# EXACT TAIL LAWS FOR NONLINEAR SGD WITH INFINITE-VARIANCE GRADIENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

With infinite-variance gradient noise, SGD can have infinite expected excess objective even when most iterates are close to the optimum. We characterize its accuracy through exact first-order tail probabilities for the last iterate of nonlinear, constant-stepsize SGD. For regularly varying innovations of index $t \in (1, 2)$, the leading measure transports one large gradient shock through the local Hessian flow. This is the same measure that governs the central stable Ornstein–Uhlenbeck limit; we prove its relative accuracy at shrinking tolerances above the central scale, with state-dependent noise and a growing observation horizon. The admissible window is nonempty for every available gradient moment $p > 1$. The main technical step is a moment bound at the central fluctuation scale and a lower-moment control of the nonlinear error. The result yields explicit projection and excess-objective tails, extreme quantiles, and a factor $b^{1-t}$ for fixed batch size $b$. For affine population gradients, a sharper comparison gives the same law at fixed tolerances, including random sample Hessians. Coupled simulations in dimensions one to one hundred examine relative calibration, objective quantiles, batch effects, and the error due to random curvature.

## 1 INTRODUCTION

The expected optimization error can be a poor description of an accurate iterate. With additive infinite-variance gradient noise, even a strongly restoring SGD recursion has infinite expected excess objective after one update, while most iterates can remain close to the optimum. Tail probabilities answer a more useful question in this setting: how often does the last iterate exceed a prescribed error tolerance, and how does that probability change with the stepsize, curvature, or batch size? We derive exact first-order asymptotics for these probabilities, including their leading constants.

Consider iid samples $\varepsilon_n$, a differentiable objective $M : \mathbb{R}^d \to \mathbb{R}$, and a jointly measurable unbiased stochastic gradient $g$:

$$m(x) = \nabla M(x) = \mathbb{E}g(x, \varepsilon), \qquad x_n = x_{n-1} - \gamma g(x_{n-1}, \varepsilon_n). \tag{1}$$

The stepsize is constant within each run. Let $x^\star$ be the target, $X_n = x_n - x^\star$, $H = \nabla m(x^\star)$, and $\eta = -g(x^\star, \varepsilon)$. When $\eta$ is regularly varying with index $t \in (1, 2)$ and radial-normalized tail measure $\mu_\eta$, our projected tail law is

$$\mathbb{P}(\ell^\top X_{n_\gamma} > u_\gamma) \sim C_\ell \frac{\mathbb{P}(|\eta| > u_\gamma/\gamma)}{\gamma}, \qquad C_\ell = \int_0^\infty \mu_\eta\{z : \ell^\top e^{-sH} z > 1\}\, \mathrm{d}s. \tag{2}$$

Here $\ell \neq 0$ is fixed, $C_\ell > 0$, and $\sim$ means that the ratio tends to one. A single large gradient shock has direction governed by $\mu_\eta$; the local flow $e^{-sH}$ transports a shock of age $s$, and $1/\gamma$ counts effective shock opportunities. The vector result also specifies the direction of a large error and the tail of the excess objective.

The main technical difficulty is that the one-big-jump principle is most direct for linear recursions and weighted sums, whereas SGD is nonlinear , and in (2) the weights, the memory length, the horizon and the threshold all depend on $\gamma$. Extreme value theory and regular variation provide precise tail tools for heavy-tailed sums, linear processes, and stochastic recursions (Bingham et al., 1987; Embrechts et al., 1997; Hult & Samorodnitsky, 2008; Mikosch & Wintenberger, 2013). Existing heavy-tailed SGD theory has established convergence guarantees, stable-limit behavior, and high-probability bounds



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.

under weak moment assumptions or clipping schemes (Wang et al., 2021; Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023; Blanchet et al., 2025). Wang & Rhee (2024) develop sample-path large deviations and metastability for heavy-tailed dynamical systems with a factorized noise term $\sigma(x)Z$. Their results hold on fixed time horizons under global Lipschitz conditions. The untruncated case further assumes bounded drift and noise coefficients. Additional estimates are needed beyond a direct application of these fixed-window results. In contrast, our result identifies the exact shrinking-threshold tail probability of the nonlinear constant-stepsize iterate, including the sharp constant.

### 1.1 Contributions

- **Relative tail asymptotics for nonlinear SGD.** Under centered dissipativity and the stated regularity assumptions, we establish a vector tail law for nonlinear constant-stepsize SGD. The result identifies the leading measure and gives vanishing relative error at shrinking tolerances over a diverging effective observation horizon.
- **Moment bounds and nonlinear comparison.** A central-scale moment bound and a lower-moment coupling control the nonlinear error at the rare-event probability scale. For a quadratic local remainder, these estimates yield $(t-p)/t < a < (t-1)/t$, a nonempty threshold window for every admissible gradient moment $p > 1$. This includes models whose noise has fewer moments than the innovation at the target. We also obtain corresponding windows for superlinear Hölder remainders.
- **Objective extremes and fixed-tolerance extensions.** The tail measure yields projection and directional laws, extreme objective quantiles, and fixed-batch scaling. For affine population gradients, a sharper comparison extends the law to fixed tolerances under an additional moment condition, including models with random sample Hessians.

### 1.2 Related literature

**SGD asymptotics and uncertainty quantification.** Classical stochastic approximation traces back to Robbins & Monro (1951), with asymptotic efficiency of averaged recursions developed by Polyak & Juditsky (1992) and in the stochastic approximation monograph of Kushner & Yin (2003). Non-asymptotic convergence theory for stochastic approximation and SGD is now extensive; see, for example, Moulines & Bach (2011). More closely related to the inferential motivation of this paper, Chen et al. (2020) develop statistical inference for model parameters computed by SGD, Fang et al. (2018) propose online bootstrap confidence intervals, Yu et al. (2021) analyze constant-stepsize SGD inference in nonconvex regimes, and Sheshukova et al. (2026) establish Gaussian approximation and multiplier bootstrap validity for averaged SGD, while Wei et al. (2025) develop Gaussian approximation and concentration results for constant-learning-rate SGD. These works are primarily finite-variance and Gaussian in nature. Most distributional analyses of SGD rely, either explicitly or implicitly, on a finite-variance picture (Zhu et al., 2023; Merad & Gaïffas, 2025).

**Heavy-tailed SGD and stochastic optimization.** Heavy-tailed stochastic gradient noise has been observed and modeled in modern learning problems (Simsekli et al., 2019; Gürbüzbalaban et al., 2021; Hodgkinson & Mahoney, 2021; Damek & Mentemeier, 2026). Theoretical work has studied convergence rates and limiting distributions of SGD under infinite-variance noise (Wang et al., 2021; Blanchet et al., 2025), as well as robust or clipped methods that achieve high-probability guarantees under weak moment assumptions (Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023; Liu & Zhou, 2025). These results show that optimization remains possible under heavy-tailed noise and provide convergence or high-probability control.

**Extreme value theory and regular variation.** The mathematical mechanism behind our result is the one-big-jump principle for regularly varying random variables. Regular variation and Potter-type bounds (Bingham et al., 1987) are classical tools in extreme value theory (Bingham et al., 1987; Embrechts et al., 1997). Denisov et al. (2008) characterize uniform big-jump domains for random walks. Tail asymptotics for heavy-tailed linear processes, infinite weighted series, and stochastic recurrence equations have been developed in Hult & Samorodnitsky (2008) and Mikosch & Wintenberger (2013). Wang & Rhee (2024) develop sample-path large deviations for recursions with state-dependent multiplicative noise under global Lipschitz conditions; related work studies exit times and multi-basin dynamics (Wang & Rhee, 2025; 2026). Our linearized recursion is related to

these weighted-sum settings, but the learning-rate-dependent weights and the nonlinear SGD transfer require additional analysis. In particular, the sharp constant in (2) arises from a Riemann-sum limit of the exponentially stable impulse response of the linearized dynamics.

**Nonconvex and dissipative constant-stepsize dynamics.** Constant-stepsize SGD can be viewed as a Markov chain, and this viewpoint has been used to study asymptotic expansions, bias, and stationary behavior (Dieuleveut et al., 2020; Yu et al., 2021). Dissipativity assumptions are also standard in nonconvex sampling and learning dynamics, including analyses of stochastic gradient Langevin dynamics (Raginsky et al., 2017; Zhang et al., 2023). Our assumptions follow this stability perspective but target a different phenomenon: rather than Gaussian fluctuations or stationary bias, we quantify the exact heavy-tailed rare-event probability of the last iterate.

## 2 Setting and main tail results

**Notation.** The dimension $d$ is fixed, $|\cdot|$ is the Euclidean norm, and $\|\cdot\|_{\mathrm{op}}$ is the matrix operator norm. For a random vector $Z$, let $\|Z\|_r = (\mathbb{E}|Z|^r)^{1/r}$. We write $I_d$ for the identity and $S^{d-1} = \{z : |z| = 1\}$. The initial point $x_0$ is fixed and deterministic; $\mathcal{F}_n = \sigma(\varepsilon_1, \dots, \varepsilon_n)$ is the natural filtration. Constants may depend on the fixed model, dimension, and exponents, but not on $n$ or $\gamma$. The exponent $p \in (1, 2)$ denotes an available gradient moment, and $t \in (1, 2)$ denotes the tail index.

**Assumption 2.1** (Centered dissipativity)**.** *The continuously differentiable objective $M$ has gradient $m$, $m(x^\star) = 0$, and, for some $\alpha > 0$,*

$$\langle x - x^\star, m(x)\rangle \geq \alpha|x - x^\star|^2, \qquad x \in \mathbb{R}^d. \tag{3}$$

Assumption 2.1 is a one-point monotonicity condition centered at the target $x^\star$. It is weaker than global strong convexity, as it does not require monotonicity between arbitrary pairs of points, but it is stronger than a purely tail dissipativity condition since (3) is imposed globally. This type of stability condition is common in the analysis of stochastic approximation and dissipative dynamical systems, and it is compatible with many nonconvex statistical objectives (Yu et al., 2021). The nonconvex models in Section 5 satisfy this condition explicitly.

**Assumption 2.2** ($L^p$ gradient regularity)**.** *For some $1 < p < 2$, $L_p, L_g < \infty$, and $\kappa > 0$,*

$$\|g(x, \varepsilon)\|_p \leq L_p(1 + |x - x^\star|), \quad x \in \mathbb{R}^d, \tag{4}$$

$$\|g(x, \varepsilon) - g(x^\star, \varepsilon)\|_p \leq L_g|x - x^\star|, \quad |x - x^\star| \leq \kappa. \tag{5}$$

The gradient has global $L^p$ growth and local $L^p$ increments; sample gradients may have random Lipschitz constants.

**Assumption 2.3** (Local quadratic remainder)**.** *The gradient is differentiable at $x^\star$, with $H = \nabla m(x^\star)$, and for some $C_R < \infty$,*

$$|m(x) - H(x - x^\star)| \leq C_R|x - x^\star|^2, \qquad |x - x^\star| \leq \kappa. \tag{6}$$

Here $H$ is symmetric and positive definite, with $H \succeq \alpha I_d$. A locally Lipschitz Hessian is sufficient. Remark 3.6 discusses any superlinear remainder with a corresponding threshold window.

**Assumption 2.4** (Regularly varying innovation)**.** *The centered innovation $\eta = -g(x^\star, \varepsilon)$ has index $t \in (1, 2)$ and a nonzero Radon tail measure $\mu_\eta$ on $\mathbb{R}^d \setminus \{0\}$ satisfying*

$$\frac{\mathbb{P}(r^{-1}\eta \in A)}{\mathbb{P}(|\eta| > r)} \longrightarrow \mu_\eta(A), \qquad r \to \infty, \tag{7}$$

*for every Borel set $A$ bounded away from zero with $\mu_\eta(\partial A) = 0$. The measure satisfies $\mu_\eta(cA) = c^{-t}\mu_\eta(A)$ for $c > 0$ and $\mu_\eta\{|z| > 1\} = 1$.*

Spherical Student-$t$ noise with $t$ degrees of freedom is an example that has $\mathbb{P}(|\eta| > r) \sim c_\eta r^{-t}$ for a constant $c_\eta > 0$ and a rotation-invariant $\mu_\eta$. The two objects in (2) are

$$L_\gamma(u) := \frac{\mathbb{P}(|\eta| > u/\gamma)}{\gamma}, \qquad \nu_H(A) := \int_0^\infty \mu_\eta\{z : e^{-sH}z \in A\}\,\mathrm{d}s. \tag{8}$$

One unit of effective time contains $1/\gamma$ iterations, and each noise step exceeds $u$ with probability $\mathbb{P}(|\eta| > u/\gamma)$, so $L_\gamma(u)$ is the rate of large shocks. The measure $\nu_H$ is finite on sets bounded away from zero and satisfies $\nu_H(cA) = c^{-t}\nu_H(A)$.

**Theorem 2.5** (Nonlinear vector tail)**.** *Assume Assumptions 2.1–2.4, with*

$$1 < p \le t, \qquad \frac{t-p}{t} < a < \frac{t-1}{t}, \qquad u_\gamma = c\gamma^a, \quad c > 0. \tag{9}$$

*As $\gamma \downarrow 0$, let $n_\gamma \in \mathbb{N}$ satisfy $n_\gamma \gamma \to \infty$. If $x_0 \neq x^\star$, also require $n_\gamma \gamma / \log(1/\gamma) \to \infty$. For every fixed Borel set $A$ satisfying*

$$\operatorname{dist}(A, 0) > 0, \qquad \nu_H(\partial A) = 0, \tag{10}$$

*we have*

$$\mathbb{P}(X_{n_\gamma}/u_\gamma \in A) = L_\gamma(u_\gamma)\{\nu_H(A) + o(1)\}. \tag{11}$$

*If $\nu_H(A) > 0$, this is a relative asymptotic equivalence.*

For a half-space $A_\ell = \{z : \ell^\top z > 1\}$, its boundary has zero tail measure. Thus (11) gives (2), with

$$C_\ell = \nu_H(A_\ell). \tag{12}$$

The theorem also applies to norm exceedances and to directional sets $\{z : |z| > 1,\ z/|z| \in C\}$ when their boundaries have zero $\nu_H$ mass. These sets describe both the size and direction of extreme deviations. More precisely, writing $D_1 = \{z : |z| > 1\}$,

$$\mathcal{L}(X_{n_\gamma}/u_\gamma \mid |X_{n_\gamma}| > u_\gamma) \Longrightarrow \frac{\nu_H(\cdot \cap D_1)}{\nu_H(D_1)}. \tag{13}$$

The symbol $\mathcal{L}$ denotes a probability law and $\Longrightarrow$ weak convergence. Mapping $z$ to $z/|z|$ gives the conditional direction of a large error (Appendix E.4).

**Corollary 2.6** (Eigenvector projections)**.** *Under Theorem 2.5, let $|v| = 1$, $H^\top v = \lambda_v v$, and $\mu_\eta\{z : v^\top z > 1\} > 0$. Then*

$$\mathbb{P}(v^\top X_{n_\gamma} > u_\gamma) \sim \frac{\mathbb{P}(v^\top \eta > u_\gamma/\gamma)}{t\lambda_v \gamma}. \tag{14}$$

For a general direction, the propagated direction $e^{-sH^\top}\ell$ varies with $s$, so the full measure in (12) matters. The eigenvector formula isolates the inverse dependence on local curvature. Appendix D also gives a projected-only result, with its own sufficient moment and threshold conditions.

**Corollary 2.7** (Excess objective value)**.** *Under Theorem 2.5, define $A_M = \{z : z^\top H z/2 > 1\}$ and $C_M = \nu_H(A_M)$. Then $0 < C_M < \infty$ and*

$$\mathbb{P}\{M(x_{n_\gamma}) - M(x^\star) > u_\gamma^2\} \sim C_M L_\gamma(u_\gamma). \tag{15}$$

This is a tail statement for the optimization objective itself. For a pure power innovation tail, the objective threshold has exponent $t/2$ in this joint limit. Tail probabilities remain informative even when expected excess objective value is infinite, as in additive infinite-variance quadratic models. The resulting quantiles describe increasingly rare objective deviations.

**Remark 2.8** (Reading the constants)**.** *Write the radial-normalized tail measure in polar form as $\mu_\eta(\mathrm{d}r, \mathrm{d}\omega) = tr^{-t-1}\mathrm{d}r\, S_\eta(\mathrm{d}\omega)$, where $S_\eta$ is a probability measure on $S^{d-1}$. For the special case $H = \lambda I_d$, direct radial integration gives*

$$C_\ell = \frac{\int (\ell^\top \omega)_+^t S_\eta(\mathrm{d}\omega)}{t\lambda}, \qquad C_M = \frac{2^{-t/2}}{t}\lambda^{t/2-1}, \qquad r_+ = \max(r, 0). \tag{16}$$

*Indeed, the flow multiplies a shock of age $s$ by $e^{-\lambda s}$, and the objective event requires its propagated radius to exceed $\sqrt{2/\lambda}$. The projected constant depends on the angular distribution, whereas this isotropic objective constant does not. Holding the innovation tail law fixed, increasing local curvature suppresses projected tails as $\lambda^{-1}$ and objective-value tails as $\lambda^{t/2-1}$. The difference comes from the objective's curvature weighting. For a nonscalar $H$, directions can change through unequal contraction rates, and* (8) *retains the full interaction between curvature and tail geometry.*

**Remark 2.9** (Three scales)**.** *For Student innovations, typical fluctuations have scale proportional to $\gamma^{1-1/t}$. The upper bound on $a$ places $u_\gamma$ above this scale, while $a > 0$ localizes the event near $x^\star$. The lower bound on $a$ makes the nonlinear comparison error negligible relative to the tail probability. For general regularly varying innovations, $L_\gamma$ retains the slowly varying factor.*

*The propagated measure $\nu_H$ is also the Lévy measure of the stationary distribution of a stable Ornstein–Uhlenbeck process with drift $z \mapsto -Hz$ and driving Lévy measure $\mu_\eta$ (Głowacki et al., 2026). Theorem 2.5 shows that this measure governs shrinking-threshold tail probabilities of nonlinear SGD with vanishing relative error.*

# 3 WHY THE LINEARIZATION IS RELATIVELY ACCURATE

We separate the argument into uniform moment control, a linear single-jump limit, and a nonlinear comparison. This order exposes where each assumption enters. All statements are uniform over $n$ for sufficiently small $\gamma$ when a finite-time bound is asserted.

## 3.1 CENTRAL-SCALE MOMENTS AND THE COUPLED RECURSION

Choose the innovation quantile and central scale

$$b_\gamma = \inf\{r \geq 1 : \mathbb{P}(|\eta| > r) \leq \gamma\}, \qquad s_\gamma = \gamma b_\gamma. \tag{17}$$

Regular variation gives $\mathbb{P}(|\eta| > b_\gamma) \sim \gamma$ and makes $s_\gamma$ regularly varying at zero with index $1 - 1/t$. For Student innovations, $s_\gamma$ is proportional to $\gamma^{1-1/t}$.

**Theorem 3.1** (Central-scale moment bound)**.** *Under Assumptions 2.1, 2.2, and 2.4, with $1 < p < t$, there exist $C, c, \gamma_0 > 0$ such that, for all $0 < \gamma < \gamma_0$ and integers $n \geq 0$,*

$$\mathbb{E}|X_n|^p \leq Ce^{-cn\gamma}|X_0|^p + Cs_\gamma^p. \tag{18}$$

The regularized Lyapunov function $V_\gamma(x) = (s_\gamma^2 + |x|^2)^{p/2}$ resolves the singular curvature of $|x|^p$ near zero. Its Taylor remainder uses truncated second moments for small innovations and $p$th moments for large ones. Both contributions are $O(\gamma s_\gamma^p)$. The state-dependent noise term is absorbed into the dissipative drift. Appendix C.1 gives the proof. Without regular variation, the weaker $O(\gamma^{p-1})$ bound remains available under global moment growth alone (Appendix A).

For linear drift $m(x) = \lambda x$, zero initialization, and symmetric $t$-stable innovation with $\mathbb{E}e^{i\theta\eta} = e^{-|\theta|^t}$,

$$X_n \overset{d}{=} \gamma \left\{ \sum_{j<n} (1 - \lambda\gamma)^{tj} \right\}^{1/t} \eta.$$

Hence $\mathbb{E}|X_n|^p \sim \mathbb{E}|\eta|^p (t\lambda)^{-p/t} \gamma^{p(1-1/t)}$ when $n\gamma \to \infty$. The moment theorem attains this polynomial order, improving the coarse exponent by $1 - p/t > 0$.

Using the same innovations, define

$$Y_n = B_\gamma Y_{n-1} + \gamma\eta_n, \quad B_\gamma = I_d - \gamma H, \quad Y_0 = X_0, \quad \eta_n = -g(x^\star, \varepsilon_n). \tag{19}$$

Then $\mathbb{E}\eta_n = 0$ and

$$Y_n = B_\gamma^n X_0 + \gamma T_{\gamma,n}, \qquad T_{\gamma,n} = \sum_{j=0}^{n-1} B_\gamma^j \eta_{n-j}. \tag{20}$$

Write $y_n = x^\star + Y_n$ and $\Delta_n = X_n - Y_n = x_n - y_n$.

## 3.2 LINEAR SINGLE-JUMP LIMIT

**Assumption 3.2** (Linear rare-event regime)**.** *The positive sequences $u_\gamma$ and integers $n_\gamma \to \infty$ satisfy*

$$n_\gamma\gamma \to \infty, \quad r_\gamma = u_\gamma/\gamma \to \infty, \quad L_\gamma(u_\gamma) \to 0, \quad |B_\gamma^{n_\gamma} X_0|/u_\gamma \to 0. \tag{21}$$

**Theorem 3.3** (Linear vector tail)**.** *Assume Assumption 2.1, differentiability of $m$ at $x^\star$, and Assumptions 2.4 and 3.2. For $A$ satisfying* (10)*,*

$$\mathbb{P}(Y_{n_\gamma}/u_\gamma \in A) = L_\gamma(u_\gamma)\{\nu_H(A) + o(1)\}. \tag{22}$$

*If $\nu_H(A) > 0$, this probability is also asymptotic to $\sum_{j<n_\gamma} \mathbb{P}(\gamma B_\gamma^j \eta \in u_\gamma A)$.*

**Corollary 3.4** (Linear-functional constant)**.** *Under Theorem 3.3, let $\ell \in \mathbb{R}^d \setminus \{0\}$ be fixed with $C_\ell > 0$. Then*

$$\mathbb{P}(\ell^\top Y_{n_\gamma} > u_\gamma) \sim \Lambda_\gamma(\ell, u_\gamma) \sim C_\ell L_\gamma(u_\gamma), \quad \Lambda_\gamma(\ell, u) = \sum_{j<n_\gamma} \mathbb{P}(\gamma \ell^\top B_\gamma^j \eta > u). \tag{23}$$

The finite sum uses the exact innovation distribution when available. The weights converge to the Hessian flow, turning the sum into the integral for $\nu_H$; truncated moments control the small jumps uniformly over the horizon. Appendix B gives the vector and projected-tail proofs. The linear theorem also allows fixed positive tolerances.

### 3.3 A LOWER-MOMENT NONLINEAR TRANSFER

**Theorem 3.5** (Central-scale linearization error)**.** *Under Assumptions 2.1–2.4, with $1 < p < t$, for every $1 \le q \le p$ there are $C_q, c_q, \gamma_q > 0$ such that*

$$\mathbb{E}|\Delta_n|^q \le C_q s_\gamma^p + C_q e^{-c_q n\gamma}(|X_0|^p + |X_0|^q), \qquad n \ge 1, \quad 0 < \gamma < \gamma_q. \tag{24}$$

The first-moment case $q = 1$, combined with the linear tail law, yields Theorem 2.5; the proof is given in Appendix D. When the gradient growth and local increment conditions hold for every $1 < p < t$, choosing a fixed $p$ sufficiently close to $t$ covers every fixed threshold $u_\gamma = c\gamma^a$ with $0 < a < (t-1)/t$.

**Remark 3.6** (Superlinear Remainders)**.** *The local quadratic remainder Assumption 2.3 can be extended to the Hölder local remainders, i.e. assuming $|m(x) - H(x - x^\star)| \le C|x - x^\star|^\beta$ for $|x - x^\star| \le \kappa$ instead, where $1 < \beta \le 2$. This further extends the scope; see Appendix E.*

## 4 CONSEQUENCES AND EXTENSIONS

The tail law quantifies extreme objective values and the effects of stepsize, batching, and observation time. We also describe when it extends to fixed tolerances and how local curvature emerges from nonlinear dynamics.

**How large are extreme objective values?** Quantiles describe the upper tail even when the expected objective is infinite. Suppose $\mathbb{P}(|\eta| > r) \sim c_\eta r^{-t}$ with $c_\eta > 0$, and define $Q_\gamma(1-\delta) = \inf\{q : \mathbb{P}(M(x_{n_\gamma}) - M(x^\star) > q) \le \delta\}$. For $1 < p < t$, Corollary 2.7 gives

$$Q_\gamma(1-\delta_\gamma) \sim \left(\frac{C_M c_\eta \gamma^{t-1}}{\delta_\gamma}\right)^{2/t}, \qquad \delta_\gamma = \gamma^\zeta, \quad 0 < \zeta < p - 1. \tag{25}$$

This formula relates an extreme loss level to its exceedance probability. The restriction on $\zeta$ places the corresponding iterate threshold within the nonlinear theorem's admissible window. Appendix E gives the inversion argument and extensions to slowly varying tails and Hölder remainders.

**What do smaller steps and larger batches change?** Averaging a fixed number $b$ of independent gradients multiplies the leading innovation tail probability by $b^{1-t}$. For pure power tails, the dependence on stepsize and batch size is therefore $(\gamma/b)^{t-1}$ at a common admissible tolerance. For example, when $t = 1.5$, batches of four and sixteen reduce the leading tail probability by factors of $1/2$ and $1/4$. Comparing $(\gamma, b)$ with $(\gamma/b, 1)$ also requires appropriate observation horizons; Appendix E states the equal-sample-budget condition.

**When can the tolerance remain fixed?** The main nonlinear theorem uses shrinking tolerances. For an affine population gradient, the following extension also covers any fixed $u > 0$.

**Corollary 4.1** (Affine gradients at fixed tolerances)**.** *Assume Assumptions 2.1, 2.2, and 2.4, with $m(x) = H(x - x^\star)$ globally and $t^2/(2t-1) < p < t$. Under the initialization and horizon conditions of Theorem 2.5, for every fixed $u > 0$ and $A$ satisfying* (10)*,*

$$\mathbb{P}(X_{n_\gamma} \in uA) = L_\gamma(u)\{\nu_H(A) + o(1)\}. \tag{26}$$

*The conclusion also holds when the complete gradient conditions hold at the finite endpoint $p = t$.*

Sample gradients may still have random curvature, as in least squares with bounded random designs and heavy-tailed labels. At $t = 1.5$, the sufficient moment condition is $p > 9/8$. Appendix E.5 gives the proof.

**How does observation time affect the tail?** Starting from $X_0 = 0$, a finite horizon $n_\gamma\gamma \to T \in (0, \infty)$ replaces the integral in $\nu_H$ by $\int_0^T$. For a unit eigenvector $Hv = \lambda_v v$, the resulting constant is $C_v(T) = C_v(1 - e^{-t\lambda_v T})$, where $C_v = \nu_H(\{z : v^\top z > 1\})$. This factor describes how the tail approaches its long-time level.

**Why does local curvature describe nonlinear dynamics?** For a locally Lipschitz nonlinear drift, let $\varphi_s(z)$ solve $\dot\varphi_s = -m(x^\star + \varphi_s)$ with $\varphi_0(z) = z$. The flow transports a large shock through the nonlinear dynamics, giving the deterministic quantity $C_A^{\mathrm{flow}}(u) = \int_0^\infty \mu_\eta\{z : \varphi_s(z) \in uA\}\,\mathrm{d}s$. Under its stated assumptions, Proposition E.3 proves $u^t C_A^{\mathrm{flow}}(u) \to \nu_H(A)$ as $u \downarrow 0$. Thus the nonlinear flow recovers the local Hessian description at shrinking tolerances.

## 5 NUMERICAL STUDIES

We examine three questions: whether the tail law calibrates nonlinear SGD in its joint limit, what it predicts for extreme objective values and batching, and how random sample curvature affects finite-step accuracy. Coupled nonlinear and population-linear iterates $X, Y$ share the same innovations. For a projected event, write

$$R_{\mathrm{lin}} = \frac{\widehat{p}_Y}{\Lambda_\gamma}, \qquad R_{\mathrm{tr}} = \frac{\widehat{p}_X}{\widehat{p}_Y}, \qquad R_{\mathrm{tot}} = \frac{\widehat{p}_X}{\Lambda_\gamma} = R_{\mathrm{lin}} R_{\mathrm{tr}}, \qquad R_{\mathrm{const}} = \frac{\Lambda_\gamma}{C_\ell L_\gamma(u)}, \tag{27}$$

where $\Lambda_\gamma$ is the exact-distribution single-jump sum in Corollary 3.4. A ratio of one indicates agreement with the benchmark. Intervals are nominal 95% pointwise Monte Carlo intervals; paired ratios retain the covariance within a coupled trajectory.

### 5.1 TAIL CALIBRATION IN A NONCONVEX MODEL

The nonlinear theorem concerns shrinking thresholds; fixed-threshold sweeps diagnose finite-level effects. We use

$$m(x) = x + \frac{4\sin^3 x}{1+x^2}, \qquad g(x,\varepsilon) = m(x) - \sigma\varepsilon, \qquad \varepsilon \sim t_{1.5}, \quad \sigma = 0.02. \tag{28}$$

Here $x^\star = 0$, $H = 1$, $xm(x) > 0.882x^2$, and $m'(2) \simeq -0.307$. The objective is nonconvex and the assumptions hold for every $1 < p < 1.5$. For $\gamma \in \{0.01, 0.003, 0.001\}$, we use 600,000 independent terminal pairs, $X_0 = Y_0 = 0$, and $n = \lceil 5\log(1/\gamma)/\gamma \rceil$.

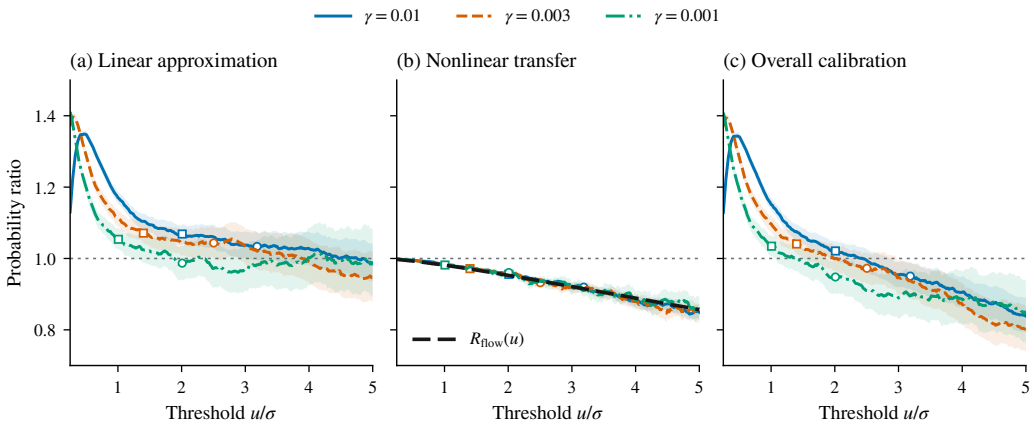


Figure 1: Nonconvex tail calibration on 241 thresholds, using the same 600,000 terminal pairs per stepsize. Axes are shared; shading gives pointwise intervals. Circles and squares mark $u_\gamma = 0.16\gamma^a$, $a = 0.2, 0.3$. The black dashed curve in (b) is the deterministic flow diagnostic $R_{\mathrm{flow}}(u)$.

Figure 1(b) compares nonlinear transfer with $R_{\mathrm{flow}}(u) = tu^t \int_u^\infty r^{-t}/m(r)\,\mathrm{d}r$, which tends to one as $u \downarrow 0$ by Proposition E.3. This is a deterministic diagnostic. It tracks the transfer deficit within 0.01 at all six prespecified points (Table 1); its finite-step counterpart does likewise and differs by less than 0.001 over the displayed range (Appendix E.7). The joint-limit test uses $a = 0.2, 0.3$, both in the theorem's window at $p = 1.45$; thresholds are respectively 14.8–20.1 and 9.33–10.1 times $\sigma\gamma^{1/3}$.

Nonlinear transfer improves along both sequences; at the smallest step the total ratios are within 5.2% of one. The $a = 0.2$ total ratios are nonmonotone because linear and transfer errors contribute in opposite directions. For $M(x) = x^2/2 + 4\int_0^x \sin^3(s)/(1+s^2)\,\mathrm{d}s$, the smallest-step objective ratios are 0.907 and 0.993; slower threshold shrinkage retains a larger correction. The constant correction $R_{\mathrm{const}} - 1$ is 0.025%–0.25% (full counts in Appendix F).

Table 1: All prespecified settings, with the flow diagnostic. The objective column compares $M(X_n) - M(0) > u_\gamma^2$ with Corollary 2.7. Brackets give pointwise intervals.

| $a$ | $\gamma$ | $R_{\text{flow}}$ | $R_{\text{tr}}$ | $R_{\text{tot}}$ [95% CI] | Objective ratio [95% CI] |
|---|---|---|---|---|---|
| 0.2 | 0.01 | 0.915 | 0.920 | 0.951 [0.914, 0.988] | 0.884 [0.852, 0.918] |
| 0.2 | 0.003 | 0.937 | 0.932 | 0.972 [0.931, 1.016] | 0.944 [0.907, 0.983] |
| 0.2 | 0.001 | 0.953 | 0.961 | 0.948 [0.903, 0.996] | 0.907 [0.866, 0.950] |
| 0.3 | 0.01 | 0.953 | 0.956 | 1.021 [0.994, 1.048] | 0.973 [0.949, 0.998] |
| 0.3 | 0.003 | 0.971 | 0.972 | 1.041 [1.013, 1.069] | 1.007 [0.982, 1.033] |
| 0.3 | 0.001 | 0.981 | 0.982 | 1.034 [1.006, 1.064] | 0.993 [0.968, 1.020] |

### 5.2 Extreme objective values and batching

We next use $d = 10$, $H = \text{diag}(1, 1 + 1/9, \ldots, 2)$, $\boldsymbol{a} = (1, \ldots, 1)^\top/\sqrt{10}$, and

$$m(x) = Hx + 6\boldsymbol{a}\frac{\sin^3(\boldsymbol{a}^\top x)}{1 + (\boldsymbol{a}^\top x)^2}, \qquad M(x) = \tfrac{1}{2}x^\top Hx + 6\int_0^{\boldsymbol{a}^\top x} \frac{\sin^3 s}{1 + s^2}\,\mathrm{d}s.$$

This objective is nonconvex and centered-dissipative (Appendix H.1). Innovations have the multivariate Student-$t_{1.5}$ law; each step averages $b$ independent samples. Six configurations use 300,000 coupled terminal pairs each, from zero, at $n = \max\{500, \lceil 5\log(1/\gamma)/\gamma \rceil\}$.

The full-objective single-jump prediction is

$$B_b(\epsilon) = b\sum_{j<n} \mathbb{P}\{M((\gamma/b)(I_d - \gamma H)^j\eta) - M(0) > \epsilon\}. \tag{29}$$

It evaluates each linearly transported shock with the exact objective. Appendix G separately reports the local quadratic approximation, the quadratic-event single-jump sum, and their numerical standard errors. The matched heavy-tail reference evaluates the same objective on $Y_n$. For comparison, we also fit a covariance-based Gaussian surrogate using $10^5$ independent innovations and repeat this fit five times. Since the true covariance is infinite, this diagnostic tests what is lost when a finite empirical covariance is used to describe the tail.

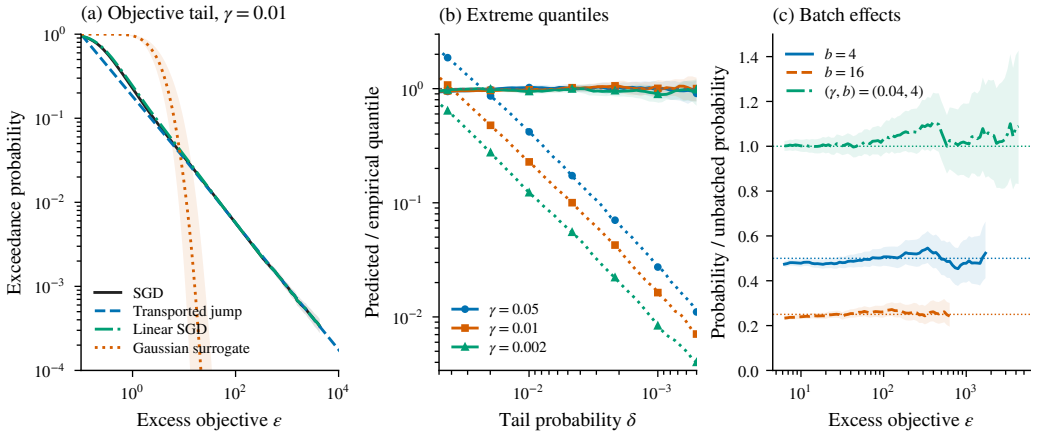


Figure 2: Objective tails and batch effects. (a) Full-objective exceedances at $(\gamma, b) = (0.01, 1)$; the Gaussian band spans five covariance fits. (b) Predicted/empirical quantiles for $b = 1$: solid curves use (29), dotted curves the median Gaussian fit. (c) Tail probabilities relative to $(0.01, 1)$; horizontal references are $1/2$, $1/4$, and 1. Empirical tail displays require at least 100 exceedances in each relevant sample; panel (c) also requires baseline probability below 0.05.

At the 99.9% level, single-jump quantile predictions differ from the empirical estimates by 0.5%–10.8% and all lie within their pointwise intervals (Table 2). The matched linear quantiles differ by

Table 2: The 99.9% objective quantile for unbatched SGD. Predictions and the matched linear reference are divided by the empirical quantile. All use the full objective.

| $\gamma$ | Empirical [95% CI] | Linear / emp. | Jump / emp. | Gaussian / emp. |
|---|---|---|---|---|
| 0.05 | 3026.1 [2582.4, 3509.2] | 0.998 | 0.977 | 0.027 |
| 0.01 | 988.0 [835.7, 1145.7] | 0.995 | 1.005 | 0.016 |
| 0.002 | 380.5 [320.9, 448.9] | 1.000 | 0.892 | 0.008 |

at most $0.5\%$. The median Gaussian fits underestimate these quantiles by factors of about 37, 61, and 120. Over the count-qualified thresholds in Figure 2(c), batches of four and sixteen give median probability ratios $0.487$ and $0.252$, close to $1/2$ and $1/4$. The matched-$\gamma/b$ ratio has median $1.021$. The comparison at matched $\gamma/b$, $(0.04, 4)$ versus $(0.01, 1)$, uses 2,000 and 2,303 gradient samples per trajectory, respectively. See Appendix G.

### 5.3 Random curvature and finite-step accuracy

Random-design regression uses $Z = \sqrt{d}H^{1/2}U$, with $U$ uniform on the sphere, and gradient $g(x; Z, \varepsilon) = Z(Z^\top x - \varepsilon) + 6\boldsymbol{a}\sin^3(\boldsymbol{a}^\top x)/(1 + (\boldsymbol{a}^\top x)^2)$, where $\varepsilon \sim t_{1.5}$. The population objective is unchanged, but the sample Hessian at the target is $ZZ^\top$. Alongside the population-Hessian benchmark, we propagate a shock through independent sample matrices $I_d - \gamma ZZ^\top$ to form the random-Hessian benchmark $\Lambda_\gamma^{\mathrm{RH}}$ (Appendix H.5). This isolates a source of finite-step error that the asymptotic population constant averages out.

Table 3: Median projected-tail ratios over four directions and 17 thresholds in $[1, 5]$. $R_{\mathrm{tot}}^{\mathrm{RH}} = \widehat{p}_X/\widehat{\Lambda}_\gamma^{\mathrm{RH}}$. The last column is the largest relative Monte Carlo standard error of $\widehat{\Lambda}_\gamma^{\mathrm{RH}}$ on this grid. Empirical sample sizes and intervals are in Appendix H.

| $d$ | $\gamma d$ | $R_{\mathrm{tot}}$ | $R_{\mathrm{tot}}^{\mathrm{RH}}$ | MC SE (%) |
|---|---|---|---|---|
| 10 | 1 | 4.264 | 1.141 | 0.16 |
| 10 | 0.5 | 1.587 | 1.045 | 0.19 |
| 10 | 0.1 | 1.080 | 1.007 | 0.10 |
| 100 | 1 | 4.328 | 1.081 | 0.08 |
| 100 | 0.5 | 1.586 | 0.998 | 0.10 |
| 100 | 0.1 | 1.117 | 1.041 | 0.06 |

At $\gamma d = 1$, retaining random curvature reduces median ratios from about $4.3$ to $1.08$–$1.14$ (Table 3). At $\gamma d = 0.1$, both denominators are closer, and the random-Hessian ratios are $1.007$ and $1.041$. A scalar random-curvature ablation in Appendix F.1 gives the same separation between curvature amplification and nonlinear drift.

## 6 Discussion

Our results identify a regime in which local curvature and the innovation law at the target determine rare errors of nonlinear SGD. The experiments show that finite-step calibration also depends on nonlinear-flow effects at fixed thresholds and on random sample curvature. A stochastic fixed-tolerance extension under our assumptions, together with quantitative control of these finite-step effects, would strengthen the practical use of the tail approximation. The present analysis assumes fixed dimension and centered dissipativity around a single equilibrium. Extending relative tail control to multiple basins and dimension-dependent regimes is a natural direction for further work.

## REPRODUCIBILITY STATEMENT

The appendix gives proofs, model definitions, sampling units, benchmarks, and uncertainty calculations. The supplement contains the simulation and analysis code with fixed seeds and configurations, together with the summary data behind every figure and table. Most displays can be regenerated from these data directly.

## AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, code development, and running and checking simulations.

We have reviewed all AI-assisted work. Experimental code and numerical results were executed and verified by the authors.

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

Nicholas H. Bingham, Charles M. Goldie, and Jozef L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, 1987.

Jose Blanchet, Aleksandar Mijatović, and Wenhao Yang. Limit theorems for stochastic gradient descent with infinite variance. arXiv:2410.16340, 2025. URL https://arxiv.org/abs/2410.16340v4. Version 4.

Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020. doi: 10.1214/18-AOS1801.

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.

Ewa Damek and Sebastian Mentemeier. Analysing heavy-tail properties of stochastic gradient descent by means of stochastic recurrence equations. *Journal of Applied Probability*, 63(2):459–483, 2026. doi: 10.1017/jpr.2025.10036.

Denis Denisov, A. B. Dieker, and Vsevolod Shneer. Large deviations for random walks under subexponentiality: The big-jump domain. *The Annals of Probability*, 36(5):1946–1991, 2008. doi: 10.1214/07-AOP382. URL https://arxiv.org/abs/math/0703265.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020. doi: 10.1214/19-AOS1850.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events: For Insurance and Finance*, volume 33 of *Applications of Mathematics*. Springer, Berlin, 1997. doi: 10.1007/978-3-642-33483-2.

Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *Journal of Machine Learning Research*, 19(78):1–21, 2018.

Bartosz Głowacki, Rafał Kulik, and Philippe Soulier. Convergence of stochastic gradient descent with mini-batching and infinite variance. arXiv:2605.07184, 2026. URL https://arxiv.org/abs/2605.07184v1. Version 1.

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.

Mert Gürbüzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3964–3975. PMLR, 2021.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4262–4274, 2021. URL https://proceedings.mlr.press/v139/hodgkinson21a.html.

Henrik Hult and Gennady Samorodnitsky. Tail probabilities for infinite series of regularly varying random vectors. *Bernoulli*, 14(3):838–864, 2008. doi: 10.3150/08-BEJ125.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, New York, 2 edition, 2003. doi: 10.1007/b97441.

Jiaqi Li, Zhipeng Lou, Stefan Richter, and Wei Biao Wu. The stochastic gradient descent from a nonlinear time series perspective. Working paper, 2024. URL https://stat.math.uni-heidelberg.de/SGD_GMC-3.pdf.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NKotdPUc3L.

Ibrahim Merad and Stéphane Gaïffas. Convergence and concentration properties of constant step-size SGD through markov chains. *Electronic Journal of Statistics*, 19(2):5843–5894, 2025.

Thomas Mikosch and Olivier Wintenberger. Precise large deviations for dependent regularly varying sequences. *Probability Theory and Related Fields*, 156(3–4):851–887, 2013. doi: 10.1007/s00440-012-0445-0.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987. doi: 10.2307/1911031. URL https://www.jstor.org/stable/1911031.

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

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

Yu. V. Prokhorov. An extremal problem in probability theory. *Theory of Probability & Its Applications*, 4(2):201–203, 1959.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703. PMLR, 2017.

Sidney I Resnick. *Extreme values, regular variation and point processes*. Springer, 1987.

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

Marina Sheshukova, Sergey Samsonov, Denis Belomestny, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for stochastic gradient descent. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1378–1386, 2026. URL https://proceedings.mlr.press/v300/sheshukova26a.html.

Umut Simsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837. PMLR, 2019.

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18866–18877, 2021.

Xingyu Wang and Chang-Han Rhee. Large deviations and metastability analysis for heavy-tailed dynamical systems. arXiv:2307.03479, 2024. URL https://arxiv.org/abs/2307.03479v4. Version 4.

Xingyu Wang and Chang-Han Rhee. Global dynamics of heavy-tailed SGDs in nonconvex loss landscape: Characterization and control. arXiv:2510.20905, 2025. URL https://arxiv.org/abs/2510.20905.

Xingyu Wang and Chang-Han Rhee. First-exit time analysis for truncated heavy-tailed dynamical systems. arXiv:2602.07968, 2026. URL https://arxiv.org/abs/2602.07968.

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. URL https://papers.nips.cc/paper_files/paper/2025/hash/7aae9e3ec211249e05bd07271a6b1441-Abstract-Conference.html.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4234–4248, 2021.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics and Optimization*, 87(2):25, 2023. doi: 10.1007/s00245-022-09932-6.

Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

# A Proof of the $L^p$ Moment Bound

**Theorem A.1** (Uniform moment bound)**.** *Assume centered dissipativity and* (4)*. For every $1 < r \le p$, there are $C_r, c_r, \gamma_r > 0$ such that*

$$\mathbb{E}|X_n|^r \le (1 - c_r\gamma)^n |X_0|^r + C_r\gamma^{r-1}, \quad n \ge 0, \quad 0 < \gamma < \gamma_r. \tag{30}$$

*In particular,*

$$\mathbb{P}(|X_n| > u) \le C_p u^{-p}\{\gamma^{p-1} + e^{-c_p n\gamma}|X_0|^p\}. \tag{31}$$

This auxiliary bound requires neither regular variation nor local linearization. We use the following elementary inequality for $1 < r \le 2$.

**Lemma A.2** (Subquadratic increment inequality)**.** *For $1 < r \le 2$, there is $C_r < \infty$ such that, for deterministic $u, w \in \mathbb{R}^d$,*

$$|u + w|^r \le |u|^r + r|u|^{r-2}u^\top w + C_r|w|^r, \tag{32}$$

*with the middle term zero at $u = 0$.*

*Proof.* The gradient of $f(u) = |u|^r$ is $(r-1)$-Hölder continuous. Integrating $\nabla f(u + sw) - \nabla f(u)$ against $w$ over $s \in [0, 1]$ proves the inequality. Its conditional-expectation form also appears in Li et al. (2024, Appendix A.1). □

*Proof of Theorem A.1.* Fix $r \in (1, p]$. Since (4) holds in $L^p$, monotonicity of $L^s$ norms gives

$$\|g(x, \varepsilon)\|_r \le L_p(1 + |x - x^\star|), \qquad x \in \mathbb{R}^d. \tag{33}$$

Applying Lemma A.2 conditionally on $\mathcal{F}_{n-1} := \sigma(\varepsilon_1, \dots, \varepsilon_{n-1})$ with $u = x_{n-1} - x^\star$ and $w = -\gamma g(x_{n-1}, \varepsilon_n)$ yields

$$\begin{aligned}\mathbb{E}[|x_n - x^\star|^r \mid \mathcal{F}_{n-1}] \le |x_{n-1} - x^\star|^r - r\gamma\,|x_{n-1} - x^\star|^{r-2}\,\langle x_{n-1} - x^\star, m(x_{n-1})\rangle \\ + C_r\gamma^r\mathbb{E}[|g(x_{n-1}, \varepsilon_n)|^r \mid \mathcal{F}_{n-1}].\end{aligned} \tag{34}$$

By centered dissipativity (3),

$$|x_{n-1} - x^\star|^{r-2}\,\langle x_{n-1} - x^\star, m(x_{n-1})\rangle \ge \alpha\,|x_{n-1} - x^\star|^r\,. \tag{35}$$

By (33),

$$\mathbb{E}[|g(x_{n-1}, \varepsilon_n)|^r \mid \mathcal{F}_{n-1}] \le C_r(1 + |x_{n-1} - x^\star|^r). \tag{36}$$

Combining (34)–(36) gives

$$\mathbb{E}[|x_n - x^\star|^r \mid \mathcal{F}_{n-1}] \le (1 - r\alpha\gamma + C_r\gamma^r)\,|x_{n-1} - x^\star|^r + C_r\gamma^r. \tag{37}$$

Since $r > 1$, for sufficiently small $\gamma$,

$$1 - r\alpha\gamma + C_r\gamma^r \le 1 - c_r\gamma$$

for some $c_r > 0$. Taking expectations and iterating yields

$$\begin{aligned}\mathbb{E}\,|x_n - x^\star|^r &\le (1 - c_r\gamma)^n\,|x_0 - x^\star|^r + C_r\gamma^r\sum_{j=0}^{n-1}(1 - c_r\gamma)^j \\ &\le (1 - c_r\gamma)^n\,|x_0 - x^\star|^r + C_r\gamma^{r-1}.\end{aligned} \tag{38}$$

This proves (30). The tail bound (31) follows from (30) with $r = p$ and Markov's inequality. □

# B Proofs for the Linearized Recursion

## B.1 Stability of the vector linearization

**Lemma B.1** (Stability of the vector linearization)**.** *Assume Assumption 2.1 and suppose that $m$ is differentiable at $x^\star$ with $H = \nabla m(x^\star)$. Then*

$$\frac{1}{2}(H + H^\top) \succeq \alpha I_d. \tag{39}$$

*Consequently, $H$ is positive stable. Moreover, there exists $\gamma_0 > 0$ such that, for all $0 < \gamma < \gamma_0$, $B_\gamma := I_d - \gamma H$ satisfies*

$$\|B_\gamma^j\|_{\mathrm{op}} \le e^{-\alpha\gamma j/2}, \qquad j \ge 0. \tag{40}$$

*Proof of Lemma B.1.* Fix $u \in \mathbb{R}^d$ and $r > 0$. Apply Assumption 2.1 with $x = x^\star + ru$. Since $m(x^\star) = 0$,

$$\left\langle u, \frac{m(x^\star + ru) - m(x^\star)}{r} \right\rangle \geq \alpha \|u\|^2.$$

Letting $r \downarrow 0$ gives $\langle u, Hu \rangle \geq \alpha \|u\|^2$. Since $\langle u, Hu \rangle = \langle u, (H + H^\top)u/2 \rangle$, this proves (39).

If $Hz = \lambda z$ for a possibly complex eigenvector $z \neq 0$, then

$$\mathrm{Re}(\lambda)\|z\|^2 = \mathrm{Re}(z^* H z) = z^* \frac{H + H^\top}{2} z \geq \alpha \|z\|^2,$$

so $H$ is positive stable. Moreover,

$$\|B_\gamma u\|^2 = \|u - \gamma H u\|^2 \leq (1 - 2\alpha\gamma + \gamma^2 \|H\|_{\mathrm{op}}^2)\|u\|^2.$$

For sufficiently small $\gamma$, the right-hand side is bounded by $(1 - \alpha\gamma)\|u\|^2$. Hence

$$\|B_\gamma u\| \leq (1 - \alpha\gamma)^{1/2}\|u\| \leq e^{-\alpha\gamma/2}\|u\|.$$

Iterating proves (40). □

### B.2 Proof of Theorem 3.3

Throughout this subsection, write

$$Q_\gamma := \mathbb{P}(|\eta_0| > r_\gamma), \qquad X_{\gamma,j} := B_\gamma^j \eta_{n_\gamma - j}, \qquad 0 \leq j < n_\gamma.$$

Then $T_{\gamma,n_\gamma} = \sum_{j=0}^{n_\gamma - 1} X_{\gamma,j}$. By Lemma B.1, there exists $c_B > 0$ such that, for all sufficiently small $\gamma$,

$$\|B_\gamma^j\|_{\mathrm{op}} \leq e^{-c_B \gamma j}, \qquad j \geq 0. \tag{41}$$

**Lemma B.2** (One-jump Riemann limit)**.** *Let $A \subset \mathbb{R}^d \setminus \{0\}$ be a Borel set satisfying* (10)*. Then*

$$\gamma \sum_{j=0}^{n_\gamma - 1} \frac{\mathbb{P}(r_\gamma^{-1} B_\gamma^j \eta_0 \in A)}{Q_\gamma} \longrightarrow \nu_H(A). \tag{42}$$

*The same conclusion holds with $A$ replaced by any $\nu_H$-continuity set bounded away from the origin.*

*Proof.* Fix $K < \infty$. Since $n_\gamma \gamma \to \infty$, the grid $\{j : \gamma j \leq K\}$ is eventually contained in $\{0, \ldots, n_\gamma - 1\}$. On $0 \leq \gamma j \leq K$,

$$\sup_{0 \leq j \leq K/\gamma} \|B_\gamma^j - e^{-\gamma j H}\|_{\mathrm{op}} \to 0. \tag{43}$$

By product vague convergence and the continuous mapping $(s, z) \mapsto e^{-sH} z$,

$$\gamma \sum_{0 \leq j \leq K/\gamma} \frac{\mathbb{P}(r_\gamma^{-1} e^{-\gamma j H} \eta_0 \in A)}{Q_\gamma} \longrightarrow \int_0^K \mu_\eta\{z : e^{-sH} z \in A\}\, ds, \tag{44}$$

where the boundary condition follows from $\nu_H(\partial A) = 0$. The replacement of $e^{-\gamma j H}$ by $B_\gamma^j$ follows from (43) and the usual inner/outer enlargement argument; because $A$ is bounded away from the origin, the contribution of $|z| > R$ is controlled uniformly by Potter's bound (Resnick, 1987) and then sent to zero as $R \to \infty$. Hence

$$\gamma \sum_{0 \leq j \leq K/\gamma} \frac{\mathbb{P}(r_\gamma^{-1} B_\gamma^j \eta_0 \in A)}{Q_\gamma} \longrightarrow \int_0^K \mu_\eta\{z : e^{-sH} z \in A\}\, ds. \tag{45}$$

It remains to let $K \to \infty$. If $r_\gamma^{-1} B_\gamma^j \eta_0 \in A$ and $a := \mathrm{dist}(A, 0) > 0$, then $|\eta_0| \geq a r_\gamma / \|B_\gamma^j\|_{\mathrm{op}}$. Potter's bound (Resnick, 1987) and (41) imply that, for every sufficiently small $\zeta > 0$,

$$\frac{\mathbb{P}(r_\gamma^{-1} B_\gamma^j \eta_0 \in A)}{Q_\gamma} \leq C_A \|B_\gamma^j\|_{\mathrm{op}}^{t - \zeta} \leq C_A e^{-(t-\zeta) c_B \gamma j}.$$

Therefore the discrete tail over $\gamma j > K$ is $O(e^{-cK})$, and the same bound controls the tail of the integral defining $\nu_H(A)$. Combining this with (45) proves (42). □

**Lemma B.3** (Negligibility of many small jumps)**.** *For every $\delta > 0$,*

$$\lim_{\theta \downarrow 0} \limsup_{\gamma \downarrow 0} \frac{\gamma}{Q_\gamma} \mathbb{P}\left( \left\| \sum_{j=0}^{n_\gamma - 1} X_{\gamma,j} \mathbf{1}_{\{\|X_{\gamma,j}\| \le \theta r_\gamma\}} \right\| > \delta r_\gamma \right) = 0. \tag{46}$$

*Proof.* Let $\widetilde{X}_{\gamma,j}^{(\theta)} := X_{\gamma,j} \mathbf{1}_{\{\|X_{\gamma,j}\| \le \theta r_\gamma\}}$ and $\overline{X}_{\gamma,j}^{(\theta)} := \widetilde{X}_{\gamma,j}^{(\theta)} - \mathbb{E}\widetilde{X}_{\gamma,j}^{(\theta)}$. Choose $\beta \in (t, 2)$ and $0 < \zeta < \beta - t$. A coordinate-wise von Bahr–Esseen inequality, equivalence of norms in finite dimension, and independence give

$$\mathbb{E}\left\| \sum_{j=0}^{n_\gamma - 1} \overline{X}_{\gamma,j}^{(\theta)} \right\|^\beta \le C_\beta \sum_{j=0}^{n_\gamma - 1} \mathbb{E}\|\widetilde{X}_{\gamma,j}^{(\theta)}\|^\beta.$$

By Karamata's theorem (Resnick, 1987) for the norm tail, Potter's bound (Resnick, 1987), and (41),

$$\sum_{j=0}^{n_\gamma - 1} \mathbb{E}\|\widetilde{X}_{\gamma,j}^{(\theta)}\|^\beta \le C r_\gamma^\beta Q_\gamma \theta^{\beta - t - \zeta} \sum_{j=0}^{n_\gamma - 1} \|B_\gamma^j\|_{\mathrm{op}}^{t+\zeta} \le C r_\gamma^\beta Q_\gamma \theta^{\beta - t - \zeta} \gamma^{-1}.$$

Markov's inequality yields

$$\mathbb{P}\left( \left\| \sum_{j=0}^{n_\gamma - 1} \overline{X}_{\gamma,j}^{(\theta)} \right\| > \frac{\delta r_\gamma}{2} \right) \le C \delta^{-\beta} \theta^{\beta - t - \zeta} \frac{Q_\gamma}{\gamma}. \tag{47}$$

Since $\mathbb{E}X_{\gamma,j} = B_\gamma^j \mathbb{E}\eta_0 = 0$,

$$\mathbb{E}\widetilde{X}_{\gamma,j}^{(\theta)} = -\mathbb{E}\left[ X_{\gamma,j} \mathbf{1}_{\{\|X_{\gamma,j}\| > \theta r_\gamma\}} \right].$$

Again by Karamata's theorem and Potter's bound (Resnick, 1987),

$$\left\| \sum_{j=0}^{n_\gamma - 1} \mathbb{E}\widetilde{X}_{\gamma,j}^{(\theta)} \right\| \le C r_\gamma \theta^{1 - t - \zeta} \frac{Q_\gamma}{\gamma} = o(r_\gamma)$$

for each fixed $\theta > 0$, because $Q_\gamma / \gamma \to 0$. Combining this with (47) and then letting $\theta \downarrow 0$ proves the lemma. □

*Proof of Theorem 3.3.* Let $A$ satisfy (10), and set $a := \mathrm{dist}(A, 0) > 0$. For $\varepsilon > 0$, define $A^\varepsilon := \{x : \mathrm{dist}(x, A) \le \varepsilon\}$ and $A_{-\varepsilon} := \{x \in A : \mathrm{dist}(x, A^c) > \varepsilon\}$. Choose $\varepsilon < a/4$ so that $A^\varepsilon$ and $A_{-\varepsilon}$ are $\nu_H$-continuity sets.

For $\theta \in (0, a/4)$, let $N_{\gamma,\theta} := \sum_{j=0}^{n_\gamma - 1} \mathbf{1}_{\{\|X_{\gamma,j}\| > \theta r_\gamma\}}$. Lemma B.2 applied to the norm-exceedance set gives $\mathbb{E}N_{\gamma,\theta} = O(Q_\gamma/\gamma)$; hence, by independence and $Q_\gamma/\gamma \to 0$,

$$\mathbb{P}(N_{\gamma,\theta} \ge 2) = o(Q_\gamma/\gamma). \tag{48}$$

For the upper bound, if $r_\gamma^{-1} \sum_j X_{\gamma,j} \in A$, if at most one summand is larger than $\theta r_\gamma$, and if the sum of the small summands has norm at most $\varepsilon r_\gamma$, then the unique large summand, if present, must lie in $r_\gamma A^\varepsilon$. Thus

$$\begin{aligned} \mathbb{P}\left( r_\gamma^{-1} \sum_j X_{\gamma,j} \in A \right) &\le \mathbb{P}\left( \bigcup_j \{ r_\gamma^{-1} X_{\gamma,j} \in A^\varepsilon \} \right) + \mathbb{P}(N_{\gamma,\theta} \ge 2) \\ &\quad + \mathbb{P}\left( \left\| \sum_j X_{\gamma,j} \mathbf{1}_{\{\|X_{\gamma,j}\| \le \theta r_\gamma\}} \right\| > \varepsilon r_\gamma \right). \end{aligned}$$

Using the union bound, Lemma B.2, (48), and Lemma B.3, then letting $\theta \downarrow 0$ and $\varepsilon \downarrow 0$, gives

$$\limsup_{\gamma \downarrow 0} \frac{\gamma}{Q_\gamma} \mathbb{P}\left( r_\gamma^{-1} \sum_j X_{\gamma,j} \in A \right) \le \nu_H(A).$$

The lower bound is analogous. On the event that one summand lies in $r_\gamma A_{-\varepsilon}$, no second summand is larger than $\theta r_\gamma$, and the small-summand contribution has norm at most $\varepsilon r_\gamma$, the full sum lies in $r_\gamma A$. Bonferroni's inequality and the same estimates yield

$$\liminf_{\gamma\downarrow 0} \frac{\gamma}{Q_\gamma}\mathbb{P}\left(r_\gamma^{-1}\sum_j X_{\gamma,j} \in A\right) \geq \nu_H(A).$$

Therefore

$$\mathbb{P}(r_\gamma^{-1}T_{\gamma,n_\gamma} \in A) = \frac{Q_\gamma}{\gamma}\{\nu_H(A) + o(1)\}. \tag{49}$$

Since $r_\gamma = u_\gamma/\gamma$, this proves the additive tail limit for $\gamma T_{\gamma,n_\gamma}/u_\gamma$, including sets with $\nu_H(A) = 0$. Lemma B.2 also gives

$$\sum_{j=0}^{n_\gamma-1} \mathbb{P}(r_\gamma^{-1}B_\gamma^j\eta_0 \in A) = \frac{Q_\gamma}{\gamma}\{\nu_H(A) + o(1)\},$$

which is used in the proof.

Finally, by (20), $Y_{n_\gamma} = B_\gamma^{n_\gamma}Y_0 + \gamma T_{\gamma,n_\gamma}$. The initialization condition in Assumption 3.2 implies $u_\gamma^{-1}B_\gamma^{n_\gamma}Y_0 \to 0$. A final inner/outer enlargement argument proves (22). When $\nu_H(A) > 0$, the two additive limits also give the stated relative single-jump equivalence. □

### B.3 Proof of Corollary 3.4

*Proof.* For $c > 0$, let $A_{\ell,c} := \{x : \ell^\top x > c\}$. This set is bounded away from the origin. Its boundary is the affine hyperplane $\partial A_{\ell,c} = \{x : \ell^\top x = c\}$. For each $s \geq 0$,

$$(e^{-sH})^{-1}\partial A_{\ell,c} = \{z : \ell^\top e^{-sH}z = c\},$$

which is an affine hyperplane not passing through the origin.

We claim that a homogeneous Radon measure on $\mathbb{R}^d \setminus \{0\}$ assigns zero mass to such a hyperplane. It is enough to check this on annuli $\{r \leq \|z\| \leq R\}$. If an affine hyperplane $L$ had positive $\mu_\eta$-mass on such an annulus, then for infinitely many distinct $\lambda_m$ sufficiently close to one, the sets $\lambda_m(L \cap \{r \leq \|z\| \leq R\})$ would be disjoint, have uniformly positive $\mu_\eta$-mass by homogeneity, and be contained in a slightly larger compact annulus away from zero. This contradicts Radon finiteness. Hence $\mu_\eta\{z : \ell^\top e^{-sH}z = c\} = 0$ for every $s \geq 0$, and therefore $\nu_H(\partial A_{\ell,c}) = 0$.

Taking $c = 1$ and applying Theorem 3.3 to $A_\ell$ gives

$$\mathbb{P}(\ell^\top\gamma T_{\gamma,n_\gamma} > u_\gamma) \sim \frac{\mathbb{P}(|\eta_0| > u_\gamma/\gamma)}{\gamma}C_\ell.$$

The deterministic shift $\ell^\top B_\gamma^{n_\gamma}Y_0$ is $o(u_\gamma)$ by Assumption 3.2. Applying the same result to $(1 \pm \varepsilon)u_\gamma$ and then letting $\varepsilon \downarrow 0$ yields (23). □

### B.4 Projected tails under one-dimensional regular variation

For a general linear functional $\ell^\top Y_n$, the direction $(B_\gamma^j)^\top\ell$ changes with $j$, so the full multivariate tail measure is the natural object. In contrast, if $v$ is a left eigenvector of $H$, the projected recursion is genuinely scalar. This allows us to replace Assumption 2.4 by a one-dimensional tail assumption on the projected innovation.

Specifically, let $v \in \mathbb{R}^d \setminus \{0\}$ be a fixed unit vector satisfying

$$H^\top v = \lambda_v v, \qquad \lambda_v > 0. \tag{50}$$

Define the original-space linearized iterate $y_n := x^\star + Y_n$. Set

$$y_n^{(v)} := \langle v, y_n - x^\star\rangle = \langle v, Y_n\rangle, \qquad \eta_n^{(v)} := \langle v, \eta_n\rangle = -\langle v, g(x^\star, \varepsilon_n)\rangle. \tag{51}$$

Then $\mathbb{E}\eta_n^{(v)} = 0$, and by (50),

$y_n^{(v)} = (1-\gamma\lambda_v)y_{n-1}^{(v)} + \gamma\eta_n^{(v)}$.

So the centered projected linear iterate is

$$S_n^{(v)} := y_n^{(v)} - \rho_{\gamma,v}^n y_0^{(v)} = \gamma \sum_{j=0}^{n-1} \rho_{\gamma,v}^j \eta_{n-j}^{(v)}, \qquad y_0^{(v)} := \langle v, x_0 - x^\star \rangle, \tag{52}$$

where $\rho_{\gamma,v} = 1-\gamma\lambda_v$. Here we also define one-big-jump intensity

$$\Lambda_\gamma(u) \;:=\; \sum_{j=0}^{n_\gamma - 1} P\Big(\gamma\rho_{\gamma,v}^j \eta_0^{(v)} > u\Big), \qquad u > 0. \tag{53}$$

**Assumption B.4** (Projected tail regular variation)**.** *Let $v \in \mathbb{R}^d \setminus \{0\}$ be fixed and define*

$$\bar{F}_v^+(u) := \mathbb{P}(\eta_0^{(v)} > u), \qquad \bar{F}_v^-(u) := \mathbb{P}(\eta_0^{(v)} < -u). \tag{54}$$

*There exist $t \in (1,2)$ and a slowly varying function $h_v$ such that*

$$\bar{F}_v^+(u) \sim h_v(u) u^{-t}, \qquad u \to \infty. \tag{55}$$

*Moreover, the left tail is dominated by the right-tail scale: there exist constants $C_- < \infty$ and $u_- < \infty$ such that*

$$\bar{F}_v^-(u) \le C_- \bar{F}_v^+(u), \qquad u \ge u_-. \tag{56}$$

**Theorem B.5** (Projected linear exact tail)**.** *Assume the left-eigenvector condition* (50) *and Assumption B.4. Let $\gamma \downarrow 0$, $n = n_\gamma \to \infty$, and $u_\gamma \downarrow 0$ satisfy*

$$0 < \gamma\lambda_v < 1, \qquad n_\gamma\gamma \to \infty, \qquad r_\gamma := \frac{u_\gamma}{\gamma} \to \infty, \qquad \frac{\bar{F}_v^+(r_\gamma)}{\gamma} \to 0, \qquad \frac{\rho_{\gamma,v}^{n_\gamma}|y_0^{(v)}|}{u_\gamma} \to 0. \tag{57}$$

*Then*

$$\mathbb{P}\left(y_{n_\gamma}^{(v)} > u_\gamma\right) \sim \Lambda_\gamma(u_\gamma) \sim \frac{1}{t\lambda_v} \frac{\bar{F}_v^+(u_\gamma/\gamma)}{\gamma} \sim \frac{1}{t\lambda_v} h_v(u_\gamma/\gamma)\gamma^{t-1}u_\gamma^{-t}. \tag{58}$$

*Equivalently,*

$$\left| \frac{\mathbb{P}(y_{n_\gamma}^{(v)} > u_\gamma)}{\Lambda_\gamma(u_\gamma)} - 1 \right| \to 0. \tag{59}$$

**Remark B.6.** *Equation* (59) *identifies the leading tail probability as a sum of single-jump probabilities.*

*The closed-form sharp constant $C_\ell$ is recovered by passing the Riemann-sum limit after replacing $\bar{F}_v^+$ with its regular-variation asymptotic: when an exact distribution function is available, one may evaluate the one-jump sum directly.*

**Remark B.7** (Relation between the vector and projected assumptions)**.** *Under the stronger Assumption 2.4, if $A_v := \{z : v^\top z > 1\}$ satisfies $\mu_\eta(A_v) > 0$, then the projected right-tail regular variation* (55) *follows with $h_v(u)u^{-t} \sim \mu_\eta(A_v)\mathbb{P}(|\eta_0| > u)$. However, the projected theorem above only needs the one-dimensional right tail of $\eta_0^{(v)}$ and the comparable left-tail bound* (56)*.*

**Remark B.8** (Power thresholds)**.** *For $u_\gamma = \gamma^a$ with $0 < a < 1$, the rare-event part of* (57) *becomes*

$$h_v(\gamma^{a-1})\gamma^{t-1-at} \to 0. \tag{60}$$

*The initialization part is implied, for example, by $n_\gamma\gamma/\log(1/\gamma) \to \infty$. Under these conditions,*

$$\mathbb{P}(y_{n_\gamma}^{(v)} > \gamma^a) \sim \frac{1}{t\lambda_v} h_v(\gamma^{a-1})\gamma^{t-1-at}. \tag{61}$$

*If $h_v(u) \to h_{v,0} \in (0,\infty)$, then* (60) *is equivalent to $a < (t-1)/t$.*

### B.5 Proof of Theorem B.5

*Proof.* For notational simplicity, write $u = u_\gamma$, $n = n_\gamma$, $\rho = \rho_{\gamma,v} = 1-\gamma\lambda_v$, $\lambda = \lambda_v$, and $\eta_j = \eta_j^{(v)}$. Put

$$Z_j := \gamma\rho^j \eta_{n-j}, \qquad 0 \le j < n,$$

so that $S_n^{(v)} = \sum_{j=0}^{n-1} Z_j$. Define the one-big-jump intensity

$$\Lambda_\gamma(u) := \sum_{j=0}^{n-1} \mathbb{P}(Z_j > u) = \sum_{j=0}^{n-1} \bar{F}_v^+\big(u/(\gamma\rho^j)\big).$$

We first establish the asymptotic

$$\Lambda_\gamma(u) \sim \frac{1}{t\lambda}\,\frac{\bar{F}_v^+(u/\gamma)}{\gamma}. \tag{62}$$

Since the multipliers $\rho^{-j}$ range over an unbounded set as $j$ varies in $\{0,\dots,n-1\}$, neither Potter's bound (Resnick, 1987) alone nor the uniform convergence theorem alone yields (62) with the sharp constant; we therefore split the sum at a cutoff $K > 0$ (assumed below to satisfy $K < n\gamma$, which holds eventually since $n\gamma \to \infty$). For $\gamma j \le K$, the values $\rho^{-j}$ lie in $[1, \rho^{-\lfloor K/\gamma \rfloor}]$, which converges to $[1, e^{\lambda K}]$ as $\gamma \downarrow 0$ and hence remains in a fixed compact subset of $[1,\infty)$ for small $\gamma$. Combining $u/\gamma \to \infty$ (from (57)) with the uniform convergence theorem for slowly varying functions (Resnick, 1987) applied to $h_v$,

$$\frac{\bar{F}_v^+(u/(\gamma\rho^j))}{\bar{F}_v^+(u/\gamma)} = \frac{h_v(u/(\gamma\rho^j))}{h_v(u/\gamma)}\,\rho^{tj} = (1+o(1))\,\rho^{tj}, \qquad \gamma \downarrow 0,$$

uniformly in $j$ with $\gamma j \le K$. Using $1-\rho^t \sim t\gamma\lambda$ and $\rho^{t\lfloor K/\gamma\rfloor} \to e^{-t\lambda K}$,

$$\lim_{\gamma\downarrow 0} \frac{\gamma}{\bar{F}_v^+(u/\gamma)} \sum_{j:\,\gamma j \le K} \bar{F}_v^+(u/(\gamma\rho^j)) = \frac{1-e^{-t\lambda K}}{t\lambda}. \tag{63}$$

For $\gamma j > K$ we use Potter's bound (Resnick, 1987): fix $\varepsilon \in (0, t-1)$, then there exist $A_\varepsilon < \infty$ and $X_\varepsilon < \infty$ such that for all $x \ge X_\varepsilon$ and $\mu \ge 1$,

$$\bar{F}_v^+(\mu x)/\bar{F}_v^+(x) \le A_\varepsilon\,\mu^{-t+\varepsilon}.$$

For $\gamma$ small enough that $u/\gamma \ge X_\varepsilon$, taking $\mu = \rho^{-j}$ gives $\bar{F}_v^+(u/(\gamma\rho^j)) \le A_\varepsilon \rho^{(t-\varepsilon)j}\bar{F}_v^+(u/\gamma)$ for every $j \ge 0$, and consequently

$$\limsup_{\gamma\downarrow 0} \frac{\gamma}{\bar{F}_v^+(u/\gamma)} \sum_{j:\,\gamma j > K} \bar{F}_v^+(u/(\gamma\rho^j)) \le \frac{A_\varepsilon\, e^{-(t-\varepsilon)\lambda K}}{(t-\varepsilon)\lambda}. \tag{64}$$

Combining (63) and (64), for every fixed $K > 0$,

$$\frac{1-e^{-t\lambda K}}{t\lambda} \le \liminf_{\gamma\downarrow 0} \frac{\gamma\Lambda_\gamma(u)}{\bar{F}_v^+(u/\gamma)} \le \limsup_{\gamma\downarrow 0} \frac{\gamma\Lambda_\gamma(u)}{\bar{F}_v^+(u/\gamma)} \le \frac{1-e^{-t\lambda K}}{t\lambda} + \frac{A_\varepsilon e^{-(t-\varepsilon)\lambda K}}{(t-\varepsilon)\lambda}.$$

Letting $K \to \infty$ with $\varepsilon$ fixed proves (62). In particular, $\Lambda_\gamma(u) \to 0$ by (57). The same argument applied to $cu$ in place of $u$, together with $\bar{F}_v^+(cu/\gamma)/\bar{F}_v^+(u/\gamma) \to c^{-t}$ (which is direct since $cu/\gamma$ and $u/\gamma$ are comparable up to a fixed ratio), yields the shift stability

$$\Lambda_\gamma(cu) \sim c^{-t}\Lambda_\gamma(u), \qquad c > 0 \text{ fixed}. \tag{65}$$

We next show that $\mathbb{P}(S_n^{(v)} > u) \sim \Lambda_\gamma(u)$. Fix $\delta \in (0, 1/2)$ and $\theta \in (0, \delta/16)$, and define the truncated sum

$$W_\theta := \sum_{j=0}^{n-1} Z_j \mathbf{1}_{\{|Z_j| \le \theta u\}}.$$

We first establish moment bounds for $W_\theta$. Set $a_j := \gamma\rho^j$ and $M_j := \theta u/a_j$, so $Z_j = a_j\eta_{n-j} \overset{d}{=} a_j\eta_0$ and $\{|Z_j| > \theta u\} = \{|\eta_0| > M_j\}$. Since $M_j \ge \theta u/\gamma \to \infty$, Karamata's theorem (Resnick, 1987) applied to $|\eta_0|$ (whose tail is $\le (1+C_-)\bar{F}_v^+$ by (56)) gives

$$\mathbb{E}\big[|\eta_0|\mathbf{1}_{\{|\eta_0|>M_j\}}\big] \le CM_j\bar{F}_v^+(M_j), \qquad \mathbb{E}\big[\eta_0^2\mathbf{1}_{\{|\eta_0|\le M_j\}}\big] \le CM_j^2\bar{F}_v^+(M_j),$$

where the second bound uses $t < 2$. Summing the corresponding bounds for $Z_j$ and using $\sum_j \bar{F}_v^+(M_j) = \Lambda_\gamma(\theta u)$,

$$\sum_j \big|\mathbb{E}[Z_j \mathbf{1}_{\{|Z_j|>\theta u\}}]\big| \le \sum_j a_j \mathbb{E}\big[|\eta_0| \mathbf{1}_{\{|\eta_0|>M_j\}}\big] \le C\theta u\, \Lambda_\gamma(\theta u),$$

$$\sum_j \mathbb{E}[Z_j^2 \mathbf{1}_{\{|Z_j|\le\theta u\}}] \le \sum_j a_j^2 \mathbb{E}\big[\eta_0^2 \mathbf{1}_{\{|\eta_0|\le M_j\}}\big] \le C(\theta u)^2 \Lambda_\gamma(\theta u).$$

Since $\mathbb{E}Z_j = 0$, $\mathbb{E}W_\theta = -\sum_j \mathbb{E}[Z_j \mathbf{1}_{\{|Z_j|>\theta u\}}]$ and $\mathrm{Var}(W_\theta) \le \sum_j \mathbb{E}[Z_j^2 \mathbf{1}_{\{|Z_j|\le\theta u\}}]$. Combining these with the shift stability $\Lambda_\gamma(\theta u) \sim \theta^{-t}\Lambda_\gamma(u)$ from (65), we conclude

$$|\mathbb{E}W_\theta| \le C\theta^{1-t} u \Lambda_\gamma(u) = o(u), \qquad \mathrm{Var}(W_\theta) \le C\theta^{2-t} u^2 \Lambda_\gamma(u), \tag{66}$$

where the $o(u)$ uses that $\Lambda_\gamma(u) \to 0$ and $\theta$ is fixed.

For the upper bound, partition the summands according to $\mathcal{L} := \{j : Z_j > \theta u\}$ and observe that on $\{S_n^{(v)} > u\}$ at least one of the following holds: some $Z_j$ exceeds $(1-\delta)u$; $|\mathcal{L}| \ge 2$; or $W_\theta > \delta u$ (since negative summands only reduce $S_n^{(v)}$). Therefore

$$\mathbb{P}(S_n^{(v)} > u) \le \Lambda_\gamma((1-\delta)u) + \mathbb{P}(|\mathcal{L}| \ge 2) + \mathbb{P}(W_\theta > \delta u). \tag{67}$$

By independence and (65), $\mathbb{P}(|\mathcal{L}| \ge 2) \le \frac{1}{2}\Lambda_\gamma(\theta u)^2 \sim \frac{1}{2}\theta^{-2t}\Lambda_\gamma(u)^2 = o(\Lambda_\gamma(u))$. For the truncated sum, $|\mathbb{E}W_\theta| \le \delta u/2$ for $\gamma$ small enough by (66), and each centered summand $Z_j \mathbf{1}_{\{|Z_j|\le\theta u\}} - \mathbb{E}[\cdot]$ is bounded by $2\theta u$. Bennett–Prokhorov's inequality (Prokhorov, 1959) therefore gives

$$\mathbb{P}(W_\theta > \delta u) \le \mathbb{P}(W_\theta - \mathbb{E}W_\theta > \delta u/2) \le \exp\left\{-\frac{\delta}{8\theta}\ \mathrm{arsinh}\left(\frac{\delta\,\theta^{t-1}}{2C\,\Lambda_\gamma(u)}\right)\right\}.$$

Since $\mathrm{arsinh}(y) \sim \log(2y)$ as $y \to \infty$ and $\Lambda_\gamma(u) \to 0$, this bound equals $\Lambda_\gamma(u)^{\delta/(8\theta)+o(1)}$. The constraint $\theta < \delta/16$ forces the exponent to exceed 1, so

$$\mathbb{P}(W_\theta > \delta u) = o(\Lambda_\gamma(u)). \tag{68}$$

Substituting (68) and (65) into (67),

$$\limsup_{\gamma\downarrow 0} \frac{\mathbb{P}(S_n^{(v)} > u)}{\Lambda_\gamma(u)} \le (1-\delta)^{-t}.$$

For the lower bound, let $A_j := \{Z_j > (1+\delta)u\}$ and $R_j := \sum_{i\ne j} Z_i$, which is independent of $Z_j$. On $A_j \cap \{R_j > -\delta u\}$ the full sum $S_n^{(v)} = Z_j + R_j$ exceeds $u$, so by Bonferroni,

$$\mathbb{P}(S_n^{(v)} > u) \ge \sum_{j=0}^{n-1} \mathbb{P}(A_j)\mathbb{P}(R_j > -\delta u) - \sum_{i<j} \mathbb{P}(A_i)\mathbb{P}(A_j),$$

and the double sum is bounded by $\frac{1}{2}\Lambda_\gamma((1+\delta)u)^2 = o(\Lambda_\gamma(u))$ by independence and (65).

It remains to show $\sup_{0\le j<n} \mathbb{P}(R_j \le -\delta u) \to 0$. Decompose the event $\{R_j \le -\delta u\}$ according to whether some summand $Z_i$ with $i \ne j$ is itself less than $-\theta u$. The probability of at least one such large negative summand is, by the union bound and the comparable left-tail condition (56),

$$\mathbb{P}(\exists\, i \ne j : Z_i < -\theta u) \le \sum_{i\ne j} \mathbb{P}(Z_i < -\theta u) \le C_-\, \Lambda_\gamma(\theta u) = O\big(\theta^{-t}\Lambda_\gamma(u)\big)$$

uniformly in $j$, which is $o(1)$ for fixed $\theta$ since $\Lambda_\gamma(u) \to 0$. On the complementary event $\{Z_i \ge -\theta u$ for all $i \ne j\}$, the sum $R_j$ coincides with the truncated sum

$$\widetilde{W}_\theta^{(j)} := \sum_{i\ne j} Z_i \mathbf{1}_{\{|Z_i|\le\theta u\}} + \sum_{i\ne j} Z_i \mathbf{1}_{\{Z_i>\theta u\}},$$

whose first piece is a bounded centered sum (after centering) and whose second piece is non-negative. Hence on this event $R_j \ge \sum_{i\ne j} Z_i \mathbf{1}_{\{|Z_i|\le\theta u\}} =: W_\theta^{(j)}$, and

$$\mathbb{P}(R_j \le -\delta u) \le \mathbb{P}(\exists\, i \ne j : Z_i < -\theta u) + \mathbb{P}\big(W_\theta^{(j)} \le -\delta u\big).$$

The stronger absolute-mean estimate

$$\sum_{i=0}^{n-1} \left|\mathbb{E}\left[Z_i \mathbf{1}_{\{|Z_i|\le\theta u\}}\right]\right| \le C\theta^{1-t} u \Lambda_\gamma(u)$$

implies, uniformly in $j$,

$$|\mathbb{E}W_\theta^{(j)}| \le C\theta^{1-t} u \Lambda_\gamma(u) = o(u).$$

Moreover, dropping one index can only decrease the variance bound:

$$\mathrm{Var}(W_\theta^{(j)}) \le C\theta^{2-t} u^2 \Lambda_\gamma(u).$$

Bennett–Prokhorov's inequality applied to $\mathbb{E}W_\theta^{(j)} - W_\theta^{(j)}$ then gives, exactly as in the upper-bound argument,

$$\mathbb{P}\big(W_\theta^{(j)} \le -\delta u\big) \le \mathbb{P}\big(\mathbb{E}W_\theta^{(j)} - W_\theta^{(j)} > \delta u/2\big) = o(\Lambda_\gamma(u)) = o(1),$$

uniformly in $j$, for $\theta < \delta/16$. Combining these two estimates,

$$\sup_{0\le j<n} \mathbb{P}(R_j \le -\delta u) \to 0,$$

and therefore

$$\mathbb{P}(S_n^{(v)} > u) \ge (1 - o(1))\Lambda_\gamma((1+\delta)u) - o(\Lambda_\gamma(u)) \sim (1+\delta)^{-t}\Lambda_\gamma(u),$$

giving $\liminf_{\gamma\downarrow 0} \mathbb{P}(S_n^{(v)} > u)/\Lambda_\gamma(u) \ge (1+\delta)^{-t}$. Letting $\delta \downarrow 0$ proves $\mathbb{P}(S_n^{(v)} > u) \sim \Lambda_\gamma(u)$, and combination with (62) gives

$$\mathbb{P}(S_n^{(v)} > u) \sim \frac{1}{t\lambda} \frac{\bar{F}_v^+(u/\gamma)}{\gamma}.$$

It remains to transfer this to $y_n^{(v)} = S_n^{(v)} + \rho^n y_0^{(v)}$. The last condition in (57) gives $\rho^n |y_0^{(v)}| = o(u)$, so for any fixed $\epsilon \in (0,1)$, eventually

$$\mathbb{P}(S_n^{(v)} > (1+\epsilon)u) \le \mathbb{P}(y_n^{(v)} > u) \le \mathbb{P}(S_n^{(v)} > (1-\epsilon)u).$$

Applying the asymptotic just established to $(1\pm\epsilon)u$ together with shift stability (65), and then letting $\epsilon \downarrow 0$, gives (58); the second equivalence follows from $\bar{F}_v^+(u/\gamma) \sim h_v(u/\gamma)\,\gamma^t u^{-t}$. □

**Relationship to the Wang–Rhee framework.** For $u_\gamma = \gamma^a$, rescaling the linearized recursion by $V_j = Y_j/u_\gamma$ gives

$$V_{j+1} = (I - \gamma H)V_j + \gamma^{1-a}\eta_{j+1}.$$

The noise exponent $1-a$ corresponds to the general-scaling framework in Theorem A.1 of Wang & Rhee (2024); its scaling restriction becomes $1 - a > 1/t$, or $a < (t-1)/t$. The two analyses therefore share this threshold-scaling restriction.

However, obtaining our last-iterate asymptotic result additionally requires control as the observation horizon diverges, including the contributions of old innovations and the initial condition. Their unclipped sample-path theorem also assumes bounded drift and noise coefficients, which are not assumed here. Furthermore, Wang & Rhee (2024) does not study general nonlinear SGD as we do in this paper. Their framework instead assumes that the stochastic noise can be factorized as the product of a function of the current iterate and an i.i.d. noise term. Our proof directly controls the matrix-weight limit, remote-past contributions, and accumulated small jumps, and removes initialization under the stated burn-in condition. These estimates identify the impulse-response constant $C_\ell$. Theorem C.3 and Proposition D.1 then provide a separate nonlinear transfer at the relative rare-event scale. Together, these estimates connect the growing observation horizon to a nonlinear final-iterate tail constant.

# C CENTRAL-SCALE MOMENTS AND NONLINEAR COMPARISON

**Lemma C.1** (Consequences of the local and global conditions)**.** *Under Assumptions 2.1–2.3, write $R(x) = H(x - x^\star) - m(x)$ and $D(x, \varepsilon) = m(x) - g(x, \varepsilon) + g(x^\star, \varepsilon)$. There are finite constants such that, for all $x$,*

$$|m(x)| \le C|x - x^\star|, \tag{69}$$

$$\|D(x, \varepsilon)\|_p \le C|x - x^\star|, \tag{70}$$

$$|R(x)| \le C\{|x - x^\star|^2 \mathbf{1}_{\{|x-x^\star|\le\kappa\}} + |x - x^\star|\mathbf{1}_{\{|x-x^\star|>\kappa\}}\}. \tag{71}$$

*Furthermore, $\|B_\gamma^j\|_{\mathrm{op}} \le e^{-c_H j\gamma}$ for small $\gamma$.*

*Proof.* Jensen's inequality, local increments, and global growth give the first two bounds as follows. Inside the local ball, $m(x^\star) = 0$ and (5) imply $|m(x)| \le L_g|x - x^\star|$ and $\|D(x, \varepsilon)\|_p \le 2L_g|x - x^\star|$. Outside this ball, global growth absorbs the constant term into $C_\kappa|x - x^\star|$. The same argument gives the linear bound on $R$ outside the ball; inside, use the local remainder. The contraction follows from Lemma B.1. □

**Lemma C.2** (Martingale moment inequality)**.** *If $(d_k, \mathcal{F}_k)$ are $\mathbb{R}^d$-valued martingale differences with finite $r$th moments, $1 < r \le 2$, then*

$$\mathbb{E}\left|\sum_{k=1}^n d_k\right|^r \le C_r \sum_{k=1}^n \mathbb{E}|d_k|^r. \tag{72}$$

*Proof.* Apply Lemma A.2 to the partial sum through $k - 1$ and $d_k$. Conditioning on $\mathcal{F}_{k-1}$ removes the linear increment term. Taking expectations and iterating proves the inequality. This argument needs the martingale-difference property, rather than independence. □

## C.1 PROOF OF THE CENTRAL-SCALE MOMENT BOUND

*Proof of Theorem 3.1.* Translate $x^\star$ to zero and write $m_0(x) = m(x^\star + x)$. The first two bounds of Lemma C.1 use only centered dissipativity and gradient regularity, so they apply without the quadratic remainder assumption. For $h > 0$, set $V_h(x) = (h^2 + |x|^2)^{p/2}$. Its Hessian satisfies

$$\nabla^2 V_h(x) = p(h^2+|x|^2)^{(p-2)/2} I_d + p(p-2)(h^2+|x|^2)^{(p-4)/2} xx^\top, \qquad 0 \preceq \nabla^2 V_h(x) \preceq ph^{p-2} I_d.$$

The gradient is also $(p - 1)$-Hölder with a constant independent of $h$. Indeed, for $w \ne 0$ put $e = w/|w|$. Along the segment $x + \theta w$, its Hessian norm is at most $p|\langle x, e\rangle + \theta|w||^{p-2}$, whose integral over $0 \le \theta \le 1$ is at most $p2^{2-p}|w|^{p-2}/(p - 1)$. Integrating the gradient difference gives the uniform Taylor bound

$$0 \le V_h(x + w) - V_h(x) - \langle \nabla V_h(x), w\rangle \le C_p \min\{h^{p-2}|w|^2, |w|^p\}. \tag{73}$$

The scalar function $\phi_h(r) = \min(h^{p-2}r^2, r^p)$ is increasing and obeys $\phi_h(2r) \le 4\phi_h(r)$. Consequently,

$$\phi_h(|u + v|) \le 4\phi_h(|u|) + 4\phi_h(|v|) \le 4\phi_h(|u|) + 4|v|^p. \tag{74}$$

Write $Z = |\eta|$ and $\bar F(r) = \mathbb{P}(Z > r)$. The generalized inverse in (17) satisfies $\bar F(b_\gamma) \sim \gamma$; monotone regular variation with negative index ensures that relative jumps vanish, so atoms cause no difficulty. Karamata's integral theorem (Bingham et al., 1987), with $1 < p < t < 2$, yields

$$\mathbb{E}[Z^2; Z \le b] \sim \frac{t}{2 - t} b^2 \bar F(b), \qquad \mathbb{E}[Z^p; Z > b] \sim \frac{t}{t - p} b^p \bar F(b).$$

These formulas follow directly by integrating $\bar F$ on $(0, b)$ and $(b, \infty)$, respectively. At $h = s_\gamma = \gamma b_\gamma$, the two branches of $\phi_h(\gamma Z)$ meet at $Z = b_\gamma$. Thus

$$\mathbb{E}\phi_{s_\gamma}(\gamma Z) \sim \left(\frac{t}{2 - t} + \frac{t}{t - p}\right) s_\gamma^p \bar F(b_\gamma) = O(\gamma s_\gamma^p). \tag{75}$$

Given the past and $X_{n-1} = x$, set $F = x - \gamma m_0(x)$. Dissipativity and $|m_0(x)| \le C|x|$ give $|F|^2 \le (1 - \alpha\gamma)|x|^2$ for small $\gamma$. Concavity of $z^{p/2}$ and $|x|^2(h^2 + |x|^2)^{p/2-1} \ge V_h(x) - h^p$ imply

$$V_h(F) \le (1 - c_0\gamma)V_h(x) + c_0\gamma h^p, \qquad c_0 = p\alpha/2. \tag{76}$$

The next iterate is $F + \gamma(\eta_n + D(x_{n-1}, \varepsilon_n))$. Both noise terms have conditional mean zero, although they may be dependent. Apply (73) at the past-measurable point $F$. The linear Taylor term has conditional expectation zero. Equations (74)–(76) and (70) give

$$\begin{aligned}\mathbb{E}[V_{s_\gamma}(X_n) \mid \mathcal{F}_{n-1}] &\le (1 - c_0\gamma)V_{s_\gamma}(X_{n-1}) + C\gamma s_\gamma^p + C\gamma^p|X_{n-1}|^p \\ &\le (1 - c\gamma)V_{s_\gamma}(X_{n-1}) + C\gamma s_\gamma^p.\end{aligned} \tag{77}$$

The last step absorbs $C\gamma^p|X_{n-1}|^p$ into the drift because $p > 1$. Iterating and using $V_{s_\gamma}(X_0) \le s_\gamma^p + |X_0|^p$ proves

$$\mathbb{E}|X_n|^p \le (1 - c\gamma)^n|X_0|^p + Cs_\gamma^p.$$

No finite second moment is used, and the initialization term vanishes exactly at $X_0 = 0$. □

### C.2 Comparison at lower moments

**Coupled error decomposition.** Subtracting the recursions gives

$$\Delta_n = \gamma\sum_{k=1}^{n} B_\gamma^{n-k}R(x_{k-1}) + \gamma\sum_{k=1}^{n} B_\gamma^{n-k}D(x_{k-1}, \varepsilon_k) = I_{1,n} + I_{2,n}. \tag{78}$$

Here $R$ and $D$ are defined in Lemma C.1. The first term contains the population remainder; the second is a martingale convolution.

**Theorem C.3** (Comparison without a radial tail assumption)**.** *Under Assumptions 2.1–2.3, for every $1 \le q \le p$ there are $C_q, c_q, \gamma_q > 0$ such that*

$$\mathbb{E}|\Delta_n|^q \le C_q\gamma^{p-1} + C_q e^{-c_q n\gamma}(|X_0|^p + |X_0|^q), \qquad n \ge 1, \quad 0 < \gamma < \gamma_q. \tag{79}$$

*Proof of Theorems 3.5 and C.3.* Use $w_\gamma = s_\gamma^p$ under the assumptions of Theorem 3.5, and $w_\gamma = \gamma^{p-1}$ for Theorem C.3. The corresponding moment theorem gives $\mathbb{E}|X_k|^p \le Cw_\gamma + Ce^{-c_p k\gamma}|X_0|^p$. For $q \in [1, p]$, (71) implies $|R(x)|^q \le C_q|x - x^\star|^p$: locally $2q \ge p$, and outside the local ball $q \le p$. Minkowski's inequality in (78) yields

$$\|I_{1,n}\|_q \le C_q\gamma\sum_{k=1}^{n} e^{-c_H(n-k)\gamma}\{w_\gamma^{1/q} + e^{-c_p(k-1)\gamma/q}|X_0|^{p/q}\}.$$

The convolution of two decaying exponentials is bounded by $Ce^{-cn\gamma}$ for any $c$ strictly below their decay rates; a coincident-rate factor $n\gamma$ is absorbed by reducing $c$. Therefore, including at $q = 1$,

$$\mathbb{E}|I_{1,n}|^q \le C_q w_\gamma + C_q e^{-c_q n\gamma}|X_0|^p. \tag{80}$$

Let $\mathcal{F}_k = \sigma(\varepsilon_1, \dots, \varepsilon_k)$. Fresh innovations and unbiasedness imply

$$\mathbb{E}[D(x_{k-1}, \varepsilon_k) \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[|D(x_{k-1}, \varepsilon_k)|^p \mid \mathcal{F}_{k-1}] \le C|X_{k-1}|^p.$$

The deterministic weights in $I_{2,n}$ preserve the martingale-difference property. Apply Lemma C.2 at order $p$:

$$\begin{aligned}\mathbb{E}|I_{2,n}|^p &\le C\gamma^p\sum_{k=1}^{n} e^{-pc_H(n-k)\gamma}\{w_\gamma + e^{-c_p(k-1)\gamma}|X_0|^p\} \\ &\le C\gamma^{p-1}w_\gamma + C\gamma^{p-1}e^{-cn\gamma}|X_0|^p.\end{aligned} \tag{81}$$

Lyapunov's inequality now gives

$$\mathbb{E}|I_{2,n}|^q \le C_q\gamma^{q(p-1)/p}w_\gamma^{q/p} + C_q\gamma^{q(p-1)/p}e^{-c_q n\gamma}|X_0|^q. \tag{82}$$

For $w_\gamma = \gamma^{p-1}$, the first term is $C_q\gamma^{2q(p-1)/p} = o(w_\gamma)$ because $2q > p$. For $w_\gamma = s_\gamma^p$, its ratio to $w_\gamma$ is regularly varying at zero with exponent

$$e(q) = \frac{q(p-1)}{p} + (q - p)(1 - 1/t) \ge e(1) = (p - 1)(1/p + 1/t - 1) > 0.$$

Thus it is again $o(w_\gamma)$, including all slowly varying factors. Combining (80) and (82), with $\gamma \le 1$, proves both claims. □

## D TRANSFER OF VECTOR AND OBJECTIVE-VALUE TAILS

**Proposition D.1** (General-threshold transfer)**.** *Suppose the linear vector tail theorem holds at $u_\gamma$ and at fixed multiples of $u_\gamma$. Suppose also that, for some $q \geq 1$,*

$$\mathbb{E}|\Delta_{n_\gamma}|^q = o\{u_\gamma^q L_\gamma(u_\gamma)\}.$$

*Then* (11) *holds for every $A$ satisfying* (10)*.*

*Proof.* For $\delta > 0$, define $A^\delta = \{z : \mathrm{dist}(z, A) \leq \delta\}$ and $A_{-\delta} = \{z \in A : \mathrm{dist}(z, A^c) > \delta\}$. Take $\delta < \mathrm{dist}(A, 0)/2$. The inclusions induced by $X = Y + \Delta$ give

$$\mathbb{P}(Y/u \in A_{-\delta}) - \mathbb{P}(|\Delta| > \delta u) \leq \mathbb{P}(X/u \in A) \leq \mathbb{P}(Y/u \in A^\delta) + \mathbb{P}(|\Delta| > \delta u).$$

Markov's inequality makes the error $o(L_\gamma(u))$ for each fixed $\delta$. Choose $\delta$ decreasing to zero through values for which the enlarged and contracted sets are continuity sets. Such values exist because the distance level sets are disjoint and the measure is finite away from zero. Divide by $L_\gamma(u)$, take upper and lower limits, and then let $\delta$ decrease to zero. Continuity from above and below, together with $\nu_H(\partial A) = 0$, gives the result, including the case $\nu_H(A) = 0$. □

*Proof of Theorem 2.5.* Let $u_\gamma = c\gamma^a$ with the exponents in (9). Then $r_\gamma \to \infty$ and $L_\gamma(u_\gamma) \to 0$ by regular variation and Potter bounds. First suppose $p < t$. If $X_0 = 0$, both the linear initial shift and the exponential terms in (24) vanish. Otherwise, the superlogarithmic time condition makes the initial shift and each exponential error smaller than every positive power of $\gamma$. Hence Assumption 3.2 holds. Regular variation gives

$$\begin{aligned}\frac{s_\gamma^p}{u_\gamma L_\gamma(u_\gamma)} &= \gamma^{p(1-1/t)-(t-1)+a(t-1)+o(1)} \\ &= \gamma^{(t-1)(a+p/t-1)+o(1)} \longrightarrow 0,\end{aligned}$$

where the exponent is positive because $a > (t-p)/t$. Theorem 3.5 with $q = 1$, together with the initialization bounds above, therefore yields

$$\mathbb{E}|\Delta_{n_\gamma}| = o\{u_\gamma L_\gamma(u_\gamma)\}.$$

Apply Theorem 3.3 and Proposition D.1. If the full gradient conditions hold at the finite endpoint $p = t$, choose $\max\{1, t(1-a)\} < p' < t$ so that the strict threshold window still holds. The gradient conditions descend to $p'$ by monotonicity of moment norms, giving the same tail conclusion. □

*Proof of Corollary 2.6.* For $H^\top v = \lambda_v v$, homogeneity gives $C_v = \mu_\eta(A_v)\int_0^\infty e^{-t\lambda_v s}\mathrm{d}s = \mu_\eta(A_v)/(t\lambda_v)$. Regular variation also gives $\mathbb{P}(v^\top\eta > r) \sim \mu_\eta(A_v)\mathbb{P}(|\eta| > r)$. This proves the assertion. □

**Proposition D.2** (Nonlinear eigenprojection with only projected regular variation)**.** *Retain Assumptions 2.1–2.3 and suppose a unit eigenvector $v$ satisfies Assumption B.4 with index $t$. If*

$$2 - 1/t < p \leq t, \qquad \frac{t-p}{t-1} < a < \frac{t-1}{t}, \qquad u_\gamma = c\gamma^a, \quad c > 0,$$

*and the initialization conditions of Theorem 2.5 hold, then* (14) *holds.*

*Proof.* Apply Theorem B.5 to $v^\top Y_n$. The $q = 1$ bound in Theorem C.3, divided by $u_\gamma\mathbb{P}(v^\top\eta > u_\gamma/\gamma)/\gamma$, has power exponent $p - t + a(t-1) > 0$. The scalar sandwich at thresholds $(1 \pm \delta)u_\gamma$ proves the claim. The lower and upper bounds on $a$ define a nonempty interval exactly when $p > 2 - 1/t$. □

This result needs no radial tail control for other coordinates. It therefore uses the moment-growth comparison, whereas the wider main-text window uses the radial regular variation in Assumption 2.4.

*Proof of Corollary 2.7.* Positive definiteness of $H$ places $A_M$ between two norm-exceedance sets, so $0 < C_M < \infty$. The quadratic homogeneity implies $\nu_H\{z : z^\top Hz/2 > s\} = s^{-t/2}C_M$ for $s > 0$. Consequently the boundary of $A_M$ has zero mass. Locally,

$$M(x^\star + x) - M(x^\star) = \tfrac{1}{2}x^\top Hx + O(|x|^3).$$

For a fixed $R$, the remainder divided by $u_\gamma^2$ vanishes uniformly on $|x| \le Ru_\gamma$. Sandwich the objective exceedance event on this ball between quadratic exceedances at $1 \pm \delta$. Theorem 2.5 applies to the truncated continuity sets. The omitted region satisfies

$$\lim_{\gamma\downarrow 0} \frac{\mathbb{P}(|X_{n_\gamma}| > Ru_\gamma)}{L_\gamma(u_\gamma)} = \nu_H\{|z| > R\} = R^{-t}\nu_H\{|z| > 1\}.$$

Let $\delta$ decrease to zero and $R$ increase to infinity. This proves (15) without asserting uniform Taylor approximation on an unbounded set. □

# E Extensions and parameter interpretation

## E.1 Hölder local remainders

**Theorem E.1** (Superlinear local remainder)**.** *Replace Assumption 2.3 by differentiability at $x^\star$ and $|m(x) - H(x - x^\star)| \le C|x - x^\star|^\beta$ for $|x - x^\star| \le \kappa$, where $1 < \beta \le 2$. Retain the other assumptions. Let $q_* = \max(1, p/\beta)$. For $1 < p < t$,* (24) *holds for $q \in [q_*, p]$. The nonlinear vector, projection, and objective-value tail results hold under*

$$\frac{(t-1)(t-p)}{t(t-q_*)} < a < \frac{t-1}{t} \tag{83}$$

*and the same initialization conditions.*

*Proof.* For $q \ge q_*$ the local relation $\beta q \ge p$ gives $|R(x)|^q \le C|x - x^\star|^p$. Outside the ball, $q \le p$ and the global linear bound give the same inequality. Thus the deterministic convolution calculation is unchanged. The martingale $p$th-moment calculation does not use $\beta$. This proves the error bound, including $q = q_*$ and $q = p$. Use $q = q_*$ in Markov's inequality. The relative-error exponent becomes $p(1 - 1/t) - (t - 1) + a(t - q_*)$, which is positive in the stated window. The interval is nonempty exactly when $p > q_*$, true for every $p > 1$ and $\beta > 1$. Proposition D.1 proves the vector conclusion. The projection is a half-space instance of the vector conclusion. For a finite $p = t$ moment, lower the moment order slightly as in the proof of Theorem 2.5; continuity of $q_* = \max(1, p/\beta)$ preserves every strict window. For the objective tail, the local Taylor remainder is now $O(|x|^{1+\beta}) = o(|x|^2)$, sufficient for the same compact-truncation proof. □

**A model outside the quadratic-remainder class.** Let $1 < \beta < 2$, $c > 0$, and

$$m(x) = x + c\frac{x|x|^{\beta-1}}{1 + |x|^{\beta-1}}, \qquad g(x, \varepsilon) = m(x) - \varepsilon, \qquad \varepsilon \sim t_t.$$

Here $xm(x) \ge x^2$, $m'(0) = 1$, and $1 \le m'(x) \le 1 + c$. The stochastic gradient satisfies the growth and increment assumptions for every $1 < p < t$. Its local remainder is $-c\,\mathrm{sign}(x)|x|^\beta(1 + o(1))$, which is not $O(|x|^2)$. This example establishes a strict regularity extension; its objective is strongly convex. Taking $M(x) = \int_0^x m(s)\mathrm{d}s$ and $G(x, \varepsilon) = M(x) - \varepsilon x$ gives an integrable sample loss.

**Available moment and threshold exponents.** For $\beta = 2$, $t = 1.5$ and $p = 1.35$ allow $0.1 < a < 1/3$. For $t = 1.5$ and $p = 1.45$, $\beta = 1.1$ permits $11/120 < a < 1/3$, while $\beta = 1.2$ permits $2/35 < a < 1/3$. The examples $a = .30$ and $a = .25$ have positive remainder exponents $5/132$ and $9/160$, respectively. These sufficient exponents establish asymptotic validity, not fast finite-step convergence. If the full gradient assumptions hold for every $p < t$, taking $p$ close enough to $t$ proves validity for every fixed $0 < a < (t-1)/t$ and every $\beta > 1$. Merely knowing that the target innovation has these moments is insufficient for this last statement.

### E.2 FIXED MINI-BATCHES

**Corollary E.2** (Fixed-batch scaling)**.** *Let $b$ be a fixed positive integer and use the average of $b$ independent fresh stochastic gradients at each step. Under the corresponding conditions of Theorem 2.5 or E.1,*

$$\mathbb{P}(X_{n_\gamma}^{(b)}/u_\gamma \in A) = b^{1-t} L_\gamma(u_\gamma)\{\nu_H(A) + o(1)\},$$

*where $L_\gamma$ and $\mu_\eta$ use the single-sample innovation normalization.*

*Proof.* Finite sums of independent regularly varying vectors have the sum of their tail measures. Rescaling the sum by $b$ therefore gives tail measure $b \cdot b^{-t}\mu_\eta$. The batch average preserves the population gradient, centered dissipativity, and local remainder. Minkowski's inequality preserves the required growth and increment bounds. Apply the nonlinear tail theorem with the batch innovation and rewrite its radial normalization in terms of the original innovation. □

At a common admissible threshold sequence, $(\gamma, b)$ and $(\gamma/b, 1)$ have the same leading tail. Indeed, regular variation gives $L_{\gamma/b}(u_\gamma) \sim b^{1-t} L_\gamma(u_\gamma)$. For a shared sample budget $S_\gamma$, their effective times are both asymptotic to $S_\gamma \gamma/b$. This comparison requires that the budget grows enough to satisfy the applicable initialization conditions. Equivalently, $(b\gamma, b)$ matches $(\gamma, 1)$. This fixed-$b$ result does not interchange the limits in $\gamma$ and $b$.

### E.3 EXTREME OBJECTIVE QUANTILES

Under Theorem E.1 and its horizon conditions, suppose $\mathbb{P}(|\eta| > r) \sim c_\eta r^{-t}$ and let $\delta_\gamma = \gamma^\zeta$ with

$$0 < \zeta < \frac{(t-1)(p-q_*)}{t-q_*}.$$

The excess-objective quantile at probability $1 - \delta_\gamma$ satisfies

$$Q_{1-\delta_\gamma} \sim \left(\frac{C_M c_\eta \gamma^{t-1}}{\delta_\gamma}\right)^{2/t}.$$

To prove this, apply the risk-tail result at fixed multiples of the proposed quantile and use monotonicity. The implied exponent $a = (t-1-\zeta)/t$ lies in the valid window. For a general slowly varying factor, the corresponding asymptotic inverse must be retained. The result concerns extreme quantiles whose tail probabilities vanish.

### E.4 CONDITIONAL SIZES AND DIRECTIONS

Put $D_1 = \{z : |z| > 1\}$ and $c_H = \nu_H(D_1) > 0$. Homogeneity gives $\nu_H\{|z| > r\} = r^{-t} c_H$ and $\nu_H\{|z| = r\} = 0$ for every $r > 0$. Dividing Theorem 2.5 for $A \cap D_1$ by its conclusion for $D_1$ gives

$$\mathcal{L}(X_{n_\gamma}/u_\gamma \mid |X_{n_\gamma}| > u_\gamma) \implies \frac{\nu_H(\cdot \cap D_1)}{c_H}.$$

Indeed, this holds for every continuity set of the probability measure on the right, which characterizes weak convergence. The continuous map $z \mapsto z/|z|$ on $D_1$ gives the conditional direction law. In particular, for a Borel set $C \subset S^{d-1}$ with boundary of zero angular mass, its probability is $\nu_H\{z : |z| > 1, z/|z| \in C\}/c_H$. These are conditional directional probabilities; the constants of separate eigenvector projections are marginal quantities.

### E.5 AFFINE POPULATION GRADIENTS AT FIXED TOLERANCES

*Proof of Corollary 4.1.* Because $R \equiv 0$, the decomposition (78) has only its martingale term. Theorem 3.1 and (81) imply

$$\mathbb{E}|\Delta_n|^p \le C\gamma^{p-1} s_\gamma^p + C\gamma^{p-1} e^{-cn\gamma} |X_0|^p. \tag{84}$$

For a fixed $u > 0$, $L_\gamma(u)$ is regularly varying at zero with exponent $t - 1$. The ratio of the first term in (84) to $u^p L_\gamma(u)$ has exponent

$$(p-1) + p(1-1/t) - (t-1) = p(2-1/t) - t > 0.$$

The strict inequality absorbs all slowly varying factors. The initialization term vanishes at $X_0 = 0$; otherwise the stated superlogarithmic horizon makes it negligible. For $u_\gamma \equiv u$, Assumption 3.2 holds: $u/\gamma \to \infty$, $L_\gamma(u) \to 0$, and the initial linear shift is negligible. The linear proof uses these conditions on $u/\gamma$ and $L_\gamma(u)$, with no requirement that $u$ itself decrease. Apply Theorem 3.3, Markov's inequality at order $p$, and Proposition D.1. If the complete gradient conditions hold at $p = t$, choose any $p' \in (t^2/(2t-1), t)$ and apply the preceding argument at order $p'$. □

For least squares with a bounded random design $Z$, $H = \mathbb{E}ZZ^\top \succ 0$, and independent centered regularly varying label noise $\epsilon$, the gradient $g(x; Z, \epsilon) = Z\{Z^\top(x - x^\star) - \epsilon\}$ has mean $H(x - x^\star)$. The sample loss contrast $\frac{1}{2}[(Z^\top(x - x^\star) - \epsilon)^2 - \epsilon^2]$, with subtraction performed before expectation, has finite population objective $\frac{1}{2}(x - x^\star)^\top H(x - x^\star)$ even when $\mathbb{E}\epsilon^2 = \infty$. Boundedness of $Z$ gives the complete gradient conditions at every $p < t$. Conditioning on $Z$ and applying regular variation with a dominated bound gives multivariate regular variation of $Z\epsilon$, provided its tail measure is nonzero. Thus the corollary includes random sample Hessians. The cutoff on $p$ is a sufficient condition from the comparison bound; at $t = 1.5$ it is $p > 9/8$.

### E.6 Finite observation horizons

Let $X_0 = 0$ and $n_\gamma \gamma \to T \in (0, \infty)$. Under the remaining assumptions and threshold window of Theorem 2.5, define

$$\nu_{H,T}(A) = \int_0^T \mu_\eta\{z : e^{-sH} z \in A\}\, \mathrm{d}s.$$

For every Borel $A$ bounded away from zero with $\nu_{H,T}(\partial A) = 0$,

$$\mathbb{P}(X_{n_\gamma}/u_\gamma \in A) = L_\gamma(u_\gamma)\{\nu_{H,T}(A) + o(1)\}. \tag{85}$$

To prove this, use the same triangular array as in Appendix B. Its one-jump Riemann sum now ends at $T$, while the truncated small-jump bounds remain valid because all finite-horizon sums are bounded by the corresponding infinite sums of contracted weights. The time-zero boundary contributes a term of order $\gamma$ after normalization and hence vanishes. The uniform-in-$n$ bound of Theorem 3.5 has no initial term, so the identical set sandwich transfers the limit to $X$. The case $p = t$ again follows by lowering the moment slightly. For a unit eigenvector $v$, radial homogeneity gives

$$C_v(T) = \frac{\mu_\eta\{z : v^\top z > 1\}}{t\lambda_v}(1 - e^{-t\lambda_v T}) = C_v(1 - e^{-t\lambda_v T}).$$

### E.7 Localization of the nonlinear flow

Assume in addition that $m$ is locally Lipschitz on $\mathbb{R}^d$. Let $\varphi_s(z)$ solve $\dot{\varphi}_s = -m(x^\star + \varphi_s)$ with $\varphi_0 = z$. Centered dissipativity gives $|\varphi_s(z)| \le e^{-\alpha s}|z|$, so the unique solution exists for all $s \ge 0$. For $u > 0$, define

$$C_A^{\mathrm{flow}}(u) = \int_0^\infty \mu_\eta\{z : \varphi_s(z) \in uA\}\, \mathrm{d}s. \tag{86}$$

**Proposition E.3** (Deterministic localization)**.** *Under Assumptions 2.1, 2.3, and 2.4, and the local Lipschitz condition above,*

$$u^t C_A^{\mathrm{flow}}(u) \longrightarrow \nu_H(A), \qquad u \downarrow 0,$$

*for every Borel set $A$ with $\mathrm{dist}(A, 0) > 0$ and $\nu_H(\partial A) = 0$. The statement also holds for the superlinear remainder in Theorem E.1.*

*Proof.* Write $\psi_u(s, z) = \varphi_s(uz)/u$. Homogeneity changes $u^t C_A^{\mathrm{flow}}(u)$ into $\int_0^\infty \int \mathbf{1}_{\{\psi_u(s,z) \in A\}}\, \mu_\eta(\mathrm{d}z)\, \mathrm{d}s$. For fixed $(s, z)$, contraction keeps the path inside the local ball for small $u$. The rescaled equation has the form $\dot{\psi}_u = -H\psi_u + \rho_u(\psi_u)$ with $|\rho_u(\psi_u)| \le Cu^{\beta-1}|z|^\beta$, where $\beta = 2$ in the quadratic case. Variation of constants therefore gives $|\psi_u(s, z) - e^{-sH} z| \le Cu^{\beta-1}|z|^\beta s \to 0$. The exceptional points whose limits lie on $\partial A$ have product-measure mass $\nu_H(\partial A) = 0$. If $d_A = \mathrm{dist}(A, 0) > 0$, the indicators are dominated by $\mathbf{1}_{\{|z| \ge d_A e^{\alpha s}\}}$. The integral of this envelope is $d_A^{-t}/(\alpha t) < \infty$. Dominated convergence proves the assertion. □

For a scalar restoring drift and a positive-tail set $A = (1, \infty)$, changing the order of integration over shock size and time gives

$$C^{\mathrm{flow}}_{(1,\infty)}(u) = \mu_\eta(1,\infty) \int_u^\infty \frac{r^{-t}}{m(x^\star + r)}\,\mathrm{d}r.$$

The normalized factor is thus $tHu^t \int_u^\infty r^{-t}/m(x^\star + r)\,\mathrm{d}r$. Figure 1(b) compares this factor with the saved paired endpoints; Appendix I gives the scalar studies. Establishing the corresponding fixed-tolerance stochastic tail formula for general nonlinear drifts requires a separate theorem.

For the finite-step check in Figure 1, put $\psi_\gamma(x) = x - \gamma m(x)$ and $z_j = \psi_\gamma^{-j}(u)$. The displayed stepsizes make $\psi_\gamma$ increasing and preserve the positive half-line. The diagnostic is the ratio of $\sum_{j<n} \mathbb{P}\{\gamma\sigma\varepsilon > z_j\}$ to $\sum_{j<n} \mathbb{P}\{\gamma\sigma\varepsilon > u(1-\gamma)^{-j}\}$, using the exact Student survival function and all $n$ shock ages. Across Figure 1's grid it differs from the continuous factor by at most $0.00063$; at the six prespecified points its discrepancy from $R_{\mathrm{tr}}$ is at most $0.00813$. The supplement supplies both calculations and independent quadrature and root-solver checks.

### E.8 Different innovation and state-dependent moment indices

Consider $g(x; \eta, \zeta) = m(x) + x\zeta - \eta$, where $m(x) = x + 4\sin^3(x)/(1 + x^2)$ and independent centered Student variables $\eta$ and $\zeta$ have tail indices $1.5$ and $1.25$, respectively. The innovation at $x^\star = 0$ is $\eta$, whereas the full gradient has moments only below $1.25$ away from zero. The regularity assumptions hold at $p = 1.2$, and Theorem 2.5 gives $0.2 < a < 1/3$. The projected-only sufficient condition $p > 2 - 1/t = 4/3$ does not apply to this example. At a fixed positive stepsize, any nondegenerate invariant distribution has infinite $1.25$th moment: conditional on any nonzero state, the next update contains a nonzero multiple of $\zeta$, while the additive innovation has a finite $1.25$th moment. This observation concerns remote tails at fixed stepsize; the theorem concerns the stated joint limit.

### E.9 When the expected objective is infinite

For additive innovations, $X_n = X_{n-1} - \gamma m(x^\star + X_{n-1}) + \gamma\eta_n$. Regular variation with $t < 2$ gives $\mathbb{E}|\eta_n|^2 = \infty$. Conditional on the past, the other term is a fixed finite vector, so $\mathbb{E}(|X_n|^2 \mid \mathcal{F}_{n-1}) = \infty$ for every $n \geq 1$. Since centered dissipativity implies $M(x^\star + x) - M(x^\star) \geq \alpha|x|^2/2$, the expected excess objective is infinite. The tail probabilities and quantiles in Corollary 2.7 remain finite and quantify the error distribution.

## F Shrinking-threshold and objective-value experiments

**Design and simulation unit.** The study uses the scalar model in (28). For each of the three stepsizes, we generate three batches of $200{,}000$ independent trajectories, totaling $600{,}000$ terminal pairs. Each trajectory starts at zero and follows

$$X_{k+1} = X_k - \gamma m(X_k) + \gamma\sigma\varepsilon_{k+1}, \qquad Y_{k+1} = (1-\gamma)Y_k + \gamma\sigma\varepsilon_{k+1}.$$

The innovations are shared within a pair and independent between trajectories. Horizon lengths are $2303$, $9682$, and $34539$ for $\gamma = 0.01$, $0.003$, and $0.001$, respectively. Both $a = 0.2$ and $a = 0.3$ use the same saved terminal states; there is no separate simulation per threshold. The simulation design, sample counts, seeds, and two threshold exponents were fixed before production sampling. No stopping rule depends on exceedance counts or observed calibration.

Figure 1 adds a descriptive sweep of the same saved endpoints over 241 equally spaced thresholds in $[0.005, 0.1]$. At each threshold, all three exceedance counts are recomputed directly from the saved paired endpoints, without smoothing or interpolating empirical probabilities. The common plotting range is $u/\sigma \in [0.25, 5]$ and the common vertical range includes every pointwise interval. The smallest nonlinear exceedance count across this grid is 360.

**Structural certificates.** With $h(x) = \sin^3(x)/(1 + x^2)$, $xh(x) \geq 0$ for $|x| \leq \pi$ and $h(x)/x \geq -1/\{\pi(1 + \pi^2)\}$ outside that interval. Consequently $xm(x) \geq [1 - 4/\{\pi(1 + \pi^2)\}]x^2 > 0.882x^2$. The derivative at $x = 2$ is approximately $-0.307$, establishing nonconvexity. The population

remainder at zero is cubic. Bounded $h'$ and Student moments give the global growth and local increment assumptions for every $1 < p < 1.5$. Choosing $p = 1.45$ makes the quadratic-remainder window $1/30 < a < 1/3$, containing both design exponents.

**Deterministic denominators.** Writing $\overline{T}_t$ for the Student survival function,

$$\Lambda_\gamma(u) = \sum_{j=0}^{n-1} \overline{T}_{1.5}\left(\frac{u}{\gamma\sigma(1-\gamma)^j}\right), \qquad A_\gamma(u) = \frac{\overline{T}_{1.5}(u/(\gamma\sigma))}{1.5\gamma}.$$

Both quantities are evaluated using the exact survival function. The second expression is the theorem's first-order formula, still retaining the exact target-innovation tail. The reported $R_{\mathrm{const}}$ isolates the finite-sum correction.

**Intervals and coupling.** Let $K_X$, $K_Y$, and $K_{XY}$ count the events $X > u$, $Y > u$, and their intersection among $N$ independent pairs. The intervals for $R_{\mathrm{tot}}$ and $R_{\mathrm{lin}}$ divide a Wilson interval by the deterministic $\Lambda_\gamma$. For the transfer ratio, set $\widehat{p}_X = K_X/N$, $\widehat{p}_Y = K_Y/N$, and $\widehat{p}_{XY} = K_{XY}/N$. The delta-method variance of its logarithm is

$$\widehat{v} = \frac{1}{N}\left(\frac{1}{\widehat{p}_X} + \frac{1}{\widehat{p}_Y} - \frac{2\widehat{p}_{XY}}{\widehat{p}_X\widehat{p}_Y}\right).$$

We use $R_{\mathrm{tr}} \exp(\pm 1.96\sqrt{\widehat{v}})$. This keeps the within-pair covariance and uses $N$ independent units. The intervals across different thresholds are dependent and are interpreted pointwise.

Table 4: Complete prespecified shrinking-threshold results. $K_X$ is the one-sided nonlinear exceedance count among 600,000 independent endpoints. The final column is the 95% pointwise interval for $R_{\mathrm{tot}}$.

| $\gamma$ | $a$ | $K_X$ | $R_{\mathrm{lin}}$ | $R_{\mathrm{tr}}$ | $R_{\mathrm{tot}}$ | 95% CI |
|---|---|---|---|---|---|---|
| 0.01 | 0.2 | 2529 | 1.034 | 0.920 | 0.951 | [0.914, 0.988] |
| 0.01 | 0.3 | 5420 | 1.069 | 0.956 | 1.021 | [0.994, 1.048] |
| 0.003 | 0.2 | 2030 | 1.043 | 0.932 | 0.972 | [0.931, 1.016] |
| 0.003 | 0.3 | 5192 | 1.071 | 0.972 | 1.041 | [1.013, 1.069] |
| 0.001 | 0.2 | 1588 | 0.987 | 0.961 | 0.948 | [0.903, 0.996] |
| 0.001 | 0.3 | 4883 | 1.054 | 0.982 | 1.034 | [1.006, 1.064] |

**Exact objective event.** The objective is even and strictly increasing on the positive half-line. For each $u$, we solve $M(r_u) = u^2$ using adaptive quadrature and a bracketed scalar root solver, and evaluate the event as $|X_n| > r_u$. This is an exact reformulation of the nonlinear objective event up to the reported root-solver tolerance. With $H = 1$ and symmetric scalar innovations, $\nu_H\{|z| > r\} = r^{-t}/t$. Hence $C_M = 2^{-t/2}/t$ and the denominator is $A_{M,\gamma} = 2^{-t/2}\mathbb{P}(|\sigma\varepsilon| > u/\gamma)/(t\gamma)$. The risk intervals use the binomial Wilson construction with the actual $N$.

Figures 3 and 4 give the complete transfer, constant, and objective comparisons for the two prespecified sequences.

There are 1588–5420 one-sided nonlinear exceedances among 600,000 independent endpoints in the six settings. All six paired transfer intervals lie below one, with point estimates moving toward one as the stepsize decreases. For $a = 0.2$, the total ratios are 0.951, 0.972, and 0.948; for $a = 0.3$, they are 1.021, 1.041, and 1.034. Thus the total error combines the opposite signs of the nonlinear-transfer and linear single-jump discrepancies in several settings. The exact-CDF sum divided by the asymptotic constant ranges from 1.000250 to 1.002522. Its small discrepancy cannot explain the larger observed differences from one.

The exact objective-event counts at decreasing stepsizes are 2791, 2342, and 1806 for $a = 0.2$, and 6128, 5971, and 5574 for $a = 0.3$. The corresponding objective-tail ratios are 0.884, 0.944, and 0.907, and 0.973, 1.007, and 0.993. The maximum absolute residual of the numerical equation $M(r_u) = u^2$ is $1.3 \times 10^{-16}$. All counts, denominators, inverse thresholds, and intervals are supplied as machine-readable tables. The objective result is accurate for the faster-shrinking sequence at the smallest step, while the slower-shrinking sequence retains a visible finite-threshold discrepancy.

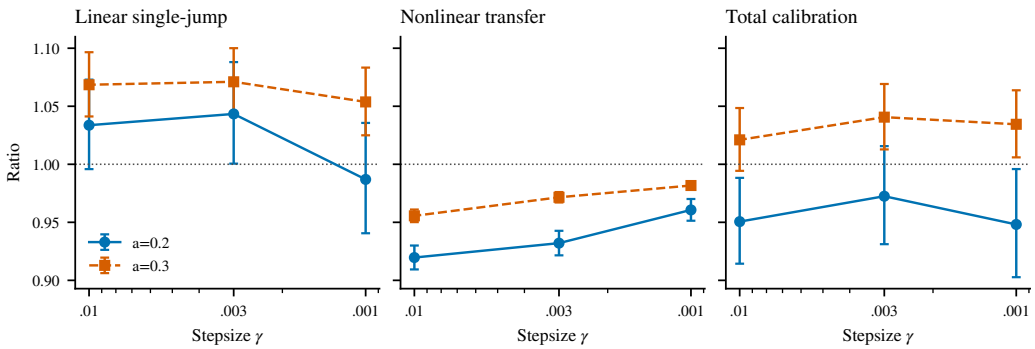


Figure 3: The two prespecified shrinking-threshold sequences. Each point uses 600,000 independent terminal pairs; stepsizes decrease from left to right. Error bars are pointwise probability or paired-ratio intervals.

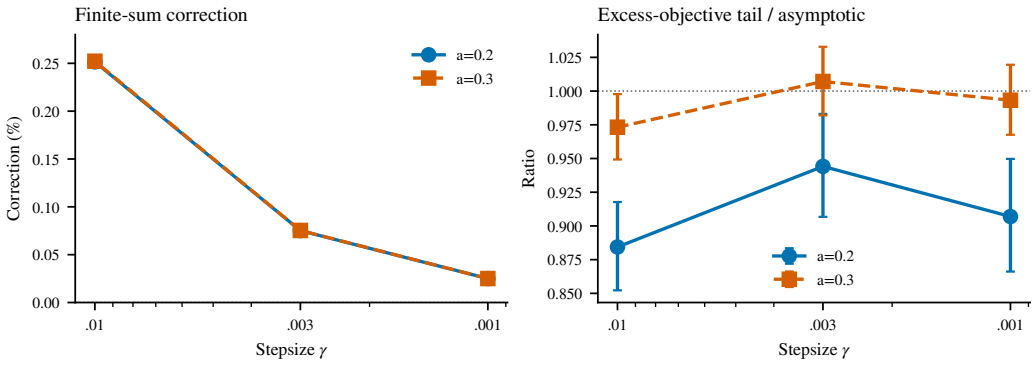


Figure 4: Finite-sum correction and objective-tail calibration for both prespecified exponents. The left panel reports the percentage correction $100(R_{\text{const}} - 1)$; the right panel retains the full range of objective-tail ratios and pointwise intervals.

**Saved artifacts.** The supplement includes every terminal pair, configuration, per-batch seed, source hash, elapsed time, aggregate count, and figure-generation script. The generator uses NumPy's PCG64 random generator; seeds are $2026092607 + 100i + j$, where $i = 0, 1, 2$ indexes the stepsize and $j = 0, 1, 2$ the batch. The simulator was checked against direct scalar updates and a closed-form weighted linear sum on identical deterministic innovations. Analysis checks reconcile count identities, Wilson constructions, paired-ratio covariance, and the objective inversion.

### F.1 Random-curvature ablation

Let $A_k = 1 + 10\zeta_k$, where $\zeta_k$ is Rademacher and independent of the Student innovation. The model in the main text has an integrable sample loss $G(x; \zeta, \varepsilon) = \frac{1}{2}(1 + 10\zeta)x^2 + 4\Phi(x) - 0.02\varepsilon x$, where $\Phi'(x) = \sin^3 x/(1 + x^2)$. Its expectation is the scalar objective. Since $A_k$ is bounded, all growth and increment assumptions hold for every $1 < p < 1.5$. The target innovation and local population Hessian are the same as in the additive study. The state-dependent martingale term is nonzero.

The two linear controls are

$$Q_{k+1} = (1 - \gamma A_{k+1})Q_k + \gamma\sigma\varepsilon_{k+1}, \qquad Y_{k+1} = (1 - \gamma)Y_k + \gamma\sigma\varepsilon_{k+1}.$$

They share $(A_k, \varepsilon_k)$ with the full nonlinear chain. All three start at zero. The same three stepsizes, two threshold exponents, and horizon schedule as the additive study are used, with $N = 200{,}000$ independent triples per stepsize. Two batches of $100{,}000$ are generated with seeds $2026092611 + 100i + j$. The configuration is fixed before sampling and all endpoints are retained. The homogeneous random-curvature update satisfies $\mathbb{E}(1 - \gamma A)^2 = 1 - 2\gamma + 101\gamma^2 < 1$ at all tested stepsizes. This check concerns the update coefficient and does not assume a finite second moment of the driven process.

Define $R_{\mathrm{curv}} = \widehat{p}_Q/\widehat{p}_Y$ and $R_{\mathrm{drift}} = \widehat{p}_X/\widehat{p}_Q$. Then $R_{\mathrm{tot}} = R_{\mathrm{lin}}R_{\mathrm{curv}}R_{\mathrm{drift}}$ exactly at every displayed point. All ratio intervals keep the paired covariance as described above. The deterministic benchmark remains the population-Hessian exact-CDF sum.

**A coefficient-based diagnostic.** For the random linear control, the coefficient product of a shock of age $j$ is $W_j = \prod_{i=1}^{j}(1 - \gamma A_i)$. All these factors are positive at the tested stepsizes, and independence gives $\mathbb{E}W_j^t = m_t(\gamma)^j$, with $m_t(\gamma) = \{(1 - 11\gamma)^t + (1 + 9\gamma)^t\}/2$. The ratio of expected sums of the $t$th powers of the random and deterministic weights is therefore exactly

$$J_{\gamma,n} = \frac{1 - (1 - \gamma)^t}{1 - m_t(\gamma)} \frac{1 - m_t(\gamma)^n}{1 - (1 - \gamma)^{nt}}.$$

For the tested horizons and $t = 1.5$, this equals $1.337$, $1.081$, and $1.026$ at the three decreasing stepsizes. It measures the amplification of a power-tail single-shock contribution by the random coefficients. The dotted curve in the middle panel is this coefficient diagnostic, not an exact finite-step probability formula. When the horizon factors are negligible, its small-$\gamma$ expansion is $1+25\gamma+O(\gamma^2)$. This expansion is an algebraic property of the weights; no finite-step error rate for the nonlinear probability is inferred from it.

Table 5: Complete random-curvature ablation. $K_X$ counts nonlinear exceedances among $200{,}000$ independent endpoints. The final column is the pointwise interval for $R_{\mathrm{tot}}$. The two middle ratios isolate random curvature and the smooth population remainder.

| $\gamma$ | $a$ | $K_X$ | $R_{\mathrm{tot}}$ | $R_{\mathrm{curv}}$ | $R_{\mathrm{drift}}$ | 95% CI |
|---|---|---|---|---|---|---|
| 0.01 | 0.2 | 1206 | 1.360 | 1.428 | 0.953 | [1.286, 1.439] |
| 0.01 | 0.3 | 2550 | 1.441 | 1.438 | 0.966 | [1.387, 1.498] |
| 0.003 | 0.2 | 724 | 1.041 | 1.079 | 0.951 | [0.968, 1.119] |
| 0.003 | 0.3 | 1930 | 1.160 | 1.101 | 0.984 | [1.110, 1.213] |
| 0.001 | 0.2 | 569 | 1.019 | 1.080 | 0.963 | [0.939, 1.106] |
| 0.001 | 0.3 | 1690 | 1.074 | 1.039 | 0.985 | [1.024, 1.126] |

The $a = 0.2$ random-curvature ratios at decreasing stepsizes are $1.428$, $1.079$, and $1.080$; their paired intervals are $[1.348, 1.512]$, $[1.026, 1.135]$, and $[1.036, 1.126]$. The last two estimates do not

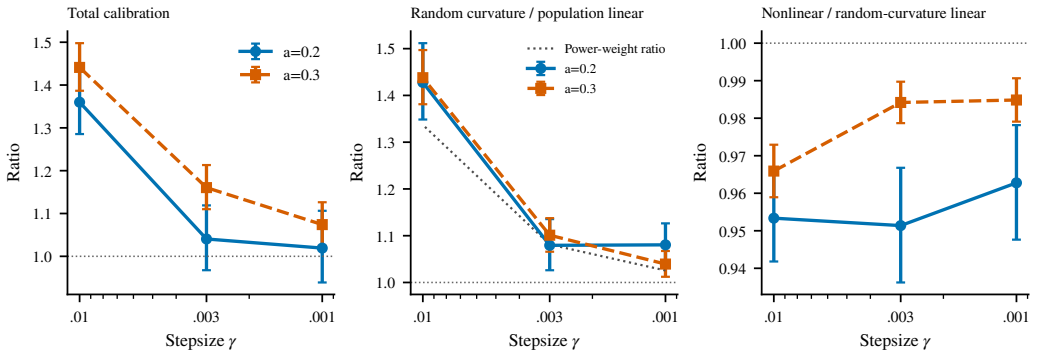


Figure 5: Random-curvature ablation with both prespecified threshold exponents. Left: total calibration against the population-Hessian single-jump sum. Middle: the random-curvature linear chain relative to the population linear chain. Right: the nonlinear chain relative to the random-curvature linear chain. Error bars are pointwise and use the common-innovation covariance where appropriate.

establish a strictly monotone finite-step trend, although both are much smaller than the largest-step amplification. At $\gamma = 0.001$, the population-remainder ratio is $0.963$ ($[0.948, 0.978]$), while the population-linear single-jump ratio is $0.980$ ($[0.901, 1.065]$). Their product with the curvature ratio yields the total $1.019$ ($[0.939, 1.106]$). For $a = 0.3$, the random-curvature ratios are $1.438$, $1.101$, and $1.039$, and the population-remainder ratios are $0.966$, $0.984$, and $0.985$. At the smallest step, the total $1.074$ ($[1.024, 1.126]$) remains detectably above one. The $t$th-weight diagnostic approaches one more rapidly than some empirical ratios; it isolates a coefficient effect and does not include accumulated small jumps or nonlinear drift. There are between $569$ and $2550$ one-sided nonlinear exceedances per setting, so the uncertainty at the rarer thresholds is visible rather than suppressed by plotting only point estimates.

## G Objective tails, quantiles, and batch effects

**Model and sampling units.** In dimension $d = 10$, let $H = \mathrm{diag}(1, 1 + 1/(d-1), \ldots, 2)$, $\boldsymbol{a} = (1, \ldots, 1)^\top/\sqrt{d}$, and

$$h(s) = \frac{\sin^3 s}{1+s^2}, \qquad \Phi(s) = \int_0^s h(v)\,\mathrm{d}v,$$
$$m(x) = Hx + 6\boldsymbol{a}h(\boldsymbol{a}^\top x), \qquad M(x) = \tfrac{1}{2}x^\top Hx + 6\Phi(\boldsymbol{a}^\top x).$$

The innovation is the multivariate Student vector $\eta = V/\sqrt{Q/t}$, with $V \sim N(0, I_d)$, $Q \sim \chi^2_t$, $t = 1.5$, and independent $V, Q$. Each step averages $b$ independent innovations. The configurations, in seed order, are $(\gamma, b) = (0.05, 1), (0.01, 1), (0.04, 4), (0.01, 4), (0.01, 16), (0.002, 1)$. Each uses $N = 300{,}000$ independent terminal pairs, $X_0 = Y_0 = 0$, and $n = \max\{500, \lceil 5\log(1/\gamma)/\gamma \rceil\}$. The nonlinear and population-linear chains share every innovation within a pair. Seeds are $20260926 + i$ for configuration index $i = 0, \ldots, 5$.

**Objective evaluation.** The integral $\Phi$ is even and bounded. We tabulate it on $[0, 64]$ by composite Simpson integration with mesh $0.002$ and use cubic Hermite interpolation with the known derivative $h$. For larger arguments, $\sin^3 s = (3\sin s - \sin 3s)/4$ gives a tail expansion after eight integrations by parts. Adaptive weighted quadrature at 59 validation points up to $10^4$ differs by less than $10^{-12}$ in $\Phi$. This procedure avoids applying a fixed quadrature rule across arbitrarily many oscillations. All comparisons, including the Gaussian surrogate and the coupled linear chain, use the same full objective $M$.

**Three distinct predictions.** Write $\eta = R\Theta$, where $\Theta$ is uniform on $S^{d-1}$, independent of $R$, and $R^2/d \sim F(d,t)$. Let $B_\gamma = I_d - \gamma H$. We distinguish

$$A_b(\epsilon) = b^{1-t} C_M \frac{\mathbb{P}(R > \sqrt{\epsilon}/\gamma)}{\gamma}, \tag{87}$$

$$\Lambda_b^{\mathrm{quad}}(\epsilon) = b \sum_{j<n} \mathbb{P}\left\{ \tfrac{1}{2} (\gamma B_\gamma^j \eta/b)^\top H (\gamma B_\gamma^j \eta/b) > \epsilon \right\}, \tag{88}$$

$$B_b(\epsilon) = b \sum_{j<n} \mathbb{P}\{ M(\gamma B_\gamma^j \eta/b) - M(0) > \epsilon \}. \tag{89}$$

The first is the local quadratic tail-measure approximation. The second uses the exact distribution of an individual gradient shock and a quadratic event. The third evaluates the full objective at the linearly transported shock. The factor $b$ counts the possible extreme sample within a batch; each transported sample has weight $\gamma/b$. For $b > 1$, these sums concern the individual independent samples in the unrolled recursion, rather than replacing the average innovation by a Student variable. The theorem controls the local approximation in its stated shrinking-tolerance regime; $B_b$ is also assessed empirically at the finite objective levels used here. The local prediction uses the same propagated constant as the stable-limit extrapolation; the theorem establishes its relative accuracy in the specified regime.

**Numerical benchmark and its uncertainty.** For each configuration, we draw $200{,}000$ independent pairs of shock age and spherical direction, with seed $20270925 + i$. The age is a geometric draw with success probability $1 - e^{-1.2\gamma}$, capped at $n - 1$; its terminal mass includes the entire capped tail. We weight by the inverse age probability and integrate the radial Student tail analytically. For $B_b$, the radial threshold is found by bracketed Newton iteration. It is unique because the model's centered dissipativity makes $M(rz)$ strictly increasing for $r > 0$ and $z \neq 0$. The solver's scaled residual is below $3 \times 10^{-11}$ over the tested grid. Standard errors are computed across the independent age–direction draws, retaining all contributions from a draw together. The constant $C_M$ is evaluated by angular Monte Carlo and 80-node Gauss–Legendre integration on $[0, 25]$; the omitted time tail is exponentially small. Its estimate and standard error, as well as the two finite-sum predictions, are supplied for every threshold.

**Empirical uncertainty and quantiles.** All tail intervals are pointwise 95% Wilson intervals for independent terminal trajectories. Ratios between independent configurations use the delta-method log-ratio variance $(1/p_1 - 1)/N_1 + (1/p_2 - 1)/N_2$. Within a coupled pair, the joint exceedance count supplies the covariance correction. Empirical quantile intervals use binomial order-statistic ranks: for level $1 - \delta$, the lower and upper ranks are the $0.025$ and $0.975$ quantiles of $\mathrm{Binomial}(N, 1 - \delta)$, with the upper rank increased by one. Predicted quantiles invert a monotone log–log interpolation of the corresponding survival curve. The 181 objective levels run from $0.1$ to $10^5$; the quantile display uses 63 tail levels between $5 \times 10^{-4}$ and $0.05$. The figure's quantile bands reflect empirical quantile uncertainty conditional on the independently estimated benchmark; the latter's standard errors are separately recorded. On the count-qualified rare-event cells, the largest relative benchmark standard error is $0.117\%$. Evaluating the survival prediction directly at each interpolated 99% and 99.9% quantile gives relative inversion error below $2.7 \times 10^{-7}$.

**Covariance-based Gaussian surrogate.** For each of five independent seeds, we estimate $\widehat{\Sigma} = K^{-1} \sum_{k=1}^{K} \eta_k \eta_k^\top$, $K = 10^5$, and solve the discrete Lyapunov covariance of the linear recursion,

$$(\widehat{\Sigma}_X)_{ij} = \frac{\gamma^2 (\widehat{\Sigma}/b)_{ij}}{1 - (1 - \gamma H_{ii})(1 - \gamma H_{jj})}.$$

We draw two million Gaussian vectors per fitted covariance and evaluate their full objective $M$. The figures and tables report the median across the five fits and retain their full range in the supplementary outputs. Common standard-normal draws are reused across stepsize/batch configurations within a fit; the five covariance fits are independent. Because the true innovation covariance is infinite, this is a deliberately misspecified covariance-based diagnostic. The matched heavy-tail reference is the coupled linear SGD chain, evaluated using the same objective and innovations as the nonlinear chain. Quadratic-only Gaussian quantiles are also retained to separate covariance misspecification from the choice of objective.

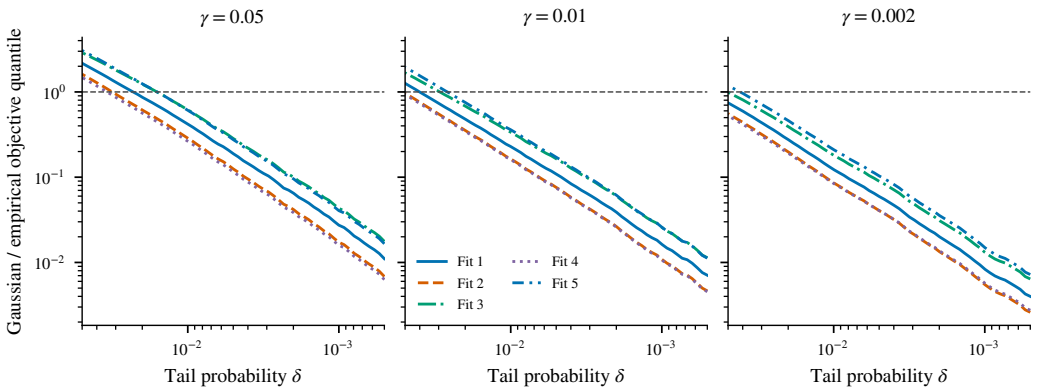


Figure 6: Sensitivity of the covariance-based Gaussian diagnostic. Every curve is one of five independent covariance fits; all predictions use the full objective. The denominator is the empirical SGD quantile from 300,000 terminal trajectories at the corresponding stepsize. The logarithmic vertical scale shows both fit variability and the discrepancy at increasingly extreme quantiles.

Table 6: All configurations at the 99% and 99.9% objective quantiles. "Linear" uses the coupled heavy-tail chain with the full objective. "Local", "Quad. jump", and "Exact-loss jump" invert (87)–(89), respectively.

| $\gamma$ | $b$ | $\delta$ | Empirical [95% CI] | Linear | Local | Quad. jump | Exact-loss jump |
|---|---|---|---|---|---|---|---|
| 0.05 | 1 | 0.001 | 3026.1 [2582.4, 3509.2] | 3021.5 | 2881.4 | 2953.8 | 2956.1 |
| 0.05 | 1 | 0.01 | 136.1 [129.3, 141.9] | 136.0 | 133.7 | 137.1 | 139.0 |
| 0.01 | 1 | 0.001 | 988.0 [835.7, 1145.7] | 983.1 | 985.6 | 990.8 | 993.0 |
| 0.01 | 1 | 0.01 | 48.7 [46.7, 51.0] | 49.0 | 45.7 | 46.0 | 47.6 |
| 0.04 | 4 | 0.001 | 1002.2 [858.1, 1225.1] | 1002.0 | 985.5 | 1006.1 | 1008.3 |
| 0.04 | 4 | 0.01 | 48.9 [46.7, 51.4] | 49.2 | 45.7 | 46.7 | 48.4 |
| 0.01 | 4 | 0.001 | 399.5 [356.8, 464.7] | 399.5 | 391.1 | 393.2 | 395.3 |
| 0.01 | 4 | 0.01 | 18.9 [18.0, 19.9] | 19.4 | 18.2 | 18.3 | 19.5 |
| 0.01 | 16 | 0.001 | 166.9 [146.8, 188.7] | 166.8 | 155.2 | 156.0 | 158.0 |
| 0.01 | 16 | 0.01 | 7.7 [7.3, 8.1] | 8.2 | 7.2 | 7.2 | 7.9 |
| 0.002 | 1 | 0.001 | 380.5 [320.9, 448.9] | 380.6 | 337.1 | 337.4 | 339.4 |
| 0.002 | 1 | 0.01 | 17.8 [17.0, 18.7] | 18.2 | 15.6 | 15.7 | 16.8 |

### G.1 Repeated plug-in pilots

We also examine estimation of the projected constant from stochastic-gradient samples. The pilot point is the coordinatewise median of 2,000 independent terminal states from $(\gamma, b) = (0.01, 1)$. At each sample size $K \in \{10^4, 10^5, 10^6\}$, five independent gradient samples yield a centered innovation sample, a Hill estimate from the largest 1% of norms, and the empirical angular measure of those extremes. A common-random-number finite difference estimates the Jacobian at the pilot point; 80-node quadrature on $[0, 30]$ evaluates the estimated propagated constant. Predictions are compared at thresholds $1, 2, 3, 5, 8$ in the four directions used in Appendix H. All pilot fits and empirical exceedance counts are retained. The additive model cancels sampling error in the finite-difference Jacobian, while the difference between curvature at the pilot point and at the target remains. These repetitions describe sensitivity to gradient-sample size and tail estimation; they do not supply an estimator-validity or coverage theorem. Across the five fits, the median empirical/predicted ratio over the 20 reporting cells ranges from $0.571$ to $1.847$ at $K = 10^4$, from $0.745$ to $1.123$ at $10^5$, and from $0.983$ to $1.064$ at $10^6$. The larger pilot stabilizes the median calibration, although individual directions and thresholds retain finite-level discrepancies. Figure 7 displays every fit and the full range across reporting cells.

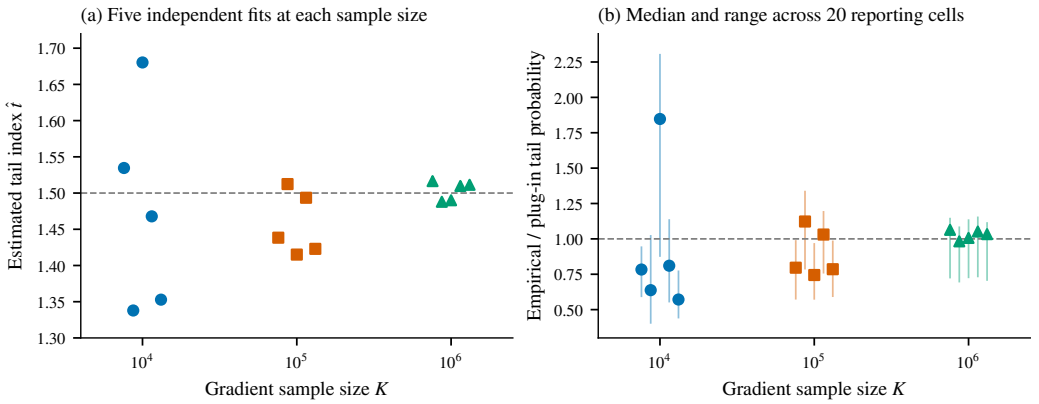


Figure 7: Repeated plug-in estimation. Each marker is one independent gradient-sample fit; horizontal offsets separate the five fits at a common sample size. In (b), markers show medians and vertical segments the minimum and maximum over four directions and five thresholds. Segments describe variation across reporting cells, not confidence intervals.

# H MULTIDIMENSIONAL PROJECTED-TAIL EXPERIMENTS

We examined projected-tail calibration in $d = 10$ and $d = 100$ for two coupled nonconvex models: additive heavy-tailed gradients (Model A) and random-design regression SGD (Model B). The latter includes a random sample Hessian and state-dependent gradient noise. These experiments test the finite-step one-big-jump approximation along three eigendirections and a dense non-eigendirection.

## H.1 POPULATION OBJECTIVE AND STOCHASTIC-GRADIENT MODELS

For each $d \in \{10, 100\}$, define

$$H = \operatorname{diag}\left(1 + \frac{i-1}{d-1}\right)_{i=1}^{d}, \qquad \boldsymbol{a} = \frac{(1,\dots,1)^\top}{\sqrt{d}}, \qquad h(s) = \frac{\sin^3(s)}{1+s^2}. \tag{90}$$

Both models have population objective and gradient

$$\begin{aligned} M(x) &= \tfrac{1}{2} x^\top H x + 6\Phi(\boldsymbol{a}^\top x), \qquad \Phi(s) = \int_0^s h(v)\,\mathrm{d}v, \\ m(x) &= Hx + 6\boldsymbol{a}\, h(\boldsymbol{a}^\top x). \end{aligned} \tag{91}$$

The nonlinear correction couples all coordinates. It is $O(|x|^3)$ near zero and bounded globally, so the same linear drift $Hx$ governs both the local and large-$|x|$ behavior. The spectrum of $H$ remains in $[1, 2]$ as the dimension changes.

**Nonconvexity and centered dissipativity.** For $|s| \le \pi$, $sh(s) \ge 0$; for $|s| \ge \pi$, $h(s)/s \ge -[\pi(1+\pi^2)]^{-1}$. Hence

$$x^\top m(x) \ge \left(1 - \frac{6}{\pi(1+\pi^2)}\right)|x|^2 > 0.824|x|^2. \tag{92}$$

Thus $x^\star = 0$ is the unique stationary point and global minimizer. Nevertheless,

$$\boldsymbol{a}^\top \nabla^2 M(2\boldsymbol{a})\boldsymbol{a} = 1.5 + 6h'(2) \simeq -0.4604 < 0,$$

so $M$ is nonconvex. These certificates hold in both dimensions.

**Model A: additive heavy-tailed gradients.** Independently at each step, let $V \sim N(0, I_d)$ and $Q \sim \chi^2_{1.5}$ be independent, and set $\eta = V/\sqrt{Q/1.5}$. We used

$$X_{k+1} = X_k - \gamma m(X_k) + \gamma\eta_{k+1}. \tag{93}$$

Every unit projection of $\eta$ has the univariate Student-$t_{1.5}$ law, but its coordinates are not independent. The sample loss $G_A(x;\eta) = M(x) - \eta^\top x$ has finite expectation and population gradient $m$.

**Model B: random-design regression SGD.** Let $U$ be uniform on the unit sphere in $\mathbb{R}^d$, set $Z = \sqrt{d}\, H^{1/2}U$, and draw $\varepsilon \sim t_{1.5}$ independently of $U$. Independent pairs $(Z, \varepsilon)$ are sampled at successive steps. We used

$$G_B(x; Z, \varepsilon) = \tfrac{1}{2}(Z^\top x)^2 - \varepsilon Z^\top x + 6\Phi(\boldsymbol{a}^\top x), \tag{94}$$

$$X_{k+1} = X_k - \gamma\left[Z_{k+1}(Z_{k+1}^\top X_k - \varepsilon_{k+1}) + 6\boldsymbol{a}h(\boldsymbol{a}^\top X_k)\right]. \tag{95}$$

Since $\mathbb{E}[ZZ^\top] = H$ and $\mathbb{E}\varepsilon = 0$, $\mathbb{E}G_B(x; Z, \varepsilon) = M(x)$. Equation (94) is the squared-error loss with the sample-dependent constant $\varepsilon^2/2$ removed *before* taking expectations. This leaves each stochastic gradient unchanged and makes the population objective finite; the uncentered expected squared loss is infinite. Model B has the random rank-one Hessian $ZZ^\top$ at the target and couples all coordinates through both the random design and the nonlinear term.

**Relation to the assumptions.** For every $1 < p < 1.5$, both innovation laws have finite $p$th moments. The bounded derivative of $h$ and the bound $|Z|^2 \le 2d$ give the global $L^p$ growth and local increment conditions at each fixed $d$. The cubic population remainder satisfies the local quadratic bound. Both innovations, $\eta$ in Model A and $Z\varepsilon$ in Model B, are multivariate regularly varying with index $t = 1.5$. Any $1 < p < 1.5$ satisfies the gradient-moment conditions, and the chosen $p$ determines the shrinking-threshold window in Theorem 2.5; the endpoint $p = 1.5$ is not available. These statements do not assert dimension-uniform constants. The fixed-threshold study below diagnoses finite-step behavior.

### H.2 Step sizes, projections and simulation protocol

Model A used $\gamma \in \{0.1, 0.05, 0.01\}$ in both dimensions. Model B used the same grid at $d = 10$ and $\gamma \in \{0.01, 0.005, 0.001\}$ at $d = 100$, preserving $\gamma d \in \{1, 0.5, 0.1\}$. The smaller regression step sizes account for the size of the random sample Hessian, rather than a change in the population spectrum.

More precisely, the spherical fourth-moment identity gives

$$\mathbb{E}[|Z|^2 ZZ^\top] = \frac{d}{d+2}\{\mathrm{tr}(H)H + 2H^2\}.$$

Consequently, for deterministic $v$,

$$\mathbb{E}|(I_d - \gamma ZZ^\top)v|^2 = v^\top \left[ I_d - 2\gamma H + \gamma^2 \frac{d}{d+2}\{\mathrm{tr}(H)H + 2H^2\} \right] v. \tag{96}$$

The homogeneous linear update is strictly contractive in Euclidean mean square for every $v \neq 0$ precisely when

$$0 < \gamma < \frac{2(d+2)}{d\{\mathrm{tr}(H) + 2\lambda_{\max}(H)\}}. \tag{97}$$

This upper bound is approximately $0.1263$ for $d = 10$ and $0.01325$ for $d = 100$. It characterizes mean-square contraction of the homogeneous linear update and motivates the $1/d$ stepsize scale for Model B. We assess the heavy-tailed forced recursion through tail probabilities; its finite-step errors and admissible stepsizes can depend on dimension.

We studied the four unit directions

$$\ell \in \{e_1, e_{d/2+1}, e_d, \boldsymbol{a}\}.$$

The first three are the slow, middle and fast eigendirections of $H$ (middle indices 6 and 51). The dense nonlinear coupling direction $\boldsymbol{a}$ is not an eigenvector, so it also tests the general projected formula, beyond the scalar autoregressive special case.

Each independent trajectory started at $X_0 = 0$ and ran for

$$n = \max\left\{500, \left\lceil \frac{5\log(1/\gamma)}{\gamma} \right\rceil \right\}. \tag{98}$$

The horizons for $\gamma = 0.1, 0.05, 0.01, 0.005, 0.001$ were respectively $500, 500, 2303, 5299, 34539$. We used $N = 200{,}000$ independent terminal trajectories per model–dimension–step-size setting, except Model B at $d = 100$, $\gamma = 0.001$, for which $N = 250{,}000$. There were $2{,}450{,}000$ trajectories in total. All generated terminal states were finite, and no trajectories were discarded. We evaluated $y = 0.001, 0.002, \ldots, 5.000$ from the same terminal projections; no additional chains were run for individual thresholds. Samples here are independent trajectories, not correlated observations taken along one chain.

### H.3 Deterministic benchmark and Monte Carlo uncertainty

Alongside each nonlinear trajectory we simulated

$$Y_{k+1} = B_\gamma Y_k + \gamma \eta_{k+1}, \qquad B_\gamma = I_d - \gamma H, \qquad Y_0 = 0,$$

using the same innovations as $X_k$: $\eta$ from Model A or $Z\varepsilon$ from Model B. This reference uses the *population* Hessian $H$, not the random sample Hessian. The deterministic denominator was the finite-horizon one-big-jump sum

$$\Lambda_\gamma(\ell, y) = \sum_{j=0}^{n-1} \mathbb{P}\{\gamma \ell^\top B_\gamma^j \eta > y\}. \tag{99}$$

It is a sum of single-jump probabilities, not the exact probability of their sum. With $w_j = (B_\gamma^j)^\top \ell$ and Student survival function $\overline{T}_{1.5}$, the summands are

$$\text{Model A:} \quad \overline{T}_{1.5}\left(\frac{y}{\gamma |w_j|}\right),$$

$$\text{Model B:} \quad \mathbb{E}_S\left[\overline{T}_{1.5}\left(\frac{y}{\gamma\sqrt{d}\,|H^{1/2}w_j|S}\right)\right], \qquad S^2 \sim \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{d-1}{2}\right).$$

The regression expectation was evaluated by 128-node Gauss–Jacobi quadrature, with shape-preserving log-scale interpolation for dense-grid evaluation. Increasing the order to 256 in the $d = 100$ regression checks at $y = 1$ and 5, for all displayed directions and step sizes, changed the denominator by less than $4.2 \times 10^{-5}$ in relative terms.

Writing $\widehat{p}_X$ and $\widehat{p}_Y$ for the projected-tail estimates, we report

$$R_{\mathrm{lin}} = \frac{\widehat{p}_Y}{\Lambda_\gamma}, \qquad R_{\mathrm{tr}} = \frac{\widehat{p}_X}{\widehat{p}_Y}, \qquad R_{\mathrm{tot}} = \frac{\widehat{p}_X}{\Lambda_\gamma}. \tag{100}$$

The transfer ratio in Model B combines random-Hessian and nonlinear-correction effects; it does not isolate either effect.

For $d = 10$, we retained the ordinary one-sided estimates $\widehat{p}_X = K_X/N$, where $K_X = \sum_{i=1}^{N} \mathbf{1}_{\{\ell^\top X_n^{(i)} > y\}}$, and analogously for $Y$. The $d = 100$ experiments used exact central symmetry to reduce Monte Carlo variance. Since $m$ is odd and both chains start at zero, negating all innovations in Model A, or all residuals while holding the design sequence fixed in Model B, negates the entire trajectory. Thus

$$p_X = \tfrac{1}{2}\mathbb{P}(|\ell^\top X_n| > y), \qquad \widehat{p}_X = \frac{K_X^{\mathrm{abs}}}{2N}, \qquad K_X^{\mathrm{abs}} \sim \mathrm{Binomial}(N, 2p_X). \tag{101}$$

The variance is $p_X(1 - 2p_X)/(2N)$, compared with $p_X(1 - p_X)/N$ for the one-sided estimator. The two signs of a trajectory are *not* independent trials. Symmetry is a property of these examples, not an assumption of the general tail theorem.

Shading in the figures gives nominal 95% *pointwise* Wilson score intervals for $p_X$, divided by $\Lambda_\gamma$. For $d = 100$, we first computed the interval for $K_X^{\mathrm{abs}}/N$ using $N$ trials, then divided its endpoints by two. To restrict shading to cells with adequate tail counts, we displayed it only where $\Lambda_\gamma \le 0.1$ and the observed nonlinear count, linear count and corresponding expected benchmark count were all at least 200. The expected count is $N\Lambda_\gamma$ for $d = 10$ and $2N\Lambda_\gamma$ for $d = 100$. Curves are shown on the entire threshold grid, without smoothing. These intervals quantify Monte Carlo uncertainty only: they neither include approximation error nor provide simultaneous coverage over directions and thresholds.

### H.4 Results and scope

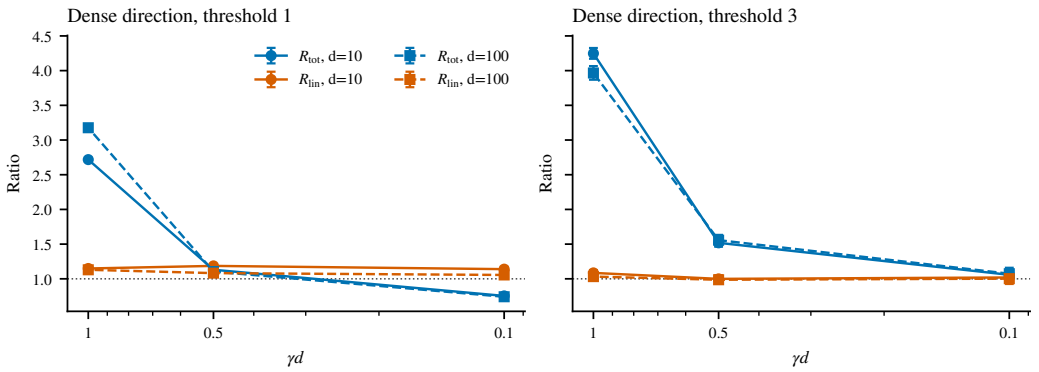


Figure 8: Random-design SGD in the dense non-eigendirection, at matched $\gamma d$. Stepsizes decrease from left to right and both panels share the same vertical range. Blue curves show total calibration and orange curves the population-Hessian linear reference. Each point uses 200,000 independent trajectories, except $d = 100$, $\gamma = 0.001$, which uses 250,000.

Table 7 uses the *same deterministic reporting domain* for every row: all four directions and $y = 1.000, 1.001, \ldots, 5.000$, comprising 16,004 direction–threshold cells. This table does not apply the count-based shading filter. The medians are descriptive summaries of correlated cells, not estimates based on 16,004 independent replications. Also, the median of $R_{\mathrm{tot}}$ need not equal the product of the other two medians, even though the pointwise identity holds.

Table 7: Median projected-tail ratios on the common domain $1 \le y \le 5$. $N$ is the actual number of independent trajectories. The last column is the median full width of the 95% pointwise interval for $R_{\mathrm{tot}}$. The $d = 10$ rows use one-sided counts; the $d = 100$ rows use symmetry pooling.

| $d$ | Model | $\gamma$ | $N/10^3$ | $R_{\mathrm{lin}}$ | $R_{\mathrm{tr}}$ | $R_{\mathrm{tot}}$ | CI width |
|---|---|---|---|---|---|---|---|
| 10 | A | 0.100 | 200 | 1.069 | 0.992 | 1.049 | 0.085 |
| 10 | A | 0.050 | 200 | 1.044 | 0.992 | 1.029 | 0.101 |
| 10 | A | 0.010 | 200 | 1.015 | 0.993 | 0.999 | 0.150 |
| 10 | B | 0.100 | 200 | 1.096 | 3.923 | 4.266 | 0.158 |
| 10 | B | 0.050 | 200 | 1.057 | 1.533 | 1.588 | 0.117 |
| 10 | B | 0.010 | 200 | 1.028 | 1.051 | 1.081 | 0.147 |
| 100 | A | 0.100 | 200 | 1.063 | 0.998 | 1.055 | 0.059 |
| 100 | A | 0.050 | 200 | 1.068 | 0.999 | 1.059 | 0.072 |
| 100 | A | 0.010 | 200 | 1.041 | 0.999 | 1.035 | 0.106 |
| 100 | B | 0.010 | 200 | 1.028 | 4.298 | 4.331 | 0.204 |
| 100 | B | 0.005 | 200 | 0.983 | 1.628 | 1.586 | 0.149 |
| 100 | B | 0.001 | 250 | 1.045 | 1.070 | 1.118 | 0.169 |

For Model A, the median total ratios were within 6% of one in both dimensions, and the transfer ratios were close to one. Model B showed larger finite-step discrepancies: at the matched values $\gamma d = 1, 0.5, 0.1$, the median total ratios decreased from $4.266$ to $1.588$ to $1.081$ in $d = 10$, and from $4.331$ to $1.586$ to $1.118$ in $d = 100$. The similar patterns at matched $\gamma d$ support the step-size scaling motivated by (97), without establishing a dimension-uniform result.

At $d = 100$ and $\gamma = 0.001$ in Model B, the direction-specific median total ratios ranged from $1.078$ to $1.152$. An approximately 12% median excess over one therefore remained at the smallest tested step size. On the common reporting domain, the median and maximum full widths of the pointwise ratio intervals were $0.169$ and $0.263$, respectively. As a sensitivity check, using ordinary one-sided counts on the same $d = 100$ trajectories gave median $R_{\mathrm{tot}} = 1.117$ in this setting, close to the symmetry-pooled value $1.118$.

Figures 9–12 show the full threshold dependence. The near-origin region is not a rare-event regime, and the one-big-jump sum can exceed one there. The zero initial state and schedule (98) satisfy the time condition of Theorem 2.5. The fixed threshold grid serves as a finite-step diagnostic; the theorem itself uses shrinking thresholds. Together with the diagonal study, these results locate the empirical discrepancies in threshold choice, single-jump calibration, and nonlinear transfer.

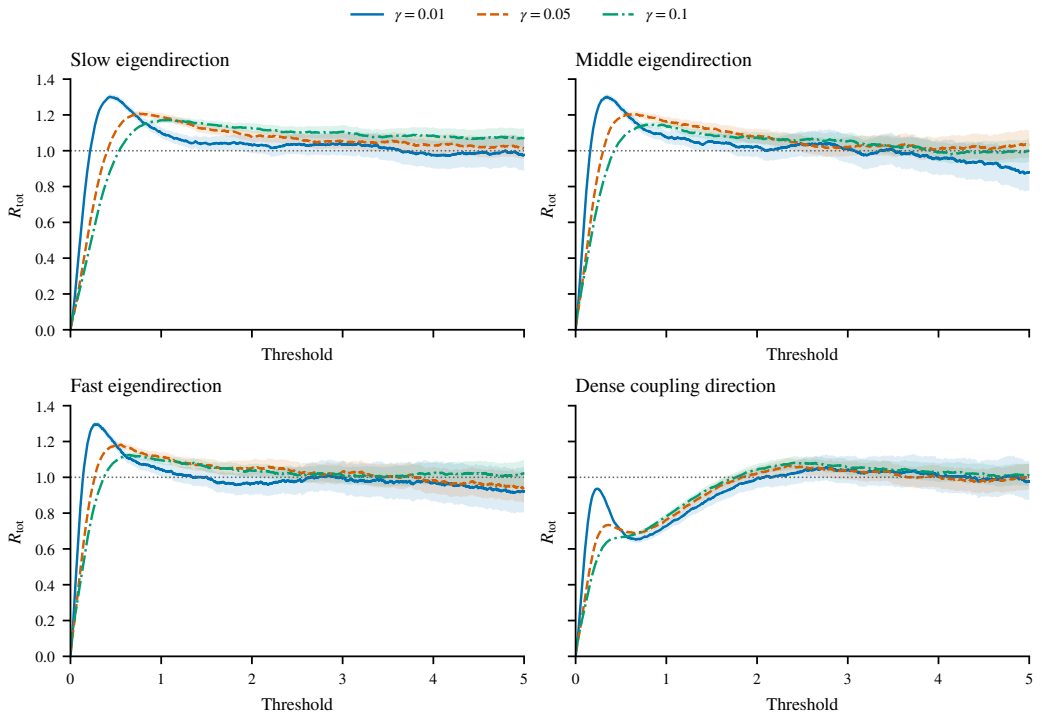


Figure 9: Additive-noise SGD (Model A) in $d = 10$. Curves show $R_{\mathrm{tot}}$ for $\gamma = 0.1, 0.05, 0.01$, with $N = 200{,}000$ independent trajectories per step size and ordinary one-sided tail estimates. Panels show $e_1$, $e_6$, $e_{10}$ and the dense nonlinear coupling direction $\boldsymbol{a}$. Shading gives 95% pointwise Wilson ratio intervals on the count-qualified cells defined in Appendix H.3. The dashed line marks ratio one. Axes are linear, and all four panels share a vertical scale.

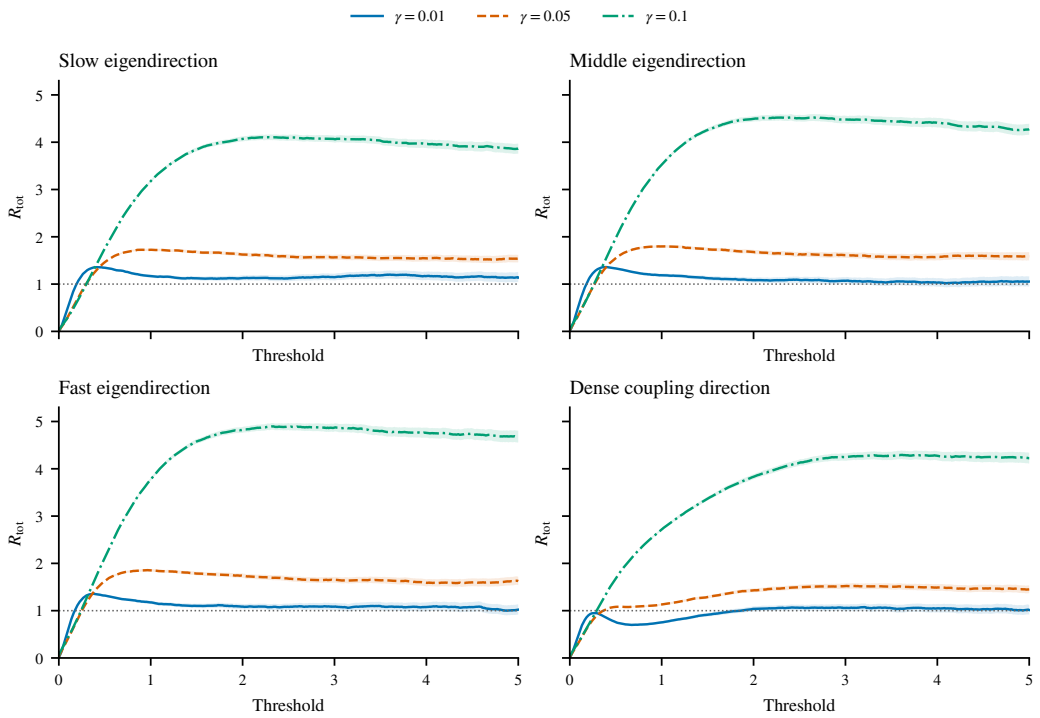


Figure 10: Random-design regression SGD (Model B) in $d = 10$, with $\gamma = 0.1, 0.05, 0.01$ and $N = 200{,}000$ independent trajectories per step size. The directions, one-sided estimator and pointwise interval convention are the same as in Figure 9. The larger ratios at larger step sizes display finite-step discrepancies from the population-Hessian one-big-jump benchmark.

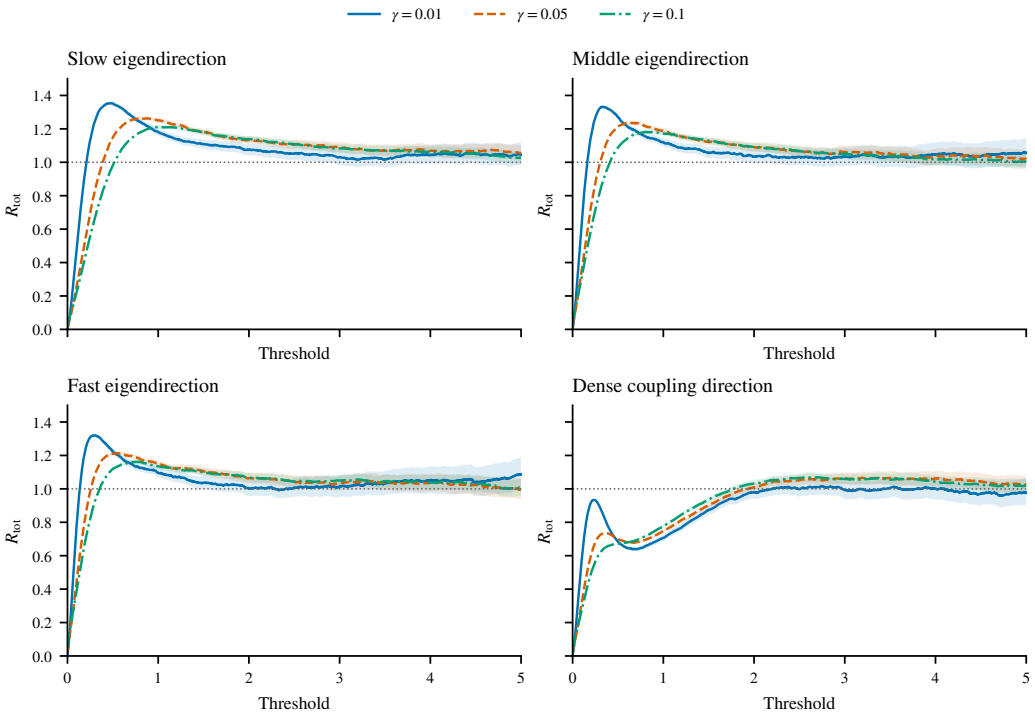


Figure 11: Additive-noise SGD (Model A) in $d = 100$, with $\gamma = 0.1, 0.05, 0.01$ and $N = 200{,}000$ independent trajectories per step size. Panels show $e_1$, $e_{51}$, $e_{100}$ and $\boldsymbol{a}$. Curves use symmetry-pooled tail estimates; shading gives 95% pointwise Wilson ratio intervals computed with the actual $N$, as described in Appendix H.3. Axes are linear, with a shared vertical scale across the four panels.

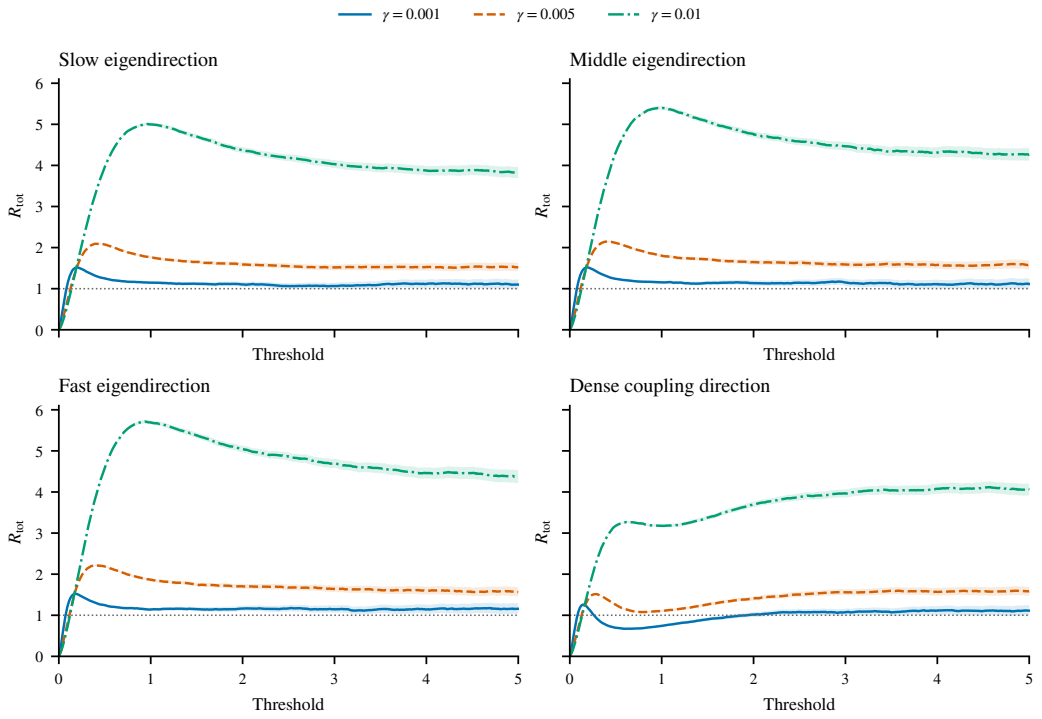


Figure 12: Random-design regression SGD (Model B) in $d = 100$, with $\gamma = 0.01, 0.005, 0.001$. The first two settings use $N = 200{,}000$ independent trajectories each, and the last uses $N = 250{,}000$. Directions, symmetry pooling and pointwise interval conventions match Figure 11. Decreasing the step size improves calibration, but a residual discrepancy remains at $\gamma = 0.001$. The intervals measure Monte Carlo precision, not approximation error or simultaneous coverage.

### H.5 A RANDOM-HESSIAN SINGLE-JUMP BENCHMARK

For Model B, let $Z = \sqrt{d}H^{1/2}U$ with $U$ uniform on the sphere and scalar label noise $\epsilon \sim t_{1.5}$. The innovation $Z\epsilon$ at the shock is independent of the subsequent sample Hessians. Define $W_0 = \ell$ and $W_{j+1} = (I_d - \gamma Z_{j+1}Z_{j+1}^\top)W_j$, with fresh designs $Z_j$. Reversal of an iid finite product shows that $W_j$ has the distribution of the direction transported backward through $j$ sample Hessians. The benchmark is

$$\Lambda_\gamma^{\mathrm{RH}}(\ell, u) = \sum_{j<n} \mathbb{E}\left[\overline{T}_{1.5}\left(\frac{u}{\gamma|W_j^\top Z|}\right)\right], \tag{102}$$

where $Z$ is independent of $W_j$ and $\overline{T}_t$ is the Student survival function. The population benchmark replaces $W_j$ by $B_\gamma^j\ell$. Thus $R_{\mathrm{tot}}^{\mathrm{RH}} = \widehat{p}_X/\widehat{\Lambda}_\gamma^{\mathrm{RH}}$ separates the effect of random sample curvature from the population-flow approximation.

**Independent sampling units.** We use 20,000 independent design paths in $d = 10$ and 8,000 in $d = 100$, at all three stepsizes. A path contributes the complete sum across ages and all four directions, so its standard error retains the dependence between ages. We integrate over the independent shock design and Student noise analytically, reducing Monte Carlo variance. The reported domain has four directions and 17 equally spaced thresholds on $[1, 5]$, giving 68 ratios per dimension–stepsize setting. The table's MC SE is the maximum relative standard error of the estimated denominator over this domain, expressed as a percentage. It does not include uncertainty in the empirical SGD numerator. Twenty independent path blocks are retained; resampling them also gives denominator-only intervals for the median ratio.

**Analytic tail integration.** Conditionally on $W_j$, write $v = u/(\gamma|H^{1/2}W_j|)$. If $U_1$ is the first coordinate of a uniform sphere point, the conditional summand is $F_d(v) = \mathbb{E}\overline{T}_t\{v/(\sqrt{d}|U_1|)\}$. Expanding the Student density and integrating its tail gives

$$F_d(v) = \sum_{k\geq 0} c_{d,k}v^{-(t+2k)}, \qquad c_{d,k} = \frac{(-1)^k t^{t/2+k}\Gamma((t+1)/2+k)}{\sqrt{\pi}\Gamma(t/2)k!(t+2k)}\mathbb{E}(\sqrt{d}|U_1|)^{t+2k},$$

with

$$\mathbb{E}(\sqrt{d}|U_1|)^r = \frac{d^{r/2}\Gamma(d/2)\Gamma((r+1)/2)}{\sqrt{\pi}\Gamma((d+r)/2)}.$$

In this design family, $\gamma d \leq 1$ and $|Z|^2 \leq 2d$ make each sample propagation matrix nonexpansive. Consequently $|W_j| \leq 1$ and $td/v^2 \leq 2td\gamma^2/u^2 \leq 0.3$ on the reported domain. The series is uniformly convergent and alternating there. We keep 14 terms and bound the remainder by the first omitted term, summed over ages. Independent adaptive spherical quadrature agrees to within $1.3 \times 10^{-12}$ in the validation cases.

**Truncation and sensitivity.** Let $h_i$ be the diagonal entries of $H$ and set

$$q_2 = \max_i \left\{1 - 2\gamma h_i + \gamma^2 \frac{d}{d+2}\big(\operatorname{tr}(H)h_i + 2h_i^2\big)\right\} < 1.$$

The exact spherical fourth-moment formula gives $\mathbb{E}|W_j|^2 \leq q_2^j|\ell|^2$, hence $\mathbb{E}|W_j|^t \leq q_2^{jt/2}$. We truncate at $J = \min\{n, \lceil 40/(t|\log q_2|)\rceil\}$. The remaining sum is bounded by

$$c_{d,0}(\gamma/u)^t\|H\|_{\mathrm{op}}^{t/2}\frac{q_2^{Jt/2}}{1 - q_2^{t/2}},$$

and is zero when $J = n$. The actual cutoffs, residual bounds, and comparison with the shorter cutoff $\min\{n, \lceil 16/(t\gamma)\rceil\}$ are supplied for every setting. This controls truncation using the random-product decay rather than assuming the population contraction rate.

The largest shorter-cutoff correction is $0.287\%$; the final omitted-age bounds are below $4.1 \times 10^{-9}$ relative to the calculated denominators. The retained block summaries provide a separate check on Monte Carlo variability without treating shock ages as independent samples.

Table 8: Numerical accuracy of the random-Hessian denominator. Columns report the largest relative standard error, largest relative change from the shorter cutoff, and largest omitted-age bound divided by the computed denominator, over the common 68-cell domain. The residual-bound column is a fraction, whereas the two preceding columns are percentages.

| $d$ | $\gamma d$ | $J$ | MC SE (%) | Old-cutoff change (%) | Residual bound |
|---|---|---|---|---|---|
| 10 | 1 | 444 | 0.162 | 0.22907 | 3.33e-09 |
| 10 | 0.5 | 400 | 0.195 | 0.00007 | 3.81e-09 |
| 10 | 0.1 | 1422 | 0.103 | 0.00001 | 4.07e-09 |
| 100 | 1 | 2303 | 0.085 | 0.28637 | 0 |
| 100 | 0.5 | 4237 | 0.102 | 0.00015 | 3.97e-09 |
| 100 | 0.1 | 14394 | 0.058 | 0.00001 | 4.08e-09 |

# I ADDITIONAL SCALAR CALIBRATION STUDIES

In this section, we conduct a simulation study on three representative stochastic recursions to examine the sharpness of the exact tail approximation in Theorem 2.5. Our goal is not merely to verify that the proposed tail expression gives an upper bound, but to check whether the empirical tail probability is well calibrated by the exact one-big-jump constant predicted by the linearized recursion. We use $\gamma \in \{0.1, 0.01, 0.005, 0.001\}$ and thresholds $y = 0.001, 0.002, \ldots, 3.000$. The denominator is the exact-CDF one-jump sum for the linearized recursion, evaluated over the finite horizon for independent terminal sampling and over the infinite horizon for stationary-chain sampling.

**Smooth nonconvex model.** This is (28) with $\sigma = 1$. Each stepsize uses $10^6$ independent terminal trajectories, with horizons $1000$, $2303$, $5299$, and $34539$. Pointwise probability intervals use the Wilson construction.

**Threshold model.** The drift is

$$m(x) = 5x - 2(x + \delta)\mathbf{1}_{\{x \le -\delta\}}, \qquad \delta = 10^{-7},$$

with independent Student-$t_{1.5}$ additive noise. It has slopes 3 and 5, a breakpoint at $-\delta$, and local derivative $m'(0) = 5$. The objective is strongly convex and locally quadratic near the target, while the drift has a derivative discontinuity at the breakpoint. The local quadratic condition holds on a ball of radius $\kappa \le \delta$. Every reported threshold is at least $10^{-3}$, far beyond that neighborhood, so this experiment probes finite-level behavior across the breakpoint. For each stepsize, 10 independent chains run for $10^6$ steps, discard the first $10^5$ steps, and retain every tenth subsequent iterate. The $900{,}000$ retained observations are dependent within chains.

**Expectile estimation.** Following Newey & Powell (1987), for $\tau \in \{0.4, 0.6\}$ and $Z \sim t_{1.2}$, define $\rho_\tau(r) = |\tau - \mathbf{1}_{\{r \le 0\}}|r^2$ and the finite loss contrast

$$\theta^\star_\tau = \arg\min_\theta \mathbb{E}[\rho_\tau(Z - \theta) - \rho_\tau(Z)].$$

Subtracting before expectation is essential because $\mathbb{E}Z^2 = \infty$. The stochastic gradient is

$$g(\theta, Z) = 2[\tau \mathbf{1}_{\{Z > \theta\}} + (1 - \tau)\mathbf{1}_{\{Z \le \theta\}}](\theta - Z).$$

We use the centered recursion $X_{k+1} = X_k - \gamma g(\theta^\star_\tau + X_k, Z_{k+1})$. The two targets are approximately $-0.831205$ and $0.831205$, and the common population derivative is $H \simeq 0.907776$. Each setting uses 100 independent chains of $10^6$ steps, discards the first $10^5$, and retains every twentieth subsequent iterate, giving $4{,}500{,}000$ observations.

**Dependence-aware uncertainty.** For the threshold and expectile studies, we use the Monte Carlo standard error across independent chain-level tail estimates. We use $\widehat{p} \pm t_{J-1,0.975}\,\widehat{\mathrm{se}}_{\mathrm{chain}}$, clipped to $[0, 1]$, with $J = 10$ and 100, respectively. Dividing by the deterministic benchmark gives the displayed ratio intervals. Thinning is not treated as producing independent trials. The normal approximation across only ten chains is less reliable in the far tail; the resulting wide intervals remain visible. These intervals quantify sampling variation across chains and do not assess residual burn-in bias.

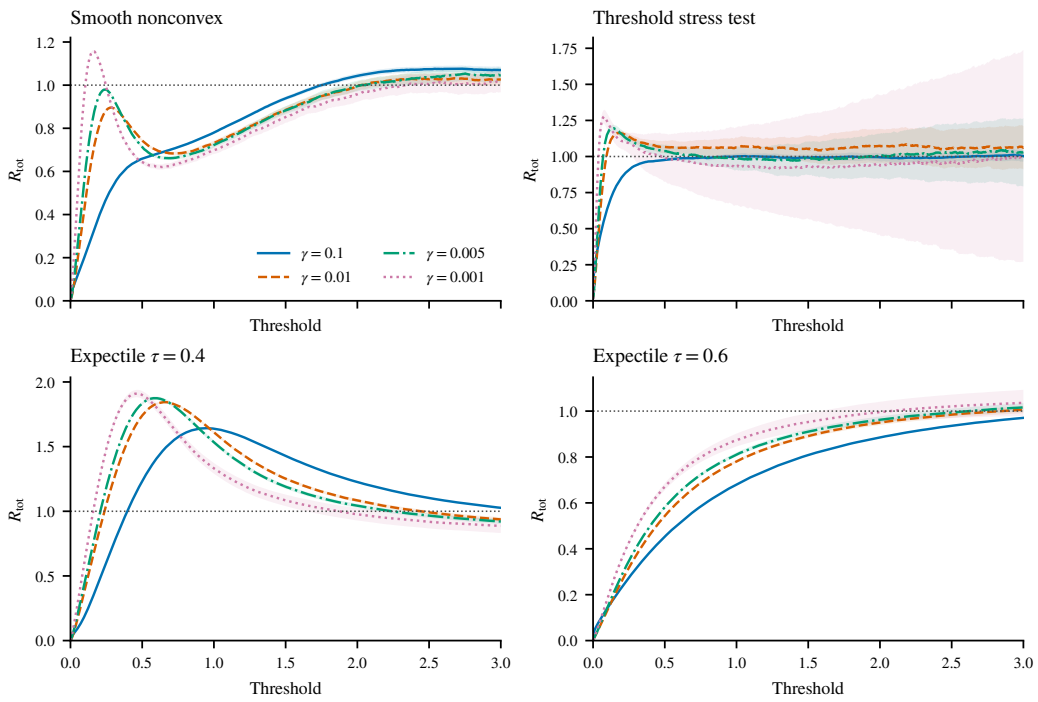


Figure 13: Scalar studies with 95% pointwise intervals. The smooth model uses independent-endpoint Wilson intervals; threshold and expectile models use independent-chain standard errors. Colors and line styles identify stepsizes. The reference line is one.

The smooth model exhibits a clear threshold effect. At $\gamma = 0.001$, $R_{\text{tot}}$ is $0.698$ with interval $[0.680, 0.717]$ at $y = 1$, but $1.016$ with interval $[0.967, 1.068]$ at $y = 3$. Thus decreasing the stepsize at a fixed threshold need not yield the local-Hessian constant. At the same stepsize and $y = 3$, the $\tau = 0.4$ expectile ratio is $0.887$ with interval $[0.832, 0.942]$, so a systematic deficit remains. For $\tau = 0.6$, the ratio is $1.036$ with interval $[0.979, 1.093]$. The threshold model's interval is much wider, $[0.271, 1.738]$, reflecting its ten independent chains. These differences motivate separating linear and transfer errors and complement the shrinking-threshold design in the main text.

**The nonlinear-flow correction.** For the smooth scalar drift, the deterministic factor in Appendix E.7 is $F(u) = tu^t \int_u^\infty r^{-t}/m(r)\,\mathrm{d}r$, with $t = 1.5$ and $H = 1$. Figure 14 shows $F(u) \to 1$ at small thresholds. Numerical integration in log coordinates and an independent physical-coordinate calculation agree within $2 \times 10^{-8}$ at the validation points. The theorem proves this deterministic localization; the comparison with fixed-threshold SGD is an empirical diagnostic of the corresponding nonlinear-flow approximation.

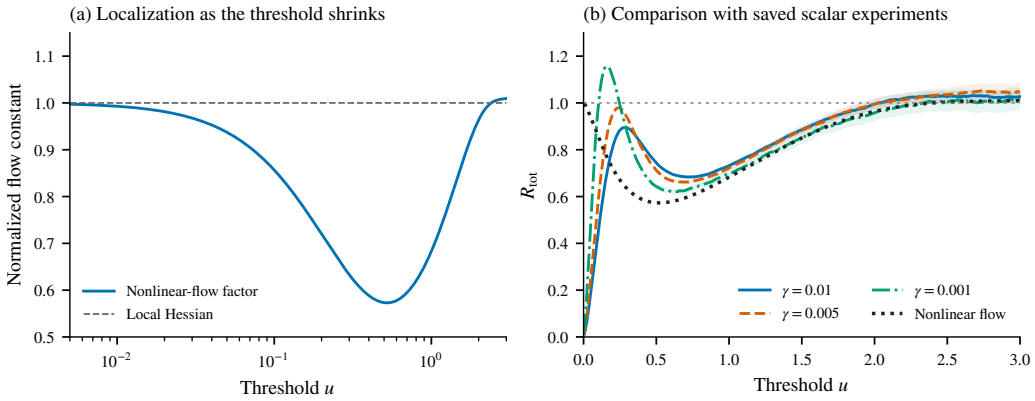


Figure 14: Local curvature and nonlinear flow in the smooth model. (a) The deterministic normalized flow integral approaches one as $u \downarrow 0$. (b) The same factor is overlaid on fixed-threshold SGD ratios, with independent-endpoint pointwise intervals. The selected stepsizes display the approach as $\gamma$ decreases; all four stepsizes remain in Figure 13.

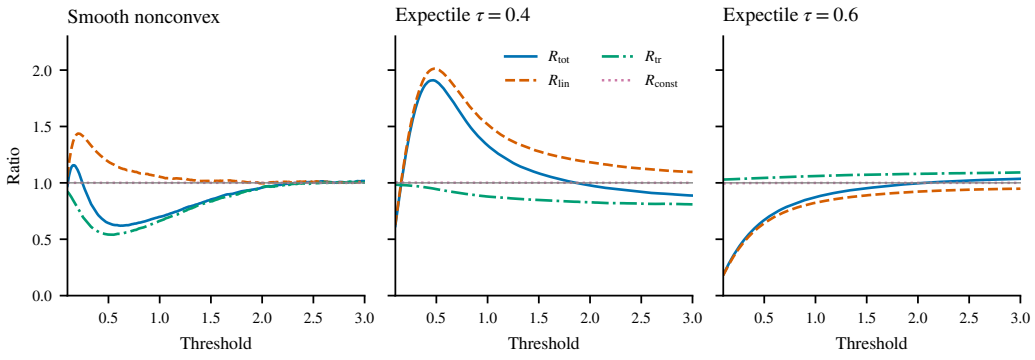


Figure 15: Error decomposition of the smooth and expectile data at $\gamma = 0.001$. Curves show total calibration, linear calibration, nonlinear transfer, and finite-sum versus asymptotic-constant replacement. These diagnostic curves share the same samples and are not independent estimates. Their purpose is to identify which factor contributes to a visible total discrepancy.