BEYOND AGGREGATE DRIFT: OPTIMAL CLIENT-SPECIFIC WINDOWS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study federated learning from continuously arriving data: each client samples from its own distribution, these distributions drift over time, and the server must track the minimizer of the clients' weighted average loss. Clients are not alike as they collect data at different rates, with different noise, and change at different speeds. The main question is how much history should each one keep? A single shared window must serve every client at once, and so makes a fast-changing client carry the sampling noise of a slow, nearly stationary one.

We give each client a memory of its own, sized to that client's noise, arrival rate, and drift, and refresh it by restarted accelerated method that starts from a common point. The tracking error is then bounded by

$$R\left(\sum_m w_m^{3/2}\,\sigma_m\sqrt{\gamma_m/\lambda_m}\right)^{2/3},$$

for clients with weight w_m, noise σ_m, arrival rate λ_m, and drift rate γ_m on a domain of diameter R. This is never worse than what any common window can achieve, and strictly better whenever noise and drift are spread unevenly across clients.

A lower bound that puts different groups of clients on different drift clocks shows this quantity is the right one. The two bounds agree up to a constant whenever the clients occupy boundedly many refresh scales, and in general differ only by the sixth root of the number of scales. Experiments show the gain comes from the memory rather than from the solver: it persists across accelerated, stochastic-gradient, and one-step methods, and against a separately tuned common-memory baseline.

1 INTRODUCTION

Many learning systems receive data continuously from several clients whose statistics change over time: sensors in a monitoring network, institutions or devices contributing to a federated model, users of a recommendation service. Because the target is the *current* population objective, an observation loses value as the distribution that produced it moves away, and the learner must decide how much history to keep. For a single stream the classical answer is a sliding window whose length balances the bias caused by drift against the variance of estimating from few samples (Gama et al., 2014; Mazzetto & Upfal, 2023a). But clients rarely change at the same speed: one sensor may supply a few noisy observations of a population that hardly moves, while another supplies many observations of a population that changes quickly. One common memory for all clients then faces a dilemma. A short memory throws away the stable client's history, which is still informative, whereas a long memory keeps the changing client's history, which is already stale. The statistical question behind this dilemma is how long each client's past observations remain useful, and what a common memory costs.

The problem. There are M clients with public weights w_m. Client m supplies independent time-stamped samples at rate λ_m from a distribution $\mathcal{D}_m^t$ that moves in time. The server tracks the weighted

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.

population objective $F^t = \sum_m w_m f_m^t$, where f_m^t is the expected loss under $\mathcal{D}_m^t$ and F^t is convex on a compact domain of diameter D, by reporting a causal trajectory $(x_t)_{t\geq 0}$: the point reported at time t may depend only on information the server has received by then, and each synchronization with the clients takes τ time units. The learner is charged the time-averaged suboptimality of x_t with respect to the current objective F^t, a continuous-time counterpart of dynamic regret. Each client is described by three public numbers: its rate λ_m, a variance bound σ_m^2 for its stochastic gradients, and a bound γ_m on the speed of its population gradient. The learner's decision is the *data memory*: which of each client's past samples enter the stochastic problem that the server solves when it refreshes its report.

What is known, and what is not. With one common memory of length h, the single-stream bias–variance balance carries over: the aggregate drift $\bar{\gamma} = \sum_m w_m \gamma_m$ is balanced against the pooled uncertainty $\Xi^2 = \sum_m w_m^2 \sigma_m^2 / \lambda_m$, and the best common length yields the upper-bound scale $D\bar{\gamma}^{1/3}\Xi^{2/3}$ (Appendix K). This scale is loose whenever the noisiest clients are the ones that barely change, since it charges their uncertainty against the drift of the others. Whether the looseness is real cannot be decided by a lower bound that lets an adversary reallocate a total drift budget among the clients, because here each client's drift bound is prescribed separately. Two further obstacles stand between the natural remedy, a separate memory per client, and a guarantee: retained samples are reused across refreshes, so the iterates depend on the data they are evaluated on, and the drift bounds may be unknown in practice.

Contributions. We characterize what client-specific memories gain over a common one, and we show how to select among them without knowing the drift bounds.

1. **A heterogeneous tracking quantity.** Separate completed data windows, a reset stochastic solve after each refresh, and disjoint batches within each solve attain
$$D\mathcal{H}_\gamma = D\Big(\sum_m w_m^{3/2}\sigma_m\sqrt{\gamma_m/\lambda_m}\Big)^{2/3} \leq D\bar{\gamma}^{1/3}\Xi^{2/3}$$
with explicit optimization, sample-count, and elapsed-time terms (Theorem 4.1). The argument reuses stored samples across solves under pointwise gradient-moment assumptions only, and it turns the uncertainty of a stationary client into a vanishing finite-horizon transient; projected SGD attains the same quantity.
2. **Lower bounds that respect the client vector.** Independent client groups with unequal clocks give a group-sensitive lower bound (Corollary 5.1) whose uniform simplification is a constant multiple of $D\mathcal{H}_\gamma/J^{1/6}$ for J dyadic memory scales, and an explicit joint regime controls every term of the upper bound. Separately, an exact periodic calculation shows that the risk ratio between common and client-specific uniform completed-window estimators is unbounded (Proposition 5.2) and yields a test for when forgetting helps.
3. **Selection without drift inputs.** Under a bounded-loss assumption, independent validation and causal expert aggregation compete with the best fixed joint memory vector (Theorem 6.1), retaining the pooled-variance comparison and paying explicit validation and source-age costs; the aggregation is standard, and the point is its validity under causal reuse.
4. **Evidence and boundaries.** Prospective tests on periodic streams confirm predicted benefits and failures of a specified memory rule, and controlled experiments separate the memory mechanism from the solver. On temperature laws inside the regime of Corollary B.3, the windows of Corollary 4.2 cut a December gap by 51% and 52% against the bound-optimal and tuned common lengths; the gain tracks the predicted heterogeneity and shrinks when the clocks agree. A clinic network with unequal drift shows the same window shape cutting the gap under exact buffered minimization, with a smaller reduction once Algorithm 1 leaves an optimization remainder. These comparisons establish no universal advantage for client-specific windows.

2 Related work

This paper differs from prior guarantees in its prescribed clientwise drift vector, simultaneously running unequal clocks, and the resulting pooled statistical quantity, all introduced in Sections 3–6; Table 5 in Appendix L compares assumptions and criteria.

Streaming and federated learning. Marfoq et al. (2023) study fixed-weight population targets with local caches and sample reuse; their laws are stationary, and their optimization analysis relies on uniform gradient-discrepancy control. Huynh et al. (2025) treat dependent Markovian streams and nonconvex objectives relative to fixed stationary laws, and Zhao & Liang (2026) optimize admission and retention under sampling and buffer budgets for stationary independent streams; we track persistently changing populations from independent samples with uncharged storage and computation, and none of these models contains the others. LASG's selective communication (Chen et al., 2020) and FedDrift's multi-model drift clustering (Jothimurugesan et al., 2023) address different mechanisms; we contact every client and track one fixed-weight objective.

Drift and adaptation. Master-FL (Ganguly & Aggarwal, 2024) uses detection and restarts to control dynamic regret under unknown aggregate variation of empirical losses, whereas our target fixes population weights and separates clientwise gradient drift from information rates. Mazzetto & Upfal (2023a) select a recent window under uniform convergence and discrepancy assumptions. Our adaptive result instead competes with fixed joint memory experts through an independent validation stream. The aggregation and the delayed-feedback reduction are standard (Freund & Schapire, 1997; Joulani et al., 2013); the contribution is that the pooled statistical comparison survives in the causal reuse model, not a new drift-detection principle.

Optimization and lower bounds. Acceleration with reciprocal-count variance control is established (Tyurin, 2025), and single-stream tracking has guarantees under strong convexity (Cutler et al., 2023); historical reweighting admits discrepancy and complexity bounds (Awasthi et al., 2023), and Mazzetto & Upfal (2023b) prove average-risk block lower bounds for drifting distributions, whereas our construction retains unequal client-group clocks under a prescribed drift vector.

3 SETTING

Objective. There are M clients with known weights $w_m > 0$, $\sum_m w_m = 1$. On a compact convex set $\mathcal{X} \subset \mathbb{R}^d$ of diameter $D > 0$, with Euclidean projection $\Pi_{\mathcal{X}}$ and Euclidean norms throughout, we track

$$F^t(x) = \sum_{m=1}^{M} w_m f_m^t(x), \qquad f_m^t(x) = \mathbb{E}_{\xi \sim \mathcal{D}_m^t}\, \ell(x;\xi), \qquad t \geq 0.$$

The sample loss ℓ is public. The population path $(\mathcal{D}_m^t)_{m,t}$ is fixed before any sample noise is drawn and is never revealed to the learner. Each F^t is convex and L-smooth, $L > 0$, and a public bound $\|\nabla F^t(x)\| \leq G$ holds for all $t \geq 0$ and $x \in \mathcal{X}$; compactness guarantees that the current minimizer exists, and the gradient bound controls the loss before the first data-based report.

Independent time-stamped samples. Client m has a locally finite arrival count $N_m(t)$ with $N_m(0) = 0$. Conditionally on the complete exogenous array of arrival times, all samples are independent across arrivals and clients, and a sample arriving at time s has law $\mathcal{D}_m^s$. A stored sample may be evaluated at any later query point, and the oracle assumptions are pointwise: for each fixed $x \in \mathcal{X}$ and $\xi \sim \mathcal{D}_m^s$,

$$\mathbb{E}\nabla\ell(x;\xi) = \nabla f_m^s(x), \qquad \mathbb{E}\|\nabla\ell(x;\xi) - \nabla f_m^s(x)\|^2 \leq \sigma_m^2. \tag{1}$$

No uniform concentration over sample-dependent query points is assumed.

Prescribed clientwise drift. A public vector $\gamma = (\gamma_m)_{m=1}^M$ with $\gamma_m \geq 0$ bounds the speed of each client's population gradient:

$$\|\nabla f_m^t(x) - \nabla f_m^s(x)\| \leq \gamma_m |t - s|, \qquad x \in \mathcal{X},\ s, t \geq 0. \tag{2}$$

The bound is imposed client by client; an adversary may not reallocate the aggregate budget $\bar{\gamma} := \sum_m w_m \gamma_m$ across clients. The fixed-memory rule of Section 4 takes certified noise and drift bounds as inputs; Section 6 removes the drift bounds from the inputs under an additional bounded-loss assumption and an independent validation stream, without estimating them from the optimization buffers.

Definition 3.1 (Regular arrivals)**.** For known positive rates $(\lambda_m)_{m=1}^M$, with $\lambda_{\min} = \min_m \lambda_m$, a realized schedule is λ-regular if, for every m and all $t - s \geq 2/\lambda_m$,

$$\tfrac{1}{2}\lambda_m(t-s) \leq N_m(t) - N_m(s) \leq 2\lambda_m(t-s).$$

The schedule $N_m(t) = \lfloor \lambda_m t \rfloor$ is λ-regular, whereas a Poisson path in general is not; the lower inequality guarantees enough buffered data, and the upper one limits the information available in a hard instance. We work conditionally on a fixed regular schedule, and all expectations concern the sample noise and the learner's private randomization.

Communication and loss. One synchronization (client messages followed by a server broadcast) costs $\tau \geq 0$ time units, regardless of the message sizes. Clients may compute arbitrarily on their samples and the broadcast state. Computation and storage carry no time charge, and raw samples may be transmitted, so raw buffers may be pooled at the server. Consequently, the round counts we report describe one implementation. The reported point may depend only on information already received at the server, and its loss is

$$\mathrm{AvgErr}_T = \frac{1}{T}\int_0^T \mathbb{E}\Big[F^t(x_t) - \min_{x\in\mathcal{X}} F^t(x)\Big]\, dt. \tag{3}$$

Two statistical-drift quantities. For each client define

$$a_m = w_m\gamma_m, \qquad b_m = \frac{w_m^2\sigma_m^2}{\lambda_m}, \qquad \mathcal{S}_\gamma = \sum_m \sqrt{a_m b_m}, \qquad \mathcal{H}_\gamma = \mathcal{S}_\gamma^{2/3}, \tag{4}$$

so that $\bar\gamma = \sum_m a_m$, $\Xi^2 := \sum_m b_m$ is the pooled uncertainty of the introduction, and $\mathcal{H}_\gamma$ is the quantity displayed there. Client-specific memories replace the common-memory scale $\bar\gamma^{1/3}\Xi^{2/3}$ (derived in Appendix K) by $\mathcal{H}_\gamma$, and by the Cauchy–Schwarz inequality

$$\mathcal{H}_\gamma \leq \bar\gamma^{1/3}\Xi^{2/3}, \tag{5}$$

with equality exactly when the vectors $(a_m)_m$ and $(b_m)_m$ are proportional. The gap can be large: for two clients with $a_1 > 0$, $a_2 = 0$ and $b_1, b_2 > 0$, the client-specific quantity is $(a_1 b_1)^{1/3}$, whereas the common-memory quantity is $[a_1(b_1+b_2)]^{1/3}$. The stationary client's uncertainty b_2 has left the drift floor altogether; Theorem 4.1 shows that it reappears only as a finite-horizon transient.

4 Client-specific memory under known drift bounds

A common memory ties the bias of a rapidly changing client to the noise of every other client. When each client's drift bound is prescribed, a different balance is possible: each client keeps its own window, and the server solves a fresh stochastic problem on the frozen windows whenever it refreshes its report.

4.1 Data buffers, reset solves, and disjoint batches

Fix an integer $K \geq 1$ (gradient rounds per solve), client memory lengths $h_m \geq 2K/\lambda_m$, and a reporting stride $\Delta > 0$ with $\Delta \geq K\tau$. The memory of client m is a *completed window*: at each time ph_m, $p = 1, 2, \ldots$, the buffer is replaced by the observations that arrived in $((p-1)h_m, ph_m]$, and before the first complete window the buffer holds the arrived prefix. A client with $\gamma_m = 0$ may use $h_m = \infty$ and keep its entire prefix; we then read $1/h_m = a_m h_m = 0$.

Solve $j = 0, 1, \ldots$ starts at $s_j = s_0 + j\Delta$ with $s_0 = 2K/\lambda_{\min}$. At s_j the server freezes a copy of every client's current buffer, which is the latest completed window or, if $s_j < h_m$, the prefix. The frozen interval of client m has length $h_{m,j} = \min\{s_j, h_m\}$, contains $n_{m,j} \geq K$ observations on a λ-regular schedule, and consists of timestamps in $(s_j - 2h_{m,j}, s_j]$; buffer replacements during the solve do not affect the frozen copy. The solve lasts $c := K\tau$ time units, its output is reported on the interval $\mathcal{I}_j = [t_j, t_{j+1})$ with $t_j := s_j + c$, and a fixed public point $x_\circ \in \mathcal{X}$ is reported before t_0.

Algorithm 1 states the solve. Because $y^1 = x_\circ$ is fixed, the batch counts travel with the first gradient message and the server knows η_j before its first update; every round contacts all clients and costs τ,

Algorithm 1 Client-specific data refresh with reset accelerated solves

Require: $x_\circ \in \mathcal{X}$, K, (h_m), $\Delta > 0$ with $\Delta \geq K\tau$, L, certified (σ_m)

for solves $j = 0, 1, \ldots$ starting at $s_j = 2K/\lambda_{\min} + j\Delta$ **do**
  freeze the client buffers; split each into K nonempty batches $\mathcal{B}_{m,j,1}, \ldots, \mathcal{B}_{m,j,K}$ of sizes differing by at most one, by timestamp order only
  reset $x^0 = v^0 = x_\circ$
  for $k = 1, \ldots, K$ **do**
    $\theta_k = 2/(k+1)$; $\quad y^k = (1-\theta_k)x^{k-1} + \theta_k v^{k-1}$
    each client returns the mean gradient of batch $\mathcal{B}_{m,j,k}$ at y^k; the server forms the w-weighted aggregate g^k
    if $k = 1$ **then**
      clients also send their batch-size vectors; the server sets $V_{j,l} = \sum_m w_m^2 \sigma_m^2 / |\mathcal{B}_{m,j,l}|$ for $l = 1, \ldots, K$, $S_j = \sum_{l=1}^K l^2 V_{j,l}$, and $\eta_j = \min\{1/(4L),\, D/\sqrt{2S_j}\}$ (with $\eta_j = 1/(4L)$ if $S_j = 0$)
    end if
    $v^k = \Pi_{\mathcal{X}}(v^{k-1} - k\eta_j g^k)$; $\quad x^k = (1-\theta_k)x^{k-1} + \theta_k v^k$
  end for
  report $x_j = x^K$ on $\mathcal{I}_j = [s_j + K\tau,\, s_{j+1} + K\tau)$
end for

and no round waits for fresh samples. Each solve resets its iterates and gradient state: an observation may reappear in later solves, but within one solve the K batches are disjoint. The reset is what permits this reuse under the pointwise assumptions (1): query y^k depends only on the earlier batches of the *same* solve, so the analysis reveals them in order without conditioning on previous outputs that used overlapping data (Appendix C). Since messages are unrestricted, the frozen buffers can instead be uploaded in one synchronization and the same updates run at the server, in which case the bounds below hold with $K\tau$ replaced by τ (Appendix B.1): the elapsed time $K\tau$ belongs to the gradient-message implementation, not to the model.

4.2 Tracking guarantee and optimal memory lengths

Theorem 4.1 (Client-specific refresh)**.** *Under the assumptions of Section 3, condition all expectations on a fixed λ-regular schedule and let $c = K\tau$ and $\beta_0 = 4\sqrt{2/3}$. For Algorithm 1, every $t \in \mathcal{I}_j$ satisfies*

$$\mathbb{E}[F^t(x_j) - \min_{\mathcal{X}} F^t] \leq \frac{4LD^2}{K(K+1)} + \beta_0 D \sqrt{\sum_m \frac{b_m}{h_{m,j}}} + 2D \sum_m a_m h_{m,j} + D\bar{\gamma}(c + \Delta), \tag{6}$$

and for every $T \geq s_0 + c$ the reported trajectory satisfies

$$\mathrm{AvgErr}_T \leq \frac{4LD^2}{K(K+1)} + D\Big(2\sum_m a_m h_m + \beta_0 \sqrt{\sum_m b_m/h_m}\Big) + D\bar{\gamma}(c+\Delta) + \mathcal{T}_T, \tag{7}$$

$$\mathcal{T}_T := \beta_0 D\, \Xi\Big(\frac{2}{\sqrt{T}} + \frac{\Delta}{T\sqrt{s_0}}\Big) + \frac{GD(s_0 + c)}{T}. \tag{8}$$

The proof (Appendix C) combines a projected accelerated potential for a stochastic oracle with bounded gradient bias and a filtration that exposes only the disjoint batches of the current solve. The bias bound, controlled client by client through the sample ages (including the solve and reporting lag), enters multiplied by D independently of the number of internal iterations, and the stochastic term comes from the independent batches. The four terms of (7) are the optimization error, the statistical-drift balance, the elapsed-time cost, and the finite-horizon transient $\mathcal{T}_T$; a client with an infinite window appears only in the transient, through Ξ, so stationary clients are never charged a permanent drift floor.

Corollary 4.2 (Optimized heterogeneous windows)**.** *Suppose $\mathcal{S}_\gamma > 0$ and $\sigma_m > 0$ for every client with $\gamma_m > 0$. For such clients set*

$$h_m^\star = (2/3)^{1/3} \mathcal{S}_\gamma^{-1/3} \sqrt{b_m/a_m}, \qquad h_m = \max\{h_m^\star,\, 2K/\lambda_m\},$$

and set $h_m = \infty$ when $\gamma_m = 0$. Then

$$\mathrm{AvgErr}_T \le 6(2/3)^{1/3} D\mathcal{H}_\gamma + \frac{4LD^2}{K(K+1)} + 4KD\sum_m \frac{a_m}{\lambda_m} + D\bar{\gamma}(K\tau + \Delta) + \mathcal{T}_T. \tag{9}$$

Before the sample-count floors are imposed, $(h_m^\star)$ exactly minimizes the statistical-drift expression in parentheses in (7).

The optimal lengths satisfy $h_m^\star/h_n^\star = \sqrt{b_m a_n/(a_m b_n)}$: a client receives a longer memory when its weighted uncertainty is large relative to its drift. For $\tau > 0$ one may take $\Delta = K\tau$, and the nontransient remainders in (9) are then at most $4LD^2/K^2 + DK(4\sum_m a_m/\lambda_m + 2\bar{\gamma}\tau)$; they record the sample requirements and elapsed time of the gradient-message implementation (the K-dependent communication charge disappears under the raw-buffer implementation), and we do not claim that they are minimax sharp; comparison (5) concerns the statistical-drift term alone. The solver is not the mechanism: reset projected SGD on the same disjoint batches, averaged over its post-update iterates, satisfies the same guarantees with LD^2/K in place of $4LD^2/[K(K+1)]$ and $\beta_{\mathrm{sgd}} = 2\sqrt{2}$ in place of β_0 (Proposition B.1), so $\mathcal{H}_\gamma$ is determined by the memories and the independent batches, and acceleration only improves the optimization remainder.

5 Is client-specific memory necessary?

Theorem 4.1 and inequality (5) compare upper bounds. This section gives two converse statements: a minimax lower bound that matches $D\mathcal{H}_\gamma$, in an explicit regime, up to a factor depending only on the number of distinct memory scales, and an exact periodic calculation in which the best shared window loses to a client-specific policy by an unbounded factor.

5.1 A minimax lower bound with fixed clientwise drift

The comparison class. For a fixed schedule N and horizon T, let $\mathfrak{P}_T(\mathsf{N};\gamma)$ be the set of admissible instances: a public sample loss with its oracle, and client stream laws satisfying Section 3 with the prescribed parameters, on the segment $\mathcal{X}_D = [-D/2, D/2] \times \{0\}^{d-1}$. Let $\mathfrak{A}_T(\mathsf{N})$ be the set of causal rules that receive every raw sample as soon as it arrives, may compute arbitrarily, and may use a private seed independent of the samples; a rule may know the loss map, the public parameters, and the schedule, is fixed before the hidden stream laws are chosen, and learns the population path only through arrived observations. Writing $\mathcal{R}_T(\mathsf{A}, \mathsf{P} \mid \mathsf{N})$ for the loss (3) of rule A on instance P, the minimax risk is

$$\mathcal{R}_T^\star(\gamma) := \inf_{\mathsf{A}\in\mathfrak{A}_T(\mathsf{N})} \sup_{\mathsf{P}\in\mathfrak{P}_T(\mathsf{N};\gamma)} \mathcal{R}_T(\mathsf{A}, \mathsf{P} \mid \mathsf{N}).$$

Since the class grants immediate access to all arrived samples, every lower bound on $\mathcal{R}_T^\star(\gamma)$ also applies to protocols that pay for communication; since the supremum ranges over the loss, the bound is not asserted for every externally fixed loss.

The construction. The hard instance runs independent triangular sign blocks on disjoint client groups, each with its own clock, on one segment (Theorem B.2, Appendix B). Posterior uncertainty about the current signs stays separate across groups, a moment argument converts it into expected linear optimization loss, and a vector integral inequality preserves the loss over the horizon, turning a hard instance at one instant into a time-average bound.

Corollary 5.1 (Comparison across memory scales)**.** *Under the positivity assumptions of Theorem B.2, suppose that at least one $\gamma_m > 0$. For the clients with $\gamma_m > 0$, let $q_m = \sqrt{b_m/a_m}$ and partition them into J nonempty dyadic bins $\mathcal{J}_1, \dots, \mathcal{J}_J$ of the values q_m. For each bin put*

$$Q_j = \sum_{m\in\mathcal{J}_j} \frac{\lambda_m \gamma_m^2}{\sigma_m^2}, \quad T_j = (128 Q_j)^{-1/3}, \quad C_j = \frac{T_j}{2}\sum_{m\in\mathcal{J}_j} w_m\gamma_m, \quad \mathcal{S}_j = \sum_{m\in\mathcal{J}_j} \sqrt{a_m b_m}.$$

If $T_j \geq 2/\min_{m\in\mathcal{J}_j}\lambda_m$ for every j, $T \geq \max_j T_j$, and $G \geq \sum_j C_j$, then

$$\mathcal{R}_T^\star(\gamma) \geq c_{\mathrm{het}} D\Big(\sum_{j=1}^{J} \mathcal{S}_j^{4/3}\Big)^{1/2} \tag{10}$$

$$\geq c_{\mathrm{het}} D\,\frac{\mathcal{H}_\gamma}{J^{1/6}}, \qquad c_{\mathrm{het}} = \frac{1/\sqrt{3}-1/4}{32\cdot 128^{1/3}}. \tag{11}$$

The bins can be chosen so that $J \leq 1 + \lceil \log_2(q_{\max}/q_{\min})\rceil$.

The group-sensitive bound (10) remembers how the mass $\mathcal{S}_\gamma = \sum_j \mathcal{S}_j$ is spread across scales; the simplification (11) is Hölder's inequality, tight for equal bin masses and sharper otherwise. Combined with (9), it gives a constant-factor match for bounded J whenever the optimization, timing, and transient terms of the upper bound are $O(D\mathcal{H}_\gamma)$; Corollary B.3 (Appendix B) makes these conditions explicit, under them $c_{\mathrm{het}} D\mathcal{H}_\gamma/J^{1/6} \leq \mathcal{R}_T^\star(\gamma) \leq [6(2/3)^{1/3}+3]D\mathcal{H}_\gamma$, and Appendix F exhibits a family that satisfies them with arbitrarily many scales and positive communication time. On that family no retuning of the disjoint-group construction closes the $J^{1/6}$ gap (a limitation of the construction, not a minimax impossibility); whether the factor is necessary for unrestricted scales remains open. Appendix E also lower-bounds the finite-horizon cost of stationary clients (Proposition E.1), and Appendix B treats samples whose source time precedes their arrival.

5.2 An exact separation for uniform completed windows

Inequality (5) does not by itself show that a common-memory method loses. We therefore compare actual risks within a specified estimator class: reset empirical minimizers that weight every observation in each client's latest completed buffer uniformly and hold their output until a buffer changes; a *shared* window uses one fixed length h at every client, and the class also contains the full prefix.

Proposition 5.2 (A strict separation within this estimator class)**.** *Let $\ell(x;Y) = x^2/2 - Yx$ on $[-2,2]$, with two equally weighted clients and $\tau = 0$. For an integer $r \geq 2$, client one has mean $\cos(t+\Phi)$, rate r, and Rademacher noise of scale $1/r$; client two has mean zero, rate one, and independent Rademacher noise of scale $\sqrt{2}$. The phase Φ is arbitrary and fixed. Every finite shared window $h \geq 1$ has limiting inferior risk at least $1/16$, and the full prefix attains $1/16$. The separate choice $h_1 = 1/r$, $h_2 = \infty$ has long-run risk*

$$\frac{1}{8}\Big[1 - \frac{\sin(1/r)}{1/r} + \frac{1}{r^2}\Big] \leq \frac{7}{48r^2}. \tag{12}$$

Hence the ratio between the best shared-window risk and the risk of this separate policy is at least $3r^2/7$, on a fixed compact domain.

The proof (Appendix G) separates the within-block variation of the signal from the variance of uniform sample averages: the stationary client's noise makes every finite shared window worse than the prefix, while the changing client benefits from short, accurate buffers. The same calculation yields an exact frequency-response criterion for when a finite window beats the prefix at a single client (the reduction in squared temporal bias must exceed the variance σ_m^2/n_m of the average of the n_m window samples), which predicts failures as well as benefits. The separation concerns uniform completed-window estimators; a learner with common access to the full history can implement client-specific weights itself and is not covered.

6 Choosing among joint memories without drift inputs

Corollary 4.2 identifies good memories when the drift bounds are available; when they are not, a finite collection of joint memory vectors can be combined on the basis of an independent validation stream, under one additional assumption: the sample loss has bounded oscillation at each client, $|\ell(x;\xi) - \ell(y;\xi)| \leq B_m$ for all $x, y \in \mathcal{X}$, and $B := \sum_m w_m B_m > 0$. Bounded-feature logistic loss on a compact domain satisfies this. The noise and rate bounds remain known, but the learner does not know γ.

Procedure. Alternate arrivals at each client are assigned to a training and a validation stream. For each of E fixed memory vectors $h^1, \dots, h^E$, a separate reset SGD *state* runs on the shared training array with disjoint batches within each solve; these experts never see validation samples or mixture weights. The server reports a convex mixture of the expert outputs. Fresh validation samples from each reporting interval score all experts, and two interleaved exponential-weight forecasters (for even and for odd intervals) ensure that a full interval of feedback has arrived before the forecaster that used it reports again. The aggregation principle is standard (Freund & Schapire, 1997); the independent split and a source-time correction make it valid here. Appendix H specifies the communication pipeline and proves the complete form of the following statement (Theorem H.2).

Theorem 6.1 (Oracle comparison for joint memory choices)**.** *Assume the model of Section 3 with the bounded oscillation above, except that the drift bounds need not be known. Let* $s_0 = 6K/\lambda_{\min}$, $c = K\tau$, $\Delta \geq \max\{c,\, 6/\lambda_{\min}\}$, *and* $T = s_0 + c + R\Delta$ *for an integer* $R \geq 1$. *Choose* $E \geq 2$ *deterministic memory vectors, independently of the data, with* $6K/\lambda_m \leq h_m^e \leq T$, *and let* $\mathcal{R}_T^e$ *be the risk of the training-stream SGD expert* e, *including initialization. The risk* $\mathcal{R}_T^{\mathrm{mix}}$ *of the causal mixture satisfies*

$$\mathcal{R}_T^{\mathrm{mix}} \leq \min_{e \in [E]} \mathcal{R}_T^e + 2B\sqrt{\frac{\Delta \log E}{T}} + D\bar{\gamma}\Delta. \tag{13}$$

Neither the updates nor the learning rate uses a drift bound.

Any interval whose length $v - u$ is at least $6/\lambda_m$ contains at least $\lambda_m(v-u)/6$ training samples of client m, so every expert enjoys the fixed-memory SGD guarantee with b_m replaced by $3b_m$. The comparator in (13) therefore retains the joint bias–variance expression $D \min_e\{2\sum_m a_m h_m^e + 2\sqrt{2}\sqrt{\sum_m 3b_m/h_m^e}\}$, plus the SGD, timing, and initialization terms of Proposition B.1: the pooled variance is not replaced by a sum of clientwise errors. A Cartesian geometric grid approximates the continuous minimum of this expression within a factor of two and contains the full prefix $h_m = T$ for every client.

What the theorem does and does not give. It compares with the best *fixed* joint memory vector over the stated horizon, not with a sequence of locally best windows; the validation split consumes observations, the E expert solves can be expensive, and keeping the union of their buffers may require the full training history. It establishes statistical adaptation under a stronger loss assumption, not an efficient storage policy.

7 Experiments: when does forgetting help?

Two networks with measured drift test Corollary 4.2 from opposite sides. At climate stations the model of Section 3 holds by construction and the windows are used exactly as prescribed; at clinics the drift is realistic, the assumptions hold only approximately, and the solver is Algorithm 1. All comparisons are paired. We write $g = 1 - \mathcal{H}_\gamma/(\bar{\gamma}^{1/3}\Xi^{2/3})$ for the relative gap in the comparison $\mathcal{H}_\gamma \leq \bar{\gamma}^{1/3}\Xi^{2/3}$ of Section 3, the gain that the upper bounds predict.

Temperature stations. Three stations of the U.S. Climate Reference Network (Diamond et al., 2013), chosen by insolation before any simulation, report every five minutes: a desert, a maritime site, and a site in polar night. A station's population path is its diurnal cycle fitted to ten Decembers, and a draw adds a resampled day-to-day residual, so the drift and noise bounds hold by construction. The loss is nearly linear, and one round per refresh suffices ($K = 1$, where Algorithm 1 and SGD coincide). The desert cycle is fast, the maritime one slow, the polar path constant: $g = 0.32$, and the prescribed windows are 0.56 and 1.28 hours and the full history, which no single length can offer. Used without a multiplier, they reduce the gap by 51% against the bound-optimal common window and by 52% against one tuned on separate seeds (Figure 1, left). In July the cycles align and g falls to 0.066. Against the bound-optimal common window the reduction is 6%, and the paired interval $[-0.00158, 0.00005]$ contains zero. Against the window tuned on separate seeds the reduction is 11%, with interval $[-0.0025, -0.0008]$ (right). Across 200 random station triplets the realized gain grows with g (Spearman 0.73), and for g near zero a tuned common window can win by a few percent: the comparison of upper bounds predicts both where client-specific memory pays and where it does not (Appendix J.1).

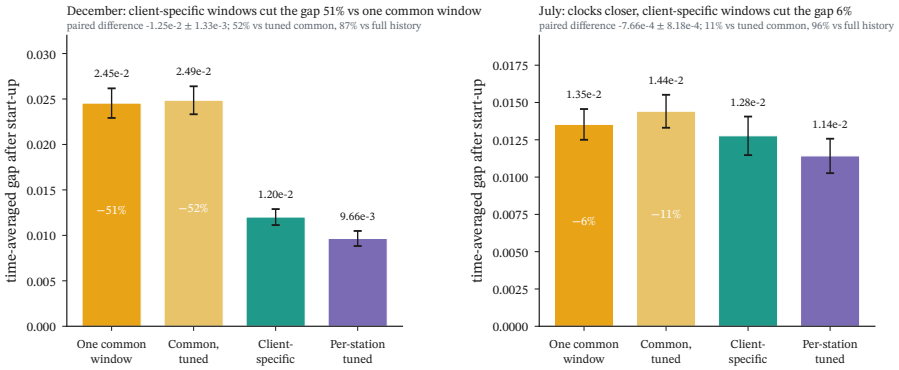

Figure 1: Temperature stations: time-averaged gap over twelve evaluation seeds with 95% intervals (Appendix J.1). "One common window" has the bound-optimal length; client-specific windows follow Corollary 4.2 without tuning; per-station tuning is an oracle. December has $g = 0.32$, July $g = 0.066$; subtitles give paired differences.

Recorded order. Replayed in recorded order, held-out Decembers leave the model: persistent weather acts as drift, and the climatological inputs, which keep the whole history at the polar site, lose in every year to a common window tuned on the other nine. A drift input on the time scale of weather, computed on the training folds, matches that window without any search (Appendix J.2): what fails is the drift input, not the window rule.

Clinics. Twelve clinics in six specialties share a logistic model with one intercept per clinic that predicts an emergency or inpatient encounter within 90 days from synthetic records (Walonoski et al., 2018). A cohort the model does not observe, patients with recent acute care, follows public influenza and COVID-19 surveillance curves, so the optimum moves at speeds that differ by two orders of magnitude across clinics, while the specialized clinics, which see the fewest patients, are stationary. The drift and noise inputs are evaluated along the exact optimal trajectory rather than certified, and the single scale of every memory is tuned on separate streams, so this experiment compares the shape of the memory, not its constant. Under exact minimization of the buffered risk, client-specific windows reduce the gap by 33% against the best common window, which is the full history, and by 26% against a common window with a per-clinic data floor, reaching 89% of the improvement of tuning all twelve windows. With Algorithm 1 ($K = 8$) the optimization error dilutes every difference: the reduction is 16% against the best common window, and the data-floored common window comes within 1% (Appendix I). Where drift is seasonal and returns to its mean, the worst-case prescription is conservative: per-clinic tuning keeps the full history for pediatrics and primary care, where Corollary 4.2 forgets.

8 Conclusion and Limitations

Client-specific memory separates the statistical uncertainty of stable clients from the forgetting that drifting clients require: $\mathcal{H}_\gamma$ is attainable with explicit finite-horizon costs, a lower construction with unequal drift clocks retains each group's contribution, an exact periodic calculation separates common from client-specific windows within a specified estimator class, and independent validation competes with a finite family of fixed memories without any drift bound as input. Several limitations remain. Whether the $J^{1/6}$ factor between the upper and lower bounds is necessary for unrestricted memory scales is open; adaptive aggregation needs a bounded-loss assumption, spends samples on validation, and may require many experts and full-history storage; the experiments are the station laws, the clinic network, and the controlled synthetic studies. Full history can be preferable under smooth, seasonal, or shared drift. Temporal dependence, efficient selection over large joint grids, and communication budgets that exclude raw-data pooling call for further analysis; the source-age extension assumes an unbiased oracle for the source time rather than inferring censored labels from silence.

AI USE STATEMENT

Generative AI tools were used in preparing this manuscript: editing text and partial proof verification via Lean. The authors review the results and take full responsibility for them.

ETHICS STATEMENT

The experiments use public air-temperature records of the U.S. Climate Reference Network, synthetic electronic health records generated with Synthea, and aggregate public surveillance series (CDC ILINet and HHS COVID-19 hospital admissions, retrieved through the Delphi Epidata service (Farrow et al., 2015)). No real patient-level data were used, no new human-subject data were collected, and no intervention was performed. The clinic experiment tests an optimization mechanism; it is not a clinical prediction rule.

REPRODUCIBILITY STATEMENT

The model and the algorithm are specified in Sections 3 and 4; a table of notation is in Appendix A. Complete proofs appear in Appendices C–H. The supplement contains the experiment and verification scripts, the scripts that retrieve the station records and the surveillance series, disjoint tuning and evaluation seeds, seed-level outputs, and the protocols of Appendices I and J. Numerical bound diagnostics evaluate every remainder and check the joint regime without simulating its long arrival paths. Appendix K summarizes the benchmark comparisons and the scope of the supplementary verification.

REFERENCES

Pranjal Awasthi, Corinna Cortes, and Christopher Mohri. Theory and algorithm for batch distribution drift problems. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9826–9851. PMLR, 2023. URL https://proceedings.mlr.press/v206/awasthi23b.html.

Centers for Disease Control and Prevention. Outpatient illness surveillance (ILINet). https://www.cdc.gov/fluview/overview/index.html, 2024.

Song Xi Chen. Beijing multi-site air quality. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5RK5G.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. LASG: Lazily aggregated stochastic gradients for communication-efficient distributed learning, 2020. URL https://arxiv.org/abs/2002.11360.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023. URL https://jmlr.org/papers/v24/21-1410.html.

Howard J. Diamond, Thomas R. Karl, Michael A. Palecki, C. Bruce Baker, Jesse E. Bell, Ronald D. Leeper, David R. Easterling, Jay H. Lawrimore, Tilden P. Meyers, Michael R. Helfert, Grant Goodge, and Peter W. Thorne. U.S. Climate Reference Network after one decade of operations: Status and assessment. *Bulletin of the American Meteorological Society*, 94(4):485–498, 2013. doi: 10.1175/BAMS-D-12-00170.1.

David C. Farrow, Logan C. Brooks, Aaron Rumack, Ryan J. Tibshirani, and Roni Rosenfeld. Delphi Epidata API. https://github.com/cmu-delphi/delphi-epidata, 2015.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

Bhargav Ganguly and Vaneet Aggarwal. Online federated learning via non-stationary detection and adaptation amidst concept drift. *IEEE/ACM Transactions on Networking*, 32(1):643–653, 2024. doi: 10.1109/TNET.2023.3294366. URL https://doi.org/10.1109/TNET.2023.3294366.

Khiem Huynh, Malcolm Egan, Giovanni Neglia, and Jean-Marie Gorce. Streaming federated learning with markovian data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ba0dd55207ef557f6774bc85332f94a8-Abstract-Conference.html.

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip B. Gibbons. Federated learning under distributed concept drift. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 5834–5853, 2023. URL https://proceedings.mlr.press/v206/jothimurugesan23a.html.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461, 2013. URL https://proceedings.mlr.press/v28/joulani13.html.

Othmane Marfoq, Giovanni Neglia, Laetitia Kameni, and Richard Vidal. Federated learning for data streams. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8889–8924. PMLR, 2023. URL https://proceedings.mlr.press/v206/marfoq23a.html.

Alessio Mazzetto and Eli Upfal. An adaptive algorithm for learning with unknown distribution drift. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://papers.nips.cc/paper_files/paper/2023/hash/1fe6f635fe265292aba3987b5123ae3d-Abstract-Conference.html.

Alessio Mazzetto and Eli Upfal. Nonparametric density estimation under distribution drift. In *International Conference on Machine Learning*, volume 202, pp. 24251–24270, 2023b. URL https://proceedings.mlr.press/v202/mazzetto23a.html.

Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, et al. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51): e2111452118, 2021. doi: 10.1073/pnas.2111452118. URL https://doi.org/10.1073/pnas.2111452118.

Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1997. URL https://www.dspguide.com/CH15.PDF.

Alexander Tyurin. Tight time complexities in parallel stochastic optimization with arbitrary computation dynamics. In *International Conference on Learning Representations*, 2025.

U.S. Department of Health and Human Services. COVID-19 reported patient impact and hospital capacity by state timeseries. HealthData.gov, 2021. URL https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018. doi: 10.1093/jamia/ocx079.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A*, 473(2205):20170457, 2017. doi: 10.1098/rspa.2017.0457. URL https://doi.org/10.1098/rspa.2017.0457.

Zhuoyi Zhao and Ben Liang. Adaptive data admission and retention for streaming federated learning. *arXiv preprint arXiv:2607.23987*, 2026. URL https://arxiv.org/abs/2607.23987.

A NOTATION

Tables 1 and 2 list the symbols used in the main text and in the proofs, with the place where each is introduced. Every symbol has one meaning throughout the paper; experiment-specific constants (introduced in experiment-specific constants (introduced in Appendices I and J) use letters that do not appear in the theory. We follow the usual conventions of the optimization literature: L is the Lipschitz constant of the gradient, η denotes stepsizes and learning rates, δ the inexactness (bias) of a gradient oracle, ε centered noise, $x^\star$ a minimizer, and iterates carry the iteration index as a superscript; no result uses strong convexity, so the strong-convexity parameter μ and the condition number κ do not appear. All norms are Euclidean, $[n] = \{1, \dots, n\}$, and i denotes the imaginary unit.

Table 1: Symbols of the model, the algorithm, and the upper bounds.

Symbol	Meaning	Where
M, m, w_m	number of clients, client index, public weights ($\sum_m w_m = 1$)	Sec. 3
$\mathcal{X}$, d, D, $\Pi_{\mathcal{X}}$	compact convex domain in $\mathbb{R}^d$, its diameter, Euclidean projection	Sec. 3
$\ell(x;\xi)$, $\mathcal{D}_m^t$	public sample loss; law of a sample of client m generated at time t	Sec. 3
f_m^t, F^t	population objective of client m and weighted objective at time t	Sec. 3
L, G	smoothness constant of F^t; public bound on $\|\nabla F^t\|$	Sec. 3
λ_m, $\lambda_{\min}$, $N_m(t)$	arrival rate, smallest rate, arrival count of client m	Sec. 3
σ_m^2, γ_m, γ, $\bar{\gamma}$	gradient-noise variance bound, clientwise drift bound, drift vector, $\sum_m w_m \gamma_m$	Sec. 3
τ, T, AvgErr_T	synchronization time, horizon, time-averaged current-objective gap	Sec. 3
a_m, b_m, $\mathcal{S}_\gamma$, $\mathcal{H}_\gamma$, Ξ	$w_m\gamma_m$, $w_m^2\sigma_m^2/\lambda_m$, $\sum_m \sqrt{a_m b_m}$, $\mathcal{S}_\gamma^{2/3}$, $\Xi^2 = \sum_m b_m$	(4)
K, h_m, $h_m^\star$, Δ	rounds per solve, memory length of client m and its ideal value, reporting stride	Sec. 4.1
s_0, s_j, c, t_j, $\mathcal{I}_j$	first solve start, start of solve j, solve duration $K\tau$, report time $s_j + c$ ($x_\circ$ is reported before t_0), report interval $[t_j, t_{j+1})$	Sec. 4.1
$h_{m,j}$, $n_{m,j}$	length $\min\{s_j, h_m\}$ and sample count of the frozen buffer of client m at solve j	Sec. 4.1
$\mathcal{B}_{m,j,k}$, $n_{m,j,k}$, $t_{m,i}$	kth batch of that buffer, its size, timestamp of sample i	Alg. 1
$x_\circ$, x^k, v^k, y^k, θ_k, g^k	reset point, accelerated iterates, momentum weight $2/(k+1)$, aggregated batch gradient	Alg. 1
$V_{j,k}$, S_j, η_j, x_j	batch variance bound, $\sum_k k^2 V_{j,k}$, stepsize, output of solve j	Alg. 1
β_0, $\mathcal{T}_T$	noise constant $4\sqrt{2/3}$, finite-horizon transient	Thm. 4.1
z_j^k, $\bar{z}_j$, U_j, η_j^{sgd}, β_{sgd}	SGD iterates and output, $\sum_k V_{j,k}$, SGD stepsize, noise constant $2\sqrt{2}$	Prop. B.1
$\mathcal{F}_k$, δ_k, δ, ε_k, V_k	filtration, gradient bias (oracle inexactness) and its bound, centered noise and its variance bound (Lemmas)	App. C
$\bar{V}_j$, $\delta_j(t)$	buffer-level variance proxy and bias bound of solve j at time t	App. C
Δ_m^{src}	source-age bound of client m	App. B
$\mathcal{C}_\gamma$	count-floor and communication charge $4\sum_m a_m/\lambda_m + 2\bar{\gamma}\tau$	Cor. B.3

Public records. Appendix J uses two public temperature records. The U.S. Climate Reference Network five-minute air-temperature product (Diamond et al., 2013) is the `subhourly01` release, `https://www.ncei.noaa.gov/pub/data/uscrn/products/subhourly01`, Decembers and Julys 2010–2019 at the named stations and, for the network study, every station with sufficient December coverage. The Beijing control uses hourly temperature (`TEMP`) from the twelve-site archive (Zhang et al., 2017; Chen, 2017); meteorological fields in that archive are matched to the nearest China Meteorological Administration station. Appendix I uses synthetic visits generated

Table 2: Symbols of the lower bounds, the selection procedure, and the periodic streams.

Symbol	Meaning	Where
N, $\mathfrak{P}_T(\mathsf{N};\gamma)$, $\mathfrak{A}_T(\mathsf{N})$	schedule, admissible instances, causal rules	Sec. 5
$\mathcal{X}_D$, $\mathcal{R}_T(\mathsf{A},\mathsf{P}\mid\mathsf{N})$, $\mathcal{R}_T^\star(\gamma)$	segment $[-D/2, D/2]\times\{0\}^{d-1}$, risk of rule A on instance P, minimax risk	Sec. 5
$\mathcal{J}_j$, J, q_m, Q_j	client groups (dyadic bins), their number, memory scale $\sqrt{b_m/a_m}$, bin information rate	Cor. 5.1
T_j, A_m, C_j, $\mathcal{S}_j$, c_{het}	block length and amplitudes of group j, $\sum_{m\in\mathcal{J}_j} w_m A_m$, bin mass $\sum_{m\in\mathcal{J}_j}\sqrt{a_m b_m}$, constant	Thm. B.2
Λ, $\Lambda_j(t)$, $\varsigma_{j,k}$, $\varsigma_j(t)$	triangular profile, its block version, block signs	App. E
$\zeta_j(t)$, $\Theta(t)$, $\mathcal{F}_t$, $\bar{\varsigma}_j(t)$	$C_j\Lambda_j(t)$, population coefficient, pooled transcript, posterior mean of the sign	App. E
B_m, B	loss-oscillation bounds, $\sum_m w_m B_m$	Sec. 6
E, e, h^e, $\mathcal{G}$, $\mathcal{G}_m$	number of experts, expert index, its memory vector, joint and per-client grids	Sec. 6
R, x_j^e, $\pi_{j,e}$, η_{ew}, $\widehat{\ell}_{j,e}$	number of report intervals, expert output, mixture weights, learning rate, validation score	App. H
$\mathcal{V}_{m,j}$, $n_{m,j}^{\mathrm{val}}$, $\overline{F}_j$, F_j^{val}	validation samples of interval j, their count, centered interval-average and validation-time objectives	App. H
$\mathcal{R}_T^e$, $\mathcal{R}_T^{\mathrm{mix}}$, $\widetilde{b}_m$, $\widetilde{\Xi}$, $\widetilde{\mathcal{T}}_T$	expert and mixture risks, $3b_m$, and the corresponding pooled uncertainty and transient	Thm. H.2
$Y_{m,i}$, $\bar{\vartheta}_m$, A_m, ω_m, Φ_m, $\varepsilon_{m,i}$	observation, mean level, amplitude, frequency, phase, Rademacher sign	(49)
r, t_{s}, $\vartheta(t)$, $\widehat{\vartheta}_m(t)$	moving-client rate, start of the data-based reports ($\max_m 1/\lambda_m$), population and empirical coefficients	App. G
$\mathcal{M}_m(t)$, $n_m(t)$, n_m, $\mathcal{K}_m(t)$	retained indices, their count, window count $\lambda_m h_m$, complex coefficient of the average	App. G
φ_m, $\mathcal{K}_n(\varphi)$, $\Upsilon_n(\varphi)$, sinc	phase increment ω_m/λ_m, window filter, bias factor, $\sin(z)/z$	Thm. G.1
$\mathcal{Q}_T$, $\mathcal{Q}_\infty$, $\mathcal{Q}_\infty^{\mathrm{sep}}$	phase-averaged risk, its limit, limit of the separate policy	App. G

with Synthea (Walonoski et al., 2018). Mix weights follow public CDC ILINet and HHS COVID-19 hospital series (Centers for Disease Control and Prevention, 2024; U.S. Department of Health and Human Services, 2021) retrieved from Delphi Epidata (Reinhart et al., 2021). No real patient record is used.

B SUPPORTING STATEMENTS

This appendix collects the statements that the main text refers to but does not display: the raw-buffer implementation, the projected-SGD counterpart of Theorem 4.1, the general unequal-clock lower bound behind Corollary 5.1, the explicit regime in which the upper and lower bounds are comparable, and the source-time extension of the upper bounds. Proofs are given in Appendices C–F.

B.1 THE RAW-BUFFER IMPLEMENTATION

Messages are unrestricted, so clients with finitely encodable samples may upload their frozen buffers, timestamps, and counts in one synchronization at the start of each solve. The server then runs the same ordered batch updates of Algorithm 1 locally and either waits to reproduce the report times of the gradient-message implementation or reports after τ. The latter schedule uses $c = \tau$ and $\Delta \geq \tau$ in the proofs, with unchanged sample-count floors and optimization error, so Theorem 4.1 and Corollary 4.2 hold for it with $K\tau$ replaced by τ. The elapsed time $K\tau$ therefore describes the gradient-message implementation and is not intrinsic to the model.

B.2 PROJECTED SGD WITH THE SAME MEMORIES

Acceleration improves the optimization remainder of Theorem 4.1 from order K^{-1} to order K^{-2}; it does not change the statistical-drift quantity. The following comparison keeps the reset rule, the disjoint batches, and the reporting times of Algorithm 1 and replaces only the solver.

Proposition B.1 (Projected SGD with the same memories)**.** *Under the assumptions of Theorem 4.1, let $V_{j,k}$ be the batch variance bounds of Algorithm 1 and set*

$$U_j = \sum_{k=1}^{K} V_{j,k}, \qquad \eta_j^{\mathrm{sgd}} = \min\{1/(2L),\, D/\sqrt{2U_j}\},$$

with $\eta_j^{\mathrm{sgd}} = 1/(2L)$ when $U_j = 0$. In solve j, reset $z_j^0 = x_\circ$ and let g_j^k be the weighted gradient of the same batches $\mathcal{B}_{m,j,k}$ evaluated at z_j^{k-1}. Define

$$z_j^k = \Pi_{\mathcal{X}}\big(z_j^{k-1} - \eta_j^{\mathrm{sgd}} g_j^k\big), \qquad \bar{z}_j = \frac{1}{K}\sum_{k=1}^{K} z_j^k. \tag{14}$$

Report $\bar{z}_j$ on $\mathcal{I}_j = [s_j + K\tau, s_{j+1} + K\tau)$ and $x_\circ$ before the first report. Then the pointwise and average guarantees (6)–(8) hold for this trajectory with x_j replaced by $\bar{z}_j$ and with the substitutions

$$\frac{4LD^2}{K(K+1)} \;\longmapsto\; \frac{LD^2}{K}, \qquad \beta_0 \;\longmapsto\; \beta_{\mathrm{sgd}} := 2\sqrt{2}. \tag{15}$$

Appendix D proves the proposition and optimizes its windows: the statistical-drift value becomes $6 \cdot 2^{-1/3} D\mathcal{H}_\gamma$, with the same count-floor and timing charges as in Corollary 4.2. The average in (14) runs over post-update points, so every batch affects the report. The first query $z_j^0 = x_\circ$ is fixed, so the batch counts and the first gradient share one message, exactly as in Algorithm 1.

B.3 THE GENERAL UNEQUAL-CLOCK CONSTRUCTION

Theorem B.2 (Different drift clocks in one dimension)**.** *Fix positive $(w_m, \lambda_m, \sigma_m)_m$, $D > 0$, the drift vector γ, and a λ-regular schedule. Choose disjoint nonempty client groups $\mathcal{J}_1, \ldots, \mathcal{J}_J$, block lengths $0 < T_j \leq T$, and amplitudes $A_m \geq 0$ for the clients in these groups such that*

$$T_j \geq \frac{2}{\min_{m\in\mathcal{J}_j} \lambda_m}, \qquad A_m \leq \frac{\gamma_m T_j}{2}\ \ (m \in \mathcal{J}_j), \qquad C_j := \sum_{m\in\mathcal{J}_j} w_m A_m, \qquad \sum_{j=1}^{J} C_j \leq G,$$

and

$$4T_j \sum_{m\in\mathcal{J}_j} \frac{\lambda_m A_m^2}{\sigma_m^2} \leq \frac{1}{128} \qquad \textit{for every } j.$$

Then

$$\mathcal{R}_T^\star(\gamma) \geq \frac{1/\sqrt{3}-1/4}{8}\, D\Big(\sum_{j=1}^{J} C_j^2\Big)^{1/2}. \tag{16}$$

For every fixed rule and horizon, one deterministic admissible drift path attains this bound.

The proof (Appendix E) runs independent triangular sign blocks with a different clock for each group on the same one-dimensional segment. Posterior uncertainty about the current signs stays separate across groups; a second- and fourth-moment argument turns it into expected linear optimization loss; and a vector integral inequality preserves the loss over the horizon. A hard bump at a single time would not justify the integrated statement.

B.4 An explicit joint comparison regime

The next statement lists the conditions under which every term of the upper bound is of order $D\mathcal{H}_\gamma$, so that the upper and lower bounds are comparable. Write

$$\mathcal{C}_\gamma = 4\sum_m \frac{a_m}{\lambda_m} + 2\bar{\gamma}\tau$$

for the per-round count-floor and communication charge.

Corollary B.3 (Joint upper and lower regime)**.** *Consider the model of Section 3 on $\mathcal{X}_D$ with positive $(\lambda_m, \sigma_m)_m$, $\mathcal{H}_\gamma > 0$, and $\tau > 0$. Fix a λ-regular schedule and the dyadic partition that defines T_j and C_j in Corollary 5.1, and suppose*

$$T_j \geq \frac{2}{\min_{m\in\mathcal{J}_j}\lambda_m} \quad (j\in[J]), \qquad G \geq \sum_{j=1}^{J} C_j.$$

Choose

$$K = \Big\lceil \max\{1,\, 2\sqrt{LD/\mathcal{H}_\gamma}\}\Big\rceil, \qquad \Delta = K\tau, \qquad s_0 = 2K/\lambda_{\min},$$

suppose $K\mathcal{C}_\gamma \leq \mathcal{H}_\gamma$, and, with $\beta_0 = 4\sqrt{2/3}$, assume

$$\begin{aligned} T &\geq \max\Big\{\max_j T_j,\; s_0 + K\tau,\; \frac{16\beta_0^2\Xi^2}{\mathcal{H}_\gamma^2}\Big\},\\ T &\geq \frac{2\big[\beta_0\Xi K\tau/\sqrt{s_0} + G(s_0+K\tau)\big]}{\mathcal{H}_\gamma}. \end{aligned} \tag{17}$$

With the windows of Corollary 4.2, Algorithm 1 satisfies $\mathrm{AvgErr}_T \leq [6(2/3)^{1/3}+3]D\mathcal{H}_\gamma$ on every admissible instance, and

$$c_{\mathrm{het}}\frac{D\mathcal{H}_\gamma}{J^{1/6}} \leq \mathcal{R}_T^\star(\gamma) \leq [6(2/3)^{1/3}+3]\, D\mathcal{H}_\gamma. \tag{18}$$

The condition $K\mathcal{C}_\gamma \leq \mathcal{H}_\gamma$ controls the sample-count and communication charges, and the horizon conditions control the startup and statistical transients; neither follows from the lower-bound hypotheses alone. Appendix F gives a family with positive communication time that satisfies all requirements for arbitrarily many memory scales, and shows that on this family no retuning of the disjoint-group construction removes the $J^{1/6}$ factor. The latter is a limitation of the construction, not a proof that the factor is minimax necessary.

B.5 Source time and usable time

Suppose that a sample usable at time t was generated at an exogenous source time in $[\max\{0, t-\Delta_m^{\mathrm{src}}\}, t]$, and that the independent unbiased oracle assumptions (1) hold relative to the population at the source time. Then the bounds of Theorem 4.1 and Corollary 4.2 gain the additive term

$$D\sum_m w_m\gamma_m\Delta_m^{\mathrm{src}}. \tag{19}$$

The proof (Appendix C) adds the mean-gradient bias $\gamma_m\Delta_m^{\mathrm{src}}$ at each client. Positive-only or censored labels would require a separate unbiased correction and variance analysis; the extension assumes such an oracle rather than deriving it.

C PROOFS FOR CLIENT-SPECIFIC REFRESH

The proof of Theorem 4.1 has two parts. We first prove a projected accelerated potential for a stochastic oracle with bounded gradient bias (Lemma C.1); as in the inexact-oracle literature, δ denotes the inexactness of the oracle, here a predictable bias of the batch gradient. We then apply the lemma to one reset solve, with a filtration that exposes only the disjoint batches of that solve; this is what allows data buffers to be reused across solves. Corollary 4.2 follows from an elementary window-functional lemma, and the section closes with the source-time extension of Appendix B.

C.1 PROJECTED ACCELERATION WITH BOUNDED GRADIENT BIAS

Lemma C.1 (A biased stochastic oracle on a compact set)**.** *Let F be convex and L-smooth, $L > 0$, on a compact convex set $\mathcal{X}$ of diameter $D > 0$. Start from $x^0 = v^0 = x_\circ \in \mathcal{X}$ and, for $k = 1, \dots, K$, let*

$$\theta_k = \frac{2}{k+1}, \qquad y^k = (1-\theta_k)x^{k-1} + \theta_k v^{k-1},$$
$$v^k = \Pi_{\mathcal{X}}(v^{k-1} - k\eta\, g^k), \qquad x^k = (1-\theta_k)x^{k-1} + \theta_k v^k.$$

Let $(\mathcal{F}_k)_{k\geq 0}$ be a filtration such that y^k, v^{k-1}, x^{k-1} are $\mathcal{F}_{k-1}$-measurable, and suppose

$$g^k = \nabla F(y^k) + \delta_k + \varepsilon_k, \qquad \|\delta_k\| \leq \delta,$$

where δ_k is $\mathcal{F}_{k-1}$-measurable and

$$\mathbb{E}[\varepsilon_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\varepsilon_k\|^2 \mid \mathcal{F}_{k-1}] \leq V_k$$

for deterministic $V_k \geq 0$. Then, for every $0 < \eta \leq 1/(4L)$,

$$\mathbb{E}[F(x^K) - \min_{\mathcal{X}} F] \leq \frac{2}{K(K+1)}\Big(\frac{D^2}{2\eta} + \eta \sum_{k=1}^{K} k^2 V_k\Big) + \delta D. \tag{20}$$

If $S = \sum_{k=1}^{K} k^2 V_k > 0$, the choice $\eta = \min\{1/(4L), D/\sqrt{2S}\}$ gives

$$\mathbb{E}[F(x^K) - \min_{\mathcal{X}} F] \leq \frac{4LD^2 + 2\sqrt{2}D\sqrt{S}}{K(K+1)} + \delta D, \tag{21}$$

and for $S = 0$ the same bound holds with $\eta = 1/(4L)$.

Proof. Fix $x^\star \in \arg\min_{\mathcal{X}} F$ and write $r_k = \|v^k - x^\star\|^2$. Since $x^k - y^k = \theta_k(v^k - v^{k-1})$ and $y^k = (1-\theta_k)x^{k-1} + \theta_k v^{k-1}$, smoothness together with the two supporting-hyperplane inequalities at y^k gives

$$F(x^k) - F(x^\star) \leq (1-\theta_k)[F(x^{k-1}) - F(x^\star)] + \theta_k\langle \nabla F(y^k), v^k - x^\star\rangle + \frac{L\theta_k^2}{2}\|v^k - v^{k-1}\|^2. \tag{22}$$

The variational inequality of the projection that defines v^k gives

$$\langle g^k, v^k - x^\star\rangle \leq \frac{r_{k-1} - r_k - \|v^k - v^{k-1}\|^2}{2k\eta}. \tag{23}$$

Substitute $\nabla F(y^k) = g^k - \delta_k - \varepsilon_k$ into (22). Since $v^k, x^\star \in \mathcal{X}$, the bias contributes $-\theta_k\langle \delta_k, v^k - x^\star\rangle \leq \theta_k \delta D$. For the noise, write $v^k - x^\star = (v^{k-1} - x^\star) + (v^k - v^{k-1})$. The stepsize restriction gives $L\theta_k \cdot k\eta \leq 2L\eta \leq 1/2$, so after (23) the coefficient of $-\|v^k - v^{k-1}\|^2$ is at least $\theta_k/(4k\eta)$, and Young's inequality yields

$$-\theta_k\langle \varepsilon_k, v^k - v^{k-1}\rangle \leq \theta_k k\eta\,\|\varepsilon_k\|^2 + \frac{\theta_k}{4k\eta}\|v^k - v^{k-1}\|^2.$$

The squared-step terms cancel, and $\mathbb{E}\langle \varepsilon_k, v^{k-1} - x^\star \rangle = 0$ by conditional centering. Hence, with $\mathcal{E}_k = \mathbb{E}[F(x^k) - F(x^\star)]$,

$$\mathcal{E}_k \le (1-\theta_k)\mathcal{E}_{k-1} + \frac{\theta_k}{2k\eta}\mathbb{E}[r_{k-1} - r_k] + \theta_k k\eta V_k + \theta_k \delta D. \tag{24}$$

Let $\Gamma_k = 2/[k(k+1)]$. For $k \ge 2$ we have $(1-\theta_k)/\Gamma_k = 1/\Gamma_{k-1}$, and for $k = 1$ the coefficient of $\mathcal{E}_0$ vanishes. Moreover,

$$\frac{\theta_k}{2k\eta\,\Gamma_k} = \frac{1}{2\eta}, \qquad \frac{\theta_k k\eta}{\Gamma_k} = k^2\eta, \qquad \frac{\theta_k}{\Gamma_k} = k.$$

Dividing (24) by Γ_k, summing over k, and using $r_0 \le D^2$, $r_K \ge 0$, and $\Gamma_K \sum_{k=1}^K k = 1$ proves (20).

If the tuned stepsize equals $D/\sqrt{2S}$, the two terms in parentheses in (20) sum to $\sqrt{2}D\sqrt{S}$. If the cap $1/(4L)$ is active, then $S \le 8L^2D^2$ and

$$\frac{D^2}{2\eta} + \eta S = 2LD^2 + \frac{S}{4L} \le 2LD^2 + \frac{D\sqrt{S}}{\sqrt{2}}.$$

Both cases give (21); the case $S = 0$ follows directly from (20). □

The bias bound uses the compactness of $\mathcal{X}$ at the projected auxiliary point v^k, not only at the reported point x^k. An unprojected acceleration recurrence therefore cannot be substituted without a separate bound on its auxiliary iterates.

C.2 Proof of Theorem 4.1

Proof. Condition on the complete exogenous arrival schedule. All clocks, buffer choices, batch partitions, and variance-based stepsizes are then fixed without revealing any sample value. Recall $s_j = s_0 + j\Delta$, $c = K\tau$, and $h_{m,j} = \min\{s_j, h_m\}$ with $\min\{s_j, \infty\} = s_j$.

Step 1: buffer ages and sample counts. If $s_j < h_m$, the frozen buffer of client m is the prefix $(0, s_j]$, of length $h_{m,j} = s_j$. If $h_m \le s_j$ and $h_m < \infty$, let $p = \lfloor s_j/h_m \rfloor$; the frozen buffer is $((p-1)h_m, ph_m]$, of length $h_{m,j} = h_m$, and since $ph_m \le s_j < (p+1)h_m$ every sample in it has age less than $2h_m$ at the solve start. Infinite windows always use the prefix. Hence every selected timestamp $t_{m,i}$ satisfies

$$s_j - 2h_{m,j} < t_{m,i} \le s_j. \tag{25}$$

By $s_0 = 2K/\lambda_{\min}$ and $h_m \ge 2K/\lambda_m$ we have $h_{m,j} \ge 2K/\lambda_m$, so the lower regularity inequality applies to the frozen interval and gives

$$n_{m,j} \ge \frac{\lambda_m h_{m,j}}{2} \ge K. \tag{26}$$

Splitting the $n_{m,j}$ observations into K batches whose sizes differ by at most one yields batch sizes

$$n_{m,j,k} \ge \left\lfloor \frac{n_{m,j}}{K} \right\rfloor \ge \frac{n_{m,j}}{2K}, \tag{27}$$

where the last inequality uses $n_{m,j}/K \ge 1$.

Step 2: the filtration of one solve. Fix a solve j. Its iterates reset to the public point $x_\circ$, and its updates use only the batches of this solve. Let $\mathcal{F}_{j,k}$ be generated by the fixed schedule and the values of batches $1, \dots, k$ of this solve at all clients. Then y^k and the earlier iterates are $\mathcal{F}_{j,k-1}$-measurable, and conditional independence of the time-stamped samples makes batch k independent of the earlier batches of the solve.

The same samples may have appeared in earlier solves. We do not condition on those uses or on earlier reported outputs: the current solve is a measurable function of its own frozen buffers, its partition, and the fixed initialization, so this marginal filtration bounds its expected risk even when buffers overlap across solves. Earlier outputs change neither the marginal laws of the buffer samples nor the fact that each batch is used once within this solve. Warm starts or sample-dependent windows would require a different argument.

Step 3: the conditional oracle at a fixed report time. Fix $t \in \mathcal{I}_j = [s_j + c, s_{j+1} + c)$. The conditional mean of the aggregate gradient at step k is

$$\mathbb{E}[g^k \mid \mathcal{F}_{j,k-1}] = \sum_m \frac{w_m}{n_{m,j,k}} \sum_{i \in \mathcal{B}_{m,j,k}} \nabla f_m^{t_{m,i}}(y^k).$$

By the clientwise drift bounds and (25),

$$\begin{aligned}\left\|\mathbb{E}[g^k \mid \mathcal{F}_{j,k-1}] - \nabla F^t(y^k)\right\| &\le \sum_m \frac{w_m \gamma_m}{n_{m,j,k}} \sum_{i \in \mathcal{B}_{m,j,k}} (t - t_{m,i}) \\ &\le 2 \sum_m a_m h_{m,j} + \bar{\gamma}(t - s_j) =: \delta_j(t).\end{aligned}$$

Conditionally on $\mathcal{F}_{j,k-1}$, the centered noises of different samples and clients are independent, so their second moments add:

$$\mathbb{E}\Big[\left\|g^k - \mathbb{E}[g^k \mid \mathcal{F}_{j,k-1}]\right\|^2 \;\Big|\; \mathcal{F}_{j,k-1}\Big] \le V_{j,k} := \sum_m \frac{w_m^2 \sigma_m^2}{n_{m,j,k}}.$$

Apply Lemma C.1 with $F = F^t$ and $\delta = \delta_j(t)$. With $S_j = \sum_{k=1}^K k^2 V_{j,k}$ and the solve output x_j, this gives the exact-count bound

$$\mathbb{E}[F^t(x_j) - \min_{\mathcal{X}} F^t] \le \frac{4LD^2 + 2\sqrt{2}D\sqrt{S_j}}{K(K+1)} + D\Big(2\sum_m a_m h_{m,j} + \bar{\gamma}(t - s_j)\Big). \tag{28}$$

The argument applies separately to every deterministic report time; it yields the pointwise expectation needed below without any uniform continuous-time high-probability statement.

Step 4: the heterogeneous count bound. Put $\bar{V}_j = \sum_m w_m^2 \sigma_m^2 / n_{m,j}$. From (27), $V_{j,k} \le 2K\bar{V}_j$, whence

$$S_j \le 2K\bar{V}_j \sum_{k=1}^K k^2 \le \frac{2}{3}\bar{V}_j K^2 (K+1)^2.$$

Thus the statistical term in (28) is at most $4D\sqrt{\bar{V}_j}/\sqrt{3}$, and by (26)

$$\bar{V}_j \le 2\sum_m \frac{b_m}{h_{m,j}}.$$

Using $t - s_j < c + \Delta$ we arrive at

$$\mathbb{E}[F^t(x_j) - \min_{\mathcal{X}} F^t] \le \frac{4LD^2}{K(K+1)} + \beta_0 D\sqrt{\sum_m \frac{b_m}{h_{m,j}}} + 2D\sum_m a_m h_{m,j} + D\bar{\gamma}(c + \Delta), \tag{29}$$

with $\beta_0 = 4\sqrt{2/3}$; this is (6).

Step 5: time averaging, including prefixes and stationary clients. For $h_m = \infty$ we use the conventions $1/h_m = 0$ and $a_m h_m = 0$, the latter only when $\gamma_m = 0$. Then

$$\frac{1}{h_{m,j}} = \frac{1}{\min\{s_j, h_m\}} \le \frac{1}{s_j} + \frac{1}{h_m}, \qquad \sum_m a_m h_{m,j} \le \sum_m a_m h_m,$$

and therefore, with $\Xi^2 = \sum_m b_m$,

$$\sqrt{\sum_m \frac{b_m}{h_{m,j}}} \le \sqrt{\sum_m \frac{b_m}{h_m}} + \frac{\Xi}{\sqrt{s_j}}.$$

In particular, a moving client's prefix count is never replaced by $\lambda_m h_m/2$ before its first complete window exists.

When $T \ge s_0 + c$, the report intervals that meet $[0, T]$ have consecutive indices $0, 1, \ldots, j_T$. Since $s \mapsto s^{-1/2}$ is decreasing,

$$\sum_{j=0}^{j_T} \frac{|\mathcal{I}_j \cap [0,T]|}{\sqrt{s_j}} \le \frac{\Delta}{\sqrt{s_0}} + \sum_{j=1}^{j_T} \frac{\Delta}{\sqrt{s_0 + j\Delta}} \le \frac{\Delta}{\sqrt{s_0}} + \int_{s_0}^{s_{j_T}} s^{-1/2}\, ds \le \frac{\Delta}{\sqrt{s_0}} + 2\sqrt{T},$$

with the usual convention that an empty sum is zero. Before the first report, the uniform gradient bound gives $F^t(x_\circ) - \min_{\mathcal{X}} F^t \le GD$. Integrating (29) over $[0, T]$, applying the preceding inequalities, and dividing by T proves

$$\begin{aligned}\mathrm{AvgErr}_T \le \frac{4LD^2}{K(K+1)} + \beta_0 D\sqrt{\sum_m \frac{b_m}{h_m}} + 2D\sum_m a_m h_m + D\bar{\gamma}(c+\Delta)\\ + \beta_0 D\,\Xi\Big(\frac{2}{\sqrt{T}} + \frac{\Delta}{T\sqrt{s_0}}\Big) + \frac{GD(s_0+c)}{T},\end{aligned}$$

which is (7)–(8). All expectations were conditional on the fixed schedule; the same inequalities hold unconditionally if the regularity assumption holds almost surely. The clocks and snapshots use arrived data only, and the K rounds cost $c = K\tau \le \Delta$, so the solves and the reporting trajectory are causal. □

C.3 Proof of Corollary 4.2

Lemma C.2 (The window functional)**.** *Let $\alpha, \beta > 0$ and suppose $\mathcal{S}_\gamma > 0$. Over all $h_m > 0$ indexed by the clients with $a_m b_m > 0$,*

$$\inf_h \Big\{\alpha\sum_m a_m h_m + \beta\sqrt{\sum_m b_m/h_m}\Big\} = 3\cdot 2^{-2/3}\alpha^{1/3}\beta^{2/3}\,\mathcal{H}_\gamma,$$

attained at $h_m = \varrho_\star\sqrt{b_m/a_m}$ with $\varrho_\star = (\beta/2\alpha)^{2/3}\mathcal{S}_\gamma^{-1/3}$.

Proof. By the Cauchy–Schwarz inequality,

$$\Big(\sum_m a_m h_m\Big)\Big(\sum_m \frac{b_m}{h_m}\Big) \ge \Big(\sum_m \sqrt{a_m b_m}\Big)^2 = \mathcal{S}_\gamma^2,$$

with equality when $h_m = \varrho\sqrt{b_m/a_m}$ for a common $\varrho > 0$. For a fixed value of the second sum, the first sum is therefore at least the value attained by these proportional windows, so the infimum is attained on them. Along the proportional windows the objective equals $\alpha\varrho\mathcal{S}_\gamma + \beta\sqrt{\mathcal{S}_\gamma/\varrho}$, whose minimizer and minimum are

$$\varrho_\star = \Big(\frac{\beta}{2\alpha}\Big)^{2/3}\mathcal{S}_\gamma^{-1/3}, \qquad 3\cdot 2^{-2/3}\alpha^{1/3}\beta^{2/3}\mathcal{S}_\gamma^{2/3}.$$ □

Proof of Corollary 4.2. First suppose $\mathcal{S}_\gamma > 0$ and consider the clients with $a_m b_m > 0$. The coefficients in (7) are $\alpha = 2$ and $\beta = \beta_0 = 4\sqrt{2/3}$, for which $\varrho_\star = (2/3)^{1/3}\mathcal{S}_\gamma^{-1/3}$. Lemma C.2 therefore gives the ideal windows

$$h_m^\star = (2/3)^{1/3}\mathcal{S}_\gamma^{-1/3}\sqrt{b_m/a_m}, \qquad a_m b_m > 0,$$

with statistical-drift value $6(2/3)^{1/3}\mathcal{H}_\gamma$.

To meet the finite-arrival constraint, set $h_m = \max\{h_m^\star, 2K/\lambda_m\}$ on these clients. Increasing a window cannot increase its contribution to $\sum_m b_m/h_m$, and the increase in the weighted drift cost is at most

$$2\sum_m a_m(h_m - h_m^\star) \le 4K\sum_m \frac{a_m}{\lambda_m}.$$

A client with $a_m > 0 = b_m$ receives $h_m = 2K/\lambda_m$; its entire drift cost is contained in the same bound. A client with $a_m = 0$ receives $h_m = \infty$; it has no limiting drift or variance contribution, while its finite-prefix variance remains in the transient of Theorem 4.1. Restoring the factor D and the remaining terms of the theorem proves (9).

If $\mathcal{S}_\gamma = 0$, every client with positive drift has zero noise. Use $h_m = 2K/\lambda_m$ for those clients and $h_m = \infty$ for the zero-drift clients. Then $\sum_m b_m/h_m = 0$ and $2\sum_m a_m h_m = 4K\sum_m a_m/\lambda_m$, so the same bound holds with $\mathcal{H}_\gamma = 0$, without interpreting a singular window formula.

Finally, inequality (5) is the Cauchy–Schwarz bound

$$\mathcal{S}_\gamma = \sum_m \sqrt{a_m b_m} \le \sqrt{\sum_m a_m}\sqrt{\sum_m b_m} = \sqrt{\bar\gamma}\,\Xi,$$

raised to the power $2/3$; equality holds exactly when $(a_m)_m$ and $(b_m)_m$ are proportional. □

C.4 Separate source and usable timestamps

For the extension of Appendix B (Section B.5), condition on the joint exogenous array of usable and source timestamps. The buffer construction and the count bounds use the usable times $t_{m,i}$. If the source time satisfies $\max\{0, t_{m,i} - \Delta_m^{\mathrm{src}}\} \le t_{m,i}^{\mathrm{src}} \le t_{m,i}$, then every selected sample obeys

$$t - t_{m,i}^{\mathrm{src}} = (t - t_{m,i}) + (t_{m,i} - t_{m,i}^{\mathrm{src}}) \le t - s_j + 2h_{m,j} + \Delta_m^{\mathrm{src}}.$$

Its conditional mean gradient is $\nabla f_m^{t_{m,i}^{\mathrm{src}}}(y^k)$ by the source-time oracle assumption, so the bias in Step 3 of the proof of Theorem 4.1 is at most $\delta_j(t) + \sum_m a_m \Delta_m^{\mathrm{src}}$. The conditional noise variance and all usable-time counts are unchanged. Lemma C.1 consequently adds exactly $D\sum_m a_m \Delta_m^{\mathrm{src}}$ to each pointwise bound, and at most that amount to the average bound. Since the source-age term does not depend on the buffer lengths, the same addition applies to Corollary 4.2. The independence used here is conditional on both timestamp arrays; the calculation does not supply an unbiased oracle for censored or outcome-dependent label selection.

D Projected SGD with client-specific memories

We first prove a projected-SGD potential for the bounded-bias oracle and then apply the same marginal filtration, age bound, and count estimates as in the accelerated analysis. This isolates the effect of the optimizer from the clientwise memory allocation.

Lemma D.1 (A projected-SGD potential with bounded bias)**.** *Let F be convex and L-smooth, $L > 0$, on a compact convex set $\mathcal{X}$ of diameter $D > 0$. Fix $z^0 = x_\circ \in \mathcal{X}$, an integer $K \ge 1$, and a deterministic $0 < \eta \le 1/(2L)$. For $k = 1, \dots, K$ let*

$$z^k = \Pi_{\mathcal{X}}(z^{k-1} - \eta g^k), \qquad g^k = \nabla F(z^{k-1}) + \delta_k + \varepsilon_k,$$

where z^{k-1} and δ_k are $\mathcal{F}_{k-1}$-measurable, $\|\delta_k\| \le \delta$, and

$$\mathbb{E}[\varepsilon_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\varepsilon_k\|^2 \mid \mathcal{F}_{k-1}] \le V_k$$

for finite deterministic $V_k \ge 0$. Then, for $\bar z = K^{-1}\sum_{k=1}^K z^k$,

$$\mathbb{E}[F(\bar z) - \min_{\mathcal{X}} F] \le \frac{D^2}{2\eta K} + \frac{\eta}{K}\sum_{k=1}^K V_k + \delta D. \tag{30}$$

Write $U = \sum_{k=1}^K V_k$. With $\eta = \min\{1/(2L), D/\sqrt{2U}\}$ if $U > 0$, and $\eta = 1/(2L)$ if $U = 0$,

$$\mathbb{E}[F(\bar z) - \min_{\mathcal{X}} F] \le \frac{LD^2}{K} + \frac{\sqrt{2}D\sqrt{U}}{K} + \delta D. \tag{31}$$

Proof. Fix $x^\star \in \arg\min_{\mathcal{X}} F$, write $r_k = \|z^k - x^\star\|^2$, and combine smoothness at z^{k-1} with the supporting-hyperplane inequality for $F(x^\star)$:

$$F(z^k) - F(x^\star) \le \langle \nabla F(z^{k-1}), z^k - x^\star\rangle + \frac{L}{2}\|z^k - z^{k-1}\|^2.$$

The projection variational inequality gives

$$\langle g^k, z^k - x^\star\rangle \le \frac{r_{k-1} - r_k - \|z^k - z^{k-1}\|^2}{2\eta}.$$

After substituting $\nabla F(z^{k-1}) = g^k - \delta_k - \varepsilon_k$, the bias term is bounded by $-\langle \delta_k, z^k - x^\star \rangle \le \delta D$, because both z^k and $x^\star$ lie in $\mathcal{X}$. Split the noise term as

$$-\langle \varepsilon_k, z^k - x^\star \rangle = -\langle \varepsilon_k, z^{k-1} - x^\star \rangle - \langle \varepsilon_k, z^k - z^{k-1} \rangle.$$

Since $\eta \le 1/(2L)$, we have $1/(2\eta) - L/2 \ge 1/(4\eta)$, and Young's inequality gives

$$-\langle \varepsilon_k, z^k - z^{k-1} \rangle \le \eta \|\varepsilon_k\|^2 + \frac{1}{4\eta} \|z^k - z^{k-1}\|^2,$$

so the squared-step terms have a nonpositive sum. Conditional centering removes $\mathbb{E}\langle \varepsilon_k, z^{k-1} - x^\star \rangle$. Therefore

$$\mathbb{E}[F(z^k) - F(x^\star)] \le \frac{\mathbb{E} r_{k-1} - \mathbb{E} r_k}{2\eta} + \eta V_k + \delta D. \tag{32}$$

Summing over k, using $r_0 \le D^2$ and $r_K \ge 0$, and applying convexity to the average of the post-update points proves (30); the accumulated bias term remains δD after division by K.

If the chosen stepsize is $D/\sqrt{2U}$, then $D^2/(2\eta) + \eta U = \sqrt{2} D \sqrt{U}$. If the cap $1/(2L)$ is active, then $U \le 2L^2 D^2$ and

$$\frac{D^2}{2\eta} + \eta U = LD^2 + \frac{U}{2L} \le LD^2 + \frac{D\sqrt{U}}{\sqrt{2}}.$$

Both cases imply (31). When $U = 0$, the cap in (30) gives $LD^2/K + \delta D$ directly. □

D.1 Proof of Proposition B.1

Proof. Condition on the fixed exogenous regular schedule. The frozen windows, timestamp-only partitions, counts, $V_{j,k}$, and η_j^{sgd} are then deterministic. In solve j, reveal only its batches $1, \dots, k$ at step k, starting from the fixed $z_j^0 = x_\circ$. Then z_j^{k-1} is measurable with respect to the earlier batches, and the next batch is independent of them. As in the accelerated proof, this is a marginal analysis of one reset solve; earlier uses of the same observations in other solves are not conditioned upon.

Fix a deterministic $t \in \mathcal{I}_j = [s_j + c, s_{j+1} + c)$, where $c = K\tau$. The pointwise oracle assumptions, independence within the current batch, and the age bound (25) give

$$\left\| \mathbb{E}[g_j^k \mid \mathcal{F}_{j,k-1}] - \nabla F^t(z_j^{k-1}) \right\| \le 2 \sum_m a_m h_{m,j} + \bar{\gamma}(t - s_j) =: \delta_j(t),$$

$$\mathbb{E}\left[\left\| g_j^k - \mathbb{E}[g_j^k \mid \mathcal{F}_{j,k-1}] \right\|^2 \,\middle|\, \mathcal{F}_{j,k-1} \right] \le V_{j,k}.$$

The mean-gradient discrepancy is predictable and bounded by the deterministic $\delta_j(t)$, so Lemma D.1 applies with $F = F^t$. Let $\bar{V}_j = \sum_m w_m^2 \sigma_m^2 / n_{m,j}$. The balanced-batch and regular-count bounds (27) and (26) imply

$$V_{j,k} \le 2K\bar{V}_j, \qquad U_j \le 2K^2 \bar{V}_j \le 4K^2 \sum_m \frac{b_m}{h_{m,j}}.$$

Thus, with $\beta_{\mathrm{sgd}} = 2\sqrt{2}$ and $t - s_j < c + \Delta$,

$$\mathbb{E}[F^t(\bar{z}_j) - \min_{\mathcal{X}} F^t] \le \frac{LD^2}{K} + \beta_{\mathrm{sgd}} D \sqrt{\sum_m \frac{b_m}{h_{m,j}}} + 2D \sum_m a_m h_{m,j} + D\bar{\gamma}(c + \Delta), \tag{33}$$

which is the asserted pointwise replacement, including the zero-variance case.

For the time average, the finite and infinite window conventions give, as before,

$$\sum_m a_m h_{m,j} \le \sum_m a_m h_m, \qquad \sqrt{\sum_m \frac{b_m}{h_{m,j}}} \le \sqrt{\sum_m b_m / h_m} + \frac{\Xi}{\sqrt{s_j}},$$

and the estimate from the proof of Theorem 4.1 gives $\sum_j |\mathcal{I}_j \cap [0,T]| / \sqrt{s_j} \le \Delta/\sqrt{s_0} + 2\sqrt{T}$. During the initial interval of length $s_0 + c$, the report is $x_\circ$ and its gap is at most GD. Integrating

(33), for every $T \geq s_0 + c$ we obtain

$$\mathrm{AvgErr}_T \leq \frac{LD^2}{K} + D\Big(2\sum_m a_m h_m + \beta_{\mathrm{sgd}}\sqrt{\sum_m b_m/h_m}\Big) + D\bar{\gamma}(c+\Delta) + \mathcal{T}_T^{\mathrm{sgd}}, \tag{34}$$

$$\mathcal{T}_T^{\mathrm{sgd}} := \beta_{\mathrm{sgd}} D\,\Xi\Big(\frac{2}{\sqrt{T}} + \frac{\Delta}{T\sqrt{s_0}}\Big) + \frac{GD(s_0+c)}{T}. \tag{35}$$

These are precisely the replacements (15) in (7)–(8). Infinite windows are used only when $a_m = 0$; their stationary noise is retained in $\mathcal{T}_T^{\mathrm{sgd}}$.

Finally, the first query $z_j^0 = x_\circ$ does not depend on the stepsize, so clients can return their batch-size vectors with its gradient, and the server computes U_j and η_j^{sgd} before the first update. The implementation therefore uses the same K gradient-message rounds and report times as Algorithm 1, and consults no sample that arrives after s_j. □

D.2 Optimizing the SGD windows

Under the positivity conditions of Corollary 4.2, set

$$h_m^{\star,\mathrm{sgd}} = 2^{-1/3}\mathcal{S}_\gamma^{-1/3}\sqrt{b_m/a_m}, \qquad h_m = \max\{h_m^{\star,\mathrm{sgd}}, 2K/\lambda_m\} \tag{36}$$

for $a_m > 0$, and $h_m = \infty$ for $a_m = 0$. Then

$$\mathrm{AvgErr}_T \leq 6\cdot 2^{-1/3} D\mathcal{H}_\gamma + \frac{LD^2}{K} + 4KD\sum_m \frac{a_m}{\lambda_m} + D\bar{\gamma}(K\tau+\Delta) + \mathcal{T}_T^{\mathrm{sgd}}. \tag{37}$$

Indeed, Lemma C.2 with $\alpha = 2$ and $\beta = \beta_{\mathrm{sgd}} = 2\sqrt{2}$ gives $\varrho_\star = 2^{-1/3}\mathcal{S}_\gamma^{-1/3}$ and the value $6\cdot 2^{-1/3}\mathcal{H}_\gamma$ for the statistical-drift expression in (34). Raising an ideal window to its count floor cannot increase the variance term and raises the weighted drift cost by at most $4K\sum_m a_m/\lambda_m$, which proves (37). As in the accelerated bound, a client with $a_m > 0 = b_m$ receives $h_m = 2K/\lambda_m$, and when $\mathcal{S}_\gamma = 0$ the moving clients receive these finite windows and the stationary clients infinite ones; the same bound then holds with $\mathcal{H}_\gamma = 0$.

The statistical-drift dependence therefore comes from the memory balance and the independent-batch oracle alone. For $\mathcal{H}_\gamma > 0$, the optimization term is at most $D\mathcal{H}_\gamma$ under the sufficient integer choices

$$K_{\mathrm{sgd}} = \Big\lceil \max\{1, LD/\mathcal{H}_\gamma\}\Big\rceil, \qquad K_{\mathrm{acc}} = \Big\lceil \max\{1, 2\sqrt{LD/\mathcal{H}_\gamma}\}\Big\rceil,$$

which for $LD/\mathcal{H}_\gamma \geq 1$ scale linearly and as a square root, respectively. In both cases the sample-count, timing, and transient terms must be controlled as well.

E Lower bounds under fixed clientwise drift

We prove Theorem B.2 on the segment $\mathcal{X}_D = [-D/2, D/2]\times\{0\}^{d-1}$ of Section 5. The vectors indexed by client groups below organize the proof; they do not enlarge the optimization domain, which uses one coordinate only. We restrict the instance class to the public loss $\ell(x;\xi) = \xi x_1$. This subclass is contained in the minimax supremum, and knowing its loss map reveals nothing about its time-indexed sample laws.

Fix the exogenous schedule N, the horizon T, and one (possibly randomized) causal rule $\mathsf{A} \in \mathfrak{A}_T(\mathsf{N})$ before introducing any prior. Every probability and divergence in this appendix is conditional on the fixed schedule. A random sign array serves only to average risks over admissible instances: each deterministic realization of the array, not the mixture over arrays, is a member of the independent-stream class $\mathfrak{P}_T(\mathsf{N};\gamma)$.

E.1 Proof of Theorem B.2

The instance and its drift bounds. Let

$$\Lambda(u) = 1 - |2u-1|, \qquad u \in [0,1],$$

be the triangular profile: it is nonnegative, vanishes at the endpoints, has Lipschitz constant 2, and satisfies $\int_0^1 \Lambda(u)\,du = 1/2$. For every group $j \in [J]$ and block $k = 0, 1, \ldots$, draw an independent uniform sign $\varsigma_{j,k} \in \{-1, +1\}$, and for $t \in [kT_j, (k+1)T_j)$ put

$$\Lambda_j(t) = \Lambda\big((t - kT_j)/T_j\big), \qquad \varsigma_j(t) = \varsigma_{j,k}.$$

A client $m \in \mathcal{J}_j$ has sample law and loss

$$\xi \sim \mathcal{N}\big(A_m \varsigma_j(t)\Lambda_j(t), \sigma_m^2\big) \quad \text{at time } t, \qquad \ell(x;\xi) = \xi x_1,$$

and a client outside the selected groups has law $\mathcal{N}(0, \sigma_m^2)$ at every time. For each fixed sign array, all samples are independent across arrivals and clients. Their laws depend on their timestamps but not on the query points, so repeated or adaptive queries reveal no more than the raw arrived sample, which the pooled transcript already contains.

For $m \in \mathcal{J}_j$ the population gradient is $A_m \varsigma_j(t)\Lambda_j(t)e_1$. The path is continuous across block boundaries and piecewise linear with slopes of magnitude at most $2A_m/T_j \le \gamma_m$, so it satisfies the clientwise Lipschitz bound (2) for all pairs of times, including pairs in different blocks. Unselected clients have zero drift. All objectives are linear, hence convex and L-smooth for every $L \ge 0$, and the gradient noise variance is exactly σ_m^2.

Set

$$\zeta_j(t) = C_j \Lambda_j(t), \qquad \zeta(t) = (\zeta_j(t))_{j=1}^J, \qquad \Theta(t) = \sum_{j=1}^J \zeta_j(t)\varsigma_j(t).$$

The population objective is $F^t(x) = \Theta(t)x_1$, and

$$\sup_{t,x} \big\|\nabla F^t(x)\big\| = \sup_t |\Theta(t)| \le \sum_{j=1}^J C_j \le G.$$

Thus every deterministic sign path meets all constraints of the class. Each client's drift bound is respected separately; no public aggregate budget is redistributed.

Information about a group's current sign. Fix a time t and a group j. Conditionally on either value of $\varsigma_j(t)$, the samples of that group in its current block form a product of Gaussians. The block has length $T_j \ge 2/\lambda_m$ for every $m \in \mathcal{J}_j$, so two-sided regularity bounds its count at client m by $2\lambda_m T_j$. Endpoint observations, if any, have mean zero and carry no information about either adjacent sign. The KL divergence between the two laws of a full block is bounded by

$$\mathrm{KL}\big(\mathbb{P}_{j,+}^{\mathrm{block}} \,\big\|\, \mathbb{P}_{j,-}^{\mathrm{block}}\big) \le \sum_{m \in \mathcal{J}_j} \sum_{s \text{ in block}} \frac{2A_m^2 \Lambda_j(s)^2}{\sigma_m^2} \le 4T_j \sum_{m \in \mathcal{J}_j} \frac{\lambda_m A_m^2}{\sigma_m^2} \le \frac{1}{128}. \tag{38}$$

The law of the samples arrived by t is a marginal of the full-block law, so it obeys the same bound, and Pinsker's inequality gives

$$\mathrm{TV}\big(\mathbb{P}_{j,+}^t, \mathbb{P}_{j,-}^t\big) \le \frac{1}{16}, \tag{39}$$

where $\mathbb{P}_{j,\pm}^t$ are the laws of the group's current-block samples arrived by t under its two signs.

Let $\mathcal{F}_t$ be the pooled transcript: all samples arrived by t together with the rule's private seed. Define the posterior mean

$$\bar{\varsigma}_j(t) = \mathbb{E}[\varsigma_j(t) \mid \mathcal{F}_t].$$

Samples from earlier blocks of group j are independent of its current sign under the prior, and all raw streams of the other groups and the private seed are independent of both that sign and its current-block samples. Hence $\bar{\varsigma}_j(t)$ is the posterior mean given the current-block samples of group j alone; the variables $\bar{\varsigma}_j(t)$ are independent across groups and satisfy $\mathbb{E}\bar{\varsigma}_j(t) = \mathbb{E}\varsigma_j(t) = 0$.

For a uniform binary prior with posterior mean $\bar{\varsigma}$, the identity

$$\mathbb{E}|\bar{\varsigma}| = \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \tag{40}$$

follows from the mixture density: if ρ_+, ρ_- are densities with respect to a common dominating measure, then $\bar{\varsigma} = (\rho_+ - \rho_-)/(\rho_+ + \rho_-)$ while the mixture density is $(\rho_+ + \rho_-)/2$. Since $|\bar{\varsigma}_j(t)| \le 1$, (39)–(40) imply

$$\mathbb{E}\bar{\varsigma}_j(t)^2 \le \mathbb{E}|\bar{\varsigma}_j(t)| \le \frac{1}{16}. \tag{41}$$

These statements use the product structure of the raw group streams; they do not assume that an algorithm's messages from different groups remain independent after adaptive processing.

Posterior uncertainty implies optimization error. The minimum of $\Theta(t)x_1$ on $\mathcal{X}_D$ is $-D|\Theta(t)|/2$. For any causal report $x_t \in \mathcal{X}_D$, conditioning its first coordinate on the pooled transcript gives

$$\mathbb{E}\big[F^t(x_t) - \min_{\mathcal{X}_D} F^t\big] = \mathbb{E}[\Theta(t)(x_t)_1] + \frac{D}{2}\mathbb{E}|\Theta(t)| \ge \frac{D}{2}\Big(\mathbb{E}|\Theta(t)| - \mathbb{E}\big|\mathbb{E}[\Theta(t) \mid \mathcal{F}_t]\big|\Big). \tag{42}$$

This is the Bayes-optimal decision bound for a linear objective; it already allows arbitrary randomized causal rules.

At a fixed time the current signs are independent Rademacher variables, so

$$\mathbb{E}\Theta(t)^2 = \|\zeta(t)\|^2, \qquad \mathbb{E}\Theta(t)^4 = 3\|\zeta(t)\|^4 - 2\sum_j \zeta_j(t)^4 \le 3\|\zeta(t)\|^4.$$

Hölder's inequality in the form $(\mathbb{E}Z^2)^{3/2} \le \mathbb{E}|Z|\,(\mathbb{E}Z^4)^{1/2}$ therefore yields

$$\mathbb{E}|\Theta(t)| \ge \frac{1}{\sqrt{3}}\|\zeta(t)\|; \tag{43}$$

this is immediate when $\zeta(t) = 0$, and otherwise follows by dividing by the positive fourth-moment bound. On the other hand, $\mathbb{E}[\Theta(t) \mid \mathcal{F}_t] = \sum_j \zeta_j(t)\bar{\varsigma}_j(t)$, and the zero means, independence across groups, and (41) imply

$$\mathbb{E}\Big|\sum_j \zeta_j(t)\bar{\varsigma}_j(t)\Big| \le \Big(\sum_j \zeta_j(t)^2\,\mathbb{E}\bar{\varsigma}_j(t)^2\Big)^{1/2} \le \frac{1}{4}\|\zeta(t)\|. \tag{44}$$

Combining (42)–(44) proves the pointwise Bayes bound

$$\mathbb{E}\big[F^t(x_t) - \min_{\mathcal{X}_D} F^t\big] \ge \frac{D}{2}\Big(\frac{1}{\sqrt{3}} - \frac{1}{4}\Big)\|\zeta(t)\|. \tag{45}$$

Different clocks and the horizon average. Because $T \ge T_j$, complete blocks of group j cover at least half of the horizon. The triangular profile has mean $1/2$ on a complete block and is nonnegative on the remaining partial block, so

$$\frac{1}{T}\int_0^T \Lambda_j(t)\,dt \ge \frac{\lfloor T/T_j \rfloor T_j}{2T} \ge \frac{1}{4}.$$

The triangle inequality for the vector integral of $\zeta(t) = (C_j\Lambda_j(t))_j$ gives

$$\frac{1}{T}\int_0^T \|\zeta(t)\|\,dt \ge \Big(\sum_{j=1}^J C_j^2\Big[\frac{1}{T}\int_0^T \Lambda_j(t)\,dt\Big]^2\Big)^{1/2} \ge \frac{1}{4}\Big(\sum_{j=1}^J C_j^2\Big)^{1/2}. \tag{46}$$

This is the step that keeps the different group clocks inside one time-averaged hard instance, instead of replacing the time average by a different hard instance at every evaluation time.

Integrating (45) and using (46) shows that the sign-averaged risk of the fixed rule A is at least

$$\frac{1/\sqrt{3} - 1/4}{8}\,D\Big(\sum_{j=1}^J C_j^2\Big)^{1/2}.$$

Only finitely many group-block signs affect $[0, T]$, so at least one deterministic realization of them attains the average risk. Extend that realization arbitrarily on later blocks, keeping the triangular profiles; the extension preserves the drift and gradient bounds and is not revealed to the rule. Because the rule and the horizon were fixed before the sign averaging, this proves

$$\forall \mathsf{A} \in \mathfrak{A}_T(\mathsf{N}) \quad \exists \mathsf{P} \in \mathfrak{P}_T(\mathsf{N}; \gamma)$$

with the asserted risk, and hence Theorem B.2. The selected path may depend on the rule and the horizon; no single path is claimed to be worst for all rules or horizons. □

E.2 Proof of Corollary 5.1

Consider the clients with positive prescribed drift and variance. Since $q_m = \sqrt{b_m/a_m}$, elementary substitution gives

$$a_m = \frac{\sqrt{a_m b_m}}{q_m}, \qquad b_m = \sqrt{a_m b_m}\, q_m, \qquad \frac{\lambda_m \gamma_m^2}{\sigma_m^2} = \frac{a_m^2}{b_m} = \frac{\sqrt{a_m b_m}}{q_m^3}.$$

In each bin set $A_m = \gamma_m T_j/2$, as in the corollary. The drift constraint is then saturated, and

$$4T_j \sum_{m \in \mathcal{J}_j} \frac{\lambda_m A_m^2}{\sigma_m^2} = T_j^3 Q_j = \frac{1}{128}.$$

The remaining hypotheses of Theorem B.2 are exactly the count, horizon, and G conditions of the corollary.

Let $\underline{q}_j$ be the lower endpoint of the jth dyadic bin, so that $\underline{q}_j \le q_m \le 2\underline{q}_j$ for all $m \in \mathcal{J}_j$. Then

$$C_j = \frac{1}{2 \cdot 128^{1/3}} \frac{\sum_{m \in \mathcal{J}_j} \sqrt{a_m b_m}/q_m}{\left(\sum_{m \in \mathcal{J}_j} \sqrt{a_m b_m}/q_m^3\right)^{1/3}} \ge \frac{1}{2 \cdot 128^{1/3}} \frac{\mathcal{S}_j/(2\underline{q}_j)}{(\mathcal{S}_j/\underline{q}_j^3)^{1/3}} = \frac{\mathcal{S}_j^{2/3}}{4 \cdot 128^{1/3}}. \tag{47}$$

Inserting this into Theorem B.2 gives the group-sensitive bound

$$\mathcal{R}_T^\star(\gamma) \ge c_{\mathrm{het}} D \Big(\sum_{j=1}^{J} \mathcal{S}_j^{4/3}\Big)^{1/2}, \qquad c_{\mathrm{het}} = \frac{1/\sqrt{3} - 1/4}{32 \cdot 128^{1/3}},$$

which is (10). Hölder's inequality gives

$$\sum_j \mathcal{S}_j \le J^{1/4} \Big(\sum_j \mathcal{S}_j^{4/3}\Big)^{3/4},$$

and raising to the power $2/3$ and rearranging yields

$$\Big(\sum_j \mathcal{S}_j^{4/3}\Big)^{1/2} \ge J^{-1/6} \Big(\sum_j \mathcal{S}_j\Big)^{2/3} = J^{-1/6} \mathcal{H}_\gamma,$$

which is (11). Dyadic bins starting at $q_{\min}$ need at most $1 + \lceil \log_2(q_{\max}/q_{\min}) \rceil$ nonempty bins, and empty bins may be omitted. The argument proves a comparison with this bin factor; it does not show that the factor is necessary. □

E.3 A finite-horizon statistical term

The fixed-vector class contains stationary objectives even when some public drift bounds are positive. The next bound retains the statistical cost of every client on a finite horizon, including clients with $\gamma_m = 0$.

Proposition E.1 (Stationary contribution at a finite horizon)**.** *Fix $D > 0$, $T > 0$, positive $(w_m, \lambda_m, \sigma_m)_m$, nonnegative $(\gamma_m)_m$, and a λ-regular schedule on $[0, \infty)$. For the fixed-vector class and the pooled causal rules of Section 5,*

$$\mathcal{R}_T^\star(\gamma) \ge \frac{D}{4} \min\Big\{G,\ \frac{1}{\sqrt{8}} \Big(\sum_m \frac{w_m^2 \sigma_m^2}{\lambda_m T + 2}\Big)^{1/2}\Big\}.$$

For every fixed rule, one of two stationary linear instances attains the bound.

Proof. Choose amplitudes $A_m \ge 0$ and one uniform sign $\varsigma \in \{-1, +1\}$. At every time, client m supplies an independent sample from $\mathcal{N}(\varsigma A_m, \sigma_m^2)$ with loss $\ell(x; \xi) = \xi x_1$. Let $C = \sum_m w_m A_m$. The objectives have zero drift and population gradient bound C, so both signs give admissible instances whenever $C \le G$.

For every $T > 0$, regularity implies $N_m(T) \le 2\lambda_m T + 4$: when $T \ge 2/\lambda_m$ the upper inequality applies directly, and otherwise $N_m(T) \le N_m(2/\lambda_m) \le 4$. The two full transcripts through time T therefore have KL divergence at most

$$\sum_m \frac{2A_m^2 N_m(T)}{\sigma_m^2} \le 4 \sum_m \frac{(\lambda_m T + 2)A_m^2}{\sigma_m^2}.$$

If the last sum, without the factor 4, is at most $1/8$, the full transcripts have total variation at most $1/2$, and so do the partial transcripts at every earlier time. At every time, the two objective gaps of any common candidate action add up to DC, so the two-point Bayes argument gives expected gap at least $DC/4$ at every time and hence for the average over $[0, T]$.

By the Cauchy–Schwarz inequality,

$$\sup\Big\{\sum_m w_m A_m : \sum_m \frac{(\lambda_m T + 2)A_m^2}{\sigma_m^2} \le \frac{1}{8}\Big\} = \frac{1}{\sqrt{8}}\Big(\sum_m \frac{w_m^2 \sigma_m^2}{\lambda_m T + 2}\Big)^{1/2},$$

with equality for nonnegative A_m proportional to $w_m \sigma_m^2/(\lambda_m T + 2)$. If this value exceeds G, scale the amplitudes down until $C = G$; the information constraint remains valid. Maximizing the Bayes bound over these amplitudes and selecting one of the two signs proves the claim. When $G = 0$ the bound is zero and there is nothing to prove. □

The drifting and stationary constructions give separate lower bounds on the same class, so their maximum is a valid lower bound. We do not add their values, which would require a joint construction.

F The joint comparison regime and the remaining scale gap

F.1 Proof of Corollary B.3

The choice of K gives $4LD/[K(K+1)] \le 4LD/K^2 \le \mathcal{H}_\gamma$. Since $\Delta = K\tau$, the sum of the count-floor and timing charges in (9) equals $DK\mathcal{C}_\gamma \le D\mathcal{H}_\gamma$. The statistical threshold in (17) gives $2\beta_0 \Xi/\sqrt{T} \le \mathcal{H}_\gamma/2$, and its second line bounds the remaining transient terms, divided by D, by $\mathcal{H}_\gamma/2$; thus $\mathcal{T}_T \le D\mathcal{H}_\gamma$. Corollary 4.2 now proves the upper guarantee uniformly over the fixed-vector class. The algorithm belongs to the pooled causal class, so its worst-case risk bounds $\mathcal{R}_T^\star(\gamma)$ from above. The count and gradient hypotheses, together with $T \ge \max_j T_j$, give the lower inequality by Corollary 5.1. □

For orientation, minimizing the continuous relaxation $4LD/K^2 + \mathcal{C}_\gamma K$ over K gives $K = (8LD/\mathcal{C}_\gamma)^{1/3}$ and the value $3(LD\,\mathcal{C}_\gamma^2)^{1/3}$. The sufficient integer condition of the corollary therefore expresses, up to constants, the two requirements $\mathcal{C}_\gamma \lesssim \mathcal{H}_\gamma$ and $LD\,\mathcal{C}_\gamma^2 \lesssim \mathcal{H}_\gamma^3$. The lower-bound count conditions yield only $\sum_m a_m/\lambda_m \le \sum_j C_j$ and impose no restriction on L or τ.

F.2 A feasible family with positive communication cost

For any integer $J \ge 1$, take $M = J$ clients and set

$$q_m = 4^{m-1}, \qquad w_m = \frac{1}{J}, \qquad \lambda_m = \frac{100 J^{1/3}}{q_m}, \qquad \gamma_m = \frac{J}{q_m}, \qquad \sigma_m^2 = 100 J^{7/3}. \tag{48}$$

Then $a_m = 1/q_m$, $b_m = q_m$, and $\mathcal{H}_\gamma = J^{2/3}$. Use $D = 1$, $L = \mathcal{H}_\gamma$, $G = J$, the regular arrivals $N_m(t) = \lfloor \lambda_m t \rfloor$, and

$$K = 2, \qquad \tau = \frac{\mathcal{H}_\gamma}{200\bar{\gamma}} > 0, \qquad \Delta = K\tau, \qquad T = 10^4 q_{\max} J^2.$$

Each nonempty dyadic bin contains exactly one client, and the lower construction has $T_m = 128^{-1/3} q_m$ and $C_m = 128^{-1/3}/2$. Hence $\lambda_m T_m = 100 \cdot 128^{-1/3} J^{1/3} > 2$, $\sum_m C_m = 128^{-1/3} J/2 < G$, and $T \ge \max_m T_m$. The upper count floors are inactive, because $\lambda_m h_m^\star =$

$100(2/3)^{1/3} > 2K$. Moreover $K\mathcal{C}_\gamma/\mathcal{H}_\gamma = 0.10$, and the three nontransient remainder terms in (9), divided by $D\mathcal{H}_\gamma$, are

$$\frac{4LD}{K(K+1)\mathcal{H}_\gamma} = \frac{2}{3}, \qquad \frac{4K\sum_m a_m/\lambda_m}{\mathcal{H}_\gamma} = 0.08, \qquad \frac{\bar{\gamma}(K\tau+\Delta)}{\mathcal{H}_\gamma} = 0.02.$$

Here $\Xi^2 = \sum_m q_m \le 4q_{\max}/3$, $s_0 = q_{\max}/(25J^{1/3})$, and $1 \le \bar{\gamma} < 4/3$. The statistical horizon threshold is at most $64\beta_0^2 T/(3\cdot 10^4) < 0.023\,T$, and each other threshold is below $0.001\,T$, so all four horizon requirements hold. We also obtain

$$\frac{\mathcal{T}_T}{D\mathcal{H}_\gamma} \le \frac{\beta_0}{25\sqrt{3}}J^{-5/3} + \frac{\beta_0}{10^5\sqrt{3}}J^{-11/6} + \frac{1}{250000J^2} + \frac{1}{10^6 J} < 0.076.$$

Consequently all upper remainders together are less than $0.843\,D\mathcal{H}_\gamma$. The parameters vary with J: the example certifies that the joint regime is nonempty, not that it models a fixed empirical population. Note also that on this segment the fixed midpoint already gives $F^t(0) - \min_{\mathcal{X}_D} F^t \le GD/2 = J/2$, since every feasible point lies within $D/2$ of zero and $\|\nabla F^t\| \le G$; this no-data guarantee is smaller than the complete upper expression at the displayed values of J. The family therefore establishes feasibility, not numerically informative constants at those scales.

F.3 A LIMITATION OF THE DISJOINT-GROUP CONSTRUCTION

Proposition F.1 (Retuning does not close the scale gap)**.** *Suppose $M = J$, $a_m = 1/q_m$, $b_m = q_m$, and $q_m = 4^{m-1}$. For any disjoint nonempty groups, block lengths, and amplitudes satisfying the drift and information constraints of Theorem B.2,*

$$\Big(\sum_j C_j^2\Big)^{1/2} \le c_{\mathrm{grp}}\sqrt{J}, \qquad c_{\mathrm{grp}} = \frac{1}{2(1-1/4)} + \frac{1}{\sqrt{512}\,(1-1/2)} < 0.756.$$

Since $\mathcal{H}_\gamma = J^{2/3}$ in this family, the lower expression of the theorem cannot be of order $D\mathcal{H}_\gamma$ uniformly in J, whatever the partition and amplitudes.

Proof. Fix one group j with block length T_j, and write $\bar{A}_m = w_m A_m$ for its clients. The drift condition gives $\bar{A}_m \le T_j/(2q_m)$. The information condition $4T_j\sum_m \bar{A}_m^2/b_m \le 1/128$ separately gives $\bar{A}_m \le \sqrt{q_m/(512T_j)}$ for every selected client. Therefore

$$C_j \le \sum_{q_m > T_j} \frac{T_j}{2q_m} + \sum_{q_m \le T_j} \sqrt{\frac{q_m}{512T_j}} \le \frac{1}{2(1-1/4)} + \frac{1}{\sqrt{512}\,(1-1/2)}.$$

Both sums are dominated by geometric series, and omitting clients only reduces them. There are at most J disjoint nonempty groups, which proves the claim. □

For the singleton groups of (48), the amplitude norm equals $128^{-1/3}\sqrt{J}/2$ exactly. The existing construction thus already achieves its best order on this family, and the $J^{1/6}$ loss in the final Hölder step of Appendix E has the correct order for it. Settling the unrestricted minimax dependence would require a different lower construction or a sharper upper method; Proposition F.1 is not a minimax upper bound.

F.4 NUMERICAL EVALUATION OF THE COMPLETE GUARANTEES

Figure 2 evaluates the displayed bounds without simulating arrival paths. The left panel uses the feasible family above; the right panel evaluates (7) with the selected heterogeneous windows of the three fixed-path experiments, which are not claimed to satisfy Corollary B.3. These calculations distinguish a proved comparison regime from an exploratory instance on which the bound can be loose.

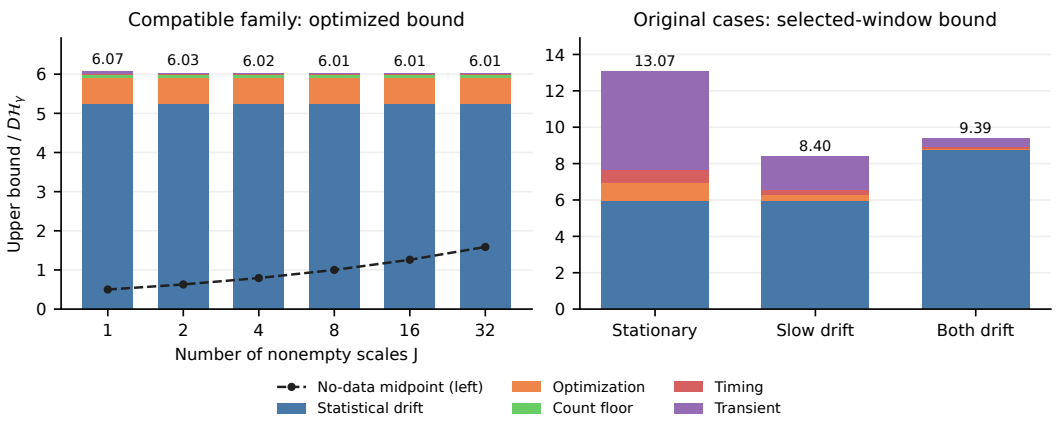

Figure 2: Complete upper guarantees divided by $D\mathcal{H}_\gamma$. Left: Corollary 4.2 on the family (48), including its conservative count-floor surcharge even though the floors are inactive; the dashed line is the no-data midpoint guarantee $GD/2$, which is smaller at all shown scales. Right: Theorem 4.1 at the selected heterogeneous windows of the three original synthetic cases, where no separate floor surcharge arises; the stationary case has a substantial finite-horizon transient. These are evaluations of inequalities, not observed errors, and not evidence that the upper and lower numerical constants are close.

G When forgetting helps on periodic streams

The drift upper bound protects against every admissible path, whereas a particular oscillating path may favor the full prefix. This appendix gives an exact diagnostic for quadratic losses and uniformly weighted completed buffers. It identifies a noise–frequency threshold for a single client and proves the separation between shared and client-specific windows stated in Proposition 5.2. The frequency response of a rectangular average is classical (Smith, 1997, Chapter 15); the calculation here includes the interval over which a completed buffer is held and the noise of the stationary clients.

Model and estimator class. Let $L > 0$ and use the public loss

$$\ell(x; Y) = \frac{L}{2}x^2 - Yx, \qquad x \in \mathcal{X} = [-D/2, D/2].$$

At times i/λ_m, $i = 1, 2, \ldots$, client m observes

$$Y_{m,i} = \bar{\vartheta}_m + A_m \cos(\omega_m i/\lambda_m + \Phi_m) + \sigma_m \varepsilon_{m,i}, \tag{49}$$

where $A_m \geq 0$, $\omega_m > 0$, and the signs $\varepsilon_{m,i}$ are independent Rademacher variables; stationary clients have $A_m = 0$. For the phase-averaged diagnostic, the phases Φ_m are independent and uniform on $[0, 2\pi]$. The phase average is a prior over fixed population paths: conditionally on any fixed phase vector, all observations are independent and the model of Section 3 applies. It does not assert unconditional independence after mixing over phases.

Assume

$$\frac{D}{2} > \frac{1}{L}\sum_m w_m(|\bar{\vartheta}_m| + A_m + \sigma_m), \tag{50}$$

so that all aggregate empirical and population minimizers are interior. The smoothness constant is L, the noise bounds are σ_m^2, the certified drift bounds are $\gamma_m = A_m\omega_m$, and an aggregate gradient bound is $G = LD/2 + \sum_m w_m(|\bar{\vartheta}_m| + A_m)$.

A finite memory at client m has length $h_m = n_m/\lambda_m$ with $n_m \in \mathbb{N}$. At time ph_m its estimate is replaced by the average of the observations with indices $(p-1)n_m + 1, \ldots, pn_m$, and that average is held until $(p+1)h_m$; before the first complete buffer, the arrived prefix is used. A prefix memory always averages all arrived observations. Set $\tau = 0$ and let the server update at the union of the client update times, reporting the exact weighted empirical minimizer $x_t = L^{-1}\sum_m w_m \widehat{\vartheta}_m(t)$, where $\widehat{\vartheta}_m(t)$ is client m's current average. A fixed point $x_\circ$ is reported until $t_\mathrm{s} = \max_m 1/\lambda_m$, when every client has at least one sample. These rules define an admissible empirical estimator; they are not the batch-splitting algorithm of Theorem 4.1.

An exact finite-horizon identity. Let $\mathcal{M}_m(t)$ be the set of observation indices used by client m at time $t \geq t_\mathrm{s}$, let $n_m(t) = |\mathcal{M}_m(t)|$, and let

$$\mathcal{K}_m(t) = \frac{1}{n_m(t)} \sum_{i \in \mathcal{M}_m(t)} e^{\mathrm{i}\omega_m i/\lambda_m}$$

be the complex coefficient of its average. Write $\mathcal{Q}_T$ for the current-objective risk averaged over both the phases and the observation noise. For $T \geq t_\mathrm{s}$,

$$\begin{aligned}\mathcal{Q}_T = \frac{t_\mathrm{s}}{2LT}\Big[\Big(Lx_\circ - \sum_m w_m\bar{\vartheta}_m\Big)^2 + \frac{1}{2}\sum_m w_m^2 A_m^2\Big] \\ + \frac{1}{2LT}\sum_m w_m^2 \int_{t_\mathrm{s}}^T \Big[\frac{A_m^2}{2}\big|\mathcal{K}_m(t) - e^{\mathrm{i}\omega_m t}\big|^2 + \frac{\sigma_m^2}{n_m(t)}\Big]dt.\end{aligned} \tag{51}$$

Finite-horizon initialization and prefix variance can therefore be retained exactly rather than approximated by a steady-state formula.

Theorem G.1 (Exact periodic risk of completed windows)**.** *Under (49)–(50), the long-run phase-averaged risk $\mathcal{Q}_\infty := \lim_{T\to\infty}\mathcal{Q}_T$ exists. Define the phase increment per sample $\varphi_m = \omega_m/\lambda_m$,*

$$\mathcal{K}_n(\varphi) = \frac{1}{n}\sum_{l=0}^{n-1} e^{-\mathrm{i}l\varphi}, \qquad \mathrm{sinc}(z) = \begin{cases}\sin(z)/z, & z \neq 0,\\ 1, & z = 0,\end{cases}$$

and, for a finite window of n samples, the bias factor

$$\Upsilon_n(\varphi) = 1 + |\mathcal{K}_n(\varphi)|^2 - 2\,\mathrm{Re}\Big\{\mathcal{K}_n(\varphi)\, e^{-\mathrm{i}n\varphi/2}\,\mathrm{sinc}(n\varphi/2)\Big\}. \tag{52}$$

If $\sin(\varphi/2) \neq 0$, an equivalent real expression is

$$\Upsilon_n(\varphi) = 1 + \Big(\frac{\sin(n\varphi/2)}{n\sin(\varphi/2)}\Big)^2 - 2\,\frac{\sin(n\varphi/2)}{n\sin(\varphi/2)}\,\mathrm{sinc}(n\varphi/2)\cos\big((2n-1)\varphi/2\big). \tag{53}$$

A client with a finite window of n_m samples contributes

$$\frac{w_m^2}{2L}\Big[\frac{A_m^2}{2}\Upsilon_{n_m}(\varphi_m) + \frac{\sigma_m^2}{n_m}\Big] \tag{54}$$

to $\mathcal{Q}_\infty$; a prefix client contributes the same expression with the bracket replaced by $\frac{1}{2}A_m^2(1+\mathbf{1}\{\varphi_m \in 2\pi\mathbb{Z}\})$. The long-run risk is the sum of these contributions over clients. In particular, a stationary prefix contributes zero, and a nonaliased moving prefix contributes $w_m^2 A_m^2/(4L)$.

Proof. The population objective is $F^t(x) = Lx^2/2 - \vartheta(t)x$ with $\vartheta(t) = \sum_m w_m[\bar{\vartheta}_m + A_m\cos(\omega_m t + \Phi_m)]$, and its gap at x is $(Lx - \vartheta(t))^2/(2L)$. For $t \geq t_\mathrm{s}$, the error of client m's empirical mean is

$$A_m\,\mathrm{Re}\Big\{e^{\mathrm{i}\Phi_m}\big[\mathcal{K}_m(t) - e^{\mathrm{i}\omega_m t}\big]\Big\} + \frac{\sigma_m}{n_m(t)}\sum_{i\in\mathcal{M}_m(t)}\varepsilon_{m,i}.$$

It has zero mean over phase and noise, and independence across clients removes all cross terms even when the replacement times differ. The identity $\mathbb{E}_\Phi(\mathrm{Re}(e^{\mathrm{i}\Phi}z))^2 = |z|^2/2$ and the independence of the noise give (51), including the fixed-initialization interval.

For a completed window of length $h = n/\lambda$ and $t = ph + u$ with $0 \leq u < h$, the noiseless complex coefficient is $\mathcal{K}(t) = e^{\mathrm{i}\omega ph}\mathcal{K}_n(\varphi)$, so the phase-averaged squared bias, divided by $A^2/2$, equals $|\mathcal{K}_n(\varphi) - e^{\mathrm{i}\omega u}|^2$. Averaging over $u \in [0, h)$ and using

$$\frac{1}{h}\int_0^h e^{-\mathrm{i}\omega u}\,du = e^{-\mathrm{i}n\varphi/2}\,\mathrm{sinc}(n\varphi/2)$$

gives (52); the geometric sum for $\mathcal{K}_n$ gives (53). The noise variance is exactly σ^2/n. The initial finite prefix and the incomplete last holding interval vanish after division by T.

For a prefix with $N(t) = \lfloor \lambda t \rfloor$ samples, the complex sample average tends to zero if $\varphi \notin 2\pi\mathbb{Z}$, by the finite geometric-sum formula, and is identically one if $\varphi \in 2\pi\mathbb{Z}$; in the latter case $\omega > 0$ and the time average of $|1 - e^{\mathrm{i}\omega t}|^2$ equals two. The prefix noise variance $\sigma^2/N(t)$ has vanishing time average. Boundedness justifies passing to these limits in (51). □

A necessary and sufficient improvement test. For a nonaliased moving client with $\sigma_m > 0$, replacing only its prefix by a finite completed window of n samples strictly improves the long-run risk if and only if

$$\frac{A_m^2 n}{2\sigma_m^2}\big[1 - \Upsilon_n(\varphi_m)\big] > 1. \tag{55}$$

This exact test includes the hold delay and the discrete arrival rate. If $\Upsilon_n \geq 1$, the window cannot help at any noise level; if $\Upsilon_n < 1$, the display gives the required signal-to-noise ratio. The weights cancel when a single client's window is changed, so the other clients may remain on their prefixes.

For interpretation, take the high-rate limit $\varphi \to 0$ with $n\varphi \to \omega h$. Then

$$\Upsilon_n(\varphi) \longrightarrow 1 + \mathrm{sinc}^2(\omega h/2)\big(1 - 2\cos(\omega h)\big),$$

and the corresponding high-rate bias–variance expression, with $\sigma^2/(\lambda h)$ for the sampling noise, favors the finite window exactly when

$$\frac{A^2\lambda}{\sigma^2\omega}\,\frac{\omega h}{2}\,\mathrm{sinc}^2(\omega h/2)\big(2\cos(\omega h) - 1\big) > 1. \tag{56}$$

Equation (56) tests the limiting expression; (55) remains the exact finite-rate criterion. A long window attenuates noise but may average over several changes of direction; a shorter window helps only if its signal gain exceeds both its remaining hold bias and its variance cost.

Proposition G.2 (A shared window can be forced to keep the prefix)**.** *Suppose one client has a nonaliased sinusoidal mean with amplitude A, frequency ω, and weight w, while all other clients are stationary, and recall $\Xi^2 = \sum_m w_m^2 \sigma_m^2/\lambda_m$. Restrict the estimator to uniformly weighted completed-window empirical minimization with one fixed shared length h, or to the full prefix at every client. For a real length h, the completed interval $((p-1)h, ph]$ contains the actual observations of each client, and its sample count may vary with p. For every real h with $\lambda_m h \geq 1$ at every client, and for every fixed phase,*

$$\liminf_{T\to\infty} \mathrm{AvgErr}_T^{\mathrm{shared}} \geq \frac{w^2A^2}{4L} + \frac{\Xi^2 - 2w^2A^2/\omega}{2Lh}. \tag{57}$$

Hence $\Xi^2 \geq 2w^2A^2/\omega$ makes every such shared window no better than the prefix, and strict inequality makes every finite shared window worse. For the phase-averaged risk at commensurate lengths $\lambda_m h \in \mathbb{N}$, the sharper sufficient condition $\Xi^2 \geq w^2A^2/\omega$ holds.

Proof. First let h be any real length satisfying the count condition, so that every completed interval contains at least one observation per client. The empirical minimizer is constant on each holding interval, even if the server reports more often. Whatever its deterministic conditional mean on that interval, its integrated squared bias is at least the within-interval variance of $wA\cos(\omega t + \Phi)$. The long-run average of $\cos^2(\omega t + \Phi)$ is $1/2$, whereas the square of the mean cosine over any interval of length h is at most $\mathrm{sinc}^2(\omega h/2)$. Therefore the long-run squared bias has lower limit at least $w^2A^2[1/2 - \mathrm{sinc}^2(\omega h/2)]$.

If $n_{m,p}$ is the count in client m's pth completed interval, the independent noise contributes variance $\sum_m w_m^2\sigma_m^2/n_{m,p}$. The average count satisfies $P^{-1}\sum_{p=1}^P n_{m,p} = \lfloor \lambda_m P h \rfloor / P \to \lambda_m h$, so Jensen's inequality gives $\liminf_P P^{-1}\sum_{p=1}^P 1/n_{m,p} \geq 1/(\lambda_m h)$. Finally,

$$h\,\mathrm{sinc}^2(\omega h/2) \leq \min\{h,\, 4/(\omega^2 h)\} \leq 2/\omega.$$

Adding bias and variance and dividing by $2L$ proves (57); bounded startup and incomplete last intervals do not affect the lower limit. Nonaliasing ensures that the full prefix has risk $w^2A^2/(4L)$ for every fixed phase.

For the sharper phase-averaged statement, use a commensurate length and write $v = e^{-\mathrm{i}\omega h/2}\,\mathrm{sinc}(\omega h/2)$. Completing the square in (52) yields

$$1 - \Upsilon_n(\varphi) = 2\,\mathrm{Re}(\mathcal{K}_n v) - |\mathcal{K}_n|^2 \leq |v|^2 = \mathrm{sinc}^2(\omega h/2).$$

The shared-window risk minus the prefix risk is therefore at least

$$\frac{1}{2L}\Big[\frac{\Xi^2}{h} - \frac{w^2A^2}{2}\big(1 - \Upsilon_n(\varphi)\big)\Big] \geq \frac{\Xi^2 - w^2A^2/\omega}{2Lh},$$

and the assertion follows. □

Corollary G.3 (An unbounded ratio within the shared-window class)**.** *Take two clients with $w_1 = w_2 = 1/2$, $L = 1$, $\mathcal{X} = [-2, 2]$ (so $D = 4$), $\bar\vartheta_1 = \bar\vartheta_2 = 0$, $A_1 = \omega_1 = 1$, and $A_2 = 0$. For an integer $r \geq 2$, set*

$$\lambda_1 = r, \quad \sigma_1 = 1/r, \qquad \lambda_2 = 1, \quad \sigma_2 = \sqrt{2}.$$

Among all fixed real shared lengths $h \geq 1$ in Proposition G.2, including the full prefix, the smallest lower limit of the average risk is $1/16$, for every fixed phase. The client-specific choice $h_1 = 1/r$ with a prefix at client two has risk

$$\mathcal{Q}_\infty^{\mathrm{sep}} = \frac{1}{8}\Big[1 - \mathrm{sinc}(1/r) + \frac{1}{r^2}\Big] \leq \frac{7}{48r^2}. \tag{58}$$

Thus the ratio between the best shared-window risk and this client-specific risk is at least $3r^2/7$. The client-specific limit also holds for every fixed phase Φ_1.

Proof. Condition (50) holds because $\frac{1}{2}(1+1/r+\sqrt{2}) < 2$. In the model of Section 3 this instance has $D = 4$, smoothness constant $L = 1$, $G = 5/2$, and drift vector $(1, 0)$. Every real shared length $h \geq 1$ has nonempty completed buffers. Here $\Xi^2 = 1/(4r^3) + 1/2 > 1/2 = 2w_1^2A_1^2/\omega_1$, so Proposition G.2 makes every finite shared window worse than the prefix, whose risk is $w_1^2A_1^2/4 = 1/16$.

For the client-specific rule, $n_1 = 1$ and $\Upsilon_1(\varphi) = 2-2\,\mathrm{sinc}(\varphi)$, while the stationary prefix contributes nothing. Substitution into (54) gives (58); the final inequality uses $1 - \mathrm{sinc}(u) \le u^2/6$, obtained by integrating $1 - \cos(v) \le v^2/2$ from zero to $u > 0$.

To remove the phase averaging for the client-specific rule, expand the moving client's squared bias over successive holding intervals. The phase-dependent remainder is a fixed multiple of $e^{2\mathrm{i}p\omega h}$ (and its conjugate), whose Cesàro mean vanishes whenever $e^{2\mathrm{i}\omega h} \neq 1$; this holds for $h_1 = 1/r$, which is not a nonzero integer multiple of π. The shared-window lower bound already holds for every phase, without any resonance exclusion on h, and the stationary noise is centered, so no cross-client bias term remains. □

The separation concerns uniformly weighted completed-window empirical estimators with one fixed common length. It does not apply to arbitrary learners that store a common full history, since such learners can implement the client-specific effective weights themselves, and it is not a communication lower bound, since $\tau = 0$. What it isolates is the concrete statistical restriction that the common-window comparison imposes.

H Adapting over a finite collection of client memories

The prescribed drift vector determines the oracle windows, but it need not be an input to a finite model-selection procedure. This appendix gives such a procedure under an additional bounded-loss assumption and proves Theorem 6.1 in its complete form (Theorem H.2). The procedure retains the pooled variance of each fixed-memory candidate and pays an explicit selection cost; it aggregates candidates and neither estimates the drift vector nor tracks the best changing candidate.

Additional assumption and data split. Assume that known constants $B_m \ge 0$ satisfy

$$|\ell(x;\xi) - \ell(y;\xi)| \le B_m, \qquad x, y \in \mathcal{X}, \tag{59}$$

almost surely for a sample from client m at any source time, and put $B = \sum_m w_m B_m > 0$. The assumption bounds loss differences, not absolute loss offsets. For logistic loss with feature vectors of norm at most $\bar{z}_m$ on a domain of diameter D one may take $B_m = D\bar{z}_m$; if the sample loss contains a known deterministic regularizer, its oscillation on $\mathcal{X}$ is added to B_m.

Assign the odd-indexed arrivals of each client to a training stream and the even-indexed arrivals to a validation stream, breaking ties in the timestamps by an arbitrary fixed order. Conditionally on the arrival schedule, the two sample arrays are independent. Validation observations are never used in an expert's gradient updates.

Fix $K \ge 1$, $c = K\tau$, and

$$s_0 = 6K/\lambda_{\min}, \qquad \Delta \ge \max\{c,\, 6/\lambda_{\min}\}, \qquad T = s_0 + c + R\Delta, \tag{60}$$

with an integer $R \ge 1$. Let $\mathcal{G} = \{h^1, \dots, h^E\}$, $E \ge 2$, be a deterministic collection of memory vectors with $6K/\lambda_m \le h_m^e \le T$, chosen without knowledge of γ; the value $h_m^e = T$ keeps the training prefix throughout the horizon. Each expert e runs the projected-SGD procedure of Proposition B.1 on its training buffers, with timestamp-only balanced batches, the same fixed reset point $x_\circ$, and solve starts $s_j = s_0 + j\Delta$; its variance inputs use its actual training-batch sizes. Write x_j^e for its output at $t_j = s_j + c$. Experts may share training observations with one another and with their own later solves, but no expert uses another expert's iterates or the aggregation weights.

Causal validation and aggregation. On the report interval $\mathcal{I}_j = [t_j, t_{j+1})$ the server reports $x_j = \sum_{e=1}^E \pi_{j,e} x_j^e$. It maintains two exponential-weight instances, one for even and one for odd j, each initialized with uniform weights, with learning rate and weights

$$\eta_{\mathrm{ew}} = \frac{2}{B}\sqrt{\frac{\log E}{R}}, \qquad \pi_{j,e} = \frac{\exp\{-\eta_{\mathrm{ew}} \sum_{0\le i\le j-2:\ i\equiv j \pmod 2} \widehat{\ell}_{i,e}\}}{\sum_{v=1}^E \exp\{-\eta_{\mathrm{ew}} \sum_{0\le i\le j-2:\ i\equiv j \pmod 2} \widehat{\ell}_{i,v}\}},$$

where empty sums are zero. The score $\widehat{\ell}_{j,e}$ uses the validation arrivals of client m in $(t_j, t_{j+1}]$; call this set $\mathcal{V}_{m,j}$ and its size $n_{m,j}^{\mathrm{val}}$. Then

$$\widehat{\ell}_{j,e} = \sum_m \frac{w_m}{n_{m,j}^{\mathrm{val}}} \sum_{(u,\xi)\in\mathcal{V}_{m,j}} \big[\ell(x_j^e;\xi) - \ell(x_\circ;\xi)\big]. \tag{61}$$

All these counts are positive by Lemma H.1. The interval $(t_j, t_{j+1}]$ has ended by the start of solve $j+2$, since $t_{j+1} = s_j + \Delta + c \le s_j + 2\Delta$. The server includes the points $(x_j^e)_{e=1}^E$ in the first broadcast of that solve, and the clients return the corresponding validation sums and counts with their first gradient messages, so $\widehat{\ell}_{j,e}$ is available before the report at t_{j+2}. No validation sample later than the current report is used. Because message size and local computation are uncharged, all experts' gradient messages and the validation statistics can be bundled into the same K rounds. The report is $x_\circ$ before t_0.

The construction uses the standard exponential-weight principle and the standard reduction of delayed feedback to parallel forecasters (Freund & Schapire, 1997; Joulani et al., 2013). The result below accounts for the independent validation stream, its nonuniform source times, and the finite-sample cost of the split.

Lemma H.1 (Counts after alternating assignment)**.** *On a λ-regular schedule, either assigned stream of client m has at least $\lambda_m(v-u)/6$ observations in $(u, v]$ whenever $v - u \ge 6/\lambda_m$.*

Proof. If the full interval contains n observations, either parity contains at least $(n-1)/2$ of them, and regularity gives

$$\frac{n-1}{2} \ge \frac{\lambda_m(v-u)}{4} - \frac{1}{2} \ge \frac{\lambda_m(v-u)}{6}.$$

Only this lower count inequality is needed for the upper analysis; we do not assert that either thinned schedule is regular with rate $\lambda_m/3$ in the two-sided sense of Definition 3.1. □

Write $\mathcal{R}_T^e$ for the expected average current-objective gap of expert e's whole reported trajectory, including initialization, and $\mathcal{R}_T^{\mathrm{mix}}$ for the same risk of the aggregate.

Theorem H.2 (Unknown-drift adaptation over joint memories)**.** *Under Section 3, except that γ need not be known to the learner, and under (59)–(60), the causal procedure above satisfies*

$$\mathcal{R}_T^{\mathrm{mix}} \le \min_{1\le e\le E} \mathcal{R}_T^e + 2B\sqrt{\frac{\Delta \log E}{T}} + D\bar{\gamma}\Delta. \tag{62}$$

In particular, with $\widetilde{b}_m = 3b_m$, $\widetilde{\Xi}^2 = \sum_m \widetilde{b}_m$, and

$$\widetilde{\mathcal{T}}_T = \beta_{\mathrm{sgd}} D\widetilde{\Xi}\Big(\frac{2}{\sqrt{T}} + \frac{\Delta}{T\sqrt{s_0}}\Big) + \frac{GD(s_0+c)}{T},$$

we have

$$\mathcal{R}_T^{\mathrm{mix}} \le \frac{LD^2}{K} + D\min_{h\in\mathcal{G}}\Big\{2\sum_m a_m h_m + \beta_{\mathrm{sgd}}\sqrt{\sum_m \widetilde{b}_m/h_m}\Big\}$$
$$+ D\bar{\gamma}(c+2\Delta) + \widetilde{\mathcal{T}}_T + 2B\sqrt{\frac{\Delta\log E}{T}}. \tag{63}$$

The algorithm uses the noise and rate bounds but not the values $a_m = w_m\gamma_m$ that appear in the oracle bound.

The selection inequality (62) requires only that the experts' reports lie in $\mathcal{X}$ and are causal functions of the training array, independent of the validation observations; it therefore also applies to fixed-memory empirical-risk minimizers. The explicit gradient-based bound (63) additionally requires the projected-SGD experts specified above.

Proof. We first prove the selection inequality, which needs no gradient concentration event. Condition on the full arrival schedule and on the entire training array. Every expert point is then fixed, the validation observations remain mutually independent, and at report j the weight vector π_j depends only on previously used validation intervals and is independent of the observations in $(t_j, t_{j+1}]$.

Population losses are centered at $x_\circ$ before integrating over time; this removes arbitrary loss offsets, which the assumptions do not bound. Define the centered average population objective on a reporting interval and the centered population objective at the validation source times by

$$\overline{F}_j(x) = \frac{1}{\Delta}\int_{t_j}^{t_{j+1}} [F^t(x) - F^t(x_\circ)]\,dt, \qquad F_j^{\mathrm{val}}(x) = \sum_m \frac{w_m}{n_{m,j}^{\mathrm{val}}} \sum_{(u,\xi)\in\mathcal{V}_{m,j}} [f_m^u(x) - f_m^u(x_\circ)];$$

the latter uses only the timestamps u. Conditionally on the stated information and on past validation feedback,

$$\mathbb{E}[\widehat{\ell}_{j,e}] = F_j^{\mathrm{val}}(x_j^e), \qquad |\widehat{\ell}_{j,e}| \le B. \tag{64}$$

This conditional expectation evaluates finitely many training-measurable points on a separate sample array; it asserts no uniform concentration over queries.

Consider either parity of reporting intervals. The exponential-weight potential and Hoeffding's lemma for losses in $[-B, B]$ give, for any expert e and n intervals of that parity,

$$\sum_j \Big(\sum_v \pi_{j,v} \widehat{\ell}_{j,v} - \widehat{\ell}_{j,e} \Big) \le \frac{\log E}{\eta_{\mathrm{ew}}} + \frac{\eta_{\mathrm{ew}} B^2 n}{2} :$$

the one-step upper bound for the log potential ratio is $-\eta_{\mathrm{ew}} \sum_v \pi_{j,v} \widehat{\ell}_{j,v} + \eta_{\mathrm{ew}}^2 B^2/2$, and the terminal weight of expert e gives the matching lower bound $-\eta_{\mathrm{ew}} \sum_j \widehat{\ell}_{j,e} - \log E$. All earlier losses of a parity are available before its next report. Adding the inequalities of the two parities (one of which may be empty) and substituting η_{ew} yields the pathwise bound

$$\sum_{j=0}^{R-1} \Big(\sum_v \pi_{j,v} \widehat{\ell}_{j,v} - \widehat{\ell}_{j,e} \Big) \le 2B\sqrt{R \log E}. \tag{65}$$

Source-time drift gives $\|\nabla F_j^{\mathrm{val}}(x) - \nabla \overline{F}_j(x)\| \le \bar{\gamma}\Delta$ uniformly on $\mathcal{X}$, because every source time and every integration time lie in an interval of length Δ. Consequently, for all $x, y \in \mathcal{X}$,

$$\big| [F_j^{\mathrm{val}}(x) - F_j^{\mathrm{val}}(y)] - [\overline{F}_j(x) - \overline{F}_j(y)] \big| \le D\bar{\gamma}\Delta,$$

and convexity of each F^t, hence of $\overline{F}_j$, gives

$$\overline{F}_j(x_j) - \overline{F}_j(x_j^e) \le \sum_v \pi_{j,v} \big[F_j^{\mathrm{val}}(x_j^v) - F_j^{\mathrm{val}}(x_j^e) \big] + D\bar{\gamma}\Delta.$$

Take expectations, use (64) and (65), and multiply by Δ/T. The current minimum cancels in the risk difference, and both procedures report $x_\circ$ before t_0. We obtain

$$\mathcal{R}_T^{\mathrm{mix}} - \mathcal{R}_T^e \le \frac{2B\Delta\sqrt{R \log E}}{T} + \frac{R\Delta}{T} D\bar{\gamma}\Delta \le 2B\sqrt{\frac{\Delta \log E}{T}} + D\bar{\gamma}\Delta.$$

This holds for every deterministic index e, so taking the minimum of the expected expert risks proves (62).

Finally, Lemma H.1 supplies the only arrival-count input used in Proposition B.1, with lower rate $\lambda_m/3$. Frozen training intervals have length $\min\{s_j, h_m^e\} \ge 6K/\lambda_m$ and contain at least K observations. Each expert retains the original reset and disjoint-batch filtration, and shared observations between experts impose no independence requirement on their marginal bounds. The proof of that proposition therefore applies with b_m replaced by $3b_m$ and with the initialization time of (60). Substituting into (62) proves (63). □

A geometric joint grid. Let $\underline{h}_m = 6K/\lambda_m$ and take

$$\mathcal{G}_m = \big\{ \min(2^i \underline{h}_m, T) : i = 0, \dots, \lceil \log_2(T/\underline{h}_m) \rceil \big\}, \qquad \mathcal{G} = \prod_m \mathcal{G}_m.$$

For every $h_m \in [\underline{h}_m, T]$, rounding upward to this grid gives $h_m \le h'_m \le 2h_m$, so the minimum in (63) is at most twice the continuous minimum of the same bias–pooled-variance expression on the box. Moreover $\log E = \sum_m \log |\mathcal{G}_m|$: the selection cost depends on this logarithm, whereas maintaining all experts explicitly can require exponentially many gradient evaluations in M. If the grid has a single point, that expert is used and the selection penalty is omitted.

For stationary clients the grid contains the full-prefix candidate, whose statistical contribution vanishes with T. The guarantee is adaptation to a fixed joint memory vector over the stated horizon, not to a sequence of locally optimal windows, and the implementation must store the union of the candidate buffers, potentially the full training history: it gives no storage or computation saving.

Other horizons and scope. For an unaligned horizon $T > s_0 + c + \Delta$, let $R = \lfloor (T - s_0 - c)/\Delta \rfloor$, use the same learning rate with this R, and keep the last mixture through the remaining fraction of an interval. The same argument on the R complete intervals gives (62)–(63) with one additional term $GD\Delta/T$, since the unmatched part contributes at most $GD\Delta$ to the cumulative risk difference; the expert bound is applied at the actual horizon T.

Unknown-drift window selection is available under uniform function-class convergence assumptions (Mazzetto & Upfal, 2023a), and federated drift adaptation is studied by Ganguly & Aggarwal (2024). The extension here is the finite joint-memory oracle inequality with independent validation, asynchronous source-time correction, and a causal communication schedule. It does not replace those results, and it does not establish efficient unrestricted adaptation under the original pointwise moment assumptions alone.

I CLINICS WITH DIFFERENT SPECIALTIES: WHEN FORGETTING PAYS

The logistic study of Section 7 marks a boundary: along smooth population paths on which both the covariates and the label mechanism change, the best common memory is the full history, and client-specific windows bring no measurable benefit. That result says what the theory does not promise. This appendix asks the complementary question. What does a network of data sources look like in which forgetting does pay, and does the memory rule of Corollary 4.2 capture that benefit when the data are realistic and the solver is Algorithm 1?

The answer we construct is a regional network of twelve clinics. Its acute services see many patients and follow epidemic waves; its specialized clinics see few patients from stable populations; and in between, cardiology drifts slowly as its population ages. The shared risk model is the same for all of them. With the solver held fixed, client-specific windows lower the time-averaged tracking error by 33% relative to the best single memory, and they attain almost nine tenths of what an oracle obtains by tuning every clinic's window separately.

I.1 THE NETWORK

Table 3 lists the clinics. There are six specialties, each with an urban and a rural site. All clinics estimate one quantity: the probability that a patient seen at a visit has an emergency or inpatient encounter within the following 90 days. They differ in three respects that matter for memory.

The *acute services*, pediatrics, the emergency department and primary care, see between 25 and 120 patients a week. Their case mix changes with the season and, for the emergency department, with the waves of COVID-19. The drift is fast, but so is the flow of new data, so a short memory is affordable. The *specialized clinics*, endocrinology and oncology, see between one and a half and eight patients a week, and the populations they serve do not change over the three years we follow. Every discarded record costs them precision and buys nothing. *Cardiology* is the intermediate case: 15 to 40 patients a week and a population whose risk profile shifts slowly as it ages, with winter peaks on top. A single memory length has to choose between these situations; a client-specific memory does not.

Table 3: The twelve clinics. The surge cohort is defined by a variable the model does not observe. Urban and rural sites follow surveillance curves from different regions. All clinics have weight $w_m = 1/12$.

Specialty	Patients seen	Surge cohort	Shape of the drift	Visits per week (urban / rural)
Pediatrics	age below 18	recent acute care	ILINet, youngest age group	60 / 25
Emergency	emergency encounters	recent acute care	COVID-19 hospital admissions	120 / 40
Primary care	adults at wellness, ambulatory and outpatient visits	recent acute care	weighted ILINet activity	80 / 30
Cardiology	age 40 or older with hypertension, heart failure or coronary disease	recent acute care	slow ageing trend with winter peaks	40 / 15
Endocrinology	diabetes or prediabetes	none	stationary	8 / 3
Oncology	active cancer diagnosis	none	stationary	3 / 1.5

I.2 PATIENTS, OUTCOME, AND WHAT THE MODEL DOES NOT SEE

The patients are synthetic electronic health records generated with Synthea (Walonoski et al., 2018). Every ambulatory, wellness, outpatient, urgent-care or emergency visit in the last eight years of the simulated period, with at least 90 days of follow-up, is a candidate index visit; at most four are kept per patient. The outcome of a visit is an emergency or inpatient encounter within the next 90 days. The features are those available when the patient is seen: age, sex, body-mass index, the latest systolic and diastolic blood pressure within three years, indicators of hypertension, diabetes or prediabetes,

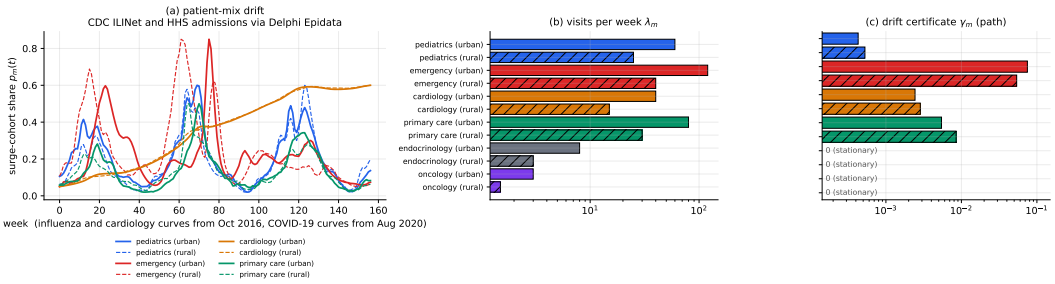

Figure 3: The network. (a) Share of the surge cohort in each drifting clinic; solid lines are urban sites, dashed lines rural ones. (b) Visits per week. (c) Drift bound γ_m along the optimal trajectory; the four specialized clinics are stationary. Drift speeds span two orders of magnitude and are not ordered like the visit rates, which is what makes a single memory length a compromise.

asthma, chronic obstructive pulmonary disease, heart failure, coronary disease, cancer and chronic kidney disease active at the visit, and indicators of missing measurements. A fifth of the patients is set aside once and used only to fit a public starting model; none of them appears in any clinic's data.

The design turns on one choice. Drift can help a forgetting rule only if it moves the *optimum*. If the patients who become more frequent during a wave were distinguished by the features, the change would be a covariate shift, and a logistic model that already accounts for those features would barely need to move. We therefore define the surge cohort by a variable the model never receives: an emergency or inpatient encounter in the 90 days *before* the visit. In the generated population, patients with such a recent episode are several times as likely to return to acute care in cardiology and in the emergency department, less so in primary care, and only mildly so among children. When a respiratory wave brings more of them to a clinic, the same observed features carry a higher risk, and the optimal parameter follows the wave. This is concept drift as the model experiences it, and its speed differs from clinic to clinic.

For every specialty the eligible visits of the remaining patients form two finite cohort banks, a baseline bank and a surge bank, each holding up to 3000 records; the stationary specialties have a single bank. A clinic's population at time t is the mixture

$$(1 - p_m(t))\,\mathrm{baseline}_m + p_m(t)\,\mathrm{surge}_m, \tag{66}$$

and each arrival draws one record from the corresponding bank. Because the banks are finite, the current objective, its minimizer, and the drift and noise of every clinic can be computed exactly at any time.

I.3 How the patient mix changes

The shares $p_m(t)$ follow public surveillance data retrieved from the Delphi Epidata service (Reinhart et al., 2021) (Figure 3a). Pediatrics follows the share of influenza-like-illness visits among the youngest children in the CDC ILINet network (Centers for Disease Control and Prevention, 2024), nationally for the urban site and for HHS Region 6 for the rural one. Primary care follows weighted ILINet activity in HHS Regions 1 and 9. The emergency department follows weekly COVID-19 hospital admissions reported to HHS (U.S. Department of Health and Human Services, 2021) in New York and in Montana, whose waves arrived at different times: the rural site peaks in autumn 2020 and during the Delta wave, the urban one in winter 2020–21 and during the Omicron wave. Cardiology combines an S-shaped ageing trend with winter peaks whose timing is taken from the regional ILINet curve, as cardiovascular admissions rise in winter. Influenza and cardiology curves start in October 2016 and COVID-19 curves in August 2020; each is smoothed over three weeks and rescaled to the range of shares given for its specialty. Time is measured in weeks, the horizon is $T = 156$, and patients arrive at the regular times n/λ_m of Definition 3.1.

I.4 Model, objective, and exact quantities

The model is a logistic regression with coefficients shared by all clinics and one intercept per clinic, the centre effect of multicentre studies. With z the observed features of a record and e_m the indicator

of clinic m, the parameter θ enters through $\theta^\top(z,\, 0.8\, e_m)$, and the features are scaled so that the whole input has norm at most one. The tracked objective is

$$F^t(\theta) = \sum_{m=1}^{12} w_m \big[(1 - p_m(t))\, L_m^{\mathrm{base}}(\theta) + p_m(t)\, L_m^{\mathrm{surge}}(\theta)\big] + \frac{\rho}{2}\|\theta\|^2, \qquad \rho = 0.005, \tag{67}$$

where L_m^{base} and L_m^{surge} are the mean logistic losses of the two banks of clinic m. Every minimizer lies in the ball of radius $1/\rho = 200$, which serves as the domain. The clinic intercept localizes a clinic's drift: when the surge share of the emergency department rises, the optimum moves mainly in that clinic's intercept, while the shared coefficients adjust only as far as the other clinics allow.

Since (67) is a finite sum, the minimizer $\theta_t^\star$ and the value $\min F^t$ are computed exactly on a grid of four points per week, and the gap of any reported parameter is evaluated without Monte Carlo error. The memory rule needs the quantities of Theorem 4.1: the drift bound γ_m and the gradient noise σ_m^2 of every clinic. Bounds that hold uniformly over the parameter ball are valid but loose by one to two orders of magnitude, and unevenly so across clinics. We therefore evaluate both exactly along the optimal trajectory,

$$\gamma_m = \max_t\, |p_m'(t)|\, \big\|\nabla L_m^{\mathrm{surge}}(\theta_t^\star) - \nabla L_m^{\mathrm{base}}(\theta_t^\star)\big\|, \qquad \sigma_m^2 = \max_t\, \mathrm{Var}_m^t\big[\nabla \ell(\theta_t^\star)\big], \tag{68}$$

where the variance is that of a single-record gradient under the clinic's mixture at time t. These are the drift and noise the algorithm actually meets near the optimum, and they are the inputs of $a_m = w_m\gamma_m$, $b_m = w_m^2\sigma_m^2/\lambda_m$ and of the windows of Corollary 4.2. Figure 3c shows the resulting drift bounds: the emergency department drifts fastest, followed by primary care and cardiology, while the hidden cohort moves children's risk only slightly.

I.5 Memories, solvers, and evaluation

Five memories are compared. *One common window* gives every clinic the same memory length, the comparator of the theory. *Common window with a data floor* uses one length but never lets a clinic's window hold fewer than 100 of its records, which protects the logistic fit of small clinics from buffers with no events; it is a stronger practical baseline. *Full history* keeps everything. *Client-specific windows* use the shape $h_m \propto \sqrt{b_m/a_m}$ of Corollary 4.2, with the full history for stationary clinics and the same data floor. The *per-clinic oracle* starts from the client-specific windows and adjusts each clinic's window separately by a coordinate search; it is an upper benchmark, not a method.

Each memory is combined with two solvers. Exact minimization of the buffered empirical risk isolates the quality of the memory. Algorithm 1 is the method of the paper: $K = 8$ accelerated rounds on disjoint batches, restarted at every solve from a public point, with solves every week and a round duration of $1/16$ week. Its public point is the model fitted once on the held-out pilot patients. Every memory has one free scale, selected on six tuning streams; the evaluation uses twenty further independent streams, on which all memories receive the same patients at the same times. Because the objective is exact, the scale is selected by the true tracking error; this compares memories at their best scale and is not a data-driven procedure for choosing windows. The score is the time-averaged gap $F^t(x_t) - \min F^t$ after the first eleven weeks, during which every memory reports the same public model.

I.6 Results

Figure 4 gives the headline. With exact minimization, client-specific windows reduce the tracking error from $3.36 \cdot 10^{-3}$ to $2.26 \cdot 10^{-3}$, by 33% relative to one common window; the paired difference over the twenty evaluation streams is $-1.10 \cdot 10^{-3}$ with a 95% half-width of $8.7 \cdot 10^{-5}$. The best common window is the full history, so the same 33% is the gain over keeping everything. Against the common window with a data floor the reduction is 26%. The per-clinic oracle reaches $2.13 \cdot 10^{-3}$; the rule of Corollary 4.2 thus obtains 89% of the improvement that separate tuning of all twelve windows achieves.

Figure 5 shows why. Tuned as one number, the common memory settles on the whole history: most of the network is stable or slow, and any shorter single window would add noise to all of it. Client-specific windows keep the whole history for the four specialized clinics and forget quickly

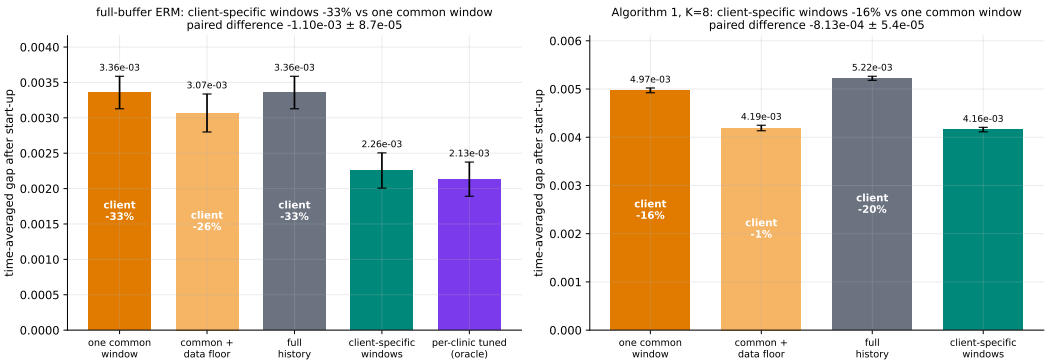

Figure 4: Time-averaged gap after start-up, mean over twenty evaluation streams with 95% intervals. Left: exact minimization of the buffered risk. Right: Algorithm 1 with $K = 8$. The label in each baseline bar is the change obtained by client-specific windows relative to that baseline.

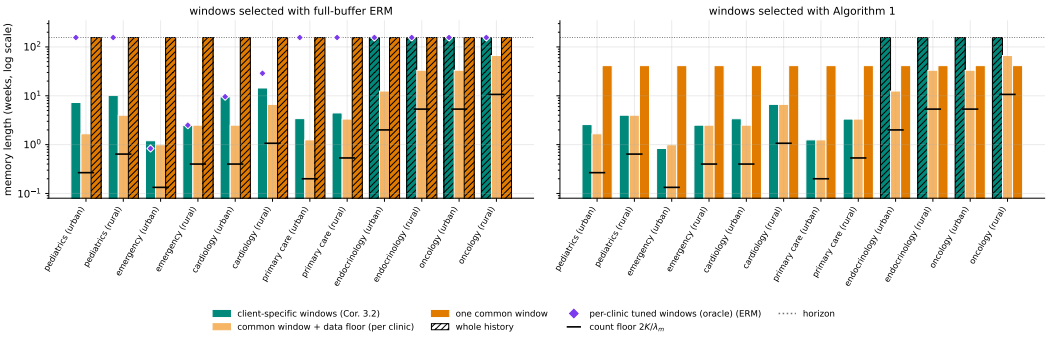

Figure 5: Memory lengths selected for each clinic. Hatched bars mean that the whole history is kept. Diamonds mark the per-clinic oracle, dashes the minimum length $2K/\lambda_m$ needed to form K batches.

where the optimum moves fast and data are plentiful: about one week for the urban emergency department, two for the rural one, three to four weeks for primary care, and between seven and fourteen weeks for pediatrics and cardiology. The oracle agrees on the emergency departments and on urban cardiology. It parts from the prescription on pediatrics and primary care, where it returns to the full history, and on rural cardiology, where it chooses a window twice as long; we return to this below.

The time course in Figure 6 shows where the error is made. The full history is worst during the waves that reach the emergency departments, around weeks 20 and 60–80, when its error rises to about 10^{-2}, and again in the last year, when cardiology keeps ageing and every old record drifts further from the present. Client-specific windows stay close to the oracle throughout and remain near $1.3 \cdot 10^{-3}$ from week 90 on. The full history is better only for short spells between waves, when the current mix happens to pass through its long-run average. The common window with a data floor tracks the waves, but pays a constant excess in the calm periods, because its short common length also shortens the memory of clinics that did not need it.

Figure 7 splits the gap into the contributions $w_m[f_m^t(x_t) - f_m^t(\theta_t^\star)]$ of the individual clinics. Almost all of the gain comes from the two emergency departments, whose contributions fall from about 10^{-3} and $1.6 \cdot 10^{-3}$ to about $3.6 \cdot 10^{-4}$ and $6 \cdot 10^{-4}$, a reduction near two thirds. Several contributions are negative. The joint optimum $\theta_t^\star$ minimizes the sum over clinics, not each clinic's loss, and through the shared coefficients it bends towards the emergency departments during their waves; a memory that ignores those waves therefore fits pediatrics, cardiology and primary care slightly better than the optimum does, at a larger cost elsewhere. The data-floor baseline shows the other side of the trade-off: it helps the emergency departments as much as the client-specific rule, but its short windows make the other drifting clinics positive.

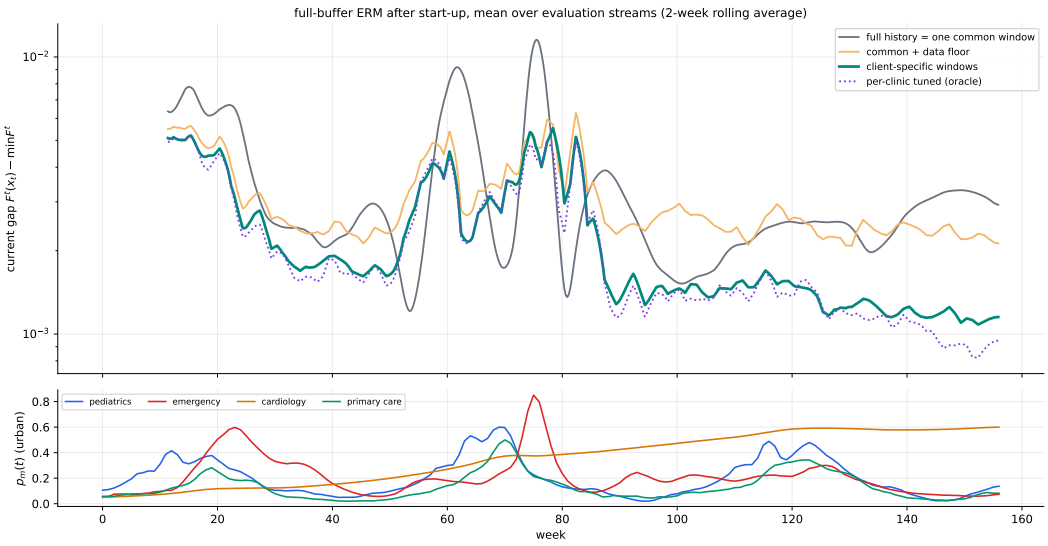

Figure 6: Current gap over time under exact minimization, after start-up; two-week rolling mean over the evaluation streams. The lower panel repeats the surge shares of the urban clinics. The full history and the one common window coincide because the common window selects the whole history.

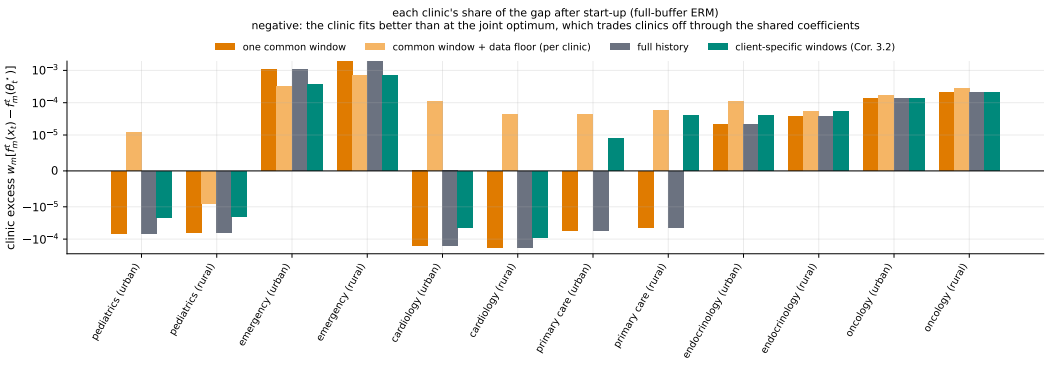

Figure 7: Contribution of each clinic to the gap after start-up under exact minimization, on a symmetric logarithmic scale. Negative values mean that the clinic fits better than at the joint optimum.

With Algorithm 1 every error is larger, between $4.2 \cdot 10^{-3}$ and $5.2 \cdot 10^{-3}$, because eight accelerated rounds do not reach the minimizer of the buffered risk. The ordering remains: client-specific windows reduce the gap by 16% relative to the best common window, which now keeps about forty weeks, and by 20% relative to the full history; the paired difference is $-8.1 \cdot 10^{-4}$ with a 95% half-width of $5.4 \cdot 10^{-5}$. Against the common window with a data floor the two memories are within one percent of each other.

I.7 Reading the results through the theory

Theorem 4.1 charges a memory h_m through two terms per clinic, a drift cost growing like $a_m h_m$ and a statistical cost falling like $\sqrt{b_m/h_m}$. Their balance is struck at $h_m \propto \sqrt{b_m/a_m}$, and it moves in opposite directions across this network: b_m is largest in the small specialized clinics, a_m largest in the emergency departments. One common length has to serve both ends. Here it is pushed to the full history, because the stable clinics and the slowly drifting ones form most of the network, and the emergency departments then carry the drift cost of every wave. This is precisely the situation in which the statistical-drift scale $\mathcal{H}_\gamma$ of the client-specific rule falls well below the common-memory scale, since the drift–noise masses $\sqrt{a_m b_m}$ are spread over very different ratios b_m/a_m.

The comparison with the data-floor baseline isolates what the rule adds. The floor $100/\lambda_m$ already makes the memory depend on the visit rate, which is one half of the prescription. It does not depend on drift. Under exact minimization this is decisive: the floored common window shortens the memory of pediatrics, cardiology and primary care to the same few weeks as the emergency departments and pays for it in noise, whereas Corollary 4.2 lengthens their windows in proportion to $\gamma_m^{-1/2}$. The remaining 26% is the value of accounting for drift, clinic by clinic.

Where the prescription and the oracle part, the reason is the shape of the drift. The theorem is built for a worst-case drift that can keep moving in one direction, so the bias of an old record grows without bound. Seasonal drift returns to its average. For pediatrics and primary care, whose shares oscillate around a stable level, the bias of a long memory saturates at the amplitude of the season, and with a moderate flow of patients the full history is the better compromise. Cardiology, whose ageing trend does not revert, is the opposite case, and there the oracle keeps a finite window as the theory prescribes. The prescription is thus conservative exactly where the drift is mean-reverting, and it still captures most of the attainable gain.

The solver changes the size of the effect but not its sign. Proposition B.1 makes precise that the memory scale does not depend on acceleration; what the solver contributes is the optimization remainder, which with $K = 8$ is comparable to the statistical term and dilutes every difference between memories. Under Algorithm 1 the two memories that shorten the emergency departments' windows, the data-floor baseline and the client-specific rule, therefore end up close, while both remain clearly better than any memory that treats the network as one.

Section 7 and this experiment differ in how the optimum moves. There the population changes smoothly in both covariates and labels, and the best common memory is the full history. Here the optimum is moved by a cohort the model cannot see, at speeds that differ by two orders of magnitude from clinic to clinic. Heterogeneous drift bounds justify different memories in both settings; only in the second does the difference between memories carry a large price, and the client-specific rule then pays off by a third of the error. Client-specific memory also holds far fewer records: the emergency departments keep a week or two of patients instead of three years, so every solve touches a small fraction of the data that the full history requires.

I.8 Scope

The records are synthetic, and each clinic's population is a mixture of two fixed cohorts whose shares follow real surveillance curves; the experiment tests the memory mechanism of the paper, not a clinical prediction rule. The drift and noise bounds are evaluated along the exact optimal trajectory, which is possible only because the objective is known; in deployment they would have to be certified or replaced by the validation-based selection of Section 6. The single free scale of each memory is chosen with the true tracking error, for every memory alike. Algorithm 1 is used unchanged, with a public starting point fitted once on patients who never enter the streams.

J Station temperature records

This appendix runs Algorithm 1 on laws fitted to measured temperature records; Figure 1 in Section 7 summarizes it. Two experiments are reported. The first is inside the model of Section 3: day-to-day weather is independent noise around a fitted diurnal clock, the drift and noise bounds are attained by construction, and Corollary 4.2 is scored against a common window. The second replays held-out months in their recorded order, where weather persists, and asks whether the same corollary improves on a common length once its drift input matches that path.

Setup. Unless stated otherwise the runs use the scalar loss $\ell(x;\xi) = (L/2)x^2 + \xi x$ with $L = 10^{-3}$ on $[-1/2, 1/2]$, so $D = 1$, where ξ is a scaled reading, and the raw-buffer implementation of Section 4 with $\tau = 0.02$ hours. Weights are $1/3$. With $\mu_m(t)$ the path of station m, the population gradient $Lx + \theta(t)$, $\theta(t) = \sum_m w_m \mu_m(t)$, is the network path up to a negligible curvature term, and the minimizer lies at an endpoint of $[-1/2, 1/2]$ whenever $|\theta(t)| > L/2$; the tables call this loss "endpoint". The curvature is kept small so that the optimization remainder is negligible: the number of rounds is $K = \lceil \max\{1, 2\sqrt{LD/\mathcal{H}_\gamma}\} \rceil = 1$, the stride is $\Delta = K\tau$, and $s_0 = 2K/\lambda_{\min}$,

the choices of Corollary B.3. With $K = 1$ the accelerated solve and averaged SGD coincide, so comparisons are between memories.

Client-specific windows are those of Corollary 4.2, without a multiplier (labelled "formula" in the plots). The bound-optimal common length minimizes the common-memory bound of Appendix K; the tuned common length is chosen on seeds 0–3 from a fixed grid of common lengths. Seeds 4–15 are evaluation, and within a seed every policy sees the same draws. The predicted gain is $g = 1 - \mathcal{H}_\gamma/(\bar{\gamma}^{1/3}\Xi^{2/3})$, the relative gap in the comparison $\mathcal{H}_\gamma \le \bar{\gamma}^{1/3}\Xi^{2/3}$ of Section 3, computed before count floors are imposed. The score inside the model is the population gap AvgErr_T. On time-ordered months it is the prequential excess over the zero report: the average, over the recorded readings, of the loss of the current report on each reading as it arrives, minus the loss of the report $x = 0$ on the same reading; lower is better.

J.1 Experiment I: inside the model

The U.S. Climate Reference Network (Diamond et al., 2013) reports air temperature every five minutes. We take the ten Decembers 2010–2019 at three stations chosen by insolation before any simulation: Stovepipe Wells (desert), Sitka (maritime), and Utqiagvik, where the sun does not rise in December (polar night). The population path of a station is the truncated Fourier series of its average day; leave-one-December-out cross-validation selects 2, 3, and 0 harmonics, so the polar-night path is constant. Paths are centred, and paths and residuals are divided by one common scale, $4.56°\mathrm{C}$; the bounds below are quoted in degrees, before this scaling, and the windows do not depend on it. The drift bounds are the largest slopes of the fitted paths, $(1.58, 0.354, 0)°\mathrm{C}$ per hour. A draw at a slot is a reading of a uniformly chosen December day at that slot, minus the slot mean, plus the fitted path. The noise scales $(3.41, 3.72, 6.64)°\mathrm{C}$ are the largest slot standard deviations. Each station reports $\lambda = 12$ draws per hour.

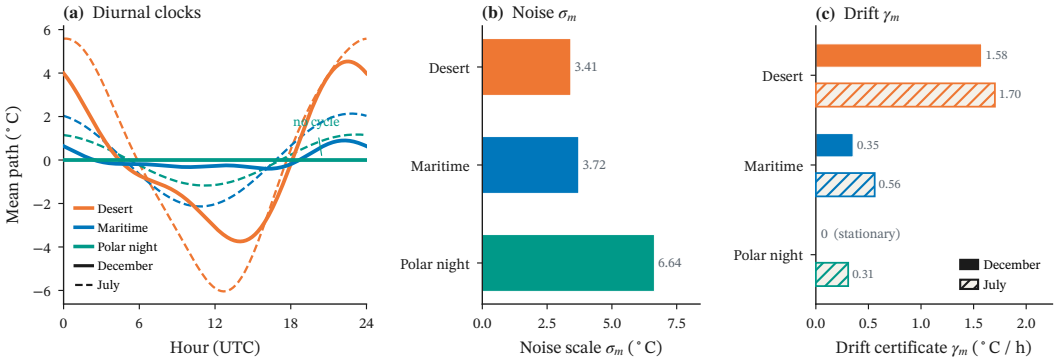

Figure 8: The two station laws. (a) Mean paths in UTC, December solid and July dashed; polar night has no December cycle. (b) Noise scales σ_m. (c) Drift bounds γ_m; polar night is stationary in December and moving in July.

On this law $\mathcal{H}_\gamma = 0.111$. The prescribed windows are $(0.556, 1.28, \infty)$ hours. The bound-optimal common length is 1.01 hours, the tuning seeds select 0.972 hours, and $g = 0.32$. The count floor $2K/\lambda = 1/6$ hour does not bind, $K\mathcal{C}_\gamma \le \mathcal{H}_\gamma$, and $T = 432$ hours exceeds both joint-regime thresholds, so every hypothesis of Corollary B.3 used by the upper guarantee holds.

The client-specific mean is 0.0120. Against the full history the paired difference is -0.0835 with interval $[-0.0843, -0.0827]$, an 87% reduction. Against the bound-optimal common length it is -0.0125 with interval $[-0.0139, -0.0112]$, a 51% reduction, and against the tuned common length it is -0.0128 with interval $[-0.0144, -0.0113]$, a 52% reduction. The client-specific windows are below both common lengths on all twelve seeds. Per-station tuning, shown for reference in Figure 9, reaches 0.0097, so the untuned windows recover about 84% of the improvement over the bound-optimal common length that separate tuning attains. A single length cannot give the desert a half-hour window and polar night its full history at once.

The same three stations in July have harmonics $(3, 1, 1)$: the sun now rises at all three. Then $g = 0.066$ and the windows are $(0.472, 0.645, 1.40)$ hours. The client-specific mean is 0.0128, a

6% reduction against the bound-optimal common length, close to the predicted g but within noise (paired difference $-7.7 \cdot 10^{-4} \pm 8.2 \cdot 10^{-4}$), and an 11% reduction against the tuned common length. The Beijing July control groups twelve districts into three shared-clock clients, $g = 0.0002$, and the client-specific windows tie the common length.

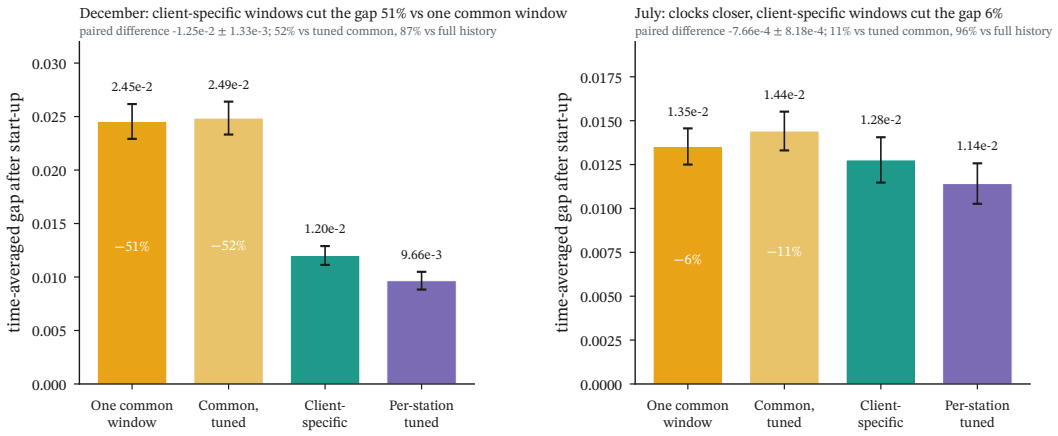

Figure 9: Mean AvgErr_T over seeds 4–15 with Student 95% intervals. In December client-specific windows cut the gap by 51% against the bound-optimal common window; in July, whose clocks are closer, by 6%, within noise. The subtitle of each panel records the paired difference and the reductions against the tuned common length and the full history.

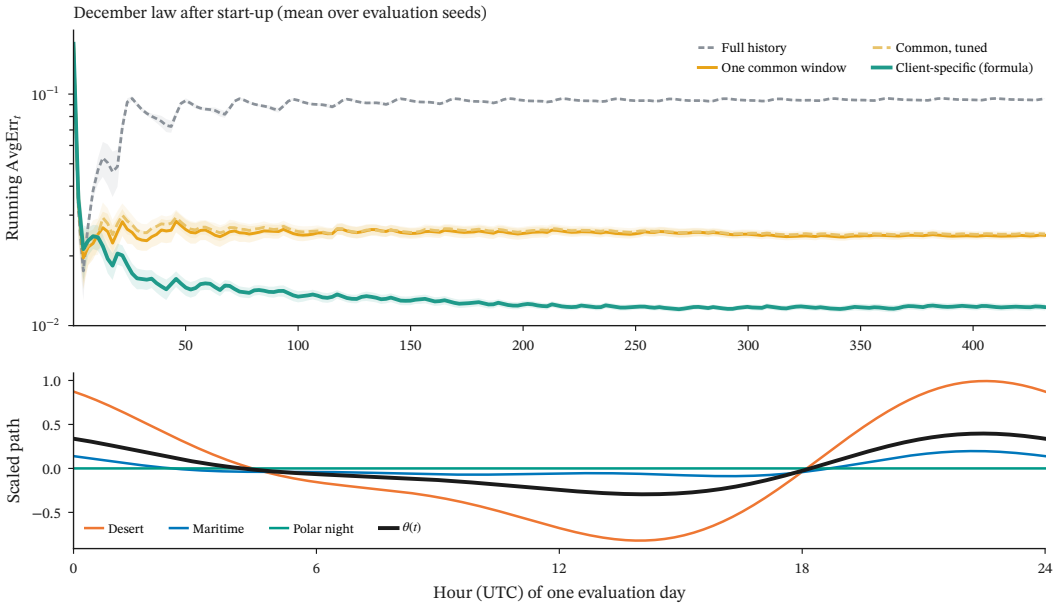

Figure 10: Top: running AvgErr_t on the December law, mean and one standard error over seeds 4–15. Client-specific windows separate from a common length within a day and stay below it. Bottom: the three station means and the network path $\theta(t) = \sum_m w_m \mu_m(t)$ on one evaluation day.

Over 200 distinct random triplets from the 134 network stations with enough December coverage, the measured relative gain of client-specific windows over the bound-optimal common length grows with the predicted g (Spearman 0.73, bootstrap interval $[0.64, 0.80]$) and, for $g \geq 5\%$, exceeds it on average: the comparison of upper bounds is a conservative predictor. For the 31 triplets with $g \geq 5\%$ the client-specific windows are below the bound-optimal common length in all and below the tuned common length in 29. When the clocks agree, g is small and a tuned common length can beat the conservative drift input by a few percent.

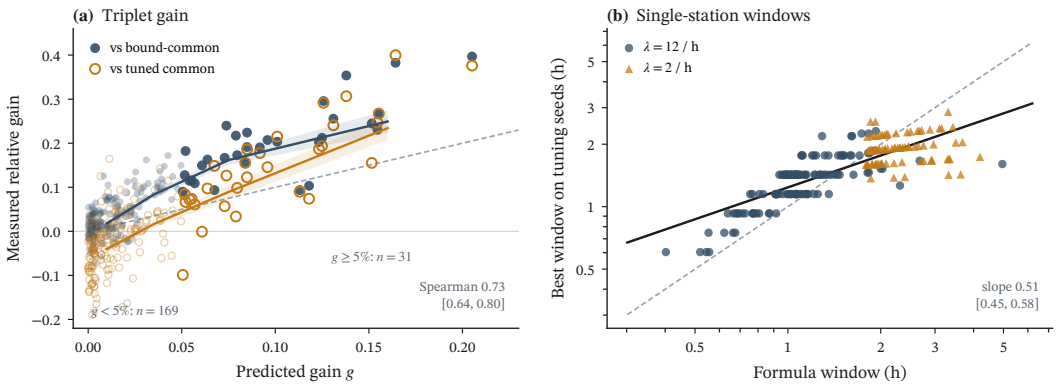

Figure 11: Left: measured relative gain of client-specific windows over the bound-optimal and the tuned common length for 200 random station triplets, against the predicted gain g. Opaque points have $g \geq 5\%$; curves are binned means with one standard error; the dashed line is equality. Right: best window on four tuning seeds against the prescribed window for single stations at two reading rates; the log–log slope is 0.51, not 1, consistent with the prescription charging the steepest slope rather than the typical one.

J.2 Experiment II: time-ordered records

The resampled law treats day-to-day weather as independent noise. To see what that choice hides, each December is held out in turn; the law is fitted on the other nine years; and the held-out month is replayed in its recorded order for $T = 744$ hours. A common length is tuned on the nine training months replayed the same way. The climatological windows, which treat weather as noise, lose to that tuned length in all ten years: mean -0.167 against -0.238, paired difference 0.071 with interval $[0.037, 0.104]$. They keep the full history at the polar site, whereas exponentially weighted memories tuned per station on the training years keep less than an hour there: in recorded order, persistent weather acts as drift.

Corollary 4.2 takes (γ_m, σ_m) as inputs. Two other certificates are scored on the same folds, both fixed on the training folds before their held-out scores were computed.

The stream-scale pair uses five-minute differences for noise and the largest hourly-mean slope for drift. Polar night is then the fastest client. Every window binds at the count floor $2K/\lambda = 1/6$ hour, personalization vanishes, and the score is -0.231 against -0.238 for the tuned length.

The two-scale pair keeps the climatological noise and sets γ_m to the sum of the clock Lipschitz constant and the largest change of successive 12-hour block means of the climate residual, per hour, a scale above the hour-scale windows and below a day. Polar night is no longer stationary. The prescribed windows, averaged over the training folds, are $(0.356, 0.665, 1.11)$ hours, $g = 0.078$, and the count floor does not bind.

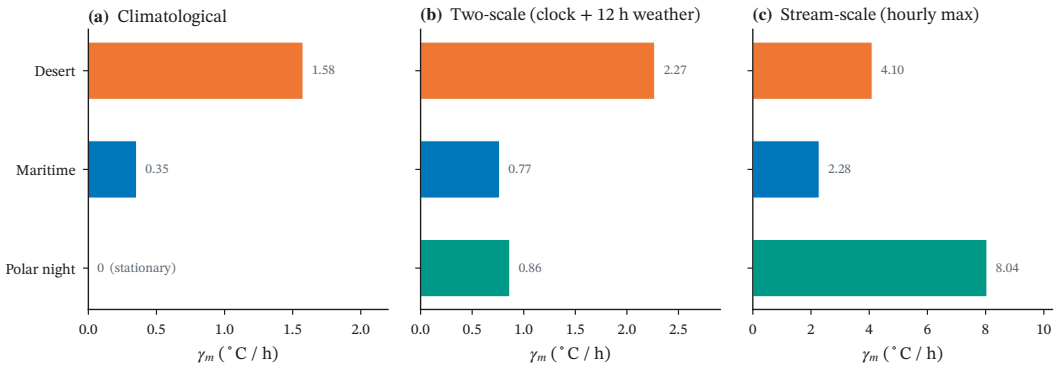

Figure 12: Drift certificates on the December training folds. (a) Climatological slopes: polar night is stationary. (b) Two-scale: clock plus the 12-hour residual Lipschitz constant. (c) Stream-scale hourly maxima, the opposite extreme.

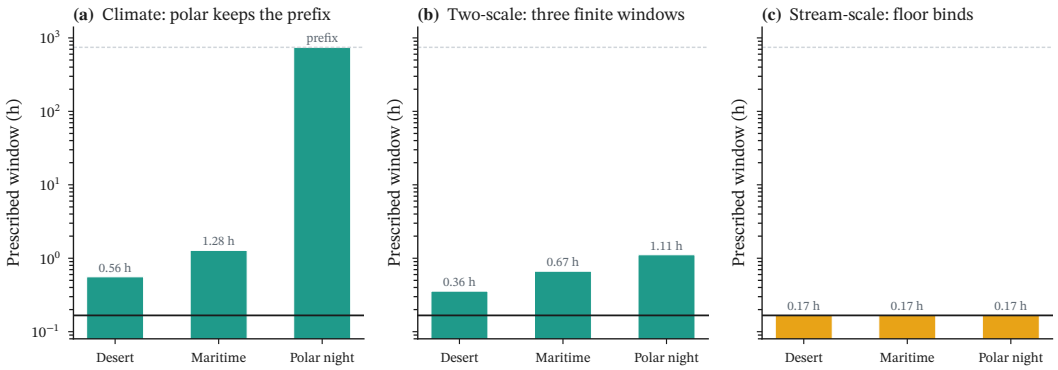

Figure 13: Windows the three certificates prescribe. Climate keeps the full history at the polar site. Two-scale gives three finite lengths $(0.36, 0.67, 1.11)$ hours. Stream-scale binds at the count floor, so personalization vanishes.

On the held-out Decembers the two-scale windows have mean -0.2385 against -0.2382 for the common length tuned on the other nine years, a paired difference of -0.0003 with interval $[-0.0008, 0.00008]$, and are lower in seven of the ten years. Against the climatological windows they win in all ten years, by 0.071 with interval $[0.038, 0.105]$. Against the stream-scale floor they win in all ten, by 0.0075 with interval $[0.0061, 0.0089]$. Squared loss repeats the pattern.

On the ten Julys the same two-scale certificate matches the tuned common length: it is lower in all ten years, by 0.00022 with interval $[0.00011, 0.00033]$, about 0.06% of the score. In July every station has a diurnal clock, so the climatological windows are already next to the tuned length. The two-scale windows use no grid: only the training-fold certificate.

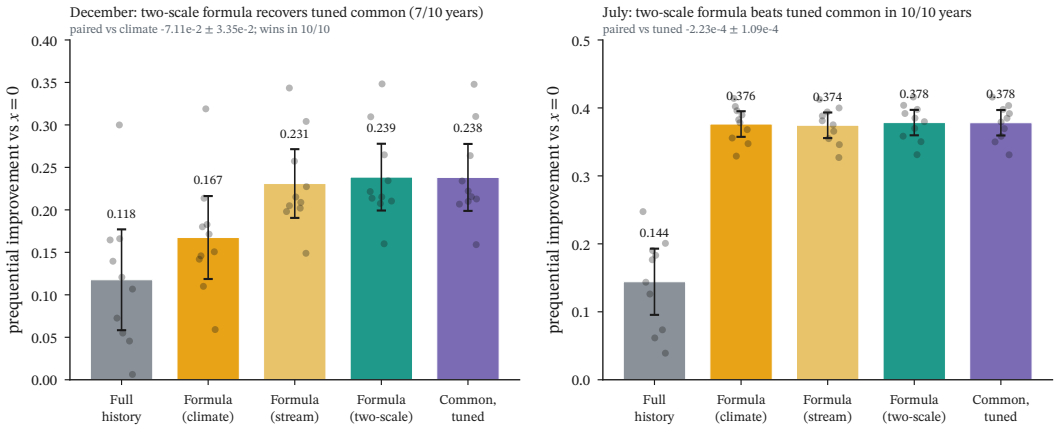

Figure 14: Prequential improvement over the zero report, December (left) and July (right); the improvement is the negative of the prequential excess reported in the text and in Table 4. Points are years; bars are means with Student 95% intervals. Higher is better. The two-scale windows recover the tuned common length in December and match it in July.

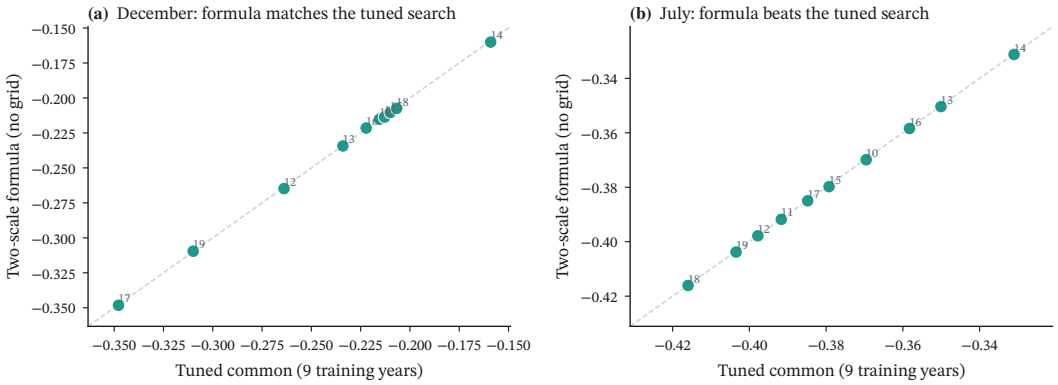

Figure 15: Two-scale windows against the common length tuned on the other nine years, one point per held-out year; lower is better. The two-scale windows use only the training-fold certificate. Both months sit on the diagonal; in July every point lies slightly below it.

The loss of the climatological windows on the recorded months is therefore a wrong input, not a failure of the window rule. A certificate that treats weather as drift, fixed on the training folds, returns finite windows and matches a search over common lengths.

J.3 Summary

Law	loss	regime	g	client-specific	minus tuned common
December stations	endpoint	yes	0.32	0.0120	-0.0128 $[-0.0144, -0.0113]$
July stations	endpoint	yes	0.066	0.0128	-0.0017 $[-0.0025, -0.0008]$
July Beijing districts	endpoint	yes	0.0002	0.0235	-0.0002 $[-0.0013, 0.0009]$
December, real order, climate	endpoint	no	0.32	-0.167	0.071 $[0.037, 0.104]$
December, real order, stream-scale	endpoint	no	0.09	-0.231	0.007 $[0.006, 0.008]$
December, real order, two-scale	endpoint	no	0.078	-0.239	-0.0003 $[-0.0008, 0.00008]$
July, real order, climate	endpoint	no	0.066	-0.376	0.0019 $[0.0015, 0.0023]$
July, real order, two-scale	endpoint	no	0.022	-0.378	-0.00022 $[-0.00033, -0.00011]$

Table 4: The two station experiments. The first three rows are population AvgErr_T over seeds 4–15; the remaining rows are prequential excess scores over ten held-out months (lower is better). Intervals are paired Student 95% intervals. "Regime" records whether every hypothesis of Corollary B.3 holds; in real order independence fails. The predicted gain g is computed before count floors, which bind for the stream-scale windows.

K STATIONARY AND COMMON-MEMORY BENCHMARKS

The stationary and common-memory specializations of Theorem 4.1 separate the roles of information arrival and client-specific drift. We record them to make the fixed-vector comparison of Section 3 explicit.

Stationary information. When $\gamma_m = 0$ for every client, all prefix memories are admissible and $\mathcal{H}_\gamma = 0$. The drift terms disappear, and the leading finite-horizon statistical term is proportional to $D\Xi/\sqrt{T}$, with

$$\Xi^2 = \sum_m \frac{w_m^2 \sigma_m^2}{\lambda_m}.$$

The complete bound retains the optimization and startup terms of (7) and (8). This information quantity weights a client's uncertainty by its contribution to the target and by its arrival rate. Proposition E.1 supplies the corresponding finite-horizon statistical lower bound, including the gradient cap and the finite-count corrections. Stationary clients therefore retain an information cost even though they contribute nothing to a persistent drift term.

Common memory and prescribed drift. Giving every client the same finite memory length h produces the window-dependent upper term

$$D\Big(2\bar{\gamma}h + \frac{\beta_0 \Xi}{\sqrt{h}}\Big), \qquad \bar{\gamma} = \sum_m w_m \gamma_m, \qquad \beta_0 = 4\sqrt{2/3},$$

and balancing its two terms yields the common-memory scale $D\bar{\gamma}^{1/3}\Xi^{2/3}$, before sample-count floors and the other remainders are imposed. Separate memories replace this scale by $D\mathcal{H}_\gamma$, as shown by Corollary 4.2 and (5). The comparison concerns the displayed upper terms; it does not establish that a common-memory policy is optimal among all causal estimators. The distinction also matters for lower bounds: an aggregate constraint lets drift be allocated among clients, whereas Theorem B.2 respects every prescribed clientwise bound, with the group-sensitive consequence in Corollary 5.1. The full proofs and the joint comparison conditions are in Appendices E and F.

L DETAILED COMPARISON OF PRIOR GUARANTEES

Table 5: Selected theorem assumptions and observation clocks. TV denotes total variation. Each guarantee retains the conditions of its cited theorem.

Reference and result	Assumptions and criterion	Clock and decisive distinction
Marfoq et al. (2023), Thms. 4.1–4.4	Fixed local laws; bounded convex smooth sample losses, bounded domain, a stated pseudo-dimension condition; terminal population excess risk.	Client arrivals and finite caches; the optimization theorem uses full participation and full batches. Our populations drift under separate clientwise bounds.
Ganguly & Aggarwal (2024), Master-FL	Bounded convex Lipschitz losses and near-stationary base guarantees; dynamic regret for changing empirical losses.	Federated rounds with drift detection and restarts. Its target uses current data-size weights; ours fixes population weights and models noisy samples arriving over time.
Tyurin (2025), Thms. H.11–H.12	Fixed convex smooth objective; unbiased local gradients with bounded variance; terminal expected objective gap.	Heterogeneous gradient-computation times. Our clocks govern reusable data from changing populations.
Cutler et al. (2023), Thm. 8	Strong convexity; fixed convex regularizer; second-moment drift/noise and decorrelation conditions; current gap of an averaged iterate.	One current stochastic gradient per iteration. We use a merely convex target and separate client drifts and arrival rates.
Mazzetto & Upfal (2023b), Thms. 4.2, 4.9	Independent finite-support observations; regular endpoint drift for terminal TV risk, bounded successive drift for long-run average TV risk.	Observation index. Our criterion is convex suboptimality of a weighted client aggregate, with unequal group clocks.
This work, Thms. 4.1, B.2	Compact convex domain; public clientwise gradient-drift bounds; expected time-average current-objective gap.	Exogenous arrivals and paid synchronization; reset solves and sample splitting permit reuse. Local computation is uncharged.