CLIPPED OR UNCLIPPED? FINITE-SAMPLE TRADE-OFFS FOR AVERAGED SGD UNDER HEAVY-TAILED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient clipping is widely used to stabilize training, but it need not improve the statistical accuracy of averaged SGD, even under heavy-tailed noise. We derive a finite-sample comparison of clipped and unclipped Polyak-Ruppert averaged SGD under finite conditional p-th moments, $p \geq 2$. Our main result gives explicit accuracy and confidence conditions under which, for $p > 2$, the Gaussian term dominates the unclipped deviation bound, so clipping need not improve its leading order. By balancing clipping bias and concentration, we obtain a bound in which the heavy-tail correction depends logarithmically rather than polynomially on the inverse failure probability. At $p = 2$, this improves the confidence dependence of the leading bound. We establish sharpness of the unclipped heavy-tail term through an exact one-dimensional quadratic recursion and extend the comparison to projected convex SGD. We also prove concrete costs of clipping: every fixed finite threshold increases asymptotic variance on a scalar Gaussian quadratic, while whole-gradient clipping can shift the limiting point under asymmetric noise.

1 INTRODUCTION

Gradient clipping is widely used to control large stochastic updates and stabilize training (Pascanu et al., 2013; Brown et al., 2020; Touvron et al., 2023). Its appeal is particularly clear under heavy-tailed noise, where rare gradients can be much larger than typical ones. For averaged stochastic gradient descent (SGD), evaluating this protection requires understanding its effect on the final estimator. Clipping suppresses extreme updates, but also changes the mean update and can attenuate the optimization signal. Its benefit therefore depends on the accuracy and confidence required from the averaged result.

We compare clipped and unclipped Polyak–Ruppert averaged SGD under finite conditional noise moments of order $p \geq 2$. The comparison turns on a distinction between two contributions to the unclipped deviation bound: a Gaussian term, governed by the variance, and a heavy-tail correction, accounting for rare large increments. For $p > 2$, the correction decays faster with the sample budget than the Gaussian term at fixed confidence. A threshold chosen to balance clipping bias and concentration improves the confidence dependence of the correction from polynomial to logarithmic. The consequence depends on which term controls the guarantee: once the Gaussian term dominates, clipping need not improve its leading order. At $p = 2$, the confidence improvement affects the leading statistical bound itself.

Our finite-sample results make this distinction quantitative through explicit conditions on the requested accuracy and confidence. They also retain the different moment scales involved in the comparison: whole-gradient clipping acts on both signal and noise, so its bias bound depends on moments of the full stochastic gradient. Clipping only the centered noise provides an oracle benchmark for separating these effects. For projected convex optimization, the complete bounds additionally account for initialization and squared-gradient terms.

A complementary question concerns thresholds held fixed as the budget grows. Such a threshold can help at a finite horizon while changing the limiting estimator. Even when the target is preserved, a reduction in the variance of individual updates need not reduce the asymptotic variance of their

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.

averaged trajectory, because clipping also changes the local mean dynamics. We study these effects alongside the finite-sample comparison.

Contributions: our contributions are following

- **Finite-sample conditions for Gaussian dominance.** We derive high-probability bounds for projected convex SGD and an epoch-averaged extension for strongly convex quadratics. An accuracy-normalized boundary identifies when the unclipped heavy-tail correction is at most of Gaussian order and clarifies when clipping's confidence improvement changes the leading stochastic guarantee.
- **Sharpness for the untrimmed averaging mechanism.** We embed a rare-outlier construction into a scalar quadratic SGD recursion, attaining the unclipped heavy-tail dependence. At $p = 2$, this gives the polynomial confidence obstruction for the specified single-run averaging mechanism. Robust final aggregation and confidence amplification remain outside this lower-bound claim.
- **Fixed-threshold effects on the target and variance.** Under the stated Polyak–Ruppert regularity conditions, every finite fixed threshold strictly increases asymptotic variance on a scalar Gaussian quadratic, while suitable thresholds can reduce it for some symmetric heavy-tailed laws. Under asymmetric noise, fixed whole-gradient clipping can shift the limiting target. For $p > 2$ in the linear noise-centered setting, we give a sufficient window for thresholds growing with the run horizon to preserve the unclipped first-order limit.

An off-policy click-prediction experiment illustrates how the effect of fixed clipping can change with the budget. We leave importance weights intact, select a gradient-clipping threshold on separate source-validation runs, and hold it fixed during evaluation. The selected method improves the upper-tail target error at an intermediate budget and loses to unclipped SGD at a larger one. This provides a concrete example of a finite-horizon benefit that does not persist as the run grows.

Related work. Clipped stochastic optimization achieves logarithmic confidence dependence under weak moment assumptions (Gorbunov et al., 2020; Sadiev et al., 2023; Nguyen et al., 2023). Robust mean estimation and confidence amplification provide related ways to control rare deviations (Catoni, 2012; Lugosi & Mendelson, 2019; Davis et al., 2021). Unmodified stochastic-gradient and mirror-descent methods also admit guarantees under weak moments (Eldowa & Paudice, 2024; Fatkhullin et al., 2025). Our comparison identifies when the confidence improvement due to clipping concerns the leading stochastic term and when it concerns a lower-order correction, while keeping track of the moment scale and bias introduced by the clipping rule.

The asymptotic efficiency and finite-time concentration of averaged stochastic approximation are well studied (Polyak & Juditsky, 1992; Mou et al., 2020; Durmus et al., 2025). Koloskova et al. (2023) quantify stochastic clipping bias, and Marshall et al. (2025) analyze distribution-dependent clipping effects in high-dimensional least-squares dynamics. We connect the finite-sample accuracy–confidence comparison to the limiting target and covariance of the averaged estimator through explicit models. Together, these results explain how the budget, confidence requirement, and threshold determine the statistical trade-offs of clipping.

Organization. Section 2 introduces the oracle and concentration tools. Section 3 gives the finite-sample optimization bounds; Sections 4 and 5 describe the asymptotic effects and the accuracy-normalized comparison. Section 6 presents the off-policy experiment. The appendices provide complete proofs and further supporting results.

2 Problem Setting and Concentration Tools

We consider the stochastic convex optimization problem

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E} f(x, \xi),$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed, and convex. Assume that the solution set $\mathcal{X}^\star$ is nonempty, and fix $x^\star \in \mathcal{X}^\star$. Unless otherwise specified, we assume F is convex, and all iterates lie in a set of diameter $D > 0$,

$$\|x - x^\star\| \le D, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm; assumption (1) is typical for bounded feasible sets, projections onto a known ball, and stopping-time arguments (see (Oymak & Soltanolkotabi, 2018, Sec. 9.4.5)).

Let $\mathcal{F}_{k-1}$ contain the information available before iteration k, for $k \geq 1$. The oracle returns

$$g_k := g(x_k, \xi_k) = s_k + \zeta_k, \quad s_k \in \partial F(x_k),$$

where x_k is $\mathcal{F}_{k-1}$-measurable and ξ_k is fresh conditionally on $\mathcal{F}_{k-1}$.

Assumption 2.1 (conditional p-th moment of the noise ζ_k)**.** For some $p \geq 2$ and deterministic constants $\sigma_2, \sigma_p < \infty$,

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_{k-1}] \leq \sigma_2^2, \quad \mathbb{E}[\|\zeta_k\|^p \mid \mathcal{F}_{k-1}] \leq \sigma_p^p$$

for every k. When $p = 2$, set $\sigma_p = \sigma_2$.

Unless a different output is specified, the averaged iterate is $\bar{x}_N := N^{-1} \sum_{k=1}^{N} x_k$.

2.1 Clipping and the relevant moment scales

For a threshold $\lambda > 0$, define the clipping map

$$\mathrm{clip}_\lambda(y) = y \min\{1, \lambda/\|y\|\} \quad (y \neq 0), \quad \mathrm{clip}_\lambda(0) = 0.$$

Whole-gradient clipping also requires control of the full stochastic gradient.

Assumption 2.2 (moment of the full stochastic gradient)**.** For constants $G_2, G_p < \infty$,

$$\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_{k-1}] \leq G_2^2, \quad \mathbb{E}[\|g_k\|^p \mid \mathcal{F}_{k-1}] \leq G_p^p.$$

Centered and full-gradient moments. The scale σ_p controls centered noise, whereas G_p includes the predictable optimization signal. For example, if F is L-smooth and (1) holds, then for x_k satisfying $\|x_k - x^\star\| \leq r$,

$$G_p(r) \leq C_p\big(\sigma_p + \|\nabla F(x^\star)\| + Lr\big).$$

For an unconstrained problem, or an interior optimizer, $\nabla F(x^\star) = 0$; it need not vanish at a boundary solution. Thus $G_p(r)$ may exceed σ_p if r is large enough.

Directions under comparison. Our main comparison is between unclipped SGD (U) and whole-gradient clipping (WG). Noise-centered clipping (NC) serves as an oracle benchmark:

$$h_k^{\mathrm{U}} = g_k, \qquad h_k^{\mathrm{WG}} = \mathrm{clip}_\lambda(g_k), \quad h_k^{\mathrm{NC}} = s_k + \mathrm{clip}_\lambda(\zeta_k) \tag{2}$$

The U and WG directions use only the observed stochastic gradient. NC requires additional information because s_k and ζ_k are not separately observable from g_k.

The smooth quadratic extension also considers difference clipping (DC). It uses an anchor $\widetilde{x}$ and an estimate $\widetilde{s} \approx \nabla F(\widetilde{x})$:

$$h_k^{\mathrm{DC}} := \widetilde{s} + \mathrm{clip}_{\lambda_k}\left(g(x_k, \xi_k) - g(\widetilde{x}, \xi_k)\right). \tag{3}$$

The anchor and its estimate are available before drawing ξ_k, and both gradients use the same sample. Constructions of this type appear in Gorbunov et al. (2024). Our guarantees for DC require the anchor and difference-moment conditions in Proposition C.12.

2.2 Fuk–Nagaev decomposition and clipping

The comparison separates variance-controlled fluctuations from rare large increments. For $\delta \in (0, 1)$, write $\ell_\delta := \log(4/\delta)$. Below, C denotes a numerical constant and C_p a constant depending only on p; their values may change between displays.

Unclipped averages. Under Assumption 2.1 with $p > 2$, the martingale Fuk–Nagaev inequality gives, with probability at least $1 - \delta$,

$$\left\| \frac{1}{N} \sum_{k=1}^{N} \zeta_k \right\| \le C\sigma_2 \sqrt{\frac{\ell_\delta}{N}} + C_p \sigma_p \delta^{-1/p} N^{-(1-1/p)}. \tag{4}$$

The general statement and its justification are given in Theorem A.1 in Appendix A.

The first term has a Gaussian dependence on the confidence level and describes the accumulation of moderate increments. The second captures the contribution of rare large increments under a finite p-th moment. For fixed δ and $p > 2$, the second term is lower order than $N^{-1/2}$. Proposition A.5 gives the boundary at which it becomes no larger than the Gaussian term.

Clipped averages. Now use a common deterministic threshold $\lambda_k = \lambda$. For either NC or WG, let

$$M_p := \begin{cases} \sigma_p, & \text{for NC,} \\ G_p, & \text{for WG.} \end{cases}$$

NC uses Assumption 2.1; WG additionally uses Assumption 2.2. For each of these choices, Proposition A.4 yields, with probability at least $1 - \delta$,

$$\left\| \frac{1}{N} \sum_{k=1}^{N} (h_k - s_k) \right\| \le C\sigma_2 \sqrt{\frac{\ell_\delta}{N}} + C\lambda \frac{\ell_\delta}{N} + \frac{M_p^p}{\lambda^{p-1}}.$$

The three terms correspond to variance, bounded increments, and clipping bias, respectively. Nonexpansiveness of clipping ensures that the conditional variance of the centered clipped direction is at most σ_2^2, including for WG. The bias bound, however, depends on the moments of the quantity being clipped; see Lemma A.3.

For $M_p > 0$, choosing $\lambda = M_p (N/\ell_\delta)^{1/p}$ balances the last two terms up to numerical constants; we call it the *balanced threshold*. It gives

$$\left\| \frac{1}{N} \sum_{k=1}^{N} (h_k - s_k) \right\| \le C\sigma_2 \sqrt{\frac{\ell_\delta}{N}} + C M_p \left(\frac{\ell_\delta}{N} \right)^{1-1/p} \tag{5}$$

with probability at least $1 - \delta$.

Equations (4) and (5) exhibit the same variance-controlled scale. Their corrections differ in their confidence dependence: the unclipped correction contains $\delta^{-1/p}$, whereas the tuned clipped correction depends on δ through ℓ_δ. For NC, both corrections involve the same noise scale σ_p. For WG, the clipped correction involves G_p, which also accounts for the signal. The optimization bounds use the same concentration arguments with predictable weights or fixed linear maps. Sections 3–5 track the resulting moment scales and confidence dependence.

The endpoint $p = 2$. Under a second-moment assumption alone, Chebyshev's inequality gives $\left\| \frac{1}{N} \sum_{k=1}^{N} \zeta_k \right\| \le \frac{\sigma_2}{\sqrt{N\delta}}$ with probability at least $1 - \delta$. In contrast, (5) for NC gives $C\sigma_2 \sqrt{\ell_\delta / N}$. The rare-outlier construction in Proposition A.6 shows why the $\delta^{-1/2}$ dependence cannot be uniformly removed from the ordinary empirical mean under a variance bound alone.

3 Finite-Sample Optimization Guarantees

We compare projected U-SGD and whole-gradient Clip-SGD on a convex, possibly nonsmooth objective. Fix an initial point $x_1 \in \mathcal{X}$. For a constant stepsize $\eta > 0$ (and a constant clipping threshold $\lambda > 0$ for Clip-SGD), both methods follow the recursion

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta h_k), \quad k = 1, \dots, N, \tag{6}$$

where $\Pi_{\mathcal{X}}$ is the Euclidean projection. We consider U-SGD with $h_k := h_k^{\mathrm{U}} = g_k$ and whole-gradient Clip-SGD with $h_k := h_k^{\mathrm{WG}} = \mathrm{clip}_\lambda(g_k)$; see (2). For the projected convex results below,

Assumptions 2.1 and 2.2 hold, F is convex, and (1) is satisfied. For U-SGD, only the second-moment part of Assumption 2.2 is used. Let $R_0 := \|x_1 - x^\star\|$.

Both methods satisfy a common deterministic inequality. Since $x^\star \in \mathcal{X}$ and $\Pi_{\mathcal{X}}$ is nonexpansive, convexity and Jensen's inequality give

$$F(\bar{x}_N) - F(x^\star) \le \frac{R_0^2}{2\eta N} + \frac{\eta}{2N}\sum_{k=1}^{N}\|h_k\|^2 + \frac{1}{N}\sum_{k=1}^{N}\langle h_k - s_k,\, x^\star - x_k\rangle. \tag{7}$$

The last term is a weighted analogue of the averages in Section 2.2: the weights $x^\star - x_k$ are $\mathcal{F}_{k-1}$-measurable and bounded by D. The martingale arguments underlying (4) and (5) therefore apply to the corresponding scalar increments, giving the same bounds with an additional factor D.

The squared-gradient term produces an additional remainder. For U-SGD, write $\|g_k\|^2 = \|s_k\|^2 + 2\langle s_k, \zeta_k\rangle + \|\zeta_k\|^2$, where $\|s_k\|^2 + \mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_{k-1}] \le G_2^2$. In particular, $\|s_k\| \le G_2$. The cross term is again a predictably weighted sum of ζ_k. For the stepsize below, its contribution is absorbed into the Gaussian and heavy-tail terms using $R_0 \le D$. The centered variable $\|\zeta_k\|^2 - \mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_{k1}]$ has conditional moment of order $p/2$ bounded by $C_p\sigma_p^p$. Its concentration gives the remainder

$$\mathcal{Q}_{p,N}(\delta) := \begin{cases} C_p\sigma_p^2\delta^{-2/p}N^{-(1-2/p)}, & 2 < p \le 4, \\ C\sigma_p^2\sqrt{\frac{\ell_\delta}{N}} + C_p\sigma_p^2\delta^{-2/p}N^{-(1-2/p)}, & p > 4. \end{cases}$$

The two cases use von Bahr–Esseen and Fuk–Nagaev concentration, respectively. The latter applies at moment order $p/2$, using $\mathbb{E}[\|\zeta_k\|^4 \mid \mathcal{F}_{k-1}] \le \sigma_p^4$. For WG, $\|h_k\| \le \lambda$, and Bernstein–Freedman concentration bounds the squared-gradient term.

The calibrated guarantees follow. Appendix B gives the arbitrary-parameter bounds and complete proofs.

Theorem 3.1 (U-SGD on a convex problem)**.** *Let $p > 2$ and consider U-SGD defined by* (6) *with $h_k = g_k$ and $\eta = R_0/(G_2\sqrt{N})$. Then, with probability at least $1-\delta$,*

$$F(\bar{x}_N) - F(x^\star) \le \frac{R_0 G_2}{\sqrt{N}} + CD\sigma_2\sqrt{\frac{\ell_\delta}{N}} + \frac{C_p D\sigma_p}{\delta^{1/p}N^{1-1/p}} + \frac{R_0}{2G_2\sqrt{N}}\mathcal{Q}_{p,N}(\delta). \tag{8}$$

For fixed δ, the final term in (8) scales as $N^{-(3/2-2/p)}$ for $2 < p \le 4$; for $p > 4$, its two components scale as N^{-1} and $N^{-(3/2-2/p)}$. Its polynomial component is lower order than the preceding heavy-tail correction for fixed δ, but not uniformly as δ decreases.

Theorem 3.2 (Whole-gradient Clip-SGD on a convex problem)**.** *Let $p > 2$ and consider whole-gradient Clip-SGD defined by* (6) *with $h_k = \mathrm{clip}_\lambda(g_k)$, $\eta = R_0/(G_2\sqrt{N})$, and $\lambda = c_p G_p (N/\ell_\delta)^{1/p}$. Then, with probability at least $1-\delta$,*

$$F(\bar{x}_N) - F(x^\star) \le \frac{R_0 G_2}{\sqrt{N}} + CD\sigma_2\sqrt{\frac{\ell_\delta}{N}} + C_p D G_p\left(\frac{\ell_\delta}{N}\right)^{1-1/p} + \frac{C_p R_0 G_p^2}{G_2\sqrt{N}}\left(\frac{\ell_\delta}{N}\right)^{1-2/p}. \tag{9}$$

In (9), the squared-gradient term proportional to $R_0 G_p \ell_\delta^{1/2-1/p} N^{-(1-1/p)}$ has been absorbed into the displayed DG_p correction using $R_0 \le D$ and $\ell_\delta \ge 1$; the remaining term is lower order in N for every $p > 2$ at fixed confidence.

For NC (2), the martingale and clipping-bias calculations replace G_p by σ_p and do not truncate the full predictable signal. DC (3) uses a certified difference scale when its anchor error and difference moments are controlled; the corresponding quadratic guarantees are described below. Difference-clipped constructions of this kind are developed in Gorbunov et al. (2024).

3.1 Second-moment calibrations

If $p = 2$, we use a confidence-dependent step for U-SGD:

$$\eta = \frac{R_0\sqrt{\delta}}{G_2\sqrt{N}}.$$

Markov's inequality for the squared-gradient average and Chebyshev's inequality for the martingale term give

$$\mathbb{P}\left(F(\bar{x}_N) - F(x^\star) > C(R_0 G_2 + D\sigma_2)\delta^{-1/2}N^{-1/2}\right) \le \delta. \tag{10}$$

See Appendix B.4 for the proof.

For whole-gradient clipping, the same argument with $p = 2$ and $\eta = R_0/(G_2\sqrt{N})$, $\lambda = cG_2(N/\ell_\delta)^{1/2}$ gives

$$F(\bar{x}_N) - F(x^\star) \le CR_0 G_2 N^{-1/2} + CDG_2\sqrt{\frac{\ell_\delta}{N}} \tag{11}$$

with probability at least $1-\delta$. This second-moment calibration is also valid for $p > 2$, but its clipping correction retains the slower $N^{-1/2}$ rate.

3.2 Comparison and sample requirements

For $p > 2$, the tuned bounds (8) and (9) share the optimization and Gaussian terms $R_0 G_2 N^{-1/2} + CD\sigma_2\sqrt{\frac{\ell_\delta}{N}}$. Their $N^{-(1-1/p)}$ corrections differ in moment scale (σ_p versus G_p) and confidence dependence ($\delta^{-1/p}$ versus $\ell_\delta^{1-1/p}$).

To compare these corrections, we first require that the squared-gradient remainders can be absorbed into them. Sufficient conditions are

$$N^{1/2-1/p} \gtrsim_p \frac{R_0}{DG_2}\max\left\{\sigma_p\delta^{-1/p}, G_p\ell_\delta^{-1/p}\right\}, \quad N^{1/p} \gtrsim_p \frac{R_0\sigma_p}{DG_2}\delta^{1/p}\ell_\delta^{1/2}, \quad p > 4. \tag{12}$$

From now on, we write $\gtrsim_z$ (resp. $\lesssim_z$) to hide the multiplicative constants depending on z. The second condition is needed only for the Gaussian component of $\mathcal{Q}_{p,N}$ when $p > 4$.

Under (12), the U-SGD correction is no larger in order when

$$\sigma_p\delta^{-1/p} \lesssim_p G_p\ell_\delta^{1-1/p}, \quad \text{equivalently} \quad \left(\frac{\sigma_p}{G_p}\right)^p \lesssim_p \delta\ell_\delta^{p-1}. \tag{13}$$

This compares the correction scales in the upper bounds.

Define the shared sample requirement $N_0(\varepsilon,\delta) = 1 + \left(\frac{R_0G_2}{\varepsilon}\right)^2 + \left(\frac{D\sigma_2}{\varepsilon}\right)^2\ell_\delta$. Since a finite sum and the corresponding maximum agree up to a numerical factor, sufficient sample sizes satisfy

$$N_{\mathrm{U}} \lesssim_p N_0(\varepsilon,\delta) + \left(\frac{D\sigma_p}{\varepsilon}\right)^{p/(p-1)}\delta^{-1/(p-1)}, \quad N_{\mathrm{C}} \lesssim_p N_0(\varepsilon,\delta) + \left(\frac{DG_p}{\varepsilon}\right)^{p/(p-1)}\ell_\delta, \tag{14}$$

provided that horizons of these respective orders satisfy (12). Outside this regime, the sample requirements must retain the squared-gradient contributions from (8) and (9).

Up to constants depending only on p, condition (13) also characterizes when the additional U-SGD sample requirement in (14) is no larger.

At $p = 2$, (10) and (11) give

$$N_{\mathrm{U}}^{(2)} \lesssim 1 + \left(\frac{R_0G_2 + D\sigma_2}{\varepsilon}\right)^2\delta^{-1}, \quad N_{\mathrm{C}}^{(2)} \lesssim 1 + \left(\frac{R_0G_2}{\varepsilon}\right)^2 + \left(\frac{DG_2}{\varepsilon}\right)^2\ell_\delta. \tag{15}$$

We write $\gtrsim$ (resp. $\lesssim$) to hide the multiplicative constants. Polynomial confidence dependence also appears in the exact quadratic Polyak–Ruppert recursion in Corollary A.7. This does not rule out improving confidence by running several independent copies and combining their outputs.

3.3 Extension to strongly convex quadratics

For unconstrained strongly convex quadratics, we analyze SGD under conditional p-th moment bounds on the noise, $p \ge 2$. The method uses constant-step epochs, averages the second half of each epoch, and restarts from this average. Appendix C gives the setup, algorithm, and proofs.

The bounds separate deterministic contraction, a Gaussian variance term, and a heavy-tail correction (Lemma C.6). For fixed $p > 2$ and fixed confidence, the heavy-tail correction is lower order as the target error decreases. Consequently, ordinary SGD and the noise-centered clipping benchmark have the same leading accuracy dependence in our bounds (Theorem C.9). Clipping improves the confidence dependence of the heavy-tail correction from polynomial to logarithmic. At $p = 2$, this improvement affects the leading statistical term itself (Corollary C.10).

The clipped guarantees extend to inputs satisfying Assumption C.1. Difference clipping is covered when its anchor error and difference moments satisfy Proposition C.12, with the additional oracle work accounted for in (75).

4 Polyak–Ruppert Asymptotics

The finite-sample bounds allow the threshold to depend on the horizon and confidence. We now compare the limiting target and covariance when a finite threshold is held fixed as the horizon grows, before considering growing thresholds. Complete assumptions, statements, and proofs appear in Appendix D.

Unclipped SGD. Consider quadratic SGD with Hessian $H = H^\top \succ 0$, and step sizes $\eta_k = a(k+k_0)^{-\gamma}$, $\gamma \in (1/2, 1)$. Under the moment-decay condition in Proposition D.1 and the martingale-CLT conditions (78)–(79),

$$\sqrt{N}(\bar{x}_N - x^\star) \Longrightarrow \mathcal{N}(0, H^{-1}\Sigma H^{-1}).$$

The exact decomposition in Proposition D.1 explains this limit: to first order, the averaged error is the empirical mean of the gradient noise transformed by $-H^{-1}$.

Fixed-threshold clipping. Let $h_\lambda(x) = \mathbb{E}[\mathrm{clip}_\lambda(g(x,\xi))]$ and let x_λ be its stable root. Write $A_\lambda = \nabla h_\lambda(x_\lambda)$ and $\Sigma_\lambda = \mathrm{Cov}(\mathrm{clip}_\lambda(g(x_\lambda,\xi)))$. Under Assumption D.2, Theorem D.4 gives

$$\sqrt{N}(\bar{x}_N - x_\lambda) \Longrightarrow \mathcal{N}(0, A_\lambda^{-1}\Sigma_\lambda A_\lambda^{-\top}).$$

Thus clipping can change both the target and the limiting covariance. In particular, $x_\lambda \neq x^\star$ produces a persistent bias; see Proposition D.5.

For the scalar model $g(x,\xi) = \mu x + \xi$ with continuous symmetric noise, the target is preserved. Under the same PR regularity conditions, Corollary D.6 yields $V_\lambda = \frac{\mathbb{E}\min\{\xi^2,\lambda^2\}}{\mu^2\mathbb{P}(|\xi|<\lambda)^2}$. Clipping reduces the noise variance but also weakens the mean-field slope. Every finite threshold strictly increases V_λ relative to unclipped averaging for nondegenerate Gaussian noise, whereas suitable finite thresholds can reduce it for some symmetric heavy-tailed distributions.

Growing thresholds. In the linear additive-noise model, noise-centered clipping can recover the unclipped limit while controlling the clipping bias. For i.i.d. noise with a finite p-th moment, $p > 2$, and bounded or polylogarithmic $\ell_N = \log(4/\delta_N)$, Theorem D.7 gives the sufficient choice

$$\lambda_N = N^\beta, \quad \frac{1}{2(p-1)} < \beta < \frac{1}{2}.$$

This makes both the clipping bias and the bounded-increment concentration term negligible on the $N^{-1/2}$ scale and preserves the covariance $H^{-1}\Sigma H^{-1}$. The accompanying finite-horizon bound, including the PR remainder, is given in (91). Here one threshold λ_N is used throughout the length-N run; online thresholds and whole-gradient clipping require separate analysis.

5 When Does Clipping Help?

We now express the stochastic comparison in terms of the requested accuracy and confidence. For $\sigma_2 > 0$, define the normalized accuracies

$$\rho_{\mathrm{cvx}} = \frac{\varepsilon}{D\sigma_2}, \quad \rho_{\mathrm{sc}} = \frac{\sqrt{\mu\varepsilon}}{\sigma_2}, \quad \chi_p = \frac{\sigma_p}{\sigma_2},$$

for the convex and strongly convex quadratic settings, respectively. We compare the Gaussian and heavy-tail components of the bounds; the remaining optimization terms require separate control.

Proposition 5.1 (Accuracy-normalized comparison). *Let $p > 2$ and let ρ denote the appropriate normalized accuracy. At $N \asymp \rho^{-2}\ell_\delta$, where $\ell_\delta = \log(4/\delta)$, the respective heavy-tail corrections are at most of the Gaussian order under*

$$\chi_p^p \rho^{p-2} \lesssim_p \delta \ell_\delta^{p-1} \qquad \textit{for U-SGD}, \tag{16}$$

$$\chi_p^p \rho^{p-2} \lesssim_p 1 \qquad \textit{for NC}. \tag{17}$$

For whole-gradient clipping, replace χ_p in (17) *by G_p/σ_2. The derivation is given in Appendix A.3.*

For fixed $p > 2$ and confidence, requiring a smaller target error drives $\chi_p^p \rho^{p-2}$ to zero. Thus the bounds eventually share a Gaussian leading order, and balanced clipping changes a lower-order correction. At $p = 2$, its confidence improvement affects the leading statistical term. Changes in the limiting target and covariance are addressed in Section 4. Table 1 in Appendix A.3 summarizes the resulting regimes and their interpretation.

6 Off-Policy Click Prediction

We test whether a threshold selected at one budget remains beneficial as the optimization horizon grows. Off-policy learning makes clipping particularly appealing: a small logging propensity produces a large importance-weighted gradient. Truncating large importance weights is a classical variance-control device in importance sampling and counterfactual learning (Ionides, 2008; Bottou et al., 2013). Here we keep the importance weights intact and clip the stochastic gradient.

We use the full `all` campaign of the Open Bandit Dataset (OBD) (Saito et al., 2020). The behavior policy is Bernoulli Thompson sampling (BTS), and the target policy is Uniform Random over 80 items. If μ_i is the released probability that BTS displayed item i, inverse propensity scoring (IPS) reweights that observation by the target-to-logging probability ratio $\rho_i = (1/80)/\mu_i = (80\mu_i)^{-1}$. The experiment uses $12{,}357{,}200$ BTS impressions and a separately collected Random-policy log of $1{,}374{,}327$ impressions; no importance weight is truncated.

We fit a structured Brier click scorer, which gives an exact quadratic objective. Write $\mathcal{D}_{\mathrm{B}}$ and $\mathcal{D}_{\mathrm{R}}$ for the BTS and Random logs, with sizes n_{B} and n_{R}. Let $Y_i \in \{0, 1\}$ denote the click, let $j_i \in \{1, \ldots, 36\}$ index the display-position–item-category cell, and let e_j be the jth standard basis vector. For $w \in [0, 1]^{36}$, we denote by $\widehat{F}_{\mathrm{B}}^{\mathrm{IPS}}$ the IPS-weighted empirical Brier risk on the BTS log and by $\widehat{F}_{\mathrm{R}}$ the empirical Brier risk on the independent Random-policy log:

$$\widehat{F}_{\mathrm{B}}^{\mathrm{IPS}}(w) = \frac{1}{2n_{\mathrm{B}}} \sum_{i \in \mathcal{D}_{\mathrm{B}}} \rho_i (w_{j_i} - Y_i)^2,$$

$$\widehat{F}_{\mathrm{R}}(w) = \frac{1}{2n_{\mathrm{R}}} \sum_{i \in \mathcal{D}_{\mathrm{R}}} (w_{j_i} - Y_i)^2,$$

$$g_i(w) = \rho_i (w_{j_i} - Y_i) e_{j_i}.$$

Thus g_i is unbiased for the empirical IPS source objective $\widehat{F}_{\mathrm{B}}^{\mathrm{IPS}}$, while $\widehat{F}_{\mathrm{R}}$ is the separately collected target-policy evaluation risk. All 36 cells have positive target frequency, so the empirical target objective is strongly convex on this feature space. We compare projected U-SGD with fixed whole-gradient clipping h^{WG} (2).

Smaller λ means more aggressive clipping, while $\lambda = \infty$ recovers U-SGD. All methods return the complete post-update Polyak–Ruppert average; let $\bar{w}_N^{(M)}$ denote the average returned by method M after N sampled BTS rows.

The stepsize and clipping threshold are selected on separate source-validation trajectories before target evaluation. All methods share the selected stepsize, and the selected threshold $\lambda_{\mathrm{val}} = 5$ remains fixed across evaluation horizons. The Random-policy log is used only for evaluation. Appendix F gives the complete selection protocol, including the threshold grid and the independent evaluation runs.

For any method M, define its target upper-tail error by

$$G_M(N) = Q_{0.95}\left(\widehat{F}_{\mathrm{R}}(\bar{w}_N^{(M)}) - \min_{w\in[0,1]^{36}} \widehat{F}_{\mathrm{R}}(w)\right),$$
$$R_M(N) = \frac{G_M(N)}{G_{\mathrm{U}}(N)}. \tag{18}$$

The figure reports the absolute error G_M, for which lower is better. Ratios R_M are used only to state paired effect sizes and confidence intervals.

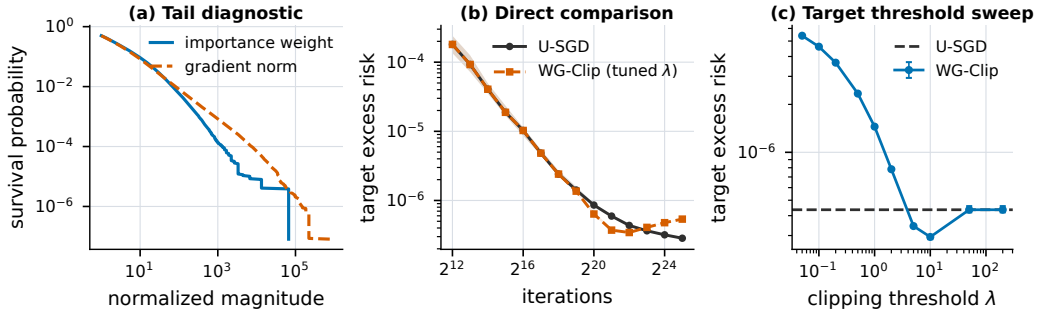

Figure 1: OBD off-policy click prediction. Lower is better in (b)–(c). (a) Empirical survival functions of the importance weight and stochastic-gradient norm at the IPS optimum, each normalized by its median; the effective-sample fraction is 0.16% and the largest weight is 12500. (b) Direct target-error comparison: $G_M(N)$ from (18) for U-SGD and WG-Clip with the single source-validation-selected threshold $\lambda_{\mathrm{val}} = 5$, held fixed throughout every trajectory. (c) Target-error sensitivity $G_M(4194304)$ across every predeclared finite threshold; the dashed horizontal line is U-SGD. Points below it favor WG-Clip, whereas points above it favor U-SGD. This post-selection sweep is a diagnostic and is not used for tuning. Bands and bars are pointwise distribution-free 95% order-statistic intervals conditional on the fixed logs; paired-bootstrap intervals are reported in the text.

Figure 1(b) shows the comparison as the budget grows. At the threshold-selection horizon, on the fresh evaluation trajectories, the frozen clipped method has ratio $R_{\mathrm{Clip}} = 0.63$ with paired 95% interval $[0.60, 0.65]$. Thus source-only threshold selection yields a finite-horizon gain on fresh evaluation trajectories. The same fixed method is still better at $N = 4194304$ ($R_{\mathrm{Clip}} = 0.79$), but the curves subsequently cross. The first displayed checkpoint after which the pointwise paired lower bound remains above one is $N = 8388608$. At $N = 33554432$, the clipped-to-unclipped ratio is 1.89 $[1.81, 1.95]$. Thus the method chosen without target information first helps and later loses, consistently with its shifted mean-field target becoming the dominant error.

Panel (c) reports sensitivity to the threshold; it is evaluated separately from the selection used in panel (b). At $N = 4194304$, the smallest error in the displayed finite grid occurs at $\lambda = 10$, with ratio 0.67 $[0.64, 0.70]$ relative to U-SGD, whereas aggressive clipping is substantially worse and very large thresholds coincide with U-SGD. Together, panels (b)–(c) show that a threshold that helps at one budget need not remain beneficial as the run grows.

7 Conclusion

The benefit of clipping depends on the requested accuracy, confidence, and clipping bias. Under a finite second moment alone, robustification improves the confidence dependence of the ordinary untrimmed average, and bias-balanced clipping provides logarithmic confidence bounds. For $p > 2$, unclipped Polyak–Ruppert averaging enters a finite-sample Gaussian-dominated regime, so clipping need not improve the leading order. Fixed whole-gradient clipping can also attenuate the signal, shift the limiting target under asymmetric noise, and change asymptotic efficiency. In the linear noise-centered setting, a growing threshold can recover the standard first-order Polyak–Ruppert limit. The bounds, exact examples, and off-policy experiment therefore support comparing clipping policies at the intended budget while accounting for both confidence gains and their effect on the target.

AI USE STATEMENT

Generative AI tools were used for language editing and to improve the clarity of the manuscript. We reviewed all resulting changes and take full responsibility for the final content.

REFERENCES

Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013. URL https://www.jmlr.org/papers/v14/bottou13a.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454.

Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38, 2021.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. In *Advances in Neural Information Processing Systems 34*, pp. 30063–30074, 2021.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for polyak–ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 50 (2):935–964, 2025. doi: 10.1287/moor.2022.0179. Published online in 2024.

Khaled Eldowa and Andrea Paudice. General tail bounds for non-smooth stochastic mirror descent. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3205–3213, 2024.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can sgd handle heavy-tailed noise?, 2025. arXiv:2508.04860.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems 33*, pp. 15042–15053, 2020.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15951–16070, 2024.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

Chuan He and Zhaosong Lu. Accelerated stochastic first-order method for convex optimization under heavy-tailed noise, 2025. arXiv:2510.11676.

Dongyan Lucy Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in markovian linear stochastic approximation with constant step sizes. *Mathematics of Operations Research*, 2026. doi: 10.1287/moor.2024.0471. Articles in Advance.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2): 295–311, 2008. doi: 10.1198/106186008X320456.

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. PMLR, 2023.

Zijian Liu. In-expectation convergence of stochastic gradient methods under heavy-tailed noise, 2026. arXiv:2606.00520.

Gabor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of sgd with gradient clipping in high-dimensions. In *International Conference on Learning Representations*, volume 2025, pp. 27381–27417, 2025.

Mattes Mollenhauer and Christian Fiedler. Fuk–nagaev inequality in smooth banach spaces: Optimum bounds for distributions of heavy-tailed martingales, 2025. arXiv:2512.10012.

Wenlong Mou, Chris Junchi Li, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. On linear stochastic approximation: Fine-grained polyak–ruppert and non-asymptotic concentration. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2947–2997, 2020.

Sergey Victorovich Nagaev. On probability and moment inequalities for supermartingales and martingales. *Acta Applicandae Mathematica*, 79(1-2):35–46, 2003.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise, 2023. arXiv:2302.05437.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path? *arXiv e-prints*, art. arXiv:1812.10004, December 2018. doi: 10.48550/arXiv.1812.10004.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, 2013.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648, 2023.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020.

Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. *Advances in Neural Information Processing Systems*, 37:12408–12460, 2024a.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4511–4547, 2024b.

Sergey Samsonov, Marina Sheshukova, Eric Moulines, and Alexey Naumov. Statistical inference for linear stochastic approximation with markovian noise. *Advances in Neural Information Processing Systems*, 38: 174565–174626, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

GUIDE TO THE APPENDICES

The appendices follow the main results. Each proof is kept with its statement or in the corresponding proof section.

Appendix	Contents
A	Concentration tools, crossover calculations, and lower constructions.
B	Complete proofs for projected convex optimization.
C	Quadratic epochs, restarts, and difference-clipping certificates.
D	Polyak–Ruppert limits, the shifted-root example, and growing thresholds.
E	Deterministic transit and constant-step covariance.
F	Protocol for the OBD experiment.
G	Hypotheses and proof targets for accelerated methods.

A CONCENTRATION TOOLS AND LOWER CONSTRUCTIONS

This appendix collects the concentration results, crossover calculations, and lower constructions used in the main text. We state the results for Hilbert-valued martingale differences; scalar versions follow immediately. Throughout, $\ell_\delta := \log(4/\delta)$. Constants denoted by C are absolute, while C_p may depend only on p; repeated uses in distinct statements need not denote the same numerical values.

Since $\log(4m/\delta) \le (1+\log m/\log 4)\,\ell_\delta$, combining m concentration events changes only numerical constants.

A.1 A MARTINGALE FUK–NAGAEV BOUND

Let $(d_k, \mathcal{F}_k)_{k=1}^N$ be martingale differences in a finite-dimensional Hilbert space. Define predictable moment budgets

$$V_N = \sum_{k=1}^N \mathbb{E}\left[\|d_k\|^2 \mid \mathcal{F}_{k-1}\right], \quad A_{p,N} = \sum_{k=1}^N \mathbb{E}\left[\|d_k\|^p \mid \mathcal{F}_{k-1}\right].$$

Theorem A.1 (martingale Fuk–Nagaev inequality)**.** *Let $p > 2$ and suppose $V_N \le v^2$ and $A_{p,N} \le a_p^p$ almost surely. Then, for every $\delta \in (0,1)$,*

$$\left\|\sum_{k=1}^N d_k\right\| \le C\, v\sqrt{\ell_\delta} + C_p a_p \delta^{-1/p}$$

with probability at least $1-\delta$, where $\ell_\delta = \log(4/\delta)$. A corresponding tail form states that, for every $t > 0$,

$$\mathbb{P}\left(\max_{j\le N}\left\|\sum_{k=1}^j d_k\right\| > t\right) \le 2\exp\left(-\frac{t^2}{Cv^2}\right) + \frac{C_p^p a_p^p}{t^p}.$$

The structure, rather than a particular numerical constant, is essential: the exponential term describes the accumulation of many moderate increments, while the power term is of the same order as the probability that one increment alone exceeds t. Results of this form go back to S. V. Nagaev and martingale extensions Nagaev (2003); sharp Banach-space formulations are based on Pinelis' smooth-norm inequalities Pinelis (1994).

Proof. Theorem 2.2 and Corollary 2.3 of Mollenhauer and Fiedler Mollenhauer & Fiedler (2025) apply under essential-supremum conditional moment budgets in a $(2, D)$-smooth Banach space. A Hilbert space is $(2,1)$-smooth, and their notation maps to ours as

$$q = p, \qquad D = 1, \qquad \sigma^2 = V_N \le v^2, \qquad C_q^q = A_{p,N} \le a_p^p, \qquad M_j = \sum_{k\le j} d_k.$$

Their confidence coefficient is

$$c_{p,1} = \frac{1}{2p} + \min\{1/p, 1/5\} + 1 + \mathbf{1}_{\{p>3\}} \frac{p}{3}.$$

The constants are therefore dimension-free in a Hilbert space, remain bounded as $p \downarrow 2$, and grow at most linearly in p for large p. The symbols C, C_p absorb the source factors $\sqrt{2}$, $2^{1/p}$, and $c_{p,1}$. □

At the endpoint $p = 2$, the decomposition does not provide two different scales. Chebyshev's inequality gives

$$\left\| \sum_{k=1}^{N} d_k \right\| \le v\delta^{-1/2}$$

with probability at least $1 - \delta$. Without a stronger assumption or a robust transformation, the factor $\delta^{-1/2}$ cannot be replaced uniformly by $\sqrt{\log(1/\delta)}$ for a single empirical average; see Catoni (Catoni, 2012, Proposition 6.2) for an explicit fixed-variance lower bound.

Corollary A.2 (weighted stochastic-gradient noise)**.** *Under Assumption 2.1, let w_k be predictable scalar weights. For $p > 2$, with probability at least $1 - \delta$,*

$$\left\| \sum_{k=1}^{N} w_k \zeta_k \right\| \le C\sigma_2 \sqrt{\ell_\delta \sum_{k=1}^{N} w_k^2} + C_p \sigma_p \delta^{-1/p} \left(\sum_{k=1}^{N} |w_k|^p \right)^{1/p}. \tag{19}$$

In particular, for $w_k = 1/N$,

$$\left\| \frac{1}{N} \sum_{k=1}^{N} \zeta_k \right\| \le r_{\mathrm{U}}^{\mathrm{mean}}(N, \delta) := C\sigma_2 \sqrt{\frac{\ell_\delta}{N}} + C_p \sigma_p \delta^{-1/p} N^{-(1-(1/p))} \tag{20}$$

with probability at least $1 - \delta$.

Proof. Take $d_k = w_k \zeta_k$ in Theorem A.1. Predictability of w_k and Assumption 2.1 give

$$V_N \le \sigma_2^2 \sum_{k=1}^{N} w_k^2, \qquad A_{p,N} \le \sigma_p^p \sum_{k=1}^{N} |w_k|^p.$$

Substitution proves (19). For $w_k = 1/N$, the two weight norms are $N^{-1/2}$ and $N^{-(1-1/p)}$, which gives (20). □

For fixed δ, the second term in (20) is lower order in N exactly when $p > 2$. Proposition A.6 shows that its (σ_p, N, δ) dependence is sharp for the empirical-mean mechanism, but not uniformly over classes with both σ_2 and σ_p fixed.

A.2 Bias and concentration after clipping

The next elementary lemma exposes all three effects of clipping.

Lemma A.3 (clipped conditional mean)**.** *Let Y be a random vector with conditional mean $m = \mathbb{E}[Y \mid \mathcal{G}]$.*

(i) *If $\mathbb{E}[\|Y\|^p \mid \mathcal{G}] \le M_p^p$, then*

$$\|\mathbb{E}[\mathrm{clip}_\lambda(Y) \mid \mathcal{G}] - m\| \le \frac{M_p^p}{\lambda^{p-1}}. \tag{21}$$

(ii) *If $Y = m + Z$, $\mathbb{E}[Z \mid \mathcal{G}] = 0$, and one clips Z rather than Y, then*

$$\|\mathbb{E}[\mathrm{clip}_\lambda(Z) \mid \mathcal{G}]\| \le \frac{\mathbb{E}[\|Z\|^p \mid \mathcal{G}]}{\lambda^{p-1}}. \tag{22}$$

Thus the raw scale M_p is replaced by the central scale σ_p.

(iii) *The projection* $\mathrm{clip}_\lambda(\cdot)$ *is nonexpansive. Consequently, for an independent conditional copy* Y'*,*

$$\mathbb{E}\Big[\|\mathrm{clip}_\lambda(Y) - \mathbb{E}[\mathrm{clip}_\lambda(Y) \mid \mathcal{G}]\|^2 \mid \mathcal{G}\Big] \le \frac{1}{2}\mathbb{E}\left[\|Y - Y'\|^2 \mid \mathcal{G}\right]. \tag{23}$$

For $Y = m + Z$ *this is at most* $\mathbb{E}\left[\|Z\|^2 \mid \mathcal{G}\right]$.

Proof. For every vector y,

$$\|\mathrm{clip}_\lambda(y) - y\| = (\|y\| - \lambda)_+ \le \|y\|\,\mathbf{1}\{\|y\| > \lambda\} \le \frac{\|y\|^p}{\lambda^{p-1}}.$$

Conditional Jensen's inequality yields (21); the same argument applied to Z proves (22).

For any square-integrable random vector U and an independent conditional copy U',

$$\mathbb{E}[\|U - \mathbb{E}[U \mid \mathcal{G}]\|^2 \mid \mathcal{G}] = \frac{1}{2}\mathbb{E}[\|U - U'\|^2 \mid \mathcal{G}]. \tag{24}$$

Indeed, conditional independence and equality of the conditional laws give

$$\mathbb{E}[\|U - U'\|^2 \mid \mathcal{G}] = 2\mathbb{E}[\|U\|^2 \mid \mathcal{G}] - 2\,\|\mathbb{E}[U \mid \mathcal{G}]\|^2.$$

Take $U = \mathrm{clip}_\lambda(Y)$. Since Euclidean projection is nonexpansive,

$$\|\mathrm{clip}_\lambda(Y) - \mathrm{clip}_\lambda(Y')\| \le \|Y - Y'\|,$$

which proves (23). If $Y = m + Z$, the right-hand side of (24) is $\mathbb{E}[\|Z\|^2 \mid \mathcal{G}]$. □

Combining these bias and variance bounds with bounded-increment martingale concentration yields the following control of clipped averages.

Proposition A.4 (concentration of clipped averages)**.** *Let* $m_k = \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]$*, and suppose that, almost surely for every* k*,* $\mathbb{E}[\|Y_k - m_k\|^2 \mid \mathcal{F}_{k-1}] \le \sigma_2^2$ *and* $\mathbb{E}[\|Y_k\|^p \mid \mathcal{F}_{k-1}] \le M_p^p$*. Then, for every* $\lambda > 0$ *and* $\delta \in (0,1)$*,*

$$\left\|\frac{1}{N}\sum_{k=1}^{N}\big(\mathrm{clip}_\lambda(Y_k) - m_k\big)\right\| \le r_{\mathrm{C}}^{\mathrm{mean}}(N,\delta;\lambda) := C\sigma_2\sqrt{\frac{\ell_\delta}{N}} + C\lambda\frac{\ell_\delta}{N} + \frac{M_p^p}{\lambda^{p-1}} \tag{25}$$

with probability at least $1-\delta$*. The non-Gaussian terms are balanced, up to constants, by*

$$\lambda_\star \asymp M_p\left(\frac{N}{\ell_\delta}\right)^{1/p}, \quad r_{\mathrm{C}}^{\mathrm{mean}}(N,\delta) \lesssim \sigma_2\sqrt{\frac{\ell_\delta}{N}} + M_p\left(\frac{\ell_\delta}{N}\right)^{1-1/p}. \tag{26}$$

Proof. Let

$$u_k = \mathrm{clip}_\lambda(Y_k) - \mathbb{E}[\mathrm{clip}_\lambda(Y_k) \mid \mathcal{F}_{k-1}].$$

Then (u_k) is a martingale-difference sequence and $\|u_k\| \le 2\lambda$. Lemma A.3(iii) gives

$$\mathbb{E}[\|u_k\|^2 \mid \mathcal{F}_{k-1}] \le \mathbb{E}[\|Y_k - m_k\|^2 \mid \mathcal{F}_{k-1}] \le \sigma_2^2.$$

Apply Pinelis' Bennett–Freedman inequality for martingales in 2-smooth Banach spaces (Pinelis, 1994, Theorem 3.4), with increment bound 2λ and quadratic-variation bound $N\sigma_2^2$, to obtain

$$\left\|\sum_{k=1}^{N} u_k\right\| \le C\sigma_2\sqrt{N\ell_\delta} + C\lambda\ell_\delta \tag{27}$$

with probability at least $1-\delta$. Divide by N and use Lemma A.3(i) for the average conditional bias to obtain (25).

To optimize the non-Gaussian part, minimize

$$\phi(\lambda) = c_1\lambda\frac{\ell_\delta}{N} + c_2\frac{M_p^p}{\lambda^{p-1}}.$$

The stationary equation is

$$c_1\frac{\ell_\delta}{N} = (p-1)c_2M_p^p\lambda^{-p},$$

and hence $\lambda \asymp M_p(N/\ell_\delta)^{1/p}$ and $\phi(\lambda) \asymp M_p(\ell_\delta/N)^{1-1/p}$, proving (26). □

For noise-centered clipping $M_p = \sigma_p$; for whole-gradient clipping the valid uniform choice is $M_p = G_p$.

A.3 CROSSOVER BOUNDARIES AND THEIR INTERPRETATION

There are two useful comparisons. First, compare the one-big-jump term of U-SGD with the Gaussian core.

Proposition A.5 (Gaussian-dominance boundary)**.** *Let $p > 2$, $\chi_p = \sigma_p/\sigma_2$, and define $K_p := \left(\frac{C_p}{C}\right)^p$, where C and C_p are the two coefficients displayed in* (20)*. The polynomial term in* (20) *is no larger than the Gaussian term whenever*

$$\delta\, \ell_\delta^{p/2} \geq K_p \chi_p^p N^{-(p-2)/2}. \tag{28}$$

Proof. The one-big-jump term is no larger than the Gaussian term if

$$C_p \sigma_p \delta^{-1/p} N^{-(1-1/p)} \leq C \sigma_2 \ell_\delta^{1/2} N^{-1/2}.$$

Divide by $\sigma_2 N^{-1/2}$ and raise both sides to the pth power. This gives (28) with $K_p = (C_p/C)^p$. □

For fixed δ and χ_p, the Gaussian-dominance condition (28) eventually holds for every $p > 2$, whereas at $p = 2$ increasing N does not improve the comparison.

For completeness, the same algebra proves the accuracy-normalized conditions in Proposition 5.1. Substituting $N \asymp \rho^{-2}\ell_\delta$ into (28) gives (16). Balancing the centered-clipping correction with the Gaussian term gives $\chi_p^p \rho^{p-2} \lesssim 1$, which is (17).

Second, compare only the non-Gaussian corrections of the unclipped and centered-clipped averages. Ignoring universal constants,

$$\frac{\sigma_p \delta^{-1/p} N^{-(1-1/p)}}{\sigma_p (\ell_\delta/N)^{1-1/p}} = \left(\frac{1}{\delta \ell_\delta^{p-1}}\right)^{1/p}.$$

Thus finite-threshold clipping improves this correction in the very-high-confidence region $\delta\ell_\delta^{p-1} < 1$, while the unclipped Nagaev correction can be smaller at moderate confidence and sufficiently large p. This comparison is independent of N because both corrections have the same homogeneity after the clipping-bias upper bound has been balanced. Their importance relative to the Gaussian core, however, decays with N when $p > 2$.

Table 1: Finite-sample guidance after the deterministic optimization error has been reduced below the target accuracy. The conclusions concern the guarantees established in this paper; instance-wise constants can still depend on the noise distribution.

Regime	Conclusion	Reason
Only $p = 2$ is known, and a distribution-free $1 - \delta$ guarantee is required	Use a target-preserving robustification	The ordinary untrimmed average has polynomial confidence dependence. Clipping is one option, but robust aggregation or robust mean estimation can provide the same type of confidence improvement.
$p > 2$ and $\chi_p^p \rho^{p-2} \gtrsim \delta \ell_\delta^{p-1}$	Finite clipping can improve the confidence correction if its bias is controlled	The heavy-tail correction of U-SGD is not yet guaranteed to be lower order. For whole-gradient clipping, the direct comparison must also account for G_p/σ_p.
$p > 2$ and $\chi_p^p \rho^{p-2} \lesssim \delta \ell_\delta^{p-1}$	U-SGD is already Gaussian dominated	Finite clipping does not improve the leading order, although it may change an instance-specific constant.
Large signal-to-noise ratio or asymmetric noise	Keep U-SGD as an explicit candidate	Whole-gradient clipping can truncate the optimization signal or change the limiting target. A finite threshold should therefore be used only after checking signal attenuation and clipping bias.

The table describes the distribution-free comparison. Fixed-law behavior can still differ: a finite threshold can improve a symmetric heavy-tailed location problem, whereas U-SGD is asymptotically efficient in the scalar Gaussian case, and fixed whole-gradient clipping can shift the limiting point under asymmetric noise.

A.4 THE ENDPOINT $p = 2$ AND MATCHING LOWER CONSTRUCTIONS

A simple scalar construction makes the obstruction transparent. For each N and δ, choose a centered distribution that equals a value of order $\sigma_2\sqrt{N/\delta}$ with probability of order δ/N, and uses a small compensating value otherwise. Its variance is of order σ_2^2. With probability of order δ, at least one of the N observations takes this exceptional value; conditional on this event, the empirical mean has magnitude of order $\sigma_2/\sqrt{N\delta}$. Therefore no distribution-free confidence radius of order $\sigma_2\sqrt{\log(1/\delta)/N}$ can hold for the ordinary mean under only a second-moment bound. Corollary A.7 embeds this mean exactly into a one-dimensional quadratic U-SGD recursion with Polyak–Ruppert averaging. If logarithmic confidence is required uniformly over the finite-variance class, some additional robustification (clipping, median-of-means, Catoni estimation, or confidence boosting) is needed.

For $p > 2$, the same construction is constrained by the higher moment: the exceptional value cannot simultaneously be as large and as frequent. Its contribution becomes $N^{-(1-(1/p))}\delta^{-1/p}$, matching the second term of (20) for this mechanism. This is the probabilistic reason why U-SGD re-enters the comparison for $p > 2$.

The following construction formalizes this argument for every $p \geq 2$.

Proposition A.6 (one-outlier lower bound for a finite p-th moment)**.** *Fix $p \geq 2$, $N \geq 2$, and $0 < \delta \leq 1/2$. There is a centered scalar law with $\mathbb{E}\,|X|^p = \sigma_p^p$ such that, for i.i.d. copies,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq c_p\sigma_p\delta^{-1/p}N^{-(1-1/p)}\right) \geq c\delta. \tag{29}$$

For $p = 2$, the same law has variance σ_2^2 and the radius becomes $c\sigma_2/\sqrt{N\delta}$.

Proof. Set $q = \delta/(4N)$ and let

$$X = \begin{cases} A, & \text{with probability } q, \\ -qA/(1-q), & \text{with probability } 1-q, \end{cases} \qquad A = \sigma_p\left(q + \frac{q^p}{(1-q)^{p-1}}\right)^{-1/p}. \tag{30}$$

Then $\mathbb{E}X = 0$ and $\mathbb{E}\,|X|^p = \sigma_p^p$. Let E be the event that exactly one of $X_1, \ldots, X_N$ equals A. Since $Nq \leq 1/8$,

$$\mathbb{P}(E) = Nq(1-q)^{N-1} \geq c\delta.$$

On E,

$$\frac{1}{N}\sum_{i=1}^{N} X_i = \frac{A(1-Nq)}{N(1-q)}.$$

For $q \leq 1/8$, the factor defining A is comparable to q, while $(1-Nq)/(1-q)$ is bounded below. The right-hand side is therefore at least $c_p\sigma_p q^{-1/p}/N$, which is the scale in (29). □

Corollary A.7 (an exact quadratic U-SGD lower bound for every $p \geq 2$)**.** *Consider the one-dimensional quadratic*

$$F(x) = \frac{x^2}{2}, \qquad g(x, X) = x + X,$$

run U-SGD with the admissible constant step $\eta_k \equiv 1$ and output $\bar{x}_N = N^{-1}\sum_{k=1}^{N} x_k$. For the law in (30)*, for every $p \geq 2$,*

$$\mathbb{P}\left(F(\bar{x}_N) - F(0) \geq c_p\sigma_p^2\delta^{-2/p}N^{-2(1-1/p)}\right) \geq c'\delta. \tag{31}$$

In particular, at $p = 2$,

$$\mathbb{P}\left(F(\bar{x}_N) - F(0) \geq c\frac{\sigma_2^2}{N\delta}\right) \geq c'\delta. \tag{32}$$

Thus the one-jump objective term is sharp for this explicit untrimmed single-run Polyak–Ruppert mechanism. The statement does not cover median-of-means, Catoni-type final aggregation, or confidence boosting across independent runs.

Proof. Index the update from x_{k-1} to x_k. The recursion is exact: $x_k = x_{k-1} - (x_{k-1} + X_k) = -X_k$. Hence $\bar{x}_N = -N^{-1}\sum_{k=1}^N X_k$. Proposition A.6, followed by $F(x) = x^2/2$, proves (31); (32) is its $p = 2$ specialization. □

B Proofs for Convex Optimization

Throughout this appendix, write $\mathbb{E}_{k-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$ and use $\ell_\delta = \log(12/\delta)$. Constants C are absolute, whereas C_p may depend only on p. All stepsizes and clipping thresholds considered below are deterministic.

B.1 The common projected inequality

We first derive the common projected inequality. For any direction h_k, write

$$h_k = s_k + b_k + u_k, \qquad \mathbb{E}_{k-1}[u_k] = 0, \tag{33}$$

where b_k is a predictable bias. Nonexpansiveness of projection gives

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &\le \|x_k - x^\star - \eta h_k\|^2 \\ &= \|x_k - x^\star\|^2 - 2\eta\,\langle s_k, x_k - x^\star\rangle - 2\eta\,\langle b_k + u_k, x_k - x^\star\rangle + \eta^2\,\|h_k\|^2 .\end{aligned}$$

Since $F(x_k) - F(x^\star) \le \langle s_k, x_k - x^\star\rangle$, we obtain

$$\begin{aligned}F(x_k) - F(x^\star) \le \frac{\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2}{2\eta} + \frac{\eta}{2}\,\|h_k\|^2 \\ - \langle u_k, x_k - x^\star\rangle - \langle b_k, x_k - x^\star\rangle .\end{aligned}$$

Summing, dividing by N, and applying Jensen's inequality proves (7). Bounding the bias term using $\|x_k - x^\star\| \le D$ also gives

$$\begin{aligned}F(\bar{x}_N) - F(x^\star) \le \frac{R_0^2}{2\eta N} + \frac{\eta}{2N}\sum_{k=1}^N \|h_k\|^2 \\ - \frac{1}{N}\sum_{k=1}^N \langle u_k, x_k - x^\star\rangle + \frac{D}{N}\sum_{k=1}^N \|b_k\| .\end{aligned} \tag{34}$$

For U-SGD, $h_k = g_k$, $b_k = 0$, and $u_k = \zeta_k$. Define

$$d_k := \langle \zeta_k, x_k - x^\star\rangle . \tag{35}$$

Predictability of x_k gives

$$\mathbb{E}_{k-1}[d_k] = 0, \qquad \mathbb{E}_{k-1}[d_k^2] \le D^2\sigma_2^2, \qquad \mathbb{E}_{k-1}[|d_k|^p] \le D^p\sigma_p^p.$$

For $p > 2$, Theorem A.1, with failure probability $\delta/3$, therefore yields

$$\frac{1}{N}\left|\sum_{k=1}^N d_k\right| \le CD\sigma_2\sqrt{\frac{\ell_\delta}{N}} + C_p D\sigma_p \delta^{-1/p} N^{-(1-1/p)}. \tag{36}$$

The deterministic moment budgets used here are $ND^2\sigma_2^2$ and $ND^p\sigma_p^p$.

B.2 Proof of Theorem 3.1

Proof. We first separate the predictable signal from the noise in the squared-gradient term. Since $\mathbb{E}_{k-1}[\zeta_k] = 0$,

$$\|s_k\|^2 + \mathbb{E}_{k-1}\,\|\zeta_k\|^2 = \mathbb{E}_{k-1}\,\|g_k\|^2 \le G_2^2. \tag{37}$$

In particular, $\|s_k\| \le G_2$.

Define $Z_k := \|\zeta_k\|^2 - \mathbb{E}_{k-1}\,\|\zeta_k\|^2$ and $c_k := \langle s_k, \zeta_k\rangle$. Both sequences are scalar martingale differences, and

$$\|g_k\|^2 = \mathbb{E}_{k-1}\,\|g_k\|^2 + 2c_k + Z_k. \tag{38}$$

For the cross term, predictability of s_k and (37) imply

$$\mathbb{E}_{k-1}[c_k^2] \le G_2^2\sigma_2^2, \qquad \mathbb{E}_{k-1}[|c_k|^p] \le G_2^p\sigma_p^p.$$

Applying Theorem A.1 with failure probability $\delta/3$ gives

$$\frac{1}{N}\left|\sum_{k=1}^N c_k\right| \le CG_2\sigma_2\sqrt{\frac{\ell_\delta}{N}} + C_pG_2\sigma_p\delta^{-1/p}N^{-(1-1/p)}. \tag{39}$$

We next control the centered squared noise. Put $q = p/2 > 1$. Conditional Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_{k-1}[|Z_k|^q] &\le 2^{q-1}\left(\mathbb{E}_{k-1}\left\|\zeta_k\right\|^{2q} + \left(\mathbb{E}_{k-1}\left\|\zeta_k\right\|^2\right)^q\right)\\ &\le 2^q\mathbb{E}_{k-1}\left\|\zeta_k\right\|^{2q} \le C_p\sigma_p^p.\end{aligned}$$

If $2 < p \le 4$, then $1 < q \le 2$. The martingale von Bahr–Esseen inequality yields

$$\mathbb{E}\left|\sum_{k=1}^N Z_k\right|^q \le C_q\sum_{k=1}^N\mathbb{E}|Z_k|^q \le C_pN\sigma_p^p. \tag{40}$$

For completeness, the first inequality follows by applying

$$|a+b|^q \le |a|^q + q|a|^{q-1}\operatorname{sgn}(a)b + C_q|b|^q$$

to the partial sums and taking conditional expectations. The linear term vanishes because Z_k is a martingale difference.

Markov's inequality applied to (40) gives, with probability at least $1-\delta/3$,

$$\frac{1}{N}\left|\sum_{k=1}^N Z_k\right| \le C_p\sigma_p^2\delta^{-2/p}N^{-(1-2/p)}, \qquad 2 < p \le 4.$$

If $p > 4$, then $q = p/2 > 2$. Moreover,

$$\mathbb{E}_{k-1}[Z_k^2] \le \mathbb{E}_{k-1}\left\|\zeta_k\right\|^4 \le \left(\mathbb{E}_{k-1}\left\|\zeta_k\right\|^p\right)^{4/p} \le \sigma_p^4.$$

Thus Theorem A.1, applied at moment order q with deterministic budgets $N\sigma_p^4$ and $C_pN\sigma_p^p$, gives, with probability at least $1-\delta/3$,

$$\frac{1}{N}\left|\sum_{k=1}^N Z_k\right| \le C\sigma_p^2\sqrt{\frac{\ell_\delta}{N}} + C_p\sigma_p^2\delta^{-2/p}N^{-(1-2/p)}.$$

Consequently, in either case,

$$\frac{1}{N}\left|\sum_{k=1}^N Z_k\right| \le \mathcal{Q}_{p,N}(\delta)$$

with probability at least $1-\delta/3$.

Combining (38), (39), and the bound on Z_k yields

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^N\left\|g_k\right\|^2 \le G_2^2 &+ CG_2\sigma_2\sqrt{\frac{\ell_\delta}{N}}\\ &+ C_pG_2\sigma_p\delta^{-1/p}N^{-(1-1/p)} + \mathcal{Q}_{p,N}(\delta).\end{aligned}$$

Intersect this event with the event in (36). The three concentration events have total failure probability at most δ. Substitution into (34) gives the arbitrary-stepsize bound

$$\begin{aligned}F(\bar{x}_N) - F(x^\star) \le \frac{R_0^2}{2\eta N} + \frac{\eta G_2^2}{2} &+ C(D+\eta G_2)\sigma_2\sqrt{\frac{\ell_\delta}{N}}\\ &+ C_p(D+\eta G_2)\sigma_p\delta^{-1/p}N^{-(1-1/p)} + \frac{\eta}{2}\mathcal{Q}_{p,N}(\delta).\end{aligned} \tag{41}$$

Finally, choose $\eta = R_0/(G_2\sqrt{N})$. Then

$$\frac{R_0^2}{2\eta N} + \frac{\eta G_2^2}{2} = \frac{R_0G_2}{\sqrt{N}}, \qquad D + \eta G_2 = D + \frac{R_0}{\sqrt{N}} \le 2D.$$

Substituting these relations into (41) proves (8). □

B.3 Proof of Theorem 3.2

Proof. We first derive an arbitrary-parameter bound. This part of the argument is valid for every $p \geq 2$.

Set $h_k = \mathrm{clip}_\lambda(g_k)$ and define

$$b_k(\lambda) := \mathbb{E}_{k-1}[h_k] - s_k, \qquad \widetilde{\zeta}_k(\lambda) := h_k - \mathbb{E}_{k-1}[h_k].$$

Thus (33) holds with $b_k = b_k(\lambda)$ and $u_k = \widetilde{\zeta}_k(\lambda)$.

Lemma A.3 gives

$$\|b_k(\lambda)\| \leq \frac{G_p^p}{\lambda^{p-1}}, \qquad \mathbb{E}_{k-1}\left\|\widetilde{\zeta}_k(\lambda)\right\|^2 \leq \sigma_2^2. \tag{42}$$

Since $\|h_k\| \leq \lambda$, we also have

$$\left\|\widetilde{\zeta}_k(\lambda)\right\| \leq 2\lambda.$$

The contribution of the predictable bias is therefore bounded by

$$\frac{D}{N}\sum_{k=1}^{N}\|b_k(\lambda)\| \leq \frac{DG_p^p}{\lambda^{p-1}}. \tag{43}$$

Define the scalar martingale differences $\widetilde{d}_k := \left\langle \widetilde{\zeta}_k(\lambda), x_k - x^\star \right\rangle$. Their increments and conditional variances satisfy

$$|\widetilde{d}_k| \leq 2D\lambda, \qquad \mathbb{E}_{k-1}[\widetilde{d}_k^2] \leq D^2\sigma_2^2.$$

The Bernstein–Freedman argument used for (27), applied to these scalar increments, gives, with probability at least $1 - \delta/3$,

$$\frac{1}{N}\left|\sum_{k=1}^{N}\widetilde{d}_k\right| \leq CD\sigma_2\sqrt{\frac{\ell_\delta}{N}} + CD\lambda\frac{\ell_\delta}{N}. \tag{44}$$

For the squared-gradient term, let

$$W_k := \|h_k\|^2 - \mathbb{E}_{k-1}\|h_k\|^2.$$

Since $0 \leq \|h_k\|^2 \leq \lambda^2$, we have $|W_k| \leq \lambda^2$. Moreover,

$$\begin{aligned}\mathbb{E}_{k-1}[W_k^2] &\leq \mathbb{E}_{k-1}\|h_k\|^4 \\ &\leq \lambda^2\mathbb{E}_{k-1}\|h_k\|^2 \leq \lambda^2 G_2^2.\end{aligned}$$

Here we used $\|h_k\| \leq \|g_k\|$. Bernstein–Freedman concentration, together with $\mathbb{E}_{k-1}\|h_k\|^2 \leq G_2^2$, gives, with probability at least $1 - \delta/3$,

$$\frac{1}{N}\sum_{k=1}^{N}\|h_k\|^2 \leq G_2^2 + CG_2\lambda\sqrt{\frac{\ell_\delta}{N}} + C\lambda^2\frac{\ell_\delta}{N}. \tag{45}$$

The intersection of the events in (44) and (45) has probability at least $1 - 2\delta/3$, hence at least $1 - \delta$. Combining these inequalities with (34) and (43) yields

$$\begin{aligned}F(\bar{x}_N) - F(x^\star) \leq{}& \frac{R_0^2}{2\eta N} + \frac{\eta G_2^2}{2} + CD\sigma_2\sqrt{\frac{\ell_\delta}{N}} \\ &+ CD\lambda\frac{\ell_\delta}{N} + \frac{DG_p^p}{\lambda^{p-1}} \\ &+ C\eta G_2\lambda\sqrt{\frac{\ell_\delta}{N}} + C\eta\lambda^2\frac{\ell_\delta}{N}.\end{aligned} \tag{46}$$

Now let $p > 2$ and choose

$$\eta = \frac{R_0}{G_2\sqrt{N}}, \qquad \lambda = c_p G_p (N/\ell_\delta)^{1/p},$$

where $c_p > 0$ depends only on p. The first two terms in (46) sum to $R_0 G_2/\sqrt{N}$. Also,

$$D\lambda\frac{\ell_\delta}{N} + \frac{DG_p^p}{\lambda^{p-1}} \le C_p D G_p \ell_\delta^{1-1/p} N^{-(1-1/p)},$$
$$\eta G_2 \lambda\sqrt{\frac{\ell_\delta}{N}} = c_p R_0 G_p \ell_\delta^{1/2-1/p} N^{-(1-1/p)},$$
$$\eta\lambda^2\frac{\ell_\delta}{N} = c_p^2\frac{R_0 G_p^2}{G_2}\ell_\delta^{1-2/p} N^{-(3/2-2/p)}.$$

Because $R_0 \le D$ and $\ell_\delta \ge 1$, the middle expression is at most

$$c_p D G_p \ell_\delta^{1-1/p} N^{-(1-1/p)}.$$

Substitution into (46) proves (9). □

B.4 Proof of (10)

Proof. The scalar martingale differences d_k defined in (35) satisfy

$$\mathbb{E}\left|\sum_{k=1}^N d_k\right|^2 = \sum_{k=1}^N \mathbb{E}[d_k^2] \le N D^2\sigma_2^2.$$

Chebyshev's inequality therefore gives, with probability at least $1 - \delta/2$,

$$\frac{1}{N}\left|\sum_{k=1}^N d_k\right| \le \frac{\sqrt{2}D\sigma_2}{\sqrt{N\delta}}.$$

Also,

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^N \|g_k\|^2\right] \le G_2^2.$$

By Markov's inequality, with probability at least $1 - \delta/2$,

$$\frac{1}{N}\sum_{k=1}^N \|g_k\|^2 \le \frac{2G_2^2}{\delta}.$$

On the intersection of these two events, (34) gives

$$F(\bar{x}_N) - F(x^\star) \le \frac{R_0^2}{2\eta N} + \frac{\eta G_2^2}{\delta} + \frac{\sqrt{2}D\sigma_2}{\sqrt{N\delta}}.$$

With $\eta = R_0\sqrt{\delta}/(G_2\sqrt{N})$, the first two terms sum to $3R_0G_2/(2\sqrt{N\delta})$. Thus, with probability at least $1 - \delta$,

$$F(\bar{x}_N) - F(x^\star) \le C\frac{R_0 G_2 + D\sigma_2}{\sqrt{N\delta}},$$

which proves (10). □

B.5 Proof of (11)

Proof. The preceding clipping argument applies with $p = 2$ and the raw second-moment bound G_2^2. In addition to (42), we have

$$\mathbb{E}_{k-1}\left\|\widetilde{\zeta}_k(\lambda)\right\|^2 \le \mathbb{E}_{k-1}\|h_k\|^2 \le G_2^2.$$

Thus the martingale bound can use G_2 as its variance scale, regardless of the stated value of σ_2.

Consequently, (46) holds with $p = 2$, with G_2^2/λ as the bias bound, and with G_2 in place of σ_2. Choose

$$\eta = \frac{R_0}{G_2\sqrt{N}}, \qquad \lambda = cG_2\sqrt{\frac{N}{\ell_\delta}},$$

where $c > 0$ is a fixed numerical constant. Then

$$D\lambda\frac{\ell_\delta}{N} + \frac{DG_2^2}{\lambda} = (c + c^{-1})DG_2\sqrt{\frac{\ell_\delta}{N}},$$

and

$$\eta G_2\lambda\sqrt{\frac{\ell_\delta}{N}} = c\frac{R_0G_2}{\sqrt{N}}, \qquad \eta\lambda^2\frac{\ell_\delta}{N} = c^2\frac{R_0G_2}{\sqrt{N}}.$$

Substituting these identities into the general bound gives, with probability at least $1 - \delta$,

$$F(\bar{x}_N) - F(x^\star) \le C\frac{R_0G_2}{\sqrt{N}} + CDG_2\sqrt{\frac{\ell_\delta}{N}},$$

which proves (11). □

B.6 Remainder conditions and sample requirements

We verify the conditions used in the comparison in Section 3. For $p > 2$, write

$$r := \frac{1}{2} - \frac{1}{p} > 0, \qquad T_{\mathrm{U}} := D\sigma_p\delta^{-1/p}N^{-(1-1/p)}, \qquad T_{\mathrm{C}} := DG_p\ell_\delta^{1-1/p}N^{-(1-1/p)}.$$

These are the heavy-tail correction scales, with constants depending only on p suppressed.

The polynomial squared-gradient remainders factor as

$$\frac{R_0\sigma_p^2}{G_2}\delta^{-2/p}N^{-(3/2-2/p)} = T_{\mathrm{U}}\left(\frac{R_0\sigma_p}{DG_2}\delta^{-1/p}N^{-r}\right),$$
$$\frac{R_0G_p^2}{G_2}\ell_\delta^{1-2/p}N^{-(3/2-2/p)} = T_{\mathrm{C}}\left(\frac{R_0G_p}{DG_2}\ell_\delta^{-1/p}N^{-r}\right).$$

Thus the first condition in (12) absorbs these two remainders into their respective heavy-tail corrections.

For $p > 4$, the Gaussian component of $\frac{R_0}{2G_2\sqrt{N}}\mathcal{Q}_{p,N}(\delta)$ has scale

$$\frac{R_0\sigma_p^2}{G_2}\frac{\sqrt{\ell_\delta}}{N} = T_{\mathrm{U}}\left(\frac{R_0\sigma_p}{DG_2}\delta^{1/p}\ell_\delta^{1/2}N^{-1/p}\right).$$

The second condition in (12) absorbs this term as well.

After these absorptions, the ratio of the two correction scales is

$$\frac{T_{\mathrm{U}}}{T_{\mathrm{C}}} = \frac{\sigma_p}{G_p}\delta^{-1/p}\ell_\delta^{-(1-1/p)}.$$

Taking the p-th power gives (13), with constants depending only on p.

To obtain the sample requirements, observe that

$$N \ge K_pN_0(\varepsilon, \delta)$$

makes the optimization and Gaussian terms at most a prescribed constant fraction of ε, for a sufficiently large K_p. Similarly,

$$N \ge K_p\left(\frac{D\sigma_p}{\varepsilon}\right)^{p/(p-1)}\delta^{-1/(p-1)}$$

controls the U-SGD correction, while

$$N \geq K_p \left(\frac{DG_p}{\varepsilon} \right)^{p/(p-1)} \ell_\delta$$

controls the WG-Clip correction.

Choosing each horizon as the ceiling of a sufficiently large p-dependent constant times the corresponding sum gives (14), provided that the selected horizons also satisfy (12). The term 1 in N_0 accounts for integer rounding. This is the stated conditional interpretation of those sample requirements; without the remainder conditions, the full bounds (8) and (9) must be used.

Finally, solving (10) for N gives a sufficient budget of order

$$1 + \left(\frac{R_0 G_2 + D\sigma_2}{\varepsilon} \right)^2 \delta^{-1}.$$

Controlling the two terms in (11) separately gives a sufficient budget of order

$$1 + \left(\frac{R_0 G_2}{\varepsilon} \right)^2 + \left(\frac{DG_2}{\varepsilon} \right)^2 \ell_\delta.$$

These are exactly (15).

C Epoch-Averaged SGD on Strongly Convex Quadratics

This appendix gives a finite-sample analysis of an explicit, unprojected epoch scheme for strongly convex quadratic objectives. We use $\ell_\delta := \log(4/\delta)$ for $0 < \delta < 1/2$.

C.1 Quadratic model and oracle

We minimize an unconstrained quadratic on $\mathbb{R}^d$:

$$F(x) := F^\star + \frac{1}{2}(x - x^\star)^\top H (x - x^\star), \qquad H = H^\top, \qquad 0 < \mu I \preceq H \preceq LI, \qquad \kappa = \frac{L}{\mu},$$

where H is deterministic and $F^\star = F(x^\star)$. In particular, $\nabla F(x) = H(x - x^\star)$ and $x^\star$ is the unique minimizer.

The smoothness and strong-convexity constants satisfy, for all points x, y in the domain,

$$\|\nabla F(x) - \nabla F(y)\| \leq L \|x - y\|,$$
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2.$$

We write $\kappa = L/\mu$ and $\Delta_0 = F(x_1) - F(x^\star)$.

Within an epoch, let $\mathcal{F}_0$ contain all information available at its start. The starting point $x_0 = z$ is $\mathcal{F}_0$-measurable. The filtration $\mathcal{F}_t$ includes the oracle response at step t, and x_{t-1} is $\mathcal{F}_{t-1}$-measurable. For a fixed $p \geq 2$, use the notation of Assumption 2.1 and write

$$g_t = H(x_{t-1} - x^\star) + \zeta_t, \qquad \mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}] = 0,$$

where, almost surely at every step,

$$\mathbb{E}[\|\zeta_t\|^2 \mid \mathcal{F}_{t-1}] \leq \sigma_2^2, \qquad \mathbb{E}[\|\zeta_t\|^p \mid \mathcal{F}_{t-1}] \leq \sigma_p^p.$$

The noise may depend on the current iterate and need not be independent across steps. The conditional bounds are assumed to hold along the entire unprojected run. Write the supplied direction as

$$h_t = H(x_{t-1} - x^\star) + u_t. \tag{47}$$

Ordinary, unclipped SGD (U-SGD) uses $h_t = g_t$, hence $u_t = \zeta_t$.

The noise-centered oracle benchmark uses $u_t = \mathrm{clip}_\lambda(\zeta_t)$, leaving the exact mean gradient outside the clipping map. This benchmark generally requires information beyond one stochastic-gradient observation.

Assumption C.1 (Certified centered input)**.** At every step of an epoch, the perturbation in (47) has a decomposition

$$u_t = v_t + b_t, \qquad b_t \text{ is } \mathcal{F}_{t-1}\text{-measurable},$$

with deterministic epoch-wise bounds $\bar{\sigma}_2 \geq 0$, $\lambda > 0$, and $B_\lambda \geq 0$ such that, almost surely,

$$\mathbb{E}[v_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\|v_t\|^2 \mid \mathcal{F}_{t-1}] \leq \bar{\sigma}_2^2,$$
$$\|v_t\| \leq 2\lambda, \qquad \|b_t\| \leq B_\lambda.$$

A p-moment certificate with scale $M_p > 0$ additionally means

$$\bar{\sigma}_2 \leq \sigma_2, \qquad B_\lambda \leq c_{\mathrm{cert}} \frac{M_p^p}{\lambda^{p-1}}, \tag{48}$$

where c_{cert} is fixed.

The term *centered* refers to v_t, not to the full perturbation: b_t is retained in all bounds. Any anchor error or error from clipping a gradient difference must be included in b_t. We verify the noise-centered certificate in Lemma C.5 and the difference-clipping certificate in Proposition C.12.

C.2 Epoch algorithm and restarts

One epoch. Given a starting point z, the length $m \geq \lceil 4\kappa \rceil$, and a choice of directions h_t, set

$$x_0 = z, \qquad \eta = \frac{1}{2L}, \qquad b = \lfloor m/2 \rfloor, \qquad n = m - b.$$

Run the m unprojected updates

$$x_t = x_{t-1} - \eta h_t = x_{t-1} - \eta\big(H(x_{t-1} - x^\star) + u_t\big), \qquad t = 1, \ldots, m, \tag{49}$$

and return

$$z^+ := \frac{1}{n} \sum_{t=b+1}^{m} x_t.$$

The first b steps are a burn-in period. A clipping threshold is constant throughout this epoch.

Geometric restarts. Let z_0 satisfy $F(z_0) - F^\star \leq \Delta_0$ almost surely, where $\Delta_0 > 0$ is a known deterministic upper bound. Fix $0 < \varepsilon < \Delta_0$ and $0 < \delta < 1/2$, and define

$$S := \left\lceil \log_2 \frac{\Delta_0}{\varepsilon} \right\rceil, \qquad \Delta_s = 2^{-s}\Delta_0, \qquad s = 0, \ldots, S.$$

Allocate failure probabilities by

$$\delta_s = \frac{6\delta}{\pi^2(s+1)^2}, \quad s = 0, \ldots, S-1, \qquad \sum_{s=0}^{S-1} \delta_s \leq \delta. \tag{50}$$

For later use, set

$$\ell_S := \log\left(\frac{2\pi^2 S^2}{3\delta}\right) = \max_{0 \leq s < S} \ell_{\delta_s}.$$

For each $s = 0, \ldots, S-1$, carry out the following three steps:

1. Choose m_s with $(\Delta, \delta) = (\Delta_s, \delta_s)$ and the appropriate input type.
2. Start a new epoch at $x_{s,0} = z_s$ and use (49) with $b_s = \lfloor m_s/2 \rfloor$ and $n_s = m_s - b_s$. For a certified p-moment input, use $\lambda_s = M_{p,s}(m_s/\ell_{\delta_s})^{1/p}$.
3. Set $z_{s+1} = n_s^{-1} \sum_{t=b_s+1}^{m_s} x_{s,t}$.

Return $\widehat{x} := z_S$ after $N_{\mathrm{core}} := \sum_{s=0}^{S-1} m_s$ core updates. If $\varepsilon \geq \Delta_0$, return z_0 without an update.

Lemma C.2 (Filter identity and norm budgets)**.** *For the epoch above, let $e_t := x_t - x^\star$, $A := I - \eta H$, $q := 1 - 1/(2\kappa)$, and $\bar{e} := n^{-1} \sum_{t=b+1}^{m} e_t$. For every realized input sequence $(u_t)_{t=1}^m$,*

$$H^{1/2}\bar{e} = d_m - \sum_{i=1}^{m} W_{i,m} u_i, \tag{51}$$

where $W_{i,m}$ are deterministic and d_m is $\mathcal{F}_0$-measurable, with

$$d_m := \frac{1}{n} \sum_{t=b+1}^{m} H^{1/2} A^t e_0, \quad W_{i,m} := \frac{\eta}{n} H^{1/2} \sum_{t=\max\{b+1,i\}}^{m} A^{t-i}.$$

For every $r \geq 1$,

$$\sum_{i=1}^{m} \|W_{i,m}\|_{\mathrm{op}}^r \leq \frac{2}{\mu^{r/2} n^{r-1}}, \qquad \max_i \|W_{i,m}\|_{\mathrm{op}} \leq \frac{1}{\sqrt{\mu}\, n}.$$

In particular,

$$\begin{aligned} \sum_i \|W_{i,m}\|_{\mathrm{op}}^2 &\leq \frac{4}{\mu m}, & \sum_i \|W_{i,m}\|_{\mathrm{op}}^p &\leq \frac{2^p}{\mu^{p/2} m^{p-1}}, \\ \max_i \|W_{i,m}\|_{\mathrm{op}} &\leq \frac{2}{\sqrt{\mu}\, m}, & \sum_i \|W_{i,m}\|_{\mathrm{op}} &\leq \frac{2}{\sqrt{\mu}}. \end{aligned} \tag{52}$$

Consequently, on $\{F(z) - F^\star \leq \Delta\}$,

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \left\| \sum_{i=1}^{m} W_{i,m} u_i \right\|^2. \tag{53}$$

Proof. The recursion is $e_t = A e_{t-1} - \eta u_t$, and induction gives $e_t = A^t e_0 - \eta \sum_{i=1}^{t} A^{t-i} u_i$. Average this equality over $t = b+1, \ldots, m$ and interchange the finite sums to obtain (51). Since A is a polynomial in H, these matrices commute, and

$$\eta \sum_{j=0}^{r-1} A^j = H^{-1}(I - A^r), \qquad r \in \mathbb{N}.$$

Hence

$$W_{i,m} = \begin{cases} n^{-1} H^{-1/2} A^{b+1-i}(I - A^n), & 1 \leq i \leq b, \\ n^{-1} H^{-1/2}(I - A^{m-i+1}), & b+1 \leq i \leq m. \end{cases}$$

All eigenvalues of A lie in $[1/2, q]$. Thus $\|I - A^r\|_{\mathrm{op}} \leq 1$ and

$$\|W_{i,m}\|_{\mathrm{op}} \leq \frac{1}{\sqrt{\mu}\, n} \begin{cases} q^{b+1-i}, & i \leq b, \\ 1, & i > b. \end{cases}$$

Because $n \geq m/2 \geq 2\kappa$ and, for $r \geq 1$, $(1 - q^r)^{-1} \leq (1-q)^{-1} = 2\kappa$, we obtain

$$\sum_i \|W_{i,m}\|_{\mathrm{op}}^r \leq \frac{n + \sum_{j=1}^{b} q^{rj}}{\mu^{r/2} n^r} \leq \frac{n + 2\kappa}{\mu^{r/2} n^r} \leq \frac{2}{\mu^{r/2} n^{r-1}}.$$

Commutation also yields $\|d_m\| \leq q^{b+1} \|H^{1/2} e_0\| \leq q^{b+1} \sqrt{2\Delta}$. Using $2(b+1) \geq m$ and $q \leq e^{-1/(2\kappa)}$, we have $\|d_m\|^2 \leq 2\Delta e^{-m/(2\kappa)}$. Finally, $F(z^+) - F^\star = \frac{1}{2}\|H^{1/2}\bar{e}\|^2$ and $\frac{1}{2}\|a-b\|^2 \leq \|a\|^2 + \|b\|^2$ give (53). □

C.3 Concentration and clipping

If (d_t) is a martingale-difference sequence and the deterministic budgets satisfy $\sum_t \mathbb{E}[\|d_t\|^2 \mid \mathcal{F}_{t-1}] \le V$ and $\sum_t \mathbb{E}[\|d_t\|^p \mid \mathcal{F}_{t-1}] \le A_p$ almost surely, then, for $p > 2$, the Fuk–Nagaev inequality gives

$$\left\| \sum_t d_t \right\| \le C\sqrt{V\ell_\delta} + C_p(A_p/\delta)^{1/p} \tag{54}$$

with probability at least $1-\delta$; see (Mollenhauer & Fiedler, 2025, Theorem 2.2). If instead $\|d_t\| \le R$ almost surely, the Hilbert-space Bennett–Freedman inequality gives

$$\left\| \sum_t d_t \right\| \le C\sqrt{V\ell_\delta} + CR\ell_\delta \tag{55}$$

with the same probability; see (Pinelis, 1994, Theorem 3.4). These inequalities also apply under the conditional law given $\mathcal{F}_0$ when the same deterministic budgets hold.

Lemma C.3 (Operator-weighted bounds with deterministic budgets)**.** *Let W_t be deterministic linear maps between Euclidean spaces, and put $\mathcal{W}_r := \sum_t \|W_t\|_{\mathrm{op}}^r$ and $\mathcal{W}_\infty := \max_t \|W_t\|_{\mathrm{op}}$. Under Assumption 2.1, for $p > 2$,*

$$\left\| \sum_t W_t\zeta_t \right\| \le C\sigma_2\sqrt{\mathcal{W}_2\ell_\delta} + C_p\sigma_p\mathcal{W}_p^{1/p}\delta^{-1/p}$$

with conditional probability at least $1-\delta$ given $\mathcal{F}_0$. For the centered component in Assumption C.1,

$$\left\| \sum_t W_t v_t \right\| \le C\bar{\sigma}_2\sqrt{\mathcal{W}_2\ell_\delta} + C\lambda\mathcal{W}_\infty\ell_\delta \tag{56}$$

with the same conditional probability. At $p = 2$, the unclipped sum satisfies

$$\mathbb{E}\left[\left\| \sum_t W_t\zeta_t \right\|^2 \,\middle|\, \mathcal{F}_0 \right] \le \sigma_2^2\mathcal{W}_2. \tag{57}$$

Proof. For $d_t = W_t\zeta_t$, the conditional second- and p-moment budgets are bounded by $\sigma_2^2\mathcal{W}_2$ and $\sigma_p^p\mathcal{W}_p$. Apply (54). For $d_t = W_t v_t$, the second-moment budget is $\bar{\sigma}_2^2\mathcal{W}_2$ and $\|d_t\| \le 2\lambda\mathcal{W}_\infty$; (55) gives the second claim. For the last claim, distinct martingale increments are orthogonal in conditional L^2. Expanding the squared norm therefore leaves only $\sum_t \mathbb{E}[\|W_t\zeta_t\|^2 \mid \mathcal{F}_0]$. □

Remark C.4. All $\mathcal{W}_r$ are deterministic. More generally, predictable random maps can be used if deterministic almost-sure upper bounds for their norm sums and maximum replace $\mathcal{W}_r$ and $\mathcal{W}_\infty$.

Lemma C.5 (Bias, variance, and the noise-centered certificate)**.** *For a random vector Y, a sigma-field $\mathcal{G}$, and $a := \mathbb{E}[Y \mid \mathcal{G}]$, suppose $\mathbb{E}[\|Y\|^p \mid \mathcal{G}] \le D^p$. Then*

$$\|\mathbb{E}[\mathrm{clip}_\lambda(Y) \mid \mathcal{G}] - a\| \le D^p/\lambda^{p-1}, \tag{58}$$

$$\mathbb{E}\Big[\|\mathrm{clip}_\lambda(Y) - \mathbb{E}[\mathrm{clip}_\lambda(Y) \mid \mathcal{G}]\|^2 \,\Big|\, \mathcal{G} \Big] \le \mathbb{E}[\|Y-a\|^2 \mid \mathcal{G}]. \tag{59}$$

Consequently, noise-centered clipping is certified with $\bar{\sigma}_2 = \sigma_2$, $M_p = \sigma_p$, and $c_{\mathrm{cert}} = 1$.

Proof. Pointwise,

$$\|\,\mathrm{clip}_\lambda(y) - y\| = (\|y\| - \lambda)_+ \le \|y\|\mathbf{1}_{\{\|y\|>\lambda\}} \le \|y\|^p/\lambda^{p-1}.$$

Conditional expectation and Jensen's inequality give (58). The clipping map is Euclidean projection onto a closed ball, hence is nonexpansive. Conditional expectation minimizes squared error, so the left-hand side of (59) is at most

$$\mathbb{E}[\|\,\mathrm{clip}_\lambda(Y) - \mathrm{clip}_\lambda(a)\|^2 \mid \mathcal{G}] \le \mathbb{E}[\|Y-a\|^2 \mid \mathcal{G}].$$

Apply these facts with $Y = \zeta_t$ and $\mathcal{G} = \mathcal{F}_{t-1}$. Take $b_t = \mathbb{E}[\mathrm{clip}_\lambda(\zeta_t) \mid \mathcal{F}_{t-1}]$ and $v_t = \mathrm{clip}_\lambda(\zeta_t) - b_t$. Then $\|v_t\| \le 2\lambda$ and all the stated certificate bounds follow. □

Lemma C.6 (One epoch for a quadratic objective)**.** *Let $m \geq \lceil 4\kappa \rceil$, and suppose $F(z) - F^\star \leq \Delta$. Each bound below holds with conditional probability at least $1-\delta$ given the information at the epoch start. For U-SGD satisfying Assumption 2.1 with $p > 2$,*

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \frac{C\sigma_2^2 \ell_\delta}{\mu m} + \frac{C_p \sigma_p^2}{\mu} \delta^{-2/p} m^{-2(1-1/p)}. \tag{60}$$

For any certified centered input,

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \frac{C\bar{\sigma}_2^2 \ell_\delta}{\mu m} + \frac{C}{\mu}\left(\frac{\lambda \ell_\delta}{m} + B_\lambda\right)^2. \tag{61}$$

If this input has a p-moment certificate, $p \geq 2$, choose

$$\lambda = M_p \left(\frac{m}{\ell_\delta}\right)^{1/p}. \tag{62}$$

Then

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \frac{C\sigma_2^2 \ell_\delta}{\mu m} + \frac{C_p M_p^2}{\mu}\left(\frac{\ell_\delta}{m}\right)^{2(1-1/p)}. \tag{63}$$

For a random starting point, the claims apply on the $\mathcal{F}_0$-measurable event $\{F(z) - F^\star \leq \Delta\}$.

Proof. For U-SGD, Lemma C.3 and (52) imply

$$\left\| \sum_{t=1}^m W_{t,m} \zeta_t \right\| \leq \frac{C\sigma_2}{\sqrt{\mu}} \sqrt{\frac{\ell_\delta}{m}} + \frac{C_p \sigma_p}{\sqrt{\mu}} \delta^{-1/p} m^{-(1-1/p)}.$$

Square this bound using $(a+b)^2 \leq 2a^2 + 2b^2$ and insert it into (53). For a certified input, apply (56) to the centered part and bound the bias pathwise:

$$\left\| \sum_t W_{t,m} b_t \right\| \leq B_\lambda \sum_t \|W_{t,m}\|_{\mathrm{op}} \leq \frac{2B_\lambda}{\sqrt{\mu}}.$$

With conditional probability at least $1-\delta$, their sum is bounded by

$$\left\| \sum_t W_{t,m}(v_t + b_t) \right\| \leq \frac{C\bar{\sigma}_2}{\sqrt{\mu}} \sqrt{\frac{\ell_\delta}{m}} + \frac{C\lambda \ell_\delta}{\sqrt{\mu}\, m} + \frac{2B_\lambda}{\sqrt{\mu}}.$$

Squaring proves (61). For (62), both $\lambda \ell_\delta / m$ and M_p^p / λ^{p-1} equal $M_p (\ell_\delta / m)^{1-1/p}$. Use (48) and $\bar{\sigma}_2 \leq \sigma_2$ to obtain (63). □

Proposition C.7 (The finite-variance endpoint)**.** *Under the same quadratic setup with $p = 2$, U-SGD satisfies*

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \frac{C\sigma_2^2}{\mu m \delta} \tag{64}$$

with conditional probability at least $1-\delta$. A certified second-moment input with

$$\lambda = M_2 \sqrt{\frac{m}{\ell_\delta}} \tag{65}$$

satisfies, with the same conditional probability,

$$F(z^+) - F^\star \leq 2\Delta e^{-m/(2\kappa)} + \frac{C(\sigma_2^2 + M_2^2)\ell_\delta}{\mu m}. \tag{66}$$

Proof. Equations (57) and (52) give $\mathbb{E}[\| \sum_t W_{t,m} \zeta_t \|^2 \mid \mathcal{F}_0] \leq 4\sigma_2^2/(\mu m)$. Conditional Markov's inequality bounds this squared norm by $4\sigma_2^2/(\mu m \delta)$ outside an event of conditional probability at most δ. Apply (53). For the certified input, use the $p = 2$ case of (63), whose proof uses bounded-increment concentration and does not require $p > 2$. □

C.4 Epoch lengths and restart complexity

We now specify the epoch lengths and total update count.

Proposition C.8 (A deterministic epoch schedule that halves the gap)**.** *Let $0 < \delta < 1/2$ and $\Delta > 0$. For $p > 2$, set $\alpha = p/[2(p-1)]$. There are sufficiently large constants K_0 and K_p, independent of Δ, δ, such that the following choices ensure $F(z^+) - F^\star \le \Delta/2$ with conditional probability at least $1-\delta$ whenever $F(z) - F^\star \le \Delta$:*

$$m_{\mathrm{U}}(\Delta,\delta) = \left\lceil K_0\kappa + K_0\frac{\sigma_2^2\ell_\delta}{\mu\Delta} + K_p\left(\frac{\sigma_p^2}{\mu\Delta}\right)^{\alpha}\delta^{-1/(p-1)}\right\rceil, \tag{67}$$

$$m_{\mathrm{C}}(\Delta,\delta;M_p) = \left\lceil K_0\kappa + K_0\frac{\sigma_2^2\ell_\delta}{\mu\Delta} + K_p\left(\frac{M_p^2}{\mu\Delta}\right)^{\alpha}\ell_\delta\right\rceil. \tag{68}$$

The clipped choice uses (62)*; its constants may depend on c_{cert}. At $p = 2$, suitable choices are*

$$m_{\mathrm{U}}^{(2)}(\Delta,\delta) = \left\lceil K_0\kappa + K_0\frac{\sigma_2^2}{\mu\Delta\delta}\right\rceil, \tag{69}$$

$$m_{\mathrm{C}}^{(2)}(\Delta,\delta;M_2) = \left\lceil K_0\kappa + K_0\frac{(\sigma_2^2+M_2^2)\ell_\delta}{\mu\Delta}\right\rceil, \tag{70}$$

with (65) *in the second case.*

Proof. Choose K_0 so that $m \ge K_0\kappa$ implies both $m \ge \lceil 4\kappa\rceil$ and $2\Delta e^{-m/(2\kappa)} \le \Delta/6$. The Gaussian term in (60) is at most $\Delta/6$ when $m \ge C\sigma_2^2\ell_\delta/(\mu\Delta)$. Solving the remaining inequality gives

$$\frac{C_p\sigma_p^2}{\mu}\delta^{-2/p}m^{-2(p-1)/p} \le \frac{\Delta}{6} \quad \text{if} \quad m \ge C_p\left(\frac{\sigma_p^2}{\mu\Delta}\right)^{p/[2(p-1)]}\delta^{-1/(p-1)}.$$

These three bounds prove the U-SGD claim. Solving the last term of (63) instead gives $m \ge C_p(M_p^2/(\mu\Delta))^\alpha\ell_\delta$. At $p = 2$, apply the same reasoning to (64) and (66). The ceiling and the sum of the required lengths preserve each inequality. □

Theorem C.9 (Restarted quadratic scheme, with explicit restart factors)**.** *Run the algorithm of Section C.2, using the epoch lengths of Proposition C.8. Suppose the conditional noise bounds, or the required certified-input bounds, hold uniformly across the run. For $p > 2$, the output of either corresponding scheme satisfies*

$$\mathbb{P}\big(F(\widehat{x}) - F^\star \le \varepsilon\big) \ge 1-\delta.$$

With $\alpha := p/[2(p-1)]$, the U-SGD update count obeys

$$N_{\mathrm{U}} \le C\kappa S + \frac{C\sigma_2^2}{\mu\varepsilon}\ell_S + C_p\left(\frac{\sigma_p^2}{\mu\varepsilon}\right)^{\alpha}\delta^{-1/(p-1)}S^{2/(p-1)}. \tag{71}$$

For certified inputs with $M_{p,s} \le \overline{M}_p$ and uniform c_{cert}, the core-update count obeys

$$N_{\mathrm{core,C}} \le C\kappa S + \frac{C\sigma_2^2}{\mu\varepsilon}\ell_S + C_p\left(\frac{\overline{M}_p^2}{\mu\varepsilon}\right)^{\alpha}\ell_S. \tag{72}$$

Noise-centered clipping has $\overline{M}_p = \sigma_p$. For difference clipping, (72) *counts only the core updates; the total oracle count is* (75)*.*

Proof. Let G_s be the event that $F(z_j) - F^\star \le \Delta_j$ for every $j \le s$. It is measurable at the start of epoch s. The conditional halving guarantee gives

$$\mathbb{P}\big(G_s \cap \{F(z_{s+1}) - F^\star > \Delta_{s+1}\}\big) \le \delta_s\,\mathbb{P}(G_s) \le \delta_s.$$

Since G_0 holds almost surely, a union bound over the first failed epoch proves $\mathbb{P}(G_S) \ge 1 - \sum_s \delta_s \ge 1-\delta$. On G_S, the final error is at most $\Delta_S \le \varepsilon$. No independence between epochs is required.

For every $r > 0$, the geometric targets satisfy

$$\sum_{s=0}^{S-1} \Delta_s^{-r} \le \frac{\varepsilon^{-r}}{1-2^{-r}}. \tag{73}$$

Indeed, $\Delta_{S-1} > \varepsilon$ and the earlier inverse powers decrease by the factor 2^{-r}. Also, $\ell_{\delta_s} \le \ell_S$ and

$$\delta_s^{-1/(p-1)} \le \left(\frac{\pi^2}{6\delta}\right)^{1/(p-1)} S^{2/(p-1)}.$$

Sum (67), applying (73) with $r = 1$ and $r = \alpha$. The ceilings add at most S, which is absorbed by $C\kappa S$ because $\kappa \ge 1$. This proves (71). For (72), sum (68), use $M_{p,s} \le \overline{M}_p$, and apply the same geometric sums. □

Corollary C.10 (Restart complexity under finite variance)**.** *At $p = 2$, the same success guarantee holds with the schedules* (69)–(70)*, and*

$$N_{\mathrm{U}}^{(2)} \le C\kappa S + \frac{C\sigma_2^2 S^2}{\mu\varepsilon\delta},$$
$$N_{\mathrm{core,C}}^{(2)} \le C\kappa S + \frac{C(\sigma_2^2 + \overline{M}_2^2)\ell_S}{\mu\varepsilon}.$$

Proof. The conditional success argument in the theorem applies without change. Sum the two $p = 2$ epoch lengths, using $\delta_s^{-1} \le \pi^2 S^2/(6\delta)$, $\ell_{\delta_s} \le \ell_S$, and $\sum_s \Delta_s^{-1} \le 2/\varepsilon$. □

Remark C.11 (Interpretation of the comparison)**.** The factor $S^{2/(p-1)}$ in (71), and its $p = 2$ counterpart S^2, come from the chosen allocation (50). They are polynomial in the number of epochs S and therefore polylogarithmic in Δ_0/ε. For fixed $p > 2$ and fixed confidence, $\alpha < 1$, so the displayed heavy-tail terms have a lower order in $1/\varepsilon$ than the variance term, even with these restart factors. At $p = 2$, both statistical terms have order $1/\varepsilon$, and the confidence dependence distinguishes the displayed bounds. These are upper bounds for the specified schemes, not a minimax comparison with all robust estimators or aggregations of multiple runs.

C.5 Difference clipping and its oracle cost

We now consider the difference-clipped direction introduced in Section 2. We allow the anchor and its gradient estimate to depend on time, writing them as $\widetilde{x}_t$ and $\widetilde{s}_t$. The following proposition gives sufficient conditions on their accuracy and on the coupled gradient differences.

Proposition C.12 (Sufficient certificate for difference clipping)**.** *Let $\widetilde{x}_t$ and $\widetilde{s}_t$ be $\mathcal{F}_{t-1}$-measurable. Suppose a coupled-sample oracle returns*

$$d_t := g(x_{t-1}, \xi_t) - g(\widetilde{x}_t, \xi_t),$$
$$a_t := \mathbb{E}[d_t \mid \mathcal{F}_{t-1}] = H(x_{t-1} - \widetilde{x}_t).$$

Use $h_t := \widetilde{s}_t + \mathrm{clip}_\lambda(d_t)$. Assume deterministic epoch-wise bounds a, S_2, D_p satisfy

$$\|\widetilde{s}_t - H(\widetilde{x}_t - x^\star)\| \le a, \quad \mathbb{E}[\|d_t - a_t\|^2 \mid \mathcal{F}_{t-1}] \le S_2^2, \quad \mathbb{E}[\|d_t\|^p \mid \mathcal{F}_{t-1}] \le D_p^p$$

almost surely at every step. Then the input is certified with

$$\bar{\sigma}_2 = S_2, \qquad B_\lambda = a + \frac{D_p^p}{\lambda^{p-1}}. \tag{74}$$

In particular, a p-moment certificate with scale M_p follows if $S_2 \le \sigma_2$, $D_p \le c_D M_p$, and $a \le c_a M_p^p/\lambda^{p-1}$, with $c_{\mathrm{cert}} = c_a + c_D^p$ independent of the epoch and its parameters.

Proof. Define

$$v_t = \mathrm{clip}_\lambda(d_t) - \mathbb{E}[\mathrm{clip}_\lambda(d_t) \mid \mathcal{F}_{t-1}],$$
$$b_t = \widetilde{s}_t - H(\widetilde{x}_t - x^\star) + \mathbb{E}[\mathrm{clip}_\lambda(d_t) \mid \mathcal{F}_{t-1}] - a_t.$$

Then $h_t - H(x_{t-1} - x^\star) = v_t + b_t$, $\mathbb{E}[v_t \mid \mathcal{F}_{t-1}] = 0$, and $\|v_t\| \le 2\lambda$. Apply Lemma C.5 to $Y = d_t$: the conditional centered variance is at most S_2^2, and the clipping bias is at most D_p^p/λ^{p-1}. Adding the anchor error gives (74) and then the claimed p-moment certificate. □

Remark C.13 (Core updates and total oracle cost)**.** The difference and anchor bounds above must be verified for the actual implementation, including each anchor refresh. Its total stochastic-gradient evaluation count is

$$N_{\mathrm{total}}^{\mathrm{DC}} = N_{\mathrm{core,C}} + N_{\mathrm{anchor}} + N_{\mathrm{refresh}} + N_{\mathrm{pair}}. \tag{75}$$

Here N_{anchor} counts initial anchor construction, N_{refresh} counts subsequent anchor work, and N_{pair} counts paired evaluations beyond the one charged to each core update. If a pair costs two evaluations, $N_{\mathrm{pair}} = N_{\mathrm{core,C}}$; it is zero only in an oracle model returning a coupled pair at unit cost. The bounds above control $N_{\mathrm{core,C}}$ after certification.

D Polyak–Ruppert Asymptotics: Assumptions and Proofs

This appendix provides the complete statements and proofs underlying Section 4. Throughout, $\|\cdot\|$ denotes the Euclidean norm or its induced operator norm, as appropriate. We write $\|X\|_{L^2} := (\mathbb{E}\|X\|^2)^{1/2}$ and $\ell_\delta := \log(4/\delta)$.

D.1 The exact linear identity and the unclipped limit

Let $e_k := x_k - x^\star$ and consider

$$e_{k+1} := e_k - \eta_k(He_k + \zeta_{k+1}), \qquad H = H^\top \succ 0,$$

where e_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\zeta_{k+1} \mid \mathcal{F}_k] = 0$. The stepsizes are deterministic and positive. Define

$$\bar{e}_N := \frac{1}{N}\sum_{k=1}^{N} e_k, \qquad \bar{x}_N := \frac{1}{N}\sum_{k=1}^{N} x_k.$$

Proposition D.1 (Exact Polyak–Ruppert decomposition)**.** *For every positive stepsize sequence,*

$$H\bar{e}_N = -\frac{1}{N}\sum_{k=1}^{N}\zeta_{k+1} + R_N, \tag{76}$$

where

$$R_N := \frac{1}{N}\left[\frac{e_1}{\eta_1} - \frac{e_{N+1}}{\eta_N} + \sum_{k=2}^{N} e_k\left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\right)\right]. \tag{77}$$

The identity is pathwise and does not require symmetry or stability of the drift matrix.

Suppose, additionally, that $\eta_k := a(k+k_0)^{-\gamma}$ with $a > 0$, $k_0 \geq 0$, $\gamma \in (1/2, 1)$, and $\|e_k\|_{L^2} \leq C_e k^{-\gamma/2}$. Then

$$\|R_N\|_{L^2} \leq CN^{\gamma/2-1} = o(N^{-1/2}).$$

Proof. Rearranging the update gives

$$He_k = \frac{e_k - e_{k+1}}{\eta_k} - \zeta_{k+1}.$$

Summation by parts yields

$$\sum_{k=1}^{N}\frac{e_k - e_{k+1}}{\eta_k} = \frac{e_1}{\eta_1} - \frac{e_{N+1}}{\eta_N} + \sum_{k=2}^{N} e_k\left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\right),$$

which proves (76)–(77).

For the stated stepsizes,

$$0 \leq \frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \leq Ck^{\gamma-1}.$$

Minkowski's inequality therefore gives

$$\|R_N\|_{L^2} \leq \frac{C}{N}\left[1 + N^{\gamma/2} + \sum_{k=2}^{N} k^{\gamma/2-1}\right] \leq CN^{\gamma/2-1}.$$

Since $\gamma < 1$, this is $o(N^{-1/2})$. □

To obtain a Gaussian limit, suppose that the noise is square integrable and satisfies

$$Q_N := \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}[\zeta_{k+1}\zeta_{k+1}^\top \mid \mathcal{F}_k] \xrightarrow{\mathbb{P}} \Sigma, \tag{78}$$

where Σ is deterministic, and, for every $\epsilon > 0$,

$$L_N(\epsilon) := \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\|\zeta_{k+1}\|^2\mathbf{1}_{\{\|\zeta_{k+1}\|>\epsilon\sqrt{N}\}}\,\middle|\,\mathcal{F}_k\right] \xrightarrow{\mathbb{P}} 0. \tag{79}$$

Under the stepsize and moment-decay assumptions of Proposition D.1, these conditions imply

$$\sqrt{N}\,\bar{e}_N \Longrightarrow \mathcal{N}(0, H^{-1}\Sigma H^{-1}). \tag{80}$$

Indeed, the martingale central limit theorem Hall & Heyde (1980) applies to $N^{-1/2}\sum_{k=1}^{N}\zeta_{k+1}$, while $\sqrt{N}R_N \to 0$ in L^2. Equation (76) and Slutsky's theorem then give (80), recovering the classical Polyak–Ruppert limit Ruppert (1988); Polyak & Juditsky (1992).

For i.i.d. centered noise independent of the initial condition, finite covariance implies both noise conditions above. Let an $\mathcal{F}_0$-measurable variable U equal 1 or 2 with equal probability, and let $\zeta_k = U\varepsilon_k$, where the ε_k are independent Rademacher variables independent of U. Then $Q_N = U^2$, and the normalized noise sum converges to

$$\tfrac{1}{2}\mathcal{N}(0,1) + \tfrac{1}{2}\mathcal{N}(0,4).$$

If, in addition, for some $p > 2$,

$$\mathbb{E}[\|\zeta_{k+1}\|^r \mid \mathcal{F}_k] \le \sigma_r^r, \qquad r \in \{2, p\},$$

uniformly in k, Theorem A.1 and (76) give, with probability at least $1-\delta$,

$$\|\bar{e}_N\| \le \|H^{-1}\|\left[C\sigma_2\sqrt{\frac{\ell_\delta}{N}} + C_p\sigma_p\delta^{-1/p}N^{-(1-1/p)}\right] + \|H^{-1}R_N\|.$$

Thus the Gaussian limit and the finite-sample heavy-tail correction follow from the same exact representation.

D.2 Fixed-threshold clipped stochastic approximation

Fix $0 < \lambda < \infty$ and define

$$\psi_\lambda(x,\xi) := \mathrm{clip}_\lambda(g(x,\xi)), \qquad h_\lambda(x) := \mathbb{E}\psi_\lambda(x,\xi).$$

Consider

$$x_{k+1} = x_k - \eta_k\psi_\lambda(x_k,\xi_{k+1}), \qquad \bar{x}_N = \frac{1}{N}\sum_{k=1}^{N}x_k,$$

where the samples are i.i.d. and independent of the initial condition. Its equilibrium is determined by h_λ.

Assumption D.2 (Fixed-threshold PR regularity)**.** The following conditions hold.

1. The mean field h_λ has a locally unique root x_λ. The matrix $A_\lambda := \nabla h_\lambda(x_\lambda)$ has eigenvalues with positive real parts. For some $\alpha \in (0,1]$,
$$\|h_\lambda(x_\lambda + y) - A_\lambda y\| \le C\|y\|^{1+\alpha} \tag{81}$$
for all sufficiently small y.
2. Define $M_\lambda(x,\xi) := \psi_\lambda(x,\xi) - h_\lambda(x)$. For some $\beta \in (0,1]$,
$$\mathbb{E}\|M_\lambda(x_\lambda + y,\xi) - M_\lambda(x_\lambda,\xi)\|^2 \le C\|y\|^{2\beta} \tag{82}$$
for all sufficiently small y.

3. For $\eta_k = a(k + k_0)^{-\gamma}$ with $a > 0$, $k_0 \geq 0$, and $\gamma \in (1/2, 1)$, the recursion converges almost surely to x_λ. Writing $y_k := x_k - x_\lambda$, assume

$$\mathbb{E}\|y_k\|^2 \leq C\eta_k, \qquad \mathbb{E}\|y_k\|^{2(1+\alpha)} \leq C\eta_k^{1+\alpha}, \qquad \gamma(1+\alpha) > 1. \tag{83}$$

Remark D.3. The three conditions separate local stability of the clipped mean field, continuity of the noise, and control of the actual iterates. They provide explicit sufficient conditions for replacing the recursion by its linearization around x_λ.

The first condition makes the linearized dynamics $\dot{y} = -A_\lambda y$ asymptotically stable and controls the error of this linear approximation. The remainder estimate holds, for example, when h_λ has a locally α-Hölder continuous Jacobian. A locally Lipschitz Jacobian gives $\alpha = 1$. The required smoothness concerns the expected clipped score h_λ: individual clipped scores may be nonsmooth at the clipping boundary.

The second condition allows the state-dependent noise to be replaced, to first order, by the noise evaluated at x_λ. A simple sufficient condition for $\beta = 1$ is local Lipschitz continuity of the stochastic gradient in mean square:

$$\mathbb{E}\|g(x_\lambda + y, \xi) - g(x_\lambda, \xi)\|^2 \leq L_g^2\|y\|^2.$$

Indeed, clipping is nonexpansive, and subtracting the expectation cannot increase the second moment of the score difference. This condition holds directly for additive linear models. Under the iterate moment bound in the third condition, every $\beta > 0$ makes the accumulated noise-replacement error negligible on the root-N scale.

The third condition controls convergence to the selected root and the size of the fluctuations around it. These properties require a separate stability argument for the recursion. The restriction $\gamma(1+\alpha) > 1$ ensures that the accumulated nonlinear drift is negligible, since

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{N} \mathbb{E}\|h_\lambda(x_\lambda + y_k) - A_\lambda y_k\| \lesssim \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \eta_k^{(1+\alpha)/2} \longrightarrow 0.$$

For $\alpha = 1$, this restriction reduces to $\gamma > 1/2$; weaker regularity requires $\gamma > 1/(1+\alpha)$. The upper restriction $\gamma < 1$ makes the summation-by-parts remainder negligible on the same scale.

The higher-moment bound in (83) is stronger than needed for the displayed CLT proof. Because $\alpha \leq 1$,

$$\mathbb{E}\|y_k\|^{1+\alpha} \leq \big(\mathbb{E}\|y_k\|^2\big)^{(1+\alpha)/2} \lesssim \eta_k^{(1+\alpha)/2}.$$

Thus the second-moment bound already supplies the required nonlinear-drift estimate. Finally, fixed clipping makes the score at the root bounded, so its Lindeberg condition is automatic.

Theorem D.4 (Fixed-threshold clipped PR limit)**.** *Under Assumption D.2, let*

$$\Sigma_\lambda := \mathrm{Cov}\big(\psi_\lambda(x_\lambda, \xi)\big).$$

Then

$$\sqrt{N}(\bar{x}_N - x_\lambda) \Longrightarrow \mathcal{N}\left(0, A_\lambda^{-1}\Sigma_\lambda A_\lambda^{-\top}\right). \tag{84}$$

If $x_\lambda \neq x^\star$, the sequence $\sqrt{N}(\bar{x}_N - x^\star)$ is not tight. In particular, $h_\lambda(x^\star) \neq 0$ excludes a root-N limit centered at the optimizer.

Proof. Define

$$\varepsilon_{k+1} := M_\lambda(x_\lambda, \xi_{k+1}), \qquad \Delta M_{k+1} := M_\lambda(x_\lambda + y_k, \xi_{k+1}) - M_\lambda(x_\lambda, \xi_{k+1}),$$

and $r_h(y) := h_\lambda(x_\lambda + y) - A_\lambda y$. Then

$$y_{k+1} = y_k - \eta_k\big(A_\lambda y_k + \varepsilon_{k+1} + \Delta M_{k+1} + r_h(y_k)\big).$$

Summation by parts gives

$$A_\lambda \bar{y}_N = -\frac{1}{N}\sum_{k=1}^{N} \varepsilon_{k+1} - \frac{1}{N}\sum_{k=1}^{N} \Delta M_{k+1} - \frac{1}{N}\sum_{k=1}^{N} r_h(y_k) + R_N,$$

where R_N is (77) with e_k replaced by y_k. The moment bound in (83) and the proof of Proposition D.1 imply $\sqrt{N}R_N \to 0$ in L^2.

The local estimates (81)–(82) extend to all y after increasing their constants. Indeed, $\|\psi_\lambda\| \le \lambda$ and $\|M_\lambda\| \le 2\lambda$. Outside any fixed ball around zero,

$$\|r_h(y)\| \le \lambda + \|A_\lambda\|\,\|y\| \le C\|y\|^{1+\alpha},$$

and the squared noise difference is bounded by $16\lambda^2 \le C\|y\|^{2\beta}$.

Consequently,

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{k=1}^{N} r_h(y_k)\right\| \le \frac{C}{\sqrt{N}}\sum_{k=1}^{N}\eta_k^{(1+\alpha)/2} \longrightarrow 0,$$

because $\gamma(1+\alpha) > 1$.

The sequence ΔM_{k+1} is a martingale difference. Orthogonality of its increments and Jensen's inequality give

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\Delta M_{k+1}\right\|^2 \le \frac{C}{N}\sum_{k=1}^{N}\mathbb{E}\|y_k\|^{2\beta} \le \frac{C}{N}\sum_{k=1}^{N}\eta_k^{\beta} \longrightarrow 0.$$

It follows that

$$\sqrt{N}(\bar{x}_N - x_\lambda) = -A_\lambda^{-1}\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\varepsilon_{k+1} + o_{\mathbb{P}}(1).$$

The root scores ε_{k+1} are i.i.d., centered, bounded, and have covariance Σ_λ. The ordinary multivariate CLT and Slutsky's theorem prove (84).

Finally,

$$\sqrt{N}(\bar{x}_N - x^\star) = \sqrt{N}(x_\lambda - x^\star) + \sqrt{N}(\bar{x}_N - x_\lambda).$$

If $x_\lambda \neq x^\star$, the first term diverges in norm and the second is tight. □

The next subsection gives an exact quadratic example with $x_\lambda \neq x^\star$.

D.3 A QUADRATIC COUNTEREXAMPLE FOR FIXED WHOLE-GRADIENT CLIPPING

Whole-gradient clipping supplies $h_t = \mathrm{clip}_\lambda(H(x_{t-1} - x^\star) + \zeta_t)$. It clips the signal as well as the noise.

The following exact quadratic example records the resulting possible shift of the target. The general stochastic clipping-bias phenomenon is established by Koloskova et al. (2023).

Proposition D.5 (Shifted equilibrium under asymmetric noise)**.** *Let*

$$F(x) = \frac{\mu x^2}{2}, \qquad g(x,\xi) = \mu x + \xi, \qquad \xi = \begin{cases} a, & \text{with probability } q, \\ -b, & \text{with probability } 1-q, \end{cases} \qquad b = \frac{qa}{1-q},$$

where $a > 0$ and $0 < q < 1/2$. Thus $\mathbb{E}\xi = 0$ and $x^\star = 0$. For a fixed threshold $b < \lambda < a$, the clipped mean field $h_\lambda(x) = \mathbb{E}[\mathrm{clip}_\lambda(\mu x + \xi)]$ has the unique root

$$x_\lambda = \frac{q(a-\lambda)}{(1-q)\mu} > 0. \tag{85}$$

For i.i.d. copies ξ_k, deterministic finite x_0, and deterministic steps $\eta_k > 0$ with $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$, the recursion

$$x_{k+1} = x_k - \eta_{k+1}\,\mathrm{clip}_\lambda(\mu x_k + \xi_{k+1})$$

satisfies $x_k \to x_\lambda$ almost surely. In particular, its arithmetic averages also converge to x_λ, and

$$F(x_\lambda) - F(x^\star) = \frac{q^2(a-\lambda)^2}{2\mu(1-q)^2} > 0. \tag{86}$$

Proof. At the proposed root, $\mu x_\lambda - b = -q\lambda/(1-q) \in (-\lambda, 0)$ and $\mu x_\lambda + a > \lambda$. In a neighborhood of this point,

$$h_\lambda(x) = q\lambda + (1-q)(\mu x - b) = q\lambda + (1-q)\mu x - qa.$$

Its zero is (85). The scalar clipping map is continuous and nondecreasing, so h_λ has these properties as well. It is strictly increasing near its displayed root; monotonicity then implies that this root is unique and that $(x - x_\lambda)h_\lambda(x) > 0$ for every $x \neq x_\lambda$.

Let $V_k = (x_k - x_\lambda)^2$ and $\phi(x) = (x - x_\lambda)h_\lambda(x) \geq 0$. The direction is bounded by λ, and hence

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \leq V_k - 2\eta_{k+1}\phi(x_k) + \eta_{k+1}^2\lambda^2.$$

Consequently $V_k + \lambda^2 \sum_{j=k+1}^{\infty} \eta_j^2$ is a nonnegative supermartingale and converges almost surely. The deterministic tail sum tends to zero, so V_k has a finite almost-sure limit. Taking expectations and summing the drift inequality also shows $\sum_k \eta_{k+1}\phi(x_k) < \infty$ almost surely. If the limit of V_k were positive, the iterates would eventually belong to a compact set bounded away from x_λ. Continuity and the strict sign property would then give $\phi(x_k) \geq c > 0$ along the tail, contradicting $\sum_k \eta_k = \infty$. Thus $V_k \to 0$, proving convergence; convergence of arithmetic averages follows. Substitution into F gives (86). □

The shrinking-error bounds in Appendix C use epoch-dependent thresholds and the stated certificates. The fixed whole-gradient recursion in this last proposition is a different policy, so its nonzero limiting error is consistent with those bounds.

D.4 Symmetric noise and the covariance comparison

Corollary D.6 (One-dimensional symmetric formula)**.** *Let $F(x) = \mu x^2/2$ with $\mu > 0$ and $g(x, \xi) = \mu x + \xi$, where ξ has a continuous distribution symmetric around zero and finite variance. Suppose $\mathbb{P}(|\xi| < \lambda) > 0$ and Assumption D.2 holds at $x_\lambda = 0$. Then $x_\lambda = x^\star = 0$ and*

$$A_\lambda = \mu\mathbb{P}(|\xi| < \lambda), \qquad \Sigma_\lambda = \mathbb{E}\min\{\xi^2, \lambda^2\}.$$

The asymptotic variance of $\sqrt{N}\bar{x}_N$ is

$$V_\lambda = \frac{\mathbb{E}\min\{\xi^2, \lambda^2\}}{\mu^2\mathbb{P}(|\xi| < \lambda)^2}. \tag{87}$$

Proof. Symmetry and oddness imply $h_\lambda(0) = \mathbb{E}\,\mathrm{clip}_\lambda(\xi) = 0$. For $|\xi| \neq \lambda$,

$$\frac{\partial}{\partial x}\,\mathrm{clip}_\lambda(\mu x + \xi)\bigg|_{x=0} = \mu\mathbf{1}_{\{|\xi|<\lambda\}}.$$

The difference quotients are bounded by μ, and continuity of the noise law excludes mass at $\pm\lambda$. Dominated convergence therefore gives the stated A_λ. The root score has mean zero and squared value $\min\{\xi^2, \lambda^2\}$, proving the formula for Σ_λ. Apply Theorem D.4. □

Gaussian noise. Let $\xi \sim \mathcal{N}(0, \sigma^2)$ with $\sigma > 0$. Gaussian integration by parts and Cauchy–Schwarz give

$$\sigma^2\mathbb{P}(|\xi| < \lambda) = \mathbb{E}[\xi\,\mathrm{clip}_\lambda(\xi)] \leq \sigma\left(\mathbb{E}\min\{\xi^2, \lambda^2\}\right)^{1/2}.$$

Consequently, $V_\lambda \geq \sigma^2/\mu^2$. For every finite $\lambda > 0$, the inequality is strict: equality would require $\mathrm{clip}_\lambda(\xi)$ to be proportional to ξ almost surely, which fails for a nondegenerate Gaussian law. Dominated convergence gives $V_\lambda \to \sigma^2/\mu^2$ as $\lambda \to \infty$.

A symmetric heavy-tailed example. The reverse comparison is possible for other noise laws. If ξ has a density f continuous and positive at zero, then

$$\frac{\mathbb{E}\min\{\xi^2, \lambda^2\}}{\lambda^2} \to 1, \qquad \frac{\mathbb{P}(|\xi| < \lambda)}{\lambda} \to 2f(0) \qquad (\lambda \downarrow 0).$$

Thus $V_\lambda \to [4\mu^2 f(0)^2]^{-1}$. For a standard Student t_3 variable,

$$f(x) = \frac{2}{\pi\sqrt{3}}\left(1 + \frac{x^2}{3}\right)^{-2}, \qquad \mathbb{E}\xi^2 = 3,$$

and hence

$$\lim_{\lambda \downarrow 0} V_\lambda = \frac{3\pi^2}{16\mu^2} < \frac{3}{\mu^2}.$$

Therefore sufficiently small positive thresholds give a smaller variance in (87) than the unclipped value. The PR interpretation of this comparison uses the regularity and stability assumptions of Corollary D.6.

D.5 Growing thresholds in the linear additive model

Let $H = H^\top \succ 0$, and write $\mu = \lambda_{\min}(H)$ and $L = \lambda_{\max}(H)$. Let (ζ_k) be i.i.d. with

$$\mathbb{E}\zeta_k = 0, \qquad \mathbb{E}[\zeta_k \zeta_k^\top] = \Sigma, \qquad \mathbb{E}\|\zeta_k\|^2 \le \sigma_2^2, \qquad \mathbb{E}\|\zeta_k\|^p \le \sigma_p^p,$$

where $p > 2$ and the moment bounds are finite and positive.

For each horizon N, use one deterministic threshold $\lambda_N > 0$ throughout the recursion

$$e_{k+1}^{(N)} = e_k^{(N)} - \eta_k \left(H e_k^{(N)} + \mathrm{clip}_{\lambda_N}(\zeta_{k+1}) \right), \qquad k = 1, \dots, N.$$

Use the common initialization $e_1^{(N)} = e_1$, independent of the noise sequence, with $\mathbb{E}\|e_1\|^2 < \infty$. Define

$$b_N := \mathbb{E}\,\mathrm{clip}_{\lambda_N}(\zeta), \qquad e_N^\star := -H^{-1} b_N, \qquad v_{N,k} := \mathrm{clip}_{\lambda_N}(\zeta_k) - b_N.$$

The vector $e_N^\star$ is the exact equilibrium of the horizon-dependent mean field. Set

$$y_k^{(N)} := e_k^{(N)} - e_N^\star, \qquad \bar{e}_N^{(N)} := \frac{1}{N} \sum_{k=1}^N e_k^{(N)}.$$

Theorem D.7 (A sufficient growing-threshold window)**.** *Let $\eta_k = a(k + k_0)^{-\gamma}$ with $a > 0$, $k_0 \ge 0$, and $\gamma \in (1/2, 1)$. Let $\delta_N \in (0, 1/2)$ and $\ell_N := \log(4/\delta_N)$. Suppose*

$$\lambda_N \to \infty, \qquad \frac{\sqrt{N}\,\sigma_p^p}{\lambda_N^{p-1}} \to 0, \qquad \frac{\lambda_N \ell_N}{\sqrt{N}} \to 0. \tag{88}$$

Then

$$\sqrt{N}\,\bar{e}_N^{(N)} \Longrightarrow \mathcal{N}(0, H^{-1}\Sigma H^{-1}). \tag{89}$$

Let $R_N^{(N)}$ be (77) applied to $(y_k^{(N)})$. There is a constant C_R, independent of N and the thresholds, such that

$$\|R_N^{(N)}\|_{L^2} \le C_R N^{\gamma/2 - 1}. \tag{90}$$

For every $\rho_N \in (0, 1/2)$, with probability at least $1 - \delta_N - \rho_N$,

$$\begin{aligned} \|\bar{e}_N^{(N)}\| \le \|H^{-1}\| \Bigg[& \frac{\sigma_p^p}{\lambda_N^{p-1}} + C\sigma_2 \sqrt{\frac{\ell_N}{N}} + C\lambda_N \frac{\ell_N}{N} \\ & + \frac{C_R N^{\gamma/2-1}}{\sqrt{\rho_N}} \Bigg]. \end{aligned} \tag{91}$$

The calibrated remainder term is $o(N^{-1/2})$ if

$$N^{(\gamma-1)/2} \rho_N^{-1/2} \to 0. \tag{92}$$

For $\rho_N = \delta_N$, this condition is equivalent to $\delta_N N^{1-\gamma} \to \infty$.

If ℓ_N is bounded or polylogarithmic, a sufficient power-law choice is $\lambda_N = N^\beta$ with

$$\frac{1}{2(p-1)} < \beta < \frac{1}{2}. \tag{93}$$

More generally, the threshold conditions are compatible whenever $\ell_N = o(N^{(p-2)/(2(p-1))})$.

Proof. The centered recursion is

$$y_{k+1}^{(N)} = (I - \eta_k H) y_k^{(N)} - \eta_k v_{N,k+1}.$$

Its noise satisfies, uniformly in N,

$$\mathbb{E} v_{N,k} = 0, \qquad \mathbb{E}\|v_{N,k}\|^2 \le \mathbb{E}\|\operatorname{clip}_{\lambda_N}(\zeta)\|^2 \le \sigma_2^2.$$

Also $\|b_N\| \le \mathbb{E}\|\zeta\| \le \sigma_2$, so the initial second moments of $y_1^{(N)}$ are uniformly bounded.

For all sufficiently large k, $\eta_k \le 1/L$, and therefore

$$\mathbb{E}\|y_{k+1}^{(N)}\|^2 \le (1 - \mu\eta_k)\mathbb{E}\|y_k^{(N)}\|^2 + \sigma_2^2 \eta_k^2.$$

Since $\gamma < 1$, $\eta_k - \eta_{k+1} = o(\eta_k^2)$. Choose K so that, for $k \ge K$,

$$\eta_k - \eta_{k+1} \le \frac{\mu}{2}\eta_k^2.$$

Taking a sufficiently large constant $C_y \ge 2\sigma_2^2/\mu$ and applying induction gives

$$\mathbb{E}\|y_k^{(N)}\|^2 \le C_y \eta_k,$$

uniformly in N. The constant can be increased to cover the finite initial segment, whose second moments are also uniformly bounded. Proposition D.1 now yields (90) and

$$H\bar{y}_N^{(N)} = -\frac{1}{N}\sum_{k=1}^{N} v_{N,k+1} + R_N^{(N)}. \tag{94}$$

Because $\mathbb{E}\zeta = 0$,

$$\|b_N\| \le \mathbb{E}\|\operatorname{clip}_{\lambda_N}(\zeta) - \zeta\| \le \frac{\mathbb{E}\|\zeta\|^p}{\lambda_N^{p-1}}.$$

Hence

$$\|e_N^\star\| \le \|H^{-1}\| \frac{\sigma_p^p}{\lambda_N^{p-1}} = o(N^{-1/2}). \tag{95}$$

Since $\lambda_N \to \infty$, $\operatorname{clip}_{\lambda_N}(\zeta) \to \zeta$ in L^2. It follows that $\operatorname{Cov}(v_{N,1}) \to \Sigma$. Moreover, $\|v_{N,k}\| \le 2\lambda_N$ and $\lambda_N/\sqrt{N} \to 0$ by (88). The triangular-array Lindeberg condition therefore holds, and

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N} v_{N,k+1} \Longrightarrow \mathcal{N}(0, \Sigma).$$

Combine this limit with (94), $\sqrt{N} R_N^{(N)} \to 0$ in L^2, and (95) to obtain (89).

For the finite-horizon bound, the Hilbert-space Bennett–Freedman inequality Pinelis (1994) gives

$$\left\| \frac{1}{N}\sum_{k=1}^{N} v_{N,k+1} \right\| \le C\sigma_2 \sqrt{\frac{\ell_N}{N}} + C\lambda_N \frac{\ell_N}{N}$$

with probability at least $1 - \delta_N$. Separately, Markov's inequality and (90) give

$$\mathbb{P}\left(\|R_N^{(N)}\| > \frac{C_R N^{\gamma/2-1}}{\sqrt{\rho_N}} \right) \le \rho_N.$$

A union bound, followed by (94) and the bound on $e_N^\star$, proves (91). No independence between these two events is required. Multiplying the remainder radius by $\sqrt{N}$ gives (92).

Finally, for $\lambda_N = N^\beta$, the bias condition requires $\beta > 1/[2(p-1)]$, while the bounded-increment condition holds for $\beta < 1/2$ when ℓ_N is bounded or polylogarithmic. More generally, the lower scale $(\sigma_p^p \sqrt{N})^{1/(p-1)}$ is asymptotically smaller than the upper scale $\sqrt{N}/\ell_N$ whenever $\ell_N = o\big(N^{(p-2)/(2(p-1))}\big)$. Choosing a threshold strictly between these scales proves the last assertion. □

Remark D.8 (Scope of the threshold window)**.** The bias and bounded-increment requirements are sufficient conditions. At $p = 2$, their lower and upper power exponents meet.

The theorem concerns linear additive noise and noise-centered clipping, with one threshold held constant throughout each horizon-N run. Extensions to online thresholds, nonlinear recursions, or whole-gradient clipping require an additional asymptotic linearization and corresponding remainder control.

E Additional Effects of Clipping and Constant Stepsizes

This appendix records the deterministic transit cost of whole-gradient clipping and the constant-step covariance calculation for linear stochastic approximation.

E.1 The deterministic transit cost of clipping

The following exact calculation isolates the deterministic cost of clipping the optimization signal.

Proposition E.1 (linear transit versus geometric contraction)**.** *Consider the noiseless one-dimensional quadratic $F(x) = \mu x^2/2$, $x_0 > \lambda/\mu$, and a step $0 < \eta\mu < 1$.*

For U-SGD, which is ordinary gradient descent here,

$$x_k = (1-\eta\mu)^k x_0, \quad T_{\mathrm{U}}(\lambda) = \left\lceil \frac{\log(\mu x_0/\lambda)}{-\log(1-\eta\mu)} \right\rceil \tag{96}$$

steps suffice to reach $x_k \le \lambda/\mu$. For whole-gradient clipping,

$$x_{k+1} = x_k - \eta\lambda \quad \textit{as long as} \;\; x_k > \lambda/\mu,$$

so the transit time is

$$T_{\mathrm{C}}(\lambda) = \left\lceil \frac{x_0 - (\lambda/\mu)}{\eta\lambda} \right\rceil. \tag{97}$$

Thus a threshold selected from the noise scale can turn logarithmic deterministic contraction into a phase linear in x_0/λ.

Difference clipping can avoid this particular cost only after its anchor error, difference-moment certificate, refresh schedule, and auxiliary oracle requirements have been controlled. Alternatively, a practical whole-gradient policy can use a large warm-up threshold and activate robust clipping only after localization. Both modifications reduce the simplicity advantage of plain Clip-SGD and should be counted when algorithms are compared.

Proof. Without clipping, $x_{k+1} = (1-\eta\mu)x_k$, and solving $(1-\eta\mu)^k x_0 \le \lambda/\mu$ gives (96). With clipping and $x_k > \lambda/\mu$, the direction equals λ, so $x_k = x_0 - k\eta\lambda$ until the threshold region is reached. Solving $x_0 - k\eta\lambda \le \lambda/\mu$ gives (97). □

E.2 The $O(\eta)$ covariance correction

We now turn to the fine-grained constant-step regime of Mou et al. Mou et al. (2020). Consider linear stochastic approximation

$$e_{t+1} = \big(I - \eta A_{t+1}\big)e_t + \eta\varepsilon_{t+1}, \quad A_{t+1} = \bar{A} + \Xi_{t+1}. \tag{98}$$

We use the following checkable conditions. The pairs (A_t, ε_t) are i.i.d. and independent of the past; A_t and ε_t are independent within each pair; $\mathbb{E}A_t = \bar{A}$, $\mathbb{E}\varepsilon_t = 0$, and for some $q > 4$, $\mathbb{E}\,\|A_t\|_{\mathrm{op}}^q + \mathbb{E}\,\|\varepsilon_t\|^q < \infty$. The matrix $\bar{A}$ is invertible with eigenvalues in the open right half-plane, and there exist $c > 0$ and $\eta_0 > 0$ such that

$$\Big(\mathbb{E}\,\|I - \eta A_t\|_{\mathrm{op}}^q\Big)^{1/q} \le 1 - c\eta, \quad 0 < \eta \le \eta_0. \tag{99}$$

Condition (99) implies a unique stationary solution with a finite q-th moment and geometric L_q contraction. We allow either stationary initialization or any $\mathcal{F}_0$-measurable e_0, independent of the future i.i.d. pairs, with $\mathbb{E}\|e_0\|^q < \infty$; the contraction makes its initial transient negligible for the averaged CLT. The condition is stronger than necessary, but avoids assuming "a stationary solution and a CLT" as hypotheses.

Let $\Sigma = \mathbb{E}\left[\varepsilon_t\varepsilon_t^\top\right]$ and let Λ_η be the covariance of the stationary iterate. Independence of the fresh pair from e_t gives

$$\bar{A}\Lambda_\eta + \Lambda_\eta\bar{A}^\top - \eta\bar{A}\Lambda_\eta\bar{A}^\top - \eta\mathbb{E}\left[\Xi\Lambda_\eta\Xi^\top\right] = \eta\Sigma. \tag{100}$$

Theorem E.2. *Under the preceding i.i.d., moment, independence, initialization, and contraction conditions, the recursion has a unique stationary law with covariance Λ_η. For stationary initialization, and also for every independent finite-q initialization described above, the averaged process satisfies a multivariate CLT with asymptotic covariance*

$$\Gamma_\eta = \bar{A}^{-1}\left(\Sigma + \mathbb{E}\left[\Xi\Lambda_\eta\Xi^\top\right]\right)\bar{A}^{-\top}. \tag{101}$$

If Λ_0 solves

$$\bar{A}\Lambda_0 + \Lambda_0\bar{A}^\top = \Sigma,$$

then, as $\eta \downarrow 0$,

$$\Lambda_\eta = \eta\Lambda_0 + O_F(\eta^2), \quad \Gamma_\eta = \bar{A}^{-1}\Sigma\bar{A}^{-\top} + \eta\bar{A}^{-1}\mathbb{E}\left[\Xi\Lambda_0\Xi^\top\right]\bar{A}^{-\top} + O_F(\eta^2), \tag{102}$$

where $O_F(\eta^2)$ is in Frobenius norm and is uniform for $0 < \eta \le \eta_1$ for some $\eta_1 \le \eta_0$.

The exact identity (101) and expansion (102) are specializations of the fixed-step i.i.d. LSA theory of Mou et al. Mou et al. (2020). Tight high-probability and moment bounds under weaker random-matrix conditions are developed by Durmus et al. Durmus et al. (2021; 2025); quantitative Gaussian approximation and multiplier-bootstrap inference for decreasing-step and Markovian LSA are given in Samsonov et al. (2024a; 2026), with temporal-difference stability in Samsonov et al. (2024b). These results delimit the assumptions under which Gaussian concentration or inferential approximations are available.

For a fixed clipped score, freeze the local linearization at x_λ and suppose that the resulting affine recursion satisfies the i.i.d. contraction conditions above. The formal substitution

$$\bar{A} \rightsquigarrow A_\lambda, \quad \Sigma \rightsquigarrow \Sigma_\lambda, \quad \Xi \rightsquigarrow J_\lambda(\xi) - A_\lambda,$$

where $J_\lambda(\xi) = \nabla_x\psi_\lambda(x_\lambda, \xi)$ away from the clipping sphere, identifies the *multiplicative-noise component* of the local $O(\eta)$ correction. Taylor curvature terms, a generally nonzero $O(\eta)$ stationary-mean displacement, and possible score–Jacobian dependence contribute additional terms. A full clipped nonlinear expansion would require a stationary bias theorem and a Poisson-equation/Markov-chain analysis uniform in the clipping parameter. Step-size extrapolation results such as Huo et al. (2026) also show that an $O(\eta)$ term is not an unimprovable risk contribution.

For the scalar illustration, take $A_t = 1 \pm 0.6$ with equal probabilities and unit additive-noise variance. Here

$$\Lambda_\eta = \frac{\eta}{2 - 1.36\eta}, \quad \Gamma_\eta = 1 + 0.36\Lambda_\eta = \frac{2 - \eta}{2 - 1.36\eta}.$$

The classical covariance is one, so this formula isolates the upward correction caused by a nonzero constant stepsize.

E.2.1 Proofs for the covariance result

Lemma E.3 (conditional covariance and Lindeberg for the effective LSA noise)**.** *Under the assumptions preceding Theorem E.2, define $\Xi_t = A_t - \bar{A}$ and $m_{t+1} = \varepsilon_{t+1} - \Xi_{t+1}e_t$. Let Λ_η be the covariance under the unique stationary law. For stationary initialization, and for every independent initial condition with finite q-th moment,*

$$\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}\left[m_{t+1}m_{t+1}^\top \mid \mathcal{F}_t\right] \xrightarrow{\mathbb{P}} \Sigma + \mathbb{E}\left[\Xi\Lambda_\eta\Xi^\top\right], \tag{103}$$

$$\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}\Big[\|m_{t+1}\|^2\mathbf{1}\left\{\|m_{t+1}\| > a\sqrt{N}\right\} \mid \mathcal{F}_t\Big] \xrightarrow{\mathbb{P}} 0 \quad \textit{for every } a > 0. \tag{104}$$

Consequently,

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} m_{t+1} \quad \Longrightarrow \quad \mathcal{N}\big(0, \Sigma + \mathbb{E}\left[\Xi\Lambda_\eta\Xi^\top\right]\big). \tag{105}$$

Proof. The fresh pair is independent of $\mathcal{F}_t$, $\mathbb{E}\Xi = 0$, and ε is independent of Ξ. Hence $(m_{t+1}, \mathcal{F}_{t+1})$ is a martingale difference and, for the positive linear operator $\mathcal{T}(B) = \mathbb{E}[\Xi B \Xi^\top]$,

$$\mathbb{E}[m_{t+1} m_{t+1}^\top \mid \mathcal{F}_t] = \Sigma + \mathcal{T}(e_t e_t^\top).$$

Under stationary initialization, geometric L_q contraction implies ergodicity, so the ergodic theorem applied entrywise to $\mathcal{T}(e_t e_t^\top)$ gives (103). For a nonstationary finite-q initial condition, couple e_t with a stationary copy $\widetilde{e}_t$ driven by the same fresh pairs. Iterating (99) gives

$$\{\mathbb{E}\|e_t - \widetilde{e}_t\|^q\}^{1/q} \le (1 - c\eta)^t \{\mathbb{E}\|e_0 - \widetilde{e}_0\|^q\}^{1/q}.$$

Because $q > 4$, Hölder's inequality, the finite q-moments of Ξ, e_t and $\widetilde{e}_t$, and $\|uu^\top - vv^\top\|_F \le (\|u\| + \|v\|)\|u - v\|$ show that the Cesaro mean of $\|\mathcal{T}(e_t e_t^\top) - \mathcal{T}(\widetilde{e}_t \widetilde{e}_t^\top)\|_F$ converges to zero in L^1. Thus the same covariance limit holds from every allowed initialization.

For Lindeberg, conditional Markov's inequality gives

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\Big[\|m_{t+1}\|^2 \mathbf{1}\left\{\|m_{t+1}\| > a\sqrt{N}\right\} \mid \mathcal{F}_t\Big] &\le \frac{1}{a^{q-2} N^{(q-2)/2}} \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}[\|m_{t+1}\|^q \mid \mathcal{F}_t] \\ &\le \frac{C}{N^{(q-2)/2}} \left(1 + \frac{1}{N}\sum_{t=1}^{N} \|e_t\|^q\right).\end{aligned}$$

The last empirical average is $O_{\mathbb{P}}(1)$ by the stationary ergodic theorem and the same coupling argument. Since $q > 4$, the right-hand side converges to zero. The multivariate martingale CLT applied with (103)–(104) proves (105). □

Proof of Theorem E.2. In stationarity, e_t is independent of the fresh pair $(A_{t+1}, \varepsilon_{t+1})$. Its mean m_η satisfies $m_\eta = (I - \eta\bar{A})m_\eta$; invertibility of $\bar{A}$ therefore gives $m_\eta = 0$. Expanding the covariance of (98) gives

$$\begin{aligned}\Lambda_\eta &= \mathbb{E}[(I - \eta A)\Lambda_\eta (I - \eta A)^\top] + \eta^2 \Sigma \\ &= \Lambda_\eta - \eta(\bar{A}\Lambda_\eta + \Lambda_\eta \bar{A}^\top) + \eta^2 \bar{A}\Lambda_\eta \bar{A}^\top + \eta^2 \mathbb{E}\left[\Xi \Lambda_\eta \Xi^\top\right] + \eta^2 \Sigma.\end{aligned}$$

Rearrangement proves (100).

The recursion also gives the exact identity

$$\bar{A} e_t = \varepsilon_{t+1} - \Xi_{t+1} e_t - \frac{e_{t+1} - e_t}{\eta}. \tag{106}$$

The effective noise

$$m_{t+1} = \varepsilon_{t+1} - \Xi_{t+1} e_t$$

has the martingale CLT established in Lemma E.3. Summing (106) gives

$$\bar{A}\sqrt{N}\,\bar{e}_N = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} m_{t+1} - \frac{e_{N+1} - e_1}{\eta\sqrt{N}}.$$

The contraction assumption gives $\sup_t \mathbb{E}\|e_t\|^q < \infty$, so the boundary term is $o_{\mathbb{P}}(1)$ for every allowed initialization. Lemma E.3 and Slutsky's theorem yield (101).

Write $\Lambda_\eta = \eta\Lambda_0 + R_\eta$ in (100). Invertibility of the continuous Lyapunov operator $X \mapsto \bar{A}X + X\bar{A}^\top$ gives $\|R_\eta\| = O(\eta^2)$. Substitution into (101) proves (102). □

F Experimental Protocol

This appendix gives the protocol for the OBD experiment in Section 6. The source and target empirical risks, the compared directions, and the output average are defined there.

F.1 Source-only selection and paired evaluation

We freeze the experimental protocol before target evaluation. First, one U-SGD stepsize is selected on 128 source-objective validation trajectories and then shared by every method, so clipping is the only algorithmic change. Next, a second set of 256 trajectories selects one threshold from the predeclared grid $\{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 50, 200, \infty\}$ by minimizing the source 0.95-quantile at $N = 2097152$; exact ties are resolved toward the largest threshold. This selects $\lambda_{\mathrm{val}} = 5$, where the subscript emphasizes that this is the validation-selected value of the same clipping threshold λ. The Random-policy log is not used in either selection stage. We then freeze λ_{val} and evaluate it against U-SGD on 512 fresh trajectories, using common sampled rows between methods. The complete grid is evaluated separately as a threshold-sensitivity diagnostic.

F.2 Uncertainty and threshold sensitivity

Figure 1 reports pointwise distribution-free 95% order-statistic intervals conditional on the fixed logs. The paired-bootstrap intervals in the main text quantify the reported clipped-to-unclipped error ratios. The threshold sweep is a diagnostic over the predeclared grid and is not used to tune the frozen method.

F.3 Tail diagnostic

Figure 1(a) characterizes the empirical tails of the importance weights and stochastic gradients. At horizontal value x, its vertical value is the fraction of observations whose weight or gradient norm exceeds x times its median. The long right tails show that a tiny fraction of observations is orders of magnitude larger than a typical one: observations at or above the empirical 0.999-quantile of the gradient norm account for 99.6% of its empirical second moment. This concentration of the empirical second moment motivates the clipping comparison.

G Directions for Accelerated Methods

Accelerated clipped methods under heavy-tailed noise already admit high-probability guarantees Gorbunov et al. (2020), while unmodified accelerated proximal schemes can be optimal in expectation under weak moments He & Lu (2025); Liu (2026). A direct counterpart of our three-level comparison is still missing. We therefore record only the weighted quantities that such a theorem would have to control.

G.1 Weighted Nagaev geometry

Suppose an accelerated output admits a predictable stochastic representation

$$Z_N = \sum_{k=1}^{N} w_{k,N}\zeta_k, \quad w_{k,N} \ge 0, \quad \sum_{k=1}^{N} w_{k,N} = 1. \tag{107}$$

Corollary A.2 then gives

$$\|Z_N\| \lesssim \sigma_2\sqrt{\ell_\delta V_{2,N}} + \sigma_p\delta^{-1/p}V_{p,N}^{1/p}, \quad V_{r,N} = \sum_{k=1}^{N} w_{k,N}^r. \tag{108}$$

Define

$$N_2 = V_{2,N}^{-1}, \quad N_p = V_{p,N}^{-1/(p-1)}, \quad w_{\max,N} = \max_k w_{k,N}.$$

Uniform averaging has $N_2 = N_p = N$. Momentum can concentrate the weights, reducing N_p and increasing the maximum influence. The resulting candidate Gaussian-dominance boundary is

$$\delta\ell_\delta^{p/2} \gtrsim \chi_p^p \frac{N_2^{p/2}}{N_p^{p-1}}. \tag{109}$$

Hypothesis G.1 (accelerated regime map)**.** If a concrete accelerated convex method has deterministic error $O(LR_0^2/N^2)$ and a localized predictable representation of the form (107), then its candidate heavy-tail stochastic term is (108). For $p > 2$, U-SGD-style untrimmed acceleration should match the leading clipped order in the region (109); outside it, clipping can improve confidence dependence but may attenuate the accelerated signal.

The hard step is the predictable representation after momentum, clipping, and localization have interacted.

G.2 Thresholds, centering, and first-order efficiency

For a common threshold, the bias term is of order σ_p^p/λ^{p-1} and the bounded-increment contribution is of order $\lambda \ell_\delta w_{\max,N}$. Balancing them suggests

$$\lambda_\star \asymp \sigma_p (w_{\max,N}\ell_\delta)^{-1/p}, \quad \text{correction } \asymp \sigma_p (w_{\max,N}\ell_\delta)^{1-1/p}.$$

A sufficient weighted analogue of the robust–efficient window is

$$\left(\frac{\sigma_p^p}{\sigma_2\sqrt{V_{2,N}}}\right)^{1/(p-1)} \ll \lambda_N \ll \frac{\sigma_2\sqrt{V_{2,N}}}{\ell_N w_{\max,N}}.$$

For uniform weights this reduces to (93); concentrated accelerated weights can shrink or eliminate the interval. An implementable difference-clipped accelerated method would additionally need a charged anchor estimator and a proof that one clipped error remains controlled through all later momentum states.

Under strong convexity, a successful construction should retain the deterministic $\sqrt{\kappa}\log(\Delta_0/\varepsilon)$ burn-in while preserving the leading variance term $\sigma_2^2/(\mu\varepsilon)$. For fine-grained limits, the augmented position–momentum state leads to a block Lyapunov operator. A proof would need exactly four ingredients: the stochastic weights, a localized weighted Nagaev decomposition, a maximum-influence bound through the augmented dynamics, and a triangular-array linearization uniform in the threshold.