NAGAEV BOUNDS FOR NONCONVEX MOMENTUM: FINITE MOMENTS AND TEMPORAL DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Momentum can smooth individual gradient errors without removing the confidence cost of rare shocks along an optimization trajectory. For ordinary nonconvex momentum under conditional second and pth moments, $p > 2$, we prove a bound on average squared-gradient stationarity that separates a logarithmic variance term from a polynomial rare-shock term. A quadratic construction matches the polynomial component in its powers of horizon, confidence and stepsize over the stated moment class, allowing the noise law to depend on horizon and confidence. Both bounds have constants uniform in momentum within an explicit stepsize region. We then replace conditional moment envelopes by a budget for accumulated noise energy. This gives $T^{-1/2}$ stationarity at fixed momentum and confidence under a uniform second-moment bound, including state-dependent martingale noise with unbounded conditional variance. A predictable correction extends the analysis to additive causal noise, covering polynomial linear filters with exponent $a > 3/2$ under finite $p > 2$ moments. Exact calculations and simulations explain the common mechanism: pointwise smoothing can coexist with a persistent normalized shock response. They also test fixed-law tails and the scope of the random-volatility and causal extensions.

1 INTRODUCTION

Momentum is a standard component of stochastic optimization for machine learning (Polyak, 1964; Bottou et al., 2018; Yan et al., 2018). Its averaging of recent gradients raises a basic reliability question: how much protection does this smoothing provide against a rare, large gradient error? Heavy-tailed gradient models are motivated by empirical observations in attention models (Zhang et al., 2020). Temporal dependence introduces a further difficulty, because errors can cluster and their conditional variance can fluctuate. Can finitely many moments yield a trajectory-level confidence guarantee for ordinary momentum that separates accumulated variance from rare shocks? Our upper and lower bounds answer this question for nonconvex average stationarity. Section 5 explains why pointwise smoothing does not remove the resulting polynomial confidence cost.

The benefits and limits of momentum depend on the criterion and noise model. Yuan et al. (2016) relate momentum to rescaled SGD in the small-step, strongly convex mean-square regime; Kidambi et al. (2018) exhibit stochastic problems on which heavy ball and Nesterov momentum fail to accelerate. Frequency-domain analysis explains momentum's filtering behavior (Li et al., 2025). These results distinguish pointwise smoothing and acceleration from the confidence of a trajectory-level stationarity guarantee.

Expected stationarity under finite variance is well understood for SGD (Ghadimi & Lan, 2013) and SGDM (Liu et al., 2020). Finite-moment analyses further cover SGD and mini-batch variants (Fatkhullin et al., 2025), and moments of average stationarity for unchanged SGD and SGDM (Liu, 2026). High-probability momentum guarantees are available under conditional sub-Gaussian (Li & Orabona, 2020) and sub-Weibull assumptions (Li et al., 2026); related SGD results include Li & Liu (2022); Madden et al. (2024). These results motivate a finite-moment confidence bound that retains the distinct contributions of variance, rare shocks and momentum.

Nagaev inequalities provide this exponential-plus-polynomial separation for independent sums (Fuk & Nagaev, 1971; Nagaev, 1979) and dependent processes (Liu et al., 2013; Wu & Wu, 2016). In

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.

stochastic approximation, this structure has been established for averaged linear regression (Lou et al., 2022) and strongly convex models (Wei et al., 2025; Zhang et al., 2026). General nonconvex stationarity introduces adaptive, potentially unbounded martingale coefficients in the descent inequality, together with an additional momentum-filtered noise energy term. Retaining this energy is essential to track how momentum changes the confidence bound.

Contributions. Our main result identifies the confidence cost of ordinary SGDM; two extensions broaden the noise models it can accommodate. **Main result: a confidence split with matching polynomial cost.** Under conditional second and pth moments, $p > 2$, Theorem 2.1 separates a logarithmic variance term from a $T^{2/p-1}\delta^{-2/p}$ rare-shock term. A quadratic construction matches the powers of horizon, confidence and stepsize in the polynomial component for both SGD and SGDM (Theorem 2.2); its law may depend on T, δ. The constants are uniform in momentum within the explicit stepsize region. Thus momentum's pointwise smoothing leaves a polynomial trajectory-level confidence cost over this moment class.

Extension 1: random conditional variance via energy control. A worst-case variance envelope can be unbounded even when accumulated noise energy is controlled. Theorem 3.1 turns a budget for realized and predictable energy into a stationarity bound. A uniform second-moment bound already yields a $T^{-1/2}$ rate at fixed momentum and confidence, including state-dependent martingale oracles with unbounded conditional variance. Higher moments and dependence control refine the budget.

Extension 2: causal dependence via a predictable correction. An additive causal error can also have nonzero conditional mean. A martingale–coboundary correction absorbs this predictable component and gives the same fixed-confidence rate (Theorem 4.2). Under square-summable coupling tails, it covers polynomial linear filters with $a > 3/2$, beyond the coefficientwise condition requiring $a > 2$. Proposition E.1 illustrates the exact decomposition boundary in this subclass.

Complementary algorithmic approaches. Clipping and normalization modify the response to extreme gradients (Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023; Sadiev et al., 2023; Hübler et al., 2025; Liu & Zhou, 2025). Momentum can also improve normalized SGD under additional smoothness (Cutkosky & Mehta, 2020). Our results characterize the confidence guarantees of ordinary momentum, giving a reference point for these robust alternatives. For dependent sampling, Duchi et al. (2012); Sun et al. (2018); Even (2023) analyze gradient queries through mixing and hitting times. We use innovation coupling (Wu, 2005) to control stationary energies and additive causal errors. Table 1 compares the scope of representative guarantees; Appendices H–I give quantitative rates and detailed comparisons.

Table 1: Representative guarantees, distinguished by algorithm, assumptions and criterion. Our momentum-uniform bounds use the stepsize region in (2). Conditional moments have deterministic envelopes. SA denotes stochastic approximation; A_T is average squared-gradient stationarity.

Result	Algorithm / noise	Relevant conclusion
Li & Orabona (2020); Li et al. (2026)	SGDM; conditional sub-Gaussian or sub-Weibull	High-probability nonconvex gradient guarantees.
Liu (2026)	SGD/SGDM; conditional central qth moment, $1 < q \leq 2$	Bounds on $\mathbb{E}[A_T^{q/2}]$.
Lou et al. (2022); Wei et al. (2025); Zhang et al. (2026)	SGD/SA; finite moments and model-specific regularity	Nagaev bounds for estimation or iterate error in linear or strongly convex models.
Nguyen et al. (2023)	Clipped SGD; conditional central moments	High-probability A_T bounds with logarithmic confidence costs.
Theorems 2.1–2.2	Ordinary SGDM; conditional $p > 2$	Variance/polynomial split for A_T; matching polynomial component with constants uniform in β.
Theorems 3.1, 4.2	Ordinary SGDM; unconditional moments and energy control	$T^{-1/2}$ stationarity at fixed β, δ; random volatility and causal extensions.

2 SETTING AND A CONFIDENCE BOUND UNIFORM IN MOMENTUM

Let $F : \mathbb{R}^d \to \mathbb{R}$ have an L-Lipschitz gradient, $L > 0$, and finite lower bound $F_\star$. Let x_1 be deterministic and $\Delta = F(x_1) - F_\star$. We study ordinary stochastic gradient descent with momentum (SGDM; Polyak, 1964) in its exponential moving-average convention:

$$m_t = \beta m_{t-1} + (1-\beta) g_t, \qquad x_{t+1} = x_t - \alpha m_t, \qquad m_0 = 0, \qquad g_t = \nabla F(x_t) + \xi_t. \tag{1}$$

Here $\alpha > 0$ and $0 \le \beta < 1$. On a probability space with increasing filtration $(\mathcal{F}_t)$, ξ_t is $\mathcal{F}_t$-measurable and x_t, m_{t-1} are $\mathcal{F}_{t-1}$-measurable. The heavy-ball stepsize is $\eta = \alpha(1-\beta)$; α is the effective stepsize. Define

$$A_T = \frac{1}{T}\sum_{t=1}^{T} \|\nabla F(x_t)\|^2, \qquad \kappa = \frac{\beta}{1-\beta}, \qquad 0 < \alpha \le \frac{1-\beta}{8L}. \tag{2}$$

Throughout, T is a positive integer and $\delta \in (0,1)$ denotes the failure probability. The statistic uses gradients before the updates. Norms are Euclidean, $\|Z\|_p = (\mathbb{E}\|Z\|^p)^{1/p}$, and $\mathbb{E}_t Z = \mathbb{E}(Z \mid \mathcal{F}_t)$. The first result separates the confidence cost of accumulated variance from that of rare shocks, using deterministic conditional moment envelopes. Section 3 replaces these envelopes by an energy budget.

Theorem 2.1 (Conditional finite-moment bound)**.** *Suppose $\sigma, \nu_p \in [0,\infty)$, $\mathbb{E}(\xi_t \mid \mathcal{F}_{t-1}) = 0$ and, almost surely,*

$$\mathbb{E}(\|\xi_t\|^2 \mid \mathcal{F}_{t-1}) \le \sigma^2, \qquad \mathbb{E}(\|\xi_t\|^p \mid \mathcal{F}_{t-1}) \le \nu_p^p, \quad p > 2. \tag{3}$$

Under (2)*, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\boxed{A_T \le C_p \left\{ \frac{\Delta}{\alpha T} + L\alpha \left[\sigma^2 \log\frac{6}{\delta} + \nu_p^2 T^{2/p-1} \delta^{-2/p} \right] \right\}.} \tag{4}$$

The constant C_p depends only on p, not on β or d.

The three contributions are the optimization transient $\Delta/(\alpha T)$, the variance cost $L\alpha\sigma^2 \log(6/\delta)$, and the rare-shock cost $L\alpha\nu_p^2 T^{2/p-1}\delta^{-2/p}$. The bound allows adaptive noise without bounded gradients, iid errors, or stationarity. Its constant is uniform in momentum within (2); comparisons use the effective stepsize α, with heavy-ball stepsize $\eta = \alpha(1-\beta)$. The next result shows that the polynomial cost cannot be removed over this moment class.

Theorem 2.2 (Momentum-uniform polynomial lower bound)**.** *Let $F(x) = \lambda x^2/2$, $\lambda > 0$, $x_1 = m_0 = 0$, $\alpha > 0$, $0 \le \beta < 1$, and suppose $a = \alpha\lambda \le \min\{1/16, (1-\beta)/8\}$. Put $N = \lfloor (8a)^{-1} \rfloor$ and let $T \ge 2N$ be even. For every $p > 2$, $\nu > 0$, and $\delta \in (0, 1/8]$, there is a centered iid noise law, allowed to depend on T, δ, with $\mathbb{E}|\xi_t|^p = \nu^p$ such that*

$$\Pr\left\{ A_T \ge \frac{\lambda}{2048\, 4^{2/p}} \alpha\nu^2 T^{2/p-1}\delta^{-2/p} \right\} \ge \delta. \tag{5}$$

The law puts total mass $4\delta/T$ at shocks of magnitude $\nu(T/(4\delta))^{1/p}$, with variance $\nu^2(4\delta/T)^{1-2/p}$. A single early shock leaves a fixed fraction of its response energy, uniformly in β within the stated stepsize region. This matches the powers of T, δ, α in the polynomial component of (4), over the stated moment class. The construction includes SGD ($\beta = 0$). Appendix B proves the result.

Together, the bounds distinguish pointwise smoothing from trajectory-level confidence. Momentum can reduce the variance of a filtered error at a fixed time, yet an early shock affects many later iterates. Its normalized response energy stays bounded away from zero, so smoothing does not remove the polynomial confidence cost for average stationarity. The exact calculations in Section 5 make this distinction explicit.

For a concrete confidence comparison, take $p = 4$, $\delta = T^{-1}$ and $\alpha = a_0/\sqrt{T}$ within (2), with fixed problem parameters. The upper bound gives $A_T = O(T^{-1/2}\log T)$, whereas applying Markov to a finite-variance $O(T^{-1/2})$ expectation bound gives $O(T^{1/2})$. Appendix H develops the IID specialization and other moment orders. The bound also allows $\beta_T \uparrow 1$ provided $1-\beta_T \ge 8La_0/\sqrt{T}$; dimension enters through σ, ν_p.

Why the momentum factor cancels. The proof separates filtered-noise energy from the adaptive martingale term. The virtual iterate $y_t = x_t - \alpha\kappa m_{t-1}$ obeys $y_{t+1} = y_t - \alpha(\nabla F(x_t) + \xi_t)$. To control the momentum factor, we retain the filtered noise $n_{t-1} = (1-\beta)\sum_{s=1}^{t-1}\beta^{t-1-s}\xi_s$ in the descent inequality. In the path space $(\mathbb{R}^d)^T$, it is a sum of deterministic linear images of martingale differences. The squared operator norm of each image is at most $c_\beta = (1-\beta)/(1+\beta)$. A Hilbert-space Fuk–Nagaev inequality (Mollenhauer & Fiedler, 2025) controls this energy and the raw-noise energy together, without a union bound over time or coordinates. The critical factor becomes

$$L\alpha\kappa^2 c_\beta \le \frac{\beta^2}{8(1+\beta)} \le \frac{1}{16}.$$

Stopping the remaining adaptive martingale at an objective level and using Fan et al. (2017) completes the argument. Appendix A gives the pathwise inequality, the two lifts, and the first-exit proof.

3 Random conditional variance: an energy-based bound

A worst-case conditional variance bound can fail even when the noise accumulated over a run is well behaved. For example, with iid centered unit-variance innovations, $e_t = \varepsilon_t\varepsilon_{t-1}$ is a martingale difference sequence (MDS), yet $\mathbb{E}_{t-1}e_t^2 = \varepsilon_{t-1}^2$ has no deterministic upper bound when the innovations have unbounded support. Its marginal pth moment is finite whenever that of ε_0 is finite. We replace the conditional envelopes by a budget for the combined realized and predictable energy: the squared errors accumulated along the run plus their conditional expectations.

Theorem 3.1 (Energy-based martingale bound)**.** *For the objective and initialization in Section 2, assume ξ_t are square-integrable martingale differences. Suppose a deterministic $\mathcal{E} = \mathcal{E}(T,\delta) \in [0,\infty)$ satisfies*

$$\Pr\left\{\sum_{t=1}^{T}(\|\xi_t\|^2 + \mathbb{E}_{t-1}\|\xi_t\|^2) > T\mathcal{E}\right\} \le \delta/2. \tag{6}$$

Let $\ell = \log(4/\delta)$ and $L_\beta = L[1+\beta^2/(8(1-\beta))]$. Under (2)*, with probability at least $1-\delta$,*

$$A_T \le \frac{4\Delta}{\alpha T} + 16\sqrt{\frac{L\Delta\mathcal{E}\ell}{T}} + 60L_\beta\alpha\mathcal{E}\ell. \tag{7}$$

The theorem converts any such energy budget into a stationarity guarantee. Its pathwise filtered-energy estimate retains the momentum factor L_β. If $\sup_t \|\xi_t\|_p \le \nu_p$ for $p > 2$, Minkowski and Markov give $\mathcal{E} = 2\nu_p^2(\delta/2)^{-2/p}$ without stationarity.

A finite-second-moment consequence. If $\sup_t \mathbb{E}\|\xi_t\|^2 \le \nu_2^2$, the mean combined energy per step is at most $2\nu_2^2$. Markov therefore gives $\mathcal{E} = 4\nu_2^2/\delta$. For fixed β, δ and $\alpha = a_0/\sqrt{T}$ satisfying (2), this already yields $A_T = O(T^{-1/2})$. The finer confidence split in Theorem 2.1 uses the additional conditional $p > 2$ moment.

Concentrating combined energy. Stationary causal dependence offers a second route: it controls energy fluctuations around their mean and gives a variance-scale budget. For a stationary causal process $Z_t = G(\ldots,\varepsilon_{t-1},\varepsilon_t)$, the innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ are iid, and the filtration is $\mathcal{F}_t = \sigma(\varepsilon_s : s \le t)$. Let Z_k^* replace innovation zero by an independent copy, and define

$$\delta_q^Z(k) = \|Z_k - Z_k^*\|_q, \quad \Theta_{m,q}^Z = \sum_{k\ge m}\delta_q^Z(k), \quad \mathcal{D}_{q,b}(Z) = \sup_{m\ge 0}(m+1)^b\Theta_{m,q}^Z. \tag{8}$$

These coefficients measure how replacing one innovation affects future observations (Wu, 2005). For stationary causal martingale noise e, put $\sigma_e^2 = \mathbb{E}\|e_0\|^2$ and $Z_t^e = \|e_t\|^2 + \mathbb{E}_{t-1}\|e_t\|^2 - 2\sigma_e^2$. Lemma C.1 gives an explicit allowance $\mathcal{Q}_p^e(T,u) = 2\sigma_e^2 + \mathfrak{b}_{p/2}(Z^e;T,u)$ for $u \in (0,1)$, where

$$\mathfrak{b}_q(Z;T,u) \lesssim \begin{cases} \mathcal{D}_{q,0}(Z)T^{1/q-1}u^{-1/q}, & 1 < q \le 2, \\ \mathcal{D}_{2,b}(Z)\sqrt{\log(C/u)/T} + \mathcal{D}_{q,b}(Z)T^{1/q-1}u^{-1/q}, & q > 2. \end{cases} \tag{9}$$

Here $\lesssim$ suppresses constants depending only on q and, in the second line, b; $C \ge 2$ has the same dependence. The first line is a projective moment bound; the second requires $b > 1/2 - 1/q$ and

follows from Wu & Wu (2016). Squaring the noise changes its moment order to $q = p/2$. Thus the Gaussian-plus-polynomial energy refinement uses $p > 4$, while $2 < p \le 4$ still gives convergence. The norms of Z^e are bounded by those of e via $\mathcal{D}_{q,b}(Z^e) \le 4\|e_0\|_{2q}\mathcal{D}_{2q,b}(e)$.

At $\alpha = a_0/\sqrt{T}$ and fixed β, δ, either energy choice gives $A_T = O(T^{-1/2})$. With the refined choice and $\delta = T^{-r}$, $0 < r \le p/2 - 1$ is sufficient for $A_T = O(\log T/\sqrt{T})$. Theorem 2.1 further removes the localization logarithm from the polynomial term when conditional envelopes are available.

Why the adaptive coefficient is admissible. Stopping at a deterministic objective level bounds the adaptive coefficient via $\|\nabla F(y_t)\|^2 \le 2L(F(y_t) - F_\star)$. A self-normalized inequality then uses (6) to control the stopped transform, and a first-exit argument removes the stop. Appendix C proves this on the intersection of the energy and martingale events, without conditioning on future energy.

A stationary process can also dominate the energy of a nonstationary, iterate-dependent oracle, as the following application shows.

Corollary 3.2 (State-dependent amplitude)**.** *Let $e_t \in \mathbb{R}^r$ be a stationary causal MDS, $\|e_0\|_p < \infty$, $p > 2$, with an available combined-energy allowance $\mathcal{Q}^e_p$. If $H : \mathbb{R}^d \to \mathbb{R}^{d\times r}$ is measurable and $\sup_x \|H(x)\|_{\mathrm{op}} \le h$, then the oracle $g_t = \nabla F(x_t) + H(x_t)e_t$ satisfies Theorem 3.1 with $\mathcal{E} = h^2\mathcal{Q}^e_p(T, \delta/2)$ and $\sup_t \mathbb{E}\|\xi_t\|^p \le h^p\mathbb{E}\|e_0\|^p$.*

The gradient error itself is generally nonstationary. The corollary uses stationarity only for a process dominating its energy, making the result applicable to the learning experiment in Section 6.

4 Causal noise beyond martingale differences

Causal dependence may create a predictable noise component as well as random variance. We treat additive centered stationary noise $\xi_t = G(\ldots, \varepsilon_{t-1}, \varepsilon_t)$, where G has no iterate argument. A martingale–coboundary decomposition (Gordin, 1969; Volný, 2017) absorbs the predictable component into a corrected virtual iterate, returning the problem to the energy analysis of Section 3. The next proposition gives a sufficient dependence condition for this correction to exist in L^p.

Proposition 4.1 (Square-tail sufficient condition)**.** *Let ξ_t be centered, stationary and causal relative to $(\mathcal{F}_t)$, with $\xi_0 \in L^p(\mathbb{R}^d)$, $p > 2$, and suppose*

$$\Theta^{\xi}_{0,p} < \infty, \qquad \mathfrak{R}^2_p = \sum_{m\ge 1} (\Theta^{\xi}_{m,p})^2 < \infty. \tag{10}$$

With $\mathcal{P}_t Z = \mathbb{E}_t Z - \mathbb{E}_{t-1} Z$, define

$$D_t = \sum_{j\ge 0} \mathcal{P}_t \xi_{t+j}, \qquad \sum_{j=1}^{N} \mathbb{E}(\xi_{t+j} \mid \mathcal{F}_t) \xrightarrow[N\to\infty]{L^p} R_t. \tag{11}$$

Then D_t is a stationary causal MDS and

$$\xi_t = D_t + R_{t-1} - R_t, \qquad \|D_0\|_p \le \Theta^{\xi}_{0,p}, \qquad \|R_0\|_p \le C_p\mathfrak{R}_p.$$

The D series is absolutely L^p convergent. The definition of R uses L^p convergence without assuming absolute summability.

The proof uses finite projections, as in Wu (2007, Theorem 1(ii)), and gives explicit L^p bounds for both terms. The square-tail condition is weaker than the coefficientwise requirement $\sum_{k\ge 1} k\delta^{\xi}_p(k) < \infty$. For $\xi_t = \sum_{j\ge 0}(j+1)^{-a}\varepsilon_{t-j}$, these conditions give $a > 3/2$ and $a > 2$, respectively. Proposition E.1 illustrates sharpness at $a = 3/2$ for stationary adapted L^2 decompositions in this linear subclass.

Put $w_t = y_t - \alpha R_{t-1}$. Predictability is preserved, and direct substitution yields

$$w_{t+1} = w_t - \alpha(\nabla F(x_t) + D_t). \tag{12}$$

The correction leaves two additional energy terms, for ξ and R. The next result states their costs explicitly.

Theorem 4.2 (Causal nonconvex bound)**.** *Assume the conditions of Proposition 4.1 and* (2)*. Write* $\nu_{\xi,p} = \|\xi_0\|_p$*,* $\rho_p = \|R_0\|_p$*,* $\epsilon = \delta/5$*,* $\ell_D = \log(2/\epsilon)$*, and let* $\mathcal{E}_D$ *be any combined-energy allowance for* D *with failure probability* ϵ*, including the weak choice in* (29)*. Define*

$$\begin{aligned} r_0 &= \rho_p \epsilon^{-1/p}, & \Delta_\alpha &= (\sqrt{\Delta} + \alpha\sqrt{L/2}\, r_0)^2, \\ \mathcal{E}_\xi &= \nu_{\xi,p}^2 \epsilon^{-2/p}, & \mathcal{E}_R &= \rho_p^2 \epsilon^{-2/p}, \\ B &= L\alpha\mathcal{E}_D + 2L^2\alpha^2\kappa^2\mathcal{E}_\xi + L^2\alpha^2\mathcal{E}_R. \end{aligned} \tag{13}$$

Then, with probability at least $1-\delta$*,*

$$A_T \le \frac{4\Delta_\alpha}{\alpha T} + 16\sqrt{\frac{L\Delta_\alpha \mathcal{E}_D \ell_D}{T}} + 32L\alpha\mathcal{E}_D\ell_D + 12B. \tag{14}$$

For fixed β, δ*,* $\alpha = a_0/\sqrt{T}$ *gives* $A_T = O(T^{-1/2})$*.*

Thus the predictable component changes the energy budget while preserving the fixed-confidence stationarity order. To let confidence increase with the horizon, we also control fluctuations of all three energies.

Corollary 4.3 (Refining all three energies)**.** *In Theorem 4.2, one may replace the allowances by* $\mathcal{E}_D = \mathcal{Q}_p^D(T,\epsilon)$*,* $\mathcal{E}_\xi = \mathcal{Q}_{p,\mathrm{raw}}^\xi(T,\epsilon)$ *and* $\mathcal{E}_R = \mathcal{Q}_{p,\mathrm{raw}}^R(T,\epsilon)$*, as defined in Lemma C.1. A sufficient condition is*

$$\sum_{k\ge 0}(k+1)^{b+1}\delta_p^\xi(k) < \infty, \quad b = 0 \text{ if } 2 < p \le 4, \quad b > 1/2 - 2/p \text{ if } p > 4. \tag{15}$$

Then $\delta = T^{-r}$*,* $0 < r \le p/2 - 1$*, permits* $A_T = O(\log T/\sqrt{T})$ *for fixed* β*.*

Condition (10) supplies the basic fixed-confidence result; (15) additionally controls dependence in all three energies for the shrinking-confidence refinement. Appendix E proves both statements.

Filtered volatility. Let $u_t = \varepsilon_t\varepsilon_{t-1}$ as above, and $\xi_t = \rho\xi_{t-1} + cu_t$ for $|\rho| < 1$, $c \ne 0$ and $\varepsilon_0 \in L^p$, $p > 2$. Then

$$D_t = \frac{c}{1-\rho}u_t, \quad R_t = \frac{\rho}{1-\rho}\xi_t, \quad \mathbb{E}_{t-1}D_t^2 = \frac{c^2}{(1-\rho)^2}\varepsilon_{t-1}^2.$$

All required dependence norms are finite for any fixed ρ, while the conditional variance is unbounded when the innovations have unbounded support. At fixed marginal noise variance σ_ξ^2, the martingale variance is $\sigma_\xi^2(1+\rho)/(1-\rho)$. This explains why matching marginal noise variances does not match the optimization noise scale under dependence.

5 Momentum smoothing and optimization noise

Three quantities explain the distinction behind Theorems 2.1–2.2: pointwise filtered variance, long-run covariance, and the optimization response energy to a shock. Their dependence on momentum differs.

For a stationary AR(1) noise process with lag-one correlation ρ, define its stationary exponential average $n_t^\infty = (1-\beta)\sum_{j\ge 0}\beta^j\xi_{t-j}$. Then

$$\frac{\mathrm{Var}(n_t^\infty)}{\mathrm{Var}(\xi_t)} = \frac{1-\beta}{1+\beta}\frac{1+\beta\rho}{1-\beta\rho}, \qquad \sum_{k\in\mathbb{Z}}\mathrm{Cov}(n_0^\infty, n_k^\infty) = \sum_{k\in\mathbb{Z}}\mathrm{Cov}(\xi_0,\xi_k). \tag{16}$$

The first expression can decrease with momentum; the second shows that this pointwise smoothing preserves the long-run variance. For iid errors ($\rho = 0$), the pointwise ratio is $(1-\beta)/(1+\beta)$.

For a single shock of magnitude $\nu(T/(4\delta))^{1/p}$ at time s on $F(x) = \lambda x^2/2$, with $\lambda, \nu > 0$, put $a = \alpha\lambda$, $o = 1-\beta$ and $h_0 = 1$, $h_1 = 1+\beta-ao$, $h_j = (1+\beta-ao)h_{j-1} - \beta h_{j-2}$ for $j \ge 2$. Starting from $x_1 = m_0 = 0$, the normalized energy is exactly

$$\frac{A_T}{\lambda\alpha\nu^2 T^{2/p-1}\delta^{-2/p}} = 4^{-2/p}ao^2\sum_{j=0}^{T-s-1}h_j^2. \tag{17}$$

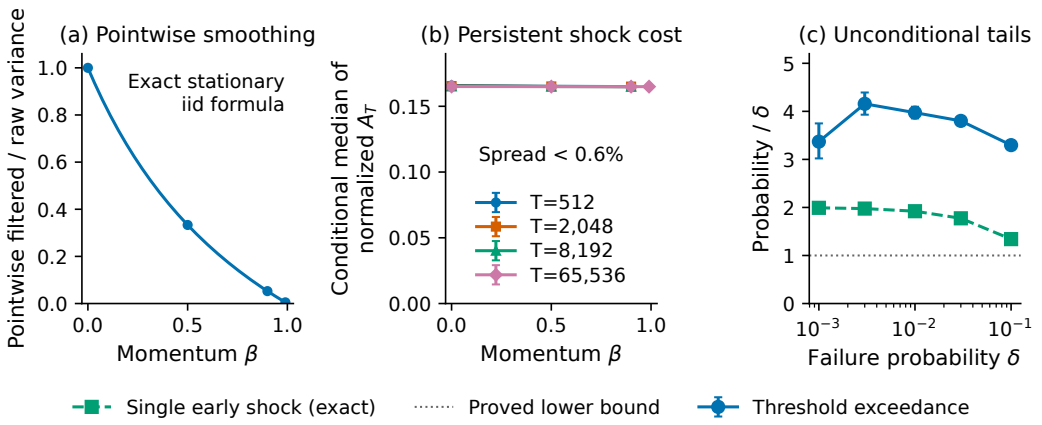

Figure 1: Pointwise smoothing and trajectory-level shock cost, $p = 2.5$. (a) Analytic stationary filtered/raw variance for iid noise, $(1-\beta)/(1+\beta)$. (b) Conditional median $A_T/(\alpha T^{2/p-1}\delta^{-2/p})$ on event E, at $\delta = .01$, with a zero baseline. (c) Unconditional lower-threshold exceedance probability divided by δ, $T = 8192$, $\beta = .9$, compared with exact $P(E)/\delta$. Here E is one shock in the first $T/2$ steps and none elsewhere. Rank/binomial intervals use the conditional subset/all 100,000 runs. Panels (b)–(c) concern finite runs.

In the stable region, replacing the finite sum by its infinite limit gives $4^{-2/p}(1+\beta)/[2(1+\beta)-a\omega]$. For fixed β and small a, this reference tends to $1/(2\,4^{2/p})$, making the persistent impulse scale explicit. Proposition F.1 proves the identity and gives the finite-response remainder. This also calibrates Figure 1 without fitting a constant to its observed values.

For iid noise on the quadratic, the stationary iterate variance has leading term $\alpha\sigma^2/(2\lambda)$ at small effective step, without a leading $(1-\beta)$ gain. Under correlated noise it can even increase: at $\lambda = 1$, $\alpha = .03$, $\rho = .8$, changing β from 0 to .9 raises it by 7.27%. Appendix F derives these values within the exact quadratic stability region; Figure 8 displays further filter responses.

6 Experiments

We test the evidence chain from rare-shock persistence and fixed-law tails to state-dependent energy control and additive causal dependence. All stationarity statistics use full gradients before each update. Momentum comparisons share noise streams. Reported uncertainty is pointwise 95% Monte Carlo uncertainty: binomial-rank intervals for quantiles and exact Clopper–Pearson intervals for probabilities. Appendix G and the code provide all configurations, per-run statistics and supporting controls.

Rare-shock energy persists across momentum values. We simulate Theorem 2.2 on $F(x) = x^2/2$ with $\alpha = .25/\sqrt{T}$, $p \in \{2.5, 6\}$ and 100,000 runs per configuration. The grid covers $T = 512, 2048, 8192, 65536$, $\delta = .1, .03, .01, .003, .001$ and $\beta = 0, .5, .9$; $\beta = .99$ is included at $T = 65536$ within the stepsize region. Figure 1 places the exact pointwise smoothing ratio beside the simulated trajectory response. Normalized impulse energy remains near .165, with less than .6% variation across the displayed settings. At $p = 2.5$, $T = 8192$, $\delta = .01$ and $\beta = .9$, the unconditional lower-threshold exceedance probability is .03973, above the proved .01. The value near .165 agrees with the small-step impulse reference $1/(2\,4^{2/p}) = .16494$ at $p = 2.5$; the saved calibration also evaluates the finite-response correction (Proposition F.1).

Fixed-law tails and the role of initialization. The lower construction allows its noise law to vary with T, δ. A separate fixed-law experiment holds each data-generating law fixed and uses the nonconvex, 1.1-smooth objective $F(x) = .05x^2 + 1 - \cos x$ and iid noise of variance $.7^2$. With $x_1 = 0$, $T = 8192$, $\alpha = .25/\sqrt{T}$, $\beta = .9$ and 100,000 runs, Gaussian and standardized t_3 noise give 99.9th percentiles .001640 [.001617, .001667] and .002760 [.002577, .002945]. The 68.3% increase

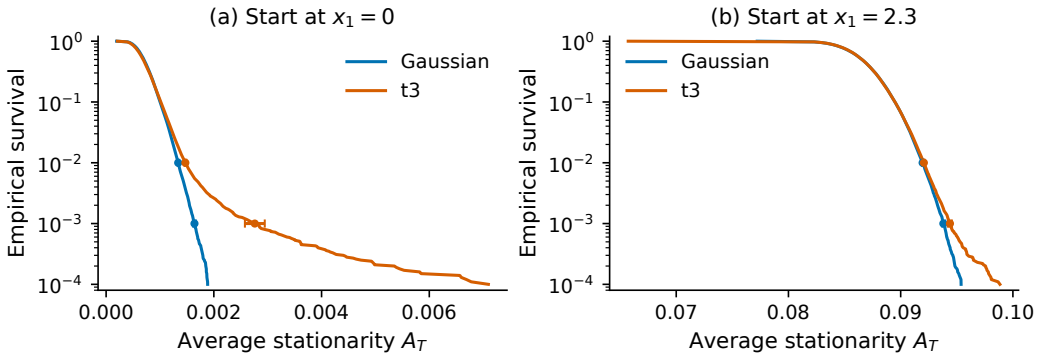

Figure 2: Fixed-law stationarity tails on $F(x) = .05x^2 + 1 - \cos x$, $T = 8192$, $\beta = .9$, with 100,000 runs per law and initialization. Points mark the 99th and 99.9th percentiles with horizontal 95% rank intervals. Curves stop at ten observed exceedances. Horizontal scales differ between the panels. The two initializations separate the noise response from a larger transient.

isolates a substantial tail effect. An $x_1 = 2.3$ control has a larger initialization transient and a smaller relative difference of .63% (Figure 2). Both laws satisfy Theorem 2.1 for $2 < p < 3$; this comparison shows how their finite-sample tails depend on initialization.

State-dependent stochastic gradients in a learning problem. We use a fixed synthetic binary dataset with $n = 64$, $d = 20$ and squared sigmoid loss with ridge parameter $\lambda = .001$:

$$f_i(\theta) = \tfrac{1}{2}(\mathrm{sigmoid}(a_i^\top \theta) - y_i)^2 + \tfrac{\lambda}{2}\|\theta\|^2, \qquad F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta).$$

A valid global smoothness bound is .05360; a Hessian witness has minimum eigenvalue $-.01512$. The oracle is $g_t = \nabla f_{I_t}(\theta_t) + .7\varepsilon_t\varepsilon_{t-1}\nabla\phi(\theta_t)$, where I_t is uniform on $\{1, \dots, n\}$, independently across time and of the innovations, $\phi(\theta) = .75v^\top\theta + .25\sin(v^\top\theta)$ and $\|v\| = 1$. Its amplitude lies in $[.5, 1]$, yielding unbounded conditional variance along the adaptive trajectory. Appendix D verifies the combined-energy bound, including sample-selection noise.

Within the common theorem stepsize cap, both methods' 90th percentiles decrease and remain close (Figure 3). At $T = 2048$ under t_3 products, SGD and SGDM give .001047 and .001051. These common-step results agree with the comparable stationarity guarantees for SGD and SGDM. Independent-pilot tuning additionally compares clipped SGD (Nguyen et al., 2023, Algorithm 1); its configurations, stationarity and terminal-loss results appear in Figure 9 and Table 3. Those results distinguish robustness gains from the effects of tuning, including steps beyond SGDM's sufficient theorem cap.

Causal noise and the additional slow-decay range. On the scalar objective $F(x) = .05x^2 + 1 - \cos x$, take $x_1 = 0$, $m_0 = 0$, $\alpha = .25/\sqrt{T}$ and $\beta = .9$. For $\xi_t = \rho\xi_{t-1} + .7\sqrt{1-\rho^2}\,\varepsilon_t\varepsilon_{t-1}$, compare $\rho = 0, .8$ with Gaussian and standardized t_3 innovations, using 8,192 runs. All four 99th-percentile curves decrease with T (Figure 4, left). At $T = 8192$, the t_3 quantile increases from .002560 to .019510 when ρ changes from zero to .8, despite equal marginal variance. This is consistent with the larger martingale variance in the causal decomposition, which increases by a factor of nine.

We also simulate the stationary infinite filter $\xi_t = .7\,\zeta(2a)^{-1/2}\sum_{j\geq 0}(j+1)^{-a}\varepsilon_{t-j}$ with Gaussian innovations and $a = 1.6, 1.8, 2.2$; ζ is the Riemann zeta function. The first two exponents satisfy (10) while failing $\sum k\delta_p^\xi(k) < \infty$. With 4,096 runs, their 99th percentiles decrease across all three horizons. For $a = 1.6$, the quantile falls from .03044 at $T = 512$ to .005755 at $T = 8192$. This tests the wider fixed-confidence scope of Theorem 4.2. Stationary covariance sampling avoids a finite-memory replacement of the infinite filter. All three filters have summable autocovariances; the additional range $3/2 < a \leq 2$ concerns the coboundary construction. Their differing long-run variances, given in Appendix G, explain why equal marginal variance permits different stationarity levels.

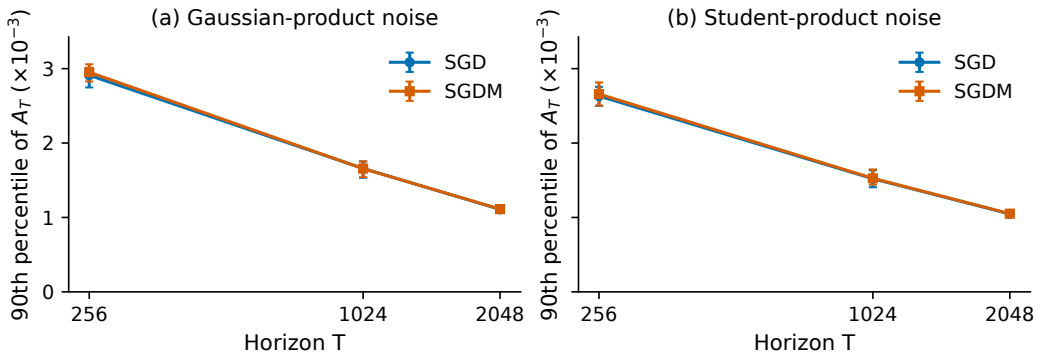

Figure 3: Vector learning with state-dependent product volatility. SGD and SGDM use a common effective step satisfying the theorem's cap. Both panels use the same zero-based vertical scale. Error bars are 95% rank intervals for the 90th percentile: 1,024 runs at $T = 256, 1024$ and 10,000 at $T = 2048$. The oracle uses controlled injected noise.

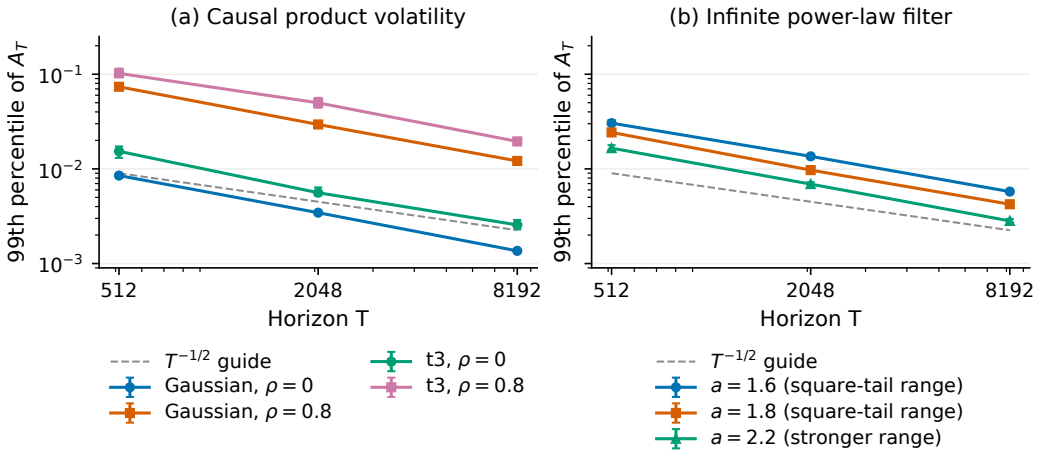

Figure 4: Average stationarity under causal noise, with common logarithmic axes, $x_1 = 0$, $\beta = .9$ and $\alpha = .25/\sqrt{T}$. Left: product volatility, 8,192 runs per setting. Right: power-law Gaussian filters, 4,096 runs; $a = 1.6, 1.8$ enter through the square-tail extension. Error bars are 95% rank intervals. The gray $T^{-1/2}$ guides show the theorem's reference order, not fitted rates. Both momentum values and initialization controls appear in Figure 7.

7 Discussion

Momentum's pointwise smoothing leaves a polynomial trajectory-level confidence cost. The matching powers hold over the stated moment class, whose laws may vary with T, δ, with constants uniform in momentum within the stepsize region. Exact impulse calculations explain this cost; fixed-law simulations show tail effects most clearly with a small initialization transient.

Accumulated energy provides a route beyond deterministic conditional envelopes. It yields fixed-confidence guarantees for random volatility and, after a predictable correction, additive causal errors. Stronger dependence control supports shrinking-confidence refinements. The learning experiment illustrates the broader scope with comparable common-step SGD and SGDM stationarity, and clipping as a complementary robust update. The guarantees concern average squared gradients along the run; terminal loss and last-iterate behavior are separate criteria. A next step is to extend the causal correction to state-dependent nonmartingale gradient fields, where products of iterate increments and the coboundary require additional control.

REPRODUCIBILITY STATEMENT

The supplement provides full proofs, experiment configurations, executable Python scripts, generated figures, and per-replication statistics. The experiments use synthetic data and require no external dataset. All displayed experimental values are generated by the accompanying code.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing and editing, literature search and identification, drafting portions of the manuscript, experimental design and code development, exploration of new research directions based on existing results, and the development and verification of mathematical arguments and proofs.

We have reviewed all AI-assisted work. Suggested references were checked against original sources, experimental code and numerical results were executed and verified by the authors, and mathematical claims, derivations, and proofs were independently checked for correctness.

We take responsibility for the final content of this work, including all text, claims, proofs, code, experiments, and other artifacts produced with the aid of generative AI.

REFERENCES

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268, 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3A): 1902–1933, 2004. doi: 10.1214/009117904000000397.

John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012. doi: 10.1137/110836043.

Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9412–9439, 2023. URL https://proceedings.mlr.press/v202/even23a.html.

Xiequan Fan, Ion Grama, and Quansheng Liu. Deviation inequalities for martingales with applications. *Journal of Mathematical Analysis and Applications*, 448(1):538–566, 2017. doi: 10.1016/j.jmaa.2016.11.023.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860v1*, 2025. URL https://arxiv.org/abs/2508.04860v1.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

D. Kh. Fuk and S. V. Nagaev. Probability inequalities for sums of independent random variables. *Theory of Probability and Its Applications*, 16(4):643–660, 1971. doi: 10.1137/1116071.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018. URL https://www.tse-fr.eu/articles/stochastic-heavy-ball.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15042–15053, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/abd1c782880cc59759f4112fda0b8f98-Abstract.html.

M. I. Gordin. The central limit theorem for stationary processes. *Doklady Akademii Nauk SSSR*, 188 (4):739–741, 1969. URL https://www.mathnet.ru/eng/dan34921.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421, 2025. URL https://proceedings.mlr.press/v258/hubler25a.html.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1803.05591.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *Proceedings of the International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12931–12963, 2022.

Shaojie Li, Pengwei Tang, Bowei Zhu, and Yong Liu. High probability bounds for non-convex stochastic optimization with momentum. In *International Conference on Learning Representations*, 2026.

Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.19671.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. *arXiv preprint arXiv:2007.14294*, 2020. URL https://arxiv.org/abs/2007.14294.

Weidong Liu, Han Xiao, and Wei Biao Wu. Probability and moment inequalities under dependence. *Statistica Sinica*, 23(3):1257–1272, 2013. doi: 10.5705/ss.2011.287.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18261–18271, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/d3f5d4de09ea19461dab00590df91e4f-Abstract.html.

Zijian Liu. In-expectation convergence of stochastic gradient methods under heavy-tailed noise. *arXiv preprint arXiv:2606.00520v1*, 2026. URL https://arxiv.org/abs/2606.00520v1.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of the Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2266–2290, 2023. URL https://proceedings.mlr.press/v195/liu23c.html.

Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-Gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23(46):1–22, 2022. URL https://jmlr.org/papers/volume23/21-0560/21-0560.pdf.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-Weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

Mattes Mollenhauer and Christian Fiedler. Fuk–Nagaev inequality in smooth Banach spaces: Optimum bounds for distributions of heavy-tailed martingales. *arXiv preprint arXiv:2512.10012v2*, 2025. URL https://arxiv.org/abs/2512.10012v2.

S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979. doi: 10.1214/aop/1176994938.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24191–24222, 2023.

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

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

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 International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648, 2023.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1809.04216.

Dalibor Volný. Martingale-coboundary decomposition for stationary random fields. *arXiv preprint arXiv:1706.07978v1*, 2017. URL https://arxiv.org/abs/1706.07978v1.

Ziyang Wei, Jiaqi Li, Zhipeng Lou, and Wei Biao Wu. Gaussian approximation and concentration of constant learning-rate stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2848.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005. doi: 10.1073/pnas.0506715102.

Wei Biao Wu. Strong invariance principles for dependent random variables. *The Annals of Probability*, 35(6):2294–2320, 2007.

Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016. doi: 10.1214/16-EJS1108.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2955–2961, 2018. doi: 10.24963/ijcai.2018/410.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. URL https://jmlr.org/papers/v17/16-157.html.

Danna Zhang, Zhipeng Lou, and Wei Biao Wu. Sharp concentration analysis of online stochastic approximation algorithms. *IEEE Transactions on Information Theory*, 2026. doi: 10.1109/TIT.2026.3702077.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/b05b57f6add810d3b7490866d74c0053-Abstract.html.

A CONDITIONAL MOMENTS: FILTERED ENERGY AND LOCALIZATION

A.1 A DETERMINISTIC DESCENT INEQUALITY

The smoothness and lower-bound assumptions imply

$$\|\nabla F(z)\|^2 \le 2L\{F(z) - F_\star\}. \tag{18}$$

Indeed, apply the descent lemma at $z - \nabla F(z)/L$ and use $F_\star$ as a lower bound. Throughout the proofs, $F_\star$ need not be attained.

For optimization prefixes, $\sum_{t\le k}$ means $\sum_{t=1}^{k}$. Split the momentum into

$$v_{t-1} = (1-\beta)\sum_{s=1}^{t-1}\beta^{t-1-s}\nabla F(x_s), \qquad n_{t-1} = (1-\beta)\sum_{s=1}^{t-1}\beta^{t-1-s}\xi_s.$$

Then $m_{t-1} = v_{t-1} + n_{t-1}$. Jensen's inequality, with the missing geometric weight placed at zero, gives for every prefix k,

$$\sum_{t\le k}\|v_{t-1}\|^2 \le \sum_{t\le k}\|\nabla F(x_t)\|^2, \qquad \sum_{t\le k}\|m_{t-1}\|^2 \le 2\sum_{t\le k}\|\nabla F(x_t)\|^2 + 2\sum_{t\le k}\|n_{t-1}\|^2. \tag{19}$$

Splitting the signal and noise contributions to momentum is also useful in the moment analysis of Liu (2026). Here the additional step is to retain the filtered-noise energy for a high-probability estimate.

Lemma A.1 (Split-filter descent)**.** *For arbitrary noise vectors, let $y_t = x_t - \alpha\kappa m_{t-1}$ and $M_k = -\sum_{t\le k}\langle\nabla F(y_t), \xi_t\rangle$. Under* (2)*, every $k \le T$ satisfies*

$$F(y_{k+1}) - F_\star + \frac{\alpha}{4}\sum_{t\le k}\|\nabla F(x_t)\|^2 \le \Delta + \alpha M_k + L\alpha^2\sum_{t\le k}\|\xi_t\|^2 + L^2\alpha^3\kappa^2\sum_{t\le k}\|n_{t-1}\|^2. \tag{20}$$

Proof. Direct substitution gives $y_{t+1} = y_t - \alpha(\nabla F(x_t) + \xi_t)$. Use $\langle a, b\rangle \ge \|b\|^2/2 - \|a-b\|^2/2$ in the descent lemma:

$$\begin{aligned} F(y_{t+1}) \le F(y_t) &- \frac{\alpha}{2}\|\nabla F(x_t)\|^2 + \frac{\alpha}{2}\|\nabla F(y_t) - \nabla F(x_t)\|^2 \\ &- \alpha\langle\nabla F(y_t), \xi_t\rangle + L\alpha^2(\|\nabla F(x_t)\|^2 + \|\xi_t\|^2). \end{aligned}$$

The gradient difference is at most $L\alpha\kappa\|m_{t-1}\|$. Summing and using (19), the coefficient of $\sum\|\nabla F(x_t)\|^2$ on the left is $\alpha(1/2 - L\alpha - L^2\alpha^2\kappa^2) \ge \alpha/4$: $L\alpha \le 1/8$ and $L\alpha\kappa \le \beta/8$. □

A.2 CONCENTRATION INPUTS AND TWO PATH-SPACE LIFTS

We use the terminal-sum consequence of Mollenhauer & Fiedler (2025, Theorem 2.2). For Hilbert-valued martingale differences X_s, if almost surely

$$\sum_s \mathbb{E}_{s-1}\|X_s\|^2 \le v, \qquad \sum_s \mathbb{E}_{s-1}\|X_s\|^p \le a,$$

then, with probability at least $1-u$,

$$\begin{aligned} \Big\|\sum_s X_s\Big\| &\le \sqrt{2v\log(2/u)} + c_p(2a/u)^{1/p}, \\ c_p &= \frac{1}{2p} + \min\{1/p, 1/5\} + 1 + \mathbf{1}_{\{p>3\}}\frac{p}{3}. \end{aligned} \tag{21}$$

The Hilbert smoothness constant equals one in every dimension. For scalar martingales we use the maximal inequality of Fan et al. (2017, Corollary 2.5). A conservative form is

$$\Pr\{\max_{k\le T}\textstyle\sum_{s\le k} X_s > z\} \le \exp\{-z^2/(K_p v)\} + 2B_p a/z^p, \tag{22}$$

Here $K_p = \frac{1}{2}(p+2)^2 e^p$, $B_p = (1+2/p)^p$, and v, a bound the conditional second and absolute pth moment totals. The factor two gives a convenient conservative form for the subsequent bounds.

Let $\ell = \log(6/\delta)$ and set

$$b_p = \max\{(2B_p)^{1/p}, c_p\}, \qquad J = \sqrt{K_p T \sigma^2 \ell} + b_p \nu_p (6T/\delta)^{1/p}. \tag{23}$$

For a scalar transform with predictable coefficients of norm at most G, (22) gives $\max_k \sum_{s\le k} X_s \le GJ$ except with probability $\delta/3$. The exponential and polynomial terms are each at most $\delta/6$. Degenerate zero-moment cases follow by continuity or directly.

Lemma A.2 (Raw and filtered energy)**.** *Under* (3)*, with probability at least* $1 - 2\delta/3$*, simultaneously*

$$\sum_{t=1}^{T} \|\xi_t\|^2 \le J^2, \qquad \sum_{t=1}^{T} \|n_{t-1}\|^2 \le c_\beta J^2, \quad c_\beta = \frac{1-\beta}{1+\beta}. \tag{24}$$

Proof. Work in the Hilbert space $\mathcal{H}_T = (\mathbb{R}^d)^T$. Let E_s insert a vector in coordinate s. The increments $E_s\xi_s$ are martingale differences, their conditional moment totals are at most $T\sigma^2, T\nu_p^p$, and the squared norm of their sum equals the raw energy. Apply (21) with $u = \delta/3$.

Next define deterministic maps

$$(B_s u)_t = \begin{cases} (1-\beta)\beta^{t-1-s}u, & 1 \le s < t \le T, \\ 0, & 1 \le t \le s \le T. \end{cases}$$

Their squared operator norms are at most $(1-\beta)^2 \sum_{j\ge 0} \beta^{2j} = c_\beta$; the formula also holds at $\beta = 0$, using $0^0 = 1$ for the lag-zero weight. Thus $B_s\xi_s$ are Hilbert-valued martingale differences with moment totals at most $c_\beta T\sigma^2$ and $c_\beta^{p/2} T\nu_p^p$. Their sum is $(n_0, \dots, n_{T-1})$. A second application of (21) gives the second energy event. A union bound completes the proof. □

A.3 Proof of Theorem 2.1

Put $S = J^2/T$ and

$$B = (L\alpha + L^2\alpha^2\kappa^2 c_\beta)S, \qquad H = 4\{\Delta + L\alpha^2 J^2 + \alpha T B\}.$$

The apparent singularity in the filtered-noise term cancels because

$$L\alpha\kappa^2 c_\beta = \frac{L\alpha\beta^2}{(1-\beta)(1+\beta)} \le \frac{\beta^2}{8(1+\beta)} \le \frac{1}{16}. \tag{25}$$

Consequently $B \le 17L\alpha S/16$ and $H \le 4\Delta + (33/4)L\alpha^2 J^2$.

Suppose first $H > 0$. The indicators $I_t = \mathbf{1}_{\{\max_{1\le s\le t}(F(y_s)-F_\star)\le H\}}$ are $\mathcal{F}_{t-1}$-measurable. The stopped transform $M_k^\circ = -\sum_{t\le k} I_t \langle \nabla F(y_t), \xi_t\rangle$ has coefficients bounded by $G = \sqrt{2LH}$, by (18). Apply (22): outside an event of probability $\delta/3$, $\max_k M_k^\circ \le GJ$. Intersect this event with (24). We do not condition a martingale inequality on an energy event that depends on future observations.

If the first exit occurs at $s+1 \le T+1$, the stopped and original transforms agree through s, including the increment producing the exit. Dropping the nonnegative gradient term in (20) gives

$$F(y_{s+1}) - F_\star \le \Delta + \alpha\sqrt{2LH}J + \alpha T B \le H/2 + \Delta + L\alpha^2 J^2 + \alpha T B = 3H/4,$$

a contradiction. At $k = T$ retain that term to obtain

$$A_T \le \frac{4\Delta}{\alpha T} + \frac{4\sqrt{2LH}J}{T} + 4B.$$

For $d_0 = \Delta/(\alpha T)$ and $e_0 = L\alpha J^2/T$, the right side is at most

$$4d_0 + 8\sqrt{2d_0 e_0} + (4\sqrt{33/2} + 17/4)e_0 \le 10d_0 + 27e_0.$$

Finally $J^2/T \le C_p[\sigma^2\ell + \nu_p^2 T^{2/p-1}\delta^{-2/p}]$. All constants depend only on p, and the three exceptional probabilities sum to at most δ. If $H = 0$, both Δ and J vanish. The noise is zero almost surely and (18) gives a constant trajectory. This completes the proof.

B A UNIFORM POLYNOMIAL LOWER BOUND

Proof of Theorem 2.2. Let $a = \alpha\lambda$, $o = 1 - \beta$, and $N = \lfloor (8a)^{-1} \rfloor$. The homogeneous quadratic recursion has characteristic polynomial $P(r) = r^2 - (1 + \beta - ao)r + \beta$. Its discriminant is

$$o\{o - 2a(1 + \beta) + a^2 o\} > 0,$$

because $a \le o/8$. Its roots satisfy $0 \le r_- < r_+ < 1$. To verify the upper bound, their sum is less than two and $P(1) = ao > 0$; two roots straddling one would give $P(1) \le 0$, while two roots above one would have sum at least two. Furthermore

$$P(1 - 2a) = a\{-o + 2a(1 + \beta)\} \le 0,$$

so $r_+ \ge 1 - 2a > 0$. Define $h_0 = 1$, $h_1 = 1 + \beta - ao$ and $h_j = (1 + \beta - ao)h_{j-1} - \beta h_{j-2}$. Then

$$h_j = r_+^j \sum_{l=0}^{j} (r_-/r_+)^l, \qquad r_-/r_+ = \beta/r_+^2 \ge \beta.$$

For $j < N$, Bernoulli's inequality gives $r_+^j \ge 1 - 2aj \ge 3/4$. Hence

$$oh_j \ge r_+^j (1 - \beta^{j+1}).$$

Since $a \le 1/16$, $N \ge (16a)^{-1}$ and $No \ge 1/2$. For $j = \lfloor N/2 \rfloor, \ldots, N - 1$, we have $(j+1)o \ge 1/4$ and therefore

$$oh_j \ge \tfrac{3}{4}(1 - e^{-1/4}) > \tfrac{1}{8}. \tag{26}$$

This reasoning includes $\beta = 0$ directly.

Set $r = 4\delta/T$, $v = \nu r^{-1/p}$ and choose independently $\xi_t = +v, -v, 0$ with probabilities $r/2, r/2, 1 - r$. The law is centered and $\mathbb{E}|\xi_t|^p = \nu^p$. Let E be the event of exactly one nonzero shock among times $1, \ldots, T$, located in $1, \ldots, T/2$. Then

$$\Pr(E) = \frac{T}{2} r(1 - r)^{T-1} = 2\delta(1 - 4\delta/T)^{T-1} \ge \delta,$$

where $(1-r)^{T-1} \ge 1-(T-1)r \ge 1-4\delta \ge 1/2$. If the shock occurs at s, then $x_{s+1+j} = -\alpha o h_j \xi_s$. All $0 \le j < N$ enter the statistic A_T, since $s + N \le T$. At least $N/2$ of these responses satisfy (26), and so

$$A_T \ge \frac{N\lambda^2\alpha^2 v^2}{128T} \ge \frac{\lambda\alpha v^2}{2048T}.$$

The simpler estimate $h_j \ge 3/4$ also gives, on the same event, $A_T \ge \lambda\alpha o^2 v^2/(64T)$. Substituting v^2 proves the stronger combined constant

$$\frac{\lambda}{4^{2/p}} \max\{(1 - \beta)^2/64, 1/2048\}.$$

The theorem displays only its beta-uniform part. The noise law depends on T, δ; the assertion is a distribution-class lower bound, not a fixed-law tail asymptotic. In this construction $\mathbb{E}\xi_t^2 = \nu^2(4\delta/T)^{1-2/p}$. Thus the variance term in Theorem 2.1, relative to its polynomial term, is of order $\delta \log(6/\delta)$ as $\delta \downarrow 0$. The polynomial contribution cannot be absorbed into that variance term uniformly over this class. □

C UNCONDITIONAL ENERGY CONTROL AND SELF-NORMALIZED DESCENT

C.1 SCALAR FUNCTIONAL-DEPENDENCE BOUNDS

For a centered stationary causal scalar process $Z \in L^q$, let $\delta_q^Z(k)$, $\Theta_{m,q}^Z$ and $\mathcal{D}_{q,b}(Z)$ be as defined in Section 3. For $u \in (0, 1)$, set

$$\mathfrak{b}_q(Z; T, u) = \begin{cases} 2^{1/q}\mathcal{D}_{q,0}(Z) T^{1/q-1} u^{-1/q}, & 1 < q \le 2, \\ C_{q,b}\left\{ \mathcal{D}_{2,b}(Z)\sqrt{\dfrac{\log(C_{q,b}/u)}{T}} + \mathcal{D}_{q,b}(Z) T^{1/q-1} u^{-1/q} \right\}, & q > 2,\ b > \tfrac{1}{2} - \tfrac{1}{q}. \end{cases} \tag{27}$$

Here and below choose $C_{q,b} \geq 2$ large enough for the stated probability bound. Then

$$\Pr\left\{T^{-1}\sum_{t=1}^{T} Z_t > \mathfrak{b}_q(Z;T,u)\right\} \leq u. \tag{28}$$

For $q > 2$, this follows from Wu & Wu (2016, Theorem 2), with all weights equal to one, by bounding its exponential and polynomial terms separately. The decay condition in (27) is retained. For $1 < q \leq 2$, expand $Z_t = \sum_{j\geq 0} \mathcal{P}_{t-j} Z_t$. For scalar u, v and $\operatorname{sgn}(0) = 0$, the elementary inequality

$$|u+v|^q \leq |u|^q + q\operatorname{sgn}(u)|u|^{q-1}v + 2|v|^q$$

gives $\mathbb{E}|\sum_t Y_t|^q \leq 2\sum_t \mathbb{E}|Y_t|^q$ for martingale differences, by conditioning and iteration. For each fixed j, this yields

$$\Big\|\sum_{t=1}^{T} \mathcal{P}_{t-j} Z_t\Big\|_q \leq 2^{1/q} T^{1/q} \delta_q^Z(j).$$

Summing over j by Minkowski, taking the L^q limit of finite projection sums, and applying Markov proves (28). This low-moment bound does not contain a Gaussian branch.

Lemma C.1 (Energy allowances). *For adapted vectors e_t with $\sup_t \|e_t\|_p \leq \nu_p$, $p > 2$, put $V_t^e = \mathbb{E}_{t-1}\|e_t\|^2$. For $u \in (0,1)$, with failure probability at most u, respectively,*

$$T^{-1}\sum_t \|e_t\|^2 \leq \nu_p^2 u^{-2/p}, \qquad T^{-1}\sum_t (\|e_t\|^2 + V_t^e) \leq 2\nu_p^2 u^{-2/p}. \tag{29}$$

If e is stationary causal, let $\sigma_e^2 = \mathbb{E}\|e_0\|^2$ and $Z_t^e = \|e_t\|^2 + V_t^e - 2\sigma_e^2$. Whenever the dependence norms are finite, a refined combined-energy allowance is

$$\mathcal{Q}_p^e(T,u) = 2\sigma_e^2 + \mathfrak{b}_{p/2}(Z^e;T,u). \tag{30}$$

For ordinary energy use $\mathcal{Q}_{p,\mathrm{raw}}^e(T,u) = \sigma_e^2 + \mathfrak{b}_{p/2}(\|e\|^2 - \sigma_e^2;T,u)$. Furthermore, for $s > 1$,

$$\mathcal{D}_{s,b}(Z^e) \leq 4\|e_0\|_{2s}\mathcal{D}_{2s,b}(e), \qquad \mathcal{D}_{s,b}(\|e\|^2 - \sigma_e^2) \leq 2\|e_0\|_{2s}\mathcal{D}_{2s,b}(e). \tag{31}$$

Proof. For $q = p/2$, Minkowski and conditional Jensen give $\|T^{-1}\sum \|e_t\|^2\|_q \leq \nu_p^2$ and $\|T^{-1}\sum(\|e_t\|^2 + V_t^e)\|_q \leq 2\nu_p^2$. Markov proves (29). For coupled vectors, Hölder gives

$$\big\|\|e_k\|^2 - \|e_k^*\|^2\big\|_s \leq (\|e_k\|_{2s} + \|e_k^*\|_{2s})\|e_k - e_k^*\|_{2s} = 2\|e_0\|_{2s}\delta_{2s}^e(k).$$

For $k \geq 1$, the analogous bound for V_k^e follows by coupling inside the conditional expectation and conditional Jensen. At $k = 0$, V_0^e is measurable before innovation zero and its coupling difference vanishes. Sum the bounds to obtain (31). Apply (28) to obtain the refined allowances. □

C.2 A MAXIMAL SELF-NORMALIZED INEQUALITY

For scalar square-integrable martingale differences X_t, define $M_k = \sum_{t\leq k} X_t$, $[M]_k = \sum_{t\leq k} X_t^2$ and $\langle M\rangle_k = \sum_{t\leq k} \mathbb{E}_{t-1} X_t^2$. The elementary inequality $e^{z-z^2/2} \leq 1 + z + z^2/2$ yields, for any deterministic $\lambda > 0$, the nonnegative supermartingale

$$\exp\{\lambda M_k - \tfrac{1}{2}\lambda^2([M]_k + \langle M\rangle_k)\}.$$

One proof of the elementary inequality is to differentiate $\log(1+z+z^2/2) - z + z^2/2$: its derivative is $z(z^2+z+2)/(z^2+2z+2)$, which has the sign of z. Its minimum is therefore its value zero at $z = 0$. We may also use the standard self-normalized construction of Bercu & Touati (2008) directly; see also the general framework in de la Peña et al. (2004); Howard et al. (2020). Ville's inequality, followed by $\lambda = u/v$ for $u, v > 0$, gives

$$\Pr\{M_k > u,\ [M]_k + \langle M\rangle_k \leq v \text{ for some } k \leq T\} \leq e^{-u^2/(2v)}. \tag{32}$$

Only finite second moments are required. This is a probability bound on an intersection, not a conditional probability given a future energy event.

Lemma C.2 (Objective-level localization)**.** *Let v_t be predictable and e_t be square-integrable vector martingale differences. Suppose for every prefix $k \le T$,*

$$F(v_{k+1}) - F_\star + \frac{\alpha}{4}\sum_{t\le k}\|\nabla F(x_t)\|^2 \le F(v_1) - F_\star + \alpha M_k + \alpha C_k, \quad M_k = -\sum_{t\le k}\langle \nabla F(v_t), e_t\rangle.$$

On an event $\mathcal{A}$, suppose deterministic $d_0, B, \mathcal{E} \ge 0$ obey $F(v_1) - F_\star \le d_0$, $\max_k C_k \le TB$ and $\sum_t(\|e_t\|^2 + V_t^e) \le T\mathcal{E}$. For $u \in (0,1)$, let $J = \sqrt{2T\mathcal{E}\log(2/u)}$ and $H = 4(d_0 + L\alpha^2 J^2 + \alpha TB) > 0$. Except with probability $\Pr(\mathcal{A}^c) + u$,

$$\max_{t\le T+1}\{F(v_t) - F_\star\} \le H, \qquad \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(x_t)\|^2 \le \frac{4d_0}{\alpha T} + \frac{4\sqrt{2LH}J}{T} + 4B. \tag{33}$$

Proof. Use the predictable stopping indicator $I_t = \mathbf{1}_{\{\max_{s\le t}(F(v_s)-F_\star)\le H\}}$. By (18), the stopped transform has coefficients bounded by $\sqrt{2LH}$. Thus square integrability of e_t makes the stopped transform square integrable. On $\mathcal{A}$ its combined quadratic variation at every prefix is at most $2LHT\mathcal{E}$. Equation (32) therefore bounds its maximum by $\sqrt{2LH}J$ except on a subset of $\mathcal{A}$ of probability at most $u/2 \le u$. At a first exit, the original and stopped transforms agree through the exit-producing step. Young's inequality gives the same contradiction $F(v_{s+1}) - F_\star \le 3H/4$ as in Appendix A. Finally retain the gradient term at $k = T$. □

C.3 Proof of the unconditional martingale result

Since Jensen also gives $\sum_{t\le k}\|n_{t-1}\|^2 \le \sum_{t\le k}\|\xi_t\|^2$, Lemma A.1 implies

$$F(y_{k+1}) - F_\star + \frac{\alpha}{4}\sum_{t\le k}\|\nabla F(x_t)\|^2 \le \Delta + \alpha M_k + \alpha K\sum_{t\le k}\|\xi_t\|^2, \quad K = L\alpha + L^2\alpha^2\kappa^2.$$

Apply Lemma C.2 on the combined-energy event with $d_0 = \Delta$, $B = K\mathcal{E}$ and $u = \delta/2$ to prove the following stronger trajectory and stationarity statements. With $J = \sqrt{2T\mathcal{E}\log(4/\delta)}$ and $U = 4(\Delta + L\alpha^2J^2 + \alpha KT\mathcal{E})$,

$$\max_{t\le T+1}\{F(y_t) - F_\star\} \le U, \qquad A_T \le \frac{4\Delta}{\alpha T} + \frac{4\sqrt{2LU}J}{T} + 4K\mathcal{E}. \tag{34}$$

If U vanishes, then on that energy event the gradient and noise vanish, and induction gives the same conclusion directly. With $L_\beta = L[1 + \beta^2/(8(1-\beta))]$, we have $K \le L_\beta\alpha$. Expanding the square root in (33) and using $2J\sqrt{T\mathcal{E}} \le J^2 + T\mathcal{E}$ gives

$$A_T \le \frac{4\Delta}{\alpha T} + 16\sqrt{\frac{L\Delta\mathcal{E}\log(4/\delta)}{T}} + 60L_\beta\alpha\mathcal{E}\log(4/\delta).$$

If $\sup_t \mathbb{E}\|\xi_t\|^2 \le \nu_2^2$, the expectation of the combined energy is at most $2T\nu_2^2$. Markov gives (6) with $\mathcal{E} = 4\nu_2^2/\delta$; no higher moment enters the preceding localization argument. For fixed β and confidence, this allowance, or either $p > 2$ allowance above, is $O(1)$, proving the $T^{-1/2}$ rate at $\alpha = a_0/\sqrt{T}$. For refined energy and $\delta = T^{-r}$, its polynomial fluctuation is $O(T^{2(1+r)/p-1})$, bounded for $r \le p/2 - 1$; the Gaussian fluctuation, when available, is $O(\sqrt{\log T/T})$. This proves the stated sufficient shrinking-confidence range.

D State-dependent stochastic-gradient applications

Proof of Corollary 3.2. Since $H(x_t)$ is predictable and e_t is an MDS, $\mathbb{E}_{t-1}[H(x_t)e_t] = 0$. The uniform operator norm bound yields pathwise

$$\|H(x_t)e_t\|^2 + \mathbb{E}_{t-1}\|H(x_t)e_t\|^2 \le h^2(\|e_t\|^2 + \mathbb{E}_{t-1}\|e_t\|^2).$$

Thus an energy event for stationary e supplies an event for the generally nonstationary noise $\xi_t = H(x_t)e_t$. Apply Theorem 3.1. Stationarity of ξ is unnecessary. □

In the learning experiment, $q_i(\theta) = \mathrm{sigmoid}(a_i^\top \theta)$ and $f_i(\theta) = \frac{1}{2}(q_i(\theta) - y_i)^2 + \frac{\lambda}{2}\|\theta\|^2$, with $\|a_i\| \le 1$, $y_i \in \{0, 1\}$. Write $F = n^{-1}\sum_i f_i$ and $\phi(\theta) = \frac{3}{4}v^\top\theta + \frac{1}{4}\sin(v^\top\theta)$, $\|v\| = 1$. Then $\nabla\phi(\theta) = (3/4 + \cos(v^\top\theta)/4)v$ has norm between $1/2$ and one. The oracle is the exact gradient of the random loss

$$f_{I_t}(\theta) + se_t\phi(\theta), \qquad e_t = \varepsilon_t\varepsilon_{t-1}, \quad s \ge 0.$$

The index I_t is uniform and independent across time and of all innovations. In the enlarged filtration containing past indices and innovations, the oracle is conditionally unbiased. Each data gradient without its ridge term has norm at most $1/4$; the sampling fluctuation $\zeta_t = \nabla f_{I_t}(\theta_t) - \nabla F(\theta_t)$ therefore satisfies $\|\zeta_t\| \le K = 1/2$. Its ridge term cancels. Consequently

$$\|\xi_t\|^2 + \mathbb{E}_{t-1}\|\xi_t\|^2 \le 4K^2 + 2s^2(e_t^2 + \mathbb{E}_{t-1}e_t^2). \tag{35}$$

Theorem 3.1 applies with $\mathcal{E} = 4K^2 + 2s^2\mathcal{Q}_p^e(T, \delta/2)$. More precisely, independence and centering give the conditional variance

$$V_t = \frac{1}{n}\sum_i \|\nabla f_i(\theta_t) - \nabla F(\theta_t)\|^2 + s^2\big(3/4 + \cos(v^\top\theta_t)/4\big)^2\varepsilon_{t-1}^2.$$

Its second term is at least $s^2\varepsilon_{t-1}^2/4$, so unbounded innovation support with $s > 0$ rules out a deterministic variance envelope even along the adaptive trajectory.

For the objective assumptions, $F \ge 0$ and its Hessian is

$$\nabla^2 F(\theta) = \lambda I + \frac{1}{n}\sum_i \left\{q_i^2(1 - q_i)^2 + (q_i - y_i)q_i(1 - q_i)(1 - 2q_i)\right\} a_i a_i^\top.$$

Let $A_{\mathrm{feat}} \in \mathbb{R}^{n\times d}$ have rows $a_i^\top$. Using $q(1-q) \le 1/4$ and $\sup_{q\in[0,1]} q(1-q)|1-2q| = 1/(6\sqrt{3})$ gives the valid global bound

$$L \le L_{\mathrm{valid}} = \lambda + \left(\frac{1}{16} + \frac{1}{6\sqrt{3}}\right)\frac{\|A_{\mathrm{feat}}\|_{\mathrm{op}}^2}{n}. \tag{36}$$

The finite dataset's negative-curvature witness and numerical derivative checks are reported in Appendix G. Sample losses need not share a deterministic smoothness constant after the volatility term is added; only F requires global smoothness in our theorem.

E Causal noise and the martingale–coboundary correction

E.1 Construction under square-summable coupling tails

Proof of Proposition 4.1. Write $\mathcal{P}_t Z = \mathbb{E}(Z \mid \mathcal{F}_t) - \mathbb{E}(Z \mid \mathcal{F}_{t-1})$. For centered causal Bernoulli shifts, approximation by functions of finitely many innovations implies $\mathbb{E}(Z \mid \mathcal{F}_{-m}) \to 0$ in L^p. Consequently the past projection expansion converges in L^p. Innovation coupling and conditional Jensen imply $\|\mathcal{P}_t\xi_{t+j}\|_p \le \delta_p^\xi(j)$. Thus $D_t = \sum_{j\ge 0}\mathcal{P}_t\xi_{t+j}$ converges absolutely in L^p, is a martingale difference, and $\|D_t\|_p \le \Theta_{0,p}^\xi$.

For completeness, the Hilbert martingale moment inequality gives

$$\Big\|\sum_k Y_k\Big\|_p \le C_p\Big(\sum_k \|Y_k\|_p^2\Big)^{1/2}, \qquad p \ge 2, \tag{37}$$

for finite Hilbert-valued martingale differences. This is the usual Burkholder square-function inequality followed by Minkowski in $L^{p/2}$; the constant does not depend on dimension (Pinelis, 1994). We combine this inequality with the projection method of Wu (2007).

Define the finite future sum $R_t^{(N)} = \sum_{j=1}^N \mathbb{E}(\xi_{t+j} \mid \mathcal{F}_t)$. For fixed N,

$$R_t^{(N)} = \sum_{k\ge 0}\sum_{j=1}^N \mathcal{P}_{t-k}\xi_{t+j}.$$

To justify this expression, first truncate k. The summands, ordered by increasing filtration time, are martingale differences with norms at most $\sum_{j=1}^{N} \delta_p^{\xi}(k+j) \le \Theta_{k+1,p}^{\xi}$. Equation (37) and the assumed square summability justify the limit of these finite past sums. The finite number of future variables allows their projection expansions to identify the limit with $R_t^{(N)}$. For $N > M$ the same argument yields

$$\|R_t^{(N)} - R_t^{(M)}\|_p \le C_p \left\{ \sum_{k\ge 0} \left(\sum_{j=M+1}^{N} \delta_p^{\xi}(k+j) \right)^2 \right\}^{1/2} \le C_p \left\{ \sum_{r\ge M+1} (\Theta_{r,p}^{\xi})^2 \right\}^{1/2} \longrightarrow 0.$$

Hence $R_t^{(N)} \to R_t$ in L^p, with $\|R_t\|_p \le C_p \mathfrak{R}_p$. The limit is a stationary causal measurable function of the innovation past, obtained by shifting the same construction at every time. The square-tail estimate establishes the required L^p limit directly.

For finite N, direct cancellation gives

$$\xi_t + R_t^{(N)} - R_{t-1}^{(N)} = \sum_{j=0}^{N-1} \mathcal{P}_t \xi_{t+j} + \mathbb{E}(\xi_{t+N} \mid \mathcal{F}_t).$$

The last term has norm at most $\Theta_{N,p}^{\xi} \to 0$. Taking L^p limits proves $\xi_t = D_t + R_{t-1} - R_t$. □

E.2 Sharpness for polynomial linear filters

The classical L^2 criterion characterizes martingale–coboundary decompositions through square-summable projection tails (Volný, 2017, Theorem 2). For the following filter, it reduces to square summability of the tail coefficients b_j. We give a direct proof to make the boundary explicit.

Proposition E.1 (Polynomial linear filters)**.** *Let ε_t be centered iid scalar innovations with variance one and finite pth moment, $p > 2$. For $a > 1$, let $\xi_t = \sum_{j\ge 0}(j+1)^{-a}\varepsilon_{t-j}$, and write $A = \sum_{j\ge 0}(j+1)^{-a}$ and $b_j = \sum_{k>j}(k+1)^{-a}$. If $a > 3/2$, a causal decomposition is*

$$D_t = A\varepsilon_t, \qquad R_t = \sum_{j\ge 0} b_j \varepsilon_{t-j}, \qquad \mathbb{E}R_t^2 = \sum_{j\ge 0} b_j^2 < \infty.$$

For $1 < a \le 3/2$, there is no decomposition $\xi_t = D_t + R_{t-1} - R_t$ with adapted square-integrable D, R and $\mathbb{E}(D_t \mid \mathcal{F}_{t-1}) = 0$, relative to the innovation filtration, such that (D, R, ε) is jointly stationary.

Proof. Here $\delta_p^{\xi}(j) = (j+1)^{-a}\|\varepsilon_0 - \varepsilon_0'\|_p$. Integral bounds give $\Theta_{m,p}^{\xi} \asymp m^{1-a}$ and $b_j \asymp (j+1)^{1-a}$. Thus the square-tail condition holds exactly when $a > 3/2$. Equation (37) gives L^p convergence of the displayed R, while orthogonality gives its variance. Substituting $b_{j-1} - b_j = (j+1)^{-a}$ verifies the identity.

For necessity, suppose such a decomposition exists. Put $d = \mathbb{E}(D_t\varepsilon_t)$ and $c_j = \mathbb{E}(R_t\varepsilon_{t-j})$; stationarity makes these independent of t. Adaptedness gives $\mathbb{E}(R_{t-1}\varepsilon_t) = 0$, and the MDS property gives $\mathbb{E}(D_t\varepsilon_{t-j}) = 0$ for $j \ge 1$. Taking inner products with the orthonormal innovation sequence therefore yields

$$c_0 = d - 1, \qquad c_j = c_{j-1} - (j+1)^{-a} = d - \sum_{k=0}^{j} (k+1)^{-a}.$$

Bessel's inequality requires $\sum_j c_j^2 \le \mathbb{E}R_t^2 < \infty$. In particular $c_j \to 0$, forcing $d = A$ and $c_j = b_j$. But $\sum_j b_j^2$ diverges for $1 < a \le 3/2$, a contradiction. □

This boundary concerns stationary adapted L^2 decompositions in the specified linear subclass. The coefficientwise sufficient condition $\sum k\delta_p^\xi(k) < \infty$ requires $a > 2$ here. In contrast, the actual future conditional expectations satisfy

$$\|\mathbb{E}(\xi_{t+j} \mid \mathcal{F}_t)\|_p \le C_p\|\varepsilon_0\|_p \left(\sum_{k\ge j}(k+1)^{-2a}\right)^{1/2} = O(j^{1/2-a}).$$

Their norms are summable for $a > 3/2$, including $3/2 < a \le 2$. Thus failure of the coefficientwise condition need not imply failure of absolute convergence of the future series.

E.3 Corrected descent and the unconditional causal bound

Let $w_t = y_t - \alpha R_{t-1}$, which is $\mathcal{F}_{t-1}$-measurable. The decomposition gives exactly

$$w_{t+1} = w_t - \alpha(\nabla F(x_t) + D_t). \tag{38}$$

The argument in Lemma A.1, now using

$$\|\nabla F(w_t) - \nabla F(x_t)\|^2 \le 2L^2\alpha^2\kappa^2\|m_{t-1}\|^2 + 2L^2\alpha^2\|R_{t-1}\|^2,$$

and $\sum_{t\le k}\|m_{t-1}\|^2 \le 2\sum_{t\le k}(\|\nabla F(x_t)\|^2 + \|\xi_t\|^2)$, proves

$$F(w_{k+1}) - F_\star + \frac{\alpha}{4}\sum_{t\le k}\|\nabla F(x_t)\|^2 \le F(w_1) - F_\star + \alpha M_k^D + \alpha C_k, \tag{39}$$

where

$$M_k^D = -\sum_{t\le k}\langle \nabla F(w_t), D_t\rangle, \qquad C_k = L\alpha\sum_{t\le k}\|D_t\|^2 + 2L^2\alpha^2\kappa^2\sum_{t\le k}\|\xi_t\|^2 + L^2\alpha^2\sum_{t\le k}\|R_{t-1}\|^2.$$

The absorbed gradient coefficient is at least $\alpha/4$, since $1/2 - L\alpha - 2L^2\alpha^2\kappa^2 \ge 1/4$.

Proof of Theorem 4.2. Markov's inequality gives $\|R_0\| \le r_0 = \rho_p\epsilon^{-1/p}$ except with probability ϵ. On this event, smoothness and (18) give

$$F(w_1) - F_\star \le \Delta + \alpha\sqrt{2L\Delta}r_0 + \tfrac{1}{2}L\alpha^2 r_0^2 = \Delta_\alpha.$$

The combined D-energy event fails with probability at most ϵ. The two ordinary energy events for ξ, R each fail with probability at most ϵ. On their intersection, $C_k \le TB$ for all k. Apply Lemma C.2 with $v_t = w_t$, $e_t = D_t$, $d_0 = \Delta_\alpha$ and $\mathcal{E} = \mathcal{E}_D$. The additional self-normalized failure event has probability at most ϵ, so the total is at most $5\epsilon = \delta$. In particular, with $J_D = \sqrt{2T\mathcal{E}_D\ell_D}$ and $W = 4(\Delta_\alpha + L\alpha^2 J_D^2 + \alpha TB)$, we obtain

$$\max_{t\le T+1}\{F(w_t) - F_\star\} \le W, \qquad A_T \le \frac{4\Delta_\alpha}{\alpha T} + \frac{4\sqrt{2LW}J_D}{T} + 4B. \tag{40}$$

Expanding the square root and applying $2\sqrt{uv} \le u + v$ gives

$$\frac{4\sqrt{2LW}J_D}{T} \le 16\sqrt{\frac{L\Delta_\alpha\mathcal{E}_D\ell_D}{T}} + (16\sqrt{2}+8)L\alpha\mathcal{E}_D\ell_D + 8B.$$

Since $16\sqrt{2} + 8 < 32$, this proves (14). When $W = 0$, the zero-noise trajectory yields the same conclusion directly. For fixed confidence and beta, $\alpha = a_0T^{-1/2}$ gives bounded energy allowances and $B = O(T^{-1/2})$, $J_D = O(\sqrt{T})$, $W = O(1)$. The rate follows. □

E.4 Refined energies require stronger dependence control

Coupling finite future sums through conditional expectations and Jensen, then taking their L^p limits, gives

$$\delta_p^R(k) \le \Theta_{k+1,p}^\xi \quad (k \ge 0), \qquad \delta_p^D(k) \le 2\Theta_{k,p}^\xi \quad (k \ge 1), \qquad \delta_p^D(0) \le 2\Theta_{0,p}^\xi.$$

The stronger condition $\sum_{k\ge0}(k+1)^{b+1}\delta_p^{\xi}(k)<\infty$ controls the dependence tails needed for the energy bounds. For example,

$$(m+1)^b\sum_{k\ge m}\delta_p^D(k)\le 2(m+1)^b\sum_{j\ge m}(j-m+1)\delta_p^{\xi}(j)\le 2\sum_{j\ge m}(j+1)^{b+1}\delta_p^{\xi}(j),\quad m\ge 1.$$

The $m=0$ term and the R tails obey analogous bounds. Together with Lemma C.1, these estimates justify the three refined energy allowances in Corollary 4.3. The same four energy and initialization events and one martingale event prove the conclusion. At $\delta=T^{-r}$ with $0<r\le p/2-1$, the refined energies remain $O(1)$. Moreover, $\alpha\delta^{-1/p}=O(T^{-1/p})$ under this restriction, so $\Delta_\alpha=O(1)$ and the stationarity bound is $O(\log T/\sqrt{T})$.

E.5 Product volatility and linear filters

For $u_t=\varepsilon_t\varepsilon_{t-1}$ with centered iid innovations of variance one, $\mathbb{E}_{t-1}u_t=0$ and $\mathbb{E}_{t-1}u_t^2=\varepsilon_{t-1}^2$. The physical-dependence coefficients vanish after lag one and

$$\delta_p^u(0)=\delta_p^u(1)=\|\varepsilon_0\|_p\|\varepsilon_0-\varepsilon_0'\|_p.$$

For $\xi_t=c\sum_{j\ge0}\rho^j u_{t-j}$, $|\rho|<1$, $\mathbb{E}(\xi_{t+j}\mid\mathcal{F}_t)=\rho^j\xi_t$ and $\mathcal{P}_t\xi_{t+j}=c\rho^j u_t$. Therefore

$$D_t=\frac{c}{1-\rho}u_t,\qquad R_t=\frac{\rho}{1-\rho}\xi_t,\quad \mathrm{Var}(\xi_t)=\frac{c^2}{1-\rho^2},\quad \mathrm{Var}(D_t)=\frac{c^2}{(1-\rho)^2}.$$

The dependence coefficients decay geometrically. If the innovation law has unbounded support, the predictable variance of D_t has no finite deterministic envelope. Setting $c=\sigma_\xi\sqrt{1-\rho^2}$ gives the long-run variance $\sigma_\xi^2(1+\rho)/(1-\rho)$. For a vector linear process $\xi_t=\sum_{j\ge0}C_j\varepsilon_{t-j}$, the corresponding formulas are $D_t=(\sum_j C_j)\varepsilon_t$ and $R_t=\sum_{j\ge0}(\sum_{k>j}C_k)\varepsilon_{t-j}$ whenever the stated series converge. Their covariance interpretation is the same.

F Exact quadratic and filter calculations

For $F(x)=\lambda x^2/2$, eliminating m gives, for $t\ge2$,

$$x_{t+1}=(1+\beta-\alpha\lambda(1-\beta))x_t-\beta x_{t-1}-\alpha(1-\beta)\xi_t.$$

The second-order Jury stability conditions reduce to $0<\alpha\lambda(1-\beta)<2(1+\beta)$ when $0\le\beta<1$. Write $a_\beta=1+\beta-\alpha\lambda(1-\beta)$ and let $\gamma_0=\mathrm{Var}(x_t)$, $\gamma_1=\mathrm{Cov}(x_t,x_{t-1})$ under iid noise. The Yule–Walker equations imply

$$\gamma_1=\frac{a_\beta}{1+\beta}\gamma_0,\qquad \gamma_0=(a_\beta^2+\beta^2)\gamma_0-2a_\beta\beta\gamma_1+\alpha^2(1-\beta)^2\sigma^2.$$

Solving these equations gives

$$\mathrm{Var}(x)=\frac{\alpha\sigma^2(1+\beta)}{\lambda\{2(1+\beta)-\alpha\lambda(1-\beta)\}}. \tag{41}$$

At small fixed effective stepsize, this is $\alpha\sigma^2/(2\lambda)+O(\alpha^2)$.

For stationary AR(1) noise with $\mathrm{Cov}(\xi_t,\xi_{t-k})=\sigma_\xi^2\rho^{|k|}$, the exponential average $n_t^\infty=(1-\beta)\sum_{j\ge0}\beta^j\xi_{t-j}$ satisfies

$$\mathrm{Var}(n_t^\infty)=(1-\beta)^2\sigma_\xi^2\sum_{i,j\ge0}\beta^{i+j}\rho^{|i-j|}=\sigma_\xi^2\frac{1-\beta}{1+\beta}\frac{1+\beta\rho}{1-\beta\rho}.$$

The filter weights sum to one and both weight and covariance sequences are absolutely summable. Interchanging their sums proves the long-run covariance identity in (16).

For the AR-noise quadratic simulations with $\lambda=1$, the state $s_t=(x_t,m_{t-1},\xi_{t-1})^\top$ obeys $s_{t+1}=As_t+b\varepsilon_t$, where, for unit marginal noise variance and $o=1-\beta$,

$$A=\begin{pmatrix}1-\alpha o & -\alpha\beta & -\alpha o\rho\\ o & \beta & o\rho\\ 0 & 0 & \rho\end{pmatrix},\qquad b=\sqrt{1-\rho^2}\begin{pmatrix}-\alpha o\\ o\\ 1\end{pmatrix}.$$

The exact covariance P solves $P = APA^\top + bb^\top$. From a zero initial state, its covariance after T updates is $P - A^T P (A^\top)^T$. The first coordinate of this state is x_{T+1}. We standardize both Gaussian and Student errors by $\sqrt{P_{11}}$, using only the innovations' variance.

For the example in Section 5, let $\lambda = 1$, $\alpha = .03$ and $\sigma_\xi^2 = 1$. Increasing β from 0 to .9 changes the exact iterate variance from .015228 to .015012 for $\rho = 0$. For $\rho = .8$, it increases from .120740 to .129522, a 7.27% change. These are stationary quadratic calculations within the exact stability region, with the effective step held fixed.

More generally, with $x_1 = m_0 = 0$, for a centered linear process $\xi_t = \sum_{j\ge 0} c_j \varepsilon_{t-j}$, the impulse coefficients from Appendix B give

$$x_{T+1} = -\alpha(1-\beta)\sum_{t=1}^{T} h_{T-t}\xi_t = -\sum_{k\le T} b_{T,k}\varepsilon_k, \qquad b_{T,k} = \alpha(1-\beta)\sum_{t=\max\{1,k\}}^{T} h_{T-t}c_{t-k}.$$

With centered iid innovations in L^p, $p > 2$, and $\sum_j |c_j| < \infty$, this equality holds in L^p. The scalar Fuk–Nagaev bound yields, for $u > 0$ and $\sigma_\varepsilon^2 = \mathbb{E}\varepsilon_0^2$, with constants depending only on p,

$$\Pr(|x_{T+1}| > u) \le 2\exp\left\{-\frac{u^2}{C_p\sigma_\varepsilon^2\sum_k b_{T,k}^2}\right\} + \frac{C_p\mathbb{E}|\varepsilon_0|^p\sum_k |b_{T,k}|^p}{u^p}.$$

For an infinite innovation past, first truncate, apply the finite-sum bound, and pass to the L^p limit. This identity calibrates terminal error in the linear model.

Proposition F.1 (Exact finite-horizon impulse energy)**.** *Let $F(x) = \lambda x^2/2$, $\lambda > 0$, $x_1 = m_0 = 0$, $\alpha > 0$ and $0 \le \beta < 1$. Write $a = \alpha\lambda$, $o = 1-\beta$, $h_0 = 1$, $h_1 = 1+\beta-ao$ and $h_j = (1+\beta-ao)h_{j-1} - \beta h_{j-2}$ for $j \ge 2$. If the only nonzero innovation is $\xi_s = v$, $1 \le s \le T$, then*

$$A_T = \frac{\lambda^2\alpha^2 o^2 v^2}{T}\sum_{j=0}^{T-s-1} h_j^2. \tag{42}$$

An empty sum is zero. In particular, for $p > 2$, $\nu > 0$, $\delta \in (0,1)$ and $|v| = \nu(T/(4\delta))^{1/p}$,

$$\frac{A_T}{\lambda\alpha\nu^2 T^{2/p-1}\delta^{-2/p}} = 4^{-2/p}ao^2\sum_{j=0}^{T-s-1} h_j^2.$$

In the stable region $0 < ao < 2(1+\beta)$, the infinite-response reference is

$$4^{-2/p}ao^2\sum_{j\ge 0} h_j^2 = 4^{-2/p}\frac{1+\beta}{2(1+\beta)-ao}. \tag{43}$$

For fixed β, this reference tends to $1/(2\,4^{2/p})$ as $a \to 0$.

Proof. Before the shock $x_t = 0$. The recurrence gives $x_{s+1+j} = -\alpha o v h_j$, so summing the squared gradients at $x_1, \dots, x_T$ proves (42); the shock at time T does not enter this statistic. Stability makes the impulse square summable. For unit-variance iid noise, its stationary output variance is $\alpha^2 o^2 \sum_j h_j^2$. Substituting (41) and using $a = \alpha\lambda$ proves (43). The limit follows directly. □

For finite $T - s$, the difference from (43) is exactly $4^{-2/p}ao^2\sum_{j=T-s}^{\infty} h_j^2$. Thus the infinite-response reference is an upper comparison, not an identity for a finite run. At $p = 2.5$ its small-step limit is 0.16493849. The code evaluates the finite sum for every early shock position in the displayed grid and validates it against the literal optimizer and an independent Lyapunov calculation.

Random output iterate. If τ is uniform on $\{1, \dots, T\}$ and independent of the run, $\mathbb{E}(\|\nabla F(x_\tau)\|^2 \mid x_1, \dots, x_T) = A_T$. For $u \in (0,1)$ and $B > 0$, if $\Pr(A_T > B) \le \delta$, conditional Markov gives $\Pr(\|\nabla F(x_\tau)\|^2 > B/u) \le \delta + u$. The extra output-randomization cost is explicit; no last-iterate guarantee is inferred from average stationarity.

G EXPERIMENTAL PROTOCOLS AND ADDITIONAL RESULTS

G.1 IMPLEMENTATION, RANDOM STREAMS AND REPORTING

The source archive contains `rare_shocks.py`, `scalar_suite.py`, `causal_suite.py`, `learning.py` and `make_figures.py`, with shared utilities in `common.py`. All paths are relative to the source directory. Each configuration stores its horizon, replication count, initialization, step convention, law, dependence parameters and seed keys. Random generators use NumPy `SeedSequence([20260912, *keys])`; chunked runs append a batch index. The chunk size is part of the saved protocol. Replications are independent; algorithms share innovations within a configuration. Pilot and final keys are disjoint. The learning dataset uses keys $[13, 0]$.

The recorded numerical environment is Python 3.12.14, NumPy 2.5.3, SciPy 1.18.1, Numba 0.67.0 and Matplotlib 3.11.2. Numba compiles numerical loops; the learning code executes independent batches with four threads. Every reported experiment saves compressed per-replication statistics in addition to JSON and CSV summaries. Figures and tables are generated from these records. Wall-clock times in the metadata describe this machine, not algorithmic complexity. Nonfinite optimization statistics are checked explicitly. Variance-normalized diagnostics are undefined for zero-noise controls and are labeled as such.

For a qth quantile from R observations, the confidence interval uses binomial ranks based on $\mathrm{Binomial}(R, q)$. If a finite sample cannot provide a finite endpoint at the requested confidence, the endpoint is reported as infinite. Tail event intervals use the exact Clopper–Pearson construction, including zero counts. These are pointwise Monte Carlo intervals, not simultaneous guarantees across the whole parameter grid. Fixed-law survival curves show no point with fewer than ten exceedances; the supplementary quadratic curve uses at least three, with exact intervals displayed at four standard deviations. No zero count is interpreted as a zero probability.

G.2 RARE-SHOCK EXPERIMENT

For each T, δ, draw $K \sim \mathrm{Binomial}(T, 4\delta/T)$. Conditional on K, choose K distinct positions uniformly and assign independent symmetric signs. This is exactly the iid three-point law in Theorem 2.2. The implementation includes zero, single and multiple shocks, and retains their complete impulse cross terms. It does not sample only the sufficient event used in the proof.

If two unit shocks occur at zero-based positions $s \le t$, their response inner product uses $\sum_{j=0}^{T-2-t} h_j h_{j+t-s}$. In the positive-root regime, for length n and lag d, this sum is

$$\frac{r_+^{d+2} G_n(r_+^2) + r_-^{d+2} G_n(r_-^2) - (r_+ r_-^{d+1} + r_- r_+^{d+1}) G_n(\beta)}{(r_+ - r_-)^2}, \qquad G_n(z) = \sum_{j=0}^{n-1} z^j.$$

The geometric sum is evaluated with `expm1` for stability. We compared this calculation against the literal momentum recursion on 956 paths, including early and last-step shocks, cancellation between adjacent shocks, multiple shocks and $\beta = .99$. The maximum relative discrepancy was 1.04×10^{-12} or less. Every stored one-early-shock path was also checked against the theorem's threshold.

The main grid uses $p = 2.5, 6$, $\nu = 1$, $T = 512, 2048, 8192, 65536$ and $\delta = .1, .03, .01, .003, .001$, with $\beta = 0, .5, .9$ and additionally $\beta = .99$ only at $T = 65536$. There are 100,000 runs per configuration. The laws have variance $(4\delta/T)^{1-2/p}$, which varies across the grid. The sparse draws are shared across momentum and p values within each T, δ configuration, with p changing only shock amplitude. The exceedance threshold uses the stronger combined constant in Appendix B. At $\beta = .9$, this threshold coefficient equals the displayed beta-uniform constant; the branch changes at $\beta = 1 - \sqrt{1/32} \simeq .8232$. For the single-early-shock event E in Figure 1, $P(E) = 2\delta(1-4\delta/T)^{T-1}$. At $p = 2.5$, $\beta = .9$, $T = 8192$ and $\delta = .01$, this is .01921588, between the theorem's probability guarantee .01 and the observed unconditional exceedance .03973. Shocks outside E can also exceed the threshold.

G.3 SCALAR BASELINE PROTOCOLS AND VOLATILITY DIAGNOSTICS

The bounded-gradient objective is $F(x) = 1 - \cos x + .05 \log\cosh x$, with $F'(x) = \sin x + .05 \tanh x$ and $L \le 1.05$. The unbounded-gradient objective is $F(x) = .05x^2 + 1 - \cos x$, with $L = 1.1$. Both

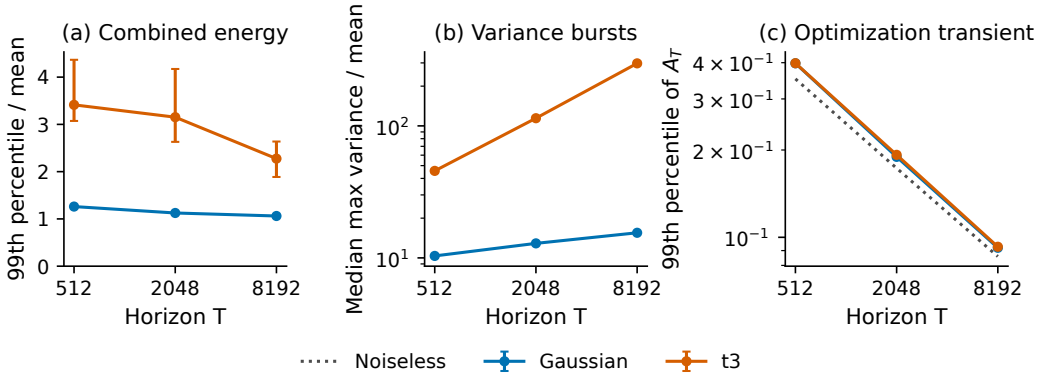

Figure 5: Product-volatility diagnostics for the unbounded-gradient objective, $\rho = 0$, $\beta = .9$, with 4,096 runs per configuration. Panels show the 99th percentile of normalized combined energy, median maximum conditional variance, and stationarity relative to the noiseless control. Error bars are 95% rank intervals.

are nonnegative and have negative curvature near π. Their baseline runs use $x_1 = 2.3$, $m_0 = 0$, $\alpha = .25/\sqrt{T}$, $\beta = 0, .9$ and $T = 512, 2048, 8192$, with 4,096 replications per configuration. Noiseless controls use the same initialization and update conventions.

The additive-noise baseline is $\xi_t = \rho\xi_{t-1} + .7\sqrt{1-\rho^2}\, e_t$, where $\rho = 0, .8$ and e_t is Gaussian, standardized t_6, or standardized t_3, all with variance one. The volatility baseline replaces e_t by $\varepsilon_t\varepsilon_{t-1}$ and uses Gaussian or standardized t_3 innovations. For $\rho = 0$ draw ε_0 from its stationary innovation law. For $\rho = .8$, the filter starts at zero and receives 512 burn-in steps, preserving the final innovation before optimization. The omitted stationary filter contribution has L^2 norm $.7(.8)^{512} < 2 \times 10^{-50}$, quantifying the stationary-initialization approximation.

For volatility, save $D_t = cu_t/(1-\rho)$, $V_t^D = c^2\varepsilon_{t-1}^2/(1-\rho)^2$, and the pathwise diagnostics

$$\frac{\sum_t V_t^D}{T\sigma_D^2}, \quad \frac{\sum_t (D_t^2 + V_t^D)}{2T\sigma_D^2}, \quad \frac{\max_t V_t^D}{\sigma_D^2}, \qquad \sigma_D^2 = \frac{c^2}{(1-\rho)^2}.$$

Figure 5 illustrates why a maximum-variance envelope is a different object from combined average energy. At $T = 8192$, $\rho = 0$, the normalized combined-energy 99th percentile is 1.062 for Gaussian products and 2.276 for t_3 products. The corresponding median normalized maximum variances are 15.50 and 297.63. The scalar stationarity 99th percentiles remain close, .09201 and .09277, because the initialization transient is substantial. These diagnostics compare accumulated energy with conditional-variance bursts; the theory specifies the admissible confidence rates.

The quadratic calibration uses $F(x) = x^2/2$, $\alpha = .03$, $T = 1024$, 50,000 runs, $\beta = 0, .9$ and $\rho = 0, .8$, with Gaussian or standardized t_3 innovations. All states start at zero. Table 2 compares simulated and exact stationary variances; finite-horizon variances and Lyapunov residuals are also saved. Figure 6 shows heavier terminal-error tails for t_3 at the same variance scale. This terminal metric is supplementary to the direct A_T experiments.

G.4 Fixed-law study

The fixed-law experiment uses the unbounded-gradient objective, iid noise of variance $.7^2$, $\alpha = .25/\sqrt{T}$, $\beta = 0, .9$ and $x_1 = 0, 2.3$. Each of $T = 512, 2048$ uses 10,000 runs; the prespecified key horizon $T = 8192$ uses 100,000 for each law and initialization, supporting the 99.9th-percentile comparison. Gaussian and standardized t_3 laws are held fixed across horizons. We use $p = 2.5$ for t_3; it does not have a finite third moment. For any discussion of the high-moment branch using t_6, a valid choice is $p = 5 < 6$.

At $T = 8192$, $\beta = .9$, the 99.9th percentiles for $x_1 = 2.3$ are .09380 (Gaussian) and .09439 (t_3), compared with the noiseless value .085906, a reference for the initialization transient. The fixed-law

Table 2: Quadratic calibration: 50,000 runs per row. Exact variances follow Appendix F. Four-SD exceedances are counted relative to the exact stationary standard deviation.

Innovations	ρ	β	Exact variance	Empirical variance	4-SD count
Gaussian	0	0	0.015228	0.015169	6
Gaussian	0	0.9	0.015012	0.015041	3
Gaussian	0.8	0	0.120740	0.120164	2
Gaussian	0.8	0.9	0.129522	0.129840	6
t3	0	0	0.015228	0.015040	82
t3	0	0.9	0.015012	0.014690	69
t3	0.8	0	0.120740	0.126028	84
t3	0.8	0.9	0.129522	0.131788	94

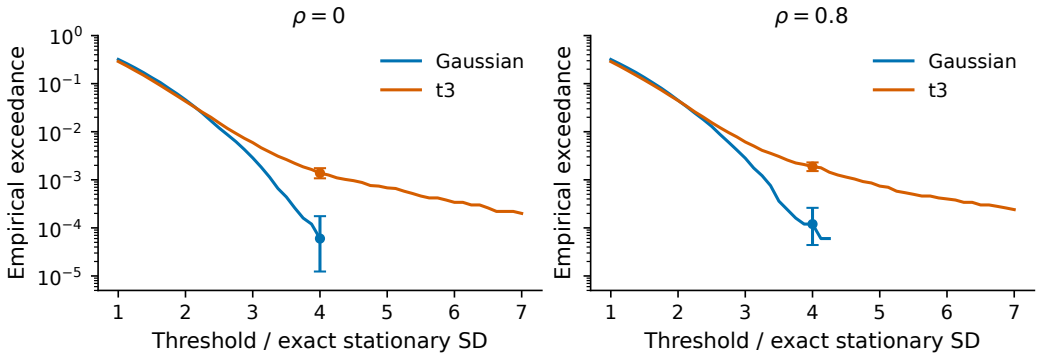

Figure 6: Terminal quadratic error, $T = 1024$, $\beta = .9$, 50,000 runs. Markers at four standard deviations show exact 95% binomial intervals. The stepsize .03 is justified by exact quadratic stability, not by the general nonconvex stepsize cap.

comparison measures quantiles rather than fitting a tail exponent. Figure 2 displays both initialization settings in the main text.

G.5 Causal optimization experiments and exact boundary diagnostics

The additional causal experiments use the unbounded-gradient objective, $x_1 = 0, 2.3$, $m_0 = 0$, $\alpha = .25/\sqrt{T}$, $\beta = 0, .9$ and $T = 512, 2048, 8192$. All configurations satisfy the common stepsize cap. Product-volatility runs use $\rho = 0, .8$, Gaussian or standardized t_3 innovations, 8,192 replications and batches of 512. The stationary initialization convention and 512-step burn-in are as above. Their seed keys are $[14, 0, l, r, x, t, b]$, indexing law, correlation, initialization, horizon and batch.

For slow-decay noise, use iid standard Gaussian innovations and

$$\xi_t = \frac{.7}{\sqrt{\zeta(2a)}} \sum_{j \geq 0} (j+1)^{-a} \varepsilon_{t-j}, \qquad a \in \{1.6, 1.8, 2.2\}.$$

There are 4,096 replications and batches of 128, with keys $[14, 1, a, x, t, b]$ using parameter indices. The noise variance is .49 in every setting; its long-run variance is $.49\zeta(a)^2/\zeta(2a)$, respectively 2.1942, 1.5555 and 1.0277. For $a = 1.6, 1.8$, the experiment tests the basic fixed-confidence theorem. The stronger sufficient condition for refining all three energies in Corollary 4.3 is a separate requirement. Figure 7 retains both momentum values and initializations for all three power exponents and the correlated product-volatility laws.

To sample the stationary infinite-filter model, compute its covariance

$$\gamma(k) = \frac{.49}{\zeta(2a)} \sum_{n=1}^{\infty} n^{-a} (n+k)^{-a}.$$

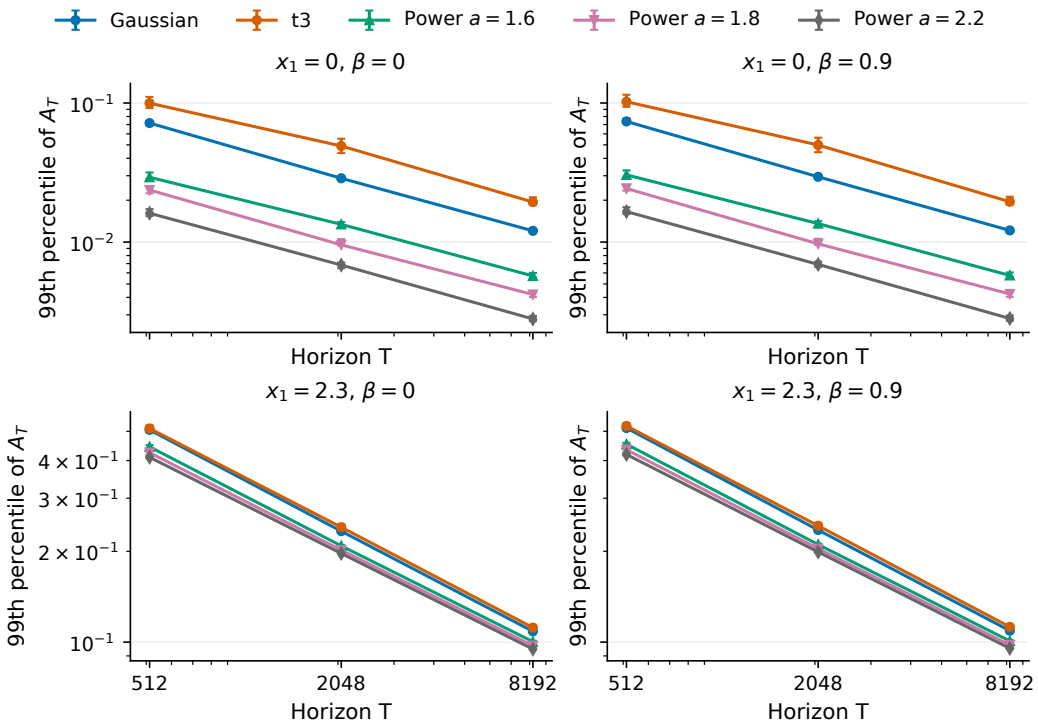

Figure 7: Causal-noise controls, $\alpha = .25/\sqrt{T}$. Gaussian and t_3 curves use product volatility with $\rho = .8$ (8,192 runs); power-filter curves use 4,096 runs. Rows vary initialization and columns vary momentum. All panels report the 99th percentile with 95% rank intervals; vertical scales are shared within each row.

The code sums $n < 256$ and evaluates the remaining infinite tail with four Euler–Maclaurin corrections. For $f_k(x) = x^{-a}(x+k)^{-a}$, the truncation remainder is bounded by $|B_8|\,|f_k^{(7)}(256)|/8!$, where B_8 is the eighth Bernoulli number. This bound is below 1.2×10^{-25} on the validated grid. Independent direct summation plus numerical quadrature agrees within 8×10^{-16}; this latter comparison also reflects floating-point error. A circulant embedding of the covariance has nonnegative eigenvalues for every configuration. Applying its Fourier square root to iid Gaussian vectors yields the required finite-horizon stationary covariance. No negative eigenvalues are clipped and no finite moving average replaces the infinite filter. A small-matrix covariance check agrees within 3×10^{-16}, and the optimizer kernel agrees with a literal recursion.

The analytic diagnostic in Figure 8 illustrates the pointwise EMA calculation and the square-tail boundary. For the unnormalized linear filter in Proposition E.1, $b_j = \zeta(a, j+2)$, where $\zeta(a, z)$ is the Hurwitz zeta function. The plotted quantity is exactly the variance of the first M terms of the coboundary series, computed for $M \le 10^6$. It converges for $a > 3/2$ and diverges for $1 < a \le 3/2$. This is the boundary of a stationary adapted L^2 decomposition in the stated subclass; it is distinct from an optimization impossibility claim.

G.6 Vector learning design and tuning

The dataset is generated once with seed keys $[13, 0]$. A Gaussian vector is normalized to $v \in \mathbb{R}^{20}$, latent scores z_i are Gaussian, and raw features are $.65 z_i v + .5\eta_i/\sqrt{20}$ with iid standard Gaussian η_i. Each feature is divided by the larger of its norm and one. Labels are $\mathbf{1}_{\{z_i > 0\}}$, with flips at six uniformly selected distinct examples among 64. All methods start at $\theta_1 = 0$. The ridge parameter is .001 and the noise scale is .7. At $\theta = -2.25v$, the smallest Hessian eigenvalue is $-.0151228$. The analytic gradient and Hessian agree with central differences to 6.01×10^{-12} and 4.33×10^{-13}, respectively. A literal implementation independently checks average stationarity, terminal loss,

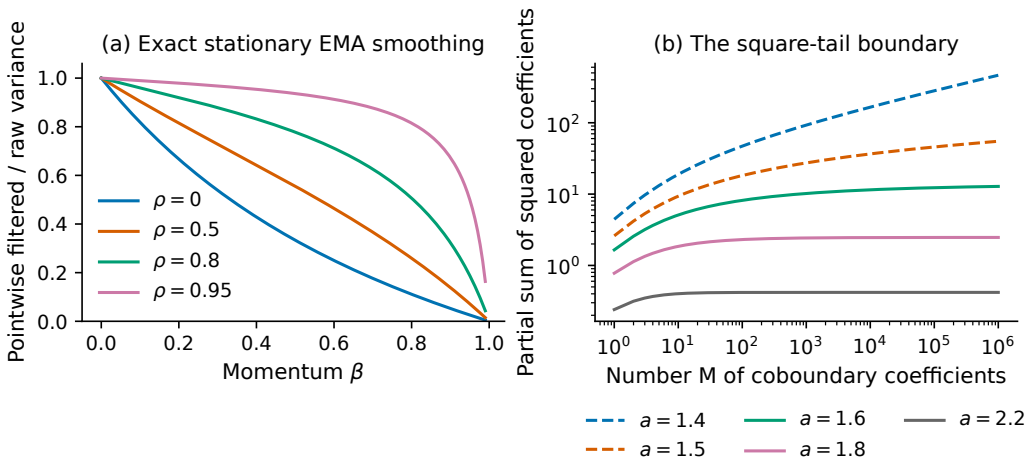

Figure 8: Analytic mechanism checks. Left: stationary EMA pointwise variance ratio from (16); the long-run variance ratio is one for every β. Right: partial sums of squared coboundary coefficients. Dashed curves have $a \leq 3/2$ and diverge; solid curves have $a > 3/2$. These are deterministic calculations, not Monte Carlo estimates.

conditional variance, combined energy and clipping updates to a maximum discrepancy below 9×10^{-16}.

For the common-α experiment, use $\alpha = a_{\text{safe}}/\sqrt{T}$, where $a_{\text{safe}} = .8(1 - .9)\sqrt{256}/(8L_{\text{valid}})$. This satisfies the strictest cap for every horizon and both momentum choices. The tuned comparison uses an independent pilot of 128 runs at $T = 1024$ for each innovation law, minimizing the empirical 90th percentile of A_T. Grid expansion used only pilot outcomes when a chosen parameter lay on a boundary; all choices were frozen before the final evaluation. The final grid has 36 stepsize coefficients geometrically spaced from .05 to 400 for each ordinary method. Clipped SGD has the same 36 candidate evaluations: six geometric stepsize coefficients over that interval crossed with thresholds $\{.0125, .05, .2, .8, 3.2, 12.8\}$. Earlier grids and their results remain in the archive; each expansion used the same number of evaluations for all three methods and reused only the pilot stream.

The final coefficients are 51.2773 for Gaussian SGD and SGDM, 51.2773 for t_3 SGD, and 85.6957 for t_3 SGDM; in each case α is the coefficient divided by $\sqrt{T}$. Clipped SGD selects coefficient 400 and threshold .0125 for both laws. This selection lies on the clipping-grid boundary and is reported as the best tested configuration. Its clipping update is implemented exactly as in Nguyen et al. (2023, Algorithm 1), with constant pilot-selected parameters. Final runs use disjoint seed keys $[13, 3, l, t]$, 1,024 replications at $T = 256, 1024$, and 10,000 at $T = 2048$.

Each method uses one sample-gradient call per update. Exact full-gradient diagnostics cost $64T$ sample equivalents per path for every method and are reported separately from the optimization budget. The saved statistics are A_T, terminal loss, combined energy, average and maximum conditional variance, average squared amplitude, training accuracy and maximum iterate norm. Figure 9 compares the tuned methods for both innovation laws. Clipping is worse at the smallest Gaussian horizon under parameters selected at $T = 1024$. The tuned SGDM steps exceed the general cap, so this comparison characterizes empirical behavior beyond that sufficient region. Clipped SGD is an empirical comparator: the conditional-moment assumptions of its cited theorem exclude this volatility oracle. At $T = 2048$ under t_3 products, the tuned 90th-percentile stationarity values are .000680, .000725 and .000137 for SGD, SGDM and clipped SGD. Tuned SGDM's lower stationarity coexists with higher median terminal loss (.1303 versus .0624 at the common step), illustrating the distinction between these criteria.

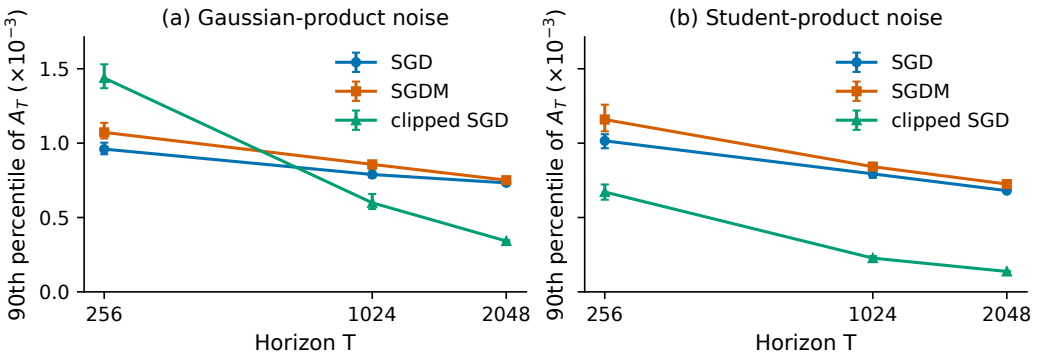

Figure 9: Independent-pilot tuning on the vector learning problem. Both innovation laws use the same zero-based vertical scale. Tuned SGDM exceeds the sufficient theorem cap. Parameters are frozen before final runs, with 1,024 replications at $T = 256, 1024$ and 10,000 at $T = 2048$. Error bars are 95% rank intervals. Clipping's favorable longer-horizon result coexists with its worse Gaussian result at $T = 256$.

Table 3: Vector learning at $T = 2048$, 10,000 runs per row. Stationarity columns and their confidence intervals are in units of 10^{-3}; loss is unscaled. These results retain the difference between small average gradient norms and low terminal loss.

Innovations	Method	$q_{.90}(A_T)$	$q_{.99}(A_T)$ [95% CI]	Median loss
Gaussian	SGD (fixed α)	1.108	1.944 [1.883, 2.048]	0.0623
Gaussian	SGDM (fixed α)	1.112	1.960 [1.884, 2.050]	0.0623
Gaussian	SGD (tuned)	0.733	0.970 [0.956, 0.994]	0.1013
Gaussian	SGDM (tuned)	0.751	1.018 [0.997, 1.036]	0.1038
Gaussian	clipped SGD (tuned)	0.342	0.600 [0.577, 0.625]	0.0593
t3	SGD (fixed α)	1.047	1.899 [1.830, 1.961]	0.0624
t3	SGDM (fixed α)	1.051	1.897 [1.834, 1.950]	0.0624
t3	SGD (tuned)	0.680	1.012 [0.972, 1.059]	0.0949
t3	SGDM (tuned)	0.725	1.302 [1.206, 1.424]	0.1303
t3	clipped SGD (tuned)	0.137	0.219 [0.211, 0.225]	0.0584

H IID SPECIALIZATION AND COMPARISON WITH RELATED LITERATURE

H.1 ADDITIVE IID NOISE AND FRESH-SAMPLE STOCHASTIC GRADIENTS

For centered additive iid noise, the conditional moments in (3) equal the marginal moments. Theorem 2.1 therefore applies with $\sigma^2 = \mathbb{E}\|\xi_1\|^2$ and $\nu_p^p = \mathbb{E}\|\xi_1\|^p$. More generally, take iid data Z_t, a measurable stochastic-gradient map G, and $g_t = G(x_t, Z_t)$, with Z_t independent of $\mathcal{F}_{t-1}$. If, uniformly in x,

$$\mathbb{E}_Z G(x, Z) = \nabla F(x), \quad \mathbb{E}_Z\|G(x, Z) - \nabla F(x)\|^2 \le \sigma^2, \quad \mathbb{E}_Z\|G(x, Z) - \nabla F(x)\|^p \le \nu_p^p,$$

then conditioning on $\mathcal{F}_{t-1}$ proves the same result. The errors can depend on the iterate; independence of the fresh data is sufficient for this conditional verification. For a fixed training sample S, the same argument can be applied conditionally on S to its empirical objective F_S and iid index draws. Population-risk or generalization conclusions require a separate link between F_S and the population objective.

H.2 EXPLICIT HORIZON AND CONFIDENCE DEPENDENCE

Choose $\alpha = a_0/\sqrt{T}$ with $a_0 > 0$ and $T \ge [8La_0/(1-\beta)]^2$. With probability at least $1 - \delta$,

$$A_T \le C_p \left[\frac{\Delta/a_0 + La_0\sigma^2 \log(6/\delta)}{\sqrt{T}} + La_0\nu_p^2 T^{2/p-3/2}\delta^{-2/p} \right]. \tag{44}$$

Table 4: Powers in (44) at $\alpha = a_0/\sqrt{T}$. At each listed confidence exponent, the polynomial term is $O(T^{-1/2})$ and the variance term is $O(T^{-1/2}\log T)$. Noise moments and other problem constants are held fixed.

Moment p	Rare-shock term, fixed δ	Largest sufficient r	$\delta = T^{-r}$
$5/2$	$T^{-7/10}$	$1/4$	$T^{-1/4}$
4	T^{-1}	1	T^{-1}
6	$T^{-7/6}$	2	T^{-2}

For fixed confidence the rare-shock term is of smaller order than $T^{-1/2}$. At shrinking confidence $\delta = T^{-r}$, $r > 0$, it becomes $La_0\nu_p^2 T^{-3/2+2(1+r)/p}$. Hence

$$0 < r \le p/2 - 1 \quad \Longrightarrow \quad A_T = O(T^{-1/2}\log T). \tag{45}$$

The constant can depend on fixed p, L, a_0, σ, ν_p and Δ. Theorem 2.2 matches the polynomial powers in (44) once its impulse-response horizon condition holds. This matching concerns that component, whereas (45) is a sufficient range for the displayed root-T guarantee. Table 4 makes the powers explicit.

H.3 Comparison with expectation and moment guarantees

Liu et al. (2020, Theorem 1) use the same effective-step convention and give the finite-variance SGDM bound

$$\mathbb{E}A_T \le B_T := \frac{2\Delta}{\alpha T} + \left(1 + \frac{\beta + 3\beta^2}{2(1+\beta)}\right) L\alpha\sigma^2 \tag{46}$$

under the step restriction

$$\alpha \le \min\left\{\frac{1-\beta}{L(4-\beta+\beta^2)}, \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}\right\},$$

where the second term is infinite at $\beta = 0$. Our cap (2) implies both restrictions, since $4-\beta+\beta^2 \le 4$ and $2\sqrt{2}\sqrt{\beta+\beta^2} \le 4$. Markov's inequality yields $\Pr(A_T > B_T/\delta) \le \delta$. Thus (44) replaces an undifferentiated δ^{-1} cost by a logarithmic variance contribution and a separate finite-p contribution. For example, at $p = 4$, $\delta = T^{-1}$ and $\alpha = a_0/\sqrt{T}$, the Markov conversion gives $O(T^{1/2})$, while (44) gives $O(T^{-1/2}\log T)$. This compares confidence certificates under the additional fourth moment, with all problem constants fixed.

To distinguish moment orders, write $q \in (1, 2]$ for the assumptions of Liu (2026, Theorem 4.3 and Theorem C.3). If $(\mathbb{E}A_T^{q/2})^{2/q} \le R_T$, then

$$\Pr\{A_T > R_T\delta^{-2/q}\} \le \delta. \tag{47}$$

For a constant effective step, their full theorem gives, up to constants depending on q,

$$R_T \lesssim \frac{\Delta}{\alpha T} + \sigma_q^2 T^{2/q-2} + \frac{L\alpha\sigma_q^2}{(1-\beta)^{2/q-1}} T^{2/q-1},$$

where σ_q bounds the conditional central qth moment in L^q. Choosing α proportional to $T^{-1/q}$ gives $R_T = O(T^{-(q-1)/q})$ for fixed β and problem parameters. These moment guarantees reach the possibly infinite-variance range; our $p > 2$ theorem uses additional moments to separate confidence costs.

The sub-Weibull momentum comparison. More precisely, Li et al. (2026, Theorem 3.1) use $m_t = \gamma m_{t-1} + \eta_t g_t$, $x_{t+1} = x_t - m_t$, with $\eta_t = c/\sqrt{t}$ and $c \le (1-\gamma)^3/[4L(3-\gamma)]$. Their conditional sub-Weibull parameter is denoted here by ϑ. The sub-Gaussian case $\vartheta = 1/2$ gives the row above. For $1/2 < \vartheta \le 1$ the rate is $O(\log^{2\vartheta}(1/\delta)\log T/\sqrt{T})$; for $\vartheta > 1$ it is

$$O\left(\frac{\log^{\vartheta-1}(T/\delta)\log(1/\delta) + \log^{2\vartheta}(1/\delta)\log T}{\sqrt{T}}\right).$$

Table 5: Quantitative comparisons for smooth nonconvex objectives. Each row retains its own algorithm, moment range and criterion. $\widetilde{O}$ suppresses logarithmic factors in T and $1/\delta$; other problem parameters and momentum are fixed.

Result	Criterion and assumptions	Confidence certificate
Liu et al. (2020) + Markov	A_T; ordinary SGDM, finite variance	$O(T^{-1/2}/\delta)$ at $\alpha \asymp T^{-1/2}$.
Liu (2026) + Markov	A_T via its $q/2$ moment; ordinary SGDM, $1 < q \leq 2$	$O(T^{-(q-1)/q}\delta^{-2/q})$ with a horizon-tuned step.
Li & Orabona (2020, Theorem 1)	$\min_{t\leq T} \|\nabla F(x_t)\|^2$; momentum, conditional sub-Gaussian noise	$\widetilde{O}(T^{-1/2})$; decreasing within-run steps.
Li et al. (2026, Theorem 3.1)	$T^{-1}\sum_t \|\nabla F_S(x_t)\|^2$; heavy ball, conditional sub-Gaussian noise, sample smoothness	$O(\log(1/\delta)\log T/\sqrt{T})$; decreasing within-run steps.
Nguyen et al. (2023, Theorem 3.1)	A_T; clipped SGD, conditional central qth moments, $1 < q \leq 2$	$\widetilde{O}(T^{-2(q-1)/(3q-2)})$; logarithmic confidence factors.
Theorem 2.1	A_T; ordinary SGDM, conditional $p > 2$	$O(T^{-1/2}\log(6/\delta) + T^{2/p-3/2}\delta^{-2/p})$.
Theorems 3.1, 4.2	A_T; unconditional moments with energy control / additive causal noise	$O(T^{-1/2})$ at fixed confidence; refined confidence range under (9) or (15).

Both latter branches additionally assume $\eta_t\|\nabla F_S(x_t)\| \leq G$. Their objective is the empirical risk F_S, with sample-gradient smoothness. These are stronger tail assumptions than finite polynomial moments. The stepsize conventions also differ: rewriting a variable-step EMA recursion in heavy-ball form changes its inertia coefficient to $\beta\alpha_t/\alpha_{t-1}$, whereas their γ is constant. The constant-step conversion in Section 2 is therefore used only for fixed within-run steps.

H.4 Reading the quantitative comparison

Conditional sub-Gaussian and sub-Weibull assumptions control exponential moments, enabling logarithmic confidence factors (Li & Orabona, 2020; Li & Liu, 2022; Madden et al., 2024; Li et al., 2026). Equation (44) retains a variance-scale term under polynomial moments alone. Its polynomial term is the price of retaining ordinary SGDM over that moment class. Clipped SGD can obtain logarithmic confidence dependence even when only $q \leq 2$ moments are available (Nguyen et al., 2023); the update changes the response to a large gradient. The two guarantees therefore serve different algorithmic choices.

The criterion also matters. An average bound controls the minimum along the same trajectory, while a minimum bound alone leaves the average uncontrolled. A bound on average gradient norms likewise differs from one on their squares. Finally, parameter-error and last-iterate results in linear or strongly convex stochastic approximation (Lou et al., 2022; Wei et al., 2025; Zhang et al., 2026) exploit additional geometry. Table 5 keeps these distinctions explicit rather than assigning all results a single nominal convergence rate.

I Broader literature and comparison details

Stochastic approximation and momentum. Stochastic approximation originates with Robbins & Monro (1951). Trajectory averaging (Polyak & Juditsky, 1992) addresses estimation efficiency, whereas Bottou et al. (2018) review stochastic optimization in large-scale learning. For stochastic heavy ball, Gadat et al. (2018) establish almost-sure convergence for nonconvex coercive objectives and finite-time results under convexity. Liu et al. (2020) obtain finite-variance expected stationarity for SGDM with constants of the SGD order. Yan et al. (2018) give a unified nonconvex analysis of stochastic gradient, heavy-ball and Nesterov-style updates. In the strongly convex, small-constant-step regime, Yuan et al. (2016) establish mean-square relations between momentum and rescaled SGD. Kidambi et al. (2018) give stochastic instances where standard heavy-ball and Nesterov schemes do

not accelerate even with tuned parameters. These statements concern convergence or mean-square behavior; our lower bound instead concerns the confidence cost over a finite-moment class. The frequency-domain view of Li et al. (2025) already explains momentum as a filter. Our analysis uses its path-space energy to retain the cancellation needed for a constant uniform in β. These are distinct guarantees from the finite-moment confidence decomposition studied here. Appendix H compares the expectation-to-confidence conversion quantitatively.

Finite moments and conditional exponential tails. Fatkhullin et al. (2025) treat different gradient-moment, central-moment, smoothness and mini-batch regimes. Liu (2026) control a fractional moment of average squared-gradient stationarity for ordinary momentum. We retain the distinction between their $q \leq 2$ setting and our $p > 2$ variance/polynomial split.

Li & Orabona (2020) study momentum with decreasing steps and a delayed AdaGrad variant under conditional sub-Gaussian noise. Li et al. (2026) extend high-probability momentum analysis to conditional sub-Weibull noise with sample-gradient smoothness and their stated stepsize conditions. The heavier sub-Weibull branch includes an additional gradient/stepsize restriction; the sub-Gaussian specialization should be read under its own assumptions. Their Polyak–Łojasiewicz and learning guarantees use additional structure. Related SGD concentration results appear in Li & Liu (2022); Madden et al. (2024). Our fixed-within-horizon step and finite-polynomial-moment assumptions address a different tradeoff between admissible noise and confidence cost.

Robustness modifications and structural acceleration. Empirical and theoretical work on attention models motivates heavy-tailed gradient noise and clipping (Zhang et al., 2020). Gorbunov et al. (2020) analyze accelerated clipping in convex optimization. For nonconvex problems, Cutkosky & Mehta (2021) combine clipping, normalization and momentum under stochastic-gradient moments, while Nguyen et al. (2023); Sadiev et al. (2023) use conditional central moments. Normalization also yields finite-moment guarantees (Hübler et al., 2025; Liu & Zhou, 2025). Momentum improves normalized SGD under second-order smoothness (Cutkosky & Mehta, 2020); additional sample smoothness and variance reduction enable faster heavy-tail rates in Liu et al. (2023). These results identify algorithmic or structural ways to improve robustness. Theorem 2.2 concerns the ordinary recursion, so its polynomial cost is compatible with those improvements.

Dependence: sampling structure and innovation coupling. Duchi et al. (2012) establish expectation and high-probability guarantees for convex mirror descent with ergodic sampling. Sun et al. (2018) analyze Markov chain gradient descent, including nonconvex objectives and nonreversible finite-state chains. Even (2023) analyze stochastic gradients sampled by a Markov chain, with bounds depending on mixing or hitting times and the problem setting. Such queries generally produce state-dependent gradient errors. Our causal theorem treats additive errors through an explicit martingale–coboundary correction, while Corollary 3.2 allows bounded state-dependent amplitudes of martingale noise. The frameworks therefore cover complementary forms of dependence. The physical-dependence coefficients of Wu (2005) link the noise-generating mechanism directly to moment bounds. Liu et al. (2013) develop Nagaev and Rosenthal inequalities in this framework, and Wu & Wu (2016) supply the weighted concentration inequality used for scalar energy averages. We apply these tools to stationary energies, after controlling the adaptive optimization coefficient by objective-level localization.

Nagaev concentration in stochastic approximation. Lou et al. (2022) give Gaussian-plus-polynomial estimation bounds for averaged SGD in linear regression. Wei et al. (2025) establish concentration and Gaussian approximation for constant-learning-rate SGD under strong convexity and gradient regularity. Zhang et al. (2026) study sharp concentration of last and averaged iterates in online stochastic approximation. The present result transfers finite-moment concentration to general nonconvex average stationarity and tracks its dependence on momentum. The filtered-energy upper bound, uniform polynomial lower bound and unconditional causal extension provide the distinctions from these estimation and iterate-error results.

Probability tools and attribution. The independent-sum foundations are Fuk & Nagaev (1971); Nagaev (1979). For martingales, classical exponential concentration controls predictable variance under increment restrictions (Freedman, 1975). Self-normalized methods instead incorporate realized variation (de la Peña et al., 2004; Bercu & Touati, 2008); nonnegative-supermartingale formulations

organize the corresponding maximal bounds (Howard et al., 2020). Our Hilbert Fuk–Nagaev input is the terminal consequence of Mollenhauer & Fiedler (2025, Theorem 2.2, version 2), and the scalar maximal input is Fan et al. (2017, Corollary 2.5). The optimization-specific work is the path-space lifting and localization. The causal construction builds on classical martingale approximation (Gordin, 1969; Volný, 2017) and dimension-free Hilbert martingale moments (Pinelis, 1994). For scalar processes, Wu (2007, Theorem 1(ii), Eq. (3)) bound the L^q approximation remainder by projection tails raised to $\min(2, q)$. Projection norms are bounded by coupling norms, connecting that estimate to our square-tail sufficient condition. Volný (2017, Proposition 1 and Theorem 2) restates the classical decomposition criteria. Proposition E.1 illustrates their sharp boundary for the specified positive linear filter. We use this structure to obtain a predictable SGDM correction and explicit bounds for its three energy costs.