DISTRIBUTED VARIANCE REDUCTION WITH COMPRESSED COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign-based methods can reduce communication costs in distributed environments, but aggregating local signs can introduce bias when data are heterogeneous. As a result, existing sign-based variance reduction algorithms fail to obtain the optimal convergence rates. In this paper, we solve this problem and obtain optimal rates for both stochastic and finite-sum optimization. We first give a counterexample showing that majority voting can fail to approach stationary points even with exact local gradients. Motivated by this limitation, we propose tracking the global gradient at the server through unbiased compression of recursive gradient increments. The proposed methods obtain the convergence rates of $O(\sqrt{d/K}+\sqrt{d}\,(a/(nK))^{1/3})$ for the ℓ_1-norm and $O(\sqrt{a/K}+\sqrt{a}/(nK)^{1/3})$ for the ℓ_2-norm. Here, K is the iteration number, n is the number of workers, d is the dimension, and $a=1+\omega$, with ω denoting the relative variance of the compressor. For finite-sum problems with M components, we combine periodic exact gradient refreshes with compressed component-gradient differences. The resulting total sample complexities are $O(M+d\sqrt{aM}\epsilon^{-2})$ and $O(M+a\sqrt{M}\epsilon^{-2})$ for ℓ_1 and ℓ_2 expected gradient norms at most ϵ, matching the corresponding bounds in centralized settings.

1 INTRODUCTION

In this paper, we focus on smooth nonconvex optimization in the distributed environment:

$$\min_{x\in\mathbb{R}^d} f(x)=\frac{1}{n}\sum_{j=1}^{n} f_j(x), \qquad f_j(x)=\mathbb{E}_{\xi\sim\mathcal{D}_j}F_j(x;\xi). \tag{1}$$

In this setting, each worker $j\in\{1,2,\cdots,n\}$ accesses its own data distribution $\mathcal{D}_j$, and these data distributions may differ for different nodes. We also consider the finite-sum formulation, in which every worker has a fixed dataset of m components, and the problem can be written in the form:

$$\min_{x\in\mathbb{R}^d} f(x)=\frac{1}{n}\sum_{j=1}^{n} f_j(x), \qquad f_j(x)=\frac{1}{m}\sum_{i=1}^{m} f_{j,i}(x). \tag{2}$$

Here, $M=nm$ is the total number of components across the system. Our goal is to find stationary points while reducing the information sent between workers and the parameter server.

Sign-based methods are highly attractive for this purpose since a vector of signs only takes one bit per coordinate (Bernstein et al., 2018; 2019). Later studies show that the momentum technique improves the small-batch convergence guarantees (Safaryan & Richtárik, 2021; Jiang et al., 2025), and variance reduction further improves the quality of the gradient estimate. In particular, the previous centralized variance reduction method SSVR attains an ℓ_1 convergence rate of $O(\sqrt{d}\,K^{-1/3})$ under mean-squared smoothness, and SSVR-FS attains a total gradient complexity of $O(M+d\sqrt{M}\,\epsilon^{-2})$ for ℓ_1 accuracy ϵ on a centralized dataset of M components (Jiang et al., 2024).

These results motivate extending the benefits of variance reduction to the distributed sign-based problem. Such an extension is not automatic. In the traditional majority-vote method, each worker first converts its local gradient estimate into a sign vector, and the server takes the majority of those signs. However, for different local objectives, this nonlinear aggregation may not align with the

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.

gradient of their average. The SSVR-MV method (Jiang et al., 2024) provides two options for server updates. Option 1 uses deterministic majority voting and has an ℓ_1 guarantee of $O(\sqrt{d/K} + d/\sqrt{n})$ with an error floor. Option 2 instead employs unbiased stochastic sign operations and obtains an ℓ_2 rate of $O(d^{1/4}K^{-1/4})$. It also requires projecting each local estimate onto a Euclidean ball before the worker-side sign quantization, which is impractical to use. More importantly, the obtained $O(K^{-1/4})$ guarantee leaves a gap to the centralized variance-reduced $O(K^{-1/3})$ rate.

In this paper, we obtain improved convergence rates with newly designed algorithms. We assume an unbiased relative-variance compressor that introduces noise proportional to the input magnitude, allowing more general compressors. We then let the server retain a global gradient estimate. Workers send compressed recursive increments, and the server accumulates these messages before choosing a global update direction. Our recursion compresses an increment consisting of a scaled gradient and a same-sample gradient difference, whose second moment can be controlled by choosing the parameters properly. Finally, we use deterministic signs for the ℓ_1 criterion and unbiased compression for the ℓ_2 guarantee.

For the finite-sum problem, we periodically compute the exact global gradient and use compressed component-gradient differences for steps between checkpoints. The same component is evaluated at two successive iterates, so component smoothness controls the update even when local gradients are large or heterogeneous. The resulting total sample complexities are $O(M + d\sqrt{aM}\epsilon^{-2})$ for the ℓ_1-norm and $O(M + a\sqrt{M}\epsilon^{-2})$ for the ℓ_2-norm, where $a = 1 + \omega$ and ω is the compressor's relative variance. These have the same oracle orders as the centralized variance-reduced methods.

Our contributions. We summarize the results and contributions of this paper below.

- We first analyze the drawback of local majority voting. We give a counterexample in Proposition 1, in which majority voting can produce a nonzero gradient floor despite exact local gradients. It isolates the bias of the final vote and motivates tracking the global gradient before taking signs.
- For stochastic problems, we attain the $O(\sqrt{d/K} + \sqrt{d}(a/(nK))^{1/3})$ rate in the ℓ_1 criterion, removing the previous error floor. Then, we also obtain the $O(\sqrt{a/K} + \sqrt{a}/(nK)^{1/3})$ rate for the ℓ_2-norm, improving the dependence on K for previous methods. Both rates ensure the linear speedup for the dominating term and are optimal for the iteration number K.
- We further analyze the finite-sum bounds for the sign-based distributed setting. Periodic exact refreshes and compressed component differences remove the need for the bounded gradient assumption. Finally, we obtain sample complexities of $O(M + d\sqrt{aM}\epsilon^{-2})$ and $O(M + a\sqrt{M}\epsilon^{-2})$ for the ℓ_1 and ℓ_2 criteria, matching the corresponding optimal centralized rates in terms of M and ϵ.

2 Related Work

Sign-based optimization. SignSGD and its majority-vote variant reduce communication by transmitting gradient signs (Bernstein et al., 2018; 2019). Their analysis has motivated several ways to handle stochastic noise and heterogeneous data. Later, Safaryan & Richtárik (2021) study randomized sign methods and distributed momentum, while Sun et al. (2023) show that momentum permits convergence with constant batch sizes under weaker noise assumptions. Stochastic-Sign SGD uses randomized sign compressors in heterogeneous federated learning (Jin et al., 2020). More recent analyses sharpen the guarantees for sign methods with momentum (Jiang et al., 2025) and study Lion in both centralized and distributed settings (Jiang et al., 2026).

Variance reduction for sign updates. SignSVRG first combines variance reduction techniques with sign updates for finite-sum optimization (Chzhen & Schechtman, 2023), and Qin et al. (2026) study sign methods with random reshuffling and variance reduction. To improve the convergence rates, SSVR and SSVR-FS are proposed to obtain the optimal stochastic and finite-sum guarantees of centralized sign methods (Jiang et al., 2024). The same work introduces SSVR-MV for distributed heterogeneous workers. However, the bound for its deterministic majority-vote option contains an

error-floor term, whereas its randomized server option obtains a worse convergence guarantee and requires projection of local estimates before quantization.

Variance reduction for nonconvex optimization. The variance reduction technique SARAH develops a recursive estimator based on component-gradient differences (Nguyen et al., 2017) and obtains an improved convergence rate for smooth convex optimization. For nonconvex objectives, the SPIDER and PAGE methods also use related estimators to obtain stochastic and finite-sum oracle guarantees (Fang et al., 2018; Li et al., 2021). Under the average smoothness assumption, these methods attain the $O(\epsilon^{-3})$ stochastic and $O(M + \sqrt{M}\,\epsilon^{-2})$ finite-sum orders for $\mathbb{E}\|\nabla f(x)\|_2 \le \epsilon$. These bounds match the lower bounds and are thus optimal in their settings (Arjevani et al., 2023). To avoid using large batches in the iteration, STORM incorporates recursive variance reduction into momentum (Cutkosky & Orabona, 2019), and the STORM+ algorithm further develops this approach with adaptive parameters (Levy et al., 2021).

3 Assumptions

We now specify the assumptions used. Write $g_j(x;\xi) = \nabla F_j(x;\xi)$ for a stochastic gradient and $\Delta_f = f(x_1) - f_*$ for the initial objective gap, where f_* is a lower bound on the objective f.

We first give the following assumption on the objective function for the stochastic problem (1).

Assumption 1. *The objective satisfies $f \ge f_* > -\infty$. For every worker and all x, y, we have*

$$\mathbb{E}_\xi g_j(x;\xi) = \nabla f_j(x), \tag{3}$$

$$\mathbb{E}_\xi \|g_j(x;\xi)\|_2^2 \le H^2, \tag{4}$$

$$\mathbb{E}_\xi \|g_j(x;\xi) - g_j(y;\xi)\|_2^2 \le L^2 \|x - y\|_2^2. \tag{5}$$

Remark: Here, Jensen's inequality implies that f_j and f are also L-smooth. Besides, inequality (4) can be replaced with bounded oracle variance σ^2 and a true-gradient bound $\|\nabla f_j(x)\|_2 \le G_2$, which imply inequality (4) with $H^2 = \sigma^2 + G_2^2$.

Next, we give the general compression assumption below.

Assumption 2 (Relative-variance compressor)**.** *For every input v, the output $Q(v)$ satisfies*

$$\mathbb{E}[Q(v) \mid v] = v, \qquad \mathbb{E}[\|Q(v) - v\|_2^2 \mid v] \le \omega \|v\|_2^2. \tag{6}$$

Note that unbiasedness also gives

$$\mathbb{E}[\|Q(v)\|_2^2 \mid v] \le (1 + \omega)\|v\|_2^2. \tag{7}$$

Remark: The general compression assumption includes the usual scaled stochastic sign $S(v)$. Let $\rho(v) = \|v\|_\infty$ and draw independent signs with $\Pr(S(v)_k = +1 \mid v) = (1 + v_k/\rho(v))/2$ and $\Pr(S(v)_k = -1 \mid v) = (1 - v_k/\rho(v))/2$. Then

$$Q(v) = \rho(v) S(v), \qquad Q(0) = 0, \tag{8}$$

is unbiased and satisfies $1 + \omega = d$ since $\mathbb{E}[\|Q(v) - v\|_2^2 \mid v] = d\|v\|_\infty^2 - \|v\|_2^2 \le (d-1)\|v\|_2^2$. Two other typical examples are the Rand-k sparsification, which retains the k random values of the vector scaled by d/k, and the natural compression proposed by Horvóth et al. (2022).

Finally, we list assumptions for the finite-sum problem (2).

Assumption 3. *The objective is lower bounded and each component satisfies*

$$\|\nabla f_{j,i}(x) - \nabla f_{j,i}(y)\|_2 \le L\|x - y\|_2 \quad \textit{for all } x, y, j, i. \tag{9}$$

Our finite-sum results use Assumptions 2 and 3, requiring no bound on gradients, oracle variance, or gradient heterogeneity.

4 The proposed method for stochastic problems

We begin with the earlier SSVR-MV method and explain the bias created by its majority vote. This motivates a global gradient estimate at the server, followed by either a sign update for the ℓ_1 criterion or an unbiased compressed update for the ℓ_2 criterion.

4.1 SSVR-MV AND THE LIMITATION OF ITS MAJORITY VOTING

The earlier method. SSVR-MV, introduced by Jiang et al. (2024), combines a local variance reduction estimator with bidirectional sign communication. Specifically, each worker j first initializes $v_1^j = g_j(x_1; \xi_1^j)$ and, for $t \geq 2$, computes

$$v_t^j = g_j(x_t; \xi_t^j) + (1-\beta)[v_{t-1}^j - g_j(x_{t-1}; \xi_t^j)]. \tag{10}$$

The two stochastic gradients in this update use the same sample ξ_t^j. For a fixed radius $R > 0$ and an input satisfying $\|v\|_\infty \leq R$, define the random sign vector $S_R(v)$ coordinatewise by

$$\Pr(S_R(v)_k = +1 \mid v) = \frac{1 + v_k/R}{2}, \qquad \Pr(S_R(v)_k = -1 \mid v) = \frac{1 - v_k/R}{2}. \tag{11}$$

Thus, the signs estimate the scaled input without any bias. In SSVR-MV with Option 1 (Jiang et al., 2024), worker j sends $q_t^j = S_R(v_t^j)$, and the server broadcasts their deterministic majority vote:

$$s_t = \operatorname{Sign}\left(\frac{1}{n}\sum_{j=1}^n q_t^j\right), \qquad x_{t+1} = x_t - \eta s_t. \tag{12}$$

The setting of their Theorem 3 uses $\beta = 1/2$, $\eta = \mathcal{O}((dK)^{-1/2})$, and $R = 4G$, keeping the inputs inside the radius. The server forms a new vote at every iteration.

Where the bias enters. The worker signs are unbiased for their scaled inputs, but the majority operation is nonlinear. This can be seen exactly with three workers. For any $q_1, q_2, q_3 \in \{-1, 1\}$,

$$\operatorname{Sign}(q_1 + q_2 + q_3) = \frac{q_1 + q_2 + q_3 - q_1 q_2 q_3}{2}. \tag{13}$$

Writing $m_j = \mathbb{E}q_j$, with all expectations conditional on these inputs, we obtain

$$\begin{aligned} \mathbb{E}\operatorname{Sign}(q_1 + q_2 + q_3) &= \frac{1}{2}\left(\sum_{j=1}^3 \mathbb{E}q_j - \mathbb{E}[q_1 q_2 q_3]\right) \\ &= \frac{1}{2}(m_1 + m_2 + m_3 - m_1 m_2 m_3). \end{aligned} \tag{14}$$

If $(m_1, m_2, m_3) = (u, u, -2u)$ with $0 < u \leq 1/2$, the average sign mean is zero, whereas the expected majority vote is $u^3 > 0$. Unbiased worker messages therefore can produce a biased vote.

Proposition 1 (A gradient floor with exact local estimates)**.** *There exist three lower-bounded smooth one-dimensional local objectives with gradients bounded by $G = 1$ such that SSVR-MV Option 1, initialized at a global minimizer and using exact gradients, satisfies*

$$\liminf_{K\to\infty} \mathbb{E}|f'(x_\tau)| \geq \frac{1}{1542} > 0 \tag{15}$$

for $R = 4$ and step size $\eta = \Theta(K^{-1/2})$, where τ is independent and uniform on $\{1, \ldots, K\}$.

Construction and intuition. Take

$$f_1(x) = f_2(x) = \tfrac{1}{2}\log\cosh x + \tfrac{1}{4}x, \qquad f_3(x) = \tfrac{1}{2}\log\cosh x - \tfrac{1}{2}x.$$

Their average is $f(x) = \frac{1}{2}\log\cosh x$, minimized at $x = 0$. At this point the worker sign means are $(u, u, -2u)$ with $u = 1/16$, so (14) gives a nonzero expected vote despite $f'(0) = 0$. Appendix A proves that this bias produces the stated gradient floor. □

4.2 TRACKING THE GLOBAL GRADIENT THROUGH COMPRESSED INCREMENTS

We retain the variance-reduction idea in (10), but change what workers transmit and what the server remembers. For $t \geq 2$, worker j forms the increment

$$\begin{aligned} h_t^j &= g_j(x_t; \xi_t^j) - (1-\beta) g_j(x_{t-1}; \xi_t^j) \\ &= \beta g_j(x_t; \xi_t^j) + (1-\beta)[g_j(x_t; \xi_t^j) - g_j(x_{t-1}; \xi_t^j)]. \end{aligned} \tag{16}$$

Algorithm 1 DVR-Sign / DVR-Q

Require: x_1, K, B_0, η, β, compressor Q.
Each worker draws B_0 samples at x_1.
Server initializes $z_1 = (nB_0)^{-1} \sum_{j,b} Q(g_j(x_1; \xi_{1,b}^j))$.
for $t = 1, \dots, K$ **do**
  if $t \geq 2$ **then**
    Each worker draws a sample ξ_t^j, forms (16), and sends $Q(h_t^j)$.
    Server updates z_t by (17).
  end if
  if variant is DVR-Sign **then**
    Server broadcasts $s_t = \mathrm{Sign}(z_t)$ to all workers.
  else
    Server broadcasts $s_t = Q(z_t)$ to all workers.
  end if
  All workers and the server set $x_{t+1} = x_t - \eta s_t$.
end for
Draw an independent uniform $\tau \in \{1, \dots, K\}$ and **return** x_τ.

The worker sends an encoding of $Q(h_t^j)$ and the server then accumulates the increments:

$$z_t = (1 - \beta) z_{t-1} + \frac{1}{n} \sum_{j=1}^{n} Q(h_t^j). \tag{17}$$

The whole algorithm is specified in Algorithm 1. The difference from (12) is substantive: the server retains z_{t-1} and corrects it with numerical increments before choosing its update direction. In particular, it does not take a majority vote of independently randomized local signs.

The increment (16) can be decomposed as follows: the fresh-gradient term $g_j(x_t; \xi_t^j)$ is multiplied by β, while mean-squared smoothness controls the same-sample difference $g_j(x_t; \xi_t^j) - g_j(x_{t-1}; \xi_t^j)$. These two terms give

$$\mathbb{E}\|h_t^j\|_2^2 \leq 2\beta^2 H^2 + 2L^2 \mathbb{E}\|x_t - x_{t-1}\|_2^2. \tag{18}$$

Unlike repeatedly compressing a full local estimate, compressing this increment introduces noise that decreases with smaller β and smaller model movements. The two methods below control this movement differently.

4.3 The compressed update for the server

For ℓ_1 measure For the ℓ_1 criterion $\mathbb{E}\|\nabla f(x)\|_1 \leq \epsilon$, the natural direction is $s_t = \mathrm{Sign}(z_t)$. With an exact estimate $z_t = \nabla f(x_t)$, its inner product with the gradient equals $\|\nabla f(x_t)\|_1$. With an approximate estimate, the corresponding inequality is

$$-\langle g, \mathrm{Sign}(z) \rangle \leq -\|g\|_1 + 2\|z - g\|_1 \leq -\|g\|_1 + 2\sqrt{d}\|z - g\|_2. \tag{19}$$

Thus the deterministic sign update converts a bound on the server's tracking error into an ℓ_1 stationarity guarantee. We call this algorithm *DVR-Sign*: distributed variance reduction with sign updates.

For ℓ_2 measure The ℓ_1-norm guarantee already implies an ℓ_2-norm guarantee via $\|g\|_2 \leq \|g\|_1$. However, this transfer retains the $\sqrt{d}$ tracking-error coefficient in (19). We therefore use the unbiased compressor defined in Assumption 2 for the server broadcast as well:

$$s_t = Q(z_t), \qquad x_{t+1} = x_t - \eta s_t. \tag{20}$$

We call this method *DVR-Q*. Conditional on the history $\mathcal{G}_t$ containing z_t and all current worker messages, fresh server randomness gives

$$\mathbb{E}[s_t \mid \mathcal{G}_t] = z_t, \qquad \mathbb{E}[\|s_t\|_2^2 \mid \mathcal{G}_t] \leq (1 + \omega)\|z_t\|_2^2. \tag{21}$$

With an exact estimate, the expected gradient inner product is $\|\nabla f(x_t)\|_2^2$. For an approximate estimate, the elementary identity

$$-2\langle g, z \rangle = -\|g\|_2^2 - \|z\|_2^2 + \|z - g\|_2^2 \tag{22}$$

connects descent to squared Euclidean tracking error. Its negative $\|z\|_2^2$ term is essential: it absorbs the additional tracking error caused by the random step length in (21). The two methods thus use different directions for different criteria. Deterministic signs give the ℓ_1 inner product in (19), while unbiased Q updates give squared ℓ_2 descent without its explicit $\sqrt{d}$ conversion.

5 STOCHASTIC GUARANTEES

We first analyze DVR-Sign under the ℓ_1 criterion and then analyze DVR-Q in the Euclidean norm. The two methods share the server recursion, but their model movements and descent arguments differ. Denote $a = 1 + \omega$, $c = a/n$, $e_t = z_t - \nabla f(x_t)$, $E_t = \mathbb{E}\|e_t\|_2^2$, and $\overline{E}_K = K^{-1}\sum_{t=1}^{K} E_t$.

5.1 THE ℓ_1 GUARANTEE FOR DVR-SIGN

Theorem 2. *Under Assumptions 1 and 2, for any $\eta > 0$, $\beta \in (0, 1]$, and integers $K, B_0 \geq 1$, the DVR-Sign variant of Algorithm 1 satisfies*

$$E_1 \leq cH^2/B_0, \qquad E_t \leq (1-\beta)E_{t-1} + 2c(\beta^2 H^2 + L^2\eta^2 d) \quad (t \geq 2), \tag{23}$$

$$\overline{E}_K \leq \frac{cH^2}{B_0\beta K} + 2c\left(H^2\beta + \frac{L^2\eta^2 d}{\beta}\right), \tag{24}$$

$$\mathbb{E}\|\nabla f(x_\tau)\|_1 \leq \frac{\Delta_f}{\eta K} + \frac{L\eta d}{2} + 2\sqrt{d\,\overline{E}_K}. \tag{25}$$

Balancing these effects gives the following result.

Corollary 3. *Under the assumptions of Theorem 2, choose*

$$u_1 = \frac{1}{\sqrt{K} + c^{1/3}K^{2/3}}, \qquad \beta = u_1, \quad \eta = \frac{Hu_1}{L\sqrt{d}}, \quad B_0 = \left\lceil \frac{1}{u_1^2 K} \right\rceil. \tag{26}$$

Then DVR-Sign satisfies

$$\begin{aligned}\mathbb{E}\|\nabla f(x_\tau)\|_1 \leq& \sqrt{d}\left[\left(\frac{L\Delta_f}{H} + \frac{H}{2}\right)K^{-1/2} + \left(\frac{L\Delta_f}{H} + 2\sqrt{5}H\right)\left(\frac{c}{K}\right)^{1/3}\right] \\ =& O\left(\sqrt{\frac{d}{K}} + \sqrt{d}\left(\frac{1+\omega}{nK}\right)^{1/3}\right).\end{aligned} \tag{27}$$

Remark: To achieve $\mathbb{E}\|\nabla f(x_\tau)\|_1 \leq \epsilon$, the required number of model updates is

$$K = O\left(1 + \frac{d}{\epsilon^2} + \frac{ad^{3/2}}{n\epsilon^3}\right).$$

5.2 THE ℓ_2 GUARANTEE FOR DVR-Q

The unbiased compressor $Q(z_t)$ has conditional mean z_t, and its conditional second moment is at most $(1+\omega)\|z_t\|_2^2$. The descent analysis therefore controls the *squared* gradient norm and retains a negative tracker-norm term. This term absorbs the tracking error caused by moving the model, which is the key difference from the sign analysis.

Theorem 4. *Under Assumptions 1 and 2, for any $\eta > 0$, $\beta \in (0, 1]$, and integers $K, B_0 \geq 1$, the DVR-Q variant of Algorithm 1 satisfies*

$$\begin{aligned}&\mathbb{E}\|\nabla f(x_\tau)\|_2^2 + \left(1 - La\eta - \frac{2caL^2\eta^2}{\beta}\right)\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|z_t\|_2^2 \\ &\qquad \leq \frac{2\Delta_f}{\eta K} + \frac{cH^2}{B_0\beta K} + 2c\beta H^2.\end{aligned} \tag{28}$$

In particular, if $La\eta \leq 1/2$ and $\beta \geq 4caL^2\eta^2$, the tracker-norm term on the left is nonnegative and can be dropped.

Corollary 5. *Under the assumptions of Theorem 4, choose*

$$r = \left(\frac{K}{n^2}\right)^{1/3}, \qquad \eta = \frac{1}{2La(1+r)}, \quad \beta = \frac{1}{n(1+r)^2}, \quad B_0 = \lceil 1+r \rceil. \tag{29}$$

Then DVR-Q satisfies

$$\begin{aligned} \mathbb{E}\|\nabla f(x_\tau)\|_2 &\le \sqrt{4L\Delta_f + H^2}\sqrt{\frac{a}{K}} + \sqrt{4L\Delta_f + 3H^2}\frac{\sqrt{a}}{(nK)^{1/3}} \\ &= O\left(\sqrt{\frac{1+\omega}{K}} + \frac{\sqrt{1+\omega}}{(nK)^{1/3}}\right). \end{aligned} \tag{30}$$

Remark: To achieve $\mathbb{E}\|\nabla f(x_\tau)\|_2 \le \epsilon$, the required number of model updates is

$$K = O\left(1 + \frac{a}{\epsilon^2} + \frac{a^{3/2}}{n\epsilon^3}\right). \tag{31}$$

Notably, the obtained rates are optimal in terms of ϵ (Li et al., 2021; Arjevani et al., 2023) and we achieve the linear speedup effect for both the ℓ_1 and ℓ_2 criteria. Specifically, the required iteration number K scales as $\mathcal{O}(1/(n\epsilon^3))$ for the dominant term, which is a factor-n reduction when n workers are used. This linear speedup does not appear in the previous rates (Jiang et al., 2024).

6 The proposed methods for finite-sum problems

For the finite-sum problem (2), an occasional full pass through the data can replace the stochastic correction used by DVR-Sign and DVR-Q. We compute exact gradients at checkpoints, followed by compressed component-gradient differences. We use a deterministic sign for the ℓ_1 criterion and employ the unbiased compressor Q for the ℓ_2 criterion.

Fix an integer refresh period q. At times $t = 1 + kq$ (where $k \ge 0$ is an integer), every worker computes and sends its local full gradient, and the server resets $z_t = \nabla f(x_t)$. At other times, each worker independently draws a uniform component $i_t^j \in \{1, \dots, m\}$ and sends a compressed paired difference, using the same component at both iterates. The server uses

$$y_t^j = \nabla f_{j,i_t^j}(x_t) - \nabla f_{j,i_t^j}(x_{t-1}), \qquad z_t = z_{t-1} + \frac{1}{n}\sum_{j=1}^{n} Q(y_t^j). \tag{32}$$

The server retains z_t itself for the next estimator update. Smoothness gives $\|y_t^j\|_2 \le L\|x_t - x_{t-1}\|_2$, which controls the increment without requiring bounded gradients or bounded heterogeneity. The whole method is described in Algorithm 2.

6.1 Deterministic sign updates for the ℓ_1 guarantee

The first version uses $s_t = \mathrm{Sign}(z_t)$. A refresh eliminates the estimation error and only the $q-1$ intervening updates can contribute to its variance. A refresh costs M component-gradient evaluations, whereas an ordinary update costs $2n$. Choosing $q = m$ makes the amortized refresh cost $M/q = n$, of the same order as the ordinary update cost.

Theorem 6. *Under Assumptions 2 and 3, let $a = 1 + \omega$. For DVR-Sign-FS, define*

$$J_1 = \frac{d}{2} + 2d\sqrt{\frac{a(q-1)}{n}}. \tag{33}$$

For an independent uniform $\tau \in \{1, \dots, K\}$,

$$\mathbb{E}\|z_t - \nabla f(x_t)\|_2^2 \le \frac{aL^2\eta^2 d(q-1)}{n}, \tag{34}$$

$$\mathbb{E}\|\nabla f(x_\tau)\|_1 \le \frac{\Delta_f}{\eta K} + L\eta J_1. \tag{35}$$

Algorithm 2 DVR-Sign-FS and DVR-Q-FS

Require: x_1, K, η, refresh period q, and compressor Q.
for $t = 1, \dots, K$ **do**
  if $t = 1 + kq$ for an integer $k \geq 0$ **then**
    Each worker computes and sends $\nabla f_j(x_t) = m^{-1} \sum_{i=1}^{m} \nabla f_{j,i}(x_t)$.
    Server sets $z_t = n^{-1} \sum_{j=1}^{n} \nabla f_j(x_t)$.
  else
    Every worker j draws an independent uniform $i_t^j \in \{1, \dots, m\}$.
    Every worker forms y_t^j in (32) and sends $Q(y_t^j)$.
    Server sets $z_t = z_{t-1} + n^{-1} \sum_{j=1}^{n} Q(y_t^j)$.
  end if
  if DVR-Sign-FS **then**
    Server broadcasts $s_t = \mathrm{Sign}(z_t)$.
  else
    Server broadcasts $s_t = Q(z_t)$.
  end if
  Set $x_{t+1} = x_t - \eta s_t$.
end for
Draw an independent uniform $\tau \in \{1, \dots, K\}$ and **return** x_τ.

Corollary 7. *Under the assumptions of Theorem 6, set $q = m$. We ensure $\mathbb{E}\|\nabla f(x_\tau)\|_1 \leq \epsilon$, with*

$$\eta = \frac{\epsilon}{2LJ_1}, \quad K = O\left(1 + \frac{L\Delta_f d}{\epsilon^2}\left[1 + \sqrt{\frac{a(m-1)}{n}}\right]\right). \tag{36}$$

$$N_{\mathrm{grad}} = O\left(M + \frac{L\Delta_f d}{\epsilon^2}[n + \sqrt{aM}]\right), \tag{37}$$

where N_{grad} is the total number of component-gradient evaluations.

Remark: For $n = O(am)$, the total oracle order is $N_{\mathrm{grad}} = O\Big(M + \frac{d\sqrt{aM}}{\epsilon^2}\Big)$.

6.2 Unbiased compressed updates for the ℓ_2 guarantee

The second version keeps the same refreshes and estimator, but uses $s_t = Q(z_t)$, where Q satisfies Assumption 2 with $a = 1 + \omega$. Unlike a sign step of fixed length, this update has conditional second moment at most $a\eta^2\|z_t\|_2^2$. Unbiasedness also makes the expected descent depend on $\langle \nabla f(x_t), z_t \rangle$. These two facts allow the tracking error to be absorbed into the descent inequality and yield a squared Euclidean stationarity guarantee.

Theorem 8. *Under Assumptions 2 and 3, for $a = 1 + \omega$, define*

$$J_Q = a\left(1 + \sqrt{\frac{q-1}{n}}\right), \qquad \eta = \frac{1}{2LJ_Q}. \tag{38}$$

Then an independent uniform $\tau \in \{1, \dots, K\}$ satisfies

$$\mathbb{E}\|\nabla f(x_\tau)\|_2^2 \leq \frac{4L\Delta_f J_Q}{K}, \qquad \mathbb{E}\|\nabla f(x_\tau)\|_2 \leq 2\sqrt{\frac{L\Delta_f J_Q}{K}}. \tag{39}$$

Corollary 9. *Under the assumptions of Theorem 8, set $q = m$. We ensure $\mathbb{E}\|\nabla f(x_\tau)\|_2 \leq \epsilon$, with*

$$K = O\left(1 + \frac{aL\Delta_f}{\epsilon^2}\left[1 + \sqrt{\frac{m-1}{n}}\right]\right), N_{\mathrm{grad}} = O\left(M + \frac{aL\Delta_f}{\epsilon^2}[n + \sqrt{M}]\right). \tag{40}$$

Remark: For $n = O(m)$, this total oracle order is $N_{\mathrm{grad}} = O\Big(M + \frac{a\sqrt{M}}{\epsilon^2}\Big)$, matching the centralized Euclidean rates of SPIDER and PAGE (Fang et al., 2018; Li et al., 2021), and is optimal in terms of ϵ and M (Li et al., 2021; Arjevani et al., 2023).

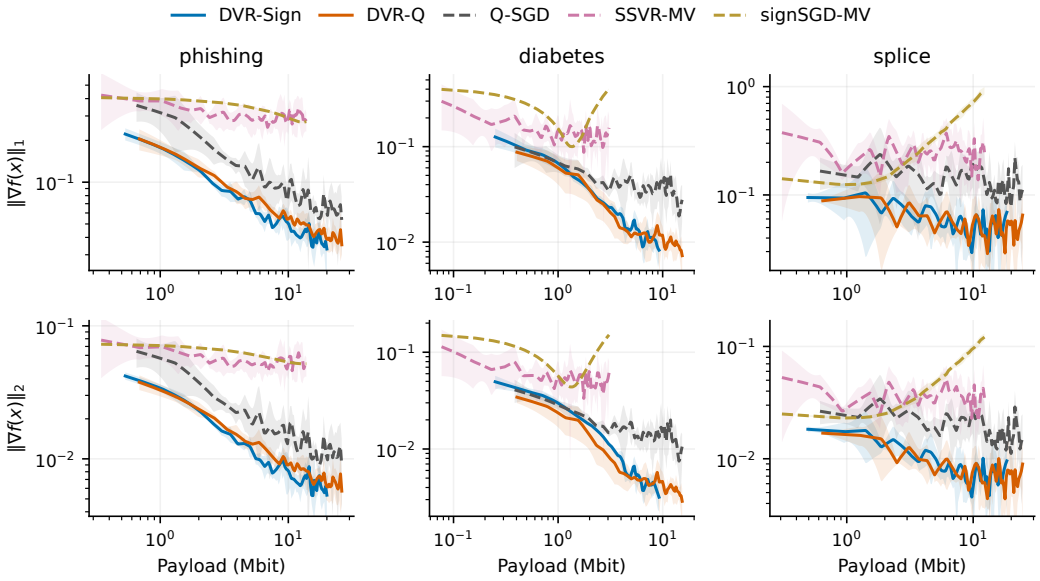


Figure 1: Gradient norms versus cumulative bidirectional payload for different methods.

Table 1: ImageNet accuracies (%) for different methods.

Model	signSGD-MV	Q-SGD	SSVR-MV	DVR-Sign	DVR-Q
ResNet-50	72.8 ± 0.4	72.9 ± 0.3	73.1 ± 0.3	$\mathbf{73.7 \pm 0.2}$	73.5 ± 0.3
ViT-S	72.7 ± 0.4	73.0 ± 0.4	73.2 ± 0.2	73.6 ± 0.3	$\mathbf{73.9 \pm 0.2}$

7 Experiments

Finally, we compare our DVR-Sign and DVR-Q methods with SSVR-MV (Jiang et al., 2024), signSGD with majority vote (Bernstein et al., 2019), and Q-SGD. Here, Q-SGD compresses stochastic gradients and uses the same compressor as DVR-Q, allowing us to assess the benefit of the variance reduction technique. The primary compressor is the scaled stochastic binary sign in (8). The step sizes are searched from the set $\{1, 3e{-}1, 1e{-}1, 3e{-}2, 1e{-}2, 3e{-}3, 1e{-}3, 3e{-}4, 1e{-}4\}$ and the momentum parameter is selected within $\{1e{-}1, 1e{-}2, 1e{-}3\}$.

Minimize logistic loss We first study binary classification with a smooth nonconvex regularizer. Let (a_i, y_i) denote a training example, where $y_i \in \{-1, +1\}$. The training objective is given by $f(x) = \frac{1}{M}\sum_{i=1}^{M} \log\big(1 + \exp(-y_i a_i^\top x)\big) + \lambda \sum_{k=1}^{d} \frac{x_k^2}{1+x_k^2}$ with $\lambda = 10^{-3}$. We minimize this loss on three LIBSVM datasets (Chang & Lin, 2011) with eight workers. Figure 1 compares the five methods using the ℓ_1 and ℓ_2 gradient norms. Here, we count payload, which denotes the cumulative communication volume across all worker uplinks and server-to-worker downlinks. As can be seen in the figure, our DVR-Sign and DVR-Q reach lower final mean norms than all three baselines on all datasets, and improvement also appears at a fixed gradient norm target. These gains illustrate the benefit of the recursive estimator under bidirectional compression.

Distributed ImageNet training To further evaluate the proposed methods on deep learning tasks, we train ResNet-50 (He et al., 2016) and ViT-S (Dosovitskiy et al., 2021) models on ImageNet (Deng et al., 2009) using eight worker nodes. We report ImageNet validation accuracy in Table 1. On ResNet-50, DVR-Sign attains the highest accuracy, reaching 73.7% and exceeding the strongest baseline by 0.6%. At the same time, DVR-Q achieves 73.5%, corresponding to a 0.4% improvement over the same baseline. For ViT-S, DVR-Q achieves the highest accuracy at 73.9%, surpassing SSVR-MV by 0.7%, while DVR-Sign reaches 73.6%, also outperforming three other baselines.

AI USE STATEMENT

In this work, we used AI tools mainly for polishing writing, searching literature, and assisting in the writing of proofs. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2): 165–214, 2023.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

Evgenii Chzhen and Sholom Schechtman. SignSVRG: Fixing SignSGD via variance reduction. *arXiv preprint arXiv:2305.13187*, 2023.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Samuel Horvóth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtarik. Natural compression for distributed deep learning. In *Proceedings of Mathematical and Scientific Machine Learning*, pp. 129–141, 2022.

Wei Jiang, Sifan Yang, Wenhao Yang, and Lijun Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. In *Advances in Neural Information Processing Systems*, 2024.

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates. *arXiv preprint arXiv:2507.12091*, 2025.

Wei Jiang, Mao Xu, Wenhao Yang, Yibo Wang, Zechao Li, and Lijun Zhang. Convergence analysis of the Lion optimizer in centralized and distributed settings. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-sign SGD for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Kfir Levy, Ali Kavis, and Volkan Cevher. STORM+: Fully adaptive SGD with recursive momentum for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2021.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6286–6295, 2021.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2613–2621, 2017.

Zhen Qin, Zhishuai Liu, and Pan Xu. Convergence of sign-based random reshuffling algorithms for nonconvex optimization. *arXiv preprint arXiv:2310.15976*, 2026.

Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9224–9234, 2021.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 33077–33099, 2023.

A PROOF OF PROPOSITION 1

We first prove Proposition 1 for the SSVR-MV Option 1. Use the three one-dimensional functions

$$f_1(x) = f_2(x) = \tfrac{1}{2}\log\cosh x + \tfrac{1}{4}x, \qquad f_3(x) = \tfrac{1}{2}\log\cosh x - \tfrac{1}{2}x. \tag{A.1}$$

Since $\cosh x \geq e^{|x|}/2$, we have $\log\cosh x \geq |x| - \log 2$. It follows that

$$f_1(x) = f_2(x) \geq \tfrac{1}{2}|x| + \tfrac{1}{4}x - \tfrac{1}{2}\log 2 \geq -\tfrac{1}{2}\log 2, \quad f_3(x) \geq \tfrac{1}{2}(|x| - x) - \tfrac{1}{2}\log 2 \geq -\tfrac{1}{2}\log 2.$$

Thus each local objective is lower bounded. Their derivatives are

$$f_1'(x) = f_2'(x) = \tfrac{1}{2}\tanh x + \tfrac{1}{4}, \qquad f_3'(x) = \tfrac{1}{2}\tanh x - \tfrac{1}{2}.$$

Because $|\tanh x| \leq 1$, all three derivatives have absolute value at most $G = 1$. Moreover, $f_j''(x) = 1/(2\cosh^2 x) \in (0, 1/2]$, so local objectives are $1/2$-smooth. Their average and its derivative are

$$f(x) = \tfrac{1}{2}\log\cosh x, \qquad h(x) := f'(x) = \tfrac{1}{2}\tanh x.$$

In particular, we observe that $x_1 = 0$ is a global minimizer and $|h(x)| \leq 1/2$. Take the deterministic oracle $g_j(x;\xi) = f_j'(x)$ whose variance is zero.

Exact initialization gives $v_1^j = f_j'(x_1)$. If $v_{t-1}^j = f_j'(x_{t-1})$, the local recursion gives

$$v_t^j = f_j'(x_t) + (1-\beta)[f_j'(x_{t-1}) - f_j'(x_{t-1})] = f_j'(x_t).$$

Induction therefore proves exact local estimates at every iteration for every $\beta \in (0, 1]$. The signs $q_t^j = S_4(f_j'(x_t))$ are well defined because $|f_j'(x_t)| \leq 1 < 4$.

Next, to keep the following scalar calculation short, write $u = 1/16$ and $z = h(x)/4$. The three sign means at $x_t = x$ are $z + u, z + u, z - 2u$. Expanding the product,

$$(z+u)^2(z-2u) = (z^2 + 2uz + u^2)(z-2u) = z^3 - 3u^2 z - 2u^3.$$

Consequently the expected vote, denoted by $M(x)$, is

$$M(x) := \mathbb{E}[s_t \mid x_t = x] = \frac{3z - (z+u)^2(z-2u)}{2} = \frac{3(1+u^2)z - z^3 + 2u^3}{2}. \tag{A.2}$$

At the minimizer, $h(0) = z = 0$ and $M(0) = u^3 > 0$. The following derivative bound controls how much this bias can change when the true gradient is small:

$$\frac{dM}{dh} = \frac{3}{8}(1 + u^2 - z^2), \qquad 0 < \frac{dM}{dh} \leq \frac{3}{8}(1+u^2) = \frac{771}{2048}. \tag{A.3}$$

Indeed $|z| = |h|/4 \leq 1/8$, so $1 + u^2 - z^2 > 0$. The mean value theorem now gives

$$|M(x) - u^3| \leq \frac{771}{2048}|h(x)|. \tag{A.4}$$

Next, we control the iterate without assuming it is bounded. Since $\tanh 1 > 1/2$, $x \geq 1$ implies that $z \geq u$, and $x \leq -1$ implies $z \leq -u$. The polynomial in (A.2) is increasing in z on $[-1/8, 1/8]$. Its values at u and $-u$ are, respectively, $(3u + 4u^3)/2 > 0$ and $-3u/2 < 0$. Thus $xM(x) \geq 0$ for $|x| \geq 1$. For $|x| < 1$, the fact that $M(x)$ is the expectation of a sign gives $|M(x)| \leq 1$ and hence $xM(x) \geq -1$. Combining the two regions yields $xM(x) \geq -1$ for every x.

The update $x_{t+1} = x_t - \eta s_t$ and $s_t^2 = 1$ give

$$\mathbb{E}[x_{t+1}^2 \mid x_t] = x_t^2 - 2\eta x_t M(x_t) + \eta^2.$$

Take expectations and sum over $t = 1, \ldots, K$. Since $x_1 = 0$,

$$\mathbb{E}x_{K+1}^2 = \sum_{t=1}^{K}\left(-2\eta\mathbb{E}[x_t M(x_t)] + \eta^2\right) \leq 2\eta K + \eta^2 K. \tag{A.5}$$

All moments are finite, since for a fixed horizon, $|x_t| \leq (t-1)\eta$ on every sample path.

Taking expectations gives $\mathbb{E}x_{t+1} - \mathbb{E}x_t = -\eta\mathbb{E}M(x_t)$. After summing and using $x_1 = 0$,

$$\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}M(x_t) = -\frac{\mathbb{E}x_{K+1}}{\eta K}.$$

Cauchy–Schwarz and (A.5) imply

$$\left|\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}M(x_t)\right| \le \frac{\sqrt{\mathbb{E}x_{K+1}^2}}{\eta K} \le \sqrt{\frac{2}{\eta K} + \frac{1}{K}}. \tag{A.6}$$

Finally, the triangle inequality and (A.6) give

$$\begin{aligned} u^3 &\le \left|\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}M(x_t)\right| + \left|u^3 - \frac{1}{K}\sum_{t=1}^{K}\mathbb{E}M(x_t)\right| \\ &\le \sqrt{\frac{2}{\eta K} + \frac{1}{K}} + \frac{1}{K}\sum_{t=1}^{K}\mathbb{E}|u^3 - M(x_t)| \\ &\le \sqrt{\frac{2}{\eta K} + \frac{1}{K}} + \frac{771}{2048}\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}|h(x_t)|. \end{aligned}$$

By independent uniform output selection, the final average is $\mathbb{E}|f'(x_\tau)|$. Since $u^3 = 1/4096$, rearranging gives the finite-horizon bound

$$\mathbb{E}|f'(x_\tau)| \ge \frac{2048}{771}\left[\frac{1}{4096} - \sqrt{\frac{2}{\eta K} + \frac{1}{K}}\right]. \tag{A.7}$$

For $\eta = \Theta(K^{-1/2})$, we have $\eta K \to \infty$, so the square-root term tends to zero. Taking the lower limit and using $2048/(771 \cdot 4096) = 1/1542$ proves (15).

B Proofs for the stochastic results

We first prove a tracking estimate that retains the actual model movement. Throughout, $a = 1 + \omega$, $c = a/n$, K counts model updates, and B_0 is the initialization batch per worker.

B.1 Moments and server tracking

For $t \ge 2$, let $\mathcal{F}_{t-1}$ contain all randomness generated before the current worker samples and compression calls. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. First, unbiased compression controls second moments. Conditional on an input v, expansion of the squared norm gives

$$\mathbb{E}[\|Q(v)\|_2^2 \mid v] = \|v\|_2^2 + 2\langle v, \mathbb{E}[Q(v) - v \mid v]\rangle + \mathbb{E}[\|Q(v) - v\|_2^2 \mid v] \le a\|v\|_2^2.$$

The middle term is zero by unbiasedness. Applying $\|u+v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2$ to the decomposition in (16), we obtain

$$\begin{aligned} \mathbb{E}_t\|h_t^j\|_2^2 &\le 2\beta^2\mathbb{E}_t\|g_j(x_t;\xi_t^j)\|_2^2 + 2(1-\beta)^2\mathbb{E}_t\|g_j(x_t;\xi_t^j) - g_j(x_{t-1};\xi_t^j)\|_2^2 \\ &\le 2\beta^2 H^2 + 2L^2\|x_t - x_{t-1}\|_2^2. \end{aligned} \tag{B.1}$$

The last inequality uses mean-squared smoothness and $(1-\beta)^2 \le 1$.

Next, we control the initial error. Write $Y_{j,b} = Q(g_j(x_1;\xi_{1,b}^j))$ for the message from sample b at worker j. Repeated conditioning gives

$$\mathbb{E}Y_{j,b} = \nabla f_j(x_1), \qquad \mathbb{E}\|Y_{j,b}\|_2^2 \le aH^2.$$

Its centered second moment therefore satisfies

$$\mathbb{E}\|Y_{j,b} - \nabla f_j(x_1)\|_2^2 = \mathbb{E}\|Y_{j,b}\|_2^2 - \|\nabla f_j(x_1)\|_2^2 \le aH^2.$$

Distinct initialization messages use independent samples and compression calls. For $(j,b) \neq (k,r)$, their centered cross term is consequently

$$\mathbb{E}\langle Y_{j,b} - \nabla f_j(x_1), Y_{k,r} - \nabla f_k(x_1)\rangle = 0.$$

Expanding the squared norm of the average gives

$$E_1 = \frac{1}{n^2 B_0^2} \sum_{j=1}^{n} \sum_{b=1}^{B_0} \mathbb{E}\|Y_{j,b} - \nabla f_j(x_1)\|_2^2 \leq \frac{nB_0 aH^2}{n^2 B_0^2} = \frac{cH^2}{B_0}.$$

This explains why each initialization sample is compressed separately. We now derive the tracking recursion. For $t \geq 2$, define the centered decoded error

$$\delta_t^j = Q(h_t^j) - [\nabla f_j(x_t) - (1-\beta)\nabla f_j(x_{t-1})].$$

Unbiasedness of the oracle and of the compressor implies

$$\mathbb{E}_t Q(h_t^j) = \mathbb{E}_t h_t^j = \nabla f_j(x_t) - (1-\beta)\nabla f_j(x_{t-1}), \qquad \mathbb{E}_t \delta_t^j = 0.$$

Subtracting $\nabla f(x_t)$ from (17) gives the exact identity

$$e_t = (1-\beta)e_{t-1} + \frac{1}{n}\sum_{j=1}^{n} \delta_t^j. \tag{B.2}$$

The conditional squared norm expands as

$$\begin{aligned}\mathbb{E}_t\|e_t\|_2^2 = (1-\beta)^2\|e_{t-1}\|_2^2 + \frac{2(1-\beta)}{n}\sum_j \langle e_{t-1}, \mathbb{E}_t\delta_t^j\rangle \\ + \frac{1}{n^2}\sum_j \mathbb{E}_t\|\delta_t^j\|_2^2 + \frac{2}{n^2}\sum_{j<k}\mathbb{E}_t\langle\delta_t^j, \delta_t^k\rangle.\end{aligned} \tag{B.3}$$

The second term is zero because e_{t-1} is fixed under conditioning. The last term is zero because the current worker errors are conditionally independent and centered. For each remaining variance, centering and (B.1) give

$$\begin{aligned}\mathbb{E}_t\|\delta_t^j\|_2^2 &= \mathbb{E}_t\|Q(h_t^j)\|_2^2 - \|\nabla f_j(x_t) - (1-\beta)\nabla f_j(x_{t-1})\|_2^2 \\ &\leq a\mathbb{E}_t\|h_t^j\|_2^2 \leq 2a\beta^2 H^2 + 2aL^2\|x_t - x_{t-1}\|_2^2.\end{aligned} \tag{B.4}$$

Taking full expectations and using $(1-\beta)^2 \leq 1-\beta$ yields

$$E_t \leq (1-\beta)E_{t-1} + 2c\beta^2 H^2 + 2cL^2\mathbb{E}\|x_t - x_{t-1}\|_2^2. \tag{B.5}$$

For completeness, summing this recursion from $t=2$ to K gives

$$\begin{aligned}\beta\sum_{t=1}^{K} E_t &\leq E_1 - (1-\beta)E_K + 2c\beta^2 H^2(K-1) + 2cL^2\sum_{t=1}^{K-1}\mathbb{E}\|x_{t+1} - x_t\|_2^2 \\ &\leq E_1 + 2c\beta^2 H^2 K + 2cL^2\sum_{t=1}^{K-1}\mathbb{E}\|x_{t+1} - x_t\|_2^2.\end{aligned} \tag{B.6}$$

After division by βK, this proves

$$\overline{E}_K \leq \frac{cH^2}{B_0\beta K} + 2c\beta H^2 + \frac{2cL^2}{\beta K}\sum_{t=1}^{K-1}\mathbb{E}\|x_{t+1} - x_t\|_2^2. \tag{B.7}$$

For DVR-Sign, $\|x_{t+1} - x_t\|_2^2 = \eta^2 d$. Substituting this identity into (B.5) and (B.7) proves (23) and (24). The DVR-Q specialization is given separately in Appendix B.3.

B.2 THE ℓ_1 GUARANTEE

Let g be a true gradient and z an arbitrary estimate. If $\mathrm{Sign}(g_k) \neq \mathrm{Sign}(z_k)$ and $g_k \neq 0$, then z_k is on the opposite side of zero or is zero, and therefore $|g_k| \leq |g_k - z_k|$. Coordinates with $g_k = 0$ contribute zero. It follows that

$$\begin{aligned}\|g\|_1 - \langle g, \mathrm{Sign}(z)\rangle &= 2\sum_{k=1}^{d} |g_k| \mathbf{1}\{\mathrm{Sign}(g_k) \neq \mathrm{Sign}(z_k)\} \\ &\leq 2\sum_{k=1}^{d} |g_k - z_k| = 2\|g - z\|_1 \leq 2\sqrt{d}\|g - z\|_2. \end{aligned} \tag{B.8}$$

Next, we sum the objective decreases. Mean-squared smoothness and Jensen's inequality give

$$\|\nabla f_j(x) - \nabla f_j(y)\|_2 = \|\mathbb{E}[g_j(x;\xi) - g_j(y;\xi)]\|_2 \leq \left(\mathbb{E}\|g_j(x;\xi) - g_j(y;\xi)\|_2^2\right)^{1/2} \leq L\|x - y\|_2.$$

Averaging over workers shows that f is also L-smooth. Applying its smoothness inequality to the deterministic sign update gives

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \eta\langle \nabla f(x_t), \mathrm{Sign}(z_t)\rangle + \frac{L\eta^2}{2}\|\mathrm{Sign}(z_t)\|_2^2 \\ &\leq f(x_t) - \eta\|\nabla f(x_t)\|_1 + 2\eta\sqrt{d}\|e_t\|_2 + \frac{L\eta^2 d}{2}. \end{aligned} \tag{B.9}$$

Taking expectations and summing from $t = 1$ to K, we have

$$\eta\sum_{t=1}^{K} \mathbb{E}\|\nabla f(x_t)\|_1 \leq \Delta_f + 2\eta\sqrt{d}\sum_{t=1}^{K} \mathbb{E}\|e_t\|_2 + \frac{KL\eta^2 d}{2}.$$

For the error sum, Jensen's inequality followed by Cauchy–Schwarz gives

$$\frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|e_t\|_2 \leq \frac{1}{K}\sum_{t=1}^{K} \sqrt{E_t} \leq \frac{1}{K}\sqrt{K\sum_{t=1}^{K} E_t} = \sqrt{\overline{E}_K}.$$

Dividing the previous descent inequality by ηK proves

$$\frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|\nabla f(x_t)\|_1 \leq \frac{\Delta_f}{\eta K} + \frac{L\eta d}{2} + 2\sqrt{d\,\overline{E}_K}. \tag{B.10}$$

Because τ is uniform on $\{1, \ldots, K\}$ and independent of the run, its expected gradient norm equals the average on the left. This proves Theorem 2. For Corollary 3, choose

$$u_1 = \frac{1}{\sqrt{K} + c^{1/3}K^{2/3}}, \qquad \beta = u_1, \quad \eta = \frac{Hu_1}{L\sqrt{d}}, \quad B_0 = \left\lceil \frac{1}{u_1^2 K} \right\rceil. \tag{B.11}$$

These are admissible because $0 < u_1 \leq K^{-1/2} \leq 1$ and $B_0 \geq 1$. Inserting the parameters into (24) controls each contribution separately:

$$\frac{cH^2}{B_0\beta K} \leq cH^2 u_1, \qquad 2cH^2\beta = 2cH^2 u_1, \qquad \frac{2cL^2\eta^2 d}{\beta} = 2cH^2 u_1.$$

Thus $\overline{E}_K \leq 5cH^2 u_1$, and (B.10) becomes

$$\mathbb{E}\|\nabla f(x_\tau)\|_1 \leq \sqrt{d}\left[\frac{L\Delta_f}{Hu_1 K} + \frac{Hu_1}{2} + 2\sqrt{5}H\sqrt{cu_1}\right].$$

The definition of u_1 gives

$$\frac{1}{u_1 K} = K^{-1/2} + c^{1/3}K^{-1/3}, \qquad u_1 \leq K^{-1/2}, \qquad u_1 \leq c^{-1/3}K^{-2/3}.$$

In particular, the third inequality implies $\sqrt{cu_1} \le c^{1/3}K^{-1/3}$. Substitution proves the explicit bound

$$\mathbb{E}\|\nabla f(x_\tau)\|_1 \le \sqrt{d}\left[\left(\frac{L\Delta_f}{H}+\frac{H}{2}\right)K^{-1/2}+\left(\frac{L\Delta_f}{H}+2\sqrt{5}H\right)\left(\frac{c}{K}\right)^{1/3}\right]. \tag{B.12}$$

Keeping L, H, Δ_f fixed and recalling $c = (1+\omega)/n$ gives (27).

Finally, we specify initialization and a target accuracy. Squaring the denominator of u_1 yields the exact batch size

$$B_0 = \left\lceil (1 + c^{1/3}K^{1/6})^2 \right\rceil = \Theta(1 + c^{2/3}K^{1/3}). \tag{B.13}$$

To verify the order including the ceiling, use $1 + b^2 \le (1+b)^2 \le 2(1+b^2)$ for $b \ge 0$, and $y \le \lceil y \rceil \le y + 1 \le 2y$ for $y \ge 1$. For any $\epsilon > 0$, the following explicit integer horizon suffices:

$$K = \max\left\{1,\ \left\lceil \frac{4d(L\Delta_f/H + H/2)^2}{\epsilon^2} \right\rceil,\ \left\lceil \frac{8cd^{3/2}(L\Delta_f/H + 2\sqrt{5}H)^3}{\epsilon^3} \right\rceil\right\}. \tag{B.14}$$

The second entry makes the first term of (B.12) at most $\epsilon/2$: square that desired inequality and solve for K. The third entry does the same for the second term by cubing it. Their sum is therefore at most ϵ. With fixed L, H, Δ_f,

$$K = O\left(1 + \frac{d}{\epsilon^2} + \frac{cd^{3/2}}{\epsilon^3}\right). \tag{B.15}$$

B.3 The ℓ_2 guarantee

Let $\mathcal{G}_t$ denote the history after the server has formed z_t and before drawing that downlink. Thus

$$\mathbb{E}[Q(z_t) \mid \mathcal{G}_t] = z_t, \qquad \mathbb{E}[\|Q(z_t)\|_2^2 \mid \mathcal{G}_t] \le a\|z_t\|_2^2. \tag{B.16}$$

For clarity, write $Z_t = \mathbb{E}\|z_t\|_2^2$ only within this proof. The actual movement satisfies

$$\mathbb{E}\|x_{t+1} - x_t\|_2^2 = \eta^2\mathbb{E}\|Q(z_t)\|_2^2 \le a\eta^2 Z_t.$$

Inserting it into the general tracking bounds gives

$$E_t \le (1-\beta)E_{t-1} + 2c\beta^2H^2 + 2caL^2\eta^2 Z_{t-1} \quad (t \ge 2), \tag{B.17}$$

$$\overline{E}_K \le \frac{cH^2}{B_0\beta K} + 2c\beta H^2 + \frac{2caL^2\eta^2}{\beta K}\sum_{t=1}^{K-1} Z_t. \tag{B.18}$$

Unlike the sign update, this movement becomes small when the tracker becomes small. We must retain its dependence on Z_t in the descent calculation.

By L-smoothness and (B.16),

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{G}_t] \le f(x_t) - \eta\langle \nabla f(x_t), z_t\rangle + \frac{La\eta^2}{2}\|z_t\|_2^2. \tag{B.19}$$

The exact inner-product identity

$$2\langle \nabla f(x_t), z_t\rangle = \|\nabla f(x_t)\|_2^2 + \|z_t\|_2^2 - \|e_t\|_2^2$$

then gives, after taking full expectations,

$$\mathbb{E}f(x_{t+1}) \le \mathbb{E}f(x_t) - \frac{\eta}{2}\mathbb{E}\|\nabla f(x_t)\|_2^2 - \frac{\eta}{2}(1 - La\eta)Z_t + \frac{\eta}{2}E_t. \tag{B.20}$$

Summing from $t = 1$ to K, using $\mathbb{E}f(x_{K+1}) \ge f_*$, and dividing by $\eta K/2$ yields

$$\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|\nabla f(x_t)\|_2^2 + \frac{1 - La\eta}{K}\sum_{t=1}^{K} Z_t \le \frac{2\Delta_f}{\eta K} + \overline{E}_K. \tag{B.21}$$

Substitute (B.18) and use $\sum_{t=1}^{K-1} Z_t \le \sum_{t=1}^{K} Z_t$. Moving this last contribution to the left gives

$$\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|\nabla f(x_t)\|_2^2 + \left(1 - La\eta - \frac{2caL^2\eta^2}{\beta}\right)\frac{1}{K}\sum_{t=1}^{K} Z_t \le \frac{2\Delta_f}{\eta K} + \frac{cH^2}{B_0\beta K} + 2c\beta H^2. \tag{B.22}$$

The independent uniform output τ turns the first average into $\mathbb{E}\|\nabla f(x_\tau)\|_2^2$. If $La\eta \le 1/2$ and $\beta \ge 4caL^2\eta^2$, both subtracted terms in the tracker coefficient are at most $1/2$, so the coefficient is nonnegative. This proves Theorem 4. The cancellation explains why we did not replace Z_t by a gradient-magnitude bound earlier.

We next prove Corollary 5. Recall $c = a/n$ and choose

$$r = (K/n^2)^{1/3}, \qquad \eta = \frac{1}{2La(1+r)}, \quad \beta = \frac{1}{n(1+r)^2}, \quad B_0 = \lceil 1+r \rceil. \tag{B.23}$$

These choices satisfy $0 < \beta \le 1$ for every $n, K \ge 1$. Direct substitution gives

$$La\eta = \frac{1}{2(1+r)}, \qquad 4caL^2\eta^2 = \frac{1}{n(1+r)^2} = \beta.$$

The tracker coefficient in (28) is consequently

$$1 - \frac{1}{2(1+r)} - \frac{1}{2} = \frac{r}{2(1+r)} \ge 0.$$

We can drop that term and bound the three remaining terms separately:

$$\begin{aligned} \frac{2\Delta_f}{\eta K} &= \frac{4La\Delta_f(1+r)}{K}, \\ \frac{cH^2}{B_0\beta K} &= \frac{aH^2(1+r)^2}{B_0 K} \le \frac{aH^2(1+r)}{K}, \\ 2c\beta H^2 &= \frac{2aH^2}{n^2(1+r)^2} = \frac{2aH^2 r^3}{K(1+r)^2} \le \frac{2aH^2 r}{K}. \end{aligned} \tag{B.24}$$

The second bound uses $B_0 \ge 1 + r$. The last uses $r^3 = K/n^2$ and $r^2 \le (1+r)^2$. Adding the bounds and using $r/K = (nK)^{-2/3}$ proves

$$\mathbb{E}\|\nabla f(x_\tau)\|_2^2 \le a\left[\frac{4L\Delta_f + H^2}{K} + \frac{4L\Delta_f + 3H^2}{(nK)^{2/3}}\right]. \tag{B.25}$$

Jensen's inequality and $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$ give

$$\begin{aligned} \mathbb{E}\|\nabla f(x_\tau)\|_2 &\le \sqrt{\mathbb{E}\|\nabla f(x_\tau)\|_2^2} \\ &\le \sqrt{4L\Delta_f + H^2}\sqrt{\frac{a}{K}} + \sqrt{4L\Delta_f + 3H^2}\,\frac{\sqrt{a}}{(nK)^{1/3}}. \end{aligned} \tag{B.26}$$

This proves (30). To obtain $\mathbb{E}\|\nabla f(x_\tau)\|_2^2 \le \epsilon^2$, and hence $\mathbb{E}\|\nabla f(x_\tau)\|_2 \le \epsilon$, it suffices to choose

$$K = \max\left\{1, \left\lceil \frac{2a(4L\Delta_f + H^2)}{\epsilon^2} \right\rceil, \left\lceil \frac{[2a(4L\Delta_f + 3H^2)]^{3/2}}{n\epsilon^3} \right\rceil\right\}. \tag{B.27}$$

The second entry makes the first term in (B.25) at most $\epsilon^2/2$. Raising the desired bound on its second term to the power $3/2$ gives the third entry. Their sum is at most ϵ^2. For fixed L, H, Δ_f, the resulting update and per-worker sample complexities are

$$K = O\left(1 + \frac{a}{\epsilon^2} + \frac{a^{3/2}}{n\epsilon^3}\right). \tag{B.28}$$

C Proofs for finite sums

C.1 Exact computation

For K updates and an integer refresh period $q \ge 1$, the times $1, 1+q, 1+2q, \dots$ give

$$r = 1 + \left\lfloor \frac{K-1}{q} \right\rfloor = \left\lceil \frac{K}{q} \right\rceil.$$

Each refresh evaluates every component once for $M = nm$ evaluations. At each of the $K - r$ ordinary updates, every worker evaluates one component at two points. Therefore

$$N_{\mathrm{grad}} = Mr + 2n(K - r), \qquad N_{\mathrm{grad},j} = mr + 2(K - r). \tag{C.1}$$

For $q = m$, the inequality $r \le 1 + K/m$ gives

$$N_{\mathrm{grad}} \le M\left(1 + \frac{K}{m}\right) + 2nK = M + 3nK. \tag{C.2}$$

C.2 Tracking error between exact refreshes

Lemma C.1. *For the estimator* (32)*, the error is zero at each refresh $t_0 = 1 + kq$. For $t_0 \le t < t_0 + q$, we have*

$$\mathbb{E}\|z_t - \nabla f(x_t)\|_2^2 \le \frac{aL^2}{n} \sum_{s=t_0+1}^{t} \mathbb{E}\|x_s - x_{s-1}\|_2^2. \tag{C.3}$$

Deterministic sign updates therefore give the bound $aL^2\eta^2 d(q-1)/n$. For updates $x_{t+1} = x_t - \eta Q(z_t)$, with a fresh common downlink satisfying the same compressor assumption,

$$\frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|z_t - \nabla f(x_t)\|_2^2 \le \frac{a^2L^2\eta^2(q-1)}{n} \frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|z_t\|_2^2. \tag{C.4}$$

Proof. At an ordinary iteration, condition on the history through x_t and before the current component indices and compression calls. Denote this conditional expectation by $\mathbb{E}_t$; the points x_t, x_{t-1} and the previous estimate z_{t-1} are fixed under this conditioning. Write

$$X_j = Q\big(\nabla f_{j,i_t^j}(x_t) - \nabla f_{j,i_t^j}(x_{t-1})\big),$$
$$\mu_j = \mathbb{E}_t X_j = \nabla f_j(x_t) - \nabla f_j(x_{t-1}).$$

Uniform component sampling and unbiased compression give the expression for μ_j. The same component is evaluated at both points. Thus

$$\mathbb{E}_t\|X_j\|_2^2 \le \frac{a}{m}\sum_{i=1}^{m}\|\nabla f_{j,i}(x_t) - \nabla f_{j,i}(x_{t-1})\|_2^2 \le aL^2\|x_t - x_{t-1}\|_2^2.$$

The first inequality uses (7); the second is component smoothness. Since $X_j - \mu_j$ is centered,

$$\mathbb{E}_t\|X_j - \mu_j\|_2^2 = \mathbb{E}_t\|X_j\|_2^2 - \|\mu_j\|_2^2 \le aL^2\|x_t - x_{t-1}\|_2^2.$$

Note that the component indices and compressor calls are conditionally independent across all n workers. For $j \ne k$, this independence and centering imply

$$\mathbb{E}_t\langle X_j - \mu_j, X_k - \mu_k\rangle = \langle \mathbb{E}_t(X_j - \mu_j), \mathbb{E}_t(X_k - \mu_k)\rangle = 0.$$

Consequently,

$$\mathbb{E}_t\left\|\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_j)\right\|^2 = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}_t\|X_j - \mu_j\|_2^2 \le \frac{aL^2}{n}\|x_t - x_{t-1}\|_2^2.$$

The averaging is over all workers and $n^{-1}\sum_j \mu_j = \nabla f(x_t) - \nabla f(x_{t-1})$ exactly.

Next, let $e_t = z_t - \nabla f(x_t)$. The estimator recursion gives

$$e_t = e_{t-1} + \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_j).$$

The new sum has zero conditional mean, and e_{t-1} is fixed under $\mathbb{E}_t$. Expanding the square yields

$$\begin{aligned}\mathbb{E}_t\|e_t\|_2^2 &= \|e_{t-1}\|_2^2 + 2\left\langle e_{t-1}, \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_t(X_j - \mu_j)\right\rangle + \mathbb{E}_t\left\|\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_j)\right\|^2 \\ &\le \|e_{t-1}\|_2^2 + \frac{aL^2}{n}\|x_t - x_{t-1}\|_2^2.\end{aligned}$$

At a refresh, $z_{t_0} = \nabla f(x_{t_0})$, so $e_{t_0} = 0$ on every sample path. Taking total expectations and applying the recursion successively at $t_0 + 1, \dots, t$ gives

$$\mathbb{E}\|e_t\|_2^2 \le \mathbb{E}\|e_{t_0}\|_2^2 + \frac{aL^2}{n} \sum_{s=t_0+1}^{t} \mathbb{E}\|x_s - x_{s-1}\|_2^2.$$

This proves (C.3). For DVR-Sign-FS, each summand is $\eta^2 d$, and there are at most $q-1$ summands.

For DVR-Q-FS, condition on the history $\mathcal{G}_s$ after all uplinks in iteration s and before the server's new compression call. Then

$$\mathbb{E}[\|x_{s+1} - x_s\|_2^2 \mid \mathcal{G}_s] = \eta^2 \mathbb{E}[\|Q(z_s)\|_2^2 \mid \mathcal{G}_s] \le a\eta^2 \|z_s\|_2^2.$$

Substituting its total expectation in (C.3) yields

$$\mathbb{E}\|e_t\|_2^2 \le \frac{a^2 L^2 \eta^2}{n} \sum_{s=t_0}^{t-1} \mathbb{E}\|z_s\|_2^2.$$

Let $t_1 = \min\{t_0 + q - 1, K\}$ be the last iteration in this interval between refreshes. Reversing the order of the finite sums gives

$$\begin{aligned}\sum_{t=t_0}^{t_1} \mathbb{E}\|e_t\|_2^2 &\le \frac{a^2 L^2 \eta^2}{n} \sum_{s=t_0}^{t_1-1} (t_1 - s)\mathbb{E}\|z_s\|_2^2 \\ &\le \frac{a^2 L^2 \eta^2 (q-1)}{n} \sum_{s=t_0}^{t_1} \mathbb{E}\|z_s\|_2^2.\end{aligned}$$

These intervals are disjoint. Summing over them and dividing by K proves (C.4). □

C.3 Descent for the ℓ_1 criterion

Proof of Theorem 6. The deterministic sign $s_t = \mathrm{Sign}(z_t)$ has squared norm d. Thus the deterministic sign case of Lemma C.1 proves (34). To translate this into stationarity, smoothness gives

$$f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), \mathrm{Sign}(z_t) \rangle + \frac{L\eta^2 d}{2}.$$

By (19),

$$\langle \nabla f(x_t), \mathrm{Sign}(z_t) \rangle \ge \|\nabla f(x_t)\|_1 - 2\sqrt{d}\|z_t - \nabla f(x_t)\|_2.$$

Substituting and rearranging before taking expectations gives

$$\eta \mathbb{E}\|\nabla f(x_t)\|_1 \le \mathbb{E} f(x_t) - \mathbb{E} f(x_{t+1}) + 2\eta\sqrt{d}\mathbb{E}\|z_t - \nabla f(x_t)\|_2 + \frac{L\eta^2 d}{2}.$$

Sum this inequality over $t = 1, \dots, K$. The objective terms telescope to $f(x_1) - \mathbb{E} f(x_{K+1}) \le f(x_1) - f_* = \Delta_f$. After division by ηK, we obtain

$$\begin{aligned}\frac{1}{K}\sum_{t=1}^{K} \mathbb{E}\|\nabla f(x_t)\|_1 &\le \frac{\Delta_f}{\eta K} + \frac{L\eta d}{2} + \frac{2\sqrt{d}}{K}\sum_{t=1}^{K} \mathbb{E}\|z_t - \nabla f(x_t)\|_2 \\ &\le \frac{\Delta_f}{\eta K} + \frac{L\eta d}{2} + 2\sqrt{d}\sqrt{\frac{aL^2\eta^2 d(q-1)}{n}} \\ &= \frac{\Delta_f}{\eta K} + L\eta\left[\frac{d}{2} + 2d\sqrt{\frac{a(q-1)}{n}}\right].\end{aligned}$$

In the second line, $\mathbb{E}\|z_t - \nabla f(x_t)\|_2 \le \sqrt{\mathbb{E}\|z_t - \nabla f(x_t)\|_2^2}$ and Lemma C.1 bound each summand. Uniformity of τ implies that the average on the left equals $\mathbb{E}\|\nabla f(x_\tau)\|_1$. This proves (35).

Finally, $J_1 \ge d/2 > 0$, so the positive stepsize is well defined and $L\eta J_1 = \epsilon/2$. If $\Delta_f > 0$, the choice of K gives

$$\frac{\Delta_f}{\eta K} = \frac{2L\Delta_f J_1}{\epsilon K} \le \frac{\epsilon}{2}.$$

□

Complexity Set $q = m$. The target-accuracy choice gives, for every $\epsilon > 0$,

$$K \le 1 + \frac{4L\Delta_f J_1}{\epsilon^2} = 1 + \frac{4L\Delta_f d}{\epsilon^2}\left[\frac{1}{2} + 2\sqrt{\frac{a(m-1)}{n}}\right].$$

For the total oracle count, multiply the coefficient by n:

$$nJ_1 = \frac{dn}{2} + 2d\sqrt{an(m-1)} \le \frac{dn}{2} + 2d\sqrt{aM}. \tag{C.5}$$

The inequality uses $n(m-1) \le nm = M$. Substituting the iteration bound into (C.2) now gives

$$\begin{aligned} N_{\mathrm{grad}} &\le M + 3n + \frac{12L\Delta_f nJ_1}{\epsilon^2} \\ &\le 4M + \frac{12L\Delta_f d}{\epsilon^2}\left[\frac{n}{2} + 2\sqrt{aM}\right]. \end{aligned}$$

Here $n \le M$ absorbs the extra update arising from the ceiling.

C.4 Descent for the ℓ_2 criterion

Proof of Theorem 8. Condition on the post-uplink history $\mathcal{G}_t$, which includes x_t and z_t and precedes the server's new call to Q. Unbiasedness and the second-moment bound give

$$\mathbb{E}[Q(z_t) \mid \mathcal{G}_t] = z_t, \qquad \mathbb{E}[\|Q(z_t)\|_2^2 \mid \mathcal{G}_t] \le a\|z_t\|_2^2.$$

Since $x_{t+1} = x_t - \eta Q(z_t)$, smoothness implies

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{G}_t] \le f(x_t) - \eta\langle \nabla f(x_t), z_t\rangle + \frac{La\eta^2}{2}\|z_t\|_2^2.$$

The inner product has the exact decomposition

$$2\langle \nabla f(x_t), z_t\rangle = \|\nabla f(x_t)\|_2^2 + \|z_t\|_2^2 - \|z_t - \nabla f(x_t)\|_2^2.$$

Substitute this identity, take total expectations, and rearrange:

$$\frac{\eta}{2}\mathbb{E}\|\nabla f(x_t)\|_2^2 + \frac{\eta}{2}(1 - La\eta)\mathbb{E}\|z_t\|_2^2 \le \mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1}) + \frac{\eta}{2}\mathbb{E}\|z_t - \nabla f(x_t)\|_2^2.$$

Sum the last inequality over $t = 1, \dots, K$ and divide by $\eta K/2$. The objective values telescope, and $f(x_1) - \mathbb{E}f(x_{K+1}) \le \Delta_f$. Applying (C.4) to the right side gives

$$\begin{aligned} \frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|\nabla f(x_t)\|_2^2 + \left[1 - La\eta - \frac{a^2L^2\eta^2(q-1)}{n}\right]\frac{1}{K}\sum_{t=1}^{K}\mathbb{E}\|z_t\|_2^2 \\ \le \frac{2\Delta_f}{\eta K}. \end{aligned} \tag{C.6}$$

Write $s = \sqrt{(q-1)/n} \ge 0$. The choice in (38) gives

$$La\eta + \frac{a^2L^2\eta^2(q-1)}{n} = \frac{1}{2(1+s)} + \frac{s^2}{4(1+s)^2} = \frac{1}{4} + \frac{1}{4(1+s)^2} \le \frac{1}{2}.$$

Thus the bracket in (C.6) is at least $1/2$, and we may discard its nonnegative contribution. Since $2\Delta_f/(\eta K) = 4L\Delta_f J_Q/K$, we obtain the squared-norm bound in (39). The independent uniform choice of τ identifies its expectation with the averaged left side. Cauchy–Schwarz then gives

$$\mathbb{E}\|\nabla f(x_\tau)\|_2 \le \sqrt{\mathbb{E}\|\nabla f(x_\tau)\|_2^2} \le 2\sqrt{\frac{L\Delta_f J_Q}{K}}.$$

□

Complexity For $q = m$, we have $J_Q = a(1 + \sqrt{(m-1)/n})$. Hence

$$K \le 1 + \frac{4L\Delta_f J_Q}{\epsilon^2} = 1 + \frac{4aL\Delta_f}{\epsilon^2}\left[1 + \sqrt{\frac{m-1}{n}}\right].$$

To count gradients, use

$$nJ_Q = a\left[n + \sqrt{n(m-1)}\right] \le a(n + \sqrt{M}).$$

Then (C.2) yields

$$\begin{aligned} N_{\mathrm{grad}} &\le M + 3n + \frac{12L\Delta_f n J_Q}{\epsilon^2} \\ &\le 4M + \frac{12aL\Delta_f}{\epsilon^2}(n + \sqrt{M}). \end{aligned}$$