# PROBFEDZO: PROBABILISTIC STEP SIZE ADAPTATION FOR FEDERATED ZEROTH-ORDER OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Zeroth-order optimization allows clients with forward-only access to participate in federated learning while uploading only a few scalars. However, it does not determine how far to move along the estimated direction. Fixed schedules cannot adapt to the current round, whereas line search tests each candidate through additional client communication. We propose two algorithms that select the step size on the server without further client contact. Both collect one shared batch of probe evaluations and fit a polynomial surrogate of the global objective. Under full participation, SurrFedZO applies an Armijo test to the surrogate and takes the largest passing step. Under partial participation, ProbFedZO corrects the test for the uncertainty caused by client sampling and certifies a step only when its predicted decrease exceeds that uncertainty. We bound the error of the surrogate by the probe radius and show that accepted steps decrease the global objective, deterministically for SurrFedZO and with a prescribed probability for ProbFedZO. On federated logistic regression and LoRA fine-tuning, both algorithms achieve competitive or improved risk at matched communication budgets, including settings where fixed and diminishing schedules stall.

## 1 INTRODUCTION

Federated learning (Konečný et al., 2016; McMahan et al., 2017; Yang et al., 2019) has emerged as a fundamental paradigm for distributed machine learning, allowing clients to train a shared model while keeping raw data local. In each round, the server broadcasts the current global model, clients compute certain quantities such as local gradients, and the server aggregates these quantities received from the clients to form the next global model. This cycle requires local computation and communication of high-dimensional models or updates. Zeroth-order (ZO) optimization (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017) instead constructs updates based on output values of the current global model, allowing clients who are only able to support forward inference to participate (Malladi et al., 2023). If the server and clients share the probe directions, it suffices for clients to upload only a few scalar changes in their empirical risks alongside the probe directions rather than full parameter updates (Li et al., 2024; Fang et al., 2022; Mhanna & Assaad, 2024).

Whether the update is built from gradients or from function values, the server moves the global objective along an estimated direction by a step size at each round, and the choice of this step size matters as much as the direction. Selecting a step size that yields stable and fast convergence to a minimum is important, since a fixed step size may not remain appropriate throughout training. A small constant step size makes slow progress, whereas a large step size can overshoot or diverge. Note that an adequate step size changes over training, because the local geometry and the quality of the estimated direction vary from round to round (see Appendix E.3).

In federated ZO optimization, it is difficult to accurately estimate the local geometry of the loss landscape using only a few function evaluations, which makes selecting an adequate step size that provides sufficient decrease in the global objective at each round challenging (Yu et al., 2026). This holds even when every client participates, since the server knows the exact change in the global objective only at the few probed points. Moreover, the problem becomes worse when only a random subset of clients participates in a round, because the changes in the local objectives reported by



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.

the participants may differ from those of the full population under client heterogeneity (Fang et al., 2022), so the selected step size may not decrease the global objective. Conventional line search addresses this lack of adaptivity by evaluating candidate steps on the current objective, but each trial requires another exchange with participating clients, since every function value must be computed by the clients. Sequential backtracking can thus turn one update into several communication rounds, which is expensive and sometimes prohibitive (Safaryan et al., 2022) in federated optimization.

The aim of this paper is twofold. First, under full participation, we seek a method that selects both the update direction and the step size without any additional communication between the server and the clients. To this end, we reuse the responses clients already send to estimate the update direction. The server fits a local surrogate of the global objective to the exact change at the probe points and tests candidate step sizes on the surrogate instead of the objective, so no candidate costs a client query. We call this approach SurrFedZO, federated zeroth-order optimization with surrogate line search. The resulting method, SurrFedZO, applies an Armijo condition to the surrogate and takes the largest passing step.

The second aim is that under partial participation, we seek a method that selects the step size with a probabilistic guarantee, since the surrogate fitted from a random subset of clients differs from the one all clients would have produced. ProbFedZO, propagates this uncertainty through the surrogate and certifies a step size only when its predicted decrease exceeds the uncertainty. If a candidate step size is not certified, a diminishing step size is retried.

Our contributions are as follows.

- **Iterative descent method based on a surrogate model in federated learning:** We propose SurrFedZO, which constructs a local surrogate from one shared batch of probe evaluations and selects the direction and step size on the server without additional communcation.
- **Uncertainty-aware step-size selection under partial participation:** We propose ProbFedZO, which quantifies the uncertainty in the surrogate due to random client participation and tests candidate step sizes through an uncertainty-corrected Armijo condition.
- **Theoretical guarantees:** Under explicit additional assumptions, we show that each step certified by ProbFedZO decreases the global objective with probability at least $\rho$, and we bound the expected total decrease over $T$ rounds in terms of the predicted progress, model error, and sampling error of the surrogate.
- **Empirical evaluation:** Experiments on nonconvex objective functions, logistic regression under full and partial participation, and fine-tuning language models using LoRA (Hu et al., 2021) show competitive or improved performance our proposed algorithms over baselines under communication budget constraints.

## 2 Related Works

**Zeroth-order optimization in federated learning.** Most federated optimization methods are first-order and require clients to compute gradients (Li et al., 2020a; Wang et al., 2020; Zhao et al., 2018), which is expensive for resource-constrained clients (Fang et al., 2022; Malladi et al., 2023) and communication-heavy (Li et al., 2024). ZO optimization avoids gradient computation: clients evaluate their local objectives at perturbed versions of the global model, and these evaluations are used to estimate an update direction. When the server and the clients share the probe directions, clients upload only scalar changes in their local objectives rather than parameter updates (Li et al., 2024). Since the number of evaluations needed to estimate a direction grows with the parameter dimension, subspace methods restrict the search to a low-dimensional subspace, typically built from random projections (Luo et al., 2016; Gower et al., 2019; Kozak et al., 2021; Kim et al., 2025). These methods still require a step size along the estimated direction. We extend this line by selecting the step size on the server from a surrogate fitted to the same evaluations, so no additional communication is needed, and by correcting the Armijo test for client sampling under partial participation.

**Step size selection in federated optimization.** Federated optimization commonly uses a fixed step size (McMahan et al., 2017; Fang et al., 2022), or a scheduled step size (Chen et al., 2019; Li et al., 2020b). Adaptive step-size methods have also been explored in first-order federated optimization (Safaryan et al., 2022), but their reliance on client-computed gradients prevents their direct

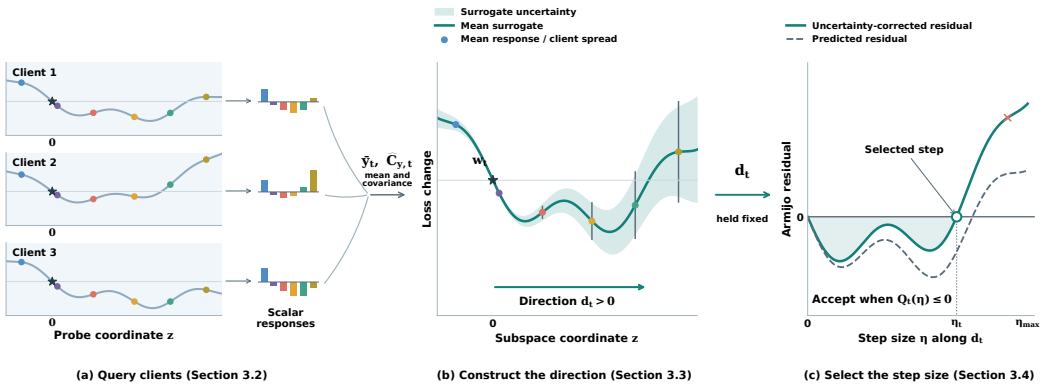



Figure 1: One round of SurrFedZO. (a) Participating clients evaluate the same probe points and return the scalar changes in their local objectives (Section 3.2). (b) The server fits a polynomial surrogate to the averaged responses and forms the update direction $\mathbf{d}_t$ (Section 3.3); under partial participation it also forms an uncertainty envelope from the variation across clients. (c) Candidate step sizes along $\mathbf{d}_t$ are tested against an Armijo condition on the surrogate. No further client queries are made after (a). Geometry is illustrative.

application to the zeroth-order setting. Line search can adapt the step size to the current update direction, but each candidate step size requires clients to evaluate the objective again. Testing several candidate step sizes therefore requires additional computation and communication. This motivates selecting an adequate step size using the information already received from participating clients to estimate the update direction. We follow this approach, retaining line-search adaptivity without requiring additional client queries.

## 3 SurrFedZO: Surrogate Line Search for Method for Federated ZO Optimization

In this section we propose SurrFedZO. Every client evaluates a shared set of probe points, the server fits a surrogate of the global objective based on the exact changes of the global objective at the probe points received from clients, and the step size is selected by Armijo test on the surrogate.

Figure 1 summarizes one round of SuPer-FedZO in three stages.

1. **Collect client evaluations (Section 3.2).** The server broadcasts the current model, a low-dimensional subspace, and a shared set of probe points. Each participating client evaluates its local objective at the probe points and returns the scalar changes.
2. **Construct the direction (Section 3.3).** The server averages the responses, fits a polynomial surrogate of the change in the global objective, and forms an update direction from the surrogate's gradient and Hessian.
3. **Select the step size (Sections 3.4).** The server tests candidate step sizes against an Armijo condition on the surrogate and takes the largest passing step.

Only the first stage communicates with clients. The other two stages are executed on the server, so testing more candidate step sizes costs no client evaluations and so no communication.

### 3.1 Setup and Notation

A population of $N$ clients tries to minimize the average of their local objectives,

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{w}), \qquad F_i(\mathbf{w}) = \frac{1}{n_i}\sum_{\xi\in\mathcal{D}_i} \ell(\mathbf{w};\xi), \tag{1}$$

where $\mathcal{D}_i$ is the local dataset of client $i$ with $n_i$ examples and $\ell$ is the loss. Each client evaluates its local objective $F_i$ exactly, and the global objective $F$ is the average over these $N$ clients.

**Algorithm 1** SurrFedZO at round $t$ (all $N$ clients participate).

**Require:** $\mathbf{w}_t$, $\mathbf{V}_t$, $K$, $\Delta_t$, $\epsilon_H$, $\eta_0$, $c$, grid $\mathcal{H}_t$
1: Draw probes $\mathbf{z}_t^1, \dots, \mathbf{z}_t^K$ with $\|\mathbf{z}_t^k\| \le \Delta_t$, broadcast $(\mathbf{V}_t, \{\mathbf{z}_t^k\})$, and receive $\mathbf{y}_{i,t}$ from every client by equation 2
2: $\bar{\mathbf{y}}_t \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{y}_{i,t}$ ▷ server side from here, no further queries
3: Fit $\widehat{m}_t$ by equation 4 with $\lambda = 0$, $\widehat{\mathbf{g}}_t \leftarrow \nabla \widehat{m}_t(\mathbf{0})$, $\widehat{\mathbf{H}}_t \leftarrow \nabla^2 \widehat{m}_t(\mathbf{0})$
4: $\mathbf{d}_t \leftarrow -|\widehat{\mathbf{H}}_t|_{\epsilon_H}^{-1}\widehat{\mathbf{g}}_t$, $\widehat{s}_t \leftarrow -\widehat{\mathbf{g}}_t^\top \mathbf{d}_t$
5: $\mathcal{H}_t^{\mathrm{acc}} \leftarrow \{\eta \in \mathcal{H}_t : \widehat{R}_t(\eta) \le 0\}$ by equation 8
6: If $\mathcal{H}_t^{\mathrm{acc}} \neq \emptyset$, $\eta_t \leftarrow \max \mathcal{H}_t^{\mathrm{acc}}$. Otherwise $\mathbf{d}_t \leftarrow -\widehat{\mathbf{g}}_t$ and $\eta_t \leftarrow \min\{\eta_0/\sqrt{t+1},\ \Delta_t/\|\widehat{\mathbf{g}}_t\|\}$ ($\eta_t \leftarrow 0$ if $\widehat{\mathbf{g}}_t = \mathbf{0}$)
7: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \mathbf{V}_t(\eta_t \mathbf{d}_t)$

The update is $\mathbf{z} = \eta_t \mathbf{d}_t$ with direction $\mathbf{d}_t \in \mathbb{R}^r$ and step size $\eta_t > 0$, restricted to the trust region $\|\eta_t \mathbf{d}_t\| \le \Delta_t$ of radius $\Delta_t$, within which the surrogate is used. The radius $\Delta_t$ is a hyperparameter set by the server. It bounds both the probes and the step, and it controls the accuracy of the surrogate (Lemma 1). We use a fixed $\Delta_t = \Delta$ in the experiments and study its effect in Appendix E.10.

**Participation.** All $N$ clients participate in every round.

**Conventions.** Vectors and matrices are bold. A bar denotes an average over the $N$ clients and a hat marks a quantity the server computes from the client responses. We fix the Armijo parameter $c \in (0, 1)$. Sections 3.2–3.4 describe a single round and omit the index $t$, except for the selected step $\eta_t$. A candidate step is $\eta$.

### 3.2 Shared probe queries

The server draws $K$ probe points $\mathbf{z}^1, \dots, \mathbf{z}^K$ in the trust region $\{\|\mathbf{z}\| \le \Delta\}$ and broadcasts them, together with $\mathbf{V}$, to every participating client. Client $i \in S$ evaluates its local objective once at the current model and once at each probe, and returns the $K$ scalar changes

$$y_i^k = F_i(\mathbf{w} + \mathbf{V}\mathbf{z}^k) - F_i(\mathbf{w}), \qquad k = 1, \dots, K, \tag{2}$$

as the vector $\mathbf{y}_i = [y_i^1, \dots, y_i^K]^\top \in \mathbb{R}^K$. This costs $K+1$ local evaluations and $K$ uploaded scalars per client. The server averages the responses,

$$\bar{\mathbf{y}} = \frac{1}{m}\sum_{i \in S} \mathbf{y}_i. \tag{3}$$

Under full participation, $\bar{y}^k = \frac{1}{N}\sum_{i=1}^N y_i^k = f(\mathbf{z}^k)$ is the exact change in the global objective at probe $k$.

### 3.3 Surrogate and update direction

**Polynomial surrogate.** The server fits a surrogate $\widehat{m}(\mathbf{z})$ of $f(\mathbf{z})$ to the averaged responses by kernel ridge regression with a polynomial kernel of degree $p$, with the anchor $\widehat{m}(\mathbf{0}) = 0$ since zero displacement leaves the objective unchanged. Because the kernel's feature space is the polynomials of degree at most $p$, the surrogate is itself a polynomial of degree $p$, and for fixed probes it is a linear function of $\bar{\mathbf{y}}$:

$$\widehat{m}(\mathbf{z}) = \mathbf{a}(\mathbf{z})^\top \bar{\mathbf{y}}, \tag{4}$$

where the coefficient map $\mathbf{a} : \mathbb{R}^r \to \mathbb{R}^K$ depends only on the probes, the kernel, and the ridge parameter $\lambda \ge 0$ (Appendix B, Equation equation 21).

**Update direction.** Differentiating the surrogate at the origin gives its gradient and Hessian,

$$\widehat{\mathbf{g}} = \nabla \widehat{m}(\mathbf{0}), \qquad \widehat{\mathbf{H}} = \nabla^2 \widehat{m}(\mathbf{0}), \tag{5}$$

which are the derivatives of the surrogate, not gradients computed by clients. The update direction rescales the gradient by the curvature,

$$\mathbf{d} = -|\widehat{\mathbf{H}}|_{\epsilon_H}^{-1}\widehat{\mathbf{g}}, \tag{6}$$

where $|\widehat{\mathbf{H}}|_{\epsilon_H}$ keeps the eigenvectors of $\widehat{\mathbf{H}}$ and replaces each eigenvalue by the maximum of its absolute value and a floor $\epsilon_H > 0$. This avoids reversing the gradient along directions of negative curvature and limits amplification along nearly flat ones. Since $|\widehat{\mathbf{H}}|_{\epsilon_H}$ is positive definite, the predicted slope along $\mathbf{d}$,

$$\widehat{s} = -\widehat{\mathbf{g}}^\top \mathbf{d} = \widehat{\mathbf{g}}^\top |\widehat{\mathbf{H}}|_{\epsilon_H}^{-1} \widehat{\mathbf{g}}, \tag{7}$$

is positive whenever $\widehat{\mathbf{g}} \neq \mathbf{0}$: the direction always descends on the surrogate.

### 3.4 Applying the Armijo test

For a candidate step $\eta$ along $\mathbf{d}$, the Armijo condition on the surrogate asks for a decrease of at least a fraction $c$ of the linear prediction. We measure it by the residual

$$\widehat{R}(\eta) = \widehat{m}(\eta \mathbf{d}) + c\,\eta\,\widehat{s}, \tag{8}$$

so that $\widehat{R}(\eta) \leq 0$ means the surrogate decreases sufficiently. In this case, we say that $\eta$ passes the Armijo test.

When every client participates, $\bar{\mathbf{y}}$ holds the exact changes in the global objective at the probes, so $\widehat{m}$, $\widehat{\mathbf{g}}$, $\widehat{s}$ and $\widehat{R}$ are not estimates: they are from the population surrogate. The only gap between the surrogate and global objective is the approximation error of the polynomial fit, which the radius $\Delta$ can control (Section 5). The server therefore applies the Armijo condition directly. Candidates are taken from a finite grid $\mathcal{H} \subset (0, \eta_{\max}]$, where $\eta_{\max} = \min\{\eta_0, \Delta/\|\mathbf{d}\|\}$ keeps every candidate locating inside the trust region, and the server takes the largest step among them that passes the Armijo test,

$$\eta_t = \max\{\eta \in \mathcal{H} : \widehat{R}(\eta) \leq 0\}. \tag{9}$$

Because $\widehat{R}(\eta) = -(1-c)\eta\,\widehat{s} + O(\eta^2)$ and $\widehat{s} > 0$, small enough steps always pass, so a grid that reaches small enough values always contains an accepted step; the diminishing fallback of Algorithm 1 is reached only if $\widehat{\mathbf{g}} = \mathbf{0}$ or no grid point is small enough.

After choosing $\eta$, the updated model becomes $\mathbf{w}_{t+1} = \mathbf{w} + \mathbf{V}(\eta_t \mathbf{d})$, and the next round of communication starts.

## 4 ProbFedZO: Uncertainty aware line search for Fed ZO Optimization

**Algorithm 2** ProbFedZO at round $t$ (only $m < N$ clients participate). Lines $\star$ are new.

**Require:** As in Algorithm 1, plus $\mathcal{S}_t$, $\lambda \geq 0$, $\kappa_t > 0$

1: Draw probes $\mathbf{z}_t^1, \ldots, \mathbf{z}_t^K$ with $\|\mathbf{z}_t^k\| \leq \Delta_t$, broadcast $(\mathbf{V}_t, \{\mathbf{z}_t^k\})$, and receive $\mathbf{y}_{i,t}$ from each $i \in \mathcal{S}_t$ by equation 2

2: $\bar{\mathbf{y}}_t \leftarrow \frac{1}{m}\sum_{i \in \mathcal{S}_t} \mathbf{y}_{i,t}$ ▷ server side from here, no further queries

3: $\star$ $\widehat{\mathbf{C}}_t \leftarrow$ covariance of $\bar{\mathbf{y}}_t$ by equation 10

4: Fit $\widehat{m}_t$ by equation 4 with ridge $\lambda$, $\widehat{\mathbf{g}}_t \leftarrow \nabla \widehat{m}_t(\mathbf{0})$, $\widehat{\mathbf{H}}_t \leftarrow \nabla^2 \widehat{m}_t(\mathbf{0})$

5: $\mathbf{d}_t \leftarrow -|\widehat{\mathbf{H}}_t|_{\epsilon_H}^{-1} \widehat{\mathbf{g}}_t$, $\quad \widehat{s}_t \leftarrow -\widehat{\mathbf{g}}_t^\top \mathbf{d}_t$

6: $\star$ $\underline{s}_t \leftarrow \widehat{s}_t - \kappa_t \widehat{\sigma}_{s,t}$ by equation 11

7: $\star$ $\mathcal{H}_t^{\mathrm{acc}} \leftarrow \{\eta \in \mathcal{H}_t : \widehat{R}_t(\eta) + \kappa_t \widehat{\sigma}_{R,t}(\eta) \leq 0\}$ if $\underline{s}_t > 0$, else $\emptyset$

8: If $\mathcal{H}_t^{\mathrm{acc}} \neq \emptyset$, $\eta_t \leftarrow \max \mathcal{H}_t^{\mathrm{acc}}$. Otherwise $\mathbf{d}_t \leftarrow -\widehat{\mathbf{g}}_t$ and $\eta_t \leftarrow \min\{\eta_0/\sqrt{t+1},\ \Delta_t/\|\widehat{\mathbf{g}}_t\|\}$ ($\eta_t \leftarrow 0$ if $\widehat{\mathbf{g}}_t = \mathbf{0}$)

9: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \mathbf{V}_t(\eta_t \mathbf{d}_t)$

In this section we propose ProbFedZO, the variant of SurrFedZO for partial client participation. Stages 1 and 2 of in SurrFedZO in Section 3 are unchanged while Stage 3 needs a modification. When only $m < N$ clients participate, $\bar{\mathbf{y}}$ is a random estimate of the exact response vector, so $\widehat{s}$ and $\widehat{R}(\eta)$ are random, and a step that passes the test of $\widehat{R}(\eta) \leq 0$ may not pass the same test on the surrogate all clients would have produced. ProbFedZO adds one ingredient to SurrFedZO: an estimate of how far $\widehat{s}$ and $\widehat{R}(\eta)$ can be from their population values, and a test that checks whether the predicted decrease to exceed this uncertainty. Below we explain how to modify Stage 3 in SurrFedZO.

### 4.1 SAMPLING UNCERTAINTY

Since all participating clients evaluate the same probes, the covariance of $\bar{\mathbf{y}}$ can be estimated from the variation of $\mathbf{y}_i, i \in S$,

$$\widehat{\mathbf{C}} = \frac{N-m}{Nm} \cdot \frac{1}{m-1} \sum_{i \in S} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top, \tag{10}$$

where the first factor accounts for sampling without replacement and vanishes at $m = N$. Because $\widehat{s}$ and $\widehat{R}(\eta)$ are linear in $\bar{\mathbf{y}}$, their standard errors follow directly:

$$\widehat{\sigma}_s = \sqrt{\mathbf{u}^\top \widehat{\mathbf{C}}\, \mathbf{u}}, \qquad \widehat{\sigma}_R(\eta) = \sqrt{\mathbf{b}(\eta)^\top \widehat{\mathbf{C}}\, \mathbf{b}(\eta)}, \tag{11}$$

with $\mathbf{u}$ and $\mathbf{b}(\eta) \in \mathbb{R}^K$ being the coefficient vectors in the formulae $\widehat{s} = \mathbf{u}^\top \bar{\mathbf{y}}$ and $\widehat{R}(\eta) = \mathbf{b}(\eta)^\top \bar{\mathbf{y}}$ (Appendix B, Equations equation 23 and equation 24). Note that these standard errors measure uncertainty due to client sampling only, not the approximation error of the surrogate.

### 4.2 PROBABILISTIC ARMIJO TEST

A multiplier $\kappa > 0$, fixed before the clients are sampled, sets the size of uncertainty margin (with respect to the standard error); larger $\kappa$ becomes the algorithm being more conservative to accept a candidate step size. The server first checks the direction,

$$\underline{s} = \widehat{s} - \kappa\, \widehat{\sigma}_s > 0, \tag{12}$$

that is, the predicted slope must exceed its uncertainty. It then tests each candidate step,

$$Q(\eta) = \widehat{R}(\eta) + \kappa\, \widehat{\sigma}_R(\eta) \le 0, \tag{13}$$

that is, the predicted Armijo margin must cover its uncertainty. From the grid $\mathcal{H}$ of Section 3.4, the server takes the largest certified step,

$$\eta_t = \max\{\eta \in \mathcal{H} : Q(\eta) \le 0\}. \tag{14}$$

If the direction test fails or no candidate is certified, Algorithm 2 uses a pre-scheduled step size for $\eta_t$. Finally, the final updated model becomes $\mathbf{w}_{t+1} = \mathbf{w} + \mathbf{V}(\eta_t \mathbf{d})$.

**Relation to SurrFedZO.** At $m = N$ the factor $(N-m)/(Nm)$ in equation 10 becomes zero, so $\widehat{\mathbf{C}} = \mathbf{0}$, both standard errors vanish, the tests reduce to $\widehat{s} > 0$ and $\widehat{R}(\eta) \le 0$, and Algorithm 2 coincides with Algorithm 1. The lines that ProbFedZO adds are marked $\star$ in Algorithm 2. In the experiments we set $\kappa = \sqrt{\rho/(1-\rho)}$ for a tuning parameter $\rho \in (0,1)$; $\rho$ is not the probability that a certified step is valid (Remark 2). Section 5 gives the guarantee for a given $\kappa$.

## 5 THEORETICAL GUARANTEES

There are two errors in the decision of the choice of step size. The *model error* between the surrogate and the global objective is deterministic and controlled by the trust radius. The *sampling error* between the surrogate fitted to randomly participating clients and the one fitted from all clients would have produced, is random and vanishes at full participation. SurrFedZO faces only the first, so we treat it first, and ProbFedZO adds the second on top of the first. Proofs and full assumptions are given in Appendix C.

### 5.1 FULL CLIENT PARTICIPATION

On the event of Proposition 1, a certified step satisfies $R_t \le 0$ and $s_t \ge \underline{s}_t > 0$, so Theorem 1 applies to the population surrogate and gives $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \le -c\, \eta_t \underline{s}_t + \varepsilon_t \Delta_t$. The proof is in Appendix C.

**Lemma 1** (Model error)**.** *Under Assumption 5, the surrogate gradient error on the trust region* $\Omega_t = \{\|\mathbf{z}\| \le \Delta_t\}$ *satisfies* $\varepsilon_t := \sup_{\mathbf{z} \in \Omega_t} \|\nabla f_t(\mathbf{z}) - \nabla \widehat{m}_t(\mathbf{z})\| \le \kappa_{eg} \Delta_t^p$*, where* $\kappa_{eg}$ *depends on* $L_p, p$ *and the probe geometry, but not on the number of probes or clients.*

The error is controlled by the radius $\Delta_t$, but not by the number of probes. Once the number of probes exceed the $\binom{r+p}{p}$ points, the surrogate can perfectly reconstruct $f_t$, and so additional probes can only reduce the constant $\kappa_{eg}$. The full proof is in Appendix D.

**Theorem 1** (True Armijo condition)**.** *Let Assumption 5 hold and let SurrFedZO accept $\eta_t$ with $\mathbf{z}_t = \eta_t \mathbf{d}_t$. Then*

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq -c\,\eta_t \widehat{s}_t + \varepsilon_t \|\mathbf{z}_t\|. \tag{15}$$

*If moreover $\varepsilon_t \|\mathbf{d}_t\| \leq \theta\, c\, \widehat{s}_t$ for some $\theta \in [0, 1)$, then $F$ strictly decreases and the Armijo condition holds for the global objective with constant at least $(1-\theta)c/(1+\theta c)$.*

### 5.2 Partial client participation

When only $m < N$ clients respond, the server sees a random estimate $\bar{\mathbf{y}}$ of the responses. Everything computed from it is then random, including $\widehat{s}_t$, $\widehat{R}_t(\eta)$, the direction $\mathbf{d}_t$ and the selected step $\eta_t$. Write $s_t$ and $R_t(\eta)$ for the values of $\widehat{s}_t$ and $\widehat{R}_t(\eta)$ under full participation, along the same direction $\mathbf{d}_t$. By Theorem 1, the global objective decreases at a step with $s_t > 0$ and $R_t(\eta_t) \leq 0$. The server cannot check these two conditions, since it does not observe $s_t$ and $R_t$. It can, however, check equation 12 and equation 13 of ProbFedZO, and these imply the two conditions

$$s_t \geq \widehat{s}_t - \kappa_t \widehat{\sigma}_{s,t}, \qquad R_t(\eta_t) \leq \widehat{R}_t(\eta_t) + \kappa_t \widehat{\sigma}_{R,t}(\eta_t). \tag{16}$$

So decrease of the global objective is guaranteed with the probability that equation 16 holds. For fixed direction and step size, this probability is at least $\rho$. The server, however, chooses $\mathbf{d}_t$ from $\bar{\mathbf{y}}$, tests every $\eta$ in the grid, and keeps one that passes the Armijo test, and so equation 16 must hold for all directions and all steps simultaneously. Proposition 1 proves this uniform guarantee.

**Proposition 1** (Simultaneous sampling control)**.** *Fix $\kappa_t > 0$ before sampling. With probability at least $[1 - d_t^{\text{eff}}/\kappa_t^2]_+$, every tested direction and step satisfies $R_t(\eta \mathbf{d}) \leq \widehat{R}_t(\eta) + \kappa_t \widehat{\sigma}_{R,t}(\eta)$ and $s_t(\mathbf{d}) \geq \widehat{s}_t - \kappa_t \widehat{\sigma}_{s,t}$ simultaneously, where $d_t^{\text{eff}} \leq \min\{K,\, N-1,\, \binom{r+p}{p} - 1\}$.*

On the event of Proposition 1, a step size certified by the tests equation 12 and equation 13 satisfies, by equation 16,

$$s_t \;\geq\; \widehat{s}_t - \kappa_t \widehat{\sigma}_{s,t} \;=\; \underline{s}_t \;>\; 0, \qquad R_t(\eta_t) \;\leq\; \widehat{R}_t(\eta_t) + \kappa_t \widehat{\sigma}_{R,t}(\eta_t) \;\leq\; 0.$$

These are the conditions of Theorem 1 for the population surrogate, whose model error is still bounded by Lemma 1. Hence $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq -c\,\eta_t \underline{s}_t + \varepsilon_t \Delta_t$, and the true Armijo condition holds with $\underline{s}_t$ in place of $\widehat{s}_t$. The proof is in Appendix C.

**Aggregate progress.** Let $\mathcal{G}_t$ be the event where the proposition 1 holds. Under the condition of Theorem 1, the objective decreases by at least $\delta_t = c\,\eta_t s_t$ up to the model error on $\mathcal{G}_t$. On the other hand, outside $\mathcal{G}_t$, the objective may increase, and we write $I_T$ for the sum of these increases over $T$ rounds. Let $M_T = \kappa_{eg} \sum_{t<T} \Delta_t^{p+1}$ be the accumulated model error. If $F$ is bounded below by $F^\star$ and $L$-smooth, Theorem 2 gives

$$\min_{0 \leq t < T} \mathbb{E}\big[\delta_t \mathbf{1}_{\mathcal{G}_t}\big] \leq \frac{F(\mathbf{w}_0) - F^\star + I_T + M_T}{T}. \tag{17}$$

The left side averages the progress of certified rounds only, so the bound says that certified rounds make progress at rate $1/T$ once $I_T$ and $M_T$ are bounded.

## 6 Experiments

We evaluate SurrFedZO and ProbFedZO on illustrative two-dimensional nonconvex functions, federated logistic regression, and language model fine-tuning to assess their optimization performance and communication efficiency. Toy examples illustrate the behavior of the surrogate line search (Appendix E.3). Logistic regression is evaluated under both complete and random participation regimes, with SurrFedZO when all clients participate and ProbFedZO when a random subset does, and a fine-tuned language model is evaluated under random participation with ProbFedZO. All comparisons are made with the same communication budget constraints. We also examine the effects of the update direction and the uncertainty threshold through ablation studies (Appendices E.8 and E.4). All ProbFedZO experiments use the uninflated plug-in rule in Algorithm 2.

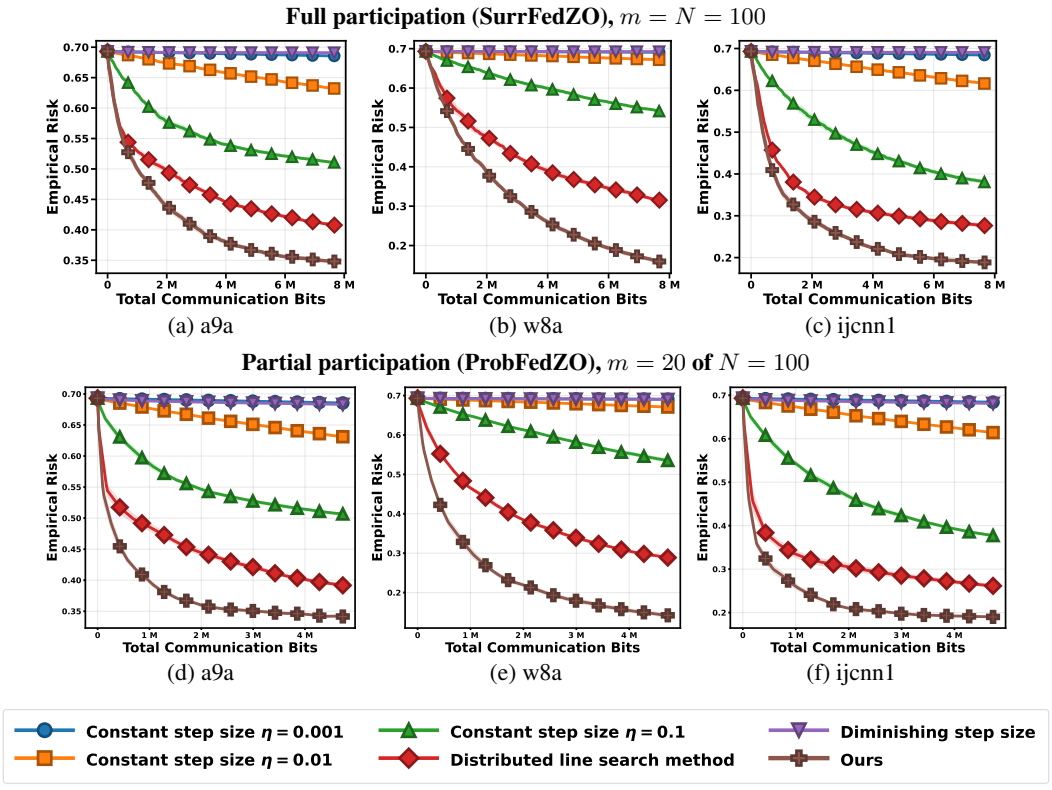


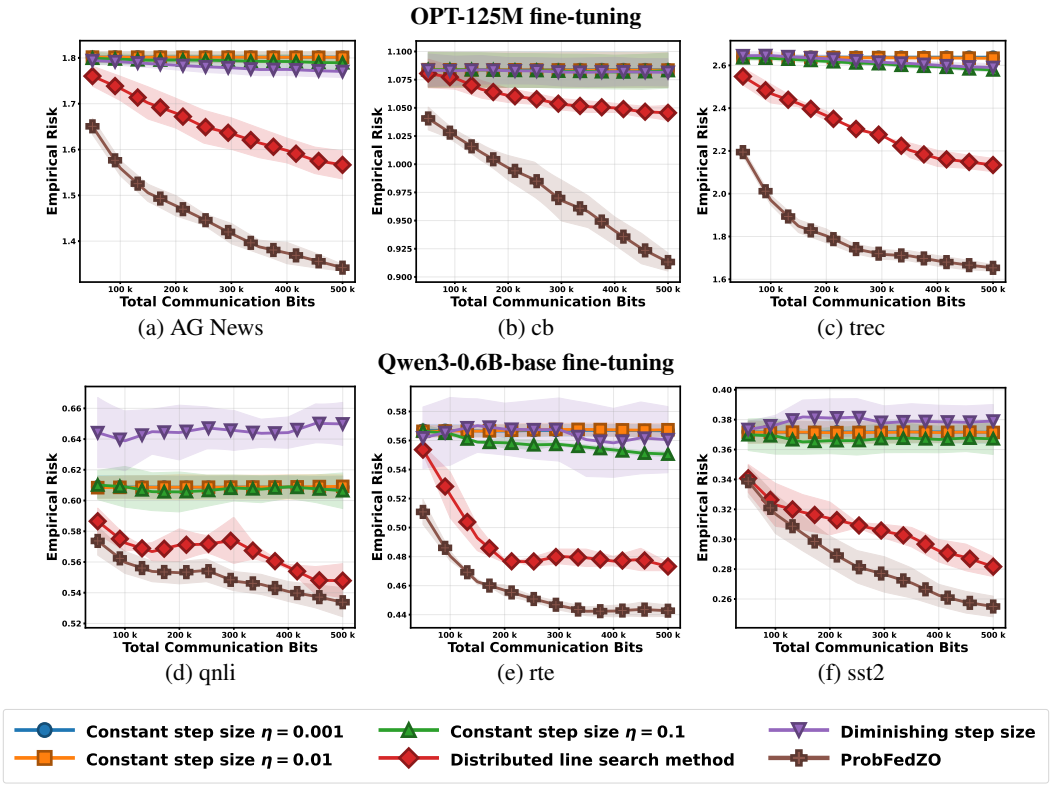

Figure 2: Federated logistic regression, mean $\pm$ standard error over five seeds. Top row: SurrFedZO with all 100 clients. Bottom row: ProbFedZO with 20 sampled clients. Three more benchmarks are in Appendix E.2.

### 6.1 Experiments on Binary Logistic Regression

Suppose that client $i$ holds $n_i$ labeled examples $(\mathbf{x}, y)$ with $y \in \{-1, +1\}$, and its local objective is the $\ell_2$-regularized logistic loss $F_i(\mathbf{w}) = \frac{1}{n_i} \sum_{(\mathbf{x},y)\in\mathcal{D}_i} \log(1 + e^{-y\mathbf{x}^\top \mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$. We analyze `a9a` (Platt, 1999), `w8a` (Platt, 1999) and `ijcnn1` (Prokhorov, 2001) datasets here, and defer `a5a`, `w1a` and `phishing` datasets to Appendix E.2, where the conclusions are similar. Each dataset split across $N = 100$ clients with label skew. For every class $c \in \{-1, 1\}$, we shuffle its training examples, draw a proportion vector $\mathbf{p}_c \sim \text{Dirichlet}(\alpha \mathbf{1}_N)$ with $\alpha = 0.1$, and assign a fraction $p_{c,k}$ of the class-$c$ examples to client $k$. Since $\alpha = 0.01$ is small, examples of each class are concentrated on few clients, so most clients hold either class label 1 or -1 as the majority label and thus the local objectives differ substantially across clients. Results for other values of $\alpha$ are reported in Appendix E.5.

**Full participation.** All 100 clients participate in every round, so the surrogate is fitted to the exact changes in the global objective and the step is selected by the test of SurrFedZO. The top row of Figure 2 reports the results. Under the constrained communication budget, SurrFedZO reaches a lower empirical risk than the fixed and diminishing schedules and the distributed line search on every benchmark, since the baselines either progress slowly or spend part of the budget on candidate evaluations. Since no client sampling is involved, this experiment isolates the value of selecting the step on the server from the surrogate.

**Random participation.** Twenty of the 100 clients are sampled randomly per round, and the step is certified by the sampling-aware test of ProbFedZO. The bottom row of Figure 2 reports the results. ProbFedZO attains competitive or superior risk at matched communication. The ablations show that uncertainty-driven rejections increase with client heterogeneity (Appendix E.5). Stricter certification thresholds increase the fraction of certified steps that reduce the global objective, but certify fewer steps (Appendix E.6).




(a) AG News (b) cb (c) trec

(d) qnli (e) rte (f) sst2

Figure 3: Empirical risk against total communication bits for OPT-125M fine-tuning in panels (a)–(c) and Qwen3-0.6B-base fine-tuning in panels (d)–(f). All panels share the legend below the plots.

### 6.2 Experiments on Language Models

We evaluate LoRA fine-tuning of OPT-125M (Zhang et al., 2022) on AG News (Zhang et al., 2015), CB (de Marneffe et al., 2019) and TREC (Li & Roth, 2002), and Qwen3-0.6B-base (Yang et al., 2025) on QNLI (Rajpurkar et al., 2016), RTE (Dagan et al., 2006) and SST-2 (Socher et al., 2013). Training data are partitioned across 100 clients with Dirichlet label skew $\alpha = 0.1$, and 10 clients are sampled per round, so these experiments use ProbFedZO. Only LoRA parameters are updated. All methods use DeComFL (Li et al., 2024). At $r = 2$, the tested fixed and diminishing schedules stall, whereas ProbFedZO reduces risk on all six tasks (Figure 3). The proposed directions outperform negative surrogate gradients on Qwen3-0.6B-base RTE and BoolQ at matched probe budgets and certification thresholds (Appendix E.8). The threshold sweep separately shows poorer progress when adaptive proposals are suppressed and logged steps match the fallback (Appendix E.4). These results support both direction construction and step selection, not certification alone. Qwen3-0.6B-base train and test accuracies are in Appendix E.9.

## 7 Conclusion

SurrFedZO constructs an update direction and selects a step size on the server from a single round of client evaluations, so candidate search adds no communication. Under full participation, SurrFedZO applies a Armijo test to a surrogate fitted to exact values of the global objective, and an accepted step satisfies the Armijo condition for the true objective under a condition on the probe radius. Under partial participation, ProbFedZO adds a sampling correction to the same test, with a guarantee that holds simultaneously for all candidates despite adaptive selection and carries the full-participation result over to the sampling event. Experiments show competitive or improved risk at matched communication budgets in both regimes. Certifying the uninflated plug-in rule, controlling fallback increases, and adapting the probe radius across rounds remain open.

## AI USE STATEMENT

We did not use generative AI tools for any task that requires disclosure. We used them for tasks with recommended disclosure: writing and editing software code, producing and editing figures, drafting the paper, editing it for readability, and formatting references. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with its aid.

## ETHICS STATEMENT

This work studies federated zeroth-order optimization. Although the proposed method operates without directly sharing clients' raw data, this alone does not provide a formal privacy guarantee, and communicated information may reveal properties of local data. Deployments involving sensitive data would therefore require additional privacy safeguards. Potential biases in the training data or pretrained models are not addressed by the proposed optimization method.

## REPRODUCIBILITY STATEMENT

The main paper describes the proposed algorithm and its theoretical assumptions, with detailed proofs provided in the appendix. The experimental sections and appendix describe the datasets, evaluation protocols, and hyperparameter settings. We will release the implementation and experiment configurations to facilitate reproduction of the reported results.

## REFERENCES

Afonso S. Bandeira, Katya Scheinberg, and Luís N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.

Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. ZO-AdaMM: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Andrew R. Conn, Katya Scheinberg, and Luís N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, PA, 2009.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pp. 177–190. Springer, 2006.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*, pp. 107–124, 2019.

Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N. Jones, and Yong Zhou. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized subspace Newton. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dongyoon Kim, Sungjae Lee, Wonjin Lee, and Kwang In Kim. Subspace-based approximate Hessian method for zeroth-order optimization. *arXiv preprint arXiv:2507.06125*, 2025.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications*, 79(2):339–368, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *International Conference on Learning Representations (ICLR)*, 2020b. URL https://openreview.net/forum?id=HJxNAnVtDS.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. DeComFL: Federated learning with dimension-free communication. In *International Workshop on Federated Foundation Models in Conjunction with NeurIPS*, 2024. URL https://openreview.net/forum?id=Vy9ltlTXXd.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53038–53075, 2023.

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 (AISTATS)*, 2017.

Elissa Mhanna and Mohamad Assaad. Communication and energy efficient federated learning using zero-order optimization technique. *arXiv preprint arXiv:2409.16456*, 2024.

Rami Mohammad and Lee McCluskey. Phishing websites. UCI Machine Learning Repository, 2012. DOI: 10.24432/C51W2X.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 185–208. MIT Press, Cambridge, MA, 1999.

Danil Prokhorov. IJCNN 2001 neural network competition. Slide presentation at IJCNN 2001, Ford Research Laboratory, 2001. URL https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ref.html.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making Newton-type methods applicable to federated learning. In *International Conference on Machine Learning (ICML)*, pp. 18959–19010. PMLR, 2022.

Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1–2): 87–110, 1953. doi: 10.1093/biomet/40.1-2.87.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7611–7623, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Ziming Yu, Shuyao Xiao, Xingyu Zhao, Sike Wang, Pan Zhou, Peiyu Zang, Xiangda Yan, Yongjie Yang, and Jia Li. SubZero+: Efficient zeroth-order LLM fine-tuning via large learning rates. *arXiv preprint arXiv:2608.15665*, 2026.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*, 2018.

# A GUIDE TO THE APPENDIX

Appendix B gives the closed form of the surrogate, its gradient and Hessian, and the coefficient vectors used in the uncertainty estimates. Appendix D proves Lemma 1 and Theorem 1, and Appendix C proves Proposition 1 and Theorem 2. Appendix E.2 reports logistic regression on `a5a`, `w1a` and `phishing`, and Appendix E.3 the two-dimensional examples behind Section 5. The ablations are: the certification threshold $\rho$ on logistic regression (Appendix E.6) and on LoRA fine-tuning (Appendix E.4), client heterogeneity $\alpha$ (Appendix E.5), the probe radius (Appendix E.10), the surrogate kernel (Appendix E.7), and the search direction (Appendix E.8).

# B CONSTRUCTION AND NOTATION

The main text presents the operations and acceptance tests. This appendix collects the linear-algebra formulas used to implement them and the population quantities used in the analysis. Unlike the fixed-round method explanation, all quantities here retain their round indices. In particular, the method's $\widehat{m}$, $\widehat{R}$, and $Q$ correspond to $\widehat{m}_t$, $\widehat{R}_t$, and $Q_t$ below. This restores the indices without changing any of the estimators or tests.

**Response aggregation.** The population and sample means are

$$\boldsymbol{\mu}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_{i,t}, \qquad \bar{\mathbf{y}}_t = \frac{1}{m}\sum_{i\in S_t} \mathbf{y}_{i,t}, \tag{18}$$

Write $\boldsymbol{\Sigma}_t$ for the population covariance of client responses and $\widehat{\boldsymbol{\Sigma}}_t$ for their sample covariance, using denominators $N-1$ and $m-1$, respectively. Set $f_m = (N-m)/(Nm)$, $\mathbf{C}_t = f_m\boldsymbol{\Sigma}_t$, and $\widehat{\mathbf{C}}_t = f_m\widehat{\boldsymbol{\Sigma}}_t$. Thus $\boldsymbol{\Sigma}_t$ describes individual clients, whereas $\mathbf{C}_t$ describes uncertainty in the sampled mean. The sample mean is unbiased, but a particular client sample can suggest a decrease that the population does not. Since the same clients answer every probe, these errors are correlated across coordinates. For $m \geq 2$ the estimated covariance of the mean is

$$\widehat{\mathbf{C}}_t = \frac{N-m}{Nm(m-1)}\sum_{i\in S_t}(\mathbf{y}_{i,t}-\bar{\mathbf{y}}_t)(\mathbf{y}_{i,t}-\bar{\mathbf{y}}_t)^\top. \tag{19}$$

This scales the sample covariance by $(N-m)/(Nm)$ to account for sampling without replacement. The full covariance retains the correlations needed for uncertainty propagation.

**Kernel construction and derivatives.** Add the deterministic zero anchor $\mathbf{z}_t^0 = \mathbf{0}$ to the probes when constructing the kernel matrix $\mathbf{K}_t$ and evaluation vector $\mathbf{k}_t(\mathbf{z})$, indexed by $0,\dots,K$. Let $\mathbf{A}_t = \mathbf{K}_t + \lambda\mathbf{I}_{K+1}$. Define the $K$-dimensional coefficient map

$$\mathbf{a}_t(\mathbf{z}) := \left[\mathbf{A}_t^{-1}\big(\mathbf{k}_t(\mathbf{z}) - \mathbf{k}_t(\mathbf{0})\big)\right]_{1:K}, \qquad \mathbf{a}_t(\mathbf{0}) = \mathbf{0}. \tag{20}$$

Here $[\cdot]_{1:K}$ discards coordinate zero. This absorbs both the zero response and output centering into the coefficient map, while preserving the original anchored KRR fit. All response vectors, coefficients, and response covariance matrices henceforth have dimension $K$. The sampled surrogate is

$$\widehat{m}_t(\mathbf{z}) := \mathbf{a}_t(\mathbf{z})^\top\bar{\mathbf{y}}_t. \tag{21}$$

Anchoring ensures $\widehat{m}_t(\mathbf{0}) = 0$. For fixed probes and kernel, the prediction is linear in $\bar{\mathbf{y}}_t$, as is its gradient. Differentiating gives

$$\widehat{\mathbf{g}}_t := \nabla\widehat{m}_t(\mathbf{0}) = \mathbf{G}_t\bar{\mathbf{y}}_t, \qquad \widehat{\mathbf{H}}_t := \nabla^2\widehat{m}_t(\mathbf{0}), \qquad \mathbf{G}_t := \nabla_{\mathbf{z}}\mathbf{a}_t(\mathbf{0})^\top. \tag{22}$$

The gradient's linear form propagates covariance as $\widehat{\mathbf{C}}_{g,t} = \mathbf{G}_t\widehat{\mathbf{C}}_t\mathbf{G}_t^\top$. For a fixed direction $\mathbf{d}$, the estimated standard error of the slope is

$$\widehat{\sigma}_{d,t}(\mathbf{d}) := \sqrt{\mathbf{d}^\top\mathbf{G}_t\widehat{\mathbf{C}}_t\mathbf{G}_t^\top\mathbf{d}}. \tag{23}$$

Any update formed by a linear transformation of the surrogate gradient can be tested this way (Appendix B). Sampling uncertainty is distinct from model error, which separates the population-surrogate gradient $\mathbf{g}_t = \mathbf{G}_t\boldsymbol{\mu}_t$ from the true projected gradient $\mathbf{p}_t = \mathbf{V}_t^\top\nabla F(\mathbf{w}_t)$.

Table 1: Shared notation for the method and descent analysis.

| Symbol | Meaning |
|---|---|
| $N$ and $m$ | Total and participating client counts |
| $d$ and $r$ | Parameter and subspace dimensions |
| $K$ and $Z_t$ | Probe count and probe set excluding the zero anchor |
| $\mathbf{w}_t$ and $\mathbf{V}_t$ | Global iterate and subspace basis |
| $\Delta_t$ | Trust-region radius |
| $\boldsymbol{\mu}_t$ and $\bar{\mathbf{y}}_t$ | Population and sampled mean response vectors |
| $\boldsymbol{e}_t$ | Sampling error $\bar{\mathbf{y}}_t - \boldsymbol{\mu}_t$ |
| $\boldsymbol{\Sigma}_t$ and $\widehat{\boldsymbol{\Sigma}}_t$ | Population and sample covariance of client responses |
| $\mathbf{C}_t = f_m \boldsymbol{\Sigma}_t$ | Covariance of the sampled mean; $\widehat{\mathbf{C}}_t = f_m \widehat{\boldsymbol{\Sigma}}_t$ |
| $f_t$, $m_t$, and $\widehat{m}_t$ | True loss change, population surrogate, and sampled surrogate |
| $\mathbf{p}_t$ | Projected analytical gradient |
| $\mathbf{g}_t$ and $\widehat{\mathbf{g}}_t$ | Population and sampled surrogate gradients |
| $\mathbf{d}_t$ and $\eta_t$ | Selected update direction and step size |
| $s_t(\mathbf{d})$ and $\widehat{s}_t(\mathbf{d})$ | Population and sampled surrogate descent slopes; omit the argument at $\mathbf{d}_t$ |
| $\sigma_{R,t}$ and $\widehat{\sigma}_{R,t}$ | Residual uncertainty using exact and estimated covariance |
| $\sigma_{d,t}$ and $\widehat{\sigma}_{d,t}$ | Directional uncertainty using exact and estimated covariance |
| $\underline{s}_t$ and $Q_t(\eta)$ | Certified directional margin and Armijo upper bound |
| $\mathcal{H}_t$ and $\mathcal{H}_t^{\mathrm{acc}}$ | Candidate and certified step-size sets |
| $E_t$ and $\varepsilon_t$ | Model error $f_t - m_t$ and its maximum gradient norm |
| $\delta_t$ | Successful-round decrease bound |
| $d_t^{\mathrm{eff}}$ | Admissible effective dimension (not the direction $\mathbf{d}_t$) |
| $\mathcal{G}_t$ | Accepted round with sampling and model-accuracy guarantees |
| $\kappa_t$ and $\rho_{\mathrm{cov},t}$ | Input multiplier and sampling coverage lower bound |

**Implemented residual uncertainty.** For fixed $\eta$ and $\mathbf{d}_t$, the sampled Armijo residual is linear in $\bar{\mathbf{y}}_t$, giving

$$\widehat{R}_t(\eta \mathbf{d}_t) = \mathbf{b}_t(\eta \mathbf{d}_t)^\top \bar{\mathbf{y}}_t, \qquad \widehat{\sigma}_{R,t}(\eta \mathbf{d}_t) = \sqrt{\mathbf{b}_t(\eta \mathbf{d}_t)^\top \widehat{\mathbf{C}}_t \mathbf{b}_t(\eta \mathbf{d}_t)}, \tag{24}$$

where $\mathbf{b}_t(\eta \mathbf{d}_t) = \mathbf{a}_t(\eta \mathbf{d}_t) - c\eta \mathbf{G}_t^\top \mathbf{d}_t$.

**Conditional sampling identities.** Let $\mathcal{F}_t$ contain the history and current subspace before drawing the probes, and let $\mathcal{F}_t^+ = \mathcal{F}_t \vee \sigma(Z_t)$ also include the realized design, with kernel parameters fixed before client sampling. For $\boldsymbol{e}_t = \bar{\mathbf{y}}_t - \boldsymbol{\mu}_t$, uniform sampling without replacement gives

$$\mathbb{E}[\boldsymbol{e}_t \mid \mathcal{F}_t^+] = 0, \qquad \mathbf{C}_t = \mathrm{Cov}(\boldsymbol{e}_t \mid \mathcal{F}_t^+) = \frac{N-m}{Nm} \boldsymbol{\Sigma}_t, \tag{25}$$

where $\boldsymbol{\Sigma}_t$ is the population covariance of the client responses (Assumption 1). Both sampling gaps are linear in this same error vector: $\widehat{R}_t - R_t = \mathbf{b}_t^\top \boldsymbol{e}_t$ and $\widehat{s}_t - s_t = -(\mathbf{G}_t^\top \mathbf{d}_t)^\top \boldsymbol{e}_t$. A single event controlling all tested coefficients therefore covers the selected direction and step without a union bound over candidates.

The following symbols have the same meanings in the method and proofs.

**Surrogates.** Using the coefficient map in equation 20, the population and sampled surrogates are

$$m_t(\mathbf{z}) := \mathbf{a}_t(\mathbf{z})^\top \boldsymbol{\mu}_t, \qquad \widehat{m}_t(\mathbf{z}) := \mathbf{a}_t(\mathbf{z})^\top \bar{\mathbf{y}}_t, \tag{26}$$

with $\mathbf{G}_t := \nabla_{\mathbf{z}} \mathbf{a}_t(\mathbf{0})^\top$, so $\mathbf{g}_t := \nabla m_t(\mathbf{0}) = \mathbf{G}_t \boldsymbol{\mu}_t$ and $\widehat{\mathbf{g}}_t := \nabla \widehat{m}_t(\mathbf{0}) = \mathbf{G}_t \bar{\mathbf{y}}_t$. The centering yields $m_t(\mathbf{0}) = \widehat{m}_t(\mathbf{0}) = f_t(\mathbf{0}) = 0$ and leaves the gradients unchanged.

**Update transformation.** Algorithm 2 uses $\mathbf{d}_t = -|\widehat{\mathbf{H}}_t|_{\epsilon_H}^{-1} \widehat{\mathbf{g}}_t$. Here $|\widehat{\mathbf{H}}_t|_{\epsilon_H}$ preserves the eigenvectors and replaces each eigenvalue $\lambda_{t,j}$ by $\max\{|\lambda_{t,j}|, \epsilon_H\}$ with $\epsilon_H > 0$. For a fixed transformation matrix the direction is linear in the surrogate gradient. The matrix itself is fitted from the round's responses.

**Residuals.** For a candidate $(\eta, \mathbf{d})$ and Armijo constant $c \in (0, 1)$,

$$\widehat{R}_t(\eta\mathbf{d}) := \widehat{m}_t(\eta\mathbf{d}) - c\eta\,\widehat{\mathbf{g}}_t^\top \mathbf{d}, \qquad R_t(\eta\mathbf{d}) := m_t(\eta\mathbf{d}) - c\eta\,(\mathbf{g}_t)^\top \mathbf{d}, \tag{27}$$

and, for $c' \in (0, c]$, the true residual

$$R_t^{\mathrm{true},c'}(\eta\mathbf{d}) := f_t(\eta\mathbf{d}) - c'\eta\,\mathbf{p}_t^\top \mathbf{d}. \tag{28}$$

**Uncertainties.** The coefficient vector shared by the residual and its uncertainty is

$$\mathbf{b}_t(\eta\mathbf{d}) := \mathbf{a}_t(\eta\mathbf{d}) - c\eta\mathbf{G}_t^\top \mathbf{d}. \tag{29}$$

Thus $\mathbf{b}_t(\mathbf{0}) = \mathbf{0}$ and $\widehat{R}_t(\eta\mathbf{d}) = \mathbf{b}_t(\eta\mathbf{d})^\top \bar{\mathbf{y}}_t$. For the population-covariance analysis the uncertainty terms are

$$\sigma_{R,t}(\eta\mathbf{d}) := \sqrt{\mathbf{b}_t(\eta\mathbf{d})^\top \mathbf{C}_t \mathbf{b}_t(\eta\mathbf{d})}, \tag{30}$$
$$\sigma_{d,t}(\mathbf{d}) := \sqrt{\mathbf{d}^\top \mathbf{G}_t \mathbf{C}_t \mathbf{G}_t^\top \mathbf{d}}.$$

The hatted, implemented versions in equation 23 and equation 24 replace $\mathbf{C}_t$ with $\widehat{\mathbf{C}}_t$. The feasibility calculation also applies with this replacement.

# C Proofs of Sampling Control and Conditional Progress

## C.1 Setting and notation

We fix a communication round $t$ and condition throughout on

$$\mathcal{F}_t^+ := \mathcal{F}_t \vee \sigma(Z_t), \tag{31}$$

the history of the algorithm up to round $t$ together with the probe design $Z_t$ drawn in that round. All conditional probabilities and expectations are with respect to $\mathcal{F}_t^+$ unless stated otherwise. Under this conditioning the only randomness remaining in the round is the client sample.

The server holds $\mathbf{w}_t \in \mathbb{R}^d$ and a subspace basis $\mathbf{V}_t \in \mathbb{R}^{d\times r}$ with $\mathbf{V}_t^\top \mathbf{V}_t = \mathbf{I}_r$ and $r \ll d$. It draws $S_t \subset \{1, \dots, N\}$ uniformly without replacement with $|S_t| = m \geq 2$ and broadcasts $K$ probes $Z_t = \{\mathbf{z}_t^1, \dots, \mathbf{z}_t^K\} \subset \Omega_t := \{\mathbf{z} \in \mathbb{R}^r : \|\mathbf{z}\| \leq \Delta_t\}$. The radius $\Delta_t$ and the multiplier $\kappa_t$ are $\mathcal{F}_t$-measurable.

**Responses.** Each participating client returns the $K$ scalars $y_{i,t}^k = \Delta_{i,t}(\mathbf{z}_t^k)$ where $\Delta_{i,t}(\mathbf{z}) = F_i(\mathbf{w}_t + \mathbf{V}_t\mathbf{z}) - F_i(\mathbf{w}_t)$. The deterministic zero anchor is absorbed into the surrogate coefficient map. We use the response vectors and their population and sample means

$$\mathbf{y}_{i,t} = [y_{i,t}^1, \dots, y_{i,t}^K]^\top \in \mathbb{R}^K, \qquad \boldsymbol{\mu}_t = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_{i,t}, \qquad \bar{\mathbf{y}}_t = \frac{1}{m}\sum_{i\in S_t} \mathbf{y}_{i,t}, \tag{32}$$

and write the sampling error and its covariance as

$$\boldsymbol{e}_t := \bar{\mathbf{y}}_t - \boldsymbol{\mu}_t, \qquad \mathbf{C}_t := \mathrm{Cov}(\boldsymbol{e}_t \mid \mathcal{F}_t^+). \tag{33}$$

**Surrogate.** Use the coefficient map $\mathbf{a}_t : \mathbb{R}^r \to \mathbb{R}^K$ from equation 20. The population and sampled surrogates are

$$m_t(\mathbf{z}) = \mathbf{a}_t(\mathbf{z})^\top \boldsymbol{\mu}_t, \qquad \widehat{m}_t(\mathbf{z}) = \mathbf{a}_t(\mathbf{z})^\top \bar{\mathbf{y}}_t, \tag{34}$$

so that $m_t(\mathbf{0}) = \widehat{m}_t(\mathbf{0}) = 0$. Throughout, $\nabla_{\mathbf{z}}\mathbf{a}_t(\mathbf{z}) \in \mathbb{R}^{K\times r}$ denotes the Jacobian of $\mathbf{a}_t$, whose $j$th row is $\nabla a_{t,j}(\mathbf{z})^\top$. With this convention $\mathbf{G}_t := \nabla_{\mathbf{z}}\mathbf{a}_t(\mathbf{0})^\top \in \mathbb{R}^{r\times K}$, so that $\nabla_{\mathbf{z}}\mathbf{a}_t(\mathbf{0}) = \mathbf{G}_t^\top$, and the two surrogate gradients at the origin are

$$\mathbf{g}_t = \nabla m_t(\mathbf{0}) = \mathbf{G}_t\boldsymbol{\mu}_t, \qquad \widehat{\mathbf{g}}_t = \nabla \widehat{m}_t(\mathbf{0}) = \mathbf{G}_t\bar{\mathbf{y}}_t. \tag{35}$$

**Population objective.** The population loss change in subspace coordinates and the projected analytical gradient are

$$f_t(\mathbf{z}) = F(\mathbf{w}_t + \mathbf{V}_t\mathbf{z}) - F(\mathbf{w}_t), \qquad \mathbf{p}_t = \nabla f_t(\mathbf{0}) = \mathbf{V}_t^\top \nabla F(\mathbf{w}_t), \tag{36}$$

and the model error and its worst-case gradient magnitude on the trust region are

$$E_t := f_t - m_t, \qquad \varepsilon_t := \sup_{\mathbf{z}\in\Omega_t} \|\nabla E_t(\mathbf{z})\|. \tag{37}$$

Note $E_t(\mathbf{0}) = 0$, since $f_t(\mathbf{0}) = m_t(\mathbf{0}) = 0$, and $\mathbf{p}_t - \mathbf{g}_t = \nabla E_t(\mathbf{0})$, so $\|\mathbf{p}_t - \mathbf{g}_t\| \le \varepsilon_t$. Because $m_t$ is determined by the probe design and the population responses, $\varepsilon_t$ is $\mathcal{F}_t^+$-measurable: it does not depend on the client sample.

**Residuals, slopes, uncertainties.** Fix the Armijo constant $c \in (0,1)$. For a displacement $\mathbf{z} \in \mathbb{R}^r$ and a direction $\mathbf{d} \in \mathbb{R}^r$ put

$$\mathbf{b}_t(\mathbf{z}) := \mathbf{a}_t(\mathbf{z}) - c\,\mathbf{G}_t^\top \mathbf{z}, \qquad \widehat{R}_t(\mathbf{z}) := \mathbf{b}_t(\mathbf{z})^\top \bar{\mathbf{y}}_t, \qquad R_t(\mathbf{z}) := \mathbf{b}_t(\mathbf{z})^\top \boldsymbol{\mu}_t, \tag{38}$$

so that $\mathbf{b}_t(\mathbf{0}) = \mathbf{0}$ and $\widehat{R}_t(\eta\mathbf{d}) = \widehat{m}_t(\eta\mathbf{d}) - c\eta\,\widehat{\mathbf{g}}_t^\top \mathbf{d}$, and for $c' \in (0, c]$ define the true residual

$$R_t^{\mathrm{true},c'}(\mathbf{z}) := f_t(\mathbf{z}) - c'\,\mathbf{p}_t^\top \mathbf{z}. \tag{39}$$

Write $s_t = s_t(\mathbf{d}_t)$ and $\widehat{s}_t = \widehat{s}_t(\mathbf{d}_t)$ when the direction argument is omitted. The slopes and uncertainty scales are

$$\begin{aligned} \widehat{s}_t(\mathbf{d}) &= -\widehat{\mathbf{g}}_t^\top \mathbf{d}, & s_t(\mathbf{d}) &= -\mathbf{g}_t^\top \mathbf{d}, \\ \sigma_{R,t}(\mathbf{z}) &= \|\mathbf{C}_t^{1/2}\mathbf{b}_t(\mathbf{z})\|, & \sigma_{d,t}(\mathbf{d}) &= \|\mathbf{C}_t^{1/2}\mathbf{G}_t^\top \mathbf{d}\|. \end{aligned} \tag{40}$$

**Acceptance rule.** A candidate $(\eta, \mathbf{d})$ with $\eta > 0$ and $\|\eta\mathbf{d}\| \le \Delta_t$ is accepted only if it passes, in order,

$$\textbf{descent:} \quad \widehat{\mathbf{g}}_t^\top \mathbf{d} \le 0, \tag{41}$$

$$\textbf{margin:} \quad \underline{s}_t := \widehat{s}_t(\mathbf{d}) - \kappa_t\,\sigma_{d,t}(\mathbf{d}) > 0, \tag{42}$$

$$\textbf{Armijo UCB:} \quad Q_t(\eta) := \widehat{R}_t(\eta\mathbf{d}) + \kappa_t\,\sigma_{R,t}(\eta\mathbf{d}) \le 0. \tag{43}$$

These are exactly the tests of Algorithm 2, and at full participation $\mathbf{C}_t = \mathbf{0}$ reduces them to the tests of Algorithm 1. Model accuracy enters the analysis only through the trust-region constraint $\|\eta\mathbf{d}\| \le \Delta_t$, which caps the loss a model error can cause in one round (Section C.7). A round is *accepted* if its selected candidate passes equation 41–equation 43, and *rejected* otherwise. On a rejected round we consider two rules: the *null step* $\mathbf{w}_{t+1} = \mathbf{w}_t$, and the diminishing fallback of Algorithm 2. Under either rule the realized displacement $\mathbf{z}_t := \eta_t\mathbf{d}_t$ satisfies $\|\mathbf{z}_t\| \le \Delta_t$.

The candidate set searched in round $t$ is $\mathcal{C}_t$, the selected pair is $(\eta_t, \mathbf{d}_t) \in \mathcal{C}_t$, and

$$\mathcal{V}_t := \operatorname{span}\Big(\{\mathbf{b}_t(\eta\mathbf{d}) : (\eta,\mathbf{d}) \in \mathcal{C}_t\} \cup \{\mathbf{G}_t^\top \mathbf{d} : (\eta,\mathbf{d}) \in \mathcal{C}_t\}\Big) \subseteq \mathbb{R}^K \tag{44}$$

is the span of the coefficient vectors the tests evaluate. Both $\mathcal{C}_t$ and $\mathcal{V}_t$ are permitted to depend on $\bar{\mathbf{y}}_t$; this is the central difficulty the analysis must accommodate, since Algorithm 2 builds $\mathbf{d}_t$ from $\widehat{\mathbf{g}}_t$ and $\widehat{\mathbf{H}}_t$, caps the grid at $\eta_{\max,t} = \min\{\eta_0, \Delta_t/\|\mathbf{d}_t\|\}$, and selects $\eta_t = \max \mathcal{H}_t^{\mathrm{acc}}$, all from the same responses the tests use.

### C.2 Assumptions

**Assumption 1** (Client sampling)**.** *The $m \ge 2$ participating clients are drawn uniformly without replacement from the $N$ clients, so that $\mathbb{E}[\bar{\mathbf{y}}_t \mid \mathcal{F}_t^+] = \boldsymbol{\mu}_t$ and*

$$\mathbf{C}_t = \frac{N-m}{Nm}\,\boldsymbol{\Sigma}_t, \qquad \boldsymbol{\Sigma}_t := \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{y}_{i,t} - \boldsymbol{\mu}_t)(\mathbf{y}_{i,t} - \boldsymbol{\mu}_t)^\top. \tag{45}$$

*No distributional form is assumed beyond the existence of this second moment. Full participation is the case $m = N$, for which $\mathbf{C}_t = \mathbf{0}$.*

**Assumption 2** (Probabilistic model accuracy). *$f_t \in C^1(\Omega_t)$ and the kernel is $C^1$ in its second argument, so that $E_t \in C^1(\Omega_t)$. There are an $\mathcal{F}_t$-measurable accuracy level $\bar{\varepsilon}_t \geq 0$ and an $\mathcal{F}_t$-measurable $p_{\mathrm{acc},t} \in [0,1]$ such that*

$$\mathcal{E}_t^{\mathrm{acc}} := \{\varepsilon_t \leq \bar{\varepsilon}_t\} \in \mathcal{F}_t^+ \qquad \textit{satisfies} \qquad \Pr\left(\mathcal{E}_t^{\mathrm{acc}} \mid \mathcal{F}_t\right) \geq p_{\mathrm{acc},t}. \tag{46}$$

*The only randomness in $\mathcal{E}_t^{\mathrm{acc}}$ is the probe design.*

**Remark 1** (Verifying Assumption 2). *For the polynomial kernels used here, Lemma 10 gives $\varepsilon_t \leq \kappa_{eg}\Delta_t^p$ deterministically on a poised probe set, so $\bar{\varepsilon}_t = \kappa_{eg}\Delta_t^p$ and $p_{\mathrm{acc},t}$ is the probability that the drawn probes are poised with constant at most $\Lambda$, which the server can force to one by checking the design before broadcasting. Randomly drawn probes are poised with probability bounded below (Bandeira et al., 2014). The abstract form equation 46 also covers kernels whose native space is infinite dimensional, where the accuracy level is driven by the probe budget rather than by the radius.*

**Assumption 3** (Lower bound and smoothness). *$F \geq F^\star > -\infty$, and $\nabla f_t$ is L-Lipschitz on $\Omega_t$ for every $t$. The latter holds whenever $F$ is L-smooth, because $\mathbf{V}_t$ has orthonormal columns. Moreover $\mathbb{E}|F(\mathbf{w}_t)| < \infty$ and $\mathbb{E}\left[\|\mathbf{p}_t\|\Delta_t + \Delta_t^2 + \bar{\varepsilon}_t\Delta_t\right] < \infty$ for every $t$.*

Assumption 3 does not require $F$ to be nonincreasing on unsuccessful rounds. That would fail for the diminishing fallback, and it would also fail on accepted rounds where the sampling or the accuracy event does not hold; such falsely certified rounds are observed in the experiments (the false-positive column of Table 3). Theorem 2 accounts for them through an explicit increase budget.

### C.3 Two elementary lemmas

**Lemma 2** (Sampling gap). *For every $\mathbf{z} \in \mathbb{R}^r$ and $\mathbf{d} \in \mathbb{R}^r$, random or deterministic,*

$$\widehat{R}_t(\mathbf{z}) - R_t(\mathbf{z}) = \mathbf{b}_t(\mathbf{z})^\top \boldsymbol{e}_t, \qquad \widehat{s}_t(\mathbf{d}) - s_t(\mathbf{d}) = -(\mathbf{G}_t^\top \mathbf{d})^\top \boldsymbol{e}_t. \tag{47}$$

*Proof.* Subtracting the two definitions in equation 38 gives $\widehat{R}_t(\mathbf{z}) - R_t(\mathbf{z}) = \mathbf{b}_t(\mathbf{z})^\top(\bar{\mathbf{y}}_t - \boldsymbol{\mu}_t) = \mathbf{b}_t(\mathbf{z})^\top \boldsymbol{e}_t$. By equation 35, $\widehat{\mathbf{g}}_t - \mathbf{g}_t = \mathbf{G}_t \boldsymbol{e}_t$, so $\widehat{s}_t(\mathbf{d}) - s_t(\mathbf{d}) = -(\widehat{\mathbf{g}}_t - \mathbf{g}_t)^\top \mathbf{d} = -(\mathbf{G}_t^\top \mathbf{d})^\top \boldsymbol{e}_t$. Both are algebraic identities holding pointwise in the sample space, so no measurability of $\mathbf{z}$ or $\mathbf{d}$ is used. This is what allows $\mathbf{z}$ and $\mathbf{d}$ to be chosen after observing $\bar{\mathbf{y}}_t$. □

**Lemma 3** (Support of the sampling error and whitening). *Let $\mathbf{\Pi}_{C,t}$ denote the orthogonal projection onto $\operatorname{range}\mathbf{C}_t$ and $(\mathbf{C}_t^\dagger)^{1/2}$ the square root of the Moore–Penrose pseudo-inverse of $\mathbf{C}_t$. Then, conditionally on $\mathcal{F}_t^+$,*

$$\boldsymbol{e}_t \in \operatorname{range}\mathbf{C}_t \text{ a.s.}, \qquad \widetilde{\boldsymbol{e}}_t := (\mathbf{C}_t^\dagger)^{1/2}\boldsymbol{e}_t \text{ satisfies } \boldsymbol{e}_t = \mathbf{C}_t^{1/2}\widetilde{\boldsymbol{e}}_t,\ \mathbb{E}[\widetilde{\boldsymbol{e}}_t] = \mathbf{0},\ \operatorname{Cov}(\widetilde{\boldsymbol{e}}_t) = \mathbf{\Pi}_{C,t}. \tag{48}$$

*Moreover* $\operatorname{rank}\mathbf{C}_t \leq \min\{K,\, N-1\}$.

*Proof.* Let $\mathbf{u} \perp \operatorname{range}\mathbf{C}_t$. Then $\mathbb{E}[\mathbf{u}^\top \boldsymbol{e}_t] = 0$ by Assumption 1 and $\operatorname{Var}(\mathbf{u}^\top \boldsymbol{e}_t) = \mathbf{u}^\top \mathbf{C}_t \mathbf{u} = 0$, so $\mathbf{u}^\top \boldsymbol{e}_t = 0$ almost surely. Applying this to each member of a finite orthonormal basis of $(\operatorname{range}\mathbf{C}_t)^\perp$ and intersecting the finitely many almost-sure events gives $\mathbf{\Pi}_{C,t}\boldsymbol{e}_t = \boldsymbol{e}_t$ almost surely. Since $\mathbf{C}_t^{1/2}(\mathbf{C}_t^\dagger)^{1/2} = \mathbf{\Pi}_{C,t}$, this yields $\mathbf{C}_t^{1/2}\widetilde{\boldsymbol{e}}_t = \mathbf{\Pi}_{C,t}\boldsymbol{e}_t = \boldsymbol{e}_t$. Linearity of the pseudo-inverse root gives $\mathbb{E}[\widetilde{\boldsymbol{e}}_t] = \mathbf{0}$ and $\operatorname{Cov}(\widetilde{\boldsymbol{e}}_t) = (\mathbf{C}_t^\dagger)^{1/2}\mathbf{C}_t(\mathbf{C}_t^\dagger)^{1/2} = \mathbf{\Pi}_{C,t}$.

For the rank bound, $\mathbf{C}_t$ is a $K \times K$ matrix. The $N$ centered response vectors sum to zero, so their span has dimension at most $N-1$. Since $\mathbf{C}_t = f_m \mathbf{\Sigma}_t$, this gives $\operatorname{rank}\mathbf{C}_t \leq \min\{K, N-1\}$. □

### C.4 Admissible effective dimensions

The next definition isolates the single property of the candidate set that the sampling argument requires.

**Definition 1** (Admissible effective dimension). *A nonnegative $\mathcal{F}_t^+$-measurable random variable $d_t^{\mathrm{eff}}$ is an* admissible effective dimension *for round $t$ if there exists a nonnegative random variable $W_t$ with*

$$\mathbb{E}\left[W_t^2 \mid \mathcal{F}_t^+\right] \leq d_t^{\mathrm{eff}} \tag{49}$$

*such that, almost surely and simultaneously for every* $\mathbf{v} \in \mathcal{V}_t$,

$$\left|\mathbf{v}^\top \boldsymbol{e}_t\right| \;\le\; W_t \left\|\mathbf{C}_t^{1/2}\mathbf{v}\right\|. \tag{50}$$

The force of equation 50 is that one scalar $W_t$, not depending on the candidate, controls every coefficient vector the tests evaluate, and does so on precisely the scale $\|\mathbf{C}_t^{1/2}\mathbf{v}\|$ that appears in the acceptance rule equation 42–equation 43. This is what permits $\mathcal{V}_t$, and hence the selected candidate, to be chosen after seeing $\bar{\mathbf{y}}_t$.

Both instances below follow from one observation: it suffices to enclose the data-dependent span $\mathcal{V}_t$ in a subspace that is fixed before the clients are sampled.

**Lemma 4** (Measurable envelope)**.** *Let* $\mathcal{U}_t \subseteq \mathbb{R}^K$ *be an* $\mathcal{F}_t^+$*-measurable subspace with* $\mathcal{V}_t \subseteq \mathcal{U}_t$ *almost surely, and let* $\mathbf{\Pi}_{\mathcal{U},t}$ *be the orthogonal projection onto* $\mathbf{C}_t^{1/2}\mathcal{U}_t$*. Then*

$$d_t^{\mathcal{U}} := \dim\left(\mathbf{C}_t^{1/2}\mathcal{U}_t\right) \le \min\{\dim \mathcal{U}_t,\ \operatorname{rank}\mathbf{C}_t\} \tag{51}$$

*is admissible, with witness* $W_t = \|\mathbf{\Pi}_{\mathcal{U},t}\widetilde{\boldsymbol{e}}_t\|$ *and* $\mathbb{E}[W_t^2 \mid \mathcal{F}_t^+] = d_t^{\mathcal{U}}$ *exactly.*

*Proof.* Let $\mathbf{v} \in \mathcal{U}_t$. By Lemma 3, $\boldsymbol{e}_t = \mathbf{C}_t^{1/2}\widetilde{\boldsymbol{e}}_t$ almost surely, and $\mathbf{C}_t^{1/2}\mathbf{v}$ lies in the range of the symmetric projection $\mathbf{\Pi}_{\mathcal{U},t}$, so

$$|\mathbf{v}^\top \boldsymbol{e}_t| = \left|(\mathbf{C}_t^{1/2}\mathbf{v})^\top \widetilde{\boldsymbol{e}}_t\right| = \left|(\mathbf{C}_t^{1/2}\mathbf{v})^\top \mathbf{\Pi}_{\mathcal{U},t}\widetilde{\boldsymbol{e}}_t\right| \le \|\mathbf{C}_t^{1/2}\mathbf{v}\|\, W_t$$

by Cauchy–Schwarz. The inequality holds pathwise on one almost-sure event and for all $\mathbf{v} \in \mathcal{U}_t$ at once, hence for every $\mathbf{v} \in \mathcal{V}_t$, including vectors that depend on $\boldsymbol{e}_t$. For the moment, $\mathbf{\Pi}_{\mathcal{U},t}$ is $\mathcal{F}_t^+$-measurable and its range lies in $\operatorname{range}\mathbf{C}_t$, so $\mathbf{\Pi}_{C,t}\mathbf{\Pi}_{\mathcal{U},t} = \mathbf{\Pi}_{\mathcal{U},t}$ and

$$\mathbb{E}[W_t^2 \mid \mathcal{F}_t^+] = \operatorname{tr}\left(\mathbf{\Pi}_{\mathcal{U},t}\mathbf{\Pi}_{C,t}\mathbf{\Pi}_{\mathcal{U},t}\right) = \operatorname{tr}\mathbf{\Pi}_{\mathcal{U},t} = \dim\left(\mathbf{C}_t^{1/2}\mathcal{U}_t\right).$$

The bound in equation 51 holds because a linear image cannot increase dimension and $\mathbf{C}_t^{1/2}\mathcal{U}_t \subseteq \operatorname{range}\mathbf{C}_t$. □

**Lemma 5** (Admissible effective dimensions)**.** *Each of the following is admissible in the sense of Definition 1, and so is their minimum.*

(i) (Rank.) $d_t^{\mathrm{rank}} := \operatorname{rank}\mathbf{C}_t \le \min\{K, N-1\}$*, with* $W_t = \|\widetilde{\boldsymbol{e}}_t\|$*. No further hypothesis is required.*

(ii) (Feature dimension.) *Suppose there are functions* $\phi_1, \dots, \phi_D : \mathbb{R}^r \to \mathbb{R}$*, differentiable at the origin with* $\phi_j(\mathbf{0}) = 0$*, and vectors* $\mathbf{k}_{t,1}, \dots, \mathbf{k}_{t,D}$ *in* $\mathbb{R}^{K+1}$ *that are* $\mathcal{F}_t^+$*-measurable, such that*

$$\mathbf{k}_t(\mathbf{z}) - \mathbf{k}_t(\mathbf{0}) = \sum_{j=1}^{D} \phi_j(\mathbf{z})\,\mathbf{k}_{t,j} \qquad \textit{for all } \mathbf{z} \in \mathbb{R}^r. \tag{52}$$

*Put* $\mathbf{h}_{t,j} := [\mathbf{A}_t^{-1}\mathbf{k}_{t,j}]_{1:K}$ *and* $\mathcal{V}_t^{\mathrm{feat}} := \operatorname{span}\{\mathbf{h}_{t,j} : 1 \le j \le D\}$*. Then*

$$d_t^{\mathrm{feat}} := \dim\left(\mathbf{C}_t^{1/2}\mathcal{V}_t^{\mathrm{feat}}\right) \le \min\{D,\ \operatorname{rank}\mathbf{C}_t\} \tag{53}$$

*is admissible for* every *candidate set* $\mathcal{C}_t$*, including directions and grids chosen after observing* $\bar{\mathbf{y}}_t$*. For the polynomial kernel* $k(\mathbf{z}', \mathbf{z}) = (\gamma\,\mathbf{z}'^\top\mathbf{z} + c_0)^p$ *the* $\phi_j$ *are the monomials of degree* $1, \dots, p$ *and* $D = \binom{r+p}{p} - 1$*, which is* $9$ *for the cubic kernel with* $r = 2$*.*

*Proof. (i)* Apply Lemma 4 with $\mathcal{U}_t = \mathbb{R}^K$. Then $\mathbf{C}_t^{1/2}\mathcal{U}_t = \operatorname{range}\mathbf{C}_t$, $\mathbf{\Pi}_{\mathcal{U},t} = \mathbf{\Pi}_{C,t}$ and $W_t = \|\mathbf{\Pi}_{C,t}\widetilde{\boldsymbol{e}}_t\| = \|\widetilde{\boldsymbol{e}}_t\|$. The bound on the rank is Lemma 3.

*(ii)* Substituting equation 52 into equation 20 gives $\mathbf{a}_t(\mathbf{z}) = \sum_{j=1}^{D}\phi_j(\mathbf{z})\mathbf{h}_{t,j} \in \mathcal{V}_t^{\mathrm{feat}}$ for every $\mathbf{z}$. Differentiating at the origin, $\mathbf{G}_t^\top\mathbf{d} = \nabla_{\mathbf{z}}\mathbf{a}_t(\mathbf{0})\,\mathbf{d} = \sum_{j=1}^{D}\left(\nabla\phi_j(\mathbf{0})^\top\mathbf{d}\right)\mathbf{h}_{t,j} \in \mathcal{V}_t^{\mathrm{feat}}$ for every $\mathbf{d}$. Hence $\mathbf{b}_t(\eta\mathbf{d}) \in \mathcal{V}_t^{\mathrm{feat}}$ and $\mathbf{G}_t^\top\mathbf{d} \in \mathcal{V}_t^{\mathrm{feat}}$ for *all* $(\eta, \mathbf{d})$, so $\mathcal{V}_t \subseteq \mathcal{V}_t^{\mathrm{feat}}$ whatever the candidate set is. The subspace $\mathcal{V}_t^{\mathrm{feat}}$ is $\mathcal{F}_t^+$-measurable because $\mathbf{A}_t$ and the $\mathbf{k}_{t,j}$ are, and Lemma 4 applies. For the

polynomial kernel, the multinomial expansion $(\gamma\,\mathbf{z}'^{\top}\mathbf{z}+c_0)^p=\sum_{|\alpha|\le p}c_\alpha\,\mathbf{z}'^{\alpha}\mathbf{z}^{\alpha}$ gives equation 52 with $\phi_\alpha(\mathbf{z})=\mathbf{z}^\alpha$ and $\mathbf{k}_{t,\alpha}=c_\alpha[(\mathbf{z}_t^k)^\alpha]_{k=0}^K$ for $1\le|\alpha|\le p$; there are $\binom{r+p}{p}-1$ such multi-indices.

The closure claim holds because admissibility is witnessed separately by each $W_t$. The index of the smaller bound is $\mathcal{F}_t^+$-measurable, so the witness attached to it is a valid witness for the minimum. □

Instance (ii) is the one that applies to Algorithm 2 as implemented: it needs no restriction on how $\mathbf{d}_t$ is built and it improves on the rank whenever $K>D$.

### C.5 The sampling guarantee at a free multiplier

**Proposition 2** (Simultaneous sampling guarantee)**.** *Let Assumption 1 hold and let $d_t^{\mathrm{eff}}$ be admissible. Fix any $\kappa_t>0$ and run the tests equation 42–equation 43 with this multiplier. Define the certified coverage*

$$\rho_{\mathrm{cov},t}(\kappa_t):=\Big[1-\frac{d_t^{\mathrm{eff}}}{\kappa_t^2}\Big]_+. \tag{54}$$

*Then there is an event $\mathcal{E}_t^{\mathrm{smp}}$ with $\Pr(\mathcal{E}_t^{\mathrm{smp}}\mid\mathcal{F}_t^+)\ge\rho_{\mathrm{cov},t}(\kappa_t)$ on which, simultaneously for every $(\eta,\mathbf{d})\in\mathcal{C}_t$,*

$$R_t(\eta\mathbf{d})\le\widehat{R}_t(\eta\mathbf{d})+\kappa_t\,\sigma_{R,t}(\eta\mathbf{d}),\qquad s_t(\mathbf{d})\ge\widehat{s}_t(\mathbf{d})-\kappa_t\,\sigma_{d,t}(\mathbf{d}). \tag{55}$$

*In particular, on $\mathcal{E}_t^{\mathrm{smp}}$, if the selected candidate $(\eta_t,\mathbf{d}_t)$ passes equation 42–equation 43, then*

$$R_t(\eta_t\mathbf{d}_t)\le 0,\qquad s_t(\mathbf{d}_t)\ge\underline{s}_t>0, \tag{56}$$

*even though $(\eta_t,\mathbf{d}_t)$ is selected after observing $\bar{\mathbf{y}}_t$. At full participation $\mathbf{C}_t=\mathbf{0}$, so $d_t^{\mathrm{eff}}=0$ and the coverage is one for every $\kappa_t>0$.*

*Proof. Step 1: a single scalar event.* Let $W_t$ witness admissibility and set $\mathcal{E}_t^{\mathrm{smp}}:=\{W_t\le\kappa_t\}$. This event is defined by one scalar inequality and references no candidate.

*Step 2: the two bounds on that event.* On $\mathcal{E}_t^{\mathrm{smp}}$, equation 50 gives $|\mathbf{v}^\top\boldsymbol{e}_t|\le\kappa_t\|\mathbf{C}_t^{1/2}\mathbf{v}\|$ for every $\mathbf{v}\in\mathcal{V}_t$ at once. Taking $\mathbf{v}=\mathbf{b}_t(\eta\mathbf{d})\in\mathcal{V}_t$ and using Lemma 2 together with $\sigma_{R,t}(\eta\mathbf{d})=\|\mathbf{C}_t^{1/2}\mathbf{b}_t(\eta\mathbf{d})\|$,

$$R_t(\eta\mathbf{d})=\widehat{R}_t(\eta\mathbf{d})-\mathbf{b}_t(\eta\mathbf{d})^\top\boldsymbol{e}_t\le\widehat{R}_t(\eta\mathbf{d})+\kappa_t\,\sigma_{R,t}(\eta\mathbf{d}).$$

Taking $\mathbf{v}=\mathbf{G}_t^\top\mathbf{d}\in\mathcal{V}_t$ and using Lemma 2 again,

$$s_t(\mathbf{d})=\widehat{s}_t(\mathbf{d})+(\mathbf{G}_t^\top\mathbf{d})^\top\boldsymbol{e}_t\ge\widehat{s}_t(\mathbf{d})-\kappa_t\,\sigma_{d,t}(\mathbf{d}).$$

This is equation 55, holding for all candidates simultaneously because the event does not depend on which candidate is considered.

*Step 3: the coverage.* Since $W_t\ge0$, Markov's inequality applied to $W_t^2$ and then equation 49 give

$$\Pr\big((\mathcal{E}_t^{\mathrm{smp}})^{\mathrm{c}}\mid\mathcal{F}_t^+\big)=\Pr\big(W_t^2>\kappa_t^2\mid\mathcal{F}_t^+\big)\le\frac{\mathbb{E}[W_t^2\mid\mathcal{F}_t^+]}{\kappa_t^2}\le\frac{d_t^{\mathrm{eff}}}{\kappa_t^2}, \tag{57}$$

so $\Pr(\mathcal{E}_t^{\mathrm{smp}}\mid\mathcal{F}_t^+)\ge[1-d_t^{\mathrm{eff}}/\kappa_t^2]_+$ since probabilities are nonnegative.

*Step 4: the conclusions.* Suppose the selected candidate passes the tests, so $\underline{s}_t>0$ and $Q_t(\eta_t)\le0$. Applying equation 55 at $(\eta_t,\mathbf{d}_t)$, which is legitimate because Step 2 covers every element of $\mathcal{C}_t$,

$$R_t(\eta_t\mathbf{d}_t)\le\widehat{R}_t(\eta_t\mathbf{d}_t)+\kappa_t\,\sigma_{R,t}(\eta_t\mathbf{d}_t)=Q_t(\eta_t)\le0,$$

and $s_t\ge\widehat{s}_t-\kappa_t\sigma_{d,t}(\mathbf{d}_t)=\underline{s}_t>0$. □

**Corollary 1** (Certification floor)**.** *$\rho_{\mathrm{cov},t}(\kappa_t)>0$ if and only if $\kappa_t>\sqrt{d_t^{\mathrm{eff}}}$, and $\rho_{\mathrm{cov},t}$ is strictly increasing on $(\sqrt{d_t^{\mathrm{eff}}},\infty)$ with $\rho_{\mathrm{cov},t}(\kappa_t)\to1$ as $\kappa_t\to\infty$. Choosing $\kappa_t=\sqrt{d_t^{\mathrm{eff}}/(1-\rho)}$ returns $\rho_{\mathrm{cov},t}=\rho$.*

*Proof.* The map $\kappa_t \mapsto 1 - d_t^{\text{eff}} \kappa_t^{-2}$ is strictly increasing on $(0, \infty)$, vanishes exactly at $\kappa_t = \sqrt{d_t^{\text{eff}}}$ and tends to 1; truncation at zero preserves these statements. The last claim is the identity $1 - d_t^{\text{eff}} / \big(d_t^{\text{eff}}/(1-\rho)\big) = \rho$. □

**Remark 2** (Relation to the per-candidate calibration)**.** *For a candidate fixed before the sample is drawn, one may instead apply Cantelli's inequality to the scalar $\mathbf{b}_t(\mathbf{z})^\top \boldsymbol{e}_t$, which has conditional mean zero and variance $\sigma_{R,t}(\mathbf{z})^2$, obtaining coverage $\rho$ at the smaller multiplier $\sqrt{\rho/(1-\rho)}$. That argument is unavailable here because the coefficient vector depends on $\boldsymbol{e}_t$, so the conditional mean is not zero; Proposition 2 replaces the mean-zero step by the pathwise bound equation 50. By Corollary 1 the price is the floor $\kappa_t > \sqrt{d_t^{\text{eff}}}$, which no choice of target probability removes.*

**Remark 3** (The floor is intrinsic; Scheffé calibration)**.** *For instance (i) the witness is optimal: since $\mathbf{C}_t^{1/2}\mathbf{v}$ ranges over* $\operatorname{range} \mathbf{C}_t \ni \widetilde{\boldsymbol{e}}_t$*,*

$$\sup_{\mathbf{v}:\, \mathbf{C}_t^{1/2}\mathbf{v} \neq \mathbf{0}} \frac{|\mathbf{v}^\top \boldsymbol{e}_t|}{\|\mathbf{C}_t^{1/2}\mathbf{v}\|} = \|\widetilde{\boldsymbol{e}}_t\|, \qquad \mathbb{E}\big[\|\widetilde{\boldsymbol{e}}_t\|^2 \mid \mathcal{F}_t^+\big] = \operatorname{rank} \mathbf{C}_t.$$

*No other choice of $W_t$ can control all of $\mathbb{R}^K$ with a smaller second moment, so the floor can only be lowered by shrinking the envelope, which is what instance (ii) does. The event $\{W_t \le \kappa_t\}$ is the distribution-free counterpart of Scheffé's simultaneous confidence region for all linear contrasts (Scheffé, 1953). Under a finite-population central limit theorem $W_t^2$ is approximately $\chi^2$ with $d_t^{\text{eff}}$ degrees of freedom, which suggests the calibration $\kappa_t^2 = \chi^2_{d_t^{\text{eff}},1-\alpha}$. For $d_t^{\text{eff}} = 3$ and $\alpha = 0.1$ this is $\kappa_t \approx 2.50$, against $\sqrt{30} \approx 5.48$ from Corollary 1. This calibration is asymptotic in $m$ and is not a finite-sample guarantee.*

### C.6 Feasibility of the certified step-size set

**Lemma 6** (Certified step-size interval)**.** *Fix $\mathbf{d}_t \neq \mathbf{0}$, set $\eta_{\max,t} = \min\{\eta_0, \Delta_t/\|\mathbf{d}_t\|\}$, and suppose $\boldsymbol{\beta}_t(\eta) := \mathbf{b}_t(\eta \mathbf{d}_t) \in \mathbb{R}^K$ is twice continuously differentiable on $[0, \eta_{\max,t}]$, which holds whenever the kernel is $C^2$ in its second argument and in particular for polynomial kernels. Put*

$$B_t^R := \sup_{0 \le \eta \le \eta_{\max,t}} \big|\boldsymbol{\beta}_t''(\eta)^\top \bar{\mathbf{y}}_t\big|, \qquad B_t^\sigma := \sup_{0 \le \eta \le \eta_{\max,t}} \big\|\mathbf{C}_t^{1/2} \boldsymbol{\beta}_t''(\eta)\big\|, \qquad B_t^Q := B_t^R + \kappa_t B_t^\sigma. \tag{58}$$

*Then for every $\eta \in [0, \eta_{\max,t}]$,*

$$Q_t(\eta) \;\le\; -(1-c)\, \underline{s}_t\, \eta + \tfrac{1}{2} B_t^Q \eta^2. \tag{59}$$

*Consequently, if the margin test equation 42 holds then $Q_t(\eta) \le 0$ for every $\eta$ in the nonempty interval $(0, \bar{\eta}_t]$ with*

$$\bar{\eta}_t := \min\Big\{\eta_{\max,t},\; \frac{2(1-c)\underline{s}_t}{B_t^Q}\Big\} > 0, \tag{60}$$

*where the second entry is read as $+\infty$ when $B_t^Q = 0$. Any grid $\mathcal{H}_t$ containing a point of $(0, \bar{\eta}_t]$ therefore has a nonempty certified set.*

*Proof.* No differentiability of the norm $\sigma_{R,t}$ is needed; the argument expands the vector $\boldsymbol{\beta}_t$ and then uses the triangle inequality.

*Step 1: first-order expansion of the coefficient vector.* By equation 38, $\boldsymbol{\beta}_t(0) = \mathbf{b}_t(\mathbf{0}) = \mathbf{0}$. By the chain rule and $\nabla_{\mathbf{z}} \mathbf{a}_t(\mathbf{0}) = \mathbf{G}_t^\top$,

$$\boldsymbol{\beta}_t'(0) = \nabla_{\mathbf{z}} \mathbf{a}_t(\mathbf{0})\, \mathbf{d}_t - c\, \mathbf{G}_t^\top \mathbf{d}_t = (1-c)\, \mathbf{G}_t^\top \mathbf{d}_t. \tag{61}$$

Taylor's formula with integral remainder gives $\boldsymbol{\beta}_t(\eta) = \eta\, \boldsymbol{\beta}_t'(0) + \mathbf{r}_t(\eta)$ with $\mathbf{r}_t(\eta) = \int_0^\eta (\eta - u)\, \boldsymbol{\beta}_t''(u)\, du$.

*Step 2: the residual.* Since $\widehat{R}_t(\eta \mathbf{d}_t) = \boldsymbol{\beta}_t(\eta)^\top \bar{\mathbf{y}}_t$ and $(\mathbf{G}_t^\top \mathbf{d}_t)^\top \bar{\mathbf{y}}_t = \mathbf{d}_t^\top \widehat{\mathbf{g}}_t = -\widehat{s}_t(\mathbf{d}_t)$,

$$\widehat{R}_t(\eta \mathbf{d}_t) = -(1-c)\, \eta\, \widehat{s}_t(\mathbf{d}_t) + \mathbf{r}_t(\eta)^\top \bar{\mathbf{y}}_t \le -(1-c)\, \eta\, \widehat{s}_t(\mathbf{d}_t) + \tfrac{1}{2} B_t^R \eta^2, \tag{62}$$

because $|\mathbf{r}_t(\eta)^\top \bar{\mathbf{y}}_t| \le \int_0^\eta (\eta - u) B_t^R\, du = \frac{1}{2} B_t^R \eta^2$.

*Step 3: the uncertainty scale.* By the triangle inequality in $\mathbb{R}^K$,

$$\sigma_{R,t}(\eta \mathbf{d}_t) = \big\|\mathbf{C}_t^{1/2}\boldsymbol{\beta}_t(\eta)\big\| \le \eta(1-c)\big\|\mathbf{C}_t^{1/2}\mathbf{G}_t^\top \mathbf{d}_t\big\| + \big\|\mathbf{C}_t^{1/2}\mathbf{r}_t(\eta)\big\| \le (1-c)\,\eta\,\sigma_{d,t}(\mathbf{d}_t) + \tfrac{1}{2}B_t^\sigma \eta^2. \tag{63}$$

*Step 4: combine.* Adding equation 62 and $\kappa_t$ times equation 63, and recalling $\underline{s}_t = \widehat{s}_t(\mathbf{d}_t) - \kappa_t\sigma_{d,t}(\mathbf{d}_t)$,

$$Q_t(\eta) \le -(1-c)\,\eta\,\big(\widehat{s}_t(\mathbf{d}_t) - \kappa_t\sigma_{d,t}(\mathbf{d}_t)\big) + \tfrac{1}{2}\big(B_t^R + \kappa_t B_t^\sigma\big)\eta^2,$$

which is equation 59. Writing the right-hand side as $\eta\big[\frac{1}{2}B_t^Q\eta - (1-c)\underline{s}_t\big]$, the bracket is affine and nondecreasing in $\eta$, negative at $\eta = 0$ when $\underline{s}_t > 0$, and nonpositive exactly for $\eta \le 2(1-c)\underline{s}_t/B_t^Q$. Intersecting with the cap $\eta_{\max,t}$ gives $Q_t(\eta) \le 0$ on $(0, \bar{\eta}_t]$, and $\bar{\eta}_t > 0$ because $\underline{s}_t > 0$ and $1 - c > 0$. □

### C.7 Transfer to the true objective

**Proposition 3** (True decrease). *Let $E_t \in C^1(\Omega_t)$ and suppose the selected step satisfies $\eta_t > 0$, $\|\eta_t\mathbf{d}_t\| \le \Delta_t$ and $R_t(\eta_t\mathbf{d}_t) \le 0$. Then*

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \;\le\; -\,c\,\eta_t\,s_t + \varepsilon_t\|\mathbf{z}_t\| \;=\; -\,\eta_t\big(c\,s_t - \varepsilon_t\|\mathbf{d}_t\|\big), \qquad \mathbf{z}_t := \eta_t\mathbf{d}_t. \tag{64}$$

*In particular $F(\mathbf{w}_{t+1}) < F(\mathbf{w}_t)$ whenever $\varepsilon_t\|\mathbf{d}_t\| < c\,s_t$. The statement is deterministic: it uses no probability and no hypothesis on how $(\eta_t, \mathbf{d}_t)$ was chosen.*

*Proof.* Write $\mathbf{z}_t := \eta_t\mathbf{d}_t$ and split $f_t = m_t + E_t$. By equation 38, $m_t(\mathbf{z}_t) = R_t(\mathbf{z}_t) + c\,\mathbf{g}_t^\top\mathbf{z}_t$, and $\mathbf{g}_t^\top\mathbf{z}_t = -\eta_t s_t$, so $R_t(\mathbf{z}_t) \le 0$ gives $m_t(\mathbf{z}_t) \le -c\,\eta_t s_t$. Because $\|\mathbf{z}_t\| \le \Delta_t$ and $\Omega_t$ is convex, the segment $\{\varsigma\mathbf{z}_t : \varsigma \in [0,1]\}$ lies in $\Omega_t$, where the gradient error is bounded by $\varepsilon_t$. With $E_t(\mathbf{0}) = 0$, the fundamental theorem of calculus and Cauchy–Schwarz give

$$E_t(\mathbf{z}_t) = \int_0^1 \nabla E_t(\varsigma\mathbf{z}_t)^\top\mathbf{z}_t\,d\varsigma \le \varepsilon_t\|\mathbf{z}_t\| = \eta_t\varepsilon_t\|\mathbf{d}_t\|. \tag{65}$$

Adding the two bounds, $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) = f_t(\mathbf{z}_t) \le -c\,\eta_t s_t + \eta_t\varepsilon_t\|\mathbf{d}_t\|$. □

The two forms of equation 64 play different roles. The sign of the second form does not depend on $\eta_t$: whether the round decreases $F$ is a property of the direction, namely $\varepsilon_t\|\mathbf{d}_t\| < c\,s_t$, and by Remark 4 no step length along a direction that violates it can be certified. The constraint $\|\mathbf{z}_t\| \le \Delta_t$ therefore cannot guarantee decrease in an individual round. What it does guarantee, via the first form, is that the loss caused by model error is at most $\varepsilon_t\Delta_t$ in any round. This is all the aggregate result needs, so no additional test is required. The true Armijo condition that follows from equation 64 is Theorem 3(ii).

**Corollary 2** (Decrease up to a model-error term). *On $\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}}$, every accepted round satisfies*

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \;\le\; -\,c\,\eta_t\,s_t + \varepsilon_t\|\mathbf{z}_t\| \;\le\; -\,c\,\eta_t\,\underline{s}_t + \bar{\varepsilon}_t\,\Delta_t. \tag{66}$$

*Proof.* On $\mathcal{E}_t^{\mathrm{smp}}$, Proposition 2 gives $R_t(\eta_t\mathbf{d}_t) \le 0$ and $s_t \ge \underline{s}_t > 0$, and acceptance gives $\eta_t > 0$ and $\|\mathbf{z}_t\| \le \Delta_t$, so Proposition 3 applies and gives the first inequality. On $\mathcal{E}_t^{\mathrm{acc}}$, $\varepsilon_t \le \bar{\varepsilon}_t$; with $\|\mathbf{z}_t\| \le \Delta_t$ and $s_t \ge \underline{s}_t$ this gives the second. □

**Remark 4** (Necessity of the margin condition). *The positive-margin condition cannot be weakened in its dependence on $\varepsilon_t$. Suppose $s_t \le \varepsilon_t\|\mathbf{d}_t\|$ and choose a model error whose center gradient points along the search direction, $\nabla E_t(\mathbf{0}) = \varepsilon_t\mathbf{d}_t/\|\mathbf{d}_t\|$, which has norm exactly $\varepsilon_t$ and is therefore consistent with equation 37. Then $\mathbf{p}_t^\top\mathbf{d}_t = \mathbf{g}_t^\top\mathbf{d}_t + \nabla E_t(\mathbf{0})^\top\mathbf{d}_t = -s_t + \varepsilon_t\|\mathbf{d}_t\| \ge 0$, so $\mathbf{d}_t$ is not a descent direction for $f_t$ and no positive $c'$ makes the true residual nonpositive for small $\eta$.*

### C.8 Per-round decrease

**Corollary 3** (Per-round decrease). *Let Assumptions 1 and 2 hold, let $d_t^{\mathrm{eff}}$ be admissible, and run the tests equation 41–equation 43 at multiplier $\kappa_t$. Put $\mathcal{G}_t := \{\text{round } t \text{ is accepted}\} \cap \mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}}$ and*

$$\delta_t := c\,\eta_t\,s_t, \qquad \nu_t := \varepsilon_t\,\|\mathbf{z}_t\|, \tag{67}$$

*the certified surrogate decrease and the model-error term, and set $\delta_t = \nu_t = 0$ off $\mathcal{G}_t$. Then*

$$\Pr\left(\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}} \mid \mathcal{F}_t\right) \;\geq\; q_t := \mathbb{E}\big[\rho_{\mathrm{cov},t}(\kappa_t)\,\mathbf{1}_{\mathcal{E}_t^{\mathrm{acc}}} \mid \mathcal{F}_t\big], \tag{68}$$

*and if $d_t^{\mathrm{eff}} \leq \bar{d}_t$ almost surely for an $\mathcal{F}_t$-measurable $\bar{d}_t$, then $q_t \geq [1 - \bar{d}_t/\kappa_t^2]_+\, p_{\mathrm{acc},t}$. On $\mathcal{G}_t$,*

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) - \delta_t + \nu_t, \qquad \delta_t \geq c\,\eta_t\,\underline{s}_t > 0, \qquad \nu_t \leq \bar{\varepsilon}_t\,\Delta_t. \tag{69}$$

*Proof.* The event $\mathcal{E}_t^{\mathrm{acc}}$ belongs to $\mathcal{F}_t^+$, so by the tower property and Proposition 2,

$$\Pr\left(\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}} \mid \mathcal{F}_t\right) = \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^{\mathrm{acc}}} \Pr(\mathcal{E}_t^{\mathrm{smp}} \mid \mathcal{F}_t^+) \mid \mathcal{F}_t\big] \geq \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_t^{\mathrm{acc}}}\,\rho_{\mathrm{cov},t}(\kappa_t) \mid \mathcal{F}_t\big].$$

If $d_t^{\mathrm{eff}} \leq \bar{d}_t$ then $\rho_{\mathrm{cov},t}(\kappa_t) \geq [1 - \bar{d}_t/\kappa_t^2]_+$, which is $\mathcal{F}_t$-measurable and factors out, leaving $\Pr(\mathcal{E}_t^{\mathrm{acc}} \mid \mathcal{F}_t) \geq p_{\mathrm{acc},t}$. The three statements in equation 69 are Corollary 2. □

The product bound is never worse than the union bound $[1-\bar{d}_t/\kappa_t^2]_+ + p_{\mathrm{acc},t} - 1$, since $ab \geq a+b-1$ on $[0,1]^2$. It is available because the accuracy event does not involve the client sample. Admissible choices are $\bar{d}_t = \min\{K, N-1\}$ and, under Lemma 5(ii), $\bar{d}_t = \min\{K, N-1, \binom{r+p}{p} - 1\}$. Neither bound controls the probability that a round is accepted.

### C.9 Conditional aggregate progress

**Lemma 7** (Bounded increase)**.** *Under Assumption 3, every realized displacement satisfies*

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \;\leq\; \iota_t := \|\mathbf{p}_t\|\,\ell_t + \tfrac{L}{2}\,\ell_t^2, \qquad \ell_t := \|\eta_t \mathbf{d}_t\| \leq \Delta_t. \tag{70}$$

*Proof.* Let $\mathbf{z}_t = \eta_t \mathbf{d}_t \in \Omega_t$. Since $f_t(\mathbf{0}) = 0$ and $\nabla f_t(\mathbf{0}) = \mathbf{p}_t$,

$$f_t(\mathbf{z}_t) = \mathbf{p}_t^\top \mathbf{z}_t + \int_0^1 \big(\nabla f_t(\varsigma \mathbf{z}_t) - \nabla f_t(\mathbf{0})\big)^\top \mathbf{z}_t\, d\varsigma \leq \|\mathbf{p}_t\|\,\|\mathbf{z}_t\| + \int_0^1 L\varsigma \|\mathbf{z}_t\|^2\, d\varsigma,$$

which is equation 70. For the fallback of Algorithm 2, $\ell_t = \eta_t \|\widehat{\mathbf{g}}_t\| \leq \Delta_t$ by the definition of $\eta_t$; for a null step $\ell_t = 0$. □

**Theorem 2** (Aggregate progress with explicit budgets)**.** *Let Assumption 3 hold together with the hypotheses of Corollary 3 at every round, and put $\Delta F_0 := F(\mathbf{w}_0) - F^\star$. Then for every $T \geq 1$,*

$$\sum_{t=0}^{T-1} \mathbb{E}\big[\delta_t \mathbf{1}_{\mathcal{G}_t}\big] \leq \Delta F_0 + I_T + M_T, \qquad \min_{0 \leq t < T} \mathbb{E}\big[\delta_t \mathbf{1}_{\mathcal{G}_t}\big] \leq \frac{\Delta F_0 + I_T + M_T}{T}, \tag{71}$$

*where the* increase budget *and the* model-error budget *are*

$$I_T := \sum_{t=0}^{T-1} \mathbb{E}\big[\iota_t \mathbf{1}_{\mathcal{G}_t^{\mathrm{c}}}\big], \qquad M_T := \sum_{t=0}^{T-1} \mathbb{E}\big[\nu_t \mathbf{1}_{\mathcal{G}_t}\big] \;\leq\; \sum_{t=0}^{T-1} \mathbb{E}\big[\bar{\varepsilon}_t\,\Delta_t\big], \tag{72}$$

*with $\iota_t$ from Lemma 7. If moreover $\delta_t \geq \underline{\delta}_t$ on $\mathcal{G}_t$ for some $\mathcal{F}_t$-measurable $\underline{\delta}_t \geq 0$, and $\Pr(\mathcal{G}_t \mid \mathcal{F}_t) \geq \pi_t$ for some $\mathcal{F}_t$-measurable $\pi_t$, then $\sum_{t<T} \mathbb{E}[\pi_t\,\underline{\delta}_t] \leq \Delta F_0 + I_T + M_T$.*

*Proof. Step 1: one round.* Fix $t$. On $\mathcal{G}_t$, Corollary 3 gives $F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \geq \delta_t - \nu_t$; on $\mathcal{G}_t^{\mathrm{c}}$, Lemma 7 gives $F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \geq -\iota_t$. The two cases combine into the single pathwise inequality

$$F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \;\geq\; (\delta_t - \nu_t)\mathbf{1}_{\mathcal{G}_t} - \iota_t \mathbf{1}_{\mathcal{G}_t^{\mathrm{c}}} \qquad \text{a.s.} \tag{73}$$

No measurability of $\delta_t$ with respect to $\mathcal{F}_t$ is needed here, because equation 73 is an inequality between random variables rather than between conditional expectations. This matters: $\eta_t$, $s_t$ and $\varepsilon_t$ all depend on the client sample or the probe design drawn in round $t$, so $\delta_t$ is not $\mathcal{F}_t$-measurable.

*Step 2: telescope.* All terms are integrable by Assumption 3. Taking expectations in equation 73 and summing over $t = 0, \ldots, T-1$,

$$\sum_{t=0}^{T-1} \mathbb{E}[\delta_t \mathbf{1}_{\mathcal{G}_t}] - M_T - I_T \leq \mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}_T)] \leq \Delta F_0,$$

the last step by $F(\mathbf{w}_T) \geq F^\star$. This is the first claim in equation 71, and the bound on $M_T$ in equation 72 is $\nu_t \leq \bar{\varepsilon}_t \Delta_t$ from equation 69. The summands are nonnegative, so their minimum is at most their average, which is the second claim.

*Step 3: the $\mathcal{F}_t$-measurable variant.* Since $\delta_t \mathbf{1}_{\mathcal{G}_t} \geq \underline{\delta}_t \mathbf{1}_{\mathcal{G}_t}$, $\mathbb{E}[\delta_t \mathbf{1}_{\mathcal{G}_t} \mid \mathcal{F}_t] \geq \underline{\delta}_t \Pr(\mathcal{G}_t \mid \mathcal{F}_t) \geq \pi_t \underline{\delta}_t$, and the tower property with Step 2 gives the claim. □

**Corollary 4** (Size of the two budgets)**.** *Let $\bar{\iota}_t := \|\mathbf{p}_t\| \Delta_t + \frac{L}{2} \Delta_t^2$, which is $\mathcal{F}_t$-measurable, and let $q_t$ be as in equation 68.*

(a) (Null step on rejection.) $I_T \leq \sum_{t<T} \mathbb{E}\big[(1 - q_t)\, \bar{\iota}_t\big]$.

(b) (Diminishing fallback of Algorithm 2.) *If $\|\mathbf{p}_t\| \leq G$ and $\|\widehat{\mathbf{g}}_t\| \leq G$ almost surely for all $t$, then*

$$I_T \leq \sum_{t<T} \mathbb{E}\big[(1 - q_t)\, \bar{\iota}_t\big] + 2\eta_0 G^2 \sqrt{T} + \tfrac{L}{2} \eta_0^2 G^2 (1 + \log T). \tag{74}$$

(c) (Model error.) *Under the polynomial accuracy level $\bar{\varepsilon}_t = \kappa_{eg} \Delta_t^p$ of Lemma 10, $M_T \leq \kappa_{eg} \sum_{t<T} \mathbb{E}[\Delta_t^{p+1}]$. For $p = 2$ and $\Delta_t = \Delta_0 (t+1)^{-a}$ this is bounded in $T$ whenever $a > \frac{1}{3}$, and equals $\kappa_{eg} \Delta^3 T$ for a constant radius $\Delta$.*

*Proof.* A round in $\mathcal{G}_t^{\mathrm{c}}$ with $\ell_t > 0$ is either accepted with $\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}}$ failing, or rejected with a fallback step. For the first kind, $\iota_t \leq \bar{\iota}_t$ and

$$\mathbb{E}\big[\bar{\iota}_t \mathbf{1}_{(\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}})^{\mathrm{c}}} \mid \mathcal{F}_t\big] = \bar{\iota}_t \Pr\big((\mathcal{E}_t^{\mathrm{smp}} \cap \mathcal{E}_t^{\mathrm{acc}})^{\mathrm{c}} \mid \mathcal{F}_t\big) \leq (1 - q_t)\, \bar{\iota}_t$$

by equation 68. Under the null-step rule rejected rounds have $\iota_t = 0$, which proves (a). Under the fallback, $\ell_t \leq \eta_0 \|\widehat{\mathbf{g}}_t\| / \sqrt{t+1} \leq \eta_0 G / \sqrt{t+1}$, so $\iota_t \leq \eta_0 G^2 (t+1)^{-1/2} + \frac{L}{2} \eta_0^2 G^2 (t+1)^{-1}$ on rejected rounds, and $\sum_{k=1}^{T} k^{-1/2} \leq 2\sqrt{T}$ and $\sum_{k=1}^{T} k^{-1} \leq 1 + \log T$ give (b). Part (c) substitutes $\bar{\varepsilon}_t$ into equation 72 and uses $\sum_{k \geq 1} k^{-a(p+1)} < \infty$ for $a(p+1) > 1$. □

**Corollary 5** (Successful-round slope bound)**.** *Suppose $\eta_t \geq \underline{\eta} > 0$ on $\mathcal{G}_t$ for all $t$. Then*

$$\min_{0 \leq t < T} \mathbb{E}\big[s_t \mathbf{1}_{\mathcal{G}_t}\big] \leq \frac{\Delta F_0 + I_T + M_T}{c\, \underline{\eta}\, T}. \tag{75}$$

*Proof.* By equation 67, $\delta_t \mathbf{1}_{\mathcal{G}_t} \geq c\, \underline{\eta}\, s_t \mathbf{1}_{\mathcal{G}_t}$; take expectations and combine with equation 71. □

**Remark 5** (How to read these bounds)**.** *The bounds concern a success-weighted quantity: they can be small because successful rounds are rare, and they do not imply stationarity of the full gradient. Under rule (a) with a fixed radius $\Delta$, a fixed $q_t = q < 1$ and a fixed level $\bar{\varepsilon}$, both budgets grow linearly and equation 71 becomes*

$$\min_{t<T} \mathbb{E}[\delta_t \mathbf{1}_{\mathcal{G}_t}] \leq \frac{\Delta F_0}{T} + (1 - q) \sup_t \mathbb{E}[\bar{\iota}_t] + \bar{\varepsilon}\, \Delta,$$

*that is, progress down to a floor set by the false-certification probability and by the model error over the trust region. In the polynomial case the last term is $\kappa_{eg} \Delta^{p+1}$, whereas the certified decrease is linear in the step; this is why a shrinking radius removes the floor. The floor disappears when $\sum_t \big[(1 - q_t)\bar{\iota}_t + \bar{\varepsilon}_t \Delta_t\big] < \infty$. The trust-region constraint is what makes $M_T$ controllable: without $\|\mathbf{z}_t\| \leq \Delta_t$ the model-error term would scale with an unbounded step. Under rule (b) the fallback contributes $O(\sqrt{T})$ to the budget and hence $O(T^{-1/2})$ to the averaged bound. A lower bound on $\Pr(\mathcal{G}_t \mid \mathcal{F}_t)$, which would turn these statements into an iteration complexity, requires controlling acceptance; the frameworks of Cartis & Scheinberg (2018) and Paquette & Scheinberg (2020) do this by adapting the step cap across rounds and are the natural next step.*

### C.10 ESTIMATED COVARIANCE

The tests above use the population covariance $\mathbf{C}_t$, whereas Algorithm 2 uses the plug-in $\widehat{\mathbf{C}}_t$. This subsection gives the one modification under which the guarantee survives with a covariance the server can compute. It requires a known uniform response bound, and it is not the rule used in the experiments.

**Assumption 4** (Bounded responses)**.** *There is a known, $\mathcal{F}_t^+$-measurable $B > 0$ such that $|y_{i,t}^k| \le B$ for every client $i$ and probe $k$.*

**Lemma 8** (One-sided covariance bound)**.** *Under Assumptions 1 and 4, fix $\delta_C \in (0,1)$ before sampling and set $L = 4B^2K$. For $2 \le m < N$ define*

$$\varepsilon_C^{\mathrm{conc}} := \frac{m}{m-1} L \left[ \sqrt{\frac{8\log(2K/\delta_C)}{m}} + \frac{\log(4K/\delta_C)}{2m} \right], \qquad \varepsilon^{\mathrm{det}} := \frac{N}{N-1} \cdot \frac{L}{4}, \tag{76}$$
$$\varepsilon_C := \min\left\{ \varepsilon_C^{\mathrm{conc}},\, \varepsilon^{\mathrm{det}} \right\},$$

*and set $\varepsilon_C := 0$ for $m = N$. Then with conditional probability at least $1 - \delta_C$, $\mathbf{\Sigma}_t \preceq \widehat{\mathbf{\Sigma}}_t + \varepsilon_C \mathbf{I}_K$, where $\widehat{\mathbf{\Sigma}}_t$ is the sample covariance of the client responses. When $\varepsilon_C = \varepsilon^{\mathrm{det}}$ the bound holds deterministically.*

*Proof. Deterministic branch.* For any $\mathbf{v}$, the scalars $\mathbf{v}^\top \mathbf{y}_{i,t}$ lie in an interval of length $2B\|\mathbf{v}\|_1$, so their population variance is at most $B^2\|\mathbf{v}\|_1^2$ and $\mathbf{v}^\top \mathbf{\Sigma}_t \mathbf{v} \le \frac{N}{N-1} B^2 \|\mathbf{v}\|_1^2$. Since $\|\mathbf{v}\|_1^2 \le K\|\mathbf{v}\|^2$, this gives $\mathbf{\Sigma}_t \preceq \varepsilon^{\mathrm{det}} \mathbf{I}_K \preceq \widehat{\mathbf{\Sigma}}_t + \varepsilon^{\mathrm{det}} \mathbf{I}_K$.

*Concentration branch.* Let $\mathbf{u}_i = \mathbf{y}_{i,t} - \boldsymbol{\mu}_t$, $\bar{\mathbf{u}} = m^{-1} \sum_{i \in S_t} \mathbf{u}_i$ and $\mathbf{M} = N^{-1} \sum_i \mathbf{u}_i \mathbf{u}_i^\top$. Since $\|\mathbf{u}_i\|^2 \le L$, direct expansion gives

$$\mathbf{\Sigma}_t - \widehat{\mathbf{\Sigma}}_t \preceq \frac{m}{m-1} \left( \lambda_{\max}\Big( \mathbf{M} - \frac{1}{m} \sum_{i \in S_t} \mathbf{u}_i \mathbf{u}_i^\top \Big) + \|\bar{\mathbf{u}}\|^2 \right) \mathbf{I}_K,$$

the discarded term being negative semidefinite. The summands $\mathbf{M} - \mathbf{u}_i \mathbf{u}_i^\top$ are centered under sampling with replacement and satisfy $\mathbf{X}_i^2 \preceq L^2 \mathbf{I}_K$, so the one-sided matrix Hoeffding bound (Tropp, 2012, Theorem 1.3), transferred to sampling without replacement by the trace-exponential comparison of Gross & Nesme (2010), bounds the first term by $L\sqrt{8\log(2K/\delta_C)/m}$ with failure probability $\delta_C/2$. Scalar Hoeffding concentration with the same comparison and a union bound over the $K$ coordinates bounds $\|\bar{\mathbf{u}}\|^2$ by $\frac{L}{2m}\log(4K/\delta_C)$ with failure probability $\delta_C/2$. Combining gives the claim, and taking the smaller of the two valid constants preserves it. □

**Remark 6** (Which branch is active)**.** *The concentration branch is smaller only if $m > 128\big(\frac{N-1}{N}\big)^2 \log\frac{2K}{\delta_C}$, which for $K = 10$ and $\delta_C = 0.05$ exceeds 700 clients per round. With $N = 100$ and $m \in \{10, 20\}$ the deterministic branch is active, so $\delta_C$ may be set to zero in Proposition 4.*

**Proposition 4** (Sampling control with estimated covariance)**.** *Under Assumptions 1 and 4, fix $\kappa_t > 0$ and $\delta_C \in (0,1)$ before sampling, let $d_t^{\mathrm{eff}}$ be admissible, and replace $\mathbf{C}_t$ in equation 42 and equation 43 by*

$$\mathbf{C}_t^+ := \widehat{\mathbf{C}}_t + f_m \varepsilon_C \mathbf{I}_K. \tag{77}$$

*Then with conditional probability at least $\big[1 - d_t^{\mathrm{eff}}/\kappa_t^2 - \delta_C\big]_+$, every accepted candidate satisfies $R_t(\eta \mathbf{d}_t) \le 0$ and $s_t(\mathbf{d}_t) \ge \underline{s}_t > 0$. When $\varepsilon_C = \varepsilon^{\mathrm{det}}$ the term $\delta_C$ may be dropped. For $m = N$ all corrected scales vanish and the conclusions are deterministic.*

*Proof.* On the event of Lemma 8, multiplying $\mathbf{\Sigma}_t \preceq \widehat{\mathbf{\Sigma}}_t + \varepsilon_C \mathbf{I}_K$ by $f_m$ gives $\mathbf{C}_t \preceq \mathbf{C}_t^+$, hence $\sigma_{R,t}(\mathbf{v}) \le \|(\mathbf{C}_t^+)^{1/2}\mathbf{v}\|$ and likewise for $\sigma_{d,t}$, simultaneously for all $\mathbf{v}$. Intersect with the sampling event of Proposition 2, whose failure probability is at most $d_t^{\mathrm{eff}}/\kappa_t^2$; a union bound costs at most $\delta_C$, and nothing when Lemma 8 is deterministic. On the intersection the inflated scales dominate the population ones, so the conclusions of Proposition 2 hold for candidates accepted with $\mathbf{C}_t^+$. Full participation follows from $f_m = 0$ and $\bar{\mathbf{y}}_t = \boldsymbol{\mu}_t$. □

**Remark 7** (Scope)**.** *Corollary 2 and Theorem 2 apply to the corrected variant with $\underline{s}_t$ computed from $\mathbf{C}_t^+$ and with $q_t$ replaced by $\mathbb{E}[\rho_{\mathrm{cov},t}^+ \mathbf{1}_{\mathcal{E}_t^{\mathrm{acc}}} \mid \mathcal{F}_t]$. The experiments use the uninflated plug-in $\widehat{\mathbf{C}}_t$, which this proposition does not certify, and which is evaluated empirically in Appendix E.6. Since $\widehat{\mathbf{C}}_t$ has rank at most $m-1$, the gap can be significant when $K \geq m$.*

# D Armijo Guarantees under Full and Partial Participation

We first prove that a step accepted by the surrogate Armijo test of SurrFedZO satisfies the Armijo condition for the true objective (Part A). The argument is deterministic: under full participation the probe responses are exact values of the global objective change, so the only discrepancy between the surrogate and the objective is the approximation error of a polynomial fitted to exact data, and this error is bounded by a Taylor argument on a poised probe set. No client count and no sample-size rate appear. We then extend the result to ProbFedZO (Part B), where the surrogate is fitted to the responses of a random subset of clients. The extension adds one ingredient, the simultaneous sampling bound of Proposition 2, and otherwise reuses Part A verbatim for the population surrogate.

Throughout, $r$ is the subspace dimension, $\Omega_t = \{\mathbf{z} \in \mathbb{R}^r : \|\mathbf{z}\| \leq \Delta_t\}$, and $f_t(\mathbf{z}) = F(\mathbf{w}_t + \mathbf{V}_t\mathbf{z}) - F(\mathbf{w}_t)$ with $\mathbf{p}_t = \nabla f_t(\mathbf{0}) = \mathbf{V}_t^\top \nabla F(\mathbf{w}_t)$. The round index $t$ is fixed and omitted where no confusion arises.

## D.1 Part A: full participation

### D.1.1 Kernel ridge regression with a polynomial kernel

Let $k(\mathbf{z}', \mathbf{z}) = (\gamma\, \mathbf{z}'^\top \mathbf{z} + c_0)^p$ with $\gamma, c_0 > 0$ be the polynomial kernel of degree $p$ on $\mathbb{R}^r$ and let $\mathcal{H}_k$ be its reproducing kernel Hilbert space. By the multinomial theorem $k(\mathbf{z}', \mathbf{z}) = \sum_{|\alpha| \leq p} c_\alpha\, \mathbf{z}'^\alpha \mathbf{z}^\alpha$ with $c_\alpha > 0$, so $k(\mathbf{z}', \mathbf{z}) = \boldsymbol{\psi}(\mathbf{z}')^\top \boldsymbol{\psi}(\mathbf{z})$ for the feature map $\psi_\alpha(\mathbf{z}) = \sqrt{c_\alpha}\, \mathbf{z}^\alpha$, $|\alpha| \leq p$. The $\psi_\alpha$ are a basis of the space $\mathcal{P}_p$ of polynomials of degree at most $p$ in $r$ variables, hence

$$\mathcal{H}_k = \mathcal{P}_p \quad \text{as a set}, \qquad q_p := \dim \mathcal{H}_k = \binom{r+p}{p}, \tag{78}$$

with $q_2 = 6$ and $q_3 = 10$ for $r = 2$. In particular every Taylor polynomial of degree at most $p$ lies in $\mathcal{H}_k$.

The probe set of a round, including the anchor, is $Y = \{\mathbf{y}_0 = \mathbf{0}, \mathbf{y}_1, \ldots, \mathbf{y}_K\} \subset \Omega_t$ with $n := K+1$ points. Write $\mathbf{K} = (k(\mathbf{y}_i, \mathbf{y}_j))_{i,j=0}^K$ for the kernel matrix, $\mathbf{k}(\mathbf{z}) = (k(\mathbf{y}_i, \mathbf{z}))_{i=0}^K$, and $\boldsymbol{\Phi} \in \mathbb{R}^{n \times q_p}$ for the matrix with rows $\boldsymbol{\psi}(\mathbf{y}_i)^\top$, so that $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$. For a response vector $\mathbf{v} \in \mathbb{R}^n$ with $v_0 = 0$ and a ridge parameter $\lambda \geq 0$, the anchored KRR fit is

$$m_\lambda(\mathbf{z}) := \big(\mathbf{k}(\mathbf{z}) - \mathbf{k}(\mathbf{0})\big)^\top (\mathbf{K} + \lambda \mathbf{I})^+ \mathbf{v} = \sum_{i=0}^K a_i^\lambda(\mathbf{z})\, v_i, \qquad a_i^\lambda(\mathbf{z}) := \big[(\mathbf{K} + \lambda \mathbf{I})^+ (\mathbf{k}(\mathbf{z}) - \mathbf{k}(\mathbf{0}))\big]_i, \tag{79}$$

where $(\cdot)^+$ is the Moore–Penrose pseudo-inverse, equal to the inverse whenever $\mathbf{K} + \lambda\mathbf{I}$ is nonsingular. This is the construction of equation 20. The fit is linear in $\mathbf{v}$ and the *KRR weight functions* $a_i^\lambda$ depend only on the kernel, the probes and $\lambda$. Each $a_i^\lambda$ is a linear combination of $k(\mathbf{y}_j, \cdot) - k(\mathbf{y}_j, \mathbf{0})$ and therefore lies in $\mathcal{H}_k = \mathcal{P}_p$. We write $a_i := a_i^0$.

**Definition 2** (Poised probe set and gradient poisedness constant)**.** *$Y$ is* poised for $k$ *if* $\operatorname{rank} \mathbf{K} = q_p$*, equivalently if the representers $k(\mathbf{y}_0, \cdot), \ldots, k(\mathbf{y}_K, \cdot)$ span $\mathcal{H}_k$, equivalently if $\boldsymbol{\Phi}$ has full column rank. This requires $n \geq q_p$. The* gradient poisedness constant *of a poised $Y$ in $\Omega_t$ is*

$$\Lambda_1(Y) := \Delta_t \max_{\mathbf{z} \in \Omega_t} \sum_{i=0}^K \|\nabla a_i(\mathbf{z})\|. \tag{80}$$

The three characterizations of poisedness agree because the evaluation map $E : \mathcal{H}_k \to \mathbb{R}^n$, $E(h) = (h(\mathbf{y}_i))_{i=0}^K$, has matrix $\boldsymbol{\Phi}$ in the basis $\boldsymbol{\psi}$, and $\operatorname{rank} \mathbf{K} = \operatorname{rank} \boldsymbol{\Phi}$. Poisedness says that $E$ is injective, that is, an element of $\mathcal{H}_k$ vanishing on all probes is zero. $\Lambda_1(Y)$ depends only on the shape of $Y$

and not on its scale, since $\mathcal{H}_k = \mathcal{P}_p$ is invariant under $\mathbf{z} \mapsto \Delta_t \mathbf{z}$ and the weight functions of $Y/\Delta_t$ in the unit ball are $\hat{a}_i(\mathbf{u}) = a_i(\Delta_t \mathbf{u})$. The constant is computable on the server from the kernel and the probes alone, before anything is broadcast, and is finite if and only if $Y$ is poised.

**Lemma 9** (KRR at $\lambda = 0$ reproduces its RKHS on a poised set)**.** *Let $Y$ be poised for $k$ and let $h \in \mathcal{H}_k$ with $h(\mathbf{0}) = 0$. If $\mathbf{v} = (h(\mathbf{y}_i))_{i=0}^K$, then the anchored fit equation 79 with $\lambda = 0$ satisfies*

$$m_0 = h \quad \text{on } \mathbb{R}^r, \qquad \text{in particular} \qquad \nabla m_0 = \nabla h. \tag{81}$$

*Equivalently, $\sum_{i=0}^K h(\mathbf{y}_i)\, a_i = h$ for every such $h$.*

*Proof.* Set $\boldsymbol{\alpha} := \mathbf{K}^+ \mathbf{v}$ and $g := \sum_{j=0}^K \alpha_j\, k(\mathbf{y}_j, \cdot) \in \mathcal{H}_k$, so that $m_0 = g - g(\mathbf{0})$. The values of $g$ on the probes are $E(g) = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}^+\mathbf{v}$, and $\mathbf{K}\mathbf{K}^+$ is the orthogonal projector onto $\operatorname{range}\mathbf{K}$. The columns of $\mathbf{K}$ are $E(k(\mathbf{y}_j, \cdot))$, and since the representers span $\mathcal{H}_k$, $\operatorname{range}\mathbf{K} = E(\mathcal{H}_k)$. As $\mathbf{v} = E(h)$ with $h \in \mathcal{H}_k$, we get $\mathbf{K}\mathbf{K}^+\mathbf{v} = \mathbf{v}$, that is, $E(g) = E(h)$. Poisedness makes $E$ injective on $\mathcal{H}_k$, so $g = h$, and $m_0 = h - h(\mathbf{0}) = h$. □

The lemma is the only property of KRR used in Part A. It does not require $n = q_p$: when $n > q_p$ the kernel matrix is singular, the pseudo-inverse selects the fit of minimal RKHS norm among those with least squared residual on $Y$, and this fit still reproduces every element of $\mathcal{H}_k$. When $n = q_p$ the fit interpolates $\mathbf{v}$ on $Y$. For $\lambda > 0$ the reproduction property fails and a bias appears, which Lemma 11 bounds by $\kappa_{\mathrm{reg}} \Delta_t^p$ under a ridge schedule tied to $\Delta_t$.

### D.1.2 Assumptions for full participation

**Assumption 5** (Full participation, smoothness, design)**.** *Round $t$:*

(i) (Exact responses.) *Every client $i \in \{1, \dots, N\}$ participates and returns $y_i^k = F_i(\mathbf{w}_t + \mathbf{V}_t \mathbf{z}^k) - F_i(\mathbf{w}_t)$ exactly, so that $\bar{y}^k = \frac{1}{N}\sum_{i=1}^N y_i^k = f_t(\mathbf{z}^k)$ for $k = 1, \dots, K$.*

(ii) (Smoothness of order $p$.) *$F \in C^p$ on a neighbourhood of $\{\mathbf{w}_t + \mathbf{V}_t \mathbf{z} : \mathbf{z} \in \Omega_t\}$ and its $p$th derivative is $L_p$-Lipschitz there.*

(iii) (Design.) *The probe set $Y$ (anchor included) is poised for $k$ (Definition 2) with gradient poisedness constant $\Lambda_1(Y) \le \Lambda$, and SurrFedZO fits the anchored polynomial KRR of Lemma 9 with $\lambda = 0$.*

Part (i) is an identity, not an estimate, because $F$ is defined as the average of the $N$ local objectives. This is the sense in which no client count enters Part A. If the local objective were evaluated on a minibatch rather than in full, the response would carry evaluation noise and (i) would fail. Part B covers noise that comes from client sampling, not from stochastic local evaluation.

In (ii), $f_t \in C^p(\Omega_t)$ with $\nabla^p f_t$ $L_p$-Lipschitz on $\Omega_t$ follows because $\mathbf{V}_t$ has orthonormal columns. For $p = 2$ this is a Lipschitz Hessian. For the regularized logistic loss $\ell(\mathbf{w}; \mathbf{x}, y) = \log(1 + e^{-y\mathbf{x}^\top \mathbf{w}}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, the third derivative of $u \mapsto \log(1 + e^{-u})$ is bounded by $1/(6\sqrt{3})$, so $L_2 \le \frac{1}{6\sqrt{3}} \max_{(\mathbf{x},y)} \|\mathbf{x}\|^3$ uniformly over the clients. For $p = 3$ one needs a Lipschitz third derivative, which holds for the logistic loss with a constant of the same form in $\max \|\mathbf{x}\|^4$.

The server chooses $Y$, so (iii) is enforced, not assumed: it draws probes, computes $\Lambda_1(Y)$ from equation 80, and redraws if the constant exceeds $\Lambda$. This costs no client evaluations since it happens before the probes are broadcast. The requirement $K + 1 \ge q_p$ is $K \ge 5$ for a quadratic model and $K \ge 9$ for a cubic model in $r = 2$.

### D.1.3 Model error

**Lemma 10** (Deterministic gradient error of the surrogate; restates Lemma 1)**.** *Under Assumption 5, the surrogate $m_t = m_0$ of SurrFedZO satisfies*

$$\varepsilon_t := \sup_{\mathbf{z} \in \Omega_t} \|\nabla f_t(\mathbf{z}) - \nabla m_t(\mathbf{z})\| \;\le\; \kappa_{eg}\, \Delta_t^p, \qquad \kappa_{eg} := L_p\Big(\frac{1}{p!} + \frac{\Lambda}{(p+1)!}\Big). \tag{82}$$

*Moreover $E_t := f_t - m_t$ satisfies $E_t(\mathbf{0}) = 0$ and $|E_t(\mathbf{z})| \le \varepsilon_t \|\mathbf{z}\|$ for $\mathbf{z} \in \Omega_t$.*

*Proof.* Let $T \in \mathcal{P}_p$ be the degree-$p$ Taylor polynomial of $f_t$ at $\mathbf{0}$ and $R := f_t - T$, so $T(\mathbf{0}) = f_t(\mathbf{0}) = 0$ and $R(\mathbf{0}) = 0$. Because $\nabla^p f_t$ is $L_p$-Lipschitz on the convex set $\Omega_t$, the integral form of the Taylor remainder gives, for all $\mathbf{z} \in \Omega_t$,

$$|R(\mathbf{z})| \le \frac{L_p}{(p+1)!}\|\mathbf{z}\|^{p+1}, \qquad \|\nabla R(\mathbf{z})\| \le \frac{L_p}{p!}\|\mathbf{z}\|^p. \tag{83}$$

By Assumption 5(i) the response vector is $\mathbf{v} = (f_t(\mathbf{y}_i))_{i=0}^K = (T(\mathbf{y}_i))_i + (R(\mathbf{y}_i))_i$ with $v_0 = 0$, and by Assumption 5(iii) the surrogate is the KRR fit equation 79 with $\lambda = 0$. Since the fit is linear in $\mathbf{v}$,

$$m_t = \sum_{i=0}^K T(\mathbf{y}_i)\, a_i + \sum_{i=0}^K R(\mathbf{y}_i)\, a_i = T + \sum_{i=0}^K R(\mathbf{y}_i)\, a_i,$$

where the first sum equals $T$ by Lemma 9, because $T \in \mathcal{P}_p = \mathcal{H}_k$ by equation 78 and $T(\mathbf{0}) = 0$. Hence $\nabla f_t - \nabla m_t = \nabla R - \sum_{i=0}^K R(\mathbf{y}_i)\nabla a_i$, and for $\mathbf{z} \in \Omega_t$, using $R(\mathbf{y}_0) = 0$ and $\|\mathbf{y}_i\| \le \Delta_t$,

$$\begin{aligned}\|\nabla f_t(\mathbf{z}) - \nabla m_t(\mathbf{z})\| &\le \|\nabla R(\mathbf{z})\| + \sum_{i=1}^K |R(\mathbf{y}_i)|\, \|\nabla a_i(\mathbf{z})\| \\ &\le \frac{L_p}{p!}\Delta_t^p + \frac{L_p}{(p+1)!}\Delta_t^{p+1}\sum_{i=0}^K \|\nabla a_i(\mathbf{z})\| \le \frac{L_p}{p!}\Delta_t^p + \frac{L_p}{(p+1)!}\Delta_t^{p+1}\cdot\frac{\Lambda_1(Y)}{\Delta_t},\end{aligned}$$

by equation 83 and equation 80. This is equation 82 with $\Lambda_1(Y) \le \Lambda$. The number of probes enters only through the weight functions $a_i$, whose gradients are summed and bounded by $\Lambda$, and the number of clients does not enter because $\mathbf{v}$ consists of exact values of $f_t$. Finally $E_t(\mathbf{0}) = f_t(\mathbf{0}) - m_t(\mathbf{0}) = 0$ by anchoring, and $E_t(\mathbf{z}) = \int_0^1 \nabla E_t(\varsigma\mathbf{z})^\top \mathbf{z}\, d\varsigma$ gives $|E_t(\mathbf{z})| \le \varepsilon_t\|\mathbf{z}\|$ because the segment lies in $\Omega_t$. □

**Lemma 11** (Ridge parameter: the bias has the same form)**.** *Apply the kernel in scaled coordinates, $k_t(\mathbf{z}', \mathbf{z}) = (\gamma_0\, \mathbf{z}'^\top \mathbf{z}/\Delta_t^2 + c_0)^p$, and choose the ridge by the schedule $\lambda_t = \lambda_0\Delta_t^p$ with $\lambda_0 \ge 0$. Assume $\Delta_t \le 1$ and that the derivatives of $f_t$ at $\mathbf{0}$ of order $1$ to $p$ satisfy $\sum_{1\le|\alpha|\le p} |\partial^\alpha f_t(\mathbf{0})|/(\alpha!\sqrt{c_\alpha}) \le G$. Let $\widehat{\mathbf{\Phi}}$ be the design matrix of $Y/\Delta_t$ in the unit ball, with smallest singular value $\hat{\sigma}_{\min} > 0$ by poisedness. Then*

$$\sup_{\Omega_t} \|\nabla m_{\lambda_t} - \nabla m_0\| \le \kappa_{\mathrm{reg}}\Delta_t^p, \qquad \kappa_{\mathrm{reg}} := \frac{\lambda_0}{\hat{\sigma}_{\min}^2}\Big(G + \frac{\sqrt{K+1}\, L_p}{(p+1)!\,\hat{\sigma}_{\min}}\Big) \sup_{\|\mathbf{u}\|\le 1} \|\nabla \boldsymbol{\psi}(\mathbf{u})\|_{\mathrm{op}}, \tag{84}$$

*so Lemma 10 holds for the ridge fit with $\kappa_{eg}$ replaced by $\kappa_{eg} + \kappa_{\mathrm{reg}}$. Setting $\lambda_0 = 0$ recovers Lemma 10.*

*Proof.* Write $\mathbf{u} = \mathbf{z}/\Delta_t$, so the scaled feature map is $\boldsymbol{\psi}(\mathbf{u})$ and $\widehat{\mathbf{\Phi}}$ has rows $\boldsymbol{\psi}(\mathbf{y}_i/\Delta_t)^\top$. The kernel identity $\mathbf{K} = \widehat{\mathbf{\Phi}}\widehat{\mathbf{\Phi}}^\top$ and the push-through identity give $m_\lambda(\mathbf{z}) = (\boldsymbol{\psi}(\mathbf{u}) - \boldsymbol{\psi}(\mathbf{0}))^\top \boldsymbol{\beta}_\lambda$ with $\boldsymbol{\beta}_\lambda = (\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Phi}} + \lambda\mathbf{I})^{-1}\widehat{\mathbf{\Phi}}^\top\mathbf{v}$ and $\boldsymbol{\beta}_0 = \widehat{\mathbf{\Phi}}^+\mathbf{v}$, hence $\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0 = -\lambda(\widehat{\mathbf{\Phi}}^\top\widehat{\mathbf{\Phi}} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}_0$ and $\|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0\| \le (\lambda/\hat{\sigma}_{\min}^2)\|\boldsymbol{\beta}_0\|$. By the chain rule $\nabla_{\mathbf{z}} m_\lambda - \nabla_{\mathbf{z}} m_0 = \Delta_t^{-1}\nabla\boldsymbol{\psi}(\mathbf{u})^\top(\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0)$, so

$$\sup_{\Omega_t} \|\nabla m_\lambda - \nabla m_0\| \le \frac{\lambda}{\Delta_t\hat{\sigma}_{\min}^2}\, \|\boldsymbol{\beta}_0\| \sup_{\|\mathbf{u}\|\le 1} \|\nabla\boldsymbol{\psi}(\mathbf{u})\|_{\mathrm{op}}. \tag{85}$$

It remains to show $\|\boldsymbol{\beta}_0\| \le \Delta_t\big(G + \sqrt{K+1}L_p/((p+1)!\hat{\sigma}_{\min})\big)$. The responses are the values of $h(\mathbf{u}) := f_t(\Delta_t\mathbf{u})$ on $Y/\Delta_t$. Split $h = \widehat{T} + \widehat{R}$ with $\widehat{T}$ the degree-$p$ Taylor polynomial of $h$ at $\mathbf{0}$, whose coefficient of $\psi_\alpha$ is $\partial^\alpha f_t(\mathbf{0})\Delta_t^{|\alpha|}/(\alpha!\sqrt{c_\alpha})$, and $|\widehat{R}(\mathbf{u})| \le L_p\Delta_t^{p+1}/(p+1)!$ on the unit ball by equation 83. Since $\widehat{\mathbf{\Phi}}^+$ reproduces $\widehat{T}$ (Lemma 9), $\boldsymbol{\beta}_0 = \mathrm{coef}(\widehat{T}) + \widehat{\mathbf{\Phi}}^+(\widehat{R}(\mathbf{y}_i/\Delta_t))_i$. With $\Delta_t \le 1$ and $|\alpha| \ge 1$, $\|\mathrm{coef}(\widehat{T})\| \le \Delta_t G$, and $\|\widehat{\mathbf{\Phi}}^+(\widehat{R}(\mathbf{y}_i/\Delta_t))_i\| \le \hat{\sigma}_{\min}^{-1}\sqrt{K+1}\, L_p\Delta_t^{p+1}/(p+1)! \le \Delta_t\sqrt{K+1}\, L_p/((p+1)!\hat{\sigma}_{\min})$. Substituting into equation 85 with $\lambda = \lambda_0\Delta_t^p$ gives equation 84. The last claim follows from the triangle inequality $\|\nabla f_t - \nabla m_{\lambda_t}\| \le \|\nabla f_t - \nabla m_0\| + \|\nabla m_0 - \nabla m_{\lambda_t}\|$ and Lemma 10. □

The scaled kernel does not change $m_0$, since reproduction is basis free, nor $\Lambda_1(Y)$, which is scale invariant. What the schedule $\lambda_t = \lambda_0 \Delta_t^p$ does is tie the shrinkage to the trust region so that it decays at the same rate as the Taylor remainder. The factor $\sqrt{K+1}$ multiplies only the remainder's contribution to the coefficients and can be dropped by bounding that contribution through $\Lambda_1(Y)$ instead, at the cost of a longer statement. Under full participation there is no statistical reason for a ridge and Assumption 5(iii) sets $\lambda = 0$.

#### D.1.4 Acceptance and the true Armijo condition

Write $\mathbf{g}_t = \nabla m_t(\mathbf{0})$, $\mathbf{H}_t = \nabla^2 m_t(\mathbf{0})$, $\mathbf{d}_t = -|\mathbf{H}_t|_{\epsilon_H}^{-1}\mathbf{g}_t$ and $s_t = -\mathbf{g}_t^\top \mathbf{d}_t$. SurrFedZO accepts $\eta$ if the surrogate Armijo condition

$$m_t(\eta \mathbf{d}_t) \;\le\; -\,c\,\eta\, s_t \qquad (\text{equivalently } R_t(\eta \mathbf{d}_t) \le 0) \tag{86}$$

holds and takes the largest such $\eta$ on its grid. Under full participation there are no hats: $m_t$, $\mathbf{g}_t$ and $s_t$ are known exactly on the server.

**Proposition 5** (The surrogate test is always satisfiable)**.** *If $\mathbf{g}_t \neq \mathbf{0}$ then $s_t = \mathbf{g}_t^\top |\mathbf{H}_t|_{\epsilon_H}^{-1}\mathbf{g}_t \ge \|\mathbf{g}_t\|^2/\lambda_{\max}(|\mathbf{H}_t|_{\epsilon_H}) > 0$, and with $B_t := \sup_{0\le\eta\le\eta_{\max,t}} \left|\mathbf{d}_t^\top \nabla^2 m_t(\eta \mathbf{d}_t)\mathbf{d}_t\right|$ the condition equation 86 holds for every $\eta \in (0, \bar\eta_t]$, where $\bar\eta_t = \min\{\eta_{\max,t},\, 2(1-c)s_t/B_t\} > 0$ (read as $\eta_{\max,t}$ if $B_t = 0$). Hence any grid $\mathcal{H}_t$ containing a point of $(0, \bar\eta_t]$ has a nonempty accepted set.*

*Proof.* The first claim is the Rayleigh quotient bound for the positive definite matrix $|\mathbf{H}_t|_{\epsilon_H}^{-1}$. Since $m_t(\mathbf{0}) = 0$ and $\nabla m_t(\mathbf{0})^\top \mathbf{d}_t = -s_t$, Taylor's theorem along the segment gives $m_t(\eta \mathbf{d}_t) \le -\eta s_t + \frac{1}{2}B_t\eta^2$, so equation 86 holds whenever $-\eta s_t + \frac{1}{2}B_t\eta^2 \le -c\eta s_t$, that is whenever $\eta \le 2(1-c)s_t/B_t$. Any grid point in $(0, \bar\eta_t]$ therefore passes. □

**Theorem 3** (True Armijo condition under full participation; restates and extends Theorem 1)**.** *Let Assumption 5 hold, let $\mathbf{g}_t \neq \mathbf{0}$, and let $\eta_t$ be accepted by SurrFedZO, so that equation 86 holds at $\eta_t$ and $\|\eta_t \mathbf{d}_t\| \le \Delta_t$. Put $\mathbf{z}_t = \eta_t \mathbf{d}_t$.*

(i) (Decrease up to model error.) $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) = f_t(\mathbf{z}_t) \le -c\,\eta_t s_t + \kappa_{eg}\Delta_t^p\,\eta_t\|\mathbf{d}_t\|$.

(ii) (True Armijo condition.) *If*

$$\kappa_{eg}\,\Delta_t^p\,\|\mathbf{d}_t\| \;\le\; \theta\, c\, s_t \qquad \textit{for some } \theta \in [0,1), \tag{87}$$

*then $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \le -(1-\theta)\,c\,\eta_t s_t < 0$ and, for every $c' \in (0, \bar c_t]$ with*

$$\bar c_t := \frac{c\,s_t - \varepsilon_t\|\mathbf{d}_t\|}{s_t + \varepsilon_t\|\mathbf{d}_t\|} \;\ge\; \frac{(1-\theta)\,c}{1+\theta c}, \tag{88}$$

*the Armijo condition for the true objective holds:*

$$F(\mathbf{w}_t + \mathbf{V}_t\mathbf{z}_t) \;\le\; F(\mathbf{w}_t) + c'\,\eta_t\,\nabla F(\mathbf{w}_t)^\top \mathbf{V}_t\mathbf{d}_t. \tag{89}$$

*Proof. (i)* Split $f_t = m_t + E_t$. By equation 86, $m_t(\mathbf{z}_t) \le -c\eta_t s_t$. By Lemma 10, $E_t(\mathbf{z}_t) \le \varepsilon_t\|\mathbf{z}_t\| = \varepsilon_t\eta_t\|\mathbf{d}_t\| \le \kappa_{eg}\Delta_t^p\eta_t\|\mathbf{d}_t\|$. Add the two.

*(ii)* Under equation 87, $\varepsilon_t\|\mathbf{d}_t\| \le \kappa_{eg}\Delta_t^p\|\mathbf{d}_t\| \le \theta c s_t$, so (i) gives $f_t(\mathbf{z}_t) \le -c\eta_t s_t + \theta c\eta_t s_t$, which is the strict decrease. For equation 89, note that $\nabla F(\mathbf{w}_t)^\top \mathbf{V}_t\mathbf{d}_t = \mathbf{p}_t^\top \mathbf{d}_t$ and $\mathbf{p}_t = \mathbf{g}_t + \nabla E_t(\mathbf{0})$ with $\|\nabla E_t(\mathbf{0})\| \le \varepsilon_t$, hence

$$c'\eta_t\,\mathbf{p}_t^\top \mathbf{d}_t = -c'\eta_t s_t + c'\eta_t \nabla E_t(\mathbf{0})^\top \mathbf{d}_t \;\ge\; -c'\eta_t\big(s_t + \varepsilon_t\|\mathbf{d}_t\|\big).$$

Combining with (i) in the form $f_t(\mathbf{z}_t) \le -\eta_t(cs_t - \varepsilon_t\|\mathbf{d}_t\|)$, the true Armijo inequality $f_t(\mathbf{z}_t) \le c'\eta_t\mathbf{p}_t^\top\mathbf{d}_t$ holds whenever $-\eta_t(cs_t - \varepsilon_t\|\mathbf{d}_t\|) \le -c'\eta_t(s_t + \varepsilon_t\|\mathbf{d}_t\|)$, that is whenever $c' \le \bar c_t$. The map $a \mapsto (cs_t - a)/(s_t + a)$ is decreasing on $a \ge 0$, so $a = \varepsilon_t\|\mathbf{d}_t\| \le \theta c s_t$ gives the lower bound in equation 88. Positivity of $\bar c_t$ follows from $\theta < 1$, and $\bar c_t \le c$ from $cs_t - a \le c(s_t + a)$. □

**Remark 8** (The condition equation 87 is checkable)**.** *Under full participation the server knows $s_t$ and $\|\mathbf{d}_t\|$ exactly, and $\Delta_t$ is its own choice, so given an upper bound on $L_p$ the condition equation 87 is a single inequality the server could evaluate. Since its left side scales as $\Delta_t^p$ while $s_t$ stays*

*bounded away from zero as long as the surrogate gradient does, shrinking the radius satisfies it whenever $\mathbf{p}_t \neq \mathbf{0}$. This is the criticality condition of model-based trust-region methods (Conn et al., 2009). Algorithm 1 does not test it, so part (i) of Theorem 3 applies to every accepted round and part (ii) to those rounds on which the condition happens to hold.*

**Remark 9** (Where the probe count enters)**.** *Nothing in Part A improves with $K$ beyond the requirement $K + 1 \geq q_p$: the bound equation 82 contains no $K$, and the number of clients $N$ appears nowhere. What controls $\varepsilon_t$ is the radius $\Delta_t$, through $\Delta_t^p$, and the geometry of the probes, through $\Lambda$. This is the reason the guarantee is available at $K = 7$ probes for $p = 2$ and at $K = 9$ for $p = 3$, whereas a kernel whose native space is infinite dimensional would need the fill distance of the probes to be small and hence $K$ large.*

### D.2 Part B: random client sampling

Now $S_t \subset \{1, \ldots, N\}$ is a uniform random subset of size $2 \leq m < N$ (Assumption 1), and the server observes $\bar{\mathbf{y}}_t = \boldsymbol{\mu}_t + \boldsymbol{e}_t$ with $\mathbb{E}[\boldsymbol{e}_t \mid \mathcal{F}_t^+] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{e}_t \mid \mathcal{F}_t^+) = \mathbf{C}_t = \frac{N-m}{Nm}\boldsymbol{\Sigma}_t$. The fitted surrogate and the population surrogate are

$$\widehat{m}_t(\mathbf{z}) = \mathbf{a}_t(\mathbf{z})^\top \bar{\mathbf{y}}_t, \qquad m_t(\mathbf{z}) = \mathbf{a}_t(\mathbf{z})^\top \boldsymbol{\mu}_t, \tag{90}$$

with the coefficient map $\mathbf{a}_t$ of equation 20. The key observation is that $\boldsymbol{\mu}_t$ is the exact response vector of Assumption 5(i): $\mu_t^k = \frac{1}{N}\sum_i y_{i,t}^k = f_t(\mathbf{z}_t^k)$. Hence $m_t$ is precisely the surrogate analysed in Part A, and Lemma 10 applies to it without change. The whole difference between the two regimes is the sampling error $\widehat{m}_t - m_t = \mathbf{a}_t(\cdot)^\top \boldsymbol{e}_t$, and that is the only quantity for which a client count, namely $m$, enters.

ProbFedZO runs the tests

$$\underline{s}_t := \widehat{s}_t - \kappa_t \sigma_{d,t}(\mathbf{d}_t) > 0, \qquad Q_t(\eta) := \widehat{R}_t(\eta \mathbf{d}_t) + \kappa_t \sigma_{R,t}(\eta \mathbf{d}_t) \leq 0, \tag{91}$$

where $\widehat{s}_t = -\widehat{\mathbf{g}}_t^\top \mathbf{d}_t$, $\widehat{R}_t(\eta \mathbf{d}_t) = \widehat{m}_t(\eta \mathbf{d}_t) - c\eta \widehat{\mathbf{g}}_t^\top \mathbf{d}_t$, and $\sigma_{d,t}, \sigma_{R,t}$ are the standard errors of $\widehat{s}_t$ and $\widehat{R}_t$ computed from $\mathbf{C}_t$ (Appendix B). The direction $\mathbf{d}_t = -|\widehat{\mathbf{H}}_t|_{\epsilon_H}^{-1} \widehat{\mathbf{g}}_t$ is built from the same responses; $s_t = -\mathbf{g}_t^\top \mathbf{d}_t$ denotes the slope of the population surrogate along that (random) direction.

**Assumption 6** (Design and model, partial participation)**.** *Assumption 5(ii)–(iii) holds with the same kernel and probe rule, and either $\lambda = 0$ or the ridge schedule of Lemma 11, so that $\varepsilon_t \leq \kappa_{eg} \Delta_t^p$, where $\kappa_{eg}$ is the constant of Lemma 10 in the first case and $\kappa_{eg} + \kappa_{\mathrm{reg}}$ in the second. Either way the bound is determined by the probes before the clients are sampled, and below $\kappa_{eg}$ denotes whichever constant applies.*

**Theorem 4** (True Armijo condition under random client sampling; the partial-participation counterpart of Theorem 1)**.** *Let Assumptions 1 and 6 hold, choose $\kappa_t > 0$ before sampling, and let $d_t^{\mathrm{eff}}$ be an admissible effective dimension (Lemma 5; for the polynomial kernel of degree $p$ one may take $d_t^{\mathrm{eff}} = \min\{\operatorname{rank} \mathbf{C}_t,\, q_p - 1\}$, which covers a direction and a grid chosen after seeing $\bar{\mathbf{y}}_t$). Then there is an event $\mathcal{E}_t^{\mathrm{smp}}$ with $\Pr(\mathcal{E}_t^{\mathrm{smp}} \mid \mathcal{F}_t^+) \geq [1 - d_t^{\mathrm{eff}}/\kappa_t^2]_+$ on which every step $\eta_t$ certified by equation 91 satisfies, with $\mathbf{z}_t = \eta_t \mathbf{d}_t$:*

(i) *$s_t \geq \underline{s}_t > 0$ and $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq -c\,\eta_t\,\underline{s}_t + \kappa_{eg}\Delta_t^p\,\eta_t \|\mathbf{d}_t\|$;*

(ii) *if the observable test*

$$\kappa_{eg}\Delta_t^p\,\|\mathbf{d}_t\| \;\leq\; \theta\, c\, \underline{s}_t, \qquad \theta \in [0, 1), \tag{92}$$

*holds, then $F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq -(1-\theta)\, c\, \eta_t\, \underline{s}_t < 0$ and the true Armijo condition equation 89 holds for every $c' \in \big(0,\ (1-\theta)c/(1+\theta c)\big]$.*

*With the estimated covariance $\mathbf{C}_t^+$ of Proposition 4 in place of $\mathbf{C}_t$ in equation 91, the same conclusions hold with probability at least $[1 - d_t^{\mathrm{eff}}/\kappa_t^2 - \delta_C]_+$ under Assumption 4.*

*Proof. Step 1: from the fitted to the population surrogate.* Proposition 2 provides $\mathcal{E}_t^{\mathrm{smp}}$ with the stated probability on which, simultaneously for every tested $(\eta, \mathbf{d})$, $R_t(\eta \mathbf{d}) \leq \widehat{R}_t(\eta \mathbf{d}) + \kappa_t \sigma_{R,t}(\eta \mathbf{d})$ and $s_t(\mathbf{d}) \geq \widehat{s}_t(\mathbf{d}) - \kappa_t \sigma_{d,t}(\mathbf{d})$. Since the bounds hold for all candidates at once, they hold at the

selected $(\eta_t, \mathbf{d}_t)$ even though it was chosen from $\bar{\mathbf{y}}_t$. Certification by equation 91 therefore gives, on $\mathcal{E}_t^{\mathrm{smp}}$,

$$R_t(\eta_t \mathbf{d}_t) \le Q_t(\eta_t) \le 0 \quad \text{and} \quad s_t \ge \underline{s}_t > 0, \tag{93}$$

that is, the population surrogate satisfies the surrogate Armijo condition equation 86 at $\eta_t$ along $\mathbf{d}_t$, and $\mathbf{d}_t$ is a descent direction for it.

*Step 2: from the population surrogate to the objective.* The population surrogate $m_t$ is the anchored KRR fit to the exact values $\mu_t^k = f_t(\mathbf{z}_t^k)$, so Lemma 10, or Lemma 11 under the ridge schedule, gives $\varepsilon_t \le \kappa_{eg} \Delta_t^p$ deterministically, whatever the client sample. The bound is uniform over $\Omega_t$, so it applies along the random direction $\mathbf{d}_t$. Theorem 3(i) used only equation 86 at $\eta_t$ and the bound $|E_t(\mathbf{z})| \le \varepsilon_t \|\mathbf{z}\|$; both are available, so $f_t(\mathbf{z}_t) \le -c\eta_t s_t + \varepsilon_t \eta_t \|\mathbf{d}_t\| \le -c\eta_t \underline{s}_t + \kappa_{eg} \Delta_t^p \eta_t \|\mathbf{d}_t\|$ by equation 93. This is (i).

For (ii), Theorem 3(ii) applies with $s_t$ replaced by the certified margin: from equation 92 and $s_t \ge \underline{s}_t$, $\varepsilon_t \|\mathbf{d}_t\| \le \kappa_{eg} \Delta_t^p \|\mathbf{d}_t\| \le \theta c \underline{s}_t \le \theta c s_t$, which is equation 87 with $s_t$ replaced by the smaller $\underline{s}_t$. The proof of Theorem 3(ii) then yields $f_t(\mathbf{z}_t) \le -(1-\theta) c \eta_t s_t \le -(1-\theta) c \eta_t \underline{s}_t$ and the true Armijo condition for $c' \le \bar{c}_t$, where $\bar{c}_t$ is given by equation 88 with the unobserved $s_t$. Since $a \mapsto (cs - a)/(s + a)$ is decreasing in $a$ and increasing in $s$, the constraints $a \le \theta c s_t$ and $s_t \ge \underline{s}_t$ give $\bar{c}_t \ge (1-\theta)c/(1+\theta c)$, a bound that does not involve $s_t$.

*Step 3: estimated covariance.* Proposition 4 replaces $\mathcal{E}_t^{\mathrm{smp}}$ by its intersection with the covariance event of Lemma 8, on which the inflated scales dominate the population ones, so equation 93 continues to hold for steps certified with $\mathbf{C}_t^+$; a union bound costs $\delta_C$. Steps 1–2 are unchanged. □

**Remark 10** (What is random and what is not)**.** *Theorem 4 separates the two errors cleanly. The model error $\varepsilon_t$ is deterministic given the probes. It is bounded by $\kappa_{eg}\Delta_t^p$ exactly as in Lemma 10, and no client count enters it. The sampling error is random, is the only place where $m$ enters, and does so through $\mathbf{C}_t = \frac{N-m}{Nm}\mathbf{\Sigma}_t$, which vanishes at $m = N$. At full participation the event $\mathcal{E}_t^{\mathrm{smp}}$ has probability one for every $\kappa_t > 0$ (because $\operatorname{rank} \mathbf{C}_t = 0$), $\underline{s}_t = s_t$, and Theorem 4 reduces to Theorem 3.*

**Remark 11** (The criticality conditions are not run)**.** *Under partial participation the server observes $\underline{s}_t$ but not $s_t$, so equation 92 is stated with the certified margin and is more conservative than equation 87 by the sampling correction. Neither condition is tested by Algorithm 1 or Algorithm 2, and neither is needed for Theorem 2. They are the price of a per-round statement of the true Armijo condition, and part (i) of each theorem holds without them.*

# E Additional experiments and ablations

## E.1 Proof of Lemma 1

Let $\Omega_t = \{\mathbf{z} \in \mathbb{R}^r : \|\mathbf{z}\| \le \Delta_t\}$, let $\mathcal{P}_p$ be the polynomials of degree at most $p$ in $r$ variables, $q_p = \dim \mathcal{P}_p = \binom{r+p}{p}$, and fix a basis $\boldsymbol{\phi} = (\phi_1, \dots, \phi_{q_p})$. The probe set including the anchor is $Y = \{\mathbf{y}_0 = \mathbf{0}, \mathbf{y}_1, \dots, \mathbf{y}_K\} \subset \Omega_t$ and $\mathbf{\Phi} \in \mathbb{R}^{(K+1)\times q_p}$ has rows $\boldsymbol{\phi}(\mathbf{y}_i)^\top$.

**Definition 3** (Poised set, Lagrange polynomials, poisedness constant)**.** *$Y$ is* poised for degree-$p$ regression *if $K+1 \ge q_p$ and $\mathbf{\Phi}$ has full column rank. Then the least-squares fit*

$$\mathcal{L}(\mathbf{v}) := \arg\min_{m \in \mathcal{P}_p} \sum_{i=0}^{K} \big(m(\mathbf{y}_i) - v_i\big)^2, \qquad \mathbf{v} \in \mathbb{R}^{K+1}, \tag{94}$$

*is well defined and linear, and $\mathcal{L}(\mathbf{v}) = \sum_{i=0}^K v_i \ell_i$ with $\ell_i(\mathbf{z}) = \boldsymbol{\phi}(\mathbf{z})^\top \mathbf{\Phi}^+ \mathbf{e}_i$, $\mathbf{\Phi}^+ = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top$. When $K+1 = q_p$ the fit interpolates and the $\ell_i$ are the Lagrange polynomials of $Y$. The* gradient poisedness constant *is*

$$\Lambda_1(Y) := \Delta_t \max_{\mathbf{z} \in \Omega_t} \sum_{i=0}^{K} \|\nabla \ell_i(\mathbf{z})\|. \tag{95}$$

Two facts are used. First, $\mathcal{L}$ reproduces $\mathcal{P}_p$: if $T \in \mathcal{P}_p$ then $\mathcal{L}((T(\mathbf{y}_i))_i) = T$, since $T$ attains the value zero in equation 94 and the minimizer is unique. Second, $\Lambda_1(Y)$ depends only on the shape of $Y$ and not on $\Delta_t$, since $\mathcal{P}_p$ is invariant under scaling. The server can compute it from the probes before broadcasting them.

**Lemma 12** (Polynomial KRR at $\lambda = 0$ is the least-squares fit)**.** *Let $k(\mathbf{z}', \mathbf{z}) = (\gamma \mathbf{z}'^{\top}\mathbf{z} + c_0)^p$ with $\gamma, c_0 > 0$ and let $Y$ be poised. For $\mathbf{v} \in \mathbb{R}^{K+1}$ with $v_0 = 0$, the anchored kernel fit $\widehat{m}(\mathbf{z}) = (\mathbf{k}(\mathbf{z}) - \mathbf{k}(\mathbf{0}))^{\top}\mathbf{K}^{+}\mathbf{v}$, where $\mathbf{K} = (k(\mathbf{y}_i, \mathbf{y}_j))_{ij}$ and $\mathbf{k}(\mathbf{z}) = (k(\mathbf{y}_i, \mathbf{z}))_i$, satisfies $\nabla\widehat{m} = \nabla\mathcal{L}(\mathbf{v})$.*

*Proof.* By the multinomial theorem $k(\mathbf{z}', \mathbf{z}) = \boldsymbol{\psi}(\mathbf{z}')^{\top}\boldsymbol{\psi}(\mathbf{z})$ with $\psi_\alpha(\mathbf{z}) = \sqrt{c_\alpha}\,\mathbf{z}^\alpha$, $|\alpha| \le p$, $c_\alpha > 0$, and the $\psi_\alpha$ form a basis of $\mathcal{P}_p$. Taking $\boldsymbol{\phi} = \boldsymbol{\psi}$ gives $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}$ and $\mathbf{k}(\mathbf{z}) = \boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{z})$. With $\boldsymbol{\Phi}$ of full column rank, $\mathbf{K}^{+} = (\boldsymbol{\Phi}^{+})^{\top}\boldsymbol{\Phi}^{+}$ and $\boldsymbol{\Phi}^{+}\boldsymbol{\Phi} = \mathbf{I}$, so $\mathbf{k}(\mathbf{z})^{\top}\mathbf{K}^{+}\mathbf{v} = \boldsymbol{\phi}(\mathbf{z})^{\top}\boldsymbol{\Phi}^{+}\mathbf{v} = \mathcal{L}(\mathbf{v})(\mathbf{z})$ by the normal equations. Subtracting the value at $\mathbf{0}$ shifts by a constant and does not change the gradient. □

*Proof of Lemma 1.* Let $T \in \mathcal{P}_p$ be the degree-$p$ Taylor polynomial of $f_t$ at $\mathbf{0}$ and $R := f_t - T$. Since $\nabla^p f_t$ is $L_p$-Lipschitz on the convex set $\Omega_t$ (Assumption 5(ii) and orthonormality of $\mathbf{V}_t$), the integral form of the Taylor remainder gives, for all $\mathbf{z} \in \Omega_t$,

$$|R(\mathbf{z})| \le \frac{L_p}{(p+1)!}\|\mathbf{z}\|^{p+1}, \qquad \|\nabla R(\mathbf{z})\| \le \frac{L_p}{p!}\|\mathbf{z}\|^{p}. \tag{96}$$

By Assumption 5(i) the response vector is $\mathbf{v} = (f_t(\mathbf{y}_i))_{i=0}^{K}$ with $v_0 = f_t(\mathbf{0}) = 0$, and by Assumption 5(iii) and Lemma 12, $\nabla\widehat{m}_t = \nabla\mathcal{L}(\mathbf{v})$. Linearity of $\mathcal{L}$ and reproduction of $T$ give

$$\nabla\widehat{m}_t = \nabla\mathcal{L}\big((T(\mathbf{y}_i))_i\big) + \nabla\mathcal{L}\big((R(\mathbf{y}_i))_i\big) = \nabla T + \sum_{i=0}^{K} R(\mathbf{y}_i)\,\nabla\ell_i.$$

Since $\nabla f_t = \nabla T + \nabla R$, $R(\mathbf{y}_0) = R(\mathbf{0}) = 0$ and $\|\mathbf{y}_i\| \le \Delta_t$, for every $\mathbf{z} \in \Omega_t$

$$\begin{aligned}\|\nabla f_t(\mathbf{z}) - \nabla\widehat{m}_t(\mathbf{z})\| &\le \|\nabla R(\mathbf{z})\| + \sum_{i=1}^{K} |R(\mathbf{y}_i)|\,\|\nabla\ell_i(\mathbf{z})\| \\ &\le \frac{L_p}{p!}\Delta_t^p + \frac{L_p}{(p+1)!}\Delta_t^{p+1}\sum_{i=0}^{K}\|\nabla\ell_i(\mathbf{z})\| \le \frac{L_p}{p!}\Delta_t^p + \frac{L_p}{(p+1)!}\Delta_t^{p+1}\cdot\frac{\Lambda_1(Y)}{\Delta_t},\end{aligned}$$

by equation 96 and equation 95. With $\Lambda_1(Y) \le \Lambda$ this is $\varepsilon_t \le L_p\big(\frac{1}{p!} + \frac{\Lambda}{(p+1)!}\big)\Delta_t^p = \kappa_{eg}\Delta_t^p$. Neither the number of probes nor the number of clients appears: the probes enter only through $\Lambda_1(Y)$, and the clients do not enter because $\mathbf{v}$ consists of exact values of $f_t$. □

### E.2 Federated logistic regression on the remaining benchmarks

Section 6.1 reports `a9a`, `w8a` and `ijcnn1`. Figure 4 gives the same comparison on `a5a` (Platt, 1999), `w1a` (Platt, 1999) and `phishing` (Mohammad & McCluskey, 2012), under identical settings: $N = 100$ clients, Dirichlet label skew $\alpha = 0.1$, five seeds, and the same baselines. `a5a` and `w1a` are the smaller variants of the Adult and Web collections whose larger variants appear in the main text, so the two groups differ in scale rather than in kind. The ordering of the methods is the same as in Figure 2 on all three benchmarks and in both participation regimes.

### E.3 Illustrative two-dimensional problems

We consider the following two-dimensional test functions

$$\begin{aligned}f_1(x, y) &= 8(x-1)^2(1.3x^2 + 2x + 1) + 0.5(y-4)^2, \\ f_2(x, y) &= (x^2 + y - 11)^2 + (x + y^2 - 7)^2, \\ f_3(x, y) &= (1-x)^2 + 100(y - x^2)^2.\end{aligned}$$

Here $f_1$ is the custom nonconvex function while $f_2$ and $f_3$ are respectively the Himmelblau and Rosenbrock functions.

Figure 5 shows optimization trajectories on the federated objectives. Constant step sizes fail to adapt to the valley's local curvature, with small steps converging slowly and large steps overshooting and diverging. By adaptively selecting the step size, our method reaches the optimum in far fewer iterations, avoiding both overly conservative and overly aggressive updates.

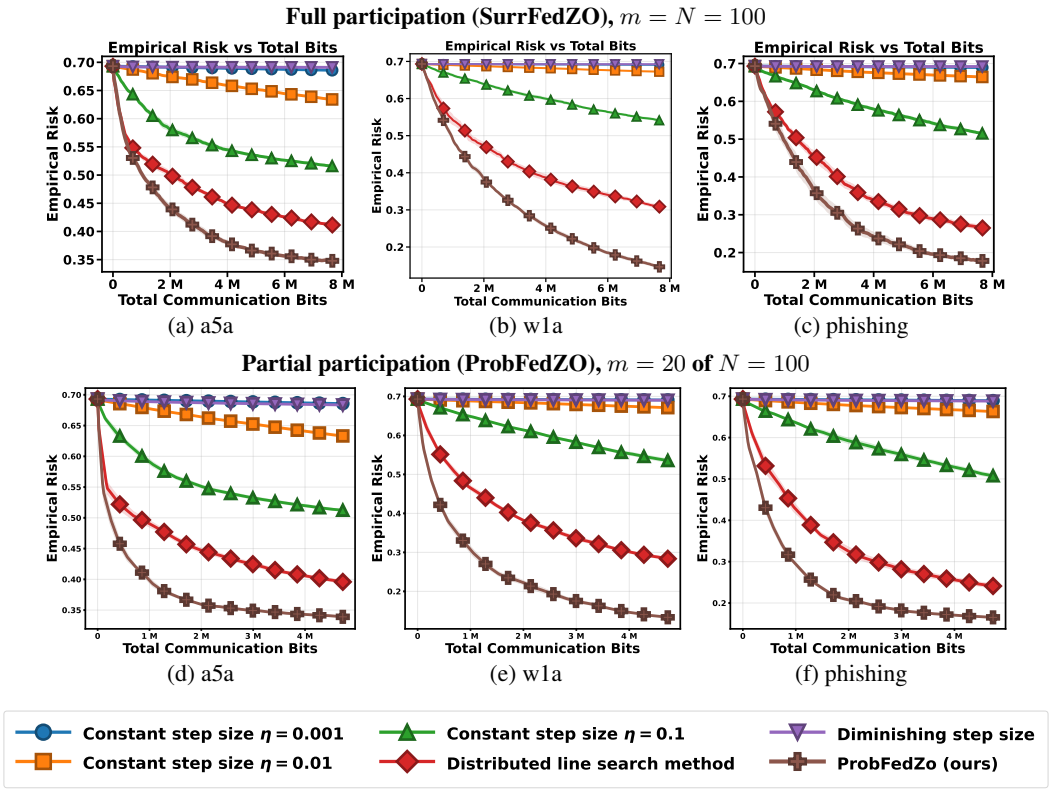
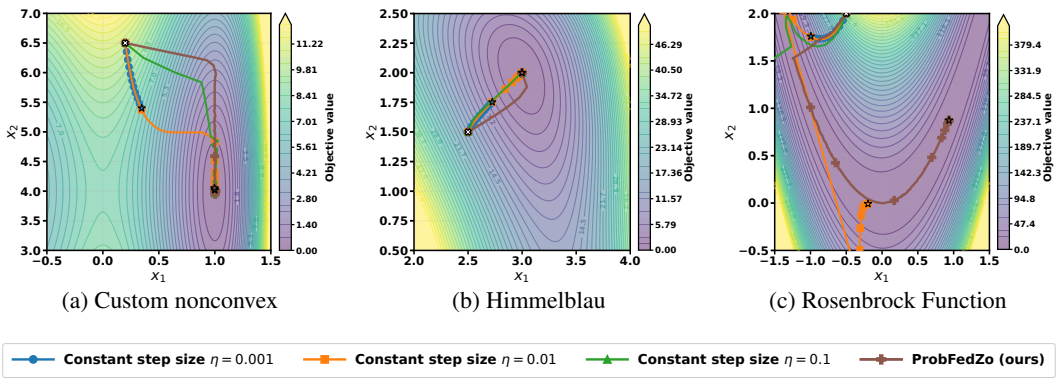



Figure 4: Federated logistic regression on the three benchmarks not shown in the main text. Empirical risk against total communication bits with mean $\pm$ standard error over five seeds. Panels (a)–(c) use SurrFedZO with all 100 clients, and panels (d)–(f) use ProbFedZO with 20 sampled clients. All panels share the legend below.




(a) Custom nonconvex (b) Himmelblau (c) Rosenbrock Function

Figure 5: Optimization trajectories on three two-dimensional test objectives. Here $\eta$ is the step-size magnitude.

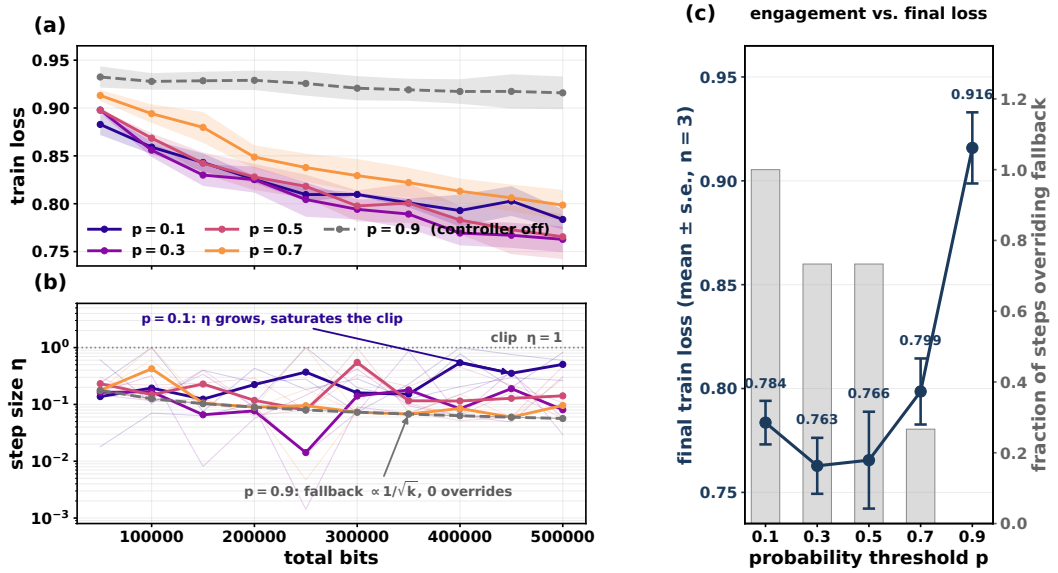



Figure 6: Ablation on the certification threshold $\rho$. (a) Training loss against total communication bits. (b) Selected step size $\eta$, with thin lines showing individual seeds and thick lines showing their means. At $\rho = 0.9$, the controller is inactive and follows the diminishing fallback. (c) Final training loss (mean $\pm$ standard error over five seeds) and the fraction of logged steps that override the fallback. Moderate thresholds engage the adaptive controller and achieve the lowest final loss.

### E.4 Ablation on the certification threshold

The threshold $\rho \in (0, 1)$ sets the uncertainty multiplier $\kappa = \sqrt{\rho/(1-\rho)}$ in the certification tests, so larger $\rho$ requires more standard errors of margin before a step is accepted in place of the diminishing fallback. We sweep $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, corresponding to $\kappa \in \{0.33, 0.65, 1.0, 1.53, 3.0\}$, holding everything else fixed at a budget of $5 \times 10^5$ bits. Figure 6 reports training loss, the realised step size $\eta$, and the rate at which the controller overrides the fallback.

The setting $\rho = 0.9$ does not behave as a conservative configuration of the method by disabling it. Its logged $\eta$ is identical across all five seeds at every logged point, and matches the fallback $\eta_t = 1/\sqrt{t}$ exactly. In other words, the controller accepts the fallback at all 30 logged opportunities and never once exercises its own proposal. This run is therefore the natural ablation of the mechanism, and it is also the worst performer by a wide margin. Its loss plateaus at 0.916 and improves by less than 0.02 over the entire budget, against 0.76–0.78 for the engaged settings. The gap to $\rho = 0.7$ is the only comparison in the sweep that is unambiguous at this sample size (paired $t = 20.0$, $\mathrm{df} = 2$).

Engagement falls monotonically in $\rho$. The fraction of logged steps on which the controller overrides the fallback is $100\%$, $73\%$, $73\%$, $27\%$ and $0\%$ for $\rho = 0.1$ through $0.9$. Final loss tracks this over the upper half of the range ($0.799$ at $\rho = 0.7$, $0.916$ at $\rho = 0.9$), and is flat over the lower half: $0.784$, $0.763$ and $0.766$ at $\rho = 0.1, 0.3, 0.5$. Pairwise comparisons among these three, taken paired by seed, give $|t| \leq 1.7$ against a critical value of $4.30$ at $\mathrm{df} = 2$, so we do not read an ordering into them. What the sweep supports is that performance is insensitive to $\rho$ throughout $[0.1, 0.5]$ and degrades once $\rho$ rises far enough to suppress the controller.

The $\eta$ traces show that this insensitivity in loss conceals two distinct failure modes at the ends of the engaged range. At $\rho = 0.1$ the step size loses its decay: $\log_{10} \eta$ has slope $+0.11$ per $10^5$ bits, against $-0.10$ for the fallback schedule, with the geometric mean rising from $0.19$ over the first half of training to $0.30$ over the second, and three of thirty logged values saturating the clip at $\eta = 1$. The controller is accepting increasingly aggressive steps and is no longer controlling anything, even though the loss penalty over this budget is small ($0.784$ vs. $0.763$) and within noise. At $\rho = 0.3$ the opposite occurs intermittently: three logged values fall to $1.4 \times 10^{-3}$, $8.1 \times 10^{-3}$ and $1.4 \times 10^{-2}$, one to two orders of magnitude below the fallback, and this setting has the largest dispersion in the sweep ($\mathrm{sd}\ \log_{10} \eta = 0.55$) despite the best mean behaviour. Both modes are visible in $\eta$ well before they would be visible in loss, which we take as an argument for monitoring the step size directly.

### E.5 Ablation on client data heterogeneity and the composition of rejections

**Setup.** The step-acceptance test is sampling-aware, so it declines to certify a step when the uncertainty induced by client sampling is too large for the Armijo condition to be resolved at the required confidence. If that mechanism is working as intended, its activity should scale with genuine disagreement between clients rather than acting as a constant brake. We test this by varying the Dirichlet concentration $\alpha$ used to partition the training data across clients, from $\alpha = 0.01$ (extreme label skew) to $\alpha = 100$ (nearly homogeneous), holding fixed the cubic kernel, $\rho = 0.7$, and seeds $42$–$46$ over $100$ rounds. We record two cumulative rates: the *fallback rate*, the fraction of rounds in which the proposed step was not certified and the diminishing-step fallback was taken, and the *uncertainty rejection rate*, the fraction of rounds in which the failure to certify was specifically attributable to sampling uncertainty rather than to the step failing the condition on its merits.

**The uncertainty rejection rate is monotone in heterogeneity.** Figure 7 plots this quantity directly. It decreases monotonically with $\alpha$ on all three benchmarks (Table 2), from roughly $0.50$ under extreme label skew to at most $0.034$ under near-homogeneity, and on `ijcnn1` it reaches exactly zero (not one of the $500$ rounds was rejected for reasons of sampling uncertainty). The curves separate by round $\sim$20 and remain separated thereafter, so the effect is a property of the regime rather than a transient of early training. Notably, the values at $\alpha = 0.01$ cluster tightly across three otherwise dissimilar datasets ($0.49$–$0.54$), suggesting that the rate saturates under extreme skew rather than growing without bound.

**Interpretation and cost.** Taken together, these results show that the uncertainty term tracks real client disagreement. It dominates the rejection budget precisely when the sampled clients disagree, and switches off almost entirely when they do not. This is the behavior required for the certification guarantee to be meaningful, since a criterion that rejected at a constant rate regardless of client agreement would provide no information about whether a step is safe. The cost is throughput. At $\alpha \leq 1$ the method spends roughly $60\%$ of rounds on the conservative fallback, and in that regime the certification threshold $\rho$ (Section E.6) is the available lever for trading reliability against progress.

### E.6 Ablation on the certification threshold $\rho$

The threshold $\rho$ is the only knob governing how much sampling uncertainty the Armijo test tolerates before it certifies a step. We ablate it over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on three LIBSVM benchmarks, running five seeds per setting for $100$ rounds each and pooling the per-round decisions across seeds ($N = 500$ per arm). A round counts as a *positive prediction* when the sampling-aware test certifies the step, and as a *positive outcome* when the exact equal-client federated objective decreases,

$$F(w_{t+1}) < F(w_t), \qquad F(w) = \frac{1}{N_c} \sum_{i=1}^{N_c} F_i(w). \tag{97}$$

Table 2: Ablation on client heterogeneity. "Rej. rate" is the final cumulative uncertainty rejection rate (mean ± standard error over five seeds, 100 rounds), expressed as a fraction of all rounds. "Share" is the percentage of all fallbacks attributable to sampling uncertainty. These are rejection statistics, not empirical risk. Cubic kernel, $\rho = 0.7$; smaller $\alpha$ is more heterogeneous.

| | a5a | | a9a | | ijcnn1 | |
|---|---|---|---|---|---|---|
| $\alpha$ | Rej. rate | Share (%) | Rej. rate | Share (%) | Rej. rate | Share (%) |
| 0.01 | 0.536 ±0.014 | 89.6 | 0.492 ±0.013 | 85.1 | 0.506 ±0.020 | 86.9 |
| 0.1 | 0.494 ±0.024 | 80.5 | 0.486 ±0.008 | 77.6 | 0.434 ±0.027 | 74.3 |
| 1 | 0.338 ±0.038 | 56.5 | 0.344 ±0.032 | 60.6 | 0.232 ±0.027 | 46.0 |
| 100 | 0.034 ±0.011 | 9.3 | 0.006 ±0.004 | 2.7 | 0.000 ±0.000 | 0.0 |

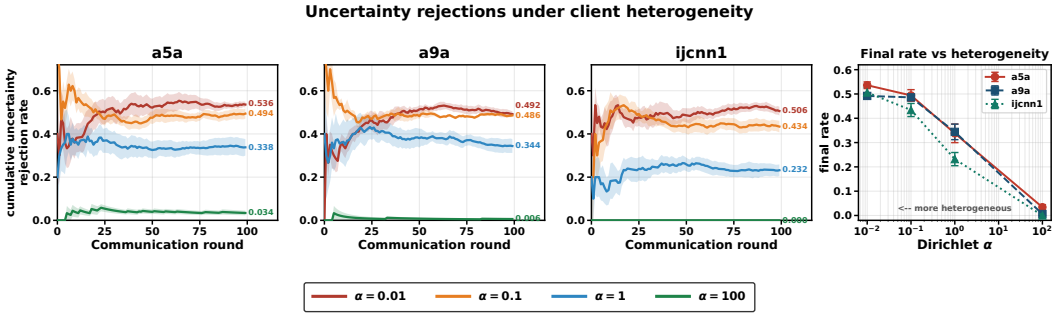



Figure 7: Cumulative uncertainty rejection rate under varying client heterogeneity, on three LIBSVM benchmarks. Curves are means over five seeds with ±1 standard error shading and end-of-run values are annotated. The right panel overlays the final rate against $\alpha$ on a logarithmic axis. The rate decreases monotonically as clients become more homogeneous and vanishes on `ijcnn1` at $\alpha = 100$, indicating that the uncertainty term responds to genuine client disagreement rather than rejecting at a fixed rate.

Precision is then the fraction of certified steps that genuinely reduced the objective. This directly tests the empirical reliability of the implemented acceptance rule, complementing the main risk-versus-communication comparisons. It does not test the true Armijo condition with the original coefficient $c$ or establish the probability guarantee under the theoretical calibration. Recall and the certification rate measure the throughput given up to obtain this reliability. All diagnostic evaluations in this section are excluded from the query and communication accounting.

**The threshold is a monotone reliability knob.** Precision increases monotonically with $\rho$ on every benchmark, from 0.771 to 0.861 on `phishing`, 0.834 to 0.925 on `ijcnn1`, and 0.889 to 0.981 on `w8a` (Table 3, Figure 8). Recall and the certification rate fall correspondingly. A practitioner can therefore position the method anywhere on the reliability/throughput curve by changing a single scalar, with no other retuning. We note that $F_1$ *decreases* with $\rho$ precisely because it penalises the conservatism the criterion is designed to supply; we report it for completeness but treat the precision/certification-rate pair as the operating curve.

**The test is selective, not merely conservative.** A threshold that simply rejected more steps would discard true and false certifications at the same rate. It does not. Moving from the permissive setting $\rho = 0.1$ to the strict end, false certifications fall by 55% on `phishing` (112 → 50), 65% on `ijcnn1` (81 → 28) and 87% on `w8a` (55 → 7), while 81–84% of the correctly certified steps are retained in every case. On `w8a` this reduces the false-certification rate from 11.0% to 1.4% of all rounds, a 7.9× reduction, at the cost of forgoing 19% of the useful steps. The uncertainty propagated from cross-client variability is thus informative about *which* candidate steps are unsafe, not merely about how many to reject.

Table 3: Ablation over the certification threshold $\rho$. Counts are pooled over five seeds ($N = 500$ decisions per arm). A round is a true positive when the probabilistic Armijo test certifies the step and the exact federated objective decreases, a false positive when it certifies but the objective does not decrease, a false negative when it does not certify yet the objective decreases, and a true negative when it neither certifies nor decreases. "Cert." is the fraction of rounds in which the test certified the step. Precision tends to rise with $\rho$.

| Dataset | $\rho$ | TP | FP | FN | TN | Precision | Recall | $F_1$ | Cert. (%) |
|---|---|---|---|---|---|---|---|---|---|
| phishing | 0.1 | 378 | 112 | 2 | 8 | 0.771 | 0.995 | 0.869 | 98.0 |
| | 0.3 | 371 | 110 | 8 | 11 | 0.771 | 0.979 | 0.863 | 96.2 |
| | 0.5 | 370 | 94 | 20 | 16 | 0.797 | 0.949 | 0.867 | 92.8 |
| | 0.7 | 359 | 86 | 35 | 20 | 0.807 | 0.911 | 0.856 | 89.0 |
| | 0.9 | 309 | 50 | 104 | 37 | **0.861** | 0.748 | 0.801 | 71.8 |
| ijcnn1 | 0.1 | 408 | 81 | 9 | 2 | 0.834 | 0.978 | 0.901 | 97.8 |
| | 0.3 | 380 | 61 | 50 | 9 | 0.862 | 0.884 | 0.873 | 88.2 |
| | 0.5 | 344 | 28 | 112 | 16 | 0.925 | 0.754 | 0.831 | 74.4 |
| | 0.7 | 380 | 61 | 50 | 9 | 0.862 | 0.884 | 0.873 | 88.2 |
| | 0.9 | 344 | 28 | 112 | 16 | **0.925** | 0.754 | 0.831 | 74.4 |
| w8a | 0.1 | 439 | 55 | 6 | 0 | 0.889 | 0.987 | 0.935 | 98.8 |
| | 0.3 | 436 | 37 | 20 | 7 | 0.922 | 0.956 | 0.935 | 94.6 |
| | 0.5 | 424 | 23 | 36 | 17 | 0.949 | 0.922 | 0.935 | 89.4 |
| | 0.7 | 355 | 7 | 125 | 13 | **0.981** | 0.740 | 0.843 | 72.4 |
| | 0.9 | 355 | 7 | 125 | 13 | **0.981** | 0.740 | 0.843 | 72.4 |

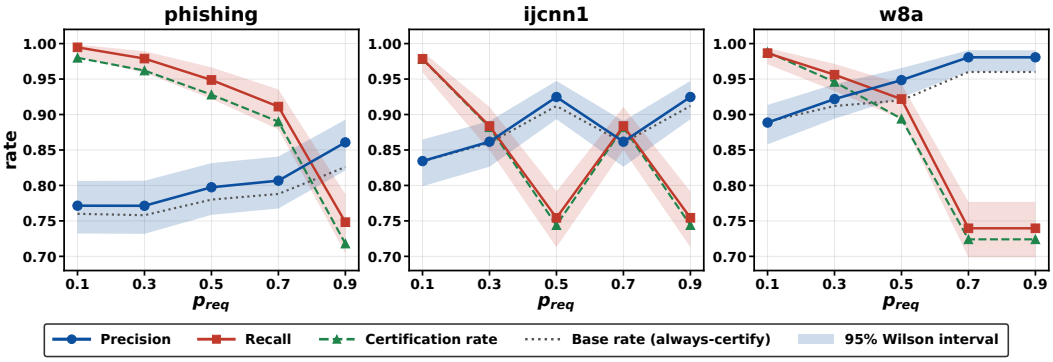



Figure 8: Ablation over $\rho$ on three LIBSVM benchmarks. Precision (blue) rises and both recall (red) and the certification rate (green) fall as the threshold tightens. Shaded regions are 95% Wilson intervals and the dotted line is the always-certify base rate. Counts are pooled over five seeds, $N = 500$ decisions per arm.

### E.7 Ablation on the surrogate kernel and accuracy of the subspace Hessian

**Setup.** The surrogate kernel determines the space in which the local model is fitted and therefore its curvature estimate $\widehat{\mathbf{H}}_t = \nabla^2_{\mathbf{z}} \widehat{m}_t(\mathbf{0})$. We compare four kernels—Matérn-5/2, RBF, and the degree-2 and degree-3 polynomial kernels—holding everything else fixed: the same probe budget, the same ridge $\lambda = 10^{-3}$, the same trust region, $\rho = 0.8$, and seeds 42–46 over 119 rounds. Accuracy is measured against the analytic projected Hessian of the selected-client aggregate $F_{S_t}$, $\mathbf{H}_t^{\text{true}} = \mathbf{V}_t^\top \nabla^2 F_{S_t}(\mathbf{w}_t)\mathbf{V}_t$, via

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \frac{\|\mathbf{H}_t^{\text{true}}\|_F^2}{\|\widehat{\mathbf{H}}_t - \mathbf{H}_t^{\text{true}}\|_F^2}, \tag{98}$$

giving $N = 500$ paired measurements per kernel and dataset. Note that all four are kernel ridge regressions. The fit remains linear in the observed loss differences, so the client-sampling covariance propagates in closed form to the joint uncertainty on value and slope, and the probabilistic Armijo machinery is unchanged by this choice.

Table 4: Kernel ablation. Surrogate Hessian SNR against the analytic selected-set reference, pooled over five seeds ($N = 500$ rounds per cell); $\rho = 0.8$, $\lambda = 10^{-3}$. "Neg." is the fraction of rounds with negative SNR, i.e. estimation error exceeding signal.

| Dataset | Kernel | Mean (dB) | Median (dB) | Min (dB) | Rel. err. | Neg. (%) |
|---|---|---|---|---|---|---|
| a5a | Matérn-5/2 | 3.97 | 3.45 | −4.28 | 63.3% | 8.0 |
| | RBF | 8.34 | 8.18 | −3.24 | 38.3% | 0.8 |
| | Quadratic | 12.79 | 12.51 | 0.93 | 22.9% | 0.0 |
| | Cubic | **15.83** | **15.76** | 1.78 | **16.2%** | **0.0** |
| a9a | Matérn-5/2 | 4.14 | 3.49 | −2.80 | 62.1% | 6.2 |
| | RBF | 8.68 | 8.53 | −1.86 | 36.8% | 0.2 |
| | Quadratic | 12.94 | 12.70 | 0.68 | 22.5% | 0.0 |
| | Cubic | **16.01** | **15.87** | 1.38 | **15.8%** | **0.0** |
| w1a | Matérn-5/2 | 0.65 | 0.68 | −15.35 | 92.8% | 43.2 |
| | RBF | 4.88 | 4.96 | −10.50 | 57.0% | 18.2 |
| | Quadratic | 11.98 | 12.09 | −3.50 | 25.2% | 1.0 |
| | Cubic | **15.08** | **15.37** | −2.39 | **17.6%** | **0.2** |

**Polynomial kernels dominate, and the ordering replicates exactly.** On all three benchmarks the ranking is cubic > quadratic > RBF > Matérn-5/2, and every pairwise gap is significant under a paired $t$-test ($p < 10^{-41}$ throughout; Table 4, Figure 9). The cubic kernel improves on Matérn-5/2 by 11.9, 11.9 and 14.4 dB on `a5a`, `a9a` and `w1a` respectively, reducing the relative Frobenius error from roughly 63% to 16%. The cubic–quadratic gap is 3.03, 3.07 and 3.11 dB across the three datasets—agreement to within 0.08 dB over five independent runs.

**The advantage is one of robustness, not only of average accuracy.** The gap widens where estimation is hardest. On `w1a` the smooth kernels degrade sharply. Matérn-5/2 falls to 0.65 dB and produces an estimate whose error exceeds its signal in 43.2% of rounds while the polynomial kernels lose less than 1 dB relative to the other two benchmarks and fail in at most 1% of rounds. Since the certified step depends on the curvature estimate, eliminating these breakdown rounds matters more for the method's reliability than the improvement in the mean.

**Interpretation.** The ordering tracks how well each reproducing space matches a low-order polynomial target. Over a trust region of radius $\delta$ the logistic objective is well approximated by a low-degree polynomial, so the quadratic space contains the leading behaviour and the cubic space additionally captures the next correction. The general-purpose smooth kernels must approximate the same target from a non-polynomial basis. With only $n_{\text{probe}}$ noisy zeroth-order evaluations per round, matching the basis to the target is worth substantially more than the additional flexibility of an infinite-dimensional smooth RKHS. We therefore adopt the cubic kernel as the default and report it throughout.

### E.8 Ablation on the search direction

**Setup.** Our method combines two components, a curvature-aware search direction built from the surrogate Hessian and a probabilistic line search that certifies the step under client-sampling uncertainty. To isolate the contribution of the first, we replace the absolute-Newton direction with the plain negative gradient while holding everything else fixed, namely the same probe budget, the same subspace sampling, the same kernel and ridge, and the same certification threshold $\rho$. Both variants therefore query the objective identically and differ only in how the accepted step is oriented. We report results on `rte` and `boolq` with Qwen3-0.6B-base at Dirichlet $\alpha = 0.1$, five seeds each.

**The curvature-aware direction accounts for the improvement.** At the 500k-bit budget the Newton direction reaches an empirical risk of 0.443 on `rte` against 0.512 for the gradient direction, a 13.5% relative reduction, and 0.527 against 0.563 on `boolq`, a 6.4% reduction (Table 5, Figure 10). The standard-error bands of the two variants do not overlap at any point on either task. The separation is present from the first logged budget, at which the Newton variant has already reached 0.511 on `rte` while the gradient variant stands at 0.559. Since the two arms share their initialization and

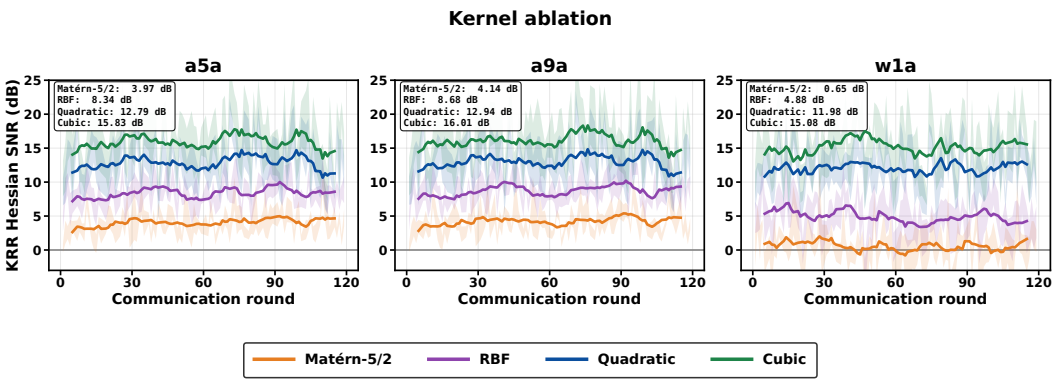



Figure 9: Surrogate subspace Hessian SNR against the analytic selected-set reference for four kernels, on three LIBSVM benchmarks. Curves are 9-round running means over five seeds. Shaded regions are $\pm 1$ standard error. The inset lists the per-kernel mean over all rounds. The ordering cubic $>$ quadratic $>$ RBF $>$ Matérn-5/2 holds on every dataset, and the smooth kernels degrade sharply on `w1a` while the polynomial kernels do not.

Table 5: Search-direction ablation on Qwen3-0.6B-base. Final empirical risk at a 500k-bit communication budget, mean $\pm$ standard deviation over five seeds. RTE is included in the main Qwen benchmark; BoolQ is an additional ablation task. These results do not cover the OPT-125M benchmark on AG News, CB and TREC. Both directions use the same probabilistic line search, probe budget and subspace sampling. Relative reduction is measured against the gradient direction.

| | | Empirical risk | | |
|---|---|---|---|---|
| Model | Dataset | Gradient | Newton (ours) | Rel. reduction |
| Qwen3-0.6B-base | `rte` | 0.512 $\pm$0.007 | **0.443** $\pm$0.007 | 13.5% |
| | `boolq` | 0.563 $\pm$0.005 | **0.527** $\pm$0.003 | 6.4% |

their query stream, this early gap cannot be attributed to a difference in starting point and instead reflects the value of the curvature estimate on the very first steps.

**Convergence behaviour differs, not only the endpoint.** The two variants also differ in shape. On `rte` the Newton direction reaches approximately 0.442 by 350k bits and is flat thereafter, whereas the gradient direction is still descending at the end of the budget. The comparison at 500k bits therefore contrasts a converged run against one that has not converged. We report the gap at the fixed budget because communication cost is the resource of interest, but we note that a larger budget would narrow it, and that the more precise statement is that the curvature-aware direction reaches a given risk level using substantially fewer communicated bits rather than that it attains a strictly better optimum.

**Scope.** This ablation supports the curvature-adaptation component of the main optimization claim, but does not isolate the contribution of certification. The complementary threshold sweeps examine its effect on optimization (Section E.4) and the empirical reliability of certified steps (Section E.6). The present direction comparison covers two Qwen3-0.6B-base tasks: RTE from the main benchmark and the additional BoolQ task. It does not evaluate OPT-125M on AG News, CB or TREC, nor Qwen3-0.6B-base on QNLI or SST-2; the direction-ablation results therefore should not be read as covering every model–dataset pair in the main study.

### E.9 Train and Test Accuracy in LLM Experiments

**Why the baseline accuracies are identical across seeds and step sizes.** Several baseline entries in Table 7 report the same accuracy on every seed, with zero standard deviation. This is not a failure to update the model. The corresponding empirical risk curves (Figure 3d) show that these methods do reduce the loss, and in some cases by a non-trivial amount, with the distributed line search dropping the training loss on `cb` from 2.567 to 2.270 over the budget. The explanation is that loss is

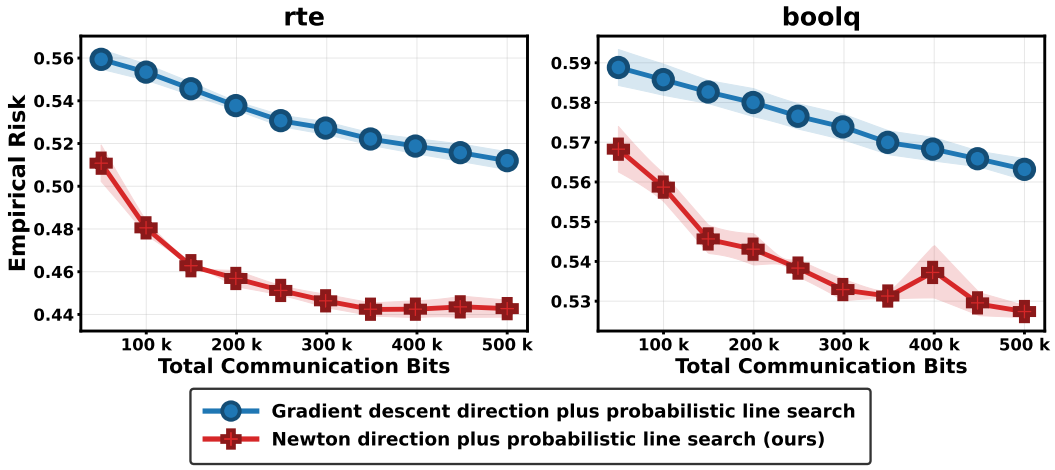



Figure 10: Empirical risk against total communication bits for the two search directions on Qwen3-0.6B-base with RTE and the additional BoolQ task, with all other components held fixed. Curves are means over five seeds with $\pm 1$ standard error shading. The bands do not overlap at any budget on either task.

continuous in the parameters whereas accuracy is a step function of the predictions. A method can move the weights, shift every logit, and lower the cross-entropy without changing the arg-max on a single evaluation example. Accuracy registers nothing until the parameters have moved far enough for some example to cross a decision boundary.

The effect is pronounced on the smaller tasks. The `cb` test split contains 56 examples, so one flipped prediction is worth 1.79 accuracy points, and the value at which the baselines are stuck, $42.86\%$, is exactly the majority-class rate. The same holds on `rte`, where the constant step sizes all report $71.12\%$, again the majority-class rate. These methods begin at a confident single-class solution and, over the communication budget available, never leave it. Their loss improves within that regime, but the predictions do not.

We regard this as the sharpest statement of the comparison rather than an artefact to be explained away. Under a fixed communication budget the relevant question is not whether an optimizer reduces its objective by some measurable amount, but whether it moves far enough to change what the model does. On `cb` and `mnli` the baselines do not, and remain at the majority-class rate, whereas our method reaches $77.4\%$ and $63.8\%$ respectively. On the tasks where the baselines do cross decision boundaries, namely `qnli`, `sst2` and `boolq`, the margins narrow correspondingly, and we report those cases without adjustment.

### E.10 Ablation on the probe radius $\delta$

**Setup.** The probe radius $\delta$ sets the scale at which the server queries the objective. All probes in a round are drawn inside a ball of radius $\delta$ in the sampled subspace, so $\delta$ controls a trade-off that has no obvious resolution a priori. A small radius keeps the queries in a region where the local model is accurate, but compresses the loss differences toward the client-sampling noise floor and so yields a poorly conditioned fit. A large radius produces well-separated function values, but the surrogate is then fitted over a region the low-order model no longer describes. We sweep $\delta \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$ on OPT-125M with the `cb` task under Dirichlet $\alpha = 0.1$, five seeds per setting, holding all other components fixed, and measure empirical risk against total communication bits.

**Progress improves with $\delta$ up to an interior optimum.** Reduction in empirical risk over the 500k-bit budget grows monotonically from $0.027$ at $\delta = 0.05$ to $0.124$ at $\delta = 0.5$, a factor of $4.7$, and then falls back to $0.079$ at $\delta = 1.0$ (Table 8, Figure 11). The optimum is therefore interior to the range tested rather than at its boundary, which is what licenses $\delta = 0.5$ as a default instead of merely the largest value we tried. The ordering is stable across seeds, and the standard-error bands of the three smallest radii are essentially disjoint.

Table 6: Final train and test accuracy (%) after a 500k-bit communication budget with Dirichlet $\alpha = 0.1$, grouped by model and its evaluation datasets. Values are means. Best available mean per column within each model in bold.

| OPT-125M | | | | | | |
|---|---|---|---|---|---|---|
| | AG News | | cb | | trec | |
| Method | Train | Test | Train | Test | Train | Test |
| $\eta = 10^{-7}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.1 | 2.6 |
| $\eta = 10^{-6}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.1 | 2.6 |
| $\eta = 10^{-5}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.1 | 2.6 |
| $\eta = 10^{-4}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.4 | 2.6 |
| $\eta = 10^{-3}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.6 | 2.6 |
| $\eta = 10^{-2}$ | 25.5 | 24.7 | 45.6 | 39.3 | 1.4 | 2.6 |
| $\eta = 10^{-1}$ | 25.5 | 24.6 | 45.3 | 39.3 | 1.4 | 2.6 |
| Distributed line search | 24.6 | 24.1 | 42.4 | 38.7 | 1.9 | 2.9 |
| Diminishing step | 25.7 | 24.5 | 44.4 | 40.5 | 1.4 | 2.6 |
| ProbFedZO (ours) | **35.5** | **33.7** | **47.8** | **45.2** | **27.8** | **26.6** |
| Qwen3-0.6B-base | | | | | | |
| | qnli | | rte | | sst2 | |
| Method | Train | Test | Train | Test | Train | Test |
| $\eta = 10^{-7}$ | 66.4 | 66.0 | 70.9 | 71.1 | 85.5 | 87.0 |
| $\eta = 10^{-6}$ | 66.4 | 66.0 | 70.9 | 71.1 | 85.5 | 87.0 |
| $\eta = 10^{-5}$ | 66.3 | 65.8 | 70.9 | 71.1 | 85.6 | 87.0 |
| $\eta = 10^{-4}$ | 65.0 | 64.7 | 70.9 | 71.1 | 85.6 | 87.0 |
| $\eta = 10^{-3}$ | 65.1 | 64.7 | 70.9 | 71.1 | 85.5 | 87.1 |
| $\eta = 10^{-2}$ | 65.0 | 64.7 | 70.9 | 71.1 | 85.4 | 87.1 |
| $\eta = 10^{-1}$ | 65.6 | 64.2 | 71.8 | 71.6 | 85.9 | 87.3 |
| Distributed line search | 73.5 | 69.2 | 78.0 | 78.2 | **90.6** | 89.7 |
| Diminishing step | 61.0 | 60.6 | 70.5 | 70.9 | 85.0 | 86.8 |
| ProbFedZO (ours) | **74.8** | **69.7** | **80.0** | **79.7** | 90.4 | **90.3** |

**The failure mode at large radius is informative.** The $\delta = 1.0$ curve does not simply underperform. It begins at the lowest risk of any setting, $0.883$ against $0.936$ for $\delta = 0.05$, and falls quickly to roughly $0.86$ within the first $90$k bits, then stalls for approximately $110$k bits before resuming a slower descent. Aggressive early progress followed by a plateau is the signature we would predict when the surrogate is fitted over a region in which the objective departs from its low-order approximation. Consistent with this reading, $\delta = 1.0$ also carries the widest standard-error band of any setting, since a trust region that exceeds the range of model validity makes the trajectory more sensitive to the particular probes drawn.

**Relation to the kernel ablation.** This ablation and the kernel comparison of Section E.7 concern the same underlying quantity, namely how far the objective departs from a low-degree polynomial over the queried region. The cubic kernel outperformed the quadratic by roughly $3\,$dB in Hessian SNR at the default radius, which we attributed to a non-negligible third-order term, and the present sweep independently shows that useful progress requires a radius large enough for that term to matter. We regard the two as consistent evidence for the same conclusion, though we note that a direct test, namely repeating the kernel comparison at $\delta = 0.05$ and confirming that the cubic advantage shrinks, would establish the connection rather than infer it.

Table 7: Final test accuracy (%) after a 500k-bit communication budget with Dirichlet $\alpha = 0.1$, grouped by model and its evaluation datasets. Available results are mean $\pm$ standard deviation over five seeds.

| OPT-125M: test accuracy (%) | | | |
|---|---|---|---|
| Method | AG News | cb | trec |
| $\eta = 10^{-7}$ | 24.7 $\pm$0.0 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0$^\dagger$ |
| $\eta = 10^{-6}$ | 24.7 $\pm$0.0 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0$^\dagger$ |
| $\eta = 10^{-5}$ | 24.7 $\pm$0.0 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0$^\dagger$ |
| $\eta = 10^{-4}$ | 24.7 $\pm$0.0 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0 |
| $\eta = 10^{-3}$ | 24.7 $\pm$2.3 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0$^\dagger$ |
| $\eta = 10^{-2}$ | 24.7 $\pm$2.3 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0 |
| $\eta = 10^{-1}$ | 24.6 $\pm$2.4 | 39.3 $\pm$0.0 | 2.6 $\pm$0.0 |
| Distributed line search | 24.1 $\pm$2.6 | 38.7 $\pm$1.0 | 2.9 $\pm$0.4 |
| Diminishing step | 24.5 $\pm$2.1 | 40.5 $\pm$2.1 | 2.6 $\pm$0.1 |
| ProbFedZO (ours) | **33.7** $\pm$4.4 | **45.2** $\pm$4.1 | **26.6** $\pm$7.2 |
| Qwen3-0.6B-base: test accuracy (%) | | | |
| Method | qnli | rte | sst2 |
| $\eta = 10^{-7}$ | 66.0 $\pm$4.2 | 71.1 $\pm$0.0 | 87.0 $\pm$1.5 |
| $\eta = 10^{-6}$ | 66.0 $\pm$4.2 | 71.1 $\pm$0.0 | 87.0 $\pm$1.5 |
| $\eta = 10^{-5}$ | 65.8 $\pm$4.0 | 71.1 $\pm$0.0 | 87.0 $\pm$1.5 |
| $\eta = 10^{-4}$ | 64.7 $\pm$3.5 | 71.1 $\pm$0.0 | 87.0 $\pm$1.5 |
| $\eta = 10^{-3}$ | 64.7 $\pm$3.8 | 71.1 $\pm$0.0 | 87.1 $\pm$1.3 |
| $\eta = 10^{-2}$ | 64.7 $\pm$3.8 | 71.1 $\pm$0.0 | 87.1 $\pm$1.3 |
| $\eta = 10^{-1}$ | 64.2 $\pm$2.4 | 71.6 $\pm$0.2 | 87.3 $\pm$0.9 |
| Distributed line search | 65.3 $\pm$3.5 | 71.5 $\pm$0.0 | 89.7 $\pm$0.9$^\dagger$ |
| Diminishing step | 60.6 $\pm$3.9 | 70.9 $\pm$2.7 | 86.8 $\pm$1.0 |
| ProbFedZO (ours) | **69.7** $\pm$4.2 | **79.7** $\pm$0.8 | **90.3** $\pm$0.3 |

Table 8: Ablation on the probe radius $\delta$. Empirical risk on OPT-125M/cb at $\alpha = 0.1$, averaged over five seeds. "Drop" is the reduction in risk over the 500k-bit budget, and "s.e." is the standard error of the final value. Best result in bold.

| $\delta$ | Initial risk | Final risk | Drop | s.e. |
|---|---|---|---|---|
| 0.05 | 0.9357 | 0.9092 | 0.0265 | 0.0081 |
| 0.1 | 0.9330 | 0.8837 | 0.0493 | 0.0107 |
| 0.2 | 0.9283 | 0.8574 | 0.0710 | 0.0007 |
| 0.5 | 0.9122 | **0.7882** | **0.1240** | 0.0146 |
| 1.0 | 0.8827 | 0.8042 | 0.0785 | 0.0202 |

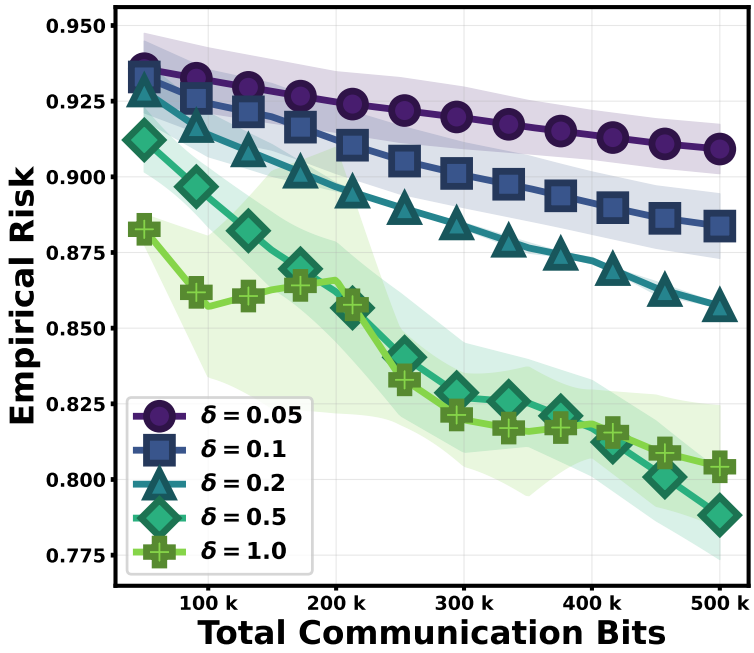



Figure 11: Empirical risk against total communication bits for five probe radii, on OPT-125M/cb at $\alpha = 0.1$. Curves are means over five seeds with $\pm 1$ standard error shading. Progress improves with $\delta$ up to $0.5$ and degrades at $1.0$, where an initially rapid descent stalls between roughly $90$k and $200$k bits.