SHARP OUTLIER AMPLIFICATION IN ERROR-FEEDBACK CLIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient clipping limits the immediate effect of an unusually large gradient. Error feedback stores the discarded part and reintroduces it in later updates. How much of clipping's protection survives this memory? We characterize the maximum subsequent loss after an isolated outlier for residual error-feedback clipping. On a scalar quadratic, an outlier of magnitude A produces a peak loss asymptotic to ηcA, despite every update having norm at most ηc. We obtain this sharp law from a support-function Lyapunov inequality that applies to projections onto arbitrary compact convex update sets. The same inequality shows that capping the stored residual preserves exact-gradient convergence for every step size $0 < \eta < 2/L$. It also gives an amplitude-independent bound after finitely many arbitrary corruptions. For a Euclidean update threshold c and residual threshold B, the single-outlier bound is $L\eta^2c^2/2 + \eta cB$; the dependence on B is asymptotically sharp on quadratics. An explicit conservation identity identifies the gradient mass sacrificed by this protection. Short, reproducible computations illustrate the bounds and their limitations.

1 INTRODUCTION

Gradient clipping and error feedback solve different problems. Clipping limits the magnitude of an update. Error feedback preserves information removed by a biased transformation by storing its residual and adding it to future inputs. Both operations are useful in optimization, but their combination raises a basic stability question: can an observation that was clipped at one iteration still cause a large excursion many iterations later?

This question concerns optimizer state as well as the update that is applied immediately. Consider the residual recursion

$$u_t = \mathrm{clip}_c(g_t + e_t), \qquad x_{t+1} = x_t - \eta u_t, \qquad e_{t+1} = g_t + e_t - u_t, \tag{1}$$

where clip_c projects onto the Euclidean ball of radius c. A large g_t cannot move x_t by more than ηc, but it can write an arbitrarily large value into e_{t+1}. The residual can then sustain updates against the current gradient. A bound on $\|x_{t+1} - x_t\|$ therefore leaves the maximum subsequent loss unresolved.

We study this distinction through a controlled intervention: start at a minimizer, replace one gradient by an arbitrary finite vector, and subsequently use exact gradients. This experiment isolates memory from persistent sampling noise and changing data. Our principal negative result is quantitative. For $f(x) = \lambda x^2/2$ and fixed $0 < \eta\lambda < 2$, an outlier of magnitude A gives

$$\sup_{t \geq 0} f(x_t) \sim \eta cA \qquad (A/c \to \infty). \tag{2}$$

Ordinary clipping, under the same intervention, has a peak bounded independently of A. Thus, residual feedback can convert outlier magnitude into future loss even on a smooth, strongly convex objective.

Our analysis also identifies a direct remedy and its cost. Clip the *stored* residual after each update. This operation bounds the energy available for a later excursion. It preserves convergence with exact gradients, but can discard information permanently. We quantify both properties rather than interpreting them as a general improvement in statistical estimation.

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

The results have three components. First, we give finite-amplitude upper and lower bounds for the quadratic intervention; their leading constants agree. Second, we establish the Lyapunov inequality behind these bounds for any compact convex update set containing a neighborhood of zero. It yields a deterministic stationarity guarantee without bounded-gradient assumptions or a global contraction property of the clipping operator. Third, we bound the effect of any finite number of corrupted gradients independently of their magnitudes, and account exactly for the information discarded by the residual cap. The numerical work tests these statements directly and takes seconds on a CPU.

Scope. The object of study is the residual recursion (1) applied to one aggregate gradient. This is different from clipping per-example gradients before aggregation, from distributed gradient tracking, and from adding a separately clipped residual to a separately clipped gradient. The theorems do not cover Adam, momentum, arbitrary communication compressors, or differential privacy. The distinction is essential when comparing error-feedback methods.

2 Related work

Clipping and stochastic bias. Zhang et al. (2020) explain advantages of clipping under gradient-dependent smoothness. Koloskova et al. (2023) characterize convergence and unavoidable stochastic bias under fixed clipping thresholds. These works establish that clipping has different consequences for deterministic and stochastic gradients. We instead hold global smoothness fixed and quantify the future loss caused by memory of an isolated corrupted gradient. We do not remove the stochastic bias of fixed-threshold clipping.

Residual feedback and gradient tracking. Karimireddy et al. (2019) analyze residual error feedback for biased gradient compression. Their standard relative-contraction framework does not apply globally to fixed-radius clipping: $\|v - \mathrm{clip}_c(v)\| / \|v\| \to 1$ as $\|v\| \to \infty$. Clip21 uses clipped differences to track local gradients (Khirirat et al., 2023); subsequent work combines tracking, clipping, and momentum (Islamov et al., 2025), or uses smoothed clipping under symmetric heavy-tailed noise (Yu et al., 2026). These are different state recursions. In particular, Clip21's single-node update is $v_t = v_{t-1} + \mathrm{clip}_c(g_t - v_{t-1})$, not (1). Its convergence results are not contradicted by our outlier construction.

DiceSGD is especially relevant because it clips a feedback signal (Zhang et al., 2024). It adds separately clipped gradients and residuals while retaining an uncapped accumulated error. Our cap instead truncates the state written to memory and explicitly records the resulting loss of conservation. Residual clipping also appears as a component of federated gradient sparsification (Abbas et al., 2026). We do not claim that modifying or clipping feedback is itself a new idea. The contribution is the sharp transient-loss characterization and the accompanying support-function analysis for the specified recursion.

Optimizer state and saturation. Failures caused by optimizer memory have motivated precise counterexamples and targeted corrections, for example in adaptive optimization (Reddi et al., 2018). In control, saturation combined with integral state motivates anti-windup designs (Cao et al., 2004). This analogy helps explain the mechanism here; our results concern discrete gradient iterations and give loss bounds, stationarity guarantees, and an explicit quadratic lower bound. The numerical studies are diagnostic, not evidence of improved performance in large-scale learning.

3 Setup and residual capping

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, lower bounded by $f_\star \in \mathbb{R}$, and L-smooth:

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\|, \qquad x, y \in \mathbb{R}^d. \tag{3}$$

Write the available gradient as $g_t = \nabla f(x_t) + \xi_t$. The errors ξ_t are deterministic or random finite vectors; no unbiasedness is imposed unless stated. Fix $0 < \eta < 2/L$ and set $\alpha = 1 - L\eta/2 > 0$.

For Euclidean clipping, we study the family indexed by $B \in [0, \infty]$:

$$\begin{aligned} v_t &= g_t + e_t, \quad u_t = \mathrm{clip}_c(v_t), \quad r_t = v_t - u_t, \\ x_{t+1} &= x_t - \eta u_t, \quad e_{t+1} = \mathrm{clip}_B(r_t), \quad d_t = r_t - e_{t+1}. \end{aligned} \tag{4}$$

Here $\mathrm{clip}_0(v) = 0$ and $\mathrm{clip}_\infty(v) = v$. The vector d_t is the discarded residual. With $e_0 = 0$, $B = 0$ gives ordinary clipped gradient descent and $B = \infty$ gives (1). A finite positive B retains part of the clipping error. All variants use the same η and c.

The general formulation replaces the ball by a compact convex set K such that

$$\rho\mathbb{B}_2 \subseteq K \subseteq R\mathbb{B}_2, \qquad 0 < \rho \le R < \infty. \tag{5}$$

Let P_K be Euclidean projection and $h_K(z) = \max_{u\in K} \langle u, z\rangle$ the support function. Given a memory budget $b \in [0, \infty]$, use

$$\begin{aligned} u_t &= \mathrm{P}_K(g_t + e_t), \quad r_t = g_t + e_t - u_t, \\ e_{t+1} &= \beta_t r_t, \qquad \beta_t = \begin{cases} 1, & h_K(r_t) \le b, \\ b/h_K(r_t), & h_K(r_t) > b, \end{cases} \\ d_t &= (1-\beta_t) r_t, \qquad x_{t+1} = x_t - \eta u_t. \end{aligned} \tag{6}$$

For $K = c\mathbb{B}_2$, $h_K(z) = c\,\|z\|$ and $b = cB$, so (6) is exactly (4). Coordinate clipping corresponds to a box, with $h_K(z) = \sum_i c_i|z_i|$. Clipping separate blocks corresponds to a product of balls, with $h_K(z) = \sum_j c_j\,\|z_j\|$. The support budget is a single budget across coordinates or blocks. It is not an arbitrary coordinatewise modification of the residual.

4 A SHARP LOSS SPIKE FROM ONE OUTLIER

We first state the result for which all constants can be exposed. Define

$$H_a(k) = c\left(k + \frac{a}{2}k(k-1)\right), \qquad k_a(z) = \max\{k \in \mathbb{Z}_{\ge 0} : H_a(k) \le z\}. \tag{7}$$

For $a > 0$, this maximum is finite. Its definition includes equality at the clipping boundary.

Theorem 1 (Sharp outlier and memory dependence). *Let $f(x) = \lambda x^2/2$, $c, \lambda > 0$, $a = \eta\lambda \in (0, 2)$, and $x_0 = e_0 = 0$. In (4), take $g_0 = A \ge c$ and $g_t = \lambda x_t$ for every $t \ge 1$. For $B \in [0, \infty]$, put $A_B = \min\{A, c + B\}$ and $k = k_a(A_B)$. Then*

$$\frac{\lambda}{2}(\eta c k)^2 \;\le\; \sup_{t\ge 0} f(x_t) \;\le\; \eta c(A_B - c) + \frac{\lambda\eta^2 c^2}{2}. \tag{8}$$

For fixed a, c, η, as $A_B/c \to \infty$,

$$\frac{\sup_{t\ge 0} f(x_t)}{\eta c A_B} \longrightarrow 1. \tag{9}$$

Consequently, uncapped EF has peak loss asymptotic to ηcA. For fixed finite B, the upper bound is independent of A once $A \ge c + B$. For $B = 0$, the peak is exactly $\lambda\eta^2 c^2/2$.

The lower bound is attained as a value along the first sequence of saturated updates, not by an adversary who keeps injecting errors. After the initial intervention, the oracle is exact. To see the mechanism, note that after the first step the state is identical to the state produced by an outlier of size A_B without a memory cap. During the next positive saturated updates the residual decreases, so the cap is inactive. Direct induction gives

$$x_j = -\eta c j, \qquad e_j = A_B - cj - \frac{ac}{2}j(j-1), \qquad 1 \le j \le k. \tag{10}$$

The current true gradient points toward zero, but the stored residual continues to push away from zero. Since $H_a(k) \le A_B < H_a(k+1)$, we have $k \sim \sqrt{2A_B/(ac)}$, which proves the asymptotic lower bound. The upper bound will follow from the next section and controls the entire subsequent trajectory, including any oscillations. Appendix A gives the complete proof.

A stochastic interpretation with precise scope. The construction also applies to an unbiased oracle with one random shock. For any $p \in (1, 2]$, $\sigma > 0$, and $\delta \in (0, 1)$, let ξ_0 equal M and $-M$ with probability $\delta/2$ each, and zero otherwise, where $M = \sigma\delta^{-1/p}$. Let $\xi_t = 0$ for $t \ge 1$. Then $\mathbb{E}\xi_0 = 0$ and $\mathbb{E}|\xi_0|^p = \sigma^p$. By symmetry and Theorem 1, for uncapped EF,

$$\Pr\left\{\max_{0\le t\le k_a(M)} f(x_t) \ge \frac{\lambda}{2}\big(\eta c k_a(M)\big)^2\right\} \ge \delta, \tag{11}$$

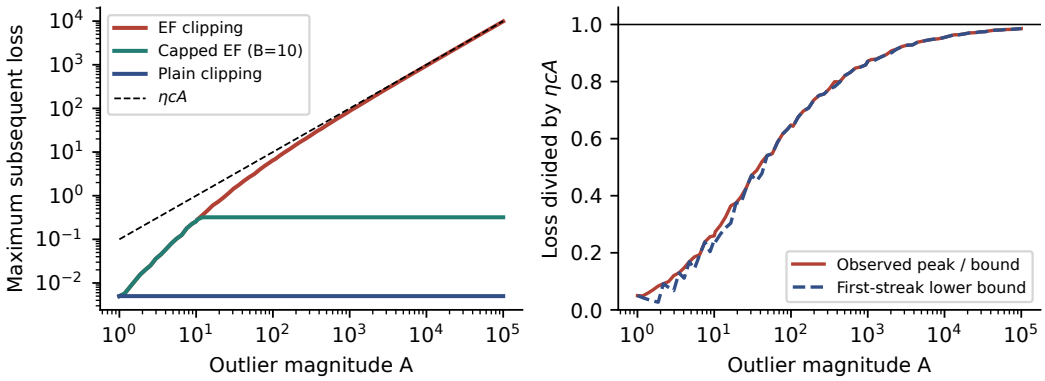

Figure 1: A single outlier on $f(x) = x^2/2$, with $\eta = 0.1$ and $c = 1$. Left: peak loss for identical update thresholds and three memory budgets. Right: the uncapped peak and the explicit lower bound approach the predicted leading term $\eta c A$. Each numerical peak is certified against later iterates using the decreasing energy and the stopping tolerance in Appendix E.

whenever $M \geq c$. The threshold on the right is asymptotic to $\eta c \sigma \delta^{-1/p}$. This is a statement about transient maximum loss, at a horizon growing with M. It is not a lower bound for final-iterate or averaged-iterate optimization rates under persistent i.i.d. noise.

5 A SUPPORT-FUNCTION ENERGY INEQUALITY

The proof uses a geometric identity specific to projection. If $u = \mathrm{P}_K(v)$ and $r = v - u$, the projection optimality condition gives

$$\langle r, w - u \rangle \leq 0 \quad (w \in K), \qquad \langle u, r \rangle = h_K(r). \tag{12}$$

Thus u maximizes the support function in the direction of r. Scaling r by a nonnegative scalar preserves this property. This remains true at $r = 0$, since every $u \in K$ maximizes $\langle u, 0 \rangle$.

Theorem 2 (Energy inequality)**.** *Under* (3)–(5)*, define*

$$\mathcal{V}_t = f(x_t) - f_\star + \eta h_K(e_t). \tag{13}$$

The iterates of (6) *satisfy, pathwise,*

$$\mathcal{V}_{t+1} \leq \mathcal{V}_t - \eta\alpha \|u_t\|^2 + \eta \langle \xi_t, u_t \rangle - \eta h_K(d_t). \tag{14}$$

The same inequality holds for any choice $\beta_t \in [0, 1]$, including choices based on the current input and past history.

Proof. By (12) and the nonnegative scaling of r_t, we have $\langle u_t, e_{t+1} \rangle = h_K(e_{t+1})$ and $\langle u_t, d_t \rangle = h_K(d_t)$. Also $\langle u_t, e_t \rangle \leq h_K(e_t)$ because $u_t \in K$. Therefore

$$\begin{aligned} h_K(e_{t+1}) - h_K(e_t) &\leq \langle u_t, e_{t+1} - e_t \rangle \\ &= \langle u_t, \nabla f(x_t) + \xi_t - u_t - d_t \rangle . \end{aligned} \tag{15}$$

Smoothness gives $f(x_{t+1}) \leq f(x_t) - \eta \langle \nabla f(x_t), u_t \rangle + L\eta^2 \|u_t\|^2 / 2$. Adding η times (15) cancels the gradient inner product and proves (14). □

For a Euclidean ball the energy is simply $f(x_t) - f_\star + \eta c \|e_t\|$. Its dependence on the residual is linear. A quadratic residual bound would obscure the linear dependence on outlier amplitude in Theorem 1. The projection optimality condition is standard; the role of the support function here is to make the gradient term cancel and to measure exactly the energy removed by discarding residuals.

Corollary 3 (Exact gradients and summable errors). *If $\xi_t = 0$ for all t, then*

$$\sum_{t=0}^{\infty} \|u_t\|^2 \le \frac{\mathcal{V}_0}{\eta\alpha}, \qquad \min_{0\le t<T} \min\{\|\nabla f(x_t)\|^2, \rho^2\} \le \frac{2\mathcal{V}_0}{\eta\alpha T} + \frac{\rho^2 \mathbf{1}\{e_0 \neq 0\}}{T}. \tag{16}$$

Moreover, there is a finite index after which $e_t = 0$ and the method follows ordinary gradient descent; $\|\nabla f(x_t)\| \to 0$. The same eventual disappearance of the residual and convergence of the gradient hold if $\sum_t \|\xi_t\| < \infty$, with ordinary gradient descent replaced by gradient descent using $\nabla f(x_t) + \xi_t$.

The bound in (16) concerns a truncated gradient measure. For a target $\epsilon < \rho$, choosing T so that the right side is less than ϵ^2 guarantees an iterate with $\|\nabla f(x_t)\| < \epsilon$. The proof does not assume that the iterates stay in a compact set. A finite number of saturated updates also does not, by itself, give a bound on the index of the last such update. We claim finite eventual disappearance, not a quantitative last-clipping time.

Proof mechanism. Summing (14) gives the first bound. If $e_t \neq 0$ for $t \ge 1$, the preceding projected update lies on the boundary of K and has norm at least ρ. If $e_t = 0$, either the current gradient is in K and equals u_t, or u_t is on that boundary. Charging the truncated gradient to $\|u_{t-1}\|^2 + \|u_t\|^2$ gives (16). Finally $u_t \to 0$ implies that all sufficiently late projections are interior, so their residuals are zero. Appendix B supplies the details, including summable errors.

6 Stability with capped memory

The energy bound separates immediate movement from retained memory. Suppose that $x_0 = x_\star$ is a global minimizer, $e_0 = 0$, and only g_0 is corrupted. Since $\|u_0\| \le R$ and $\nabla f(x_\star) = 0$, smoothness yields

$$f(x_1) - f(x_\star) \le \frac{L\eta^2 R^2}{2}, \qquad \mathcal{V}_1 \le \frac{L\eta^2 R^2}{2} + \eta b. \tag{17}$$

Subsequent exact-gradient iterations cannot increase $\mathcal{V}_t$. For the ball, this is the amplitude-independent bound $L\eta^2 c^2/2 + \eta cB$. It holds in any dimension and without convexity. Theorem 1 shows that the coefficient of ηcB cannot be improved asymptotically as $B/c \to \infty$ within this family.

The same argument extends to multiple corruptions, including corruptions chosen after observing the current state.

Theorem 4 (A bound in terms of the number of corruptions). *Assume (3)–(5), use a finite support budget b, and let $f_\star = \inf_x f(x) > -\infty$. Let m_t be the number of indices $s < t$ for which $\xi_s \neq 0$. Then*

$$\sqrt{\mathcal{V}_t} \le \sqrt{\mathcal{V}_0} + m_t\kappa, \qquad \kappa = \sqrt{\frac{L\eta^2 R^2}{2} + \eta b}. \tag{18}$$

In particular, $f(x_t) - f_\star \le (\sqrt{\mathcal{V}_0} + m_t\kappa)^2$, independently of the magnitudes of the corrupted gradients. If only finitely many gradients are corrupted, then $\|\nabla f(x_t)\| \to 0$.

Proof. At an uncorrupted iteration, Theorem 2 gives $\mathcal{V}_{t+1} \le \mathcal{V}_t$. At a corrupted iteration use the standard smoothness consequence $\|\nabla f(x_t)\|^2 \le 2L(f(x_t) - f_\star)$. Since the newly stored residual has support at most b,

$$\begin{aligned} \mathcal{V}_{t+1} &\le f(x_t) - f_\star + \eta R \|\nabla f(x_t)\| + L\eta^2 R^2/2 + \eta b \\ &\le \mathcal{V}_t + \eta R\sqrt{2L\mathcal{V}_t} + \kappa^2 \le (\sqrt{\mathcal{V}_t} + \kappa)^2. \end{aligned}$$

Taking square roots and inducting proves the bound. Following the last corruption, apply Corollary 3 to the remaining exact-gradient trajectory. □

This result concerns finite corruptions, not persistent stochastic noise. It also does not assert that retaining memory is preferable to ordinary clipping. Setting $b = 0$ gives the smallest bound in (18). A positive budget is justified only when preserving part of the discarded gradient is useful for a separate objective. For a desired single-outlier loss budget $\Delta \ge L\eta^2 R^2/2$, choosing $b \le (\Delta - L\eta^2 R^2/2)/\eta$ certifies (17); this prescription requires an appropriate smoothness bound.

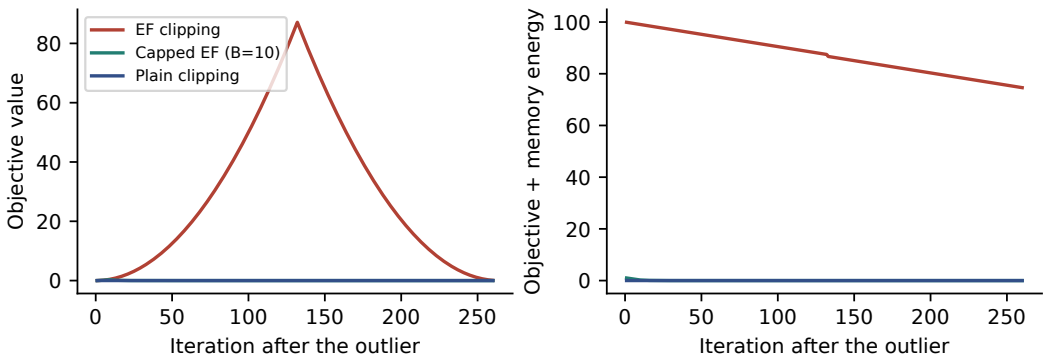

Figure 2: The single-outlier trajectory for $A = 1000$. The loss can increase while the objective-plus-memory energy decreases. The two curves diagnose the difference between a bounded update and the future effect of retained residuals.

6.1 What the memory cap discards

The exact identity

$$\sum_{t=0}^{T-1} u_t = \sum_{t=0}^{T-1} g_t + e_0 - e_T - \sum_{t=0}^{T-1} d_t \tag{19}$$

follows by telescoping $e_{t+1} = e_t + g_t - u_t - d_t$. Without a cap, $d_t = 0$. With a finite budget and $h_K(e_0) \le b$, the discrepancy from the input average is bounded by

$$\left\| \frac{1}{T} \sum_{t<T} (u_t - g_t) \right\| \le \frac{2b}{\rho T} + \frac{1}{T} \sum_{t<T} \|d_t\| . \tag{20}$$

Thus a small final residual is not evidence of small bias unless the discarded mass is also controlled.

For a concrete example, take scalar inputs $A, -s, \dots, -s$, with m copies of $-s$, $A = ms > c + B$, and $0 < s \le c$. Start with $e_0 = 0$. The inputs sum to zero. The initial discarded mass is $A - c - B$; the subsequent negative inputs drain the nonnegative residual without further overflow. The final residual is zero, and (19) gives

$$\sum_{t=0}^{m} u_t = c + B - A < 0. \tag{21}$$

Repeating this input cycle produces a persistent discrepancy. This is a statement about the input-output filter, not a claim about an i.i.d. gradient oracle. It makes explicit why the stability result cannot be advertised as universal stochastic debiasing.

7 Numerical illustrations and checks

The supplied script implements the recursions directly in double precision. It generates Figures 1–2 and saves all measurements. No training data or GPU is used. The purpose is to test the formulas, clipping boundaries, and geometric extension; the proofs provide the guarantees.

Outlier magnitude and memory budget. We use $f(x) = x^2/2$, $c = 1$, $\eta = 0.1$, and $x_0 = e_0 = 0$, with a single positive outlier. We compare $B = 0$, $B = 10$, and $B = \infty$ over amplitudes from 1 to 10^5. The algorithms share the step size and update threshold; this is a controlled intervention rather than a hyperparameter-tuned performance comparison. For $A = 10^2, 10^3, 10^4, 10^5$, uncapped peak losses are 6.48, 87.12, 954.845, and 9856.08, respectively. At these amplitudes, capped EF has peak 0.32, while ordinary clipping has peak 0.005. The uncapped peak at $A = 10^5$ is 0.985608 times $\eta c A$. These observations agree with the finite bounds and the asymptotic prediction, without suggesting that the asymptotic bound is tight at small A.

Geometric and endpoint checks. With seed 20260918, we checked (14) on 6000 random six-dimensional quadratic states, split equally among balls, unequal coordinate boxes, and products of two balls. These checks use nonzero initial residuals, injected errors, and step-size factors $L\eta$ sampled in $(0.01, 1.99)$. No violation above the numerical tolerance was observed. Fifty additional cases test equality at $A = H_a(k)$, including $\eta\lambda > 1$. A smooth nonconvex separable objective checks the energy inequality and the truncated-gradient summation bound along complete trajectories. Appendix E describes all parameters. These finite checks do not substitute for the analytical proof or establish behavior on neural networks.

8 Discussion and Limitations

The sharp quadratic law identifies an optimizer-state mechanism: residual feedback can preserve precisely the magnitude that clipping was intended to suppress. The support-function energy quantifies that mechanism and yields a cap with deterministic convergence and finite-corruption stability guarantees. The discarded-mass identity describes the price of the cap.

Several boundaries matter. The analysis uses a fixed step size, global smoothness, and Euclidean projection onto a fixed convex update set. Persistent stochastic gradients are covered only by the pathwise inequality, not by a statistical convergence theorem. Separate per-example clipping and heterogeneous distributed tracking need different arguments. The finite-corruption bound is not claimed to be tight in the number of corruptions. Finally, the experiments verify controlled mechanisms rather than practical accuracy or training speed. Extending the analysis to persistent noise and measuring the tradeoff between residual retention and statistical error remain open directions.

Reproducibility Statement

The update equations and assumptions are stated in Section 3. Complete proofs and implementation details appear in the appendix. The accompanying source package contains `code/verify.py`, the numerical results, and the generated vector figures. Running the script regenerates all reported computations from the disclosed seed. No external data are required.

AI Use Statement

AI was used for literature discovery, formulation and analysis of mathematical claims, computational verification, and manuscript preparation.

References

Moneeb Abbas, Wen Chung Kuo, Khalid Mahmood, Muhammad Asad Saleem, Yonis Gulzar, and Jenhui Chen. CIWGS-FedMed: Channel-importance weighted gradient sparsification for communication efficient federated medical image segmentation. *IEEE Transactions on Consumer Electronics*, 2026. doi: 10.1109/TCE.2026.3706981. URL https://doi.org/10.1109/TCE.2026.3706981. Early access.

Yong-Yan Cao, Zongli Lin, and David G. Ward. Anti-windup design of output tracking systems subject to actuator saturation and constant disturbances. *Automatica*, 40(7):1221–1228, 2004. doi: 10.1016/j.automatica.2004.02.012. URL https://doi.org/10.1016/j.automatica.2004.02.012.

Rustem Islamov, Samuel Horváth, Aurélien Lucchi, Peter Richtárik, and Eduard Gorbunov. Double momentum and error feedback for clipping with fast rates and differential privacy, 2025. URL https://arxiv.org/abs/2502.11682. arXiv:2502.11682.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, Rustem Islamov, Fakhri Karray, and Peter Richtárik. Clip21: Error feedback for gradient clipping, 2023. URL https://arxiv.org/abs/2305.18929. arXiv:2305.18929.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363, 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Shuhua Yu, Dušan Jakovetić, and Soummya Kar. Smoothed gradient clipping and error feedback for decentralized optimization under symmetric heavy-tailed noise. *SIAM Journal on Optimization*, 36(2):703–728, 2026. doi: 10.1137/24M170747X. URL https://doi.org/10.1137/24M170747X.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.11881.

Xinwei Zhang, Zhiqi Bu, Zhiwei Steven Wu, and Mingyi Hong. Differentially private SGD without clipping bias: An error-feedback approach. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uFbWHyTlPn.

A Proof of the sharp quadratic bounds

For the first update, $u_0 = c$, $x_1 = -\eta c$, and $e_1 = \min\{B, A - c\} = A_B - c$. Thus the state at time one is the same as the state of uncapped EF initialized at zero with outlier A_B.

For the latter recursion, suppose j positive saturated updates have occurred. Summing the exact gradients $\lambda x_i = -aci$ for $1 \le i < j$ gives

$$x_j = -\eta cj, \qquad e_j = A_B - cj - ac\sum_{i=1}^{j-1} i = A_B - H_a(j).$$

At time j, a further positive saturated update occurs if and only if

$$e_j + \lambda x_j = A_B - H_a(j) - acj \ge c,$$

or equivalently $A_B \ge H_a(j+1)$. Starting with the first update, this induction is valid through $j = k_a(A_B)$. During this sequence $e_j \ge 0$ and $e_{j+1} = e_j - c - acj \le e_j$. For finite B, all these residuals remain below $e_1 \le B$, so the additional cap does not change the induction. Evaluating f at x_k proves the lower bound. If equality holds at $A_B = H_a(k)$, then $e_k = 0$; all k saturated updates still occurred. No strict inequality is needed at that boundary.

For the upper bound, Theorem 2 applies to every iteration after time zero, with $K = [-c, c]$ and $\xi_t = 0$. Hence

$$f(x_t) \le \mathcal{V}_t \le \mathcal{V}_1 = \frac{\lambda\eta^2 c^2}{2} + \eta c(A_B - c), \qquad t \ge 1.$$

The value at $t = 0$ is zero. This upper bound therefore covers the infinite horizon, independently of subsequent sign changes or oscillations.

Finally, write $z = A_B/c$. The inequalities defining k are

$$\frac{a}{2}k^2 + \left(1 - \frac{a}{2}\right)k \le z < \frac{a}{2}(k+1)^2 + \left(1 - \frac{a}{2}\right)(k+1).$$

For fixed $a \in (0, 2)$ and $z \to \infty$, they imply $k \to \infty$ and $ak^2/(2z) \to 1$. Divide the bounds in (8) by $\eta c A_B$: their lower and upper sides are respectively

$$\frac{ack^2}{2A_B} \longrightarrow 1, \qquad 1 - \frac{c(1 - a/2)}{A_B} \longrightarrow 1.$$

The squeeze theorem proves (9). If $B = 0$, then $A_B = c$ and $k = 1$, so the lower and upper bounds agree exactly. A negative outlier gives the sign-reversed state trajectory and the same loss.

B Convergence proof

We first record the geometric facts needed to translate a bound on updates into a bound on gradients. Projection onto a nonempty closed convex set is uniquely defined. Its variational inequality proves (12). If $r = v - \mathrm{P}_K(v) \ne 0$, then $\mathrm{P}_K(v)$ is on the boundary of K: otherwise one could move a small distance from it in the direction of r while remaining in K, contradicting (12). By (5), every boundary point has norm at least ρ. These arguments hold without central symmetry of K.

For exact gradients, sum (14) and use $h_K(d_t) \ge 0$ and $\mathcal{V}_t \ge 0$ to obtain

$$\eta\alpha \sum_{t=0}^{T-1} \|u_t\|^2 \le \mathcal{V}_0.$$

Taking $T \to \infty$ proves the first part of (16). Let $q_t = \min\{\|\nabla f(x_t)\|^2, \rho^2\}$. If $e_t = 0$, then $u_t = \mathrm{P}_K(\nabla f(x_t))$. If the input is in K, u_t equals that input and $q_t \le \|u_t\|^2$. If it is outside K, the projection is on the boundary and $q_t \le \rho^2 \le \|u_t\|^2$.

If $e_t \ne 0$ and $t \ge 1$, then $r_{t-1} \ne 0$ because $e_t = \beta_{t-1} r_{t-1}$. The preceding update is on the boundary, and $q_t \le \rho^2 \le \|u_{t-1}\|^2$. Thus, for $t \ge 1$,

$$q_t \le \|u_{t-1}\|^2 + \|u_t\|^2.$$

For $t = 0$, $q_0 \le \|u_0\|^2 + \rho^2 \mathbf{1}\{e_0 \neq 0\}$. Summing proves the stated bound on the average of q_t, and hence on its minimum.

Since $\sum_t \|u_t\|^2 < \infty$, $\|u_t\| \to 0$. There is therefore a finite N such that $\|u_t\| < \rho$ for all $t \ge N$. Such a point lies in the interior of K, so $r_t = 0$ and $e_{t+1} = 0$ for all $t \ge N$. For $t \ge N + 1$, the method has $e_t = 0$, $u_t = \nabla f(x_t)$, and the ordinary gradient descent recursion follows. The gradient tends to zero. This conclusion is first-order stationarity; it does not assert convergence of x_t to a unique point or to a global minimum.

For summable errors, use $\langle \xi_t, u_t \rangle \le h_K(\xi_t) \le R \|\xi_t\|$ in (14). For every T,

$$\mathcal{V}_T + \eta\alpha \sum_{t<T} \|u_t\|^2 + \eta \sum_{t<T} h_K(d_t) \le \mathcal{V}_0 + \eta R \sum_{t<T} \|\xi_t\| . \tag{22}$$

The right side is bounded uniformly in T, giving $u_t \to 0$ and the same eventual disappearance of the residual. At sufficiently late times, $u_t = \nabla f(x_t) + \xi_t$. Since summability implies $\xi_t \to 0$, the gradient also tends to zero. This does not provide a convergence rate for nonsummable stochastic noise.

C Additional details for finite-corruption stability

The gradient bound used in Theorem 4 follows directly by evaluating the smoothness inequality at $y = x - \nabla f(x)/L$:

$$f_\star \le f(y) \le f(x) - \frac{\|\nabla f(x)\|^2}{2L}.$$

It requires only global smoothness and a finite lower bound; the infimum need not be attained. The inequality $2\kappa \ge \eta R\sqrt{2L}$ follows from the definition of κ. Every use of a square root in (18) is therefore justified by nonnegativity of $f(x_t) - f_\star$ and $h_K(e_t)$.

For the isolated-outlier statement (17), attainment is assumed explicitly: $x_\star$ is a global minimizer and therefore has zero gradient. The cap gives $h_K(e_1) \le b$ regardless of outlier direction. The bound also applies if the corruption is smaller than the clipping threshold. In the Euclidean quadratic example with $A \ge c$, the stronger amplitude-dependent formula in Theorem 1 follows from the exact value of e_1.

If the last corrupted index is s, then $\mathcal{V}_{s+1}$ is finite by (18). Restarting the proof of Corollary 3 at time $s + 1$ is valid with its potentially nonzero residual. In particular, for $T \ge 1$,

$$\min_{s+1 \le t < s+1+T} \min\{\|\nabla f(x_t)\|^2, \rho^2\} \le \frac{2\mathcal{V}_{s+1}}{\eta\alpha T} + \frac{\rho^2}{T}.$$

This is a recovery guarantee after finitely many errors, not a claim of immunity to infinitely many corruptions.

D Conservation and the zero-sum cycle

Equation (19) is obtained by summing the residual recursion. Since $\rho\mathbb{B}_2 \subseteq K$, maximizing over that ball gives $h_K(z) \ge \rho \|z\|$. Under a finite support budget, $\|e_T\| \le b/\rho$ for $T \ge 1$. If the same condition holds initially, the triangle inequality proves (20).

For (21), the first input has output c, residual B, and discard $A - c - B$. For any subsequent input $-s$, a nonnegative incoming residual e gives

$$e_{\text{new}} = \max\{e - s - c, 0\}.$$

Indeed $e - s \ge -c$ because $s \le c$. If $e - s \le c$, projection returns $e - s$ and the new residual is zero. If $e - s > c$, the output is c and the new residual is $e - s - c \le e \le B$. No further residual is discarded. After m such inputs the residual is zero because $m(c + s) > ms = A > B$. Substitution into (19) yields (21). Each repetition starts at zero residual, so the discrepancy persists under repeated cycles.

E NUMERICAL PROTOCOL

Reproduction. From the extracted source directory run `python code/verify.py`. The run used Python 3.13.5, NumPy 2.2.1, and Matplotlib 3.10.0. The measured numerical run took approximately 5.2 seconds, excluding interpreter startup and document compilation, on the preparation machine. This is a reproducibility observation, not a hardware-normalized benchmark. The script writes two PDF figures, `results/frontier.csv`, and `results/verification.json`.

Peak calculation. For each amplitude, maintain the largest loss observed so far, P_t. Once all future gradients are exact, $f(x_s) \le \mathcal{V}_t$ for every $s \ge t$. The scalar simulation stops when $\mathcal{V}_t \le P_t + 10^{-10}\max\{1, P_t\}$. The observed peak then differs from the infinite-horizon peak by at most this tolerance in exact arithmetic; double-precision roundoff is additional. Every peak is also checked against both analytical bounds with a numerical tolerance. The amplitude grid consists of 75 logarithmically spaced values between 1 and 10^5, with the reported powers of ten added explicitly.

General-set checks. For each of three shapes we generate 2000 six-dimensional cases with $H = M^\top M + 0.1I$ and M standard Gaussian. Set $L = \lambda_{\max}(H)$ and choose ηL uniformly in $(0.01, 1.99)$. Initial points and residuals are Gaussian vectors multiplied by independent log-uniform scales between 10^{-2} and 10^2. Errors have analogous scales between 10^{-3} and 10^3. Support budgets are chosen from $\{0, 0.1, 1, 100, \infty\}$. The sets are the unit ball, the box with coordinate half-widths $(0.2, 0.4, 0.6, 0.8, 1, 1.2)$, and a product of two three-dimensional balls of radii 0.7 and 1.3. The check computes both sides of (14), including its error and discard terms. It also verifies $\langle u_t, d_t\rangle = h_K(d_t)$. The largest normalized energy-inequality residual was approximately -1.46×10^{-9}, below zero.

Boundary cases. The grid $a \in \{0.01, 0.1, 0.9, 1, 1.9\}$ and $k \in \{1, 2, 5, 20, 100\}$ sets $A = H_a(k)$ with $c = \lambda = 1$. Each case is checked with $B = \infty$ and $B = A - 1$, for 50 cases. The induction predicts exactly k initial positive saturated updates and zero residual immediately afterward. A separate zero-sum input cycle uses $(c, B, s, m) = (1, 2, 0.5, 20)$; its measured output sum is -7 and its discarded mass is 7.

Nonconvex trajectories. We use $f(x) = \sum_{i=1}^{3}(0.05x_i^2 + 1 - \cos x_i)$, whose Hessian eigenvalues lie in $[-0.9, 1.1]$. Set $L = 1.1$, $\eta = 0.5$, $c = 1$, $x_0 = (5, -4, 3)$, $e_0 = 0$, and $B \in \{0, 2, \infty\}$. Across 1200 exact-gradient iterations, the script checks the energy inequality and the cumulative charging inequality underlying (16). The final gradient norms were below 2×10^{-15}. This verifies a finite collection of trajectories, not a general empirical rate.

F ASSUMPTION BOUNDARIES

- **Smoothness and a constant step size.** Smoothness controls objective changes. Constant η permits cancellation with the memory term; varying it introduces additional terms.
- **Projection and radial scaling.** These preserve the support maximizer in (12). Arbitrary compressors or coordinatewise residual truncations need not do so.
- **Positive inner radius.** Nonzero residuals then require nonvanishing updates. Without it, stationarity can fail: $K = \{0\}$ and $f(x) = (x-1)^2/2$ leave initialization zero unchanged.
- **Strict step-size bound.** At $\eta = 2/L$, ordinary gradient descent on $f(x) = Lx^2/2$ can oscillate even with clipping inactive. The summability argument requires $\alpha > 0$.
- **One aggregate gradient.** With heterogeneous local objectives, local memory terms need not cancel the global gradient inner product. The distributed case requires a separate proof.
- **Finite-corruption scope.** The bounds imply neither persistent-noise debiasing nor privacy. The zero-sum cycle shows why bounded memory alone cannot guarantee conservation.