STOCHASTIC SUBGRADIENT DESCENT AT SHARPLY REPULSIVE POINTS: A BOUNDED NOISE COUNTEREXAMPLE AND GAUSSIAN NOISE AVOIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the open question of Bianchi, Hachem, and Schechtman (Bianchi et al., 2024, Remark 4), we show that in the non-weakly convex definable setting, omnidirectional noise with conditional fourth-moment control does not suffice for universal almost-sure avoidance of sharply repulsive critical points. We demonstrate this failure by constructing a globally Lipschitz, coercive, definable objective that is not weakly convex. For this objective and its sufficiently small linear perturbations, stochastic subgradient descent with independent noise uniformly distributed on a ball in $\mathbb{R}^2$ converges *with probability* 1 to a nonminimal sharply repulsive critical point from every initial point in a specified ball. A two-cycle argument establishes uniform bounds on the rescaled iterates, yielding convergence to the sharply repulsive critical point. In contrast, for locally Lipschitz definable objectives, Gaussian noise ensures simultaneous almost-sure avoidance of any finite collection of sharply repulsive points. Through a construction of difference quotient functions, we show that convergence to any such point with positive probability would produce a stationary distribution with expected descent 0, whereas a global positive lower bound on the Fréchet subgradient norms of the limiting functions forces strictly positive expected descent under the same distribution. This contradiction proves avoidance, with the everywhere positive Gaussian density playing a key role.

1 INTRODUCTION

Convergence to a critical point does not ensure local optimality in nonconvex optimization. This distinction matters for stochastic methods used to train neural networks, whose parameterized losses are generally nonconvex (LeCun et al., 2015; Bottou et al., 2018). A central question is whether their iterates avoid nonminimal critical points, and which assumptions on the objective and the noise suffice for such avoidance.

For smooth objectives, saddle points have received considerable attention in the study of neural network optimization (Dauphin et al., 2014). There are now several ways to obtain avoidance or escape guarantees under suitable assumptions. Stable-manifold arguments show that gradient descent with random initialization almost surely avoids strict saddles when the update map has the required regularity (Lee et al., 2016). Stochastic perturbations provide another mechanism: under strict-saddle assumptions, noisy gradient methods can escape saddles and reach approximate local minima (Ge et al., 2015); carefully timed perturbations also yield efficient convergence to approximate second-order stationary points (Jin et al., 2017). These results connect local geometry with algorithmic behavior. They do not, however, make every critical point benign, and their smooth geometric arguments do not directly cover the nonsmooth operations used in many learning models.

ReLU activations, max pooling, and nonsmooth regularizers lead naturally to objectives that are both nonconvex and nonsmooth. At such points, gradients are replaced by subdifferentials, and even the meaning of criticality becomes important. Definable geometry offers a broad setting in which these objectives can be studied without allowing arbitrary oscillatory behavior. It includes

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.

semialgebraic models and, in suitable o-minimal structures, many other functions used in learning. The stochastic subgradient convergence theory of Davis et al. (2018) exploits this structure to obtain guarantees of criticality. The remaining question is whether the critical points permitted by such guarantees are local minima, or whether the stochastic iteration can converge to points with lower objective values arbitrarily nearby.

The nonsmooth setting introduces an obstruction beyond the familiar smooth saddle. The Clarke subdifferential convexifies limiting slopes, so it can contain zero even when the limiting subdifferential excludes zero. A point can consequently be Clarke critical without possessing a limiting subgradient equal to zero. Sharply repulsive critical points combine this discrepancy with an active manifold on which the point is a local minimizer. For a nonminimal point of this kind, favorable behavior along the manifold does not imply local optimality in the ambient space. Nor does the word repulsive, which describes subgradient geometry, itself prove that a discrete stochastic algorithm avoids the point. The algorithm moves in finite steps, and the interaction between those steps and the noise must be analyzed explicitly.

There are positive results under additional structure. Generic convergence to local minima is known for definable weakly convex objectives (Bianchi et al., 2024) and, more broadly, under subdifferential regularity (Davis et al., 2026). Bianchi et al. (2024, Remark 4) ask whether this conclusion extends beyond weak convexity to locally Lipschitz definable functions; the required avoidance concerns both active strict saddles and sharply repulsive critical points.[1] Related work identifies pathological Lipschitz and path-differentiable subgradient dynamics (Daniilidis & Drusvyatskiy, 2019; Rios-Zertuche, 2020), studies constant-step instability (Josz & Lai, 2022) and deterministic convergence rates (Chzhen & Schechtman, 2026), and develops pathwise saddle avoidance under locally invertible sampled updates (Qiu et al., 2026). These results motivate a closer examination of the nonregular case, where Clarke criticality and limiting criticality can differ.

We study the sharply repulsive part of this question by comparing bounded and Gaussian noise, focusing on what the noise distribution contributes beyond centering, moment bounds, and excitation in every direction. The results concern nonconvergence to specified targets.

1.1 Contributions

A counterexample under bounded noise. We resolve the open question of Bianchi et al. (2024, Remark 4) in the negative: omnidirectional noise with conditional fourth-moment control does not ensure avoidance of nonminimal sharply repulsive points. Our counterexample is an explicit globally Lipschitz, coercive, semialgebraic objective on $\mathbb{R}^2$: with independent noise uniform on a ball, the iterates converge to such a point with probability 1, for every initial point in a specified ball and all sufficiently small linear perturbations of the objective (Theorem 2.1 and Corollary 2.3). A two-cycle argument establishes uniform boundedness of the rescaled iterates under bounded noise, yielding $\|x_n\| = O(\gamma_n)$.

Gaussian noise avoidance through difference quotient functions. For locally Lipschitz definable objectives, Gaussian noise prevents convergence to any prescribed point with $0 \notin \partial_L f(x_*)$. This holds for every adapted Clarke selection and yields avoidance of finite target sets and, for almost every linear perturbation, of all sharply repulsive points (Corollary 3.7). Our difference quotient construction transfers a local lower bound on Fréchet subgradient norms to a global bound for limiting functions and converts convergence with positive probability into a stationary distribution with mean descent 0. The inherited bound and Gaussian escape force strictly positive mean descent under the same distribution, giving a contradiction.

1.2 Setting and subdifferentials

For a locally Lipschitz definable $f : \mathbb{R}^d \to \mathbb{R}$, consider

$$X_{n+1} = X_n - \gamma_n V_n + \gamma_n \eta_{n+1}, \qquad V_n \in \partial_C f(X_n). \tag{1}$$

The iterates and noises are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Appendix A.3 specifies the spaces and measures for all expectations below. A function is definable if its graph belongs to a fixed

[1] The question is problem 75 in the Open Problems in Operations Research collection: `https://ai-in-math-or.github.io/open-problems-in-or/`. The repository's regular-case progress note treats the subdifferentially regular subclass.

o-minimal expansion of the ordered real field, with real parameters allowed (van den Dries & Miller, 1996, Section 2). Such a structure is closed under Boolean operations, products, and projections, and its subsets of $\mathbb{R}$ are finite unions of points and intervals. Semialgebraic functions, including the construction below, are definable; Appendix A.2 gives the complete definition.

Definition 1.1 (Subdifferentials; Rockafellar & Wets, 1998, Definition 8.3 and Theorem 9.61)**.** For locally Lipschitz f, the Fréchet subdifferential consists of the vectors v satisfying $f(x+h) \geq f(x)+\langle v, h\rangle + o(\|h\|)$. The limiting subdifferential $\partial_L f(x)$ consists of limits of Fréchet subgradients at points tending to x. The Clarke subdifferential is $\partial_C f(x) = \operatorname{conv} \partial_L f(x)$.

Thus $\partial_F f(x) \subseteq \partial_L f(x) \subseteq \partial_C f(x)$, but $0 \in \partial_C f(x)$ need not imply $0 \in \partial_L f(x)$.

Definition 1.2 (Sharply repulsive critical point; Schechtman, 2021, Definition 6.2.4)**.** An active manifold M at a Clarke critical point x is a smooth manifold through x such that $f|_M$ is smooth and, near x off M, all Clarke subgradients have norm bounded below by a common positive constant. The point is sharply repulsive if it admits such an M, is a local minimizer of $f|_M$, and satisfies $0 \in \partial_C f(x) \setminus \partial_L f(x)$.

2 A BOUNDED-NOISE COUNTEREXAMPLE

We answer the open question of Bianchi et al. (2024, Remark 4) in the negative: bounded noise can lead to almost-sure convergence to a nonminimal sharply repulsive point. As summarized in Figure 1, the proof combines subdifferential calculations at 0 with uniform bounds on the rescaled iterates from the two-cycle argument; the vanishing step size then forces convergence. A coercive modification preserves both the local geometry and these trajectories.

2.1 THE COUNTEREXAMPLE AND ITS GEOMETRY

We define the following function based on the Euclidean norm and some piecewise linear functions:

$$f(x, y) = \sqrt{x^2 + y^2} - \min\{x, y\} - 2\max\{0, -x - y\}. \tag{2}$$

It is easy to check that 0 is Clarke critical without being limiting critical. For $u \in \mathbb{R}^2$, set

$$f_u(z) = f(z) - \langle u, z\rangle, \qquad z = (x, y). \tag{3}$$

The function f_u is a linear tilt of f, obtained by subtracting the linear function $z \mapsto \langle u, z\rangle$ (Rockafellar, 1970, Section 12, p. 104). We apply (1) to f_u, write its iterates and subgradients as x_n and v_n, and set $a = (1, 0)$ and $b = (0, 1)$.

Theorem 2.1 (A bounded-noise counterexample)**.** *Fix $u \in \mathbb{R}^2$ with $\|u\| < 1/200$. The function f_u in* (3) *is globally Lipschitz and definable, but is not weakly convex. In* (1) *applied to f_u, use $\gamma_n = 1/(n + 100)$ and independent noise $\eta_n \sim \mathrm{Unif}(B(0, 1/200))$. Let x_0 be independent of the noise sequence and satisfy $x_0 \in \gamma_0 \overline{B}(a, 1/8)$ almost surely. Then every Clarke subgradient selection adapted to $\mathcal{F}_n = \sigma(x_0, \eta_1, \ldots, \eta_n)$ gives*

$$\mathbb{P}\Big(\lim_{n\to\infty} x_n = 0\Big) = 1,$$

where 0 is a sharply repulsive Clarke critical point with active manifold $M = \{0\}$ and is not a local minimizer.

Theorem 2.1 resolves the open question of Bianchi et al. (2024, Remark 4) **in the negative**. The proof of Theorem 2.1 is completed in Appendix B.3.2, using the geometry and two-cycle argument in Appendices B.1–B.3.1. The probability is taken for each fixed admissible tilt. The initial region contains the nonempty open set $\gamma_0 B(a, 1/8)$, and the tilt region is open. Thus the failure is not confined to a null set of tilts or a single initial point.

Lemma B.1 establishes the required function class. Lemma B.2 bounds all Clarke subgradients away from zero off $M = \{0\}$, making M an active manifold. Since $f_u|_M$ is constant, 0 minimizes this restriction. Lemmas B.3 and B.4 establish Clarke criticality, exclude limiting criticality, and give a descending ray:

$$\partial_F f_u(0) = \varnothing, \qquad 0 \in \partial_C f_u(0) \setminus \partial_L f_u(0), \qquad f_u(-t, -t) < f_u(0) \quad (t > 0).$$

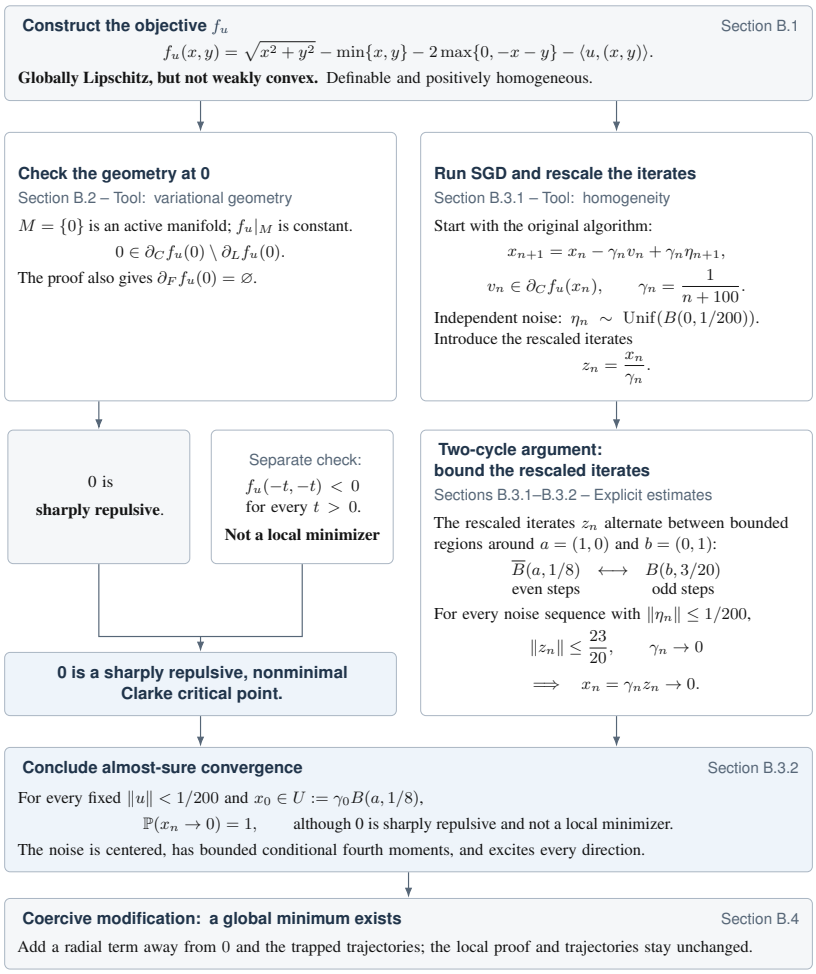

Figure 1: Proof of the bounded-noise counterexample. Left: 0 is sharply repulsive but not a local minimizer. Right: a two-cycle argument keeps the rescaled iterates bounded, so the original iterates converge to 0 almost surely. A coercive modification preserves these trajectories.

These conclusions identify 0 as a nonminimal sharply repulsive critical point. The remaining task is to prove convergence to 0.

2.2 Convergence via rescaled iterates

Set $z_n = x_n/\gamma_n$. We use a two-cycle argument to keep these rescaled iterates alternately in two bounded regions for every admissible noise sequence. Returning to the original scale gives $\|x_n\| = O(\gamma_n) \to 0$.

Lemma 2.2 (Two-cycle argument for the rescaled iterates)**.** *Fix $\|u\| < 1/200$, $x_0 \in \gamma_0\overline{B}(a,1/8)$, and any noise sequence with $\|\eta_n\| \le 1/200$ for all $n \ge 1$. For $\gamma_n = 1/(n+100)$, the rescaled iterates satisfy*

$$z_{2j} \in \overline{B}(a,1/8), \qquad z_{2j+1} \in B(b,3/20), \qquad j \ge 0. \tag{4}$$

In particular, $\|z_n\| \le 23/20$, and every physical iterate lies in a smooth cone, so the Clarke subgradient along the trajectory is unique.

The proof of the two-cycle argument Lemma 2.2 is in Appendix B.3.1. Together these results give $\|x_n\| \le 23/[20(n+100)] \to 0$.

Lemma 2.2 thus gives convergence to the nonminimal point 0 for f_u. However, f_u is not coercive. To obtain a coercive counterexample with the same trajectories, we add a radial penalty that vanishes near 0 and all trapped iterates. Specifically, set

$$\widehat{f}_u(z) = f_u(z) + 10\max\{0, \|z\| - 1/10\}.$$

Corollary 2.3 (The counterexample can be coercive)**.** *The function $\widehat{f}_u$ is globally Lipschitz, definable, coercive, and attains a global minimum. It is not weakly convex. Under the step-size and noise assumptions of Theorem 2.1, every initial point in $\gamma_0\overline{B}(a, 1/8)$ produces the same trapped trajectory for $\widehat{f}_u$ as for f_u. The point 0 remains sharply repulsive and nonminimizing.*

The complete proof is in Appendix B.4. The radial term vanishes on a neighborhood of every trapped iterate, so the local geometry and trajectories coincide. The full arguments also record the extensions to power steps in $(1/2, 1]$.

3 Gaussian noise avoidance

In this section, we show that simply requiring the noise to have an everywhere positive density, e.g., Gaussian noise, ensures avoidance of the sharply repulsive points considered in the open question of Bianchi et al. (2024, Remark 4). For locally Lipschitz definable objectives, the following theorem rules out convergence to any prescribed x_* with $0 \notin \partial_L f(x_*)$ for every adapted Clarke subgradient selection.

Theorem 3.1 (Gaussian avoidance of a specified point)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz and definable, and fix $x_* \in \mathbb{R}^d$ with $0 \notin \partial_L f(x_*)$. In* (1)*, let X_0 be finite almost surely and let X_n, V_n be adapted to a filtration $\mathcal{F}_n$. Suppose η_{n+1} is independent of $\mathcal{F}_n$ and the noise vectors are identically distributed as $N(0, \Sigma)$ for a fixed positive-definite matrix Σ. If*

$$\gamma_n = \kappa(n + n_0)^{-\alpha}, \qquad \kappa, n_0 > 0, \qquad \tfrac{1}{2} < \alpha \leq 1, \tag{5}$$

then $\mathbb{P}(X_n \to x_*) = 0$ for every adapted choice $V_n \in \partial_C f(X_n)$.

Under hypothetical convergence, summing decreases of a Lipschitz extension h over late stopped windows yields vanishing mean descent for the distributions of actual difference quotient states. Compactness and continuity give a limit with mean descent 0, and a separate state-test identity proves its stationarity. The inherited subgradient bound and Gaussian escape force positive mean descent under this same distribution, a contradiction (Figure 2).

3.1 Difference quotient functions

We analyze objective decreases at the scale of each update. Translate the target to 0 and subtract $f(x_*)$, so $f(0) = 0$. Since Gaussian updates can leave the local neighborhood, first extend f globally: for sufficiently small $R > 0$ and a Lipschitz constant L on $\overline{B}(0, R)$, set

$$h(x) = \inf_{y \in \overline{B}(0,R)} \{f(y) + L\|x - y\|\}. \tag{6}$$

Then $h = f$ on $\overline{B}(0, R)$ and h is globally L-Lipschitz (Lemma A.4). At center x and scale $s > 0$, define

$$H_{x,s}(u) = \frac{h(x + su) - h(x)}{s}, \qquad u \in \mathbb{R}^d. \tag{7}$$

Here u measures displacement in units of s. An update $x^+ = x + s(-v + \eta)$ therefore has scaled decrease

$$\frac{h(x) - h(x^+)}{s} = H_{x,s}(0) - H_{x,s}(-v + \eta).$$

The uniform Lipschitz bound and $H_{x,s}(0) = 0$ give locally uniform subsequential limits. Let $\mathcal{C}$ consist of all limits obtained as $x_j \to 0$ and $s_j \downarrow 0$:

$$\sup_{\|u\| \leq M} |H_{x_j,s_j}(u) - H(u)| \longrightarrow 0 \qquad \text{for every } M > 0. \tag{8}$$

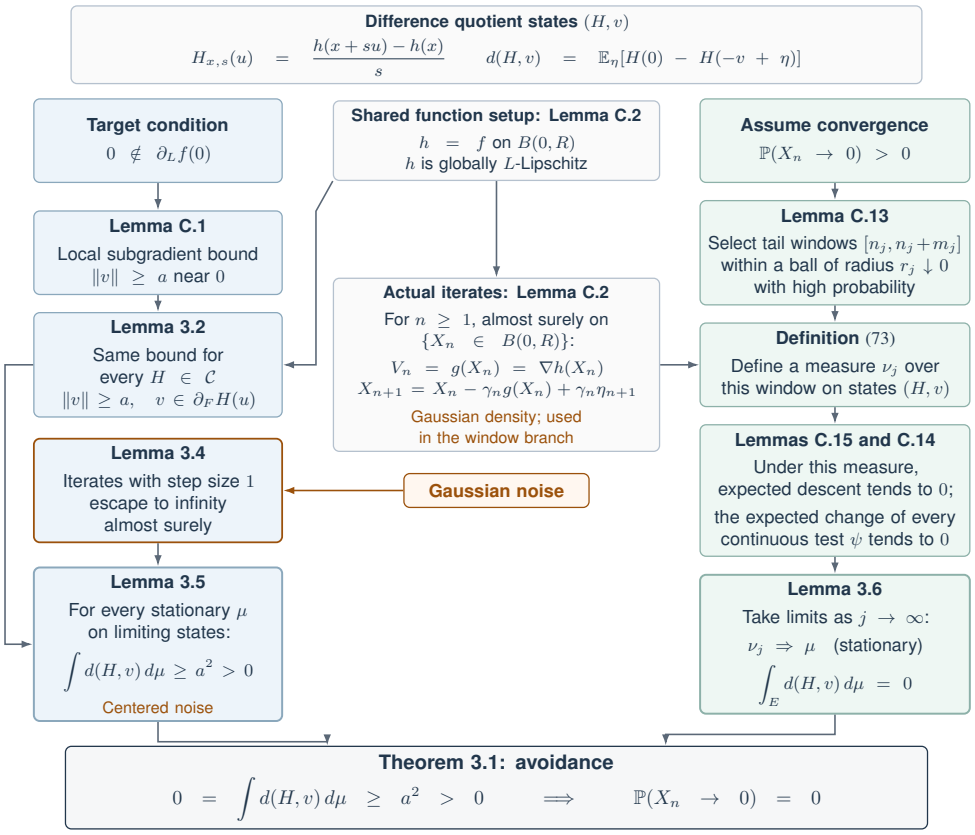

Figure 2: Proof of Gaussian noise avoidance via difference quotient functions, which capture objective changes at a different scale. Left: $0 \notin \partial_L f(0)$ yields a subgradient bound inherited by all limiting functions; Gaussian escape then forces positive mean descent under every stationary distribution. Right: hypothetical convergence yields such a distribution with zero mean descent, a contradiction.

The extension h remains fixed; the limit H may depend on the centers and scales. Lemma C.3 proves compactness of $\mathcal{C}$.

To pass to limits of the recorded pairs $(H_{X_n,s}, V_n)$, we use states $(H, v) \in \mathcal{C} \times \overline{B}(0, L)$. We record v alongside H because locally uniform convergence need not preserve gradients at 0. To express the function from a new center, define

$$(T_z H)(u) = H(u + z) - H(z). \tag{9}$$

For actual quotients, $T_z H_{x,s} = H_{x+sz,s}$: this changes the center while keeping the scale fixed. Given (H, v), draw $\eta \sim N(0, \Sigma)$ and put $\Delta = -v + \eta$. The updated state is $(T_\Delta H, \nabla H(\Delta))$. Since H is Lipschitz and Δ has a density, Rademacher's theorem (Rockafellar & Wets, 1998, Theorem 9.60) implies that the gradient exists with probability 1; set it to 0 at exceptional points. After changing the center to z, $T_z H$ still belongs to $\mathcal{C}$. The updated distribution defines the transition kernel $P((H, v), \cdot)$. The expected descent is

$$d(H, v) = \mathbb{E}_{\eta \sim N(0,\Sigma)}[H(0) - H(-v + \eta)]. \tag{10}$$

For a probability measure μ on this state space, sample $(H, V) \sim \mu$ independently of $\eta \sim N(0, \Sigma)$, and set $\Delta = -V + \eta$. The updated distribution is μP, where for every Borel set A,

$$(\mu P)(A) = \int P((H, v), A)\, d\mu(H, v).$$

Stationarity means $\mu P = \mu$: one update preserves the joint distribution of the function and vector. The mean descent is

$$\int d(H, v)\, d\mu(H, v) = \mathbb{E}_{\substack{(H,V)\sim\mu \\ \eta\sim N(0,\Sigma)}}[H(0) - H(\Delta)]. \tag{11}$$

3.2 POSITIVE EXPECTED DESCENT

The condition $0 \notin \partial_L f(0)$ gives a local Fréchet subgradient bound that becomes global for limiting difference quotient functions. Combined with Gaussian escape, it yields strictly positive stationary mean descent. For some $a > 0$ and sufficiently small $R > 0$,

$$\|w\| \geq a \qquad (x \in \overline{B}(0, R),\ w \in \partial_F f(x)). \tag{12}$$

Lemma 3.2 (Difference quotients preserve the same lower bound)**.** *For every difference quotient $H_{x,s}$,*

$$\partial_F H_{x,s}(u) = \partial_F h(x + su). \tag{13}$$

In particular, whenever $x + su \in B(0, R)$, $\|v\| \geq a$ for every $v \in \partial_F H_{x,s}(u)$, with the same a as in (12)*. Every limiting difference quotient function $H \in \mathcal{C}$ satisfies*

$$\|v\| \geq a \qquad (u \in \mathbb{R}^d,\ v \in \partial_F H(u)). \tag{14}$$

Appendix C.1 gives the complete proof. The translation class $\{T_z H : z \in \mathbb{R}^d\}$ groups functions differing only by a change of center. The following decomposition separates invariant directions from the remaining motion.

Lemma 3.3 (Linear and normal parts of a difference quotient function)**.** *For $H \in \mathcal{C}$ set*

$$S = \{z : T_z H = H\}, \qquad N = S^{\perp}. \tag{15}$$

Then S is a linear subspace. There are $c \in S$ and a definable L-Lipschitz function $K : N \to \mathbb{R}$ such that

$$H(p + q) = \langle c, p\rangle + K(q), \qquad p \in S,\ q \in N. \tag{16}$$

Moreover,

$$\partial_F H(p + q) = \{c + w : w \in \partial_F^N K(q)\}, \tag{17}$$

Here P_S and P_N denote the orthogonal projections onto S and N, and ∂_F^N uses increments within N. The map $z \mapsto T_z H$ is a continuous bijection from N onto the orbit of H, with a Borel inverse.

Within a translation class, the normal position follows $Z^+ = Z - \nabla_N K(Z) + P_N \eta$ on N. Appendix C.3 proves the decomposition and transfers stationarity to these normal positions. The next lemma excludes a stationary distribution under a global positive subgradient bound.

Lemma 3.4 (Fixed-step escape from a global subgradient bound)**.** *Let $H : \mathbb{R}^d \to \mathbb{R}$ be definable and globally L-Lipschitz, where $d \geq 1$. Suppose that, for some $a > 0$,*

$$\|v\| \geq a \qquad \text{for every } x \in \mathbb{R}^d \text{ and } v \in \partial_F H(x). \tag{18}$$

Let Y_0 be finite almost surely and let

$$Y_{n+1} = Y_n - V_n + \eta_{n+1}, \qquad V_n \in \partial_C H(Y_n), \tag{19}$$

where the states and selections are adapted to $(\mathcal{F}_n)$. Assume that the noise vectors are independent and identically distributed with law $N(0, \Sigma)$, that Σ is positive definite, and that η_{n+1} is independent of $\mathcal{F}_n$. Then

$$\mathbb{P}\big(\|Y_n\| \longrightarrow \infty\big) = 1. \tag{20}$$

In particular, for any fixed measurable rule choosing a Clarke subgradient from the current state, one update of (19) *changes every probability distribution.*

The Gaussian density makes an open region with positive escape probability accessible from every bounded set; repeated access yields almost-sure escape (Appendix C.4). If $N \neq \{0\}$ and $\|c\| < a$, the normal potential K inherits the positive subgradient bound $\sqrt{a^2 - \|c\|^2}$, contradicting stationarity by Lemma 3.4. Thus $\|c\| \geq a$; for $N = \{0\}$, the inherited bound gives this directly. Stationarity cancels the expected increment of K, and centering of the noise leaves mean descent $\|c\|^2$.

Lemma 3.5 (Positive expected descent)**.** *Every stationary probability μ for P on $\mathcal{C} \times \overline{B}(0, L)$ satisfies*

$$\int d(H, v)\, d\mu(H, v) \geq a^2 > 0. \tag{21}$$

Appendix C.5 gives the complete argument, including the averaging over translation classes.

3.3 EXPECTED DESCENT EQUAL TO 0 UNDER CONVERGENCE

Assuming convergence to the target with positive probability, we construct a stationary distribution with mean descent 0. The construction uses stopped averages of actual difference quotient states and separate telescoping identities for the values of h and for state tests.

Lemma 3.6 (Expected descent equal to 0 under convergence)**.** *If $\mathbb{P}(X_n \to 0) > 0$, there is a probability measure μ on $\mathcal{C} \times \overline{B}(0, L)$ such that*

$$\int d(H, v)\, d\mu(H, v) = 0, \qquad \mu P = \mu. \tag{22}$$

The full proof is in Appendices C.6–C.7. Here we give a proof sketch following the window branch of Figure 2. First, assuming $\mathbb{P}(X_n \to 0) > 0$, we apply Lemma C.13 to select late windows $[n_j, n_j + m_j]$ whose iterates remain in the ball of radius j^{-2} with conditional probability tending to one. There are increasingly many recorded steps in conditional expectation, and the fixed scale $s_j = \gamma_{n_j}$ makes $\gamma_n / s_j \to 1$ uniformly over each window.

Next, we define a probability ν_j by averaging the distributions of the recorded pairs $(H_{X_n, s_j}, g(X_n))$. We normalize the conditional expected visit counts by the conditional expected number of records. These measures live on the compact space $E = \mathcal{D} \times \overline{B}(0, L)$, where $\mathcal{D}$ is the locally uniform closure of all actual difference quotients. Appendix C.2 extends P and d to this space.

Under these measures, we can show that the mean descent and the mean change of each continuous state test tend to zero. Summing the decreases of h leaves only the difference between its initial and final values. The chosen radius and window length make the conditional expectation of this difference negligible after normalization, even if the last update leaves the ball. Similarly, successive changes of any continuous state test ψ telescope to two bounded endpoint values. Their average vanishes as the expected number of records grows. Replacing the nearly unit step ratios by one, Lemmas C.15 and C.14 give

$$\int_E d\, d\nu_j \longrightarrow 0, \qquad \int_E (P\psi - \psi)\, d\nu_j \longrightarrow 0 \quad (\psi \in C(E)).$$

Finally, by compactness we take a common subsequence $\nu_{j_\ell} \Rightarrow \mu$. The shrinking centers and scales put its support in $\mathcal{C} \times \overline{B}(0, L)$. Bounded continuity of d and $P\psi$ passes both identities to this same limit: its mean descent is zero and it is stationary. The measure μ may involve different limiting functions H; no single common function limit is required.

Applying Lemma 3.5 to this μ gives $0 = \int d\, d\mu \geq a^2 > 0$, the required contradiction.

Corollary 3.7 (Finite sets and generic linear tilts)**.** *Under the function, initialization, noise, and step-size assumptions of Theorem 3.1, every adapted Clarke selection satisfies*

$$\mathbb{P}(\exists x \in A : X_n \to x) = 0$$

for each finite $A \subset \mathbb{R}^d$ with $0 \notin \partial_L f(x)$ for all $x \in A$. Moreover, outside a definable exceptional set of dimension less than d, each fixed tilt u gives a finite set R_u of sharply repulsive points of $f_u(x) = f(x) - \langle u, x \rangle$. Under the same initialization, noise, and step-size assumptions, every adapted Clarke selection for f_u satisfies $\mathbb{P}(\exists x \in R_u : X_n^u \to x) = 0$.

The corollary follows from a finite union and generic finiteness of the Clarke critical set; Appendix C.8 gives the proof. It establishes sharply repulsive-point avoidance in the generic question of (Bianchi et al., 2024, Remark 4).

Remark 3.8 (Noise assumptions)**.** The conclusion also holds for iid noise independent of the past with $\mathbb{E}[\eta_1] = 0$ and $\mathbb{E}[\|\eta_1\|^2] < \infty$, whose joint density is bounded, continuous, and everywhere positive. This includes independent centered Laplace or logistic coordinates and fixed nonsingular linear transforms of these vectors. Centering and the second moment control expected descent and the window boundary terms; the density supplies gradient identification, continuity of the transition, and access to escape regions. The projected noise retains absolute continuity, full support, centering, and finite second moments, which suffice for the normal-chain argument; hence the same proof applies.

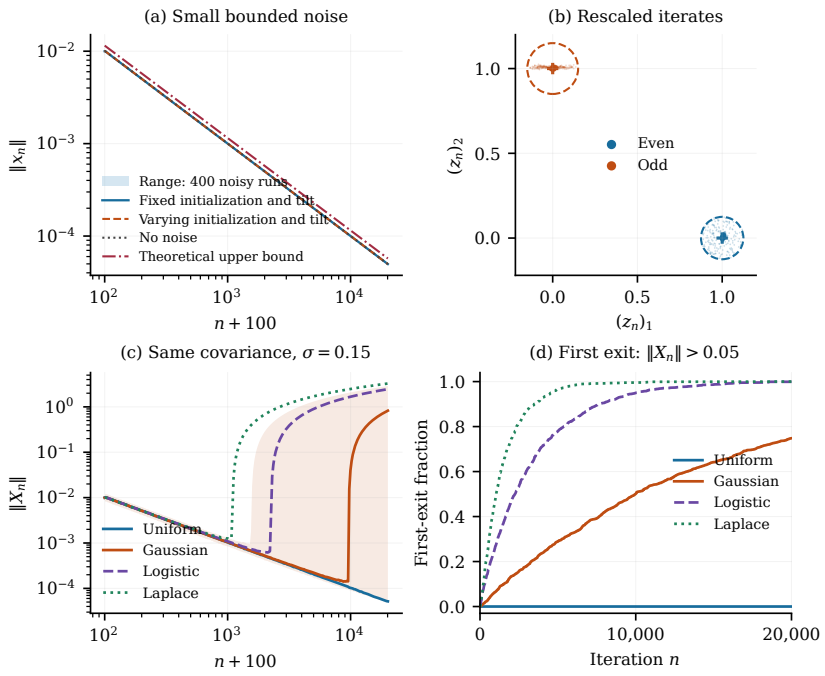

Figure 3: Experiments on the original counterexample. (a,b) Uniform noise of radius 0.005: physical distances and rescaled iterates. (c,d) Four noise laws with covariance $0.15^2 I_2$: median distance and cumulative first exit from radius 0.05, using all $1{,}024$ trajectories per law. Gaussian shading in (c) marks the pointwise 10th–90th percentiles.

4 Numerical Experiments

On the original objective (2), we examine the bounds on rescaled iterates and compare finite-time escapes under different noise distributions. For uniform noise of radius $1/200$, we run two groups of 200 trajectories for $20{,}000$ updates: one fixes $u = 0$ and $x_0 = a/100$; the other varies the admissible initial point and tilt, with each tilt fixed along its trajectory. Figure 3(a,b) shows distances shrinking with the step size and rescaled iterates alternating between two bounded regions, as predicted by the pathwise bounds.

For the noise comparison, we fix $u = 0$, $X_0 = (0.01, 0)$, and $\gamma_n = 1/(n + 100)$. Gaussian, disk-uniform, logistic, and Laplace noises all have covariance $0.15^2 I_2$. For each distribution, we run four independent groups of 256 trajectories for $20{,}000$ updates. Figure 3(c,d) shows escapes under the full-support noises, while uniform-noise distances remain small. Logistic and Laplace escapes occur earlier, and the Gaussian band includes trajectories that have not yet escaped. Equal covariance controls the noise scale, but support and tails both vary, so neither effect is isolated.

The first exit $\|X_n\| > 0.05$ is a finite-horizon statistic; it neither requires permanent escape nor is equivalent to $\mathbb{P}(X_n \to 0) = 0$. Appendix D gives individual trajectories and escape counts.

5 Conclusion

Our bounded-noise counterexample resolves the open question of Bianchi et al. (2024) in the negative, showing that stochastic subgradient descent can converge almost surely to a nonminimal sharply repulsive point. Under Gaussian noise, however, we prove avoidance of prescribed finite sets with $0 \notin \partial_L f(x)$ and, under generic linear perturbations, all sharply repulsive points. A two-cycle argument and a stationary descent contradiction establish the respective results. Building on the Gaussian result, a future direction is to establish avoidance of active strict saddles beyond weak convexity.

AI USE STATEMENT

Generative AI tools assisted with developing, checking, and writing mathematical proofs. The authors reviewed and revised the arguments and take responsibility for the mathematical claims and the final manuscript.

REFERENCES

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research*, 49(3): 1761–1790, 2024. doi: 10.1287/moor.2021.0194.

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2 edition, 1999. Theorem 5.1 (Prohorov's theorem).

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

Evgenii Chzhen and Sholom Schechtman. On convergence rates of subgradient descent on semialgebraic functions. arXiv:2604.17060, 2026.

Aris Daniilidis and Dmitriy Drusvyatskiy. Pathological subgradient dynamics. arXiv:1910.13604, 2019.

Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. arXiv:1804.07795v3, 2018.

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. *Foundations of Computational Mathematics*, 26:779–861, 2026. doi: 10.1007/s10208-025-09691-0. Published online in 2025. Section numbering cited here follows arXiv:2108.11832v2, 9 January 2023.

D. Drusvyatskiy and A. S. Lewis. Semi-algebraic functions have small subdifferentials. *Mathematical Programming*, 140:5–29, 2013. doi: 10.1007/s10107-012-0624-x.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019.

M. Faure and G. Roth. Stochastic approximations of set-valued dynamical systems: convergence with positive probability to an attractor. *Mathematics of Operations Research*, 35(3):624–640, 2010. doi: 10.1287/moor.1100.0455. Corollary 4.10 refers to the author's version 3.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 797–842, 2015.

J. Heinonen. Lectures on lipschitz analysis. Technical Report 100, Department of Mathematics and Statistics, University of Jyväskylä, 2005.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732, 2017.

Cédric Josz and Lexiao Lai. Sufficient conditions for instability of the subgradient method with constant step size. arXiv:2211.14852, 2022.

A. S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, 1995. doi: 10.1007/978-1-4612-4190-4.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1246–1257, 2016.

J.-M. Lion and P. Speissegger. A geometric proof of the definability of Hausdorff limits. *Selecta Mathematica*, 10:377–390, 2004. doi: 10.1007/s00029-004-0360-z.

Junwen Qiu, Bohao Ma, Andre Milzarek, and Junyu Zhang. Stochastic saddle avoidance beyond unit excitation and smoothness: A pathwise Lyapunov–Perron framework. arXiv:2608.03001v2, August 2026. First posted 4 August 2026; version 2 dated 6 August 2026.

Rodolfo Rios-Zertuche. Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions. arXiv:2007.11699, 2020.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1998. doi: 10.1007/978-3-642-02431-3.

S. Schechtman. The gradient's limit of a definable family of functions admits a variational stratification. arXiv:2402.08272v3, 2025. Theorem 4.3 and Corollary 4.5 refer to this fixed version.

Sholom Schechtman. *Some Problems in Nonconvex Stochastic Optimization*. PhD thesis, Université Gustave Eiffel, December 2021.

L. van den Dries. *Tame Topology and O-minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1998.

L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84:497–540, 1996. Section numbers refer to the authors' revised and extended 2001 text.

T. Zell. Topology of definable Hausdorff limits. *Discrete & Computational Geometry*, 33:423–443, 2005. doi: 10.1007/s00454-004-1112-8. Theorem 5.

CONTENTS OF THE APPENDIX

A Preliminaries **12**

A.1 Notation and nonsmooth geometry 12

A.2 Definability and variational tools 13

A.3 Probability and stationary distributions 14

B Proofs for the bounded-noise counterexample **16**

B.1 The objective is globally Lipschitz but not weakly convex 17

B.2 The point 0 is a nonminimal sharply repulsive Clarke critical point 17

B.3 Almost-sure convergence of stochastic subgradient descent to 0 20

B.4 A coercive counterexample preserving convergence to the nonminimal point 0 . . . 25

C Proofs for Gaussian noise avoidance **25**

C.1 Properties of the difference quotient functions 25

C.2 The state transition and its stability 28

C.3 Translation orbits and stationary distributions 31

C.4 Fixed-step escape 36

C.5 Proof of positive expected descent 42

C.6 Convergence windows and stopped averages 43

C.7 Proof of zero expected descent 48

C.8 Finite sets and generic linear tilts 49

D Experimental details and additional results **50**

A PRELIMINARIES

We collect the notation, nonsmooth and definable geometry, and probabilistic tools used in both Sections 2 and 3. Results specific to the counterexample or the Gaussian argument are developed in Appendices B and C, respectively.

A.1 NOTATION AND NONSMOOTH GEOMETRY

We use the Euclidean inner product $\langle x, y\rangle = x^\top y$ and norm $\|x\|$. The open and closed balls are $B(c, r)$ and $\overline{B}(c, r)$; $\|A\|_{\mathrm{op}} = \sup_{\|h\|=1}\|Ah\|$ is the operator norm. The vectors e_i are coordinate vectors, and $e = (1, 1)$ in the counterexample. We write $D\Phi$ for a derivative matrix, ∇f for a gradient, $\nabla^2 f$ for a Hessian, and $\operatorname{conv} S$ for a convex hull.

Definition A.1 (Fréchet, limiting, and Clarke subdifferentials; Rockafellar & Wets, 1998, Definition 8.3 and Theorem 9.61)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. The Fréchet subdifferential consists of the vectors p such that

$$\liminf_{\substack{h\to 0\\ h\neq 0}} \frac{f(x+h) - f(x) - \langle p, h\rangle}{\|h\|} \geq 0. \tag{23}$$

The limiting subdifferential $\partial_L f(x)$ consists of all limits $p_j \to p$ with $x_j \to x$ and $p_j \in \partial_F f(x_j)$. The Clarke subdifferential is

$$\partial_C f(x) = \operatorname{conv}\left\{\lim_j \nabla f(x_j) : x_j \to x,\ f \text{ is differentiable at } x_j,\ \text{the limit exists}\right\}.$$

A point is Clarke critical if $0 \in \partial_C f(x)$.

Continuity makes function-value convergence automatic in the limiting definition. In this setting,

$$\partial_F f(x) \subseteq \partial_L f(x) \subseteq \partial_C f(x), \qquad \partial_C f(x) = \operatorname{conv} \partial_L f(x).$$

The limiting subdifferential has a closed graph (Rockafellar & Wets, 1998, Proposition 8.7). All three subdifferentials equal $\{\nabla f(x)\}$ when f is C^1 near x, and linear perturbations satisfy (Rockafellar & Wets, 1998, Exercise 8.8)

$$\partial_\bullet(f - \langle u, \cdot\rangle)(x) = \partial_\bullet f(x) - u, \qquad \bullet \in \{F, L, C\}.$$

Subdifferential regularity is equivalent here to equality of the three subdifferentials (Rockafellar & Wets, 1998, Definition 7.25 and Corollary 8.11). Weak convexity means that $f + (\lambda/2)\|\cdot\|^2$ is convex for some $\lambda \geq 0$ (Bianchi et al., 2024, Section 2). It implies regularity because convex functions are regular and smooth perturbations preserve regularity (Rockafellar & Wets, 1998, Proposition 8.12 and Exercise 8.8(c)).

Definition A.2 (Active manifold and sharply repulsive critical point)**.** An *active manifold* M at a Clarke critical point $x^\star$ is a smooth manifold through $x^\star$ such that $f|_M$ is smooth near $x^\star$ and there are $\varepsilon, \delta_0 > 0$ for which (Bianchi et al., 2024, Definition 3)

$$\|v\| \geq \delta_0 \quad \text{whenever } x \in B(x^\star, \varepsilon) \setminus M, \quad v \in \partial_C f(x).$$

Following Schechtman (2021, Definition 6.2.4), a Clarke critical point $x^\star$ is called *sharply repulsive* if it admits an active manifold M such that

$$x^\star \text{ is a local minimizer of } f|_M, \qquad 0 \in \partial_C f(x^\star) \setminus \partial_L f(x^\star).$$

The counterexample in Section 2 uses $M = \{0\}$. A singleton is a 0-dimensional smooth manifold; the restriction of any function to it is constant and hence has a local minimum.

A.2 Definability and variational tools

The arguments use definability to control limits and variational tools to transfer subgradient bounds. We collect these facts together with the Lipschitz extension used to handle updates outside a local neighborhood. The notation $\mathcal{L}^d$ denotes Lebesgue measure on $\mathbb{R}^d$.

Definition A.3 (Definable sets and functions; van den Dries & Miller, 1996, Section 2)**.** An o-minimal expansion of the ordered real field is a collection $\mathcal{S} = (\mathcal{S}_m)_{m \geq 1}$ of families of subsets of $\mathbb{R}^m$ that contains all semialgebraic sets, is closed under finite unions, intersections, complements, Cartesian products, and coordinate projections, and whose sets in $\mathcal{S}_1$ are finite unions of points and intervals. A semialgebraic set is a finite Boolean combination of polynomial equalities and inequalities. A set is definable if it belongs to the fixed collection $\mathcal{S}$; a function is definable if its graph does. Real parameters are allowed.

We fix one such structure throughout; see (van den Dries, 1998, Chapter 1) and (van den Dries & Miller, 1996, Section 2). Semialgebraic functions, including polynomials, norms, and finite maxima of polynomials, are definable. An expansion may contain additional functions while retaining the closure properties and the 1-dimensional finiteness condition.

A function $F : \mathbb{R}^d \to \mathbb{R}$ that is definable and locally Lipschitz admits a finite partition of its domain into smooth pieces on each of which it is C^1. In particular, there is a definable open set O with

$$F|_O \in C^1(O), \qquad \mathcal{L}^d(\mathbb{R}^d \setminus O) = 0, \qquad \partial_C F(x) = \{\nabla F(x)\} \quad (x \in O).$$

These statements follow from smooth cell decomposition (van den Dries, 1998, Chapters 3 and 7), (van den Dries & Miller, 1996, Section 4.2), and the Clarke gradient formula (Rockafellar & Wets, 1998, Theorem 9.61). Rademacher's theorem gives almost-everywhere differentiability for any locally Lipschitz function (Rockafellar & Wets, 1998, Theorem 9.60); definability provides the open C^1 set above.

We also use a uniform version of 1-dimensional monotonicity. For a fixed definable family of functions of one real variable, each member is monotone on finitely many intervals, with a bound on the number of intervals independent of the parameters (van den Dries & Miller, 1996, Sections 4.1, 4.2 and 4.4).

The generalized Fermat rule says that a local minimizer has 0 as a Fréchet subgradient (Rockafellar & Wets, 1998, Theorem 10.1). Together with (23), it gives the smooth perturbation and change-of-scale rules used below.

We will use the following two variational facts.

Lemma A.4 (Exact Lipschitz extension)**.** *Let $A \subset \mathbb{R}^d$ be nonempty and let $F : A \to \mathbb{R}$ be L-Lipschitz. Then*

$$\widetilde{F}(x) = \inf_{y \in A} \{F(y) + L\|x - y\|\}$$

is a finite, globally L-Lipschitz function equal to F on A. If A and F are definable, so is $\widetilde{F}$.

This is the McShane–Whitney theorem (Heinonen, 2005, Theorem 2.3); definability follows by expressing the infimum with real quantifiers.

Lemma A.5 (A smooth lower support for a subgradient)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. For $v \in \partial_F F(x)$ there is a C^1 function ϕ on a neighborhood of x such that*

$$\phi(x) = F(x), \qquad \nabla\phi(x) = v, \qquad \phi(y) < F(y) \quad \textit{for } y \neq x \textit{ in that neighborhood.}$$

The lower support may be chosen C^1 and strict away from the contact point. We use this form of (Rockafellar & Wets, 1998, Proposition 8.5, p. 302) to transfer subgradient bounds through convergence of local minimizers.

Definition A.6 (Conservative field; Schechtman, 2025, Definition 2.13)**.** Let $\mathcal{V} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map, so that $\mathcal{V}(x) \subset \mathbb{R}^d$, and assume it has closed graph, nonempty compact values, and is locally bounded. It is conservative for a locally Lipschitz function F if, along every absolutely continuous curve x, for almost every t,

$$(F \circ x)'(t) = \langle v, \dot{x}(t) \rangle \quad \text{for every } v \in \mathcal{V}(x(t)).$$

This is the chain-rule form of a conservative field; see (Schechtman, 2025, Section 2.4, Definition 2.13). A variational stratification of $(F, \mathcal{V})$ is a partition into smooth manifolds satisfying the Whitney compatibility condition, on each of which F is smooth and every vector in $\mathcal{V}(x)$ projects onto the gradient of the restriction to that manifold. The Whitney condition requires limiting tangent spaces of an approaching stratum to contain the tangent space of the stratum it meets; see (Schechtman, 2025, Definitions 2.3 and 3.1). For definable locally Lipschitz F, the pair $(F, \partial_C F)$ admits such a stratification and $\partial_C F$ is conservative in the sense of Definition A.6 (Schechtman, 2025, Section 2.4, Proposition 2.15).

Lemma A.7 (A conservative limit for one parameter)**.** *Let $F_s : \mathbb{R}^d \to \mathbb{R}$, $0 < s \leq 1$, be jointly definable, locally Lipschitz, and locally uniformly convergent to a locally Lipschitz function F_0. Let $\mathcal{V}_s$ be jointly definable closed fields with nonempty values, and suppose each pair $(F_s, \mathcal{V}_s)$ admits a definable variational stratification. If the field values are uniformly bounded on each compact set for small s, then the outer graphical limit of $\mathcal{V}_s$, and its convexification, are conservative for F_0. At x, this limit consists of all v obtained from sequences $s_j \downarrow 0$, $x_j \to x$, and $v_j \in \mathcal{V}_{s_j}(x_j)$ with $v_j \to v$.*

This bounded form follows from (Schechtman, 2025, Theorem 4.3 and Corollary 4.5): local uniform convergence gives the required potential limits and the bounded field has horizon field $\{0\}$: limits of vanishing positive multiples of its values can only be 0. Only a single scalar parameter is used. The limit field need not equal $\partial_C F_0$.

A.3 Probability and stationary distributions

The original stochastic iteration is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with the standard probability-space and random-variable conventions of Durrett (2019, Sections 1.1–1.3). Here $\omega \in \Omega$ specifies a realization of the initialization, the entire noise sequence, and any randomness in the adapted subgradient selections; $\mathcal{F}$ is the sigma-field of events, and $\mathbb{P}$ is their probability measure. A filtration $(\mathcal{F}_n)_{n \geq 0}$ is an increasing sequence of sub-sigma-fields of $\mathcal{F}$; $\mathcal{F}_n$ represents the information available through time n. A process is adapted if its value at time n is $\mathcal{F}_n$-measurable (Durrett, 2019, Section 4.2). In particular, X_n, V_n, and the noises already observed are measurable with

respect to the current history. The sigma-field $\mathcal{F}$ also contains future events, whereas $\mathcal{F}_n$ records only the current information. For an integer stopping time τ, $\mathcal{F}_\tau$ consists of events $A \in \mathcal{F}$ such that $A \cap \{\tau \le n\} \in \mathcal{F}_n$ for every n (Durrett, 2019, Section 4.8).

Expectations and conditioning on the original space. Unsubscripted expectations of path variables use $\mathbb{P}$: $\mathbb{E}[Y] = \int_\Omega Y(\omega)\,\mathbb{P}(d\omega)$. For a probability Q on $(\Omega, \mathcal{F})$ and an integrable Y, $\mathbb{E}_Q[Y \mid \mathcal{G}]$, where $\mathcal{G} \subseteq \mathcal{F}$, is a $\mathcal{G}$-measurable random variable satisfying (Durrett, 2019, Section 4.1)

$$\int_A \mathbb{E}_Q[Y \mid \mathcal{G}]\, dQ = \int_A Y\, dQ \qquad (A \in \mathcal{G}).$$

This specifies it up to equality Q-almost surely. It is generally a random variable, while $\mathbb{E}_Q[Y]$ is a number. For nonnegative Y, the same notation is also used for extended expectations, without assuming integrability. Conditional probabilities mean $\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}[\mathbf{1}_A \mid \mathcal{G}]$. For an event B with $\mathbb{P}(B) > 0$, conditioning instead defines the probability space $(\Omega, \mathcal{F}, \mathbb{P}_B)$, where

$$\mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad \mathbb{E}_{\mathbb{P}_B}[Y] = \frac{1}{\mathbb{P}(B)} \int_B Y\, d\mathbb{P} = \mathbb{E}[Y \mid B].$$

Section C.6 displays this conditioning explicitly as $\mathbb{E}[Y \mid B_j]$. Its notation $\mathbb{E}[Y \mid \mathcal{F}_t, B_j]$ means conditional expectation given $\mathcal{F}_t$ on $(\Omega, \mathcal{F}, \mathbb{P}(\,\cdot \mid B_j))$; these identities are understood almost surely on B_j. We use the tower property and the rule for taking known factors outside a conditional expectation as stated in Durrett (2019, Theorems 4.1.13–4.1.14), applied on this conditional probability space.

A single noise vector and a concrete trajectory space. Write $\mathcal{B}(S)$ for the Borel sigma-field of a metric space S and $\Gamma_\Sigma = N(0, \Sigma)$ for the Gaussian probability measure (Durrett, 2019, Section 3.10). In the Gaussian argument, $\mathbb{E}_\eta$ and $\mathbb{P}_\eta$ use $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Gamma_\Sigma)$ with coordinate random variable $\eta(z) = z$. Thus, with all other arguments fixed,

$$\mathbb{E}_\eta[\phi(\eta)] = \int_{\mathbb{R}^d} \phi(z)\, \Gamma_\Sigma(dz) = \int_{\mathbb{R}^d} \phi(z) \varphi_\Sigma(z)\, dz.$$

A block of k independent noise vectors uses $((\mathbb{R}^d)^k, \mathcal{B}(\mathbb{R}^d)^{\otimes k}, \Gamma_\Sigma^{\otimes k})$; the product-measure construction is recalled in Durrett (2019, Section 1.7). For initialization with distribution ρ_0, independent of the noise, and subgradient selections determined measurably by the observed history, one concrete realization of the whole iteration is

$$\Omega = \mathbb{R}^d \times (\mathbb{R}^d)^{\mathbb{N}}, \qquad \mathcal{F} = \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d)^{\otimes \mathbb{N}},$$
$$\mathbb{P} = \rho_0 \otimes \Gamma_\Sigma^{\otimes \mathbb{N}}, \qquad \mathcal{F}_n = \sigma(X_0, \eta_1, \ldots, \eta_n).$$

Here X_0 and η_i are coordinate projections, and the updates define X_n as functions of these coordinates. Deterministic initialization corresponds to a point mass ρ_0. The countable product is justified by the extension theorem (Durrett, 2019, Theorem 2.1.21). The theorems allow any filtered probability space satisfying their adaptation and independence assumptions, including additional randomness in the selections. For the bounded-noise example, replace Γ_Σ by the uniform ball distribution; for Remark 3.8, replace it by the specified noise distribution.

Expectations over states and independent noise. For a probability ρ on a measurable space $(S, \mathcal{A})$, an integral $\int_S \phi\, d\rho$ is an expectation on $(S, \mathcal{A}, \rho)$, with coordinate random variable $Z(z) = z$. In particular, for the state space E of Section C.2, the integrals against ν_j, μ, or μ_θ use $(E, \mathcal{B}(E), \nu_j)$, $(E, \mathcal{B}(E), \mu)$, or $(E, \mathcal{B}(E), \mu_\theta)$, respectively. The underlying sample is a single function and vector pair. The integral over orbit labels uses $(\Theta, \mathcal{B}(\Theta), \nu)$, where $\mathcal{B}(\Theta)$ denotes its standard Borel sigma-field; this ν is distinct from the stopped averages ν_j. When a state (H, V) with distribution ρ is combined with independent Gaussian noise, the probability space is

$$\big(E \times \mathbb{R}^d,\ \mathcal{B}(E) \otimes \mathcal{B}(\mathbb{R}^d),\ \rho \otimes \Gamma_\Sigma\big),$$

and its coordinate random variables are (H, V) and η. Fubini's theorem (Durrett, 2019, Theorem 1.7.2) gives, for integrable Φ,

$$\mathbb{E}_{\rho \otimes \Gamma_\Sigma}[\Phi(H, V, \eta)] = \int_E \left(\int_{\mathbb{R}^d} \Phi(H, v, z)\, \Gamma_\Sigma(dz) \right) \rho(d(H, v)).$$

This is the joint expectation in (11) when $\rho = \mu$. The normal-position expectations in Section C.5 use the corresponding product space on $N \times \mathbb{R}^d$, specified there. We write $\mathbf{1}_A$ for an indicator and $\mathcal{L}(Z)$ for the distribution of Z under the probability measure specified locally.

For a metric space K, $C(K)$ denotes the real-valued continuous functions on K, $\|\psi\|_{\infty,K} = \sup_{x\in K} |\psi(x)|$, and $\mathrm{Lip}(\psi)$ denotes the Lipschitz constant. We omit K when the domain is clear. The notation $\mu_j \Rightarrow \mu$ means weak convergence of probability measures, that is, convergence of integrals against every bounded continuous test function; the same notation for random variables refers to their distributions (Durrett, 2019, Section 3.10).

Definition A.8 (Stationary probability; Durrett, 2019, Section 5.2 and Example 6.1.2)**.** On a measurable state space K, a Markov kernel P assigns a probability measure $P(x,\cdot)$ to each $x \in K$, with $x \mapsto P(x,A)$ measurable for every measurable $A \subset K$. Thus $P(x,A)$ is the conditional probability that the next state belongs to A. Write

$$P\psi(x) = \int \psi(y)P(x,dy), \qquad (\mu P)(A) = \int P(x,A)\,d\mu(x).$$

A probability μ is stationary if $\mu P = \mu$, equivalently

$$\int P\psi\,d\mu = \int \psi\,d\mu \quad \text{for every bounded measurable } \psi.$$

The transition kernel P is distinct from the path probability measure $\mathbb{P}$. For a fixed input x, $P\psi(x)$ can also be read as an expectation on $(K, \mathcal{B}(K), P(x,\cdot))$ when K is a Borel state space. Thus a state with law μ still has law μ after one transition; see (Durrett, 2019, Example 6.1.2). The general-state-space Markov kernel convention is used throughout.

Lemma A.9 (Upward conditional-expectation convergence)**.** *Let $(\mathcal{F}_n)_{n\ge 0}$ be a filtration on $(\Omega,\mathcal{F},\mathbb{P})$ and put $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$. For an integrable random variable Z,*

$$\mathbb{E}[Z \mid \mathcal{F}_n] \longrightarrow \mathbb{E}[Z \mid \mathcal{F}_\infty] \quad \textit{almost surely and in } L^1.$$

In particular, for $A \in \mathcal{F}_\infty$,

$$\mathbb{P}(A \mid \mathcal{F}_n) \longrightarrow \mathbf{1}_A \quad \textit{almost surely and in } L^1.$$

These are (Durrett, 2019, Theorems 4.6.8–4.6.9). The event form will turn conditional probabilities based on longer histories into indicators of the corresponding eventual behavior.

Lemma A.10 (Doob's L^2 maximal inequality)**.** *Let $(\mathcal{F}_k)_{k=0}^m$ be a filtration and let $(M_k)_{k=0}^m$ be a square-integrable Euclidean-valued martingale: M_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = M_k$. Then*

$$\mathbb{E}\left[\max_{0\le k\le m} \|M_k\|^2 \mid \mathcal{F}_0\right] \le 4\mathbb{E}[\|M_m\|^2 \mid \mathcal{F}_0].$$

Consequently, for every positive $\mathcal{F}_0$-measurable r,

$$\mathbb{P}\left(\max_{0\le k\le m} \|M_k\| > r \mid \mathcal{F}_0\right) \le \frac{4\mathbb{E}[\|M_m\|^2 \mid \mathcal{F}_0]}{r^2}.$$

Apply (Durrett, 2019, Theorem 4.4.4), with $p = 2$, to the nonnegative submartingale $\|M_k\|$. Testing against events in $\mathcal{F}_0$ gives the conditional form; conditional Markov inequality gives the second display. An integer horizon known at time 0 is handled on each of its level events.

B Proofs for the bounded-noise counterexample

The proof of the counterexample in Section 2 combines two arguments: subgradient calculations establish a nonminimal sharply repulsive point 0, and a pathwise two-cycle argument proves convergence to it. A radial modification then makes the objective coercive while preserving its local geometry and the convergent trajectories.

B.1 THE OBJECTIVE IS GLOBALLY LIPSCHITZ BUT NOT WEAKLY CONVEX

The representation as a sum of a norm and piecewise linear terms gives global Lipschitz continuity, definability, and positive homogeneity. A downward jump in a directional derivative excludes weak convexity, while homogeneity supplies the rescaling used in the convergence proof.

Lemma B.1 (Function class and positive homogeneity). *For every $u \in \mathbb{R}^2$, the function f_u is globally Lipschitz with constant at most $2 + 2\sqrt{2} + \|u\|$, definable, and positively homogeneous of degree 1. It is not weakly convex.*

Proof. The Euclidean norm is 1-Lipschitz. Also,

$$|\min\{x, y\} - \min\{\widetilde{x}, \widetilde{y}\}| \le \max\{|x - \widetilde{x}|, |y - \widetilde{y}|\} \le \|(x, y) - (\widetilde{x}, \widetilde{y})\|.$$

The map $(x, y) \mapsto \max\{0, -x - y\}$ is $\sqrt{2}$-Lipschitz. Thus f is globally Lipschitz with constant at most $2 + 2\sqrt{2}$, and f_u has constant at most $2 + 2\sqrt{2} + \|u\|$.

The graph of the norm is described by $r \ge 0$ and $r^2 = x^2 + y^2$. Minimum and maximum of finitely many linear functions have graphs described by finite combinations of polynomial equalities and inequalities. Such descriptions are definable in the real field. Closure of definable functions under finite sums and linear tilts then shows that f and f_u are definable, with u as a real parameter, in the real field and hence in any o-minimal expansion of it. Each term is positively homogeneous of degree 1, so

$$f(tz) = tf(z), \qquad f_u(tz) = tf_u(z) \qquad (t \ge 0).$$

Only nonnegative scaling is asserted; the identities need not hold for negative t.

Restrict f to the line $z(t) = (1 + t, -1 + t)$. Here $\min\{1 + t, -1 + t\} = -1 + t$ and $x + y = 2t$. Therefore

$$f(z(t)) = \begin{cases} \sqrt{2 + 2t^2} + 1 + 3t, & t < 0, \\ \sqrt{2 + 2t^2} + 1 - t, & t \ge 0. \end{cases}$$

The derivative of $\sqrt{2 + 2t^2}$ at $t = 0$ is 0. Hence the one-sided derivatives of this restriction are

$$\frac{d}{dt^-} f(z(t))\bigg|_{t=0} = 3, \qquad \frac{d}{dt^+} f(z(t))\bigg|_{t=0} = -1.$$

For a convex function of one variable, the left derivative cannot exceed the right derivative. The downward jump $3 > -1$ violates this condition. Adding $\lambda\|z(t)\|^2/2$ adds a differentiable function whose derivative is continuous at $t = 0$; it cannot remove the jump. The linear tilt likewise adds the same derivative to both sides. Since the restriction of a convex function to a line is convex, no quadratic correction can make f_u convex. □

The objective therefore has the required Lipschitz and definability properties without weak convexity. We next examine the geometry of its critical point.

B.2 THE POINT 0 IS A NONMINIMAL SHARPLY REPULSIVE CLARKE CRITICAL POINT

A uniform subgradient bound away from 0 supplies the active-manifold condition for $M = \{0\}$. We combine this bound with explicit subdifferential calculations at 0 to establish sharp repulsiveness, and use a descent ray to verify nonminimality.

Set

$$\delta = 1 - \frac{1}{\sqrt{2}} > 0. \tag{24}$$

Lemma B.2 (A uniform Clarke subgradient bound away from 0). *For the function f in (2),*

$$\|v\| \ge \delta \qquad (z \ne 0,\ v \in \partial_C f(z)). \tag{25}$$

Consequently, for $\|u\| < 1/200$, every $z \ne 0$ and $v_u \in \partial_C f_u(z)$ satisfy $\|v_u\| \ge \delta - \|u\| > 0$.

Proof. Write $z = (x, y)$, $\rho = \|z\|$, $w = z/\rho$, and $e = (1, 1)$. Thus $\|w\| = 1$. For the base function $f = f_0$ in (2), at a point off the two lines $x = y$ and $x + y = 0$, differentiation of its three terms gives

$$v = w - p + 2\mathbf{1}_{\{x+y<0\}}e, \qquad p = \begin{cases} e_1, & x < y, \\ e_2, & y < x. \end{cases} \tag{26}$$

Indeed, the norm contributes w, the negative minimum contributes $-p$, and $-2\max\{0, -x - y\}$ contributes $2e$ exactly when $x + y < 0$.

At a nonzero point of $x = y$, only the minimum term changes branch, so its two neighboring gradient limits are convexified: p ranges over $\mathrm{conv}\{e_1, e_2\}$. At a nonzero point of $x + y = 0$, the minimum branch is fixed, while 0 and $2e$ are convexified. Thus the last term becomes te, $0 \le t \le 2$. This t is a convexification parameter, not an iteration index or the line parameter of Section B.1. The lines intersect only at 0, which is excluded here.

Case 1: $x + y > 0$. For every admissible p, including the tie $x = y$, we have $\langle p, w\rangle = \min\{w_1, w_2\}$. It follows that

$$\langle v, w\rangle = 1 - \min\{w_1, w_2\} \ge 1 - \frac{1}{\sqrt{2}} = \delta.$$

The inequality follows because both coordinates of a unit vector cannot exceed $1/\sqrt{2}$. By Cauchy–Schwarz,

$$\|v\| = \|v\|\|w\| \ge |\langle v, w\rangle| \ge \langle v, w\rangle \ge \delta.$$

Case 2: $x + y < 0$. Every admissible p has coordinate sum 1. Since $\langle e, e\rangle = 2$,

$$\langle v, e\rangle = \langle w, e\rangle - 1 + 4 = 3 + \langle w, e\rangle.$$

Cauchy–Schwarz gives $\langle w, e\rangle \ge -\|w\|\|e\| = -\sqrt{2}$. Applying it once more,

$$\|v\| \ge \frac{\langle v, e\rangle}{\|e\|} \ge \frac{3 - \sqrt{2}}{\sqrt{2}} > \delta.$$

Case 3: $x + y = 0$ *and* $z \ne 0$. Since $w_1 + w_2 = 0$ and $w_1^2 + w_2^2 = 1$, we have

$$w = (c, -c) \quad \text{or} \quad w = (-c, c), \qquad c = \frac{1}{\sqrt{2}}.$$

In the first case $y < x$, so $p = e_2$ and

$$v = (c, -c - 1) + t(1, 1), \qquad 0 \le t \le 2.$$

Its first coordinate is $c + t \ge c$, and hence $\|v\| \ge c > \delta$. In the second case $p = e_1$ and

$$v = (-c - 1, c) + t(1, 1), \qquad 0 \le t \le 2;$$

its second coordinate is at least c. This proves (25) everywhere off 0.

For the tilted function, every $v_u \in \partial_C f_u(z)$ has the form $v - u$ with $v \in \partial_C f(z)$. The reverse triangle inequality yields

$$\|v_u\| \ge \|v\| - \|u\| \ge \delta - \|u\| > 0.$$

In particular, no nonzero point is Clarke critical for the stated tilts. □

The lower bound excludes Clarke critical points away from 0. We now show that 0 itself remains Clarke critical under all the allowed tilts.

Lemma B.3 (Clarke criticality persists under small linear tilts)**.** *Let* $c = 1/\sqrt{2}$ *and* $k = c - 1/2 > 1/5$. *Then*

$$(1, -1), \qquad (-1, 1), \qquad k(1, 1), \qquad -\tfrac{1}{2}(1, 1) \quad \textit{belong to } \partial_C f(0). \tag{27}$$

Moreover, $D = \mathrm{conv}\{\pm(1, -1),\ \pm k(1, 1)\} \subset \partial_C f(0)$, *and* $0 \in \partial_C f_u(0)$ *for every* $\|u\| < 1/200$.

Proof. At every differentiability point, (26) is unchanged by positive scaling. Thus, for a fixed smooth direction d, the sequence $d/j \to 0$ has constant gradient $\nabla f(d)$.

For $d = a$, (26) gives $\nabla f(a) = (1, -1)$, while $d = b$ gives $\nabla f(b) = (-1, 1)$. Hence both are gradient limits at 0. They are also limiting subgradients because these points have smooth neighborhoods and their gradients are Fréchet subgradients.

For the positive diagonal, choose $\varepsilon_j = 1/j$ and the smooth points

$$z_j^+ = \frac{1}{j}(1 + \varepsilon_j, 1), \qquad z_j^- = \frac{1}{j}(1, 1 + \varepsilon_j).$$

Both sequences approach 0 with normalized direction (c, c). The first has $y < x$ and the second has $x < y$, giving

$$\nabla f(z_j^+) \to (c, c - 1), \qquad \nabla f(z_j^-) \to (c - 1, c).$$

Their average is $(c - 1/2, c - 1/2) = k(1, 1)$. Convexity of $\partial_C f(0)$ therefore proves membership of the third vector. We do not claim that this average is itself a limiting gradient.

For the fourth vector, take $j \geq 2$ and

$$\widetilde{z}_j^+ = \frac{1}{j}(1, -1 + \varepsilon_j), \qquad \widetilde{z}_j^- = \frac{1}{j}(-1 + \varepsilon_j, 1).$$

Both have strictly positive coordinate sum, so the $2e$ term in (26) is absent. Their normalized directions approach $(c, -c)$ and $(-c, c)$, respectively. Consequently,

$$\nabla f(\widetilde{z}_j^+) \to (c, -c - 1), \qquad \nabla f(\widetilde{z}_j^-) \to (-c - 1, c).$$

The average is $(-1/2, -1/2)$, proving the fourth membership by convexification. This specifies both the approach side and the points tending to 0.

The first two vectors in (27) have midpoint 0. Also, $0 < k < 1/2$, and

$$-k(1, 1) = 2k\big[-\tfrac{1}{2}(1, 1)\big] + (1 - 2k)0.$$

Hence the diamond

$$D = \operatorname{conv}\{\pm(1, -1),\ \pm k(1, 1)\}$$

is contained in $\partial_C f(0)$. To test membership, write

$$u = t(1, -1) + s(1, 1), \qquad t = \frac{u_1 - u_2}{2}, \quad s = \frac{u_1 + u_2}{2}.$$

In this basis, the diamond is the weighted unit ℓ_1 ball: $u \in D$ if and only if $|t| + |s|/k \leq 1$. Indeed, the four vertices have basis coordinates $(\pm 1, 0)$ and $(0, \pm k)$, and their convex hull is exactly this inequality. By Cauchy–Schwarz,

$$|t| \leq \frac{\|u\|}{\sqrt{2}}, \qquad |s| \leq \frac{\|u\|}{\sqrt{2}}.$$

Since $1/k < 5$, we obtain

$$|t| + \frac{|s|}{k} \leq \frac{1 + 1/k}{\sqrt{2}}\|u\| \leq \frac{6}{\sqrt{2}}\|u\| < 1.$$

Thus $u \in D \subset \partial_C f(0)$, and the tilt rule implies $0 \in \partial_C f_u(0)$. □

It remains to exclude limiting criticality at 0 and to find nearby points with lower objective values.

Lemma B.4 (Sharp repulsiveness and nonminimality)**.** *For every $\|u\| < 1/200$, we have $\partial_F f_u(0) = \varnothing$ and $0 \in \partial_C f_u(0) \setminus \partial_L f_u(0)$. The point 0 is a sharply repulsive Clarke critical point with active manifold $M = \{0\}$, and is not a local minimizer.*

Proof. Clarke criticality follows from Lemma B.3. Suppose, for contradiction, that $p \in \partial_F f_u(0)$. Fix any nonzero $h \in \mathbb{R}^2$ and use increments th, $t \downarrow 0$, in the definition. Since $f_u(0) = 0$ and $f_u(th) = t f_u(h)$,

$$\frac{f_u(th) - f_u(0) - \langle p, th\rangle}{\|th\|} = \frac{t f_u(h) - t\langle p, h\rangle}{t\|h\|} = \frac{f_u(h) - \langle p, h\rangle}{\|h\|}.$$

The right-hand side is independent of t. Its limit must be nonnegative, so

$$\langle p, h\rangle \le f_u(h) \qquad \text{for every } h \in \mathbb{R}^2.$$

The case $h = 0$ is immediate. The passage from a local condition to all fixed directions is justified by positive homogeneity, not by convexity.

Take $h = (1, 1)$ and $-h = (-1, -1)$. Direct evaluation gives

$$f_u(h) = \sqrt{2} - 1 - u_1 - u_2, \qquad f_u(-h) = \sqrt{2} - 3 + u_1 + u_2.$$

Adding the two supposed lower-support inequalities yields

$$0 = \langle p, h\rangle + \langle p, -h\rangle \le f_u(h) + f_u(-h) = 2\sqrt{2} - 4 < 0, \tag{28}$$

a contradiction. Therefore $\partial_F f_u(0) = \varnothing$.

If $0 \in \partial_L f_u(0)$, there would be $x_j \to 0$ and $p_j \in \partial_F f_u(x_j)$ with $p_j \to 0$. No x_j can equal 0, because $\partial_F f_u(0)$ is empty. At every remaining point, the inclusion $\partial_F \subseteq \partial_C$ and Lemma B.2 give

$$\|p_j\| \ge \delta - \|u\| > 0,$$

contradicting $p_j \to 0$. We have proved

$$0 \in \partial_C f_u(0) \setminus \partial_L f_u(0). \tag{29}$$

Thus f_u is not subdifferentially regular at 0. The singleton $M = \{0\}$ is a smooth active manifold by the uniform bound away from 0, and $f_u|_M$ is constant. This verifies the stated definition of sharp repulsiveness.

Finally,

$$f_u(-1, -1) = \sqrt{2} - 3 + u_1 + u_2 \le \sqrt{2} - 3 + \sqrt{2}\|u\| < 0 = f_u(0).$$

Positive homogeneity gives $f_u(-t, -t) < 0$ for every $t > 0$. There are therefore points arbitrarily close to 0 with strictly lower function values, so 0 is not a local minimizer. □

This completes the geometric part: 0 is a nonminimal sharply repulsive Clarke critical point. We next show that the stochastic iteration actually converges to it.

B.3 Almost-sure convergence of stochastic subgradient descent to 0

We first rescale the iterates by the step size. Positive homogeneity gives a recursion on two smooth cones, and an explicit two-cycle argument keeps the rescaled iterates in bounded neighborhoods of a and b for every allowed noise sequence. Returning to the original scale then forces convergence to 0. Finally, the noise support condition turns this pathwise conclusion into almost-sure convergence.

B.3.1 The rescaled iterates remain in two bounded regions

For the iteration (1) applied to f_u, define the rescaled iterates and successive step-size ratios by

$$z_n = \frac{x_n}{\gamma_n}, \qquad q_n = \frac{\gamma_n}{\gamma_{n+1}}.$$

We introduce the open cones

$$C_a = \{(x, y) : x > y,\ x + y > 0\}, \qquad C_b = \{(x, y) : y > x,\ x + y > 0\}.$$

They are cones because multiplication by any positive scalar preserves their defining inequalities. On C_a, the minimum is y and the negative-part term vanishes, so $f(x, y) = \rho - y$, where $\rho = \|z\|$. Hence $\nabla f(z) = z/\rho - b$ and

$$T(z) := z - \nabla f(z) = b + (\rho - 1)\frac{z}{\rho} \qquad (z \in C_a). \tag{30}$$

Similarly, on C_b,

$$f(x, y) = \rho - x, \qquad T(z) = a + (\rho - 1)\frac{z}{\rho}. \tag{31}$$

Here T is the unit-step gradient update map for the unperturbed objective on these regions; it is an auxiliary map, not the physical iteration with step γ_n. It is used below to give the two-cycle argument proving Lemma 2.2. Since $\|a\| = \|b\| = 1$, we have $T(a) = b$ and $T(b) = a$. In fact, the unit-circle arc in C_a maps to b, and the arc in C_b maps to a. This explains the choice of the two points.

As long as z_n lies in either cone, $x_n = \gamma_n z_n$ lies in that cone too. Formula (26) gives $\nabla f(\gamma_n z_n) = \nabla f(z_n)$, and $v_n = \nabla f(z_n) - u$. Dividing (1) by γ_{n+1} therefore gives

$$z_{n+1} = q_n\{T(z_n)+u+\eta_{n+1}\}, \qquad 1 \le q_n = 1+\frac{1}{n+100} \le \frac{101}{100}, \qquad \|u+\eta_{n+1}\| \le \frac{1}{100}. \tag{32}$$

The noise in this rescaled recursion has fixed bounded size. The small physical noise $\gamma_n\eta_{n+1}$ must not be treated as a vanishing perturbation after rescaling.

The statement is given in Lemma 2.2.

Proof of Lemma 2.2. We prove (4) by induction. The induction also justifies the smooth formulas at each successive physical iterate; smoothness of the whole future path is not assumed in advance.

Set $r = 1/8$ and $R = 3/20$. If $z = a + h$ with $\|h\| \le r$, then

$$z_1 - z_2 \ge 1 - \sqrt{2}r > 0, \qquad z_1 + z_2 \ge 1 - \sqrt{2}r > 0.$$

Thus $\overline{B}(a, r) \subset C_a$. The same calculation with b and R shows that $\overline{B}(b, R) \subset C_b$. Both balls avoid 0, since their radii are less than 1. Positive rescaling keeps them inside the same smooth cones.

Suppose $z \in \overline{B}(a, r)$, $r = 1/8$. Set $\rho = \|z\|$, $s = \rho - 1$, and $w = z/\rho$. The reverse triangle inequality gives $|s| \le r$ and $\rho \ge 1 - r$. Since the second coordinate of a is 0,

$$|w_2| = \frac{|z_2|}{\rho} \le \frac{r}{1-r} = \frac{1}{7}, \qquad |sw_2| \le \frac{1}{56}.$$

Consider any update $z' = q(T(z) + \xi)$ with $1 \le q \le 101/100$ and $\|\xi\| \le 1/100$. Here ξ represents the combined perturbation $u + \eta$, and the prime labels the next point; it is not a derivative. Put $e' = z' - b$. From $T(z) = b + sw$,

$$e' = (q - 1)b + qsw + q\xi.$$

Taking norms gives

$$\|e'\| \le \frac{1}{100} + \frac{101}{100}\left(\frac{1}{8} + \frac{1}{100}\right) = \frac{2927}{20000} < \frac{3}{20} =: R. \tag{33}$$

Taking the second coordinate instead gives the stronger componentwise estimate

$$|e'_2| \le \frac{1}{100} + \frac{101}{100}\left(\frac{1}{56} + \frac{1}{100}\right) < \frac{1}{25}. \tag{34}$$

This additional information is important. An arbitrary point in $B(b, R)$ need not satisfy (34); an odd iterate produced by the previous phase does.

The next update map is controlled by the radial error $|\|z'\| - 1|$, the distance of z' from the unit circle. This is different from $\|z' - b\|$, the distance from the single point b. The second coordinate estimate improves the radial error because $b = (0, 1)$ is the radial direction at that point.

To estimate the radial error, let $\psi(z) = \|z\|$. For $z \ne 0$,

$$\nabla\psi(z) = \frac{z}{\|z\|}, \qquad \nabla^2\psi(z) = \frac{1}{\|z\|}I - \frac{1}{\|z\|^3}zz^\top = \frac{1}{\|z\|}(I - ww^\top), \quad w = \frac{z}{\|z\|}.$$

Here $zz^\top$ is the outer product, not the scalar $z^\top z$. The matrix $I - ww^\top$ is the orthogonal projection onto the line perpendicular to w. Its eigenvalues in $\mathbb{R}^2$ are 0 and 1, so

$$\|\nabla^2\psi(z)\|_{\mathrm{op}} = \frac{1}{\|z\|}.$$

This identity is a property of the Euclidean norm in dimension two, not the definition of a Hessian operator norm.

Every point $b + te'$, $0 \le t \le 1$, satisfies

$$\|b + te'\| \ge \|b\| - t\|e'\| \ge 1 - R > 0.$$

Thus ψ is twice continuously differentiable along the segment, and its Hessian operator norm there is at most $1/(1 - R)$.

To apply one-variable calculus, set $g(t) = \psi(b + te')$. The vector e' is fixed while t varies. The multivariable chain rule yields

$$g'(t) = \sum_{i=1}^{2} \frac{\partial \psi}{\partial z_i}(b + te')e'_i = \langle \nabla\psi(b + te'), e' \rangle,$$

and differentiating once more gives

$$g''(t) = \sum_{i,j=1}^{2} e'_i \frac{\partial^2 \psi}{\partial z_i \partial z_j}(b + te')e'_j = (e')^\top \nabla^2 \psi(b + te')e'.$$

The fundamental theorem of calculus, applied twice, gives the integral remainder:

$$\begin{aligned} g(1) - g(0) - g'(0) &= \int_0^1 \big(g'(t) - g'(0)\big)\, dt \\ &= \int_0^1 \int_0^t g''(s)\, ds\, dt = \int_0^1 (1 - s)g''(s)\, ds. \end{aligned}$$

Since $\psi(b) = 1$ and $\nabla\psi(b) = b = (0, 1)$, the first-order term is $g'(0) = \langle b, e' \rangle = e'_2$. For any matrix A, Cauchy–Schwarz and the definition of the operator norm give

$$|(e')^\top Ae'| \le \|e'\|\|Ae'\| \le \|A\|_{\mathrm{op}}\|e'\|^2.$$

Consequently,

$$\begin{aligned} |\|b + e'\| - 1 - e'_2| &\le \int_0^1 (1 - t)\left|(e')^\top \nabla^2 \psi(b + te')e'\right| dt \\ &\le \frac{\|e'\|^2}{1 - R} \int_0^1 (1 - t)\, dt = \frac{\|e'\|^2}{2(1 - R)}. \end{aligned}$$

The factor $1/2$ comes from the integral, not from an additional matrix estimate. Combining the last bound with (33)–(34) proves

$$\big|\|z'\| - 1\big| \le |e'_2| + \frac{\|e'\|^2}{2(1 - R)} < \frac{1}{25} + \frac{(3/20)^2}{2(1 - 3/20)} = \frac{1}{25} + \frac{9}{680} < \frac{3}{50}. \tag{35}$$

By (33), $z' \in C_b$. Formula (31) gives

$$\|T(z') - a\| = \left\|(\|z'\| - 1)\frac{z'}{\|z'\|}\right\| = \big|\|z'\| - 1\big| < \frac{3}{50}.$$

Allow a new factor $\widetilde{q} \in [1, 101/100]$ and a new perturbation $\widetilde{\xi}$ with $\|\widetilde{\xi}\| \le 1/100$. They need not equal those in the preceding step. For $z'' = \widetilde{q}(T(z') + \widetilde{\xi})$,

$$\|z'' - a\| \le \frac{101}{100}\left(\frac{3}{50} + \frac{1}{100}\right) + \frac{1}{100} = \frac{807}{10000} < \frac{1}{8}. \tag{36}$$

The base case $z_0 = a$ lies in $\overline{B}(a, r)$. The estimates show that each even iterate generates an odd iterate satisfying both (33) and (34), followed by an even iterate in $B(a, r)$. This proves (4). The same induction starts from any $z_0 \in \overline{B}(a, r)$.

We have not asserted that the two entire balls map into each other in one step. The return estimate uses the extra coordinate information carried by the odd iterate. Finally,

$$\|z_{2j}\| \le 1 + \frac{1}{8} < \frac{23}{20}, \qquad \|z_{2j+1}\| < 1 + \frac{3}{20} = \frac{23}{20}.$$

The cone inclusions and positive scaling give the claimed smoothness and uniqueness. □

Remark B.5. The tangent maps explain why a two-cycle argument works. Using the Hessian formula for $\psi(z) = \|z\|$ derived in the preceding proof, we have $DT(z) = I - \nabla^2\psi(z)$ on either smooth cone. In particular, on the unit circle, $DT(z) = zz^\top$. Thus

$$DT(a) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad DT(b) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad D(T \circ T)(a) = DT(b)DT(a) = 0. \tag{37}$$

The tangent maps $DT(a)$ and $DT(b)$ describe the first-order transport of a small displacement, not its exact evolution. At a, only the horizontal component survives; at b, only the vertical component survives. Hence a displacement passes through the linearized updates as

$$(h_1, h_2) \xmapsto{DT(a)} (h_1, 0) \xmapsto{DT(b)} (0, 0).$$

Since $T(a) = b$, $T(b) = a$, and $T \circ T$ is C^2 near a, Taylor's formula and (37) give

$$T(T(a + h)) - a = O(\|h\|^2) \qquad (h \to 0).$$

Thus the first-order error cancels, but a second-order remainder may remain. This statement concerns the auxiliary unperturbed unit-step map, not the actual rescaled recursion with q_n, u, and η_{n+1}. It explains the two-cycle argument; the explicit estimates above establish trapping for that noisy, time-varying recursion.

Thus the rescaled iterates remain bounded for every noise path in the prescribed ball. We now transfer this deterministic control to the original stochastic iteration.

B.3.2 Returning to the original scale and almost-sure convergence

The vanishing step size turns bounded rescaled iterates into convergence to 0. The next lemma also verifies the moment and excitation properties of the uniform noise.

Lemma B.6 (Uniform ball noise and almost-sure convergence)**.** *Under the setting of Theorem 2.1, the noise is conditionally centered with* $\mathrm{Cov}(\eta_{n+1} \mid \mathcal{F}_n) = I_2/160000$*, and satisfies*

$$\begin{aligned} &\sup_{n \geq 0} \mathbb{E}[\|\eta_{n+1}\|^4 \mid \mathcal{F}_n] \leq (1/200)^4, \\ &\inf_{\|h\|=1} \mathbb{E}[|\langle h, \eta_{n+1}\rangle| \mid \mathcal{F}_n] = \frac{4}{600\pi} > 0 \qquad (n \geq 0). \end{aligned} \tag{38}$$

Moreover, (39) *holds with probability* 1*. In particular,* $\mathbb{P}(\lim_{n\to\infty} x_n = 0) = 1$ *for every fixed admissible tilt and initial point, and for an independent random initialization supported in the stated region.*

Proof. Since $x_n = \gamma_n z_n$, Lemma 2.2 gives (39). The cone inclusions show simultaneously that every physical iterate is nonzero and is a differentiability point. All estimates hold for every perturbation path in the specified support; no averaging or probabilistic cancellation was used.

Let $r_\eta = 1/200$ and let U_{r_η} denote the uniform probability measure on $\overline{B}(0, r_\eta) \subset \mathbb{R}^2$. For deterministic initialization, a concrete probability space is

$$(\Omega, \mathcal{F}, \mathbb{P}) = \big((\mathbb{R}^2)^{\mathbb{N}},\, \mathcal{B}(\mathbb{R}^2)^{\otimes\mathbb{N}},\, U_{r_\eta}^{\otimes\mathbb{N}}\big).$$

The variable η_i is the ith coordinate; take $\mathcal{F}_n = \sigma(\eta_1, \ldots, \eta_n)$. By induction, x_n and its uniquely determined subgradient are measurable with respect to this past. Independence and symmetry of the noise give

$$\mathbb{E}[\eta_{n+1} \mid \mathcal{F}_n] = 0.$$

Thus the noise is a martingale difference sequence. Its density is $1/(\pi r_\eta^2)$ on the ball of radius $r_\eta = 1/200$ centered at 0, and 0 outside. It is absolutely continuous, although its density is not positive everywhere in $\mathbb{R}^2$.

The single-vector moment calculations use $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), U_{r_\eta})$ with coordinate $\eta(z) = z$; write $\mathbb{E}_{U_{r_\eta}}$ for this expectation. For $\eta = (\eta^{(1)}, \eta^{(2)})$, symmetry gives mean 0 and off-diagonal covariance 0. In polar coordinates,

$$\mathbb{E}_{U_{r_\eta}}[(\eta^{(1)})^2] = \frac{1}{\pi r_\eta^2} \int_0^{r_\eta} \int_0^{2\pi} \rho^2 \cos^2\theta \, \rho \, d\theta \, d\rho = \frac{r_\eta^2}{4}.$$

The same holds for the second coordinate. Hence $\mathrm{Cov}(\eta) = r_\eta^2 I_2/4 = I_2/160000$, giving the stated covariance. Also $\mathbb{E}_{U_{r_\eta}} \|\eta\|^m \le r_\eta^m$ for every $m > 0$, so all positive moments are bounded.

Every unit direction receives nontrivial noise. By rotational invariance, for any unit h,

$$\mathbb{E}_{U_{r_\eta}} |\langle h, \eta \rangle| = \frac{1}{\pi r_\eta^2} \int_0^{r_\eta} \int_0^{2\pi} \rho |\cos\theta| \, \rho \, d\theta \, d\rho = \frac{4 r_\eta}{3\pi} = \frac{4}{600\pi} > 0.$$

By symmetry, the expected positive part of this projection is half that number. Thus the noise is omnidirectional, with a common positive lower bound for the expected absolute projection over all unit directions. Independence makes these same identities and bounds conditional on $\mathcal{F}_n$, proving (38).

The step sizes of Theorem 2.1 are positive, decrease to 0, and satisfy

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \qquad \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

With probability 1, every noise vector lies in the closed ball $\overline{B}(0, 1/200)$, because the intersection of these countably many probability-1 events still has probability 1. On that event, the deterministic estimates of Section B.3.1 apply at all times. Therefore $x_n \to 0$ almost surely. The stronger pathwise statement is that any prescribed sequence inside this ball gives the same trapping conclusion, irrespective of independence.

$$\Omega_0 = \bigcap_{n=1}^{\infty} \{\|\eta_n\| \le 1/200\}, \qquad \mathbb{P}(\Omega_0) = 1,$$

$$\Omega_0 \subseteq \{\lim_{n\to\infty} x_n = 0\} \quad \Longrightarrow \quad \mathbb{P}\Big(\lim_{n\to\infty} x_n = 0\Big) = 1.$$

The bounds use u only through $\|u\| \le 1/200$. Consequently, the same noise law and initial point work for every $u \in V := B(0, 1/200)$. Since V is a nonempty open set, it has positive Lebesgue measure. The failure is not confined to a null set of tilt parameters.

Likewise, any initial point in the nonempty open set $U := \gamma_0 B(a, 1/8)$ gives $z_0 \in B(a, 1/8)$ and starts the same induction; Lemma 2.2 also includes its closure. If x_0 has distribution ρ_0, is independent of the entire noise sequence, and is supported in U or its closure, use $\Omega = \mathbb{R}^2 \times (\mathbb{R}^2)^{\mathbb{N}}$ with the product Borel sigma-field and $\mathbb{P} = \rho_0 \otimes U_{r_\eta}^{\otimes \mathbb{N}}$. Take $\mathcal{F}_n = \sigma(x_0, \eta_1, \ldots, \eta_n)$. All conclusions then hold almost surely. More generally, an independent initialization assigning positive probability to U has a failure probability at least $\mathbb{P}(x_0 \in U)$. □

Proof of Theorem 2.1. Lemmas B.1 and B.4 give the stated properties of f_u and 0. On the probability-1 event $\bigcap_{n\ge1}\{\|\eta_n\| \le 1/200\}$, Lemma 2.2 gives smoothness along the trajectory and

$$\|x_n\| = \gamma_n \|z_n\| \le \frac{23}{20(n+100)} \longrightarrow 0. \tag{39}$$

The estimate holds for every allowed initial point and noise path, so it also applies to an independent random initialization supported in the stated set. Lemma B.6 verifies the conditional centering, fourth-moment bound, and excitation of every direction. □

The same argument already covers

$$\gamma_n = (n+100)^{-\alpha}, \qquad x_0 = \gamma_0 a, \qquad \alpha \in (1/2, 1].$$

Indeed,

$$q_n = \left(1 + \frac{1}{n+100}\right)^{\alpha} \in [1, 101/100],$$

which is the only step-ratio property used in the two-cycle argument. Now $\|x_n\| \le (23/20)(n+100)^{-\alpha}$. The two series conditions also hold. In particular, the counterexample does not rely on the endpoint $\alpha = 1$ and also applies to the usual stochastic-approximation step range $(1/2, 1)$.

B.4 A COERCIVE COUNTEREXAMPLE PRESERVING CONVERGENCE TO THE NONMINIMAL POINT 0

The following modification makes the objective coercive without changing the local geometry at 0 or any of the trapped trajectories. Define

$$\widehat{f}_u(z) = f_u(z) + 10\max\{0, \|z\| - 1/10\}. \tag{40}$$

The statement is given in Corollary 2.3.

Proof. The radial term vanishes throughout the open ball $B(0, 1/10)$. From (39),

$$\|x_n\| \le \frac{23}{2000} < \frac{1}{10}.$$

Thus it vanishes on a neighborhood of each trapped iterate and of 0. The two objectives have identical local values and subdifferentials there. Starting from the same initial point and using the same noise, their recursions coincide inductively along the whole trapped trajectory. The active-manifold property, (29), and nonminimality at 0 are all unchanged.

To prove coercivity, note that $\min\{z_1, z_2\} \le \|z\|$, so $\|z\| - \min\{z_1, z_2\} \ge 0$. Also,

$$\max\{0, -z_1 - z_2\} \le |z_1 + z_2| \le \sqrt{2}\|z\|, \qquad -\langle u, z\rangle \ge -\|u\|\|z\|.$$

Consequently $f_u(z) \ge -(2\sqrt{2} + \|u\|)\|z\|$. Since $\max\{0, \|z\| - 1/10\} \ge \|z\| - 1/10$,

$$\widehat{f}_u(z) \ge (10 - 2\sqrt{2} - \|u\|)\|z\| - 1 \longrightarrow +\infty \quad \text{as } \|z\| \to \infty.$$

The coefficient is positive for the stated tilts. Continuity and coercivity imply that a sufficiently low nonempty sublevel set is compact; minimizing over it gives an attained global minimum.

The radial term is globally 10-Lipschitz and definable, since the norm and the maximum of two definable functions are definable. Thus the modified function retains both properties. To see that weak convexity still fails, fix a sufficiently small $r_0 > 0$ and consider $(r_0 + t, -r_0 + t)$ for small $|t|$. This entire short line segment lies in $B(0, 1/10)$, where the radial modification is 0. Its norm term has derivative 0 at $t = 0$, and the same downward derivative jump from 3 to -1 persists, up to the common derivative of the linear tilt. The argument of Section B.1 therefore still excludes weak convexity. □

For $\alpha \in (1/2, 1)$, use $\gamma_n = (n + 1000)^{-\alpha}$ and $x_0 \in \gamma_0 \overline{B}(a, 1/8)$. Then $q_n \le 101/100$ and

$$\|x_n\| \le \frac{23}{20}\gamma_n \le \frac{23}{20}\, 1000^{-\alpha} < \frac{23}{20\sqrt{1000}} < \frac{1}{10}.$$

The same coercive argument applies. Thus the attained-global-minimum version is compatible with the full open interval of exponents used in the related stochastic subgradient setting.

C PROOFS FOR GAUSSIAN NOISE AVOIDANCE

This appendix supplies the proofs for Section 3. We first establish the difference quotient construction and the continuity of its state transition, including the measurability and stability arguments. Translation orbits and fixed-step escape then lead to positive expected descent. The convergence windows and their stopped identities lead to expected descent 0. We finish with the finite-set and generic-tilt corollary.

C.1 PROPERTIES OF THE DIFFERENCE QUOTIENT FUNCTIONS

We prove compactness of the difference quotient family and transfer the local subgradient bound to every limiting function on all of $\mathbb{R}^d$. Local gradient identification connects these functions to the original stochastic iteration; the continuity of their noisy transition is established in Section C.2.

Lemma C.1 (The local Fréchet subgradient bound)**.** *There are a definable open neighborhood U of 0 and $a > 0$ such that*

$$\|v\| \ge a \qquad (x \in U,\ v \in \partial_F f(x)). \tag{41}$$

Proof. If the assertion failed, for every $j \geq 1$ we could choose

$$x_j \in B(0, 1/j), \qquad v_j \in \partial_F f(x_j), \qquad \|v_j\| < 1/j.$$

Then $(x_j, f(x_j), v_j) \to (0, f(0), 0)$, so the definition of the limiting subdifferential gives $0 \in \partial_L f(0)$, a contradiction. A sufficiently small open ball may be used for U. □

Choose $R > 0$ so that $\overline{B}(0, R) \subset U$ and f is L-Lipschitz on this ball. These are the constants used in Section 3.1, and (41) gives (12). Fix $0 < r < R$, set $K_0 = \overline{B}(0, r)$, and define

$$g(x) = \begin{cases} \nabla h(x), & h \text{ is differentiable at } x, \\ 0, & \text{otherwise.} \end{cases}$$

The extension supplies a global Lipschitz bound while preserving f locally, as Figure 4 illustrates. We next identify the actual selections with its gradient near the target.

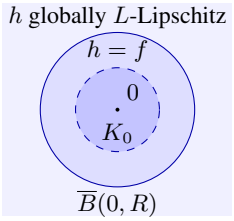

Figure 4: The extension agrees with f around K_0.

Lemma C.2 (Local extension and gradient identification)**.** *The extension h in* (6) *is definable and globally L-Lipschitz, and $h = f$ on an open neighborhood of K_0. Its gradient representative g, defined above, is bounded, definable, and Borel. Almost surely, for every $n \geq 1$,*

$$X_n \in K_0 \quad \Longrightarrow \quad V_n = g(X_n). \tag{42}$$

The same assertion holds after conditioning on any positive-probability event in a past sigma-field.

Proof. The extension formula is Lemma A.4. For $x \in \overline{B}(0, R)$, the Lipschitz inequality gives $h(x) \geq f(x)$, while choosing $y = x$ gives the reverse inequality. Hence $h = f$ on $B(0, R)$, an open neighborhood of K_0, and $\|g\| \leq L$. The infimum and the derivative limits are definable, so h and g are definable and g is Borel.

Choose a definable open set O on which h is C^1, with $\mathcal{L}^d(O^c) = 0$, using the regularity recalled in Section A.2. For $n \geq 1$,

$$\mathcal{L}(X_n \mid \mathcal{F}_{n-1}) = N(X_{n-1} - \gamma_{n-1} V_{n-1}, \gamma_{n-1}^2 \Sigma).$$

Its nondegenerate density implies $X_n \in O$ for every $n \geq 1$ almost surely. Consequently, on $\{X_n \in K_0\}$,

$$V_n \in \partial_C f(X_n) = \partial_C h(X_n) = \{\nabla h(X_n)\} = \{g(X_n)\}.$$

Conditioning on a positive-probability event $B \in \mathcal{F}_m$ preserves every null event. Independence also preserves the joint law of the noise vectors after time m under this conditioning. □

For later use, (7) and (9) give

$$T_z H_{x,s} = H_{x+sz,s}, \qquad H_{x,s}(z) = \frac{h(x + sz) - h(x)}{s}. \tag{43}$$

Introduce the space containing both actual difference quotient functions and their limits,

$$\mathcal{D} = \overline{\{H_{x,s} : x \in \mathbb{R}^d,\ s > 0\}},$$

where the closure is taken for uniform convergence on compact sets. Fix a compatible metric d_0 for this topology.

Lemma C.3 (Compactness and recentering)**.** *The spaces $\mathcal{D}$ and $\mathcal{C} \subset \mathcal{D}$ are nonempty and compact. Their members vanish at 0 and are L-Lipschitz. The map $(H, z) \mapsto T_z H$ is continuous, and for every $z \in \mathbb{R}^d$,*

$$H \in \mathcal{D} \implies T_z H \in \mathcal{D}, \qquad H \in \mathcal{C} \implies T_z H \in \mathcal{C}.$$

Every sequence H_{x_j,s_j} with $x_j \to 0$ and $s_j \to 0$, $s_j > 0$, has a subsequence converging locally uniformly to a member of $\mathcal{C}$.

Proof. The normalization and Lipschitz bound follow directly from (7). Arzelà–Ascoli, followed by a diagonal argument over balls, makes their local-uniform closure $\mathcal{D}$ a compact metric space. The sequence definition (8) is equivalently

$$\mathcal{C} = \bigcap_{j\geq 1} \overline{\{H_{x,s} : \|x\| \leq 1/j,\ 0 < s \leq 1/j\}},$$

with closures taken in $\mathcal{D}$. Each set in this decreasing intersection is nonempty and compact, so $\mathcal{C}$ is nonempty and compact. The equivalence follows by choosing an approximating difference quotient function from the jth set with distance less than $1/j$ in a metric for local-uniform convergence.

The map $(H, z) \mapsto T_z H$ is continuous: local-uniform convergence of H_j and convergence of z_j imply local-uniform convergence of $T_{z_j} H_j$. Also $T_z T_w = T_{z+w}$. For actual difference quotient functions, (43) gives invariance of $\mathcal{D}$ under recentering. If $H_{x_j,s_j} \to H \in \mathcal{C}$ with vanishing centers and scales, then

$$T_z H_{x_j,s_j} = H_{x_j+s_j z,s_j} \longrightarrow T_z H, \qquad x_j + s_j z \to 0.$$

Thus $T_z H \in \mathcal{C}$ as well. Finally, for any $x_j \to 0$ and $s_j \to 0$ with $s_j > 0$, compactness of $\mathcal{D}$ gives a convergent subsequence of H_{x_j,s_j}. Its limit belongs to $\mathcal{C}$ by definition; the scales may be made decreasing by a further subsequence.

□

We now prove Lemma 3.2, which gives the exact subgradient scaling and the inherited bound used in the main text.

Proof of Lemma 3.2. Substituting $y = x + sw$ in (23) cancels the scale in the function value against the scale in the displacement. This proves (13). Local equality $h = f$ and (41) give the stated bound for actual difference quotients.

To pass the same constant to the limit, take actual difference quotient functions $H_j = H_{x_j,s_j} \to H$, with $x_j \to 0$ and $s_j \to 0$, and fix $v \in \partial_F H(u)$. Lemma A.5 supplies a C^1 function ϕ and a small closed ball B about u such that

$$\phi(u) = H(u), \qquad \nabla\phi(u) = v, \qquad \phi(w) < H(w) \quad (w \in B \setminus \{u\}).$$

Choose $u_j \in \arg\min_B (H_j - \phi)$ and put $\varepsilon_j = \|H_j - H\|_{\infty,B} \to 0$. Then

$$0 \leq H(u_j) - \phi(u_j) \leq H_j(u_j) - \phi(u_j) + \varepsilon_j \leq H_j(u) - \phi(u) + \varepsilon_j \leq 2\varepsilon_j.$$

Compactness and the unique zero of $H - \phi$ on B give $u_j \to u$. The minimizers are eventually interior, so the smooth perturbation rule and the scaling in (7) give

$$\nabla\phi(u_j) \in \partial_F H_j(u_j) = \partial_F h(x_j + s_j u_j).$$

Indeed, $y = x_j + s_j w$ in (23) cancels the two scale factors. Since $x_j + s_j u_j \to 0$, local equality and (41) imply, for all large j,

$$\nabla\phi(u_j) \in \partial_F f(x_j + s_j u_j), \qquad \|\nabla\phi(u_j)\| \geq a.$$

Continuity of $\nabla\phi$ gives $\|v\| \geq a$. □

The conclusion is global for each limit H, although the original bound was local for f. There is no loss in the constant a. The global definability of every $H \in \mathcal{C}$, needed for the escape argument, is established in Lemma C.6 by the definable compactification described in Section C.3.

We next relate these difference quotient functions to the original iteration. For $n \geq 1$, $X_n \in K_0$, and a fixed scale $s > 0$, write $H = H_{X_n,s}$, $v = V_n = g(X_n) = \nabla H(0)$, and $\lambda = \gamma_n/s$. Lemma C.2 justifies these gradient identities almost surely. The scaled displacement is

$$\Delta = \frac{X_{n+1} - X_n}{s} = \lambda(-v + \eta_{n+1}). \tag{44}$$

Thus $s\Delta$ is the original displacement, and $H(0) - H(\Delta) = [h(X_n) - h(X_{n+1})]/s$ is its value decrease in units of s.

The same displacement identity holds at every pre-exit update of a window, using its fixed spatial scale s. Choosing $s = \gamma_n$ makes the first rescaled step a unit step; later ratios are controlled in Lemma C.13.

C.2 THE STATE TRANSITION AND ITS STABILITY

We define the noisy transition on difference quotient states and prove its continuity after averaging over the noise, which is needed to pass approximate stationarity to a weak limit. Set $E = \mathcal{D} \times \overline{B}(0, L)$ with its product metric and Borel sigma-field, generated by the open sets. An actual state has $v = \nabla H(0)$, but we retain the two components separately because uniform convergence of functions need not preserve the gradient at the center. For $(H, v) \in E$, $\lambda > 0$, use the noise space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Gamma_\Sigma)$, where $\Gamma_\Sigma = N(0, \Sigma)$ and $\eta(z) = z$. With (H, v) and λ fixed, define

$$\Delta = \lambda(-v + \eta), \qquad H^+ = T_\Delta H, \qquad v^+ = \nabla H(\Delta). \tag{45}$$

The components in (45) are determined by the current state and the noise. In particular,

$$H^+(u) = H(u + \Delta) - H(\Delta), \qquad v^+ = \nabla H(\Delta) = \nabla H^+(0)$$

whenever the derivative exists. Thus the update records the same function viewed from the new point, together with its gradient there. The gradient is assigned 0 at nondifferentiability points; Lemma C.4 gives a jointly Borel choice. The Gaussian displacement has a density and therefore avoids these exceptional points almost surely. By Lemma C.3, recentering preserves $\mathcal{D}$ and $\mathcal{C}$.

For a fixed input (H, v) and a Borel set $A \subset E$, define

$$P_\lambda((H, v), A) = \mathbb{P}_\eta\big((T_{\lambda(-v+\eta)}H, \nabla H(\lambda(-v + \eta))) \in A\big). \tag{46}$$

This is the probability that one update from (H, v) lands in A; the notation $\mathbb{P}_\eta$ means probability under Γ_Σ. Likewise, $\mathbb{E}_\eta$ below integrates only this noise coordinate. Joint measurability makes P_λ a Markov kernel in the sense of Definition A.8. Its action on a bounded Borel test function $\psi : E \to \mathbb{R}$ is

$$\begin{aligned} P_\lambda\psi(H, v) &= \int_E \psi(J, w)\, P_\lambda((H, v), d(J, w)) \\ &= \mathbb{E}_\eta\psi(H^+, v^+). \end{aligned} \tag{47}$$

In particular, if the current state has law μ and the noise is independent, its next-state law is given by

$$(\mu P_\lambda)(A) = \int_E P_\lambda((H, v), A)\, d\mu(H, v).$$

Stationarity for the unit-step kernel $P = P_1$ means that this next-state law equals μ. It does not require a realized state to remain unchanged after the update.

The expected descent at ratio λ is defined by

$$d_\lambda(H, v) = \mathbb{E}_\eta[H(0) - H(\lambda(-v + \eta))]. \tag{48}$$

These definitions extend those of Section 3.1: $P = P_1$ and $d = d_1$. Probabilities on $\mathcal{C} \times \overline{B}(0, L)$ are viewed as probabilities on E assigning mass 1 to that subset. The Lipschitz bound gives

$$|d_\lambda(H, v)| \le \lambda L(L + \mathbb{E}_\eta\|\eta\|).$$

The next result supplies the continuity needed to take limits in the difference quotient function transition (45). We then verify that stationarity identifies the recorded vector with the current gradient.

Uniform monotonicity for the definable family bounds the number of monotone pieces on every coordinate line. This gives local L^1 control of gradient differences by function-value differences. Integration against the Gaussian density then gives continuity of the update. We give the measurability and continuity proofs next. The following arguments supply the two regularity facts used to define the kernel in (45) and to pass stopped occupation measures to a stationary limit. The gradient representative is Borel by countable difference-quotient tests. Stability then follows from uniform definable monotonicity and Gaussian averaging.

C.2.1 A MEASURABLE GRADIENT REPRESENTATIVE

Lemma C.4 (A jointly measurable gradient)**.** *On the compact family $\mathcal{D}$ of normalized L-Lipschitz difference quotient functions, the map $(H, u) \mapsto \nabla H(u)$, with value 0 whenever H is not differentiable at u, is jointly Borel and takes values in $\overline{B}(0, L)$. Moreover,*

$$G = \{(H, v) \in E : H \text{ is differentiable at } 0,\ v = \nabla H(0)\}$$

is a Borel subset of E.

Proof. Let e_i denote the ith standard basis vector. For each coordinate form the limits

$$b_i(H,u) = \lim_{m\to\infty} m[H(u+e_i/m) - H(u)],$$

assigning 0 when a limit does not exist. Evaluation is continuous, so the limits and their existence sets are Borel. Where all coordinate limits exist, differentiability with derivative b is equivalent to

$$\lim_{k\to\infty} \sup_{\substack{q\in\mathbb{Q}^d \\ 0<\|q\|\le 1/k}} \frac{|H(u+q) - H(u) - \langle b, q\rangle|}{\|q\|} = 0.$$

For fixed H, u, b, the quotient is continuous in each nonzero increment. Density of rational increments therefore extends the test to all real increments. The test is Borel as a countable supremum followed by a countable limit. Using b on this Borel set and 0 elsewhere gives the required joint representative; the Lipschitz bound implies $\|\nabla H(u)\| \le L$.

All difference quotient functions satisfy $H(0) = 0$. Thus a pair $(H, v) \in E$ belongs to G exactly when

$$\lim_{k\to\infty} \sup_{\substack{u\in\mathbb{Q}^d \\ 0<\|u\|\le 1/k}} \frac{|H(u) - \langle v, u\rangle|}{\|u\|} = 0.$$

Continuity in the increment again identifies this rational test with differentiability at 0 with derivative v. Each expression is Borel in (H, v), so G is Borel. □

C.2.2 Continuity after Gaussian averaging

Proposition C.5 (Stability of one noisy update)**.** *Let $H_j, H \in \mathcal{D}$, $v_j, v \in \overline{B}(0, L)$, and $\lambda_j, \lambda > 0$. Suppose $H_j \to H$ uniformly on compact sets, $v_j \to v$, and $\lambda_j \to \lambda$. Apply* (45) *to these inputs. Then*

$$(H_j^+, v_j^+) \;\Rightarrow\; (H^+, v^+), \qquad d_{\lambda_j}(H_j, v_j) \longrightarrow d_\lambda(H, v).$$

In particular, $P_\lambda\psi(H, v)$ is jointly continuous in (H, v, λ) for every continuous $\psi : E \to \mathbb{R}$.

Proof of Proposition C.5. We first estimate gradient errors directly from function errors. Fix $R > 0$, set $Q_R = (-R, R)^d$, and take two actual difference quotient functions H, J from (7). Their differences along coordinate lines form one definable family; both sets of difference quotient function parameters and all transverse coordinates are included as parameters. Uniform monotonicity gives a common upper bound M on the number of monotone intervals of each such difference (van den Dries & Miller, 1996, §§4.1, 4.2 and 4.4).

Write $q(t) = H(y + te_i) - J(y + te_i)$ on a coordinate segment in Q_R, with $y \in e_i^\perp$ and $|y_k| < R$ for $k \ne i$. This function is Lipschitz and hence absolutely continuous. For its monotone partition $-R = t_0 < \cdots < t_m = R$, $m \le M$,

$$\int_{-R}^{R} |q'(t)|\, dt = \sum_{\ell=1}^{m} |q(t_\ell) - q(t_{\ell-1})| \le 2M \sup_{z\in Q_R} |H(z) - J(z)|.$$

For each fixed pair H, J, the slice derivatives agree with the corresponding gradient components for almost every t on almost every line. Integrating over y and summing over coordinate directions gives

$$\int_{Q_R} \|\nabla H - \nabla J\| \le C_R \sup_{z\in Q_R} |H(z) - J(z)|, \qquad C_R = 2Md(2R)^{d-1}. \tag{49}$$

The constant is independent of the two difference quotient functions.

To extend this estimate to $\mathcal{D}$, let actual difference quotient functions $F_k \to F$ locally uniformly. Equation (49) makes ∇F_k Cauchy in $L^1(Q_R)$ for every R; let its local limit be G. For a smooth compactly supported test function $\phi \in C_c^\infty(\mathbb{R}^d)$,

$$\int G_i\phi = \lim_k \int \partial_i F_k \phi = -\lim_k \int F_k \partial_i\phi = -\int F\partial_i\phi.$$

The limit F is L-Lipschitz, so $G = \nabla F$ almost everywhere. Approximating each of $H, J \in \mathcal{D}$ by actual difference quotient functions now extends (49) to all of $\mathcal{D}$. In particular, $H_j \to H$ locally uniformly implies $\nabla H_j \to \nabla H$ in local L^1. This argument does not require the limits to be definable.

For the random update, use a metric d_0 on $\mathcal{D}$ for local uniform convergence and the sum product metric on E. Let φ_Σ denote the density of $N(0, \Sigma)$. In the common displacement variable z, the Gaussian densities are

$$p_j(z) = \lambda_j^{-d} \varphi_\Sigma(z/\lambda_j + v_j), \qquad p(z) = \lambda^{-d} \varphi_\Sigma(z/\lambda + v).$$

They converge pointwise and integrate to 1. Thus

$$\|p_j - p\|_{L^1} = 2 - 2 \int \min\{p_j, p\} \longrightarrow 0,$$

by dominated convergence with dominating function p. For a bounded Lipschitz test ψ on E, both updates can be integrated over the same displacement z:

$$P_{\lambda_j} \psi(H_j, v_j) = \int \psi(T_z H_j, \nabla H_j(z)) p_j(z)\, dz, \quad P_\lambda \psi(H, v) = \int \psi(T_z H, \nabla H(z)) p(z)\, dz.$$

Add and subtract the first integrand with density p, then separate Q_R from its complement. This gives

$$\begin{aligned}
&|P_{\lambda_j} \psi(H_j, v_j) - P_\lambda \psi(H, v)| \\
&= \Big| \int \Big\{ \psi(T_z H_j, \nabla H_j(z))[p_j(z) - p(z)] \\
&\qquad\qquad + [\psi(T_z H_j, \nabla H_j(z)) - \psi(T_z H, \nabla H(z))] p(z) \Big\}\, dz \Big| \\
&\le \|\psi\|_\infty \|p_j - p\|_{L^1} \\
&\quad + \int_{Q_R} |\psi(T_z H_j, \nabla H_j(z)) - \psi(T_z H, \nabla H(z))| p(z)\, dz + 2\|\psi\|_\infty \int_{Q_R^c} p(z)\, dz \\
&\le \|\psi\|_\infty \|p_j - p\|_{L^1} + \mathrm{Lip}(\psi) \int_{Q_R} d_0(T_z H_j, T_z H) p(z)\, dz \\
&\quad + \mathrm{Lip}(\psi) \int_{Q_R} \|\nabla H_j(z) - \nabla H(z)\| p(z)\, dz + 2\|\psi\|_\infty \int_{Q_R^c} p(z)\, dz \\
&\le \|\psi\|_\infty \|p_j - p\|_{L^1} + \mathrm{Lip}(\psi) \sup_{z \in Q_R} d_0(T_z H_j, T_z H) \\
&\quad + \mathrm{Lip}(\psi) \|p\|_\infty C_R \sup_{z \in Q_R} |H_j(z) - H(z)| + 2\|\psi\|_\infty \int_{Q_R^c} p(z)\, dz.
\end{aligned}$$

The four terms respectively control the change of density, the translated functions, the gradients through (49), and the tail outside Q_R. For fixed R, joint continuity of translation on the compact closure $\overline{Q}_R$ makes the supremum tend to 0. First let $j \to \infty$, then $R \to \infty$. Bounded Lipschitz functions are uniformly dense in $C(E)$; for any such approximation ψ_ϵ,

$$\begin{aligned}
&|P_{\lambda_j} \psi(H_j, v_j) - P_\lambda \psi(H, v)| \\
&\quad \le 2\|\psi - \psi_\epsilon\|_\infty + |P_{\lambda_j} \psi_\epsilon(H_j, v_j) - P_\lambda \psi_\epsilon(H, v)|.
\end{aligned}$$

This proves continuity for every continuous test, and hence the claimed convergence in distribution.

For each fixed η, local uniform convergence and the common Lipschitz bound give

$$H_j(0) - H_j(\lambda_j(-v_j + \eta)) \to H(0) - H(\lambda(-v + \eta)),$$

and

$$|H_j(0) - H_j(\lambda_j(-v_j + \eta))| \le L\big(\sup_j \lambda_j\big)(L + \|\eta\|).$$

The bound is integrable, so dominated convergence proves the assertion about the expected descent. □

The current gradient is correct after one step: for every input, the displacement has a density and H is differentiable there almost surely. Hence

$$v^+ = \nabla H(\Delta) = \nabla H^+(0) \quad \text{almost surely.}$$

Let $G = \{(H, v) : H \text{ is differentiable at } 0,\ v = \nabla H(0)\}$, which is Borel by Lemma C.4. Since $P((H, v), G) = 1$ for every input, every stationary probability satisfies

$$\mu(G) = \mu P(G) = \int_E P(y, G)\, d\mu(y) = 1. \tag{50}$$

Thus stationarity restores the gradient relation that need not be preserved by a deterministic limit at the center of a difference quotient function.

Joint continuity on compact sets also gives

$$\sup_{(H,v)\in E} |P_\lambda \psi(H, v) - P\psi(H, v)| \to 0, \qquad \|d_\lambda - d\|_\infty \to 0 \quad \text{as } \lambda \to 1. \tag{51}$$

C.3 Translation orbits and stationary distributions

To compute stationary mean descent, we decompose a stationary distribution into components supported on individual translation orbits. Definability provides measurable orbit representatives, and each component reduces to a position process in the normal directions.

C.3.1 Definable limits and measurable orbit representatives

A definable parametrization of the limiting functions permits measurable choices of representatives for their translation orbits. For nonempty compact sets A, B, $d_H(A, B)$ denotes their Hausdorff distance, the larger of $\sup_{a\in A} \inf_{b\in B} \|a - b\|$ and $\sup_{b\in B} \inf_{a\in A} \|a - b\|$. A standard Borel space is a measurable space isomorphic to a Borel subset of a complete separable metric space. Such spaces admit the conditional distributions used below.

Lemma C.6 (A jointly definable family of limiting functions)**.** *There are a compact definable parameter set P_0 and a jointly definable family $(H_p)_{p\in P_0}$ such that*

$$\mathcal{C} = \{H_p : p \in P_0\}, \qquad p_j \to p \implies H_{p_j} \to H_p \text{ locally uniformly.}$$

In particular, every $H \in \mathcal{C}$ is globally definable.

Proof. We use the Lion–Speissegger compactification Lion & Speissegger (2004), in the form of (Zell, 2005, Theorem 5): a bounded definable family of nonempty compact sets A_t is contained in a definable family B_p parametrized by a compact definable set P, with $p \mapsto B_p$ Hausdorff continuous. To apply the compactification theorem, we represent each difference quotient function by a compact set that also records its center and scale. For $\|x\| \le 1$ and $0 < s \le 1$, define

$$A_{x,s} = \overline{\left\{ \left(\frac{u}{\sqrt{1 + \|u\|^2}}, \frac{H_{x,s}(u)}{\sqrt{1 + \|u\|^2}}, x, s \right) : u \in \mathbb{R}^d \right\}}. \tag{52}$$

The closure is taken in $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ with (x, s) fixed. Thus $A_{x,s}$ is the closure of the transformed graph of $H_{x,s}$, with the same center x and scale s attached to every point. Its first two components record the transformed argument and function value; its last two components retain the parameters from which the difference quotient function was obtained. Since $H_{x,s}(0) = 0$ and $H_{x,s}$ is L-Lipschitz, these nonempty compact sets satisfy

$$A_{x,s} \subset \overline{B}(0, 1) \times [-L, L] \times \{x\} \times \{s\},$$

and form a bounded definable family over the stated parameter range. The first component has norm strictly below 1 at every finite argument u; points with first-component norm 1 are added by the closure to record behavior as $\|u\| \to \infty$. This theorem gives a compact parameter set P and a Hausdorff-continuous definable family B_p containing all these sets.

Let P_1 consist of the parameters representing an original set $A_{x,s}$, and let P_0 consist of their limiting parameters for which the recorded center and scale are both 0. Explicitly, set

$$P_1 = \{p \in P : \exists x \in \overline{B}(0,1),\ \exists s \in (0,1],\ B_p = A_{x,s}\},$$
$$P_0 = \{p \in \overline{P_1} : B_p \subset \overline{B}(0,1) \times [-L, L] \times \{(0,0)\}\}.$$

These sets are definable, and P_0 is compact. If $p_j \in P_1$ converges to $p \in P_0$, write $B_{p_j} = A_{x_j,s_j}$. Hausdorff continuity and the recorded parameters give

$$d_H(A_{x_j,s_j}, B_p) \to 0, \qquad (x_j, s_j) \to (0,0).$$

Conversely, a sequence with $(x_j, s_j) \to (0,0)$ has representing parameters in P_1 with a subsequence converging into P_0.

For $p \in P_0$, the part of B_p with spatial norm strictly below one determines a unique global L-Lipschitz function H_p with $H_p(0) = 0$. More precisely, a point $(y, r, 0, 0) \in B_p$, where $y \in \mathbb{R}^d$, $\|y\| < 1$, and $r \in \mathbb{R}$, recovers an argument and its function value by

$$u = \frac{y}{\sqrt{1 - \|y\|^2}}, \qquad H_p(u) = \frac{r}{\sqrt{1 - \|y\|^2}}.$$

To justify this reconstruction, extract a locally uniformly convergent subsequence of the approaching difference quotient functions. The displayed inverse is continuous when $\|y\| < 1$, so Hausdorff convergence identifies the recovered value with that of the difference quotient function limit. The interior graph determines the function, and hence

$$\{H_p : p \in P_0\} = \mathcal{C}, \qquad p_j \to p \quad \Longrightarrow \quad H_{p_j} \to H_p \text{ locally uniformly.}$$

The convergence follows from compactness of the normalized Lipschitz family: every subsequential limit has the same interior graph. Evaluation $(p, u) \mapsto H_p(u)$ is definable by the inverse formula. Different parameters may still encode different boundary data at infinity for the same function.

□

The joint family is needed below: it makes equality and translation equivalence definable relations on a single parameter set, rather than separate assertions about individual functions.

Lemma C.7 (A measurable quotient by translations)**.** *There are a standard Borel space Q, a Borel map $\rho : \mathcal{C} \to Q$ whose fibers are exactly the translation orbits, and a Borel choice of a representative function for each $q \in Q$.*

Proof. Use the family from Lemma C.6. Strong definable choice (van den Dries, 1998, Chapter 6, Proposition 1.2) states that a definable family of nonempty sets D_t admits a definable choice $b(t) \in D_t$ with $D_t = D_{t'} \Rightarrow b(t) = b(t')$. We apply it first to remove duplicate function parameters and then to choose one function from each translation class.

Apply strong choice to the definable relation $p \sim p'$ if $H_p(u) = H_{p'}(u)$ for all u. Let $P_* \subset P_0$ be the resulting definable set containing exactly one parameter for each distinct function. The map

$$\chi : P_* \to \mathcal{C}, \qquad \chi(p) = H_p,$$

is a continuous bijection. Its inverse is Borel by the Lusin–Souslin theorem (Kechris, 1995, Corollary 15.2). On P_* the orbit relation is

$$p \simeq p' \quad \Longleftrightarrow \quad \exists z\, \forall u \quad H_{p'}(u) = H_p(u + z) - H_p(z).$$

A second application of strong choice gives a definable representative map $b : P_* \to Q \subset P_*$ with

$$b(p) = b(p') \quad \Longleftrightarrow \quad p \simeq p'.$$

Here $Q = b(P_*)$ contains one parameter for each translation orbit. Thus $\rho = b \circ \chi^{-1}$ is Borel and

$$\rho(H) = \rho(J) \quad \Longleftrightarrow \quad J = T_z H \text{ for some } z \in \mathbb{R}^d.$$

The representative of $q \in Q$ is $\chi(q)$. The definable Euclidean set Q is standard Borel, completing the construction. □

C.3.2 THE GEOMETRY OF A TRANSLATION ORBIT

The translations satisfying $T_z H = H$ form a subspace S. Splitting $\mathbb{R}^d = S \oplus S^\perp$ separates a linear part of H from the normal function that determines its translation orbit. For example, in two dimensions,

$$H(u_1, u_2) = cu_1 + k(u_2) \quad \Longrightarrow \quad (T_{(a,b)}H)(u_1, u_2) = cu_1 + k(u_2 + b) - k(b).$$

The position a does not enter the new difference quotient function. The construction below also includes any additional translation symmetries of k.

Proof of Lemma 3.3. Continuity and the translation group law give

$$z_i \in S,\ z_i \to z \implies T_z H = H, \qquad T_{z+w}H = T_z T_w H, \qquad T_{-z}T_z H = H.$$

Thus S is a closed additive subgroup. It is definable by Lemma C.6. For $z \in S$, the definable subgroup $\{t \in \mathbb{R} : tz \in S\}$ contains $\mathbb{Z}$, hence contains an interval I of positive length. Consequently

$$\mathbb{R} = \bigcup_{m \geq 1} m(I - I) \subset \{t : tz \in S\},$$

so S is a linear subspace.

The identity $T_p H = H$ gives

$$H(u + p) = H(u) + H(p), \qquad p \in S, \qquad H(tp) = tH(p), \qquad t \in \mathbb{Q}.$$

For $t_i \in \mathbb{Q}$ with $t_i \to t \in \mathbb{R}$, continuity yields $H(tp) = \lim_i H(t_i p) = tH(p)$. In an orthonormal basis $e_1, \ldots, e_{\dim S}$, set $c = \sum_i H(e_i)e_i$. Then

$$H(p) = \sum_i \langle p, e_i \rangle H(e_i) = \langle c, p \rangle, \qquad H(p + q) = \langle c, p \rangle + H(q).$$

Taking $K = H|_N$ proves (16); restriction preserves definability and the Lipschitz bound.

Let $v \in \partial_F H(p + q)$. For $0 \neq b \in S$, increments tb give

$$0 \leq \liminf_{t \to 0} \frac{t\langle c - P_S v, b \rangle}{|t|\|b\|} = -\frac{|\langle c - P_S v, b \rangle|}{\|b\|}.$$

Hence $P_S v = c$. With increments in N,

$$\liminf_{b \to 0,\ b \in N \setminus \{0\}} \frac{K(q + b) - K(q) - \langle P_N v, b \rangle}{\|b\|} \geq 0,$$

which gives $P_N v \in \partial_F^N K(q)$ and the first inclusion in (17). Conversely, let $w \in \partial_F^N K(q)$. For every $\epsilon > 0$ and sufficiently small $b \in N$, $K(q + b) - K(q) - \langle w, b \rangle \geq -\epsilon\|b\|$, also when $b = 0$. For a sufficiently small orthogonal increment $a + b \in S + N$,

$$\begin{aligned} &H(p + q + a + b) - H(p + q) - \langle c + w, a + b \rangle \\ &\quad = K(q + b) - K(q) - \langle w, b \rangle \geq -\epsilon\|b\| \geq -\epsilon\|a + b\|. \end{aligned}$$

This proves $c + w \in \partial_F H(p + q)$ and the reverse inclusion.

Finally,

$$T_{p+q}H = T_q H, \qquad T_{q_1}H = T_{q_2}H \iff q_1 - q_2 \in S \cap N = \{0\}.$$

Thus $z \mapsto T_z H$ is a continuous bijection from N onto its orbit. Its inverse is Borel by the Lusin–Souslin consequence (Kechris, 1995, Corollary 15.2), applied to this injection into the compact metric space $\mathcal{C}$. □

Thus (16) is an exact decomposition, not a Taylor linearization. Movement along S contributes a linear change in value but does not need to be recorded as a new difference quotient function position. This observation turns the difference quotient function kernel into an ordinary chain on N.

Lemma C.8 (The stationary chain in normal coordinates)**.** *Let a stationary probability for P assign probability 1 to pairs whose difference quotient function is a translation of $H \in \mathcal{C}$. Under the identification in Lemma 3.3, its normal position has a stationary law for*

$$Z^+ = Z - \nabla_N K(Z) + P_N \eta. \tag{53}$$

Its full displacement and its difference quotient function decrease are

$$\Delta = -c - \nabla_N K(Z) + \eta, \tag{54}$$

$$(T_Z H)(0) - (T_Z H)(\Delta) = \|c\|^2 - \langle c, P_S \eta \rangle + K(Z) - K(Z^+). \tag{55}$$

If N has positive dimension, the stationary normal law is absolutely continuous on N.

Proof. Let π be the distribution on N obtained by taking the normal position of a state with the given stationary distribution. Realize the position and the next noise on

$$\big(N \times \mathbb{R}^d,\, \mathcal{B}(N) \otimes \mathcal{B}(\mathbb{R}^d),\, \pi \otimes \Gamma_\Sigma\big), \qquad Z(z,e) = z, \quad \eta(z,e) = e.$$

Here $\Gamma_\Sigma = N(0, \Sigma)$, and H, N, K, and c are fixed. The product measure makes Z and η independent. The stationary-gradient identity (50) identifies the mark at the state $T_Z H$ almost surely. Differentiating (16) gives

$$v = \nabla (T_Z H)(0) = \nabla H(Z) = c + \nabla_N K(Z), \qquad \Delta = -c - \nabla_N K(Z) + \eta.$$

Differentiability of H at Z and of K on N at Z are equivalent by the same decomposition. For the output difference quotient function,

$$H^+ = T_\Delta T_Z H = T_{Z+\Delta} H = T_{P_N(Z+\Delta)} H,$$
$$Z^+ = P_N(Z + \Delta) = Z - \nabla_N K(Z) + P_N \eta.$$

The Borel inverse in Lemma 3.3 therefore takes a stationary difference quotient function law to a stationary law for (53). Also,

$$\begin{aligned}(T_Z H)(0) - (T_Z H)(\Delta) &= H(Z) - H(Z + \Delta) \\ &= -\langle c, P_S \Delta \rangle + K(Z) - K(Z + P_N \Delta) \\ &= \|c\|^2 - \langle c, P_S \eta \rangle + K(Z) - K(Z^+),\end{aligned}$$

which proves (55).

If $\dim N > 0$, then $u^\mathsf{T} \Sigma u > 0$ for every $u \in N \setminus \{0\}$. Write φ_N for the Gaussian density of $P_N \eta$ on N. For every Lebesgue-null set $A \subset N$,

$$\begin{aligned}\pi(A) &= (\pi \otimes \Gamma_\Sigma)(Z^+ \in A) \\ &= \int_N \left(\int_A \varphi_N(z - y + \nabla_N K(y))\, dz \right) \pi(dy) = 0.\end{aligned}$$

This proves absolute continuity. If $N = \{0\}$, the normal chain has one state and all formulas hold with $\nabla_N K = 0$. □

Only the marginal distribution of $P_N \eta$ enters the normal chain; the two orthogonal noise projections may be correlated.

C.3.3 Conditioning a stationary distribution on its orbit

A stationary distribution may involve several translation orbits, while the normal position process is defined within one orbit. Because recentering preserves each orbit, disintegration yields stationary orbit components to which the fixed-step escape lemma can be applied.

Proposition C.9 (Stationary analysis of a fixed shape)**.** *Let μ be stationary for P and satisfy $\mu(\mathcal{C} \times \overline{B}(0, L)) = 1$. There exist a standard Borel space Θ, a probability ν on Θ, a Borel map $\theta \mapsto H_\theta \in \mathcal{C}$, and a measurable family of probabilities μ_θ on E such that, for every $\theta \in \Theta$, H_θ is globally definable and*

$$\mu_\theta P = \mu_\theta, \qquad \mu_\theta\{(J, v) : J = T_z H_\theta \text{ for some } z \in \mathbb{R}^d\} = 1.$$

For every bounded Borel function $\psi : E \to \mathbb{R}$,

$$\int_E \psi \, d\mu = \int_\Theta \left(\int_E \psi \, d\mu_\theta \right) d\nu(\theta).$$

In particular, the mean descent can be computed by this identity with $\psi = d$.

Recentering preserves a translation class. Conditioning on that class will therefore preserve stationarity as well.

Proof. Use Lemma C.7 and write $\rho(H, v) = \rho(H)$ on $\mathcal{C} \times \overline{B}(0, L)$. All fibers $\rho^{-1}(q)$ below are taken in this domain and viewed as Borel subsets of E. The standard Borel conditional distribution theorem (Durrett, 2019, Theorems 4.1.17–4.1.18) gives

$$\bar{\mu} = \rho_\# \mu, \qquad \mu = \int_Q \mu_q \, d\bar{\mu}(q), \qquad \mu_q(\rho^{-1}(q)) = 1 \quad \text{for } \bar{\mu}\text{-almost every } q.$$

Here $\rho_\# \mu$ is the pushforward probability, $(\rho_\# \mu)(A) = \mu(\rho^{-1}(A))$, and (μ_q) is a measurable family of conditional probabilities on E. The measure identity holds on every Borel set and against every bounded Borel test.

For bounded Borel b on Q, extend $b \circ \rho$ by 0 outside $\mathcal{C} \times \overline{B}(0, L)$. The kernel preserves each orbit, so on this set, for $\psi \in C(E)$,

$$P[(b \circ \rho)\psi] = (b \circ \rho) P\psi.$$

Stationarity of μ gives

$$\begin{aligned} &\int_Q b(q) \left[\int_E (P\psi - \psi) \, d\mu_q \right] d\bar{\mu}(q) \\ &\qquad = \int_E \big(P[(b \circ \rho)\psi] - (b \circ \rho)\psi \big) \, d\mu = 0. \end{aligned}$$

Since b is arbitrary, the inner integral vanishes almost everywhere. Let $\mathcal{A} \subset C(E)$ be countable and uniformly dense, and set

$$\Theta = \{q : \mu_q(\rho^{-1}(q)) = 1\} \cap \bigcap_{\psi \in \mathcal{A}} \left\{ q : \int_E (P\psi - \psi) \, d\mu_q = 0 \right\}.$$

The set $\{(q, H, v) : \rho(H) = q\}$ is Borel. Integration against the measurable family μ_q therefore makes the first condition Borel; the other conditions are Borel as well. Thus $\Theta \subset Q$ is standard Borel and $\bar{\mu}(\Theta) = 1$.

For $\theta \in \Theta$ and $\psi \in C(E)$, choose $\psi_i \in \mathcal{A}$ with $\|\psi_i - \psi\|_\infty \to 0$. Since P is a contraction in the supremum norm,

$$\left| \int_E (P\psi - \psi) \, d\mu_\theta \right| \le 2\|\psi_i - \psi\|_\infty + \left| \int_E (P\psi_i - \psi_i) \, d\mu_\theta \right| \longrightarrow 0.$$

Continuous tests determine probability measures on the compact space E, so $\mu_\theta P = \mu_\theta$ for every $\theta \in \Theta$.

Restrict $\bar{\mu}$ to Θ and call the resulting probability ν. Let H_θ be the Borel representative supplied by Lemma C.7. Each H_θ is globally definable, and

$$\rho^{-1}(\theta) = \{(J, v) \in E : J = T_z H_\theta \text{ for some } z \in \mathbb{R}^d\}.$$

The defining property of Θ gives probability 1 to this orbit. Finally, for every bounded Borel ψ,

$$\int_E \psi \, d\mu = \int_Q \left(\int_E \psi \, d\mu_q \right) d\bar{\mu}(q) = \int_\Theta \left(\int_E \psi \, d\mu_\theta \right) d\nu(\theta).$$

The function d is bounded and Borel by Proposition C.5, so it is an admissible test in this identity. □

C.4 FIXED-STEP ESCAPE

We prove Lemma 3.4: a global positive lower bound on Fréchet subgradient norms forces almost-sure escape for Gaussian subgradient iterations with step size 1. The bound gives linear decrease on spheres; a homogeneous limit at infinity identifies a region supporting outward motion, and the Gaussian density makes this region accessible from every bounded set. The resulting exclusion of stationary position distributions is the escape ingredient in the positive-descent proof of Section C.5. The hypothesis implies $L \geq a > 0$, by evaluating a gradient at a differentiability point. It allows empty Fréchet subdifferentials and points at which $0 \in \partial_C H$. The proof therefore cannot replace (18) by a lower bound on the particular Clarke selection V_n. Instead, it first extracts a direction of linear decrease from the Fréchet gap.

Lemma C.10 (A lower value on every sphere)**.** *Under* (18)*, every $x \in \mathbb{R}^d$ and $r > 0$ satisfy*

$$\min_{\|z-x\|=r} H(z) \leq H(x) - \frac{a}{2} r. \tag{56}$$

Proof. For $\varepsilon > 0$, minimize

$$H(z) + \frac{a}{2}\sqrt{\|z-x\|^2 + \varepsilon^2}$$

over the closed ball $\overline{B}(x, r)$. By the generalized Fermat rule (Rockafellar & Wets, 1998, Theorem 10.1), an interior minimizer would give

$$-\frac{a}{2}\frac{z-x}{\sqrt{\|z-x\|^2+\varepsilon^2}} \in \partial_F H(z).$$

The displayed vector has norm less than $a/2$, contrary to (18). A minimizer z_ε thus lies on the boundary. Comparison with x gives

$$H(z_\varepsilon) \leq H(x) + \frac{a}{2}\varepsilon - \frac{a}{2}\sqrt{r^2+\varepsilon^2}.$$

Take a convergent subsequence on the compact sphere and let ε decrease to 0. □

We next look at H from distances tending to infinity. The resulting homogeneous potential describes radial movement, while an outer limit of the original Clarke fields retains the possible drift directions. Lemma C.11 makes this construction precise. Retaining this field is necessary because convergence of potentials does not identify their Clarke subdifferentials pointwise.

Lemma C.11 (The homogeneous potential and its conservative field)**.** *There is a definable, globally L-Lipschitz, positively homogeneous function J, where positive homogeneity means $J(tu) = tJ(u)$ for $t \geq 0$, such that*

$$\frac{H(Ru) - H(0)}{R} \longrightarrow J(u) \quad \textit{locally uniformly as } R \longrightarrow \infty, \qquad \min_{\|u\|=1} J(u) \leq -\frac{a}{2}. \tag{57}$$

Define

$$\begin{aligned} A(u) = \{v : &\textit{there exist } R_j \to \infty,\ u_j \to u,\ v_j \to v \\ &\textit{with } v_j \in \partial_C H(R_j u_j)\}, \qquad \mathcal{V}(u) = \operatorname{conv} A(u). \end{aligned} \tag{58}$$

Then $\mathcal{V}$ is definable, has closed graph and nonempty compact convex values, and is bounded by L. It is conservative for J and satisfies

$$\mathcal{V}(tu) = \mathcal{V}(u), \qquad \langle v, u\rangle = J(u) \quad \textit{for } t > 0 \textit{ and } v \in \mathcal{V}(u). \tag{59}$$

Proof. Set $f_s(u) = s[H(u/s) - H(0)]$ for $s > 0$. For every u, w,

$$|f_s(u)| \leq L\|u\|, \qquad |f_s(u) - f_s(w)| \leq L\|u - w\|.$$

For each fixed u, the one-variable definable function $s \mapsto f_s(u)$ has a finite partition into intervals on each of which it is constant or strictly monotone (van den Dries & Miller, 1996, Section 4.1). It is therefore constant or monotone on some interval $(0, \varepsilon_u)$. The bound $|f_s(u)| \leq L\|u\|$ gives a

finite limit $J(u) = \lim_{s\downarrow 0} f_s(u)$. The second inequality passes to the limit, so J is L-Lipschitz. If $u_1, \dots, u_m$ form an ε-net in a compact set K, then

$$\sup_{u\in K} |f_s(u) - J(u)| \le 2L\varepsilon + \max_{1\le i\le m} |f_s(u_i) - J(u_i)|.$$

First let $s \downarrow 0$, then $\varepsilon \downarrow 0$. This proves local uniform convergence. Definability follows from

$$J(u) = z \quad \Longleftrightarrow \quad \begin{array}{l} \forall \varepsilon > 0\ \exists s_0 > 0\ \forall s \in (0, s_0), \\ |f_s(u) - z| < \varepsilon. \end{array}$$

For $t > 0$, homogeneity and the sphere bound of Lemma C.10 give

$$\begin{aligned} J(tu) &= \lim_{s\downarrow 0} f_s(tu) = t \lim_{s\downarrow 0} f_{s/t}(u) = tJ(u), \\ \min_{\|u\|=1} f_s(u) &= s\left[\min_{\|z\|=1/s} H(z) - H(0)\right] \le -\frac{a}{2}, \\ \min_{\|u\|=1} J(u) &= \lim_{s\downarrow 0} \min_{\|u\|=1} f_s(u) \le -\frac{a}{2}. \end{aligned}$$

The last limit uses uniform convergence on the sphere.

The Clarke values are nonempty and bounded by L. Choosing $v_j \in \partial_C H(ju)$ and taking a subsequence proves $A(u) \neq \varnothing$. The graph identity

$$\operatorname{graph} A = \{(u, v) : (u, 0, v) \in \overline{\{(w, s, z) : 0 < s \le 1,\ z \in \partial_C H(w/s)\}}\}$$

gives definability and closedness. Boundedness then gives compact values. To check convexification, take $u_j \to u$ and $w_j \to w$ with $w_j \in \mathcal{V}(u_j)$. Carathéodory's theorem (Rockafellar & Wets, 1998, Theorem 2.29) gives

$$w_j = \sum_{i=0}^{d} \theta_i^{(j)} v_i^{(j)}, \qquad v_i^{(j)} \in A(u_j),$$

$$\theta_i^{(j)} \ge 0, \qquad \sum_{i=0}^{d} \theta_i^{(j)} = 1.$$

After a subsequence, $\theta_i^{(j)} \to \theta_i$ and $v_i^{(j)} \to v_i \in A(u)$ for every i. Therefore $w = \sum_i \theta_i v_i \in \mathcal{V}(u)$. This proves closedness of the graph of $\mathcal{V}$ and, with the same bound, compactness of its values.

Apply Lemma A.7 to f_s, $0 < s \le 1$, with $f_0 = J$. This is a single jointly definable parameter family, and

$$\varnothing \neq \partial_C f_s(u) = \partial_C H(u/s) \subset \overline{B}(0, L).$$

Its positive-parameter Clarke fields admit the required variational stratifications. The joint potential limit follows from

$$|f_s(u') - J(u)| \le L\|u' - u\| + |f_s(u) - J(u)| \longrightarrow 0 \quad \text{as } (u', s) \to (u, 0).$$

The horizon term is 0 because

$$\varepsilon_j \to 0, \quad v_j \in \partial_C f_{s_j}(u_j) \quad \Longrightarrow \quad \|\varepsilon_j v_j\| \le L|\varepsilon_j| \longrightarrow 0.$$

Schechtman's Theorem 4.3 and Corollary 4.5 thus give conservativity of A and $\mathcal{V} = \operatorname{conv} A$ for J Schechtman (2025).

For $t > 0$, a sequence defining $v \in A(tu)$ also satisfies

$$v_j \in \partial_C H\big((tR_j)(u_j/t)\big), \qquad tR_j \to \infty, \quad u_j/t \to u.$$

Hence $A(tu) \subset A(u)$. The same argument with $1/t$ gives $A(tu) = A(u)$ and consequently $\mathcal{V}(tu) = \mathcal{V}(u)$. Fix $v \in \mathcal{V}(u)$. Along the ray $t \mapsto tu$, conservativity and homogeneity give

$$J(u) = \frac{d}{dt}[tJ(u)] = \frac{d}{dt} J(tu) = \langle v, u\rangle \quad \text{for almost every } t \in (1, 2).$$

The first and last expressions do not depend on t, proving (59) for this arbitrary v. □

For $\epsilon > 0$, write

$$\mathcal{V}_\epsilon(u) = \{w : \|u - z\| \le \epsilon,\ \|w - v\| \le \epsilon \text{ for some } v \in \mathcal{V}(z)\}.$$

On any compact annulus, (58) implies the uniform inclusion

$$\partial_C H(Ru) \subset \mathcal{V}_\epsilon(u) \quad \text{for all sufficiently large } R. \tag{60}$$

Indeed, a sequence violating the inclusion has convergent state and field subsequences, whose limit belongs to the graph of A. This is the approximation of the original drift used below.

Use the potential and field of Lemma C.11. Write $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$ for the unit sphere. For $u \in \mathbb{S}^{d-1}$, use $J(u)$ as the angular value. Along the limiting differential inclusion $\dot{x} \in -\mathcal{V}(x)$, a negative angular value forces outward radial movement. The next lemma also gives a strict angular margin at the next radius. That margin makes the conclusion stable under discretization and small cumulative noise.

The polygon recursion below comes directly from the fixed-step chain. At the start of a phase, let $r = \|Y_n\|$ and use the space and time scales

$$x_k = Y_{n+k}/r, \qquad t_k = k/r, \qquad v_k = V_{n+k}, \qquad e_k = \frac{1}{r}\sum_{i=1}^{k} \eta_{n+i}, \quad e_0 = 0.$$

Dividing (19) by r gives exactly

$$x_{k+1} - x_k = -\frac{v_k}{r} + \frac{\eta_{n+k+1}}{r} = -(t_{k+1} - t_k)v_k + e_{k+1} - e_k.$$

Thus e_k is the cumulative rescaled noise; its difference is the single-step perturbation. For large r, (60) places the actual drift in $\mathcal{V}_\epsilon(x_k)$ on the relevant annulus. The next lemma allows a general small mesh and controls the whole cumulative error, which is the quantity bounded by Lemma A.10 in the stochastic argument.

Lemma C.12 (A robust outward phase)**.** *There are $b < 0$, $\delta > 0$, $T > 0$, and $\epsilon \in (0, 1)$ with the following property. Consider an Euler polygon with vertices x_k at times $0 = t_0 < t_1 < \cdots$, with mesh at most ϵ, and increments*

$$x_{k+1} = x_k - (t_{k+1} - t_k)v_k + e_{k+1} - e_k, \qquad v_k \in \mathcal{V}_\epsilon(x_k). \tag{61}$$

Assume $e_0 = 0$, $\sup_k \|e_k\| \le \epsilon$, $\|x_0\| = 1$, and $J(x_0) \le b$. The mesh is defined through the first time at least T. Before a mesh time exceeding T by at most ϵ, the polygon reaches a first vertex with radius at least two. At every preceding vertex its radius is greater than $1/2$ and its angular value is less than $b + \delta/2$. At that vertex,

$$2 \le \|x_k\| \le 3, \qquad J(x_k/\|x_k\|) \le b - \delta/2. \tag{62}$$

It is enough to impose the field inclusion in (61) *before this vertex or the first violation of either preceding bound.*

Proof. Let x satisfy $\dot{x} = -v$, $v \in \mathcal{V}(x)$, on an interval where $r = \|x\| > 0$, and put $u = x/r$. Since $\mathcal{V}(x) = \mathcal{V}(u)$, the conservative chain rule and (59) give, almost everywhere,

$$\begin{aligned} \frac{d}{dt}J(x) &= \langle v, -v\rangle = -\|v\|^2, \\ \dot{r} &= \langle u, -v\rangle = -J(u), \\ r\frac{d}{dt}J(u) &= \frac{d}{dt}J(x) - \dot{r}J(u) = -\|v\|^2 + J(u)^2. \end{aligned}$$

Because $\langle v, u\rangle = J(u)$ and $\|u\| = 1$, this is

$$\dot{r} = -J(u), \qquad \frac{d}{dt}J(u) = -\frac{\|v - J(u)u\|^2}{r} \le 0. \tag{63}$$

The set

$$\{u \in \mathbb{S}^{d-1} : J(u)u \in \mathcal{V}(u)\}$$

is definable and has finitely many definably path connected components (van den Dries, 1998, Chapter 6, §3). Along a piecewise C^1 path $u(t)$ in one component,

$$\frac{d}{dt}J(u(t)) = \langle J(u(t))u(t), \dot{u}(t)\rangle = \frac{J(u(t))}{2}\frac{d}{dt}\|u(t)\|^2 = 0 \quad \text{almost everywhere.}$$

Thus its image under J is finite. Choose b strictly between $\min_{\|u\|=1} J(u)$ and 0 and outside this finite set. Compactness and the closed graph of $\mathcal{V}$ give $\delta_0 > 0$ and $\mu > 0$ such that $[b-2\delta_0, b+2\delta_0] \subset (-\infty, 0)$ and

$$\|v - J(u)u\| \geq \mu \quad \text{if } |J(u) - b| \leq 2\delta_0, \quad v \in \mathcal{V}(u).$$

If this band has no points on the sphere, the assertion is vacuous and any $\mu > 0$ can be used. Set

$$0 < \delta \leq \min\{\delta_0, \mu^2/(8L)\}, \qquad T = 2/(-b).$$

For an exact solution with $r(0) = 1$ and $J(u(0)) \leq b$,

$$J(u(t)) \leq b < 0, \qquad -b \leq \dot{r}(t) \leq L, \qquad 1 - bt \leq r(t) \leq 1 + Lt.$$

Its first radius-two hit τ therefore satisfies $1/L \leq \tau \leq 1/(-b) < T$. Once J reaches $b - \delta$, it stays below that value. Otherwise $b - \delta < J(u(t)) \leq b$ for $0 \leq t \leq \tau$, and the band estimate gives

$$J(u(\tau)) - J(u(0)) \leq -\int_0^\tau \frac{\mu^2}{r(t)}\,dt \leq -\frac{\mu^2\tau}{2} \leq -\frac{\mu^2}{2L} \leq -4\delta.$$

This contradicts $J(u(\tau)) > b - \delta$. Thus every exact solution reaches radius two with $J \leq b - \delta$.

Suppose no uniform tolerance exists. Take $\epsilon_j \downarrow 0$ and violating polygons. Stop them at the first hit, guard failure or timeout, and extract $\sigma_j \to \sigma \leq T$. Interpolate the cumulative errors linearly and freeze both paths after stopping. Ascoli gives a subsequence for which

$$z_j = x_j - e_j, \qquad \mathrm{Lip}(z_j) \leq L + 1, \qquad x_j, z_j \longrightarrow x \quad \text{uniformly.}$$

The strict initial guard margins and uniform convergence imply $\sigma > 0$. For the preceding vertex $\bar{x}_j$ and every fixed $t < \sigma$, for all large j,

$$\|z_j - \bar{x}_j\|_{[0,t]} \leq (L+4)\epsilon_j, \qquad \dot{z}_j \in -\mathcal{V}_{(L+5)\epsilon_j}(z_j).$$

Faure–Roth (Faure & Roth, 2010, Corollary 4.10) applies to $-\mathcal{V}$: it is bounded, with closed graph and nonempty compact convex values. Enlarge the tolerance to $(L+6)\epsilon_j$ for their open graph neighborhoods. The corollary gives $\dot{x} \in -\mathcal{V}(x)$ before σ. The guards keep the limit away from 0 there, so the exact-solution bounds apply.

A violating subsequence and uniform convergence would give one of

$$r(\sigma) \leq 1/2, \qquad J(u(\sigma)) \geq b + \delta/2,$$
$$\sigma = T,\ r(T) \leq 2, \qquad r(\sigma) = 2,\ J(u(\sigma)) \geq b - \delta/2.$$

These contradict $r \geq 1$, $J(u) \leq b$, $r(T) \geq 1 - bT = 3$, and the strict angular estimate at the first radius-two hit, respectively. The overshoot bound $(L+3)\epsilon_j < 1$ also excludes a terminal radius above three. □

The deterministic part is now complete. At physical radius r, spatial rescaling by r turns one step of (19) into an Euler step of duration $1/r$. The cumulative centered noise over order r steps has size $o(r)$ with high probability. We use this observation first to continue outward phases, and then to rule out recurrent visits to a bounded set.

Proof of Lemma 3.4. Use the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and filtration $(\mathcal{F}_n)$ carrying the auxiliary chain Y_n, its adapted selections, and its noise sequence. All path expectations and conditional expectations in this proof use $\mathbb{P}$. Let $M_2 = \mathbb{E}\|\eta_1\|^2 = \operatorname{tr}\Sigma$. At a finite integer stopping time n, future independence follows by conditioning on $\{n = \ell\}$, $\ell = 0, 1, \ldots$. In particular, for every bounded Borel test ϕ and every fixed k,

$$\mathbb{E}[\phi(\eta_{n+1}, \ldots, \eta_{n+k}) \mid \mathcal{F}_n] = \int_{(\mathbb{R}^d)^k} \phi(z_1, \ldots, z_k)\,\Gamma_\Sigma^{\otimes k}(d(z_1, \ldots, z_k)).$$

The right side is expectation on the k-noise product space and is constant as a function of the observed history. If $r = \|Y_n\| \geq 1$, put $m = \lceil Tr \rceil$. Both are $\mathcal{F}_n$-measurable. Centering and independence give

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m} \eta_{n+i}\right\|^2 \middle| \mathcal{F}_n\right] = mM_2.$$

The conditional maximal inequality in Lemma A.10 therefore yields

$$\begin{aligned}\mathbb{P}\left(\max_{0\leq k\leq m}\left\|r^{-1}\sum_{i=1}^{k}\eta_{n+i}\right\| > \epsilon \middle| \mathcal{F}_n\right) \\ \leq \frac{4mM_2}{\epsilon^2 r^2} \leq \frac{4M_2(Tr+1)}{\epsilon^2 r^2} \leq \frac{C}{r}, \qquad C = \frac{4M_2(T+1)}{\epsilon^2}.\end{aligned} \tag{64}$$

The inequality is applied conditionally for the known horizon m and radius r.

Choose R_0 large enough that $1/R_0 \leq \epsilon$, (60) holds on $1/4 \leq \|u\| \leq 4$ for every scale $R \geq R_0$, and $2C/R_0 \leq 1/2$. Start a phase at any state y with radius $r \geq R_0$ and $J(y/r) \leq b$. The rescaled vertices $x_k = Y_{n+k}/r$ obey (61), with mesh $1/r$ and cumulative perturbation $e_k = r^{-1}\sum_{i=1}^{k}\eta_{n+i}$. Until the stopping conditions of Lemma C.12, the drift inclusion follows from (60).

Let B be the event that the actual path reaches its first vertex of radius at least $2r$ within m steps, before a guard violation, and that this endpoint satisfies the scaled version of (62). The end of the attempt is an integer stopping time, and B is measurable at that time. Lemma C.12 and (64) give

$$\left\{\max_{k\leq m}\|e_k\| \leq \epsilon\right\} \subset B,$$
$$\mathbb{P}(B^c \mid \mathcal{F}_n) \leq \mathbb{P}\left(\max_{k\leq m}\|e_k\| > \epsilon \mid \mathcal{F}_n\right) \leq \frac{C}{r}.$$

Restart at the actual successful endpoint. Its radius is in $[2r, 3r]$ and its angular value is at most b. Conditioning uses the actual endpoint sigma-field, rather than the sufficient small-noise event over the full future horizon.

Write τ_j for the start of phase j, with $\tau_0 = n$, and B_j for its success event. Put $G_0 = \Omega$ and $G_j = \bigcap_{i<j} B_i$. After a failed phase, keep all subsequent start times fixed at the failure time and declare their success events empty. Thus the τ_j are finite stopping times and $G_j \in \mathcal{F}_{\tau_j}$. Set $r_j = \|Y_{\tau_j}\|$ on G_j and $r_j = 1$ otherwise. On G_j,

$$r_j \geq 2^j r, \qquad \mathbb{P}(B_j^c \mid \mathcal{F}_{\tau_j}) \leq \frac{C}{r_j}.$$

The conditional probability of first failure in phase j is bounded by

$$\begin{aligned}\mathbb{P}(G_j \cap B_j^c \mid \mathcal{F}_n) &= \mathbb{E}\big[\mathbf{1}_{G_j}\mathbb{P}(B_j^c \mid \mathcal{F}_{\tau_j})\big|\mathcal{F}_n\big] \\ &\leq \mathbb{E}\left[\mathbf{1}_{G_j}\frac{C}{r_j}\middle|\mathcal{F}_n\right] \leq \frac{C}{2^j r}.\end{aligned}$$

The first-failure events are disjoint, so

$$\begin{aligned}\mathbb{P}\left(\bigcup_{j\geq 0}(G_j \cap B_j^c)\middle|\mathcal{F}_n\right) \leq \sum_{j=0}^{\infty}\frac{C}{2^j r} = \frac{2C}{r} \leq \frac{1}{2}, \\ \mathbb{P}\left(\bigcap_{j\geq 0} B_j\middle|\mathcal{F}_n\right) \geq 1 - \frac{2C}{r} \geq \frac{1}{2}.\end{aligned} \tag{65}$$

On $\bigcap_j B_j$, each phase has positive finite duration, and

$$\tau_j \geq n + j \longrightarrow \infty, \qquad \inf_{\tau_j \leq k < \tau_{j+1}} \|Y_k\| \geq \frac{r_j}{2} \geq 2^{j-1} r \longrightarrow \infty.$$

Thus all subsequent radii stay above $r/2$ and tend to infinity. No independence of successive phases has been assumed.

It remains to reach those states from an arbitrary bounded region. Fix an integer $M \geq 1$. Choose u_* on the unit sphere with $J(u_*) < b$. A sufficiently small open ball W around a sufficiently large multiple of u_* has

$$\|z\| \geq R_0, \qquad \|z\| > 2M, \qquad J(z/\|z\|) < b \quad \text{for every } z \in W.$$

Let φ_Σ be the Gaussian density and let $|W|$ denote the volume of W. At any deterministic time n with $\|Y_n\| \leq M$,

$$\|-Y_n + V_n\| \leq M + L, \qquad W - Y_n + V_n \subset \overline{W} + \overline{B}(0, M + L).$$

Continuity and strict positivity of the density on the latter compact set give

$$q_M := |W| \min_{z \in \overline{W} + \overline{B}(0,M+L)} \varphi_\Sigma(z) > 0.$$

Consequently,

$$\mathbb{P}(Y_{n+1} \in W \mid \mathcal{F}_n) = \int_{W - Y_n + V_n} \varphi_\Sigma(z)\, dz \geq q_M \qquad \text{whenever } \|Y_n\| \leq M. \tag{66}$$

Let $\mathcal{E}_n$ be the event that all radii after time n exceed M and tend to infinity. On $\{Y_{n+1} \in W\}$, the starting radius is greater than $2M$, so (65) gives $\mathbb{P}(\mathcal{E}_n \mid \mathcal{F}_{n+1}) \geq 1/2$. On $\{\|Y_n\| \leq M\}$, the tower property yields

$$\begin{aligned} \mathbb{P}(\mathcal{E}_n \mid \mathcal{F}_n) &\geq \mathbb{E}\big[\mathbf{1}_{\{Y_{n+1} \in W\}} \mathbb{P}(\mathcal{E}_n \mid \mathcal{F}_{n+1}) \big| \mathcal{F}_n\big] \\ &\geq \tfrac{1}{2}\mathbb{P}(Y_{n+1} \in W \mid \mathcal{F}_n) \geq q_M/2 =: p_M. \end{aligned}$$

Define

$$A_M = \{\|Y_n\| \leq M \text{ infinitely often}\}, \qquad Z_n = \mathbb{E}[\mathbf{1}_{A_M} \mid \mathcal{F}_n].$$

The event A_M belongs to $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$. Lemma A.9 gives

$$Z_n \longrightarrow \mathbb{E}[\mathbf{1}_{A_M} \mid \mathcal{F}_\infty] = \mathbf{1}_{A_M} \quad \text{almost surely.}$$

Since $A_M \cap \mathcal{E}_n = \varnothing$, the preceding conditional bound gives

$$\|Y_n\| \leq M \quad \Longrightarrow \quad Z_n \leq 1 - \mathbb{P}(\mathcal{E}_n \mid \mathcal{F}_n) \leq 1 - p_M.$$

On A_M, this inequality holds infinitely often while $Z_n \to 1$, an impossibility. Hence $\mathbb{P}(A_M) = 0$ for every integer $M \geq 1$, and

$$\mathbb{P}(\|Y_n\| \to \infty) = \mathbb{P}\left(\bigcap_{M=1}^{\infty} A_M^c\right) = 1.$$

Finally, if a measurable time-homogeneous selection admitted an invariant probability π, start the chain with law π and independent future noise vectors. For this final argument, take

$$\begin{aligned} &\Omega_\pi = \mathbb{R}^d \times (\mathbb{R}^d)^{\mathbb{N}}, \qquad \mathcal{F}_\pi = \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d)^{\otimes \mathbb{N}}, \\ &\mathbb{P}_\pi = \pi \otimes \Gamma_\Sigma^{\otimes \mathbb{N}}, \end{aligned}$$

with Y_0 and the noises as coordinate projections. Invariance and bounded convergence would imply

$$\pi(\overline{B}(0, M)) = \mathbb{P}_\pi(\|Y_n\| \leq M) \longrightarrow 0 \quad \text{for every } M,$$
$$1 = \lim_{M \to \infty} \pi(\overline{B}(0, M)) = 0.$$

This contradiction proves the last assertion. □

The absence of a stationary law is the conclusion used in Lemma 3.5: a stationary difference quotient function orbit cannot carry a globally gapped normal component. The construction also identifies the two probabilistic ingredients. Centering and a second moment control accumulated perturbations within an outward phase. Gaussian full support supplies the one-step access in (66); a noise law supported in a small ball can assign probability 0 to that event.

C.5 Proof of positive expected descent

On each stationary translation orbit, the inherited subgradient bound and Gaussian escape force a nonzero linear component. Stationarity cancels the normal component's expected increment, leaving mean descent equal to the squared norm of the linear coefficient. Averaging over orbits gives the positive lower bound in Lemma 3.5.

Proof of Lemma 3.5. Proposition C.9 gives $\mu = \int_\Theta \mu_\theta \, d\nu(\theta)$, where ν is a probability on a parameter space Θ and each μ_θ is stationary on one translation orbit, the set of all recenterings of a fixed difference quotient function. Fix $\theta \in \Theta$ and write $H = H_\theta$. The measure μ_θ is stationary and assigns probability 1 to translations of H, so Lemma C.8 applies. Explicitly, Lemma 3.3 gives

$$S = \{p : T_p H = H\}, \quad N = S^\perp, \qquad H(p+z) = \langle c, p\rangle + K(z), \quad c \in S.$$

For a linear subspace A, write P_A for its orthogonal projection. The notation $\nabla_N K$, $\partial_F^N K$, and $\partial_C^N K$ refers to the gradient and subdifferentials computed within N. The stationary orbit law induces a stationary normal position Z for

$$Z^+ = Z - \nabla_N K(Z) + P_N \eta.$$

The word normal refers to the orthogonal complement of the translation symmetries S. The construction and its measurability are proved in Section C.3. Let π_θ denote this stationary distribution on N. Throughout the calculation on this fixed orbit, $\mathbb{E}_\theta$ means expectation on the probability space

$$\big(N \times \mathbb{R}^d,\, \mathcal{B}(N) \otimes \mathcal{B}(\mathbb{R}^d),\, \pi_\theta \otimes \Gamma_\Sigma\big), \qquad Z(z,e) = z, \quad \eta(z,e) = e.$$

Thus θ, H, S, N, c, and K are fixed, while Z and η are independent random variables. The updated Z^+ is the function of these coordinates displayed above. Expectations $\mathbb{E}_\eta$ involving only the noise use $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Gamma_\Sigma)$.

We next show that the linear part satisfies $\|c\| \geq a$. Lemma 3.2 and (17) give

$$\|c\|^2 + \|w\|^2 \geq a^2 \qquad (z \in N,\ w \in \partial_F^N K(z)). \tag{67}$$

Suppose $\|c\| < a$ and $\dim N > 0$. Then

$$\inf\{\|w\| : z \in N,\ w \in \partial_F^N K(z)\} \geq \sqrt{a^2 - \|c\|^2} > 0.$$

Let $D \subset N$ be the differentiability set of K. Rademacher's theorem gives that $N \setminus D$ has Lebesgue measure 0 in N, and the ordinary gradient at every point of D belongs to the Clarke subdifferential. Definable choice as used in Lemma C.7 supplies a Borel selection on the remaining set, so there is a Borel map w with

$$w(z) \in \partial_C^N K(z) \quad (z \in N), \qquad w(z) = \nabla_N K(z) \quad (z \in D).$$

The stationary normal position in Lemma C.8 satisfies

$$\pi_\theta(D) = 1, \qquad \mathcal{L}\big(Z - w(Z) + P_N \eta\big) = \mathcal{L}(Z^+) = \mathcal{L}(Z).$$

This is an invariant probability for a chain satisfying Lemma 3.4, a contradiction. If $N = \{0\}$, then $\partial_F^N K(0) = \{0\}$ and (67) directly gives

$$\|c\| \geq a \tag{68}$$

as well. Thus (68) holds on every stationary orbit under consideration.

It remains to compute the expected descent. By (55), the decrease on this orbit is

$$(T_Z H)(0) - (T_Z H)(\Delta) = \|c\|^2 - \langle c, P_S \eta\rangle + K(Z) - K(Z^+).$$

In the stationary normal law,

$$K(Z^+) \overset{d}{=} K(Z), \qquad |K(Z) - K(Z^+)| \leq L\|-\nabla_N K(Z) + P_N \eta\| \leq L(L + \|\eta\|) \in L^1.$$

To take expectations without assuming $\mathbb{E}_\theta|K(Z)| < \infty$, define $\chi_R(t) = \max\{-R, \min\{t, R\}\}$. Equal distributions and the 1-Lipschitz property of χ_R give

$$\mathbb{E}_\theta\big[\chi_R(K(Z)) - \chi_R(K(Z^+))\big] = 0,$$
$$\big|\chi_R(K(Z)) - \chi_R(K(Z^+))\big| \leq |K(Z) - K(Z^+)|.$$

Dominated convergence therefore yields

$$\mathbb{E}_\theta[K(Z) - K(Z^+)] = \lim_{R\to\infty} \mathbb{E}_\theta\big[\chi_R(K(Z)) - \chi_R(K(Z^+))\big] = 0.$$

The three terms in (55) now give

$$\begin{aligned}\int_E d\, d\mu_\theta &= \|c\|^2 - \langle c, P_S \mathbb{E}_\eta \eta\rangle + \mathbb{E}_\theta[K(Z) - K(Z^+)] \\ &= \|c\|^2 \geq a^2.\end{aligned}$$

Finally, $|d(H, v)| \leq L(L + \mathbb{E}_\eta\|\eta\|)$ uniformly on E, so

$$\int_E d\, d\mu = \int_\Theta \left(\int_E d\, d\mu_\theta\right) d\nu(\theta) \geq a^2,$$

which proves (21). □

C.6 Convergence windows and stopped averages

Stopped occupation averages provide vanishing mean descent and approximate stationarity for the same sequence of probability measures. We construct them from finite windows of the original iteration, selecting the windows using past information and stopping at the first exit from a small ball. The two conclusions pass to one weak limit in the proof of Lemma 3.6.

Throughout this section, the original random variables are defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$, and $\mathbb{E}Z = \int_\Omega Z(\omega)\, \mathbb{P}(d\omega)$ for integrable Z. We display conditioning explicitly: $\mathcal{F}_n$ records the available history, while a past event $B_j \in \mathcal{F}_{n_j}$ selects the windows to be averaged.

The observation windows must contain many steps to make their endpoint terms negligible, while remaining short relative to their starting times so that their step lengths are nearly constant. Their spatial radius must also be small relative to their total step length. The following choice meets all three requirements.

Lemma C.13. *Assume $\mathbb{P}(X_n \to 0) > 0$. For each integer $j \geq 2$, there are increasing deterministic times $n_j \geq j^3$ and events $B_j \in \mathcal{F}_{n_j}$ with $\mathbb{P}(B_j) > 0$ such that, on B_j,*

$$\|X_{n_j}\| < j^{-2}, \qquad \mathbb{P}\big(\|X_n\| < j^{-2} \text{ for all } n \geq n_j \mid \mathcal{F}_{n_j}\big) \geq 1 - \frac{1}{j}. \tag{69}$$

For the window length and fixed scale

$$m_j = \lceil n_j/j \rceil, \qquad s_j = \gamma_{n_j}, \tag{70}$$

one has

$$m_j \to \infty, \qquad \frac{m_j}{n_j} \to 0, \qquad \frac{j^{-2}}{m_j s_j} \to 0, \qquad \max_{n_j \leq n \leq n_j + m_j} \left|\frac{\gamma_n}{s_j} - 1\right| \to 0. \tag{71}$$

Proof. Fix j. The events $\bigcap_{n\geq N}\{\|X_n\| < j^{-2}\}$ increase with N, and their union contains $\{X_n \to 0\}$. Hence, for some deterministic N, the event

$$A = \bigcap_{n\geq N} \{\|X_n\| < j^{-2}\}$$

has positive probability. Since $A \in \sigma(\bigcup_n \mathcal{F}_n)$, upward convergence of conditional expectations (Durrett, 2019, Theorems 4.6.8–4.6.9) gives

$$\mathbb{P}(A \mid \mathcal{F}_n) \longrightarrow \mathbf{1}_A \quad \text{almost surely.}$$

On A, both $\|X_n\| < j^{-2}$ and $\mathbb{P}(A \mid \mathcal{F}_n) > 1 - 1/j$ therefore hold for all sufficiently large n. Bounded convergence yields

$$\mathbb{P}\big(A \cap \{\|X_n\| < j^{-2},\ \mathbb{P}(A \mid \mathcal{F}_n) > 1 - 1/j\}\big) \longrightarrow \mathbb{P}(A) > 0.$$

Choose an integer $n_j \geq \max\{j^3, N, n_0\}$, larger than the preceding chosen time, for which

$$B_j = \{\|X_{n_j}\| < j^{-2},\ \mathbb{P}(A \mid \mathcal{F}_{n_j}) > 1 - 1/j\}$$

has positive probability. This event belongs to $\mathcal{F}_{n_j}$. Moreover, A implies remaining in the ball from time n_j onward, which proves (69).

The deterministic estimates are

$$m_j \geq j^2 \longrightarrow \infty, \qquad \frac{m_j}{n_j} \leq \frac{1}{j} + \frac{1}{n_j} \longrightarrow 0,$$

$$n_j s_j = \kappa n_j (n_j + n_0)^{-\alpha} \geq \kappa 2^{-\alpha} n_j^{1-\alpha} \geq \kappa 2^{-\alpha}, \qquad \frac{j^{-2}}{m_j s_j} \leq \frac{1}{j n_j s_j} \leq \frac{2^\alpha}{\kappa j} \longrightarrow 0.$$

Finally, since the steps decrease,

$$\max_{n_j \leq n \leq n_j + m_j} \left| \frac{\gamma_n}{s_j} - 1 \right| = 1 - \left(1 + \frac{m_j}{n_j + n_0} \right)^{-\alpha} \leq \frac{\alpha m_j}{n_j + n_0} \longrightarrow 0.$$

□

The number of observed transitions grows, even though the original total step length $m_j s_j$ need not grow. For $\alpha = 1$, it is of order $1/j$. The smaller radius j^{-2} ensures that endpoint value differences remain negligible after division by $m_j s_j$.

For the selected event B_j, we use the ordinary conditional average

$$\mathbb{E}[Z \mid B_j] = \frac{1}{\mathbb{P}(B_j)} \int_{B_j} Z \, d\mathbb{P}.$$

This is defined whenever $\int_{B_j} |Z| \, d\mathbb{P} < \infty$. Within these selected windows, $\mathbb{E}[Z \mid \mathcal{F}_n, B_j]$ further conditions on the history $\mathcal{F}_n$: it is the conditional expectation given $\mathcal{F}_n$ on $(\Omega, \mathcal{F}, \mathbb{P}(\cdot \mid B_j))$. Its identities are understood almost surely on B_j. For $n \geq n_j$, the event B_j already belongs to $\mathcal{F}_n$. Independence of η_{n+1} from $\mathcal{F}_n$ therefore gives, for bounded Borel b,

$$\mathbb{E}[b(\eta_{n+1}) \mid \mathcal{F}_n, B_j] = \int_{\mathbb{R}^d} b(z) \, N(0, \Sigma)(dz).$$

Thus future noise retains its Gaussian law after a window is selected, even when the history up to the current step is given.

Stop recording at the first exit or at the end of the window:

$$\sigma_j = \min \left(\{ k \in \{1, \ldots, m_j\} : \|X_{n_j + k}\| \geq j^{-2} \} \cup \{ m_j \} \right). \tag{72}$$

On B_j, the recorded states have indices $n_j \leq n < n_j + \sigma_j$. The terminal state $X_{n_j + \sigma_j}$ may lie outside the ball. It is excluded from the occupation average, but the transition reaching it is included in both telescoping sums below; see Figure 5. Equation (69) implies

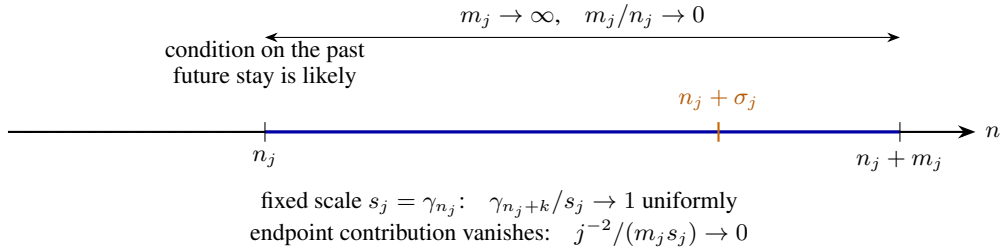

Figure 5: The stopped window used to obtain expected descent 0. Average the states from n_j to $n_j + \sigma_j - 1$, where σ_j is the first exit index or m_j. The final update may leave the ball. Horizontal lengths are schematic.

$$\mathbb{E}[\sigma_j \mid B_j] \geq m_j \, \mathbb{P}(\sigma_j = m_j \mid B_j) \geq (1 - 1/j) m_j \geq \frac{m_j}{2} \longrightarrow \infty.$$

The expected number of records within a selected window therefore diverges. The second-moment constant satisfies

$$M_2 = \mathbb{E}\|\eta_1\|^2 = \operatorname{tr}\Sigma, \qquad \mathbb{E}[\|\eta_{n_j+i}\|^2 \mid B_j] = M_2 \quad (i \geq 1).$$

The second equality again uses independence of future noise from B_j.

For all sufficiently large j, $\overline{B}(0, j^{-2}) \subset K_0$. Lemma C.2 then gives $V_n = g(X_n)$ at every recorded state on B_j, almost surely. Throughout a fixed window, we write its actual state explicitly as $(H_{X_n,s_j}, g(X_n))$, with n denoting the original iteration index. Define a probability ν_j on E through the following identity for every bounded Borel test function $\psi : E \to \mathbb{R}$:

$$\int_E \psi(H,v)\,\nu_j(dH,dv) = \frac{\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1} \psi(H_{X_n,s_j}, g(X_n)) \,\middle|\, B_j\right]}{\mathbb{E}[\sigma_j \mid B_j]}. \tag{73}$$

Taking $\psi = \mathbf{1}_D$ for Borel $D \subset E$ defines a nonnegative measure; countable additivity follows from monotone convergence. Its mass is

$$\nu_j(E) = \frac{\mathbb{E}[\sum_{n=n_j}^{n_j+\sigma_j-1} 1 \mid B_j]}{\mathbb{E}[\sigma_j \mid B_j]} = 1.$$

This is the ratio of the expected count in each set of states to the expected total count, both conditional on B_j. It is not an average of separately normalized sample paths. The same normalizing factor will apply to both telescoping identities.

On B_j, before termination, dividing the original update by the fixed scale gives

$$\frac{X_{n+1} - X_n}{s_j} = \frac{\gamma_n}{s_j}\big(-g(X_n) + \eta_{n+1}\big).$$

The difference quotient identities are exact:

$$T_{(X_{n+1}-X_n)/s_j} H_{X_n,s_j} = H_{X_{n+1},s_j}, \qquad \nabla H_{X_n,s_j}\left(\frac{X_{n+1}-X_n}{s_j}\right) = g(X_{n+1}) \quad \text{almost surely.}$$

The gradient on the right is defined from the global extension h, so the second identity remains valid if the next point leaves K_0. Thus the actual transition uses P_{γ_n/s_j}; only its first step has ratio exactly one.

The next lemma combines the transition telescope with its limiting consequence. The sum of successive changes of a test function reduces to two endpoint values. Their contribution vanishes after division by the expected number of recorded steps, and stability of the kernel replaces the actual ratios by one.

Lemma C.14. *For every $\psi \in C(E)$, the probabilities in* (73) *satisfy*

$$\int_E \big[(P\psi)(H,v) - \psi(H,v)\big]\,\nu_j(dH,dv) \longrightarrow 0.$$

Proof. Fix j sufficiently large. For $n_j \leq n < n_j + m_j$,

$$\{n < n_j + \sigma_j\} = \bigcap_{\ell=n_j+1}^{n} \{\|X_\ell\| < j^{-2}\} \in \mathcal{F}_n,$$

where the intersection is empty when $n = n_j$. The stopping indicator is therefore known at time n. Recall that $P = P_1$, and that, for a fixed state (H, v),

$$(P_\lambda\psi)(H,v) = \int_{\mathbb{R}^d} \psi\big(T_{\lambda(-v+z)}H, \nabla H(\lambda(-v+z))\big)\,N(0,\Sigma)(dz).$$

Thus $P_\lambda\psi$ averages the test function after one update of ratio λ. Given $\mathcal{F}_n$ within B_j, the current state $(H_{X_n,s_j}, g(X_n))$ and the stopping indicator are fixed, while η_{n+1} still has law $N(0,\Sigma)$. Before termination, the update and recentering identities above identify the next state with the two arguments

of ψ in this integral, with $H = H_{X_n,s_j}$, $v = g(X_n)$, and $\lambda = \gamma_n/s_j$. Integrating over the next noise and taking the known indicator outside the conditional expectation therefore gives

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}\psi(H_{X_{n+1},s_j}, g(X_{n+1}))\,\big|\,\mathcal{F}_n, B_j\big] \\
&\quad= \mathbf{1}_{\{n<n_j+\sigma_j\}}\int_{\mathbb{R}^d}\psi\Big(T_{\frac{\gamma_n}{s_j}(-g(X_n)+z)}H_{X_n,s_j}, \\
&\qquad\qquad \nabla H_{X_n,s_j}\big(\tfrac{\gamma_n}{s_j}(-g(X_n)+z)\big)\Big)\,N(0,\Sigma)(dz) \\
&\quad= \mathbf{1}_{\{n<n_j+\sigma_j\}}(P_{\gamma_n/s_j}\psi)(H_{X_n,s_j}, g(X_n)).
\end{aligned}$$

The first equality integrates the next noise, replacing η_{n+1} by the integration variable z; the second is exactly the definition of $P_\lambda\psi$ with the three substitutions above. On the event that recording has already stopped, both sides vanish. Only the next noise is averaged in this conditional identity; the right-hand side is still a function of the current random state. The tower property used here is

$$\mathbb{E}[\mathbb{E}[Z \mid \mathcal{F}_n, B_j] \mid B_j] = \mathbb{E}[Z \mid B_j],$$

for every Z integrable under $\mathbb{P}(\,\cdot \mid B_j)$. Both expectations use this same conditional probability measure: the inner expectation averages given the current history, and the outer expectation averages over that history within B_j. Applying this identity to the stopped next-state test value gives, for each fixed n,

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}(P_{\gamma_n/s_j}\psi)(H_{X_n,s_j}, g(X_n))\,\big|\,B_j\big] \\
&= \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}\psi(H_{X_{n+1},s_j}, g(X_{n+1}))\,\big|\,\mathcal{F}_n, B_j\big]\,\big|\,B_j\big] \\
&= \mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}\psi(H_{X_{n+1},s_j}, g(X_{n+1}))\,\big|\,B_j\big].
\end{aligned}$$

Subtract the conditional average of the stopped current test value from both sides and sum over the deterministic indices $n_j \le n < n_j + m_j$. The indicators restrict each sum to $n < n_j + \sigma_j$, giving

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1}\big[(P_{\gamma_n/s_j}\psi)(H_{X_n,s_j}, g(X_n)) - \psi(H_{X_n,s_j}, g(X_n))\big]\,\middle|\,B_j\right] \\
&= \mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1}\big[\psi(H_{X_{n+1},s_j}, g(X_{n+1})) - \psi(H_{X_n,s_j}, g(X_n))\big]\,\middle|\,B_j\right] \\
&= \mathbb{E}\Big[\psi(H_{X_{n_j+\sigma_j},s_j}, g(X_{n_j+\sigma_j})) - \psi(H_{X_{n_j},s_j}, g(X_{n_j}))\,\Big|\,B_j\Big].
\end{aligned}$$

All terms are integrable because ψ is bounded. Hence

$$\begin{aligned}
&\frac{\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1}\big[(P_{\gamma_n/s_j}\psi)(H_{X_n,s_j}, g(X_n)) - \psi(H_{X_n,s_j}, g(X_n))\big]\,\middle|\,B_j\right]}{\mathbb{E}[\sigma_j \mid B_j]} \\
&\quad= \frac{\mathbb{E}\Big[\psi(H_{X_{n_j+\sigma_j},s_j}, g(X_{n_j+\sigma_j})) - \psi(H_{X_{n_j},s_j}, g(X_{n_j}))\,\Big|\,B_j\Big]}{\mathbb{E}[\sigma_j \mid B_j]}.
\end{aligned} \tag{74}$$

This identity holds for every bounded Borel test function. Its right-hand side is bounded in absolute value by

$$\frac{2\|\psi\|_\infty}{\mathbb{E}[\sigma_j \mid B_j]} \le \frac{4\|\psi\|_\infty}{m_j} \longrightarrow 0.$$

Here $\|\psi\|_\infty = \sup_{(H,v)\in E}|\psi(H,v)|$. For the fixed continuous ψ in the statement, (73) and (74) imply

$$\begin{aligned}
\left|\int_E (P\psi - \psi)\,d\nu_j\right| &\le \frac{4\|\psi\|_\infty}{m_j} \\
&\quad + \max_{n_j\le n<n_j+m_j}\sup_{(H,v)\in E}|(P_{\gamma_n/s_j}\psi)(H,v) - (P\psi)(H,v)|.
\end{aligned}$$

The second term tends to zero by (71) and the joint continuity in Proposition C.5, applied on a compact interval of ratios containing one. This proves the conclusion. □

Lemma C.14 gives approximate stationarity of the occupation measures. The value telescope must now be controlled for these same measures. Unlike the bounded test function above, h need not be bounded globally. We therefore use its Lipschitz bound and account explicitly for the last step, whose endpoint may be outside the ball.

Lemma C.15. *The same probabilities ν_j satisfy*

$$\int_E d(H,v)\,\nu_j(dH,dv) \longrightarrow 0.$$

Proof. Fix j sufficiently large. For every pre-termination update on B_j, the definition of the actual difference quotient gives

$$H_{X_n,s_j}(0) - H_{X_n,s_j}\left(\frac{X_{n+1}-X_n}{s_j}\right) = \frac{h(X_n)-h(X_{n+1})}{s_j}.$$

Recall that $d = d_1$, where

$$d_\lambda(H,v) = \int_{\mathbb{R}^d} [H(0) - H(\lambda(-v+z))]\,N(0,\Sigma)(dz).$$

This is the expected decrease of H after one update of ratio λ from the fixed state (H,v). Before termination, $(X_{n+1}-X_n)/s_j = (\gamma_n/s_j)(-g(X_n)+\eta_{n+1})$. Given $\mathcal{F}_n$ within B_j, the current state is fixed and the next noise still has law $N(0,\Sigma)$. Substituting this displacement in the difference quotient identity above and integrating over the noise therefore gives $d_{\gamma_n/s_j}(H_{X_n,s_j}, g(X_n))$. The stopping indicator is $\mathcal{F}_n$-measurable, so it can be taken outside the conditional expectation:

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{n<n_j+\sigma_j\}}\frac{h(X_n)-h(X_{n+1})}{s_j}\,\middle|\,\mathcal{F}_n, B_j\right] \\
&\quad= \mathbf{1}_{\{n<n_j+\sigma_j\}}\int_{\mathbb{R}^d}\Big[H_{X_n,s_j}(0) \\
&\qquad\qquad - H_{X_n,s_j}\big(\tfrac{\gamma_n}{s_j}(-g(X_n)+z)\big)\Big]\,N(0,\Sigma)(dz) \\
&\quad= \mathbf{1}_{\{n<n_j+\sigma_j\}}d_{\gamma_n/s_j}(H_{X_n,s_j}, g(X_n)).
\end{aligned}$$

The first equality uses the exact difference quotient identity and the conditional Gaussian law. The second uses the definition of d_λ with $H = H_{X_n,s_j}$, $v = g(X_n)$, and $\lambda = \gamma_n/s_j$. When recording has already stopped, both sides are zero. The stopped terms are integrable under $\mathbb{P}(\,\cdot\mid B_j)$: on B_j, before termination,

$$\frac{|h(X_n)-h(X_{n+1})|}{s_j} \le L\frac{\gamma_n}{s_j}(L+\|\eta_{n+1}\|) \le L(L+\|\eta_{n+1}\|).$$

Apply the same tower property to the stopped value difference. The preceding integrability bound justifies both conditional averages, and for each fixed n it gives

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}d_{\gamma_n/s_j}(H_{X_n,s_j}, g(X_n))\,\big|\,B_j\big] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{n<n_j+\sigma_j\}}\frac{h(X_n)-h(X_{n+1})}{s_j}\,\middle|\,\mathcal{F}_n, B_j\right]\,\middle|\,B_j\right] \\
&= \frac{1}{s_j}\mathbb{E}\big[\mathbf{1}_{\{n<n_j+\sigma_j\}}[h(X_n)-h(X_{n+1})]\,\big|\,B_j\big]\,.
\end{aligned}$$

Summing over $n_j \le n < n_j+m_j$, then using the stopping indicators to write the finite stopped sums, gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1} d_{\gamma_n/s_j}(H_{X_n,s_j}, g(X_n))\,\middle|\,B_j\right] \\
&= \frac{1}{s_j}\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1}[h(X_n)-h(X_{n+1})]\,\middle|\,B_j\right] \\
&= \frac{1}{s_j}\mathbb{E}\big[h(X_{n_j})-h(X_{n_j+\sigma_j})\,\big|\,B_j\big]\,.
\end{aligned}$$

Consequently,

$$\frac{\mathbb{E}\left[\sum_{n=n_j}^{n_j+\sigma_j-1} d_{\gamma_n/s_j}(H_{X_n,s_j}, g(X_n)) \,\middle|\, B_j\right]}{\mathbb{E}[\sigma_j \mid B_j]} = \frac{\mathbb{E}[h(X_{n_j}) - h(X_{n_j+\sigma_j}) \mid B_j]}{s_j \mathbb{E}[\sigma_j \mid B_j]}. \tag{75}$$

On B_j, both the initial point and the point just before termination lie in the ball. Including the last update, and using decreasing steps, gives almost surely on this event

$$\begin{aligned}\|X_{n_j+\sigma_j} - X_{n_j}\| &\le \|X_{n_j+\sigma_j} - X_{n_j+\sigma_j-1}\| + \|X_{n_j+\sigma_j-1}\| + \|X_{n_j}\| \\ &\le \gamma_{n_j+\sigma_j-1}(L + \|\eta_{n_j+\sigma_j}\|) + 2j^{-2} \\ &\le s_j \left(L + \max_{1\le i\le m_j} \|\eta_{n_j+i}\|\right) + 2j^{-2}.\end{aligned}$$

Cauchy–Schwarz under $\mathbb{P}(\,\cdot \mid B_j)$ and the unchanged second moments give

$$\begin{aligned}\mathbb{E}\left[\max_{1\le i\le m_j} \|\eta_{n_j+i}\| \,\middle|\, B_j\right] &\le \left(\mathbb{E}\left[\sum_{i=1}^{m_j} \|\eta_{n_j+i}\|^2 \,\middle|\, B_j\right]\right)^{1/2} \\ &= \sqrt{m_j M_2}.\end{aligned}$$

Thus the Lipschitz bound and $\mathbb{E}[\sigma_j \mid B_j] \ge m_j/2$ yield

$$\begin{aligned}\left|\frac{\mathbb{E}[h(X_{n_j}) - h(X_{n_j+\sigma_j}) \mid B_j]}{s_j \mathbb{E}[\sigma_j \mid B_j]}\right| &\le \frac{L}{\mathbb{E}[\sigma_j \mid B_j]} \left(\frac{2j^{-2}}{s_j} + L + \sqrt{m_j M_2}\right) \\ &\le \frac{4Lj^{-2}}{m_j s_j} + \frac{2L^2}{m_j} + \frac{2L\sqrt{M_2}}{\sqrt{m_j}} \longrightarrow 0.\end{aligned} \tag{76}$$

The first term vanishes by the spatial-scale condition in (71); the remaining terms control the possible exit step using only the second moment of the noise.

Finally, uniformly over $(H, v) \in E$,

$$\begin{aligned}|d_\lambda(H, v) - d(H, v)| &\le L|\lambda - 1| \int_{\mathbb{R}^d} \|-v + z\| \, N(0, \Sigma)(dz) \\ &\le L(L + \sqrt{M_2})|\lambda - 1|.\end{aligned}$$

Applying this estimate to the recorded states in (73), and then using (75)–(76), gives

$$\begin{aligned}\left|\int_E d(H, v) \, \nu_j(dH, dv)\right| &\le L(L + \sqrt{M_2}) \max_{n_j \le n < n_j+m_j} \left|\frac{\gamma_n}{s_j} - 1\right| \\ &\quad + \frac{4Lj^{-2}}{m_j s_j} + \frac{2L^2}{m_j} + \frac{2L\sqrt{M_2}}{\sqrt{m_j}} \longrightarrow 0.\end{aligned}$$

□

The two lemmas concern the same occupation measures: one controls their change under the unit-step kernel, and the other controls their mean descent for that kernel. Compactness now provides a common limit with both properties.

C.7 Proof of zero expected descent

Compactness of the state space yields a common weak limit of the stopped occupation probabilities. Shrinking centers and scales place its support in $\mathcal{C} \times \overline{B}(0, L)$, and continuity of d and $P\psi$ transfers both conclusions of Lemmas C.15 and C.14: mean descent 0 and stationarity.

Proof of Lemma 3.6. Use the windows from Lemma C.13 and the probabilities ν_j in (73). The state space E is compact by Lemma C.3. Prohorov's theorem (Billingsley, 1999, Theorem 5.1) therefore gives a subsequence, still indexed by j, for which $\nu_j \Rightarrow \mu$.

Let $\operatorname{dist}(H, \mathcal{C}) = \inf_{J \in \mathcal{C}} d_0(H, J)$, where d_0 is the metric for local uniform convergence defined earlier. We first show

$$\sup_{\|x\| < j^{-2}} \operatorname{dist}(H_{x,s_j}, \mathcal{C}) \longrightarrow 0.$$

Otherwise, along a subsequence there would be centers $\|x_j\| < j^{-2}$ whose corresponding distances exceed a fixed $\epsilon > 0$. Compactness of $\mathcal{D}$ gives a further locally uniformly convergent subsequence of H_{x_j,s_j}. Since $x_j \to 0$ and $s_j = \gamma_{n_j} \to 0$, its limit belongs to $\mathcal{C}$, contradicting the distance bound. All states entering ν_j have such centers. Since the distance function is continuous and bounded,

$$\int_E \operatorname{dist}(H, \mathcal{C})\, \mu(dH, dv) = \lim_j \int_E \operatorname{dist}(H, \mathcal{C})\, \nu_j(dH, dv) = 0.$$

The integrand is nonnegative and $\mathcal{C}$ is closed, so $\mu(\mathcal{C} \times \overline{B}(0, L)) = 1$.

For $\psi \in C(E)$, Proposition C.5 gives $P\psi \in C(E)$. Lemma C.14 and weak convergence imply

$$\int_E (P\psi - \psi)\, d\mu = \lim_j \int_E (P\psi - \psi)\, d\nu_j = 0.$$

Continuous tests determine Borel probabilities on the compact metric space E, hence $\mu P = \mu$.

The function d is also bounded and continuous by Proposition C.5. For the same weak limit, Lemma C.15 gives

$$\int_E d(H, v)\, \mu(dH, dv) = \lim_j \int_E d(H, v)\, \nu_j(dH, dv) = 0.$$

This proves Lemma 3.6. □

In the proof of the avoidance theorem, the positive-descent result applies to every stationary probability on this same limiting state space. It therefore applies to the law constructed here, whose mean descent is zero; no uniqueness of the stationary law is needed.

Proof of Theorem 3.1. Suppose $\mathbb{P}(X_n \to 0) > 0$. Take the stationary law μ supplied by Lemma 3.6. Lemma 3.5 gives

$$0 = \int d(H, v)\, d\mu(H, v) \geq a^2 > 0,$$

a contradiction. Translating back to x_* proves the theorem. □

C.8 Finite sets and generic linear tilts

Proof of Corollary 3.7. For finite A, apply Theorem 3.1 to each point and take a finite union. For the generic statement, observe that $\{x : 0 \in \partial_C f_u(x)\} = \{x : u \in \partial_C f(x)\}$. The definable graph of $\partial_C f$ has dimension d, so its nonempty fibers over u are 0-dimensional, hence finite, outside a definable set of dimension less than d (Drusvyatskiy & Lewis, 2013, Theorem 3.5 and Corollary 4.4); the extension to definable functions is stated in that paper's introduction. Thus R_u is finite, and every $x \in R_u$ satisfies $0 \notin \partial_L f_u(x)$. The first assertion applied to f_u completes the proof. □

Definability alone does not ensure finiteness of the Clarke critical set: $f(x_1, x_2) = x_1^2$ has the critical line $\{0\} \times \mathbb{R}$. The generic tilt supplies the finiteness used above. The definition of sharply repulsive points is given in Appendix A.1. The corollary addresses sharply repulsive-point avoidance; general active-strict-saddle avoidance and generic convergence to local minima require further conclusions.

D EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

We supplement the noise comparison in Section 4 with individual trajectories and escape counts, using the same objective, initialization, step sizes, and covariance $\sigma^2 I_2$, where $\sigma = 0.15$. Uniform noise is sampled by area on the disk of radius 2σ; Gaussian, logistic, and Laplace coordinates are independent, with standard deviation σ and respective logistic and Laplace scales $\sigma\sqrt{3}/\pi$ and $\sigma/\sqrt{2}$.

Figure 6 shows the first eight trajectory indices from each of the four repeats, giving 32 paths per law. These indices were fixed independently of the outcomes. The individual paths complement the aggregate statistics in Figure 3 by showing the variation in escape times. The dashed line marks radius 0.05.

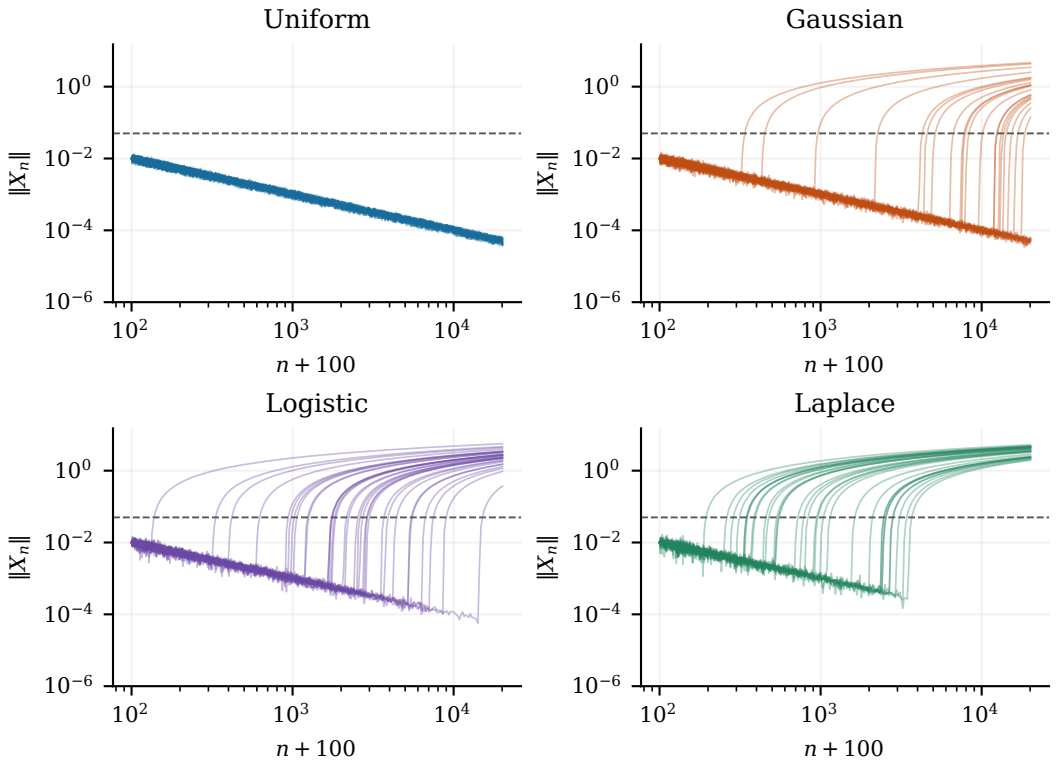

Figure 6: Individual trajectories under the four noise laws. Each panel shows 32 paths with indices fixed in advance, plotted at the recorded time indices. The dashed line marks radius 0.05.

Table 1 counts trajectories satisfying $\|X_n\| > 0.05$ at least once within 20,000 updates, using all updates of all 1,024 trajectories per law. These are finite-time escapes; an observed exit need not be permanent.

Table 1: Trajectories escaping radius 0.05 within 20,000 updates.

Noise law	Escapes / total trajectories
Uniform	0/1,024
Gaussian	766/1,024
Logistic	1,023/1,024
Laplace	1,024/1,024