training.

Table 6: Computational overhead for ISP intermediate communication and overal FL training.

Method	**Intermediate Computational Overhead**	**Overall Overhead**
ISP-UNIFORM	$\times$1.633	$\times$1.032
ISP-FEDCBS	$\times$2.244	$\times$1.062
ISP-DELTA	$\times$1.443	$\times$1.026
ISP-POW	$\times$3.380	$\times$1.093
ISP-FEDCOR	$\times$1.308	$\times$1.018

Table 7: Computational overhead for ISP intermediate communication and overal FL training under sparsing intermediate communication.

Method	**Intermediate Computational Overhead**	**Overall Overhead**
ISP-UNIFORM	$\times$1.633	$\times$1.032
ISP-UNIFORM with $\hat{f}$	$\times$1.144	$\times$1.006
ISP-UNIFORM with $w = 3$	$\times$1.213	$\times$1.021
ISP-UNIFORM with $\Delta = 100$	$\times$1.633	$\times$1.013

Communication Analysis. Table 8 summarizes the communication savings and overheads of ISP across all domains and experiments. The intermediate search introduces additional client participation, but the total number of communications still decreases in most settings because the selected client count in subsequent rounds becomes smaller. This effect is especially pronounced in large-scale settings, where adapting the participation budget avoids repeatedly communicating with an overly conservative fixed number of clients. This overhead is tunable through the auxiliary expansion ratio p and the adaptation window Δ: Appendix E shows that reducing the auxiliary pool or running the intermediate search less frequently lowers the overhead while preserving downstream performance.

Table 8: Communication efficiency and overhead for ISP technique.

Domain	Method	Overall Communications	Intermediate Communications	Overall Overhead	Overall Savings
IMAGE	UNIFORM	17,767 ± 1,937	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-UNIFORM	14,343 ± 556	6,241 ± 316	×1.43 ± 0.03	**-19%** ± 9%
	POW	10,347 ± 493	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-POW	10,864 ± 546	2,223 ± 135	×1.19 ± 0.01	+5% ± 6%
	FEDCOR	19,360 ± 557	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-FEDCOR	15,037 ± 468	7,766 ± 381	×1.51 ± 0.03	**-22%** ± 4%
	FEDCBS	19,207 ± 837	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-FEDCBS	16,611 ± 816	6,744 ± 332	×1.40 ± 0.03	**-14%** ± 6%
	DELTA	19,700 ± 191	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-DELTA	14,791 ± 214	7,312 ± 367	×1.49 ± 0.03	**-24%** ± 2%
	OORT	19,500 ± 724	0 ± 0	×1.00 ± 0.00	0% ± 0%
	ISP-OORT	15,488 ± 611	6,613 ± 352	×1.30 ± 0.02	**-21%** ± 3%
	RANDK	25,040	0	×1.00	0%
	ISP-RANDK	17,162	6,213	×1.36	**-31%**
	TOPK	25,080	0	×1.00	0%
	ISP-TOPK	21,436	6,123	×1.29	**-15%**
	IMAGENET	1,520	0	×1.00	0%
	ISP-IMAGENET	1,344	300	×1.22	**-12%**
ECG	UNIFORM	106,400	0	×1.00	0%
	ISP-UNIFORM	34,975	8000	×1.24	**-67%**
TEXT	UNIFORM	18,000	0	×1.00	0%
	ISP-UNIFORM	13,015	3850	×1.30	**-28%**

C.5 SIMULATION AND COMPUTE RESOURCES

Our implementation is written in Python 3.10 and uses CUDA 12.8 with PYTORCH 2.6.0 on Debian GNU/Linux 12 (*bookworm*). All experiments in Section 5 and Appendices D–E are simulated on a single server with an AMD EPYC 7513 32-core CPU and two NVIDIA A100 SXM4 GPUs with 80GB memory each, interconnected via NVLink. A detailed description of libraries, versions, and reproduction scripts is provided in the codebase.

The federated simulation launches client worker processes through Python multiprocessing. Each client process has a unique rank and a bidirectional pipe endpoint for communication with the server. The central manager batches client execution to respect GPU-memory constraints. Unless explicitly stated otherwise, all clients receive identical CPU/GPU allocations; system heterogeneity is introduced only in the OORT experiment described in Appendix C.3.

For image and ECG experiments, each client's local data are split into training and validation subsets in a stratified 4/1 ratio. This local validation split is used to estimate convergence-related quantities and preserves class proportions under heterogeneous Dirichlet partitioning. If a client has fewer than four samples of a minority class, these samples are kept in the training split to avoid unstable stratification. For SHAKESPEARE, we instead use the chronological split described in Appendix C.2.

Hyperparameters. To ensure fair comparisons, we keep training hyperparameters fixed across methods within each dataset. Table 9 summarizes the main parameters; complete configurations, including normalization constants, random seeds, and run commands, are provided in the codebase.

C.6 $\mathcal{GP}$ DETAILS

Algorithm 8 presents the compact posterior-guided query procedure. The implementation additionally refines the response-query step without changing this overall workflow. A newly selected candidate count is first evaluated with $N_{\min}$ sampled cohorts; if the posterior continues to select the

Table 9: Parameters for FL experiments across datasets.

Parameter	CIFAR-10	ECG	Tiny-ImageNet	Shakespeare
Input Type	RGB Image	12-channel Signal	RGB Image	Character sequence
Num Classes	10	6 multi-label	200	Character vocabulary
Model	ResNet-18	ResNet1D-18	Swin-T	2-layer LSTM
Loss Function	CrossEntropyLoss	BCEWithLogitsLoss	CrossEntropyLoss	CrossEntropyLoss
Optimizer	Adam SGD for DELTA	Adam	AdamW	Adam
Learning Rate	Adam: 0.003 SGD: 0.003	0.003	0.001	0.01
Weight Decay	0	0	0.01	0
Batch Size	64	128	256	256
Preprocessing	$\mathcal{N}(\mu_1, \sigma_1)$	$\mathcal{N}(0, 1)$	$\mathcal{N}(\mu_2, \sigma_2)$	Character encoding
Augmentations	RandomCrop (32×32) HorizontalFlip	No	No	No

same unresolved count, its response is recomputed after adding w_N samples, up to $N_{\max} = N$. The refined estimate replaces the previous current-audit observation at that count rather than creating an additional GP observation, and its empirical response uncertainty is supplied to the likelihood as a fixed heteroskedastic noise variance. Accordingly, the budget Q limits the number of distinct candidate counts queried during an audit, while repeated selections allocate additional Monte Carlo depth to the current decision boundary. The confidence-feasible set and its fallback rule defined below specify the compact selection operations in lines 9 and 17 of Algorithm 8.

$\mathcal{GP}$ posterior sample. We approximate the restricted-pool response $\widehat{\Phi}_\tau(I_{\tau+1/2}, m)$ with an exact $\mathcal{GP}$ defined over the round-dependent contexts $\boldsymbol{z}_\tau(m)$. After feature standardization, we use a constant mean function c_θ and an ARD Matérn-5/2 kernel $\mathbf{K}_\theta$. Let $\widetilde{\mathbf{Z}}_\tau^{(q)}$, $\widetilde{\mathbf{y}}_\tau^{(q)}$, and $\widetilde{\mathbf{V}}_\tau^{(q)} = \operatorname{diag}\left(v_1, \ldots, v_{J_q}\right) + \sigma_{\text{add}}^2 \mathbf{I}$. collect the contexts, response labels, and observation-noise variances from the historical response dataset and the active observations obtained during the first q posterior-path steps. For the candidate-context matrix $\mathbf{Z}_\tau = [\boldsymbol{z}_\tau(m)]_{m \in \mathbf{m}_\tau}$,the latent posterior moments are

$$\boldsymbol{\mu}_\tau^{(q)} = c_\theta(\mathbf{Z}_\tau) + \mathbf{K}_\theta\left(\mathbf{Z}_\tau, \widetilde{\mathbf{Z}}_\tau^{(q)}\right) \mathbf{A}_q^{-1}\left(\widetilde{\mathbf{y}}_\tau^{(q)} - c_\theta\left(\widetilde{\mathbf{Z}}_\tau^{(q)}\right)\right),$$

$$\boldsymbol{\Sigma}_\tau^{(q)} = \mathbf{K}_\theta\left(\mathbf{Z}_\tau, \mathbf{Z}_\tau\right) - \mathbf{K}_\theta\left(\mathbf{Z}_\tau, \widetilde{\mathbf{Z}}_\tau^{(q)}\right) \mathbf{A}_q^{-1} \mathbf{K}_\theta\left(\widetilde{\mathbf{Z}}_\tau^{(q)}, \mathbf{Z}_\tau\right),$$

where

$$\mathbf{A}_q = \mathbf{K}_\theta\left(\widetilde{\mathbf{Z}}_\tau^{(q)}, \widetilde{\mathbf{Z}}_\tau^{(q)}\right) + \widetilde{\mathbf{V}}_\tau^{(q)}, \qquad \boldsymbol{\sigma}_\tau^{(q)} = \sqrt{\operatorname{diag}\left(\boldsymbol{\Sigma}_\tau^{(q)}\right)}.$$

We define the confidence-feasible set

$$\mathcal{F}_\tau^{(q)} = \left\{ m \in \mathbf{m}_\tau : \mu_\tau^{(q)}(m) + \beta_\sigma \sigma_\tau^{(q)}(m) < 0 \right\}.$$

If $\mathcal{F}_\tau^{(q)} \neq \varnothing$, the next query is its smallest element $m_q = \min \mathcal{F}_\tau^{(q)}$. Otherwise, we query the candidate with the smallest confidence-adjusted predicted response:

$$m_q = \arg\min_{m \in \mathbf{m}_\tau} \left[\mu_\tau^{(q)}(m) + \beta_\sigma \sigma_\tau^{(q)}(m) \right].$$

Adaptive posterior path. A newly selected candidate count m_q is initially evaluated using $N_{\min} = 3$ independently sampled candidate cohorts. After conditioning on the resulting response estimate, the acquisition rule is evaluated again. If the updated posterior continues to select the same unresolved count m_q, we sample an additional $w_N = 2$ candidate cohorts and recompute its response using all accumulated samples. Thus, its Monte Carlo depth follows

$$N_q(m_q) \leftarrow \min\left\{N_q(m_q) + w_N,\, N_{\max}\right\}, \qquad N_{\min} = 3, \quad w_N = 2, \quad N_{\max} = N.$$

The refined estimate replaces the previous current-audit observation at m_q instead of being added as a separate GP observation. This allocation avoids spending the full Monte Carlo budget on candidate counts that cease to be decision-relevant after the first posterior update, while concentrating additional evaluations around a persistent decision boundary. Since each additional candidate evaluation requires transmitting the candidate state to the evaluator cohort, the adaptive depth directly reduces the evaluator component of the downlink cost in Definition C.2. The posterior path terminates after at most Q distinct candidate counts have been evaluated or after the currently selected count reaches $N_{\max}$. We then augment $\mathcal{D}_{\mathrm{GP}}$ with the active current-audit observations, refit θ by marginal likelihood maximization, and obtain $m^\star$ from the final posterior using the same feasible-set rule.

Observation uncertainty. For a posterior-path query q, let $\boldsymbol{x}_q^{(\tau+1/2)n}$ denote the candidate model produced by the n-th sampled cohort C^n, where $n = 1, \ldots, N_q$. Its observation sampled variance $\hat{v}_{N,q}$ is estimated from the each sample of the mean next-loss changes $\{r_{q,n}\}$

$$r_{q,n} = \frac{1}{|E_\tau|} \sum_{i \in E_\tau} \left[\ell_i \left(\boldsymbol{x}_q^{(\tau+1/2)n} \right) - \ell_i(\boldsymbol{x}^\tau) \right].$$

To additionally characterize the response dispersion across evaluators, we define

$$d_{q,i} = \frac{1}{N_q} \sum_{n=1}^{N_q} \left[\ell_i \left(\boldsymbol{x}_q^{(\tau+1/2)n} \right) - \ell_i(\boldsymbol{x}^\tau) \right], \qquad i \in E_\tau,$$

and compute an empirical evaluator-dispersion term $\widehat{v}_{E,q}$ from $\{d_{q,i}\}_{i \in E_\tau}$. Since E_τ is selected deterministically, this term is used as an uncertainty proxy rather than as a design-unbiased estimator of evaluator-selection variance. The resulting heteroskedastic observation variance is $v_q = \max\{\hat{v}_{N,q} + \widehat{v}_{E,q}, \varepsilon\}$, and is supplied to the GP likelihood as the fixed noise parameter associated with the observation at $\boldsymbol{z}_\tau(m_q)$. This is a dispersion-based likelihood proxy. For a fixed number N_q of independent cohort responses, the sampling variance of their mean is $\mathrm{Var}(r_{q,n} \mid I, E_\tau)/N_q$; a sample variance of individual responses must be divided by N_q to estimate that quantity. The displayed likelihood proxy, which also includes evaluator dispersion and is used along an adaptive posterior path, is not asserted to be a calibrated sampling variance or a simultaneous confidence bound.

Proxy-coverage evaluator cohort. We construct one evaluator cohort E_τ per adaptation round using the bias-update information proposed as a proxy for client label distributions (Chen & Vikalo, 2024). For each client i, the available or most recent update produces

$$\widehat{\mathbf{p}}_i = \mathrm{softmax}\left([\mathbf{u}_i]_{\mathrm{bias}} / T \right),$$

where $\widehat{p}_{i,c}$ approximates the fraction of class c in the local dataset and n_i is the corresponding local dataset size. Unlike the entropy-weighted sampling policy in (Chen & Vikalo, 2024), we use these distributions to construct a deterministic cohort satisfying a proxy class-count coverage target. Starting from the class-wise $\mathbf{d} = k\mathbf{1}_C$, we iteratively select

$$i^\star = \arg \max_{i \in [M] \setminus E_\tau} \sum_{c=1}^{C} \min\{n_i \widehat{p}_{i,c},\, d_c\},$$

and update $E_\tau \leftarrow E_\tau \cup \{i^\star\}$, $\mathbf{d} \leftarrow [\mathbf{d} - n_{i^\star} \widehat{\mathbf{p}}_{i^\star}]_+$. If $\sum_{i=1}^{M} n_i \widehat{p}_{i,c} \geq k$ for every class c, the finite greedy procedure terminates when all proxy deficits vanish, yielding at least k estimated examples per class. This feasibility condition is required for the stated coverage property; an absent or insufficiently represented class cannot meet the target through any evaluator selection. The resulting E_τ is fixed throughout the posterior path and is used to evaluate both the baseline and all candidate models, producing paired objective responses. The effect of this proxy-coverage construction and its comparison with alternative evaluator-selection policies are analyzed in Appendix D.2.

For SHAKESPEARE, the class-wise minimum-coverage rule is unnecessarily restrictive because of the large character vocabulary ($C = 127$) and the diffuse proxy distributions across characters. We instead target the population-level proxy distribution $\overline{\mathbf{p}} = M^{-1} \sum_{i=1}^{M} \widehat{\mathbf{p}}_i$ and greedily add the client minimizing $\left\| |E_\tau \cup \{i\}|^{-1} \sum_{j \in E_\tau \cup \{i\}} \widehat{\mathbf{p}}_j - \overline{\mathbf{p}} \right\|_1$.

D Additional Experiments

This section presents experiments and analyses. Appendix D.1 evaluates compatibility with federated optimization algorithms, while Appendix D.2 studies the optional $\mathcal{GP}$ add-on, including its response and evaluator-cohort approximations. Appendix D.3 reports the Tiny-ImageNet experiment, Appendix D.4 studies compatibility with gradient compression, and Appendix D.5 compares ISP with linear client-count schedules. Appendix D.6 provides additional client-count dynamics and convergence plots, while Appendix D.7 analyzes fixed baseline participant counts.

D.1 Compatibility with Federated Optimization Algorithms

We further evaluate whether ISP depends on the underlying federated optimization algorithm. We combine the client-count controller with FedAvg, FedProx, FedNova, and FedDyn under the same CIFAR-10 setup with Uniform client sampling. In each case, ISP modifies only the requested number of participating clients, while the local objective and optimizer-specific server update remain unchanged.

Table 10: Compatibility of ISP with federated optimizers on CIFAR-10; communications include intermediate ISP exchanges.

Method	**Communications** (↓)	**Test Loss** (↓)	**Accuracy** (↑)	**Time** (↓)
FedAvg	17,767	0.461	0.840	08h 54m 49s
ISP-FedAvg	**14,343**	0.477	0.834	06h 14m 21s
FedProx	16,500	0.448	0.848	07h 36m 18s
ISP-FedProx	**14,254**	0.450	0.848	06h 29m 16s
FedNova	16,960	0.431	0.853	07h 23m 49s
ISP-FedNova	**16,604**	0.434	0.852	07h 29m 12s
FedDyn	29,440	0.539	0.792	13h 41m 52s
ISP-FedDyn	**18,098**	0.597	0.780	08h 47m 41s

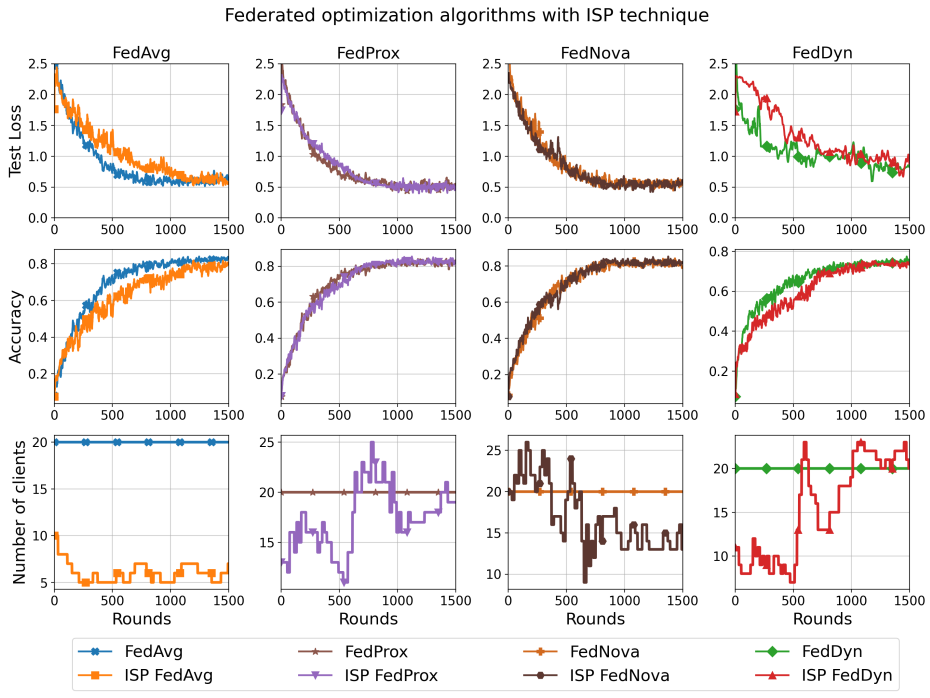


Figure 4: Optimizer-specific test loss, accuracy, and client-count dynamics on CIFAR-10; intermediate $\tau + 1/2$ communications are omitted.

Table 10 and Figure 4 show that the adaptive client-count rule remains stable across optimization algorithms with different local and server-update structures. For FEDPROX, ISP reduces communication while preserving the downstream metrics, and the FEDNOVA configuration exhibits comparable behavior with a smaller communication reduction. With FEDDYN, the controller produces a substantially lower communication profile with a moderate decrease in downstream performance. The distinct client-count trajectories further indicate that ISP reacts to the optimization dynamics induced by each algorithm rather than imposing a common participation schedule. Overall, these results support the use of ISP as an optimizer-agnostic client-count controller that does not require modifications to the underlying federated optimization rule.

D.2 GAUSSIAN PROCESS ADD-ON

Algorithms 7 and 8 in Appendix A specify the $\mathcal{GP}$ context construction, evaluator-cohort selection, and posterior-guided query procedure, while the corresponding response construction and additional implementation details are provided in Appendix C.6.

End-to-end results. In addition to FEDAVG on CIFAR-10 and SHAKESPEARE datasets in the main part, we consider FEDCBS as an alternative client-selection strategy and FEDNOVA as an alternative federated optimization algorithm on CIFAR-10. We use proxy coverage $k = 400$, temperature $T = 0.0075$, query budget $Q = 6$, maximum Monte Carlo depth $N = 10$, and $\beta_\sigma = 0$. Table 11 reports the end-to-end results.

Table 11: End-to-end comparison of the base method, ISP, and the GP-guided variant across image and text settings. Communications count client-update exchanges. Relative downlink additionally includes evaluator-cohort model transmissions and is normalized by the corresponding base method. ISP-GP uses proxy coverage $k = 400$.

Domain	Method	Communications (↓)	Test Loss (↓)	Accuracy (↑)	Time (↓)	Rel. downlink (↓)
IMAGE	FEDAVG	17,767	0.461	**0.842**	08h 54m 49s	×1.00
	ISP-FEDAVG	14,343	0.464	0.836	06h 14m 21s	—
	ISP-GP-FEDAVG	**13,620**	0.469	0.832	**06h 01m 36s**	×1.29
	FEDNOVA	16,960	**0.431**	**0.853**	07h 23m 49s	×1.00
	ISP-FEDNOVA	16,604	0.434	0.852	07h 29m 12s	—
	ISP-GP-FEDNOVA	**16,558**	0.440	0.848	**06h 51m 54s**	×1.26
	FEDCBS	19,207	0.507	0.830	11h 18m 09s	×1.00
	ISP-FEDCBS	16,611	**0.487**	**0.836**	09h 28m 14s	—
	ISP-GP-FEDCBS	**14,711**	0.533	0.828	**08h 13m 46s**	×1.32
TEXT	UNIFORM	18,000	1.676	**0.546**	15h 10m 25s	×1.00
	ISP-UNIFORM	13,015	**1.623**	0.541	13h 27m 48s	—
	ISP-GP-UNIFORM	**12,261**	1.684	0.532	**13h 05m 44s**	×1.29

Across the considered configurations, ISP-GP preserves the performance of the corresponding ISP runs while consistently reducing communications and execution time. The results remain stable across client-selection strategies, federated optimization algorithms, and data domains, showing that the $\mathcal{GP}$ approximation is not specific to a single FEDAVG setup on CIFAR-10 and can be applied on top of common scenarios on par with ISP technique. The GP slightly underestimates the restricted-pool response, causing the posterior path to terminate earlier and shifting the communication–accuracy trade-off toward communication efficiency. This yields substantially lower communication and runtime costs at the expense of small accuracy reductions in some configurations.

$\mathcal{GP}-$approximation. We additionally compare the GP approximation with an oracle estimate of the response for candidate cohorts sampled from the auxiliary pool and evaluated on all clients. At intermediate round $\tau + 1/2$, each candidate size m is evaluated using independent cohorts C_n:

$$\widehat{\Phi}^{(n)}_{M,\tau}(m) = \mathrm{ObjectiveEval}\big(x^{(\tau+1/2)_n}, [M]\big) - \mathrm{ObjectiveEval}\big(x^{\tau}, [M]\big),$$

$$\widehat{\Phi}_{M,\tau}(m) = \frac{1}{N}\sum_{n=1}^{N} \widehat{\Phi}^{(n)}_{M,\tau}(m).$$

Table 12: Approximation of the full-client server-oracle response by the GP. Here, SA_Φ denotes response-sign agreement, BE is the decision-boundary error, and $\mathrm{Cov}_{.95}$ is the empirical coverage of the post-hoc 95% GP interval. For each metric, we report the first quartile (Q_1), median, and third quartile (Q_3) across adaptation rounds. Boundary error is computed only for audits in which the evaluated oracle points resolve a sign change. The final metric is the fraction of oracle responses contained in the post-hoc 95% GP interval.

Domain	Method	RMSE_Φ (↓)			SA_Φ (↑)			BE (↓)			$\mathrm{Cov}_{.95}$ (↑)		
		Q_1	**Median**	Q_3	Q_1	**Median**	Q_3	Q_1	**Median**	Q_3	Q_1	**Median**	Q_3
IMAGE	ISP-GP-FEDAVG	0.103	**0.166**	0.257	0.600	**0.800**	1.000	1.0	**1.0**	1.25	0.750	**0.900**	1.000
	ISP-GP-FEDNOVA	0.029	**0.071**	0.130	0.400	**0.750**	1.000	1.0	**1.5**	3.0	0.600	**0.900**	1.000
	ISP-GP-FEDCBS	0.052	**0.076**	0.226	0.400	**0.800**	1.000	1.0	**1.0**	2.0	0.650	**0.800**	1.000
TEXT	ISP-GP-UNIFORM	0.005	**0.010**	0.027	0.500	**0.700**	1.000	1.0	**1.0**	2.0	0.700	**0.900**	1.000

The oracle evaluates up to five candidate sizes in a neighborhood of the final sign change predicted by the GP posterior, using up to $N = 40$ independent candidate cohorts at each size for more robust approximation.

We evaluate the GP approximation along three dimensions. Let RMSE_Φ and SA_Φ denote, respectively, the response RMSE and sign agreement between the GP posterior mean $\mu^\star_\tau(m)$ and the server-oracle response $\widehat{\Phi}_{M,\tau}(m)$ over the oracle-evaluated candidate sizes. We measure the decision-boundary error as $\mathrm{BE} = |m^\star_{\mathrm{GP}} - m^\star_M|$, where $m^\star_{\mathrm{GP}} = \min\{m : \mu^\star_\tau(m) < 0\}$ and $m^\star_M = \min\{m : \widehat{\Phi}_{M,\tau}(m) < 0\}$. Finally, $\mathrm{Cov}_{.95}$ denotes the fraction of server-oracle responses contained in the post-hoc interval $\mu^\star_\tau(m) \pm 1.96\sigma^\star_\tau(m)$. Table 12 reports the first quartile, median, and third quartile of these metrics across intermediate rounds $\tau + 1/2$.

The GP consistently tracks the full-client server-oracle response across client-selection strategies, federated optimizers, and domains. The median SA_Φ ranges from 0.70 to 0.80, while the median BE remains between one and 1.5 clients. The median RMSE_Φ ranges from 0.071 to 0.166 for the image-domain configurations and reaches 0.010 on Shakespeare. The median $\mathrm{Cov}_{.95}$ ranges from 0.80 to 0.90. Overall, the GP preserves the local response structure and reliably identifies the server-oracle decision boundary.

Evaluator-cohort approximation. We isolate the evaluator-cohort approximation from the remaining components of ISP-GP. Checkpoints were randomly sampled from three training regimes before inspecting the cohort-approximation metrics. The initial stage is represented by the FEDAVG checkpoint at round 16. The middle stage includes FEDAVG, FEDNOVA, and FEDCBS at rounds 995, 408, and 574, respectively, whereas the late stage uses rounds 1638, 814, and 867. These method-specific checkpoints represent comparable positions in their respective convergence trajectories rather than identical numerical rounds. At each checkpoint, we evaluate candidate counts in a neighborhood of the contemporaneous GP decision boundary. Across the seven checkpoints, these neighborhoods contain 72 candidate counts in total; each checkpoint–candidate-count pair is treated as one evaluation case.

We compare PROXYCOHORT with five evaluator-selection policies. UNIFORM samples clients uniformly without replacement, while ENTROPY-WEIGHTED follows the proxy-entropy sampling strategy (Chen & Vikalo, 2024). PROXY BEST-OF generates 128 uniformly sampled cohorts and retains the cohort whose mean proxy-label distribution is closest to the population proxy mean. ORACLE BEST-OF applies the same procedure using the true client label distributions. Finally, ORACLE MIN-k finds a minimum-cardinality cohort satisfying true class-count coverage and then balances the covered class counts among panels of that cardinality. For each checkpoint and candidate count, the four fixed-size baselines use exactly the same cohort size selected by PROXYCOHORT; ORACLE MIN-k selects its own cohort size. The stochastic policies are evaluated over 1000 independently generated panels.

For a candidate count m, we compare the evaluator response $\widehat{\Phi}_{E_\tau}(m)$ with the corresponding full-client response $\widehat{\Phi}_{[M]}(m)$. We report RMSE_Φ and SA_Φ as above, together with $Q_{.95}$, the 95th percentile of the absolute response error. Sign agreement indicates whether the two responses place m on the same side of the decision boundary $\hat{\Phi} = 0$. This experiment isolates evaluator-cohort

selection: the GP kernel, posterior fitting, acquisition path, and candidate-cohort estimator are held fixed. Table 13(a) reports the aggregate policy comparison over the 72 evaluation cases.

Compared with UNIFORM, PROXYCOHORT reduces the median RMSE_Φ from 0.202 to 0.111 and $Q_{.95}$ from 0.411 to 0.206, while increasing SA_Φ from 0.798 to 0.920. Its response errors are also close to the privileged oracle references. To verify that the aggregate result is not driven by a particular training stage or configuration, Table 13(b) disaggregates PROXYCOHORT by checkpoint. Each row reports medians over the candidate counts audited around the corresponding GP decision boundary. The checkpoint-level results confirm that the aggregate behavior is preserved across configurations and training stages: SA_Φ remains at least 0.800 at every sampled checkpoint, while each evaluator cohort contains only five or six clients.

Table 13: Evaluator-cohort approximation. Panel (a) compares evaluator-selection policies over 72 checkpoint–candidate-count pairs; values are medians over the pairs. Panel (b) reports stage-wise PROXYCOHORT results, with each row summarized over candidate counts around the corresponding GP decision boundary. Fixed-size baselines match the cohort size selected by PROXYCOHORT; oracle policies use privileged true-label information. Bold denotes the best non-oracle result in panel (a).

(a) Evaluator-policy comparison

Policy	$\lvert E_\tau\rvert$	RMSE_Φ (↓)	$Q_{.95}$ (↓)	SA_Φ (↑)
UNIFORM	6	0.202	0.411	0.798
ENTROPY-WEIGHTED	6	0.180	0.370	0.811
PROXY BEST-OF	6	0.137	0.282	0.852
ORACLE BEST-OF	6	0.099	0.198	0.894
PROXYCOHORT	6	**0.111**	**0.206**	**0.920**
ORACLE MIN-k	5	0.073	0.132	0.900

(b) Stage-wise PROXYCOHORT results

Stage	Setup	Round	$\#m$	$\lvert E_\tau\rvert$	RMSE_Φ(↓)	$Q_{.95}$ (↓)	SA_Φ (↑)
Initial	FEDAVG	16	10	6	0.104	0.195	0.817
Middle	FEDAVG	995	10	6	0.133	0.215	0.900
	FEDNOVA	408	12	5	0.083	0.154	0.967
	FEDCBS	574	11	6	0.121	0.262	0.800
Late	FEDAVG	1638	10	5	0.199	0.346	0.942
	FEDNOVA	814	10	6	0.136	0.266	0.848
	FEDCBS	867	9	5	0.094	0.199	0.920

Finally, we vary the PROXYCOHORT coverage threshold k along the recorded ISP-GP-FEDAVG posterior path through round 1721, comprising 30 audits and 526 candidate queries. Table 14 jointly reports evaluator-cohort composition, proxy and true-label coverage, and end-to-end performance. The reported minimum counts are the smallest class-wise counts accumulated by the selected cohort, and *True/Proxy* denotes the ratio of the true minimum count to its proxy counterpart.

Table 14: Effect of the PROXYCOHORT coverage threshold k on evaluator-cohort composition and end-to-end ISP-GP-FEDAVG performance. Cohort sizes and minimum class counts are summarized across 30 audits.

k	$\lvert E_\tau\rvert$				Median minimum class count		True/ Proxy	Comm. (↓)	Loss (↓)	Acc. (↑)	Rel. downlink (↓)
	Min	Median	Mean	Max	Proxy	True					
50	2	3	3.07	4	60.8	22.0	0.362	13,241	0.472	0.826	×1.15
100	3	4	3.77	5	118.8	36.5	0.307	14,244	0.470	0.825	×1.18
400	5	6	6.07	16	420.9	173.0	0.411	13,620	0.469	0.832	×1.29
1000	9	11	12.07	40	1019.0	457.5	0.449	14,127	0.465	0.839	×1.58
2500	25	38	52.80	100	2500.8	2049.0	0.819	14,461	0.462	0.840	×2.17

The proxy distributions overestimate the minimum class support measured from true labels. At the default threshold $k = 400$, PROXYCOHORT uses a median of six evaluators, while the median true minimum count is 173.0 compared with the proxy estimate of 420.9. Increasing k enlarges the evaluator cohort and raises the True/Proxy ratio from 0.411 at $k = 400$ to 0.819 at $k = 2500$. Over the same range, accuracy increases from 0.832 to 0.840 and loss decreases from 0.469 to 0.462, while relative downlink rises from $\times 1.29$ to $\times 2.17$. Thus, k directly controls the trade-off between evaluator cost, proxy-coverage fidelity, and downstream performance.

D.3 TINY-IMAGENET

We evaluate ISP on the 200-class TINY-IMAGENET task using visual transformer SWIN-T and Dirichlet partitioning with $\alpha = 0.5$; the remaining dataset and optimization settings are reported in Appendix C.2 and Table 9. To limit the cost of this larger vision setting, we use FEDCOR as the underlying client-selection strategy, with ISP modifying only the requested cohort size. Table 15

shows that ISP reduces client-update communications from $1{,}520$ to $1{,}344$ (11.6%) while slightly improving test loss from 0.8849 to 0.8787 and accuracy from 0.7945 to 0.7963.

Table 15: TINY-IMAGENET results with FEDCOR client selection. Communications include intermediate ISP exchanges.

Method	**Communications** (↓)	**Test Loss** (↓)	**Accuracy** (↑)
FEDCOR	1,520	0.8849	0.7945
ISP-FEDCOR	**1,344**	**0.8787**	**0.7963**

D.4 GRADIENT COMPRESSION

We further evaluate whether ISP can be combined with communication-efficient update transmission. This experiment is complementary to the client-selection results in the main text: instead of changing which clients are selected, we reduce the size of each transmitted update and let ISP adapt the number of communicating clients on top of the compressed pipeline.

Table 16: ISP combined with gradient compression on CIFAR-10.

Method	**Communications** (↓)	**Test Loss** (↓)	**Accuracy** (↑)
RANDK	25,040	0.4506	0.8438
ISP-RANDK	**17,162**	**0.4447**	**0.8568**
TOPK	25,080	**0.4435**	**0.8650**
ISP-TOPK	**21,436**	0.4644	0.8538

We use the same CIFAR-10/RESNET-18 setup as in Section 5.1, but use a nearly uniform client partition with Dirichlet concentration $\alpha = 1000$. This choice isolates the interaction between ISP and update compression, since aggressive sparsification can be unstable in non-convex federated optimization under strong statistical heterogeneity (Sahu et al., 2021). We consider two standard sparsification operators: TOPK, which keeps the largest coordinates by magnitude, and RANDK, which keeps a random coordinate subset. Following the compression setup, we set K to 5% and 15% of the parameter dimension for TOPK and RANDK, respectively. All reported communication counts include the intermediate ISP communications.

Table 16 shows that ISP remains effective when the per-client update is already compressed. For RANDK, ISP reduces the number of client update calls by 31.5% and improves both test loss and accuracy. For TOPK, ISP still reduces communication by 14.5%, with only a modest accuracy decrease. These results indicate that adaptive client-count selection is not specific to uncompressed federated updates: it can be layered on top of update-compression mechanisms and still provide additional communication savings.

Figure 5 further illustrates that the selected client count evolves differently under different compression operators. These differences are consistent with sensitivity to the complete update pipeline, including the sampler, local training, and compression. Both the current response estimate and historical stabilization contribute to the resulting counts.

D.5 ADAFL COMPARISON

We continue to develop the results of the Table 19 by addressing to the ISP dynamic of m_τ. As we see from Figure 6, ISP-UNIFORM has a clear non-linear dynamic with growth of m_τ at the end, which is consistent with the ADAFL (Li et al., 2024) mentioned in Related Work (see Section 2). However, the proposed technique represents a linear schedule with a positive trend; as we approach the optimum, we become more sensitive to the heterogeneity of local updates. The original work exploits the schedule from the initial $m_0 = 10$ to the *full*-round communication $m_T = 100$ in a close CIFAR-10 setup. However, ADAFL does not directly address communication efficiency and, as we have seen from the results in Table 19, a significant increase in the number of participants does not provide the same gain in downstream performance that it takes away for communication

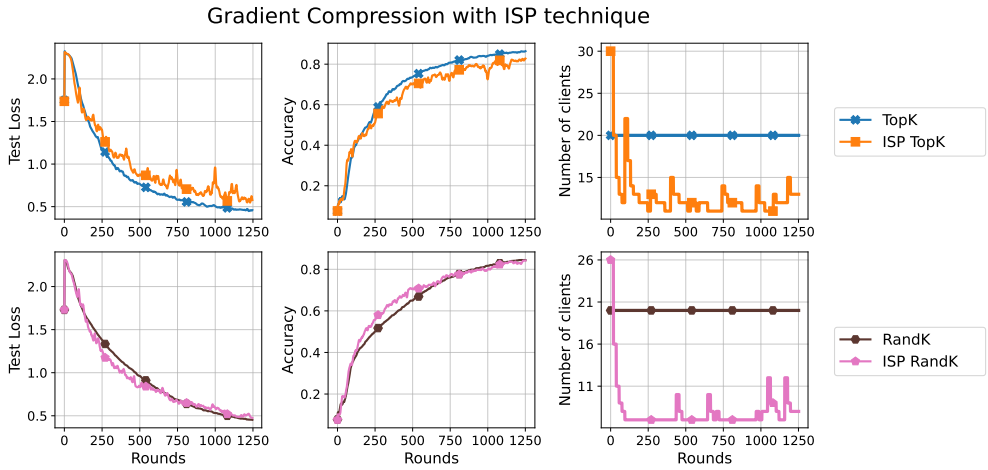


Figure 5: Test loss, accuracy, and client count dynamics for ISP under gradient-compression methods. We set K to 5% and 15% of the parameter dimension for TopK and RandK, respectively. For clarity, results are shown through round 1250, omitting $\tau + 1/2$ ISP communications.

exchanges. To address the communication bottleneck in addition to the baseline schedule, we will consider a communication-quality trade-off zone ranging clients from $m = 5$ to $m = 20$ every 50 rounds and then reaching a plateau. We call these schedules *full* and *econom*, respectively. Table 17 summarizes the results of the comparison.

Table 17: Comparison of ISP with linear scheduler dynamics.

Method	**Communications**	**Test Loss**	**Accuracy**
Uniform	17,767	0.461	0.842
AdaFL (econom)	16,191	0.508	0.827
AdaFL (full)	63,876	**0.421**	**0.855**
ISP-Uniform	**14,343**	0.464	0.836

We observe that the linear increasing *econom* scheduler does not achieve the same efficiency as the ISP technique. This demonstrates the need for an intelligent approach to adaptive choice of the number of clients. At the same time, there is a significant spread depending on the chosen schedule scenario and AdaFL does not relieve us from the choice of the boundaries m_0 and m_T and the scheduler step, which also significantly affect the federated learning process, as well as the number of clients in the current round.

D.6 Additional Plots

In this section, we present additional test loss, accuracy, and client-count dynamics for the experiments from Section 5. In addition to the trajectories reported below, Table 18 provides the complete five-run results corresponding to Table 2, additionally reporting the test-loss values omitted from the main text. Because the results in Table 2 are averaged over five runs, Figure 6 shows one representative run selected at random. For clarity, all plots omit the intermediate $\tau + 1/2$ ISP communications and visualize only the main training rounds.

Figure 6 shows that ISP adapts the number of participating clients in a strategy-dependent way. For several client-selection strategies, the selected count remains in a relatively small range for a large part of training, which is consistent with the communication–quality trade-off observed in Appendix D.7. However, unlike fixed low-budget baselines, ISP adjusts this count online according to the estimated objective decrease in (2). This allows it to retain downstream performance while reducing communication across a broad set of client-selection rules.

Table 18: Complete CIFAR-10 results corresponding to Table 2, additionally reporting test loss. Metrics are averaged over 5 runs. Communications count client updates, including intermediate ISP communications.

Method	**Communications** (↓)	**Test Loss** (↓)	**Accuracy** (↑)	**Avg. Time** (↓)
Uniform	17,767 ± 1,937	0.461 ± 0.009	0.842 ± 0.004	08h 54m 49s
ISP-Uniform	**14,343 ± 556**	0.464 ± 0.022	0.836 ± 0.013	**06h 14m 21s**
POW-D	**10,347 ± 493**	0.573 ± 0.012	0.812 ± 0.009	**03h 10m 33s**
ISP-POW-D	10,864 ± 546	0.526 ± 0.030	0.830 ± 0.011	03h 20m 14s
FedCor	19,360 ± 557	0.449 ± 0.017	0.848 ± 0.006	05h 44m 05s
ISP-FedCor	**15,037 ± 468**	0.439 ± 0.011	0.848 ± 0.010	**04h 34m 57s**
FedCBS	19,207 ± 837	0.507 ± 0.018	0.830 ± 0.007	11h 18m 09s
ISP-FedCBS	**16,611 ± 816**	0.487 ± 0.007	0.836 ± 0.001	**09h 28m 14s**
DELTA	19,700 ± 191	0.816 ± 0.019	0.721 ± 0.007	32h 05m 21s
ISP-DELTA	**14,791 ± 214**	0.814 ± 0.029	0.722 ± 0.010	**21h 45m 26s**
OORT	19,500 ± 724	0.478 ± 0.013	0.834 ± 0.007	08h 27m 30s
ISP-Oort	**15,448 ± 611**	0.454 ± 0.025	0.835 ± 0.011	**08h 02m 46s**

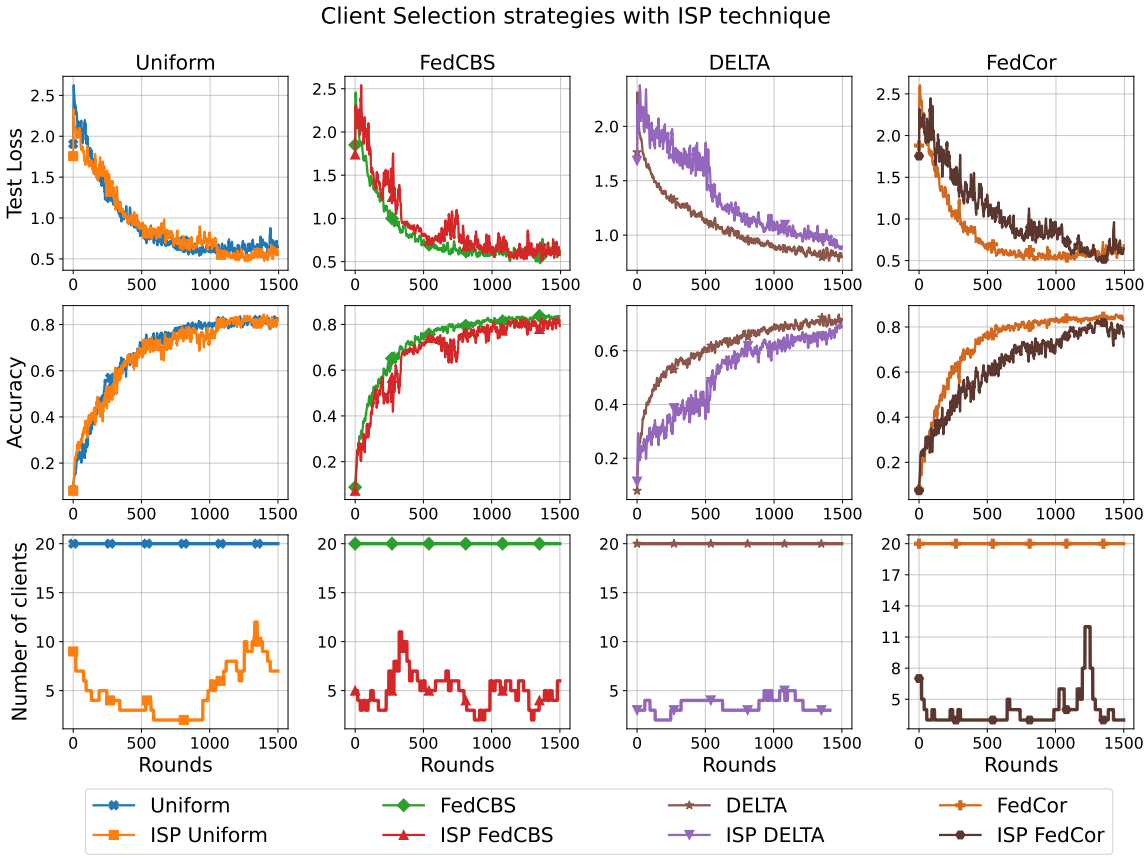


Figure 6: Test loss, accuracy, and client-count dynamics for ISP under different client selection strategies on CIFAR-10 (2nd run). For clarity, results are shown through round 1500, omitting $\tau + 1/2$ ISP communications.

The dynamics also illustrate that different client-selection strategies interact differently with adaptive client-count control. For example, methods such as FEDCOR and DELTA often operate with smaller selected client counts, which explains their strong communication savings but also reflects noisier optimization signals. In contrast, strategies that induce more structured or loss-biased subsets can require larger or more variable client counts during parts of training. These trends support the main claim that the number of clients should not be treated as a fixed external hyperparameter: even for

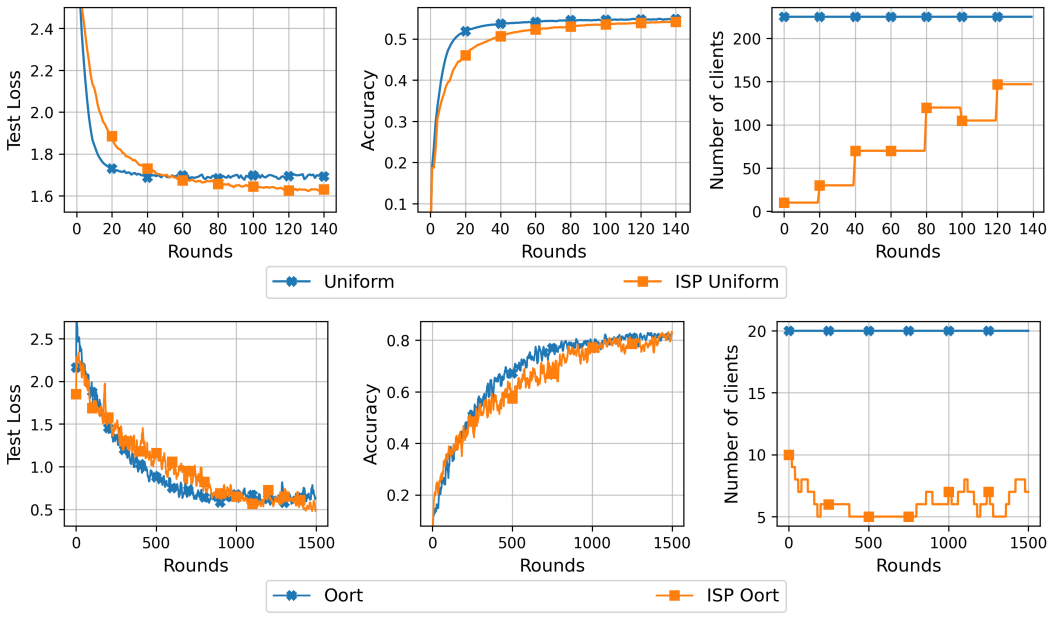


Figure 7: Loss, accuracy, and client-count dynamics for ISP with (top) UNIFORM sampling on SHAKESPEARE and (bottom) OORT sampling on CIFAR-10. Results are shown through rounds 140 and 1500, respectively, excluding $\tau + 1/2$ communications. ISP-UNIFORM and ISP-OORT reach optimum at rounds 132 and 1334, versus 80 and 975 for their baselines.

the same dataset and model, the appropriate participation budget depends on the underlying client-selection rule and changes over training.

Figure 7 (top) complements the text-domain results from Section 5.3. The trajectory shows that ISP starts from a larger participation budget and then adapts it over training, eventually operating with substantially fewer clients than the fixed-budget baseline. This behavior explains the communication savings reported in Table 4. At the same time, the loss and accuracy curves remain stable, which is consistent with the observation that ISP preserves downstream quality in the next-character prediction task while improving test loss and perplexity.

Figure 7 (bottom) provides the dynamics for the OORT experiment discussed in Section 5.2. The adaptive client-count trajectory confirms that ISP remains effective even when combined with a system-aware client selection strategy. Compared with fixed-budget OORT, ISP-OORT maintains a lower communication profile throughout most of training while preserving the final accuracy and improving test loss. This supports the main conclusion that ISP can be layered on top of heterogeneous client-selection mechanisms without requiring changes to their internal scoring rules.

D.7 BASELINE NUMBER OF PARTICIPANTS

To correctly determine the number of participants m in communication rounds for non-adaptive baselines, we performed an extensive comparison of it for the UNIFORM client selection strategy under the same conditions as described in Section 5.1 for Table 2. For $m = 20$, we take the average results from the main part of the article (see Table 2).

Table 2 reports the fixed-budget sweep, while Figure 8 visualizes the corresponding communication–accuracy trade-off together with the adaptive ISP operating point from Table 19.

- **Diminish returns beyond $m = 20$:** We see a clear non-optimal amount of communication costs for $m > 20$ to obtain the same level of target performance as for $m = 20$. This indicates the well-known statement about the suboptimal nature of a large number of participants in a setup with strong heterogeneity of local client data.
- **Communication-Quality trade-off:** As we move to $m < 20$ we cannot say for sure which of the values m is optimal in terms of communication-quality performance. At

Table 19: The number of participants m for downstream CIFAR-10 task with UNIFORM client selection strategy.

Number of clients	Communications	Test Loss	Accuracy
$m = 5$	12,065	0.624	0.791
$m = 10$	15,830	0.536	0.825
$m = 15$	18,315	0.473	0.838
$m = 20$	17,767	0.461	0.842
$m = 25$	30,350	0.458	0.842
$m = 40$	46,720	0.468	0.840
$m = 50$	62,250	0.474	0.836
ISP m_τ	14,343	0.464	0.836

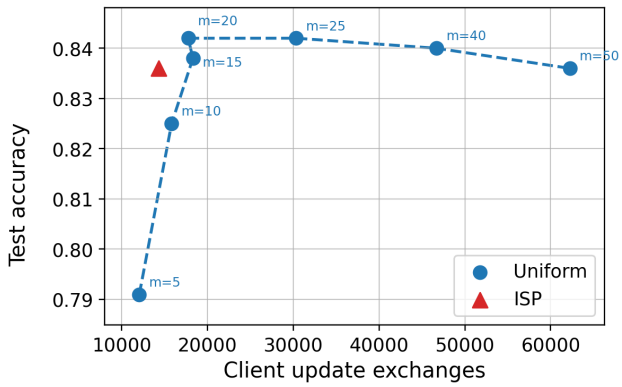


Figure 8: Communication–accuracy trade-off for fixed-budget UNIFORM sampling and adaptive ISP on CIFAR-10. Fixed-budget points are labeled by the corresponding client count m; the dashed line connects the discrete sweep as a visual guide.

> $m = 10$ we see an acceptable compromise: 11% reduction in communication exchanges while degrading the final accuracy by 1.7%. This is especially true in the case where communication efficiency plays a decisive role.

Despite the communication efficiency at $m = 10$, we preferred to choose $m = 20$, focusing on the best final target quality for ease of comparison: while the ISP technique demonstrates a significant reduction in communication costs, downstream performance remains at the same level. In the case of $m = 10$, we have the described trade-off, in which excessive communication savings lead to accuracy degradation, which requires additional mention in the main part of the work. Despite the communication-quality trade-off, ISP demonstrates clear superiority in communication savings and maintains downstream performance, which supports the similar results of Table 3 in the ECG domain.

E Ablation Study

In this section, we provide an extensive ablation study of the main components of the ISP technique. We discussed them in detail in the main part of Research Questions 1-3. The goal of this ablation is to systematically adjust these parameters compared to those chosen in the main part of the experiments (Table 1) to better understand their impact on communication-efficiency caused by the optimization of (2). By tuning these parameters, we also aim to identify which ones are more or less sensitive (see Appendix E.1-E.4). For hyperparameters that directly affect the optimization time (2) (see Appendix E.5, E.6) we perform a comparison of the intermediate round communication times similar to the measurements in Appendix C.4. In Appendix E.7 we ablate the optimization problem (2) itself, while in Appendix E.8 we address the relaxation of the surrogate loss function. Table 20 summarizes representative operating points for the two primary ISP hyperparameters (p, Δ), complementing the Pareto-front analysis in the main text. The complete sweeps and corresponding client-count dynamics are reported in Appendices E.1 and E.2.

Table 20: Representative operating points from the (p, Δ) ablations on CIFAR-10 with UNIFORM sampling.

Method	**Comm.** (↓)	**Test Loss** (↓)	**Acc.** (↑)	**Overhead** (↓)
UNIFORM	17,767	0.461	0.842	$\times$ 1.00
ISP ($p = 4, \Delta = 20$)	14,343	0.464	0.836	$\times$ 1.43
ISP ($p = 2, \Delta = 20$)	15,900	0.467	**0.845**	$\times$ 1.252
ISP ($p = 1, \Delta = 20$)	16,873	0.495	0.834	$\times$ 1.099
ISP ($p = 4, \Delta = 100$)	12,652	0.472	0.831	$\times$ 1.089
ISP ($p = 4, \Delta = 250$)	**10,952**	0.473	0.827	$\times$ 1.038

E.1 Expansion Ratio

The expansion ratio p is one of the main hyperparameters of ISP. As introduced in Research Question 1, it controls the size of the auxiliary pool used for the intermediate client-count search:

$$b_{\tau+1/2} = \min\{M, \lceil (1+p)m_\tau \rceil\}.$$

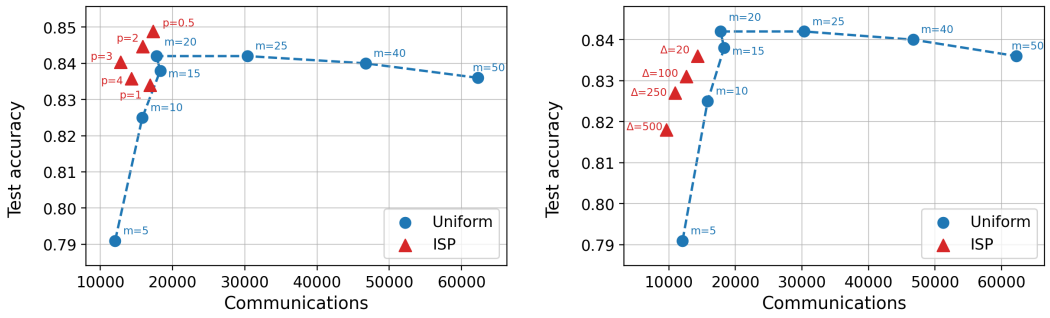
Thus, p has a direct operational interpretation: larger values allocate a wider auxiliary pool for estimating the descent condition in (2), while smaller values reduce intermediate participation and make the procedure more conservative in terms of communication overhead. This parameter therefore controls the fidelity–cost trade-off of the intermediate-pool surrogate.

We ablate p under the same CIFAR-10/RESNET-18 setup with UNIFORM client sampling used in Section 5.2. All other ISP hyperparameters are kept fixed, including $\Delta = 20$, $N = 10$, momentum $\beta = 0.5$, and the same stopping criterion based on validation performance. We consider $p \in \{4, 3, 2, 1, 0.5\}$, which ranges from a broad auxiliary search regime to a substantially more communication-restricted one. The results are reported in Table 21, and the corresponding client-count dynamics are shown in Figure 10(a).

Table 21: Expansion ratio p ablation on CIFAR-10 with UNIFORM client sampling.

Method	**Communications**	**Test Loss**	**Accuracy**	**Overhead**
UNIFORM	17,766	**0.4605**	0.8416	$\times$1.000
ISP $p = 4$	14,342	0.4635	0.8358	$\times$1.432
ISP $p = 3$	**12,842**	0.4613	0.8403	$\times$1.311
ISP $p = 2$	15,900	0.4666	0.8446	$\times$1.252
ISP $p = 1$	16,873	0.4947	0.8340	$\times$1.099
ISP $p = 0.5$	17,319	0.4673	**0.8488**	$\times$1.059

Table 21 reports the complete expansion-ratio sweep, while Figure 9 (a) places the resulting ISP configurations against the communication–accuracy profile of fixed-budget UNIFORM sampling. The results show that ISP is not overly sensitive to the exact choice of p. Across all considered



(a) Expansion ratio p **(b) Adaptation window Δ**

Figure 9: Communication–accuracy trade-offs for the primary ISP communication controls on CIFAR-10 with UNIFORM sampling: (a) auxiliary expansion ratio p and (b) intermediate communication window Δ. ISP points are labeled by the corresponding hyperparameter values, while fixed-budget points are labeled by the client count m. Dashed lines connect the discrete fixed-budget sweep as a visual guide.

values, the method preserves downstream performance at a level comparable to the fixed-budget UNIFORM baseline, while most configurations reduce communication. The broad-search setting $p = 4$, used as the default in the main CIFAR-10 experiments, already reduces communication by approximately 19% while maintaining similar test loss and accuracy. The setting $p = 3$ achieves the best communication efficiency in this sweep and preserves accuracy almost identical to the baseline, showing that this reduction in the intermediate pool preserves endpoint performance in the reported run. At the same time, smaller values such as $p = 2$ and $p = 0.5$ can slightly improve accuracy, so the observed accuracy does not decline monotonically with auxiliary-pool size.

These results support the intended role of p as a practical knob rather than a fragile tuning parameter. Larger p values make the intermediate search more exploratory by exposing ISP to a wider range of candidate client subsets, while smaller p values reduce intermediate overhead and can be preferable when client availability, inference cost, or communication constraints are strict. All configurations use the same objective (2). The sweep supports robustness to the expansion ratio in this setting, while leaving the accuracy of the response estimates unmeasured.

Figure 10(a) further illustrates that different expansion ratios induce different adaptive client-count trajectories. Together with Table 21, these trajectories show how the expansion ratio changes participation while retaining comparable endpoint performance in this sweep.

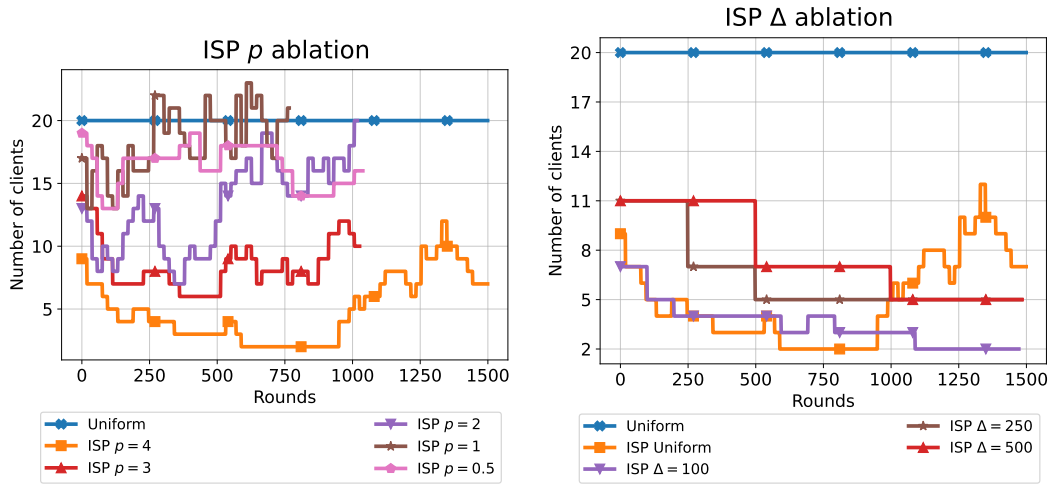
E.2 ADAPTATION WINDOW

The window Δ specifies the number of training rounds between adaptations in Algorithm 2. Increasing it reduces the frequency of auxiliary communication while reusing each selected count longer. We vary Δ from the default 20 to 100, 250, and 500 under UNIFORM sampling in the setting of Appendix E.1. Table 22 reports the results; Figure 9(b) compares them with fixed-budget sampling.

Table 22: Adaptation-window Δ ablation.

Method	Communications	Test Loss	Accuracy
UNIFORM	17,767	**0.461**	**0.842**
ISP-UNIFORM $\Delta = 20$	14,343	0.464	0.836
ISP-UNIFORM $\Delta = 100$	12,652	0.472	0.831
ISP-UNIFORM $\Delta = 250$	10,952	0.473	0.827
ISP-UNIFORM $\Delta = 500$	**9,602**	0.480	0.818

Increasing the window from 20 to 100 reduces client-update exchanges by about 12%, with accuracy changing from 0.836 to 0.831. At $\Delta = 250$, exchanges are about 38% below the fixed-count



(a) Expansion ratio p **(b) Adaptation window Δ**

Figure 10: Client-count dynamics for the primary ISP communication controls on CIFAR-10 with UNIFORM sampling: (a) auxiliary expansion ratio p and (b) adaptation window Δ. Intermediate $\tau + 1/2$ communications are omitted.

baseline, and accuracy is 0.827 versus 0.842 for that baseline. Extending the window to 500 reduces exchanges further but lowers accuracy to 0.818. The trajectories in Figure 10(b) show the longer intervals over which each count is retained. These observations expose the communication–accuracy trade-off from delayed adaptation; they do not measure the temporal response drift appearing in the theoretical bounds. Using the mean relative timing in Table 6, the amortized adaptation overhead at $\Delta = 250$ is below 1.5%.

E.3 MOMENTUM β

The hyperparameter β addresses the greedy nature of problem (2) by adjusting the contribution of the resulting $m_{\tau+1}$ compared to the previous history. We measure the impact of β by disabling it ($\beta = 1$) in the same setup setting (see Section 5.2 and Table 1) with the UNIFORM client strategy. The results are summarized jointly in Table 23 and Figure 11, which cover the momentum and EMA ablations from Appendices E.3 and E.4..

We see more stochastic client-count dynamics in Figure 11, which are reflected in a partial loss of communication efficiency without a corresponding improvement in downstream performance (Table 23). Figure 12(a) places both momentum configurations against the fixed-budget communication–accuracy profile. The two operating points remain close, indicating that momentum primarily regularizes the client-count trajectory rather than changing the attainable communication–quality regime. Thus, the β-momentum in the updated $m_{\tau+1}$ provides greater stability to the greedy optimization in (2).

E.4 EXPONENTIAL MOVING AVERAGE

We employ Exponential Moving Average (EMA) with window size $w_{ema} = 5$ to smooth global loss histories. Figure 6 shows noticeable noise in the test loss due to significant statistical heterogeneity in local clients data. This can potentially compromise loss-based optimization (2). This study ablates EMA's impact on the ISP procedure. Table 23 and Figure 11 summarize the complete history-usage ablation across momentum and EMA configurations.

Without EMA smoothing, we observe more unstable client selection patterns (see Figure 11). This stochasticity increases communication costs beyond the UNIFORM baseline despite comparable convergence. To measure the impact of the greedy optimization solving 2, we conduct an additional experiment that disables both EMA smoothing and participant momentum ($\beta = 1$), eliminating historical context. Figure 12(b) shows that EMA produces a more communication-efficient operating point while preserving downstream accuracy. Removing both EMA and count momentum leads to a

substantially less favorable regime, supporting their complementary role in stabilizing noisy round-level decisions. The obtained results continue the intuition about the importance of the previous context: we see a more unstable dynamic of client selection, which further leads to communication overhead.

Table 23: History-based stabilization ablation on CIFAR-10 with UNIFORM sampling.

Method	**Communications** (↓)	**Test Loss** (↓)	**Accuracy** (↑)
UNIFORM	17,767	0.461	0.842
ISP-UNIFORM (EMA, $\beta = 0.5$)	14,343	0.464	0.836
ISP-UNIFORM (EMA, $\beta = 1$)	15,786	0.462	0.834
ISP-UNIFORM (no EMA, $\beta = 0.5$)	24,445	0.466	0.839
ISP-UNIFORM (no EMA, $\beta = 1$)	26,905	0.469	0.822

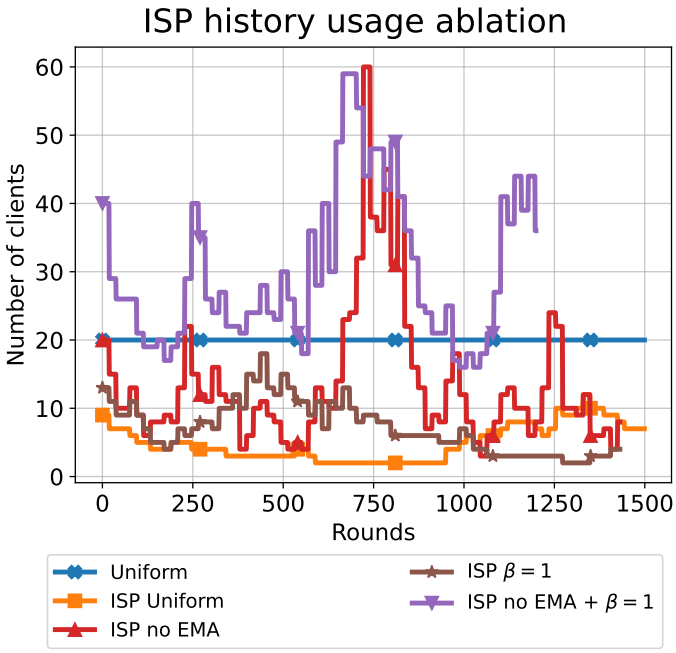
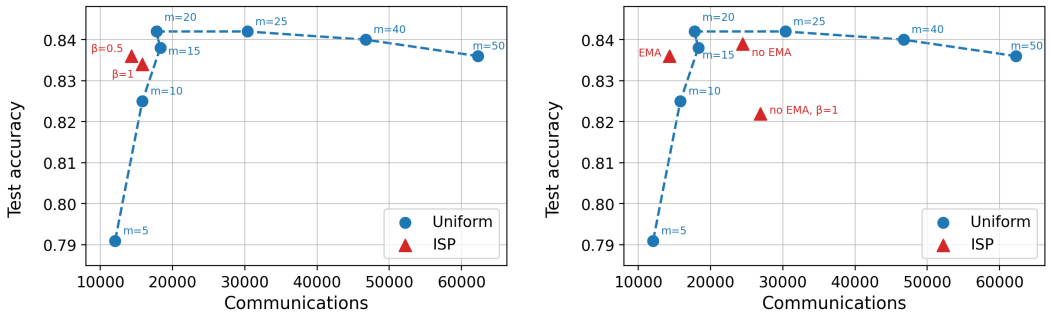


Figure 11: Client count dynamics for the ISP technique with various history usage technique with UNIFORM sampling strategy. Results are shown up to the best round of each experiment run by mean clients validation, omitting $\tau + 1/2$ ISP communications.



(a) Momentum β **(b) EMA smoothing**

Figure 12: Communication–accuracy trade-offs for the history-based stabilization components of ISP on CIFAR-10 with UNIFORM sampling: (a) momentum applied to the selected client count and (b) EMA smoothing of objective estimates. Fixed-budget points are labeled by the client count m, and dashed lines connect the discrete sweep as a visual guide.

E.5 COVERAGE DEPTH

Let us move on to the ablation of the coverage depth N, which is defined in Research Question 3. As we noted in the discussion of Table 2, the baseline choice of $N = 10$ may have been excess, since even for the UNIFORM client sampling strategy, we observed a reduction in variance across 5 experimental runs. This choice could have a negative impact on the final computational performance (see Table 6), since the coverage depth directly determines the number of client inferences and the complexity of the intermediate communication. The purpose of this ablation is to determine the necessary but sufficient value N of the coverage depth.

Table 24: Coverage-depth N ablation. Mean relative time compares an intermediate ISP round with a standard training round.

Method	**Communications (↓)**	**Loss (↓)**	**Accuracy (↑)**	**Rel. Time (↓)**
UNIFORM	17,767	0.461	**0.842**	–
ISP-UNIFORM $N = 10$	14,343	0.464	0.836	3.933
ISP-UNIFORM $N = 7$	14,365	**0.453**	**0.842**	2.753
ISP-UNIFORM $N = 5$	14,290	0.462	0.839	1.966
ISP-UNIFORM $N = 3$	**14,102**	0.476	0.838	1.180
ISP-UNIFORM $N = 1$	15,102	0.872	0.696	0.393

Table 24 reports the downstream results together with the relative intermediate-round time, while Figure 13(a) shows the corresponding client-count trajectories. The table confirms the assumption of redundancy of the choice of $N = 10$ even in the case of a uniform sampling strategy. We see that a gradual reduction of the coverage depth to $N = 3$ does not significantly affect the change in the results. At $N = 1$, a single favorable sampled subset can make the estimated loss difference negative at a small count even when its full-pool expected response is positive. The first-negative search then accepts this count, although the next training subset is drawn afresh. Averaging more subsets reduces this source of selection error. Example F.5 isolates the mechanism analytically; it explains why fixed-count unbiasedness alone does not guarantee progress after the count has been selected.

The final column of Table 24 reports the mean relative time as N decreases. From Table 24, we have determined an appropriate sufficient coverage depth $N = 3$ that optimizes the number of computations of the solution (2), making them close to the original round of communication. Figure 14(a) highlights the qualitative difference between an insufficient estimator depth and the remaining configurations. A single sampled subset does not reliably represent the candidate-count response, whereas larger values of N occupy a compact high-accuracy region with similar commu-

Table 25: Candidate-count resolution w ablation. Mean relative time compares an intermediate ISP round with a standard training round.

Method	**Comm.**	**Loss**	**Acc.**	**Rel. Time** (↓)
UNIFORM	17,767	0.461	0.842	–
ISP-UNIFORM $w = 1$	14,343	0.464	0.836	3.933
ISP-UNIFORM $w = 2$	15,771	0.467	0.842	1.814
ISP-UNIFORM $w = 3$	14,515	0.483	0.837	1.206
ISP-UNIFORM $w = 5$	19,739	0.467	0.852	1.068
ISP-UNIFORM $w = 7$	21,939	0.464	0.856	1.011
ISP-UNIFORM $w = 10$	29,984	0.458	0.860	1.137

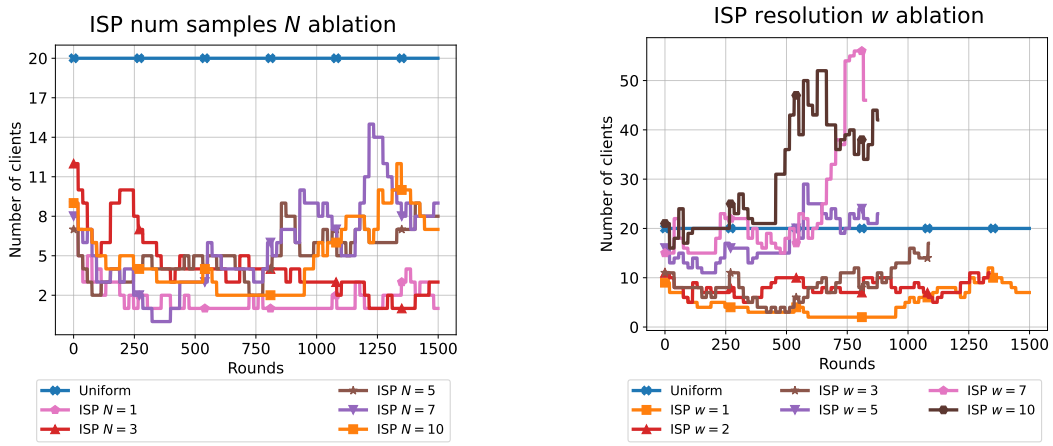
nication requirements. Less frequent adaptation reduces the amortized computation overhead. The accuracy cost of increasing Δ is evaluated separately in Appendix E.2.

E.6 ENUMERATING RESOLUTION

Another hyperparameter that directly affects the internal computation of optimization (2) (see Research Question 3) is the resolution w, which specifies the enumeration step. This step generates a window $2w$ within which the ISP technique is not sensitive to the selected $m_{\tau+1}$. However, the results of the time measurements in Table 6 demonstrate a direct relationship between the value of $m_{\tau+1}$ and the time of the intermediate communication. Increasing the resolution window clearly allows us to address this challenge.

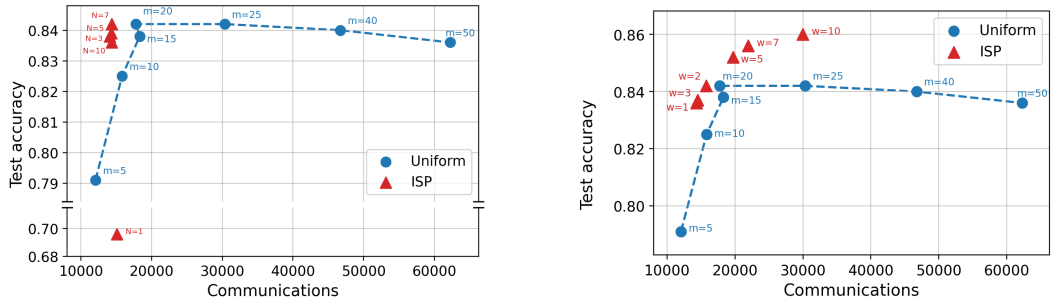
The results are reported in Table 25, while Figure 13(b) illustrates the corresponding client-count trajectories. We see that for $w > 3$ the communication costs increase significantly. This is also reflected in the client count dynamics: the average value for $m_{\tau+1}$ increases for $w > 3$. We address this to a narrow communication-quality zone from $m = 5$ to $m = 10$ (see the discussion in Appendix D.7). When we choose a resolution of $w = 5$ or more, the ISP technique becomes insensitive to it, which leads to redundant communications at the start, which then generate even more redundant ones due to the greedy structure of the optimized functional (2). Thus, the resolution w depends on the statistical indicators of the adaptive intelligence dynamics of the participants. However, for $w < 5$ we do not observe significant deviations, while even $w = 2$ greatly increases the efficiency of internal computations during optimization of (2).

The final column of Table 25 shows the computational effect of increasing the resolution. For $w = 3$, the intermediate communication time becomes comparable to the usual one without a significant decrease in performance. At the same time, a subsequent increase in w does not provide such a noticeable speedup in the computations. Figure 14(b) shows that coarsening the candidate grid does not inherently reduce downstream quality. Across the considered range, larger resolutions retain competitive communication–accuracy operating points, confirming that fine enumeration of every client count is unnecessary in this setting. This is due to the increase in the selected $m_{\tau+1}$ (see Figure 13(b)), which was discussed in the above paragraph.



(a) Coverage depth N **(b) Candidate-count resolution w**

Figure 13: Client-count dynamics for the response-estimation controls of ISP on CIFAR-10 with UNIFORM sampling: (a) Monte Carlo coverage depth N and (b) candidate-count resolution w. Intermediate $\tau + 1/2$ communications are omitted.



(a) Coverage depth N **(b) Resolution w**

Figure 14: Communication–accuracy trade-offs for the response-estimation controls of ISP on CIFAR-10 with UNIFORM sampling: (a) Monte Carlo coverage depth N and (b) candidate-count resolution w. The broken vertical axis in panel (a) exposes the degradation at insufficient estimation depth. Fixed-budget points are labeled by m, and dashed lines connect the discrete sweep as a visual guide.

E.7 ISP RELATIVE IMPROVEMENT

We revisit the client-count optimization problem in (2) from a different communication–quality perspective. The default full-pool formulation (2) is communication-oriented: it seeks the smallest client count with a negative expected loss response. Algorithm 2 approximates this criterion using a restricted pool and finite samples. This greedy feasibility rule is effective for minimizing communication, but it does not explicitly reward larger expected improvements once the descent condition is satisfied. We therefore consider an alternative formulation, ISP Relative Improvement (ISP-RI), which selects the client count that maximizes the expected loss decrease per communication unit.

Recall that $\delta f_\tau(m, C) = f(\boldsymbol{x}^\tau - \gamma g^\tau(C)) - f(\boldsymbol{x}^\tau)$, so a model improvement corresponds to a negative value of δf_τ. Thus, the full-pool relative-improvement objective can be written as

$$m_{\tau+1}^{\mathrm{RI}} = \arg\max_{m \in [M]} \frac{-\Phi_\tau([M], m)}{m^\alpha}, \qquad \Phi_\tau([M], m) = \mathbb{E}_{C \sim \mathcal{S}_\tau([M],m)} \delta f_\tau(m, C), \tag{9}$$

where $\alpha > 0$ controls the trade-off between expected improvement and communication cost. In the experiment below, we set $\alpha = 1$, which corresponds to maximizing the expected loss decrease per participating client.

As in the default ISP formulation, directly evaluating (9) over the full client pool is impractical. Therefore, ISP-RI uses the same intermediate-pool approximation introduced in RQ 1. Given the auxiliary pool $I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})$, we solve

$$m_{\tau+1}^{\mathrm{RI}} = \arg \max_{m \in [b_{\tau+1/2}]} \frac{-\widehat{\Phi}_\tau(I_{\tau+1/2}, m)}{m^\alpha}, \tag{10}$$

where $\widehat{\Phi}_\tau(I_{\tau+1/2}, m)$ is estimated using the same Monte Carlo procedure as in Algorithm 2. The main difference from the default ISP rule is that ISP-RI cannot stop after the first feasible descent point: it must evaluate all candidate counts in $[b_{\tau+1/2}]$ in order to compare their relative improvements. Consequently, ISP-RI is more quality-oriented but also more computationally demanding. When the surrogate loss $\hat{f}$ is unavailable, this complete enumeration can increase the client-side inference burden. Reducing this overhead can rely on the same practical controls as the default method, including the communication window Δ (Appendix E.2), coverage depth N (Appendix E.5), and enumeration resolution w (Appendix E.6).

Table 26: Comparison of ISP-RI with UNIFORM client sampling.

Method	Communications	Test Loss	Accuracy
UNIFORM	17,767	0.461	0.842
ADAFL (ECONOM)	16,191	0.508	0.827
ADAFL (FULL)	63,876	0.421	0.855
ISP UNIFORM	**14,343**	0.464	0.836
ISP-RI UNIFORM	49,594	**0.397**	**0.873**

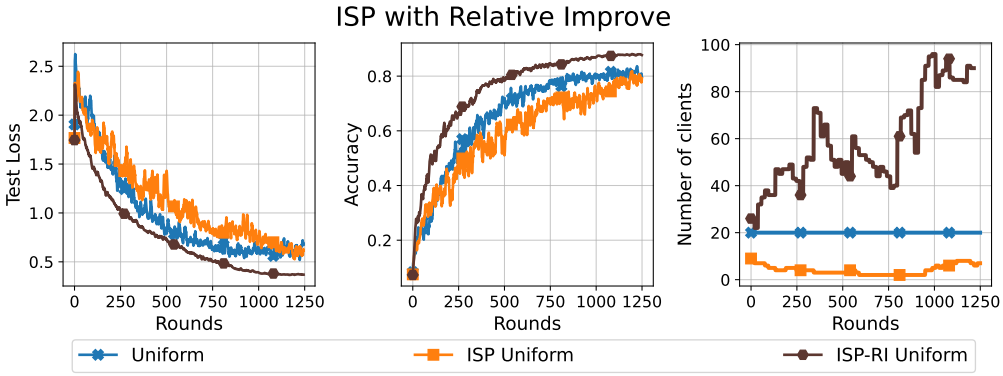


Figure 15: Test loss, accuracy, and client-count dynamics for ISP and ISP-RI with UNIFORM client selection. For clarity, results are shown through round 1250, omitting $\tau + 1/2$ ISP communications.

We compare ISP-RI with fixed-budget UNIFORM sampling, the default ISP objective (2), and the linearly increasing schedules of ADAFL. The full ADAFL schedule is particularly relevant here because it also favors downstream quality at the cost of additional communication. Table 26 summarizes the comparison, while Figure 15 shows the corresponding convergence and client-count dynamics.

The results highlight the expected trade-off. The default ISP objective is the most communication-efficient, reducing client exchanges while preserving the baseline quality. In contrast, ISP-RI shifts the operating point toward stronger optimization progress: it uses more communication than the default ISP, but achieves the best test loss and accuracy among the compared methods. Compared with the full ADAFL schedule, ISP-RI reaches higher downstream quality with substantially fewer communications. Moreover, its client-count trajectory is nonlinear, unlike the fixed linear schedule of ADAFL, which shows that the relative-improvement objective adapts the participation budget to the current optimization state rather than following a predefined schedule. Figure 17(a) illustrates that modifying the response criterion shifts ISP toward a different communication–quality regime.

The relative-improvement variant prioritizes downstream performance more strongly, whereas the default formulation retains a substantially more communication-efficient operating point.

E.8 RELAXING TRUST ASSUMPTION

Building upon our client computation reduction strategies mentioned in Section 5.2 (specifically partial participation where $m_{\tau+1/2} < M$, analyzed in Appendix E.1), we now examine an orthogonal approach: relaxing requirements for the server-side surrogate loss $\hat{f}$ introduced in Research Question 3 and Appendix C.4. While $\hat{f}$ effectively transfers computational burden from clients to the server and accelerates optimization, its assumed availability represents a strong practical constraint. This ablation study quantifies our method's sensitivity to $\hat{f}$ quality, specifically investigating whether performance degradation occurs when using non-ideal surrogate loss functions.

Using the UNIFORM strategy setup from Table 2, we implement two configurations: NORMAL $\hat{f}$: 100-example training subset and CORRUPTED $\hat{f}$: Same subset with aggressive augmentations (for a detailed description of augmentations, see Table 27).

Results in Table 28 and Figure 16 demonstrate similar downstream performance and client-count dynamics across client-side, standard-surrogate, and corrupted-surrogate evaluation. Figure 17(b) shows that the default client-side evaluation and the available surrogate $\hat{f}$ lead to similar operating points. A better-aligned surrogate can further improve downstream performance without materially increasing communication, whereas corruption of the objective signal changes the selected regime and confirms that ISP depends on the informativeness of the response estimate. This minimal sensitivity to $\hat{f}$ quality indicates our method maintains operational integrity even with substantially degraded surrogate $\hat{f}$.

Table 27: Augmentation parameters for corrupted $\hat{f}$

Augmentation Technique	Parameters	Description
RANDOMGRAYSCALE	p=0.8	Randomly converts images to grayscale with high probability. Significantly reduces color information.
COLORJITTER	brightness=0.5 contrast=0.5 saturation=0.5 hue=0.3	Applies strong random color distortions. Severely alters color distribution and visual appearance.
RANDOMRESIZEDCROP	size=32 scale=(0.5, 1.0)	Crops random portions of image and resizes. May remove critical features and distort object shapes.
RANDOMHORIZONTALFLIP	p=0.5	Flips image horizontally with 50% probability. Preserves content but alters spatial orientation.
RANDOMVERTICALFLIP	p=0.3	Flips image vertically with 30% probability. Dramatically changes scene orientation (uncommon in natural images).
GAUSSIANBLUR	kernel size=3 σ=(0.1, 2.0)	Adds variable blurring effect. Reduces high-frequency details and sharpness.
NORMALIZATION	μ=(0.485, 0.456, 0.406) σ=(0.229, 0.224, 0.225)	Standard channel-wise normalization. Transforms pixel distributions but preserves visual content.

Table 28: Surrogate loss $\hat{f}$ ablation.

Method	Communications	Test Loss	Accuracy
UNIFORM	17,767	0.461	0.842
ISP NO $\hat{f}$	14,343	0.464	0.836
ISP NORMAL $\hat{f}$	14,525	0.462	0.837
ISP CORRUPTED $\hat{f}$	14,740	0.466	0.849

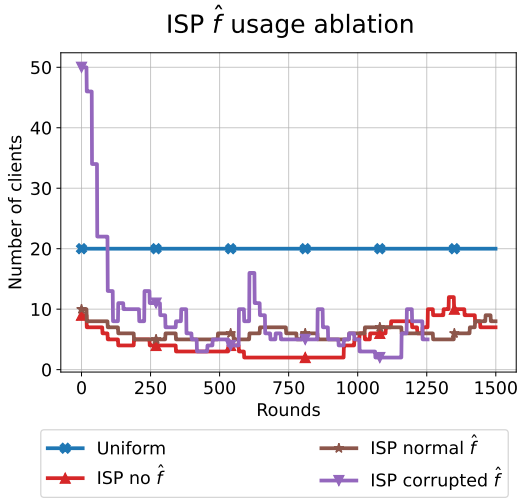
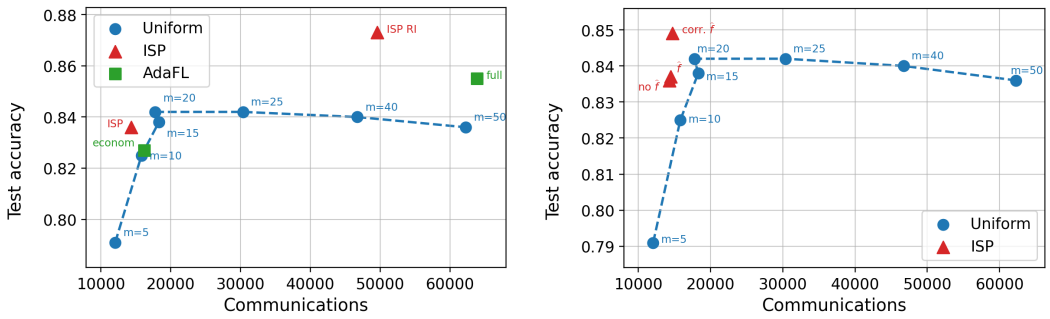


Figure 16: Client count dynamics for the ISP technique with various $\hat{f}$ surrogate loss usage with UNIFORM sampling strategy. For clarity, results are shown through round 1500, omitting $\tau + 1/2$ ISP communications.



(a) Relative-improvement objective **(b) Surrogate objective $\hat{f}$**

Figure 17: Communication–accuracy trade-offs for alternative response formulations and objective-evaluation signals on CIFAR-10: (a) the relative-improvement variant ISP-RI and (b) client-side and surrogate objective evaluation. Fixed-budget points are labeled by the client count m, and dashed lines connect the discrete sweep as a visual guide.

F THEORY OF ADAPTIVE PARTICIPANT-COUNT SELECTION

The analysis follows the controller's decisions: evaluate candidate counts, select the first negative estimate, and reuse the returned count with fresh training subsets. The first three subsections establish the corresponding progress, selection, and reuse bounds. The remaining three give concrete regimes, extend the sampling analysis, and connect the reference model to practical updates. The stationarity result uses additional response conditions; it is not obtained by simply iterating the finite-block bound.

1. Appendix F.1: how participation controls the variance and progress of a fresh update.
2. Appendix F.2: how pool restriction and finite depth affect the count actually selected.
3. Appendix F.3: what can be guaranteed while that count is reused, and which additional conditions give stationarity.
4. Appendix F.4: exact examples and constructions that make these conditions concrete.
5. Appendix F.5: what changes for other selectors and what the inclusion diagnostics measure.
6. Appendix F.6: how local training, stochastic updates, and evaluation errors enter the bounds.

We retain $g^\tau(C)$, $\Phi_\tau(A,m)$, and $\widehat{\Phi}_\tau(A,m)$ from the method. When a result concerns an arbitrary model x, a subscript x indicates evaluation at that model rather than at a numbered round. Except in Appendix F.5, sampling is uniform. All search statements condition on the history before adaptation, so the auxiliary budget may depend on previous selected counts. We fix the initial model x^0; rates also hold conditional on a random initialization with finite initial loss. Table 29 collects the shared notation, including the diagnostics in Appendix G.

Table 29: Key notation for Appendices F and G. Proof-specific quantities are defined where used.

Symbol	Description
M	Total number of clients.
$f_i(x),\ f(x)$	Local objective and global objective $f = M^{-1}\sum_i f_i$.
$\gamma,\ L$	Stepsize and smoothness constant.
$m_{\tau+1}$	Client count returned at round τ.
$b_{\tau+1/2},\ I_{\tau+1/2}$	Auxiliary-pool size and the pool itself.
p	Auxiliary-pool expansion ratio.
N	Monte Carlo samples per candidate count.
Δ	Number of rounds between count adaptations.
$\sigma_c^2,\ \rho$	Additive and relative average gradient-heterogeneity parameters.
$\Phi_x(A,m)$	Expected loss change at model x using m clients from pool A.
$\widehat{\Phi}_x(A,m)$	Monte Carlo estimate of $\Phi_x(A,m)$.
$\epsilon_m,\ \epsilon_\tau$	Candidate-specific and common response-error bounds.
δ	Failure probability for a simultaneous accuracy statement.
$\Gamma(m,b)$	Expected absolute response gap between auxiliary and full pools.
$\pi_i,\ \psi_i(I)$	Inclusion probabilities under full-pool and restricted-pool selection.
$p_i,\ q_i(I)$	Normalized inclusion weights π_i/m and $\psi_i(I)/m$.
ESS, D, Mass	Effective support size, squared inclusion distortion, and captured mass.

F.1 REFERENCE PROGRESS AND ADDITIVE HETEROGENEITY

The finite-population factor $q_m = (M-m)/(m(M-1))$ measures the variance reduction from sampling m clients without replacement. We take $M \geq 2$; with a single client there is no count-selection problem. Assume f is L-smooth. For stationarity bounds, also assume $f \geq f_{\inf} > -\infty$. The variance identity below needs no heterogeneity assumption. The subsequent progress bound allows an additive component of client disagreement. It allows nonzero client gradients at a stationary point of the global loss.

Lemma F.1 (Exact finite-population variance)**.** *For a uniform subset $C \subseteq [M]$ of size m, let $g_x(C) = m^{-1} \sum_{i \in C} \nabla f_i(x)$. Then*

$$\mathbb{E} g_x(C) = \nabla f(x),$$
$$\mathbb{E}\|g_x(C) - \nabla f(x)\|^2 = \frac{q_m}{M} \sum_{i=1}^{M} \|\nabla f_i(x) - \nabla f(x)\|^2.$$

Proof. Put $d_i = \nabla f_i(x) - \nabla f(x)$, so that $\sum_i d_i = 0$. The inclusion probabilities of one and two distinct clients are m/M and $m(m-1)/(M(M-1))$. Expanding the squared norm and using $\sum_{i \neq j} \langle d_i, d_j \rangle = -\sum_i \|d_i\|^2$ gives

$$\mathbb{E}\left\| \frac{1}{m} \sum_{i \in C} d_i \right\|^2 = \frac{1}{m^2} \left(\frac{m}{M} - \frac{m(m-1)}{M(M-1)} \right) \sum_i \|d_i\|^2$$
$$= \frac{q_m}{M} \sum_{i=1}^{M} \|\nabla f_i(x) - \nabla f(x)\|^2.$$

□

Proof of Theorem 4.3. For an arbitrary model x and count m, put $a_m = 1 - (L\gamma/2)(1 + \rho q_m)$. Under Assumptions 4.1 and 4.2, we prove

$$\Phi_x([M], m) \leq -\gamma a_m \|\nabla f(x)\|^2 + \frac{L\gamma^2}{2} \sigma_c^2 q_m. \tag{11}$$

Proof. Smoothness and the lemma yield the exact-variance upper bound

$$\Phi_x([M], m) \leq -\gamma(1 - L\gamma/2)\|\nabla f(x)\|^2 + \frac{L\gamma^2 q_m}{2M} \sum_{i=1}^{M} \|\nabla f_i(x) - \nabla f(x)\|^2. \tag{12}$$

Substitute Assumption 4.2 to obtain the stated bound. These statements also hold conditionally for a random count, provided that the production subset is sampled freshly after that count has been determined. They do not require independence of the count from the earlier search. □

For $\sigma_c = 0$, $0 < \gamma = \alpha/L < 2/L$, and any $0 < c < 1 - \alpha/2$, a sufficient condition for $a_m \geq c$ is

$$m \geq \max\left\{1, \left\lceil \frac{\alpha\rho M}{[2(1-c) - \alpha](M-1) + \alpha\rho} \right\rceil \right\}. \tag{13}$$

When $a_m = 0$, the conclusion is nonincrease, not strict descent. For $a_m > 0$, strict descent follows at a nonstationary point in the zero-additive-noise case. The threshold is sufficient, not necessary; Algorithm 2 does not explicitly enforce it.

At a nonstationary point, the additive bound also gives a state-dependent sufficient count by replacing ρ in (13) with $\rho + \sigma_c^2/\|\nabla f(x)\|^2$. This follows by absorbing the additive term into the squared-gradient coefficient in (11). It explains why a participation level sufficient far from stationarity can become inadequate as the optimization signal decreases. The controller need not know these constants; they describe the sufficient progress envelope. At $\nabla f(x) = 0$, the absolute variance bound must be used instead of this ratio.

Corollary F.2 (Progress for a prescribed count sequence)**.** *Under the preceding assumptions, with $f \geq f_{\inf} > -\infty$, let $\mathcal{F}_t$ contain the selected count m_{t+1} and all information before drawing the fresh production subset. Suppose $a_{m_{t+1}} \geq c > 0$ almost surely at every round, and the indicated expectations are finite. Then*

$$\frac{1}{T} \sum_{t<T} \mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{f(x^0) - f_{\inf}}{\gamma c T} + \frac{L\gamma\sigma_c^2}{2cT} \sum_{t<T} \mathbb{E} q_{m_{t+1}}. \tag{14}$$

Proof. Condition on $\mathcal{F}_t$, apply (11), take total expectations, sum over t, and use $\mathbb{E}f(x^T) \geq f_{\inf}$. $\square$

For $\sigma_c = 0$ and a fixed $\gamma = \alpha/L$, this gives an $O(L/T)$ bound with the initial gap and $1/(\alpha c)$ retained. For $\sigma_c > 0$, the residual term remains unless additional conditions make its average vanish. This corollary describes quantitative progress for a given count sequence; the later search results establish when Algorithm 2 itself returns a count with a suitable margin.

F.2 Finite-sample count selection

Probability model. Work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying the initial state and all search and training draws. Let $\mathcal{F}_0$ contain the initial state and any fixed problem data. At an adaptation round τ, define

$$\mathcal{H}_\tau = \sigma\big(\mathcal{F}_0, \{I_{s+1/2}, (C^n_{s,m})_{m,n} : s < \tau,\ s \text{ is an adaptation round}\}, \{C^{\mathrm{train}}_s : s < \tau\}\big).$$

Here $C^n_{s,m}$ is the nth search subset for candidate count m, and C^{train}_s is the method's C_{s+1}, used to update x^s to x^{s+1}. The model $\boldsymbol{x}^\tau$ and auxiliary budget $b_{\tau+1/2}$ are functions of this history. Conditional on $\mathcal{H}_\tau$, the auxiliary pool is a uniform subset of $[M]$ of size $b_{\tau+1/2}$. Conditional on this pool and history, for each candidate count the N search subsets are independent uniform subsets of that size. No client is repeated within a subset; independent draws can produce the same subset.

To state simultaneous bounds, realize samples for every candidate in advance and reveal them only when queried. Unqueried arrays do not affect the algorithm. Dependence between arrays for different counts is allowed. Each production subset is uniform of the selected size, conditional on all preceding search and training draws. All probabilities in the selection analysis are conditional on $\mathcal{H}_\tau$ unless another conditioning is indicated. For $b_{\tau+1/2} = 1$, Algorithm 2 returns the fallback immediately, so there is no selection error to analyze. Search confidence concerns the pool and probe draws at a fixed preceding history. The reuse bound then fixes the completed search and averages over fresh training subsets. The periodic rate averages over both kinds of randomness, including failed searches.

At an adaptation round τ, condition on the preceding history. Write $x = x^\tau$, $b = b_{\tau+1/2}$, and $I = I_{\tau+1/2}$ locally. We use the existing responses $\Phi_\tau([M], m)$, $\Phi_\tau(I, m)$, and $\widehat{\Phi}_\tau(I, m)$ throughout the search analysis. The constants

$$D = \max_{i,j} \|\nabla f_i(x) - \nabla f_j(x)\|,$$
$$G \geq \max_i \|\nabla f_i(x)\|, \qquad K_x = \|\nabla f(x)\| + L\gamma G.$$

give the gradient diameter, a finite gradient bound, and objective sensitivity at candidate models. One may take $G = \max_i \|\nabla f_i(x)\|$ or use any larger bound chosen before the search. They are fixed after conditioning on the current state; no bounded-gradient assumption over the entire optimization trajectory is needed here. In particular, K_x bounds the objective gradient on the convex hull of the candidate models. This converts differences between sampled updates into differences between their loss responses.

We use the conditional sampling laws of Appendix F.2. In particular, the fresh production response after the search is $\Phi_\tau([M], m_{\tau+1})$.

Lemma F.3 (Uniform finite-population transfer)**.** *For every fixed $m \leq b$, $\mathbb{E}_I \Phi_\tau(I, m) = \Phi_\tau([M], m)$. Let $r = \min\{b, M - b\}$ and $\delta_A \in (0, 1)$. For $b < M$, define*

$$a_b(\delta_A) = \frac{\gamma K_x D}{b} \sqrt{\frac{r}{2} \log \frac{2b}{\delta_A}}. \tag{15}$$

For $b = M$, set $a_b = 0$. Then

$$\mathbb{P}_I\left(\max_{1 \leq m \leq b} |\Phi_\tau(I, m) - \Phi_\tau([M], m)| \leq a_b(\delta_A)\right) \geq 1 - \delta_A. \tag{16}$$

For a deterministic full pool the event has probability one.

Proof. A uniform b-subset followed by a uniform m-subset has a uniform m-subset of $[M]$ as its marginal law. This proves the mean identity.

Let I and I' differ by replacing client i with client j. Couple their uniform m-subsets by replacing i with j whenever i is included. This occurs with probability m/b. On this event the aggregate-gradient difference has norm at most D/m, so the difference between the candidate responses is at most $\gamma K_x D/m$. Therefore, for every $m \le b$,

$$|\Phi_\tau(I,m) - \Phi_\tau(I',m)| \le \frac{\gamma K_x D}{b}. \tag{17}$$

Reveal an ordered sample generating either I or its complement, using whichever has r elements. Consider the Doob martingale of $\Phi_\tau(I,m)$ under these revelations. At each reveal, conditional completions following two possible next clients can be coupled by transposing those two clients. The resulting final pools are identical or differ by one swap. Consequently, the conditional expectations associated with the two possible revelations differ by at most $c_b = \gamma K_x D/b$. Each martingale difference is therefore supported in a conditional interval of length at most c_b. Conditional Hoeffding's lemma and exponential Markov give

$$\mathbb{P}_I\big(|\Phi_\tau(I,m) - \mathbb{E}_I\Phi_\tau(I,m)| \ge t\big) \le 2\exp\left(-\frac{2t^2}{rc_b^2}\right)$$

when $rc_b > 0$. If $c_b = 0$, the response is constant over all pools. A union bound over $m \in [b]$ proves (16). For $b = M$ the pool is deterministic and the error is zero directly. □

The pool error concerns the loss response itself. Its size dependence is $O(\sqrt{\log b/b})$ for $b \le M/2$, and $O(\sqrt{(M-b)\log b}/b)$ near full participation. Uniformity over m is essential when the same pool selects the count.

Lemma F.4 (Conditional Monte Carlo accuracy)**.** *Suppose $b > 1$. For $m < b$, put*

$$\begin{aligned} w_m &= \gamma K_x D \min\{1, (b-m)/m\}, \\ e_m &= w_m\sqrt{\frac{1}{2N}\log\frac{2(b-1)}{\delta_N}}, \qquad \delta_N \in (0,1). \end{aligned} \tag{18}$$

Then, conditional on any auxiliary pool,

$$\mathbb{P}\Big(|\widehat{\Phi}_\tau(I,m) - \Phi_\tau(I,m)| \le e_m \textit{ for all } m < b \,\Big|\, I\Big) \ge 1 - \delta_N. \tag{19}$$

Proof. Two m-subsets of a b-pool differ in at most $\min\{m, b-m\}$ clients. Pair their differing clients. Their aggregate gradients differ by at most $D\min\{m, b-m\}/m$, so their scalar responses differ by at most w_m. Conditional Hoeffding's inequality for the N independent responses gives

$$\mathbb{P}\big(|\widehat{\Phi}_\tau(I,m) - \Phi_\tau(I,m)| > e_m \mid I\big) \le \delta_N/(b-1).$$

A union bound proves the claim. A zero range means that the estimate is exact, and introduces no exceptional case. □

A variance-sensitive alternative is available. For $g_x(I) = b^{-1}\sum_{i\in I}\nabla f_i(x)$ and $S_I^2 = b^{-1}\sum_{i\in I}\|\nabla f_i(x) - g_x(I)\|^2$, the same Lipschitz bound gives

$$\mathrm{Var}(\delta f_\tau(m,C) \mid I) \le \gamma^2 K_x^2 \frac{b-m}{m(b-1)} S_I^2 =: V_m(I). \tag{20}$$

Indeed, the variance is at most the mean squared deviation from the constant $f(x - \gamma g_x(I)) - f(x)$; apply the Lipschitz bound and the exact sample-mean variance identity. Conditional Bernstein and a union bound therefore also establish (19) with

$$e_m = \sqrt{\frac{2V_m(I)}{N}\log\frac{2(b-1)}{\delta_N}} + \frac{2w_m}{3N}\log\frac{2(b-1)}{\delta_N}.$$

Either chosen family of radii gives its stated simultaneous guarantee.

For general failure allocations δ_A, δ_N, the combined Hoeffding radius is, for each $m < b$,

$$\begin{aligned}\epsilon_m = a_b + e_m = \frac{\gamma K_x D}{b}\sqrt{\frac{\min\{b, M-b\}}{2}\log\frac{2b}{\delta_A}} \\ + \gamma K_x D \min\{1, (b-m)/m\}\sqrt{\frac{1}{2N}\log\frac{2(b-1)}{\delta_N}}.\end{aligned} \tag{21}$$

The pool term is zero at $b = M$. Set $\delta_A = \delta_N = \delta/2$ and $G = \max_i \|\nabla f_i(x^\tau)\|$. Replacing $\min\{1, (b-m)/m\}$ by 1 and $\log(4(b-1)/\delta)$ by $\log(4b/\delta)$ gives $\epsilon_m \le \epsilon_\tau$, where ϵ_τ is the common radius in (6). The two lemmas give simultaneous accuracy for all candidate counts with conditional probability at least $1 - \delta$. In particular this holds for every tested count, proving (7).

Proof of Theorem 4.4: selecting the count. The selection argument is deterministic once the event

$$|\widehat{\Phi}_\tau(I, m) - \Phi_\tau([M], m)| \le \epsilon_m \quad \text{for every } m < b \tag{22}$$

holds. It therefore applies to any response target and selector for which this simultaneous event is established. In the uniform setting above take $\epsilon_m = a_b + e_m$; the event has probability at least $1 - \delta_A - \delta_N$ by the two lemmas.

For clarity, define

$$\begin{aligned} m_- &= \min\big(\{m < b : \Phi_\tau([M], m) < \epsilon_m\} \cup \{b\}\big), \\ m_+ &= \min\big(\{m < b : \Phi_\tau([M], m) < -\epsilon_m\} \cup \{b\}\big). \end{aligned} \tag{23}$$

If $\Phi_\tau([M], m) \ge \epsilon_m$, then $\widehat{\Phi}_\tau(I, m) \ge 0$, so this count cannot be accepted. Thus the returned count $m_{\tau+1}$ is at least m_-. If $\Phi_\tau([M], m) < -\epsilon_m$, then $\widehat{\Phi}_\tau(I, m) < 0$, so the algorithm stops no later than m. This proves

$$m_- \le m_{\tau+1} \le m_+. \tag{24}$$

If $m_{\tau+1} < b$, its estimate is negative, and the same event gives

$$\Phi_\tau([M], m_{\tau+1}) \le \widehat{\Phi}_\tau(I, m_{\tau+1}) + \epsilon_{m_{\tau+1}} < \epsilon_{m_{\tau+1}}.$$

For every $j < m_{\tau+1}$, the algorithm queried j and observed a nonnegative estimate; therefore $\Phi_\tau([M], j) \ge -\epsilon_j$. The fallback $m_{\tau+1} = b$ is not tested and has no acceptance-sign conclusion. For $b = 1$, outside the main theorem's stated range, localization is trivial and the returned count is one. These arguments prove the selected-count conclusions without monotonicity of $m \mapsto \Phi_\tau([M], m)$ or independence between candidate sizes. In particular, an empirical negative implies an upper error bound on the true response, rather than a negative true response by itself; Example F.5 gives an exact failure case. □

Corollary F.5 (Recovery of a separated first negative)**.** *On* (22)*, suppose* $m_\star \le b$ *satisfies* $\Phi_\tau([M], m) \ge \kappa > \epsilon_m$ *for every* $m < m_\star$ *and* $\Phi_\tau([M], m_\star) \le -\eta$ *for some* $\kappa, \eta > 0$*. If* $m_\star < b$*, also assume* $\eta > \epsilon_{m_\star}$*. Then* $m_{\tau+1} = m_\star$*.*

Proof. All earlier estimates are nonnegative, so none is accepted. If $m_\star < b$, its estimate is negative and it is accepted. If $m_\star = b$, the search returns the fallback after rejecting every earlier count. □

When $m_\star = b$, its negative response is an assumption about the fallback; Algorithm 2 does not check it. This corollary describes a response regime of the existing stopping rule and adds no acceptance test.

Example F.1 (A search cap below every useful count)**.** For M scalar clients, let $f_i(x) = a_i x^2/2$, where $M-1$ coefficients equal $1/2$ and the remaining coefficient is $(M+1)/2$. Then $f(x) = x^2/2$ and $M^{-1}\sum_i |\nabla f_i(x) - \nabla f(x)|^2 = (M-1)x^2/4$. For $\gamma = 1$, the exact uniform response is

$$\Phi_x([M], m) = \frac{M - 5m}{8m} x^2.$$

If $b < M/5$ and $x \ne 0$, every permissible production count has positive expected response, including the fallback. Taking $N \to \infty$ cannot change this fact. Thus covering a useful count is a separate condition from estimating its response accurately.

Expected response when the selected count can be wrong. Exact count recovery can be stronger than needed for progress. The following result instead weights a possible selection error by its loss consequence. Continue to condition on a fixed pre-adaptation history. Fix a count $k \le b$ with $\Phi_\tau([M],k) \le -\eta$, where $\eta > 0$, and put $B_\phi = \max_{1\le m\le b} \Phi_\tau([M],m)$. For $m < k$ define

$$P_m^- = \exp\left[-\frac{2N(\Phi_\tau([M],m) - a_b)_+^2}{w_m^2}\right]. \tag{25}$$

For $k < b$ define

$$P_k^+ = \exp\left[-\frac{2N(-\Phi_\tau([M],k) - a_b)_+^2}{w_k^2}\right],$$

and set $P_b^+ = 0$ when $k = b$. If a range w_m is zero, use zero for a strictly positive separated numerator and one otherwise; this conservative convention avoids division by zero. For a deterministic full pool, one can take $a_b = 0$ and $\delta_A = 0$.

Proposition F.6 (Expected response of the adaptive search)**.** *Under the conditional sampling assumptions above,*

$$\mathbb{E}\Phi_\tau([M], m_{\tau+1}) \le -\eta + \sum_{m<k} (\Phi_\tau([M],m) + \eta)_+ P_m^- + (B_\phi + \eta)_+ (P_k^+ + \delta_A). \tag{26}$$

The expectation is over the pool and Monte Carlo search. With a fresh production cohort, its left side is the actual expected next-round loss change.

Proof. Let $\mathcal{A}$ be the auxiliary-good event in (16). On $\mathcal{A}$, $\Phi_\tau(I,m) \ge \Phi_\tau([M],m) - a_b$. Selection at any $m < k$ implies $\widehat{\Phi}_\tau(I,m) < 0$, so conditional Hoeffding gives

$$\mathbb{P}(\mathcal{A} \cap \{m_{\tau+1} = m\}) \le P_m^-.$$

This bound does not condition on the search reaching m: accepting m implies the wrong-sign event for its pre-generated estimate. If $k < b$, selection after k implies $\widehat{\Phi}_\tau(I,k) \ge 0$. On $\mathcal{A}$, $\Phi_\tau(I,k) \le \Phi_\tau([M],k) + a_b$, so the other Hoeffding tail gives

$$\mathbb{P}(\mathcal{A} \cap \{m_{\tau+1} > k\}) \le P_k^+.$$

For $k = b$ overshoot is impossible.

Pathwise on $\mathcal{A}$,

$$\Phi_\tau([M], m_{\tau+1}) + \eta \le \sum_{m<k} (\Phi_\tau([M],m) + \eta)_+ \mathbf{1}_{\{m_{\tau+1}=m\}} + (B_\phi + \eta)_+ \mathbf{1}_{\{m_{\tau+1}>k\}}.$$

When $m_{\tau+1} = k$, the left side is nonpositive. On $\mathcal{A}^c$ it is at most $(B_\phi + \eta)_+$. Take expectations and use $\mathbb{P}(\mathcal{A}^c) \le \delta_A$ to obtain the result. Finally, conditional independence of the production cohort gives its expected loss change as $\Phi_\tau([M], m_{\tau+1})$. □

The bound certifies expected next-round descent when its two nonnegative error terms sum to less than η. Checking this requires the true response margins or valid bounds on them. The earlier-count term includes both harmful positive responses and weakly negative responses. A positive response separated from the auxiliary uncertainty has an exponentially small acceptance bound. A shallow negative response need not become unlikely as N grows. The last term accounts for missing the negative reference count and for auxiliary failure. Thus increasing N controls sampling error without imposing a progress margin on the first negative itself.

If Monte Carlo arrays at different sizes are also conditionally independent, the selection law can be written exactly. With $p_m(I) = \mathbb{P}(\widehat{\Phi}_\tau(I,m) < 0 \mid I)$,

$$\begin{aligned} \mathbb{P}(m_{\tau+1} = m \mid I) &= p_m(I) \prod_{j<m} [1 - p_j(I)], \qquad m < b, \\ \mathbb{P}(m_{\tau+1} = b \mid I) &= \prod_{j<b} [1 - p_j(I)]. \end{aligned} \tag{27}$$

Indeed, selection at $m < b$ requires all earlier estimates to be nonnegative and the mth estimate to be negative; the fallback requires all estimates to be nonnegative. Conditional independence factors these probabilities. Multiplying by $\Phi_\tau([M],m)$, summing over counts, and averaging over I gives the exact expected production response. Fixed-count unbiasedness does not remove these selection weights.

F.3 Reusing a count between adaptations

Selecting a count fixes its response only at the current model. We first bound the change in that response over a finite reuse window, then identify the extra conditions that allow progress to accumulate into a stationarity guarantee. The next theorem retains a sharper drift sum; bounding this sum gives the compact main-text Theorem 4.5.

Theorem F.7 (Expected progress over a reuse block)**.** *Suppose Assumptions 4.1 and 4.2 hold, with $M \geq 2$ and $L > 0$. In addition, suppose every f_i has an L-Lipschitz gradient. Fix an adaptation round τ and an integer $\Delta \geq 1$. Let $\mathcal{F}_\tau^+ = \sigma(\mathcal{H}_\tau, I_{\tau+1/2}, (C_{\tau,m}^n)_{m,n})$ contain exactly the history and current search draws defined in Appendix F.2. Write $\mathbb{E}_\tau[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_\tau^+]$ and $m = m_{\tau+1}$. Hold this count fixed for the next Δ updates. At each update, conditional on $\mathcal{F}_\tau^+$ and the preceding production draws, sample a fresh uniform m-subset of $[M]$ and use its exact average gradient with the fixed stepsize $\gamma > 0$. Define*

$$h = L\gamma, \qquad q_m = \frac{M-m}{m(M-1)},$$
$$V_{\tau,m} = (1+\rho q_m)\|\nabla f(\boldsymbol{x}^\tau)\|^2 + \sigma_c^2 q_m,$$
$$\mathcal{D}_\Delta(h) = \sum_{j=0}^{\Delta-1} \left[(2+h)((1+h)^j - 1) + (1+h)((1+h)^j - 1)^2\right].$$

Then

$$\mathbb{E}_\tau[f(\boldsymbol{x}^{\tau+\Delta}) - f(\boldsymbol{x}^\tau)] \leq \Delta\Phi_\tau([M], m) + \gamma V_{\tau,m}\mathcal{D}_\Delta(h).$$

On the simultaneous accuracy event in (7)*, an accepted count $m < b_{\tau+1/2}$ therefore satisfies*

$$\begin{aligned}\mathbb{E}_\tau[f(\boldsymbol{x}^{\tau+\Delta}) - f(\boldsymbol{x}^\tau)] \leq \Delta\left[\widehat{\Phi}_\tau(I_{\tau+1/2}, m) + \epsilon_\tau\right] \\ + \gamma V_{\tau,m}\mathcal{D}_\Delta(L\gamma).\end{aligned} \tag{28}$$

In particular, the conditional expected block loss decreases strictly if

$$-\widehat{\Phi}_\tau(I_{\tau+1/2}, m) > \epsilon_\tau + \frac{\gamma V_{\tau,m}}{\Delta}\mathcal{D}_\Delta(L\gamma).$$

The count-specific radius ϵ_m from (21) *may replace ϵ_τ on its corresponding simultaneous event.*

The drift sum bounds how far future responses can move from the initial one. At fixed positive $L\gamma$, it grows exponentially with the reuse length, so the displayed margin condition can cease to certify progress. Failure of that sufficient condition gives no conclusion about the actual block. Example F.2 verifies a useful regime; Proposition F.9 gives a sharper exact calculation.

Proof. Condition throughout on $\mathcal{F}_\tau^+$, so that $\boldsymbol{x}^\tau$ and m are fixed; no independence between the count and the search is assumed. Let U denote a uniform m-subset used only to average functions of a subset, independently of the production draws under this conditioning. The exact finite-population variance identity and Assumption 4.2 give

$$\begin{aligned}S_{\tau,m}^2 := \mathbb{E}_U\|g^\tau(U)\|^2 &= \|\nabla f(\boldsymbol{x}^\tau)\|^2 + \frac{q_m}{M}\sum_i \|\nabla f_i(\boldsymbol{x}^\tau) - \nabla f(\boldsymbol{x}^\tau)\|^2 \\ &\leq V_{\tau,m}.\end{aligned}$$

Also $\|\nabla f(\boldsymbol{x}^\tau)\| \leq S_{\tau,m}$ by Jensen's inequality. For any model y, individual smoothness gives $\|g_y(U) - g^\tau(U)\| \leq L\|y - \boldsymbol{x}^\tau\|$. Let C_j be the production subset at round $\tau + j$ and set $r_j = (\mathbb{E}_\tau\|\boldsymbol{x}^{\tau+j} - \boldsymbol{x}^\tau\|^2)^{1/2}$. Its conditional uniform law implies $\mathbb{E}_\tau\|g^\tau(C_j)\|^2 = S_{\tau,m}^2$, even though the production update at round $\tau + j$ also depends on earlier draws. Minkowski's inequality yields

$$r_{j+1} \leq (1+h)r_j + \gamma S_{\tau,m}, \qquad r_j \leq \frac{S_{\tau,m}}{L}((1+h)^j - 1),$$

since $r_0 = 0$. These bounds also establish the needed finite second moments.

For a deterministic model y, put $d = \|y - \boldsymbol{x}^\tau\|$. The difference between its subset response and the initial response is

$$\begin{aligned}&f(y - \gamma g_y(U)) - f(y - \gamma g^\tau(U)) \\ &\quad + f(y - \gamma g^\tau(U)) - f(y) - f(\boldsymbol{x}^\tau - \gamma g^\tau(U)) + f(\boldsymbol{x}^\tau).\end{aligned}$$

On the segment in the first line, the gradient of f has norm at most $\|\nabla f(\boldsymbol{x}^\tau)\| + Ld + h(\|g^\tau(U)\| + Ld)$. The segment length is at most γLd. Integrating the gradient difference along the segment from $\boldsymbol{x}^\tau$ to y bounds the second line in absolute value by $\gamma Ld\|g^\tau(U)\|$. Average these bounds over U, which has the same uniform law for both responses, and use $\mathbb{E}_U\|g^\tau(U)\| \le S_{\tau,m}$ to obtain

$$|\Phi_y([M],m) - \Phi_\tau([M],m)| \le \gamma Ld[\|\nabla f(\boldsymbol{x}^\tau)\| + (1+h)S_{\tau,m}] + \gamma L^2(1+h)d^2.$$

This inequality holds for every y, so substituting the random $y = \boldsymbol{x}^{\tau+j}$ does not require independence of that model and the search. Taking $\mathbb{E}_\tau$, using Cauchy–Schwarz and the bounds on r_j, gives

$$\mathbb{E}_\tau|\Phi_{\tau+j}([M],m) - \Phi_\tau([M],m)| \le \gamma V_{\tau,m}\big[(2+h)((1+h)^j - 1) + (1+h)((1+h)^j - 1)^2\big].$$

Conditional on $\mathcal{F}_\tau^+$ and $C_0,\dots,C_{j-1}$, the next expected loss change is $\Phi_{\tau+j}([M],m)$ by fresh uniform sampling. The tower property and telescoping therefore give the first bound. On the initial search event, substitute $\Phi_\tau([M],m) \le \widehat{\Phi}_\tau(I_{\tau+1/2},m) + \epsilon_\tau$. The event involves only search randomness and is not a condition on any future production trajectory. This proves (28). □

Proof of Theorem 4.5. For $j \ge 1$, the geometric sum gives $(1+h)^j - 1 \le jh(1+h)^{j-1}$. Hence

$$\begin{aligned}&(2+h)((1+h)^j - 1) + (1+h)((1+h)^j - 1)^2\\ &\quad = ((1+h)^j - 1)(1 + (1+h)^{j+1}) \le 2jh(1+h)^{2j}.\end{aligned}$$

The $j = 0$ summand is zero. Summing over $0 \le j < \Delta$ yields $\mathcal{D}_\Delta(h) \le h\Delta(\Delta-1)(1+h)^{2(\Delta-1)}$. Divide (28) by Δ, apply this inequality with $h = L\gamma$, and expand $V_{\tau,m}$ and q_m to obtain (8) on the same search event. □

Here $V_{\tau,m}$ controls only the initial aggregate-gradient second moment. No uniform gradient bound over the reuse trajectory is assumed. For fixed Δ, $\mathcal{D}_\Delta(h) = h\Delta(\Delta-1) + O(h^2)$ as $h \to 0$, and $\mathcal{D}_1(h) = 0$. Thus the per-round drift penalty is $O(\gamma^2)$ as $\gamma \to 0$ with the block length and initial model fixed. This finite-window scaling does not assert decay with the training horizon T. This is a certificate of expected progress over one finite block; it does not imply descent at every round or a stationarity rate. Long-run stationarity requires additional conditions, introduced below. The first bound above also covers the fallback, whose response has no acceptance certificate.

Example F.2 (A nonvacuous finite-block certificate)**.** Let $M = 4$, $f_1(x) = f_2(x) = 3x^2/8$ and $f_3(x) = f_4(x) = 5x^2/8$, so $f(x) = x^2/2$. Take $L = 5/4$, $\sigma_c = 0$, $\rho = 1/16$, $\gamma = 1/10$, a full auxiliary pool, $N = 10$, $\delta = 1/20$, and $\Delta = 2$. Both assumptions and individual client smoothness hold globally. Every singleton probe has response at most $-231(x^\tau)^2/3200$, so Algorithm 2 accepts $m_{\tau+1} = 1$ at a nonzero model. The main radius and block constants satisfy

$$\frac{\epsilon_\tau}{(x^\tau)^2} = \frac{37}{640}\sqrt{\frac{\log 320}{20}} < 0.03105, \qquad V_{\tau,1} = \frac{17}{16}(x^\tau)^2, \qquad \mathcal{D}_2(1/8) = \frac{145}{512}.$$

Consequently, on the $1-\delta$ search event, the right-hand side of (28) is bounded above by

$$\left[-\frac{231}{1600} + \frac{37}{320}\sqrt{\frac{\log 320}{20}} + \frac{2465}{81920}\right](x^\tau)^2 < -0.0521(x^\tau)^2.$$

Thus the general certificate can confirm expected two-round progress with nonzero heterogeneity and finite search depth. The simpler main bound (8) also certifies progress: its per-round drift term is $1377(x^\tau)^2/81920$, so its right-hand side is below $-0.0243(x^\tau)^2$. This calculation establishes nonvacuity of the bound; the benchmark constants are not estimated here.

Example F.3 (Negative responses without stationarity)**.** Let $f(x,y) = x^2/2 + 1 - \cos y$, take two clients $f_1 = f + h(x)$, $f_2 = f - h(x)$, and define

$$h(x) = \tfrac{1}{2}\arctan(x^2 - 1), \qquad r(x) = \frac{1}{1 + (x^2-1)^2}, \qquad h'(x) = r(x)x.$$

Both local losses are bounded below, and $\|\nabla f_i - \nabla f\|^2 = r(x)^2 x^2 \le \|\nabla f\|^2$. With $\gamma = 1$, $y_0 = 0$, $x_0 = 1.1$, and $b = M = 2$, either single-client update has magnitude $|x^+| = r(x)|x|$. For $1 < z \le 1.1$,

$$1 < \frac{z}{1 + (z^2 - 1)^2} < z,$$

because $(z-1)(z+1)^2 < 1$. Hence every one-client candidate has the same strictly negative response, so Algorithm 2 selects $m = 1$ for any N. Nevertheless, $|x_t|$ decreases to 1: a limit above 1 would contradict the fixed-point equation $r(x)|x| = |x|$. Therefore $\|\nabla f(x_t, 0)\|^2 \to 1$. The sign of progress is correct at every round, but its magnitude vanishes without reaching stationarity. This motivates the quantitative response margin in Theorem F.8.

From block progress to stationarity. To obtain a long-run rate, the progress margin must persist relative to the squared global gradient. We now state sufficient conditions for this stronger conclusion. Unlike the finite-window bound above, this result uses pointwise relative heterogeneity. In addition to L-smoothness and $f \ge f_{\inf}$, assume $L > 0$. Impose the additional pointwise relative bound, with its own parameter $\varrho \ge 0$,

$$\max_i \|\nabla f_i(x) - \nabla f(x)\| \le \sqrt{\varrho}\, \|\nabla f(x)\| \tag{29}$$

at all possible iterates. This is stronger than the average additive bound in Assumption 4.2; it forces every local gradient to vanish at a stationary global point.

Let $\mathcal{H}_s$ be the history immediately before the controller at adaptation start s, and write $b_s = b_{s+1/2}$. Suppose that, whenever $\nabla f(x^s) \ne 0$, there is a count $m_\star(s) \le b_s$ such that

$$\begin{aligned} \Phi_s([M], m) &\ge \gamma\kappa \|\nabla f(x^s)\|^2 \quad \text{for every } m < m_\star(s), \\ \Phi_s([M], m_\star(s)) &\le -\gamma\eta \|\nabla f(x^s)\|^2, \end{aligned} \tag{30}$$

where $\kappa, \eta > 0$ are fixed. This assumption states that the response curve has a separated first negative inside the search cap. It does not assume that the random count returned by Algorithm 2 descends.

Set

$$K = 1 + \sqrt{\varrho}, \quad Q = 1 + \gamma L K, \quad U = \max\{0, -1 + (L\gamma/2)(1+\varrho)\}. \tag{31}$$

Let $\delta_A, \delta_N \in (0,1)$ with $\delta = \delta_A + \delta_N < 1$. Require, at every nonstationary adaptation, that the simultaneous response radii satisfy

$$a_{b_s} + e_m < \gamma \min\{\kappa, \eta\} \|\nabla f(x^s)\|^2 \qquad \text{for every } m < b_s. \tag{32}$$

For $b_s > 1$, an explicit sufficient condition using (15) and (18) is

$$\begin{aligned} 2\sqrt{\varrho}\,[1 + L\gamma K] \left[\sqrt{\frac{\min\{b_s, M - b_s\}}{2 b_s^2} \log \frac{2 b_s}{\delta_A}} + \sqrt{\frac{1}{2N} \log \frac{2(b_s - 1)}{\delta_N}} \right] \\ < \min\{\kappa, \eta\}. \end{aligned} \tag{33}$$

The first term is zero for $b_s = M$. To verify this condition, use $D \le 2\sqrt{\varrho}\|\nabla f(x^s)\|$, choose $G = \max_i \|\nabla f_i(x^s)\| \le K \|\nabla f(x^s)\|$, and use $K_{x^s} \le [1 + L\gamma K]\|\nabla f(x^s)\|$. When $m_\star(s) = b_s$, only the earlier positive margin needs estimation accuracy. For $b_s = 1$ there are no estimates and no accuracy requirement; the negative fallback response is supplied by (30).

An additional normalized-response stability condition. Suppose there is a constant $\omega \in [0, \eta)$ such that, for every adaptation start s and every possible production trajectory obtained by holding the oracle count $m_\star(s)$ for the next block of length $\ell_s \le \Delta$,

$$\frac{\Phi_{s+j}([M], m_\star(s))}{\gamma \|\nabla f(x^{s+j})\|^2} \le \frac{\Phi_s([M], m_\star(s))}{\gamma \|\nabla f(x^s)\|^2} + \omega. \tag{34}$$

The inequality is required for $0 \le j < \ell_s$ whenever $\nabla f(x^{s+j}) \ne 0$. It controls how the progress per squared gradient changes between adaptations. It is a further response-regime assumption, not a consequence asserted from smoothness. At stationary points (29) makes the trajectory absorbing, so no normalization is needed there. Set

$$a = \eta - \omega, \qquad c_\Delta = a - (a + U)\delta\, \Delta Q^{2(\Delta - 1)}. \tag{35}$$

Theorem F.8 (Finite-depth periodic ISP in a stable-response regime)**.** *Assume f is L-smooth, $f \geq f_{\inf} > -\infty$, and the additional relative condition* (29)*, response separation* (30)*, sampling accuracy* (32)*, and normalized response stability* (34) *hold. Let production cohorts be fresh uniform subsets, and reuse each selected count for a predetermined block of at most Δ rounds. If $c_\Delta > 0$, then the unchanged Algorithm 2 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{f(x^0)-f_{\inf}}{\gamma c_\Delta T}. \tag{36}$$

Proof. Fix an adaptation start s. The two simultaneous concentration lemmas, Corollary F.5, and (30)–(32) give a recovery event $\mathcal{G}_s = \{m_{s+1} = m_\star(s)\}$ with

$$\mathbb{P}(\mathcal{G}_s^c \mid \mathcal{H}_s) \leq \delta.$$

For $b_s = 1$ recovery is deterministic. If $\nabla f(x^s) = 0$, all local gradients vanish and the subsequent block contributes zero stationarity error. We therefore consider $\nabla f(x^s) \neq 0$.

At any production point, for every possible count, the reference progress bound and (29) give

$$\Phi_t([M], m) \leq \gamma\left[-1 + \frac{L\gamma}{2}(1+\varrho q_m)\right]\|\nabla f(x^t)\|^2 \leq \gamma U\|\nabla f(x^t)\|^2. \tag{37}$$

On $\mathcal{G}_s$, the chosen count is the oracle count. Its initial normalized response is at most $-\eta$, and (34) ensures throughout the block that

$$\Phi_{s+j}([M], m_{s+1}) \leq -\gamma(\eta-\omega)\|\nabla f(x^{s+j})\|^2 = -\gamma a\|\nabla f(x^{s+j})\|^2.$$

At a zero gradient this inequality holds without normalization. On $\mathcal{G}_s^c$, use (37).

It remains to handle dependence between the recovery event and the subsequent trajectory. Every realized cohort gradient has norm at most $K\|\nabla f(x^t)\|$. Smoothness implies

$$\|\nabla f(x^{t+1})\| \leq (1+\gamma LK)\|\nabla f(x^t)\| = Q\|\nabla f(x^t)\|.$$

Consequently, for the block energy $S_s = \sum_{j=0}^{\ell_s-1}\|\nabla f(x^{s+j})\|^2$,

$$\|\nabla f(x^s)\|^2 \leq S_s \leq \ell_s Q^{2(\ell_s-1)}\|\nabla f(x^s)\|^2 \tag{38}$$

on every production trajectory. The predetermined block length ℓ_s is $\mathcal{H}_s$-measurable, so it is fixed under this conditioning. Therefore

$$\begin{aligned}\mathbb{E}[S_s\mathbf{1}_{\mathcal{G}_s^c} \mid \mathcal{H}_s] &\leq \delta\,\ell_s Q^{2(\ell_s-1)}\|\nabla f(x^s)\|^2 \\ &\leq \delta\,\ell_s Q^{2(\ell_s-1)}\mathbb{E}[S_s \mid \mathcal{H}_s].\end{aligned} \tag{39}$$

The second inequality uses the lower bound in (38); it does not assume independence between S_s and $\mathcal{G}_s$.

Condition on each production history, which includes the count-selection event, and use its fresh cohort to replace the loss change by $\Phi_{s+j}([M], m_{s+1})$ in conditional expectation. Summing the good- and bad-event bounds and applying (39) gives

$$\begin{aligned}\mathbb{E}[f(x^{s+\ell_s}) - f(x^s) \mid \mathcal{H}_s] &\leq -\gamma a\,\mathbb{E}[S_s \mid \mathcal{H}_s] + \gamma(a+U)\mathbb{E}[S_s\mathbf{1}_{\mathcal{G}_s^c} \mid \mathcal{H}_s] \\ &\leq -\gamma\big[a - (a+U)\delta\,\ell_s Q^{2(\ell_s-1)}\big]\mathbb{E}[S_s \mid \mathcal{H}_s] \\ &\leq -\gamma c_\Delta\,\mathbb{E}[S_s \mid \mathcal{H}_s].\end{aligned}$$

The last inequality uses $\ell_s \leq \Delta$ and $Q \geq 1$. Take total expectations, truncate the final block at T if necessary, sum over blocks, and use $f \geq f_{\inf}$. The same bounds hold on every prefix of a block. The block energies partition $\sum_{t=0}^{T-1}\|\nabla f(x^t)\|^2$, proving the result. □

The theorem states an explicit regime for the unchanged controller. Its response gaps and stability conditions are not enforced by the empirical-negative stopping rule. The failure penalty includes the possible dependence between an incorrect selection and a subsequently larger gradient norm. Explicitly, $c_\Delta > 0$ requires $\delta < (\eta-\omega)/[(\eta-\omega+U)\Delta Q^{2(\Delta-1)}]$. For fixed $Q > 1$, this sufficient failure tolerance shrinks exponentially with the window. Greater depth reduces Monte Carlo error, while the pool term in (33) also depends on auxiliary coverage. At $\Delta = 1$ one may take $\omega = 0$, giving the coefficient $\eta - (\eta+U)\delta$. For longer blocks, the normalized response condition allows the progress magnitude to vary with the current squared gradient.

An alternative drift bound under relative heterogeneity. For exact reference updates under the pointwise relative condition above, one can also bound absolute response drift by the initial squared gradient. This optional comparison uses the preceding constants K and Q; Theorem F.8 does not require it. Converting an absolute bound to normalized stability additionally requires control of the changing gradient denominator. If each client gradient is additionally L_c-Lipschitz, define

$$D_j = \frac{Q^j - 1}{L}, \qquad W_j = [LK + L_c(1 + L\gamma K)]D_j + LL_c(1 + \gamma L_c/2)D_j^2.$$

For any block obeying the pointwise relative condition, every count m satisfies the pathwise bound

$$|\Phi_{s+j}([M], m) - \Phi_s([M], m)| \le \gamma W_j \|\nabla f(x^s)\|^2. \tag{40}$$

To prove it, first sum the update-norm bounds from the proof above:

$$\|x^{s+j} - x^s\| \le \gamma K \sum_{i=0}^{j-1} Q^i \|\nabla f(x^s)\| = D_j \|\nabla f(x^s)\|.$$

For a fixed cohort C and two points x, y, put $u = g_x(C)$, $e = g_y(C) - g_x(C)$, and $d = \|y - x\|$. Client smoothness gives $\|e\| \le L_c d$. Write the subset response as $\delta f_z(m, C) = f(z - \gamma g_z(C)) - f(z)$, where $m = |C|$, and split

$$\begin{aligned}\delta f_y(m, C) - \delta f_x(m, C) = {} & f(y - \gamma(u + e)) - f(y - \gamma u) \\ & + f(y - \gamma u) - f(y) - f(x - \gamma u) + f(x).\end{aligned}$$

Integrating the gradient difference along $y - x$ bounds the absolute value of the second line by $L\gamma\|u\|d$. The two-sided smoothness remainder bounds the first line by

$$\gamma\|\nabla f(y - \gamma u)\|\|e\| + \frac{L\gamma^2}{2}\|e\|^2.$$

Use $\|\nabla f(y-\gamma u)\| \le \|\nabla f(x)\| + Ld + L\gamma\|u\|$ and $\|u\| \le K\|\nabla f(x)\|$ at the starting point to obtain

$$\begin{aligned}|\delta f_y(m, C) - \delta f_x(m, C)| \le {} & \gamma\|\nabla f(x)\|[LK + L_c(1 + L\gamma K)]d \\ & + \gamma LL_c(1 + \gamma L_c/2)d^2.\end{aligned} \tag{41}$$

Uniform subset laws do not change with the model. Average over C, take $x = x^s$ and $y = x^{s+j}$, and substitute the displacement bound to prove (40).

The absolute modulus can be large even when normalized responses are stable. The next examples calculate the responses directly.

F.4 Exact examples and attainable regimes

The quadratic family gives an exact depth–reuse tradeoff; its nonconvex extension isolates one-step selection error. The curvature construction verifies response separation and stability from primitive assumptions. Finally, the adaptive-pool construction combines convergence with auxiliary-update accounting. These are separate demonstrations of attainable regimes, with their own proofs.

Exact depth–reuse calculation. The preceding general bounds are conservative. An exactly solvable reference family makes the interaction between finite depth and periodic reuse explicit, including regimes with a nontrivial selected count.

Proposition F.9 (Exact block criterion for scalar quadratic clients)**.** *Let $f_i(x) = a_i x^2/2$, where $a_i \ge 0$ and $M^{-1}\sum_i a_i = 1$. Use uniform sampling, exact evaluations, a fixed auxiliary budget b, and independent fresh search randomness between adaptation blocks. The budget is constant, for example, when the expansion ratio makes $b = M$ at every adaptation. For $x \ne 0$, let P_m be the probability that the unchanged search returns $m \le b$, including its fallback, and define*

$$r_m = \mathbb{E}_{|C|=m}\left(1 - \gamma\frac{1}{m}\sum_{i\in C} a_i\right)^2, \qquad R_\Delta = \sum_{m\le b} P_m r_m^\Delta.$$

Then P_m does not depend on x, and at block boundaries

$$\mathbb{E}[x^2_{(k+1)\Delta} \mid x_{k\Delta}] = R_\Delta x^2_{k\Delta}. \tag{42}$$

For $x_0 \neq 0$, mean-square convergence at those boundaries occurs if and only if $R_\Delta < 1$. In that case, writing $B_\Delta = \sum_{j=0}^{\Delta-1} \sum_m P_m r_m^j$ gives, for $T = K\Delta$,

$$\frac{1}{T} \sum_{t<T} \mathbb{E}|\nabla f(x_t)|^2 \leq \frac{x_0^2 B_\Delta}{T(1 - R_\Delta)}. \tag{43}$$

Proof. Every candidate response is x^2 times a quantity depending only on its sampled coefficients, so the comparison with zero and the law of the returned count are independent of nonzero x. Conditional on that count, fresh training subsets multiply the squared iterate by independent factors with mean r_m. Reusing the same count for Δ steps therefore gives r_m^Δ, and averaging over the search proves (42). At $x = 0$, the point is absorbing and the same identity holds. Iteration gives $\mathbb{E}x^2_{k\Delta} = R_\Delta^k x_0^2$. The expected sum of squared gradients within a block is $B_\Delta \mathbb{E}x^2_{k\Delta}$. Sum the geometric series to obtain (43). □

Example F.4 (Depth and reuse interact)**.** Take $M = 4$, $a = (1/4, 1/4, 1/4, 13/4)$, $\gamma = 1$, and $b = M$. Every two-client response is negative, so the search returns either one or two clients. For a one-client test, if $Z \sim \mathrm{Binomial}(N, 1/4)$ counts the large-curvature client among the probes, the first count is accepted exactly when $72Z < 7N$. Consequently,

$$P_1 = p_N := \mathbb{P}(72Z < 7N), \quad P_2 = 1 - p_N, \qquad r_1 = 27/16, \quad r_2 = 9/16.$$

The exact condition becomes

$$p_N < \frac{1 - (9/16)^\Delta}{(27/16)^\Delta - (9/16)^\Delta}. \tag{44}$$

For $N = 4$, $p_N = (3/4)^4 = 81/256$: the condition holds at $\Delta = 1$ but fails at $\Delta = 2$. For $N = 5$, $p_N = 243/1024$, and it holds also at $\Delta = 2$. Thus the unchanged controller can converge with a finite search depth and reused counts even when the smallest count has positive expected response. The same example shows why increasing the reuse period can turn a mean-square contracting finite-depth rule into a mean-square expanding one.

For this family, $\Phi_x([M], m)/(\gamma|\nabla f(x)|^2) = (r_m - 1)/(2\gamma)$ is constant in x. The normalized drift condition in Theorem F.8 therefore holds with zero drift; its other conditions hold for sufficiently large finite N at any fixed Δ. This supplies a nontrivial regime for that theorem, while (44) gives a substantially sharper exact criterion. These are analytic examples, not additional training experiments or estimates of the constants in our benchmarks.

Remark F.10 (Conservative general certificate)*.* The same family satisfies Theorem F.8 with $\varrho = 81/16$, $\kappa = 11/32$, $\eta = 7/32$, $\omega = 0$, $\Delta = 2$, $\delta_A = \delta_N = 1/2000$, and $N = 40000$. These give $K = 13/4$, $Q = 17/4$, $U = 65/32$, $c_\Delta = 4399/32000 > 0$, and an accuracy bound $(153/8)\sqrt{\log(12000)/80000} < 7/32$. The exact calculation already permits $N = 5$: the general theorem supplies a sufficient regime, while the exact response law gives a sharper threshold.

The next example uses the same client coefficients to isolate the one-step selection error in a nonconvex objective. Example F.4 shows how reuse can amplify its effect.

Example F.5 (Finite-depth acceptance need not imply expected descent)**.** Let $f(x, y) = x^2/2 + 1 - \cos y$ and $f_i = c_i f$, where $M = 4$ and $c = (1/4, 1/4, 1/4, 13/4)$. The global objective is f, which is nonconvex, bounded below, and 1-smooth. Moreover, $\|\nabla f_i - \nabla f\|^2 \leq (81/16)\|\nabla f\|^2$ globally. At round $\tau = 0$, take $\gamma = 1$, $(x, y) = (1, 0)$, $b = M$, $N = 1$, and uniform sampling. For $m = 1$, a sampled candidate has response $-7/32$ with probability $3/4$ and $65/32$ with probability $1/4$. Thus $\Phi_0([M], 1) = 11/32 > 0$. Every two-client candidate has response $-7/32$, so $\Phi_0([M], 2) = -7/32$. Algorithm 2 returns one client with probability $3/4$ and two clients otherwise: its accepted estimate is always negative, whereas the fresh production response has expectation

$$\mathbb{E}\Phi_0([M], m_1) = \frac{3}{4}\frac{11}{32} - \frac{1}{4}\frac{7}{32} = \frac{13}{64} > 0.$$

The failure occurs with a full auxiliary pool and exact losses and gradients. It is caused by selecting the count using the finite sample.

A nonlinear certificate for the periodic assumptions. The normalized-response premise need not be postulated directly for every family. A curvature bound supplies one checkable sufficient condition, including nonlinear strongly convex objectives.

Proposition F.11 (Curvature-band certificate)**.** *Suppose $f_i(x) = a_i h(x)$, with $a_i \geq 0$, $M^{-1}\sum_i a_i = 1$, and h twice differentiable with $L_-\mathbf{I} \preceq \nabla^2 h(x) \preceq L_+\mathbf{I}$ everywhere, where $0 < L_- \leq L_+$. Write $v_a = M^{-1}\sum_i (a_i - 1)^2$ and $A_m = 1 + v_a q_m$. At any nonstationary point, the uniform response satisfies*

$$-1 + \frac{\gamma L_-}{2} A_m \leq \frac{\Phi_x([M], m)}{\gamma \|\nabla h(x)\|^2} \leq -1 + \frac{\gamma L_+}{2} A_m. \tag{45}$$

If the left endpoint is at least $\kappa > 0$ for every $m < m_\star$ and the right endpoint at $m_\star$ is at most $-\eta < 0$, the separated-response condition holds globally. Its normalized drift condition holds with

$$\omega = \frac{\gamma(L_+ - L_-)}{2} A_{m_\star}, \tag{46}$$

provided $\omega < \eta$, for any length of an oracle-count block.

Proof. For a sampled subset, write $a_C = m^{-1}\sum_{i \in C} a_i$. The update is $a_C \nabla h(x)$. The two-sided Taylor bound from the Hessian inequalities places its normalized response between $-a_C + (\gamma L_-/2)a_C^2$ and $-a_C + (\gamma L_+/2)a_C^2$. Uniform sampling gives $\mathbb{E}a_C = 1$ and $\mathbb{E}a_C^2 = A_m$, proving the interval. At any two nonstationary points, the difference between responses normalized by their own squared gradients is at most the width of this same interval. This proves (46), including along every possible production trajectory. At stationary points all proportional local gradients vanish. □

Example F.6 (A nonconstant curvature family)**.** Let M be divisible by four. Take three quarters of the coefficients equal to $1/4$ and one quarter equal to $13/4$, and let $h(x) = x^2/2 + (1 - \cos x)/100$, which has nonconstant curvature in $[99/100, 101/100]$. For $\gamma = 100/101$ and $m_\star = 2$,

$$\kappa = 1025/3232, \qquad \eta = 5/64, \qquad \omega \leq 59/3232 < \eta.$$

Indeed, $A_1 = 43/16$ and $A_2 \leq 59/32$ give these constants directly. The pointwise relative bound holds with $\varrho = 81/16$. Thus separation and normalized stability follow from primitive curvature and coefficient bounds in this nonlinear family. Search-cap coverage, estimation accuracy, and positive c_Δ remain the other requirements of Theorem F.8; this proposition does not assert them for arbitrary auxiliary budgets or experimental objectives.

Which communication cost is bounded? The mathematical count bounds concern client-update exchanges. For a realized trajectory, let $\mathcal{T}_\Delta = \{s \in \{0, \ldots, T-1\} : s \equiv 0 \pmod{\Delta}\}$ denote the adaptation rounds. Their total is

$$C_{\text{updates}} = \sum_{t=0}^{T-1} m_{t+1} + \sum_{s \in \mathcal{T}_\Delta} b_{s+1/2}. \tag{47}$$

Relative to a constant $\bar{m}$ over the same T rounds, fewer such exchanges occur exactly when $\sum_t (\bar{m} - m_{t+1}) > \sum_s b_{s+1/2}$. Candidate-model downlinks, scalar loss reports, and wall-clock costs are additional quantities accounted for in the experimental communication model. In particular, the baseline objective evaluation over $[M]$ in Algorithm 4 entails M scalar loss reports at an adaptation and may require model downlinks when the evaluated state is not already available locally. These evaluations are separate from the client update exchanges in (47). The theory explains when small counts preserve a specified progress margin; it does not establish universal cost dominance over a tuned fixed-count run, whose trajectory and time to a target can differ. The periodic result makes the relevant trade-off explicit: a larger Δ amortizes the auxiliary term, while increasing the possible response drift between adaptations.

Adaptive partial pools with auxiliary-update accounting. We next exhibit a reference regime in which the auxiliary budget follows Algorithm 2, finite-depth selection converges, and auxiliary update exchanges are included in a cost bound. This calculation uses the response distribution directly, giving a sharper sufficient condition than the general simultaneous worst-case radii. The parameters are conservative existence choices. The cost comparison is with full participation over the same number of rounds; a fixed two-client rule also converges in this family.

Theorem F.12 (Progress and auxiliary-update cost for the unchanged controller)**.** *Let $M \geq 10000$ be divisible by eight. Set $f_i(x) = a_i x^2/2$, where $3M/4$ coefficients equal $1/4$ and $M/4$ equal $13/4$, so $f(x) = x^2/2$. Use Algorithm 2 with uniform sampling, exact responses, fresh production subsets, $\gamma = M/(M+1)$, $N = 800$, $p = 499$, $\Delta = 2$, initial count $m_0 = 2$, and $x^0 \neq 0$. Write $\widehat{m}_k$ for the count returned at round $2k$, put $\widehat{m}_{-1} = 2$, and let $b_k = \min\{M, 500\widehat{m}_{k-1}\}$ be its actual auxiliary budget. For T even, the following statements hold, with $R_0 = 38637/51200 < 1$:*

$$\mathbb{E}[(x^{2k+2})^2 \mid \mathcal{H}_{2k}] \leq R_0 (x^{2k})^2, \tag{48}$$

$$\frac{1}{T}\sum_{t<T} \mathbb{E}|\nabla f(x^t)|^2 \leq \frac{43(x^0)^2}{16(1-R_0)T}, \tag{49}$$

$$\sup_k \mathbb{E}\widehat{m}_k \leq \frac{32}{7}, \qquad \sup_k \mathbb{E}b_k \leq \frac{16000}{7}, \tag{50}$$

$$\mathbb{E}C_{\text{updates}}(T) \leq \frac{8032}{7}T < 0.115\,MT. \tag{51}$$

The first truly negative full-pool count is two, while every $b_k \geq 500$. The initial auxiliary pool has size $1000 < M$; for every $k \geq 1$, $\mathbb{P}(b_k = M) < 1/4000$. Thus the analysis includes the cap's occasional full-pool events rather than replacing the adaptive budget by a fixed pool.

Proof. Nonzero finite iterates. For a production subset of size m containing h high-coefficient clients, $\bar{a}_C = (m+12h)/(4m)$, and $x^+ = (1-\gamma\bar{a}_C)x$. A zero multiplier would require $12Mh = m(3M+4)$. Since M is divisible by eight, $\gcd(12M, 3M+4) = 4$, so this equality requires $3M$ to divide m, which is impossible for $1 \leq m \leq M$. Thus $x^t \neq 0$ at every finite round. This condition matters because at exact stationarity the strict sign test would always take its fallback.

Controlling harmful one-client acceptance. Condition on the pre-search history, including $b = b_k \geq 500$. Let θ_I be the high-coefficient fraction in the uniform auxiliary pool. Its population fraction is $1/4$. Sampling without replacement gives

$$\mathbb{P}(\theta_I < 3/20) \leq e^{-b/50} \leq e^{-10}.$$

For completeness, exchanging one pool element changes θ_I by at most $1/b$. The reveal argument of Lemma F.3, with at most b revelations, gives the one-sided bound $\mathbb{P}(\theta_I - 1/4 \leq -u) \leq e^{-2bu^2}$ used here. Conditional on I, the count-one probes have independent high-client indicators. If Z is their total, the estimate is negative exactly when

$$\frac{Z}{N} < \frac{8-\gamma}{168\gamma - 96} \leq \frac{1}{10},$$

where the last inequality holds for $\gamma \geq 88/89$. On $\theta_I \geq 3/20$, conditional Hoeffding bounds this event by $e^{-N/200} = e^{-4}$. Therefore

$$\mathbb{P}(\widehat{m}_k = 1 \mid \mathcal{H}_{2k}) \leq e^{-10} + e^{-4} < \frac{1}{50}. \tag{52}$$

Production contraction. The coefficient population variance is $27/16$. For any count,

$$r_m = \mathbb{E}(1-\gamma\bar{a}_C)^2 = (1-\gamma)^2 + \gamma^2\frac{27}{16}q_m.$$

Here $r_1 \leq 27/16$ and $r_m \leq 27/32$ for every $m \geq 2$, using $q_m \leq 1/2$ and $\gamma \geq 5/59$ in the second inequality. In particular, $r_1 > 1$ for $\gamma > 32/43$, whereas $r_m < 1$ for $m \geq 2$. Thus the first negative expected response is at count two. Conditional on a selected count, two independent production subsets give the squared multiplier r_m^2 in expectation. Applying (52) yields

$$\mathbb{E}[(x^{2k+2})^2 \mid \mathcal{H}_{2k}] \leq \left[\frac{1}{50}\left(\frac{27}{16}\right)^2 + \frac{49}{50}\left(\frac{27}{32}\right)^2\right](x^{2k})^2 = R_0(x^{2k})^2.$$

The expected squared-gradient sum within a block is at most $(1+27/16)(x^{2k})^2 = (43/16)(x^{2k})^2$. Iterate the block contraction and sum the geometric series to obtain (49).

Controlling large returned counts. The same pool concentration gives $\mathbb{P}(\theta_I \notin [3/20, 7/20]) \le 2e^{-10}$. For a four-client probe and $\gamma = 1$, its conditional mean response divided by x^2 is

$$-\bar{a}_I + \tfrac{1}{2}\bar{a}_I^2 + \tfrac{1}{2}\frac{b-4}{4(b-1)}\,9\theta_I(1-\theta_I), \qquad \bar{a}_I = \tfrac{1}{4} + 3\theta_I.$$

The within-pool finite-population factor $(b-4)/(4(b-1))$ is at most $1/4$. Replacing it by $1/4$ bounds the expression by $-637/3200$: the resulting convex quadratic in θ_I takes values $-997/3200$ and $-637/3200$ at the two endpoints. For $\gamma = M/(M+1)$, any coefficient average a changes the normalized response relative to $\gamma = 1$ by

$$[-\gamma a + \gamma^2 a^2/2] - [-a + a^2/2] \le (1-\gamma)a \le \frac{13}{4(M+1)}.$$

Consequently the conditional four-client mean is at most $-19x^2/100$. Every individual probe response divided by x^2 lies in $[-1/2, 65/32]$, a range of length $81/32$. Conditional Hoeffding therefore bounds rejection of this count by $\exp\{-1600(152/2025)^2\}$ on the pool event. Returning a count above four requires rejecting its pre-generated estimate, whether or not the search reaches that estimate. Hence

$$\mathbb{P}(\widehat{m}_k > 4 \mid \mathcal{H}_{2k}) \le q_0 := 2e^{-10} + \exp\{-1600(152/2025)^2\} < \frac{1}{4000}. \tag{53}$$

This also covers the untested fallback, since $b_k \ge 500 > 4$.

Adaptive budget and communication. Pathwise, $\widehat{m}_k \le 4 + b_k \mathbf{1}_{\{\widehat{m}_k > 4\}}$. Because b_k is fixed before its search and is at most $500\widehat{m}_{k-1}$, the preceding bound gives

$$\mathbb{E}\widehat{m}_k \le 4 + 500 q_0 \mathbb{E}\widehat{m}_{k-1} \le 4 + \tfrac{1}{8}\mathbb{E}\widehat{m}_{k-1}.$$

Starting at two, this recurrence implies $\mathbb{E}\widehat{m}_k \le 32/7$, and hence $\mathbb{E}b_k \le 16000/7$. Each block uses $2\widehat{m}_k + b_k$ client-update exchanges. Summing these expectations proves (51). For $k \ge 1$, $b_k = M$ requires $\widehat{m}_{k-1} \ge M/500 \ge 20 > 4$, so (53) gives the stated full-pool probability. □

This construction proves coexistence of finite-depth selection, periodic reuse, an adaptive partial pool, and bounded auxiliary-update overhead. The bound counts update exchanges as defined in (47). Candidate-evaluation traffic and inference cost require separate accounting. This establishes an attainable regime; quantifying the experimental savings or an advantage over a tuned fixed-count rule requires different evidence.

F.5 General selectors and inclusion diagnostics

The uniform-pool result exploits the marginal identity of nested uniform sampling. A general CSS need not have that identity. We instead retain both its mean update and its variance, which are distinct aspects of the response. All quantities in this subsection are conditional on a fixed model and selector state. Keep the method's aggregate $g_x(C)$ and define its mean and variance under the chosen selector:

$$\bar{g}_x(A,m) = \mathbb{E}_{C\sim\mathcal{S}(A,m)} g_x(C), \quad V_A(m) = \mathbb{E}\|g_x(C) - \bar{g}_x(A,m)\|^2.$$

For the variance, subscripts F and I denote the full population and auxiliary pool. The auxiliary-pool law may itself be nonuniform. Define the absolute mean distortion $D_g(m,b) = \mathbb{E}_I\|\bar{g}_x(I,m) - \bar{g}_x([M],m)\|^2$ and the fixed-count gap $\Gamma(m,b) = \mathbb{E}_I|\Phi_x(I,m) - \Phi_x([M],m)|$.

Proposition F.13 (Response comparison with explicit second moments)**.** *For L-smooth f,*

$$\Gamma(m,b) \le \gamma(\|\nabla f(x)\| + L\gamma\|\bar{g}_x([M],m)\|)\sqrt{D_g(m,b)} + \frac{L\gamma^2}{2}D_g(m,b) + \frac{L\gamma^2}{2}\big(\mathbb{E}_I V_I(m) + V_F(m)\big). \tag{54}$$

Proof. For either pool A, expand at $x - \gamma\bar{g}_x(A,m)$. The linear term in $g_x(C) - \bar{g}_x(A,m)$ has expectation zero, so

$$|\Phi_x(A,m) - [f(x - \gamma\bar{g}_x(A,m)) - f(x)]| \le \frac{L\gamma^2}{2}V_A(m).$$

For the two mean updates, expand at $y = x - \gamma \bar{g}_x([M], m)$ with displacement $-\gamma(\bar{g}_x(I, m) - \bar{g}_x([M], m))$. The two-sided smoothness remainder gives

$$\begin{aligned}|f(x - \gamma \bar{g}_x(I, m)) - f(x - \gamma \bar{g}_x([M], m))| \\ \leq \gamma(\|\nabla f(x)\| + L\gamma \|\bar{g}_x([M], m)\|)\|\bar{g}_x(I, m) - \bar{g}_x([M], m)\| \\ + \frac{L\gamma^2}{2}\|\bar{g}_x(I, m) - \bar{g}_x([M], m)\|^2.\end{aligned}$$

Take expectations and use Cauchy–Schwarz and the preceding variance bounds. □

The response gap has two sources: a change in the mean update and variation around each mean. The sum of variances is conservative: when the two subset laws coincide, the true gap is exactly zero even if this bound is positive. In particular, $b = M$ gives zero gap directly. For uniform sampling, Lemma F.3 is sharper and controls the selected count without these variance envelopes.

From fixed-count gaps to adaptive counts. For each candidate count choose a pool-error radius $a_m^{\mathrm{pool}} > 0$. Markov's inequality and a union bound give

$$\mathbb{P}_I\big(\exists m < b : |\Phi_x(I, m) - \Phi_x([M], m)| > a_m^{\mathrm{pool}}\big) \leq \sum_{m<b} \frac{\Gamma(m, b)}{a_m^{\mathrm{pool}}}. \tag{55}$$

The gap bound (54) can upper-bound each numerator. A pool confidence level $1 - \delta_A$ requires this sum to be at most δ_A. Combining it with a separately established simultaneous Monte Carlo event of probability at least $1 - \delta_N$ gives total radii $a_m^{\mathrm{pool}} + e_m$ and failure probability at most $\delta_A + \delta_N$. The deterministic selection argument in Appendix F.2 then applies to the same full-pool CSS. An accepted negative margin larger than its total radius certifies descent; otherwise this transfer can be uninformative. A fixed-count expectation cannot be evaluated at a pool-dependent selected count. Agreement with the full-pool CSS alone also does not establish its alignment with the global gradient.

What inclusion probabilities identify. For a fixed count, let $\pi_i = \mathbb{P}(i \in C \mid [M], m)$ and $\psi_i(I) = \mathbb{P}(i \in C \mid I, m)$, extending ψ_i by zero outside I. Put $p_i = \pi_i/m$, $q_i(I) = \psi_i(I)/m$. Then $\bar{g}_x([M], m) = \sum_i p_i \nabla f_i(x)$, $\bar{g}_x(I, m) = \sum_i q_i \nabla f_i(x)$, and both weight vectors sum to one. These formulas describe uniform aggregation of the chosen updates. A selector coupled with importance reweighting has different effective weights, which must be used in its own analysis.

Proposition F.14 (Absolute inclusion bounds)**.** *The squared mean errors satisfy*

$$\begin{aligned}\|\bar{g}_x([M], m) - \nabla f(x)\|^2 &\leq \left[\sum_i (p_i - 1/M)^2\right] \sum_{j=1}^M \|\nabla f_j(x) - \nabla f(x)\|^2, \\ D_g(m, b) &\leq \left[\mathbb{E}_I \sum_i (q_i(I) - p_i)^2\right] \sum_{j=1}^M \|\nabla f_j(x) - \nabla f(x)\|^2.\end{aligned} \tag{56}$$

These bounds remain meaningful at $\nabla f(x) = 0$.

Proof. The weight differences sum to zero. Subtract $\nabla f(x)$ from each $\nabla f_i(x)$, then apply Cauchy–Schwarz to the weighted sum. The sum of squared client deviations is $\sum_{j=1}^M \|\nabla f_j(x) - \nabla f(x)\|^2$. The second inequality follows after averaging over I. □

At a nonstationary point, define the actual normalized mean errors

$$\begin{aligned}\delta^{\mathcal{S}}(m) &= \frac{\|\bar{g}_x([M], m) - \nabla f(x)\|^2}{\|\nabla f(x)\|^2}, \\ \beta^{\mathcal{S}}(m, b) &= \frac{D_g(m, b)}{\|\nabla f(x)\|^2}.\end{aligned} \tag{57}$$

Under the special relative bound $\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)-\nabla f(x)\|^2 \le \rho\|\nabla f(x)\|^2$, the inclusion certificates are

$$\delta^{\mathcal{S}}(m) \le \rho\left(M\sum_i p_i^2 - 1\right), \qquad \beta^{\mathcal{S}}(m,b) \le \rho M \mathbb{E}_I \sum_i (q_i - p_i)^2. \tag{58}$$

These are upper bounds on the actual errors, not identities defining them. For uniform nested sampling, $\bar{g}_x([M],m) = \nabla f(x)$ and $\bar{g}_x(I,m) = b^{-1}\sum_{i\in I}\nabla f_i(x)$; Lemma F.1 gives the sharper identity

$$D_g(m,b) = q_b \frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)-\nabla f(x)\|^2, \qquad \beta^{\mathrm{unif}}(m,b) \le \rho\frac{M-b}{b(M-1)}. \tag{59}$$

In contrast, the inclusion discrepancy is $\mathbb{E}\sum_i (q_i-p_i)^2 = (M-b)/(bM)$, so its general certificate is larger by a factor $M-1$. This distinction reconciles the geometric inclusion diagnostic with the tighter variance calculation.

Second-order information. For $\pi_{ij} = \mathbb{P}(i,j\in C)$ with $\pi_{ii} = \pi_i$, the update variance is

$$V_F(m) = \frac{1}{m^2}\sum_{i,j}(\pi_{ij} - \pi_i\pi_j)\langle \nabla f_i(x), \nabla f_j(x)\rangle. \tag{60}$$

The analogous formula holds conditionally on I. This follows by expanding $\mathbb{E}\|g_x(C)\|^2 - \|\mathbb{E}g_x(C)\|^2$. First-order inclusion alone does not identify it.

Example F.7 (Equal inclusion does not determine the response)**.** Take $f_i(z) = z^2/2 + a_i z$ with $a = (1,1,-1,-1)$ and evaluate at $z = 0$ with $m = 2$. Selecting $\{1,2\}$ or $\{3,4\}$ with equal probability has the same $\pi_i = 1/2$ as selecting $\{1,3\}$ or $\{2,4\}$. Their mean updates are both zero, but the expected global loss responses are $\gamma^2/2$ and zero, respectively. Consequently, the existing inclusion diagnostics assess mean-update distortion; they do not by themselves certify response signs or convergence.

F.6 Practical updates and response evaluation

Biased selection and local training. The preceding results use exact reference gradients and losses. We now separate errors in the production update from errors in the search response. Let $\mathcal{F}_t$ contain the history after all controller decisions and before production randomness, and let the actual update be $x^{t+1} = x^t - \gamma U_t$. If

$$\langle \nabla f(x^t), \mathbb{E}[U_t \mid \mathcal{F}_t]\rangle \ge a_t\|\nabla f(x^t)\|^2 - d_t, \qquad \mathbb{E}[\|U_t\|^2 \mid \mathcal{F}_t] \le b_t\|\nabla f(x^t)\|^2 + s_t,$$

where the nonnegative coefficients are measurable at that history, then

$$\mathbb{E}[f(x^{t+1}) - f(x^t) \mid \mathcal{F}_t] \le -\gamma(a_t - L\gamma b_t/2)\|\nabla f(x^t)\|^2 + \gamma d_t + \frac{L\gamma^2}{2}s_t. \tag{61}$$

This is the smoothness inequality with conditional first and second moments. If $a_t - L\gamma b_t/2 \ge c > 0$, telescoping gives

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f(x^t)\|^2 \le \frac{f(x^0) - f_{\inf}}{\gamma c T} + \frac{1}{cT}\sum_{t<T}\mathbb{E}[d_t + (L\gamma/2)s_t].$$

It explicitly separates directional alignment from stochastic variation. The inclusion formulas bound mean bias for uniform aggregation; they do not establish favorable alignment for every utility-driven selector.

One concrete local-training remainder can be obtained without treating local epochs as unbiased gradients. Let client i take E deterministic gradient steps $z_{e+1} = z_e - \eta\nabla f_i(z_e)$ from $z_0 = x$, and set $u_i^{\mathrm{loc}} = (x - z_E)/(\eta E)$. If its gradient is L_i-Lipschitz and $\|\nabla f_i(z_e)\| \le G_i$ along those steps, then

$$\|u_i^{\mathrm{loc}} - \nabla f_i(x)\| \le \frac{1}{E}\sum_{e<E} L_i\|z_e - x\| \le \frac{L_i\eta G_i(E-1)}{2}. \tag{62}$$

The second inequality follows from $\|z_e - x\| \le \eta e G_i$. For uniform averaging of these normalized updates, the same averaged residual bounds the difference from the reference mean update. For ordinary averaging of the final local models, the corresponding server scale is $\gamma = \eta E$. Different local optimizers, momentum, compression and reweighting require their own residual or moment bounds. Equation (62) provides a concrete route for plain local gradient steps, not a claim that every experimental update has the reference convergence rate.

Shared auxiliary-update error in the search. The production-noise bound and the accuracy of the search are separate. All probabilities in this paragraph condition on the pre-adaptation history. Suppose the auxiliary states supply approximate directions $\widetilde{u}_i$ and, on an event of probability at least $1 - \delta_u$, satisfy $\max_{i \in I} \|\widetilde{u}_i - \nabla f_i(x)\| \le r_u$. Use the same fixed-depth uniform subset draws as in Appendix F.2. If objective evaluations additionally have uniform error at most ϵ_f at the base and candidate points, on an event of probability at least $1 - \delta_f$, then simultaneously for all candidate counts the additional response error is at most

$$e_{\mathrm{shared}} = \gamma K_x r_u + \frac{L\gamma^2}{2} r_u^2 + 2\epsilon_f. \tag{63}$$

Here $K_x = \|\nabla f(x)\| + L\gamma G$ is the exact-gradient constant in Appendix F.2. Indeed, averaging any subset preserves the error bound r_u. Expand f at its exact candidate point $x - \gamma g_x(C)$, whose gradient norm is at most K_x, and apply the two-sided smoothness remainder. The objective-evaluation errors contribute at most $2\epsilon_f$. Averaging over any number of probes preserves this pathwise bound, even though their auxiliary errors are shared.

Consequently, the selected-count argument of Appendix F.2 applies to this perturbed search with radii $\epsilon_m + e_{\mathrm{shared}}$ and failure probability at most $\delta_A + \delta_N + \delta_u + \delta_f$. No independence among these error events is needed for their union bound. The original first-negative comparison is unchanged; its uncertainty grows by an explicit term. For auxiliary mean-square errors at most σ_u^2 per client, conditional on the realized pool and preceding history, a concrete valid choice is $r_u = \sigma_u \sqrt{b/\delta_u}$ by Markov's inequality and a union bound, without requiring independent client errors. Tighter tail assumptions can improve this conservative choice. Local mini-batching can reduce σ_u^2, and (62) can bound a deterministic local-training component. Increasing N alone does not reduce e_{shared}. This search-error extension does not, by itself, give the fresh production updates the moments or alignment required for convergence.

Objective evaluation. The response target in the preceding results is the same uniform client average f that defines the optimization objective. Exact evaluation across all clients matches this target, independently of which clients constructed the candidate. If a surrogate satisfies $|\widetilde{f}(z) - f(z)| \le \epsilon_f$ at the base point and every candidate under consideration, its loss-change error is at most $2\epsilon_f$ by the triangle inequality. This is the same contribution already included in e_{shared}; it is counted once when both update and evaluation errors are present. With exact auxiliary gradients, it is the sole addition to the reference search radius. With sampled evaluators, validity requires simultaneous control over the candidate family or an appropriate independent-evaluation argument; unbiasedness for an individual fixed model is insufficient after adaptive candidate selection. A changing available-client population can also change the target objective and must be treated as a sampling or evaluation error.

Smoothing, coarse grids, and the GP add-on. The first-negative theorem applies unchanged to an ordered fixed candidate grid, with the number of potentially tested points in that fixed grid replacing $b - 1$ in the Monte Carlo union bound; its minimum is over that grid. The realized number of queries cannot replace this fixed cardinality. Count momentum can return a different, untested count. Without a bound on $|\Phi_\tau([M], m') - \Phi_\tau([M], m)|$, its response does not follow from the tested count. Likewise, an EMA of responses changes their signs unless its discrepancy from the current response is controlled. The GP add-on uses adaptive queries and predictions at unqueried counts; its posterior variance is not, by itself, the simultaneous error bound required here. If such an error bound is established for an extension, the deterministic selection argument still applies. We do not assume that the reported GP diagnostics establish it. The finite-depth and periodic convergence theorems concern Algorithm 2 as written, for the reference controller; the default stabilization is specified in Appendix C.1 and evaluated in Appendices E.3 and E.4.

Applying the periodic theorem numerically to a particular training run would require bounds on its response margins and cap coverage, relative client deviations, simultaneous search/evaluation errors, normalized response drift, and any practical-update or count-smoothing discrepancy. The existing experiments assess performance and robustness; they do not supply this collection of certificates. The alignment and local-drift inequalities above identify the additional quantities, without asserting that they are small for every experimental optimizer.

Stochastic production gradients and finite local samples.

Proposition F.15 (One-step stochastic updates)**.** *Suppose that, conditional on the production history and selected cohort, fresh client gradients are unbiased and their noises are independent with variance at most σ_s^2. Assume all conditions of Corollary F.2 hold. Then*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f(x^t)\|^2 \le \frac{f(x^0)-f_{\inf}}{\gamma c T} + \frac{L\gamma}{2cT}\sum_{t<T}\mathbb{E}\left[\sigma_c^2 q_{m_{t+1}} + \frac{\sigma_s^2}{m_{t+1}}\right]. \tag{64}$$

Proof. Conditional on the selected clients, the average local noise has mean zero and variance at most σ_s^2/m. Its cross term with the client-sampling deviation is zero. Therefore the aggregate second moment is bounded by $\|\nabla f(x)\|^2 + q_m \frac{1}{M}\sum_{i=1}^M \|\nabla f_i(x) - \nabla f(x)\|^2 + \sigma_s^2/m$. Smoothness and telescoping give the claim. □

Corollary F.16 (Finite local mini-batches)**.** *If each local objective is the average of $B \ge 2$ sample losses and its sample-gradient population variance is at most σ_s^2, then under the preceding proposition's conditions a fresh uniform mini-batch of size $s \in [B]$, sampled without replacement and independently across clients, replaces σ_s^2 in* (64) *by $\sigma_s^2(B-s)/(s(B-1))$.*

Proof. Apply Lemma F.1 within each client, then use independence across clients as in the preceding proposition. □

The intra-client term vanishes at $s = B$; the inter-client term need not. With horizon-independent bounds on the initial gap, smoothness, and variances, a uniformly positive progress coefficient c, and γ of order $T^{-1/2}$, these conditional bounds give the usual $O(T^{-1/2})$ average stationarity rate. They concern one-step production gradients. Reusing stochastic auxiliary updates during many candidate evaluations creates an additional pool/update error; increasing the number of subset draws does not resample those updates. Nor does this proposition turn multiple local training steps into an unbiased gradient. The local-training and shared auxiliary-update errors are bounded in (62) and (63), respectively.

G EMPIRICAL INCLUSION DIAGNOSTICS

We now measure the inclusion structure appearing in the general-selector bounds of Appendix F.5. The diagnostics describe how auxiliary restriction changes the mean update. The normalized coefficients $\delta^{\mathcal{S}}$ and $\beta^{\mathcal{S}}$ below refer to the actual mean errors in (57); their relative bounds apply at nonstationary points in the special case

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x)-\nabla f(x)\|^2 \le \rho\|\nabla f(x)\|^2.$$

Equation (56) supplies the corresponding absolute bounds without that restriction.

For a target budget m, define the normalized full-support inclusion weights $p^{\mathcal{S}}_{i,\tau}(m) := \frac{\pi^{\mathcal{S}}_{i,\tau}(m)}{m}$, which form a distribution over clients by the $\sum_{i=1}^{M}\pi^{\mathcal{S}}_{i,\tau}(m)=m$. We define its Effective Support Size (ESS) as

$$\mathrm{ESS}^{\mathcal{S}}_{\tau}(m) := \left(\sum_{i=1}^{M}\left(p^{\mathcal{S}}_{i,\tau}(m)\right)^2\right)^{-1}.$$

A small $\mathrm{ESS}^{\mathcal{S}}_{\tau}(m)$ indicates that CSS $\mathcal{S}$ concentrates its inclusion mass on a smaller subset of clients, while $\mathrm{ESS}^{\mathcal{S}}_{\tau}(m)=M$ corresponds to a UNIFORM marginal inclusion profile. With this notation, Proposition F.14 gives

$$\delta^{\mathcal{S}}_{\tau}(m) \leqslant \rho\left(\frac{M}{\mathrm{ESS}^{\mathcal{S}}_{\tau}(m)}-1\right).$$

Thus, we use $M/\mathrm{ESS}^{\mathcal{S}}_{\tau}(m)-1$ as an empirical proxy for the inclusion-dependent term in the bound for $\delta^{\mathcal{S}}_{\tau}(m)$, due to the unknown heterogeneity factor ρ.

To diagnose the truncation coefficient $\beta^{\mathcal{S}}_{\tau}(m,b_{\tau+1/2})$, we define normalized restricted inclusion weights $q^{\mathcal{S}}_{i,\tau}(I,m) := \psi^{\mathcal{S}}_{i,\tau}(I,m)/m$ for a fixed auxiliary pool $I\subseteq[M]$, and quantify their deviation from the full-support weights by

$$\mathrm{D}^{\mathcal{S}}_{\tau}(b_{\tau+1/2},m) := \mathbb{E}_{I_{\tau+1/2}\sim\mathcal{S}_\tau([M],b_{\tau+1/2})}\sum_{i=1}^{M}\left(q^{\mathcal{S}}_{i,\tau}(I_{\tau+1/2},m)-p^{\mathcal{S}}_{i,\tau}(m)\right)^2.$$

This distortion captures the marginal-coverage loss caused by auxiliary-pool truncation. Then Proposition F.14 implies

$$\beta^{\mathcal{S}}_{\tau}(m,b_{\tau+1/2}) \leqslant \rho M\,\mathrm{D}^{\mathcal{S}}_{\tau}(b_{\tau+1/2},m).$$

Hence, we evaluate $\mathrm{D}^{\mathcal{S}}_{\tau}(b_{\tau+1/2},m)$ as the observable inclusion factor in an upper bound on mean-update distortion; the bound also contains the unknown heterogeneity factor ρ. As $\mathrm{D}^{\mathcal{S}}_{\tau}(b_{\tau+1/2},m)$ is a squared discrepancy between the inclusion distributions, $\sqrt{\mathrm{D}^{\mathcal{S}}_{\tau}(b_{\tau+1/2},m)/M}$ can be read as the RMS per-client probability shift. For the UNIFORM CSS, the actual mean-update coefficient admits the sharper finite-population bound

$$\beta^{\mathrm{Unif}}_{\tau}(m,b_{\tau+1/2}) \leqslant \rho\,\frac{M-b_{\tau+1/2}}{b_{\tau+1/2}(M-1)}.$$

Here the uniform bound follows from the exact sample-mean variance in (59); it is tighter than the general inclusion certificate $\rho M\mathrm{D}$ by a factor $M-1$. At fixed auxiliary size b and large M, this mean-gradient error scales as $\frac{1}{bM}\sum_{i=1}^{M}\|\nabla f_i(x)-\nabla f(x)\|^2$. Thus a proportional growth of b with the network size is not necessary to keep this particular absolute error bounded.

For a deterministic selector using the same fixed scores and tie-breaking on every pool, the auxiliary top-b set contains the full-pool top-m set whenever $b\ge m$. The subset laws then coincide exactly, and both D and the response gap Γ are zero. This property applies to such fixed-ranking implementations; it is not implied by a method name when scores or rankings are recomputed on the restricted pool. The near-zero distortions below describe the evaluated selector configurations.

To complement the distortion diagnostic, we also measure how much of the full-support inclusion distribution is captured by the auxiliary pool before restricted selection is performed. For $I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})$, define

$$\mathrm{Mass}_\tau^{\mathcal{S}}(b_{\tau+1/2}, m) := \mathbb{E}_{I_{\tau+1/2} \sim \mathcal{S}_\tau([M], b_{\tau+1/2})} \sum_{i \in I_{\tau+1/2}} p_{i,\tau}^{\mathcal{S}}(m).$$

This quantity captures the expected fraction of the target inclusion mass that is available inside the auxiliary pool. Thus, Mass measures coverage of the full-support CSS, while D measures the distortion introduced by running the selector on the restricted pool.

The inclusion diagnostics are estimated with a post-hoc Monte Carlo routine summarized in Algorithm 9 under the same CIFAR-10 setup as in Section 5.2, ensuring consistency with the client-selection experiments. At ISP adaptation rounds, $\tau \equiv 0 \pmod{\Delta}$, we run Algorithm 9 over the diagnostic grid of feasible budget pairs $m \leq b$. For each pair, the routine estimates the inclusion probabilities $\hat{\pi}_{i,\tau}^{\mathcal{S}}(m)$ and $\hat{\psi}_{i,\tau}^{\mathcal{S}}(I, m)$, and then computes the plug-in diagnostics $M/\widehat{\mathrm{ESS}}_\tau^{\mathcal{S}}(m) - 1$, $\widehat{\mathrm{Mass}}_\tau^{\mathcal{S}}(b, m)$, and $\sqrt{\widehat{\mathrm{D}}_\tau^{\mathcal{S}}(b, m)/M}$. These quantities are used only to analyze the auxiliary-budget approximation gap and do not affect the ISP training procedure.

The first-round diagnostics in Figures 18 and 19 are computed for the first intermediate ISP communication round in the CIFAR-10 client-selection experiment. We focus on statistical-utility CSS and omit OORT, whose selection rule also depends on runtime and availability constraints. Figure 18(a) reports the concentration proxy $M/\mathrm{ESS}_\tau^{\mathcal{S}}(m) - 1$. The exact UNIFORM concentration proxy is zero at every m, since $p_i = 1/M$. The positive finite-sample estimates in the plot approach zero as m grows. More generally, these plug-in squared-discrepancy estimates contain Monte Carlo estimation error and are descriptive diagnostics, not confidence bounds for the response gap. Structured selectors exhibit larger concentration at small budgets, reflecting their preference for informative, representative, or diverse clients. For most selectors the proxy decreases with m; FEDCBS shows a small non-monotone region at low budgets, but also converges to a low-concentration regime for larger m. For score-based selectors such as POWER-OF-CHOICE and FEDCOR, the proxy decreases as the selected set expands, indicating that their inclusion mass becomes less concentrated when more clients are requested.

Figure 19(a) shows the corresponding restricted-pool distortion $\sqrt{\mathrm{D}_\tau^{\mathcal{S}}(b, m)/M}$ over feasible pairs $m \leq b$. The distortion decreases as the auxiliary budget b grows, which is consistent with the interpretation of $b_{\tau+1/2}$ as the fidelity parameter of the auxiliary-pool approximation. The UNIFORM selector provides the least structured reference and therefore exhibits the largest distortion for small auxiliary pools. In contrast, FEDCBS and DELTA reduce the distortion in large parts of the grid by concentrating the inclusion mass on more structured client subsets. For POWER-OF-CHOICE and FEDCOR, the distortion is essentially zero in this first-round diagnostic, matching the nested deterministic case discussed above: once the auxiliary pool contains the clients that would be selected by the full-support rule, restricted and full-support selection coincide.

Figure 18(b) gives a direct one-dimensional view of auxiliary-pool mass and restricted-pool distortion at the fixed target budget $m = 21$. It confirms the same trade-off: structured CSS capture the target inclusion mass faster than UNIFORM, and this improved coverage translates into smaller restricted-pool distortion. The effect is strongest for deterministic score-based selectors, where the auxiliary pool recovers the relevant target mass once the top-ranked set is included. Figure 19(b) provides the full (b, m) view of the captured inclusion mass. For UNIFORM, the captured mass is mainly controlled by b because all clients have equal marginal weight. For structured selectors, the mass depends on both b and m, and large regions of the feasible grid reach high coverage with substantially smaller auxiliary pools. These diagnostics support the role of the expansion ratio p in preserving inclusion structure: larger auxiliary pools reduce the measured inclusion discrepancy, while structured CSS can achieve smaller discrepancies before full coverage. This controls one factor in the mean-distortion bound, rather than measuring the complete response gap.

To complement the one-round inclusion diagnostics, we also quantify how empirical inclusion profiles change across ISP adaptation rounds. Let

$$\mathcal{T}_\Delta := \{\tau : \tau \equiv 0 \pmod{\Delta}\}$$

be the set of rounds at which ISP performs intermediate communication, and let $\tau_0 := \min \mathcal{T}_\Delta$ denote the first such round. For the full-support selector, define

$$\hat{\mathbf{p}}_\tau(m) := \big(\hat{p}^{\mathcal{S}}_{1,\tau}(m), \dots, \hat{p}^{\mathcal{S}}_{M,\tau}(m)\big),$$

and measure its drift from the first adaptation round by

$$d_p(\tau, m) := \frac{1}{2} \left\| \hat{\mathbf{p}}_\tau(m) - \hat{\mathbf{p}}_{\tau_0}(m) \right\|_1 .$$

Thus, $d_p(\tau, m)$ is the total-variation shift of the empirical full-support inclusion distribution relative to the first ISP update. We visualize this quantity in Figure 20(a).

For the restricted selector, the empirical inclusion profile additionally depends on the sampled auxiliary pool. To remove this dependence, for each pair (b, m) we average over sampled auxiliary pools $I_s \sim \mathcal{S}_\tau([M], b)$ and define

$$\bar{\mathbf{q}}_\tau(b, m) := \frac{1}{R_{\psi_1}} \sum_{s=1}^{R_{\psi_1}} \hat{\mathbf{q}}_\tau(I_s, m), \qquad \hat{\mathbf{q}}_\tau(I_s, m) := \big(\hat{q}^{\mathcal{S}}_{1,\tau}(I_s, m), \dots, \hat{q}^{\mathcal{S}}_{M,\tau}(I_s, m)\big).$$

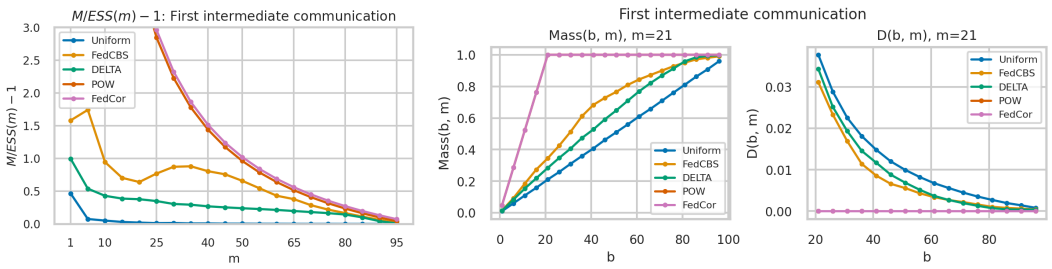
We then define the restricted-profile drift

$$d_q(\tau, b, m) := \frac{1}{2} \left\| \bar{\mathbf{q}}_\tau(b, m) - \bar{\mathbf{q}}_{\tau_0}(b, m) \right\|_1 .$$

Figure 20(b) reports the corresponding restricted-profile drift averaged over adaptation rounds.

The drift diagnostics show that the inclusion structure is not static throughout training, especially for score-based selectors. For UNIFORM, the full-support drift d_p is nearly zero by construction, while the restricted drift d_q mainly reflects the randomness of auxiliary-pool truncation and decreases with larger b and m. For FEDCBS and DELTA, both full-support and restricted profiles change moderately and become more stable at larger budgets. For POWER-OF-CHOICE and FEDCOR, the drift is larger at small m because their rankings depend on evolving loss or score estimates; nevertheless, the drift decreases as the selected set or auxiliary budget grows. This behavior is expected for adaptive score-based CSS and does not invalidate the auxiliary-pool approximation. Rather, it shows where the approximation is most demanding: small target budgets combined with rapidly changing client rankings.

Overall, the diagnostics characterize the mean-distortion component in Proposition F.13. In these experiments, larger auxiliary budgets improve coverage and reduce the observed inclusion distortion; structured selectors often preserve their target inclusion profiles within smaller pools. The temporal plots identify small-budget regimes with more rapidly changing sampling profiles. These observations complement the performance and hyperparameter ablations. First-order inclusion estimates do not measure the update variances or candidate-response errors required for a numerical convergence certificate, and marginal-profile drift is not the response drift in Theorem F.8.



(a) Inclusion concentration **(b) Mass–distortion profile**

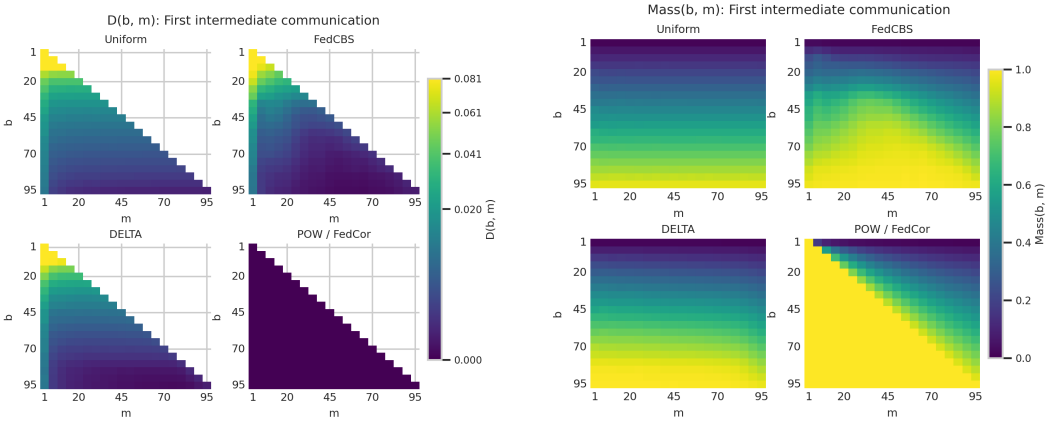
Figure 18: One-dimensional inclusion diagnostics for statistical CSS at the first intermediate ISP round: (a) the inclusion-concentration proxy $M/\mathrm{ESS}^{\mathcal{S}}_\tau(m) - 1$ across target budgets m, and (b) auxiliary-pool mass and restricted-pool distortion at the fixed target budget $m = 21$.

Algorithm 9 Inclusion Diagnostics Estimation

Require: Round τ, target budget m, auxiliary budget $b \geqslant m$, CSS $\mathcal{S}_\tau$, depths $R_\pi, R_{\psi_1}, R_{\psi_2}$.

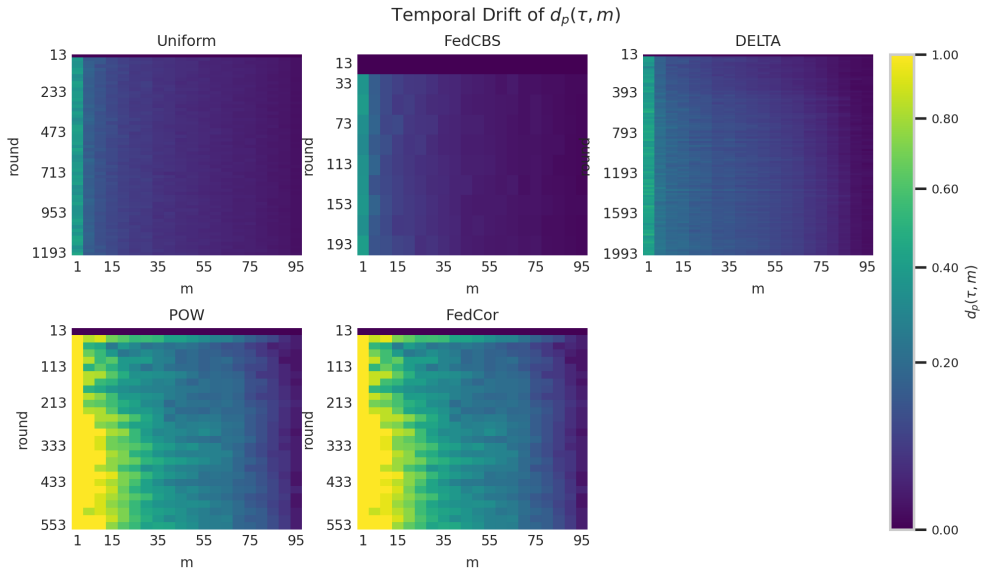
for $r = 1$ to R_π **do**
  $C_r \sim \mathcal{S}_\tau([M], m)$
end for
$\widehat{\pi}^{\mathcal{S}}_{i,\tau}(m) \leftarrow \frac{1}{R_\pi} \sum_{r=1}^{R_\pi} \mathbf{1}\{i \in C_r\}$ ▷ $i \in [M]$
$\widehat{p}^{\mathcal{S}}_{i,\tau}(m) \leftarrow \widehat{\pi}^{\mathcal{S}}_{i,\tau}(m)/m$ ▷ $i \in [M]$
for $s = 1$ to R_{ψ_1} **do**
  $I_s \sim \mathcal{S}_\tau([M], b)$
  for $t = 1$ to R_{ψ_2} **do**
    $C_{s,t} \sim \mathcal{S}_\tau(I_s, m)$
  end for
  $\widehat{\psi}^{\mathcal{S}}_{i,\tau}(I_s, m) \leftarrow \frac{1}{R_{\psi_2}} \sum_{t=1}^{R_{\psi_2}} \mathbf{1}\{i \in C_{s,t}\}$ ▷ $i \in [M]$
  $\widehat{q}^{\mathcal{S}}_{i,\tau}(I_s, m) \leftarrow \widehat{\psi}^{\mathcal{S}}_{i,\tau}(I_s, m)/m$ ▷ $i \in [M]$
end for

$\widehat{\mathrm{ESS}}^{\mathcal{S}}_\tau(m) \leftarrow \left(\sum_{i=1}^{M} \left(\widehat{p}^{\mathcal{S}}_{i,\tau}(m)\right)^2\right)^{-1}$
$\widehat{\mathrm{Mass}}^{\mathcal{S}}_\tau(b, m) \leftarrow \frac{1}{R_{\psi_1}} \sum_{s=1}^{R_{\psi_1}} \sum_{i \in I_s} \widehat{p}^{\mathcal{S}}_{i,\tau}(m)$
$\widehat{\mathrm{D}}^{\mathcal{S}}_\tau(b, m) \leftarrow \frac{1}{R_{\psi_1}} \sum_{s=1}^{R_{\psi_1}} \sum_{i=1}^{M} \left(\widehat{q}^{\mathcal{S}}_{i,\tau}(I_s, m) - \widehat{p}^{\mathcal{S}}_{i,\tau}(m)\right)^2$
return $M/\widehat{\mathrm{ESS}}^{\mathcal{S}}_\tau(m) - 1, \quad \widehat{\mathrm{Mass}}^{\mathcal{S}}_\tau(b, m), \quad \sqrt{\widehat{\mathrm{D}}^{\mathcal{S}}_\tau(b, m)/M}$

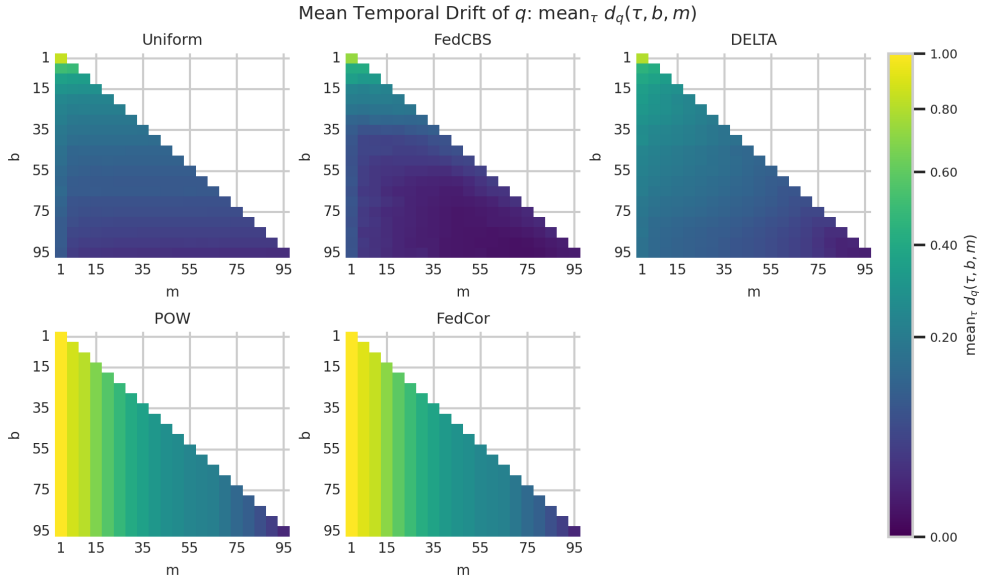


(a) Restricted-pool distortion **(b) Auxiliary-pool mass**

Figure 19: Full-grid inclusion diagnostics over auxiliary budget b and target budget m at the first intermediate ISP round: (a) restricted-pool distortion $\sqrt{\mathrm{D}^{\mathcal{S}}_\tau(b, m)/M}$ and (b) captured auxiliary-pool mass $\mathrm{Mass}^{\mathcal{S}}_\tau(b, m)$.



(a) Full-support inclusion drift



(b) Restricted-profile drift

Figure 20: Temporal stability of the inclusion profiles across ISP adaptation rounds: (a) drift $d_p(\tau, m)$ of the full-support inclusion profile relative to the first adaptation round, and (b) restricted-profile drift $d_q(\tau, b, m)$ averaged over subsequent adaptation rounds.