STABILITY AND GENERALIZATION OF NOISY signSGD: FROM SIGN DISCONTINUITY TO GAUSSIAN SMOOTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient and scalable stochastic optimization algorithms have made the training of large foundation models possible. More recently, sign-based optimizers such as signSGD have attracted growing attention for their low storage and communication costs, as they update models using only 1-bit gradient signs. Although some studies have begun to investigate the convergence properties of signSGD, its generalization properties remain poorly understood. We establish a lower bound on the generalization error of signSGD. For Lipschitz-continuous and smooth objectives, this lower bound is $\Omega(L\eta\sqrt{d}/n)$ in convex, nonconvex, and strongly convex settings. Establishing a meaningful upper bound is more challenging: for gradient coordinates near zero, even arbitrarily small perturbations can flip their signs and cause non-negligible changes in the updates. To address this difficulty, we interpret Gaussian signSGD as a smoothed surrogate for signSGD. Adding Gaussian noise to the gradients before applying the sign operator smooths the resulting update distribution and enables us to derive a noise-dependent upper bound on the generalization error. Through this Gaussian-smoothed counterpart, the resulting bound provides an indirect upper-bound characterization of signSGD. Its explicit dependence on the noise scale further allows us to quantify how Gaussian perturbations affect generalization. We also conduct experiments on synthetic data, full-parameter fine-tuning of large language models, and computer vision tasks. The results across different models and tasks support our theoretical findings.

1 INTRODUCTION

The rapid development of large foundation models relies heavily on efficient and scalable stochastic optimization algorithms. Among these algorithms, adaptive gradient methods, represented by AdamW (Kingma & Ba, 2017; Loshchilov & Hutter, 2019), have long been the mainstream choice for training foundation models. However, the growing scale of large language models (LLMs) makes distributed training increasingly necessary. Consequently, AdamW's substantial memory footprint and communication overhead become major bottlenecks (Shoeybi et al., 2020; Rajbhandari et al., 2020; Ren et al., 2021; Tang et al., 2021; Li et al., 2021). Sign-based optimizers such as signSGD use only 1-bit gradient signs for model updates and have recently gained attention for their low storage and communication costs (Bernstein et al., 2019).

The sign operator retains only the sign of each gradient coordinate, preventing extreme gradient values from directly amplifying the update step (Bernstein et al., 2018). Under large or heavy-tailed gradient noise, signSGD may also be more robust than methods with unbiased compression (Compagnoni et al., 2025). However, sign-based updates discard the magnitude information in stochastic gradients, which may cause convergence and generalization issues. Various approaches have been proposed to improve sign-based updates, including error compensation (Karimireddy et al., 2019; Zheng et al., 2019), stochastic sign compression (Chen et al., 2019; Safaryan & Richtárik, 2021; Jin et al., 2024; Tang et al., 2023), variance reduction (Chzhen & Schechtman, 2023), local momentum (Sun et al., 2023), and adaptive methods (Crawshaw et al., 2022). Among these approaches, a simple and natural idea is to add Gaussian noise to the gradients before taking their signs (Chen et al., 2019), yielding the Gaussian signSGD algorithm.

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.

So far, many existing studies have primarily focused on the convergence and optimization properties of signSGD and its variants, while their generalization behavior, particularly in terms of generalization error, remains comparatively less understood. Although Huang et al. (2026) established a generalization upper bound of $\mathcal{O}(1/(n\tau^T))$ for signSGD in the nonconvex setting, their analysis relies on a positive lower bound $\tau > 0$ on the absolute values of nonzero gradient coordinates throughout the iterations. As these coordinates approach zero, the bound becomes increasingly loose, limiting its ability to meaningfully characterize the sensitivity of the sign operator near zero.

To overcome this challenge, we study the generalization error of signSGD. The discontinuity of the sign operator at zero allows arbitrarily small perturbations to near-zero gradient coordinates to flip their signs and cause non-negligible changes in the updates. We first derive a lower bound on the generalization error of signSGD to characterize this inherent limitation. Our key observation is that Gaussian perturbations can smooth the update distribution. Motivated by this observation, we add Gaussian noise to the gradients before applying the sign operator and derive a generalization upper bound for the resulting Gaussian signSGD, a smoothed surrogate for signSGD. The bound explicitly depends on the noise scale, allowing us to characterize how Gaussian perturbations affect generalization.

Table 1: Summary of generalization bounds for signSGD-type methods. Our upper bounds are displayed for constant step size $\eta_t = \eta$, and our lower bounds concern $T = 1$.

Algorithm	**Reference**	**Convexity**	**Bound**	ℓ_1	ℓ_2
signSGD	Huang et al. (2026)	Non-convex	$\mathcal{O}\left(\frac{1}{n\tau^T}\right)$	✗	✓
Noisy Sign-SGD	Banerjee et al. (2022b)	Non-convex	$\mathcal{O}\left(\frac{\sqrt{T}}{n}\right)$	✗	✓
signSGD	Theorem 1	Convex	$\Omega\left(\frac{L_2\eta\sqrt{d}}{n}\right)$	✗	✓
		Non-convex		✗	✓
		Strongly convex (Lipschitz on $\mathcal{W}_R$)		✗	✓
Gaussian signSGD	Theorem 3	Convex	$\mathcal{O}\left(\frac{L_1 D_\sigma \eta T}{n}\right)$	✓	✗
	Theorem 2	Non-convex	$\mathcal{O}\left(\frac{L_1 D_\sigma}{a_\sigma n}\left[(1+a_\sigma\eta)^T - 1\right]\right) \quad (a_\sigma > 0)$	✓	✗
	Theorem 4	Strongly convex	$\mathcal{O}\left(\frac{L_K D_{\sigma,K}}{m_{\sigma,K} n}\left[1 - \left(1 - \frac{\eta m_{\sigma,K}}{2}\right)^T\right]\right)$	✗	✓

Our main contributions can be summarized as follows.

1. We expose a fundamental generalization limitation of signSGD: even a single update can incur an expected generalization error of $\Omega(L\eta\sqrt{d}/n)$ across convex, nonconvex, and strongly convex settings, under the respective smoothness and global or bounded-domain Lipschitz conditions.
2. We establish noise-dependent generalization guarantees for Gaussian signSGD and prove a strict worst-case advantage over clean signSGD in a common one-step setting. By making the dependence on the Gaussian noise scale explicit, our analysis further characterizes how the choice of noise level affects the resulting generalization bound. We also derive sharper bounds under convex structure and a time-uniform bound for a lazy projected variant under coordinate curvature.
3. Our synthetic experiments demonstrate that suitable Gaussian noise can simultaneously reduce estimated generalization error and population risk. The evaluation framework also includes full-parameter fine-tuning of large language models and computer vision tasks to assess the broader applicability of this approach.

2 Related Work

Sign-based optimization. SignSGD compresses stochastic gradient directions to one bit per coordinate, reducing the communication costs associated with transmitting gradient information in

distributed optimization (Bernstein et al., 2018). Safaryan & Richtárik (2021) study stochastic sign descent in heterogeneous distributed settings, while Sun et al. (2023) and Jiang et al. (2025a) establish convergence guarantees for momentum-based sign methods under weaker noise or batch-size requirements. Other studies examine noise and geometry: Kornilov et al. (2025) and Yu et al. (2026b) analyze convergence under heavy-tailed noise; Xiao et al. (2025) characterize preconditioning and noise-compression effects through high-dimensional risk dynamics; and Tao et al. (2026) identify complexity advantages over SGD through ℓ_1-stationarity and ℓ_∞-geometry. These results motivate studying how the same sign operation affects sensitivity to individual training examples.

Randomized and noisy SignSGD. Randomization changes the information retained by one-bit updates. StoSignSGD (Yu et al., 2026a) uses coordinate-scaled uniform perturbations to obtain an unbiased, scaled gradient response. Gaussian perturbations instead make the expected sign response vary smoothly with the input gradient. Jin & Dai (2025) analyze privacy amplification and convergence for Gaussian noisy SignSGD in heterogeneous distributed learning. Chen & Wang (2026) study small-batch convergence, annealed Gaussian perturbations before taking signs of momentum, and a calibrated transition to SGD. These methods differ in noise distribution, scaling, and momentum use; our main analysis focuses on fixed-scale Gaussian perturbations of stochastic gradients.

Stability and generalization of sign-based methods. Algorithmic stability connects sensitivity to sample replacement with generalization (Bousquet & Elisseeff, 2002; Hardt et al., 2016). The exponential family Langevin dynamics framework of Banerjee et al. (2022a) includes noisy SignSGD and establishes data-dependent expected-stability bounds with $O(1/n)$ sample dependence. Its noisy-sign instance uses logistic-type sign probabilities, whereas our analysis studies Gaussian perturbations through coupled parameter trajectories. Huang et al. (2026) relate Lion and SignSGD generalization bounds to the minimum nonzero magnitude of the pre-sign gradient estimator and introduce a cautious sign rule.

3 PRELIMINARIES

We use D to denote the data distribution. The training dataset $S = (z_1, \ldots, z_n)$ is a sequence of n independent and identically distributed samples drawn from D. $S, S' \in \mathcal{Z}^n$ are called neighboring datasets if and only if they differ at exactly one data point (we could assume without loss of generality that $z_n \neq z'_n$). For $n \in \mathbb{N}$, we use $[n]$ to denote $\{1, \ldots, n\}$. For $d \in \mathbb{N}$, we use 0_d to denote $(0, \ldots, 0) \in \mathbb{R}^d$ and 1_d to denote $(1, \ldots, 1) \in \mathbb{R}^d$. We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is separable if there exist functions $f_1, \ldots, f_d : \mathbb{R} \to \mathbb{R}$ such that $f(x_1, \ldots, x_d) = \sum_{i=1}^d f_i(x_i)$. Let $\mathbb{R}^d$ be the parameter space, $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ be a loss function, and $A_f : \mathcal{Z}^n \to \mathbb{R}^d$ be a learning algorithm.

Definition 1 (Generalization error) *Let $S = (z_1, \ldots, z_n) \sim D^n$ and assume that the relevant loss evaluations are integrable. Let $R_D(w) := \mathbb{E}_{Z\sim D}[f(w; Z)]$ and $R_S(w) := n^{-1} \sum_{j=1}^n f(w; z_j)$. The expected generalization error is*

$$\mathrm{gen}(A_f; D) := \left| \mathbb{E}_{S, A_f} \left[R_D(A_f(S)) - R_S(A_f(S)) \right] \right|,$$

where $Z \sim D$ is independent of S and the algorithmic randomness, and the outer expectation is over both S and the algorithmic randomness. We write $\mathrm{gen}(A_f)$ when D is clear.

Generalization error measures the absolute expected gap between population and empirical risks, while expected optimization error measures empirical-risk suboptimality in expectation. When the relevant quantities are finite, their sum bounds the expected excess population risk, so small optimization error alone does not guarantee good population performance.

Definition 2 (SignSGD and Gaussian signSGD) *Given $S = (z_1, \ldots, z_n)$, step sizes $\eta_t \geq 0$, and $\sigma \geq 0$, let*

$$W_{t+1} = W_t - \eta_t \operatorname{sign}\big(\nabla f(W_t, z_{i_t}) + \sigma G_t\big), \tag{1}$$

where $i_t \sim \mathrm{Unif}([n])$ and $G_t \sim \mathcal{N}(0, I_d)$ are sampled independently at each step. The sign acts coordinate-wise, with $\mathrm{sign}(0) = 0$. The case $\sigma > 0$ is Gaussian signSGD; $\sigma = 0$ is signSGD.

Definition 3 (Convexity) *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if*

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle \quad \textit{for all } u, v \in \mathbb{R}^d.$$

It is μ-strongly convex if the right-hand side additionally contains $\frac{\mu}{2}\|u - v\|_2^2$.

Assumption 1 (ℓ_p-Lipschitz continuity). *Fix $p \in \{1, 2\}$. There exists a constant $L_p > 0$ such that, for all $w, w' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$|f(w, z) - f(w', z)| \leq L_p \|w - w'\|_p.$$

Assumption 2 (Gradient smoothness in ℓ_p). *Fix $p \in \{1, 2\}$. There exists a constant $\beta_p \geq 0$ such that, for all $w, w' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(w, z) - \nabla f(w', z)\|_p \leq \beta_p \|w - w'\|_p.$$

Algorithmic stability. Intuitively, a stable learning algorithm, meaning that a small perturbation of the training data does not significantly affect its output, can generalize well. Specifically, for a symmetric learning algorithm A_f, whose output distribution is invariant under permutations of the training samples, the replace-one identity in Lemma 1 expresses the expected generalization gap in terms of the expected loss difference between outputs trained on neighboring datasets.

Lemma 1 (Lemma 7 in Bousquet & Elisseeff, 2002) *Let $S = (z_1, \ldots, z_n) \sim D^n$, and let $S^{(i)}$ be obtained by replacing z_i with a fresh sample $z_i' \sim D$, independent of S and the algorithmic randomness. Assume that the relevant loss evaluations are integrable. If A_f is symmetric, then, for every $i \in [n]$,*

$$\mathrm{gen}(A_f; D) = \left| \mathbb{E}_{S, z_i', A_f} \left[f(A_f(S); z_i') - f(A_f(S^{(i)}); z_i') \right] \right|.$$

4 Generalization Error of SignSGD and Gaussian SignSGD

In this section, we examine the generalization limitations of signSGD through lower bounds and investigate Gaussian smoothing as a way to address the instability of sign updates. By smoothing the update distribution, we derive generalization upper bounds for Gaussian signSGD that explicitly characterize the effect of noise intensity, and identify conditions under which it improves upon signSGD in worst-case one-step generalization.

4.1 Generalization Error of SignSGD

We establish lower bounds on the generalization error of signSGD for smooth loss functions in the convex, nonconvex, and strongly convex settings.

Theorem 1 (Generalization lower bounds for signSGD) *Let $n, d \in \mathbb{N}$ and $\eta, L_2 > 0$. For each setting below, under its stated conditions, there exist a distribution D over $\{0, 1\}$ and a differentiable loss function $f : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}$ such that signSGD, initialized at $w_0 = \mathbf{0}_d$ and run for $T = 1$ iteration with step size η, satisfies*

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \Omega\left(\frac{L_2 \eta \sqrt{d}}{n} \right).$$

The loss has the following properties:

1. ***Convex setting.*** *If $\eta\sqrt{d} \leq 1/2$, then, for every $z \in \{0, 1\}$, $f(\cdot, z)$ is convex, L_2-Lipschitz, and β_2-smooth on $\mathbb{R}^d$, with $\beta_2 = 2L_2$.*
2. ***Nonconvex setting.*** *For any $\eta > 0$ and every $z \in \{0, 1\}$, $f(\cdot, z)$ is L_2-Lipschitz and β_2-smooth on $\mathbb{R}^d$, with $\beta_2 = 2L_2$. Moreover, $f(\cdot, 1)$ is nonconvex.*
3. ***Strongly convex setting.*** *For any $R > 0$ satisfying $\eta\sqrt{d} \leq \min\{1/2, R\}$ and every $z \in \{0, 1\}$, $f(\cdot, z)$ is 1-strongly convex and β_2-smooth on $\mathbb{R}^d$, with $\beta_2 = 2L_2 + 1$, and is $(L_2 + R)$-Lipschitz on $\mathcal{W}_R := \{w \in \mathbb{R}^d : \|w\|_2 \leq R\}$.*

All Lipschitz, smoothness, and strong convexity properties are defined with respect to the ℓ_2 norm.

Theorem 1 establishes a worst-case generalization lower bound for signSGD: there exist a distribution D and a convex, L_2-Lipschitz, and $2L_2$-smooth loss in ℓ_2 such that $\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \Omega(L_2 \eta \sqrt{d}/n)$ when $\eta\sqrt{d} \leq 1/2$, $T = 1$, and $w_0 = \mathbf{0}_d$. Since this bound already holds after a

single sign update, the generalization error cannot be attributed solely to accumulation over multiple iterations.

The nonconvex case extends the same-order lower bound without the small-step-size condition, while the strongly convex case shows that it persists for 1-strongly convex losses on a bounded domain. In the latter construction, an additional quadratic term ensures strong convexity but cancels in the generalization gap. We present the proofs of the convex, nonconvex, and strongly convex cases in Sections A.1, A.2, and A.3 of Appendix A, respectively.

Taken together, these results show that convexity, smoothness, or strong convexity alone cannot rule out such worst-case generalization lower bounds. This motivates our study of Gaussian signSGD, where we examine whether injecting Gaussian noise before the sign operation can smooth the discontinuous sign mapping and improve algorithmic stability and generalization.

4.2 Generalization Error of Gaussian SignSGD

Here, we study Gaussian signSGD as a smoothed surrogate for signSGD to address the stability challenge posed by the discontinuous sign operator. Gaussian perturbations smooth the update distribution, enabling us to derive generalization upper bounds through algorithmic stability. These bounds explicitly capture the role of the noise scale, with sharper guarantees under additional structural assumptions.

Definition 4 (Global Data-Sensitivity Constant) *For noise level $\sigma > 0$, define*

$$\kappa_\sigma := \frac{\sqrt{2/\pi}}{\sigma}, \qquad D_\sigma := \min\{2d,\ \kappa_\sigma D_1\},$$

where

$$D_1 := \sup_{w\in\mathbb{R}^d,\, z,z'\in\mathcal{Z}} \|\nabla f(w,z) - \nabla f(w,z')\|_1.$$

Definition 5 (Global Same-Sample Stability Coefficient) *For ℓ_1-smoothness constant $\beta_1 \geq 0$, define $a_\sigma := \kappa_\sigma\beta_1 = \frac{\sqrt{2/\pi}\,\beta_1}{\sigma}$.*

Definition 6 (Coupled Parameter Distance) *Fix a replace-one position r. Let W_t and $W_t^{(r)}$ be the iterates on S and $S^{(r)}$, respectively, coupled using the same initialization, sample indices, and Gaussian noises. Define*

$$\Delta_t^{(r)} := \|W_t - W_t^{(r)}\|_1, \qquad \delta_t^{(r)} := \mathbb{E}\left[\Delta_t^{(r)}\right],$$

where the expectation is over the data and algorithmic randomness.

We use the replace-one stability approach to relate coupled iterates to expected generalization Bousquet & Elisseeff (2002); Hardt et al. (2016). Gaussian perturbation makes the expected sign response Lipschitz in its gradient input, allowing a bound without convexity.

For the next two theorems, let $S \sim D^n$, fix $T \geq 1$ and $\sigma > 0$, and consider equation 5 with deterministic nonnegative step sizes. Assume that the loss and its gradient are jointly measurable, and that S, W_0, the i.i.d. indices $I_t \sim \mathrm{Unif}([n])$, and the i.i.d. noises $\xi_t \sim \mathcal{N}(0,\sigma^2 I_d)$ are mutually independent. For some fixed $w_\star \in \mathbb{R}^d$, assume

$$\mathbb{E}_{Z\sim D}|f(w_\star, Z)| < \infty, \qquad \mathbb{E}\|W_0\|_1 < \infty.$$

Each replacement sample $Z_r' \sim D$ is independent of S and all algorithmic randomness. Couple runs on S and $S^{(r)}$ using the same initialization, indices, and noises, and set $A_T(S) = W_T$. Under Assumption 1, these conditions ensure integrability of the loss evaluations, since $\|W_t - W_0\|_1 \leq d\sum_{s<t}\eta_s$.

Theorem 2 *Under the preceding conditions and Assumptions 1 and 2 with $p = 1$, no convexity is required, and*

$$\mathrm{gen}(A_T) = \mathcal{O}\left(\frac{L_1 D_\sigma}{n}\sum_{t=0}^{T-1}\eta_t\prod_{j=t+1}^{T-1}(1+a_\sigma\eta_j)\right).$$

For $a_\sigma > 0$, the sum telescopes to give

$$\operatorname{gen}(A_T) = \mathcal{O}\left(\frac{L_1 D_\sigma}{a_\sigma n}\left[\prod_{t=0}^{T-1}(1 + a_\sigma \eta_t) - 1\right]\right).$$

Replacing the product by $\exp(a_\sigma \sum_{t=0}^{T-1} \eta_t)$ gives a further upper bound. If $a_\sigma = 0$, the bound reduces to $\operatorname{gen}(A_T) = \mathcal{O}((L_1 D_\sigma / n) \sum_{t=0}^{T-1} \eta_t)$.

The theorem separates data sensitivity from state amplification, measured by a_σ. The bound remains valid at zero gradient coordinates for fixed $\sigma > 0$, but not by setting $\sigma = 0$. Shared Gaussian noise yields expected sign disagreement $2|\Phi(a/\sigma) - \Phi(b/\sigma)| \le \kappa_\sigma |a - b|$. An intermediate same-sample update separates state and data sensitivity, yielding a stability recursion. Unrolling it and applying the replace-one identity gives the generalization bound. The amplification factor is $1 + \kappa_\sigma \beta_1 \eta_t$, versus $1 + \beta \eta_t$ for SGD. This noise-dependent guarantee is not uniformly tighter than SGD's and can grow exponentially in T for constant steps. The next theorem identifies additional structure that removes this amplification.

Theorem 3 *Under the preceding sampling, initialization, integrability, and coupling conditions, suppose every sample loss is coordinate-separable and convex as in Assumption 5. Assume also Assumptions 1 and 2 with $p = 1$, and Assumption 4. Then, for every $r \in [n]$ and $0 \le t < T$,*

$$\delta_{t+1}^{(r)} \le \delta_t^{(r)} + \frac{\eta_t D_\sigma}{n}, \qquad \operatorname{gen}(A_T) = \mathcal{O}\left(\frac{L_1 D_\sigma}{n} \sum_{t=0}^{T-1} \eta_t\right).$$

More generally, if deterministic $b_t \ge 0$ satisfies $\mathbb{E}[H_\sigma(\nabla f(W_t^{(r)}, Z_r), \nabla f(W_t^{(r)}, Z_r'))] \le b_t$ for every $r \in [n]$ and $0 \le t < T$, then

$$\delta_{t+1}^{(r)} \le \delta_t^{(r)} + \frac{\eta_t b_t}{n}, \qquad \operatorname{gen}(A_T) = \mathcal{O}\left(\frac{L_1}{n} \sum_{t=0}^{T-1} \eta_t b_t\right).$$

Under separable convex losses and nested steps, Gaussian signSGD avoids multiplicative amplification. Constant steps yield $\mathcal{O}(L_1 D_\sigma \eta T / n)$, while dyadic steps give logarithmic growth in T. Trajectory-dependent b_t can sharpen the bound when sign-disagreement probabilities are small. Nested steps ensure $W_{t,j} - W_{t,j}^{(r)} \in 2\eta_t \mathbb{Z}$ almost surely. Convexity orders shared-noise signs, and the lattice spacing prevents disagreements from increasing distance. Separability preserves equal coordinates, establishing pathwise non-expansiveness on reachable pairs. Unlike smooth convex SGD, this argument requires additional structure: convexity alone and arbitrary decreasing steps do not suffice. It also requires $\sigma > 0$ to exclude zero signs, so the proof does not transfer unchanged to clean signSGD.

Theorem 4 *Consider the lazy projected noisy-sign algorithm equation 9 under the conditions of Section G and Assumption 8. In particular, each sample loss has coordinate-separable curvature on the fixed box K. Let $0 < \eta m_{\sigma,K}/2 \le 1$ and couple neighboring runs as in equation 15. Write $q := 1 - \eta m_{\sigma,K}/2 \in [0, 1)$, $d_t^{(r)} := \mathbb{E}\|X_t - X_t^{(r)}\|_1$, and $A_T^{\mathrm{lazy}}(S) := X_T$. Then, for every replace-one position r,*

$$d_{t+1}^{(r)} \le q d_t^{(r)} + \frac{\eta D_{\sigma,K}}{2n}, \qquad d_T^{(r)} \le \frac{D_{\sigma,K}}{m_{\sigma,K} n}(1 - q^T).$$

Consequently,

$$\operatorname{gen}(A_T^{\mathrm{lazy}}) = \mathcal{O}\left(\frac{L_K D_{\sigma,K}}{m_{\sigma,K} n}(1 - q^T)\right).$$

For fixed K, η, σ, and curvature and gradient bounds, this is uniformly $\mathcal{O}(L_K D_{\sigma,K}/(m_{\sigma,K} n))$ in T.

The theorem establishes strict contraction for lazy projected Gaussian signSGD under coordinate curvature. With fixed box and gradient constants, it yields a time-uniform loss-gap bound, not convergence to a population minimizer. Common-uniform inverse-CDF coupling orders the lazy kernels, including at adjacent grid points. Expected absolute distance then equals a difference of means,

allowing coordinate curvature to yield contraction. Its strength is $m_{\sigma,K} = \mu_c \kappa_\sigma \exp(-G_\infty^2/(2\sigma^2))$, while clipping preserves the distance bound. Geometric contraction replaces accumulating perturbations by a finite stability level. This requires the modified algorithm and coordinate structure; ordinary strong convexity alone does not suffice. Noise affects both $D_{\sigma,K}$ and $m_{\sigma,K}$: low noise can loosen their ratio, while large noise slows contraction.

4.3 A ONE-STEP SEPARATION BETWEEN GAUSSIAN AND CLEAN SIGNSGD

The following result compares a uniform upper bound with a worst-case lower bound over the same loss class. Both algorithms use uniform sampling with replacement, the same n, d, η, the initialization $w_0 = 0_d$, and one update. For clean signSGD we set $\mathrm{sign}(0) = 0$ coordinate-wise. Gaussian signSGD adds independent $\mathcal{N}(0, \sigma^2 I_d)$ noise before taking signs. Throughout this subsection, gen includes expectation over the algorithmic randomness. Put $c = \sqrt{2/\pi}$ and $\mathcal{W}_R = \{w \in \mathbb{R}^d : \|w\|_2 \le R\}$. For a problem class $\mathcal{C}$, abbreviate the two worst-case errors by

$$\mathcal{E}_\sigma(\mathcal{C}) := \sup_{(f,D)\in\mathcal{C}} \mathrm{gen}(A_f^{\mathrm{noisy}}; D), \qquad \mathcal{E}_0(\mathcal{C}) := \sup_{(f,D)\in\mathcal{C}} \mathrm{gen}(A_f^{\mathrm{signSGD}}; D).$$

Theorem 5 *Let $n, d \in \mathbb{N}$ and $\eta, L, \sigma > 0$. Both clean and Gaussian signSGD use uniform sampling with replacement, step size η, $T = 1$, and $w_0 = 0_d$. The Gaussian algorithm additionally uses fresh pre-sign noise with standard deviation σ.*

Let $\mathcal{C}_{\mathrm{nc}}$ consist of problems (f, D) whose sample losses are differentiable, globally L-Lipschitz and $2L$-smooth, without a convexity requirement. Let $\mathcal{C}_{\mathrm{cv}}$ be the subclass in which every sample loss is convex. The data distribution is arbitrary, and all loss evaluations below are assumed integrable. Then

$$\mathcal{E}_\sigma(\mathcal{C}_{\mathrm{nc}}) \le \frac{2cL^2\eta\sqrt{d}}{\sigma n} < \frac{L\eta\sqrt{d}}{4n} \le \mathcal{E}_0(\mathcal{C}_{\mathrm{nc}}) \qquad \text{if } \sigma > 8cL, \tag{2}$$

$$\mathcal{E}_\sigma(\mathcal{C}_{\mathrm{cv}}) \le \frac{2cL^2\eta\sqrt{d}}{\sigma n} < \frac{L\eta\sqrt{d}}{8n} \le \mathcal{E}_0(\mathcal{C}_{\mathrm{cv}}) \qquad \text{if } \eta\sqrt{d} \le \tfrac{1}{2} \text{ and } \sigma > 16cL. \tag{3}$$

For $R > 0$, let $\mathcal{C}_{\mathrm{sc}}(L, R)$ consist of problems (f, D) whose sample losses are differentiable, 1-strongly convex and $(2L+1)$-smooth on $\mathbb{R}^d$, and $(L+R)$-Lipschitz on $\mathcal{W}_R$. Assume again that the relevant loss evaluations are integrable. If $\eta\sqrt{d} \le \min\{1/2, R\}$ and $\sigma > 16c(L+R)^2/L$, then

$$\mathcal{E}_\sigma(\mathcal{C}_{\mathrm{sc}}(L,R)) \le \frac{2c(L+R)^2\eta\sqrt{d}}{\sigma n} < \frac{L\eta\sqrt{d}}{8n} \le \mathcal{E}_0(\mathcal{C}_{\mathrm{sc}}(L,R)). \tag{4}$$

Discussion of Theorem 5. *What the result establishes.* The theorem makes a direct comparison of worst-case expected generalization gaps over the same problem class, with the same initialization, step size, and horizon. Above the stated noise thresholds, a uniform Gaussian-sign upper bound lies strictly below a lower bound on the worst-case clean-sign error. Thus the comparison is stronger than observing that one upper bound decreases with noise, because it separates the actual worst-case errors of the two methods. The statement concerns one update and does not establish a multi-step separation.

At the common initial point, coupled outputs can differ only when the replaced example is sampled, which contributes the factor $1/n$. Gaussian sign sensitivity controls the conditional output difference, while the ℓ_2 gradient bound and $\|v\|_1 \le \sqrt{d}\|v\|_2$ give the explicit dimension dependence. The resulting uniform upper bound is compared with the concrete clean-sign constructions from Theorem 1. Keeping the constants explicit is essential for obtaining strict thresholds.

The separation is between suprema over a class, not a claim that noise improves the generalization gap for every individual loss and data distribution. It also does not compare population risks or establish an advantage over SGD. Larger noise can reduce dependence on the training sample while discarding gradient information needed for optimization. For small noise, the displayed $1/\sigma$ upper bound may be too loose to certify a separation, even if an improvement occurs on a particular problem. The synthetic experiments therefore report population risk together with the generalization gap to distinguish these two effects.

The noise scale affects what these bounds can establish. For small σ, the one-step upper bound grows as $1/\sigma$ and may be too loose to certify a separation; the general multi-step bound can also become loose through its expansion factor $a_\sigma \propto 1/\sigma$. Larger noise can reduce the expected generalization gap without improving population risk.

5 EXPERIMENTS

Our theory characterizes the noise dependence of Gaussian signSGD's generalization bounds. We examine its empirical effects through synthetic experiments, full-parameter LLM fine-tuning, and computer vision tasks, evaluating how noise scale affects generalization and predictive performance.

5.1 SYNTHETIC EXPERIMENTS

To examine how Gaussian noise affects generalization and optimization, we compare clean and Gaussian signSGD under fixed schedules and separately tuned configurations. Experimental details and additional results are provided in Appendix J.1.

Table 2 reports representative results after separate tuning. On both convex and nonconvex objectives, Gaussian signSGD reduces the estimated absolute expected gap and population risk at $p = 0.75$, but increases both at $p = 0.5$. The fixed-schedule experiments further show that a smaller absolute expected gap need not imply lower population risk. These findings complement our stability analysis: the practical effect of Gaussian smoothing depends on the noise scale and data distribution, and its effects on optimization and generalization must be considered jointly.

Table 2: Tuned original signSGD results on S1 (convex) and S3 (nonconvex), selected separately on development data and evaluated on independent confirmation datasets. Gap estimates show clean $\rightarrow$ Gaussian. $\Delta\overline{R}_{\mathrm{pop},T}$ denotes Gaussian minus clean population risk, with dataset-level paired 95% bootstrap CIs.

Task	p	ρ	$\widehat{\mathrm{gen}}_T$	$\Delta\overline{R}_{\mathrm{pop},T}$ [95% CI]
S1	0.50	1.0	.000283 $\rightarrow$ .000520	$+.001124$ $[+.001079, +.001171]$
S1	0.75	0.3	.001596 $\rightarrow$ .001499	$-.005194$ $[-.005440, -.004977]$
S3	0.50	1.0	.000794 $\rightarrow$ .001596	$+.000265$ $[+.000092, +.000457]$
S3	0.75	1.0	.000896 $\rightarrow$ .000844	$-.003937$ $[-.004620, -.002931]$

5.2 FULL FINE-TUNING EXPERIMENTS

To validate our theoretical analysis and better understand how Gaussian perturbations affect the generalization performance of sign-based optimizers, we conduct full-parameter fine-tuning experiments on Qwen3-8B (Yang et al., 2025) across mathematical reasoning tasks. We train exclusively on a 10,000-example split from NuminaMath-CoT (Li et al., 2024) and evaluate the resulting models on multiple mathematical reasoning benchmarks outside the training distribution. We primarily compare AdamW (Loshchilov & Hutter, 2019), SignSGD, and Gaussian SignSGD with different noise scales. We evaluate mathematical reasoning on MathQA (Amini et al., 2019) and BBH Object Counting (Suzgun et al., 2023). On MathQA, we report both the standard accuracy (Acc) and normalized accuracy (Acc-norm); on BBH Object Counting, we report accuracy. All methods use the same data partition, training budget, and evaluation protocol. Detailed hyperparameters, hardware configurations, data partitioning, and implementation details are provided in Appendix J.2.

Table 3 reports the zero-shot accuracy of Qwen3-8B after full-parameter fine-tuning with AdamW, signSGD, and Gaussian signSGD at different noise scales. Gaussian signSGD with $\rho = 0.25$ achieves the best performance on all three metrics, reaching 52.66% MathQA accuracy, 51.79% MathQA normalized accuracy, and 86.80% BBH Object Counting accuracy. Performance improves as ρ increases from 0.15 to 0.25 and then decreases slightly at $\rho = 0.30$, indicating that a moderate noise scale provides the most favorable trade-off in these experiments. Our stability analysis provides a theoretical perspective on these gains: Gaussian smoothing reduces the sensitivity of sign updates to gradient perturbations and yields noise-dependent generalization guarantees. The performance decline at larger noise scales further suggests that translating this smoothing effect into predictive gains requires preserving useful gradient information.

Table 3: Full-parameter fine-tuning results on Qwen3-8B. We report accuracy (%) on MathQA and BBH Object Counting. MathQA is evaluated using both accuracy and normalized accuracy.

Method	MathQA		BBH Counting
	Acc	Acc-norm	
AdamW	50.72 ± 0.92	50.18 ± 0.92	81.60 ± 2.46
SignSGD	51.76 ± 0.91	50.92 ± 0.92	83.60 ± 2.35
Gaussian SignSGD ($\rho = 0.15$)	51.62 ± 0.91	50.62 ± 0.92	85.20 ± 2.25
Gaussian SignSGD ($\rho = 0.20$)	52.16 ± 0.91	51.19 ± 0.92	86.00 ± 2.20
Gaussian SignSGD ($\rho = 0.25$)	**52.66 ± 0.91**	**51.79 ± 0.91**	**86.80 ± 2.15**
Gaussian SignSGD ($\rho = 0.30$)	52.36 ± 0.91	51.42 ± 0.91	85.60 ± 2.22

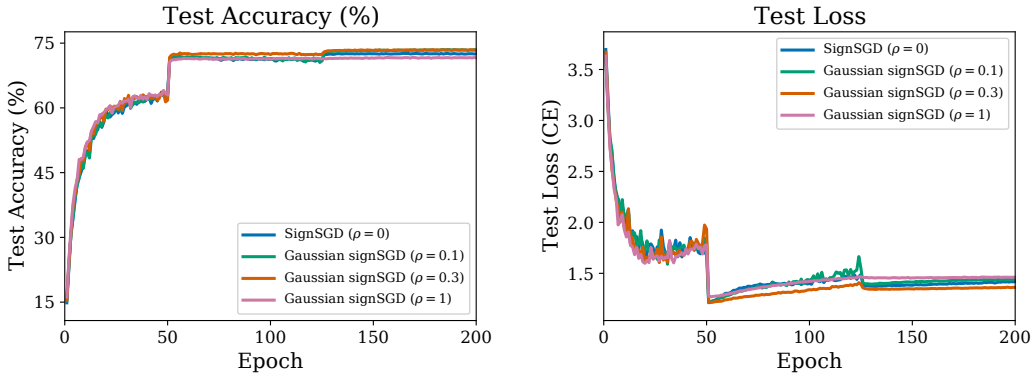


Figure 1: Test accuracy and cross-entropy loss on CIFAR-100 throughout training for signSGD and Gaussian signSGD with different relative noise scales ρ.

5.3 Computer Vision Experiments

For completeness, we also evaluate Gaussian signSGD on a standard computer vision task to examine its broader applicability beyond language model training. Specifically, we perform multiclass image classification on the CIFAR-100 dataset (Krizhevsky, 2009) using the ResNet-18 architecture (He et al., 2015), following the experimental settings of Yuan et al. (2025) and Jiang et al. (2025b). Detailed experimental configurations are provided in Appendix J.3.

The evolution of test accuracy and cross-entropy loss is illustrated in Figure 1. Among all variants, Gaussian signSGD with a moderate relative noise scale ($\rho = 0.3$) achieves the highest final test accuracy and the lowest final test loss, outperforming vanilla signSGD. In contrast, excessive noise ($\rho = 1$) degrades both metrics, indicating that the generalization benefit of Gaussian smoothing depends on an appropriate noise scale.

6 Conclusion

We studied the generalization of Gaussian signSGD. A single clean update can incur worst-case error $\Omega(L\eta\sqrt{d}/n)$ under the stated convex, nonconvex, and strongly convex conditions. Our analysis gives noise-dependent upper bounds, with sharper convex and time-uniform guarantees under additional conditions. In a common one-step setting, sufficiently large noise separates the Gaussian upper bound from the clean worst-case lower bound. Synthetic results show distribution- and noise-dependent gains in population risk and sometimes the generalization gap. Gaussian signSGD improves selected Qwen3-8B results and is broadly comparable on CIFAR-100, highlighting the need to assess optimization alongside generalization.

AI USE DISCLOSURE

We used AI-assisted tools only for language refinement and proofreading during the preparation of this manuscript. Specifically, AI tools were used to improve grammar, readability, and overall writing clarity. All research ideas, methodology, experiments, analysis, and conclusions were developed and verified by the authors. And the authors take full responsibility for the content of this manuscript.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2357–2367. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245/.

Arindam Banerjee, Tiancong Chen, Xinyan Li, and Yingxue Zhou. Stability based generalization bounds for exponential family Langevin dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1412–1449. PMLR, 2022a. URL https://proceedings.mlr.press/v162/banerjee22a.html.

Arindam Banerjee, Tiancong Chen, Xinyan Li, and Yingxue Zhou. Stability based generalization bounds for exponential family langevin dynamics, 2022b. URL https://arxiv.org/abs/2201.03064.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant, 2019. URL https://arxiv.org/abs/1810.05291.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. doi: 10.1162/153244302760200704.

Haoran Chen and Wentao Wang. Enhancing SignSGD: Small-batch convergence analysis and a hybrid switching strategy, 2026. URL https://arxiv.org/abs/2604.25550v1. arXiv preprint, version v1.

Xiangyi Chen, Tiancong Chen, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. Distributed training with heterogeneous data: Bridging median- and mean-based algorithms, 2019. URL https://arxiv.org/abs/1906.01736.

Evgenii Chzhen and Sholom Schechtman. Signsvrg: fixing signsgd via variance reduction, 2023. URL https://arxiv.org/abs/2305.13187.

Enea Monzio Compagnoni, Rustem Islamov, Frank Norbert Proske, and Aurelien Lucchi. Unbiased and sign compression in distributed learning: Comparing noise resilience via sdes, 2025. URL https://arxiv.org/abs/2502.17009.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd, 2022. URL https://arxiv.org/abs/2208.11195.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Feihu Huang, Guanyi Zhang, and Songcan Chen. CLion: Efficient cautious lion optimizer with enhanced generalization, 2026. URL https://arxiv.org/abs/2604.14587v1. arXiv preprint, version v1.

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

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates, 2025b. URL https://arxiv.org/abs/2507.12091.

Richeng Jin and Huaiyu Dai. Noisy SIGNSGD is more differentially private than you (Might) think. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27957–27978. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/jin25b.html.

Richeng Jin, Xiaofan He, Caijun Zhong, Zhaoyang Zhang, Tony Q. S. Quek, and Huaiyu Dai. Magnitude matters: Fixing signsgd through magnitude-aware sparsification and error feedback in the presence of data heterogeneity. *IEEE Transactions on Signal Processing*, 72:5140–5155, 2024. doi: 10.1109/TSP.2024.3454986.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Nikita Kornilov, Philip Zmushko, Andrei Semenov, Mark Ikonnikov, Alexander Gasnikov, and Alexander Beznosikov. Sign operator for coping with heavy-tailed noise in non-convex optimization: High probability bounds under (l_0, l_1)-smoothness, 2025. URL https://arxiv.org/abs/2502.07923.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Conglong Li, Ammar Ahmad Awan, Hanlin Tang, Samyam Rajbhandari, and Yuxiong He. 1-bit lamb: Communication efficient large-scale large-batch training with lamb's convergence speed, 2021. URL https://arxiv.org/abs/2104.06069.

Jia Li et al. NuminaMath. Hugging Face repository, 2024. URL https://huggingface.co/datasets/AI-MO/NuminaMath-CoT.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training, 2021. URL https://arxiv.org/abs/2101.06840.

Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9224–9234. PMLR, 2021. URL https://proceedings.mlr.press/v139/safaryan21a.html.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33077–33099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sun23l.html.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-acl.824/.

Hanlin Tang, Shaoduo Gan, Ammar Ahmad Awan, Samyam Rajbhandari, Conglong Li, Xiangru Lian, Ji Liu, Ce Zhang, and Yuxiong He. 1-bit adam: Communication efficient large-scale training with adam's convergence speed. In *International Conference on Machine Learning*, pp. 10118–10129. PMLR, 2021.

Zhiwei Tang, Yanmeng Wang, and Tsung-Hui Chang. z-signfedavg: A unified stochastic sign-based compression for federated learning, 2023. URL https://arxiv.org/abs/2302.02589.

Hongyi Tao, Dingzhi Yu, and Lijun Zhang. When and why SignSGD outperforms SGD: A theoretical study based on ℓ_1-norm lower bounds. *arXiv preprint arXiv:2605.06615*, 2026.

Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signSGD in high-dimensions: quantifying preconditioning and noise-compression effects. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68391–68439. PMLR, 2025. URL https://proceedings.mlr.press/v267/xiao25c.html.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dingzhi Yu, Rui Pan, Yuxing Liu, Difan Zou, and Tong Zhang. StoSignSGD: Unbiased structural stochasticity fixes SignSGD for training large language models, 2026a. URL https://arxiv.org/abs/2604.15416v2. arXiv preprint, version v2.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026b.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. Mars: Unleashing the power of variance reduction for training large models, 2025. URL https://arxiv.org/abs/2411.10438.

Shuai Zheng, Ziyue Huang, and James T. Kwok. Communication-efficient distributed blockwise momentum sgd with error-feedback, 2019. URL https://arxiv.org/abs/1905.10936.

A GENERALIZATION ERROR OF SIGNSGD

A.1 PROOF OF THEOREM 1: CONVEX SETTING

Take $D = \mathrm{Ber}(1/2)$ and define $f(w, 0) = 0$ and $f(w, 1) = H(w)$, where H is given in Lemma 2. By that lemma, $f(\cdot, z)$ is convex, L-Lipschitz, and $2L$-smooth for every $z \in \{0, 1\}$.

Let $S_1 = (1, z_2, \ldots, z_n)$ and $S_0 = (0, z_2, \ldots, z_n)$, where $z_2, \ldots, z_n \overset{\text{i.i.d.}}{\sim} D$. With the replace-one identity, the contribution vanishes in expectation when the replaced samples are equal; it also vanishes when $z_1' = 0$ because $f(w, 0) = 0$. Thus,

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{1}{4}\left|\mathbb{E}\Big[H(A_f^{\mathrm{signSGD}}(S_1)) - H(A_f^{\mathrm{signSGD}}(S_0))\Big]\right|.$$

Conditional on $k = \sum_{i=2}^n z_i$, Lemma 3 gives $A_f^{\mathrm{signSGD}}(S_j) \overset{d}{=} \eta B_j \mathbf{1}_d$, where $B_1 \sim \mathrm{Ber}((k+1)/n)$ and $B_0 \sim \mathrm{Ber}(k/n)$. Hence

$$\mathbb{E}\Big[H(A_f^{\mathrm{signSGD}}(S_1)) - H(A_f^{\mathrm{signSGD}}(S_0)) \mid k\Big] = \frac{1}{n}\left[H(\eta \mathbf{1}_d) - H(\mathbf{0}_d)\right].$$

The right-hand side does not depend on k. Therefore, Lemma 2 and $\eta\sqrt{d} \le 1/2$ yield

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n}(1 - \eta\sqrt{d}) \ge \frac{L\eta\sqrt{d}}{8n}.$$

Lemma 2 *Let $d \in \mathbb{N}$ and $L > 0$. Define*

$$H(w) = Lh\left(\left\|w - \frac{1}{2\sqrt{d}}\mathbf{1}_d\right\|_2\right),$$

where

$$h(r) = \begin{cases} r^2, & r \le \frac{1}{2}, \\ r - \frac{1}{4}, & r > \frac{1}{2}. \end{cases}$$

Then H is convex, L-Lipschitz, and $2L$-smooth. Moreover, $\nabla H(\mathbf{0}_d) = -\frac{L}{\sqrt{d}}\mathbf{1}_d$.

Furthermore, for any $\eta > 0$ satisfying $0 \le \eta\sqrt{d} \le 1$,

$$H(\eta \mathbf{1}_d) - H(\mathbf{0}_d) = -L\eta\sqrt{d}\left(1 - \eta\sqrt{d}\right).$$

Lemma 3 *Let $S = (z_1, \ldots, z_n) \in \{0, 1\}^n$ and $K := \sum_{i=1}^n z_i$. Suppose that SignSGD is initialized at $w_0 = \mathbf{0}_d$, with $\nabla f(\mathbf{0}_d, 0) = \mathbf{0}_d$ and $\mathrm{sign}(\nabla f(\mathbf{0}_d, 1)) = s\mathbf{1}_d$ for some $s \in \{-1, 1\}$. Then, conditioned on S, the output of one-step SignSGD satisfies*

$$A_f^{\mathrm{signSGD}}(S) \overset{d}{=} -\eta s B \mathbf{1}_d, \qquad B \mid K \sim \mathrm{Ber}\left(\frac{K}{n}\right).$$

For the three constructions considered in this paper, this gives:

- *For the convex construction, $f(w, 0) = 0$ and $f(w, 1) = H(w)$. Thus $s = -1$ and $A_f^{\mathrm{signSGD}}(S) \overset{d}{=} \eta B \mathbf{1}_d$.*
- *For the non-convex construction, $f(w, 0) = 0$ and $f(w, 1) = -H(w)$. Thus $s = 1$ and $A_f^{\mathrm{signSGD}}(S) \overset{d}{=} -\eta B \mathbf{1}_d$.*
- *For the strongly convex construction, $f(w, z) = zH(w) + \frac{1}{2}\|w\|_2^2$. Thus $s = -1$ and $A_f^{\mathrm{signSGD}}(S) \overset{d}{=} \eta B \mathbf{1}_d$.*

A.1.1 Proof of Lemma 2

Set $c_d = (2\sqrt{d})^{-1}\mathbf{1}_d$. The derivative of h increases as $2r$ up to $r = 1/2$ and equals 1 thereafter. Thus h is convex and nondecreasing, which implies that $H(w) = Lh(\|w - c_d\|_2)$ is convex.

Inside the ball of radius $1/2$ centered at c_d, the gradient is $2L(w - c_d)$; outside the ball, it is $L(w - c_d)/\|w - c_d\|_2$. These expressions agree at the boundary and can be written as

$$\nabla H(w) = 2L\Pi_{1/2}(w - c_d),$$

where $\Pi_{1/2}$ is projection onto the ball of radius $1/2$ centered at the origin. The projection is bounded and non-expansive, so $\|\nabla H(w)\|_2 \le L$ and ∇H is $2L$-Lipschitz. Hence H is L-Lipschitz and $2L$-smooth. Since $\|c_d\|_2 = 1/2$, this formula also gives $\nabla H(\mathbf{0}_d) = -L\mathbf{1}_d/\sqrt{d}$.

For $0 \le \eta\sqrt{d} \le 1$, both $\mathbf{0}_d$ and $\eta\mathbf{1}_d$ lie in the quadratic region. Thus $H(\mathbf{0}_d) = L/4$ and $H(\eta\mathbf{1}_d) = L(\eta\sqrt{d} - 1/2)^2$, yielding

$$H(\eta\mathbf{1}_d) - H(\mathbf{0}_d) = -L\eta\sqrt{d}(1 - \eta\sqrt{d}).$$

A.1.2 Proof of Lemma 3

Fix a dataset S containing K samples equal to 1. At $w_0 = \mathbf{0}_d$, sampling a point with label 0 gives zero gradient and leaves the iterate at $\mathbf{0}_d$. Sampling a point with label 1 gives sign vector $s\mathbf{1}_d$ and moves the iterate to $-\eta s\mathbf{1}_d$. Since the index is sampled uniformly, the latter event has probability K/n. Therefore,

$$A_f^{\mathrm{signSGD}}(S) \overset{d}{=} -\eta s B\mathbf{1}_d, \qquad B \mid K \sim \mathrm{Ber}\left(\frac{K}{n}\right).$$

It remains to identify s in each construction. Lemma 2 gives $\nabla H(\mathbf{0}_d) = -L\mathbf{1}_d/\sqrt{d}$, so $s = -1$ for the convex construction and $s = 1$ when H is replaced by $-H$. In the strongly convex construction, the added quadratic term has zero gradient at the initialization; hence its sign vector is the same as in the convex construction, with $s = -1$. Substituting these three values of s gives the claimed output distributions.

A.2 Proof of Theorem 1: Nonconvex setting

Take $D = \mathrm{Ber}(1/2)$ and define $f(w, 0) = 0$ and $f(w, 1) = g(w) := -H(w)$, where H is given in Lemma 2. By that lemma, g is concave and non-convex, while remaining L-Lipschitz and $2L$-smooth.

Let $S_1 = (1, z_2, \ldots, z_n)$ and $S_0 = (0, z_2, \ldots, z_n)$, where $z_2, \ldots, z_n \overset{\text{i.i.d.}}{\sim} D$. As in the proof of Theorem 1, the replace-one identity gives

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{1}{4}\left|\mathbb{E}\Big[g(A_f^{\mathrm{signSGD}}(S_1)) - g(A_f^{\mathrm{signSGD}}(S_0))\Big]\right|.$$

Conditional on $k = \sum_{i=2}^n z_i$, Lemma 3 gives $A_f^{\mathrm{signSGD}}(S_j) \overset{d}{=} -\eta B_j\mathbf{1}_d$, where $B_1 \sim \mathrm{Ber}((k+1)/n)$ and $B_0 \sim \mathrm{Ber}(k/n)$. Since $\|-\eta\mathbf{1}_d - c_d\|_2 = \eta\sqrt{d} + 1/2$, the linear branch of h applies and gives $H(-\eta\mathbf{1}_d) = L(\eta\sqrt{d} + 1/4)$ and $H(\mathbf{0}_d) = L/4$. Thus $g(-\eta\mathbf{1}_d) - g(\mathbf{0}_d) = -L\eta\sqrt{d}$, and hence

$$\mathbb{E}\Big[g(A_f^{\mathrm{signSGD}}(S_1)) - g(A_f^{\mathrm{signSGD}}(S_0)) \mid k\Big] = -\frac{L\eta\sqrt{d}}{n}.$$

The right-hand side does not depend on k. Therefore,

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n} = \Omega\left(\frac{L\eta\sqrt{d}}{n}\right).$$

A.3 Proof of Theorem 1: Strongly convex setting

Take $D = \mathrm{Ber}(1/2)$ and define $f(w, z) = zH(w) + \frac{1}{2}\|w\|_2^2$ for $z \in \{0, 1\}$, where H is given in Lemma 2. By Lemma 4, this loss has the required strong convexity, Lipschitzness, and smoothness properties.

Let $S = (z_1, \ldots, z_n)$ and $K := \sum_{i=1}^n z_i$. Since $z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2)$, we have $K \sim \mathrm{Bin}(n, 1/2)$. Conditioned on K, Lemma 3 gives

$$A_f^{\mathrm{signSGD}}(S) \overset{d}{=} \eta B \mathbf{1}_d, \qquad B \mid K \sim \mathrm{Ber}\left(\frac{K}{n}\right).$$

The empirical and population risks are $f(w; S) = \frac{K}{n}H(w) + \frac{1}{2}\|w\|_2^2$ and $f(w; D) = \frac{1}{2}H(w) + \frac{1}{2}\|w\|_2^2$, respectively. Hence

$$f(w; S) - f(w; D) = \left(\frac{K}{n} - \frac{1}{2}\right) H(w),$$

so the common strongly convex quadratic term cancels from the generalization gap.

Write $\Delta_H := H(\eta \mathbf{1}_d) - H(\mathbf{0}_d)$. Conditioning on K gives

$$\mathbb{E}\Big[H(A_f^{\mathrm{signSGD}}(S)) \mid K\Big] = H(\mathbf{0}_d) + \frac{K}{n}\Delta_H.$$

Let $X := K/n$. Since $K \sim \mathrm{Bin}(n, 1/2)$, $\mathbb{E}[X] = 1/2$ and $\mathrm{Var}(X) = 1/(4n)$. Therefore,

$$\mathbb{E}\left[\left(X - \frac{1}{2}\right) X\right] = \mathrm{Var}(X) = \frac{1}{4n}.$$

Using $\mathbb{E}[X - 1/2] = 0$, we obtain

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \left|\mathbb{E}\left[\left(X - \frac{1}{2}\right)(H(\mathbf{0}_d) + X\Delta_H)\right]\right| = \frac{|\Delta_H|}{4n}.$$

By Lemma 2 and $\eta\sqrt{d} \le 1/2$,

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n}(1 - \eta\sqrt{d}) \ge \frac{L\eta\sqrt{d}}{8n}.$$

Finally, $\eta\sqrt{d} \le R$ ensures that the nonzero one-step iterate $\eta\mathbf{1}_d$ belongs to $\mathcal{W}_R$. This completes the proof of the strongly convex case of Theorem 1.

Lemma 4 *For every $z \in \{0, 1\}$, the function $f(\cdot, z)$ is 1-strongly convex and $(2L + 1)$-smooth on $\mathbb{R}^d$. Moreover, on*

$$\mathcal{W}_R = \left\{w \in \mathbb{R}^d : \|w\|_2 \le R\right\},$$

$f(\cdot, z)$ is $(L + R)$-Lipschitz. At the initialization $w = \mathbf{0}_d$,

$$\nabla f(\mathbf{0}_d, 0) = \mathbf{0}_d,$$

and

$$\nabla f(\mathbf{0}_d, 1) = -\frac{L}{\sqrt{d}}\mathbf{1}_d.$$

A.3.1 Proof of Lemma 4

By Lemma 2, H is convex, L-Lipschitz, and $2L$-smooth. For every $z \in \{0, 1\}$, the function

$$zH(w)$$

is convex. Since

$$\frac{1}{2}\|w\|_2^2$$

is 1-strongly convex, their sum

$$f(w,z) = zH(w) + \frac{1}{2}\|w\|_2^2$$

is 1-strongly convex.

Moreover,

$$\nabla f(w,z) = z\nabla H(w) + w.$$

Therefore, for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla f(u,z) - \nabla f(v,z)\|_2 &\le z\|\nabla H(u) - \nabla H(v)\|_2 + \|u-v\|_2 \\ &\le (2L+1)\|u-v\|_2,\end{aligned}$$

which shows that $f(\cdot, z)$ is $(2L+1)$-smooth.

For $w \in \mathcal{W}_R$,

$$\|\nabla f(w,z)\|_2 \le z\|\nabla H(w)\|_2 + \|w\|_2 \le L + R.$$

Hence $f(\cdot, z)$ is $(L+R)$-Lipschitz on $\mathcal{W}_R$.

Finally,

$$\nabla f(\mathbf{0}_d, z) = z\nabla H(\mathbf{0}_d).$$

Therefore,

$$\nabla f(\mathbf{0}_d, 0) = \mathbf{0}_d,$$

while, by Lemma 2,

$$\nabla f(\mathbf{0}_d, 1) = \nabla H(\mathbf{0}_d) = -\frac{L}{\sqrt{d}}\mathbf{1}_d.$$

This completes the proof of Lemma 4.

B GAUSSIAN-SMOOTHED SIGN SENSITIVITY

If f is L-Lipschitz and β-smooth in ℓ_2, then one may take $L_1 = L$ and $\beta_1 = \sqrt{d}\,\beta$. The first choice follows from $\|v\|_2 \le \|v\|_1$; the second follows from $\|v\|_1 \le \sqrt{d}\|v\|_2$. We consider the Gaussian signSGD update

$$W_{t+1} = W_t - \eta_t \operatorname{sign}\big(\nabla f(W_t, Z_{I_t}) + \xi_t\big), \qquad \xi_t \sim \mathcal{N}(0, \sigma^2 I_d), \tag{5}$$

where $\eta_t \ge 0$ and $\sigma > 0$.

Write Φ and ϕ for the CDF and density of a standard normal random variable. The value assigned to $\operatorname{sign}(0)$ is immaterial because each fresh Gaussian perturbation has a continuous conditional law.

Theorem 6 (Vector sensitivity) *Let $a, b \in \mathbb{R}^d$ and let $\xi \in \mathbb{R}^d$ have marginal distribution $\xi_j \sim \mathcal{N}(0, \sigma^2)$ for every coordinate. Then*

$$\mathbb{E}\|\operatorname{sign}(a+\xi) - \operatorname{sign}(b+\xi)\|_1 \le \kappa_\sigma \|a-b\|_1,$$

and more precisely,

$$\mathbb{E}\|\operatorname{sign}(a+\xi) - \operatorname{sign}(b+\xi)\|_1 \le \sum_{j=1}^{d} \min\{2, \kappa_\sigma |a_j - b_j|\} \le 2d.$$

No independence between the coordinates of ξ is needed for this result.

B.1 PROOF OF THEOREM 6

By the definition of the ℓ_1 norm and linearity of expectation,

$$\begin{aligned}&\mathbb{E}\|\operatorname{sign}(a+\xi) - \operatorname{sign}(b+\xi)\|_1 \\ &\quad= \sum_{j=1}^{d} \mathbb{E}\left|\operatorname{sign}(a_j+\xi_j) - \operatorname{sign}(b_j+\xi_j)\right|.\end{aligned}$$

Applying Lemma 6 to each marginal gives

$$\begin{aligned}\mathbb{E}\|\operatorname{sign}(a+\xi)-\operatorname{sign}(b+\xi)\|_1 &\le \sum_{j=1}^{d}\min\{2,\kappa_\sigma|a_j-b_j|\}\\ &\le \kappa_\sigma\sum_{j=1}^{d}|a_j-b_j| = \kappa_\sigma\|a-b\|_1.\end{aligned}$$

The first line is also at most $2d$, since every coordinate difference is bounded by 2. No step uses joint independence of the coordinates; only the one-dimensional marginal law of each ξ_j is used.

Lemma 5 *If $\xi \sim \mathcal{N}(0,\sigma^2)$, then, for every $a \in \mathbb{R}$,*

$$\mathbb{E}[\operatorname{sign}(a+\xi)] = 2\Phi(a/\sigma) - 1.$$

Lemma 6 *For any $a, b \in \mathbb{R}$ and $\xi \sim \mathcal{N}(0,\sigma^2)$,*

$$\mathbb{E}\left|\operatorname{sign}(a+\xi)-\operatorname{sign}(b+\xi)\right| \le \min\left\{2,\kappa_\sigma|a-b|\right\}, \qquad \kappa_\sigma := \frac{\sqrt{2/\pi}}{\sigma}.$$

Corollary 1 (Conditional sensitivity)**.** *Let $\mathbb{F}_t$ be a history sigma-algebra, let A_t, B_t be $\mathbb{F}_t$-measurable, and let $\xi_t \sim \mathcal{N}(0,\sigma^2 I_d)$ be independent of $\mathbb{F}_t$. Then*

$$\mathbb{E}\left[\|\operatorname{sign}(A_t+\xi_t)-\operatorname{sign}(B_t+\xi_t)\|_1 \mid \mathbb{F}_t\right] \le \kappa_\sigma\|A_t-B_t\|_1.$$

B.1.1 Proof of Lemma 5

Since $\mathbb{P}(a+\xi=0)=0$,

$$\begin{aligned}\mathbb{E}[\operatorname{sign}(a+\xi)] &= 2\mathbb{P}(a+\xi\ge 0)-1\\ &= 2\mathbb{P}(\xi\ge -a)-1 = 2\Phi(a/\sigma)-1.\end{aligned}$$

B.1.2 Proof of Lemma 6

The two signs differ exactly when ξ lies between $-a$ and $-b$. Consequently,

$$\mathbb{E}\left|\operatorname{sign}(a+\xi)-\operatorname{sign}(b+\xi)\right| = 2\left|\Phi(a/\sigma)-\Phi(b/\sigma)\right|.$$

By the mean value theorem, there is a point u between a/σ and b/σ such that

$$|\Phi(a/\sigma)-\Phi(b/\sigma)| = \phi(u)\left|\frac{a-b}{\sigma}\right| \le \frac{|a-b|}{\sigma\sqrt{2\pi}},$$

because $\sup_x \phi(x) = \phi(0) = 1/\sqrt{2\pi}$. Therefore,

$$\mathbb{E}\left|\operatorname{sign}(a+\xi)-\operatorname{sign}(b+\xi)\right| \le \frac{2|a-b|}{\sigma\sqrt{2\pi}} = \kappa_\sigma|a-b|.$$

The two signs take values in $\{-1,1\}$, so their absolute difference is at most 2 pointwise. Taking the minimum of the two bounds proves the claim.

B.1.3 Proof of Corollary 1

Let

$$X_t := \|\operatorname{sign}(A_t+\xi_t)-\operatorname{sign}(B_t+\xi_t)\|_1.$$

For every realization of $\mathbb{F}_t$, the vectors A_t and B_t are fixed, while independence implies that the conditional law of ξ_t remains $\mathcal{N}(0,\sigma^2 I_d)$. Theorem 6 can therefore be applied pointwise under the conditional law, yielding

$$\mathbb{E}[X_t \mid \mathbb{F}_t] \le \kappa_\sigma\|A_t-B_t\|_1.$$

Define the data sensitivity

$$D_1 := \sup_{w,z,z'}\|\nabla f(w,z)-\nabla f(w,z')\|_1, \qquad D_\sigma := \min\{\kappa_\sigma D_1, 2d\}.$$

Assumption 1 also implies $|\partial_j f(w,z)| \le L_1$ for every coordinate. Indeed, for $h \neq 0$ and the jth standard basis vector e_j,

$$\frac{|f(w+he_j,z)-f(w,z)|}{|h|} \le L_1.$$

Letting $h \to 0$ and using differentiability gives $|\partial_j f(w,z)| \le L_1$. Consequently,

$$\|\nabla f(w,z)\|_1 = \sum_{j=1}^{d} |\partial_j f(w,z)| \le dL_1,$$

and, by the triangle inequality,

$$\|\nabla f(w,z) - \nabla f(w,z')\|_1 \le 2dL_1.$$

Thus $D_1 \le 2dL_1 < \infty$ and $D_\sigma \le 2d < \infty$.

B.2 Exact disagreement and margin refinements

Definition 7 (Gaussian sign-disagreement) *For $a, b \in \mathbb{R}^d$, define*

$$H_\sigma(a,b) := 2\sum_{j=1}^{d} |\Phi(a_j/\sigma) - \Phi(b_j/\sigma)|. \tag{6}$$

Equivalently, under the shared-noise coupling,

$$H_\sigma(a,b) = \mathbb{E}_\xi \|\operatorname{sign}(a+\xi) - \operatorname{sign}(b+\xi)\|_1.$$

Thus $H_\sigma/2$ is the expected Hamming distance between the two noisy-sign vectors. We will retain this quantity instead of immediately replacing it by $\kappa_\sigma\|a-b\|_1$ or $2d$.

Lemma 7 (Margin-sensitive disagreement) *For $a, b \in \mathbb{R}^d$, let*

$$\gamma_j(a,b) := \inf\{|u| : u \text{ lies between } a_j \text{ and } b_j\}.$$

Then

$$H_\sigma(a,b) \le \kappa_\sigma \sum_{j=1}^{d} \exp\left(-\frac{\gamma_j(a,b)^2}{2\sigma^2}\right) |a_j - b_j|.$$

In particular, if the segment between a_j and b_j stays outside $(-\gamma,\gamma)$ on every coordinate where $a_j \neq b_j$, the global $1/\sigma$ factor is multiplied by $\exp(-\gamma^2/(2\sigma^2))$.

B.2.1 Proof of Lemma 7

For each coordinate,

$$2|\Phi(a_j/\sigma) - \Phi(b_j/\sigma)| = \left| \int_{b_j}^{a_j} \frac{\sqrt{2/\pi}}{\sigma} e^{-u^2/(2\sigma^2)}\, \mathrm{d}u \right|.$$

The integrand on the interval between a_j and b_j is at most $\kappa_\sigma e^{-\gamma_j(a,b)^2/(2\sigma^2)}$. Summing the resulting coordinate bounds proves the claim.

C Synchronized stability analysis

Let

$$S = (Z_1,\dots,Z_n) \sim D^n, \qquad I_t \overset{\text{i.i.d.}}{\sim} \operatorname{Unif}([n]), \qquad \xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2 I_d), \quad t \ge 0.$$

Assume that the data S, the initialization W_0, and the complete sequences $\{I_t\}_{t\ge0}$ and $\{\xi_t\}_{t\ge0}$ are mutually independent. The step sizes $\eta_t \ge 0$ are deterministic.

Fix $r \in \{1,\dots,n\}$. Let $Z_r' \sim D$ be independent of S and define the replace-one dataset

$$S^{(r)} = (Z_1,\dots,Z_{r-1},Z_r',Z_{r+1},\dots,Z_n).$$

The replacement sample is also independent of all algorithmic randomness. For this fixed r, set

$$\mathbb{F}_t^{(r)} = \sigma(S, Z'_r, W_0, I_0, \xi_0, \ldots, I_{t-1}, \xi_{t-1}).$$

We abbreviate this as $\mathbb{F}_t$ in proofs. Both iterates at time t are $\mathbb{F}_t$-measurable, and (I_t, ξ_t) is independent of $\mathbb{F}_t$. Run the two algorithms with the same W_0, I_t, and ξ_t:

$$W_{t+1} = W_t - \eta_t \operatorname{sign}\big(\nabla f(W_t, Z_{I_t}) + \xi_t\big),$$
$$W_{t+1}^{(r)} = W_t^{(r)} - \eta_t \operatorname{sign}\big(\nabla f(W_t^{(r)}, Z_{I_t}^{(r)}) + \xi_t\big).$$

Set

$$\Delta_t^{(r)} := \|W_t - W_t^{(r)}\|_1, \qquad \delta_t^{(r)} := \mathbb{E}[\Delta_t^{(r)}].$$

C.1 Global baseline

Define the global same-sample coefficient

$$a_\sigma := \kappa_\sigma \beta_1 = \frac{\sqrt{2/\pi}\,\beta_1}{\sigma}.$$

The following baseline uses only the standing assumptions and (a_σ, D_σ).

Theorem 7 (Global one-step stability) *For $t = 0, \ldots, T-1$,*

$$\mathbb{E}[\Delta_{t+1}^{(r)} \mid \mathbb{F}_t] \le (1 + a_\sigma \eta_t)\, \Delta_t^{(r)} + \frac{\eta_t D_\sigma}{n},$$

and consequently

$$\delta_{t+1}^{(r)} \le (1 + a_\sigma \eta_t)\, \delta_t^{(r)} + \frac{\eta_t D_\sigma}{n}.$$

By Theorem 6 and gradient smoothness, the conditional expectation of the first sign difference is at most $\kappa_\sigma \beta_1 \Delta_t^{(r)} = a_\sigma \Delta_t^{(r)}$. The second difference is zero for $i \neq r$. For $i = r$, its conditional expectation is at most $\min\{2d, \kappa_\sigma D_1\} = D_\sigma$. Averaging over the independent uniform index gives the conditional recursion. The tower property gives the unconditional recursion. This argument is valid for all $n \ge 1$.

Corollary 2 (Multi-step parameter stability)**.** *Since $\delta_0^{(r)} = 0$,*

$$\delta_T^{(r)} \le \frac{D_\sigma}{n} \sum_{t=0}^{T-1} \eta_t \prod_{j=t+1}^{T-1} (1 + \eta_j \kappa_\sigma \beta_1) \le \frac{D_\sigma}{n} \sum_{t=0}^{T-1} \eta_t \exp\left(\kappa_\sigma \beta_1 \sum_{j=t+1}^{T-1} \eta_j\right).$$

C.2 Proof of Theorem 7

Conditional on $\mathbb{F}_t$ and $I_t = i$, define

$$G_i = \operatorname{sign}(\nabla f(W_t, Z_i) + \xi_t),$$
$$\widetilde{G}_i = \operatorname{sign}(\nabla f(W_t^{(r)}, Z_i) + \xi_t),$$
$$G'_i = \operatorname{sign}(\nabla f(W_t^{(r)}, Z_i^{(r)}) + \xi_t).$$

Insert $\widetilde{G}_i$ between G_i and G'_i. The update formula and the triangle inequality give

$$\Delta_{t+1}^{(r)} \le \Delta_t^{(r)} + \eta_t \|G_i - \widetilde{G}_i\|_1 + \eta_t \|\widetilde{G}_i - G'_i\|_1.$$

C.2.1 Proof of Corollary 2

Set

$$a_t := 1 + \eta_t \kappa_\sigma \beta_1, \qquad c_t := \frac{\eta_t D_\sigma}{n}.$$

The recursion is $\delta_{t+1}^{(r)} \le a_t \delta_t^{(r)} + c_t$. Repeated substitution gives

$$\delta_T^{(r)} \le \delta_0^{(r)} \prod_{j=0}^{T-1} a_j + \sum_{t=0}^{T-1} c_t \prod_{j=t+1}^{T-1} a_j.$$

Since $\delta_0^{(r)} = 0$, substituting a_j, c_t gives the first bound. Finally, $1 + x \le e^x$ for $x \ge 0$ implies

$$\prod_{j=t+1}^{T-1} (1 + \eta_j \kappa_\sigma \beta_1) \le \exp\left(\kappa_\sigma \beta_1 \sum_{j=t+1}^{T-1} \eta_j \right),$$

which proves the exponential form.

C.3 Trajectory and effective-coordinate refinements

The baseline can be sharpened by retaining exact disagreement probabilities before applying global bounds.

Assumption 3 (Trajectory sign sensitivity)**.** *There are deterministic numbers $\alpha_t, b_t \ge 0$ such that, for every replace-one position r and every training index i,*

$$\mathbb{E} H_\sigma\big(\nabla f(W_t, Z_i), \nabla f(W_t^{(r)}, Z_i)\big) \le \alpha_t \delta_t^{(r)},$$

and

$$\mathbb{E} H_\sigma\big(\nabla f(W_t^{(r)}, Z_r), \nabla f(W_t^{(r)}, Z_r')\big) \le b_t.$$

The expectations include the training data and the coupled trajectory up to time t. This is an on-average condition along reachable iterates, not a uniform condition over all w, z, z'. The global assumptions already made in the paper imply the valid but possibly loose choices

$$\alpha_t = \kappa_\sigma \beta_1, \qquad b_t = D_\sigma.$$

Corollary 3 (Trajectory-sensitive stability refinement)**.** *Under Assumption 3,*

$$\delta_{t+1}^{(r)} \le (1 + \eta_t \alpha_t) \delta_t^{(r)} + \frac{\eta_t b_t}{n}.$$

Consequently,

$$\delta_T^{(r)} \le \frac{1}{n} \sum_{t=0}^{T-1} \eta_t b_t \prod_{j=t+1}^{T-1} (1 + \eta_j \alpha_j).$$

Corollary 4 (Dimension-free margin specialization)**.** *Suppose that, along the coupled reachable trajectories, the following hold. For every same-sample pair of gradients, at most s_{par} coordinates differ; on each such coordinate the segment between the two gradient values stays outside $(-\gamma, \gamma)$; and*

$$\max_j |\partial_j f(w, z) - \partial_j f(w', z)| \le \beta_\infty \|w - w'\|_1. \tag{7}$$

For a replace-one pair at a common parameter, suppose that at most s_{data} gradient coordinates differ. Then Assumption 3 holds with

$$\alpha_t = s_{\mathrm{par}} \beta_\infty \kappa_\sigma \exp\left(-\frac{\gamma^2}{2\sigma^2} \right), \qquad b_t = 2 s_{\mathrm{data}}.$$

In particular, if $s_{\mathrm{par}}, s_{\mathrm{data}}, \beta_\infty$, and γ are independent of d, then the refined multi-step bound is dimension-free. If $\gamma > 0$, its same-sample expansion coefficient tends to zero rather than diverging as $\sigma \downarrow 0$, provided the same positive margin and structural constants hold uniformly for that family of trajectories. A dimension-free generalization bound additionally requires a dimension-independent L_1 and step-size schedule.

C.3.1 Proof of Corollary 3

For every sampled index, insert the sign vector evaluated at $(W_t^{(r)}, Z_{I_t})$ between the two original sign vectors. Conditional on $\mathbb{F}_t$ and I_t, take the expectation over the fresh shared noise. The triangle inequality and Definition 7 give

$$\mathbb{E}[\Delta_{t+1}^{(r)} \mid \mathbb{F}_t] \le \Delta_t^{(r)} + \frac{\eta_t}{n}\sum_{i=1}^{n} H_\sigma(\nabla f(W_t, Z_i), \nabla f(W_t^{(r)}, Z_i)) + \frac{\eta_t}{n} H_\sigma(\nabla f(W_t^{(r)}, Z_r), \nabla f(W_t^{(r)}, Z_r')).$$

Taking total expectations and using both parts of the assumption proves the refined recursion. This step uses unconditional trajectory bounds only after averaging over the independent current index. Starting from $\delta_0^{(r)} = 0$ and repeatedly substituting the recursion gives the refined multi-step bound.

C.3.2 Proof of Corollary 4

Apply Lemma 7 on the at most s_{par} active coordinates and then use equation 7. For a replace-one pair, each of the at most s_{data} active coordinates contributes at most 2 to equation 6.

Remark 1 (What dimension-free means)**.** *The preceding corollary replaces the ambient dimension by an effective number of sign-sensitive coordinates. Without such structure, a replace-one update can disagree in all d coordinates, producing an ℓ_1 discrepancy of $2\eta_t d$ on that event. Thus the coordinate sensitivity estimates may scale linearly in d. This observation concerns this parameter-distance analysis; it is not by itself a lower bound on every possible generalization guarantee.*

D From replace-one stability to generalization

Theorem 8 (Expected generalization bound) *Let the two outputs lie in a common set K_0, and suppose $|f(w,z) - f(w',z)| \le L_1\|w - w'\|_1$ for all $w, w' \in K_0$ and z. Assume that the loss evaluations appearing below are integrable. For every fixed pair of datasets, the conditional law of each coupled output must equal the algorithm's law on its own dataset; preserving unconditional output marginals alone is not sufficient. Then*

$$\mathrm{gen}(A) \le \frac{L_1}{n}\sum_{r=1}^{n}\mathbb{E}\|A(S) - A(S^{(r)})\|_1.$$

Theorem 9 (Global baseline generalization bound) *Under the standing assumptions for the original algorithm,*

$$\mathrm{gen}(A_T) \le \frac{L_1 D_\sigma}{n}\sum_{t=0}^{T-1}\eta_t\prod_{j=t+1}^{T-1}(1 + a_\sigma\eta_j).$$

Corollary 5 (Exact-disagreement refinement)**.** *If Assumption 3 also holds, then*

$$\mathrm{gen}(A_T) \le \frac{L_1}{n}\sum_{t=0}^{T-1}\eta_t b_t\prod_{j=t+1}^{T-1}(1 + \eta_j\alpha_j).$$

The global choices in the specialization to (a_σ, D_σ) recover Theorem 9. Corollary 4 gives effective-coordinate and margin refinements. The trajectory coefficients may themselves depend on d and σ; a dimension-free statement requires uniform control of those coefficients, L_1 and the step-size schedule.

The baseline uses no convexity and thus also covers globally Lipschitz, smooth convex losses. Global strong convexity on $\mathbb{R}^d$ is incompatible with global Lipschitz continuity; Section G uses a separate bounded-domain setting. The factor $a_\sigma \propto 1/\sigma$ may make the global bound large at small noise. This is a property of the upper bound, not a comparison theorem between the noisy and noiseless algorithms.

D.1 Proof of Theorem 8

Let Z_r' be an independent copy of Z_r. Swapping the i.i.d. pair (Z_r, Z_r') gives

$$\mathbb{E}f(A(S), Z_r) = \mathbb{E}f(A(S^{(r)}), Z_r').$$

Hence,

$$\begin{aligned}
\mathrm{gen}(A) &= \left| \frac{1}{n} \sum_{r=1}^{n} \mathbb{E}\left[f(A(S), Z_r') - f(A(S), Z_r) \right] \right| \\
&= \left| \frac{1}{n} \sum_{r=1}^{n} \mathbb{E}\left[f(A(S), Z_r') - f(A(S^{(r)}), Z_r') \right] \right| \\
&\leq \frac{1}{n} \sum_{r=1}^{n} \mathbb{E}\left| f(A(S), Z_r') - f(A(S^{(r)}), Z_r') \right| \\
&\leq \frac{L_1}{n} \sum_{r=1}^{n} \mathbb{E}\|A(S) - A(S^{(r)})\|_1.
\end{aligned}$$

D.2 Proof of Theorem 9

Apply Theorem 8 with $K_0 = \mathbb{R}^d$ and Corollary 2. The parameter-distance bound is the same for every replaced position, so averaging over r leaves it unchanged.

D.2.1 Proof of Corollary 5

Combine Theorem 8 with Corollary 3.

E Generalization Error of Gaussian SignSGD

E.1 Proof of Theorem 2

For every r, Corollary 2 gives the same upper bound, because its right-hand side does not depend on the replaced position. Applying Theorem 8 and averaging over r therefore gives the general nonconvex bound. For $a_\sigma > 0$, define

$$Q_t := \prod_{j=t}^{T-1} (1 + a_\sigma \eta_j), \qquad Q_T := 1.$$

Then

$$\eta_t \prod_{j=t+1}^{T-1} (1 + a_\sigma \eta_j) = \frac{Q_t - Q_{t+1}}{a_\sigma}.$$

Summing from $t = 0$ to $T - 1$ gives

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t \prod_{j=t+1}^{T-1} (1 + a_\sigma \eta_j) &= \frac{1}{a_\sigma} \sum_{t=0}^{T-1} (Q_t - Q_{t+1}) \\
&= \frac{Q_0 - Q_T}{a_\sigma} \\
&= \frac{1}{a_\sigma} \left[\prod_{t=0}^{T-1} (1 + a_\sigma \eta_t) - 1 \right].
\end{aligned}$$

This proves the product form. Since $a_\sigma \eta_t \geq 0$, $1 + a_\sigma \eta_t \leq e^{a_\sigma \eta_t}$, which proves the exponential form. If $a_\sigma = 0$, Theorem 7 reduces to $\delta_{t+1}^{(r)} \leq \delta_t^{(r)} + \eta_t D_\sigma / n$; summing from zero to $T - 1$ and applying Theorem 8 proves the zero-coefficient bound.

Corollary 6 (Common step-size schedules). *If $\eta_t = C/T$, then, for $a_\sigma > 0$,*

$$\mathrm{gen}(A_T) \le \frac{L_1 D_\sigma}{a_\sigma n}(e^{a_\sigma C} - 1),$$

and, if $a_\sigma = 0$, then

$$\mathrm{gen}(A_T) \le \frac{C L_1 D_\sigma}{n}.$$

If $\eta_t = C/(t+1)$ and $T \ge 1$, then, for $a_\sigma > 0$,

$$\mathrm{gen}(A_T) \le \frac{L_1 D_\sigma}{a_\sigma n}\left[(eT)^{a_\sigma C} - 1\right],$$

while for $a_\sigma = 0$,

$$\mathrm{gen}(A_T) \le \frac{C L_1 D_\sigma}{n}\log(eT).$$

E.1.1 Proof of Corollary 6

For $\eta_t = C/T$, the total step size is C. For $\eta_t = C/(t+1)$, $\sum_{t=0}^{T-1}\eta_t \le C\log(eT)$. Substitute these estimates into Theorem 2.

F Convex losses of Gaussian signSGD: non-expansiveness from structure

Throughout this section, we study the original, unprojected Gaussian signSGD update. The sample losses are globally Lipschitz and smooth, and all loss evaluations are integrable. The data, initialization, sample indices, and Gaussian noises satisfy the stated independence conditions. Runs on neighboring datasets share the same initialization, sample indices, and Gaussian noises. General convex losses satisfy the baseline bound of Theorem 2. Improved bounds below are proved from a step-size condition and explicit loss structure.

F.1 Why arbitrary-pair non-expansiveness is unnecessary

Convexity alone does not make the sign update non-expansive for arbitrary inputs. For $f(w) = w^2/2$, take $w = \varepsilon$, $w' = 0$ and $0 < \varepsilon < \eta$. With shared Gaussian noise, the signs disagree with probability $p_\varepsilon = \Phi(\varepsilon/\sigma) - 1/2 > 0$, so

$$\mathbb{E}_\xi|w^+ - w'^+| = \varepsilon + 2p_\varepsilon(\eta - \varepsilon) > \varepsilon.$$

The quadratic is strongly convex but not globally Lipschitz. The globally convex, 1-Lipschitz and 1-smooth loss $f(w) = \log\cosh w$ gives the same counterexample with $p_\varepsilon = \Phi(\tanh(\varepsilon)/\sigma) - 1/2$. These pairs need not be reachable under the synchronized coupling. The proofs below exploit precisely this distinction.

Define

$$\mathcal{T}_{t,z}(w;\xi) = w - \eta_t \,\mathrm{sign}(\nabla f(w,z) + \xi).$$

Let $\mathcal{R}_t$ be the set of coupled pairs reachable at time t, over all replace-one positions and admissible data realizations. All reachable-pair statements can equivalently be read as almost-sure statements for each coupled random pair.

F.2 Proof of Theorem 3

Fix r. Lemma 8 makes Lemma 9 applicable to the current pair. For each sampled index i, insert $\mathcal{T}_{t,Z_i}(W_t^{(r)};\xi_t)$ between the two updates. The expected same-sample distance is at most $\Delta_t^{(r)}$. The other term is zero if $i \ne r$; if $i = r$, its conditional expectation is exactly

$$\eta_t H_\sigma(\nabla f(W_t^{(r)}, Z_r), \nabla f(W_t^{(r)}, Z_r')) \le \eta_t D_\sigma.$$

Average over the independent uniform index and then over the history. This proves the stated recursion, or the recursion with b_t after taking total expectations of the exact sensitivity. Sum from zero, use $\delta_0^{(r)} = 0$, and apply Theorem 8.

Assumption 4 (Nested positive step sizes)**.** *For a fixed horizon $T \geq 1$, all steps are positive and deterministic, and*

$$\frac{\eta_t}{\eta_{t+1}} \in \mathbb{N}_+, \qquad 0 \leq t < T-1.$$

Lemma 8 (Lattice of synchronized coordinate differences) *Under Assumption 4, for every replace-one position r, coordinate j, and $0 \leq t < T$, almost surely*

$$W_{t,j} - W_{t,j}^{(r)} \in 2\eta_t \mathbb{Z}.$$

Consequently, a nonzero coordinate difference at time t has absolute value at least $2\eta_t$.

Assumption 5 (Coordinate-separable convex losses)**.** *For every sample z,*

$$f(w, z) = \sum_{j=1}^{d} f_j(w_j, z),$$

where every $f_j(\cdot, z)$ is differentiable and convex. An additive sample-dependent constant can be included in one summand.

Lemma 9 (Pathwise non-expansiveness on the difference lattice) *Under Assumption 5, fix t and any pair w, w' with $w - w' \in (2\eta_t \mathbb{Z})^d$. Then, for every sample z, almost surely in the shared Gaussian noise,*

$$\|\mathcal{T}_{t,z}(w; \xi) - \mathcal{T}_{t,z}(w'; \xi)\|_1 \leq \|w - w'\|_1.$$

Corollary 7 (Fixed, horizon-normalized and dyadic step sizes)**.** *Under the loss assumptions of Theorem 3, each of the following schedules satisfies Assumption 4. For fixed steps $\eta_t = \eta > 0$,*

$$\operatorname{gen}(A_T) \leq \frac{L_1 D_\sigma \eta T}{n}.$$

For a prescribed horizon and $\eta_t = C/T$ with $C > 0$,

$$\operatorname{gen}(A_T) \leq \frac{C L_1 D_\sigma}{n}.$$

For the dyadic staircase schedule

$$\eta_t = c\, 2^{-\lfloor \log_2(t+1) \rfloor}, \qquad c > 0,$$

we have

$$\operatorname{gen}(A_T) \leq \frac{c L_1 D_\sigma}{n}(1 + \lfloor \log_2 T \rfloor).$$

F.2.1 Proof of Lemma 8

The initial difference is zero. At step s, the increment in the difference belongs to $\{-2\eta_s, 0, 2\eta_s\}$ almost surely, since the two noisy signs are in $\{-1, 1\}$. Thus, for integers $\nu_{s,j} \in \{-1, 0, 1\}$,

$$W_{t,j} - W_{t,j}^{(r)} = 2 \sum_{s=0}^{t-1} \eta_s \nu_{s,j}.$$

For $s < t < T$, the ratio $\eta_s/\eta_t = \prod_{k=s}^{t-1} (\eta_k/\eta_{k+1})$ is a positive integer. Dividing the displayed sum by $2\eta_t$ proves the claim. The null events of zero noisy gradients have probability zero simultaneously over the finitely many times and coordinates.

F.2.2 Proof of Lemma 9

Fix j and write $h = w_j - w'_j$. If $h = 0$, separability makes the two gradient coordinates equal, so the coordinate outputs coincide. If $h > 0$, the lattice gives $h \geq 2\eta_t$. Convexity gives $f'_j(w_j, z) \geq f'_j(w'_j, z)$. Under shared noise their sign difference is either zero or two. The updated difference is therefore h or $h - 2\eta_t$, both in $[0, h]$. For $h < 0$, interchange w and w'. Summing the coordinate inequalities proves the result.

F.2.3 Proof of Corollary 7

The first two schedules have successive ratio one. The dyadic schedule has successive ratio one or two. For $k = 0, \dots, \lfloor \log_2 T \rfloor$, the block $2^k \le t + 1 < 2^{k+1}$ contains at most 2^k relevant indices, each with step $c2^{-k}$. Each block contributes at most c.

F.3 Weak coordinate coupling via column diagonal dominance

The preceding pathwise coordinate argument uses separability. We next allow nonzero cross derivatives and prove expected non-expansiveness from a verifiable derivative condition. Set

$$F_{\sigma,z}(w) = 2\Phi(\nabla f(w,z)/\sigma) - \mathbf{1},$$
$$r_j(w,z) = \kappa_\sigma \exp\left(-\frac{(\partial_j f(w,z))^2}{2\sigma^2}\right).$$

The CDF in this definition acts coordinate-wise.

Assumption 6 (Weighted column diagonal dominance)**.** *Each $f(\cdot, z)$ is convex and twice continuously differentiable on an open neighborhood of a deterministic convex set $\mathcal{U}_T$. Almost surely, $\mathcal{U}_T$ contains both coupled iterates at every $t < T$, for every replaced position. For all $w \in \mathcal{U}_T$, z, and $k = 1, \dots, d$,*

$$r_k(w,z)\,\partial^2_{kk} f(w,z) \ge \sum_{j \ne k} r_j(w,z)\,|\partial^2_{jk} f(w,z)|. \tag{C}$$

The condition must hold along the entire line segments between coupled states, not only at the iterates.

Theorem 10 (Convex stability under weak coordinate coupling) *Under the standing conditions of this section, Assumptions 4 and 6, every reachable pair satisfies*

$$\mathbb{E}_\xi \|\mathcal{T}_{t,z}(w;\xi) - \mathcal{T}_{t,z}(w';\xi)\|_1 \le \|w - w'\|_1.$$

Consequently, the one-step recursion, the resulting generalization bound, and the optional replacement by b_t hold. The schedules of Corollary 7 are admissible, provided the weighted dominance condition is verified on their corresponding reachable regions.

F.4 Proof of Theorem 10

Apply Lemma 8, then substitute the dominance-dissipation estimate into the lattice-distance identity. This proves expected non-expansiveness. The same intermediate-update decomposition used in Theorem 3 now gives the stability and generalization statements. No pathwise coordinate non-expansiveness is asserted in the coupled case.

This is column diagonal dominance of the Jacobian $J_{\sigma,z}(w) = \nabla F_{\sigma,z}(w)$, since $(J_{\sigma,z})_{jk} = r_j \partial^2_{jk} f$. In particular, it has a nonnegative diagonal. Convexity alone does not imply the weighted dominance condition.

Lemma 10 (Exact expected distance on the lattice) *Fix w, w' with $h = w - w' \in (2\eta_t \mathbb{Z})^d$, and set $v = F_{\sigma,z}(w) - F_{\sigma,z}(w')$, $I = \{j : h_j \ne 0\}$ and $I^c = \{j : h_j = 0\}$. Then*

$$\mathbb{E}_\xi \|\mathcal{T}_{t,z}(w;\xi) - \mathcal{T}_{t,z}(w';\xi)\|_1 = \|h\|_1 - \eta_t \sum_{j \in I} \operatorname{sign}(h_j) v_j + \eta_t \sum_{j \in I^c} |v_j|.$$

Lemma 11 (Column dominance controls both nonzero and zero coordinates) *Under Assumption 6, for any $w, w' \in \mathcal{U}_T$ and h, v, I as above,*

$$\sum_{j \in I} \operatorname{sign}(h_j) v_j \ge \sum_{j \in I^c} |v_j|.$$

The lattice restriction is not needed for this lemma.

Corollary 8 (A sufficient condition using ordinary Hessian entries)**.** *Suppose $|\partial_j f(w,z)| \le G$ on $\mathcal{U}_T$ for all j, z, and for all $w \in \mathcal{U}_T$, z, k,*

$$\sum_{j \ne k} |\partial^2_{jk} f(w,z)| \le e^{-G^2/(2\sigma^2)} \partial^2_{kk} f(w,z). \tag{8}$$

Then the weighted dominance condition holds.

F.4.1 Proof of Lemma 10

For $j \in I$, $|h_j| \geq 2\eta_t$. Subtracting η_t times a difference of two signs cannot move h_j strictly across zero, regardless of the gradient ordering. Thus its expected absolute value is $|h_j| - \eta_t \operatorname{sign}(h_j) v_j$. For $j \in I^c$, the expected absolute difference is $\eta_t |v_j|$ by the exact one-dimensional disagreement identity. Sum over the coordinates. The second sum cannot be dropped when cross derivatives are present.

F.4.2 Proof of Lemma 11

If $h = 0$, then $v = 0$ and both sides of the dominance-dissipation inequality vanish. Hence assume $h \neq 0$. Throughout the proof, j denotes a row index and k a column index.

The fundamental theorem of calculus gives $v = Mh$, where

$$M = \int_0^1 J_{\sigma,z}(w' + sh)\,\mathrm{d}s.$$

The convexity of $\mathcal{U}_T$ ensures that the entire segment lies in the region where the weighted dominance condition holds. For each fixed column k,

$$\begin{aligned} M_{kk} &= \int_0^1 (J_{\sigma,z}(w' + sh))_{kk}\,\mathrm{d}s \\ &\geq \sum_{j \neq k} \int_0^1 |(J_{\sigma,z}(w' + sh))_{jk}|\,\mathrm{d}s \\ &\geq \sum_{j \neq k} \left| \int_0^1 (J_{\sigma,z}(w' + sh))_{jk}\,\mathrm{d}s \right| = \sum_{j \neq k} |M_{jk}|. \end{aligned}$$

The first inequality uses column diagonal dominance at every point of the segment. The second uses the triangle inequality for each scalar integral.

Now expand the left-hand side of the desired inequality by columns:

$$\begin{aligned} \sum_{j \in I} \operatorname{sign}(h_j)(Mh)_j &= \sum_{j \in I} \operatorname{sign}(h_j) \sum_{k=1}^{d} M_{jk} h_k \\ &= \sum_{k \in I} h_k \sum_{j \in I} \operatorname{sign}(h_j) M_{jk} \\ &= \sum_{k \in I} \left[M_{kk} |h_k| + \sum_{\substack{j \in I \\ j \neq k}} \operatorname{sign}(h_j) M_{jk} h_k \right]. \end{aligned}$$

The second equality interchanges the two finite sums and discards the columns $k \in I^c$, exactly because $h_k = 0$ there. The last equality separates the $j = k$ term and uses $\operatorname{sign}(h_k) h_k = |h_k|$.

For every off-diagonal term with $j, k \in I$,

$$\operatorname{sign}(h_j) M_{jk} h_k \geq -|M_{jk}|\,|h_k|.$$

Applying this scalar absolute-value bound term by term in the column expansion yields

$$\sum_{j \in I} \operatorname{sign}(h_j)(Mh)_j \geq \sum_{k \in I} \left(M_{kk} - \sum_{\substack{j \in I \\ j \neq k}} |M_{jk}| \right) |h_k|.$$

For each $k \in I$, split its off-diagonal rows into $I \setminus \{k\}$ and I^c. The column dominance in the integrated-column-dominance estimate then gives

$$M_{kk} - \sum_{\substack{j \in I \\ j \neq k}} |M_{jk}| = \left(M_{kk} - \sum_{j \neq k} |M_{jk}| \right) + \sum_{j \in I^c} |M_{jk}| \geq \sum_{j \in I^c} |M_{jk}|.$$

Combining the off-diagonal bound and the column-surplus estimate, and rearranging finite sums once more, gives

$$\begin{aligned} \sum_{j \in I} \operatorname{sign}(h_j)(Mh)_j &\geq \sum_{k \in I} \sum_{j \in I^c} |M_{jk}|\,|h_k| \\ &= \sum_{j \in I^c} \sum_{k \in I} |M_{jk} h_k| \\ &\geq \sum_{j \in I^c} \left| \sum_{k \in I} M_{jk} h_k \right| = \sum_{j \in I^c} |(Mh)_j|. \end{aligned}$$

The last inequality is the triangle inequality within each row $j \in I^c$. The last equality again uses $h_k = 0$ for $k \in I^c$. Finally substitute $v = Mh$ to obtain the dominance-dissipation inequality. Thus the contributions from nonzero-difference coordinates control all newly created differences in the zero-difference coordinates.

F.4.3 Proof of Corollary 8

The weights satisfy $\kappa_\sigma e^{-G^2/(2\sigma^2)} \leq r_k \leq \kappa_\sigma$. Multiplying equation 8 by κ_σ bounds the weighted off-diagonal sum by $\kappa_\sigma e^{-G^2/(2\sigma^2)} \partial^2_{kk} f \leq r_k \partial^2_{kk} f$. The diagonal Hessian entry is nonnegative by convexity.

F.5 Cumulative expansivity as a unifying extension

Theorem 11 (Cumulative-expansivity extension) *Under the standing Lipschitz and sampling conditions and Assumption 7,*

$$\operatorname{gen}(A_T) \leq \frac{L_1}{n} \sum_{t<T} \eta_t b_t \prod_{j=t+1}^{T-1} (1 + \lambda_j) \leq \frac{L_1 e^{\Lambda_T}}{n} \sum_{t<T} \eta_t b_t.$$

F.6 Proof of Theorem 11

Insert the same-sample intermediate update for each index, average over the index and history, and use the assumed expansion and data sensitivity: $\delta^{(r)}_{t+1} \leq (1 + \lambda_t)\delta^{(r)}_t + \eta_t b_t / n$. Unroll from zero, apply Theorem 8, and use $1 + x \leq e^x$.

Assumption 7 (Reachable-pair expansivity)**.** *There are deterministic $\lambda_t \geq 0$ such that, for every sample z and every reachable pair,*

$$\mathbb{E}_\xi \|\mathcal{T}_{t,z}(w;\xi) - \mathcal{T}_{t,z}(w';\xi)\|_1 \leq (1 + \lambda_t)\|w - w'\|_1.$$

Let $\Lambda_T = \sum_{t<T} \lambda_t$, and let b_t satisfy the data-sensitivity condition under the same synchronized coupling.

G Modified lazy projected algorithm under coordinate-separable curvature

This section studies a modified lazy projected algorithm under coordinate-separable curvature, with a separate set of assumptions. The argument uses a fixed bounded grid, Bernoulli coordinate updates,

projection, and an inverse-CDF coupling of the transition kernels. Global Lipschitz continuity and smoothness on $\mathbb{R}^d$ are not imposed. The coordinate structure below is stronger than ordinary strong convexity.

For comparison, an unprojected noisy-sign algorithm with a fixed positive step can reach arbitrarily distant states as its horizon increases. Indeed, conditional on a fixed dataset and initial point, every finite sequence of signs has positive probability under fresh Gaussian noise. In particular, $w_0 + \eta T\mathbf{1}$ and $w_0 - \eta T\mathbf{1}$ are reachable. If coordinate strong monotonicity with constant μ_c holds between these points, a common bound on their gradient coordinates must satisfy $G_\infty \geq \mu_c \eta T$. Thus such a bound cannot be independent of T for the unprojected algorithm. The fixed box used below resolves this issue. The laziness of the modified transition also permits an order-preserving coupling on a grid of spacing η.

G.1 Algorithm, assumptions, and constants

Fix a deterministic box $K = \prod_{j=1}^d [\ell_j, u_j]$, a step $\eta > 0$, a noise level $\sigma > 0$, and integers $N_j \geq 1$ satisfying $u_j - \ell_j = N_j \eta$. Set

$$\mathcal{G} = \prod_{j=1}^d \{\ell_j, \ell_j + \eta, \ldots, u_j\},$$

and let Π_K denote coordinate-wise clipping to K. Let $S = (Z_1, \ldots, Z_n)$ consist of independent samples from D. The initial point $X_0 \in \mathcal{G}$ has a fixed distribution independent of the data. At every step, independently of the data and all previous randomness, sample

$$I_t \sim \mathrm{Unif}([n]), \qquad \xi_t \sim \mathcal{N}(0, \sigma^2 I_d), \qquad B_{t,j} \sim \mathrm{Bernoulli}(1/2), \quad j \in [d],$$

with all these variables mutually independent. The algorithm is

$$X_{t+1} = \Pi_K(X_t - \eta B_t \odot \mathrm{sign}(\nabla f(X_t, Z_{I_t}) + \xi_t)). \tag{9}$$

Here $\odot$ denotes coordinate-wise multiplication. Each coordinate either stays put or takes one noisy-sign step before clipping. Hence $\mathcal{G}$ is invariant. The value assigned to $\mathrm{sign}(0)$ is immaterial because $\sigma > 0$.

Assumption 8 (Coordinate curvature and bounded gradients on the box)**.** *The loss and its gradient are jointly measurable in (w, z). Each $f(\cdot, z)$ is continuously differentiable on a neighborhood of K. There exist $\mu_c > 0$ and $G_\infty < \infty$ such that, for all $w, w' \in K$, z, and j,*

$$(\partial_j f(w, z) - \partial_j f(w', z))(w_j - w'_j) \geq \mu_c |w_j - w'_j|^2,$$
$$|\partial_j f(w, z)| \leq G_\infty.$$

If $w_j = w'_j$, then $\partial_j f(w, z) = \partial_j f(w', z)$. For some deterministic $w_\star \in K$, $\mathbb{E}_{Z\sim D}|f(w_\star, Z)| < \infty$.

The equal-coordinate condition says that each gradient coordinate depends only on its corresponding parameter coordinate. On the box, this implies coordinate separability of the loss up to a sample-dependent additive term. Integration along a line segment gives

$$|f(w, z) - f(w', z)| \leq L_K \|w - w'\|_1, \qquad L_K := G_\infty. \tag{10}$$

A smaller valid value of L_K may be used. In particular, $|f(w, z)| \leq |f(w_\star, z)| + L_K \sum_j (u_j - \ell_j)$ for every $w \in K$. Thus the sample and population loss evaluations used below are integrable, including when w depends on the training sample.

Write Φ for the standard normal CDF and define

$$\kappa_\sigma := \frac{\sqrt{2/\pi}}{\sigma}, \qquad D_{1,K} := \sup_{w\in K, z, z'} \|\nabla f(w, z) - \nabla f(w, z')\|_1,$$
$$D_{\sigma,K} := \min\{\kappa_\sigma D_{1,K}, 2d\}, \qquad \underline{\kappa}_{\sigma,K} := \kappa_\sigma \exp\left(-\frac{G_\infty^2}{2\sigma^2}\right),$$
$$m_{\sigma,K} := \mu_c \underline{\kappa}_{\sigma,K}.$$

The gradient bound implies $D_{1,K} \leq 2dG_\infty$. We use $0 < \eta m_{\sigma,K}/2 \leq 1$. All parameters and bounds in this section are fixed independently of the horizon when asserting a bound uniform in T.

G.2 Lazy transition kernel and inverse-CDF coupling

For $a \in \mathbb{R}$, put

$$p_\sigma(a) := \Phi(a/\sigma), \qquad q_\sigma(a) := 2\Phi(a/\sigma) - 1.$$

For a scalar state x, let $V_{x,a}$ have the unprojected transition law

$$\mathbb{P}(V_{x,a} = x - \eta) = \frac{p_\sigma(a)}{2}, \quad \mathbb{P}(V_{x,a} = x) = \frac{1}{2}, \quad \mathbb{P}(V_{x,a} = x + \eta) = \frac{1 - p_\sigma(a)}{2}. \tag{11}$$

This is the law of $x - \eta B \operatorname{sign}(a + \xi)$ for independent $B \sim \text{Bernoulli}(1/2)$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. Its mean is

$$\mathbb{E}V_{x,a} = x - \frac{\eta}{2} q_\sigma(a). \tag{12}$$

Define $F_{x,a}(s) := \mathbb{P}(V_{x,a} \le s)$ and $Q_{x,a}(v) := \inf\{s \in \mathbb{R} : F_{x,a}(s) \ge v\}$ for $0 < v < 1$. Explicitly,

$$Q_{x,a}(v) = \begin{cases} x - \eta, & 0 < v \le p_\sigma(a)/2, \\ x, & p_\sigma(a)/2 < v \le [1 + p_\sigma(a)]/2, \\ x + \eta, & [1 + p_\sigma(a)]/2 < v < 1. \end{cases} \tag{13}$$

For $U = (U_1, \dots, U_d)$ with independent uniform coordinates on $(0, 1)$, define the already projected update

$$\mathcal{P}_z(w; U) := \Pi_K\Big(\big(Q_{w_j, \partial_j f(w,z)}(U_j)\big)_{j=1}^d \Big). \tag{14}$$

Clipping is applied exactly once in this definition.

Lemma 12 (Correct conditional transition law) *For every fixed $w \in \mathcal{G}$ and sample z, $\mathcal{P}_z(w; U)$ has the same distribution as the right-hand side of equation 9 conditional on $X_t = w$ and $Z_{I_t} = z$.*

Fix a replace-one position r, and let $Z'_r \sim D$ be independent of S and all algorithmic randomness. Write $S^{(r)}$ for the sample obtained by replacing Z_r with Z'_r, and $Z_i^{(r)}$ for its ith element. Couple the two runs with the same initial point $X_0^{(r)} = X_0$, the same indices I_t, and the same fresh vectors $U_t \sim \text{Unif}((0, 1)^d)$:

$$X_{t+1} = \mathcal{P}_{Z_{I_t}}(X_t; U_t), \qquad X_{t+1}^{(r)} = \mathcal{P}_{Z_{I_t}^{(r)}}(X_t^{(r)}; U_t). \tag{15}$$

The vectors U_t are independent across time and independent of the data, X_0, and the entire index sequence. In particular, with

$$\mathcal{F}_t^{(r)} = \sigma(S, Z'_r, X_0, I_0, U_0, \dots, I_{t-1}, U_{t-1}),$$

the pair (I_t, U_t) is independent of $\mathcal{F}_t^{(r)}$. Lemma 12 and induction in t show that, conditional on the pair of datasets, each trajectory has the law of the original lazy algorithm on its own dataset. Sharing the uniforms couples the outputs without altering either of these conditional laws.

For use below, recall that

$$H_\sigma(a, b) = \sum_{j=1}^d |q_\sigma(a_j) - q_\sigma(b_j)| = 2 \sum_{j=1}^d |\Phi(a_j/\sigma) - \Phi(b_j/\sigma)|. \tag{16}$$

Since $|q_\sigma| \le 1$ and $0 < q'_\sigma \le \kappa_\sigma$,

$$H_\sigma(a, b) \le \min\{2d, \kappa_\sigma \|a - b\|_1\}. \tag{17}$$

G.2.1 Proof of Lemma 12

For each coordinate, equation 13 assigns probabilities $p_\sigma(a)/2$, $1/2$, and $[1 - p_\sigma(a)]/2$ to the three possible unprojected states, exactly as in equation 11. The coordinates of U are independent, matching the conditional independence of the coordinate pairs $(B_{t,j}, \xi_{t,j})$ in the algorithm. Consequently, the full unprojected vectors have the same distribution. Applying the same deterministic clipping map preserves equality in law, including at boundary states where clipping merges atoms.

G.3 Order and contraction of the coupled transitions

Lemma 13 (Order of the lazy transitions) *If $x > y$ lie on the same grid of spacing η, then, for any $a, b \in \mathbb{R}$ and $0 < v < 1$, $Q_{x,a}(v) \geq Q_{y,b}(v)$. At a common state x, if $a \geq b$, then $Q_{x,a}(v) \leq Q_{x,b}(v)$.*

Lemma 14 (Same-sample contraction and data sensitivity) *Under Assumption 8, for every $w, w' \in \mathcal{G}$ and every pair of samples z, z',*

$$\mathbb{E}_U \|\mathcal{P}_z(w; U) - \mathcal{P}_z(w'; U)\|_1 \leq (1 - \eta m_{\sigma,K}/2)\|w - w'\|_1, \tag{18}$$

$$\mathbb{E}_U \|\mathcal{P}_z(w; U) - \mathcal{P}_{z'}(w; U)\|_1 \leq \frac{\eta}{2} H_\sigma(\nabla f(w, z), \nabla f(w, z')) \leq \frac{\eta D_{\sigma,K}}{2}. \tag{19}$$

G.3.1 Proof of Lemma 13

If $x - y \geq 2\eta$, then $x - \eta \geq y + \eta$, so every point in the first support is at least every point in the second support. It remains to consider $x - y = \eta$. For $s < y$, $F_{x,a}(s) = 0 \leq F_{y,b}(s)$. For $y \leq s < x$,

$$F_{x,a}(s) = \frac{p_\sigma(a)}{2} \leq \frac{1}{2} \leq \frac{1 + p_\sigma(b)}{2} = F_{y,b}(s).$$

For $s \geq x$, $F_{y,b}(s) = 1 \geq F_{x,a}(s)$. Thus $F_{x,a} \leq F_{y,b}$ everywhere, which gives the claimed quantile ordering. At a common state x, increasing a increases both nontrivial values of $F_{x,a}$, since p_σ is increasing. The corresponding quantiles therefore decrease.

G.3.2 Proof of Lemma 14

Fix a coordinate j and write $x = w_j$, $y = w'_j$, $a = \partial_j f(w, z)$, and $b = \partial_j f(w', z)$. If $x = y$, the equal-coordinate condition gives $a = b$, so the coupled coordinate outputs coincide. If $x > y$, Lemma 13 gives pointwise ordering before clipping. For a scalar $U \sim \mathrm{Unif}(0, 1)$,

$$\begin{aligned}\mathbb{E}|Q_{x,a}(U) - Q_{y,b}(U)| &= \mathbb{E}Q_{x,a}(U) - \mathbb{E}Q_{y,b}(U) \\ &= (x - y) - \frac{\eta}{2}\{q_\sigma(a) - q_\sigma(b)\}.\end{aligned}$$

The curvature condition implies $a - b \geq \mu_c(x - y) > 0$. Both gradients lie in $[-G_\infty, G_\infty]$, where

$$q'_\sigma(s) = \kappa_\sigma e^{-s^2/(2\sigma^2)} \geq \underline{\kappa}_{\sigma,K}.$$

Integration from b to a yields

$$q_\sigma(a) - q_\sigma(b) \geq \underline{\kappa}_{\sigma,K}(a - b) \geq m_{\sigma,K}(x - y).$$

The unprojected expected distance is therefore at most $(1 - \eta m_{\sigma,K}/2)(x - y)$. Interchanging the two states handles $x < y$. Scalar clipping is 1-Lipschitz, so it cannot increase any of these distances. Summing over coordinates proves equation 18.

For data sensitivity, fix a common scalar state x and any gradients a, b. The second part of Lemma 13 shows that the quantile difference has a fixed sign. Consequently,

$$\mathbb{E}|Q_{x,a}(U) - Q_{x,b}(U)| = |\mathbb{E}Q_{x,a}(U) - \mathbb{E}Q_{x,b}(U)| = \frac{\eta}{2}|q_\sigma(a) - q_\sigma(b)|.$$

Clipping, summing over coordinates, and applying equation 16–equation 17 at the common state $w \in K$ prove equation 19.

G.4 Proof of Theorem 4

Fix r and abbreviate $\mathcal{F}_t^{(r)}$ as $\mathcal{F}_t$. Set $\Delta_t^{(r)} = \|X_t - X_t^{(r)}\|_1$ and $d_t^{(r)} = \mathbb{E}\Delta_t^{(r)}$. Conditional on $\mathcal{F}_t$ and $I_t = i$, insert $\mathcal{P}_{Z_i}(X_t^{(r)}; U_t)$ between the two updates and use the triangle inequality. Lemma 14 gives

$$\mathbb{E}[\Delta_{t+1}^{(r)} \mid \mathcal{F}_t, I_t = i] \leq (1 - \eta m_{\sigma,K}/2)\Delta_t^{(r)} + \mathbf{1}_{\{i=r\}} \frac{\eta D_{\sigma,K}}{2}.$$

Here the same-sample contraction also applies when $i = r$; the additional term accounts for replacing Z_r by Z'_r at the common state $X_t^{(r)}$. Independence and uniformity of I_t yield

$$\mathbb{E}[\Delta_{t+1}^{(r)} \mid \mathcal{F}_t] \le (1 - \eta m_{\sigma,K}/2)\Delta_t^{(r)} + \frac{\eta D_{\sigma,K}}{2n}.$$

Taking expectations proves the stated recursion. Let $\alpha = \eta m_{\sigma,K}/2 \in (0, 1]$. The shared initialization gives $d_0^{(r)} = 0$, so, for every $T \ge 1$,

$$\begin{aligned} d_T^{(r)} &\le \frac{\eta D_{\sigma,K}}{2n} \sum_{s=0}^{T-1} (1-\alpha)^s \\ &= \frac{D_{\sigma,K}}{m_{\sigma,K} n} [1 - (1 - \eta m_{\sigma,K}/2)^T]. \end{aligned}$$

The zero-step bound is immediate. This argument also covers $n = 1$ and $\alpha = 1$.

For completeness, define $R_D(w) = \mathbb{E}_{Z\sim D} f(w, Z)$, $R_S(w) = n^{-1} \sum_{r=1}^n f(w, Z_r)$, and $\mathrm{gen}(A) = |\mathbb{E}\left[R_D(A(S)) - R_S(A(S))\right]|$. Swapping the independent identically distributed pair (Z_r, Z'_r), and using the correct conditional algorithm laws established above, gives the replace-one identity

$$\mathrm{gen}(A_T^{\mathrm{lazy}}) = \left| \frac{1}{n} \sum_{r=1}^n \mathbb{E}\big[f(X_T, Z'_r) - f(X_T^{(r)}, Z'_r)\big] \right|.$$

All terms are integrable by equation 10 and the anchor-point integrability assumption. Both outputs lie in K, so

$$\begin{aligned} \mathrm{gen}(A_T^{\mathrm{lazy}}) &\le \frac{L_K}{n} \sum_{r=1}^n d_T^{(r)} \\ &\le \frac{L_K D_{\sigma,K}}{m_{\sigma,K} n} \left[1 - (1 - \eta m_{\sigma,K}/2)^T\right] \\ &\le \frac{L_K D_{\sigma,K}}{m_{\sigma,K} n}. \end{aligned}$$

This proves the theorem.

The final bound is uniform in T for fixed K, η, σ, and the stated curvature and gradient constants. It concerns the absolute value of the expected signed generalization gap. It does not assert a high-probability bound or a bound on the expected absolute gap.

H Proof of Theorem 5

We first establish a uniform one-step bound for Gaussian signSGD. Let $\mathcal{V}$ be either $\mathbb{R}^d$ or $\mathcal{W}_R$, and suppose every sample loss is differentiable and M-Lipschitz with respect to $\|\cdot\|_2$ on $\mathcal{V}$. In the latter case, assume $\eta\sqrt{d} \le R$. Since 0_d is an interior point of $\mathcal{V}$,

$$\|\nabla f(0_d, z)\|_2 \le M \qquad \text{for every } z.$$

Both one-step outputs belong to $\mathcal{V}$, because $\|W_1\|_2 = \eta\sqrt{d}$ almost surely.

Fix a replace-one position r, and couple the runs on S and $S^{(r)}$ using the same uniform index I_0 and Gaussian noise ξ_0. Their outputs coincide when $I_0 \ne r$. Conditional on S and Z'_r, Theorem 6

therefore gives

$$\begin{aligned}\mathbb{E}\Big[\|W_1 - W_1^{(r)}\|_1 \mid S, Z_r'\Big] &= \frac{\eta}{n}\,\mathbb{E}_{\xi_0}[\|\mathrm{sign}(\nabla f(0_d, Z_r) + \xi_0) - \mathrm{sign}(\nabla f(0_d, Z_r') + \xi_0)\|_1] \\ &\le \frac{\eta c}{\sigma n}\|\nabla f(0_d, Z_r) - \nabla f(0_d, Z_r')\|_1 \\ &\le \frac{\eta c\sqrt{d}}{\sigma n}\left(\|\nabla f(0_d, Z_r)\|_2 + \|\nabla f(0_d, Z_r')\|_2\right) \\ &\le \frac{2cM\eta\sqrt{d}}{\sigma n},\end{aligned}$$

where $c = \sqrt{2/\pi}$.

Using the replace-one identity, integrability, and Lipschitz continuity on $\mathcal{V}$, we obtain

$$\begin{aligned}\mathrm{gen}(A_f^{\mathrm{noisy}}; D) &\le \frac{1}{n}\sum_{r=1}^{n}\mathbb{E}\left|f(W_1, Z_r') - f(W_1^{(r)}, Z_r')\right| \\ &\le \frac{M}{n}\sum_{r=1}^{n}\mathbb{E}\|W_1 - W_1^{(r)}\|_2 \\ &\le \frac{M}{n}\sum_{r=1}^{n}\mathbb{E}\|W_1 - W_1^{(r)}\|_1 \\ &\le \frac{2cM^2\eta\sqrt{d}}{\sigma n}.\end{aligned}$$

This bound is uniform over all problems satisfying the stated conditions.

For $\mathcal{C}_{\mathrm{nc}}$ and $\mathcal{C}_{\mathrm{cv}}$, take $\mathcal{V} = \mathbb{R}^d$ and $M = L$. Taking suprema yields

$$\mathcal{E}_\sigma(\mathcal{C}_{\mathrm{nc}}),\ \mathcal{E}_\sigma(\mathcal{C}_{\mathrm{cv}}) \le \frac{2cL^2\eta\sqrt{d}}{\sigma n}.$$

The construction in Section A.2 belongs to $\mathcal{C}_{\mathrm{nc}}$ and gives

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n}.$$

The construction in Section A.1 belongs to $\mathcal{C}_{\mathrm{cv}}$ and, when $\eta\sqrt{d} \le 1/2$, gives

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n}(1 - \eta\sqrt{d}) \ge \frac{L\eta\sqrt{d}}{8n}.$$

These particular problems lower-bound the corresponding suprema $\mathcal{E}_0$. Combining their bounds with the uniform Gaussian bound proves equation 2 when $\sigma > 8cL$ and equation 3 when $\sigma > 16cL$.

Finally, for $\mathcal{C}_{\mathrm{sc}}(L, R)$, take $\mathcal{V} = \mathcal{W}_R$ and $M = L + R$. The condition $\eta\sqrt{d} \le R$ ensures that all one-step outputs remain in this ball, so the same argument gives

$$\mathcal{E}_\sigma(\mathcal{C}_{\mathrm{sc}}(L, R)) \le \frac{2c(L+R)^2\eta\sqrt{d}}{\sigma n}.$$

The construction in Section A.3 belongs to $\mathcal{C}_{\mathrm{sc}}(L, R)$ and, under $\eta\sqrt{d} \le \min\{1/2, R\}$, satisfies

$$\mathrm{gen}(A_f^{\mathrm{signSGD}}; D) = \frac{L\eta\sqrt{d}}{4n}(1 - \eta\sqrt{d}) \ge \frac{L\eta\sqrt{d}}{8n}.$$

Thus $\sigma > 16c(L+R)^2/L$ gives the strict inequality in equation 4, completing the proof.

I Bounded non-convex losses and stopping times

Following the first-hit stopping-time strategy used for non-convex SGD , we separate the event that the replaced example is visited early from the subsequent stability evolution. We now additionally assume

$$0 \le f(w, z) \le M, \qquad \forall w \in \mathbb{R}^d,\ z \in \mathcal{Z}. \tag{20}$$

This assumption is used only in this section. Fix a replace-one position r and define

$$\tau_r := \inf\{t \geq 0 : I_t = r\}. \tag{21}$$

Since the indices are i.i.d. uniform,

$$\mathbb{P}(\tau_r < t) = 1 - \left(1 - \frac{1}{n}\right)^t \leq \frac{t}{n}. \tag{22}$$

Theorem 12 (Stopping-time bound) *Suppose Assumptions 1–2 hold and the loss satisfies equation 20. Then, for every integer $t_0 \in \{0, \dots, T\}$,*

$$\operatorname{gen}(A_T) \leq \frac{Mt_0}{n} + \frac{L_1 D_\sigma}{n} \sum_{t=t_0}^{T-1} \eta_t \prod_{j=t+1}^{T-1} (1 + a_\sigma \eta_j). \tag{23}$$

If $a_\sigma > 0$, then

$$\operatorname{gen}(A_T) \leq \frac{Mt_0}{n} + \frac{L_1 D_\sigma}{a_\sigma n} \left[\prod_{t=t_0}^{T-1} (1 + a_\sigma \eta_t) - 1\right]. \tag{24}$$

I.1 Proof of Theorem 12

Use the representation from Theorem 8 at the fresh point Z'_r:

$$\operatorname{gen}(A_T) \leq \frac{1}{n} \sum_{r=1}^{n} \mathbb{E}|f(W_T, Z'_r) - f(W_T^{(r)}, Z'_r)|.$$

Set $C_r = \{\tau_r \geq t_0\}$ and $p_{t_0} = \mathbb{P}(C_r) = (1 - 1/n)^{t_0}$, with $p_0 = 1$. On C_r^{c} the absolute loss difference is at most M. On C_r, the two iterates coincide at time t_0.

To avoid conditioning on a possibly null event (in particular when $n = 1$), put $e_t = \mathbb{E}[\Delta_t^{(r)} \mathbf{1}_{C_r}]$ for $t \geq t_0$. The event C_r belongs to $\mathbb{F}_t$ and $e_{t_0} = 0$. Multiplying the conditional inequality of Theorem 7 by $\mathbf{1}_{C_r}$ and taking expectations gives

$$e_{t+1} \leq (1 + a_\sigma \eta_t) e_t + p_{t_0} \frac{\eta_t D_\sigma}{n}.$$

Unrolling, using Lipschitz continuity on C_r, and averaging over r proves the sharper inequality

$$\operatorname{gen}(A_T) \leq M(1 - p_{t_0}) + p_{t_0} \frac{L_1 D_\sigma}{n} \sum_{t=t_0}^{T-1} \eta_t \prod_{j=t+1}^{T-1} (1 + a_\sigma \eta_j). \tag{25}$$

This remains valid if $p_{t_0} = 0$. The inequalities $1 - p_{t_0} \leq t_0/n$ and $p_{t_0} \leq 1$ give equation 23. The telescoping identity in Theorem 2 gives equation 24. One may also take the minimum with M, since $\operatorname{gen}(A_T) \leq M$.

J Details of Experiments

J.1 Synthetic Experiments

J.1.1 Experimental Setup

We study the original clean and Gaussian signSGD updates, without projection or lazy coordinate updates:

$$W_{t+1} = W_t - \eta_t \operatorname{sign}\big(\nabla f(W_t, Z_{I_t}) + \xi_t\big), \qquad \xi_t \sim \mathcal{N}(0, \sigma^2 I_d),$$

where I_t is sampled uniformly with replacement. For clean signSGD, $\xi_t = 0$ and $\operatorname{sign}(0) = 0$. Unless otherwise stated, $W_0 = 0$, $d = 20$, $n = 512$, and $T = 4096$. The coordinates of each data vector are independent. We consider $p \in \{0.1, 0.25, 0.5, 0.75\}$ and $\sigma = \rho/d$, with $\rho \in \{0, 0.1, 0.3, 1, 3\}$.

Each evaluated configuration uses 100 independent training datasets and two algorithm seeds per dataset.

For a training dataset S, define

$$R_S(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, Z_i), \qquad R_D(w) = \mathbb{E}_{Z\sim D} f(w, Z).$$

Both risks are recorded at every iteration. Population risks are evaluated analytically, so their evaluation introduces no additional test-sample error. With $B = 100$ datasets and $K = 2$ seeds, we estimate

$$\mathrm{gen}_t = \left| \frac{1}{BK}\sum_{b=1}^{B}\sum_{k=1}^{K} \left[R_D(W_t^{(b,k)}) - R_{S_b}(W_t^{(b,k)}) \right] \right|.$$

J.1.2 Objectives and step-size schedules

S1: Separable convex loss. Let $v_j(z) \in \{-1, 1\}$ with $\Pr(v_j(z) = 1) = p$, and define

$$f_{\mathrm{S1}}(w, z) = \frac{1}{d}\sum_{j=1}^{d} \log\cosh(w_j - v_j(z)).$$

Its population risk is

$$R_{D,\mathrm{S1}}(w) = \frac{1}{d}\sum_{j=1}^{d} \left[p \log\cosh(w_j - 1) + (1-p)\log\cosh(w_j + 1) \right].$$

This loss is coordinate-separable, convex, globally Lipschitz, and globally smooth. Under the ℓ_1 assumptions, one may take $L_1 = \beta_1 = 1/d$. We use the dyadic step-size schedule

$$\eta_t = c\, 2^{-\lfloor \log_2(t+1) \rfloor}, \qquad t = 0, \ldots, T-1,$$

with $c = 0.125$ in the fixed-coefficient noise sweep. Successive step-size ratios η_t/η_{t+1} belong to $\{1, 2\}$, so the nested-step condition holds. Thus, the objective and schedule satisfy the structural conditions of Theorem 3.

S2: Strongly convex quadratic loss. Let $v_j(z) \in \{-1, 1\}$ with $\Pr(v_j(z) = 1) = p$, and define

$$f_{\mathrm{S2}}(w, z) = \frac{1}{2d}\sum_{j=1}^{d} (w_j - v_j(z))^2.$$

Its population risk is

$$\begin{aligned} R_{D,\mathrm{S2}}(w) &= \frac{1}{2d}\sum_{j=1}^{d} \left[p(w_j - 1)^2 + (1-p)(w_j + 1)^2 \right] \\ &= \frac{1}{2d}\sum_{j=1}^{d} \left[(w_j - (2p-1))^2 + 4p(1-p) \right]. \end{aligned}$$

Consequently, its population minimizer and minimum risk are

$$w^\star = (2p-1)\mathbf{1}_d, \qquad R^\star_{D,\mathrm{S2}} = 2p(1-p),$$

and the population excess risk is

$$R_{D,\mathrm{S2}}(w) - R^\star_{D,\mathrm{S2}} = \frac{\|w - (2p-1)\mathbf{1}_d\|_2^2}{2d}.$$

S2 uses the same dyadic schedule as S1:

$$\eta_t = c\, 2^{-\lfloor \log_2(t+1) \rfloor}, \qquad t = 0, \ldots, T-1,$$

with $c = 0.125$ in the fixed-coefficient noise sweep.

This objective allows parameter error to be separated from the minimum population risk. For a scalar coordinate $u \in (-1, 1)$, the expected clean sign direction is

$$\mathbb{E}_v \left[\operatorname{sign}\left(\frac{u - v}{d} \right) \right] = 1 - 2p.$$

For $p \in (0, 1) \setminus \{1/2\}$, this direction does not vanish at the population minimizer $u = 2p - 1$. For $p = 1/2$, it vanishes throughout $(-1, 1)$. Thus, the expected clean sign direction does not generally identify the quadratic risk minimizer.

For $\sigma > 0$, Gaussian perturbations change the expected sign direction to

$$p\, q_\sigma\left(\frac{u-1}{d} \right) + (1-p)\, q_\sigma\left(\frac{u+1}{d} \right), \qquad q_\sigma(a) = 2\Phi(a/\sigma) - 1,$$

where Φ is the standard normal cumulative distribution function. This response motivates examining whether noise improves optimization, without guaranteeing improvement at every noise level or training horizon.

S2 uses the original unprojected algorithm. The quadratic objective is globally smooth and $1/d$-strongly convex with respect to ℓ_2, but is not globally Lipschitz. Consequently, satisfying the nested-step condition alone does not make S2 a direct validation of the globally Lipschitz theory. It also does not validate the lazy projected algorithm of Theorem 4.

S3: Bounded non-convex loss. Let $v_j(z) \in \{-0.8, 0.8\}$ with $\Pr(v_j(z) = 0.8) = p$, and define

$$f_{\mathrm{S3}}(w, z) = \frac{1}{d} \sum_{j=1}^{d} \left[1 - \cos(w_j - v_j(z)) \right].$$

Its population risk is

$$R_{D,\mathrm{S3}}(w) = \frac{1}{d} \sum_{j=1}^{d} \left[1 - p\cos(w_j - 0.8) - (1-p)\cos(w_j + 0.8) \right].$$

The loss is globally smooth and non-convex, satisfies $0 \le f_{\mathrm{S3}} \le 2$, and admits $L_1 = \beta_1 = 1/d$ under the ℓ_1 assumptions. We use the harmonic step-size schedule

$$\eta_t = \frac{C}{t+1}, \qquad t = 0, \ldots, T-1,$$

with $C = 0.5$ in the fixed-coefficient noise sweep.

To compute population excess risk, let

$$a = \cos(0.8), \qquad b = (2p-1)\sin(0.8).$$

Then

$$R_{D,\mathrm{S3}}(w) = 1 - \frac{1}{d} \sum_{j=1}^{d} \left[a\cos(w_j) + b\sin(w_j) \right],$$

and its minimum value is

$$R^\star_{D,\mathrm{S3}} = 1 - \sqrt{a^2 + b^2}.$$

The population excess risk is therefore $R_{D,\mathrm{S3}}(w) - R^\star_{D,\mathrm{S3}}$.

S3 permits comparison with both the general non-convex bound and the bounded-loss stopping-time bound. We record replace-one trajectories, sign disagreement, and the first time the replaced example is sampled. Additional diagnostics vary the sample size at $p = 0.5$ and $p = 0.75$, and initialize all coordinates at 2.5 to examine a region of negative curvature. The harmonic schedule does not generally satisfy the nested-step condition required for the sharper convex bound.

J.1.3 Results of the fixed-schedule noise sweeps

The results in this subsection use the fixed step-size coefficients specified above and vary $\rho \in \{0, 0.1, 0.3, 1, 3\}$, with $\sigma = \rho/d$. They should be distinguished from the independently evaluated results obtained after separate hyperparameter selection. For each setting, we report the final empirical risk, exact population risk, and absolute mean generalization gap gen.

S1: Separable convex loss. Moderate Gaussian noise reduces the final population risk under the fixed dyadic schedule, although the best noise scale depends on p. Among the evaluated values of ρ, the lowest population risk occurs at $\rho = 0.3$ for $p \in \{0.1, 0.25, 0.75\}$ and at $\rho = 1$ for $p = 0.5$. For $p = 0.75$, changing from clean signSGD to $\rho = 0.3$ reduces population risk from 0.304227 to 0.297794 and gen from 1.914×10^{-3} to 1.499×10^{-3}. Increasing the noise to $\rho = 3$ further reduces the gap to 0.361×10^{-3}, but raises population risk to 0.391440. The accompanying increase in empirical risk shows why a smaller gap need not yield a better population-risk outcome.

The gap and population risk can also move in opposite directions at moderate noise levels. For $p = 0.1$, $\rho = 0.3$ reduces population risk from 0.132303 to 0.128381, while gen increases from 0.104×10^{-3} to 0.272×10^{-3}. Thus, even for the same convex objective, an improvement in population risk cannot always be attributed to a smaller generalization gap.

S2: Strongly convex quadratic loss. S2 uses the original signSGD updates with the same dyadic schedule as S1. Moderate noise again lowers the final population risk: the best value in the sweep is $\rho = 0.3$ for $p = 0.1$ and $\rho = 1$ for the other three values of p. For $p = 0.75$, $\rho = 1$ reduces population risk from 0.418978 to 0.389471 and gen from 2.022×10^{-3} to 0.952×10^{-3}. At $\rho = 3$, the gap decreases further to 0.453×10^{-3}, whereas population risk rises to 0.439655, exceeding the clean result.

For $p = 0.1$, $\rho = 0.3$ lowers population risk from 0.199349 to 0.183336, while gen increases from 0.681×10^{-3} to 0.846×10^{-3}. The quadratic objective therefore provides another example in which changes in optimization and generalization must be assessed separately. These observations concern the original algorithm and do not directly test the bound for the lazy projected algorithm in Theorem 4.

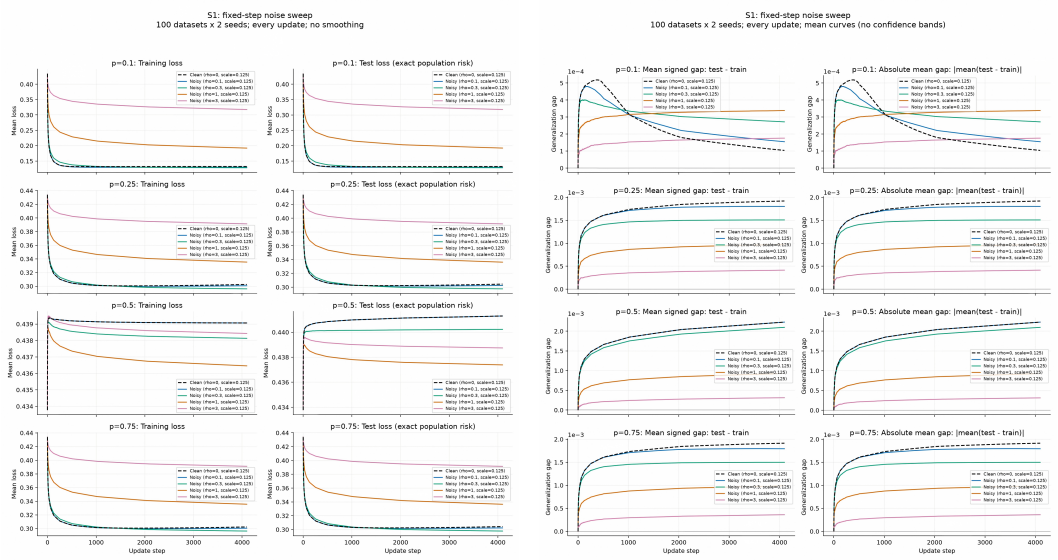
S3: Bounded non-convex loss. For S3, $\rho = 1$ gives the lowest final population risk among the evaluated noise scales for all four values of p. The corresponding gap changes depend on the data distribution. At $p = 0.5$, $\rho = 1$ reduces population risk from 0.397208 to 0.310339 and gen from 4.585×10^{-3} to 1.472×10^{-3}; empirical risk also decreases. At $p = 0.75$, population risk decreases from 0.257185 to 0.220276, although genincreases from 0.162×10^{-3} to 0.885×10^{-3}. Empirical risk falls by more than the gap increases, yielding a lower population risk. A similar combination occurs at $p = 0.25$, where genrises from 0.053×10^{-3} to 0.638×10^{-3} while population risk falls from 0.257010 to 0.220286.

Some *signed* mean gaps in S3 are negative, including the clean results for $p = 0.1$ and $p = 0.25$. Accordingly, plots of the signed gap retain a zero reference line. The quantity in our theoretical definition and in the comparisons above is the absolute value of the mean signed gap; a downward movement of a signed-gap curve alone does not establish an improvement in this quantity.

Interpretation. Across the three objectives, the fixed-schedule sweeps show that the effect of Gaussian noise depends on its scale and on the data distribution. Moderate noise can lower population risk, whereas larger noise can impair empirical fit and population risk even when the absolute mean gap shrinks. The noise scale minimizing population risk therefore need not minimize the generalization gap. The fixed-schedule results can also differ from the separately tuned comparison in the main text; in particular, an improvement at $p = 0.5$ under a fixed schedule does not imply that Gaussian signSGD wins after each method selects its own configuration.

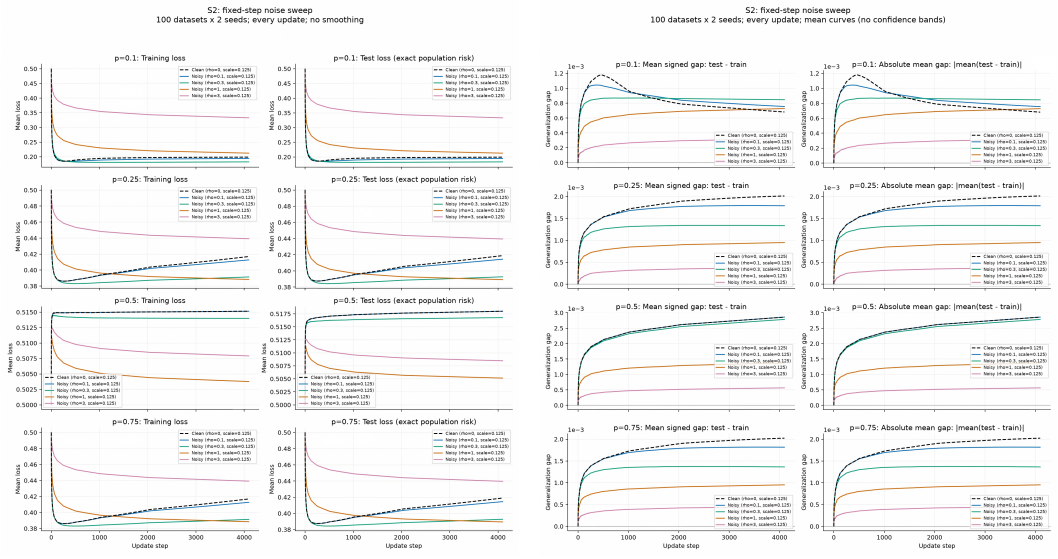
J.2 Full-Parameter Fine-Tuning Experiments

The full-parameter fine-tuning experiments use Qwen3-8B as the base model, with approximately 8.19 billion trainable parameters, all of which are updated during training. We use NuminaMath-CoT for supervised fine-tuning and partition the selected data into 10,000 training examples, 1,000 validation examples, and 1,000 test examples according to the original problem groups. This grouping prevents different answers or reformulations of the same problem from appearing in multiple splits. During training, the next-token cross-entropy loss is computed only on response tokens; prompt and padding tokens are excluded. The maximum sequence length is 2,048 tokens. All external benchmarks are used exclusively for evaluation and are not included in training. We evaluate the fine-tuned models on MathQA (Amini et al., 2019) and BBH Object Counting (Suzgun et al., 2023).



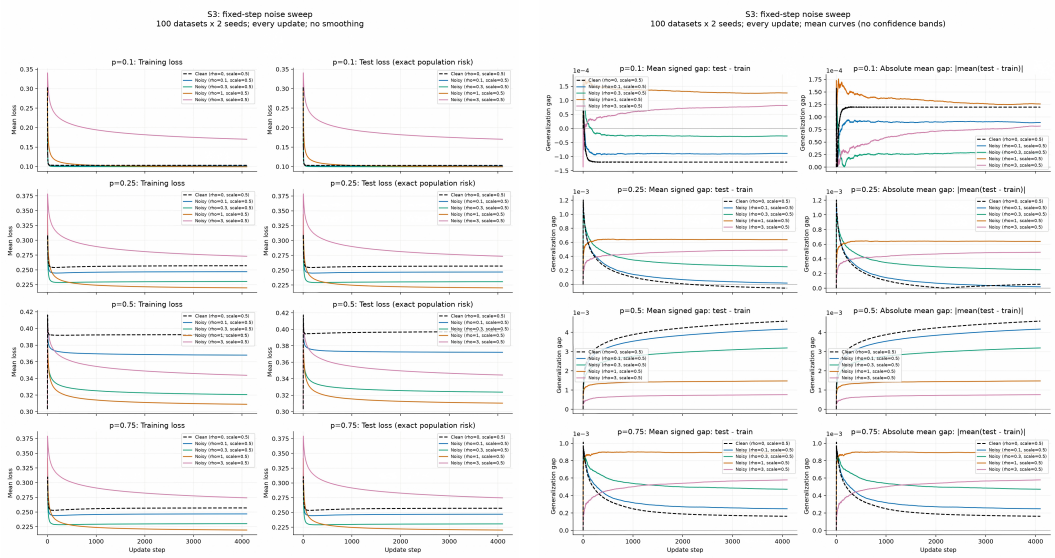
(a) Training and test loss (b) Generalization gap

Figure 2: Fixed-schedule noise scan for the separable convex log-cosh loss (S1). The dyadic schedule uses $c = 0.125$.



(a) Training and test loss (b) Generalization gap

Figure 3: Fixed-schedule noise scan for the strongly convex quadratic loss (S2). The dyadic schedule uses $c = 0.125$.



(a) Training and test loss

(b) Generalization gap

Figure 4: Fixed-schedule noise scan for the bounded non-convex cosine loss (S3). The harmonic schedule uses $C = 0.5$.

For MathQA, we report both the standard accuracy (Acc) and normalized accuracy (Acc-norm), whereas for BBH Object Counting we report accuracy.

All experiments are conducted using PyTorch Distributed Data Parallel on seven NVIDIA A800 80GB GPUs with BF16 model parameters and computation. The micro-batch size on each GPU is 1, with two gradient-accumulation steps, giving an effective global batch size of 14. The learning rate of Gaussian SignSGD is 6×10^{-6}. For the noise-scale comparison in Table 3, we evaluate $\rho \in \{0.15, 0.20, 0.25, 0.30\}$. Noise addition and sign computation are performed in FP32, after which the BF16 parameters are updated in the sign direction. To prevent small signed updates from being lost through BF16 rounding, we additionally store the rounding residuals that are not immediately reflected in the parameters and accumulate them in subsequent updates.

J.3 Computer Vision Experiments

Implementation details. For the CIFAR-100 experiments, all models are trained for 200 epochs with a training batch size of 128 and an evaluation batch size of 256. Each epoch consists of 352 optimization steps, with the final incomplete batch retained. We use the cross-entropy loss and set the initial learning rate to 10^{-3}. Following the MultiStepLR schedule (Paszke et al., 2019), the learning rate is reduced by a factor of 0.1 after epochs 50 and 125, resulting in learning rates of 10^{-3}, 10^{-4}, and 10^{-5} for epochs 1–50, 51–125, and 126–200, respectively. An ℓ_2 penalty with coefficient 10^{-4} is applied to the weights of all convolutional layers and the final linear layer, while batch-normalization parameters and bias terms are excluded.