SUM-RATE CONSTRAINT INDUCES PARTIAL PARTICIPATION AND BIDIRECTIONAL COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication-efficient distributed optimization typically treats compression, partial worker participation, and downlink communication as separate design choices. We show that they can instead arise from a single resource constraint: a hard per-round sum-rate (**SR**) constraint on the aggregate communication rate of the workers, $\sum_{i=1}^{n} R_t^i \leq R$. The budget R creates a direct tradeoff between participation and compression fidelity: involving more workers leaves fewer bits available for each update. Under a common-broadcast downlink of rate R, the server can then rebroadcast the received worker codewords so that all workers reconstruct the global state without an additional lossy server-side compressor. We instantiate this principle in **EF21-SR**, a resource-constrained variant of the popular EF21 error-feedback method. We establish $\mathcal{O}(1/T)$ convergence to stationarity under heterogeneous compression. Experiments show that EF21-SR requires the least cumulative communication to approach the target among the compared methods with tuned stepsizes on both classification and least-squares problems.

1 INTRODUCTION

Modern distributed optimization considers learning problems in which data are distributed across n workers and the global objective is

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right\}, \tag{1}$$

where worker i has access to its local objective $f_i\colon \mathbb{R}^d \to \mathbb{R}$. In a parameter-server architecture, workers compute local information and transmit it to a server, which aggregates the messages and returns the state needed for the next iteration. Repeated exchange of high-dimensional vectors can dominate bandwidth, latency, and energy consumption, particularly in federated and edge-learning systems (McMahan et al., 2017; Kairouz et al., 2021; Cao et al., 2023). Two standard responses are to reduce the information in each active message through compression or to reduce the number of communicating workers through partial participation. Under a shared worker-rate budget these mechanisms are inseparable: spreading a fixed budget across more workers broadens participation but lowers the fidelity of each message, whereas concentrating it on fewer workers improves their messages but leaves more local trackers stale.

Several different constraints are called a "sum-rate constraint" in the literature, and they should not be identified with one another. In multiterminal source coding, a sum-rate–distortion function characterizes the aggregate information needed to reconstruct a statistic or model update to a prescribed distortion; the federated-learning formulation of Zhang et al. (2023b) treats correlated worker observations in this information-theoretic sense. In wireless learning, a multiple-access channel supplies a coupled physical rate region or an over-the-air aggregation mechanism, so power, fading, interference, and simultaneous transmission determine the feasible aggregate service (Mohammadi Amiri & Gündüz, 2020; Yang et al., 2020; Jeon et al., 2022). Resource-allocation work instead constrains bandwidth, airtime, latency, or long-term energy while jointly selecting clients (Xu & Wang, 2021; Ji et al., 2022). Finally, communication-complexity results constrain or count the cumulative information used to reach an optimization target rather than the payload in one round (Arjevani & Shamir, 2015; Korhonen & Alistarh, 2021; Huang et al., 2022; He et al., 2023b). Our

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.

constraint is different and deliberately narrower: R_t^i is a source-coding rate in bits per coordinate and

$$\sum_{i=1}^{n} R_t^i \leq R, \qquad t = 0, 1, \ldots, \tag{2}$$

holds pathwise in every round. The extended review in Appendix B develops these distinctions.

Communication compression includes quantization, sparsification, and low-rank approximation, as exemplified by QSGD, sparsified SGD, signSGD, and PowerSGD (Alistarh et al., 2017; Stich et al., 2018; Bernstein et al., 2018; Vogels et al., 2019). Because aggressive biased compression can accumulate error, error-feedback methods retain information lost in earlier messages (Karimireddy et al., 2019). Of particular relevance here, EF21 uses a contractive Markov gradient tracker and obtains the standard $\mathcal{O}(1/T)$ stationarity rate for smooth nonconvex optimization (Richtárik et al., 2021). Related tracking, randomized-refresh, and lazy-communication frameworks include 3PC, MARINA, BEER, and MURANA (Richtarik et al., 2022; Gorbunov et al., 2021; Zhao et al., 2022; Condat & Richtárik, 2022).

Partial participation (PP) refreshes only a subset of workers in each round. It arises from availability and stragglers, but can also be imposed to control communication and computation. Its effects on drift, variance, and memory are studied by SCAFFOLD, FedVARP, and the broader client-sampling literature (Karimireddy et al., 2020; Jhunjhunwala et al., 2022; Fraboni et al., 2023; Chen et al., 2022). Combining compression with partial participation requires additional care because skipped workers retain stale compression memories or trackers (Li & Li, 2023; Tyurin & Richtarik, 2023; Redie et al., 2026). Most directly, EF21-PP samples a proper subset of workers and refreshes only their local EF21 trackers (Fatkhullin et al., 2025).

Bidirectional methods (BC) address a different bottleneck by compressing both worker-to-server and server-to-worker communication. Examples include DoubleSqueeze, EF21-BC, EF21-P, and 2Direction (Tang et al., 2019; Fatkhullin et al., 2025; Gruntkowska et al., 2023; Tyurin & Richtárik, 2023). Downlink compression is important when repeatedly distributing a large model or optimization state is itself expensive, as in DoCoFL and GlueFL (Dorfman et al., 2023; He et al., 2023a). In EF21-BC, workers compress local tracker innovations and the server applies a second contractive compressor to the aggregate discrepancy Fatkhullin et al. (2025). By contrast, EF21-PP changes which worker trackers are refreshed but does not derive participation from a resource constraint. These variants retain the same underlying tracker and therefore provide controlled comparisons for the resource question studied here.

1.1 Contributions

First, we formulate the pathwise resource model in (2) and introduce EF21-SR. When each active message needs at least R_c bits per coordinate to be strictly contractive, the budget determines how many worker trackers can be refreshed. The static realization chooses a feasible fixed-size active set and divides the budget uniformly; the adaptive realization permits heterogeneous positive rates and zero-rate decisions, jointly determining fidelity and participation.

Second, we show that the same codeword bundle can synchronize the global state. After receiving the active-worker codewords, the server broadcasts that bundle to all workers. If its source rate is at most R, apart from separately counted metadata, every worker reconstructs the aggregate tracker increment and updates the common surrogate and iterate locally, so no additional lossy server compressor is required anymore unlike bidirectional compression.

Third, we establish smooth nonconvex convergence directly in terms of the realized worker- and iteration-dependent contractions δ_t^i. The definition is post-allocation: a zero-rate worker has $\delta_t^i = 0$ in that round. Random-participation results average this realized quantity only at the stage of the proof where the active-set randomness is integrated out. This resolves the distinction between a realized contraction and a participation-averaged refresh strength. The analysis retains the arithmetic mean of the local smoothness constants. We also prove a result for exact-current adaptive allocation by minimizing the current conditional compression certificate before taking the final expectation.

1.2 Organization

Section 2 defines the communication model and the static EF21-SR realization. Section 3 states the convergence results and communication comparison. Section 4 reports the experimental results, and Section 5 concludes. The appendices summarize the relevant EF21 baselines, give complete proofs, develop adaptive rate allocation, translate the sum-rate constraint to other compressors, specify additional experiments and implementation details, and provide an extended literature review.

2 Proposed Sum-Rate Constraint and Its EF21 Variant

We begin with the communication constraint rather than with a particular compression algorithm. At round t, let R_t^i denote the source-coding rate assigned to worker i, and impose the pathwise worker sum-rate budget in (2). The active set is the support of the rate vector:

$$\mathcal{S}_t \coloneqq \{i \in [n] : R_t^i > 0\}. \tag{3}$$

For a chosen compressor family, let $R_c > 0$ denote the minimum rate required for a strictly contractive worker message. An active worker must therefore receive a rate $R_t^i \geq R_c$, whereas $R_t^i = 0$ represents no communication and no refresh of that worker's local tracker. Since every active worker consumes at least R_c units of rate, (2) implies

$$|\mathcal{S}_t| \leq m \coloneqq \min\left\{n, \left\lfloor \frac{R}{R_c} \right\rfloor\right\}. \tag{4}$$

If the budget $R \geq nR_c$, full participation is feasible. If $R < nR_c$, every nonempty feasible round necessarily induces partial participation. Unlike a collection of independent per-worker limits, the shared budget couples participation and compression quality: activating more workers broadens participation but leaves less rate for each message, whereas concentrating the budget on fewer workers permits higher-fidelity transmissions.

The direction in which the constraint is imposed is equally important. Since the received worker codewords have a total source rate at most R, a server whose downlink permits a common rate-R broadcast can retransmit the same codeword bundle. Every worker can then reconstruct the global iterate without a second lossy server compressor. This is different from imposing only a downlink constraint: a downlink-only limit does not constrain worker participation and instead requires a server-side compression mechanism such as that used by EF21-BC. The rebroadcast argument also relies on a common broadcast channel, where one packet is charged once rather than once per recipient; any nonreproducible metadata must be counted separately, as discussed in Appendix D.

Two natural realizations follow from this principle. The first is a static realization: it selects an active set of fixed cardinality s and divides the available budget uniformly,

$$R_t^i = \begin{cases} R/s & : i \in \mathcal{S}_t \\ 0 & : i \notin \mathcal{S}_t \end{cases}, \qquad |\mathcal{S}_t| = s,\ 1 \leq s \leq m. \tag{5}$$

Every active worker then receives at least R_c. Choosing $s = m$ maximizes participation, while choosing a smaller s increases the rate, and hence potentially the compression quality, of the few selected messages. The second is adaptive realization and optimizes the rate vector directly subject to

$$R_t^i \in \{0\} \cup [R_c, \infty), \qquad \sum_{i=1}^{n} R_t^i \leq R. \tag{6}$$

The support of the resulting rate vector determines $\mathcal{S}_t$, so worker selection and compression quality are determined jointly.

This principle is not specific to EF21. The sum-rate layer can be placed around any distributed method with a rate-dependent worker compressor and a well-defined zero-rate action. The optimization method specifies what each worker would communicate; the sum-rate policy specifies which workers communicate and at what rates. We use the prevalent EF21 (Richtárik et al., 2021) error-feedback. The static implementation is given in Algorithm 1. The active sets may be generated by a shared rule, such as a predetermined schedule or public random seed, to avoid additional communication.

Algorithm 1: `EF21-SR`: static uniform-allocation realization

Input: stepsize η, budget R, cardinality $s \le m$, shared fixed-cardinality rule for generating $\mathcal{S}_t$
Initialize: starting point $\mathbf{x}_0$, local-gradient tracker $\widehat{\mathbf{g}}_{-1}^i = \mathbf{0}\ \forall\, i$, surrogate gradient $\widehat{\mathbf{g}}_{-1} = \mathbf{0}$
for $t = 0, 1, \ldots, T-1$ **do**
  Worker (parallel for every $i \in [n]$**)**
    $\widehat{\mathbf{g}}_{t-1} = \widehat{\mathbf{g}}_{t-2} + \frac{1}{n}\sum_{i\in\mathcal{S}_{t-1}} \mathbf{q}_{t-1}^i$;
    $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta\, \widehat{\mathbf{g}}_{t-1}$;
    $\mathbf{q}_t^i = \begin{cases} \mathcal{C}\big(\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i\big) & : i \in \mathcal{S}_t \\ \mathbf{0} & : i \notin \mathcal{S}_t \end{cases}$
    $\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i$;
    Workers $i \in \mathcal{S}_t$, send $\mathbf{q}_t^i$ to the server.
  Server
    $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i\in\mathcal{S}_t} \mathbf{q}_t^i$;
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\, \widehat{\mathbf{g}}_t$; `// store locally`
    Broadcast the bundle $\{\mathbf{q}_t^i : i \in \mathcal{S}_t\}$ to all workers ; `// broadcast rate` $\le R$
  `//` All workers and the server can reconstruct $\widehat{\mathbf{g}}_t$ and $\mathbf{x}_{t+1}$.

Algorithm 1 couples participation and compression through a single rate budget. Interestingly, the server neither transmits $\mathbf{x}_{t+1}$ as a separate full-precision vector nor applies a second lossy compressor. Instead, every worker decodes the same rate-R bundle, reconstructs the aggregate tracker increment, and updates the common surrogate and iterate locally. Consequently, inactivity makes only the worker-specific tracker $\widehat{\mathbf{g}}_t^i$ stale; the global surrogate $\widehat{\mathbf{g}}_t$ and iterate $\mathbf{x}_{t+1}$ remain common to all workers. When worker i becomes active again, it evaluates $\nabla f_i(\mathbf{x}_t)$ at the current iterate and corrects that gradient against its own stale tracker.

Fixed-cardinality sampling is particularly natural under a pathwise budget because it fixes the number of positive-rate messages in every round. By contrast, independent Bernoulli sampling with the same expected cardinality can select more than m workers and thereby violate (2). Moreover, changing s affects both the participation frequency and the active compressor through the assigned rate R/s.

The adaptive realization in Algorithm 2 selects the active workers and their heterogeneous rates by solving an instance of (6) using the compression difficulty of the worker messages. For a fixed active set, this is a heterogeneous rate-allocation problem; allowing zero rates makes the active set part of the optimization. Exact same-round allocation based on the current innovation norms requires those norms before the allocation is chosen and therefore introduces a scalar control phase unless the server can infer them. Predictable envelopes, delayed state information, or a static robust allocation can avoid that same-round exchange. The allocation objective, the reverse-waterfilling solution for the fixed-rate quantizer, and the accounting of side information and control metadata are described in Appendix D for the interested reader.

In summary, partial participation and bidirectional compression deal with different locations of the communication bottleneck. For workers, the sum-rate constraint (2) can induce partial participation. Since retransmitting the worker codeword bundle satisfies payload constraints, one obtains finite-rate communication in both directions without an additional server compressor. Within the `EF21` family, `EF21-PP` and `EF21-BC` are different architectural responses to where the communication constraint is imposed, while the same resource viewpoint in `EF21-SR` places them within a common communication model.

3 Theoretical Convergence Results

Throughout, $\|\cdot\|$ denotes the Euclidean norm, $f^* := \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$, and all local gradients are evaluated exactly. The only randomness arises from worker allocation and compression.

(A1) *Smooth finite sum.* Each f_i is differentiable and L_i-smooth:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L_i\|\mathbf{x}-\mathbf{y}\|, \qquad \mathbf{x},\mathbf{y}\in\mathbb{R}^d. \tag{7}$$

Moreover, $f = n^{-1}\sum_i f_i$ is L-smooth. We write

$$\bar{L} \coloneqq \frac{1}{n}\sum_{i=1}^n L_i, \qquad \Delta_0 \coloneqq f(\mathbf{x}_0) - f^*. \tag{8}$$

(A2) *Biased contractive compressor.* Let $\mathcal{G}_t$ be the information available after the round-t allocation and immediately before the worker compressors are applied, as formalized in Appendix C.1. Define

$$\mathbf{u}_t^i \coloneqq \nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i. \tag{9}$$

For every worker i, there is a $\mathcal{G}_t$-measurable $\delta_t^i \in [0,1]$ such that

$$\mathbb{E}\Big[\big\|\mathcal{C}_t^i(\mathbf{u}_t^i) - \mathbf{u}_t^i\big\|^2 \,\Big|\, \mathcal{G}_t\Big] \le (1-\delta_t^i)\|\mathbf{u}_t^i\|^2. \tag{10}$$

Here $\mathcal{C}_t^i$ denotes the complete post-allocation worker action. If worker i is inactive, $\mathcal{C}_t^i(\mathbf{u}_t^i) = \mathbf{0}$ and $\delta_t^i = 0$. If it is active, examples satisfying (10) include the fixed-rate vector quantizer

$$\big\|\mathcal{C}_t^i(\mathbf{u}_t^i) - \mathbf{u}_t^i\big\|^2 \le \rho_z^2 2^{-2R_t^i}\|\mathbf{u}_t^i\|^2, \tag{11}$$

applied block-wise on a block of length z. If $z = d$, one recovers the quantizer employed by (Lin et al., 2022). For integer rates and covering efficiency $\rho_z \ge 1$, a rate of

$$R_c \coloneqq \lfloor \log_2 \rho_z \rfloor + 1 \tag{12}$$

is sufficient for strict contractivity of the fixed-rate quantizer at integer rates. Other compressor families are discussed in Appendix E.

For `EF21-SR`, define the $\mathcal{G}_t$-measurable conditional error certificate

$$E_t \coloneqq \frac{1}{n}\sum_{i=1}^n \sqrt{1-\delta_t^i}\,\|\mathbf{u}_t^i\|. \tag{13}$$

Theorem 3.1 (Smooth nonconvex convergence of `EF21-SR`)**.** *Suppose* (A1)–(A2) *hold and*

$$\sup_{\tau\ge 1}\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_\tau^i}\right\} < 1. \tag{14}$$

If $\eta > 0$ *satisfies*

$$\frac{1}{\eta}\left(\frac{1}{\eta} - L\right) \ge \left(\frac{\sup_{\tau\ge 1}\operatorname{ess\,sup}\left\{\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_\tau^i}\right\}}{1 - \sup_{\tau\ge 1}\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_\tau^i}\right\}}\right)^2, \tag{15}$$

then, for every $T \ge 1$,

$$\min_{0\le t<T}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] \le \frac{2}{\eta T}\left(\Delta_0 + c\,\mathbb{E}[E_0^2]\right), \tag{16}$$

where

$$c \coloneqq \frac{\eta}{2}\left(1 - \sup_{\tau\ge 1}\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_\tau^i}\right\}\right)^{-1}. \tag{17}$$

Proof sketch. Conditional Minkowski's inequality and *(A2)* give $(\mathbb{E}[\|\widehat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\|^2 \mid \mathcal{G}_t])^{1/2} \le E_t$. The smoothness descent inequality therefore contains the additive term $\eta E_t^2/2$. Smoothness of every f_i gives a one-step recursion for E_{t+1} in terms of E_t and $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|$. The optimized two-term square inequality turns the essential supremum in (14) into a strict contraction. Adding $c\mathbb{E}[E_t^2]$ to the objective gap, using (15), and telescoping yields (16). The complete proof is in Appendix C.2. □

Theorem 3.1 covers heterogeneous, time-varying realized rates whenever every worker has a strictly positive contraction in every round. It does not turn a realized zero-rate action into a positive contraction. Random partial participation is instead averaged over its fresh active-set draw.The fixed-rate worker-allocation specialization is stated later.

Corollary 3.2 (Static uniform-allocation realization)**.** *Let $s \in \{1, \dots, m\}$ and sample $\mathcal{S}_t$ uniformly without replacement from the subsets of $[n]$ of cardinality s, independently of the preceding* EF21-SR *history. Set R_t^i according to* (2) *and let*

$$I_t^i := \mathbb{I}\{i \in \mathcal{S}_t\}. \tag{18}$$

For the fixed-rate compressor, the realized post-allocation contraction is

$$\delta_t^i = I_t^i \left(1 - \rho_z^2 2^{-2R/s}\right). \tag{19}$$

With exact tracker initialization, a stepsize satisfying

$$\frac{1}{\eta}\left(\frac{1}{\eta} - L\right) \geq \left(\frac{\bar{L}\sqrt{1 - \frac{s}{n}\left(1 - \rho_z^2 2^{-2R/s}\right)}}{1 - \sqrt{1 - \frac{s}{n}\left(1 - \rho_z^2 2^{-2R/s}\right)}}\right)^2 \tag{20}$$

gives

$$\min_{0 \leq t < T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leq \frac{2\Delta_0}{\eta T}. \tag{21}$$

Consequently, it is sufficient to take

$$T_{\text{EF21-SR}} = \mathcal{O}\left(\frac{n\bar{L}\Delta_0}{s\left(1 - \rho_z^2 2^{-2R/s}\right)\varepsilon}\right) \tag{22}$$

to ensure the left-hand side of (21) *is at most ε.*

Proof sketch. For EF21-SR, $\mathcal{F}_{t-1}$ is the complete history before the fresh round-t active-set draw. Although (19) equals zero when worker i is inactive, conditioning one stage earlier gives

$$\mathbb{E}\left[\delta_t^i \mid \mathcal{F}_{t-1}\right] = \frac{s}{n}\left(1 - \rho_z^2 2^{-2R/s}\right). \tag{23}$$

Applying conditional Minkowski's inequality after this averaging produces the factor in (20); the Lyapunov argument of Theorem 3.1 then applies at the coarser conditioning stage. □

The corresponding EF21-PP and EF21-BC results are derived in Appendix C in terms of arithmetic mean of smoothness constants $\bar{L}$, with the published analyses cited there (Fatkhullin et al., 2025). Their leading dependencies and the EF21-SR statement are summarized in Table 1.

For communication accounting, all rates are in bits per coordinate. Thus a vector at rate r carries dr vector-payload bits. Let R_w denote the per-worker rate in the classical compressed methods, let R denote the server-compressor rate in EF21-BC and the worker sum-rate in EF21-SR, and let R_{fp} denote the bits per coordinate of the full-precision iterate/model representation. The common-broadcast model charges a common server packet once. Worker identifiers, rate descriptions, quantizer scales, support indices, and other noncommon metadata are excluded from the vector-payload column and are reported separately in the experiments.

The original second-last-column formulas have a factor n because they count *total* gradients across all workers; they are not per-worker counts. With the corrected heading in Table 1, that factor is appropriate. Per worker, divide the homogeneous-method totals by n. Exact-current adaptive allocation is an important exception to the general EF21-SR entry: if current innovation norms are required before rates are chosen, every worker computes its gradient, so the count is nT_{SR} even though only $|\mathcal{S}_t|$ vector messages are compressed and transmitted. Static or predictable-envelope allocation needs only the selected workers' current gradients. In every EF21-SR realization the server performs no compressor call, whereas EF21-BC performs server-compression per round.

Table 1: Smooth nonconvex round, total gradient-computation, and cumulative two-way vector-payload costs for $\min_{0\le t<T}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2]\le\varepsilon$. Here $p=s/n$, $\widetilde{L}=(n^{-1}\sum_i L_i^2)^{1/2}$, R_{fp} means the bits per coordinate used for a full-precision vector, and $T_{\mathrm{GD}}, T_{\mathrm{PP}}, T_{\mathrm{BC}}, T_{\mathrm{SR}}$ denote the respective certified horizons. The gradient column counts all workers together after the common one-time exact initialization. The EF21-BC row reports the standard published product dependence, while the EF21-SR general row retains the theorem-dependent stepsize η and the static row substitutes (22). Payloads use common broadcast and omit metadata.

Method	Active workers	Round complexity	Total gradient computations	Cumulative vector-payload bits
EF21-GD	n	$\mathcal{O}(\bar{L}\Delta_0/(\delta\varepsilon))$	nT_{GD}	$d(nR_w+R_{\mathrm{fp}})T_{\mathrm{GD}}$
EF21-PP	$s=np$	$\mathcal{O}(\bar{L}\Delta_0/(p\delta\varepsilon))$	sT_{PP}	$d(sR_w+R_{\mathrm{fp}})T_{\mathrm{PP}}$
EF21-BC	n	$\mathcal{O}(\bar{L}\Delta_0/(\delta\delta_s\varepsilon))$	nT_{BC}	$d(nR_w+R)T_{\mathrm{BC}}$
EF21-SR	$\lvert\mathcal{S}_t\rvert\le m$	$\mathcal{O}(\Delta_0/(\eta\varepsilon))$	$\sum_{t<T_{\mathrm{SR}}}\mathbb{E}[\lvert\mathcal{S}_t\rvert]$ if preselected; sT_{SR} if $\lvert\mathcal{S}_t\rvert=s$	at most $2dRT_{\mathrm{SR}}$

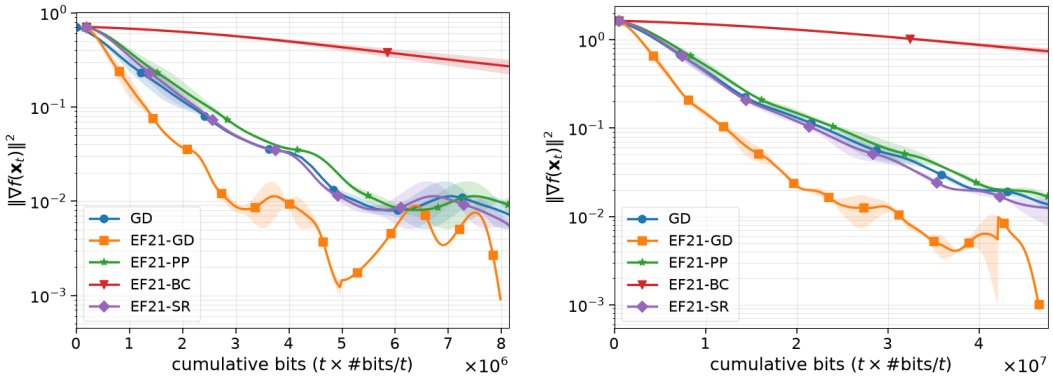
At a fixed horizon T, static EF21-SR uses sT worker gradients rather than the nT of EF21-BC; exact-current adaptive allocation does not obtain this gradient saving. Matching EF21-SR and EF21-PP by $R_w=R/s$ and $p=s/n$ makes their worker refreshes and certified horizons identical. Their per-round two-way payloads are $2dR$ and $d(R+R_{\mathrm{fp}})$, respectively, so EF21-SR uses fewer bits exactly when $R<R_{\mathrm{fp}}$. Against EF21-BC, the fixed-T payloads are $2dRT$ and $d(nR_w+R)T$, so EF21-SR uses fewer bits exactly when $R<nR_w$.

At fixed accuracy ε, the horizon must also be counted. For matched static EF21-SR and EF21-PP, the horizons agree, so the condition $R<R_{\mathrm{fp}}$ is unchanged. In general, the EF21-SR and EF21-BC payloads are $2dRT_{\mathrm{SR}}(\varepsilon)$ and $d(nR_w+R)T_{\mathrm{BC}}(\varepsilon)$, respectively. Thus EF21-SR communicates fewer bits precisely when $T_{\mathrm{SR}}(\varepsilon)/T_{\mathrm{BC}}(\varepsilon)<(nR_w+R)/(2R)$. Its static gradient count is smaller precisely when $sT_{\mathrm{SR}}(\varepsilon)<nT_{\mathrm{BC}}(\varepsilon)$. These conditions state when EF21-SR outperforms without assuming that a smaller per-round payload automatically implies a smaller cost to reach ε.

4 Experimental Results

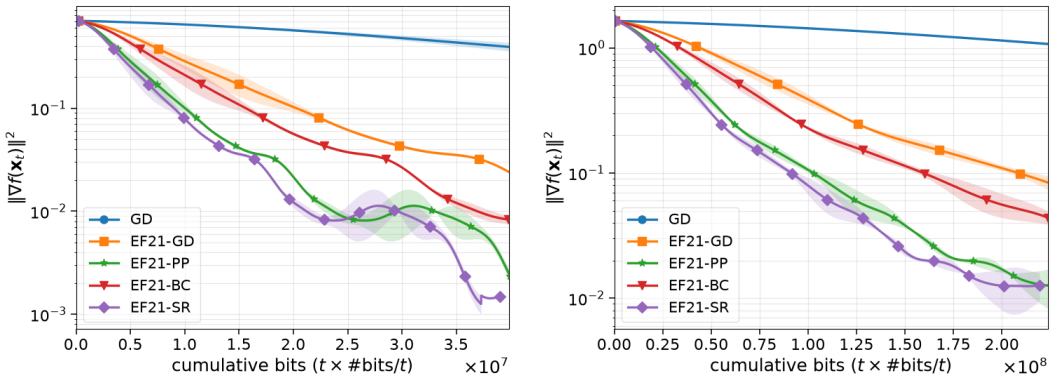
We evaluate the pathwise worker sum-rate model on two publicly available problems with different dimensions and sample-to-worker sizes: binary logistic regression on the `w8a` dataset from the LIBSVM repository, and least-squares on the `illc1850` matrix from the SuiteSparse Matrix Collection. We compare the proposed EF21-SR with full-precision distributed GD, worker-compressed EF21-GD, partial-participation EF21-PP, and bidirectional EF21-BC. All compressed methods use the group-wise fixed-rate uniform quantizer of block length $z=127$ (covering efficiency $\rho_z=\sqrt{z}$, minimum rate $R_c=\lfloor\log_2\sqrt{z}\rfloor+1=4$ bits/dim), unless stated otherwise. Data are partitioned across $n=20$ workers (label-sorted shards for classification, random shards for least squares). Each run stops at $\|\nabla f(\mathbf{x}_t)\|^2\le 10^{-3}$ rather than at a fixed iteration count. Curves average two seeds with $\pm$std bands. The evaluation metric is $\|\nabla f(\mathbf{x}_t)\|^2$; smaller values indicate closer approach to first-order stationarity. The horizontal axis is cumulative two-way communication in bits (uplinks, one common downlink packet, and group-scale metadata), not average per-worker uplink alone. Complete objectives, partitions, compressors, stepsize rules, and bit-accounting are given in Appendix F.

Figures 1 to 3 report the three stepsize regimes on both datasets. Under maximum admissible steps (Figure 1), EF21-GD typically reaches the tolerance early: it refreshes every tracker at a mild encoder and therefore retains a relatively large theory step. Matched EF21-SR and EF21-PP follow nearly identical optimization trajectories because they share p, R_w, and η; EF21-SR nevertheless spends fewer cumulative bits whenever the common rebroadcast of the rate-R codeword bundle is cheaper than a full-precision iterate downlink. Full-precision GD needs few iterations but pays $(n+1)dR_{\mathrm{fp}}$ bits every round, so its communication curve sits close to, and often slightly above, EF21-SR.



(a) `w8a`, maximum admissible η. (b) `illc1850`, maximum admissible η.

Figure 1: Nonconvex main comparison under per-method maximum admissible stepsizes as per the theory. Maximum feasible participation $m = \lfloor R/R_c \rfloor = 6$, worker rate $R_w = R_c$. ... The default design uses the maximum participation that remains feasible under a pathwise budget R when every active message is at least R_c-contractive: $s = \lfloor R/R_c \rfloor$ with $R = sR_c$ (so $s = 6$ and $R = 24$ in the default panels). Worker encoders are held at rate $R_w = R_c$ unless a sweep explicitly changes that quantity. `EF21-PP` is matched by $p = s/n$ and the same R_w.



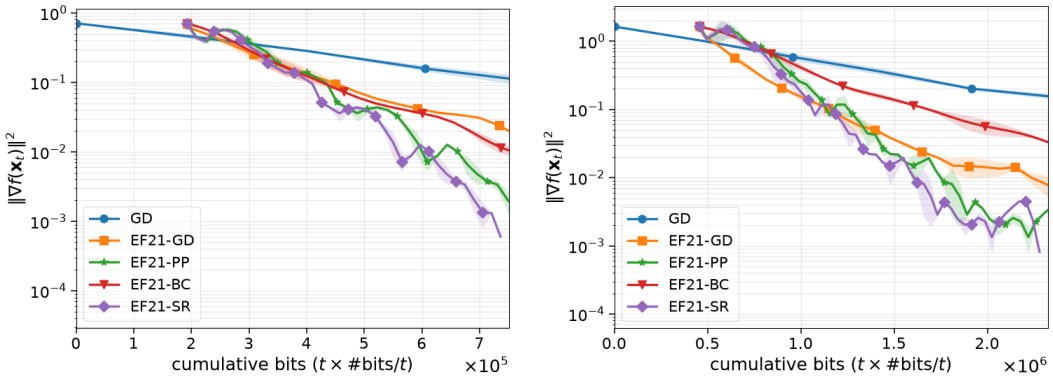
(a) `w8a`, common η. (b) `illc1850`, common η.

Figure 2: Same setting as Figure 1, but all methods use one common stepsize equal to the lowest of the five theory-max values. This isolates architecture and bit accounting from per-method stepsize selection.

`EF21-BC` is limited by the stepsize and a second compressor, and is the most expensive method on both panels.

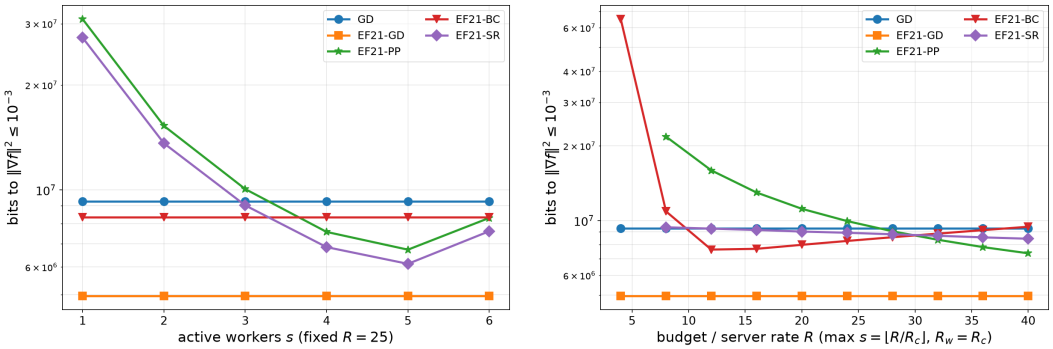
The common-η panels (Figure 2) remove stepsize heterogeneity by forcing every method to the most conservative theory-max step among the five (typically the `EF21-BC` value). Absolute progress is therefore slower, and on `illc1850` the shared step may not reach 10^{-3} within the cap, but the bit advantage of `EF21-SR` over matched `EF21-PP` remains a pure downlink effect whenever both curves are defined. Tuned steps (Figure 3) improve absolute communication for several methods without overturning that comparison. Taken together, the three regimes show that `SR` meets the standard.

Figure 4 isolates the two resource axes that the theory treats as coupled. At fixed pathwise budget $R = 25$, sweeping the active cardinality s affects only `EF21-SR`/`EF21-PP`: larger s raises participation probability $p = s/n$ but lowers worker rate $R_w = R/s$. Classical worker encoders stay at R_c, and `EF21-BC` keeps its server rate equal to the same fixed budget $R = 25$, so `GD`/`EF21-GD`/`EF21-BC` appear as flat references. The partial-participation bits-to-tolerance curve has an interior minimum near $s = 5$; the largest feasible $s = \lfloor R/R_c \rfloor$ is not optimal. In the second panel we sweep the



(a) `w8a`, tuned η. (b) `illc1850`, tuned η.

Figure 3: Same setting with independently tuned stepsizes (short grid search over multipliers of the theory-max step, scored by final $\|\nabla f\|^2$).



(a) Active set s at fixed budget $R = 25$. (b) Budget / server rate R at max $s = \lfloor R/R_c \rfloor$.

Figure 4: Controlled sweeps on `w8a` (nonconvex). Vertical axis: bits to $\|\nabla f\|^2 \le 10^{-3}$. Methods whose knobs do not move with the sweep axis appear as flat references; missing markers did not reach the tolerance within the iteration cap.

numerical budget R from $R_c = 4$ to 40 (just above the full-precision coordinate rate $R_{\text{fp}} = 32$), always taking the maximum feasible active set $s = \lfloor R/R_c \rfloor$ so matched `EF21-PP`/`EF21-SR` keep worker rate $R_w = R/s = R_c$ while participation grows. `EF21-BC` raises only its server rate; `GD`/`EF21-GD` stay flat at fixed worker rate R_c.

A scaled-`sign` compressor mirror of the three stepsize regimes, convex and strongly convex natures, two additional datasets (`a9a`, `bcsstk15`), the static-versus-adaptive `EF21-SR` comparison, the symbolic bit-cost table, and full implementation details appear in Appendix F.

5 Conclusion and Future Directions

We studied a pathwise worker sum-rate constraint that directly couples the number of participating workers with the fidelity of their compressed tracker innovations. `EF21-SR` spends this resource on worker refreshes and uses a common rebroadcast of the received codewords to synchronize the global surrogate and iterate without an additional lossy server compressor.

The analysis distinguishes the realized post-allocation contraction from its participation average. This makes zero-rate actions consistent across the algorithm, the allocation problem, and the proofs. The general theorem covers heterogeneous time-varying positive contractions, the static result averages fresh fixed-cardinality participation at the appropriate conditioning stage, and the exact-current adaptive result compares the minimum conditional certificate with feasible uniform allocations.

The principal empirical question is whether a fixed resource is better spent on more strongly compressed workers, fewer higher-fidelity workers, or a state-dependent allocation. EF21-SR is empirically shown to be effective.

Extensions to stochastic-gradient theory, time-varying budgets, and long-window allocation constraints remain open.

Reproducibility Statement

All assumptions, rate conventions, compressor models, and convergence claims are stated in the main text and appendices. Complete proofs appear in Appendix C; the rate-allocation rule, compressor translations, and experimental protocol are given in Appendices D to F.

References

Jayadev Acharya, Clement Canonne, Prathamesh Mayekar, and Himanshu Tyagi. Information-constrained optimization: can adaptive processing of gradients help? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7126–7138. Curran Associates, Inc., December 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/398475c83b47075e8897a083e97eb9f0-Paper.pdf.

Alyazeed Albasyoni, Mher Safaryan, Laurent Condat, and Peter Richtárik. Optimal Gradient Compression for Distributed and Federated Learning, October 2020. URL https://arxiv.org/abs/2010.03246.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., December 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6c340f25839e6acdc73414517203f5f0-Paper.pdf.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems*, NIPS'15, pp. 1756–1764, Cambridge, MA, USA, December 2015. MIT Press. URL https://proceedings.neurips.cc/paper/2015/hash/7fec306d1e665bc9c748b5d2b99a6e97-Abstract.html.

Toby Berger. Multiterminal source coding. In G. Longo (ed.), *The Information Theory Approach to Communications*, volume 229 of *CISM Courses and Lectures*, pp. 171–231. Springer, 1978.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed Optimisation for Non-Convex Problems. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569, Long Beach, CA, USA, July 2018. PMLR. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Xuanyu Cao, Tamer Başar, Suhas Diggavi, Yonina C. Eldar, Khaled B. Letaief, H. Vincent Poor, and Junshan Zhang. Communication-Efficient Distributed Learning: An Overview. *IEEE Journal on Selected Areas in Communications*, 41(4):851–873, April 2023. doi: 10.1109/JSAC.2023.3242710. URL https://doi.org/10.1109/JSAC.2023.3242710.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal Client Sampling for Federated Learning. *Transactions on Machine Learning Research*, August 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8GvRCWKHIL.

Yifei Cheng, Li Shen, Linli Xu, Xun Qian, Shiwei Wu, Yiming Zhou, Tie Zhang, Dacheng Tao, and Enhong Chen. Communication-efficient distributed learning with Local Immediate Error Compensation. *Neural Networks*, 188:107471, August 2025. ISSN 0893-6080. doi: 10.1016/j.neunet.2025.107471. URL https://doi.org/10.1016/j.neunet.2025.107471.

Laurent Condat and Peter Richtárik. MURANA: A Generic Framework for Stochastic Variance-Reduced Optimization. In *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pp. 81–96. PMLR, 2022. URL https://proceedings.mlr.press/v190/condat22a.html.

Laurent Condat, Artavazd Maranjyan, and Peter Richtárik. LoCoDL: Communication-efficient distributed learning with local training and compression. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PpYy0dR3Qw.

Laurent Condat, Artavazd Maranjyan, and Peter Richtárik. BiCoLoR: Communication-efficient optimization with bidirectional compression and local training. *arXiv preprint arXiv:2601.12400*, 2026. URL https://arxiv.org/abs/2601.12400.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Ron Dorfman, Shay Vargaftik, Yaniv Ben-Itzhak, and Kfir Yehuda Levy. DoCoFL: Downlink compression for cross-device federated learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8356–8388. PMLR, 2023.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with Bells & Whistles: Six Algorithmic Extensions of Modern Error Feedback. *Journal of Machine Learning Research*, 26(189):1–50, June 2025. URL http://jmlr.org/papers/v26/24-0059.html.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3407–3416. PMLR, 2021. URL https://proceedings.mlr.press/v139/fraboni21a.html.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. A General Theory for Client Sampling in Federated Learning. In Randy Goebel, Han Yu, Boi Faltings, Lixin Fan, and Zehui Xiong (eds.), *Trustworthy Federated Learning*, pp. 46–58, Cham, March 2023. Springer International Publishing. ISBN 978-3-031-28996-5. doi: 10.1007/978-3-031-28996-5_4. URL https://doi.org/10.1007/978-3-031-28996-5_4.

Eduard Gorbunov, Konstantin P. Burlachenko, Zhize Li, and Peter Richtarik. MARINA: Faster Non-Convex Distributed Learning with Compression. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3788–3798. PMLR, July 2021. URL https://proceedings.mlr.press/v139/gorbunov21a.html.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11761–11807. PMLR, July 2023. URL https://proceedings.mlr.press/v202/gruntkowska23a.html.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2818. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a1d20cc72a21ef971d7e49a90d8fa56f-Abstract-Conference.html.

Shiqi He, Qifan Yan, Feijie Wu, Lanjun Wang, Mathias Lécuyer, and Ivan Beschastnikh. GlueFL: Reconciling client sampling and model masking for bandwidth efficient federated learning. In *Proceedings of Machine Learning and Systems*, volume 5, 2023a.

Yutong He, Xinmeng Huang, and Kun Yuan. Unbiased compression saves communication in distributed optimization: When and how much? In Alice Oh, Tristan Neumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 47991–48020, 2023b. doi: 10.52202/075280-2081. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9602d22a8c791f23f8e4d1398e3fb5be-Abstract-Conference.html.

Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower Bounds and Nearly Optimal Algorithms in Distributed Learning with Communication Compression. In *Advances in Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 9781713871088. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77f2d0c271e508278ea13e24cd8773d5-Abstract-Conference.html.

Berivan Isik, Francesco Pase, Deniz Gunduz, Sanmi Koyejo, Tsachy Weissman, and Michele Zorzi. Adaptive compression in federated learning via side information. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 487–495. PMLR, 2024. URL https://proceedings.mlr.press/v238/isik24a.html.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10351–10375. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/jee-cho22a.html.

Yo-Seb Jeon, Mohammad Mohammadi Amiri, and Namyoon Lee. Communication-Efficient Federated Learning Over MIMO Multiple Access Channels. *IEEE Transactions on Communications*, 70(10):6547–6562, October 2022. doi: 10.1109/TCOMM.2022.3198433. URL https://doi.org/10.1109/TCOMM.2022.3198433.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 906–916. PMLR, 2022. URL https://proceedings.mlr.press/v180/jhunjhunwala22a.html.

Yun Ji, Zhoubin Kou, Xiaoxiong Zhong, Sheng Zhang, Hangfan Li, and Fan Yang. Client selection and bandwidth allocation for federated learning: An online optimization perspective. In *IEEE Global Communications Conference (GLOBECOM)*, pp. 5075–5080. IEEE, 2022. doi: 10.1109/GLOBECOM48099.2022.10001492.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, June 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, June 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, July 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Janne H. Korhonen and Dan Alistarh. Towards tight communication lower bounds for distributed optimisation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7254–7266, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3b92d18aa7a6176dd37d372bc2f1eb71-Abstract.html.

Fan Lai, Xiangfeng Zhu, Harsha V. Madhyastha, and Mosharaf Chowdhury. Oort: Efficient Federated Learning via Guided Participant Selection. In *15th USENIX Symposium on Operating Systems Design and Implementation*, pp. 19–35. USENIX Association, July 2021. ISBN 978-1-939133-22-9. URL https://www.usenix.org/conference/osdi21/presentation/lai.

Xiaoyun Li and Ping Li. Analysis of Error Feedback in Federated Non-Convex Optimization with Biased Compression: Fast Convergence and Partial Participation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19126–19155. PMLR, July 2023. URL https://proceedings.mlr.press/v202/li23ab/li23ab.pdf.

Chung-Yi Lin, Victoria Kostina, and Babak Hassibi. Differentially Quantized Gradient Methods. *IEEE Transactions on Information Theory*, 68(9):6078–6097, September 2022. doi: 10.1109/TIT.2022.3171173. URL https://doi.org/10.1109/TIT.2022.3171173.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, April 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Mohammad Mohammadi Amiri and Deniz Gündüz. Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air. *IEEE Transactions on Signal Processing*, 68:2155–2169, March 2020. doi: 10.1109/TSP.2020.2981904. URL https://doi.org/10.1109/TSP.2020.2981904.

Mohammad Mohammadi Amiri, Sanjeev R. Kulkarni, and H. Vincent Poor. Federated learning with downlink device selection. In *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 306–310. IEEE, 2021. doi: 10.1109/SPAWC51858.2021.9593132.

Andrei Panferov, Yury Demidovich, Ahmad Rammal, and Peter Richtárik. Correlated quantization for faster nonconvex distributed optimization. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3361–3387. PMLR, 2025. URL https://proceedings.mlr.press/v286/panferov25a.html.

C Philippenko and A Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: Tight convergence guarantees. arXiv:2006.14591, June 2020. URL https://arxiv.org/abs/2006.14591.

Constantin Philippenko and Aymeric Dieuleveut. Preserved central model for faster bidirectional compression in distributed settings. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/13d63838ef1fb6f34ca2dc6821c60e49-Abstract.html.

Dawit Kiros Redie, Reza Arablouei, and Stefan Werner. SA-PEF: Step-ahead partial error feedback for efficient federated learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=ejnVWfknCm. Accepted manuscript.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2021–2031. PMLR, 2020. URL https://proceedings.mlr.press/v108/reisizadeh20a.html.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34 of *NIPS '21*, pp. 4384–4396, Red Hook, NY, USA, December 2021. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.

Peter Richtarik, Igor Sokolov, Elnur Gasanov, Ilyas Fatkhullin, Zhize Li, and Eduard Gorbunov. 3PC: Three Point Compressors for Communication-Efficient Distributed Training and a Better Theory for Lazy Aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18596–18648. PMLR, July 2022. URL https://proceedings.mlr.press/v162/richtarik22a.html.

Peter Richtárik, Elnur Gasanov, and Konstantin Burlachenko. Error feedback reloaded: From quadratic to arithmetic mean of smoothness constants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ch7WqGcGmb.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Interspeech*, pp. 1058–1062, September 2014. doi: 10.21437/Interspeech.2014-274. URL https://doi.org/10.21437/Interspeech.2014-274.

David Slepian and Jack K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, July 1973. doi: 10.1109/TIT.1973.1055037.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, NIPS'18, pp. 4452–4463, Red Hook, NY, USA, December 2018. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3327345.3327357.

Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6155–6165. PMLR, 2019. URL https://proceedings.mlr.press/v97/tang19d.html.

S.-Y. Tung. *Multiterminal Source Coding*. PhD thesis, Cornell University, 1978.

Alexander Tyurin and Peter Richtarik. A computation and communication efficient method for distributed nonconvex problems in the partial participation setting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/778ff1fcfb6d6707fc015908a1845b62-Abstract-Conference.html.

Alexander Tyurin and Peter Richtárik. 2direction: Theoretically faster distributed training with bidirectional communication compression. In Alice Oh, Tristan Neumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11737–11808, 2023. doi: 10.52202/075280-0516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2717ad172c5495837582d70a8519abfb-Abstract-Conference.html.

H. Viswanathan and Toby Berger. The quadratic gaussian CEO problem. *IEEE Transactions on Information Theory*, 43(5):1549–1559, September 1997. doi: 10.1109/18.623155.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization. In *Advances in Neural Information Processing Systems*, pp. 10159–10170, December 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d9fbed9da256e344c1fa46bb46c34c5f-Abstract.html.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, NIPS'17, pp. 1508–1518, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 9781510860964. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Feijie Wu, Song Guo, Zhihao Qu, Shiqi He, Ziming Liu, and Jing Gao. Anchor sampling for federated learning with partial client participation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37379–37416. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wu23e.html.

Jihao Xin, Ivan Ilin, Shunkang Zhang, Marco Canini, and Peter Richtárik. Kimad: Adaptive Gradient Compression with Bandwidth Awareness. In *Proceedings of the 4th International Workshop on Distributed Machine Learning*, DistributedML '23, pp. 35–48, New York, NY, USA, December 2023. Association for Computing Machinery. ISBN 9798400704475. doi: 10.1145/3630048.3630184. URL https://doi.org/10.1145/3630048.3630184.

Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2): 1188–1200, 2021. doi: 10.1109/TWC.2020.3031503.

Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated Learning via Over-the-Air Computation. *IEEE Transactions on Wireless Communications*, 19(3):2022–2035, March 2020. doi: 10.1109/TWC.2019.2961673. URL https://doi.org/10.1109/TWC.2019.2961673.

Liping Yi, Wang Gang, and Liu Xiaoguang. QSFL: A two-level uplink communication optimization framework for federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25501–25513. PMLR, 17–23 Jul 2022.

Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, and Hai Li. Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41354–41381. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/zhang23y.html.

Naifu Zhang, Meixia Tao, Jia Wang, and Fan Xu. Fundamental limits of communication efficiency for model aggregation in distributed learning: A rate-distortion approach. *IEEE Transactions on Communications*, 2023b. doi: 10.1109/TCOMM.2022.3224977.

Haoyu Zhao, Boyue Li, Zhize Li, Peter Richtárik, and Yuejie Chi. BEER: fast $\mathcal{O}(1/T)$ rate for decentralized nonconvex optimization with communication compression. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, December 2022. Curran Associates Inc. ISBN 9781713871088. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cd86c6a804d925c4cbc5a7b96843f6d5-Abstract-Conference.html.

Guangxu Zhu, Yuqing Du, Deniz Gündüz, and Kaibin Huang. One-Bit Over-the-Air Aggregation for Communication-Efficient Federated Edge Learning: Design and Convergence Analysis. *IEEE Transactions on Wireless Communications*, 20(3):2120–2135, March 2021. doi: 10.1109/TWC.2020.3039309. URL https://doi.org/10.1109/TWC.2020.3039309.

Alphan Şahin and Rui Yang. A Survey on Over-the-Air Computation. *IEEE Communications Surveys & Tutorials*, 25(3):1877–1908, July 2023. doi: 10.1109/COMST.2023.3264649. URL https://doi.org/10.1109/COMST.2023.3264649.

A Technical Summary of Relevant EF21 Variants

This section reviews the classical EF21 mechanism and the two extensions most closely related to the present work: partial participation (EF21-PP) and bidirectional compression (EF21-BC). The aim is to make three distinctions precise: which states are maintained by the workers and the server, what is communicated in each direction, and how each architectural modification enters the smooth nonconvex convergence guarantee. Throughout, we use the notation of this paper instead of using the notation of the original papers verbatim.

For the distributed objective in (1), gradient descent (GD) updates

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t). \tag{24}$$

A direct distributed implementation of (24) requires every worker to transmit its full d-dimensional local gradient at every iteration.

The classical EF21 method replaces these exact transmissions with recursively maintained gradient trackers generated by a δ-contractive compressor $\mathcal{C}$ (Richtárik et al., 2021). Under the indexing convention of this paper, worker i forms the innovation

$$\mathbf{u}_t^i = \nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i, \tag{25}$$

compresses it as

$$\mathbf{q}_t^i = \mathcal{C}_t^i(\mathbf{u}_t^i), \tag{26}$$

and refreshes its local tracker according to

$$\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i. \tag{27}$$

Thus, only the discrepancy between the current local gradient and the previous local tracker (surrogate gradient) is compressed and the message $\mathbf{q}_t^i$ is sent to the server. Consequently, the server can update the averaged tracker recursively and perform the optimization step:

$$\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n} \sum_{i=1}^{n} \mathbf{q}_t^i, \qquad \mathbf{x}_{t+1} = \mathbf{x}_t - \eta \widehat{\mathbf{g}}_t. \tag{28}$$

Equations (25)–(27) constitute the Markov-compressor construction of Richtárik et al. (2021, Alg. 2). As the tracker follows the local gradient, the input to the compressor is itself recursively corrected rather than accumulated independently across iterations. Indeed, if $\mathbf{e}_t^i = \mathbf{q}_t^i - \mathbf{u}_t^i$ denotes the compression error, then

$$\mathbf{u}_t^i = \nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}_{t-1}) - \mathbf{e}_{t-1}^i, \tag{29}$$

$$\widehat{\mathbf{g}}_t^i = \nabla f_i(\mathbf{x}_t) + \mathbf{e}_t^i. \tag{30}$$

The first identity (29) nicely exposes the error-feedback in $\mathbf{u}_t^i$. The second identity (30) makes the tracking interpretation explicit: the local surrogate equals the current local gradient plus the current compression error. Without the error term in (29), the surrogate gradient $\widehat{\mathbf{g}}_t^i$ would have a history sum of all past compression errors.

The two extensions considered here modify different parts of this architecture. EF21-PP retains the ordinary server transmission but refreshes only the trackers belonging to a sampled worker subset. Its analysis permits arbitrary *proper* sampling, meaning that every worker has a strictly positive marginal participation probability. Without this condition, a worker may never refresh its tracker, which is generally incompatible with a heterogeneous finite-sum objective unless additional relationships among the local functions are imposed (Fatkhullin et al., 2025, Sec. 3, Alg. 3). By contrast, EF21-BC keeps every worker active and introduces a second EF21-type tracking layer on the server-to-worker path. The worker and server compressors may therefore have different contraction parameters (Fatkhullin et al., 2025, Alg. 4).

The algorithms are summarized in Table 2. Consider the standard assumptions that every f_i has an L_i-Lipschitz gradient and that $f^* := \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$. Initialize the local trackers with the exact local gradients so that $\mathbf{g}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_0^i$ and the initial tracking error vanishes.

Table 2: Classical EF21 variants and their smooth nonconvex guarantees in terms of the stationarity target $\mathbb{E}[\|\nabla f(\mathbf{x}_{\text{out}})\|^2] \le \varepsilon$. The bounds suppress numerical constants and assume exact initialization of the local trackers (Richtárik et al., 2021; Fatkhullin et al., 2025).

	Comments	**Theoretical guarantees**
EF21-GD (Richtárik et al., 2021, Alg. 2) **Worker i (for all $i \in [n]$):** $\mathbf{q}_t^i = \mathcal{C}\left(\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i\right),$ $\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i,$ Send $\mathbf{q}_t^i$ to the server. **Server:** $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{q}_t^i$ $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\,\widehat{\mathbf{g}}_t$ Broadcast $\mathbf{x}_{t+1}$ to all workers.	Every worker refreshes its local tracker in every iteration. The tracker $\widehat{\mathbf{g}}_t^i$ is a surrogate for $\nabla f_i(\mathbf{x}_t)$, and only the innovation in (25) is compressed. With the identity compressor, $\widehat{\mathbf{g}}_t = \nabla f(\mathbf{x}_t)$, so the method reduces to distributed GD.	For a δ-contractive worker compressor, $\mathsf{T}_{\text{EF21-GD}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{\delta\varepsilon}\right).$
EF21-PP (Fatkhullin et al., 2025, Alg. 3) **Worker i (for all $i \in [n]$):** $\mathbf{q}_t^i = \begin{cases} \mathcal{C}\left(\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i\right) & : i \in \mathcal{S}_t \\ \mathbf{0} & : i \notin \mathcal{S}_t \end{cases},$ $\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i,$ Workers $i \in \mathcal{S}_t$, send $\mathbf{q}_t^i$ to the server. **Server:** $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i\in\mathcal{S}_t} \mathbf{q}_t^i,$ $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\,\widehat{\mathbf{g}}_t,$ Sample $\mathcal{S}_{t+1}$. Send $\mathbf{x}_{t+1}$ to only $i \in \mathcal{S}_{t+1}$.	The server selects a proper random subset $\mathcal{S}_t \subseteq [n]$ with $\Pr(i \in \mathcal{S}_t) = p_i > 0$. Only selected workers perform gradient computation and uplink communication. Every inactive worker retains its previous tracker, which becomes stale until that worker next participates. Classical EF21-PP specifies the participation law through $\{p_i\}$ and permits arbitrary proper sampling.	For equal marginals $p_i = p$, $\mathsf{T}_{\text{EF21-PP}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{p\delta\varepsilon}\right).$ The expected number of local gradient computations is $p\,\mathsf{T}_{\text{EF21-PP}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{\delta\varepsilon}\right).$ Thus partial participation increases the round complexity by order $1/p$, while reducing the expected per-iteration worker activity by p (Fatkhullin et al., 2025, Cor. 18).
EF21-BC (Fatkhullin et al., 2025, Alg. 4) **Worker i (for all $i \in [n]$):** $\widehat{\widehat{\mathbf{g}}}_{t-1} = \widehat{\widehat{\mathbf{g}}}_{t-2} + \mathbf{q}_{t-1}^s,$ $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta\,\widehat{\widehat{\mathbf{g}}}_{t-1},$ $\mathbf{q}_t^i = \mathcal{C}\left(\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i\right),$ $\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i,$ Send $\mathbf{q}_t^i$ to the server. **Server:** $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{q}_t^i,$ $\mathbf{q}_t^s = \mathcal{C}^s\left(\widehat{\mathbf{g}}_t - \widehat{\widehat{\mathbf{g}}}_{t-1}\right),$ $\widehat{\widehat{\mathbf{g}}}_t = \widehat{\widehat{\mathbf{g}}}_{t-1} + \mathbf{q}_t^s,$ Broadcast $\mathbf{q}_t^s$ to all workers.	Every worker refreshes its local tracker in every iteration. EF21-BC applies the EF21 principle twice: $\widehat{\mathbf{g}}_t^i$ tracks the local gradient, while $\widehat{\widehat{\mathbf{g}}}_t$ tracks the worker-aggregated quantity $\widehat{\mathbf{g}}_t$. Since every worker receives the same server message, all workers maintain the common gradient surrogate needed to reconstruct the next iterate. With the identity server compressor $\mathcal{C}^s$, the method reduces to ordinary EF21.	For worker and server contraction parameters δ and δ_s, respectively, $\mathsf{T}_{\text{EF21-BC}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{\delta\delta_s\varepsilon}\right).$ Every worker computes one full local gradient per round, so the per-worker gradient complexity has the same order. The product $\delta\delta_s$ reflects the two nested contractive tracking mechanisms (Fatkhullin et al., 2025, App. F).

We next state the classical guarantees more precisely. Define

$$\widetilde{L} \coloneqq \left(\frac{1}{n}\sum_{i=1}^{n} L_i^2\right)^{1/2}, \qquad \Delta_0 \coloneqq f(\mathbf{x}_0) - f^*. \tag{31}$$

The original EF21 analysis is expressed in terms of the quadratic mean $\widetilde{L}$. For a δ-contractive compressor, Richtárik et al. (2021, Th. 2) show that a uniformly selected output iterate $\mathbf{x}_{\mathrm{out}} \in \{\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}\}$ satisfies

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\mathrm{out}})\|^2\right] \le \frac{2\Delta_0}{\eta T} + \frac{\mathbb{E}[G_0]}{(1-\sqrt{1-\delta})T}, \tag{32}$$

provided that

$$0 < \eta \le \left(L + \widetilde{L}\frac{1-\delta+\sqrt{1-\delta}}{\delta}\right)^{-1}. \tag{33}$$

Here, G_t is the mean of the squares of the norm of worker-side tracking errors:

$$G_t \coloneqq \frac{1}{n}\sum_{i=1}^{n} \left\|\widehat{\mathbf{g}}_t^i - \nabla f_i(\mathbf{x}_t)\right\|^2. \tag{34}$$

Exact tracker initialization gives $G_0 = 0$, and (32) then yields

$$\mathsf{T}_{\texttt{EF21-GD}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{\delta\varepsilon}\right) \tag{35}$$

to reach $\mathbb{E}[\|\nabla f(\mathbf{x}_{\mathrm{out}})\|^2] \le \varepsilon$. The essential improvement established by Richtárik et al. (2021) is the standard $\mathcal{O}(1/T)$ stationarity rate under smoothness and contractive compression; earlier error-feedback analyses obtained the slower $\mathcal{O}(T^{-2/3})$ rate in this setting and required stronger bounded-gradient assumptions.

Partial participation changes the tracker recursion because a local tracker moves only when its worker is selected. For proper sampling with $p_i = \Pr(i \in \mathcal{S}_t) > 0$, Fatkhullin et al. (2025, Lem. 15) establish a recursion of the form

$$\mathbb{E}[G_{t+1}] \le (1-\theta_p)\mathbb{E}[G_t] + B\,\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right], \tag{36}$$

where $\theta_p > 0$ and $B > 0$ depend on the compressor, the sampling probabilities, and the local smoothness constants. The resulting bound is

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_{\mathrm{out}})\|^2\right] \le \frac{2\Delta_0}{\eta T} + \frac{\mathbb{E}[G_0]}{\theta_p T}, \qquad 0 < \eta \le \left(L + \sqrt{\frac{B}{\theta_p}}\right)^{-1}. \tag{37}$$

For homogeneous marginal probabilities $p_i = p$, the choices in Fatkhullin et al. (2025, Lem. 16) satisfy

$$\theta_p \ge \frac{p\delta}{2}, \qquad \frac{B}{\theta_p} \le \left(\frac{4\widetilde{L}}{p\delta}\right)^2. \tag{38}$$

Consequently,

$$\mathsf{T}_{\texttt{EF21-PP}} = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{p\delta\varepsilon}\right). \tag{39}$$

Because a worker computes a gradient only when selected, its expected number of local-gradient evaluations is pT and therefore remains of order $\mathcal{O}(\widetilde{L}\Delta_0/(\delta\varepsilon))$. The arbitrary-sampling theorem depends on the individual probabilities $\{p_i\}$ (Fatkhullin et al., 2025, Th. 17) and (39) is the homogeneous specialization relevant to a direct comparison with fixed-cardinality participation (Fatkhullin et al., 2025, Cor. 18). Thus classical EF21-PP exchanges additional rounds for fewer worker computations and uplink messages per round.

For EF21-BC, let the server compressor be δ_s-contractive, in addition to the δ-contractive compressor for workers. So, the analysis must control both the worker-tracking error G_t and the additional server-tracking error $\|\widehat{\widehat{\mathbf{g}}}_t - \widehat{\mathbf{g}}_t\|^2$. This two-level decomposition is the proof strategy used by Fatkhullin et al. (2025, App. F). Under exact initialization, Fatkhullin et al. (2025, Cor. 2) obtain

$$\mathsf{T}_{\texttt{EF21-BC}}(\varepsilon) = \mathcal{O}\left(\frac{\widetilde{L}\Delta_0}{\delta\delta_s\varepsilon}\right). \tag{40}$$

Since all workers participate in every round, each worker performs one full local-gradient evaluation per iteration; hence its gradient-computation complexity has the same order as (40). The factor $\delta\delta_s$ records the two nested compression mechanisms: δ controls worker-to-server tracking, while δ_s controls server-to-worker tracking. In particular, $\delta_s = 1$ recovers the uplink-only EF21-GD dependence (35), while $\delta = \delta_s = 1$ reduces both algorithms to full-precision distributed GD.

The quadratic mean $\widetilde{L}$ arises from the analyses in (Richtárik et al., 2021; Fatkhullin et al., 2025); it is not intrinsic to the EF21 update itself. Richtárik et al. (2024) subsequently sharpened this dependence to the arithmetic mean

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i. \tag{41}$$

This distinction matters in the present paper: the results in Section 3 and Appendix C are developed directly in terms of $\bar{L}$ (41) without requiring any methods from Richtárik et al. (2024) through a simpler proof.

The classical variants clarify the communication mechanisms referred as baselines in the main text. In EF21-PP, participation is prescribed externally through a sampling law. In EF21-BC, every worker participates, but the server introduces a second compressor to communicate at finite rate. The sum-rate model studied here is different: the constraint (2) couples worker participation and compression fidelity through one shared resource budget, leading to the discussion in Section 2.

B Extended Literature Review and Positioning

The literature uses "rate," "budget," and "sum rate" for mathematically different resources.

B.1 Source coding and sum-rate formulations

Classical distributed source coding studies separate encoders observing correlated sources. The Slepian–Wolf region characterizes lossless rates through subset inequalities, while lossy multiterminal formulations such as Berger–Tung replace exact recovery by distortion constraints (Slepian & Wolf, 1973; Berger, 1978; Tung, 1978). The Gaussian chief-executive-officer problem is a canonical remote-source instance in which several terminals observe noisy versions of a latent quantity and the decoder reconstructs that quantity from their descriptions (Viswanathan & Berger, 1997). In federated learning, Zhang et al. (2023b) use this viewpoint to derive a sum-rate–distortion limit for model aggregation with correlated worker observations. These are coding theorems about an achievable rate region or the minimum total rate for a target distortion. They do not by themselves specify an iterative optimization algorithm, a contractive tracker, or a per-round active set.

The present paper uses a simpler operational source constraint: independently decodable worker codewords have realized lengths whose per-coordinate sum is at most R in each round. Correlation-aware or distributed codes may achieve a smaller sum rate than separate fixed-rate codebooks, but then the decoder, side information, and rate–contraction proof must be changed accordingly. This is why Appendix E does not assert a universal map from bits to contraction.

B.2 Physical-layer sum rates and resource allocation

In a digital multiple-access channel, feasible user rates belong to a joint capacity region and its sum-rate face depends on channel gains, power, interference, and decoding (Cover & Thomas, 2006). Wireless federated-learning work optimizes learning over such channels or reconstructs compressed

gradients from MIMO observations (Jeon et al., 2022). Over-the-air computation is more different still: workers transmit simultaneously so that the channel superposition approximates the desired aggregate rather than carrying separately decoded codewords (Mohammadi Amiri & Gündüz, 2020; Yang et al., 2020; Zhu et al., 2021; Şahin & Yang, 2023). The number of source bits in (2) is therefore not a wireless sum capacity or an AirComp channel use count.

Systems-oriented federated learning often optimizes client selection jointly with bandwidth, airtime, power, latency, or energy. Long-term and online bandwidth-allocation formulations make client availability a resource decision (Xu & Wang, 2021; Ji et al., 2022); Oort combines statistical utility with system speed (Lai et al., 2021); downlink-aware selection addresses server-to-device scarcity (Mohammadi Amiri et al., 2021); and QSFL combines resource-aware selection with model compression (Yi et al., 2022). These works strongly motivate resource-induced participation, but a physical service allocation does not equal a source-coding contraction unless the channel code and latency model are also specified.

B.3 Compression, error feedback, and tracking

Low-bit and sparse distributed training includes 1-bit SGD, QSGD, TernGrad, signSGD, sparsified SGD, and PowerSGD (Seide et al., 2014; Alistarh et al., 2017; Wen et al., 2017; Bernstein et al., 2018; Stich et al., 2018; Vogels et al., 2019). These works vary substantially in whether the compressor is biased, whether a scale and support are transmitted, and whether the guarantee is stated in coordinates, bits, or a variance parameter. Error feedback corrects accumulated biased-compression error (Karimireddy et al., 2019). EF21 replaces the classical residual with a Markov gradient tracker driven by a contractive compressor (Richtárik et al., 2021); the later arithmetic-mean analysis sharpens its dependence on heterogeneous smoothness constants (Richtárik et al., 2024). 3PC unifies contractive and lazy updates (Richtarik et al., 2022); MARINA uses randomized full refreshes and compressed differences (Gorbunov et al., 2021); and BEER and MURANA give related decentralized and stochastic-operator perspectives (Zhao et al., 2022; Condat & Richtárik, 2022).

Rate-aware compressor design is adjacent but not identical to allocating a hard sum rate. Optimal gradient-compression and information-constrained optimization study the best estimation or optimization accuracy permitted by a communication budget (Albasyoni et al., 2020; Acharya et al., 2021). Fixed-rate differential quantization supplies the covering-type law used in (11) (Lin et al., 2022). Adaptive compression can exploit decoder side information or measured network conditions (Isik et al., 2024; Xin et al., 2023), while correlated quantization couples worker messages to improve total complexity (Panferov et al., 2025). EF21-SR is distinguished by making zero rate a tracker-preserving participation action and enforcing the aggregate source rate pathwise.

B.4 Partial participation and sampling

FedPAQ combines quantization, periodic averaging, and partial participation (Reisizadeh et al., 2020); SCAFFOLD controls client drift (Karimireddy et al., 2020); and FedVARP reduces the variance caused by missing clients (Jhunjhunwala et al., 2022). Clustered, biased, class-balanced, anchor, general, and optimized sampling show that worker identities and marginal probabilities can affect convergence beyond the active fraction alone (Fraboni et al., 2021; Jee Cho et al., 2022; Zhang et al., 2023a; Wu et al., 2023; Fraboni et al., 2023; Chen et al., 2022). Under compression, stale residuals or trackers require explicit treatment, as in biased-compression analysis with partial participation, PP-MARINA, and SA-PEF (Li & Li, 2023; Tyurin & Richtarik, 2023; Redie et al., 2026). EF21-PP is the closest tracker-level baseline: it refreshes a proper sampled subset but takes the sampling law as given rather than deriving it from a pathwise rate budget (Fatkhullin et al., 2025).

B.5 Bidirectional and downlink compression

Bidirectional compression changes the server-to-worker mechanism rather than only the active set. DoubleSqueeze applies error-compensated compression in both directions (Tang et al., 2019); Artemis and MCM analyze bidirectional communication with heterogeneity or a preserved central model (Philippenko & Dieuleveut, 2020; 2021). Within the closest family, EF21-BC applies a second contractive tracker at the server, while EF21-P and 2Direction refine or accelerate bidirectional communication (Fatkhullin et al., 2025; Gruntkowska et al., 2023; Tyurin & Richtárik,

2023). MARINA-P/M3 use correlated downlink compression (Gruntkowska et al., 2024); DoCoFL and GlueFL focus on cross-device downlink cost (Dorfman et al., 2023; He et al., 2023a); and LoCoDL, BiCoLoR, and LIEC-SGD combine local training or alternative compensation with two-way compression (Condat et al., 2025; 2026; Cheng et al., 2025). These methods remain appropriate when the downlink itself is the constrained physical resource. The EF21-SR rebroadcast argument instead asks whether a received worker codeword bundle already fits the declared common source-payload budget.

B.6 Per-round and end-to-end communication

The final literature distinction concerns the time horizon. Communication lower bounds and near-optimal compressed methods count all messages required to reach a target (Arjevani & Shamir, 2015; Korhonen & Alistarh, 2021; Huang et al., 2022). Work asking when unbiased compression actually saves communication makes the same point directly: smaller messages may be offset by more iterations (He et al., 2023b). Accordingly, (2) is a per-round feasibility constraint, while the fixed-accuracy comparison following Table 1 is end-to-end. Both are needed to state a communication advantage.

C Convergence Proofs

C.1 Probability space and filtration

The algorithms use exact local gradients, so only the allocation, worker-compressor, and server-compressor randomness is needed. Let $(\mathrm{A}, \Sigma^a, \mathbb{P}^a)$ carry one allocation seed, let $(\mathrm{H}_i^w, \Sigma_i^w, \mathbb{P}_i^w)$ carry one worker-compressor seed for worker i, and let $(\mathrm{H}^s, \Sigma^s, \mathbb{P}^s)$ carry one server-compressor seed. For EF21-SR, H^s is a singleton because the server applies no randomized compressor. Define

$$(\mathrm{H}^w, \Sigma^w, \mathbb{P}^w) \coloneqq \left(\prod_{i=1}^n \mathrm{H}_i^w, \bigotimes_{i=1}^n \Sigma_i^w, \bigotimes_{i=1}^n \mathbb{P}_i^w \right), \tag{42}$$

$$(\Theta, \Sigma, \mathbb{P}_\Theta) \coloneqq (\mathrm{A} \times \mathrm{H}^w \times \mathrm{H}^s, \Sigma^a \otimes \Sigma^w \otimes \Sigma^s, \mathbb{P}^a \otimes \mathbb{P}^w \otimes \mathbb{P}^s). \tag{43}$$

The global probability space is

$$(\Omega, \mathcal{F}, \mathbb{P}) \coloneqq \left(\Theta^{\mathbb{N}_0}, \Sigma^{\otimes \mathbb{N}_0}, \mathbb{P}_\Theta^{\otimes \mathbb{N}_0} \right). \tag{44}$$

For $\omega \in \Omega$, write the round-t coordinate as

$$\theta_t(\omega) = \big(a_t(\omega), h_t^w(\omega), h_t^s(\omega)\big), \qquad h_t^w = (h_{t,1}^w, \dots, h_{t,n}^w). \tag{45}$$

The product construction makes these reference seeds independent across rounds and makes the three factors in (43) independent. This statement concerns only the seeds. The active indicators, rates, compressor outputs, and compression errors may be dependent across workers and rounds because they are measurable functions of the common allocation seed and the algorithmic history.

The initial point and all initial trackers are deterministic, and $\mathcal{F}_{-1} \coloneqq \{\varnothing, \Omega\}$. For every $t \geq 0$, define

$$\mathcal{G}_t \coloneqq \mathcal{F}_{t-1} \vee \sigma(a_t), \tag{46}$$

$$\mathcal{H}_t \coloneqq \mathcal{G}_t \vee \sigma(h_t^w), \tag{47}$$

$$\mathcal{F}_t \coloneqq \mathcal{H}_t \vee \sigma(h_t^s). \tag{48}$$

Thus $\mathcal{F}_{t-1} \subseteq \mathcal{G}_t \subseteq \mathcal{H}_t \subseteq \mathcal{F}_t$. The sigma-field $\mathcal{F}_{t-1}$ is the complete history before round t; $\mathcal{G}_t$ additionally contains the realized allocation but not the current worker-compressor seeds; $\mathcal{H}_t$ additionally contains the current worker-compressor outputs; and $\mathcal{F}_t$ additionally contains the current server-compressor output. Since the gradients are exact, $\mathbf{x}_t$, $\widehat{\mathbf{g}}_{t-1}^i$, and $\mathbf{u}_t^i$ are $\mathcal{F}_{t-1}$-measurable. The variables I_t^i, R_t^i, and δ_t^i are $\mathcal{G}_t$-measurable; $\widehat{\mathbf{g}}_t^i$ and $\widehat{\mathbf{g}}_t$ are $\mathcal{H}_t$-measurable. For EF21-BC, the server input is $\mathcal{H}_t$-measurable and the server output and $\mathbf{x}_{t+1}$ are $\mathcal{F}_t$-measurable. For EF21-SR, $\mathcal{F}_t = \mathcal{H}_t$, so $\widehat{\mathbf{g}}_t$ and $\mathbf{x}_{t+1}$ are already $\mathcal{H}_t$-measurable.

Conditional on $\mathcal{G}_t$, the worker inputs, realized allocation, and compressor laws are fixed, while the worker-compressor seeds are fresh. Assumption *(A2)* is therefore exactly

$$\mathbb{E}\big[\|\mathbf{e}_t^i\|^2 \,\big|\, \mathcal{G}_t\big] = \mathbb{E}\big[\|\mathcal{C}_t^i(\mathbf{u}_t^i) - \mathbf{u}_t^i\|^2 \,\big|\, \mathcal{G}_t\big] \leq (1 - \delta_t^i)\|\mathbf{u}_t^i\|^2. \tag{49}$$

No unbiasedness and no conditional independence among the realized worker errors are assumed. For the fixed-rate quantizer, the post-allocation value is

$$\delta_t^i = I_t^i \left(1 - \rho_z^2 2^{-2R_t^i}\right). \tag{50}$$

Thus it is zero on an inactive realization and equals the active compressor contraction otherwise. If the active set is random, quantities such as $\mathbb{E}[\delta_t^i \mid \mathcal{F}_{t-1}]$ may be used later in a proof, but that expectation is not the definition of δ_t^i.

For EF21-BC, conditional on $\mathcal{H}_t$ the server input and compressor law are fixed and h_t^s is fresh. The server contraction assumption is

$$\mathbb{E}\left[\|\mathcal{C}_t^s(\mathbf{v}) - \mathbf{v}\|^2 \,\middle|\, \mathcal{H}_t\right] \le (1-\delta_s)\|\mathbf{v}\|^2, \qquad \delta_s \in (0,1]. \tag{51}$$

A deterministic allocation or compressor is represented by a measurable rule that ignores the corresponding coordinate variable. All expectations below are with respect to $\mathbb{P}$.

C.2 Convergence proof for EF21-SR

The proof is carried as follows: first obtain a conditional descent inequality, next propagate the current error envelope compression certificate with the actual δ_t^i, then telescope a Lyapunov inequality.

For the proof, define the scalar certificate

$$E_t := \frac{1}{n}\sum_{i=1}^n \sqrt{1-\delta_t^i}\,\left\|\mathbf{u}_t^i\right\| \ge \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}\left[\left\|\mathbf{e}_t^i\right\|^2 \mid \mathcal{G}_t\right]\right)^{1/2} \tag{52}$$

$$\ge \left(\mathbb{E}\left[\left\|\widehat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2 \mid \mathcal{G}_t\right]\right)^{1/2}. \tag{53}$$

which envelopes the compression errors.

Proof of Theorem 3.1. The proof has three steps.

Step 1: Conditional descent. Using L-smoothness of f, we obtain

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{\eta}{2}\left\|\widehat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2. \tag{54}$$

Taking expectation in (54) conditional on $\mathcal{G}_t$ and using (53) proves

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) \mid \mathcal{G}_t\right] \le f(\mathbf{x}_t) - \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \mid \mathcal{G}_t\right] + \frac{\eta}{2}E_t^2. \tag{55}$$

Keeping (55) conditional is important: E_t contains the actual current values of δ_t^i and $\mathbf{u}_t^i$, and is therefore the compressor-dependent term that may be optimized using the information available at round t.

Step 2: Recursion for the compression certificate. Using $\widehat{\mathbf{g}}_t^i = \nabla f_i(\mathbf{x}_t) + \mathbf{e}_t^i$, the next compressor input satisfies

$$\begin{aligned}\|\mathbf{u}_{t+1}^i\| &= \|\nabla f_i(\mathbf{x}_{t+1}) - \widehat{\mathbf{g}}_t^i\| \\ &\le \|\nabla f_i(\mathbf{x}_{t+1}) - \nabla f_i(\mathbf{x}_t)\| + \left\|\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_t^i\right\| \\ &\le L_i\|\mathbf{x}_{t+1} - \mathbf{x}_t\| + \left\|\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_t^i\right\|.\end{aligned} \tag{56}$$

It follows pathwise that

$$\begin{aligned}E_{t+1} &= \frac{1}{n}\sum_{i=1}^n \sqrt{1-\delta_{t+1}^i}\,\|\mathbf{u}_{t+1}^i\| \\ &\le \frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_{t+1}^i}\,\|\mathbf{x}_{t+1} - \mathbf{x}_t\| + \frac{1}{n}\sum_{i=1}^n \sqrt{1-\delta_{t+1}^i}\,\|\mathbf{e}_t^i\|.\end{aligned} \tag{57}$$

The quantities δ_{t+1}^i, $\mathbf{x}_{t+1}$, and $\mathbf{e}_t^i$ are $\mathcal{G}_{t+1}$-measurable. Since δ_{t+1}^i may be selected adaptively from this information, the factors $\sqrt{1-\delta_{t+1}^i}$ cannot in general be separated from the random variables multiplying them. Using their round-$(t+1)$ essential suprema in (57), we obtain

$$E_{t+1} \le \operatorname{ess\,sup}\left\{\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_{t+1}^i}\right\}\|\mathbf{x}_{t+1}-\mathbf{x}_t\| + \operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_{t+1}^i}\right\}\frac{1}{n}\sum_{i=1}^n\|\mathbf{e}_t^i\|. \tag{58}$$

For every $\varepsilon>0$, Young's inequality therefore gives

$$E_{t+1}^2 \le (1+\varepsilon^{-1})\left(\operatorname{ess\,sup}\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_{t+1}^i}\right)^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + (1+\varepsilon)\left(\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_{t+1}^i}\right\}\right)^2\left(\frac{1}{n}\sum_{i=1}^n\|\mathbf{e}_t^i\|\right)^2. \tag{59}$$

To control the second term by the preceding certificate E_t, the same conditional Minkowski argument used in (53), followed by total expectation, yields

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n\|\mathbf{e}_t^i\|\right)^2\right] \le \mathbb{E}[E_t^2]. \tag{60}$$

Consequently,

$$\mathbb{E}[E_{t+1}^2] \le (1+\varepsilon^{-1})\left(\operatorname{ess\,sup}\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_{t+1}^i}\right)^2\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] + (1+\varepsilon)\left(\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_{t+1}^i}\right\}\right)^2\mathbb{E}[E_t^2]. \tag{61}$$

Step 3: Lyapunov descent. Define

$$\Psi_t \coloneqq \mathbb{E}\left[f(\mathbf{x}_t)-f^*\right]+c\,\mathbb{E}[E_t^2]. \tag{62}$$

Taking total expectation of (55) and combining with (61) gives

$$\Psi_{t+1} \le \Psi_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] - \left(\frac{1-L\eta}{2\eta} - c(1+\varepsilon^{-1})\left(\operatorname{ess\,sup}\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_{t+1}^i}\right)^2\right)\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] - \left(c-\frac{\eta}{2}-c(1+\varepsilon)\left(\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_{t+1}^i}\right\}\right)^2\right)\mathbb{E}[E_t^2]. \tag{63}$$

The coefficient of $\mathbb{E}[E_t^2]$ in (63) is nonpositive for the choice

$$c \coloneqq \frac{\eta}{2}\left(1-(1+\epsilon)\left(\sup_{\tau\ge1}\operatorname{ess\,sup}\left\{\max_i\sqrt{1-\delta_\tau^i}\right\}\right)^2\right)^{-1}, \tag{64}$$

provided that the denominator is positive.

The displacement term in (63) is nonpositive whenever

$$\frac{1}{\eta}\left(\frac{1}{\eta}-L\right) \ge \frac{\left(\sup_{\tau\ge1}\operatorname{ess\,sup}\left\{\frac{1}{n}\sum_{i=1}^n L_i\sqrt{1-\delta_\tau^i}\right\}\right)^2(1+\varepsilon^{-1})}{1-(1+\varepsilon)\left(\sup_{\tau\ge1}\operatorname{ess\,sup}\left\{\max_{1\le i\le n}\sqrt{1-\delta_\tau^i}\right\}\right)^2}. \tag{65}$$

The multiplicative factor on the right-hand side of (65) is minimized by

$$\varepsilon = \frac{1}{\sup_{\tau \geq 1} \operatorname{ess\,sup} \left\{ \max_{1 \leq i \leq n} \sqrt{1 - \delta_\tau^i} \right\}} - 1, \tag{66}$$

when the denominator in (66) is nonzero; the case in which this essential supremum is zero follows by continuity. Therefore,

$$\Psi_{t+1} \leq \Psi_t - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \right]. \tag{67}$$

Summing from $t = 0$ to $T - 1$ gives

$$\Psi_T \leq \Psi_0 - \frac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \right]. \tag{68}$$

Since f^* is a global lower bound and $c\,\mathbb{E}[E_T^2] \geq 0$, we have $\Psi_T \geq 0$. Hence,

$$\frac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \right] \leq \Psi_0. \tag{69}$$

Therefore,

$$\min_{0 \leq t < T} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \right] \leq \frac{2\Psi_0}{\eta T}. \tag{70}$$

This completes the proof. □

For the uniform-contraction specialization, suppose that $\delta_t^i \geq \delta > 0$ almost surely for every t and i. Then

$$\sup_{\tau \geq 1} \operatorname{ess\,sup} \left\{ \frac{1}{n} \sum_{i=1}^{n} L_i \sqrt{1 - \delta_\tau^i} \right\} \leq \bar{L} \sqrt{1 - \delta},$$
$$\sup_{\tau \geq 1} \operatorname{ess\,sup} \left\{ \max_{1 \leq i \leq n} \sqrt{1 - \delta_\tau^i} \right\} \leq \sqrt{1 - \delta}. \tag{71}$$

Hence, the stepsize condition is satisfied whenever

$$\frac{1}{\eta} \left(\frac{1}{\eta} - L \right) \geq \left(\frac{\bar{L}\sqrt{1 - \delta}}{1 - \sqrt{1 - \delta}} \right)^2. \tag{72}$$

Equivalently, it is sufficient to choose

$$0 < \eta \leq \frac{2}{L + \sqrt{L^2 + 4 \left(\frac{\bar{L}\sqrt{1 - \delta}}{1 - \sqrt{1 - \delta}} \right)^2}}. \tag{73}$$

Now, let

$$I_t^i \coloneqq \mathbb{I}\left\{ R_t^i > 0 \right\}. \tag{74}$$

If $I_t^i = 0$, the tracker is unchanged, so its error is the complete pre-refresh input. If $I_t^i = 1$, fixed-rate model bounds the squared error by $\rho_z^2 2^{-2R_t^i}$ times the squared input norm. Conditioning on $\mathcal{G}_t$ therefore gives

$$\mathbb{E}\left[\left\| \widehat{\mathbf{g}}_t^i - \nabla f_i(\mathbf{x}_t) \right\|^2 \mid \mathcal{G}_t \right] \leq \left[1 - \mathbb{E}\left[I_t^i \left(1 - \rho_z^2 2^{-2R_t^i} \right) \mid \mathcal{G}_t \right] \right] \left\| \mathbf{u}_t^i \right\|^2. \tag{75}$$

Thus,

$$\delta_t^i = \mathbb{E}\left[I_t^i \left(1 - \rho_z^2 2^{-2R_t^i} \right) \mid \mathcal{G}_t \right]. \tag{76}$$

If $\delta_t^i \geq \underline{\delta}$, then

$$\max_{1\leq i\leq n} \sqrt{1-\delta_t^i} \leq \sqrt{1-\underline{\delta}},$$
$$\frac{1}{n}\sum_{i=1}^{n} L_i \sqrt{1-\delta_t^i} \leq \bar{L}\sqrt{1-\underline{\delta}}. \tag{77}$$

Substituting (77) into the general stepsize condition completes the analysis.

C.3 Convergence proof for EF21-PP

The result for EF21-PP is now a direct specialization of a calculation similar to that in (75). If every worker is selected with conditional probability p and the active-worker compressor has contraction parameter δ, then in conditional expectation $\delta_t^i = p\delta$, $t \geq 0$, $i \in [n]$. Indeed, an inactive worker contributes no contraction, whereas an active worker contributes δ. Therefore,

$$E_t = \frac{1}{n}\sqrt{1-p\delta}\sum_{i=1}^{n} \left\|\mathbf{u}_t^i\right\|, \tag{78}$$

and (72) with δ replaced by $p\delta$ becomes

$$\frac{1}{\eta}\left(\frac{1}{\eta} - L\right) \geq \left(\frac{\bar{L}\sqrt{1-p\delta}}{1-\sqrt{1-p\delta}}\right)^2. \tag{79}$$

Since

$$1-\sqrt{1-p\delta} \geq \frac{p\delta}{2}, \tag{80}$$

the simpler choice

$$0 < \eta \leq \left(L + \frac{2\bar{L}}{p\delta}\right)^{-1} \tag{81}$$

is sufficient. Consequently,

$$\min_{0\leq t<T} \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_t)\right\|^2\right] \leq \frac{1}{T}\left[\frac{2\left(f(\mathbf{x}_0)-f^*\right)}{\eta} + \frac{\mathbb{E}\left[E_0^2\right]}{1-\sqrt{1-p\delta}}\right]. \tag{82}$$

Under the exact-tracker initialization

$$\widehat{\mathbf{g}}_{-1}^i = \nabla f_i(\mathbf{x}_0), \qquad i \in [n], \tag{83}$$

we have $E_0 = 0$. Thus, it is sufficient to take

$$T = \mathcal{O}\left(\frac{\left(L + \bar{L}/(p\delta)\right)\left(f(\mathbf{x}_0)-f^*\right)}{\varepsilon}\right). \tag{84}$$

Since $L \leq \bar{L}$, the communication-round and expected per-worker gradient complexities are

$$T = \mathcal{O}\left(\frac{\bar{L}\left(f(\mathbf{x}_0)-f^*\right)}{p\delta\,\varepsilon}\right), \tag{85}$$
$$pT = \mathcal{O}\left(\frac{\bar{L}\left(f(\mathbf{x}_0)-f^*\right)}{\delta\,\varepsilon}\right). \tag{86}$$

C.4 Convergence proof for EF21-BC

We provide an alternate and simpler proof of convergence of EF21-BC in terms of the arithmetic mean $\bar{L}$ instead of the quadratic mean $\widetilde{L}$. We use the same proof pattern as before, but the server compressor adds one genuine error level. The proof therefore needs two scalar root-mean-square

envelopes: one for the local trackers and one for the server surrogate. No additional worker or server vector state is introduced.

Let $\delta \in (0,1]$ and $\delta_s \in (0,1]$ be the worker-side and server-side contraction parameters, respectively. Define

$$E_t \coloneqq \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-\delta_t^i}\,\left\|\mathbf{u}_t^i\right\| \geq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\|\mathbf{e}_t^i\|^2 \mid \mathcal{G}_t\right]\right)^{1/2} \tag{87}$$

$$E_t^s \coloneqq \left(\mathbb{E}\left[\left\|\widehat{\widehat{\mathbf{g}}}_t - \widehat{\mathbf{g}}_t\right\|^2\right]\right)^{1/2}. \tag{88}$$

Theorem C.1 (Smooth nonconvex convergence of EF21-BC)**.** *Suppose* (A1) *holds, the worker compressors satisfy* (A2) *with the common contraction δ, and the server compressor satisfies* (51) *with contraction δ_s. Define*

$$c_s \coloneqq \frac{\eta}{1-\sqrt{1-\delta_s}}, \tag{89}$$

$$c \coloneqq \frac{\eta}{1-\sqrt{1-\delta}}\left[1+\frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\sqrt{1-\delta}\left(1-\sqrt{1-\delta_s}\right)^2}\right]. \tag{90}$$

If

$$\frac{1}{\eta}\left(\frac{1}{\eta}-L\right) \geq \frac{2\bar{L}^2}{\left(1-\sqrt{1-\delta}\right)^2}\left[1-\delta+\frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\left(1-\sqrt{1-\delta_s}\right)^2}\right], \tag{91}$$

then, for every $T \geq 1$,

$$\min_{0\leq t<T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leq \frac{2}{\eta T}\left(\Delta_0 + c\,\mathbb{E}\left[E_0^2\right] + c_s\left(E_0^s\right)^2\right). \tag{92}$$

With exact initialization, both E_0 and E_0^s vanish, and (92) *reduces to $\frac{2\Delta_0}{\eta T}$.*

Proof for EF21-BC. The proof has three steps.

Step 1: Descent with bidirectional compression. Let

$$\mathbf{e}_t^i \coloneqq \widehat{\mathbf{g}}_t^i - \nabla f_i(\mathbf{x}_t), \qquad \mathbf{e}_t^s \coloneqq \widehat{\widehat{\mathbf{g}}}_t - \widehat{\mathbf{g}}_t. \tag{93}$$

Conditionally on $\mathcal{G}_t$, Minkowski's inequality and worker-side contractivity give

$$\left(\mathbb{E}\left[\|\widehat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\|^2 \mid \mathcal{G}_t\right]\right)^{1/2} \leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\left\|\mathbf{e}_t^i\right\|^2 \mid \mathcal{G}_t\right]\right)^{1/2} \leq E_t. \tag{94}$$

Since

$$\widehat{\widehat{\mathbf{g}}}_t - \nabla f(\mathbf{x}_t) = \mathbf{e}_t^s + \widehat{\mathbf{g}}_t - \nabla f(\mathbf{x}_t), \tag{95}$$

the smoothness identity used in (54), followed by $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, yields

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] \leq \mathbb{E}\left[f(\mathbf{x}_t)\right] &- \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\ &- \left(\frac{1}{2\eta}-\frac{L}{2}\right)\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right] + \eta\mathbb{E}\left[E_t^2\right] + \eta\left(E_t^s\right)^2.\end{aligned} \tag{96}$$

Step 2: Recursions for the compression certificates. As in (56),

$$\left\|\mathbf{u}_{t+1}^i\right\| \leq L_i\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\| + \left\|\mathbf{e}_t^i\right\|. \tag{97}$$

Using (60) and Young's inequality with parameter $1/\sqrt{1-\delta}-1$ gives

$$\mathbb{E}\left[E_{t+1}^2\right] \leq \sqrt{1-\delta}\,\mathbb{E}\left[E_t^2\right] + \frac{1-\delta}{1-\sqrt{1-\delta}}\bar{L}^2\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right], \tag{98}$$

Moreover,

$$\widehat{\mathbf{g}}_{t+1} - \widehat{\mathbf{g}}_t = \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}_{t+1}^i\left(\mathbf{u}_{t+1}^i\right). \tag{99}$$

Minkowski's inequality, worker-side contractivity, and (97) imply

$$\begin{aligned}&\left(\mathbb{E}\left[\|\widehat{\mathbf{g}}_{t+1} - \widehat{\mathbf{g}}_t\|^2\right]\right)^{1/2} \\ &\quad\le \left(1+\sqrt{1-\delta}\right)\left[\bar{L}\left(\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right]\right)^{1/2} + \left(\mathbb{E}\left[E_t^2\right]\right)^{1/2}\right].\end{aligned} \tag{100}$$

The next server-side compressor input satisfies

$$\widehat{\mathbf{g}}_{t+1} - \widehat{\widehat{\mathbf{g}}}_t = \widehat{\mathbf{g}}_{t+1} - \widehat{\mathbf{g}}_t - \mathbf{e}_t^s. \tag{101}$$

Applying server-side contractivity to (101) and Young's inequality with parameters $1/\sqrt{1-\delta_s}-1$ and $(1-\sqrt{1-\delta})/\sqrt{1-\delta}$, and then using (100), gives

$$\begin{aligned}\left(E_{t+1}^s\right)^2 &\le \sqrt{1-\delta_s}\left(E_t^s\right)^2 + \frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\sqrt{1-\delta}\left(1-\sqrt{1-\delta_s}\right)}\mathbb{E}\left[E_t^2\right] \\ &\quad + \frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\left(1-\sqrt{1-\delta_s}\right)\left(1-\sqrt{1-\delta}\right)}\bar{L}^2\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right].\end{aligned} \tag{102}$$

Step 3: Lyapunov descent. Define

$$c_s \coloneqq \frac{\eta}{1-\sqrt{1-\delta_s}}, \tag{103}$$

$$c \coloneqq \frac{\eta}{1-\sqrt{1-\delta}}\left[1 + \frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\sqrt{1-\delta}\left(1-\sqrt{1-\delta_s}\right)^2}\right], \tag{104}$$

$$\Psi_t \coloneqq \mathbb{E}\left[f(\mathbf{x}_t) - f^*\right] + c\,\mathbb{E}\left[E_t^2\right] + c_s\left(E_t^s\right)^2. \tag{105}$$

Combining (96), (98), and (102) gives

$$\begin{aligned}\Psi_{t+1} &\le \Psi_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\ &\quad - \left(\frac{1-\eta L}{2\eta} - \frac{\eta\bar{L}^2}{\left(1-\sqrt{1-\delta}\right)^2}\left(1-\delta+\frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\left(1-\sqrt{1-\delta_s}\right)^2}\right)\right)\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right].\end{aligned} \tag{106}$$

Hence, if

$$\frac{1}{\eta}\left(\frac{1}{\eta} - L\right) \ge \frac{2\bar{L}^2}{\left(1-\sqrt{1-\delta}\right)^2}\left[1-\delta+\frac{(1-\delta_s)\left(1+\sqrt{1-\delta}\right)^2}{\left(1-\sqrt{1-\delta_s}\right)^2}\right], \tag{107}$$

then (106) reduces to

$$\Psi_{t+1} \le \Psi_t - \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right]. \tag{108}$$

Summing (108) from $t=0$ to $T-1$ and using $\Psi_T \ge 0$ gives

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \le \Psi_0. \tag{109}$$

Therefore,

$$\begin{aligned}\min_{0\le t<T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] &\le \frac{2\Psi_0}{\eta T} \\ &= \frac{2}{\eta T}\left(f(\mathbf{x}_0) - f^* + c\,\mathbb{E}\left[E_0^2\right] + c_s\left(E_0^s\right)^2\right).\end{aligned} \tag{110}$$

If $\delta = 1$ or $\delta_s = 1$, the corresponding compression error vanishes, and the same conclusion follows after removing that certificate term.

This completes the proof. □

D HETEROGENEOUS RATE ALLOCATION FOR SR IMPLEMENTATION

This appendix develops the rate-allocation rule induced by the sum-rate constraint in (2) and continues the discussion in Section 2. Because a worker may receive no communication resource in a given round, its admissible rate is either zero or at least the minimum rate required for a strictly contractive transmission. Thus,

$$R_t^i \in \{0\} \cup [R_c, \infty), \qquad \sum_{i=1}^{n} R_t^i \le R. \tag{111}$$

Here, $R_t^i = 0$ means that worker i transmits no vector and does not refresh its local tracker. For the fixed-rate quantizer, and for integer rates with covering efficiency $\rho_z \ge 1$, the choice $R_c := \lfloor \log_2 \rho_z \rfloor + 1$ ensures strict contractivity. The active set is therefore precisely the support of the rate vector:

$$\mathcal{S}_t := \{i \in [n] : R_t^i > 0\}. \tag{112}$$

Since every active worker consumes at least R_c units of rate, (111) implies

$$|\mathcal{S}_t| \le m, \qquad m := \min\left\{n, \left\lfloor \frac{R}{R_c} \right\rfloor\right\}. \tag{113}$$

Hence partial participation is not imposed independently: it is a direct consequence of the shared rate budget whenever $R < nR_c$.

At round t, let r_t^i be an envelope for the norm of worker i's innovation:

$$r_t^i \ge \|\mathbf{u}_t^i\| = \left\|\nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i\right\|. \tag{114}$$

The coefficient r_t^i measures the current compression difficulty of worker i for allocation purposes. In the fixed-rate quantizer, assigning zero rate leaves the entire innovation uncommunicated, while assigning a positive rate R_t^i bounds the residual by $r_t^i \rho_z 2^{-R_t^i}$. The joint participation-and-rate decision is therefore the mixed discrete–continuous problem

$$\underset{\{R_t^i\}_{i=1}^n}{\text{minimize}} \quad \frac{1}{n}\left(\sum_{i:R_t^i=0} r_t^i + \sum_{i:R_t^i>0} r_t^i \rho_z 2^{-R_t^i}\right) \tag{115a}$$

$$\text{subject to} \qquad R_t^i \in \{0\} \cup [R_c, \infty), \quad i = 1, \dots, n, \tag{115b}$$

$$\sum_{i=1}^{n} R_t^i \le R \tag{115c}$$

The support of an optimizer determines which workers participate, while its positive entries determine the fidelity of their compressed messages. The objective is a one-round upper bound on the aggregate root-mean-square compression error; it is not an additional communication constraint.

The *static realization* in Algorithm 1 is the simplest feasible realization and deliberately avoids solving (115). It samples an active set of fixed cardinality $s \le m$ and divides the entire budget uniformly:

$$R_t^i = \begin{cases} R/s, & i \in \mathcal{S}_t, \\ 0, & i \notin \mathcal{S}_t. \end{cases} \tag{116}$$

For the fixed-rate quantizer, the realized post-allocation contraction is

$$\delta_t^i = \begin{cases} 1 - \rho_z^2 2^{-2R/s}, & i \in \mathcal{S}_t, \\ 0, & i \notin \mathcal{S}_t. \end{cases} \tag{117}$$

Thus δ_t^i is zero on an inactive realization; it is not defined as a participation average. If $\mathcal{S}_t$ is sampled uniformly without replacement, independently of the preceding history, then worker i is selected with probability s/n, and conditioning before the fresh active-set draw gives

$$\mathbb{E}\left[\delta_t^i \,\middle|\, \mathcal{F}_{t-1}\right] = \frac{s}{n}\left(1 - \rho_z^2 2^{-2R/s}\right). \tag{118}$$

Algorithm 2: `EF21-SR`: heterogeneous sum-rate allocation

Input: stepsize η, sum-rate budget R, and a common allocation rule;
Initialize: $\widehat{\mathbf{g}}_{-1}^i = \mathbf{0}$ for all i, $\widehat{\mathbf{g}}_{-1} = \mathbf{0}$, starting point $\mathbf{x}_0$, and $r_0^i = \|\nabla f_i(\mathbf{x}_0)\|$;
for $t = 0, 1, \ldots, T-1$ **do**
  Solve (115) using the common envelopes $\{r_t^i\}_{i=1}^n$ (119) and set $\mathcal{S}_t = \{i : R_t^i > 0\}$;
  `// Exact alternative:  r_t^i = ||u_t^i|| after a scalar-report phase.`
  Worker (parallel for every $i \in [n]$**)**
    $\widehat{\mathbf{g}}_{t-1} = \widehat{\mathbf{g}}_{t-2} + \frac{1}{n}\sum_{i\in\mathcal{S}_{t-1}} \mathbf{q}_{t-1}^i$;
    $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta\, \widehat{\mathbf{g}}_{t-1}$;
    $\mathbf{u}_t^i = \nabla f_i(\mathbf{x}_t) - \widehat{\mathbf{g}}_{t-1}^i$;
    $\mathbf{q}_t^i = \begin{cases} \mathcal{C}\big(\mathbf{u}_t^i\big) & : i \in \mathcal{S}_t \\ \mathbf{0} & : i \notin \mathcal{S}_t \end{cases}$ $\widehat{\mathbf{g}}_t^i = \widehat{\mathbf{g}}_{t-1}^i + \mathbf{q}_t^i$;
    Workers $i \in \mathcal{S}_t$, send $\mathbf{q}_t^i$ to the server.
  Server
    $\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n}\sum_{i\in\mathcal{S}_t} \mathbf{q}_t^i$;
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\, \widehat{\mathbf{g}}_t$; `// store locally`
    Broadcast the bundle $\{\mathbf{q}_t^i : i \in \mathcal{S}_t\}$ to all workers ; `// broadcast rate` $\leq R$
  `// All workers and server decode the common codeword bundle, reconstruct` $\widehat{\mathbf{g}}_t$ `and` $\mathbf{x}_{t+1}$`, and update` $\{r_{t+1}^i\}_{i=1}^n$ `using (119).`

Equations (117) and (118) describe different stages of the randomness and are therefore compatible: the former is the realized contraction used by the allocation problem, whereas the latter is the conditional average used in the static convergence proof. The factor in (118) also exposes the rate–participation tradeoff: increasing s refreshes workers more frequently but reduces the rate, and therefore the contraction quality, of every active message.

The *adaptive realization* in Algorithm 2 solves (115) with heterogeneous positive rates. An exact-current policy may substitute $r_t^i = \|\mathbf{u}_t^i\|$. Operationally, however, doing so requires every worker to evaluate its current local gradient before allocation and to report a scalar score unless the server can infer the innovation norms. Exact-current allocation therefore adds a control phase and does not save gradient evaluations at workers that subsequently receive zero rate.

The one-call implementation avoids this same-round score exchange by maintaining predictable envelopes that are known before round t. Smoothness of f_i and the `EF21` tracker update give the recursion

$$r_{t+1}^i = L_i\|\mathbf{x}_{t+1} - \mathbf{x}_t\| + \begin{cases} r_t^i \rho_z 2^{-R_t^i}, & R_t^i > 0, \\ r_t^i, & R_t^i = 0, \end{cases} \tag{119}$$

which preserves $r_{t+1}^i \geq \|\mathbf{u}_{t+1}^i\|$ under the fixed-rate error bound. For an active worker, the second term bounds its post-compression tracking error; for an inactive worker, it bounds the discrepancy caused by retaining a stale tracker. When the constants L_i, the initial envelopes, and the allocation rule are shared, (119) depends only on common iterates, earlier envelopes, and assigned rates. Every worker can then reconstruct the same coefficients and, with common deterministic tie-breaking, the same allocation. No extra communication is required. The static realization removes the online optimization as well, at the cost of ignoring trajectory-dependent differences among workers.

For completeness, we next derive the closed-form allocation associated with the fixed-rate quantizer. This derivation is compressor-specific; Algorithm 2 itself requires only a common rule for solving the appropriate version of (115). Conditional on a fixed nonempty active set $\mathcal{S}_t$, the inactive contribution

in (115) is constant, and the positive-rate subproblem becomes

$$\begin{aligned} \underset{\{R_t^i\}_{i\in\mathcal{S}_t}}{\text{minimize}} \quad & \sum_{i\in\mathcal{S}_t} r_t^i \rho_z 2^{-R_t^i} && \text{(120a)} \\ \text{subject to} \quad & R_t^i \geq R_c, \quad i \in \mathcal{S}_t, && \text{(120b)} \\ & \sum_{i\in\mathcal{S}_t} R_t^i \leq R && \text{(120c)} \end{aligned}$$

Whenever $R \geq |\mathcal{S}_t| R_c$, the continuous-rate relaxation admits the floor-constrained reverse-waterfilling solution

$$R_t^{i,*} = R_c + \left[\log_2\left(\frac{r_t^i \rho_z 2^{-R_c}}{\theta_t^*}\right)\right]_+ = \max\left\{R_c, \log_2\left(\frac{r_t^i \rho_z}{\theta_t^*}\right)\right\}, \qquad i \in \mathcal{S}_t, \tag{121}$$

where $\theta_t^* > 0$ is selected so that

$$\sum_{i\in\mathcal{S}_t} \left[\log_2\left(\frac{r_t^i \rho_z 2^{-R_c}}{\theta_t^*}\right)\right]_+ = R - |\mathcal{S}_t| R_c. \tag{122}$$

Except when every remaining coefficient is zero, the optimizer exhausts the available budget. For a fixed cardinality $s \leq m$, an exchange argument shows that an optimal active set consists of the s largest coefficients r_t^i. The continuous mixed solution can therefore be found by sorting $\{r_t^i\}_{i=1}^n$, considering $s = 1, \ldots, m$, applying (121)–(122), and retaining the smallest value of (115). For integer rates, initialize every active worker at R_c and assign each remaining bit to a worker maximizing $r_t^i \rho_z 2^{-R_t^i}$. The corresponding marginal improvement halves after every additional bit, giving the discrete counterpart of reverse waterfilling.

The preceding reverse-waterfilling formulas are specific to the fixed-rate quantizer and should not be interpreted as intrinsic to `EF21-SR` itself. For instance, the Top-k sparsifier retains the k largest-magnitude coordinates and satisfies

$$\|\mathrm{Top}_k(\mathbf{u}) - \mathbf{u}\|^2 \leq \left(1 - \frac{k}{d}\right) \|\mathbf{u}\|^2, \tag{123}$$

and therefore has contraction $\delta = k/d$ (Stich et al., 2018). Its counterpart to (120) replaces $\rho_z 2^{-R_t^i}$ by $\sqrt{1 - k_t^i/d}$ and constrains the total coded cost of the retained coordinates. If each transmitted value uses R_{fp} bits and the support is shared, this cost is $k_t^i R_{\mathrm{fp}}$; if the support must also be encoded, its index cost must also be included as described in Appendix E. Thus, the fixed-rate reverse-waterfilling equations (121)–(122) are simply replaced by a compressor-specific discrete sparsification problem. This is why Algorithm 2 is stated at the level of an admissible compressor family and a shared communication budget, with fixed-rate quantization serving only as one concrete example.

The optimization in (115) is intentionally a one-step communication criterion and does not by itself guarantee long-run coverage. A greedy policy can repeatedly assign zero rate to a worker with a persistently small coefficient and leave its local tracker stale indefinitely. An adaptive policy must additionally satisfy the effective refresh condition, for example through explicit exploration or another anti-starvation mechanism. The static construction yields the corresponding per-round effective contraction directly through (117).

Finally, the broadcasts in Algorithms 1 and 2 synchronize the global state without sending $\mathbf{x}_{t+1}$ as a separate full-precision vector. After decoding the common codeword bundle, every worker performs

$$\widehat{\mathbf{g}}_t = \widehat{\mathbf{g}}_{t-1} + \frac{1}{n} \sum_{i\in\mathcal{S}_t} \mathbf{q}_t^i, \tag{124}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \widehat{\mathbf{g}}_t. \tag{125}$$

Only the worker-specific tracker $\widehat{\mathbf{g}}_t^i$ becomes stale when worker i is inactive; the global surrogate $\widehat{\mathbf{g}}_t$ and the iterate $\mathbf{x}_{t+1}$ remain common to all workers. When an inactive worker is selected again, it evaluates its local gradient at the current global iterate and corrects it against its own stale tracker.

Under the fixed-rate representation, worker i's vector codeword contains dR_t^i bits, hence

$$\sum_{i\in\mathcal{S}_t} dR_t^i \le dR. \tag{126}$$

The server may therefore rebroadcast the same collection of worker codewords using at most dR vector-payload bits; no second lossy rate-R compressor is needed. Equivalently, under idealized fixed-rate codebook accounting, the corresponding product codebook has cardinality

$$\prod_{i\in\mathcal{S}_t} 2^{dR_t^i} = 2^{d\sum_{i\in\mathcal{S}_t} R_t^i} \le 2^{dR}. \tag{127}$$

Thus the complete tuple has at most 2^{dR} possible realizations and can itself be indexed within the same vector-payload budget. In the one-call implementation, the active set and assigned rates need not be transmitted when they are reproducible from the common envelopes and deterministic allocation rule. Otherwise, active-worker identities, assigned rates, quantizer scales, support indices, codebook, and other protocol information are metadata and must be counted separately under an adopted communication model.

E Translating the Sum-Rate Constraint to Compressors

The optimization result and the communication constraint require two different properties from a compressor. Assumption *(A2)* needs a valid conditional bound on the compressor error $\mathcal{C}_t^i(\mathbf{u}_t^i) - \mathbf{u}_t^i$, while the sum-rate model needs the actual encoded length. A compressor can be inserted into the present framework without changing the proof only when its error obeys (10); it can be compared under the same resource model only when values, supports, scales, headers, and any allocation messages are included in its bit cost. An operator described solely by an unbiased variance bound is not automatically a contractive operator, so changing the symbol attached to such an operator is not a valid translation.

There is no universal deterministic relation from R_t^i to δ_t^i. A deterministic relation exists only after the compressor family, codebook, normalization, decoder side information, and accounting convention have been fixed. For the fixed-rate vector quantizer, the worst-case guarantee (11) permits the certified choice $\delta_t^i = 1 - \rho_z^2 2^{-2R_t^i}$ on an active realization. The actual error can be smaller, and a different quantizer using the same rate can have a different contraction. With randomized quantization, variable-length coding, data-dependent ranges, entropy coding, or decoder side information, the useful relation may instead be conditional or distribution-dependent. In those cases, *(A2)* uses the contraction certified after the current rate and side information are fixed, while the communication accounting uses either the realized code length or an expected length according to the resource claim. Rate and contraction must therefore remain distinct objects even when a convenient formula connects them.

If R_t^i is measured in bits per coordinate, the corresponding fixed-length vector payload is dR_t^i bits, and the hard constraint is

$$\sum_{i=1}^{n} dR_t^i \le dR. \tag{128}$$

For deterministic Top-k,

$$\|\mathrm{Top}_k(\mathbf{u}) - \mathbf{u}\|^2 \le \left(1 - \frac{k}{d}\right)\|\mathbf{u}\|^2, \tag{129}$$

so its contraction is k/d. If the support is known to the decoder and each retained value uses R_{fp} bits, a per-worker payload dR_t^i permits

$$k_t^i = \min\left\{d, \left\lfloor \frac{dR_t^i}{R_{\mathrm{fp}}} \right\rfloor\right\}. \tag{130}$$

If the support is not shared, a fixed-cardinality support alone requires at least $\lceil \log_2 \binom{d}{k_t^i} \rceil$ bits, so feasibility requires

$$k_t^i R_{\mathrm{fp}} + \left\lceil \log_2 \binom{d}{k_t^i} \right\rceil \le dR_t^i \tag{131}$$

before any additional scale or header. Unscaled random-k has the same contraction in conditional expectation when its shared randomness is fresh; the commonly used unbiased scaled random-k has a different error model and must be checked against *(A2)* rather than assigned $\delta_t^i = k_t^i/d$ by analogy. Structured sparsifiers replace the support cost by the cost of their structure, low-rank compressors must encode both factors, and finite-level quantizers must include the scale or range whenever it is not common side information.

The form of the rate constraint may also change, but different forms make different systems claims. The main paper uses the almost-sure, per-round constraint (2); therefore no realization exceeds the worker payload budget. An average constraint such as

$$\mathbb{E}\left[\sum_{i=1}^{n} R_t^i \,\middle|\, \mathcal{F}_{t-1}\right] \le R \tag{132}$$

allows individual rounds to exceed R. A probabilistic or chance constraint likewise permits violations with a prescribed probability. Such models can be meaningful for a buffered or long-term service, but they are not pathwise equivalents of (2); queueing, overflow, latency, or outage consequences must then be analyzed. The optimization proof can still condition on the realized allocation and its realized δ_t^i, but the hard-budget corollary and the fixed per-round bit comparisons no longer follow unchanged. For a variable-length prefix code, treating R_t^i as the realized length divided by d preserves the hard constraint; treating it as an expected length instead produces an expected-rate model.

Source rate is also not physical channel rate. A source payload of dR bits transmitted over a service of C bits per second has serialization time dR/C only when that service can be assigned to the messages being compared and other protocol costs are negligible. Separate uplink and downlink channels can have different capacities, a multiple-access channel has a coupled rate region rather than an additive collection of independent source links, and over-the-air aggregation communicates a function through simultaneous analog transmissions. Likewise, a common broadcast packet is charged once in the paper's accounting, while point-to-point replication can multiply the network-wide downlink traffic. These alternatives do not invalidate the source-rate formulation; they specify different physical layers to which it must be mapped.

F Additional Experimental Results and Implementation Details

This appendix specifies the datasets, objectives, partitions, compressors, bit accounting, stepsize rules, and execution environment used to produce the figures, and reports the supporting experiments deferred from Section 4.

F.1 Datasets and objectives

We use four publicly available matrices/datasets. The main text focuses on `w8a` and `illc1850`; the appendix adds `a9a` and `bcsstk15` as classification and least-squares counterparts.

Dataset	Task	N (approx.)	d
`illc1850`	least squares	1,850	712
`bcsstk15`	least squares	3,948	3,948
`w8a`	binary logistic	49,749	300
`a9a`	binary logistic	32,561	123

No samples are discarded. Features are row-normalized so that $\|a_j\|_2 = 1$, which yields conservative spectral bounds and stable curvature estimates across shards. For least-squares SuiteSparse matrices the collection supplies the design matrix A; the response is generated once as $y \sim \mathcal{N}(0, I)$ with a fixed NumPy seed (matching the companion repository). Classification labels are mapped to $\{0, 1\}$.

Three objective natures are used throughout. Convex runs minimize the unregularized empirical risk. Strongly convex runs add a ridge term $\frac{\mu_{\mathrm{reg}}}{2}\|x\|^2$ with $\mu_{\mathrm{reg}} = 0.1$. Nonconvex runs add the smooth EF21 penalty $\mu_{\mathrm{reg}} \sum_j x_j^2/(1 + x_j^2)$ with the same coefficient $\mu_{\mathrm{reg}} = 0.1$ as in Richtárik et al. (2021).

Table 3: Symbolic two-way bit cost used by the plotting ledger. $\mathrm{msg}(R_k) = \lfloor dR_k \rfloor + \lceil d/z \rceil R_{\mathrm{fp}}$ for group-wise uniform quantization; $\mathrm{msg}_{\mathrm{sign}} = d + R_{\mathrm{fp}}$ for scaled sign; $R_{\mathrm{fp}} = 32$.

Method	Init B_{init}	Per-round uplink	Per-round downlink
GD	0	$n\,d\,R_{\mathrm{fp}}$	$d\,R_{\mathrm{fp}}$
EF21-GD	$n\,d\,R_{\mathrm{fp}}$	$n \cdot \mathrm{msg}(R_w)$	$d\,R_{\mathrm{fp}}$
EF21-PP	$n\,d\,R_{\mathrm{fp}}$	$s \cdot \mathrm{msg}(R_w)$	$d\,R_{\mathrm{fp}}$
EF21-SR	$n\,d\,R_{\mathrm{fp}}$	$s \cdot \mathrm{msg}(R/s)$	$s \cdot \mathrm{msg}(R/s)$ (rebroadcast)
EF21-BC	$n\,d\,R_{\mathrm{fp}}$	$n \cdot \mathrm{msg}(R_w)$	$\mathrm{msg}(R_s)$

F.2 Distributed partitions and oracles

For each worker count n, sample indices are randomly permuted and split into n shards. Classification uses label-sorted shards (heterogeneous); least squares uses random shards. Local objectives are written so that $\frac{1}{n}\sum_k f_k = f$ even under unequal shard sizes. The initial iterate is drawn once as $x_0 \sim \mathcal{N}(0, I_d)$ and shared by every method in a comparison. Each configuration is repeated for two independent partition/algorithm seeds; plots show the mean curve with a $\pm$std band.

F.3 Compressors and communication accounting

The default encoder is the group-wise mid-rise uniform quantizer of block length $z = 127$ (the last block may be shorter). For rate R_k the integer budget $\lfloor dR_k \rfloor$ is spread as evenly as possible across coordinates, and each block carries one float32 scale. The analytic contraction is $\delta = 1 - z \cdot 2^{-2R_k} = 1 - \rho_z^2 2^{-2R_k}$ with $\rho_z = \sqrt{z}$. The contractive floor is $R_c = \lfloor \frac{1}{2}\log_2 z \rfloor + 1$. We also report a scaled-sign mirror: $C(u) = (\|u\|_1/d)\,\mathrm{sign}(u)$, charged as $d + R_{\mathrm{fp}}$ bits with $R_{\mathrm{fp}} = 32$, using the empirical contraction estimated by the companion compressor layer.

A full-precision vector costs dR_{fp} bits. Cumulative communication counts every recorded uplink, a single common downlink of the same payload (broadcast/multicast accounting), and scale metadata. Exact tracker initialization at x_0, when used, is charged once as n full-precision uplinks. The per-round ledger is summarized below; here $\mathrm{msg}(R_k) = \lfloor dR_k \rfloor + \lceil d/z \rceil R_{\mathrm{fp}}$ for the group-wise quantizer.

A run is accepted only if $\sum_i R_t^i \le R$ whenever a sum-rate budget is imposed, and the plotted cumulative-bit axis equals the ledger above.

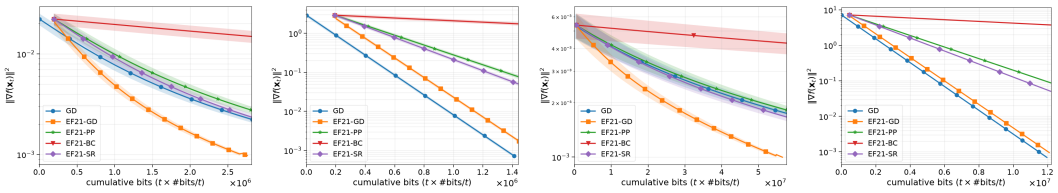
F.4 Stepsizes

Unless otherwise noted, every method uses the paper's maximum admissible stepsize under arithmetic-mean smoothness $\bar{L}$. For EF21-GD/PP/SR, η solves $(1/\eta)(1/\eta - L) \ge C(\delta_{\mathrm{eff}})^2$ with $C = \bar{L}\sqrt{1-\delta}/(1-\sqrt{1-\delta})$ and $\delta_{\mathrm{eff}} = p\delta$ under participation p. EF21-BC uses the corresponding two-level bound. At $p = 1$, EF21-PP recovers EF21-GD. The main text additionally reports (i) a common stepsize equal to the minimum of the five theory-max values, and (ii) independently tuned stepsizes obtained by a short grid search over multipliers of the theory-max step, scored by final $\|\nabla f\|^2$ on a 40-iteration probe. No other hyperparameters are tuned.

Runs stop at $\|\nabla f(\mathbf{x}_t)\|^2 \le 10^{-3}$, subject to a large iteration cap (2500 for classification, 6000 for least squares) as a safety net. Plotted x-axes are clipped to the first method that reaches the tolerance.

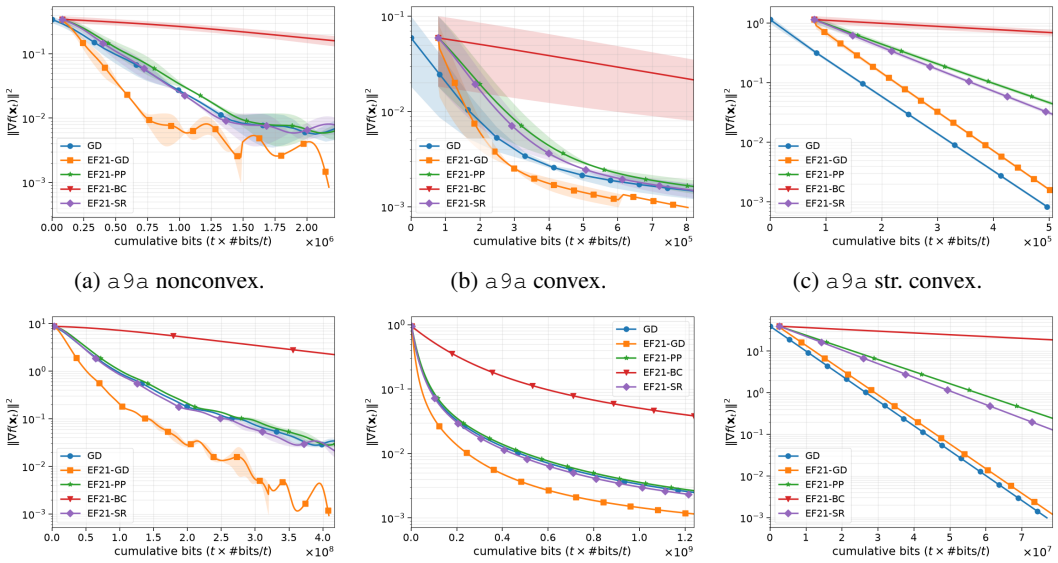
F.5 Convex and strongly convex counterparts of the main panels

Figure 5 reports the convex and strongly convex analogues of Figure 1 on w8a and illc1850, still at maximum feasible participation and maximum admissible η. The qualitative ordering is stable across natures: matched EF21-SR/EF21-PP share trajectories with a consistent bit saving for EF21-SR, while EF21-BC remains the most expensive method.



(a) `w8a` convex. (b) `w8a` str. convex. (c) `illc1850` convex. (d) `illc1850` str. convex.

Figure 5: Convex and strongly convex natures on the main datasets ($\|\nabla f\|^2$ vs cumulative bits, max admissible η).



(a) `a9a` nonconvex. (b) `a9a` convex. (c) `a9a` str. convex.

(d) `bcsstk15` nonconvex. (e) `bcsstk15` convex. (f) `bcsstk15` str. convex.

Figure 6: Additional datasets of comparable scale, all three natures.

F.6 Additional datasets

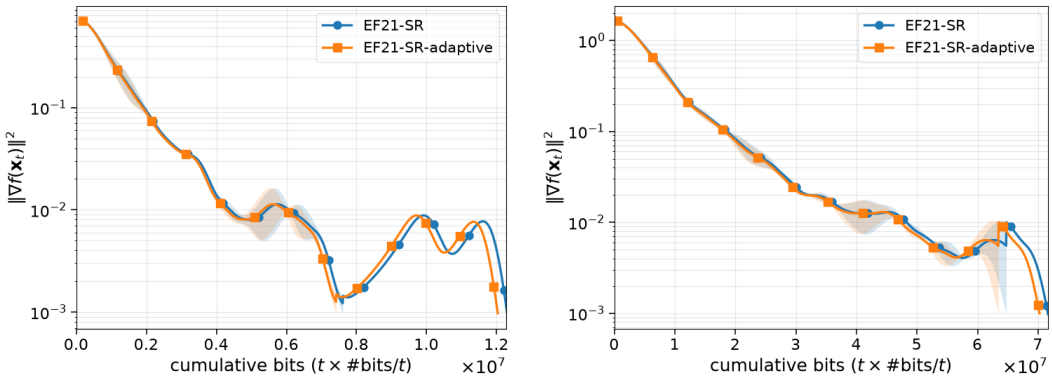
Figure 6 repeats the nonconvex/convex/strongly-convex matrix on `a9a` (classification) and `bcsstk15` (least squares). These datasets have dimensions and sample counts comparable to the main pair and serve as a reproducibility check that the ranking is not tied to a single matrix.

F.7 Static versus adaptive `EF21-SR`

Figure 7 compares static uniform allocation with the adaptive envelope rule (predictable envelopes and integer reverse waterfilling) on `w8a` and `illc1850` at the same budget. Because features are row-normalized, the per-worker smoothness constants are nearly homogeneous, so reverse waterfilling reduces to an almost uniform split: the two realizations track each other closely and adaptivity incurs no penalty at matched budget. Exact-current adaptive allocation is analyzed in the proofs but not run here.

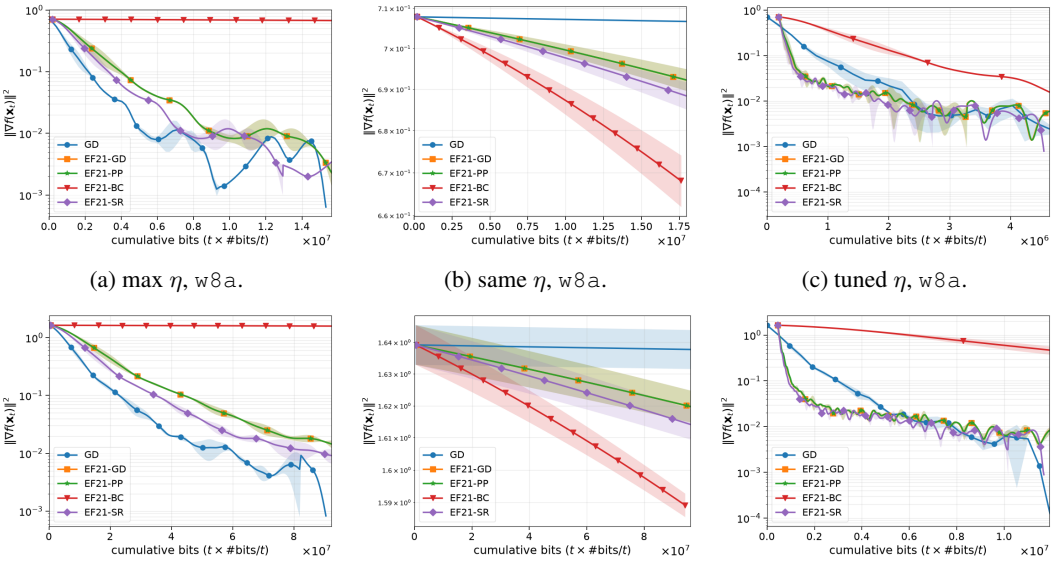
F.8 Scaled-sign compressor

Figure 8 repeats the three stepsize regimes of the main text with scaled `sign` compression and full participation. The same architectural comparison applies: when participation and the encoder are matched, `EF21-SR` and `EF21-PP` share trajectories, and any bit gap is again attributable to the



(a) `w8a`. (b) `illc1850`.

Figure 7: Static versus adaptive `EF21-SR` at matched budget.



(a) max η, `w8a`. (b) same η, `w8a`. (c) tuned η, `w8a`.

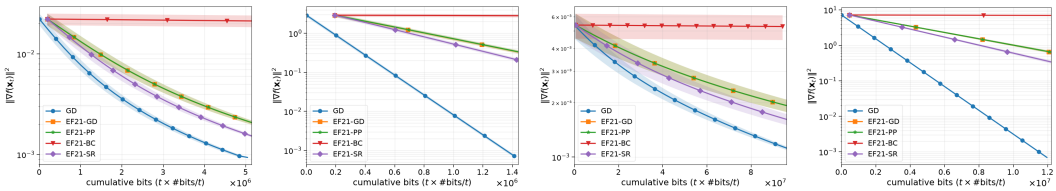
(d) max η, `illc1850`. (e) same η, `illc1850`. (f) tuned η, `illc1850`.

Figure 8: Scaled-`sign` compressor: three stepsize regimes on the main datasets.

downlink model (rebroadcast of the received sign bundle versus a full-precision iterate). Convex and strongly convex sign panels are collected in Figure 9.

F.9 Numerical bits-to-tolerance tables

Table 4 lists cumulative bits to $\varepsilon \in \{10^{-2}, 10^{-3}\}$ for the six main nonconvex panels (three stepsize regimes $\times$ two datasets). Sweep tables for s and R are written alongside the corresponding summary plots.



(a) `w8a` convex. (b) `w8a` SC. (c) `illc1850` convex. (d) `illc1850` SC.

Figure 9: Scaled-`sign` compressor, convex and strongly convex natures.

Table 4: Bits to accuracy ε on the main nonconvex panels.

Dataset	ε	GD	EF21-GD	EF21-PP	EF21-BC	EF21-SR
w8a (nonconvex, max η)	0.01	5.44×10^6	2.90×10^6	5.70×10^6	3.69×10^7	5.12×10^6
w8a (nonconvex, max η)	0.001	1.53×10^7	7.98×10^6	1.61×10^7	$--$	1.44×10^7
illc1850 (nonconvex, max η)	0.01	6.08×10^7	3.14×10^7	6.38×10^7	$--$	5.67×10^7
illc1850 (nonconvex, max η)	0.001	9.04×10^7	4.66×10^7	9.48×10^7	$--$	8.43×10^7
w8a (nonconvex, same η)	0.01	2.72×10^8	4.81×10^7	2.36×10^7	3.69×10^7	2.12×10^7
w8a (nonconvex, same η)	0.001	$--$	$--$	$--$	$--$	$--$
illc1850 (nonconvex, same η)	0.01	$--$	$--$	$--$	$--$	$--$
illc1850 (nonconvex, same η)	0.001	$--$	$--$	$--$	$--$	$--$
w8a (nonconvex, tuned η)	0.01	2.42×10^6	8.41×10^5	6.09×10^5	7.64×10^5	5.65×10^5
w8a (nonconvex, tuned η)	0.001	7.86×10^6	2.12×10^6	8.00×10^5	1.99×10^6	7.36×10^5
illc1850 (nonconvex, tuned η)	0.01	8.13×10^6	2.32×10^6	1.77×10^6	3.71×10^6	1.62×10^6
illc1850 (nonconvex, tuned η)	0.001	1.20×10^7	3.40×10^6	2.51×10^6	5.44×10^6	2.28×10^6